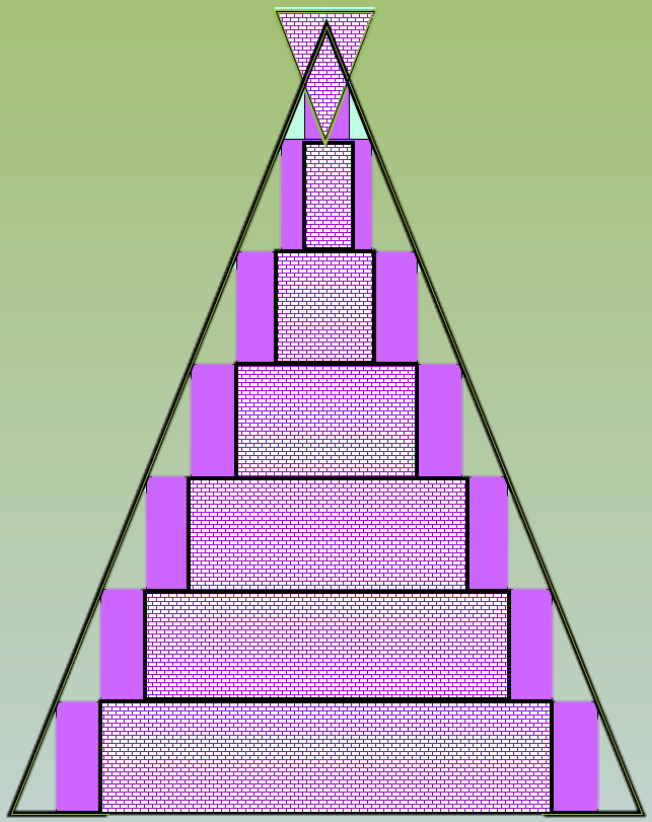
Diseño de la investigación, análisis y redacción de los resultados

Editores:

Dolores Frías-Navarro

Marcos Pascual-Soler



Edición 2ª, septiembre de 2021

Valencia

Diseño de la investigación, análisis y redacción de los resultados

Disseny de la investigació, anàlisi i redacció dels resultats

Research design, analysis and writing of results

Editores:

Dolores Frías-Navarro

Marcos Pascual-Soler

*Universidad de Valencia, España

** ESIC Business & Marketing School, España

septiembre de 2021

(2ª edición revisada y ampliada)

Edición: septiembre de 2021. https://doi.org/10.17605/osf.io/hetw2

España: Valencia

Portada y contraportada diseñada por Dolores Frías Navarro

Las fotografías pertenecen al albúm particular de Dolores Frías Navarro

El material se puede adquirir en:
"Copias y Revelados" (Palmero Ediciones). Calle: Menéndez Pelayo, 29 46010 Valencia.
Correo: trabajos@copiayrevelados.com

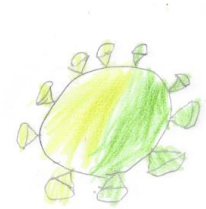
Laboratory: REsearch MEthods and design in applied psychology (**REME**)

https://www.uv.es/friasnav/

Comentarios y anotaciones en: M.Dolores.Frias@uv.es

A las alumnas y alumnos
que iniciaron su curso académico 2020/2021,
marcado para ser histórico
por la presencia del coronavirus y la COVID-19

PRESENTACIÓN 1. Septiembre, 2020



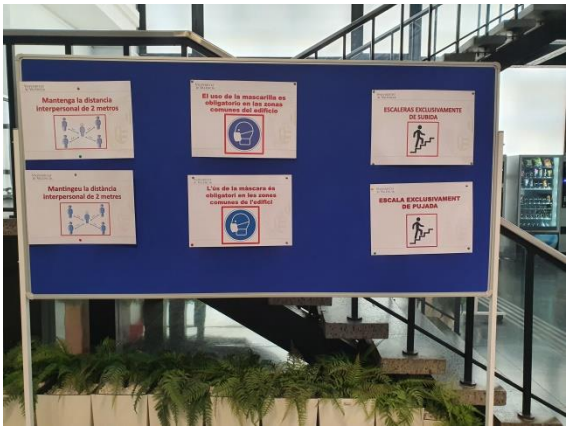
14 de marzo de 2020, el Gobierno español decreta el estado de alarma en todo el territorio nacional durante 15 días. El Congreso de los Diputados autorizó al Gobierno a prorrogar el estado de alarma en seis ocasiones. La medida del estado de alarma se extendió hasta el 21 de junio. Hubo ciertas concesiones a partir de finales de abril. Así, desde el 26 de abril se permitió la salida de menores de catorce años acompañados por un adulto durante una hora, pero solamente en las inmediaciones de sus domicilios. La medida se amplió a los mayores de dicha edad desde el 2 de mayo, solo para pasear o hacer deporte de manera individual. El 21 de mayo se hizo obligatorio el uso de mascarillas en espacios públicos a partir de los 6 años. Se trataba de una pandemia mundial provocada por un nuevo coronavirus (CoV) conocido como SARS-CoV-2 y la enfermedad que causa se conoce por la COVID-19 (acrónimo en inglés de "COronaVIrus Disease").

El resto de la historia ya la hemos vivido todos los españoles y el mundo en general. La aparición de la COVID-19 y la declaración de pandemia por la Organización Mundial de la Salud (OMS, 11 de marzo de 2020) fue muy rápido. Fue tan rápido que se tomaron medidas drásticas en muy pocos días. Tan drásticas como el estado de alarma nacional que supuso el confinamiento de la población en sus hogares. La libre circulación de los ciudadanos y ciudadanas se limitó a unas circunstancias muy concretas como adquirir alimentos y medicamentos, acudir a un centro médico o acudir al lugar de trabajo en unos puestos muy específicos y siguiendo unas rigurosas normas de seguridad. Y, por supuesto, los centros escolares y las universidades se cerraron. Después de unas semanas para afrontar el grave problema que supuso el coronavirus, comenzaron a tomarse medidas como pasar a una enseñanza virtual (*on-line*). Los profesores y las profesoras comenzaron a impartir docencia utilizando Internet, desde sus casas y con los medios personales que disponían. Se hizo todo lo posible por acabar el curso académico 2019/2020 en las mejores condiciones. Nadie podía prever al comienzo del curso, en septiembre de 2019, esta situación. Situación que ya es histórica.

Han pasado seis meses desde el estado de alarma, y la situación sanitaria sigue siendo muy crítica. Se ha pasado a lo que se denomina “la nueva normalidad”. La distancia social, el uso de mascarillas y de gel hidroalcohólico se han convertido en conductas obligatorias. Las aperturas del curso académico 2020/2021 en todas las universidades españolas se ha

caracterizado por planificar dichas medidas durante el acto institucional. Una imagen que nunca pudimos imaginar. Las directrices sobre las medidas de actuación y prevención para el profesorado y alumnado durante el primer cuatrimestre del curso siguen describiendo una situación de alarma sanitaria. Por supuesto, si se siente algún tipo de sintomatología

relacionado con la COVID-19 no acudir al centro e iniciar el aislamiento social y hacerse la prueba PCR. Y, la mascarilla siempre puesta. Distancia social de 2 metros y permanecer en la facultad el menor tiempo posible. La docencia, en gran parte de las facultades, será semi-presencial con clases presenciales y clases *on-line* sincrónicas. Nada que ver con el inicio tradicional del curso académico, con reencuentros, saludos y abrazos y muchas cosas que contar en los bancos de los pasillos y en las mesas de la cafetería. El curso académico 2020/2021 que ahora comienza no sabemos en qué condiciones acabará, ni siquiera sabemos cómo acabará el trimestre. Quizás la Universidad cierre sus puertas, quizás sea necesario pasar a la docencia totalmente *on-line* si las circunstancias sanitarias lo exigen. No lo sabemos. Ahora tenemos la experiencia anterior, pero la incertidumbre permanece debido a los datos sanitarios.



Ante esta situación de la COVID-19, decidimos ponernos a trabajar en la elaboración del material de este libro, "*Diseño de la investigación, análisis y redacción de los resultados*". Su objetivo es recoger en un formato libro los principales apuntes de la materia de diseños de investigación y permitir que el alumnado pueda consultar y estudiar los aspectos más relevantes para llevar a cabo todo el proceso del diseño de una investigación. Se trata de un material de docencia dirigido a la formación de alumnos y alumnas principalmente de

Psicología. El libro está depositado en un repositorio en abierto y, por lo tanto, el acceso y su descarga es gratuito. Además, decidimos que debía ser un material en constante actualización, añadiendo temas y reflexiones necesarias para la formación en los contenidos del área de la metodología de investigación. En esta primera edición se han elaborado los contenidos básicos del diseño de investigación.

La elaboración de esta primera edición del libro "*Diseño de la investigación, análisis y redacción de los resultados*" permite que el alumnado acceda a una explicación detallada de los conceptos y técnicas que se van a desarrollar en la materia de diseños de investigación. Este material, junto con las diapositivas de las clases y las lecturas de otros materiales recomedados por el profesorado, permitirá que puedan llevar a cabo un aprendizaje activo. Este tipo de aprendizaje implica estudiar de forma autónoma y acceder a las clases con cierto nivel de conocimiento, favoreciendo dinámicas de trabajo en clase dirigidas a potenciar la solución de problemas y avanzar en la formación. Con ello, se evita la dinámica de clase donde el alumnado es un receptor de información sincrónica, pasivo y solo puede comenzar su formación una vez ha recibido la clase.

El contenido del libro "*Diseño de la investigación, análisis y redacción de los resultados*" se ha planificado en cinco bloques o partes: 1) Elementos del proceso de diseño de la investigación, 2) Diseños de investigación, 3) Herramientas de trabajo de consulta del conocimiento previo (bases de datos) y manejo de programas estadísticos, 4) Lectura crítica de investigaciones y análisis de la presentación y redacción de resultados, 5) Ejemplos de investigaciones. Análisis de su estructura y lectura crítica.

La primera edición del material (septiembre de 2020) recoge el primer bloque (Elementos del proceso de diseño de la investigación) y del bloque dos (Diseños de investigación) se ha redactado el tema del diseño entre-grupos unifactorial univariado junto con las pruebas de contraste de hipótesis específicas. El resto de contenidos se irán preparando y formarán parte del libro en la siguiente actualización del material. Durante este curso académico, el alumnado recibirá la información necesaria para completar el temario a través de apuntes y materiales recomendados por el profesorado.

Desde el área de la metodología de investigación y la calidad de los resultados científicos, el comienzo del siglo XXI supuso reabrir de nuevo el debate sobre la replicación de los resultados, su reproducibilidad, el sesgo de publicación, los falsos positivos y los usos y abusos de las pruebas tradicionales de significación estadística junto con la necesidad de la educación y la re-educación estadística. Los elementos que rodean a la actual cultura de la investigación, especialmente la necesidad de publicar y publicar en revistas de impacto, ha conducido a un escenario que no favorece la búsqueda de la mejor evidencia (prueba) y,

en algunos casos, ha sesgado las actividades del científico o científica. Por ello, la competencia de la lectura crítica o activa se ha convertido en esencial para el investigador o investigadora y para el lector o lectora de literatura científica. Dicha lectura requiere aproximarse a los resultados científicos desde una perspectiva de análisis metodológico y con cierto grado de escepticismo que mantega activa la mirada crítica. La formación en los contenidos del área de la metodología de investigación ha sido siempre necesaria para el científico o científica que lleva a cabo estudios empíricos y para los lectores y lectoras de ese tipo de información. Y, actualmente, sigue siendo esencial, primero para poder seleccionar, de la gran cantidad de información que se publica, los artículos o informes que aportan resultados con las mejores pruebas o evidencias y, sobre todo, porque se ha demostrado que los investigadores e investigadoras hacen, en ocasiones, un uso abusivo y sesgado de sus grados de libertad fomentando el sesgo de publicación que tantos problemas ha creado a la Ciencia y a la interpretación de la realidad de los fenómenos. La revolución de la denominada Ciencia en Abierto ("Open Science"), junto con otras conductas relacionadas como el pre-registro de los protocolos de investigación, el depósito de los datos obtenidos en el estudio y sus códigos y el fomento y consideración de los trabajos de meta-investigación, podría ayudar al cambio en la cultura de la investigación, a la forma de hacer Ciencia y a la forma de comunicar los resultados. Sin embargo, para cambiar la conducta de los investigadores y las investigadoras es necesario, en primer lugar, cambiar esa cultura y el sistema de incentivos académicos dirigidos a la promoción académica y profesional. Y, en segundo lugar, fomentar desde las aulas (ya desde las primeras etapas de la formación escolar) la educación en el razonamiento científico, en el método científico y en la elaboración y uso de hipótesis científicas. Por supuesto, la comunidad científica debe aposter fuerte por la ciencia básica y la elaboración de teorías que expliquen y den sentido a los hallazgos.

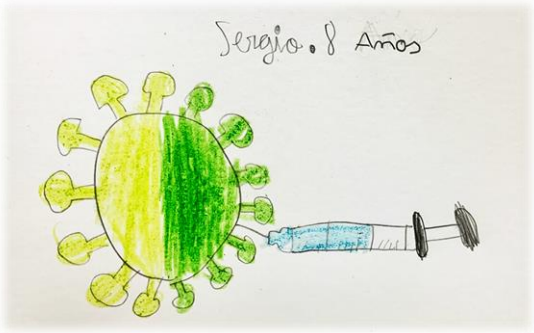
En definitiva, en el contexto actual de la Ciencia, la formación en los contenidos del área de la metodología de investigación (análisis de datos, psicometría y diseños de investigación) aporta unas herramientas fundamentales para acercarse a la elaboración de estudios y a la lectura de sus resultados con una actitud activa de rastreo y análisis crítico. Y, buscando el desarrollo de esas dos competencias se ha elaborado el material del libro "*Diseño de la investigación, análisis y redacción de los resultados*".

Dolores Frías Navarro

Marcos Pascual Soler

11 de septiembre de 2020

PRESENTACIÓN 2. EDICIÓN REVISADA Y AMPLIADA



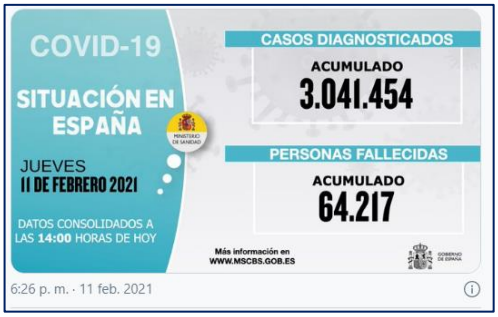
El 11 de marzo de 2020 se declaraba la pandemia mundial por parte de la Organización Mundial de la SALUD (OMS) debido al coronovirus y la enfermedad de la COVID-19. El 14 de marzo de 2021, un año después del Decreto del estado de alarma en España por el problema del coronavirus COVID-19 (supuso el confinamiento de la población española desde el domingo 15 de marzo a las 00.00h. de 2020 hasta el 21 de junio de 2020 con el final de la última prórroga y el inicio de la "nueva normalidad"), la situación no había mejorado de forma significativa en el mundo.

Como se puede observar en el dibujo de Sergio, la preocupación ahora era la vacuna contra el coronavirus. Cómo vacunar a toda la población de la mejor forma y de la forma más rápida, debatiendo sobre los grupos de riesgo que era necesario comenzar a vacunar después de vacunar a las personas mayores que viven en las residencias, el personal de acompañamiento y el personal sanitario. El 27 de diciembre de 2020 comenzó la vacunación contra la COVID-19 en toda la Unión Europea. El proceso de vacunación comenzó lento, muy lento, y las polémicas por conductas indebidas llevadas a cabo por personas que se vacunaban sin seguir el protocolo eran constantes junto con el debate sobre la eficacia de las diferentes vacunas que iban llegado al mercado ¿Cuántas dosis utilizar para aumentar su eficacia? ¿Una dosis es suficiente?, ¿Se deben dar dos dosis? Y tres vacunas en el mercado hasta el 10 de marzo de 2021: vacuna de BioNTech de Pfizer, la vacuna Moderna y la vacuna de Oxford, AstraZeneca. Y un año después, el 11 de marzo de 2021, se incorpora la vacuna Janssen del grupo Johnson & Johnson. Y sigue el desarrollo de otras vacunas y el proceso de aprobación de otras vacunas por la Agencia Europea del Medicamento (EMA) como las vacunas Novavax y CureVac.

Parecen lejanas las noticias sobre el primer paciente registrado en España con coronavirus COVID-19 (31 de enero de 2020). Se trataba de un paciente alemán ingresado en La Gomera que tuvo positivo en coronavirus. Su estado era leve y se contagió,

presuntamente, al contactar en Alemania con un infectado. Nueve días después se detectó otro caso de coronavirus Covid-19 en un turista británico en Palma de Mallorca. Y el 24 de febrero de 2020 el virus saltó a la península, detectando los primeros casos en la Comunidad de Madrid, Cataluña y la Comunidad Valenciana y de ahí a todas las comunidades españolas.

En enero de 2021 casi todas las comunidades españolas están en un semiconfinamiento, con cierre perimetral y unos horarios especiales para el comercio no alimentario. Se trataba de la denominada "segunda ola" de contagios. La situación es extrema para la hosteleria y para el comercio en general. La situación es tal desconcertante que, por ejemplo, el Tribunal Superior del País Vasco emite un auto judicial para abrir de nuevo la hosteleria. El 11 de febrero de 2021, el número de contagiados asciende a más de 107 millones de personas en el mundo. El Ministerio de Sanidad español informó ese mismo día que la COVID-19 ya había dejado 64.217 muertos y un total de 3.041.454 casos diagnosticados de coronavirus en España desde el inicio de la pandemia. El Ministerio de Sanidad también señala que hasta ese 11 de febrero de 2021, 943.278 personas estaban inmunizadas contra la COVID-19 al haber recibido las dos dosis de las vacunas de Pfizer-BioNtech y Moderna, mientras que las primeras dosis inoculadas de AstraZeneca superaban las 9.000. Los datos sobre las personas diagnosticadas y las personas fallecidas son escalofriantes.

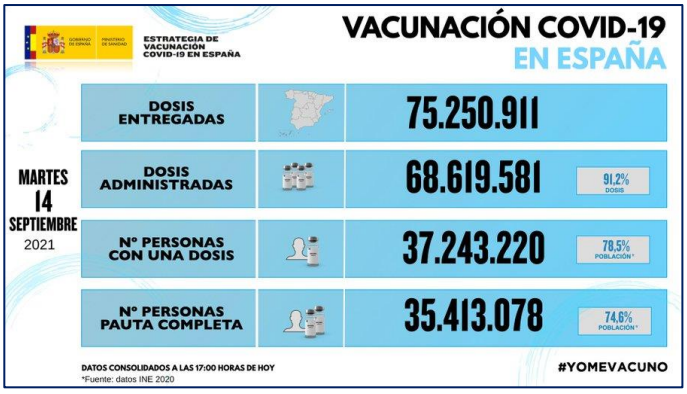


Y así, esperando que lleguen las vacunaciones, la Universidad de Valencia imparte sus clase on-line durante el mes de febrero de 2021. El segundo cuatrimestre del curso anterior 2019/2020 fue totalmente on-line, incluso los exámenes. El curso 2020/2021 comenzó de forma presencial en unos casos (especialmente en los estudios de máster) y semipresencial en otros, pero después de navidades la situación era tan grave que hubo que volver a la situación de clases on-line. El gabinete de prensa de la Universidad de Valencia así lo señalaba: "El Consejo de Gobierno ha ratificado la resolución de la rectora por la que el segundo cuatrimestre del curso se inicia con la mínima presencialidad posible, aplicando el acuerdo alcanzado por el Sistema Universitario Público Valenciano con la Generalitat, para contribuir a los objetivos del gobierno valenciano de reducción de la movilidad de las

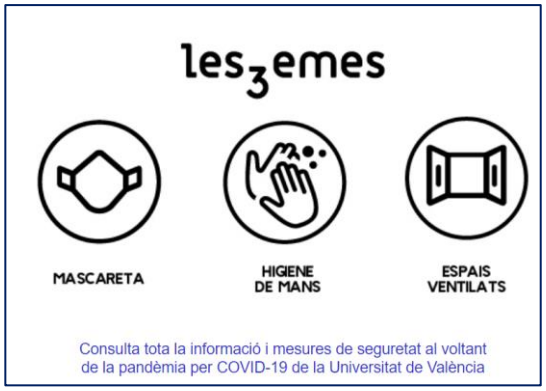
personas como medida para controlar la expansión de la COVID-19 en nuestro territorio". La información de la Universidad de Valencia concluye señalando que: "El modelo de mínima presencialidad estará vigente durante el mes de febrero, y se revisará en función de la evolución epidemiológica en la Comunitat y las medidas adoptadas por las autoridades, en coordinación con el conjunto de universidades que integran el Sistema Universitario Público Valenciano y las autoridades de la Generalitat competentes en sanidad y salud pública y en universidades". El 1 de marzo de 2021 se vuelve a la docencia semipresencial. Y durante estos meses nuestra Facultad pasa a denominarse Facultad de Psicología y Logopedia:

Durante este año, la población ha ido incoporando nuevos términos al lenguaje de la vida diaria que poco o nada se utilizaban anteriormente: pandemia, virus, distancia social, EPI, coronavirus, PCT, antígenos, Wuhan..., hasta el nombre de las marcas de las vacunas se han convertido en palabras cotidianas y se habla de ellas con naturalidad cuando se muestra preferencia por una u otra.

Y comienza el curso 2021 / 2022 con una enseñana totalmente presencial. Los datos sobre la COVID-19 en España el 14 de septiembre de 2021 son los siguientes:



De todas las medias adoptadas en el curso anterior solo quedan las mascarillas y la higiene de manos, la ventilación de las aulas, la distancia social, evitar aglomeraciones en las instalaciones del centro educativo y no acudir a la facultad si se presentan síntomas de COVID-19. El Real Decreto 463/2020, de 14 marzo de 2020, donde se declaraba el estado de alarma para la gestión de la situación de crisis sanitaria ocasionada por el COVID-19, ya ha sido superado por la nueva realidad en septiembre de 2021. En España, el 18 de septiembre de 2021 comienzan a poner la tercera dosis de la vacuna a las personas con factores de riesgo.



Y así comenzamos el curso académico 2021/2022, con mucho ánimo para afrontar todo aquello que la pandemia aún tiene que decir.

En este segunda edición revisada y ampliada se han incorporado varios capítulos nuevos. Nuestro agradecimiento a las profesoras y profesores que han aportado su trabajo. Los profesores José Perezgonzalez y Nicholas Vincent de la Universidad Massey (Nueva Zelanda) presentan un análisis paralelo del modelo frecuentista (modelo de la hipótesis nula tradicional, NHST) y el modelo bayesiano junto el desarrollo de los análisis con el programa JASP. Los profesores José Berríos-Riquelme (Universidad de Tarapacá, Chile), Manuel Martín-Fernández (Universidad Autónoma de Madrid, España), Viviana Vargas-Salmas Universidad de Valencia, España), Carla Vidal-Figueroa (Universidad de Concepción, Chile) y Cristóbal Pulido Oparraguirre (Universidad de Tarapacá, Chile) han desarrollado un informe de investigación para que los lectores y lectoras lleven a cabo una evaluación activa de su estructura y contenido, facilitando un ejemplo para trabajar la temática de elaboración de un informe de investigación tipo artículo.

Dolores Frías Navarro

Marcos Pascual Soler

18 de septiembre de 2021

INDICE

Capítulo I. Investigación científica en Psicología

Dolores Frías-Navarro

Universidad de Valencia

Índice

Citar el capítulo como:

Frías-Navarro, D. (2021). Investigación científica en Psicología. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

Este primer capítulo del libro, “Investigación científica en Psicología”, introduce a lectores y lectoras en conceptos y reflexiones que son básicos para entender los principios de la metodología de investigación y, especialmente, la materia de Diseños de Investigación dado el vínculo estrecho que mantiene con los elementos implicados en el proceso de diseño de un estudio y con las competencias que rodean a la lectura crítica o activa de los informes y artículos de investigación empírica. La valoración de los **aspectos éticos en el manejo de la investigación psicológica** es un tema importante que debe estar incluido en las guías docentes o programas de las materias que están relacionadas con el diseño de la investigación y el análisis de sus datos. La ética y la integridad científica son los pilares de la ejecución válida del método científico. Los resultados científicos se producen gracias al trabajo de investigadores e investigadoras, cuya conducta se rige por los principios éticos y de integridad de los códigos que subrayan la importancia y la necesidad de la conducta responsable del científico o científica (y profesionales en general).

Mejorar la práctica estadística y la educación estadística es algo más que una necesidad urgente para todo el alumnado universitario de cualquier disciplina científica y, sobre todo, para el de Ciencias Sociales y de la Salud. Gran parte del cuerpo de conocimiento de estas disciplinas, por ejemplo la Psicología o la Medicina, se alcanza con la aplicación de lo que genéricamente se llama el ‘método científico’ donde forman parte inexcusable la Estadística, el Diseño de la Investigación, la Psicometría y la Matemática.

La necesidad de información y de formación continua constituyen dos condiciones básicas de la práctica del profesional de la Psicología, la Medicina, la Psiquiatría, la Sociología, el Trabajo Social, la Economía, la Biología, el Derecho, la Veterinaría, la Farmacía, la Enfermería… La actualización constante de los conocimientos adquiridos tras la formación académica es un requisito imprescindible para realizar la labor profesional con el mayor grado de éxito y calidad, asegurando de este modo que no haya deterioro del saber o conocimiento y, como consecuencia, también del ejercicio profesional. La formación continua en la especialidad y la re-educación estadística son dos cuestiones básicas de la calidad del profesional.

Si se reflexiona sobre las características de un buen profesional, se podría llegar a la conclusión de que ser un buen profesional es cada vez más difícil dado el gran

número de tareas que implica y la constante actualización que requiere. La sociedad está en un momento de gran evolución del conocimiento científico que, además, avanza muy rápido. Todo ello exige estar actualizado en los avances que se producen en la especialidad o campo de conocimiento del profesional donde la elaboración y publicación de investigaciones aplicadas y/o empíricas supone de forma inexcusable un referente básico que hay que leer, consultar y valorar desde una perspectiva crítica o activa.

Actualmente, para ser un buen profesional, ya no basta con "tener una gran experiencia" o una gran antigüedad en su campo de trabajo. Ahora es necesario formarse de forma continua ('aprendizaje a lo largo de la vida') y conocer los avances que se producen (diariamente), aplicando criterios de valoración crítica o lectura activa de la calidad de las pruebas empíricas o hallazgos ('evidencia') que se consultan. La calidad de la evidencia se puede jerarquizar en función de la calidad metodológica del proceso de diseño de investigación.

Desde un modelo de Práctica Basada en la Evidencia que impera en la Ciencia del siglo XXI (la denominada 'Ciencia Basada en la Evidencia'), la autoridad de los años y de la supuesta experiencia del profesional es sustituida por la autoridad de la investigación científica basada en la mejor evidencia o en las mejores pruebas (investigación realizada con el diseño de investigación que aporte los resultados más válidos y, por lo tanto, más próximos a la realidad del fenómeno estudiado). Por este motivo, para obtener pruebas empíricas de calidad (resultados o evidencia) es necesario dominar de forma adecuada todos los elementos implicados en el proceso de diseño de investigación.

Los elementos o herramientas que forman parte del diseño de una investigación incluyen cuestiones como necesidad de conocimiento (problema de investigación), constructos, variables, tipos de hipótesis, sesgo, control, estadística, diseño, inferencia, estadísticos, error, aleatorio, distribuciones, análisis, contraste de hipótesis, validez, resultados, informe, redacción, conclusiones..., cuya presencia en el diseño está dirigida a dar solución al problema de investigación planteado o necesidad de conocimiento. Se podrían comparar dichos elementos, por ejemplo, con las herramientas que los cirujanos o las cirujanas utilizan para llevar cabo con éxito una operación (ver Imagen 1).

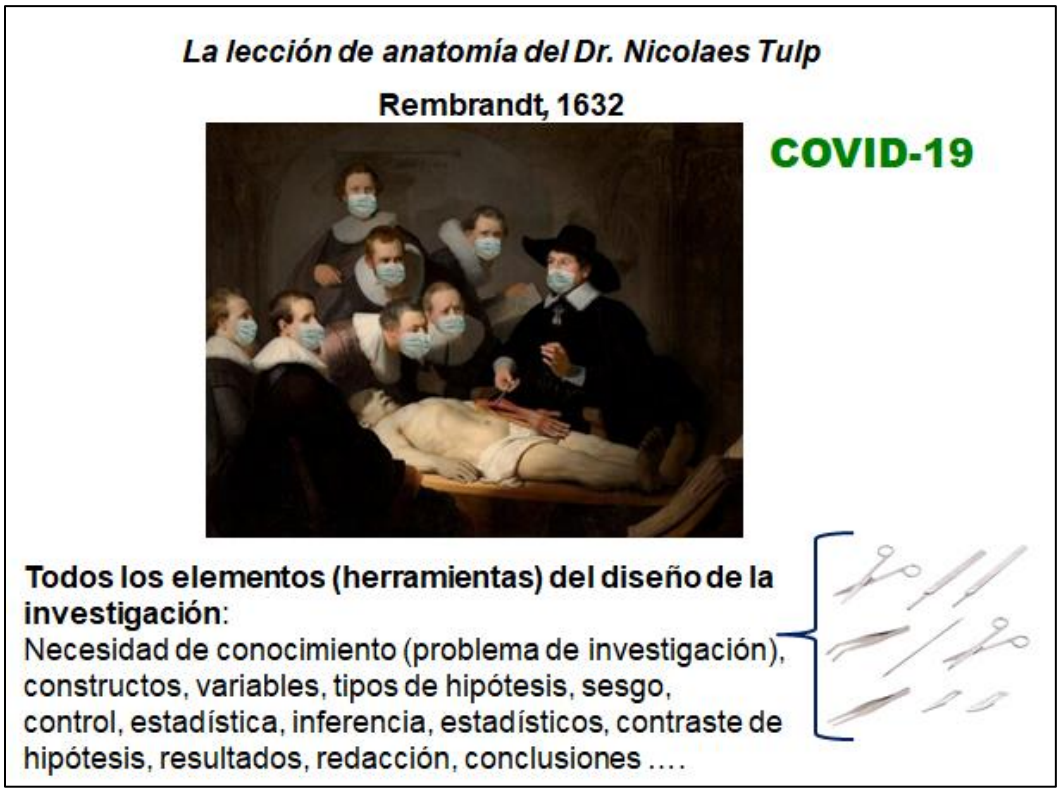


Imagen 1. Elementos necesarios para ejecutar el diseño del estudio. *Imagen disponible en el siguiente enlace:* https://ih1.redbubble.net/image.1179174666.1308/mp,504x498,matte,f8f8f8,t-pad,600x600,f8f8f8.jpg

El modelo de la Práctica Basada en la Evidencia no sólo se aplica para obtener pruebas válidas con la elaboración y ejecución de una investigación propia, pues dicho modelo de actuación también se aplica cuando el lector o lectora hace una lectura 'crítica' o activa de los hallazgos que otros investigadores o investigadoras han publicado. En ambas situaciones (como investigador o investigadora y como lector o lectora) es necesario dominar los contenidos del ámbito de la metodología de investigación. En otras palabras, valorar un resultado de un estudio sólo puede ser realizado si el lector o lectora chequea un conjunto de elementos relacionados con el proceso del diseño de investigación que podrían invalidar los resultados del estudio (amenazas a la validez de los resultados) o, por el contrario, podrían otorgar calidad a los hallazgos o pruebas (validez) y, de este modo, los resultados sí aportarían evidencia válida al campo de trabajo del que forma parte el estudio.

El psicólogo o la psicóloga (especialista o profesional en general) es el responsable de su formación y debe garantizar que está preparado profesionalmente para abordar la intervención o el estudio de las variables con la mayor calidad

disponible, evitando así errores en las decisiones e intervenciones clínicas o sociales. La valoración personal de las propiedades o calidad de los hallazgos (lectura crítica o activa de los resultados de la investigación) debe realizarla desde una perspectiva científica apoyada en el método científico y en la calidad de la evidencia o las pruebas encontradas (validez o calidad de los resultados o hallazgos).

Los resultados científicos que son publicados en las revistas científicas deben leerse de forma crítica o activa. Se ha demostrado que la falta de transparencia en el proceso de investigación, la elaboración de informes selectivos que destacan sólo lo que es publicable (son más publicables los resultados estadísticamente significativos) o es más interesante / llamativo del estudio, la elaboración de informes incompletos que ocultan el porqué de las decisiones adoptadas, la falta de protocolos y su adherencia, la falta de conocimientos metodológicos y de diseño de investigación y las malas prácticas de investigación son cuestiones que persisten en el mundo científico (Frías-Navarro y cols., 2020, 2021). Además, ha sido ampliamente estudiado el problema del sesgo de publicación en la literatura científica, donde se destaca que la mayoría de los estudios que se publican tienen resultados estadísticamente significativos (conocidos como “resultados positivos”) frente a la no publicación de los que son estadísticamente no significativos (conocidos como “resultados negativos” o resultados nulos). Ese sesgo de publicación provoca que la realidad que se lee en los informes y artículos estaría sesgada y es un problema ético importante. Ese problema ético distorsiona seriamente la evidencia y, por lo tanto, el conocimiento que tenemos sobre la efectividad de las intervenciones o sobre las relaciones entre las variables que están vinculadas en la manifiestación de ciertos trastornos o hechos sociales y, además, es un problema porque se desperdician fondos públicos. Esa situación se agrava con la cuestión demostrada de que se citan en mayor medida los resultados positivos (provoca el sesgo de publicación) y así la distorsión de la realidad queda aumentada (Duyx y cols., 2017).

Por ejemplo, desde el área de la Medicina, en el famoso artículo de Doug Altman de 1994, ya se alertaba de esta situación de crisis de la calidad de los resultados con un llamativo título que destacaba el escándalo de la mala investigación médica (“The scandal of poor medical research”). Posteriormente, en 2005 Ioannidis utiliza en su trabajo un título sorpredente que alerta de la calidad de los resultados científicos: por

qué la mayoría de las investigaciones publicadas son falsas (“Why most published research findings are false”).

En definitiva, el lector o lectora de literatura científica debe acercarse a los informes y artículos adoptando una mirada crítica, es decir, debe conocer suficientemente los elementos que forman parte del proceso de diseño de investigación y la técnica estadística para valorar los resultados científicos que lee y así, jerarquizar la calidad de la evidencia que describen. La lectura crítica es fundamental para que el profesional o la profesional que consume literatura científica pueda avanzar en su trabajo y asumir los principios de la práctica basada en la evidencia

Desde el ámbito del investigador o investigadora, como creador de evidencia científica, el problema se relaciona con el mal uso de la técnica estadística, la falta de conocimientos sobre los elementos implicados en el proceso de diseño de investigación, especialmente sobre los elementos que rodean a: la planificación del estudio: grupo de control inadecuado, sesgos de selección en la muestra, uso de tamaños de muestra pequeña, falta de planificación de la potencia estadística, uso de pruebas estadísticas inadecuadas o falta de control de ciertas variables extrañas así como las prácticas de investigación cuestionables. Como consecuencia de ese tipo de actuaciones se procede con la elaboración de un diseño de investigación deficiente que junto con la incomprensión del proceso de inferencia estadística y su alcance en la producción de conocimiento fiable y válido, provoca la ejecución de una investigación de mala calidad que repercutirá en la salud y en la sociedad en general y, tambien, en el gasto inútil de los fondos públicos (Frías-Navarro y cols., 2021).

El proceso del diseño de investigación debe basarse en una metodología sólida desde el comienzo, iniciando con la revisión del conocimiento previo con palabras clave y criterios objetivos de búsqueda y localización del conocimiento, hasta la presentación del informe de resultados, discusión y conclusiones. Por todo ello, la competencia del investigador o investigadora basada en su educación y re-educación en metodología es fundamental para producir conocimiento científico fiable y con evidencias de validez.

Crear grupos de trabajo con un especialista en metodología es fundamental para actuar desde el mismo momento en el que se plantea la necesidad de conocimiento

que va a investigar dicho equipo, hasta trabajar en el proceso de redacción y elaboración del informe o manuscrito final. No tiene sentido consultar a un metodólogo cuando los datos recogidos no cuadran, no se saben analizar o no se saben interpretar; el hecho fatal se ha producido y ya no se puede curar con prácticas de investigación no cuestionables. También conviene destacar que el o la especialista en metodología que forma parte del equipo de investigación debe involucrarse en el contenido de la temática que va a ser investigada, no es un técnico que desconoce los modelos teóricos que van a ser analizados en el estudio. Es importante que conozca bien el problema sustantivo de investigación para tomar decisiones adecuadas sobre la metodología requerida para abordar la pregunta de investigación y el diseño del estudio, operacionalizando adecuadamente los constructos o variables y reflexionando sobre la recogida, análisis e interpretación de los datos. Cada área de investigación tiene sus propias peculiaridades y contexto y no se pueden generalizar las actuaciones metodológicas. Y, también, todo el personal científico del equipo debería tener formación básica en metodología y evaluación crítica de la literatura para evitar falacias y rituales que pueden guiar consciente o inconscientemente a una conducta incorrecta o no ética.

No se debe olvidar que para lograr publicaciones con calidad científica también deben implicarse, por una parte, los editores y las editoras de las revistas que deben exigir calidad metodológica a los manuscritos enviados para su posible publicación, así como escoger de forma adecuada a los revisores y revisoras de los manuscritos requiriendo que tengan formación en metodologia junto a ciertos conocimientos teóricos del fenómeno objeto de evaluación.

Probablemente ya ha llegado el momento de valorar qué es ser revisor o revisora de manuscritos científicos, qué competencias debe tener, qué obligaciones tiene y qué recompensas debería recibir por su trabajo. Potenciar la formación en las competencias básicas de los revisores y revisoras es fundamental para proteger a los resultados científicos de la mala práctica; formación que puede ser recibida desde las propias universidades o desde las revistas. Por ejemplo, ‘Nature Masterclasses’ forma parte de la editorial Springer (publica la revista Nature desde 1869, entre otras) y tiene como principal objetivo proporcionar materiales de formación para el desarrollo de competencias entre los y las profesionales. Los materiales están dirigidos a la enseñanza de la escritura científica y la publicación científica, a formar

a revisores y revisoras competentes, a cómo trabajar de forma colaborativa y cómo gestionar los datos (https://masterclasses.nature.com). Concretamente, 'Focus on Peer Review' está dedicado a la formación de revisor o revisora y es un curso gratuito en línea que contiene videos de 1 a 5 minutos donde se ofrecen recomendaciones sobre los puntos clave para llevar a cabo una revisión de forma adecuada (Glezerson y Bryson, 2019). Por ejemplo, es importante que los revisores y revisoras estén atentos a la interpretación errónea que los autores o autoras de las investigaciones podrían realizar sobre los resultados estadísticamente no significativos (resultados nulos) como evidencia de ausencia de efecto o de ausencia de relación entre las variables. Conviene tener presente que cuando un solo estudio informa de un resultado nulo solo se puede concluir que proporciona evidencia no concluyente sobre el efecto o sobre la relación hallada entre las variables, concluyendo que no hay evidencia de un efecto y no se debe concluir que hay evidencia de ausencia de efecto. Como señalan Altman y Bland (1995), ausencia de evidencia, no es evidencia de ausencia. Y esa interpretación errónea de los resultados nulos o no estadísticamente significativos suele ser muy común en las publicaciones e informes. Además, un resultado no concluyente ejecutado con un diseño adecuadamente planificado y ejecutado (con especial atención a la planificación de la potencia estadística) debe ser publicado y debería formar parte de los estudios de revisión sistemática y de meta-análisis para aumentar la credibilidad de sus resultados resumen de los efectos registrados en las publicaciones e informes primarios.

Por otra parte, también las instituciones que financian los proyectos de investigación deben valorar como elemento clave la calidad metodológica del proyecto de investigación y junto a las propias universidades deberían reflexionar sobre el sistema de incentivos que proyectan en la comunidad científica basado principalmente en la cultura de la publicación (publicar o perecer), donde se favorece la cantidad de publicaciones y los factores de impacto de las revistas y no su calidad.

Un ejemplo que se ha sometido a debate científico, debido a la calidad de sus hallazgos, es el de las publicaciones sobre la COVID-19. La alarma social y sanitaria que ha supuesto el problema del coronavirus tuvo una rápida respuesta desde el mundo científico y rápidamente se publicaron miles de artículos. Schwab y Held (2020) señalan que más de veinte mil artículos fueron indexados en el motor de búsqueda de PubMed en los primeros 167 días después de la declaración de

pandemia por la Organización Mundial de la Salud (OMS). Esos momentos de pandemia y de necesidad de conocimiento sobre la COVID-19, probablemente condujo a realizar las revisiones por pares (revisión crítica de los manuscritos por científicos y científicas antes de su publicación) de una manera rápida, con mayor presión y quizás esas circunstancias afectaron a la calidad de la revisión. Ese proceso de revisión en las revistas suele durar meses y era un momento de emergencia mundial que requería tener pruebas o evidencias científicas. Una herramienta de revisión muy útil en esas circunstancias es la plataforma 'Outbreak Science Rapid PREreview' (outbreaksci.prereview.org), ya que ofrece revisiones rápidas de los preprints (manuscritos que aún no han sido revisados por pares) dado que son necesarias en una situación de emergencia como la del coronavirus. En esa plataforma los investigadores e investigadoras pueden llevar a cabo revisiones y valorar si los datos y el código están disponibles de forma gratuita en abierto y si el diseño del estudio incluye la suficiente información para llevar a cabo una replicación del estudio. La consulta de este tipo de revisiones de los preprints es fundamental para poder acceder a información sometida al filtro de la revisión por pares ya que se trata de un filtro o control de la evidencia científica.

La meta-investigación o los estudios sobre la calidad de las investigaciones realizadas destaca que una gran cantidad de investigaciones tiene problemas de calidad científica. Glasziou y cols. (2020) subrayan el alto número de ensayos clínicos COVID-19 registrados en ClinicalTrials.gov y señalan que algunos han sido útiles, pero muchos de ellos se han realizado con tamaños de muestra pequeña y no se han diseñado adecuadamente. Destacan los autores que de 145 ensayos registrados sobre los efectos de la hidroxicloroquina para tratar la COVID-19, 32 planifican un tamaño de muestra ≤ 100, 10 no tienen grupo de control y 12 utilizan grupo de control no equivalente o grupos no formados por asignación aleatoria del tratamiento. Además, solo uno de los estudios proporciona un protocolo y luego los investigadores o investigadoras llevan a cabo cambios injustificados. Glasziou y cols. (2020) subrayan de forma especial el escaso número de ensayos sobre intervenciones no farmacológicas. Solo 2 estudios se centran en el efecto de las mascarillas y ninguno analiza el efecto del distanciamiento social, la cuarentena, la higiene de las manos u otro tipo de intervenciones no farmacológicas. Se trata de cuestiones fundamentales para evitar la transmisión del virus que no fueron investigadas. Además, y después

de consultar la base de datos “Covid-19 Research Project Tracker” (https://www.ukcdr.org.uk/covid-circle/covid-19-research-project-tracker/), los autores corroboran que practicamente no existía investigación primaria sobre los efectos de las intervenciones no farmacológicas en la transmisión del virus, centrándose la financiación en los proyectos de intervención farmacológica con al menos 74 millones de dólares.

Teniendo en cuenta lo expuesto anteriormente, se puede llegar a la conclusión de que la ética y la conducta de integridada científica son aspectos esenciales de investigadores e investigadoras y de lectores y lectoras. La planificación y ejecución de un estudio que no ha sido sometido al rigor del método científico puede conducir a engaños y decisiones dañinas que empañan la literatura con información poco fiable que podría afectar a la salud de la población y al mal uso de la financiación con dinero público de los proyectos de investigación.

Aspectos éticos en el manejo de la investigación psicológica: código de conducta y responsabilidad científica

El Código de Conducta Europeo para la Integridad de la Investigación (2017) (European Code of Conduct for Research Integrity) (Código ALLEA, Federación Europea de Academias de Ciencias y Humanidades, ALLEA en sus siglas en inglés), se utiliza en la comunidad de investigación europea como marco para la autorregulación en todas las disciplinas científicas y académicas y para todos los entornos de investigación y es de cumplimiento obligado en el contexto de los Programas de Investigación financiados por la Unión Europa (De Lecuona, 2020).

La Comisión Europea reconoce al Código de Conducta Europeo para la Integridad de la Investigación como el documento de referencia para la integridad de la investigación para todos los proyectos de investigación financiados por la Unión Europea y como modelo para organizaciones, investigadores e investigadoras de toda Europa. Destaca el Código que las buenas prácticas de investigación se basan en principios fundamentales de integridad en la investigación, orientan a los investigadores e investigadoras en su trabajo, así como en lo referente a su compromiso con los desafíos prácticos, éticos e intelectuales inherentes a la

investigación. La conducta responsable en investigación incluye la adecuada gestión y conservación de la información derivada de la investigación (Grupo Europeo de Ética de la Ciencia y las Nuevas Tecnologías, 2018).

Dicho Código define los principios fundamentales de la integridad en la investigación como fiabilidad, honradez, respeto y responsabilidad ya que representan la base de la actividad investigadora que se considera íntegra (Código de Conducta Europeo para la Integridad de la Investigación, 2017) (ver figura 1).

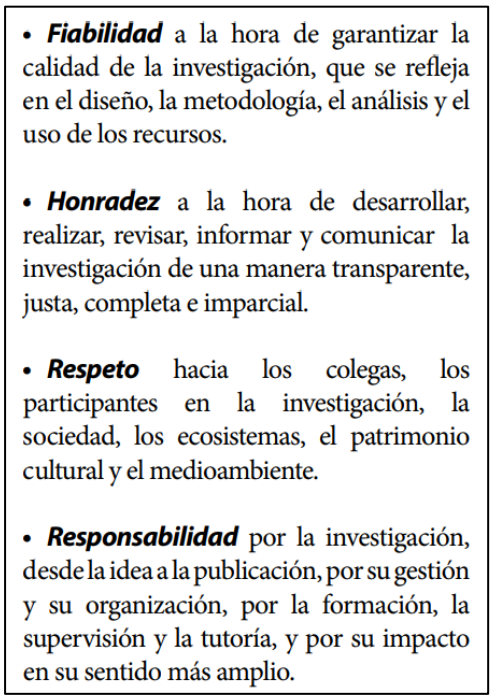


Figura 1. Principios fundamentales de integridad en la investigación. Código de Conducta Europeo para la Integridad de la Investigación (2017)

Desde la perspectiva de la 'integridad en la conducta' del investigador o investigadora, conviene resaltar una serie de cuestiones e informaciones que son fundamentales para la reflexión que todo profesional de la Psicología o de todas las Ciencias en general debería abordar cuando se está formando académicamente y, por supuesto, cuando ejerce su profesión.

1. En primer lugar, los profesionales de la Salud y las Ciencias Sociales se encuentran en un continuo desarrollo profesional y la responsabilidad ética individual exige la actualización constante de los conocimientos que garantice la justificación de las decisiones adoptadas dentro de un modelo basado en las mejores pruebas, resultados o evidencia disponible (modelo de Práctica Basada en la Evidencia). La

relación entre los años que han transcurrido desde la graduación y los conocimientos actualizados que tiene el profesional sobre el mejor tratamiento es estadísticamente significativa y negativa ("curva resbaladiza" de la evolución del conocimiento tras la graduación). De ahí la necesidad de una formación continua y actualizada a lo largo de la trayectoria profesional.

En España el Código Deontológico del Psicólogo (2010) señala que "la autoridad profesional del Psicólogo/a se fundamenta en su capacitación y cualificación para las tareas que desempeña. El/la Psicólogo/a ha de estar profesionalmente preparado y especializado en la utilización de métodos, instrumentos, técnicas y procedimientos que adopte en su trabajo. Forma parte de su trabajo el esfuerzo continuado de actualización de su competencia profesional. Debe reconocer los límites de su competencia y las limitaciones de sus técnicas" (artículo 17).

En el artículo 18 del Código Deontológico del Psicólogo (2010) se señala: "Sin perjuicio de la legítima diversidad de teorías, escuelas y métodos, el/la Psicólogo/a no utilizará medios o procedimientos que no se hallen suficientemente contrastados, dentro de los límites del conocimiento científico vigente. En el caso de investigaciones para poner a prueba técnicas o instrumentos nuevos, todavía no contrastados, lo hará saber así a sus clientes antes de su utilización".

Por lo tanto, la necesidad de la actualización continua de los conocimientos y la necesidad de aplicar los tratamientos que se han contrastado de forma válida y con pruebas científicas que avalen su efecto son aspectos fundamentales a tener en cuenta en la práctica profesional del psicólogo o psicóloga.

Desde el ámbito internacional destaca el Código Ético de la American Psychological Association, APA (2002) que afirma que los psicólogos y psicólogas deben mantener y desarrollar su competencia (artículo 2.03: "Maintaining Competence: Psychologists undertake ongoing efforts to develop and maintain their competence") y su trabajo debe estar basado en el conocimiento profesional y científico establecido de la disciplina (artículo 2.04: "Bases for Scientific and Professional Judgments: Psychologists' work is based upon established scientific and professional knowledge of the discipline"). El primer código ético de la APA se elaboró en 1953 y posteriormente fue modificado y se convirtió en las primeras pautas éticas que guiarían el ejercicio de la profesión (American Psychological Association, 1953).

2. En segundo lugar, los investigadores e investigadoras que llevan a cabo estudios empíricos deben someterse a unas pautas de integridad científica y responsabilidad personal que son la base fundamental del proceso del diseño de investigación. Si la planificación del proceso de investigación no tiene como raíz principal la ética y la integridad del investigador o investigadora, entonces poco se podrá hacer después durante el proceso del diseño de investigación y, por supuesto, los resultados y su interpretación quedarán contaminados.

A veces, la falta de integridad es por desconocimiento de las implicaciones metodológicas que su conducta tendrá sobre el resultado final al procesar y analizar los datos (conocido como 'grados de libertad del investigador o investigadora') y otras veces se debe directamente a una conducta maliciosa e intencional dirigida a crear y/o falsificar los datos y obtener unos resultados concretos independientemente de la realidad del fenómeno: se trata del fraude científico: fabricación o invención, falsificación y plagio, FFP (figura 2).


• ***Invención*** se refiere a inventar resultados y registrarlos como si fueran reales.

• ***Falsificación*** se refiere a manipular materiales, equipos o procesos de la investigación o a cambiar, omitir o suprimir datos o resultados sin justificación.

• ***Plagio*** se refiere a utilizar el trabajo y las ideas de otras personas sin citar adecuadamente la fuente original, violando así los derechos del autor o autores originales respecto a su producción intelectual.

Figura 2. Fabricación (invención), falsificación y plagio. Código de Conducta Europeo para la Integridad de la Investigación

Por ejemplo, desde el punto de vista de la formación en investigación (investigador o investigadora y también del lector o lectora activo), desconocer las implicaciones que tendrá sobre los resultados la denominada baja potencia estadística es una cuestión que afecta a la conducta del investigador o investigadora y también a la competencia de los lectores y lectoras. Desde un punto de vista ético, ese desconocimiento podría ser causa de mala conducta profesional, ya que el

profesional desconocía su herramienta de trabajo y sus consecuencias sobre el paciente. El desconocimiento de las cuestiones de diseño de la investigación y estadística no justifican el trabajo incompetente de investigadores, investigadoras, lectores y lectoras. La formación de esos profesionales debe ser continua a lo largo de toda su trayectoria profesional, con un continuo repaso y reciclaje de todas las cuestiones de formación académica que están relacionadas con su trabajo.

El investigador / la investigadora es el responsable de su formación, así como responsable de poner todos los elementos metodológicos que aseguren la calidad del proceso de diseño de investigación del estudio o de la intervención que realice con sus pacientes. Al mismo tiempo, es responsable de tener las suficientes competencias para llevar a cabo una lectura crítica o activa de los informes y artículos que consulta para estar actualizado en los avances que se producen en su área de trabajo. En este último punto, se exigen competencias vinculadas al área de la metodología de investigación junto con las competencias teóricas propias del campo de trabajo donde se actúa.

Como lector o lectora de literatura científica, el profesional debe tener la suficiente formación en metodología que asegure que puede llevar a cabo una lectura de los informes científicos más allá de una comprensión sustantiva o ingenua; es decir, es necesario que sepa valorar los hallazgos de una forma crítica o activa que le permita emitir un juicio sobre la validez de los resultados y sobre las afirmaciones que se mencionan en los informes. Se trata de realizar una lectura científica, no una lectura ligera y sin cuestionar el relato. En definitiva, los lectores y las lectoras deben ser capaces de ‘separar el grano de la paja’, es decir, deben seleccionar para su lectura a los artículos con diseños adecuadamente planificados y ejecutados del conjunto amplio de publicaciones e informes que diariamente se producen.

El Comité de Ética del Consejo Superior de Investigaciones Científicas (CSIC) español también ha desarrollado un Código de Buenas Prácticas Científicas (Consejo Superior de Investigaciones Científicas, 2021) donde se reflexiona sobre la conducta responsable en investigación y la integridad científica. En dicho Código se señala: “La integridad científica —fundamento esencial de las buenas prácticas— se identifica con un patrón de conducta que conlleva la observancia y promoción de los más elevados estándares profesionales y principios morales en el ejercicio de la

investigación" y "Es responsabilidad personal del investigador que la integridad científica oriente el ejercicio de su actividad, si bien el fomento y establecimiento de una cultura de integridad incumbe a la comunidad científica en su conjunto y, en particular, a las instituciones en las que se desarrolla la investigación" (p. 9). Instituciones y profesionales son responsables de fomentar y llevar a cabo una conducta responsable y moral relacionadas con las tareas de investigación científica que "que asegure la calidad y el rigor en las distintas facetas de la investigación (propuesta, ejecución, difusión y evaluación), el cumplimiento de la normativa aplicable, y la consideración de posibles cuestiones éticas". (p. 9). Además, se destaca la importancia de planificar y ejecutar investigaciones que se sustenten en una adecuada planificación del diseño y de su ejecución: "Los experimentos y observaciones deben estar cuidadosamente diseñados, con rigor e inteligencia, con el fin último de asegurar la obtención de información veraz y completa, y el mejor uso de los recursos disponibles, siempre teniendo en cuenta las particularidades propias de cada actividad. Esto es exigible en mayor medida cuando el objeto de la investigación son seres humanos, sus muestras y datos; animales, o cuando la seguridad humana o del medio ambiente puede estar en juego. Se aplicarán métodos estadísticos idóneos para el análisis e interpretación de los datos generados, así como para el diseño experimental cuando este lo requiera" (p. 10). Y "Es esencial prevenir la ocurrencia de sesgos en la obtención, tratamiento e interpretación de los resultados. A este fin, el personal investigador extremará el celo y el rigor metodológico en las distintas etapas del proceso de investigación, y utilizará las estrategias adecuadas en cada caso para evitar, minimizar y controlar posibles sesgos" (p. 10).

En resumen, el saber científico requiere una constante actualización y requiere las competencias necesarias para elaborar un análisis de la anatomía metodológica del proceso del diseño de investigación. No todo lo que se publica tiene la calidad suficiente para aportar pruebas o evidencia de un fenómeno y, además, la calidad de los estudios no es del todo o nada, sino que se puede jerarquizar y exige que los y las profesionales sepan valorar el grado de confianza que puede tener la información que aporta un artículo o un informe de investigación.

Ética e Integridad en la investigación

Hay dos exigencias que se deben demandar de forma contundente a la investigación científica: que sea de alto nivel científico y que sea relevante para la sociedad. La primera medida es necesaria y no hay dudas sobre su necesidad y forma parte tanto de la valoración de los proyectos de investigación como de la evaluación académica. Respecto a la segunda medida de relevancia social, ya no parece que sea tan evidente y de ahí el debate que existe en la literatura (Bouter, 2008).

La ética de la investigación está relacionada con las cuestiones que tienen que ver con la consideración ética de la investigación con humanos y animales. Hay leyes dirigidas a controlar la ética de la investigación, pero no hay leyes específicas para la integridad de la investigación. La integridad de la investigación está relacionada con la investigación responsable, es decir, con el área de la relevancia social de la investigación, vinculada a los beneficios y daños que la investigación puede causar a la sociedad y al medio ambiente (Bouter, 2019).

Respecto a la integridad de la investigación, el profesor Lex Bouter (profesor de "Metodología e Integridad" en la Facultad de Humanidades de la Vrije Universiteit de Amsterdam), señala que la integridad de la investigación se solapa, en cierto grado, con la ética de la investigación y ambos conceptos tienen, a su vez, cierto solapamiento con lo que se denomina "investigación e innovacion responsables" (Bouter, 2019). Desde su punto de vista, la integridad de la investigación se refiere a la conducta del científico o científica como individuo que puede obstaculizar la calidad o validez de los resultados de la Ciencia y amenazar la confianza de la sociedad en los científicos y científicas. En este punto se alude a la fabricación, falsificación y plagio (*fabrication*, *falsification*, *plagiarism*, FFP), pero también a otras formas más sutiles de la conducta del científico o científica que distorsionarán los hallazgos de los estudios buscando el apoyo de las hipótesis científicas y todas las recompensas que la cultura de la investigación ofrece. Se trata de las denominadas prácticas de investigación cuestionables (*Questionable Research Practices*, QRP) o, quizás, sería mejor llamarlas "conductas de investigación cuestionables".

En definitiva, la integridad en la conducta está relacionada con la conducta responsable e implica que los investigadores y las investigadoras deben cumplir los

estandares éticos y de buenas prácticas científicas cuando llevan a cabo sus tareas de investigación (Shamoo y Resnik 2015). Y esa conducta implica tener formación y competencias en metodología de investigación.

Mala conducta científica o fraude

La "mala conducta científica" se define como el comportamiento del investigador o investigadora, intencional o no, que no alcanza los buenos estándares éticos y científicos (Kakuk, 2009). Esta definición incluye las conductas indebidas de FFP ya mencionadas (fabricación o invención, falsificación y plagio) y las prácticas o conductas de investigación cuestionable. La mala conducta en la investigación erosiona la integridad de la investigación, la reputación, la confianza del público y el apoyo social a la Ciencia (Breen 2016; Martinson y cols, 2005).

En el artículo de Mariño-Hernández (2016) se resumen los criterios que las revistas suelen encuadrar en el término de fraude científico. El artículo de opinión de Perez y Sevilla (2019) también alerta de los peligros del fraude y las malas prácticas en Ciencia y se puede consultar en https://www.jakiunde.eus/blog/2019/08/el-fraude-y-las-malas-practicas-en-ciencia/. En el blog (blog de Jakiunde; https://www.jakiunde.eus) también se puede leer una colección de materiales relacionados con "Los males de la ciencia" como "El marco en que se desarrolla la ciencia", "Las publicaciones científicas", "El ethos de la ciencia", "Los valores en la filosofía de la ciencia", "Los propietarios del conocimiento", "El papel de los gobiernos en el desarrollo científico", "No todos tienen las mismas oportunidades de hacer ciencia", "El fraude y las malas prácticas en ciencia", "Ciencia patológica", "Sesgos cognitivos que aquejan a la ciencia" y "Sesgos ideológicos que aquejan a la ciencia". Y en el artículo de Tudela y Aznar (2013) se reflexiona sobre el fraude y el problema que rodea a la Ciencia rotulado como "publicar o morir" (*Publish or Perish?*).

Un ejemplo de presunto fraude, detectado en 2016 y que saltó a los medios de comunicación en 2017, está relacionado con el trabajo de una bióloga que nunca aportó los datos originales de sus estudios y le retiraron los 1,86 millones de euros que la Unión Europea le había concedido para sus investigaciones sobre las enfermedades del corazón. Más información sobre este caso y otros se puede consultar en Internet con las palabras clave: "fraudes científicos en España"; ella siempre ha defendido su inocencia aunque el Centro Nacional de Investigaciones

Cardiovasculares (CNIC) la despidió de manera fulminante tras comprobar ciertas irregularidades en sus publicaciones y participaciones en congresos que nunca pudo aclarar. En el reportaje de Ariza publicado en el País Semanal se pueden consultar otros ejemplos (Ariza, 2020).

Conductas cuestionables

Las conductas cuestionables (pueden realizarse de forma consciente o de forma inconsciente) no tienen como objetivo fabricar datos, falsificar los datos del estudio o plagiar otros datos. Sin embargo, son también dañinas porque falsean la realidad de los fenómenos e impiden el avance del conocimiento científico, ya que engañan con resultados que difícilmente podrán ser replicados. A veces, puede suceder que el investigador o investigadora lleve a cabo esa práctica cuestionable pensando que enriquece su estudio, por ejemplo, aumentando la muestra y parando ya de recoger datos cuando rechaza la hipótesis nula, sin tener un plan previo de tamaño de la muestra necesario para su estudio, debido a que desconoce el tamaño del efecto mínimo que es relevante para su estudio o sin pensar que un resultado estadísticamente significativo no otorga de forma directa importancia o relevancia al hallazgo de su estudio.

El desconocimiento en los fundamentos del diseño de la investigación y la estadística puede ser la causa de esas decisiones que sin criterio adopta el investigador o investigadora a lo largo del proceso de diseño de su estudio y que, en ocasiones, pueden derivar en resultados 'falsos positivos' que distorsionan la calidad del conocimiento científico. Los falsos positivos son resultados que indican que hay un efecto o una relación entre las variables y, sin embargo, realmente no existe.

Esta problemática se estudia con el término de "grados de libertad del investigador o investigadora" y su concepto se ampliará a continuación. Se trata de decisiones no fundamentadas o planificadas, basadas en la libertad que tiene el científico o la científica a lo largo del proceso de diseño y análisis estadístico y que provocan un gran daño a la Ciencia y a su reputación y, además, desperdicia el tiempo y el dinero de la investigación (Martinson y cols., 2005). Incluso podría dañar a la salud de los pacientes o a la economía de un país si las decisiones de los y las profesionales o la de los políticos o políticas se toman en función de hallazgos que no se han obtenido garantizando la calidad del proceso de obtención de los

resultados (validez) de forma válida y, por lo tanto, no reflejan la realidad (Antonelli y Sandroni, 2013).

Los datos sobre este tipo de conductas cuestionables deben conducir a la reflexión personal de los científicos y científicas, lectores y lectoras y también del alumnado en general. Se ha estimado que la prevalencia de la mala conducta en la investigación es de aproximadamente el 2% en las cuestiones de fabricación y falsificación y del 1.7% respecto al plagio (Fanelli, 2009; Pupovac y Fanelli, 2015). Cuando se trata de las prácticas de investigación cuestionables (no incluyen las conductas FFP) se estima una mayor prevalencia, estando en torno al 34% (Fanelli, 2009). Datos que preocupan a la comunidad científica y a la sociedad en general.

En el estudio de Frías-Navarro y cols. (2021) se analizan las opiniones de 348 académicos y académicas españoles acerca de la presencia de determinadas prácticas de investigación cuestionables entre los científicos y científicas. El 5.8% de los y las participantes creen firmemente que sí existe fraude en la Ciencia, el 30% señalan que “podría” existir fraude y el 64.2% afirma de forma contundente que no hay fraude.

En el estudio de Hofmann y cols. (2020) se llevó a cabo la investigación con estudiantes de doctorado. Sus resultados señalan que aproximadamente un 10% de los encuestados opinan que la mala conducta de FFP es común en su área de investigación, mientras que un poco más señalan estar de acuerdo con la existencia de otras malas conductas. Además, aproximadamente un 1% de los encuestados reconoce haber cometido una mala conducta grave (FFP) durante el último año, en torno al 10% informa de haber plagiado, aproximadamente un 10%-20% informa que ha alterado los datos de su investigación de una forma cuestionable y aproximadamente un 30% decidió recopilar más datos para obtener un resultado estadísticamente significativo.

Es muy importante reflexionar sobre los datos anteriores como alumnos y alumnas y como lectores y lectoras de los informes científicos. Por su parte, el profesorado, los tutores y tutoras de los futuros investigadores e investigadoras y los académicos y académicas en general, y las propias universidades y organismos que financian la investigación también deberían reflexionar sobre qué modelo ofrecen al alumnado que serán los futuros profesionales y científicos y científicas, cómo se

transmiten los valores de integridad y éticas personal desde el sistema educativo y de investigación, la importancia de fomentar la integridad en la investigación y la necesidad de ser lectores y lectoras competentes, enseñando y cuestionando la cultura de la investigación y la publicación actual (que refuerza en gran medida la competencia académica y el publicar o perecer) así como debatiendo en las clases la importancia de la integridad científica como norma saludable que fomenta la calidad de la Ciencia (Frías-Navarro y cols., 2020).

La alta competencia académica y otros factores estresantes que se viven en el mundo académico y científico como la presión por publicar en revistas de primer nivel y el estrés por obtener fondos económicos que permitan llevar a cabo la investigación se perciben como las principales causas de mala conducta científica junto con la baja probabilidad de que se detecte una mala conducta y, además, sea sometida a una pena con consecuencias graves (Holtfreter, 2020).

Grados de libertad del investigador o investigadora

La creciente sofisticación estadística, metodológica y técnica facilita que ahora sea el momento de abrir (de nuevo) el debate sobre la calidad de los resultados de la Ciencia, aportando pruebas o evidencia empírica sobre los elementos que podrían afectar a las decisiones del investigador o investigadora a lo largo del proceso de diseño de investigación y que se denomina como “grados de libertad del investigador o investigadora”. No se trata de fraude científico o las conductas FFP (en el sentido de intencionalidad en la conducta para crear resultados falsos, para falsificarlos o para plagiarlos), pero sí de un tema vinculado con la ética y la integridad del investigador o investigadora tal y como ya se ha comentado. Se trata de decisiones que puede tomar el investigador o investigadora basadas en la flexibilidad que tiene a lo largo del proceso del diseño de investigación y que trae de nuevo a la luz la importancia de la ética personal, la integridad y los valores morales en la Ciencia; tres piezas clave para hacer buena Ciencia.

Por lo tanto, los denominados “grados de libertad del investigador o de la investigadora” son decisiones que toman no dirigidas intencionadamente al fraude, a falsificar los datos o al plagio, pero sí que podrían ir dirigidas hacia la búsqueda del resultado estadísticamente (conocido como “*p*-hacking”, se podría traducir como la piratería del valor *p*) porque así se garantiza, en gran medida, la publicación del

trabajo (provocando a su vez el 'sesgo de publicación': se publican mucho más los estudios con resultados estadísticamente significativos), pensando (justificándose el propio investigador o investigadora) que se trata de una mejora del diseño del estudio, pero tomando las decisiones a posteriori (una vez ya se dispone de resultados y, por lo tanto, ajustando los datos para lograr que el hallazgo sea estadísticamente significativo). La conducta de *p*-hacking supone que las decisiones están determinadas única y exclusivamente por el valor *p* obtenido en el estudio y se dirige todo el esfuerzo o conducta hacia lograr obtener un valor de *p* estadísticamente significativo ($p \leq .05$), sin reflexionar sobre la calidad o validez de los resultados.

Con la conducta del *p*-hacking los investigadores y las investigadoras no son conscientes de que sus decisiones a lo largo del proceso del diseño de investigación podrían conducir a lo que se llama un "resultado falso positivo": informar de un resultado estadísticamente significativo cuando en realidad es falso.

Cultura de la publicación y cultura de la investigación

La conducta de *p*-hacking en sí misma no es el problema que tiene la Ciencia, sino que es un síntoma de la 'cultura de la publicación'. Se publican en mayor medida los estudios con "resultados positivos" o estudios donde se detectan efectos o relaciones estadísticamente significativos frente a los denominados "resultados negativos o resultados nulos" donde se concluye que el efecto o la relación entre las variables no es estadísticamente significativo, guardando estos hallazgos en el cajón o archivador sin tratar de publicarlos. Los dos términos no son apropiados ya que se vincula el éxito al resultado positivo y el fracaso al resultado negativo, potenciando actitudes y creencias erróneas sobre la significación estadística como un oráculo de la verdad.

Y la conducta de *p*-hacking también es un síntoma de la 'cultura de la investigación' por la forma de planificar, hacer diseño de investigación, analizar los datos e interpretar los resultados. Y esa cultura de la investigación también incluye la forma de publicar los hallazgos, pues muchas veces los artículos se parecen a cajas oscuras que impiden observar cómo se llevó a cabo el estudio y qué decisiones se tomaron. Todo el proceso de diseño se mantiene oculto.

Por ejemplo, cuestiones metodológicas que empañan la calidad de las investigaciones y que están relacionadas con el diseño de los estudios suelen caracterizarse por:

- Estudios con muestras pequeñas (problemas con la validez de conclusión estadística).
- Escasa potencia estadística para detectar el efecto o la relación entre las variables (problemas con la validez de conclusión estadística).
- En muchas ocasiones tamaños del efecto pequeños (problemas con la validez de conclusión estadística).
- Sesgo de selección de los participantes (problemas de validez interna y externa).
- Falta de planificación de los elementos fundamentales que forman parte del diseño y el análisis estadístico: alfa, beta, efecto estimado, potencia deseada, *N* o número de observaciones (muestra) a recoger (problemas de validez de conclusión estadística).
- Desconocimiento de la magnitud del efecto que se desea detectar (efecto estimado) y de ahí su falta de consideración en la planificación del estudio y en la estimación de la potencia estadística a priori (problemas con la validez de conclusión estadística).
- Instrumentos de medida con escasa fiabilidad o consistencia interna (problemas con la validez de constructo).
- Falta de información sobre la elaboración, ejecución del estudio y redacción de los resultados (problemas en el informe o reporte de los hallazgos de investigación).
- Mala comprensión de lo que representa un resultado nulo o resultado negativo (mantener la hipótesis nula en el contraste estadístico) y lo que significa un resultado estadísticamente significativo o resultado positivo (rechazar la hipótesis nula en el contraste estadístico) dentro del procedimiento clásico de significación de la hipótesis nula (NHST). De ahí las interpretaciones incorrectas que se realizan de los resultados obtenidos con las pruebas de contraste estadístico (problemas de comprensión y educación estadística) ya que se vincula el resultado estadísticamente con ser un resultado importante y válido y esta idea es errónea.

No todos los grados de libertad del investigador o investigadora son prácticas de investigación cuestionables, pero sí es cierto que en la toma de decisiones del investigador o investigadora a lo largo del proceso del diseño de investigación se puede optar por una decisión cuestionable desde el punto de vista de integridad científica, ya que distorsiona la realidad buscando los resultados estadísticamente significativos. Y estas conductas también afectan a la probabilidad de replicar dichos hallazgos, dado que generalmente sus resultados no se replican.

Es cierto que hay muchas razones basadas en el rigor metodológico que permiten eliminar, por ejemplo, una hipótesis una vez ejecutado el estudio, pues por ejemplo, si la consistencia interna de un instrumento no fue la adecuada entonces se puede argumentar por qué esa variable no fue estudiada o si, finalmente no se pudo completar la recogida de muestra planificada en un grupo se puede argumentar por qué las puntuaciones de ese grupo no se tuvieron en cuenta en el análisis de los datos. Pero todas esas conductas deben justificarse en la redacción del propio estudio desde un punto de vista metodológico, escribiendo un informe que debe ser transparente y aportar toda la información. Además, esas conductas deben basarse en criterios fijados priori (es decir, antes de ejecutar el estudio) que el investigador o investigadora deberá dejar reflejado en el denominado 'protocolo' de su estudio (elaborado antes de realizar la investigación) y nunca se deben realizar maniobras o ajustes forzados de los datos (por ejemplo, eliminando o añadiendo datos sin dejar constancia de argumentos que lo justifiquen) para poder concluir que se ha comprobado la hipótesis de investigación, ya que se rechazó la hipótesis nula.

El registro público de los protocolos de investigación indicando qué decisiones metodológicas se adoptarán cuando se lleve a cabo el estudio es una buena medida para evitar (o al menos, ejercer cierto control) sobre las conductas de investigación cuestionables.

En resumen, el investigador o investigadora cuando lleva a cabo un estudio quiere confirmar sus hipótesis sustantivas y piensa que a lo largo del proceso del diseño de su investigación puede tomar ciertas decisiones para mejorar dicho diseño, ya que con ello aportará resultados válidos y, además, aumenta la probabilidad de que su trabajo sea publicado (Smaldino y McElreath, 2016). Sin embargo, no es consciente, a veces, de que se trata de una mala conducta, pues sus decisiones

deben estar registradas previamente en el protocolo y, sobre todo, justificadas metodológicamante y narradas en el informe para que puedan ser valoradas críticamente por la comunidad científica. En cualquier caso, que haya o no voluntariedad en las malas prácticas de investigación, lo importante es que se dan y pervierte el alcance y la validez del conocimiento científico (Frías-Navarro y cols., 2021).

Ante esta situación, la denominada 'meta-investigación' o meta Ciencia (investigación sobre la investigación ya realizada) trata de describir la realidad de las actuaciones que se llevan a cabo en la Ciencia para aportar herramientas de educación y re-educación metodológica que garanticen la ejecución de investigaciones fiables y válidas (Ioannidis, 2018; Ioannidis y cols., 2015). Otra cuestión también importante para reflexionar es cómo valorar el trabajo de los científicos que hacen meta-investigación (es decir, ¿quién controla al controlador?).

En resumen, elaborar una investigación exige una conducta íntegra y ética del investigador o investigadora que comienza con tener una preparación metodológica adecuada que garantice la calidad del proceso del diseño de investigación que ejecuta, puesto que su preparación repercute en la calidad de sus hallazgos y en su correcta interpretación y redacción. Por otra parte, leer la literatura científica exige acercarse a su contenido desde una perspectiva crítica o activa que requiere también una formación metodológica adecuada que garantice la posibilidad de detectar sesgos o problemas que amenazan a la validez de los resultados presentados en el artículo o informe de investigación.

Por todo ello, la transparencia en la redacción del informe de investigación o artículo en todos sus apartados (también aportando la base de datos del estudio y la sintaxis de análisis en repositorios abiertos o públicos, en plataformas especializadas en abierto o en anexos *on-line* que mantienen las propias revistas donde se publica el artículo), el pre-registro del estudio (elaborar un protocolo y publicarlo) y la publicación del informe final en abierto en los repositorios o en las revistas (donde se somete de nuevo y de forma perpetua a la evaluación de los investigadores e investigadoras y los lectores y lectoras) son fundamentales para evitar el sesgo en la conducta del investigador o investigadora y para posibilitar los estudios de meta-

ciencia o de análisis secundario de los resultados científicos aportados en los informes y las publicaciones.

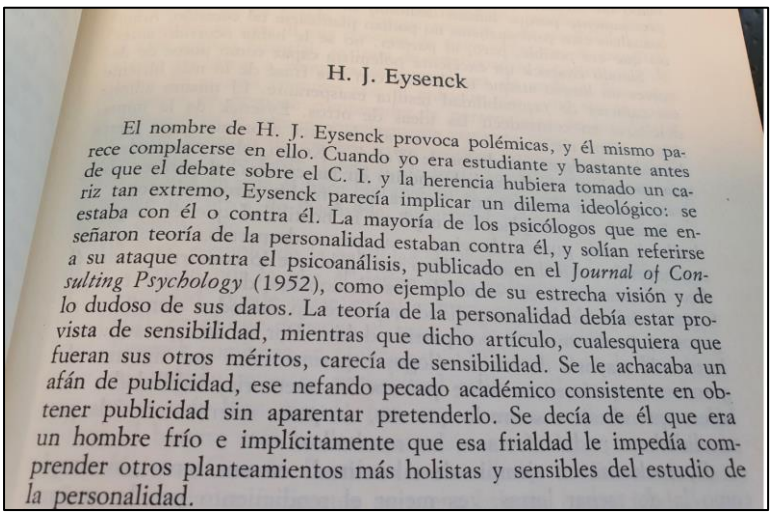
Los estudios de meta-análisis mejorarían de forma destacada también con esas prácticas de transparencia en la información estadística aportada en el informe, pues muchas veces se debe recurrir a estimaciones sesgadas de los estadísticos, ya que el artículo no detalla los datos necesarios para hacer un cómputo directo del tamaño del efecto y su intervalo de confianza. Por ello, en los informes o artículos es necesario informar siempre de las puntuaciones medias junto con su desviación típica y tamaño muestral de cada grupo e informar de los resultados de correlación junto con su tamaño muestral. De este modo, se facilita realizar estimaciones más precisas del tamaño del efecto así como estimar su intervalo de confianza, mejorando la interpretación sustantiva de los hallazgos y la calidad del tamaño del efecto medio y su intervalo en los estudios de meta-análisis. Además, disponer de esa información permite que los lectores y lectoras actúen como meta-investigadores ya que pueden comprobar los resultados de los autores y autoras y se posibilita que se detecten errores y sesgos.

La perspectiva de la Ciencia Abierta (‘Open Science’) es una de las herramientas que permite el cambio en la conducta del investigador o investigadora e Internet es la vía que la hace posible, ya que proporciona una infraestructura que antes no había y facilita que la comunidad científica acceda a todos los elementos del proceso del diseño de investigación y con ello a su valoración. Sin embargo, no solo con la Ciencia Abierta se puede acabar con la cultura de la publicación y la cultura de la investigación. La educación y la re-educación estadística entre el alumnado, investigadores y profesionales son fundamentales y requisitos imprescindibles para que lleven a cabo una reflexión crítica sobre los elementos del diseño que afectan al resultado final del proceso de diseño de investigación. También las agencias de investigación y la propia filosofía de las universidades están implicadas en ese cambio de la cultura de la investigación y promoción académica.

La integridad científica tambié se aprende con la educación y la re-educación estadística, ya que se visualiza y se comprende el alcance de las decisiones que se toman a lo largo del proceso de diseño de un estudio y se reconoce (se comprende y se valora) cómo las conductas del investigador o investigadora deben estar siempre

justificadas metodológicamente y, sobre todo, se comprende la debilidad del ser humano que le puede llevar a tomar decisiones (conscientes o inconscientes) que quedan fuera de la integridad científica.

En este punto, se propone un ejercicio para la reflexión y la formación personal como psicólogos y psicólogas. Se trata de indagar sobre la labor científica del profesor Hans Eysenck. Entre sus aportaciones a la Ciencia, Eysenck informó que existían ciertos tipos de personalidad que tienen un mayor riesgo de morir por cáncer o por problemas cardíacos. Tradicionalmente, su modelo psicobiológico de personalidad ha sido considerado uno de los más sólidos, mostrando una gran cantidad de evidencia empírica. Sus resultados han sido utilizados en un amplio número de trabajos de meta-análisis (como por ejemplo, el de Chida y cols., 2008) y han sido la base de terapias dirigidas a mejorar la salud, especialmente de los enfermos de cáncer y de corazón. Pero actualmente, en el siglo XXI, qué ha ocurrido en el mundo científico que atribuye a Eysenck una conducta de fraude y mala conducta científica y como consecuencia se han retirado numerosas publicaciones suyas de las revistas científicas (Pelosi, 2019). La falta de transparencia del proceso de diseño de sus investigaciones y de su propio trabajo como investigador junto con la incapacidad para replicar sus hallazgos han sido determinantes para la retirada de algunos de sus artículos. En 1977 David Cohen ya iniciaba su entrevista con Eysenck planteando la duda que cierta parte de la comunidad científica tenía sobre la calidad de sus datos (ver Imagen 2).



H. J. Eysenck

El nombre de H. J. Eysenck provoca polémicas, y él mismo parece complacerse en ello. Cuando yo era estudiante y bastante antes de que el debate sobre el C. I. y la herencia hubiera tomado un cariz tan extremo, Eysenck parecía implicar un dilema ideológico: se estaba con él o contra él. La mayoría de los psicólogos que me enseñaron teoría de la personalidad estaban contra él, y solían referirse a su ataque contra el psicoanálisis, publicado en el Journal of Consulting Psychology (1952), como ejemplo de su estrecha visión y de lo dudoso de sus datos. La teoría de la personalidad debía estar provista de sensibilidad, mientras que dicho artículo, cualesquiera que fueran sus otros méritos, carecía de sensibilidad. Se le achacaba un afán de publicidad, ese nefando pecado académico consistente en obtener publicidad sin aparentar pretenderlo. Se decía de él que era un hombre frío e implícitamente que esa frialdad le impedía comprender otros planteamientos más holistas y sensibles del estudio de la personalidad.

Imagen 2. Entrevista a Eysenck (Cohen, 1977; castellano 1980)

Capítulo 2. Método científico y diseño de la investigación

Marcos Pascual-Soler*

Dolores Frías-Navarro**

Irene Gómez-Frías**

*ESIC Business & Marketing School, España

**Universidad de Valencia

Índice

- Método científico. Definición de investigación científica y características.
- Diseño de una investigación.
- Reforma estadística y Práctica Basada en la Evidencia.

Citar el capítulo como:

Pascual-Soler, M., Frías-Navarro, D. y Gómez-Frías, I. (2021). Método científico y diseño de la investigación. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

Este capítulo del libro "Investigación científica en Psicología" tiene como objetivo presentar una descripción o **definición de la investigación científica y sus características**. Es decir, presenta los aspectos esenciales que describen al método científico y a los contenidos básicos del proceso de diseño de una investigación empírica.

La Psicología como Ciencia empírica adquiere su conocimiento sobre los constructos que le son propios, planteando hipótesis y aplicando una serie de técnicas para controlar, recoger y analizar los datos que le permitirán contrastar dichas hipótesis con la realidad del fenómeno.

Las tareas que el método científico implica se manifiestan en diversas modalidades[1] de investigación o métodos dependiendo del planteamiento teórico de la investigación que guiará el tipo de hipótesis formuladas, así como las técnicas para recoger los datos. En la elaboración de este material se utiliza el término de metodología para abordar las diferentes modalidades de expresión del método científico.

La Ciencia es un conocimiento organizado basado en la observación sistemática. Su meta es aplicar el método científico en la búsqueda de la 'mejor' solución para los problemas de investigación y también, en la búsqueda de la 'mejor' evidencia o pruebas científicas disponible hasta el momento.

Método científico. Definición de investigación científica y características.

El denominado método científico (del griego *-meta*, hacia, a lo largo y de *-odos*, camino y del latín *scientia*: conocimiento; 'camino hacia el conocimiento') supone buscar un hallazgo mediante la observación sistemática siguiendo una serie de etapas ordenadas que forman el proceso de la investigación. Desde el punto de vista de la Ciencia[2], se trata de un procedimiento que abarca el ciclo completo de la

[1] En general, el término *método* se utiliza indistintamente para referirse al método científico y a la modalidad que adopta el método para estudiar el problema concreto objeto de investigación. Para una revisión del tema consultar Mayor y Pérez (1989) quienes optan por mantener el término 'método' para denominar a la globalidad del método científico-positivo y para sus grandes alternativas: método experimental *versus* métodos no experimentales.

[2] Ciencia en griego "episteme", en latín "scientia", en anglo-francés "science", en italiano "scienza" y en alemán: "wissenschaft".

investigación de un determinado problema, avanzando hacia el conocimiento de forma sistemática, racional, rigurosa y crítica.

Muy brevemente, el ciclo del método científico incluye formular un problema de investigación de forma precisa planteando con ello una 'necesidad de conocimiento' que desemboca en la 'formulación de una hipótesis' que habrá que 'contrastarla con la realidad' que se observa del fenómeno en el estudio, buscando evidencias o pruebas que ayudarán al científico o científica a determinar el valor de las hipótesis y del conocimiento producido con la investigación. Los hallazgos obtenidos tras aplicar el proceso del método científico (proceso del diseño de investigación) serán la base de la formulación de un nuevo problema o de una nueva idea (nueva necesidad de conocimiento) que será también sometido al estudio científico con las pautas de dicho método (ver Figura 3). De este modo, se produce el proceso cíclico de la acumulación del conocimiento científico y el progreso de la Ciencia. La integridad científica propia de la conducta del investigador o investigadora estará presente durante todo el proceso del diseño de investigación.

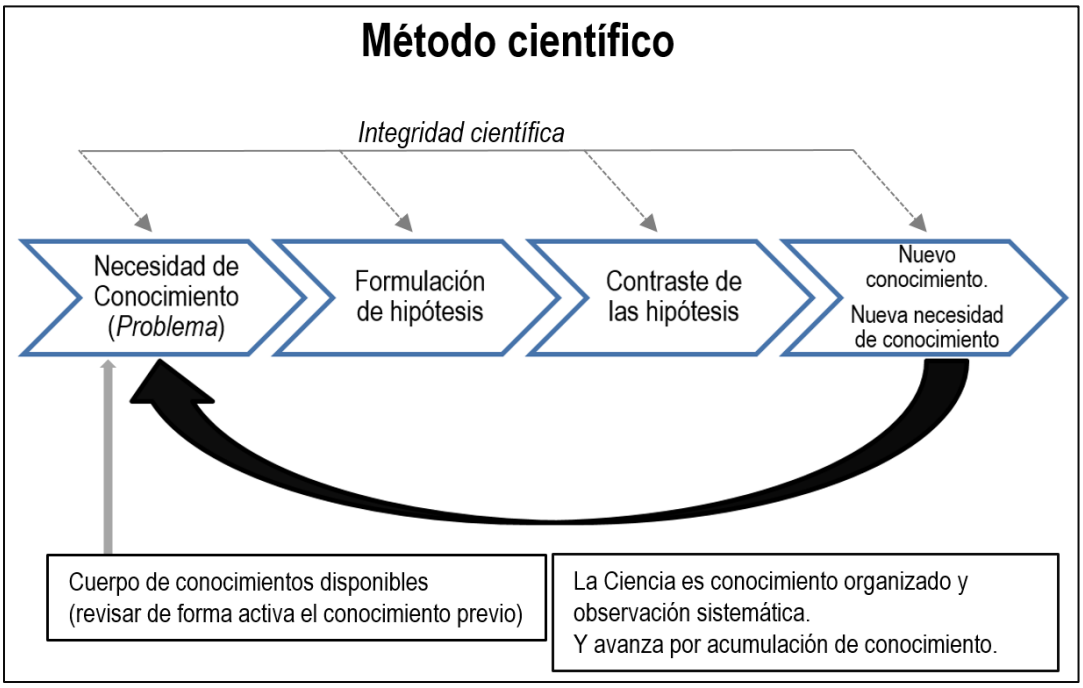


Figura 3. Método científico

Una investigación científica, por lo tanto, se caracteriza por ser *generalizable* (los resultados se pueden aplicar a otras muestras de sujetos, a otros momentos o a otros contextos), *empíricamente replicable* (se pueden replicar los efectos de los

experimentos), *transparente* (se detalla lo que se hizo y por qué se hizo), por apoyarse en los resultados de la *investigación anterior* (conocimiento previo) y por *generar nuevas ideas* para aumentar el conocimiento de un fenómeno y/o eliminar incertidumbre y avanzar así en el conocimiento científico de los fenómenos.

No todos los profesionales de la Psicología o de los diferentes ámbitos de actuación van a realizar experimentos o investigaciones en su trabajo profesional, pero sí todos los profesionales deben buscar la mejor evidencia científica disponible dentro de su área de trabajo (lecturas activas de artículos, informes, base de datos, protocolos, resultados...). Y para decidir qué evidencia científica es la de mayor calidad científica necesariamente necesitan conocimientos de metodología de investigación y del proceso del diseño de investigación. La lectura de los textos científicos (artículos, informes …) se ha convertido en una tarea de 'valoración crítica o activa' que utiliza criterios objetivos de calidad de los hallazgos que están vinculados con todos los elementos del proceso de investigación que implica un diseño.

Hay muchas clases de investigación que aportan conocimiento a las Ciencias Sociales y de la Salud y no todas implican el uso de la estadística, pero también es cierto que la mayoría de las revistas incluyen un gran número de artículos con trabajos empíricos que hacen uso del diseño de investigación y la estadística.

La calidad del conocimiento científico generado en una disciplina requiere que los investigadores e investigadoras planifiquen adecuadamente su investigación, la ejecuten eficientemente, analicen los datos correctamente, interpreten bien los resultados y presenten de forma clara las conclusiones en la redacción de sus informes o artículos. Por supuesto, la calidad del conocimiento científico ya elaborado o publicado también depende de las lecturas y valoraciones de los científicos / científicas y profesionales, ya que, entre otras funciones, son evaluadores de la validez y credibilidad de los hallazgos publicados y ante hechos sospechosos de fraude o contaminación de los resultados deben informar a la comunidad científica para que se determine si un artículo o trabajo es retirado de la revista y justificar el porqué de dicha decisión. La literatura publicada o accesible a los lectores y lectoras estará sometida a evaluación de forma perpetua por los miembros de la comunidad científica.

El desarrollo de la denominada ‘Práctica Basada en la Evidencia’ exige que el investigador o investigadora, el profesional o el consumidor de literatura valoren los resultados de forma crítica y sean conscientes de que toda la información no vale lo mismo, es decir, sus resultados no tienen la misma calidad o validez (ver Figura 4) (Frías-Navarro y Pascual-Llobell, 2003; Pascual-Llobell y cols, 2004). Es necesario “separar el grano de la paja”, es decir, valorar la calidad de los hallazgos para poder desechar los estudios cuyos resultados tienen problemas de validez (presentan sesgos) y atender de forma detenida a los hallazgos de las investigaciones que pasan el filtro de la lectura crítica del profesional o investigador/investigadora.

La validez de los resultados de los trabajos se puede jerarquizar en función del control del sesgo que se haya realizado en el diseño del estudio. Y es importante tener en cuenta que la opinión de la autoridad (‘opiniones de los expertos’) basada en la propia experiencia sin aportar pruebas o evidencia científica con calidad ha dejado de ser una fuente de información con garantías de veracidad.

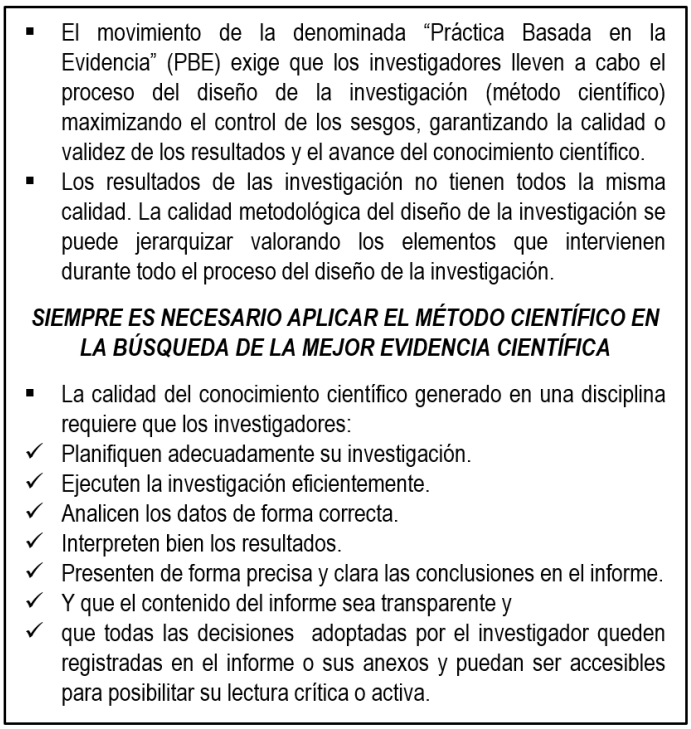


Figura 4. Método científico y Práctica Basada en la Evidencia

Por lo tanto, las decisiones de los y las profesionales exige cambiar el uso de las opiniones de experiencias personales de los expertos y expertas por 1) el uso de opiniones basadas en experiencias llevadas a cabo con la aplicación del método

científico y 2) por saber valorar la calidad de las evidencias o pruebas halladas para abordar posteriormente los problemas y obtener las conclusiones sobre intervención, diagnóstico, etiología o pronóstico de un fenómeno. Conviene, por ello, ser muy cauteloso cuando se valoran terapias sin pruebas de eficacia o efectividad realizadas con el método científico, ya que, si no se demuestra la validez científica de sus resultados, deberán ser clasificadas como pseudoterapias y consideradas como hechos no probados, sin el sello de calidad que otorga la ejecución escrupulosa de todos los pasos del proceso del método científico (proceso del diseño de investigación).

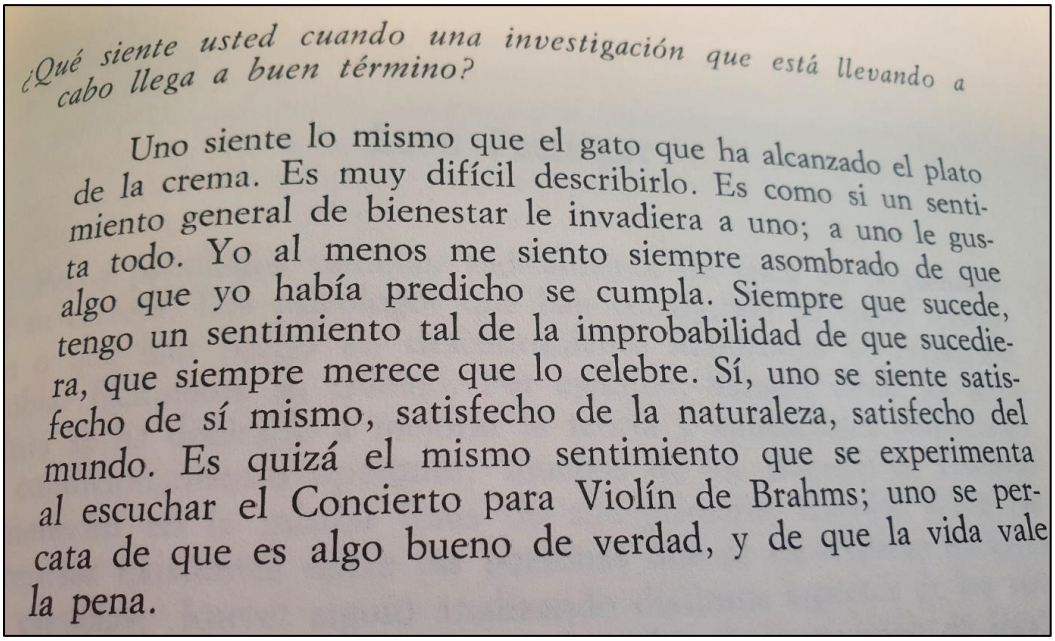
Como conclusion, destacar, en primer lugar, que comprender cómo se ha construido la investigación científica exige conocer en profundidad los elementos que determinan el método o metodología de investigación y el alcance de las interpretaciones causales o no de los hallazgos. Conocer los fundamentos de la metodología de investigación es un requisito para poder producir investigaciones con resultados válidos y acumular conocimiento científico. Y, también, es un requisito para realizar lecturas críticas o activa de los informes de investigación y artículos.

En segundo lugar, hay que tener siempre en cuenta que la calidad de la producción científica no siempre cumple los criterios de validez (Imai y cols., 2008; Onwuegbuzie, 2001). Los problemas de comprensión de la herramienta estadística y del diseño de investigación son una de las principales causas que invalidan los resultados de la investigación (Belia y cols., 2005; delMas y Liu, 2005; Falk, 1986). Productores y consumidores de la información científica necesitan la educación estadística y metodológica y necesitan conocer y aplicar los elementos que la nueva reforma estadística ha introducido desde la década de los noventa del siglo XX (Hancock y Mueller, 2010; Huck, 2007).

En tercer lugar, comprender el proceso de contraste de hipótesis estadísticas es fundamental para el investigador o investigadora, quien debe conocer y saber aplicar, y también para el lector o lectora de informes y artículos de naturaleza empírica, ya que es el procedimiento de análisis estadístico más utilizado descrito en dichos informes. Por ejemplo, saber interpretar los valores *p* de probabilidad, el tamaño del efecto y sus intervalos de confianza son competencias básicas del profesional de

aquellas disciplinas donde se aplica la inferencia estadística tradicional y la estimación de efectos.

Por último, resaltar la vivencia de los investigadores y las investigadoras cuando tras meses de trabajo llevan a cabo los análisis de sus datos y comprueban que aquella hipótesis o hipótesis que plantearon han sido comprobadas. Ha valido la pena todo el esfuerzo. Esas emociones y sentimientos están perfectamente reflejadas en las palabras de Eysenck tal y como son recogidas en la entrevista que Cohen (1977) le realiza y que se pueden leer en la imagen 3 en la edición en castellano del libro (1980, p. 153). En este punto totalmente de acuerdo con las palabras de Eysenck.



¿Qué siente usted cuando una investigación que está llevando a cabo llega a buen término?

Uno siente lo mismo que el gato que ha alcanzado el plato de la crema. Es muy difícil describirlo. Es como si un sentimiento general de bienestar le invadiera a uno; a uno le gusta todo. Yo al menos me siento siempre asombrado de que algo que yo había predicho se cumpla. Siempre que sucede, tengo un sentimiento tal de la improbabilidad de que sucediera, que siempre merece que lo celebre. Sí, uno se siente satisfecho de sí mismo, satisfecho de la naturaleza, satisfecho del mundo. Es quizá el mismo sentimiento que se experimenta al escuchar el Concierto para Violín de Brahms; uno se percata de que es algo bueno de verdad, y de que la vida vale la pena.

Imagen 3. Opinión de Eysenck sobre qué siente cuando su investigación llega a buen término (entrevista de Cohen, 1977). Concierto que se puede escuchar y disfrutar en el siguiente link: https://www.youtube.com/watch?v=_ioc6sdgugo&t=38s

Diseño de una investigación

Desde la perspectiva del ámbito metodológico, cuando se habla de diseño de investigación existe cierto consenso en caracterizarlo como un conjunto de actividades dirigidas a resolver un problema concreto (*necesidad de conocimiento*), incluyendo como elementos propios del diseño desde el planteamiento teórico del problema de investigación a partir del conocimiento previo y formulación conceptual de las hipótesis hasta su análisis estadístico, interpretación y discusión de resultados (ver Figura 5) (Arnau, 1981; Ato, 1991; Frías-Navarro, 2011; García y cols., 2006;

Kirk, 1995; Maxwell y Delaney, 2004; Maxwell y cols., 2018; Pascual-Llobell y cols., 1996).

En definitiva, partiendo del concepto de Psicología como una actividad científica, cuyo campo de estudio es el comportamiento humano, y del método de investigación como uno de los caminos válidos para contrastar los enunciados científicos con la realidad, el término "diseño de investigación" implica la planificación de todos los elementos necesarios para contrastar de forma correcta (sin sesgo) la o las hipótesis del estudio con dicho método o alcanzar los objetivos propuestos en el estudio.

Desde el ámbito de la investigación empírica cuantitativa, la formulación de las hipótesis teóricas es, por supuesto, inherente al proceso de investigación dando sentido a la ejecución del experimento o estudio como un procedimiento de contrastación de hipótesis estadísticas y comparación de modelos explicativos, cuyo fin es poder determinar si las hipótesis teóricas se confirman o no (siempre con un riesgo de error en la decisión estadística).

Es decir, el fin último del contraste de hipótesis estadísticas es aportar evidencia empírica sobre las relaciones teóricas entre las variables. Por lo tanto, trasladar una idea o un problema de investigación a un plan de trabajo requiere comprender los fundamentos del diseño de investigación y sus técnicas de análisis.

Por lo tanto, cuando el investigador o investigadora plantea una 'necesidad de conocimiento' (problema de investigación) vinculada posteriormente a una hipótesis de investigación se inicia el proceso de diseño de investigación (ver Figura 5).

El proceso de diseño de investigación se inicia con la *necesidad de conocimiento* que tiene el investigador o investigadora sobre un determinado constructo o problema psicológico (objetivo del estudio, por qué se lleva a cabo la investigación) y no termina hasta lograr cierto *conocimiento* válido sobre la realidad del fenómeno estudiado (ver Figura 5a).

Después de una etapa previa, donde se generan ideas sobre el fenómeno objeto de estudio, se concreta una "necesidad de conocimiento" vinculada al problema de investigación. Es decir, el investigador o investigadora define lo que va a investigar y restringe el campo a una pregunta concreta, delimitando las variables y operacionalizando la pregunta que hay que resolver. Posteriormente se pone en

marcha la técnica estadística y el análisis metodológico que requiera el problema de investigación planteado (ver Figura 5b).

Finalmente la valoración y la interpretación de los resultados obtenidos permiten elaborar el nuevo conocimiento, continuando de este modo el proceso cíclico de formación del saber científico (ver Figura 5c).

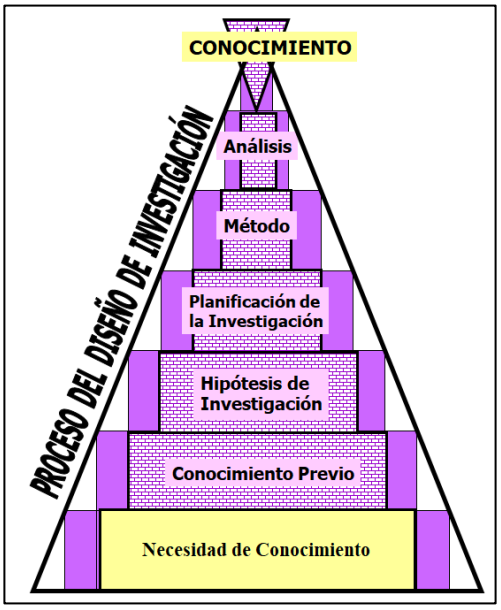


Figura a

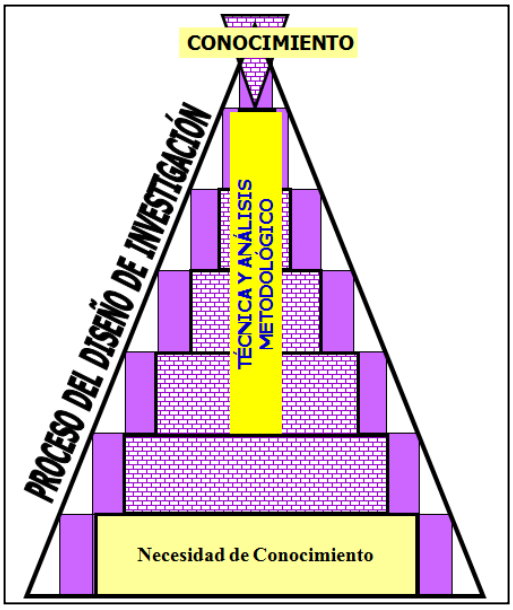


Figura b

Figura c

Figura 5. Pirámide del proceso del diseño de investigación

La pirámide invertida que aparece en las representaciones de la Figura 5 en la parte superior donde se ubica el 'Conocimiento' adquirido señala la dirección hacia una nueva pirámide (un nuevo estudio con su proceso de diseño de investigación) que será construida sobre la base del conocimiento previo obtenido con la investigación que se ha finalizado junto con los hallazgos ya publicados en el campo científico. De este modo, el saber científico se va acumulando con estudios que aplican el método científico y va formando el conocimiento científico sobre un determinado fenómeno o constructo.

Por lo tanto, la investigación científica implica un proceso de análisis de la realidad con el objetivo de dar respuesta a la necesidad de conocimiento planteada en una determinada investigación.

Conviene tener siempre muy presente que la formulación de las hipótesis teóricas, la recogida de los datos y el planteamiento de su análisis estadístico son los elementos que guían la elección del diseño de investigación más apropiado. Por lo tanto, es conveniente dedicar tiempo y esfuerzo a los pasos previos a la recogida de los datos (es decir, a la fase de planificación del estudio), ya que una vez recogidos y ejecutado el estudio poco se podrá hacer para mejorar el diseño del estudio.

La contrastación con la realidad supone poner en marcha técnicas de comprobación de hipótesis estadísticas que unidas a la metodología de investigación seleccionada por el investigador o investigadora, como la más apropiada para dar respuesta a los objetivos del estudio y sus hipótesis, forman los elementos esenciales del proceso de diseño de una investigación. Se trata de alcanzar una explicación válida de los resultados obtenidos en el estudio, comparando lo predicho por la teoría (hipótesis) y lo manifestado en la realidad (datos obtenidos), valorando y controlando en el diseño aquellas variables que podrían confundir los resultados y amenzar la calidad de las pruebas o evidencias halladas.

Como posteriormente se detallará, con la denominada metodología experimental y la cuasi-experimental el propósito es potenciar el efecto de aquella variable que se ha manipulado o *variable independiente de tratamiento* para estudiar su efecto en una determinada variable medida objeto de observación o *variable dependiente* y controlar la posible influencia de *variables extrañas* que no forman parte de las hipótesis y cuyo efecto sistemático (sesgo) sobre la variable dependiente confundiría

la interpretación de los resultados. Cuando se trabaja con una metodología no experimental el propósito es potenciar la magnitud de la relación entre las variables que se van a estudiar, y aquí también, por supuesto, evitando / controlando la presencia de variables contaminadoras o extrañas (también denominadas 'terceras variables' que están unidas al sesgo o a la contaminación de los datos).

Reforma estadística y Práctica Basada en la Evidencia

Conocer el tamaño del efecto y la precisión de su estimación puntual con el intervalo de confianza permite realizar interpretaciones sobre la magnitud de las diferencias detectadas y no sólo sobre el grado de significación estadística. Las políticas editoriales de prácticamente todas las revistas científicas se han sumado a la denominada 'reforma estadística', modificando las instrucciones a los autores y autoras para el envío de sus manuscritos, reclamando acompañar el valor *p* de probabilidad asociado al resultado de la prueba estadística aplicada con el valor del tamaño del efecto y su intervalo de confianza, si es posible, y realizando una interpretación conjunta de todos esos resultados en el informe de investigación.

El movimiento de la Medicina Basada en la Evidencia (MBE) y, en general, el de la Práctica Basada en la Evidencia (PBE) tiene un punto de arranque en la toma de conciencia de los problemas vinculados al procedimiento de significación de la hipótesis nula como único medio para producir descubrimientos, enfatizando el uso de las revisiones sistemáticas y la estimación del tamaño del efecto (Sackett y cols., 2000). Además, el auge del movimiento de la Práctica Basada en la Evidencia (Frías-Navarro y Pascual-Llobell, 2003) centra la atención en obtener las mejores pruebas o evidencia facilitando así la interpretación de los hallazgos dentro del contexto específico de efectos al que pertenece la hipótesis de investigación. Es decir, un tamaño del efecto pequeño en términos estadísticos podría ser grande en un sentido sustantivo y en determinado contexto, del mismo modo que un tamaño del efecto grande en términos estadísticos podría ser pequeño en el sentido sustantivo, clínico o de utilidad para el profesional o investigador (Pek y Flora, 2018; Rosenthal, 1994; Rosnow y Rosenthal, 1989).

La reforma estadística cambia el punto de mira desde "cómo es de probable o improbable el resultado muestral" hacia dos cuestiones principalmente: "cómo es de grande el tamaño del efecto detectado" y "si se puede replicar". Es decir, hay que

‘evaluar’ el valor del tamaño del efecto estimado y su utilidad (su grado de importancia práctica, clínica o sustantiva) y para ello es necesario considerar el contexto concreto de la investigación y comparar de forma explícita y directa los resultados y efectos de un estudio con los obtenidos en el área de investigación donde se enmarca el trabajo.

En definitiva, se trata de contextualizar los efectos en el campo propio de cada fenómeno psicológico y no generalizar sobre el uso de sus magnitudes pequeñas, medias o grandes para todos los constructos o ámbitos de investigación (Cumming y Finch, 2005; Frías-Navarro y cols., 2000; Wilkinson & the Task Force on Statistical Inference, 1999). Además, la replicabilidad del efecto supone evaluar cómo de estables son los efectos en la literatura revisada y por lo tanto evaluar en qué medida son efectos directamente comparables (Pascual-Llobell y cols., 2000).

En definitiva, el nuevo comportamiento del investigador o investigadora supone desarrollar el denominado “pensamiento meta-analítico” (Cumming y Finch, 2001). En este sentido, los investigadores e investigadoras son quienes deben planificar sus hipótesis en función de los efectos previos detectados en la literatura o en el conocimiento teórico que tengan e interpretar sus hallazgos dentro de dicho contexto de efectos. El ‘pensamiento dicotómico’ (vinculado a la comprobación tradicional de la hipótesis nula: mantenerla o rechazarla) es reemplazado por un ‘pensamiento meta-analítico’ (valoración de la magnitud del tamaño del efecto y su intervalo de confianza y contextualizar los efectos) que facilita llegar a una interpretación sustantiva de los hallazgos integrada en el contexto de efectos que hasta el momento se han obtenido en las investigaciones sobre una temática concreta. Dicha interpretación sustantiva deberá ser, además, valorada como útil (grado de utilidad) por el o la profesional desde su juicio clínico.

La elaboración de estudios de meta-análisis y sobre todo el uso de sus resultados para planificar el tamaño del efecto de un estudio primario y contextualizar los efectos, adquieren relevancia entre las indicaciones de la reforma estadística (Sánchez Meca y Botella, 2010). El cambio implica que los investigadores y las investigadoras planteen su diseño estadístico en términos de estimación de efectos y no solo para lograr la significación estadística. Este cambio no está siendo fácil para los investigadores y las investigadoras tal y como demuestran los estudios sobre el

uso del tamaño del efecto y sus intervalos de confianza en las publicaciones científicas (Frías-Navarro, 2011).

La edición del Manual de la *American Psychological Association* de 2010 mantuvo su énfasis en la denominada reforma estadística, destacando el uso de los tamaños del efecto y sus intervalos de confianza y las técnicas bayesianas tratando de minimizar la confianza excesiva que los investigadores tienen sobre las pruebas de significación estadística y las decisiones dicotómicas apoyadas en los valores *p* de probabilidad (Hoekstra y cols., 2006).

El grupo de trabajo de inferencia estadística de la American Psychological Association (Wilkinson & APA Task Force on Statistical Inference, 1999) y las cuatro últimas ediciones del Manual de la APA (American Psychological Association, 2001, 2010, 2020) señalan de forma destacada que los investigadores e investigadoras deben estimar el valor del tamaño del efecto y su intervalo de confianza junto al valor *p* de probabilidad, siempre que sea posible.

En definitiva, el Manual de la American Psychological Association recomienda a los investigadores e investigadoras estimar y tomar en consideración el tamaño del efecto en el ámbito teórico y en el aplicado. Afirma el Manual que una característica esencial de la buena investigación es la de interpretar los tamaños del efecto en relación con los efectos previamente estimados e informados por la investigación que ya se ha realizado dentro del campo de estudio (pensamiento meta-analítico).

Las nuevas recomendaciones formuladas en la reforma estadística tardaron en consolidarse y cambiar la conducta de los investigadores e investigadoras (Vacha-Haase y cols., 2000). Actualmente es bastante común que la redacción de los resultados estadísticos de las pruebas de significación estadística se acompañen del valor del tamaño del efecto. Aún no es muy común que ese tamaño del efecto vaya acompañado de su intervalo de confianza, pero se va avanzando.

Sin embargo, permanece el debate de la falta de comprensión de las implicaciones que las pruebas de contraste estadístico suponen, destacando la presencia de falacias estadísticas que debilitan las conclusiones de los estudios y la calidad de sus pruebas o resultados.

Muy probablemente los investigadores e investigadoras aún no han reflexionado sobre la cuestión de que “seguramente Dios ama al .06 (nivel de significación estadística) tanto como al .05” (Rosnow y Rosenthal, 1989, p.1277).

Además, los programas estadísticos, como el SPSS hasta la versión 26, no han incorporado la estimación directa de los nuevos estadísticos como la *d* de Cohen o los intervalos de confianza de forma generalizada en todos las estimaciones puntuales de los estadísticos (por ejemplo, cuando se estima un coeficiente de correlación no aporta su intervalo de confianza), dificultando con ello el cambio de la práctica estadística. Finalmente, en la versión 27 del SPSS ya se puede computar el tamaño del efecto de la *d* de Cohen. Otros programas como Epidat (Santiago y cols., 2010) y los programas estadísticos gratuitos JASP o JAMOVI ya han incorporado la estimación del tamaño del efecto *d* de Cohen junto a otros índices y lo han hecho en gran parte de los estadísticos, ofreciendo en algunos casos la estimación de su intervalo de confianza.

La reforma estadística también destaca la capacidad de leer la literatura científica de una forma crítica o activa como una competencia esencial del consumidor o consumidora de informes de investigación o artículos que exige tener conocimientos de metodología para valorar los trabajos con investigación empírica. La literatura es el repositorio escrito del conocimiento. Su revisión ayuda a perfilar de manera más elaborada la hipótesis de investigación, destacando que una buena hipótesis debe ser factible o abordable, interesante, novedosa, ética y relevante para generar conocimiento científico.

El propósito de la lectura crítica o activa es doble:

1. Por una parte, para comprobar si los métodos de investigación aplicados producen información útil.
2. Para analizar si las conclusiones de los trabajos de investigación realmente pueden ser obtenidas a partir de los hallazgos obtenidos. Es decir, se trata de valorar la calidad o validez de los hallazgos tal y como han sido obtenidos con el diseño de investigación aplicado en el estudio que se lee.

Capítulo 3. Variables del estudio

Dolores Frías-Navarro*

Marcos Pascual-Soler**

José Berríos-Riquelme***

*Universidad de Valencia

**ESIC Business & Marketing School, España

***Universidad de Tarapacá (Chile)

Índice

Citar el capítulo como:

Frías-Navarro, D., Pascual-Soler, M., y Berríos-Riquelme, J. (2021). Variables del estudio. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

Este capítulo, “Variables del estudio”, tiene como principal objetivo describir las características que identifican a las **variables de una investigación**, ya que es un punto esencial para identificar las hipótesis del estudio y el control del sesgo que se ha planificado. Además, se resumen las propiedades y diferentes entre las variables en función de su nivel de medida.

Una variable es cualquier entidad que puede tomar diferentes valores. Es decir, una variable es una característica de un sujeto o de un objeto que puede tener diferentes valores y por ello es lo opuesto a una constante. La conducta de un fenómeno suele estar provocada por multitud de factores o variables, definiéndose en el diseño de una investigación un reducido número de posibles variables que provocan su comportamiento, siendo el estudio de dichas variables el objetivo de la investigación.

En muchos campos, como la Psicología, no se mide una característica física, sino un concepto teórico inobservable denominado constructo. Por ejemplo, el constructo de conocimiento, de aprendizaje, de amor, de empatía, de agresividad, de depresión, de racismo, de xenofobia, de homofobia, de felicidad, entre otros tantos posibles. Estos constructos deben ser operacionalizados o ‘atrapados’ en variables que puedan ser medidas con instrumentos adecuados que tengan adecuadas propiedades psicométricas de fiabilidad y evidencias de validez en la muestra de participantes que colaboran el estudio.

Es importante tener en cuenta que es imposible medir un constructo sin cierta cantidad de error. Ese error puede ser debido al propio sujeto objeto de medición (por ejemplo, lee mal la pregunta y se equivoca), al contexto de evaluación (hace mucho frío y eso afecta a las respuestas del sujeto) o al mismo procedimiento utilizado para la medición (por ejemplo, las puntuaciones de la muestra tienen escasa fiabilidad). No hay una medición exacta del constructo. Y lo que se desea es que el error de medición sea el menor posible. Por lo tanto, una buena medición del constructo debe ser fiable y también debe ser válida.

La fiabilidad se refiere a la consistencia en las mediciones de un constructo. Si un constructo se mide varias veces con un mismo instrumento es esperable que los resultados de las mediciones sean muy similares si la medición es fiable. Por supuesto, a menos que haya ocurrido un hecho que produzca un efecto que

modifique al constructo objeto de estudio. En otras ocasiones, se habla de fiabilidad entre los jueces para comprobar que el registro de los observadores de una determinada conducta de los sujetos no dependan del evaluador y observador. Es decir, se desea una alta fiabilidad o acuerdo entre las respuestas de los jueces.

La fiabilidad de la medida es esencial para poder comparar una medida con otra. Así, una medida poco fiable nunca podrá tener una relación estadística sólida con ninguna otra medida. De ahí que los investigadores y las investigadoras comprueben siempre la fiabilidad de la medición con cada muestra de participantes como paso fundamental previo al análisis de datos para pasar a comprobar posteriormente los efectos o las relaciones entre las variables. La consistencia interna de los ítems que forman una puntuación total se suele medir con el alfa de Cronbach o con el coeficiente omega de McDonald que es más robusto y no exige que los errores no estén correlacionados (Frías-Navarro, 2021; Viladrich y cols., 2017).

La fiabilidad de la medición es importante, pero por sí sola no es suficiente: por ejemplo, se podría crear una medida de personalidad perfectamente fiable al volver a codificar en un segundo pase cada respuesta con el mismo número u opción de respuesta, independientemente de cómo responda realmente la persona. Por lo tanto, también se quiere que las medidas demuestren evidencias válidas, es decir, hay que asegurarse de que realmente se está midiendo el constructo que se cree que está midiendo (en el ejemplo se trataría de la personaldiad).

Las evidencias de validez permiten realizar inferencias e interpretaciones correctas de las puntuaciones de un test o escala y establecer su relación con el constructo/variable que se trata de medir (Gómez-Benito e Hidalgo, 2015). Hay muchos tipos de evidencias de validez que se discuten comúnmente. Por ejemplo, la validez aparente, la validez de constructo (validez convergente y validez discriminante) o la validez predictiva. Y todas ellas son pruebas de validez como concepto unitario que debe tener la medición, entendida como el grado en que la evidencia apoya las inferencias que se hacen a partir de las puntuaciones del test o escala. La validez no se resume en un solo indicador o valor numérico como la consistencia de las puntuaciones con el valor de alfa de Cronbach o la omega de McDonald. Conviene remarcar y tener en cuenta que lo que se valida no es el test en sí mismo, sino las puntuaciones del test obtenidas con una muestra concreta, por lo

tanto, la pregunta que se trata de responder: ¿es válido el uso o la interpretación de las puntuaciones de ese test con esta muestra concreta del estudio? De ahí la necesidad de comprobar ese uso válido en cada muestra.

Clasificación de las variables del estudio

Las variables del diseño de investigación pueden ser clasificadas por diferentes criterios, destacando:

1) El "criterio metodológico", que es uno de los más utilizados e inicialmente unido a los diseños con una metodología propiamente experimental, y se refiere a la función que tiene cada variable en el diseño de la investigación (variable independiente, variable dependiente y variables controladas y extrañas).

2) El "criterio manipulativo" de la variable de intervención (variable independiente manipulada o variable independiente asignada).

3) El "criterio del nivel de medición" de la variable o escala de medida (nominal, ordinal, de intervalo y de razón).

4) El "criterio estadístico" que está relacionado con la naturaleza de los aspectos a medir (variables cualitativas / variables cuantitativas (discretas / variables continuas).

Criterio metodológico

Utilizando un *criterio metodológico*, vinculado a la metodología experimental, se distinguen tres tipos de variables en una investigación:

1. la o las variables independientes (VI): son las variables de tratamiento, variables de grupos o los 'factores' del diseño de investigación. Se representan como A, B, C … También se conocen como variables predictoras.
2. la o las variables dependientes (VD): son las variables que se miden en el estudio (el resultado). Se representan en el diseño por Y_1, Y_2 … También se conocen como variables predichas.
3. la o las variables extrañas a los objetivos o hipótesis de investigación (VE): son las variables que se controlan en el diseño de investigación (variables controladas por ejemplo por la técnica de la aleatorización,

la eliminación o la constancia) y también pueden ser controladas con el tipo de diseño elegido, ya que pueden actuar como factores del diseño (fuentes de varianza) cuyo objetivo es controlar el efecto de posibles variables extrañas. Se trataría de factores o variables que no son objeto de estudio en la hipótesis teórica planteada, ya que su función en el diseño es de control del efecto de 'terceras variables' (por ejemplo, el diseño de bloques o el diseño con variables covariadas) y no la de aportar una explicación teórica del fenómeno.

El papel metodológico que tiene cada tipo de variable en el diseño del estudio es asignado por el investigador o investigadora en función de sus intereses e hipótesis concretas. Así, la ansiedad puede ser una variable dependiente si se utiliza como variable medida o resultado, puede ser una variable independiente categorizada en diferentes condiciones (por ejemplo, ansiedad baja, ansiedad media y ansiedad alta) o podría ser una variable extraña que contaminaría los resultados si no se controla con alguna técnica (por ejemplo, se puede utilizar como factor de bloqueo o como una variable covariada de ajuste estadístico).

Conviene tener muy presente que dado que la Psicología trabaja con constructos o fenómenos no observables como la depresión, la ansiedad, la autoestima, la anorexia, el prejuicio, el racismo, la homofobia…, es necesario valorar las variables del estudio en su componente de constructo y en su componente operacionalizado, tal y como es tratado de forma empírica en la investigación. Por ejemplo, si se desea estudiar el nivel de ansiedad que produce la cercanía de una serpiente será necesario que en el estudio se operacionalice la medida del constructo de ansiedad, por ejemplo, por la tasa cardíaca (variable dependiente operacionalizada), y la cercanía al estímulo se puede medir en centímetros y crear tres grupos con diferentes distancias (variable independiente operacionalizada).

Variable independiente

Las variables independientes o factores del diseño de investigación (simbolizados por letras mayúsculas como factor A, factor B, factor C…) representan la intervención o tratamiento (o el grupo de comparación), es decir, los grupos que forman parte del diseño de investigación o, también, pueden representar las fases de

medición de una variable si se trata de un diseño de investigación con medidas repetidas. El investigador o investigadora / lector o lectora debe describir / detectar la variable constructo y la variable tal y como se ha operacionalizado en el diseño de esa investigación para iniciar la planificación del estudio o para iniciar la lectura activa de un informe de investigación.

En el diseño de investigación, cada variable independiente está formada por al menos:

1- Dos condiciones o grupos, en el caso del diseño de comparación de grupos más simple con una sola medición por sujeto (conocido como '*diseño-entregrupo*'). Las condiciones o grupos se simbolizan por las letras de los factores, pero en minúsculas, por ejemplo a_1 y a_2, b_1 y b_2, c_1 y c_2…. En el caso de un diseño con a_1 y a_2, b_1 y b_2, c_1 y c_2 se trataría de un estudio con tres factores y cada uno de ellos tendría dos condiciones.

2- O puede tratarse de una variable independiente con al menos dos mediciones por sujeto, en el caso de los '*diseños de medidas repetidas o diseños intra-sujetos*' más simples (a_1 y a_2), actuando cada medición como una condición del factor (es decir, como medición primera a_1 y medición segunda a_2).

3- El diseño también puede estar formado, al menos, por una variable o factor 'entre-sujetos' y por, al menos, una variable o factor de medidas repetidas o intra-sujetos, formando los denominados '*diseños mixtos o de medidas parcialmente repetidas*'. Este tipo de diseños tendrá fuentes de varianza o variabilidad entre-grupos, fuentes de varianza intra-sujetos y un término de error (variabilidad intra-celdilla o intra-grupo) diferente para cada una de esas dos fuentes de varianza (entre e intra).

En función del número de variables independientes o factores que tiene el diseño de investigación, los diseños se pueden clasificar en:

1. diseños '*unifactoriales'*, diseños con una sola variable independiente y

2. diseños '*factoriales'*, diseños con más de una variable independiente o factor.

Teniendo en cuenta el número de observaciones que se recogen en cada condición de la variable independiente, los diseños entre-sujetos se pueden clasificar,

además, en '*diseños ortogonales*' cuando el número de observaciones es el mismo en los dos grupos ($n_1 = n_2$) y '*diseños no ortogonales*' ($n_1 \neq n_2$) cuando el número de observaciones es diferente. Lógicamente, en los diseños intra-sujetos se recoge el mismo número de observaciones en la medición 1 y en la medición 2 para poder ejecutar los análisis de medidas repetidas.

Variable independiente: manipulada / no manipulada

Las variables independientes del experimento pueden ser clasificadas utilizando diferentes criterios como el criterio manipulativo que distingue entre:

1- Variables independientes *'activas*' (variables manipuladas), cuyos niveles o condiciones son provocados por el investigador o investigadora, como el tipo de fármaco o el carácter de las instrucciones dadas a los sujetos. Es decir, sus condiciones o niveles son seleccionados por el investigador o investigadora en función de los objetivos del estudio.

2- Variables independientes '*asignadas*' (variables no manipuladas, también conocidas como variable observada, variable seleccionada o variable de clasificación) cuyas condiciones reflejan diferentes aspectos de la variable que no son manipuladas por el investigador o investigadora en el estudio, sino que son medidas o registradas tal y como han sucedido en el sujeto, como por ejemplo la edad, el sexo, la profesión o el nivel de glucosa en sangre.

A continuación, se describen los dos tipos de variables independientes con más detalle.

Variables independientes manipuladas / activas

Las variables independientes pueden denominarse de dos maneras. Primeramente serían variables independientes "activas", esta situación ocurre cuando son variables manipuladas, es decir, son variables cuyas condiciones son creadas deliberadamente por la persona a cargo de la investigación. Se trata de variables cuyos efectos están bajo el control del investigador o investigadora, es decir, puede ponerlas o quitarlas del diseño para ver sus efectos sobre las variables dependientes.

La segunda opción se denomina variables independientes ‘asignadas’, ya que el investigador o investigadora no controla sus efectos, solamente selecciona sus condiciones o niveles tal y como se encuentran en la realidad de los fenómenos (por ejemplo, el sexo asignado al nacer).

Una variable independiente manipulada puede ser un tratamiento farmacológico, una terapia psicológica, un programa de intervención social o una tarea experimental que se crea para observar cómo reacciona el sujeto ante un determinado estímulo. En definitiva, supone introducir una variable en la vida del sujeto para ver su efecto o resultado sobre sus respuestas o variable medida (variable dependiente). Es decir, esa variable manipulada debe provocar un cambio en la vida de los participantes o sujetos del estudio.

Por lo tanto, los efectos de las variables independientes activas están bajo el control del investigador o investigadora que introduce deliberadamente diferentes condiciones de tratamiento para observar cómo afectan a la variable dependiente, es decir, a la variable medida. Esto permite la generación controlada de datos, ya que la variabilidad de la variable independiente de tratamiento está determinada por el investigador o investigadora que puede hacer variar los valores de la variable independiente o decidir qué valores adoptará e incluso decidir cuándo la introduce y cuándo la retira del estudio.

Variables independientes no manipuladas / asignadas

Las variables independientes “asignadas” son variables no manipuladas directamente por el investigador o investigadora sino que sus niveles o condiciones son seleccionados por el investigador o investigadora para observar su relación con otras variables. Aquí el investigador o investigadora selecciona ciertos niveles de una variable según sus intereses de investigación, pero no puede manipularlos.

Cuando no existe una intervención activa (manipulación) de la variable independiente (metodología no experimental) por parte del investigador o investigadora, sino que se trata de un rasgo o de una exposición a una condición que ya posee el sujeto (sin manipulación) y que es seleccionada por el investigador o investigadora como objetivo de estudio, entonces la variable independiente se conoce como la variable *predictora* y la variable medida como la variable *predicha*.

En definitiva, los niveles de la variable independiente asignada los selecciona el investigador o investigadora (por cuestiones teóricas o de interés teórico) del conjunto de niveles que ya posee la variable predictora o factor. Por ejemplo, el sexo asignado al nacer, el nivel educativo, la edad, la profesión, el estatus económico, la estructura familiar… son variables predictoras asignadas que son seleccionadas por el investigador o investigadora debido a su interés teórico.

Variable dependiente

La variable dependiente es la variable medida objeto de observación en el diseño de investigación; son los resultados del estudio. Se representa con la letra Y. Es el resultado que se mide en el estudio. Es el desenlace del experimento.

Por ejemplo, la variable dependiente puede ser el grado de eficacia del tratamiento medido con el nivel de colesterol en la sangre, medido con una escala de ansiedad o ~~medido~~ con un cuestionario de auto-eficacia percibida después de recibir un determinado tratamiento o intervención. O quizás, puede ser el nivel de cortisol secretado que aparece en la orina después de someter al sujeto a una situación de estrés como puede ser un examen tal y como se desarrolla en el trabajo de García de la Banda y cols. (2004).

El investigador o investigadora debe describir en el informe o artículo todas las variables constructo (independientes, dependientes y extrañas controladas) y su operacionalización en el diseño de esa investigación. El lector o lectora debe detectar o reconocer las variables constructo y su operacionalización en el estudio para poder llevar a cabo una lectura activa o crítica y comprender los elementos implicados en el proceso de diseño.

En función del número de variables dependientes o variables medidas que tiene el diseño de investigación se pueden clasificar los diseños en:

1. diseños ‘univariados’ que son diseños con una sola variable dependiente utilizada en el análisis y

2. diseños ‘multivariados’, es decir, diseños con más de una variable dependiente medida que se utilizan de forma simultánea en el análisis de los datos.

Los diseños unifactoriales y los factoriales pueden ser univariados o multivariados.

Variable extraña

Las variables extrañas a los objetivos de la investigación, son aquellos otros factores que podrían tener algún impacto importante en los resultados del estudio. Es decir, es cualquier otra variable que también podría afectar a la variable dependiente medida o interactuar con la variable independiente, pero que no son objeto de estudio en la hipótesis de investigación.

Las variables extrañas incluyen:

1- Las variables 'aleatorias' cuyo efecto no es sistemático y representan el error aleatorio que está presente en cualquier puntuación del diseño de investigación.

2- Las denominadas variables 'perturbadoras' (terceras variables) con efecto sistemático que contamina a los datos y necesariamente deben ser controladas para potenciar las evidencias de validez o la calidad de los resultados.

Las "variables extrañas aleatorias" (el error) están siempre presentes en el diseño de investigación y es objetivo de la planificación del estudio minimizar el error aleatorio para crear grupos de comparación lo más equivalentes posibles en todas las variables excepto en aquella variable cuyo efecto se desea analizar en el estudio.

Las "variables perturbadoras" son las variables extrañas no controladas en el diseño de investigación y su presencia podría distorsionar la relación entre la variable independiente y la variable dependiente (ambas variables forman la hipótesis de investigación) y podrían provocar desequilibrios en los diferentes grupos o condiciones de intervención, creando grupos no equivalentes y dando lugar a un sesgo sistemático en los datos. En este caso, la validez interna del diseño estará claramente cuestionada por la presencia de una 'tercera variable'.

Por lo tanto, la clave de las variables extrañas está en si el sesgo que provocan es aleatorio (error no sistemático que hay que minimizar en el diseño) o es error sistemático (error sistemático que siempre hay que controlar con alguna técnica).

Es necesario que esas variables perturbadoras sean controladas (y pasarían a denominarse "variables sistemáticas controladas") con el diseño de investigación, ya

sea con la técnica de la aleatorización (variables extrañas aleatorizadas), con la técnica de la eliminación de su efecto si es posible (mantener constante su efecto sobre todos los datos), con la técnica del control matemático de su efecto mediante el diseño (diseño con variables covariadas donde se realiza un ajuste matemático de los datos) o bien incluyéndolas en el modelo de investigación como un factor más en la ecuación estructural para controlar su efecto por constancia (variables extrañas controladas por bloqueo) tal y como sucede en el diseño de bloques.

Siguiendo el planteamiento de una tipología de las variables que forman parte del plan de investigación, Kish (1975) distingue dos clases de variables: las *variables explicativas o experimentales* y las *variables extrañas* (ver Figura 6).

1- Las "variables explicativas" o experimentales constituyen el objetivo del estudio y son la o las variables independientes (o variables predictoras, variables estímulo, variables de tratamiento) y la o las variables dependientes (variables pronosticadas o predichas, variables de respuesta, variables criterio) que forman la hipótesis de investigación. Como el propio autor señala, estas variables surgen del conocimiento y la comprensión del campo de estudio.

2- Además, señala el autor, existen otras tres clases de variables que son "extrañas a los objetivos de la investigación": las variables controladas, las variables perturbadoras y las variables aleatorias. Se trata de otras fuentes extrañas de variación que el investigador o investigadora debe identificar para separar su variabilidad de la producida por las variables explicativas.

Las 'variables controladas' son variables extrañas a los objetivos de la investigación que se controlan con el diseño de investigación con el objeto de disminuir los efectos del error aleatorio o eliminar los efectos de sesgo de las variables perturbadoras o para disminuir ambas situaciones.

Las 'variables perturbadoras', en cambio, son variables extrañas no controladas en el diseño de investigación y, por lo tanto, su efecto se puede confundir con las variables explicativas y provocar un efecto de confundido, interfiriendo la relación entre la variable independiente y la dependiente.

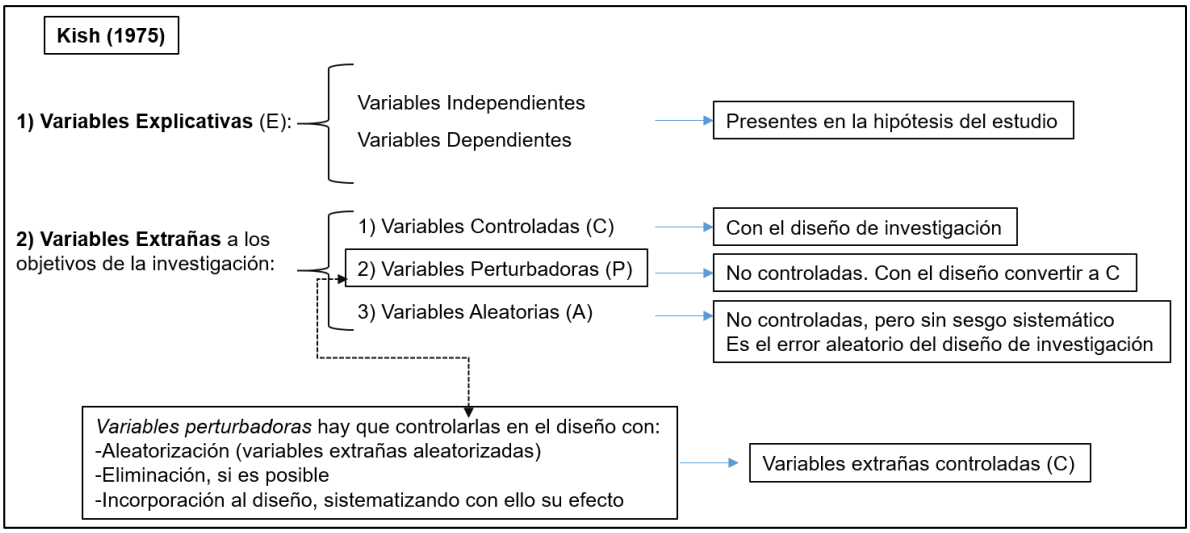


Figura 6. Variables de un diseño de investigación

El objetivo de un eficiente diseño de investigación es controlar las variables perturbadoras, convirtiéndolas en variables controladas o, si es posible, convertirlas en variables aleatorizadas. Es importante reflexionar y detectar este tipo de variables (por ejemplo, tras efectuar la revisión del conocimiento previo) en la fase de planificación del diseño de investigación para que sea posible su control, tomando las decisiones antes de recoger los datos del estudio.

Por último, las 'variables aleatorias' son variables extrañas no controladas que se tratan en el diseño como errores del muestreo aleatorio (o de la asignación aleatoria del tratamiento) y se considera que no ejercen un sesgo sistemático sobre la relación entre la variable independiente y dependiente y el diseño de investigación debe minimizar la presencia de error aleatorio. Se trata del denominado error aleatorio que siempre está presente en la medida de la variable dependiente.

En resumen, la variable independiente y la dependiente son las denominadas 'variables explicativas' del diseño de investigación y forman la hipótesis de investigación que describe el posible comportamiento del fenómeno investigado. Las variables controladas no se detallan en la hipótesis del estudio sino que se incorporan o se redactan cuando el investigador o investigadora describe el diseño de su investigación. Es muy importante leer (escribir) con detalle la hipótesis del estudio y el apartado de diseño de la investigación para poder detectar qué función tiene cada variable en el diseño de ese estudio.

La investigación puede plantear una “relación causal” entre las variables del estudio donde el resultado en la variable dependiente está causado por la variable independiente manipulada (metodología experimental) o bien puede tratarse de un estudio de “asociación” o de análisis de la magnitud de la relación entre las variables sin ningún tipo de interpretación causal entre las variables (metodología no experimental). Los resultados de un estudio con una metodología cuasi-experimental deben ser interpretados con cautela, ya que no utilizan la asignación aleatoria del tratamiento y es necesario reflexionar y controlar posibles variables extrañas para poder interpretar las relaciones como causales, pero siempre destacando la limitación del tipo de metodología.

En definitiva, las variables perturbadoras con efecto sistemático deben convertirse siempre en variables extrañas controladas dentro del diseño de investigación para garantizar la validez de los hallazgos. El sesgo sistemático provocado por el efecto de las variables extrañas no controladas es el enemigo número uno de la validez o calidad de los resultados. Si se conoce el efecto sistemático de la o las variables extrañas y no se puede eliminar, emparejar o aleatorizar de forma directa, entonces se debe incorporar al diseño como un factor más del estudio para controlar su presencia, ya que, en caso contrario, distorsionaría la verdadera relación entre las variables explicativas que forman la hipótesis de investigación (independiente y dependiente), confundiendo los datos obtenidos. Se trata de mantener constante el efecto de las variables extrañas en todos los grupos del diseño tal y como sucede en el diseño de bloques o tras aplicar la técnica de la asignación aleatoria o residualizar estadísticamente la variabilidad que provocan utilizando en este caso diseños con variables covariadas (ANCOVA).

Criterio estadístico y nivel de medición

Por otra parte, el criterio estadístico clasifica a las variables por el tipo de unidad de medida y distingue entre *variables cualitativas* y *variables cuantitativas* y entre *variables discretas* y *variables continuas*.

Dependiendo de las relaciones que se puedan establecer entre los valores o categorias de una variable, se obtienen distintos niveles de medida y diferentes escalas de medida y, por ello, se puede hablar de distintos tipos de variables.

Según el nivel de medición y el tipo de escala de medida, las variables se denominan: variables nominales, variables ordinales, variables de intervalo y variables de razón. Las escalas nominales y las ordinales miden variables de naturaleza cualitativa, mientras que las escalas de intervalo y de razón miden variables de naturaleza cuantitativa (numéricas).

Las variables cuantitativas o numéricas pueden ser discretas y continuas.

Variables cualitativas

Las variables cualitativas se miden en términos de atributos o cualidades de las variables que no representan a valores numéricos, como por ejemplo tipos de suplementos vitamínicos diferentes, color de los ojos... Sus valores o niveles no se asocian de forma natural a un número.

Con los valores de las variables cualitativas no se pueden realizar operaciones aritméticas. Por ejemplo, sexo, tipo de ocupación laboral, estado civil, lugar de nacimiento...

Las variables nominales y ordinales son variables cualitativas.

En cambio, las variables cuantitativas son aquellas variables cuyos valores se pueden expresar numéricamente, como por ejemplo la altura, la edad, los ingresos económicos, el peso, número de calzado, temperatura, rendimiento… y con ellas se pueden realizar cálculos como la media.

Conviene tener presente que no todas las variables expresadas numéricamente son cuantitativas, por ejemplo el código postal o las matrículas de los coches. Las variables de intervalo y de razón son variables cuantitativas.

Escala nominal

Los valores o categorías de las “variables cualitativas” que se representan en escala nominal se diferencian por sus cualidades y no por sus cantidades. Por lo tanto, no representan valores cuantitativos, es decir, sus valores no implican un orden numérico, solo permiten clasificar sus valores o categorias, ya que expresan un atributo o cualidad de la variable nominal. Por ejemplo, el partido político, el color del cabello, la religión, el estado civil, el sexo, la orientación sexual, la ciudad donde se vive, el color de los ojos, la profesión, la película favorita, el partido político, los

colores… Se les puede asignar un número para poder diferenciar las categorías, como por ejemplo 0 hombres y 1 mujeres, pero sus categorías no tienen valor numérico. En este sentido, por ejemplo, la variable de religión difiere cualitativamente, pero no implica ningún ordenamiento por el valor numérico asignado a cada categoría. Si las variables nominales tienen dos categorías se les llama variables categóricas (por ejemplo el sexo asignado al nacer de hombre y mujer) y si tienen más de dos categorías entonces son politómicas (por ejemplo el estado civil de soltero/a, casado/a, viudo/a, nacionalidad, ocupación laboral...).

En resumen, las variables cualitativas medidas en escala nominal implica que cuando se miden simplemente se nombran o se categorizan las respuestas y no se puede establecer ningún orden entre las respuestas. No tiene sentido clasificar como mayor o menor a las personas por su sexo asignado al nacer o por su partido político. Sólo se puede decir si una categoria es igual o distinta a otra (se pueden clasificar en diferentes categorias), pero no existe ni orden, ni relación de medida entre ellas.

Escala ordinal

Respecto a las variables cualitativas medidas en escala ordinal, se trata de variables que pueden tomar valores ordenados según una escala prefijada. Sus valores permiten clasificar a las variables y, además, establecer un orden lógico entre las variables permitiendo afirmar si un valor es mayor o menor que otro valor. Por ejemplo, el rango militar, el nivel de estudios, la clase social...

El tipo de análisis que se puede realizar con los valores en escala ordinal es solamente de orden, pues no tiene sentido realizar operaciones aritméticas con los datos. Es decir, permiten hacer comparaciones de grado en la variable medida.

En este punto conviene realizar una reflexión, cuando la escala ordinal tiene varios niveles de respuesta (por ejemplo, el grado de satisfacción medido desde nada satisfecho hasta muy satisfecho con una escala de cuatro puntos: nada, algo, bastante, mucho), ¿es adecuado suponer que la diferencia entre dos niveles de una escala ordinal (nada de acuerdo y algo de acuerdo, por ejemplo) es igual a la diferencia entre otras dos opciones de respuesta (algo y bastante de acuerdo)? La respuesta es no. El procedimiento de medición no permite determinar si las dos diferencias reflejan la misma diferencia en la variable medida (por ejemplo, satisfacción con un producto). Por lo tanto, las diferencias entre los valores de la

escala no representan necesariamente intervalos iguales en la escala de medida de la variable. Por ejemplo, los valores de las variables de clase social entendida como baja (1) - media (2) - alta (3) o el grado de acuerdo con un determinado ítem de un instrumento (desde nada de acuerdo hasta muy de acuerdo con 5 opciones de respuesta) no se puede saber si representan la misma cantidad en cada intervalo de medida.

Señalamos, por tanto, que a nivel teórico las escalas tipo *Likert* o de valoración subjetiva de determinados conceptos (puntuar un objeto, indicar el nivel de satisfacción, valorar una situación, etc.) que se suelen utilizar habitualmente en Ciencias Sociales son esencialmente ordinales. En la práctica, sin embargo, observamos que cuando se analizan los datos, una amplia mayoría de investigadores e investigadoras están asumiento el carácter cuantitativo de la información al identificar este tipo de variables como “semi-cuantitativas” o “cuasi-cuantitativas”.

En resumen, tal y como se ha comentado, la escala de medida de las variables psicológicas suele realizarse con una escala tipo *Likert* con varias opciones de respuesta. Por ejemplo, se le puede pedir a los sujetos que valoren su grado de satisfacción con un producto o que manifiesten su actitud ante un grupo minoritario. Y se suele utilizar escalas con 5 ó 7 puntos para manifestar su grado de opinión. Este tipo de escalas son ordinales, ya que no existe garantía de que la diferencia entre las opciones de respuesta represente el mismo grado en todos los valores de respuesta. Por ejemplo, la diferencia en la satisfacción entre un nivel de 1 (poco satisfecho) a 2 (algo satisfecho) podría no representar lo mismo que la diferencia entre un nivel de 3 (satisfecho medianamente) y 4 (bastante satisfecho). A pesar de ello, este tipo de escalas de medida se utilizan como cuasi escalas de intervalo para poder aplicar las técnicas de inferencia estadística, ya que requieren que la escala de medida de la variable dependiente sea realizada, al menos, en escala de intervalo. Esta situación tiene sus detractores quienes manifiestan que no se debería utilizar ese tipo de medición para llevara cabo un contraste de hipótesis estadísticas. De ahí, la importancia de valorar con detalle los puntos de la escala tipo *Likert* que se utilizarán en el instrumento de medida, reflexionando sobre si la amplitud de los diferentes intervalos son comparables entre sí, aunque nunca se podrá saber si el intervalo entre esas mediciones es realmente uniforme, y se recomienda ampliar las opciones de respuesta, al menos utilizar una escala de 6-7 puntos y utilizar intervalos de la

variable que puedan estar representando distancias aproximadas, tratando de conseguir que el intervalo entre cada opción de respuesta pueda ser comparable.

Variables cuantitativas

Los valores de las “variables cuantitativas” se miden de forma numérica y sus valores se corresponden con cantidades y tiene sentido hacer operaciones matemáticas con ellos. Por ejemplo, la edad, los ingresos económicos o el número de años recibidos de educación.

Las variables cuantitativas pueden ser discretas y continuas.

Variables discretas

Una variable discreta es aquella que tiene un conjunto de valores particulares y puede tomar un número finito de valores entre dos valores. Por ejemplo, cuántos amigos tiene uno en Facebook). En este tipo de variables cuantitativas discretas, es importante destacar que no hay valores de término medio entre las mediciones; no tiene sentido decir que uno tiene 33,7 amigos, que uno tiene un perro de ¾ partes de perro alemán o que una persona tiene 1,5 hermanos, se lanza una moneda y se obtienen dos caras y media, son 30,5 alumnos y alumnas en el aula o se ha atendido a 4,7 personas en el ambulatorio. Por lo tanto, las “variables cuantitativas discretas” sólo pueden tomar valores enteros y tienen un número finito de valores entre dos valores dados. Por ejemplo, número de personas en el hogar, número de hijos, tamaño del municipio, número de goles que ha metido un jugador en el partido, las calificaciones académicas de 0 a 10. Es decir, las posibles puntuaciones de las variables son puntos discretos en la escala.

Variables continuas

Una variable continua se define en términos de un número real y puede tomar un número infinito de valores intermedios entre dos números. Se trata de una escala de valores ininterrumpida.

Es decir, podría tener cualquier valor de determinado rango de valores, aunque normalmente las herramientas de medición limitarán la precisión con la que se puede medir; por ejemplo, una báscula de peso podría medir el peso al kilogramo más cercano, aunque en teoría el peso podría medirse con mucha más precisión.

Por lo tanto, en las “variables cuantitativas continuas” el número de valores posibles entre dos valores dados es infinito (valores con infinitos decimales). Por ejemplo, la altura, la longitud, el peso, la edad. Por ejemplo, el tiempo para responder una prueba de aprendizaje es una variable continua ya que la escala es un continuo de valores y no está formada por valores discretos.

En la práctica, las variables continuas que son el resultado de medir se redondean y se convierten en número finito de números enteros y las variables cuantitativas discretas son aquellas que tienen pocas categorías.

Como principal conclusión de los diferentes tipos de variables, hay que destacar que las estadísticas no tienen sentido en algunos tipos de datos.

Por ejemplo, si se van a recopilar datos del código postal de varias personas, esos números se representan como números enteros, pero en realidad no se refieren a una escala numérica. Cada número del código postal sirve básicamente como etiqueta para una región diferente y, por ello, no tendría sentido hablar del código postal promedio, por ejemplo.

Las variables cuantitativas continuas pueden estar en escala de intervalo y en escala de razón.

Conviene anotar que en la práctica se utilizan las escalas tipo *Likert* (variables con opciones de respuesta que oscilan en torno a unos valores que escalan la respuesta como por ejemplo desde muy en desacuerdo (1) hasta muy de acuerdo (7)) como variables cuasi-cuantitativas, a pesar de ser variables que se miden de forman ordinal y se considerarían cualitativas (escala ordinal).

Escala de intervalo

En las variables con escala de intervalo ya se dispone de una unidad de medida que es el intervalo entre dos valores y representa una distancia exacta entre los valores en cada uno de los intervalos. Es decir, los intervalos de la escala de medida tienen la misma interpretación. Se trata de una escala cuantitativa en sentido estricto. Por ejemplo, la diferencia entre 30 grados y 40 grados en la escala de temperatura Fahrenheit es la misma que la diferencia entre 10 grados y 20 grados ya que cada intervalo de 10 grados tiene el mismo significado físico. Por ejemplo, ingresos

económicos de 10000 euros, de 20000 o de 30000 euros. Así, se puede calcular la distancia o el intervalo entre cualquier par de valores de la variable.

Con las variables de intervalo se permite la clasificación de las variables, se permite establecer un orden lógico entre las variables y existe una distancia exacta o uniforme entre cada intervalo de medida. Por lo tanto, con dichas variables se pueden realizar operaciones aritméticas como la suma o la resta. Conviene tener presente que el origen de la escala de medida es arbitrario o se fija por convención; es decir, no hay un verdadero punto cero, aunque se pueda nombrar. Por ejemplo, cero grados Fahrenheit no representa la ausencia total de temperatura, sino que se trata de una decisión arbitraria sobre donde "comenzar" la escala de temperatura que será nombrada como temperatura cero.

Escala de razón

Con las variables de razón se permite la clasificación de las variables, se permite establecer un orden lógico entre las variables, existe una distancia exacta o uniforme entre cada intervalo de medida y tienen un valor de ausencia de la característica. Con otras palabras, son variables con el nivel de medición de intervalo, pero además se puede establecer un origen o punto cero que representa la ausencia absoluta de la caracetrística que se desea medir.

En las escalas de razón existe una unidad de medida y un valor de cero absoluto (por ejemplo los ingresos realizados o los gastos, el peso, la edad, la estatura…). Con este tipo de escalas se admite cualquier tipo de operación aritmética. Es decir, se trata de una escala de intervalo con la propiedad adicional de que su posición cero sí indica la ausencia de la cantidad que se mide. La escala Fahrenheit para medir la temperatura tiene un punto cero arbitrario y por lo tanto no es una escala de razón, pero en la escala de Kelvin sí hay un cero absoluto y se mide en escala de razón. Otro ejemplo podría ser la cantidad de dinero que se ingresa al mes en una cuenta bancaria (si no se realiza un ingreso la cantidad será cero y ese valor implica la ausencia de dinero), la edad, el número de hijos, el peso, el número de usuarios de una red social, el número de seguidores.

En resumen, la escala nominal proporciona un nombre o una categoría para cada objeto, ya que los números se utilizan como etiquetas. En la escala ordinal los objetos están ordenados, pues los números asignados permiten establecer dicho orden de

mayor o menor. En la escala de intervalo también existe ese orden y, además, la diferencia entre dos valores de la escala tiene el mismo significado o es uniforme. Por último, en la escala de razón se mantiene la interpretación de la escala de intervalo y, además, hay un valor cero absoluto de ausencia de la cantidad medida.

Capítulo 4. Proceso del diseño de investigación

Dolores Frías-Navarro

Universidad de Valencia

Índice

Citar el capítulo como:

Frías-Navarro, D., (2021). Proceso del diseño de investigación. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

El proceso del diseño de investigación representa los pasos implicados en el método científico. A continuación se detallarán los elementos más importantes en cada una de las fases de dicho proceso o método científico.

Necesidad de conocimiento

La necesidad de conocimiento implica definir de forma clara, precisa y concreta el problema de investigación que se debe resolver, permitiendo así iniciar el proceso de búsqueda y localización de la información que facilitará la respuesta más adecuada, precisa y actual del problema planteado.

El problema de investigación (*necesidad de conocimiento*) debe expresar una relación (causal o no causal) entre dos o más variables y su planteamiento será claro y sin ambigüedades y debe permitir su verificación empírica.

Además, el problema de investigación debe ser relevante, justificando el esfuerzo y la inversión económica y de trabajo humano que se invierte en su resolución.

El objetivo del estudio puede ser eliminar incertidumbre sobre el conocimiento de cierta temática o puede ser modificar y/o añadir nuevos conocimientos al área concreta de investigación (ver Figura 7).

También, es muy importante enmarcar el problema dentro de una teoría teniendo en cuenta que la Ciencia busca generalizar los hallazgos (que además sean replicables) y no se construye con hechos aislados.

El conocimiento teórico proporciona la explicación de los hechos empíricos investigados.

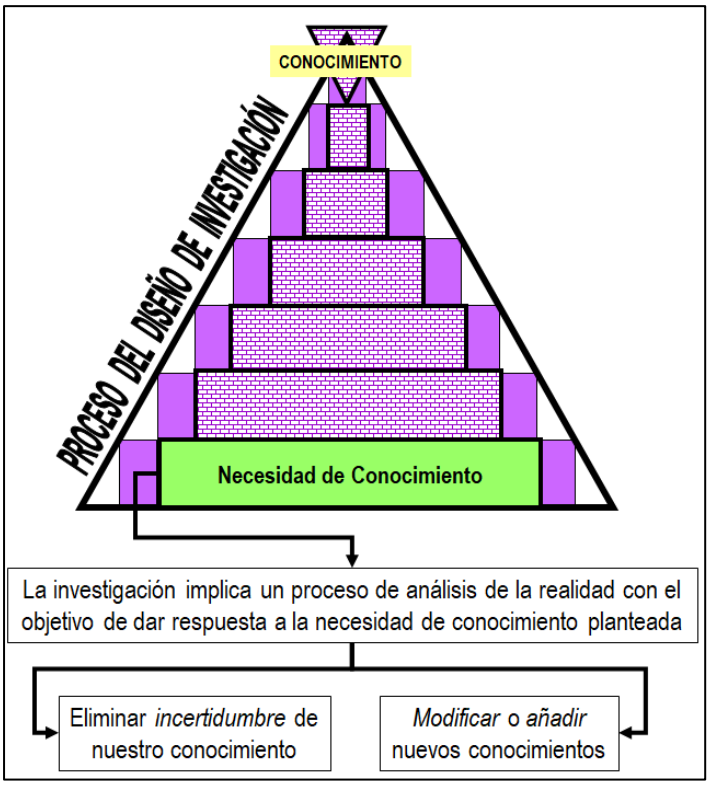


Figura 7. Necesidad de conocimiento

Pregunta PICO

En los estudios de intervención, la estructura de una "pregunta" clínica de investigación o planteamiento de la necesidad de conocimiento suele adoptar la forma conocida como PICO, definiendo de forma precisa (ver Figura 8):

- El tipo de *P*acientes o problema que se valorará en el estudio (P).
- El tipo de *I*ntervención cuyo efecto se quiere estudiar (I).
- El tipo de grupo de *C*omparación que se utilizará (C).
- Y el resultado ('*O*utcome') que se valorará (O).

La evaluación de la calidad de la pregunta de investigación se realiza reflexionando sobre sus características y valorando que sea una pregunta factible, interesante, novedosa, ética y relevante (FINER, *Feasible*, *Interesting*, *Novel*, *Ethical* y *Relevant*).

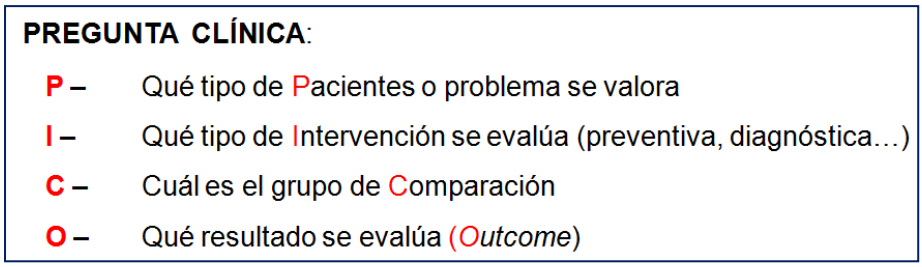


Figura 8. Estructura de la pregunta clínica de investigación

La evaluación de la pregunta clínica permite redefinirla si fuese necesario, ya que no se habrían representado adecuadamente las características anteriores de FINER.

Es recomendable tener en cuenta los elementos de la pregunta PICO para elaborar un buen "título" del estudio realizado, aportando en dicho título todos los detalles que son interesantes para que el lector o lectora tenga con una información muy breve expresada en dicho título una idea lo más completa posible del contenido que se puede encontrar si decide leer el artículo o informe de investigación.

Conocimiento previo

Una vez ya se ha planteado la necesidad de conocimiento (pregunta de investigación), para construir la investigación es necesario que el investigador o investigadora desarrolle en primer lugar, tareas de búsqueda y revisión de la información previa ('*conocimiento previo*') para posteriormente valorar críticamente la calidad de las evidencias o pruebas aportadas en la literatura y utilizar en su estudio aquella información que ha pasado el filtro de la calidad de su evidencia. En este punto es muy importante que el investigador o investigadora y el lector o lectora dispongan de competencias de lectura crítica o activa fundamentadas en el modelo de la Práctica Basada en la Evidencia así como ser competentes en conocimiento metodológico para poder desarrollar esa lectura y conocimiento sobre las principales bases de datos donde se almacenan los registros de las publicaciones científicas.

La literatura es el repositorio escrito del conocimiento. Su revisión ayuda a perfilar de manera más elaborada la hipótesis de investigación. Como ya se ha comentado, una buena hipótesis debe ser factible o abordable empíricamente, interesante, novedosa, ética y relevante para generar conocimiento científico.

Delimitar de forma clara, precisa y concreta el problema que se debe resolver ("necesidad de conocimiento") permite iniciar el proceso de búsqueda y localización

de la información (“conocimiento previo”) que facilitará la respuesta más adecuada, precisa y actual al problema planteado.

La revisión del conocimiento previo (revisión de la literatura) puede proceder del análisis de experiencias anteriores de investigación o estudios piloto y/o de la revisión bibliográfica de información (estudios primarios y estudios secundarios como las revisiones sistemáticas y los trabajos de meta-análisis o los estudios de meta-investigación) (ver Figura 9). En el video de youtube presentado Jesús López Alcalde (24 de junio de 2020), director del Centro Cochrane Asociado de Madrid, titulado “Introducción a las revisiones sistemáticas” se describen qué es y qué pasos implican una revisión sistemática que, a veces, puede acabar en un trabajo de meta-análisis: https://youtu.be/NM-TIzrpxOE (56:16 minutos). Vídeo

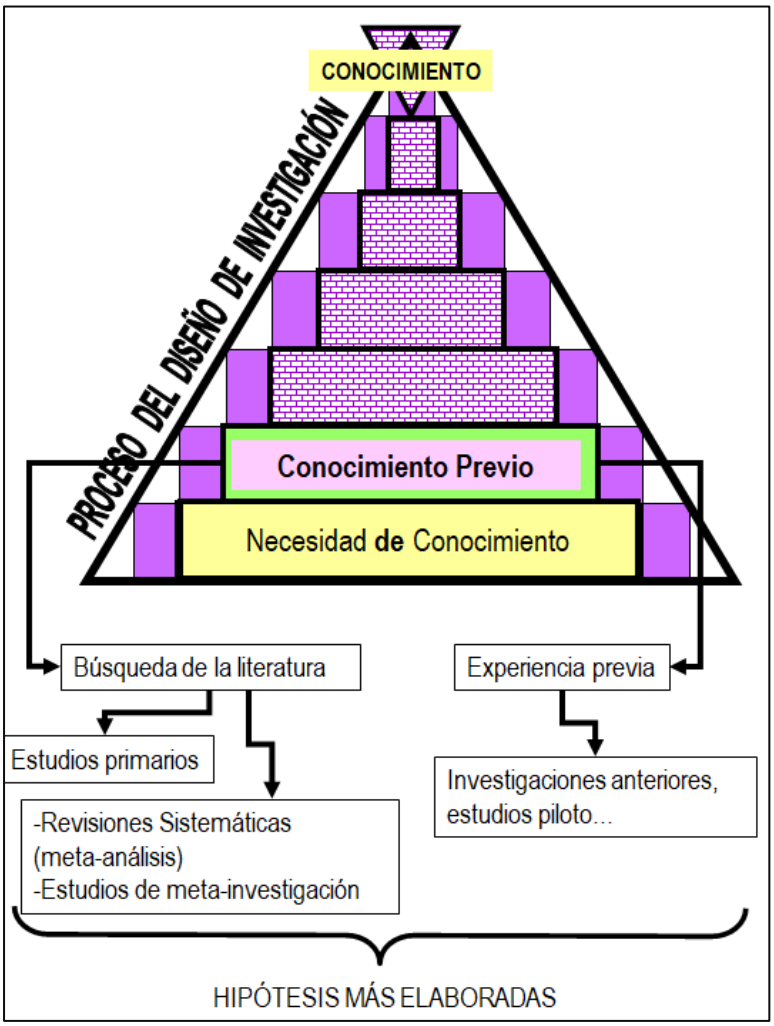


Figura 9. Conocimiento previo

Por lo tanto, una vez que se ha delimitado el problema de investigación es necesario iniciar un proceso de búsqueda de información y de valoración de dicho conocimiento previo que existe sobre el fenómeno objeto de estudio.

Actualmente, y dada la cantidad de información a la que se puede recurrir, exige utilizar una técnica de búsqueda bibliográfica que sea sobre todo metódica: definir la cuestión, seleccionar la fuente o las fuentes de información (bases de datos), formular el perfil de búsqueda con las palabras claves y ejecutar la búsqueda. Si los resultados no son satisfactorios entonces será necesario modificar el perfil de búsqueda o las palabras claves y/o las bases de datos seleccionadas. Las bases de datos (por ejemplo, WOS, SCOPUS, PubMed, PsycInfo…) son una herramienta muy valiosa para obtener el conocimiento previo.

Actualmente, las bases de datos facilitan enormemente el trabajo de búsqueda y localización del conocimiento previo y permite su almacenamiento en gestores de referencias bibliográficas. En los años 80 esta tarea era muy ardua y, por ejemplo en la Universidad de Valencia, la consulta de los abstracts o resúmenes de los artículos publicados se realizaba de forma manual, consultando los tomos denominados "Psychological Abstracts" que de forma periódica se recibían en la biblioteca para buscar el conocimiento previo a través de las palabras clave que se indexaban en su índice. Todo era manual, pues después para conservar esa información se escribía en una ficha que finalmente se guardaba en el fichero para poder ser consultada en cualquier momento (ver imagen 4). Ya en la década de los 90 la Universidad de Valencia tiene contratado el servicio bibliográfico con versión electrónica y solo con un clic del ratón se puede acceder a miles de artículos con su referencia completa y resumen y solo con un clic se puede guardar la información de aquellos artículos que nos interesen en nuestro gestor de referencias bibliográficas. Los tomos del Psychological Abstracts dieron paso a la base de datos electrónica "APA PsycInfo" (https://www.apa.org/pubs/databases/psycinfo), base esencial del ámbito de la Psicologia y ciencias afines que está producida por la American psychologial Association y ofrece a los usuarios y usuarias muchas herramientas para llevar a cabo un análisis detallado de la información especificada en las palabras clave de consulta. En la siguiente dirección se ofrece una breve historia de esta base de datos: https://ahp.apps01.yorku.ca/2011/02/brief-history-of-psycinfo y también se puede consultar el artículo de Benjamin y VandenBos (2006) y Evans y cols., (1992).

Imagen 4. Fichero de registros de las búsquedas bibliográficas (fichero de 1986)

Después de la evaluación de la información obtenida con la revisión de la literatura se continúa el proceso de diseño con otras cuestiones metodológicas como delimitar las variables que van a ser investigadas y que formarán parte de las hipótesis de investigación, plantear las posibles amenazas a la validez de los hallazgos y planificar su control ya sea con el propio diseño de investigación (por ejemplo con la técnica de control de la aleatorización, con la técnica de la constancia, con el diseño de bloques …) o con la herramienta estadística de diseño más apropiada (por ejemplo con diseños con variables covariadas). Además, también se valorará el tipo de metodología de investigación que puede ofrecer la mejor respuesta a los objetivos de investigación. La respuesta podría ser aplicar un diseño cuantitativo donde se aplique la metodología experimental, la cuasi-experimental o la metodología no experimental, un estudio cualitativo, o, quizas, un estudio con un método mixto donde se combine una parte cuantitativa con una cualitativa. Se trata de cubrir las fases de técnica y análisis metodológico del proceso de diseño de investigación descritos anteriormente en la pirámide del proceso del diseño de investigación.

Las fuentes de información que contienen el conocimiento previo pueden ser fuentes *primarias* (por ejemplo un artículo de revista, una conferencia, una tesis doctoral…) y fuentes *secundarias* donde se recogen y organizan los datos de la literatura primaria (bases de datos bibliográficas, base de datos de libros con ISBN,

base de datos TESEO de tesis doctorales españolas, estudios de meta-análisis, estudios de meta-investigación…). Los artículos, libros, tesis, conferencias… son ya tan numerosos y su producción es tan rápida que se convierte en una ardua labor revisar y evaluar la información.

Las bases de datos y el acceso a Internet son una gran ayuda, pero al mismo tiempo favorecen la acumulación de información que muchas veces se queda en el cajón (en el disco duro del ordenador) sin llegar a ser leída nunca. Además, no toda la información tiene el mismo valor científico, pues tal y como ya se ha comentado la calidad de la evidencia aportada por un estudio se puede jerarquizar. Por lo tanto, existe una jerarquía en la calidad de las pruebas aportadas por las investigaciones. Por ello, es necesario evaluar de forma crítica o activa la calidad de los hallazgos de las publicaciones e informes de investigación dado que la evidencia o las pruebas aportadas por el estudio tienen diferentes grados de validez.

Es muy importante que los programas de formación universitaria desarrollen este tipo de competencias entre los alumnos y alumnas, formando profesionales que sepan acceder a la información con las nuevas tecnologías, pero que también conozcan cómo valorar la información y jerarquizarla en función de su calidad y rigurosidad y cómo integrar la información obtenida en su tarea profesional. Competencias que el movimiento de la Práctica Basada en la Evidencia destaca como esenciales para el profesional actual.

El término de Práctica Basada en la Evidencia procede del modelo de Medicina Basada en la Evidencia cuya definición clásica señala que se trata de la utilización consciente, explícita y juiciosa por parte del profesional de la mejor evidencia clínica disponible hasta el momento, para tomar decisiones sobre el cuidado de los pacientes. Por lo tanto, integra la maestría clínica individual de los profesionales (juicio clínico) con las mejores evidencias científicas externas disponibles a partir de una investigación. Este último componente relacionado con la identificación de la evidencia o las pruebas parece, en teoría, relativamente fácil de adquirir, pero en la práctica los profesionales se ven desbordados por una cantidad de información imposible de manejar en muchas ocasiones. Hay muchas fuentes de publicación o revistas, además su calidad es muy desigual, sus contenidos se pueden quedar obsoletos en un lapsus de tiempo breve y se publica muchísima información cada

día, agravándose la situación con el hecho de que el tiempo disponible para la búsqueda de información y el estudio es cada vez más limitado (Shaughnessy, 2009). Como señala Shaughnessy (2009), "*cada año los médicos deben decidir cuál de los miles de artículos recién publicados podrá tener tiempo para leer. Para determinar qué artículos son los más útiles clínicamente, los médicos deben evaluar su relevancia, validez e importancia clínica. El uso de estos criterios puede reducir drásticamente la cantidad de artículos que los médicos necesitan leer*" (p. 668). Como consecuencia de lo anterior, los conocimientos del profesional se deterioran rápidamente tras finalizar su graduación universitaria. Como señalan Gisberta y Bonfill (2004), este hecho está magníficamente representado en la gráfica clásica donde el eje de ordenadas representa los conocimientos actualizados sobre el mejor tratamiento y el eje de abscisas los años desde la graduación (ver Figura 10). La representación gráfica señala lo que se ha denominado una peligrosa "pendiente resbaladiza" dada la relación negativa entre las variables de conocimientos actualizados y años que han pasado desde la graduación universitaria.

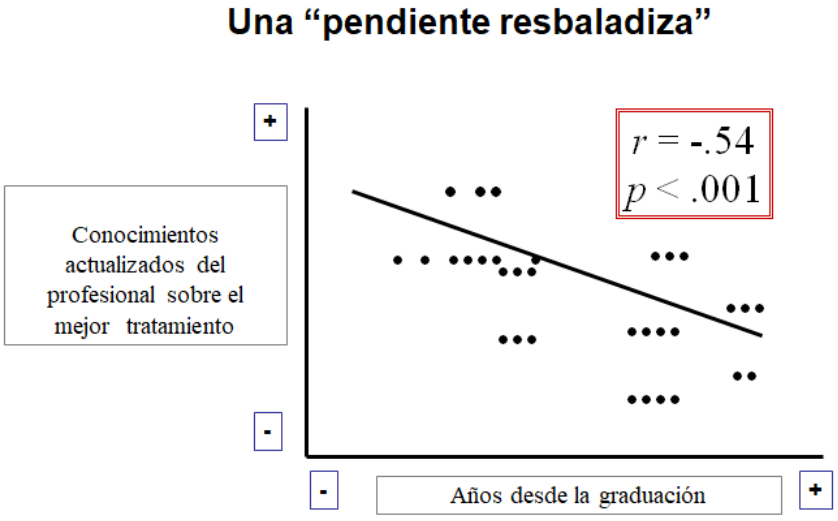


Figura 10. Relación entre conocimientos actualizados y años que han transcurrido desde la graduación

Cuando se planifica una investigación, una adecuada revisión y evaluación de la información disponible en los informes y trabajos de investigación permitirá plantear hipótesis más elaboradas, mejorando el proceso de diseño de investigación y la acumulación de conocimiento válido. Tal y como ya se ha comentado, en este punto, las bases de datos y la consulta vía Internet han ayudado de forma considerable a la

tarea de revisión de la literatura y actualmente la dificultad no estriba tanto en obtener información sobre una temática sino más bien en cómo valorar de forma crítica la calidad de las evidencias que aportan los estudios. Y esa es una competencia indispensable del investigador o investigadora moderno y del profesional como consumidor de literatura científica.

Las denominadas "listas de comprobación" o "guías de publicación" son una gran ayuda, tanto para que el investigador o investigadora chequee o compruebe la calidad de los resultados de su estudio antes de enviar el manuscrito a la revista para su futura publicación como para que los revisores y las revisoras de las revistas tengan un protocolo de comprobación del proceso de diseño de investigación llevado a cabo por el autor del artículo. También es una gran ayuda para los lectores y lectoras que desean realizar una lectura activa o crítica ya que destacan los elementos clave del proceso de diseño de investigación que deben chequear para valorar el proceso de diseño de investigacion y la calidad de la evidencia o pruebas aportadas en el estudio.

En el área de los estudios primarios, destacan las guías denominadas (González de Dios y cols., 2014):

- JARS (Journal Article Reporting Standards), elaborada desde el grupo de trabajo de la American Psychological Association (JARS Group) (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; Appelbaum y cols., 2018; Levitt y cols., 2018).
- CONSORT (CONsolidated Standards fOr Reporting of Trials) para ensayos clínicos aleatorios o estudios realizados con una metodología experimental.
- TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) para estudios con una metodología cuasi-experimental sin asignación aleatoria del tratamiento.
- STROBE (STrengthening the Reporting of Observational studies in Epidemiology) para estudios no experimentales u observacionales.

- STRAND (STandards for Reporting of Diagnostic Accuracy) para estudios de precisión diagnóstica.
- ROBINS. La herramienta ROBINS-I es una herramienta de riesgo de sesgo para evaluar estudios de intervenciones no aleatorizados. ROBINS-I es la herramienta preferida para ser utilizada en las revisiones Cochrane para estudios no aleatorios de intervenciones, aunque no es obligatoria. Una opción alternativa es la escala Newcastle-Ottawa.

En el área del análisis de las revisiones sistemáticas o estudios de meta-análisis destacan los siguientes listados o listas de comprobación:

- PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) para valorar la calidad de las revisiones sistemáticas y los estudios de meta-análisis. Consta de 27 ítems.
- AMSTAR 2. (A Measurement Tool to Assess Systematic Reviews). AMSTAR fue desarrollado para evaluar revisiones sistemáticas de ensayos aleatorizados. AMSTAR-2 incluyen también el análisis de los estudios no aleatorizados (Ciapponi, 2018).
- MOOSE (Meta-analysis of Observational Studies in Epidemiology). Listado para chequear meta-análisis de estudios observacionales (no experimentales).
- REGEMA (REliability GEneralization Meta-Analysis), listado de 30 ítems (Sánchez-Meca, y cols., 2021). Meta-análisis de generalización de la fiabilidad (reliability generalization). Este tipo de estudios meta-analíticos sobre la generalización de la fiabilidad proporciona una estimación media de la fiabilidad de las puntuaciones de un instrumento, determina si los coeficientes de fiabilidad obtenidos en diferentes investigaciones son heterogéneos entre sí (si varían cuando el instrumento se aplica a diferentes muestras) y, si lo fuesen, analiza qué características de los instrumentos y de los participantes podrían explicar dicha heterogeneidad (Sánchez-Meca y cols., 2016). Consultar en: https://www.um.es/metaanalysis/REGEMA.php

Desde el área de herramientas generales con diferentes listados de comprobación en función del objetivo del chequeo destaca:

- Los listados CASPE (Critical Appraisal Skills Programme Español; Programa de Habilidades en Lectura Crítica Español) para realizar lecturas críticas de estudios con diferentes metodologías de investigación (http://www.redcaspe.org). CASP International es una organización que proporciona materiales para la lectura crítica de la evidencia clínica como plantillas o listados de comprobación (listas de verificación) para evaluar diferentes tipos de evidencia (ensayos clínicos, revisiones sistemáticas, estudios de diagnóstico, estudios cualitativos, estudios de predicción clínica, estudios de casos y controles, estudios de cohorte y análisis de evaluaciones económicas) y calculadoras de determinados estadísticos.

Como se ha comentado, los consumidores y las consumidoras de literatura o lectores y lectoras pueden hacer uso de dichas guías para realizar una lectura crítica de los trabajos de investigación, chequeando y valorando cada uno de sus puntos cuando realiza la lectura de un artículo o informe. En todos los casos, las listas de comprobación facilitan una valoración crítica de los hallazgos de una investigación, enmarcando la conducta del profesional y del investigador o investigadora en el modelo de 'Práctica Basada en la Evidencia'.

En resumen, el objetivo de la fase de elaboración del conocimiento previo es revisar de forma sistemática la documentación que existe sobre el tema y esto exige definir un protocolo de consulta que incluya los repertorios bibliográficos o las bases de datos más relevantes dentro de la temática de investigación (fuentes de información secundaria).

Una de las bases de datos más utilizada es MEDLINE de la *National Library of Medicine* junto con su versión PubMed en Internet. Dentro del área médica destaca también la base de datos Excerpta Medica (denominada EMBASE). En Psicología sobresale la base de datos PsicInfo, producida por la *American Psychological Association* (APA) y en Educación se encuentra ERIC (*Education Resources Information Center*).

Una plataforma de reconocido prestigio de búsqueda multidisciplinaria que permite hacer uso de varias bases de datos de forma simultánea en una misma interface es Web of Science (WoS). A partir de ella se elaboran los conocidos factores de impacto de las revistas (JCR, Journal Citation Report) y los índices de citas de los artículos.

Por otra parte, una fuente de información bibliográfica primaria de gran importancia es la revista de *Annual Reviews* (http://www.annualreviews.org/). El Annual Reviews es un conjunto de revistas especializadas en una temática como la Psicología (*Annual Reviews of Psychology*), la Psicología Clínica (*Clinical Psychology*), la Estadística y sus Aplicaciones (*Annual Reviews of Statistics and its Application*), la Economía (*Annual Reviews of Economics*), la Sociología (*Annual Reviews of Sociology*), la Criminología (*Annual Reviews of Criminology*), la Salud Pública (*Annual Reviews of Public Health*) o la Medicina (*Annual Reviews of Medicine*), por ejemplo.

Conviene tener en cuenta que las bases de datos informatizadas no consiguen localizar todos los trabajos publicados ni están completamente actualizadas, ya que se requiere un tiempo desde la publicación del artículo hasta la introducción de la información en la base. Actualmente esta cuestión se ha mejorado ya que las revistas suelen incorporar los artículos on line en sus páginas Wen antes de paginarlos definitivamente en el número y volumen correspondiente. Las bases de datos tampoco son exhaustivas. La denominada 'literatura gris', es decir, aquella que se publica en resúmenes de congresos, informes de organismos o instituciones… no aparecen en dichos repertorios. Por lo tanto, es conveniente completar la búsqueda por otras vías como el contacto directo con el investigador o investigadora experto en el tema o la consulta directa de la documentación que suelen ofrecer los centros de investigación en sus páginas Web.

Actualmente el servicio que ofrecen las revistas de alertar al usuario vía correo electrónico cuando se produce la aparición de un nuevo número de la revista es una de las mejores opciones para maximizar el acceso rápido a los últimos avances que se producen en el campo científico de interés. También, el servicio de alertas de citación de un artículo que proporcionan las bases de datos es realmente muy útil.

En definitiva, la revisión del conocimiento es una ardua tarea que el investigador o investigadora debe realizar, sobre todo, de forma sistemática con el objetivo de representar las diferentes perspectivas de los autores y el desarrollo de la temática que se desea estudiar.

Hay que evitar una búsqueda de la información sesgada, ya sea por los intereses del investigador o investigadora (solamente interesan publicaciones que apoyan una teoría por ejemplo, o las que siguen un determinado modelo), o por acceder solamente a la literatura que está disponible de forma sencilla para el investigador o investigadora como la información depositada en una biblioteca o centro de trabajo.

Ciertamente, las revisiones sistemáticas y los trabajos de meta-análisis representaron un hito importante en la formación teórica sobre un fenómeno, pero de nuevo la lectura crítica y activa de este tipo de trabajos, como el de los estudios primarios, es necesario para seguir filtrando la calidad de los hallazgos publicados. El mismo control de calidad y jerarquía de la evidencia que se aplica cuando se lee un estudio primario debe ser aplicado cuando se lee un estudio de revisión sistemática o de meta-análisis.

A las bases de datos, los estudios primarios encontrados en las revistas, los informes de las asociaciones e instituciones y la literatura gris, en general, hay que añadir ahora los denominados pre-prints o documentos preimpresos que día a día inundan de nuevo la información que hay sobre un tema y los pre-registros de los estudios. Conviene estar muy alerta cuando se lee un trabajo de pre-print depositado o almacenado en un respositorio, ya que dichos trabajos aún no han pasado por el filtro de la revisión por pares que es indispensable para publicar un artículo en una revista científica (revisión crítica por parte de otros investigadores como fase previa para que un manuscrito sea aceptado para su publicación en una revista) y, por lo tanto, la calidad de su diseño de investigación no ha sido valorada.

Esa revisión por pares o por expertos de los manuscritos enviados para su posible publicación tiene como objetivo identificar los puntos débiles del diseño del estudio y puede conducir a la decisión de rechazar el manuscrito, proponer cambios importantes para mejorar la calidad de la presentación y redacción del estudio, proponer cambios menores o directamente su publicación (este último hecho se produce en muy pocas ocasiones). Por supuesto, este filtro científico es fundamental

para obtener conocimiento válido, pero no es garantía absoluta de que los artículos publicados carezcan de errores graves. En la última década ha aumentado la tasa de artículos retirados de las revistas, una vez publicados, ya que se ha demostrado, por el análisis de meta-investigación o por la revisión de los lectores o lectoras expertos, que habían errores que no se detectaron en el proceso de análisis por los pares o expertos antes de su publicación. Esta es una de las razones que aconseja comenzar la lectura de un artículo con una postura crítica de análisis del proceso del diseño en todas sus fases, desde la revisión del conocimiento previo (calidad y actualización de la información presentada en el apartado de introducción) hasta la interpretación de los resultados y el conocimiento teórico que se aporta en el apartado de discusión. Dicha lectura exige activar las competencias de análisis activo y los conocimientos del área de la metodología de investigación. Si hay carencias en la formación metodológica entonces no se podrá llevar a cabo una lectura crítica o activa eficaz. Y, además, esas competencias requieren una re-educación constante para estar actualizados en cómo se presenta y se redactan los resultados científicos en cada contexto de investigación. El lector o lectora es el responsable para decidir qué valor tiene el artículo o informe que lee.

Ya se ha comentado que el auge de la Ciencia Abierta probablemente impulsará cambios destacados en la forma de hacer, leer y consultar la información. Los deseos de transparencia que demanda la Ciencia en sus informes y trabajos para facilitar la observación, el control de la calidad de sus hallazgos, evitando los falsos positivos y la mala interpretación de los resultados nulos (cuando se mantiene la hipótesis nula) y que la cualidad auto-correctiva que tiene la Ciencia realmente funcione exigen un cambio en todos los agentes implicados en el desarrollo de los hechos científicos. Los agentes incluyen a los mismos científicos y científicas y su práctica de investigación, a las políticas de los y las editores de las revistas, a las instituciones, al software estadístico, a la ética del investigador o investigadora y su re-educación en el proceso del diseño de investigación que afecta de forma directa dentro del mundo académico, a las enseñanzas que reciben los alumnos y alumnas y, también, entre los lectores y lectoras de material científico que a pesar de no construir investigaciones necesitan saber y comprender si aquello que leen se construyó con fiabilidad y validez. La sociedad del conocimiento avanza si el conocimiento está sometido al proceso riguroso del método científico.

Hipótesis de investigación

Después de la evaluación de la información obtenida con la revisión de la literatura se continúa el proceso de diseño con otras cuestiones metodológicas como delimitar las variables que van a ser investigadas y que formarán parte de las hipótesis teóricas o sustantivas de la investigación y su operacionalización, así como plantear el control de otras variables cuyos efectos deben mantenerse constantes en el diseño para evitar que contaminen los resultados.

La hipótesis de investigación es la denominada hipótesis científica o hipótesis sustantiva del estudio.

Para poder completar el proceso del método científico es necesario operacionalizar la hipótesis científica en un enunciado contrastable empíricamente (hipótesis científica operacionalizada) y así pasar a la fase de 'contrastación estadística' con datos empíricos del estudio. En esta fase de contrastación estadística mediante los procedimientos clásicos de contraste de hipótesis se formulan las dos hipótesis estadísticas: la *hipótesis nula* (hipótesis de nulidad del efecto, que generalmente plantea un efecto de cero del tratamiento) y la *hipótesis alternativa* (que plantea, generalmente, un efecto diferente a cero).

Planificación de la investigación

Llegados a este punto del proceso de diseño de investigación es necesario detenerse y planificar de forma cuidadosa el diseño del estudio donde se reflexionará sobre las posibles amenazas a la validez interna de los resultados (así como prestar atención al resto de tipos de validez: de conclusión estadística, de constructo y externa) y, si es el caso, se planificará su control ya sea con la propia metodología del estudio como por ejemplo la asignación aleatoria del tratamiento en la metodología experimental (técnica de la aleatorización), con técnicas concretas de control como la constancia o la eliminación de la variable contaminadora, el apareamiento y/o con la factorización de la variable contaminadora en la propia ecuación estructural del diseño (por ejemplo, los diseños de bloques o los diseños de cuadrado latino o los diseños grecolatinos) o con la herramienta de ajuste estadístico más apropiada (por ejemplo los diseños con variables covariadas).

En la fase de planificación del estudio también es necesario reflexionar sobre la calidad de la medida de los constructos (fiabilidad y validez) que van a ser estudiados (validez de constructo), sobre el grado de generalización de los hallazgos a otras situaciones y momentos temporales (validez externa) y sobre la calidad del proceso de inferencia estadística que se realizará (validez de conclusión estadística), siendo especialmente relevante en este último punto planificar el tamaño de la muestra en función de los criterios de alfa, potencia estadística deseada y tamaño del efecto esperado así como la comprobación de los supuestos estadísticos que las pruebas estadísticas requieren para que funcionen de forma adecuada y no estén sesgadas por artefactos metodológicos que podrían dañar la calidad de sus resultados (validez).

Por lo tanto, una fase de especial relevancia dentro del proceso del diseño de la investigación es la etapa de la planificación de los elementos que forman parte del diseño del estudio.

Durante la planificación de la investigación, el investigador o investigadora toma una serie de decisiones sobre:

1- La elección de las variables que forman el experimento y que operacionalizan a las variables constructo objeto de estudio.

2- La estrategia de recogida de datos.

3- El tipo de diseño más adecuado para contrastar las hipótesis estadísticas planteadas, controlando las posibles variables que podrían contaminar los resultados.

Se trata de los tres elementos básicos que darán forma al diseño de la investigación que se aplicará en un determinado estudio.

En opinión de Kerlinger (1986), el principal propósito de la fase de planificación del estudio es asegurar el objetivo metodológico conocido como *MAX-MIN-CON* que está relacionado con la variabilidad que se produce en la variable dependiente:

- *MAXimizar la varianza sistemática primaria* de la variable dependiente: vinculada al efecto de la variable independiente de tratamiento o a la magnitud de la relación entre las variables.

- *MINimizar la varianza del error* de la variable dependiente: variabilidad debida al azar, errores de medidas o provocados por diferencias aleatorias relacionadas con los propios sujetos, la situación experimental o el medio ambiente.
- *CONtrolar la varianza sistemática secundaria* de la variable dependiente: se genera por la influencia de alguna variable extraña relevante que actúa de forma sistemática por sí sola o junto con la variable de tratamiento.

La "varianza sistemática" es aquella desviación o variabilidad que presentan los datos (puntuaciones en la variable dependiente) hacia una dirección más que otra. Es decir, los datos varían de forma sistemática en un determinado sentido. Y la varianza sistemática que se produce en los datos (variable dependiente) puede ser:

-Varianza Sistemática Pimaria: provocada por la variable independiente cuyo efecto se desea estudiar en la investigación o

-Varianza Sistemática Secundaria: provocada por las terceras variables o variables extrañas cuyo efecto provoca variabilidad que contamina a los resultados.

Por otra parte, la "varianza no sistemática" se vincula a la varianza de error que es la variabilidad que presentan las puntuaciones en la variable dependiente debido a factores aleatorios. No es una variabilidad sistemática sino que se trata de una varianza aleatoria. Este tipo de variabilidad suele deberse a errores de medida debido a la utilización de instrumentos que no tienen unas buenas propiedades psicométricas o están mal calibrados, a factores relacionados con los propios individuos (diferencias individuales), a la propia situación experimental o a las condiciones ambientales. La varianza del error se conoce como variabilidad intra-grupo o variabilidad intra-tratamiento, ya que diferencia a los sujetos o a las unidades experimentales a pesar de que puedan recibir el mismo tratamiento o pertenezcan al mismo grupo. Por lo tanto, dado que se trata de una varianza aleatoria no se puede controlar, pero el investigador o investigadora sí puede minimizarla o reducirla a través de un diseño de investigación que vigile posibles fuentes aleatorias de error.

El diseño más eficaz es aquel que controla la varianza sistemática secundaria, extrayéndola de la varianza total de la variable dependiente de modo que se pueda

comparar la varianza sistemática primaria y la varianza del error (aleatoria) cuando se ejecuta una prueba estadística sin la presencia de variables extrañas que contaminen la verdadera variación de las observaciones, maximizando de este modo el efecto de la varianza sistemática primaria. La aplicación de los diseños de investigación que posibiliten el principio *MAX-MIN-CON* permitirá obtener un conocimiento válido del fenómeno planteado en las hipótesis de investigación.

Método

En la fase de método se desarrolla el diseño de investigación planificado previamente: participantes, diseño, tamaño de la muestra, instrumentos de medida, procedimiento llevado a cabo para recoger los datos y para poder ejecutar el experimento y diseño. Además, se valora qué tipo de metodología de investigación ofrecerá la mejor respuesta a los objetivos e hipótesis de investigación y a la estrategia de recogida de los datos: metodología experimental, cuasi-experimental, no experimental o $N = 1$. O quizás se llegue a la decisión de optar por una metodología cualitativa o, quizás, lo más adecuado sea aplicar una metodología mixta donde se combiene una parte de análisis cuantitativo y otra de análisis cualitativo. Estas cuestiones serán tratadas en profundidad cuando se analicen las diferentes metodologías de investigación.

Análisis de los datos

Siguiendo con los pasos ejemplificados en la pirámide del proceso de investigación, el análisis de los datos se ajustará a los objetivos de las hipótesis de investigación y a su planificación gracias al contraste de las hipótesis estadísticas mediante la prueba estadística que se considere más adecuada, acompañada de información sobre la magnitud del efecto detectado (estadístico del tamaño del efecto y su intervalo de confianza) así como de otros análisis que podrían ser relevantes (por ejemplo, el valor del Factor Bayes (FB) en los contrastes de hipótesis) y la interpretación clínica o sustantiva de los hallazgos.

Conviene tener muy presente que en los procedimientos habituales de análisis de datos basado en el modelo de comprobación de la significación de la hipótesis nula (Null Significance Statistical Testing, NHST), el contraste de hipótesis se inicia siempre asumiendo que la hipótesis nula es cierta (y se conoce su distribución) y

sobre dicho supuesto se obtiene la probabilidad del resultado obtenido (o resultados más extremos) en la distribución de la hipótesis nula. Posteriormente, el tema del contraste estadístico será tratado con detalle, ya que su comprensión y uso adecuado son competencias básicas del investigador o investigadora y también del lector o lectora consumidores de literatura científica.

Conocimiento adquirido

Finalmente, los resultados de los análisis estadísticos aportaran unos hallazgos que serán debatidos o discutidos teóricamente dentro del campo científico donde se desarrolla la investigación, dando lugar a un nuevo conocimiento científico que pasará a formar parte del conocimiento previo de un futuro estudio. Así, una vez obtenidos los resultados, la teoría propuesta en el estudio es aceptada o modificada, incorporando las nuevas evidencias detectadas por deducción y se iniciará de nuevo el proceso de investigación científica, formulando vía inducción suposiciones teóricas aceptables.

En este punto es muy importante destacar otras competencias básicas de los investigadores y las investigadoras relacionadas con la redacción cuidadosa de los resultados y con la valoración correcta del alcance de los resultados, así como de sus limitaciones o problemas que podrían hacer cambiar los hallazgos o moderar el alcance de su generalización. También los lectores y las lectoras deberán valorar de forma activa (lectura crítica) la calidad de la redacción de los resultados y la calidad de la discusión y conclusiones que se detallan en el artículo o informe.

En resumen, la Psicología es una Ciencia que construye enunciados contrastables empíricamente acerca del comportamiento humano y el método tiene la finalidad general de contrastar dichos enunciados o construcciones teóricas con los resultados de sus comprobaciones como un proceso sistemático de validación.

La Psicología es una Ciencia más cuyo método de conocimiento es el propio de la Ciencia en general: el método hipotético-deductivo que muy esquemáticamente supone formular un enunciado general (es decir, una hipótesis) y contrastarla con la realidad. Así, el método de la investigación científica supone un proceso iterativo de razonamiento inductivo y razonamiento deductivo (Nesselroade y Cattell, 1988), donde a partir de la observación de ciertos hechos se infiere por inducción suposiciones teóricas aceptables que expliquen cierta regularidad de los mismos que

a su vez permiten derivar consecuencias por deducción que podrán ser contrastadas con los datos. Si el procedimiento es satisfactorio entonces el enunciado se considera válido, pero si no se considera satisfactorio entonces el enunciado debe reformularse. Se trata de la conocida espiral del método hipotético-deductivo donde la inducción representa el comienzo del proceso científico con una teoría inicial acerca de la naturaleza de los datos.

El desarrollo de la Ciencia es siempre progresivo a través de un proceso cíclico de aprendizaje guiado (Box y cols., 1978, 2005). La unión entre el diseño de la investigación y la estadística es necesaria; la estadística constituye la "*tecnología del método científico*" (Mood y Graybill, 1972) y el procedimiento de comprobación de hipótesis estadísticas requiere trabajar con un diseño de la investigación que garantice que las conclusiones que se obtengan no estarán invalidadas por factores no controlados en el diseño o por sesgos que amenazan la validez de los resultados del estudio.

Por lo tanto, un diseño de investigación es una estrategia de estudio que supone manipular o seleccionar variables, medir variables y también controlar variables de sesgo para poder encontrar una respuesta lo más cercana posible a la realidad de la cuestión de investigación. Así, es necesario que el investigador o investigadora tenga en cuenta en la fase de planificación de su estudio las posibles variables extrañas que podrían afectar a la calidad de los hallazgos ya sea para controlar su efecto, para reducir la varianza del error de las puntuaciones (en el diseño debe ser varianza de error aleatorio) o para provocar ambas cosas a la vez.

La validez interna del estudio (la calidad de los resultados) sólo podrá estar garantizada si las variables extrañas son variables controladas con el diseño de investigación, manteniendo constante su efecto sobre la relación entre las variables que forman la hipótesis de investigación y que se conocen como las 'variables explicativas' del modelo (variable independiente y variable dependiente).

Realizar una adecuada revisión del conocimiento previo permite descubrir posibles variables extrañas cuyo efecto sistemático distorsionaría los resultados y dicha revisión posibilita planificar su control en el diseño. Por ello, la reflexión crítica sobre la documentación que aporta el conocimiento previo es una competencia del

investigador o investigadora que favorece la planificación de la investigación y evita sesgos sistemáticos que amenazan la validez de los hallazgos.

Capítulo 5. Metodologías de investigación

Dolores Frías-Navarro

Universidad de Valencia

Índice

Citar el capítulo como:

Frías-Navarro, D. (2021). Metodologías de investigación. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

Toda investigación debe estar adecuadamente diseñada, eficientemente ejecutada, correctamente analizada, bien interpretada y claramente presentada y redactada. Todas estas competencias requieren un gran esfuerzo por parte de los investigadores y las investigadoras y la calidad de los hallazgos o resultados (su grado de validez) está en gran medida relacionada con la metodología de investigación que se ha podido aplicar en el estudio y con el diseño que se ha planificado junto a su correcto desarrollo, análisis e interpretación.

Como ya se ha comentado, el proceso de la investigación científica (método científico) comienza porque el investigador o investigadora se plantea un problema concreto ('necesidad de conocimiento') que debe resolver ofreciendo una explicación del fenómeno lo más válida posible o, quizás, porque plantea una nueva explicación de los hechos que requiere recoger datos para avanzar en esa línea de investigación o, tal vez, podría tratarse de la necesidad de replicar un estudio o de reproducir los resultados de una investigación. Este problema le conduce a elaborar un plan de investigación cuyo objetivo principal es obtener observaciones y datos que sean relevantes y válidos para comprobar la hipótesis empírica planteada que operacionaliza a la hipótesis científica o sustantiva del estudio.

El plan de investigación (la planificación del diseño del estudio) requiere que el investigador o investigadora tome decisiones respecto a la estrategia de recogida de datos que depende, fundamentalmente, de aspectos metodológicos y técnicos como "*la forma de operativizar las variables de la hipótesis, la posibilidad o no de manipular la variable independiente, la capacidad de controlar las variables extrañas y de confundido, el grado de selección aleatoria de las unidades de observación (por lo general, sujetos), su asignación a las diferentes condiciones o niveles de actuación de la variable independiente, la estructuración interna del procedimiento, etc.*" (Arnau, 1989, p. 585).

Una vez resueltos estos aspectos, y muy especialmente la posibilidad o no de asignar aleatoriamente los niveles o condiciones de la variable independiente a los grupos de estudio y la de su manipulación, el diseñador o diseñadora de la investigación debe entonces "seleccionar la estrategia más adecuada para obtener los datos".

Es decir, se trata de determinar la metodología del estudio así como los sistemas de registro para la medida de la variable dependiente. Son los dos aspectos clave del proceso del diseño de investigación.

Por lo tanto, como consecuencia del planteamiento de la solución del problema, el investigador o investigadora tendrá que utilizar la modalidad de investigación (metodología) que sea más adecuada o posible para solucionar el problema planteado (necesidad de conocimiento).

El método científico de investigación como procedimiento general de obtener información es único y el cumplimiento de sus principios permite alcanzar el conocimiento científico, pero cada problema de investigación requiere una determinada actuación científica que hace que exista una pluralidad de modalidades o métodos de investigaciones vinculados directamente con la elección de la estrategia de recogida de datos. Metodologías que pueden ser cuantitativas, cualitativas o también mixtas donde se produce una combinación de análisis cuantitativo y cualitativo.

Para poder contrastar adecuadamente la validez de los enunciados propuestos en un estudio, los métodos (metodologías) de investigación tienen que garantizar que los procedimientos que se aplican aseguran dicho fin, no siendo indispensable que utilicen procedimientos matemáticos, aunque su uso puede reducir las ambigüedades a la hora de la contrastación con los datos así como clarificar la relación entre las variables.

Diferencias entre las metodologías de investigación

Pero, ¿qué diferencia a las distintas metodologías de investigación cuantitativa? Las diferencias entre las metodologías de investigación no es ni el contenido de estudio, ni la población de referencia que se estudia, ni el contexto en el que se estudia; las diferencias radican en la respuesta dada a dos interrogantes:

1) *Manipulación* (sí / no) de las condiciones de la variable independiente o factor de tratamiento. Es decir, manipulación (variación controlada por el investigador o investigadora, se trata de una variable activa) de las condiciones de la variable independiente o, en cambio, se trata de una variable asignada no manipulada (variable seleccionada por el investigador o investigadora).

2) *Asignación aleatoria* (sí / no) de las condiciones de la variable independiente. Es decir, aleatorización en la asignación de los tratamientos a las unidades experimentales (sujetos) en los diseños de medidas independientes o entre-sujetos (entre-grupos) o, quizás, aleatorización en el orden de presentación de las condiciones de la variable objeto de estudio (medida) en los diseños de medidas repetidas o intra-sujetos.

Hay que destacar que la denominada metodología *experimental* es la única que incluye ambos requisitos. La investigación *cuasi-experimental* únicamente cumple el requisito de manipulación de la variable independiente y la metodología *no experimental* (diseños de encuesta y los diseños observacionales) no cumple ninguno de los dos.

Los *verdaderos diseños experimentales* (en términos de Campbell y Stanley, 1966) son los experimentos aleatorizados, es decir aquellos en los que además de la manipulación de la variable independiente existe una asignación aleatoria de los sujetos a las condiciones de tratamiento.

Por lo tanto, los dos elementos clave que identifican la naturaleza de la metodología empleada en el diseño de una investigación y la naturaleza causal o no de las relaciones encontradas entre las variables (ver Figura 11) son las características de:

1) manipulación o no manipulación de las condiciones de la variable independiente (tratamiento) y

2) la posibilidad de la asignación aleatoria del tratamiento a las unidades experimentales (generalmente son sujetos en los diseños entre grupos o la asignación aleatoria en el orden de administración del tratamiento en los diseños de medidas repetidas)

Nos encontramos en el apartado de Método de investigación en la pirámide del proceso del diseño de investigación donde se describe la metodología del estudio: *experimental*, *cuasi-experimental* o *no experimental*.

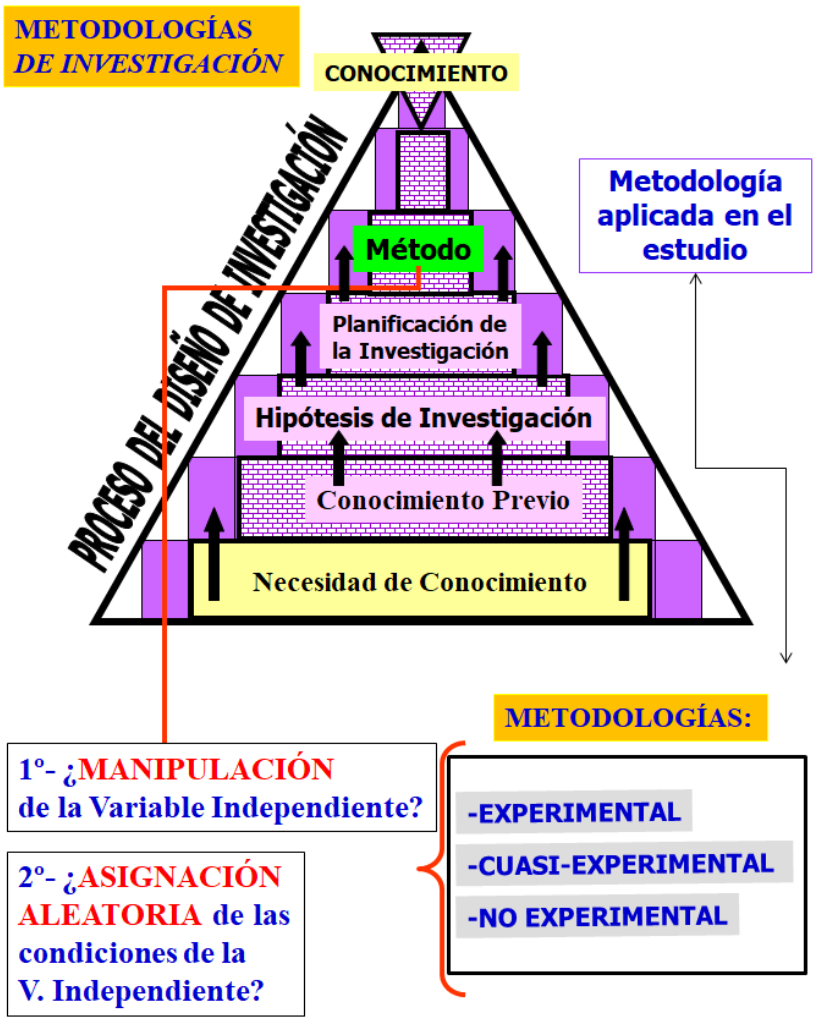


Figura 11. Metodologías de investigación

La conocida subdivisión de Campbell y Stanley (1966) clasifica las metodologías de investigación en experimentales, cuasi-experimentales y correlacionales. Correlacional no en el sentido analítico como técnica estadística sino como búsqueda de relaciones de covariación o asociación entre las variables, donde la variable independiente es únicamente observada y no es manipulada. Cook y Campbell (1979) introducen el término cuasi-experimental y con el trabajo de Pedhazur y Schmelkin (1991) se presenta una acertada clasificación tripartita que obvia el término correlacional, eliminando así la ambigüedad que el término conlleva al vincularlo a la técnica estadística, distinguiendo entre metodología experimental, cuasi-experimental y no experimental (incluyendo en esta última categoría a los estudios observacionales y los de encuesta). Los diseños de *N* = 1 son tratados como un método de investigación con características particulares de tamaño muestral, recogida y análisis de los datos y control de la calidad de los hallazgos.

En resumen, el criterio de clasificación de las metodologías de investigación en Psicología de Pedhazur y Schmelkin (1991) de metodología *experimental*, metodología *cuasi-experimental* y metodología *no experimental* se basa en dos cuestiones que se deben poder identificar de forma clara en el informe de investigación o artículo (Frías-Navarro, 2011):

1) Presencia o no de la manipulación de las condiciones de la variable de tratamiento o variable de intervención (variable independiente).

2) Utilización o no de la asignación aleatoria del tratamiento a las unidades experimentales (generalmente sujetos) que formarán los grupos de comparación o condiciones del diseño de investigación. Es decir, asignación aleatoria del tratamiento a los sujetos en los diseños entre-sujetos. O, quizás, asignación aleatoria en el orden de la presentación de las medidas o condiciones de la variable en los diseños intra-sujetos o de medidas repetidas.

Metodologías: experimental, cuasi-experimental y no experimental

En la metodología experimental la manipulación de la variable independiente se efectúa sobre las condiciones bajo las que se generarán los datos, es decir, implica que el investigador o investigadora cree situaciones de investigación (condiciones de la variable independiente; por ejemplo configuración de las pautas de intervención del grupo experimental que recibe la intervención y elaboración de las características del grupo de control que podría ser, por ejemplo, permanecer en una lista de espera) para observar cómo funciona la variable medida (variable dependiente; por ejemplo sintomatología depresiva) en cada condición de investigación.

Además, en la metodología experimental debe existir una asignación aleatoria del tratamiento o condiciones de la variable independiente o una asignación aleatoria en el orden de presentación de las medidas repetidas.

En la tabla 1 se resumen las principales características de la metodología experimental, cuasi-experimental y no experimental y, además se sitúa al ensayo clínico controlado aleatorio como sinónimo de un diseño con metodología experimental.

Tabla 1. Metodologías de investigación

Características	Metodología		
	Experimental *Ensayo clínico controlado aleatorio*	*Cuasi-experimental*	*No experimental* (Encuesta y Observacional)
Manipulación de variables	Sí	SÍ	No
Asignación aleatoria	Sí	No	No

Por lo tanto, en los estudios con una metodología experimental, el investigador o investigadora asigna aleatoriamente las condiciones de tratamiento manipualdas a los sujetos (se conoce también como '*ensayo clínico controlado aleatorio* ECCA o '*ensayo clínico aleatorio'*, ECA).

En cambio, en los estudios con una metodología cuasi-experimental aunque sí que existe una intervención/manipulación de la variable independiente, no existe la asignación aleatoria del tratamiento a los sujetos. Es decir, en los diseños con una metodología cuasi-experimental, los grupos o condiciones de la investigación no se forman de forma aleatoria sino que un criterio no aleatorio determina la configuración de los miembros que forman la condición o grupo de participantes.

Y, en los diseños con una metodología no experimental solamente se observan las variables tal y como ocurren en la realidad bajo las condiciones que ya tienen asignadas o son propias del grupo, pues no existe manipulación / intervención sobre la variable considerada independiente y por ello no es necesario plantearse la posibilidad de la asignación aleatoria o no del tratamiento; por ejemplo, la edad, ser hombre o mujer, la orientación sexual, el nivel socioeconómico, la personalidad o el bienestar percibido. Se puede identificar a la metodología no experimental como la metodología "NiNi": "*Ni manipulación Ni asignación aleatoria*". Se trataría de 'grupos intactos', ya que no reciben ningún tipo de intervención que les pueda modificar su conducta. En el diseño de este tipo de estudios el objetivo es registrar las características de los grupos tal y como están formados y analizar sus posibles diferencias en las hipótesis que se plantean en la investigación. Utilizando esa terminología, la metodología experimental puede ser representada como "SíSí", es decir, sí tiene manipulación de la variable independiente y sí tiene asignación aleatoria del tratamiento a las unidades experimentales. Y la metodología cuasi-

experimental se puede representar como “CasiCasi”, es decir casi es una metodología experimental porque tiene manipulación de la variable independiente y casi es no experimental porque no hay asignación aleatoria del tratamiento o variable independiente ya que carece de intervención.

En resumen, cuando se habla del experimento como ‘experimento verdadero’, es decir, en el sentido de un plan de investigación sobre la naturaleza de un cierto fenómeno cuyo diseño incluye los elementos de la metodología experimental propiamente dicha, se está infiriendo que se dan las siguientes condiciones (Pascual-Llobell y cols., 1996):

1. La existencia de, al menos, una variable manipulada, denominada variable independiente. La manipulación supone que el investigador o investigadora puede seleccionar varios valores de dicha variable, delimitando para cada valor una condición experimental distinta.

2. La asignación al azar de las distintas condiciones experimentales a las unidades experimentales en los diseños entre-grupos. Es muy importante no confundir número de observaciones (*N*) con número de sujetos (*S*), ya que únicamente es verdad en los denominados diseños entre-sujetos. En los diseños intra-grupos la condición son cada una de las mediciones que se realizan (medición 1, medición 2…) y se produce la asignación al azar en el orden de presentación de los tratamientos. Por ejemplo, en un diseño con 5 sujetos y 2 mediciones, el diseño tiene 10 observaciones ($N = 10$), siendo $S = 5$.

3. La comprobación del efecto de la manipulación de la variable independiente sobre la variable de medida, conocida como variable dependiente.

4. Además, se controla cualquier otra fuente de variación que no habiendo sido manipulada deliberadamente por el investigador o investigadora puede afectar de forma sistemática a la situación experimental. Se trata de las denominadas ‘variables extrañas’ o ‘terceras variables’ que deben ser eliminadas o mantenidas constantes cuando se conocen y es posible, o aleatorizarlas con el propósito de homogeneizar su efecto en las diferentes condiciones o grupos, convirtiéndose así en ‘variables controladas’ en el diseño de la investigación.

Las hipótesis causales (causa $\rightarrow$ efecto) son una clase de hipótesis muy importantes dentro de los intereses de investigación, pero no son las únicas, por

ejemplo las hipótesis de asociación (covariación entre las variables) cubren un área muy amplia de la investigación (fumar y cáncer; autoestima y éxito académico; ansiedad y rendimiento académico). Cook y Campbell (1979) señalan que la relación causal indica covariación entre la variable independiente y dependiente, precedencia temporal de la variable independiente o causa y explicaciones alternativas del cambio no plausibles. Siguiendo estos requisitos, el método experimental es el único que permite plantear y comprobar hipótesis de causalidad.

Por lo tanto, para inferir que A causa B se requieren tres condiciones (Kenny, 1979):

1) Establecer la asociación estadística entre los valores de A y B de manera que cuando se da la presunta causa aparece el presunto efecto y cuando no se da la presunta causa no aparece el presunto efecto.

2) Establecer la dirección de la causalidad por ejemplo de A a B, basado por ejemplo en un criterio temporal: precedencia temporal de la causa (manipulación previa de la variable independiente de tratamiento).

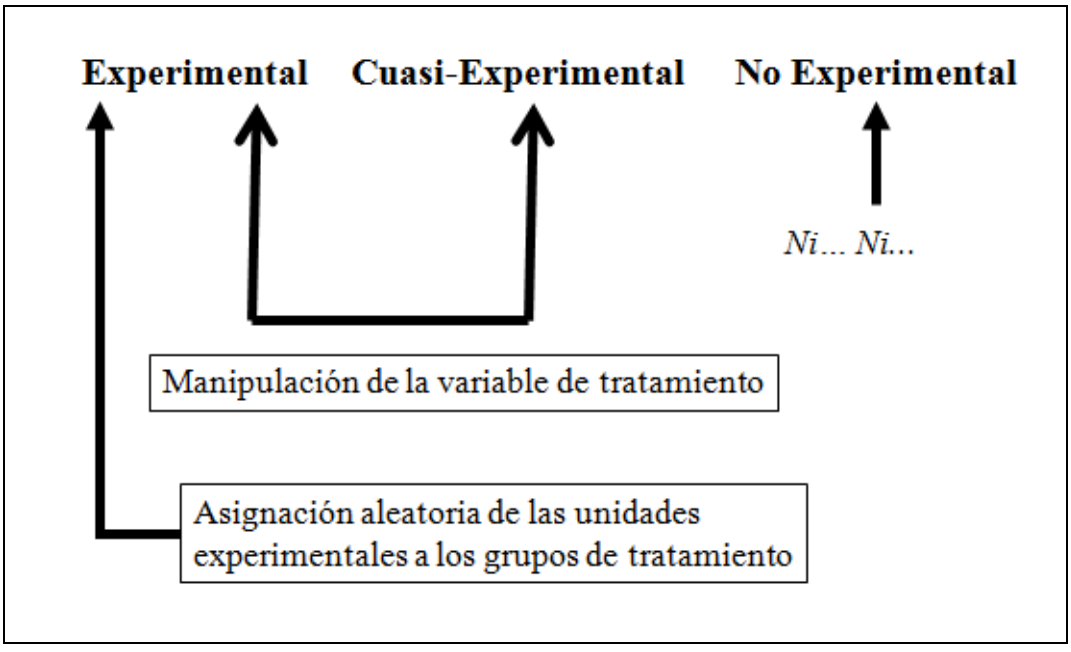
3) Eliminar los efectos de todas las posibles causas comunes de A y de B (ausencia de espuriedad), lo que implica el control previo de las variables extrañas, evitando explicaciones alternativas de los hallazgos.

Eliminar esas causas comunes (variables extrañas perturbadoras o contaminadoras) es un elemento clave en la definición del tipo de metodología. Generalmente los diseños emplean el control y la aleatorización para eliminar esas causas comunes de A y B. El control experimental supone mantener los valores de las posibles causas comunes constantes en todas las condiciones de tratamiento para que afecten por igual a todos los datos del estudio y se trata de que sí hay un efecto extraño a los intereses del investigador o investigadora sea un efecto constante en todas las condiciones o grupos que tenga el estudio. La aleatorización supone asignar los tratamientos aleatoriamente a las unidades experimentales y de esta forma las posibles causas comunes no afectan de forma sistemática a los resultados de la investigación, ya que se distribuyen de manera equilibrada. Es decir, no contaminan la relación entre las variables explicativas (variable independiente y variable dependiente) implicadas en la hipótesis de investigación. De este modo se

elimina / controla el sesgo de selección (error sistemático o varianza sistemática secundaria), ya que los efectos de las posibles causas comunes se distribuyen de forma azarosa en las diferentes condiciones de tratamiento y se supone por lo tanto que su presencia está equilibrada o compensada en los diferentes grupos o condiciones de tratamiento (grupos equilibrados).

Solamente la metodología experimental permite plantear y comprobar hipótesis de causalidad propiamente dichas. En la metodología cuasi-experimental y no experimental el control del sesgo de selección requiere mantener constantes todas aquellas variables que téoricamente podrían atentar contra la validez de los resultados, pues diferenciarían a los sujetos y, por lo tanto, ocultarían se hubo o no un efecto del tratamiento o quizás lo confundirían.

En resumen, en la *metodología experimental* es el investigador o investigadora quién decide qué tipo de intervenciones va a evaluar en su estudio, provocando con ello la manipulación de la variable independiente con la intención de modificar la vida de los sujetos. Posteriormente, asigna aleatoriamente las condiciones de estudio creadas a los sujetos cuando se trata de un diseño entre-grupos (*diseños entre-grupos*) o, cuando los participantes reciben más de un tratamiento, aleatoriza el orden de presentación de los tratamientos cuando se trabaja con un *diseño intra-sujetos* o de medidas repetidas (ver Figura 12).



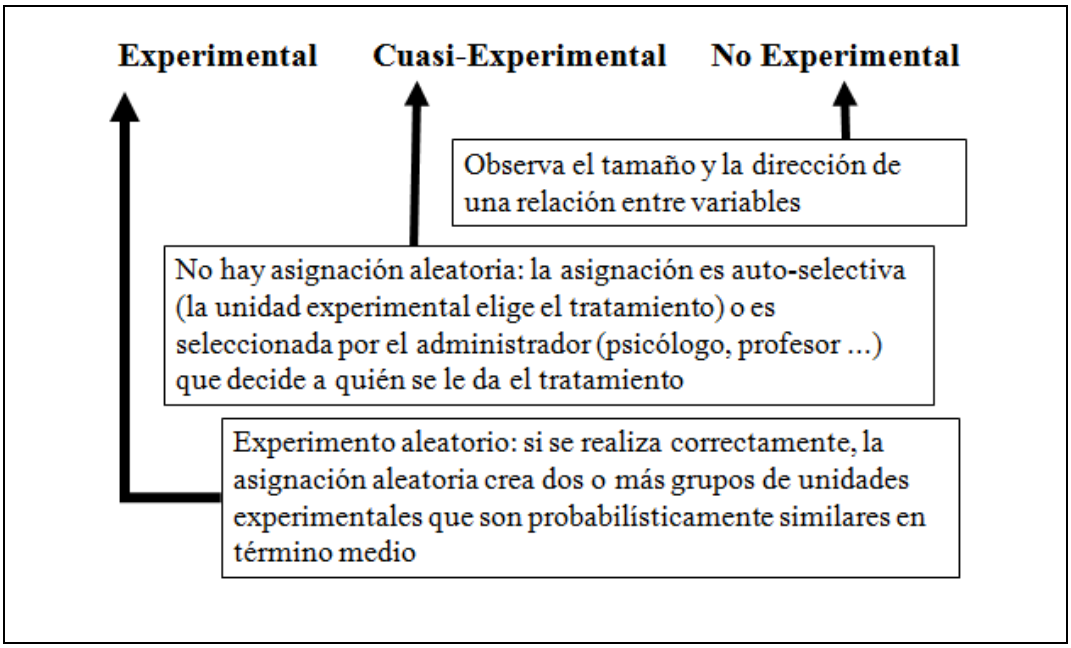


Figura 12. *Características de las metodologías de investigación*

En la metodología *cuasi-experimental* sí existe una manipulación de las causas o variables independientes, pero el tratamiento no se asigna de forma aleatoria a las unidades experimentales o sujetos o el orden de presentación de los tratamientos no es aleatorio. Se trabaja con grupos ya creados previamente, conocidos como 'grupos intactos' que no pueden formarse al azar. Aquí el control de las posibles variables de sesgo es menor que en la metodología experimental y requerirá técnicas de control específicas como la constancia (*diseños de bloques* por ejemplo), la eliminación de ciertas variables elaborando de forma detallada determinados criterios de inclusión y exclusión de los participantes o del contexto de investigación o diseños más sofisticados como el *diseño con variables covariadas* o la utilización de *diseños con puntuaciones de propensión*.

En la metodología *no experimental* no existe manipulación de variables (por lo tanto no es posible la asignación aleatoria y como consecuencia tampoco la asignación aleatoria), solamente se observa el tamaño y la dirección de la relación encontrada entre las variables. No se produce ningún tipo de intervención o cambio deliberado sobre el sujeto, solamente se observa el constructo o variable tal y como la manifiesta el sujeto.

Otras clasificaciones de las metodologías de investigación

Dentro del área de la biomedicina, la clasificación más común de la metodología de investigación distingue entre estudios experimentales y estudios observacionales, diferenciándose por cómo se asigna el tratamiento (variable independiente): con aleatorización o sin aleatorización.

En los estudios experimentales el investigador o investigadora asigna aleatoriamente las condiciones de tratamiento a los sujetos (conocido como 'ensayo clínico aleatorio' o aleatorizado, ECA).

En los estudios observacionales no existe asignación aleatoria de manera que la formación de los grupos o los criterios de selección de los grupos de la población se basan en (ver Figura 13):

1. La exposición al factor de estudio (estudio de cohortes).
2. La presencia (casos) o no (controles) de la enfermedad o efecto (resultado) que es el objetivo a investigar (estudio de casos y controles).

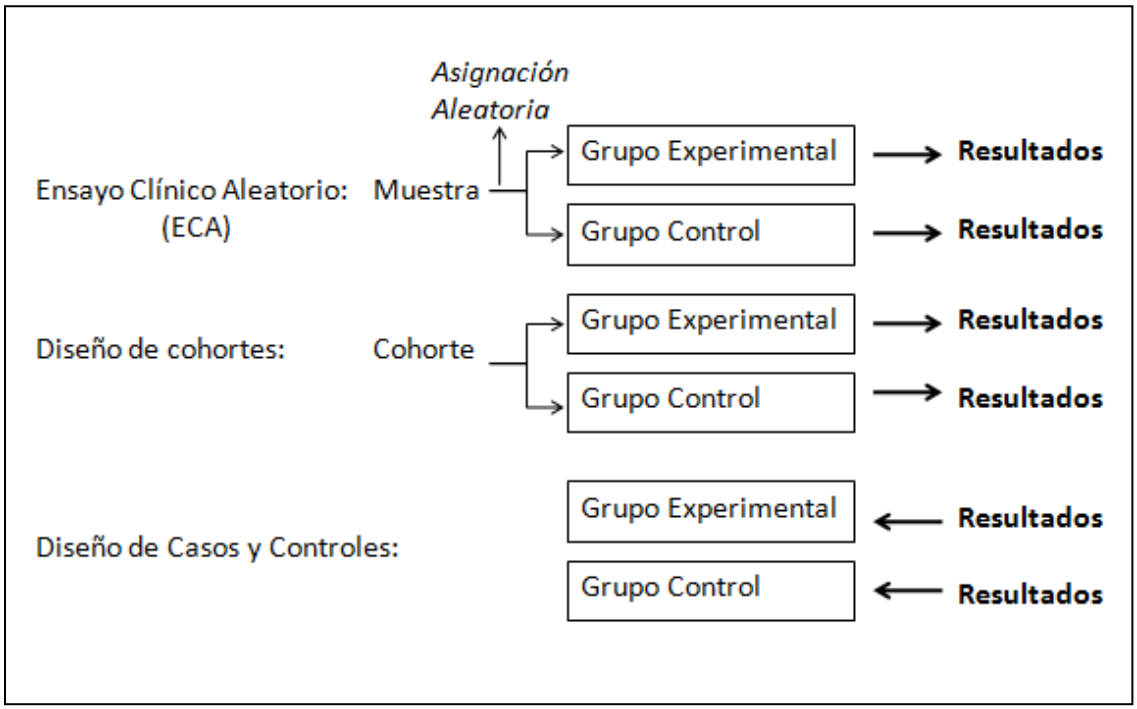


Figura 13. Configuración de los diseños ECA, cohortes y casos y controles

La principal diferencia entre el diseño ECA y un *diseño de cohortes* es la presencia o no de asignación aleatoria dado que en ambos casos el tratamiento antecede al efecto del tratamiento que se pretende evaluar.

En los diseños de cohortes se mantiene la secuencia temporal de los acontecimientos tratamiento $\rightarrow$ resultado y suele requerir tamaños muestrales grandes que son seguidos durante largos períodos de tiempo para establecer conexiones entre las variables. El diseño de cohortes no es muy recomendable cuando se trata de analizar acontecimientos que se presentan con baja frecuencia porque resulta muy costoso. Se trata de seleccionar una cohorte de sujetos que han sido expuestos a la variable de interés para poder comparar sus puntuaciones con otra cohorte de sujetos que no han sido expuestos a dicha variable y medirlos, preferiblemente durante periodos largos de tiempo para comparar sus puntuaciones en la variable dependiente.

A diferencia de los diseños ECA y los de cohortes, en *el diseño de casos y controles* se parte del resultado, es decir, si los sujetos tienen ('casos') o no tienen ('controles') el resultado de interés (el efecto o la enfermedad), y se evalúa retrospectivamente la presencia o no de los posibles tratamientos o exposiciones. Este diseño es especialmente útil para estudiar resultados poco frecuentes. Esa inversión temporal en el desarrollo de los acontecimientos es una de sus principales debilidades a la hora de poder establecer relaciones causales entre las variables de causa y efecto. Este diseño está limitado a una sola variable de resultado que es sobre la que se muestrea la población de casos y controles.

En resumen, se habla de 'experimentos aleatorios' o 'ensayo clínico aleatorizado' cuando hay manipulación de variables y asignación aleatoria y el 'experimento aleatorio controlado' señala que es un diseño aleatorio con un grupo de control para efectuar comparaciones entre el grupo experimental y el grupo de control. El clásico ensayo clínico o experimento aleatorio controlado se corresponde con la metodología experimental en las Ciencias Sociales y de la Salud y los diseños con un grupo de tratamiento y otro de comparación o control. En ambos casos es posible hablar de inferencia causal dado que hay control de la estrategia de asignación aleatoria de las diferentes condiciones de tratamiento.

Estudios de superioridad, estudios de equivalencia y estudios de no inferioridad

Los objetivos o propósitos de la investigación pueden estar dirigidos a evaluar si una intervención es más eficaz que otra, si su efecto es no inferior o si el efecto de ambas es similar (*superiority, non-inferiority, and equivalence trials*) (Frías-Navarro y cols, 2007). Se trata de los estudios de:

1) Superioridad. La superioridad de un tratamiento respecto a otros tipos de intervención o grupos de comparación (grupos de control o grupos de comparación), (intervención superior al control)).

2) Equivalencia. La equivalencia de los efectos de los tratamientos o grupos analizados (intervención similar al control).

3) No inferioridad. La no inferioridad de un determinado tratamiento (intervención no es peor que el control).

Los diseños de no inferioridad y de equivalencia consisten en comparar una nueva intervención con otra previa que se considera estándar, con el objetivo de demostrar que la nueva intervención no es inferior o es equivalente en sus beneficios clínicos. Los estudios de superioridad son los más aplicados en el ámbito de las Ciencias Sociales y de la Salud y sigue en aumento su aplicación mientras que los de equivalencia se han estabilizado y los estudios de no inferioridad tambien siguen en alza (Martínez-Franco y cols, 2021). Este tipo de estudios suele plantearse con estudios controlados aleatorios (ECA, metodología experimental), pero también podrían tener sentido con estudios cuasi-experimentales y no experimentales.

No siempre el objetivo de una investigación es determinar la superioridad de un tratamiento respecto a otro grupo de comparación o demostrar que existe una relación entre las variables. Además, desde un punto de vista ético, los estudios de superioridad puede que no sea adecuado llevarlos a cabo en algunas situaciones. Por ejemplo, puede ocurrir que exista un tratamiento que ya ha demostrado ser eficaz y seguro para tratar un determinado trastorno o enfermedad y no se considera ético formar un grupo control de pacientes a los que se le ofrece un nuevo tratamiento que podría ser peor u ofrecerles un placebo que no tendrá efecto. En esta situación no sería ético crear un grupo de control ya que el equipo de investigación es consciente

de que existe un tratamiento que ha demostrado ser eficaz y seguro para los pacientes y no administrar dicho tratamiento puede suponer un riesgo elevado para dichos pacientes. En esta situación si se podría plantear un estudio con un nuevo tratamiento que pretende tener una efectividad similar a la del tratamiento ya probado (estudios de equivalencia) o que pretende demostrar que su efecto no es inferior al del tratamiento ya probado (estudios de no inferioridad), añadiendo algunos beneficios como la adherencia al tratamiento, menos costo, más cómodo de administrar, más seguro o menos efectos secundarios e incluso porque simplemente representa una nueva alternativa terapéutica que no requiere demostra de forma necesaria su superioridad en eficacia.

Estudio de superioridad

Generalmente, como ya se ha comentado anteriormente, en el ámbito de la Psicología, los investigadores y las investigadoras tienen como objetivo comprobar que un tratamiento es mejor que otro o que existe una relación entre las variables. Con el procedimiento clásico de inferencia estadística que habitualmente se aplica, se plantean hipótesis de diferencias entre los grupos o de relación entre las variables y nunca hipótesis de igualdad entre las puntuaciones de los grupos o de no diferencias. Puede resultar paradójica esta situación ya que dicho procedimiento se inicia asumiendo que no hay efecto de la intervención o que la relación entre las variables es nula. La mayor parte de los estudios que se publican en el ámbito de la Psicología se enmarcan en los estudios de superioridad.

En primer lugar, para comprobar el efecto de un tratamiento la mejor elección es planificar el diseño con una metodología experimental o ensayo clínico aleatorio (ECA; *Randomized Clinical Trial*, RCT) y si está dirigido a mostrar la superioridad de un tratamiento respecto a un grupo de comparación (grupo de control u otro tipo de intervención) recibe el nombre de estudio de superioridad.

El estudio de superioridad consta de un grupo de tratamiento o grupo experimental (donde se lleva a cabo la intervención que es de especial interés teórico para el investigador o investigadora) y se compara con un grupo de control o con un grupo de comparación que recibe otro tipo de intervención. Si hay asignación aleatoria se aplica una metodología experimental y si no es posible y se trabaja con grupos intactos (donde no existe una asignación aleatoria de la variable de

intervención o variable manipulada) se trata de una metodología cuasi-experimental, pero en ambos casos existe una manipulación de la variable independiente.

En un estudio de superioridad el proceso de contraste estadístico sigue siempre 2 pasos (hipótesis bidireccionales) (Chow y Liu, 2004):

1. En primer lugar, hay que demostrar que los grupos difieren de forma estadísticamente significativa.
2. Si se rechaza la hipótesis nula entonces hay que comprobar si el valor de la media del grupo de tratamiento es mayor que la del grupo control (si se valora que a mayor puntuación mayor es la mejoría en los sujetos) y entonces se podrá concluir que el grupo de tratamiento es superior al grupo de control.

En este tipo de estudios de superiordad, resulta más interesante comprobar la superioridad del tratamiento del grupo experimental respecto a un grupo de comparación que recibe otro tratamiento (grupo de control activo) que utilizar un grupo de control no activo (el tratamiento C es mejor que el tratamiento Z). La hipótesis nula (H_0) del procedimiento clásico de inferencia estadística frecuencial plantea, en general, que los efectos son iguales o que la diferencia entre los efectos es cero. En cambio, la hipótesis alternativa (H_1) plantea que el efecto es diferente a cero:

$$H_0: \mu_1-\mu_2 = \delta$$

$$H_1: \mu_1-\mu_2= = >\delta$$

Siendo $\delta = 0$

Cuando se utiliza un grupo de control tipo lista de espera o simplemente sin ningún tipo de intervención (grupo de control no activo) es más fácil detectar un efecto estadísticamente significativo, pero quizás de escaso valor sustantivo ya que un mínimo de mejoraría en el grupo experimental podría indicar esa diferencia, pero quizás no es suficiente desde el punto de vista de la magnitud del efecto o desde la perspectiva del juicio clínico o diferencia sustantiva. Es más interesante comparar las puntuaciones del grupo experimental con un grupo que recibe otra intervención (grupo de control activo), por ejemplo la que habitualmente se utiliza en ese ámbito.

También se puede planificar un diseño con una metodología experimental o ensayo clínico aleatorio (ECA) dirigido a demostrar que el efecto de una intervención es igual (equivalente) o es no inferior a otro grupo de comparación (estudios de no inferioridad), por ejemplo, la terapia habitual cuya eficacia ya ha sido demostrada con estudios de superioridad. En este caso se trata de un estudio de equivalencia o, también, de un estudio de no inferioridad.

Estudio de equivalencia

En segundo lugar, en un estudio de equivalencia el objetivo es plantear hipótesis de efectos mínimos, es decir, se trata de determinar si la diferencia entre los efectos de dos intervenciones se encuentra dentro de un intervalo pequeño (−Δ a +Δ, valores delta) identificado por los valores del tamaño del efecto que el investigador o investigadora considera triviales. La clave de este tipo de estudios es establecer ese margen de equivalencia (valor de delta): máxima diferencia entre los tratamientos que se considera irrelevante o con escasa relevancia clínica.

Por lo tanto, el estudio o ensayo de equivalencia se diseña para confirmar la ausencia de una diferencia concreta entre los tratamientos. En este caso es necesario establecer un margen de equivalencia clínica (denominado delta, Δ o D) definido como la diferencia máxima entre los grupos que es clínicamente aceptable de tal modo que una diferencia mayor que esa sería importante clínicamente. Si dos tratamientos son declarados equivalentes entonces el intervalo de confianza del 95% (el cual define la amplitud de diferencias plausibles entre los dos tratamientos) estaría enteramente dentro del intervalo −Δ a +Δ.

Cuando se habla de estudios de equivalencia no se afirma que las puntuaciones de los grupos sean idénticas o iguales. Se trata de demostrar que son similares ya que su diferencia se encuentra dentro de un intervalo que el investigador o la investigadora formula como una diferencia no sustantiva o trivial. El efecto de las intervenciones es lo suficientemente similar como para poder concluir que son equivalentes desde el punto de vista clínico. Desde el área de la Medicina, los fármacos conocidos como genéricos son fármacos que han demostrado tener un efecto similar al de los fármacos tradicionales asociados a determinadas farmacéuticas, siendo su coste económico menor y por ello prescritos en mayor medida.

Los estudios de equivalencia suelen ser usuales en el área de la Medicina y la elaboración de nuevos fármacos. Así, sería relevante plantear un estudio de equivalencia si un nuevo fármaco (una nueva intervención) fuera más fácil de elaborar o aplicar, tuviera menor costo y menos efectos secundarios. En esos casos sería interesante estudiar cómo funciona ese fármaco o esa intervención ya que podría sustituir a otros fármacos o intervenciones aunque su efecto terapéutico no fuera mayor. A veces, también puede ser interesante analizar que la relación entre dos variables es escasa, planteando un intervalo de efectos de pequeña magnitud y el objetivo de la hipótesis sería comprobar si el efecto detectado en ese estudio se encuentra en ese intervalo y, por lo tanto, la relación entre las variables es trivial.

En los estudios de equivalencia la clave es tomar la decisión de qué es un tamaño del efecto trivial o pequeño. Su valor debe ser más pequeño que el del efecto que pueda ser sustantivo o clínicamente significativo. Algunos autores señalan que el valor del efecto trivial que se plantea en el estudio no debe ser mayor de la mitad del valor que se puede utilizar en un estudio de superioridad (Jones y cols, 1996). Se podrá concluir que existe equivalencia entre los grupos si el intervalo de confianza para la diferencia entre los efectos de las intervenciones se encuentra dentro de dicho intervalo. En este tipo de estudios de equivalencia, la hipótesis nula plantea que existe una diferencia de al menos la específica en el tamaño del efecto trivial planteado y el objetivo es refutar esa hipótesis y aceptar la hipótesis alternativa que plantea que no existe esa diferencia. Como se observa, los roles de las hipótesis estadísticas se han invertido.

Estudio de no inferioridad

En tercer lugar, en un estudio de no inferioridad el objetivo es mostrar que una intervención no es peor que otra intervención que se lleva a cabo en un grupo de comparación. Es decir, no es menos efectiva dentro de un margen de no inferioridad. No se trata de demostrar la equivalencia sino de mostrar la no inferioridad, es decir, demostrar que el tratamiento A no es peor que el tratamiento B. En estos estudios no se plantea como objetivo comprobar si el tratamiento experimental es más efectivo que el del grupo control. La hipótesis nula plantea que la diferencia entre los efectos de las intervenciones es inferior a un determinado valor de tamaño del efecto que será especificado por el investigador o la investigadora en el margen de no

inferioridad. Si la hipótesis nula se rechaza se concluye que el efecto del tratamiento A no es inferior al del tratamiento B.

Por lo tanto, en el estudio o diseño de no inferioridad se desea demostrar que un nuevo tratamiento no es menos efectivo que un tratamiento existente (podría ser más efectivo o tener un efecto similar). En el desarrollo de los fármacos, los diseños de no inferioridad son más comunes en la fase III que los ensayos de equivalencia. Con los estudios de no inferioridad las diferencias sólo son posibles en una sola dirección. El intervalo de confianza debe encontrarse a la derecha del valor –Δ.

La principal dificultad para llevar a cabo este tipo de estudios es escoger el margen de no inferioridad adecuado. Esta cuestión también se plantea cuando el investigador o investigadora tiene que definir el margen de similaridad en los estudios de equivalencia. En ambos casos el margen no debe ser muy amplio y debe basarse en un planteamiento teórico contextualizado en el área del fenómeno a estudiar. Por ejemplo, se puede consultar una revisión sistemática donde el estudio de meta-análisis ofrezca el valor del efecto de la intervención estándar o ya probada respecto a un placebo y utilizarlo para fijar la fracción (tamaño del efecto) que se desea obtener con el nuevo tratamiento como magnitud de no inferioridad. Si el margen es muy amplio se podría concluir que un tratamiento no es inferior en eficacia, no siendo realmente mucho mejor que el grupo placebo, por ejemplo. Por lo tanto, el margen del intervalo de no inferioridad debe ser lo suficientemente pequeño para poder valorar que el nuevo tratamiento tiene un verdadero valor terapéutico. Los autores señalan que en los estudios de no inferioridad el margen no debe ser menor del 50% del efecto que hubiese logrado la intervención estándar o ya probada si pudiera compararse de manera simultánea con el placebo en ese mismo estudio (Estrada-Pérez y Jaimes-Barragán, 2013).

Por lo tanto, este tipo de estudio es apropiado cuando se desea evaluar la eficacia de un tratamiento en fase experimental respecto a un control activo (tratamiento ya probado) y se plantea en la hipótesis que el tratamiento experimental no es necesariamente más efectivo que el tratamiento ya probado, pero su efecto sí es ‘no inferior’ estadísticamente hablando. Con otras palabras, el tratamiento experimental no es menos efectivo que el tratamiento de referencia o ya probado.

Si el límite inferior del intervalo de confianza de la diferencia entre los efectos de las intervenciones está por encima del valor del efecto planteado (se rechaza la hipótesis nula) entonces se demuestra la no inferioridad del tratamiento respecto al grupo de control. Los estudios de no inferioridad suelen aplicarse cuando el nuevo tratamiento es menos costoso, menos invasivo o tienen menos efectos secundarios tal y como sucedia con los estudios de equivalencia. Y si se demuestra la no inferioridad del nuevo tratamiento en cuanto a eficacia sería interesante valorar su aplicación como una nueva intervención ya que se evitarían los efectos indeseados de la intervención tradicional.

En este tipo de estudios de no inferioridad el valor del intervalo superior no es un tema que sea de interés para la investigación y por ello se ejecuta como una prueba unidireccional (una cola), requiriendo un número menor de observaciones que en los estudios de equivalencia (generalmente son bidireccionales o contrastes a dos colas). Es decir, en este tipo de estudios no es necesario que el tratamiento sea mejor que el del grupo control, solo que no sea inferior y es más fácil demostrar la no inferioridad que la superioridad de un tratamiento.

Los estudios de equivalencia y los estudios de no inferioridad requieren una planificación tan exhaustiva como la de los estudios de superioridad así como planificar el tamaño de la muestra de forma adecuada para lograr los objetivos plantedos en sus hipótesis. Es muy importante tener en cuenta la adecuada planificación de la potencia estadística para detectar un determinado efecto ya que si no se detecta un efecto como estadísticamente signficativo podría ser debido a que dicho efecto no existe (efecto similar en ambos grupos), pero también a la escasa potencia o sensibilidad de la prueba estadística para detectarlo. Por ello, y de forma especial en los estudios de equivalencia, es necesario que el diseño tenga la suficiente sensibilidad para detectar alguna diferencia o efecto de los tratamientos que se comparan.

En resumen, los objetivos de la investigación y la formulación de las hipótesis dirigen qué tipo de estudio y análisis es el adecuado y los investigadores e investigadoras deben planificar la fase de análisis teniendo en cuenta el planteamiento de sus hipótesis. No es lo mismo preguntarse si un tratamiento es mejor que otro o que el efecto de un tratamiento no es inferir a otro que plantear si

un tratamiento es tan bueno como otro tal y como sucede en los diseños de equivalencia.

El grupo de trabajo CONSORT (Consolidated Standards of Reporting Trials) ha desarrollado el listado de comprobación CONSORT para los ensayos clínicos aleatorios junto con dos extensiones específicas para los estudios de equivalencia y no inferioridad (Piaggio y cols, 2006, 2012). Estos materiales están disponibles en http://www.consort-statement.org/ y son una herramienta muy útil para llevar a cabo una lectura crítica o activa de estos estudios y también son muy útiles como guía para que el investigador o investigadora planifique adecuadamente el proceso de diseño de investigación. En la dirección http://www.consort-statement.org/downloads/extensions se encuentran todas las extensiones que se llevan a cabo por el grupo de trabajo CONSORT.

Proceso de diseño de un estudio con metodología experimental

El diseño más sencillo de una investigación con una metodología experimental utilizando un *diseño entre-grupos unifactorial univariado* (diseño donde se trabaja con grupos de comparación formados por sujetos diferentes que tienen una sola medición por sujeto, con una sola variable independiente y una sola variable dependiente) incluye:

1) una variable independiente (A) con dos condiciones o grupos (a_1 y a_2)

2) una variable dependiente (Y)

Las unidades experimentales (generalmente se trata de sujetos) que quedan ubicadas dentro de cada condición o grupo deben ser todo lo similares que se pueda en todas las variables previas a la introducción o presentación del tratamiento o intervención. Con ello se trata de controlar las posibles variables contaminadoras o extrañas que afectan a la relación de las variables explicativas (independiente - dependiente) implicadas en la hipótesis de trabajo (amenazaría a la validez interna de los hallazgos).

En un diseño adecuadamente planificado, el objetivo es que exista homogeneidad entre las unidades experimentales que forman los grupos antes de la

introducción del tratamiento. Es decir, el objetivo es disponer de “grupos equilibrados o grupos homogéneos”, ya que deberían ser homogéneos en todas la variables antes de la introducción de la variable de tratamiento, dado que se formaron al azar (técnica de control de la aleatorización). Si así fuese entonces se podrá observar de forma correcta la magnitud del efecto o la magnitud de la relación entre las variables que forman la hipótesis del estudio.

Esa situación de investigación solo se produce cuando las condiciones de la investigación permitan que el azar actúe de forma adecuada, por ejemplo, con muestras grandes esa configuración de la homogeneidad es más probable. Sin embargo, a pesar de que se pueda disponer de unas condiciones óptimas de asignación aleatoria conviene tener presente que no es garantía absoluta de que los grupos sean equivalentes.

Por ello, si el investigador o investigadora conoce (por estudios anteriores o después de revisar la literatura) que una tercera variable o variable extraña puede ejercer un efecto contaminante sobre la relación entre las variables explicativas de la hipótesis de investigación (variable independiente y dependiente) entonces debe planificar su control directo en el diseño del estudio ya sea por ejemplo, manteniendo constante su efecto (diseño de bloques o apareamiento de dicha variable en los grupos), eliminando dicha variable, si es posible, del estudio (quitar esa condición, por ejemplo solo utilizar hombres) o quizás introduciéndola como una variable covariada que ajuste matemáticamente la relación entre la variable dependiente y la independiente (diseño de covarianza).

Para llevar a cabo la asignación aleatoria del tratamiento se puede utilizar, por ejemplo, la plataforma on-line ‘research randomizer’ (https://www.randomizer.org/) que es una herramienta gratuita que permite generar números aleatorios o asignar los participantes de forma aleatoria a las condiciones de la investigación.

En resumen, la técnica de control de la asignación aleatoria trata de homogeneizar a los grupos, ya que gracias al azar se supone que se distribuye aleatoriamente todas las variables (observadas y no observadas) que podrían crear diferencias entre los grupos o condiciones de la variable independiente y, solamente, el efecto de la variable independiente debería ser la causa de la diferencia entre ellos, si tal efecto existe.

Por supuesto, el éxito del procedimiento de la asignación aleatoria de la intervención está muy relacionado con el tamaño de la muestra, pues es necesario tener suficiente muestra para que actúe eficazmente el azar. Y, además, el azar es caprichoso y podría no ser efectivo, ya que, a pesar de utilizar la aleatorización, podría dar lugar a grupos no homogeneos. De ahí, la importancia de considerar el control directo de la variable extraña o contaminadora que teóricamente podría afectar por parte del investigador o investigadora utilizando otras técnicas de control, tal y como ya se ha comentado, como la eliminación de la variable extraña si es posible, o factorizando su presencia en el diseño como un factor más en la ecuación estructural tal y como ocurre en los diseños de bloques no aleatorios o controlando su efecto matemáticamente tal y como se lleva a cabo en los diseños con variables covariadas.

Cuando se trabaja con una metodología cuasi-experimental (hay una intervención o una variable manipulada, pero no hay asignación aleatoria de las condiciones) también es necesario controlar que las unidades experimentales o sujetos que se van a comparar en cada grupo sean lo más homogéneas posibles en todas las variables (control del sesgo) para poder observar con validez las relaciones o la diferencia en la variable que es objeto de estudio en la investigación. También la metodología no experimental está expuesta al problema de la falta de control de las diferencias entre los sujetos de los grupos que se comparan más allá de la diferencia que es objeto de estudio en la investigación.

Especialmente en aquellas situaciones de investigación donde hay ausencia de aleatorización en la creación de los grupos es importante definir de forma clara y precisa un listado de variables de inclusión y exclusión dirigido a construir grupos o condiciones donde se hayan controlado en la fase de planificación aquella variables que la literatura ha destacado como relevantes y que afectarían al estudio de la relación entre la variable independiente y dependiente que forman la hipótesis de investigación del estudio. Una vez ejecutados los criterios de inclusión y exclusión, el investigador o investigadora podría proceder con la recogida de los datos de su investigación. Esos criterios deben justificarse teóricamente y aplicarse siempre a priori, es decir, antes de recoger los datos. Y, por supuesto, deberán ser explicados en el informe de investigación con argumentos.

Por lo tanto, una vez se ha seleccionado una muestra de participantes (si el muestreo es probabilístico mucho mejor, ya que actúa la técnica de control de la aleatorización y se proporciona validez externa a los resultados), si es posible, se introduce la asignación aleatoria del tratamiento a dichas unidades experimentales o sujetos (las condiciones de la denominada variable de tratamiento o variable independiente A se asignan al azar a cada sujeto). Es decir, se trabajaría con una metodología experimental donde un grupo de sujetos recibe el tratamiento objeto de estudio (a_1 grupo experimental) y el otro grupo recibe otro tratamiento (a_2 grupo de control o grupo de comparación) que puede ser un tratamiento de comparación, un tratamiento de placebo, un tratamiento de lista de espera, otro tipo de tratamiento… Al final del estudio y si los grupos están balanceados en todas las variables previas al tratamiento (grupos previamente homogéneos o equivalentes, es decir, se han controlado las variables extrañas), cualquier diferencia entre las puntuaciones (variable dependiente Y) de los grupos posterior a la introducción del tratamiento se asume que estará causada por el efecto del tratamiento. Se trataría en este caso de un diseño con una metodología experimental, ya que tiene 1) manipulación de la variable independiente y 2) asignación aleatoria del tratamiento a los sujetos y, solamente con este tipo de metodología se podrán realizar interpretaciones causales entre las variables, es decir, interpretar que la variable independiente (manipulada y asignada al azar) es la causa del efecto detectado en la variable dependiente medida.

A continuación se detallan dos ejemplos de investigación donde se trabaja con una metodología experimental.

Ejemplo 1. En el año 2006 Dar-Nimrod y Heine realizaron un experimento para tratar de analizar si la exposición a teorías científicas afectaba al rendimiento de las mujeres en matemáticas, publicando sus hallazgos en la revista *Science*. En concreto, analizaron si tener la creencia de baja aptitud para las matemáticas de las mujeres afectaba a su rendimiento.

La hipótesis sustantiva mantiene que dicha creencia se convierte en realidad cuando la mujer cree que lo es. Los investigadores crearon tres grupos (A = 3, variable independiente o factor A con tres condiciones: a_1, a_2, y a_3) y asignaron de forma aleatoria a las mujeres a uno de los grupos (*randomization*, R, o aleatorización).

La variable que se manipuló (variable independiente A) fue la creencia sobre el origen de las diferencias entre sexos en el rendimiento en matemáticas (a través de la lectura de un texto): grupo a_1 origen genético, grupo a_2 origen basado en la experiencia o aprendizaje y grupo a_3 ideas estereotipadas generales.

Posteriormente se midió el rendimiento (porcentaje de respuestas correctas) en una prueba de matemáticas (variable dependiente Y).

Diseño con grupo de control equivalente / no equivalente

El diseño tuvo un pre-test y un post-test. Se trata de un *diseño pre-test/post-test con grupo de control equivalente*, ya que los tratamientos (las condiciones de la variable independiente) fueron asignados al azar y de ahí que se utilice el término equivalente para describir al grupo de control. Si no hay asignación aleatoria del tratamiento (metodología cuasi-experimental) se rotularía como un *diseño pre-test/post-test con grupo de control no equivalente*, ya que al decir grupo de control no equivalente se señala que no hubo asignación aleatoria en la distribución del tratamiento. El diseño de la investigación del ejemplo 1 se representa gráficamente en la Figura 14.

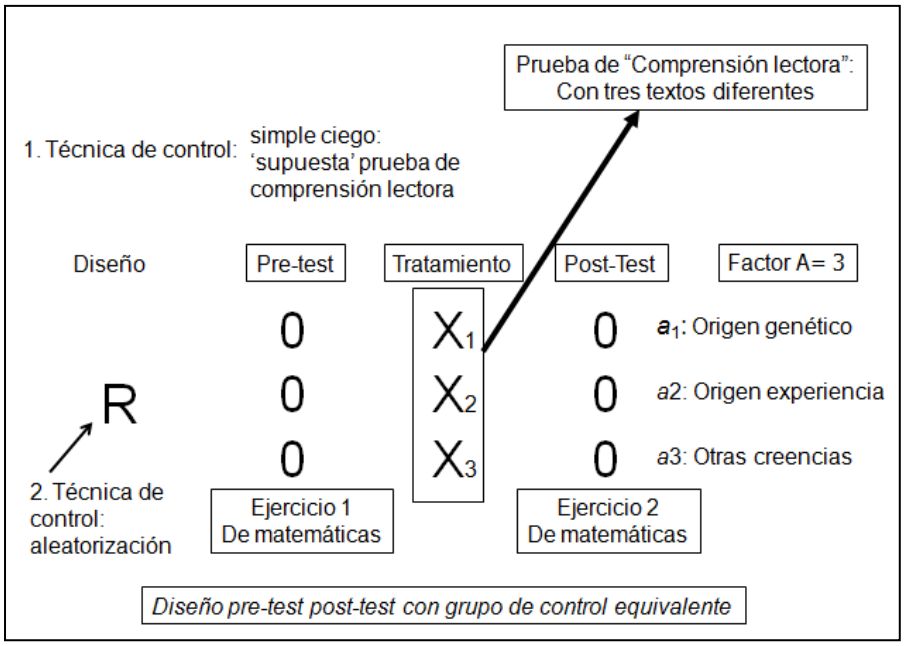


Figura 14. Configuración del diseño pre-test/post-test con grupo de control equivalente

Los resultados del estudio de Dar-Nimrod y Heine (2006) señalan que las mujeres sacan peor nota si creen sufrir una dificultad innata para las matemáticas.

Por el contrario, su rendimiento mejora cuando creen que las diferencias entre los sexos está provocada por la experiencia o el aprendizaje.

Como se ha comentado anteriormente, si la asignación hubiese sido no azarosa (metodología cuasi-experimental) entonces el diseño se formularía como *diseño pre-test/post-test con grupo de control no equivalente* (*non randomization*, nR, no aleatorizado). Es decir, se trata de una metodología donde hay manipulación de la variable independiente, pero no se produce la asignación aleatoria del tratamiento a los sujetos para formar los grupos de comparación.

En la investigación de Dar-Nimrod y Heine (2006), junto con la técnica de control de la asignación aleatoria, se utilizó la técnica de control denominada “simple ciego”, es decir, los sujetos o participantes del estudio desconocían el verdadero objeto de análisis del estudio que eran sus creencias sobre el origen de las diferencias en aptitud matemática entre los hombres y las mujeres y se les indicaba que se trataba de una prueba para medir la comprensión lectora.

En ocasiones, se pueden utilizar otras técnicas de control de cegamiento como el ‘doble ciego’ donde ni el sujeto ni el investigador o investigadora (por ejemplo) conocen qué tratamiento reciben los sujetos o el ‘triple ciego’ donde el sujeto, el investigador / investigadora y el evaluador / evaluadora (por ejemplo) no conocen el tratamiento qué recibe cada participante. En el informe de investigación es necesario detallar si se aplicó la técnica de control del cegamiento, quienes fueron cegados y cómo se llevó a cabo dicha técnica.

Ejemplo 2. Iván Petróvich Pávlov (26/09/1849-27/02/1936), fisiólogo y psicólogo ruso, fue Premio Nobel de Medicina en 1904 por su investigación sobre las glándulas digestivas y es reconocido muy especialmente por el descubrimiento del condicionamiento y los reflejos condicionados. En los estudio de Pavlov sobre el condicionamiento clásico se demostró que los perros tenían respuestas condicionadas ante un estímulo neutro (por ejemplo el plato de comida, un ruido o incluso los pasos del encargado al acercarse a los perros para ofrecer la comida) que había sido repetidamente asociado a un estímulo incondicionado como la comida. La respuesta condicionada de los perros

Iván Petróvich Pávlov
Imagen en:
https://www.psicologia-online.com/ivan-pavlov-biografia-y-teoria-del-condicionamiento-clasico-4680.html

ante el plato vacío de comida, ante el ruido de las llaves del encargado o al oír los pasos del encargado, hacía que los perros segregasen saliva sin la presencia del estímulo incondicionado de la comida. Se puede visionar el siguiente video en youtube donde se representa el experimento de Pávlov con su perro: https://youtu.be/kuAVOQixBl8 (3:06 minutos). Vídeo

Para mejorar el efecto del condicionamiento, los resultados de la literatura recomiendan que el estímulo condicionado y el estímulo incondicionado sean de la misma modalidad sensorial. La modalidad sensorial común podría ser por ejemplo el sentido de la vista tanto para el estímulo condicionado como para el estímulo incondicionado o podría ser el sentido del oído para captar ambos estímulos.

Supongamos que una investigadora desea replicar los hallazgos de Pavlov y planifica un estudio sobre la segregación salivar en perros. Elabora dos situaciones experimentales (tratamiento o variable independiente A = 2) con 10 perros en cada condición ($n_1 = n_2 = 10$; por lo tanto $N = 20$; 'diseño ortogonal', ya que hay el mismo número de obseraciones en cada condición o grupo; si no hubiese el mismo número de obseraciones se llamaría 'diseño no ortogonal'). En una de las situaciones asocia el estímulo incondicionado de visionado de la comida con una luz intermitente (condición del tratamiento a_1). En la otra situación vincula el estímulo incondicionado de visionado de la comida con un ruido intermitente (condición del tratamiento a_2). Posteriormente, asigna aleatoriamente el tratamiento a_1 a diez animales y el tratamiento a_2 a otros diez animales (metodología experimental; diseño entre-grupos o entre-sujetos).

Teniendo en cuenta el conocimiento teórico previo que existe sobre la relación entre la modalidad sensorial y el efecto del condicionamiento clásico (mejora si el estímulo condicionado e incondicionado son de la misma modalidad sensorial) se plantea la siguiente hipótesis de investigación, hipótesis científica o teórica o hipótesis sustantiva: los perros que reciben el tratamiento a_1 (misma modalidad sensorial del estímulo incondicionado y el estímulo condicionado, (visual-visual) mostrarán la respuesta condicionada más rápidamente que el grupo que recibe el tratamiento a_2 donde los estímulos son de diferente modalidad sensorial (visual-auditiva).

En este diseño de investigación del estudio de Pavlov con dos grupos de tratamiento la variable independiente manipulada es el tipo de estímulos apareados (A) con dos condiciones (a_1 misma modalidad sensorial y a_2 distinta modalidad sensorial) y la variable medida o variable dependiente (Y) es el tiempo que tarda el perro en manifestar la respuesta condicionada ante el estímulo condicionado. El procedimiento de asignación aleatoria de las condiciones de tratamientos a_1 y a_2 a los perros (todos los perros tenían la misma probabilidad de recibir el tratamiento a_1 que el a_2, es decir, un 50%) es la técnica que controla las posibles diferencias previas entre los animales que podrían afectar a los resultados de forma sistemática y es propia de la metodología experimental. La técnica de control de la aleatorización distribuye de forma equilibrada todas aquellas posibles variables que podrían diferenciar a los animales de forma sistemática antes de la introducción del tratamiento o condición de intervención, creando 'grupos equivalentes' u homogéneos. De este modo, la comparación de las puntuaciones que se realizará después de las intervenciones de condicionamiento entre los animales de la condición a_1 y la de los animales en a_2 permite estimar el efecto de ambas intervenciones y si hubiese una diferencia estadísticamente significativa en el tiempo que tarda en manifestarse el condicionamiento (variable dependiente, Y) se podría concluir que ello es debido a la vinculación entre modalidades sensoriales de la misma naturaleza o de distinta naturaleza (variable independiente A).

Los argumentos que apoyan que la metodología experimental ofrece interpretaciones de causalidad se basan en la presencia de la aleatorización en la creación de los grupos, el control del sesgo y la manipulación de variables. Así, dadas las características de asignación aleatoria y control de la influencia de posibles variables extrañas –se trabaja con grupos previamente homogéneos– cualquier diferencia que se produzca en los grupos después de realizar el experimento podrá ser atribuida a la manipulación experimental (Alvira y cols, 1980).

En otras palabras, para poder hablar de relaciones causales entre las variables, la asignación aleatoria de los sujetos a los grupos experimentales es un elemento clave dado que permite asumir que las condiciones experimentales creadas o manipuladas no están relacionadas, excepto por azar, con otras causas alternativas. Y la metodología experimental incluye por definición la asignación aleatoria, la manipulación de variables y el control de sesgos por aleatorización. Fisher (1935)

introdujo la asignación aleatoria en los diseños experimentales para asegurar que la única diferencia entre los grupos era la variable de tratamiento o variable independiente –eran grupos equivalentes estadísticamente hablando antes de la manipulación–, y por lo tanto si aparecían diferencias, su explicación vendría dada por su presencia.

En resumen, cuando se dice que un estudio se ha realizado con una metodología experimental se quiere señalar que se han creado, manipulado o controlado diferentes valores o condiciones de la variable independiente, a cuyos niveles las unidades experimentales (generalmente individuos humanos) son asignadas aleatoriamente, con objeto de observar si se producen diferencias en la variable dependiente medida. Además, los diseños experimentales son los únicos que permiten realizar interpretaciones de causalidad donde la variable independiente es la causa del efecto observado en la variable dependiente.

Metodología cuasi-experimental

Establecer relaciones causa-efecto es muy difícil cuando no es posible asignar aleatoriamente las diferentes condiciones del tratamiento a los sujetos. En esta situación determinar el efecto de un tratamiento es problemático dadas las diferencias existentes entre el grupo experimental y el grupo de control en la fase de línea base o antes de la administración del tratamiento. En estos casos se trabaja con la denominada metodología cuasi-experimental donde sí existe manipulación de la variable independiente (está sometida al control del investigador o investigadora), pero no es posible llevar a cabo la técnica de la asignación aleatoria (se trata de un diseño con grupo de control no equivalente). Se trabaja con grupos ya creados previamente, conocidos como ‘grupos intactos’ y por lo tanto con posibles causas comunes que se desconocen o no están controladas. El control estadístico de las variables contaminadoras (diseño con variables covariadas) o la introducción de las variables en el modelo de diseño de investigación (diseño de bloques no aleatorios) pueden ayudar a lograr una cuasi-aleatorización y con ello plantear hipótesis cercanas a la causalidad, pero que siempre deberán ser interpretadas con cautela. Esto supone trabajar con diseños más complejos y con una técnica estadística más sofisticada que de alguna manera atrape las diferencias previas entre los grupos, controlando el sesgo de selección; por ejemplo, con diseños basados en la

puntuación de propensión, diseños de bloques o análisis de la covarianza. Es decir, la falta de asignación aleatoria trata de ser compensada con una mayor sofisticación del diseño de investigación.

En opinión de Kirk (1995), el método cuasi-experimental se utiliza cuando no es posible la asignación aleatoria o cuando por razones prácticas o éticas es necesario utilizar grupos naturales o grupos ya formados como por ejemplo sujetos con una determinada enfermedad, sujetos que han sido sometidos a abuso sexual o sujetos maltratadores que reciben tratamiento de forma voluntaria.

En resumen, en la metodología experimental y en la cuasi-experimental hay una intervención bajo el control del investigador o investigadora dado que decide qué tipo de tratamientos va a evaluar en su experimento (manipulación de la variable independiente), pero sólo en la metodología experimental hay asignación aleatoria de las condiciones de la variable independiente a los sujetos o grupos. Qué supone esto. Pues que en la metodología cuasi-experimental la posible presencia de diferencias previas entre los grupos (sesgo de selección) es una fuente de sesgo que será necesario controlar mediante la planificación del diseño y la introducción de las posibles causas comunes en el modelo de investigación (por ejemplo con el diseño de bloques) o con el uso de herramientas estadísticas como por ejemplo el diseño de covarianza.

Metodología no experimental

En la metodología no experimental se encuentran los denominados estudios observacionales y los estudios de encuesta. En ambos casos no existe ni manipulación de la variable independiente -los valores de la variable independiente son sólo observados-–, ni por supuesto asignación aleatoria de las condiciones de la variable independiente a los sujetos ya que no hay intervención o tratamiento. Se trata de estudiar los fenómenos tal y como ocurren de forma natural (Anguera, 2010). El planteamiento de las relaciones entre las variables es de covariación y nunca causal. Como señalan Shadish y Cook (2002), un principio muy conocido en la investigación es que “la correlación no prueba la causación”. Es decir, si no se conoce qué ocurre en primer lugar (antecedente vinculado a la variable independiente) entonces las posibles explicaciones alternativas del desenlace (o consecuencia que se observa en la variable dependiente) se multiplican.

En los diseños no experimentales (en ocasiones referidos como diseños correlacionales aunque el término conduce a la confusión entre la técnica estadística y la metodología de investigación) se plantea la estimación de efectos, pero su estructura no permite identificar la causa y el efecto. No hay variables manipuladas ni tampoco efectos identificados como antecedente y consecuente. No hay asignación aleatoria. Su ejecución sólo permite conocer la magnitud y la dirección de los efectos estimados con el diseño de la investigación.

En los estudios llevados a cabo con una metodología observacional, la manipulación de los grupos es una característica propia de los individuos cuya naturaleza es precisamente el objetivo de investigación del estudio, configurándose entonces el grupo de tratados como grupo de 'expuestos' a la variable de interés y el grupo de no tratados como grupo de 'no expuestos' a dicha variable. Ahora ya no es el investigador o investigadora la persona que decide qué tipo de condiciones tiene el factor o la variable independiente pero sí es la persona que decide qué niveles de la variable independiente va a estudiar. Es decir, qué grupos de expuestos o de no expuestos a las condiciones de la variable independiente va a analizar en su trabajo. Por lo tanto, ahora los individuos ya tienen la característica que define al grupo y es justamente esa característica (sus diferencias) la que define el objetivo de la investigación. Por ejemplo, sujetos que toman alcohol frente a sujetos que no ingieren alcohol, sujetos fumadores o no fumadores, sujetos que reciben cursos para búsqueda de empleo o sujetos que no reciben el curso, una mujer o un hombre que trabajan fuera del hogar o están desempleados, madre o padre fumadora o madre o padre no fumadora, familia nuclear o familia con padres divorciados.

Por lo tanto, en los estudios observacionales (no experimentales) no existe una manipulación directa de la variable de tratamiento tal y como sí sucede en los diseños con metodología experimental y cuasi-experimental. Se trata de 'diseños prospectivos', ya que los sucesos o resultados han ocurrido antes de que el estudio comenzara. Por ejemplo, el sujeto ya es fumador o ya está desempleado. Por ello, la metodología no experimental se aplica en aquellos casos donde el investigador o investigadora no puede presentar los niveles de la variable independiente a voluntad propia ni puede crear los grupos experimentales por aleatorización, aunque sí puede introducir algo similar al diseño experimental en su programación de procedimientos para la recopilación de datos como el cuándo y el a quién de la medición (Campbell

y Stanley, 1963). Clásicos ejemplos de estudios observacionales son los de la salud y el tabaco o la ingesta de vitamina C y la prevención del cáncer.

Como ya se ha comentado, la gran ventaja de los experimentos aleatorios (metodología experimental) es la posibilidad que ofrecen para estimar de forma no sesgada los efectos causales dado que la asignación aleatoria del tratamiento permite que el grupo control y el grupo experimental sean iguales en términos medios en todas las características, tanto observadas como no observadas, siendo la variable de tratamiento la única diferencia entre ellos. Por supuesto, si la técnica de la asignación aleatoria funcionó de forma correcta.

Tanto en la metodología cuasi-experimental como en los estudios observacionales o no experimentales el investigador o investigadora no controla la asignación de los tratamientos a los sujetos participantes y por ello no puede estar seguro de si sujetos similares reciben tratamientos diferentes, es decir, si son grupos comparables entre sí. En estos casos el denominado 'sesgo de selección' contamina los resultados y por lo tanto amenaza a la validez interna de los hallazgos.

En los estudios de encuesta no existe intervención o tratamiento y por lo tanto no cabe preguntarse por el control de la asignación aleatoria. Se trata de estudios de opinión donde sí existe un control de la selección de la muestra si se aplica el muestreo probabilístico. Ejemplos de encuesta son las realizadas para conocer la intención de voto y poder con ello predecir los resultados de las elecciones, predecir el uso de productos o medir los cambios de opinión. Las cuestiones vinculadas con el muestreo aleatorio (probabilístico) son la clave metodológica que garantiza la validez de los resultados obtenidos. Definir la población objeto de estudio, determinar el tamaño de la muestra y el método de extracción de los elementos de dicha muestra y fijar el error muestral asumible en la investigación son las decisiones más importantes que el investigador o investigadora debe adoptar ante el diseño de un estudio con metodología de encuesta o selectiva.

Asignación aleatoria del tratamiento

La asignación aleatoria del tratamiento a cada uno de los sujetos (unidades experimentales) que forman la muestra de participantes tiene como objetivo garantizar que la posible función causal hallada al final del estudio no está provocada

por otras causas (denominadas 'terceras variables' o variables extrañas) ajenas al efecto de la variable independiente.

La ventaja de la técnica de la aleatorización es que trata de prevenir diferencias sistemáticas entre los grupos de cualquier tipo ya estén identificadas por el investigador o investigadora (conocidas por el investigador, denominadas variables covariadas observadas), o no identificadas (variables covariadas no observadas).

La asignación aleatoria implica que cada unidad experimental o participante tiene la misma probabilidad de formar parte de un grupo u otro. Por ejemplo, si el diseño tiene dos grupos (A = 2) entonces cada sujeto tiene una probabilidad de ½ de pertenecer al grupo a_1 y la misma probabilidad (½) de pertenecer al grupo a_2 (ver Figura 15). Gracias a la técnica de la aleatorización disminuye la probabilidad de diferencias sistemáticas previas al tratamiento entre los grupos experimental y control, es decir, hay un control del error sistemático o sesgo (varianza sistemática secundaria).

Figura 15. Proceso de asignación aleatoria de los sujetos (randomización)

En ocasiones puede ocurrir que la asignación aleatoria se produzca después de estratificar a los sujetos en alguna variable no aleatoria (por ejemplo en un diseño con variables bloqueadas donde la variable bloqueada es el factor de bloqueo) que se desea mantener controlada por constancia, configurando diseños 'parcialmente aleatorios' si la variable independiente o factor del diseño vinculado a la varianza sistemática primaria o del efecto se asigna al azar.

Por ejemplo, los sujetos pueden clasificarse previamente a la introducción del tratamiento en una determinada variable como el nivel de glucosa en sangre en la

línea base: alta, media y baja, y posteriormente se procede a la asignación aleatoria del tratamiento a los grupos experimental y control dentro de cada bloque. Es decir, con ello se controla que en todos los grupos de tratamiento hayan sujetos con los tres bloques de glucosa previa.

Así, los sujetos que forman el bloque de nivel alto de glucosa son asignados al azar a cada una de las dos condiciones de tratamiento, del mismo modo los del grupo de glucosa media y finalmente los del grupo de glucosa baja (ver Figura 16).

De este modo se garantiza la distribución homogénea de los distintos niveles de glucosa en el grupo experimental y en el grupo de control. Se trata en este caso de un *diseño de bloques con restricciones en la aleatorización* porque no todas las variables o factores del diseño se han asignado al azar (la glucosa previa en sangre es una variable asignada y por lo tanto no es posible aleatorizarla).

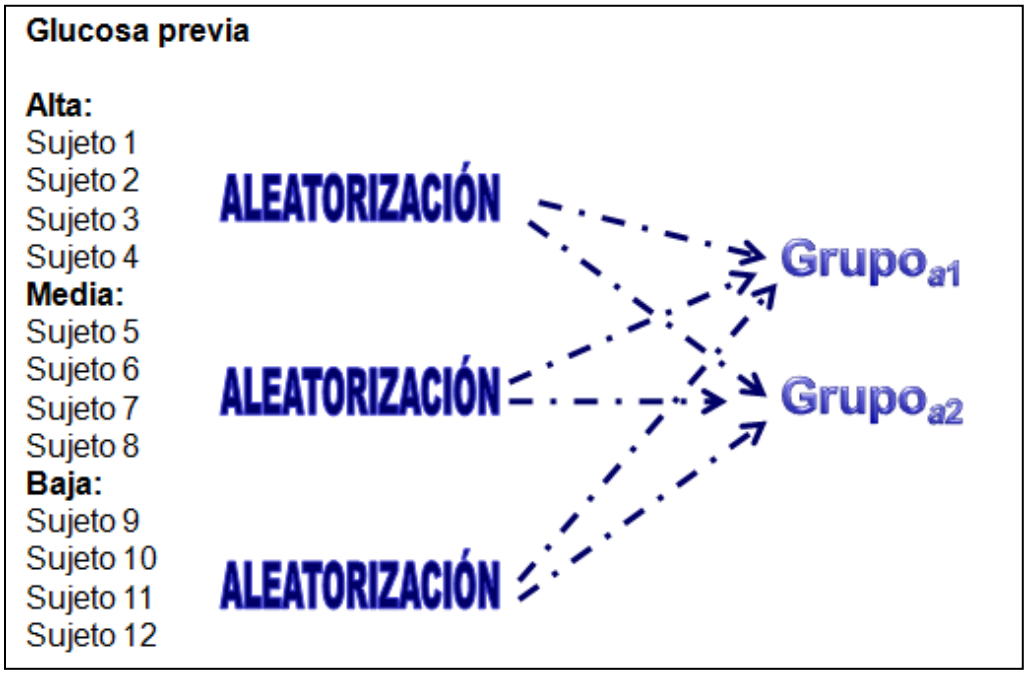


Figura 16. Bloqueo de la variable nivel de glucosa previa

Y el diseño *diseño de bloques con restricciones en la aleatorización* tendrá, al menos, dos factores: factor A vinculado a la variable independiente Grupo (factor de 'varianza sistemática primaria') y el Factor B vinculado a la variable de bloqueo que se controla (factor de 'varianza sistemática secundaria' controlada), configurando un *diseño entre grupos de bloques A x B parcialmente aleatorio*. En este punto conviene tener en cuenta que la descripción de la metodología en el informe o artículo será

como metodología experimental porque al menos hay una variable manipulada y asignada al azar, pero los lectores y lectoras deben estar atentos a cómo se interpreta el efecto de la variable que no ha sido manipuladas.

Evidentemente, no hay dos grupos que sean exactamente iguales antes del tratamiento. Incluso aunque se utilice la técnica de la asignación aleatoria pueden existir diferencias individuales o problemas relacionados con el muestreo (errores de muestreo) que provocan diferencias entre los grupos. Se trata del error muestral (varianza no sistemática).

La técnica estadística determina la probabilidad de que las diferencias observadas entre los grupos podrían ser debidas al mismo proceso de asignación aleatoria de las unidades experimentales que forman los grupos, es decir, al proceso aleatorio de extracción de las muestras. Si la probabilidad (valor *p* de probabilidad) del estadístico utilizado en la prueba del contrate de hipótesis del estudio es baja (generalmente $p \leq .05$) entonces se asume que las diferencias entre los grupos están causadas por el tratamiento interpretándose como 'diferencias estadísticamente significativas'.

Por el contrario, si la probabilidad de que las diferencias encontradas entre los grupos debidas al error de muestreo es alta ($p > .05$) entonces se asume que las diferencias observadas entre los grupos son debidas al mismo proceso de extracción aleatoria aleatoria de la muestra (error aleatorio o diferencias individuales) y se concluye que las diferencias encontradas entre las puntuaciones de ambos grupos 'no son estadísticamente significativas'.

Es decir, asumiendo que las grandes o las pequeñas diferencias entre las puntuaciones de las condiciones son debidas al azar o al error de muestreo (planteamiento de la hipótesis nula), se obtiene el valor *p* de probabilidad de los datos o resultado del estudio que se ha realizado y si es un valor pequeño se puede concluir que la variabilidad detectada se puede atribuir a algo más que al error de muestreo o al azar, concluyendo con el rechazo de la hipótesis nula que conduce, como consecuencia, a la aceptación de otra hipótesis llamada hipótesis alternativa.

Posteriormente se explicarán con detalle las hipótesis estadísticas de hipótesis nula e hipótesis alternativa.

Por lo tanto, disponer de grupos lo más homogéneos posibles en todas las variables antes de la introducción del tratamiento o variable independiente es fundamental para abordar con calidad los resultados aportados por el estudio (validez de los resultados). Y es la metodología de investigación del estudio la que determina la jerarquización de las pruebas o la evidencia aportada por los estudios basándose en la técnica empleada para controlar las variables y para formar grupos equilibrados o balanceados, es decir, grupos equivalentes.

En resumen, se habla de 'experimentos aleatorios', 'ensayo clínico aleatorizado', 'experimentos aleatorios controlados' o 'metodología experimental' cuando hay manipulación de variables y asignación aleatoria del tratamiento (mayor control de la validez interna). El clásico ensayo clínico o experimento aleatorio controlado se corresponde con la metodología experimental en las Ciencias Sociales y de la Salud. En ambos casos es posible hablar de inferencia causal dado que hay control de la estrategia de asignación aleatoria de las diferentes condiciones de tratamiento y, por lo tanto, se asume que las diferencias entre los grupos o condiciones se pueden atrbuir al efecto de la variable independiente manipulada.

Manipulación y aleatorización y, por supuesto, el mayor grado de control de la situación que implican ambos elementos, resume las características más sobresalientes de la estrategia experimental.

El concepto de manipulación de una variable supone una intervención deliberada o el control de los valores que se seleccionan y administran. Administrar un tratamiento equivale a definir un valor de la variable independiente y en el contexto del diseño de investigación no tiene connotación ética alguna. El elemento de la aleatorización permite controlar fuentes de sesgo y asegura la aplicación correcta de las técnicas estadísticas del contraste de hipótesis.

Si la aleatorización es completa se produce tanto en la selección de las unidades experimentales (relacionada con la validez externa de los resultados) como en la asignación a las condiciones de la variable independiente (relacionada con la validez interna de los resultados). De este modo, gracias a la 'selección aleatoria' de la muestra, el diseño permite trabajar con muestras representativas cuya información refleja estimaciones exactas de los parámetros que describen estadísticamente a la población (Arnau, 1989a). Las pruebas estadísticas están basadas en la asunción

del muestreo aleatorio a partir de la población, siendo esencial el muestreo aleatorio o la asignación aleatoria de las unidades observacionales a los grupos de tratamiento para la aplicación correcta de dichas pruebas, asegurando la independencia de las observaciones o de los errores así como la estimación correcta de los parámetros. En otras palabras, una prueba de significación estadística utilizada sin aleatorización no facilitará información válida sobre la probabilidad de un resultado bajo el modelo de la hipótesis nula (Shaver, 1993).

Diseño de *N* = 1

Existen otras alternativas metodológicas, como por ejemplo, el diseño de caso único o diseño de *N* = 1, que resulta difícil de incluir en cualesquiera de las distintas opciones de clasificación de los métodos de investigación que se han señalado anteriormente. Este diseño surgió dentro de la orientación experimentalista propia del conductismo, y debido a sus peculiaridades representa un método de investigación propio que es especialmente útil y muy utilizado dentro del ámbito de la psicología aplicada y en especial en la investigación clínica.

Un criterio amplio de clasificación de los diseños experimentales distingue entre diseño experimental clásico o fisheriano (diseño de *N* > 1) y diseño de *N* = 1. Cook y Campbell (1986) diferencian dos modelos de diseño experimental en función de la naturaleza del control que aplican, distinguiendo entre la tradición del control estadístico, propio de los diseños de comparación de grupos o *N* > 1, y la tradición del control y aislamiento experimental que caracteriza a los diseños de replicación intrasujeto o *N* = 1.

El diseño de *N* = 1 o diseño conductual ha recibido diversos rótulos; como señala Arnau (1995a), el diseño ha sido etiquetado como diseño de replicación intra-sujeto (Arnau, 1984; Gentile y cols., 1972), especialmente dentro del ámbito clínico, como, diseño de caso único (Barlow y Hersen, 1984; Kazdin, 2010), diseño operante (Sidman, 1973) o diseño de un sólo sujeto (Bordens y Abbot, 1988; Cozby, 1993). Nosotros nos hemos quedado con el título menos comprometedor de diseños de *N* = 1 (Pascual-Llobell, Frías-Navarro, y García-Pérez, 1995).

En el análisis histórico que realizan Barlow y Hersen (1984) se señala que la investigación experimental se desarrolló en dos direcciones:

- las investigaciones sobre grupos diferentes donde se estudia el grupo y su respuesta media como reflejo de la acción de la variable de tratamiento, actuando la aleatorización como principio fundamental de homogeneización de los grupos y
- las investigaciones sobre una sola unidad de observación (sujeto o grupo reducido) que es sometida a los distintos tratamientos en una situación de estricto control experimental donde la fase de línea base proporciona la puntuación de comparación y control de variables.

En estas últimas investigaciones, la replicación de los datos se basa en la reversibilidad de los tratamientos y, a diferencia de las investigaciones de grupos, la generalización de los resultados a la población de origen se hace difícil. Sin embargo, también es cierto que el objetivo prioritario de un experimento verdadero no es realizar estimaciones probabilísticas del efecto en la población (Berkowitz y Donnerstein, 1982) sino que se centra en inferir relaciones entre variables, eliminando posibles causas alternativas, posibilitando con ello la creación de teorías.

Desde una perspectiva vinculada con la tradición social aplicada y en conexión directa con la incorporación de la dimensión temporal que intrínsecamente el propio diseño incluye, los diseños de $N = 1$ son conocidos como diseños de series temporales (Arnau, 1995b) rotulados como diseño de series temporales interrumpidas por Campbell y Stanley (1966), quienes lo sitúan como método cuasi-experimental dada la ausencia de aleatorización, o identificados simplemente como diseños de series temporales (Glass, Willson y Gottman, 1975; Hayes, 1981; Kratochwill, 1978).

La ubicación de la naturaleza de la metodología de caso único no es unánime entre los investigadores y hay quienes lo sitúan en el ámbito cuasi-experimental y otros en el propiamente experimental. Por supuesto, existen ejemplos de aplicación de diseño de $N = 1$ que son perfectamente experimentales del mismo modo que también hay trabajos cuya metodología es cuasi-experimental (Cook y Campbell, 1979; Edgington, 1987).

Siguiendo la proposición de los autores que sitúan este diseño dentro del paradigma experimental (por ejemplo Arnau, 1995b; Barlow y Hersen, 1984; Kazdin,

1992), el diseño de *N* = 1 reúne las condiciones de método experimental de investigación siempre que su planteamiento incluya las condiciones de control estricto de los efectos nocivos provocados por variables extrañas, donde la fase de línea base tiene un protagonismo destacado como elemento de comparación. Además, y siguiendo los comentarios de Arnau (1995b, p. 177), el carácter experimental de los diseños de *N* = 1 se expresa en los siguientes términos:

"… se trata de estructuras donde no sólo tiene que estar presente el control de las posibles fuentes de confusión, sino la manipulación de la variable de tratamiento así como la correcta especificación de la variable de medida o de resultado. Comparte, pues con el diseño experimental clásico un objetivo común: evaluar la acción causal de la variable independiente y establecer el grado de impacto o efecto que ejerce dicha variable sobre alguna medida de respuesta del sujeto. Quizá podríamos destacar, como característica propia de esta estructura, el hecho que se suele utilizar como unidad de análisis a un solo individuo o a un reducido grupo de individuos, con la consiguiente repercusión que ello tiene en la generalización de resultados".

Los problemas de validez externa pueden ser obvios desde el momento que se utiliza un sólo sujeto (o pocos sujetos) y la generalización de los datos puede ser ampliamente cuestionada. En los diseños de *N* = 1 la replicación es la clave de la generalización y la validez interna del diseño.

La replicación implica generar nuevos estudios que especifican claramente las condiciones de tratamiento y medida que se desean replicar. Barlow y Hersen (1984) describieron tres tipos distintos de replicación: *directa* donde se utilizan los mismos tratamientos con nuevos sujetos, *sistemática* donde hay un cambio de las variables de interés como contexto, tipo de desajuste o trastorno… y *clínica* donde se comprueba el tratamiento con sujetos que presentan problemas conductuales semejantes.

Es evidente, que el primer tipo de replicación tiene que ver con la fiabilidad de la investigación y, por tanto, se relaciona principalmente con la validez interna de la misma. En cambio, los otros dos tipos de replicación tienen que ver con la validez externa. Los autores a afirmar que *"… una serie de diseños de caso único con sujetos semejantes a los que se les aplica el mismo tratamiento original -replicación-, puede*

sobrepasar en tres o cuatro veces el diseño experimental con grupos de tratamiento / grupo de control" (Barlow y Hersen, 1984, p. 57).

Otras amenazas contra la validez interna, por ejemplo, la interacción selección por tratamiento, son difíciles de controlar en este tipo de diseños: un sujeto particular con unas características individuales, puede reaccionar de forma específica al tratamiento, haciendo igualmente difícil la atribución causal que se pretende.

En estos diseños de $N = 1$, al igual que en los diseños cuasi-experimentales, hay que extremar las condiciones de control y tomar las precauciones necesarias para que factores tales como la historia, la maduración… no afecten seriamente a la validez de los resultados.

Otro tipo de amenaza contra la validez interna, y que puede afectar también a la validez de conclusión estadística, procede del hecho mismo de la repetición de la secuencia tratamiento - no tratamiento. Cuando a un sujeto se le dan múltiples tratamientos de manera serial o el mismo tratamiento en múltiples etapas, ¿hay interferencia? ¿Hay facilitación? ¿Los efectos de un tratamiento dependen del orden en que se hayan administrado? ¿Es diferente el efecto según sea en presencia o ausencia de un segundo elemento? De nuevo la salida airosa que, como solución general, se puede aportar parece venir de la replicación de secuencias alternativas.

Por otro lado, también las estrategias de reversibilidad del tratamiento, cuando es éticamente posible, son otro elemento importante cuya presencia apoya la validez interna del diseño. Por ejemplo, con los denominados diseños de reversión o diseños de retirada del tratamiento se puede obtener un alto grado de certeza del efecto de la intervención como agente responsable de los cambios observados en la conducta (Arnau, 1995c). Si con la retirada de la variable de tratamiento en la fase crítica se produce un cambio que se comprueba a lo largo de una serie de retiradas sucesivas (modificándose la conducta del sujeto cuando se aplica el tratamiento mientras que retorna al nivel de línea base cuando se retira dicho tratamiento) entonces se puede inferir la efectividad del tratamiento como causante de cambio.

Por último, hay una serie de creencias incorrectas respecto a los diseños de $N = 1$ que crean confusión, destacando especialmente dos:

En primer lugar, no hay que confundir los diseños de $N = 1$ con los estudios de caso tradicionales, ya que en este último tipo de estudios el análisis se lleva a cabo

a través de un único sujeto, pero sus objetivos son exploratorios, no existiendo suficiente control de la situación.

En segundo lugar, en ocasiones se asocian las investigaciones que utilizan un amplio número de sujetos con análisis estadísticos mientras que las que incluyen un único sujeto (pocos sujetos) tienen que estar asociadas necesariamente a análisis visual de los resultados. Es un grave error. Los estudios de $N = 1$ pueden ser analizados con modelos estadísticos y los que incluyen un $N > 1$ también pueden ser abordados con modelos no estadísticos como las técnicas de inspección visual (Kratochwill, 1978; Martínez, 1988).

Capítulo 6. Validez de los resultados de la investigación

Dolores Frías-Navarro*

Marcos Pascual-Soler**

*Universidad de Valencia

**ESIC Business & Marketing School, España

Índice

Citar el capítulo como:

Frías-Navarro, D. y Pascual-Soler, M. (2021). Validez de los resultados. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

El contenido nuclear de la metodología de investigación está definido por el concepto de validez de los resultados junto con los criterios para definirla y evaluarla. En consecuencia, un programa de formación en metodología se ha de desarrollar en torno a este núcleo y sus variantes, donde la validez interna y la validez de conclusión estadística son quizá las más importantes, si nos atenemos a la necesidad de rigor, de certeza y precisión, pero sin ignorar la validez externa y la de constructo como elementos esenciales para planificar y ejecutar de forma correcta el diseño de la investigación.

La validez asociada a los resultados (a la inferencia de los resultados) de la investigación determina la calidad de los hallazgos de un estudio. Dicha calidad va a poder ser jerarquizada y no es una cuestión de todo o nada, como ya anteriormente se ha comentado.

La validez está vinculada a la calidad de las pruebas o evidencia encontrada en el estudio. Es decir, la validez está relacionada con la credibilidad de los resultados de la investigación y su aplicación a la población general de interés. Por lo tanto, un resultado es válido cuando se aproxima a la realidad verdadera del fenómeno objeto de estudio.

Planificar adecuadamente un diseño de investigación es la tarea principal del investigador o investigadora para conseguir resultados válidos. Valorar la validez de los resultados de investigación publicados es la tarea principal del lector o lectora que lleva a cabo una lectura crítica. Para una revisión completa del tema de la validez de los resultados en castellano se pueden consultar las obras de Arnau y Balluerka (1998), Ato (1991), Balluerka (1999), García-Pérez, Frías-Navarro y Pascual-Llobell (2006) y Vallejo (1991).

Los problemas de validez, en cualquiera de sus dimensiones, no son de todo o nada, porque todas las investigaciones están expuestas a amenazas o fallos. La validez es una cuestión de grados y es una propiedad de la inferencia que se realiza a través de los resultados de la muestra a la población.

La validez no es una propiedad del diseño o de la metodología de investigación, pues un mismo diseño o la aplicación de una determinada metodología pueden producir inferencias con mayor o menor grado de validez. Por ejemplo, aplicar la

metodología experimental no implica necesariamente que las inferencias sean válidas, pues factores como la pérdida selectiva de muestra o la escasa potencia de la prueba estadística podrían afectar a los hallazgos.

En definitiva, la metodología aplicada en el estudio no garantiza por sí misma la validez de una inferencia (Shadish, Cook, y Campbell, 2002).

El proceso del diseño de investigación debe garantizar la validez de los resultados y, por lo tanto, implica planificar un estudio que considere las posibles variables extrañas sistemáticas que podrían contaminar o amenazar la calidad de los hallazgos y exige controlar dichas variables para evitar su efecto sobre la variable dependiente y/o la variable independiente.

El lector o lectora consumidor de información se aproxima a la lectura de un trabajo de investigación desde la credibilidad de los hallazgos, pero su lectura siempre debe ser crítica (activa), es decir, debe valorar si los resultados que ofrece el trabajo realmente reflejan la situación del fenómeno estudiado. Y en ese punto, la experiencia profesional y el juicio clínico ayudan a valorar la evidencia o las pruebas del estudio dentro de un contexto o diseño de investigación concreto que podría limitar la calidad de los hallazgos. Sin embargo, los conocimientos sobre metodología de investigación son necesarios si el lector realiza una lectura crítica basada en todos los elementos que forman el proceso de diseño de investigación: conocimiento previo, hipótesis, planificación, metodología, resultados y discusión. Y adelantando cuestiones que se verán con más detalle próximamente, la credibilidad de los hallazgos, su importancia y su utilidad nunca se encuentra en el valor del denominado valor *p* de probabilidad del resultado.

Un aspecto importante relacionado con la validez de los hallazgos tiene que ver con la ética del investigador o investigadora tal y como ya se ha comentado anteriormente. Afortunadamente no es la norma pero, el fraude científico existe. Se considera fraude la fabricación, falsificación y el plagio de experimentos. Muy conocido dentro de la investigación biomédica es el fraude llevado a cabo por el científico surcoreano Hwang Woo Suk, conocido como “el doctor clon”, quien en 2005 afirmaba que su equipo había obtenido células madre de embriones humanos clonados. Los resultados de sus investigaciones fueron publicados en la prestigiosa revista *Science*. Sin embargo, en el año 2006 una comisión de investigación de la

Universidad de Seúl descubrió que todo era falso, concluyendo que "los datos del artículo de 2005 no fueron el resultado de simples errores, sino de la fabricación intencional", señalando que "la manipulación es un acto grave que afecta de forma negativa a los fundamentos de la Ciencia". El equipo de investigación había falsificado los datos del experimento y, además, no había conseguido esas células madre. En Octubre de 2009 el científico fue condenado a tres años de vigilancia por las autoridades.

La honestidad del científico es el eslabón que inicia el proceso de la Ciencia y representa el primer elemento de calidad de los hallazgos (Bosch, 2008).

El proceso de revisión por expertos (‘revisión por pares’) de los trabajos de investigación enviados a publicación, el control de las propias instituciones o de las asociaciones profesionales son claves en la corrección del fraude, el plagio y la falsificación de los hallazgos.

Existen cuatro componentes de la validez que deben planificarse en el proceso de diseño de investigación (Cook y Campbell, 1979; Shadish, Cook y Campbell, 2002):

- la validez interna
- la validez de constructo
- la validez de conclusión estadística
- la validez externa

Cuando se ejecutan investigaciones, el objetivo básico es detectar relaciones causales (si se aplica la metodología experimental) entre las variables explicativas de las hipótesis o detectar la covariación y su magnitud entre las variables (*validez interna*), así como ser capaz de captar el efecto experimental o la posible covariación de las variables si dicho efecto realmente existe en la población (*validez de conclusión estadística*) (ver Figura 17).

Además, el experimento también debe ser capaz de inferir o medir los fenómenos no observables o constructos, operacionalizando adecuadamente las variables que son objeto de estudio (*validez de constructo*) y debe proporcionar resultados que puedan ser generalizados (*validez externa*) más allá de los sujetos

experimentales investigados, más allá de la propia situación del experimento y más allá del momento temporal estudiado, validez poblacional, validez ecológica y validez histórica respectivamente.

La selección aleatoria de la muestra y la replicación de los experimentos (introduciendo distintas muestras de sujetos, estímulos, tareas, mediciones…) juegan un papel destacado en el ámbito de la validez externa para confirmar la consistencia del fenómeno estudiado e incluso para idear nuevas hipótesis alternativas producto de los nuevos datos de replicación.

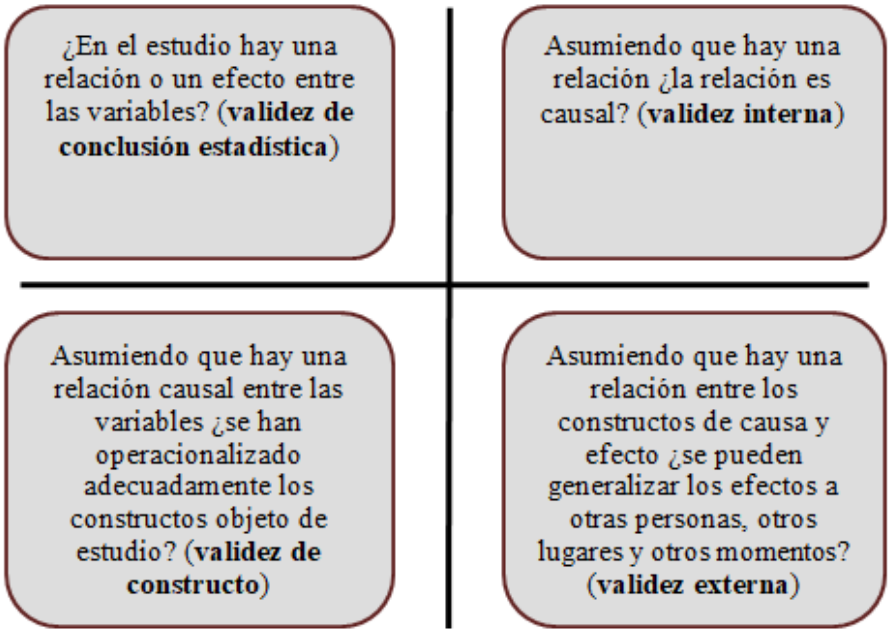


Figura 17. Validez de los resultados

Validez interna

La *validez interna* tiene que ver con la probabilidad de obtener una conclusión correcta acerca de la función de la variable independiente sobre la variable dependiente. Es decir, está relacionada con las amenazas que afectan a la inferencia causal (si se trata de una metodología experimental) o a la relación detectada entre las variables. Su propósito es asegurar que los cambios en la variable dependiente están provocados por el efecto de la o las variables independientes, si se trata de un estudio con una metodología experimental, y no por variables extrañas que intervienen en esa relación. En el resto de metodologías, la validez interna está

relacionada con el efecto de sesgo que podrían producir terceras variables o variables extrañas sobre la relación no causal detectada entre las variables.

La validez interna supone valorar en qué grado los resultados obtenidos pueden ser atribuidos a la manipulación de la variable independiente (o a la asociación entre las variables) y no al efecto de variables de confundido o contaminadoras (variables perturbadoras o extrañas no controladas).

Por lo tanto, la validez interna está relacionada con el control de todas las influencias sistemáticas o contaminadoras (varianza sistemática *secundaria*) que afectan a las unidades experimentales, maximizando los efectos de la variable independiente que sí son objeto de comprobación en el estudio (varianza sistemática *primaria*) y minimizando el error aleatorio (*varianza no sistemática*).

En resumen, la validez interna está relacionada con la posibilidad de inferir conclusiones precisas y exactas sobre la relación entre las variables independientes y dependientes detectadas en la investigación.

Según la terminología de Cook y Campbell (1979), la validez interna está relacionada con la "*validez causal*", o lo que es lo mismo, que la investigación demuestre, sin lugar a dudas, que no hay otras interpretaciones alternativas plausibles de los fenómenos objeto de estudio.

Es decir, los cambios observados en la variable dependiente se deben exclusivamente a la condición definida por el investigador o investigadora en la variable independiente. El problema de la validez interna es el problema de la "tercera variable" entendida como posible explicación del fenómeno: *¿existe una tercera variable (Z) responsable de la supuestamente relación evidente entre A e Y?*

Las principales amenazas a la validez interna se detallan en la Tabla 2 (Cook y Campbell, 1979).

Tabla 2. Amenazas a la validez interna

Amenaza a la validez interna	*¿Hay una relación entre las variables independientes (A) y dependientes (Y)?*
1. Historia	Sucesos externos al tratamiento que ocurren durante el estudio y que pueden afectar a la variable dependiente
2. Maduración	Cambios biológicos y psicológicos de los sujetos que afectarán a sus respuestas

3. Administración de pruebas	Los efectos del pre-test pueden alterar las respuestas en el post-test independientemente del tratamiento
4. Instrumentación	Cambios en la instrumentación, o en los observadores (dificultades de calibración)
5. Regresión estadística	Las puntuaciones extremas tienden acercarse a la media en el post-test a pesar del efecto del tratamiento
6. Selección de la muestra	Diferencias sistemáticas en los sujetos anteriores al tratamiento
7. Mortalidad experimental (atrición)	Perdida selectiva de sujetos a lo largo del estudio
8. Interacciones con selección	Algunas características de los sujetos de la muestra producen errores en el efecto del tratamiento en el post-test; efectos diferenciales en los factores de selección de la muestra
9. Ambigüedades sobre la direccionalidad de la influencia causal	Falta de claridad en la dirección de la causalidad
10. Difusión/imitación de tratamientos	Los miembros de los grupos de tratamiento comparten las condiciones de tratamiento con cada uno de los demás grupos de comparación o intentan copiar el tratamiento
11. Igualación compensatoria de tratamientos	Determinar que todos los sujetos tanto del grupo *experimental* como del grupo de *control* reciban un tratamiento que les proporcionen efectos beneficiosos
12. Frustración de los sujetos	Los miembros de los grupos que no reciben tratamiento (grupos de control) se perciben como inferiores

El control de las amenazas a la validez interna supone planificar los elementos que podrían afectar a las relaciones postuladas en la hipótesis científica. La técnica de control de la aleatorización es el principal elemento del diseño de investigación que si es correctamente ejecutada elimina las amenazas a la validez interna (tanto de variables observables como inobservables) y con ello las explicaciones alternativas ante los resultados obtenidos en el estudio. Por supuesto, hay otras técnicas de control dirigidas al control de las variables extrañas como la eliminación / constancia de la variable extraña, su factorización en el propido diseño de la ecuación estructural del estudio o su inclusión en la ecuación como una variable covariada, tal y como ya se ha comentado anteriormente. En la tabla 3 se detalla cómo afrontar algunas de las amenazas a la validez interna con el diseño de investigación.

Tabla 3. Control de las amenazas a la validez interna

Amenaza a la validez interna	*Control de la amenaza*
Historia	Selección aleatoria, asignación aleatoria
Maduración	Apareamiento de los sujetos, aleatorización
Administración de pruebas	Grupo de control
Instrumentación	Fiabilidad y consistencia interna de los instrumentos
Regresión estadística	Omitir puntuaciones extremas, aleatorización

Selección de la muestra	Selección aleatoria, asignación aleatoria
Mortalidad experimental	Aparear a los sujetos y eliminar en ambos grupos
Ambigüedades sobre la direccionalidad de la influencia causal	Comprobar de forma repetida como actúa la presencia del tratamiento y su ausencia
Sesgos del experimentador y/o de los participantes	Enmascaramiento o cegamiento de los sujetos o de los sujetos y el experimentador (técnica de simple ciego o técnica del doble ciego)

Desde la lógica de la investigación experimental, dos factores pueden afectar seriamente a la validez interna: el sesgo de selección y el efecto de confundido.

El "sesgo de selección" ocurre siempre que el proceso mediante el que se configuran los distintos grupos experimentales o mediante el que se asignan los tratamientos a los sujetos no garantiza que estos sean equivalentes. El sesgo de selección debe ser evitado siempre y en todos los casos, o corregido una vez recogidos los datos y antes de pasar a analizarlos, por ejemplo realizando un ajuste estadístico de los datos del estudio mediante un análisis de covarianza (ANCOVA).

El sesgo de selección se puede producir en cualquiera de las fases de la investigación. En la fase del diseño y planificación se puede introducir a través de los procesos complejos de identificación de sujetos, esquema de muestreo, instrumentos de diagnóstico utilizados, etc. En la fase de implementación del experimento se puede introducir a partir de la existencia de fenómenos como la atrición o pérdida de datos. Finalmente, en la etapa de análisis de datos también se puede filtrar, por ejemplo, al analizar datos imputados y no datos reales.

Desde la perspectiva del diseño y planificación de la investigación experimental, la solución universal para hacer frente al sesgo de selección es la aleatorización: mediante el proceso de asignación aleatoria del tratamiento a los sujetos o unidades experimentales (o asignación aleatoria en el orden de administración de los tratamientos si el diseño es de medidas repetidas o de medidas dependientes) ya que se garantiza que los grupos de tratamiento o condiciones experimentales sean equivalentes entre sí previamente a la introducción del tratamiento (si el azar actúa de forma eficaz). Gracias a la aleatorización la equivalencia de los grupos se logra tanto para variables observadas como para variables no observadas.

El concepto de "sesgo de confundido" mantienen una cierta ambigüedad y es por ello que algunos autores lo tratan al hablar de la validez de constructo y otros lo hacen al hablar de la validez interna. En principio, el efecto de confundido se comete por la

manipulación inadvertida de una segunda variable (o por la covariación inadvertida de una segunda variable) o bien por la evaluación de una variable teórica relevante que no es la variable que el investigador o investigadora desea estudiar. Si la manipulación de la variable independiente conlleva la covariación sistemática (inadvertida) de otra variable, estamos atentando contra la *validez interna* del experimento. Del mismo modo, si la relación entre la variable independiente y la variable dependiente depende de la presencia de terceras variables estamos también atentando contra la *validez interna*. Por el contrario, cuando se atribuye la causalidad a un constructo teórico que no es el representado en el experimento mediante su operacionalización empírica entonces se atenta contra la denominada *validez de constructo*. La calidad de la medida (fiabilidad) realizada de los constructos o fenómenos inobservables es fundamental para garantizar la validez de constructo.

Una variable de confundido es una variable que está relacionada tanto con la variable independiente como con la variable dependiente y su presencia puede disminuir o aumentar falsamente la relación encontrada entre dichas variables de interés (Meinert, 1986). En este sentido, las hipótesis de confundido plantean que una tercera variable explica la relación entre la variable independiente y la variable dependiente

Distinguir entre los términos de variables mediadores y variables moderadores es básico para plantear el control de dichos efectos mediante el análisis estadístico y el diseño de investigación. Sin embargo, los investigadores y las investigadoras no siempre saben diferenciar adecuadamente entre las variables mediadoras y las variables moderadoras (Ato y Vallejo, 2011; Holmbeck, 1997). La mediación no se debe confundir con una interacción o efecto de moderación donde la magnitud del efecto de A sobre Y está determinado por el valor o condición que adopte la variable moderadora. Las estrategias de análisis serán diferentes según se postule el control de una variable de confundido relacionado con el efecto de una variable mediadora o relacionado con el efecto de una variable moderadora.

Validez de conclusión estadística

Un experimento válido debe garantizar la producción del fenómeno y su generalidad, pero eso no quiere decir que, aun estando presente el efecto, el diseño de investigación tenga la capacidad de detectarlo, ya sea por el pequeño tamaño del

efecto o porque el diseño no tiene la suficiente sensibilidad estadística para captar el efecto, por ejemplo por escaso tamaño de la muestra.

Se trata de la *validez de conclusión estadística* que está relacionada con el uso apropiado de las pruebas estadísticas (inferencia estadística) como instrumento para inferir si las variables del diseño covarían entre sí.

La validez de conclusión estadística tiene que ver con el componente de covariación que se establece cuando afirmamos relaciones de causa y efecto o relaciones meramente funcionales entre variables.

Se pueden plantear dos preguntas clave dentro de la validez de conclusión estadística: ¿existe covariación / correlación entre la causa y el efecto? Y ¿de qué magnitud es?

Y se pueden cometer error cuando se llevan a cabo las decisiones. Incorrectamente se puede concluir que la causa y el efecto covarían (las variables analizadas en la hipótesis de investigación covarían) cuando no es así (*error de Tipo I*) o incorrectamente concluir que no covarían cuando de hecho sí lo hacen en la población (*error de Tipo II*).

Las dos incorrecciones anteriores están asociadas a la primera de las preguntas anteriores y hay que establecer las medidas oportunas para que no ocurran. En cuanto a la segunda pregunta, una infraestimación o una sobreponderación de la magnitud del tamaño del efecto, puede igualmente afectar a la inferencia estadística de los datos de la investigación.

Ahora ya no se trata de valorar si la relación entre las variables dependientes e independientes implicadas en la hipótesis de investigación se ha detectado sin la contaminación de posibles variables extrañas perturbadoras (validez interna). Ahora se trata de valorar si el diseño tiene las características adecuadas para detectar un efecto o una relación entre las variables, si realmente existe dicho efecto en la población.

La validez de conclusión estadística trata de garantizar que la prueba estadística empleada es la más adecuada para detectar una relación entre las variables, si dicha relación existe. Es una cuestión vinculada a si la inferencia estadística se aplica correctamente, es decir, si tiene sensibilidad la prueba estadística seleccionada por

el investigador o investigadora para detectar un efecto, si es la más adecuada para las hipótesis de investigación planteadas y si se aplica en las condiciones correctas relacionadas con sus supuestos estadísticos. La validez de conclusión estadística está relacionada con lo que se podría denominar el "diseño estadístico" de la investigación.

Controlar los denominados errores estadísticos (*error de Tipo I* y *error de Tipo II*), garantizar el cumplimiento de los supuestos de la técnica de análisis aplicada, aumentar la potencia de la prueba estadística (probabilidad de detectar un efecto si realmente existe), estimar de forma precisa el tamaño del efecto y el tamaño de la muestra… son algunas de las cuestiones que necesariamente se deben afrontar en el apartado de la validez de conclusión estadística y que se abordarán con más detalle posteriormente al tratar el tema del procedimiento de significación de la hipótesis nula (NHST) y los diferentes diseños de investigación.

Las amenazas a la validez de conclusión estadística se detallan en la Tabla 4 (Cook y Campbell, 1979).

Tabla 4. Amenazas a la validez de conclusión estadística

Validez de conclusión estadística	*¿El estudio es sensible para detectar si las variables covarían?*
1. La baja potencia estadística puede provocar erróneamente un resultado estadísticamente no significativo	El error de *Tipo II* aumenta cuando el valor de alfa es bajo y la muestra es pequeña
2. Violación de los supuestos de las pruebas estadísticas pueden provocar una sobreestimación o una infravaloración del tamaño del efecto y su significación estadística	Todos los supuestos de las pruebas estadísticas deben ser conocidos y comprobados cuando sea necesario
3. Tasa de Error de Tipo I: repetidas pruebas estadísticas pueden aumentar la significación estadística	Se incrementa el error de Tipo I, a menos que se ajuste al número de contrastes posibles
4. Fiabilidad de medición	Fiabilidad baja implica más errores que constituyen un problema serio en los estadísticos inferenciales
5. Fiabilidad de la administración de los tratamientos	Los tratamientos necesitan ser administrados del mismo modo de una persona a otra, de un lugar a otro y a lo largo del tiempo
6. Irrelevancias aleatorias en la situación experimental	Los efectos situacionales aleatorios pueden ser causa o interactuar con los efectos del tratamiento
7. Heterogeneidad aleatoria de las unidades experimentales	Ciertas características de los sujetos pueden correlacionar con las variables dependientes

Un aspecto muy importante vinculado a la validez de conclusión estadística es el tamaño de la muestra ¿Cuántos datos son necesarios en una investigación?, es decir, ¿cuántos sujetos necesita el estudio para que sea sensible a captar un efecto

o una relación entre las variables objeto de estudio? Un investigador o investigadora podría pensar que cuantos más sujetos mejor, pero no es una respuesta correcta.

Evidentemente cuantos más datos tenga el estudio más probable será llegar a una decisión correcta dado que habrá menos error de estimación de los parámetros. Lo ideal sería tener todos los datos de la población y de este modo el investigador o investigadora podría obtener conclusiones con absoluta confianza y es más, ya no necesitaría la técnica de la inferencia estadística. Pero esto no suele ser la práctica común.

Generalmente en el mundo científico se trabaja con muestras aleatorias y se necesita de la inferencia estadística para obtener conclusiones sobre la población y por ello la estimación de la potencia estadística a priori en la fase de planificación del estudio es fundamental.

En la fase de planificación de una investigación es primordial plantearse qué tamaño muestral es necesario para garantizar de forma razonable el equilibrio entre la probabilidad de rechazar la hipótesis nula (siendo realmente verdadera, *error de Tipo I*) y el riesgo de no rechazarla (siendo realmente falsa, *error de Tipo II*).

Es necesario, por lo tanto, argumentar o planificar los riesgos de error de Tipo I (seleccionar a priori el valor de alfa), de error de Tipo II (planificar un nivel aceptable de beta) y el tamaño del efecto poblacional esperado (planificar el efecto estimado). Con esas tres cantidades ya puede el investigador o investigadora decidir qué tamaño muestral (*N*) deberá tener el estudio para garantizar la validez de conclusión estadística. Esta cuestión se tratará posteriormente cuando se analizan los diferentes índices del tamaño del efecto y el tamaño de la muestra.

En el libro de Shadish, Cook y Campbell (2002) se añade una nueva amenaza a la validez de conclusión estadística denominada “estimación imprecisa del tamaño del efecto”. Conviene tener en cuenta que algunos estadísticos sobreestiman o infraestiman de forma sistemática el tamaño de un efecto. Posteriormente al abordar la reforma estadística y la estimación del tamaño del efecto se retomará también dicha cuestión.

Validez de constructo

La *validez de constructo* alude a la relación existente entre la manipulación de la variable independiente y el constructo teórico que presumiblemente se supone que se está manipulando y también a la relación entre la variable dependiente y el constructo que se supone se está midiendo (Cook y Campbell, 1976, 1979).

Como el término constructo es sinónimo de concepto o de construcción teórica (es un invento que la comunidad científica utilizada para referirse al concepto no observable directamente, por ejemplo la ansiedad, la depresión, la atención, el aprendizaje, el prejuicio), es evidente que la validez de constructo tiene que ver con la validez de las inferencias que se hacen acerca de los fenómenos no observables o constructos, apoyándonos en la evidencia que nos ofrecen las variables observadas que actúan como indicadores empíricos (representan una operacionalización de los constructos).

Es decir, la validez de constructo se centra en la interpretación correcta de los constructos implicados en la relación entre las variables que forman el diseño de investigación. La labor teórica de enlazar conceptos, generar hipótesis, asociar manipulaciones a variables teóricas, operacionalizar variables inobservables… es tan fundamental a la ciencia empírica como el análisis de datos.

La tarea no es fácil, dado que un mismo constructo se puede manifestar a través de múltiples variables observadas y a la vez, una misma variable puede reflejar distintos constructos. Por ello, el problema de la validez de constructo implica problemas tanto de validez convergente (relación entre las variables) como de validez discriminante (ausencia de relación entre las variables).

Las amenazas a la validez de constructo se detallan en la Tabla 5 (Cook y Campbell, 1979).

Tabla 5. Amenazas a la validez de constructo

Validez de Constructo	*¿Qué variables teóricas o implícitas están siendo estudiadas?*
1. Explicación preoperacional inadecuada	Escasa definición de los constructos
2. Sesgo debido al empleo de operacionalizaciones únicas	Medida de una sola variable dependiente
3. Sesgo debido al empleo de un método de operacionalización único.	Medida de la variable dependiente mediante un sólo método

4. Adivinación de la hipótesis	Los sujetos intentan adivinar la hipótesis experimenta y actúan de la forma que creen que el investigador o investigadora quiere que actúen
5. Recelo de evaluación	Los sujetos manifiestan cierto recelo ante la situación de evaluación
6. Expectativas del experimentador	Los experimentadores producen sesgos en el estudio a causa de sus expectativas en y durante el estudio
7. Confusión entre constructos y niveles de constructo	No se implementan todos los niveles del constructo y pueden presentarse de forma débil o no existir
8. Interacción de tratamientos intrasujeto	Los sujetos forman parte también de otros tratamientos (intrasujetos)
9. Interacción de administración de pruebas y tratamientos	La administración de las pruebas puede facilitar o inhibir el efecto del tratamiento
10. Generalidad restringida entre constructos	Punto en que un constructo puede ser generalizado de un estudio a otro

Validez externa

El tema de la validez externa tal y como es entendida por Cook y Campbell (1976, 1979), está relacionado con la generalización de los hallazgos más allá de los sujetos del estudio, del contexto de investigación y del momento del experimento. La replicación de las relaciones halladas entre los constructos otorgará validez externa a los hallazgos.

La validez externa está relacionada con la selección aleatoria de la muestra y no se debe confundir con la asignación aleatoria del tratamiento.

La asignación aleatoria del tratamiento a los sujetos está relacionada con el control del sesgo y, por lo tanto, con la validez interna de los resultados. La selección aleatoria de las unidades experimentales (generalmente sujetos) está relacionada con la validez externa de los hallazgos.

Conviene tener en cuenta que la validez externa no es una cuestión de todo o nada. Se trata de una cuestión estrictamente empírica, pues lo que es válido para una población puede no serlo para otra. Por ello, la replicación sistemática de los hallazgos es clave para otorgar validez externa a los resultados de la investigación (Underwood y Shaughessy, 1978).

Cuando el diseño de investigación pretende comprobar teorías o buscar conductas regulares de los fenómenos entonces la replicación de los efectos elaborando nuevas investigaciones facilita el control de calidad de la validez externa. Introducir nuevas variantes del mismo fenómeno, estimar el efecto con nuevas

muestras de sujetos, con nuevos estímulos y contextos permite confirmar la consistencia del fenómeno estudiado (validez interna) y al mismo tiempo permite valorar el alcance o generalización del fenómeno estudiado.

La comprobación de la consistencia de los fenómenos más allá del experimento (validez externa) también puede ponerse a prueba con la elaboración de estudios de meta-análisis.

Las amenazas a la validez externa se detallan en la Tabla 6 (Cook y Campbell, 1979).

Tabla 6. *Amenazas a la validez externa*

Validez Externa	*¿Pueden generalizarse los efectos y causas de un estudio a otros sujetos, situaciones o contextos?*
1. Interacción selección-tratamiento (validez de población)	Capacidad para generalizar el tratamiento a personas que no pertenezcan al grupo estudiado
2. Interacción contexto-tratamiento (validez ecológica)	Capacidad de generalización del tratamiento a situaciones más allá de la estudiada
3. Interacción historia-tratamiento (validez histórica)	Capacidad para generalizar el tratamiento a otras ocasiones temporales (pasado o futuro)

En definitiva, el diseño o plan de investigación incluye valorar los aspectos relacionados con la maximización de la *varianza sistemática primaria* (efecto del tratamiento o variable independiente), el *control de la varianza sistemática secundaria* (variables extrañas perturbadoras no controladas) y la *minimización de la variabilidad atribuida al error aleatorio* (variables extrañas aleatorias no sistemáticas) (principio MX-MIN-CON), lo que supone reflexionar sobre las diferentes dimensiones de la validez del diseño de investigación que se pueden finalmente resumir en cuatro consideraciones.

En primer lugar, evaluar todos los factores que potenciarán la adecuada medición de la o las variables dependientes utilizando instrumentos de medida fiables y válidos que aseguren que se está midiendo la característica de interés de manera consistente, siendo sensible al tipo y magnitud de cambio que el investigador o investigadora espera encontrar.

En segundo lugar, asegurar la manipulación correcta de la o las variables independientes que permitan que el tratamiento repercuta positivamente sobre la varianza sistemática primaria, impidiendo que esta variabilidad quede reducida por factores como la ineficaz administración de tratamientos o la dificultad (facilidad) de

la tarea experimental, que de no controlarlos pueden provocar la presencia de la varianza sistemática secundaria. Dos efectos que suelen ir asociados a la variable de elaboración de la tarea experimental son los denominados efectos de techo (*ceilling effect*) y efectos de suelo (*floor effect*). El 'efecto de techo' sucede cuando la ejecución de los sujetos en la tarea es tan buena en todos los casos que resulta imposible hallar alguna diferencia entre las condiciones experimentales (por ejemplo, con tareas muy sencillas). La respuesta de los sujetos está limitada por el límite impuesto por la tarea experimental y no por las capacidades o las propias respuestas de los sujetos. En el 'efecto de suelo' la ejecución de todos los sujetos es tan pobre que no permite discriminar entre las condiciones de tratamiento (por ejemplo, con tareas muy difíciles). En ambos tipos de efectos no se ha operacionalizado adecuadamente la manipulación de la variable independiente impidiendo observar efectos del tratamiento si los hubiese.

En tercer lugar, es objetivo del plan de investigación reflexionar sobre aquellas otras variables que no siendo teóricamente relevantes para la explicación del fenómeno objeto de interés (no se encuentran en la hipótesis del estudio), afectan de forma sistemática a la relación entre las variables explicativas del diseño al mismo tiempo que también aumentan la varianza de error del modelo. La presencia de estas variables perturbadoras debe mantenerse siempre bajo la potestad del diseño de investigación, siendo por lo tanto variables extrañanas controladas.

Y en cuarto lugar, la reducción de la varianza de error no sistemático o error aleatorio (relacionada con errores de medida y diferencias individuales) constituye también un objetivo a tener en cuenta en la planificación de la investigación. La homogeneidad de las unidades de observación y la fiabilidad de los instrumentos de medida ayudarán a su disminución.

En definitiva, todas estas consideraciones tienen el propósito de obtener datos válidos (reflejo de la realidad) con los que someter a prueba la hipótesis teórica o científica (sustantiva) propuesta derivada de la necesidad de conocimiento. En este proceso, la estrategia de recogida de los datos y la naturaleza manipulada o no de la variable independiente de tratamiento marcará el tipo de metodología empleada en la investigación y la naturaleza causal o no de las relaciones encontradas, todo ello guiado siempre y primeramente por el planteamiento teórico y las hipótesis

sustantivas de la investigación. La estadística está al servicio del diseño del estudio y el diseño necesita de una labor estadística realizada de forma adecuada para resolver sus hipótesis de investigación.

BLOQUE 2. ANÁLISIS DE LOS RESULTADOS DE LA INVESTIGACIÓN

Capítulo 7. Diseño entre-grupos, unifactorial, univariado

Dolores Frías-Navarro*
Marcos Pascual-Soler**

*Universidad de Valencia
**ESIC Business & Marketing School, España

Índice

- Hipótesis científica, hipótesis estadísticas (H_0, H_1).
- Práctica Basada en la evidencia.
- Lectores y lectoras.
- Informe transparente.
- El Factor bayes (BF).
- Resultados nulos.

Citar el capítulo como:

Frías-Navarro, D. y Pascual-Soler, M. (2021). Diseño entre-grupos, unifactorial, univariado. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

La herramienta de la estadística junto con la planificación y ejecución del diseño de la investigación empírica que se realiza permiten que el investigador o investigadora aborde el análisis de los datos de la muestra recogidos y proceder a continuación con la elaboración del informe de los resultados de la investigación.

En la pirámide del proceso de diseño de investigación, que se ha ido desarrollando en este libro, se procede ahora con la fase de '*análisis de los datos*' (y se obtienen los resultados de la investigación) donde se realiza un contraste estadístico con los datos empíricos de la muestra del estudio a través del procedimiento clásico de significación de la hipótesis nula (NHST). En el procedimiento NHST se asume desde el principio que el modelo planteado por la hipótesis nula de efectos cero o nula relación entre las variables es cierto.

Dicho procedimiento estadístico implica formular la *hipótesis estadística nula* (Hipótesis nula, H_0) y asumirla como cierta. Es decir, se parte del modelo estadístico que plantea que el efecto es cero en la población (o relación nula entre las variables del estudio. Así:

$$H_0 \equiv \mu_1 = \mu_2$$

Es decir,

$$H_0 \equiv \mu_1 - \mu_2 = 0$$

Conviene tener muy presente que en el proceso de contraste de hipótesis estadísticas, la hipótesis nula se asume desde el principio del análisis que es cierta y, además, se conoce su distribución muestral. Dicha distribución es la que aparece en las tablas teóricas de los estadísticos donde se presentan sus datos según diferentes grados de libertad y nivel de significación seleccionado a priori por el investigador o investigadora. Gracias al conocimiento que se tiene de su distribución teórica de datos se puede comparar la probabilidad observada (valor *p* del estadístico) de un determinado resultado (empírico) (o un resultado más extremo) asociado a un estadístico calculado en el estudio con la probabilidad que tendría en una distribución donde la hipótesis nula es cierta (valor téorico de las tablas o valor crítico de comparación asociado a una determinada probabilidad). Y, a partir de esa

comparación se procedería con la decisión dicotómica: mantener o rechazar la hipótesis nula.

Por lo tanto, en el proceso de inferencia estadística solamente se trabaja con la distribución de la hipótesis nula, que se asume como cierta desde el comienzo del proceso de análisis. Y en ningún momento se va a demostrar que es cierta o que es falsa. Se asume como cierta y nuestro resultado empírico se compara con su distribución teórica, teniendo en cuenta los valores de alfa y grados de libertad vinculados a la prueba estadística ejecutada.

La *hipótesis alternativa* (Hipótesis alternativa, H_1), en cambio, plantea la existencia de un efecto o una relación diferente a cero de la o las variables independientes sobre la o las variables dependientes (sí que hay relación entre las variables estudiadas) y su distribución muestral no se valora en el proceso de contraste estadístico a la hora de tomar la decisión dicotómica. Solamente si se rechaza la hipótesis nula entonces se podrá aceptar la hipótesis alternativa.

$$H_1 \equiv \mu_1 \neq \mu_2$$

Por lo tanto,

$$H_1 \equiv \mu_1 - \mu_2 \neq 0$$

Hipótesis científica, hipótesis estadísticas (H_0, H_1)

Tal y como se observa en la figura 18, para obtener los resultados del análisis de una hipótesis concreta es necesario realizar una serie de pasos durante el proceso del diseño de investigación.

1) En primer lugar, en la mayoría de las ocasiones, es necesario establecer un modelo teórico (teoría) que dará sentido al planteamiento de la hipótesis del estudio. De ahí la importancia de los dos primeros pasos de la pirámide del proceso de diseño de investigación: '*necesidad de conocimiento*' (problema a investigar) y revisión del '*conocimiento previo*' (teorías y resultados de los estudios ya realizados), ya explicados anteriormente cuando se desarrolló el proceso del diseño de una investigación.

2) En segundo lugar, ese conocimiento teórico permite elaborar la hipótesis científica del estudio (conocida también como hipótesis teórica o hipótesis sustantiva

y formulada en términos de constructos no observables), apoyada por la revisión teórica realizada.

El planteamiento de la hipótesis científica puede ser ampliar información sobre el fenómeno objeto de estudio, introduciendo por ejemplo nuevas variables en el modelo teórico explicativo (generando conocimiento nuevo), replicar un estudio, reproducir un hallazgo previo o tratar de solucionar ciertas ambigüedades que puedan haber en ese campo de trabajo.

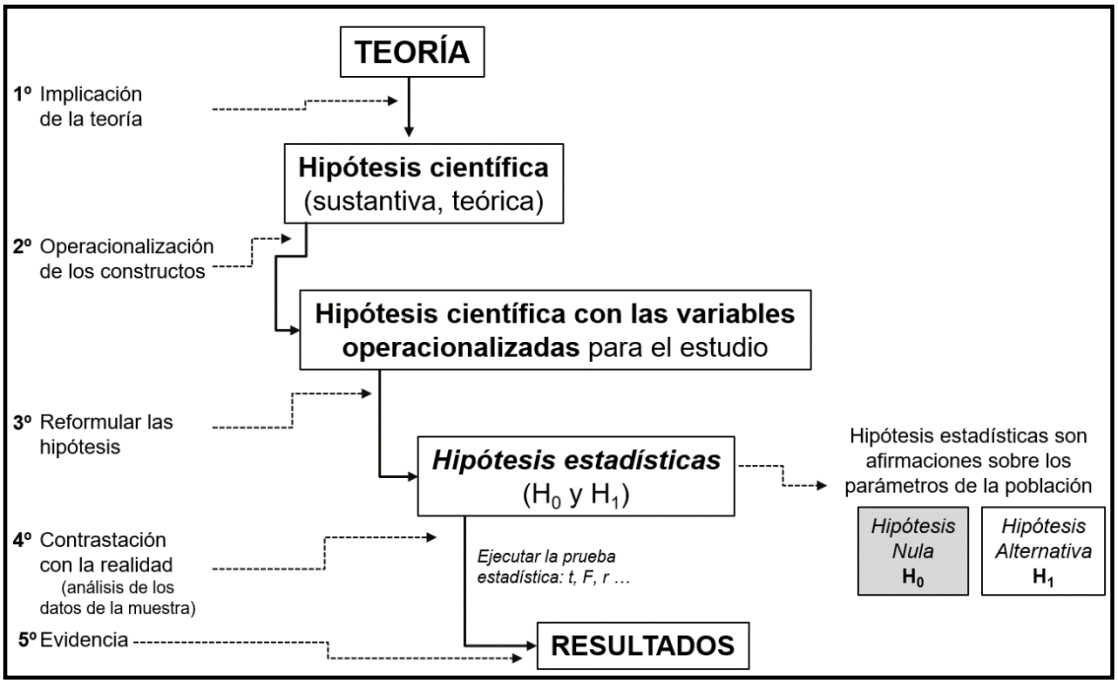


Figura 18. Hipótesis científica, hipótesis estadística y proceso de diseño de investigación

3) En tercer lugar, esa hipótesis teórica o científica se transforma en una hipótesis científica, pero con las variables operacionalizadas para el estudio concreto que se va a realizar. Es decir, las variables se operacionalizan para el caso concreto del estudio que se lleva a cabo (se definen las variables de forma que puedan ser medibles / observables en el estudio). Se trata de la hipótesis experimental de un estudio concreto. Así, es necesario que la hipótesis científica se transforme para ser sometida a su estudio empírico y, para ello, el investigador o investigadora operacionaliza los constructos de la hipótesis sustantiva en variables que pueden ser medidas para poder captar en la realidad cómo se comporta el fenómeno que está siendo objeto de estudio (hipótesis científica experimental operacionalizada).

Por lo tanto, conviene tener muy presente que todas las variables del estudio (independientes, dependientes y extrañas) tiene 2 dimensiones:

- como constructo teórico no observable y
- operacionadas en el diseño para poder ser medidas y recoger los datos del estudio o experimento

Por ejemplo, si el investigador o investigadora desea analizar si la indefensión aprendida provoca depresión (se trata de dos constructos no observables) será necesario que en el diseño del estudio operacionalice esos constructos para poderlos medir y comprobar la relación entre ellos. Podría operacionalizar el constructo de la indefensión con una tarea experimental donde un determinado efecto se mostrará haga lo que haga el sujeto. No hay control de la situación por parte del sujeto y, por lo tanto, se producirá indefensión aprendida. Y, además, el investigador o investigadora tendrá que medir la depresión a través de, por ejemplo, un autoinforme o una entrevista que mida la presencia del trastorno de forma fiable y válida o utilizar por ejemplo el tiempo invertido por una rata en recorrer un laberinto como variable dependiente, planteando que a mayor tiempo mayor es la indefensión aprendida. Por supuesto, debe haber una base teórica que respalde las hipótesis del estudio y las tareas experimentales deben realizarse sobre dicho fundamento.

En definitiva, es necesario operacionalizar o transformar los enunciados de las hipótesis científicas teóricas en enunciados que puedan ser contrastados empíricamente con la realidad. Y, ese proceso de transformación necesariamente implica que el investigador o investigadora describa de forma detallada en el informe de investigación la calidad de sus decisiones, aportando información sobre la fiabilidad y consistencia interna de la medida de los constructos (variable dependiente) así como argumentar por qué la tarea experimental (variable independiente) representa al fenómeno o constructo objeto de estudio. Por lo tanto, la hipótesis teórica de la indefensión aprendida provoca depresión podría ser operacionalizada en una investigación como: “las ratas sometidas a una situación de indefensión aprendida como puede ser recibir un shock eléctrico de forma continua durante su permanencia en un laberinto (perciben que ese efecto no está relacionado con su conducta) tardarán más tiempo en recorrer un laberinto que las ratas que

pueden actuar para eliminar el shock eléctrico (perciben que su conducta sí controla la situación y pueden dejar de recibir el shock)".

4) En cuarto lugar, se procede al análisis de los datos de la muestra y con ello a la contrastación del modelo teórico con la realidad. Se trata de ejecutar el denominado 'contraste de hipótesis estadísticas'. En el procedimiento tradicional de contraste de hipótesis NHST se asume, como principio, que la hipótesis nula es cierta (en la población el efecto o la relación entre las variables es cero) y gracias a conocer su distribución teórica (distribución teórica o muestral del estadístico) se podrá llegar en el estudio a una conclusión estadística al comparar la probabilidad del resultado obtenido con la muestra, o un valor más extremo, (valor *p* del estadístico o nivel de significación utilizado en la prueba estadística ejecutada en el estudio) con la probabilidad que tendría dicho resultado en esa distribución de H_0. En la distribución de la hipótesis nula, como ya se ha comentado, se asume un efecto de cero o ausencia de relación entre las variables (distribución teórica o muestral del estadístico que se encuentra en las tablas de los manuales de estadística y de diseño de investigación).

La hipótesis nula es la protagonista del contraste de hipótesis y solamente ella interviene al tomar la decisión estadística: se *mantiene* la hipótesis nula o se *rechaza* la hipótesis nula (decisión estadística dicotómica). No se debe decir 'se acepta la hipótesis nula', ya que, desde el principio del contraste estadístico, se acepta dicha hipótesis como cierta.

En ese proceso de contraste estadístico, la hipótesis alternativa tiene un papel secundario, pues sólo entra en acción cuando se rechaza la hipótesis nula, ya que, como consecuencia, se aceptará la hipótesis alternativa de efecto diferente a cero. Cuando se acepta la hipótesis alternativa es conveniente tener en cuenta y reflexionar sobre la calidad científica de esa hipótesis alternativa que se defiende en el estudio. Una vez rechazada la hipótesis nula, hay cientos de hipótesis alternativas que podrían explicar el hallazgo. Y la calidad del diseño de investigación que se planificó y ejecutó determina que se pueda enlazar ese resultado estadísticamente significativo con el efecto o relación planteada en la hipótesis del estudio. No es una tarea sencilla. En este punto la competencia del investigador o investigadora en el tema del diseño de investigación es fundamental.

Cuando se mantiene la hipótesis nula (también se conoce como resultado nulo o resultado negativo), tras finalizar la decisión dicotómica, no se debe interpretar la hipótesis alternativa ni tampoco concluir que no hay efecto o no hay relación entre las variables, pues en ese caso se trata de un resultado no concluyente y, en todo caso, se necesita seguir investigando sobre él. Es un error muy común creer que mantener la hipótesis nula significa que las medias son iguales, que las puntuaciones no correlacionan o que no hay efecto del tratamiento (Badenes-Ribera y cols., 2016, 2018). Y no es así.

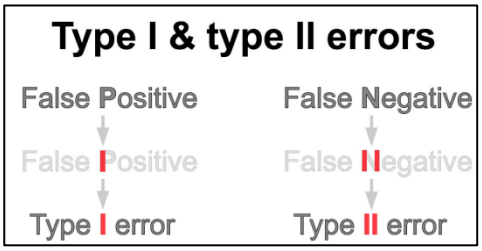
Práctica Basada en la Evidencia

Sin embargo, desde el punto de vista del diseño de la investigación y la Práctica Basada en la Evidencia es importante ir más allá del valor *p* de probabilidad. Así, el investigador o investigadora no debe centrar todos sus esfuerzos únicamente en 1) decidir si mantiene o no la hipótesis nula (significación estadística), pues es fundamental llevar a cabo una interpretación más allá del valor *p*, recurriendo al 2) análisis e interpretación del tamaño del efecto y su intervalo de confianza (significación del tamaño del efecto), contextualizando la magnitud de los efectos hallados dentro del área de investigación del fenómeno objeto de estudio, 3) a la interpretación clínica o sustantiva del hallazgo, basada en el juicio del experto o juicio clínico, respecto a su utilidad dentro de los objetivos del estudio (significación clínica) y, 4) también, a la calidad del diseño de investigación que se ha ejecutado, llevando a cabo una valoración crítica del control efectuado con el diseño del estudio sobre las amenazas a la validez interna, externa, de constructo y de conclusión estadística (valorar conscientemente y de forma juiciosa la calidad de la evidencia hallada).

Lectores y lectoras

Por otra parte, el lector o lectora activo que lleva a cabo una lectura crítica o activa del informe de investigación también debe poner en marcha esas reflexiones sobre la técnica y el análisis metodológico que se ha llevado a cabo en el artículo o informe y valorar todas las facetas de la significación de los hallazgos y no solo la faceta de la significación estadística. Gracias a esa valoración, el lector o lectora puede decidir qué grado de validez tienen las conclusiones de ese trabajo de investigación.

Lograr que un resultado sea estadísticamente significativo (rechazar la hipótesis nula) no significa de forma directa que dicho resultado es válido y fiable, pues hay más causas que podrían llevar a dicha significación (si fuese así se habría cometido un error estadístico Tipo II o un falso positivo). Ni tampoco significa que sea un resultado importante, ni tampoco útil. E incluso, a veces, el investigador o investigadora puede dar un resultado como estadísticamente significativo cuando se puede leer claramente en el texto que no es así (se trataría de un grave error de educación estadística por parte del investigador) y, de ahí, la importancia de leer críticamente o de forma activa los resultados del trabajo o informe y no dejarse llevar solamente por las conclusiones que ha redactado el autor o autora del informe.



De alguna manera, todos los lectores y lectoras activos se convierten en meta-investigadores, es decir, deben llevar a cabo una lectura rigurosa y crítica que valore cada apartado del informe, interpretando de nuevo el diseño del estudio y, si es necesario, utilizando herramientas o programas que le permitan comprobar los resultados que se aportan en el trabajo o añadir resultados que el investigador / investigadora o autor / autora del artículo no ha incorporado en su estudio.

Por ejemplo, si el autor o autora del artículo no lo ha realizado en su informe, podría ser interesante estimar el tamaño del efecto y su intervalo de confianza y ampliar la interpretación de la significación estadística (valor *p*) con la aportada por la significación del tamaño del efecto y valorar la incertidumbre de su estimación puntual gracias a la interpretación de su intervalo de confianza. También, el juicio clínico del investigador o investigadora podría acompañar a los resultados del estudio y, por ejemplo, en el apartado de discusión del artículo o informe se podría debatir los hallazgos desde un punto de vista sustantivo y/o aplicado y no solamente desde la significación estadística y del tamaño del efecto obtenido.

Por lo tanto, en ocasiones, puede ser necesario que el mismo lector o lectora activo lleve a cabo una estimación de otros indicadores que no han sido incluidos en el artículo o informe, pero sería interesante disponer de dicha información para comprender y valorar mejor los hallazgos y su alcance.

Informe transparente

Por todo ello, es muy importante que el estudio primario presente un informe transparente y con toda la información fundamental de los estadísticos básicos (también en los estudios secundarios como los trabajos de meta-análisis). Por ejemplo, es esencial informar siempre, al menos, del número total de observaciones que se ha empleado en el estudio (*N*), número de observaciones utilizadas en cada condición experimental (*n*), medias y desviaciones típicas de todas las puntuaciones mencionadas en el estudio y valores de las pruebas de contraste estadístico efectuadas junto con sus características como los grados de libertad y el valor de *p* exacto del estadístico calculado. Toda esa información es imprescindible cuando se redacta el informe.

Además, disponer de un informe transparente y completo ayuda a realizar los procesos de revisión de los artículos ya publicados por otros investigadores desde un punto de vista crítico (se trata de actividades de meta-investigación: investigadar sobre la investigación publicada) y, con ello se potencia la calidad de la Ciencia, se promueve que la Ciencia pueda autocorregirse y se detectan casos de fraude y de prácticas de investigación cuestionables (Frías-Navarro y cols., 2019). La acumulación del conocimiento científico necesita del filtro del control de los miembros de la propia comunidad científica.

Los denominados estudios de meta-investigación han demostrado que en la literatura publicada hay más errores en los diseños de investigación de lo que sería deseable y muchos problemas de comprensión y educación estadística entre los y las investigadores o autores y autoras de los artículos e informes. Y, si a estas circunstancias se une el hecho de que hay problemas de redacción y de transparencia en los artículos e informe, entonces la tarea de la lectura crítica se hace mucho más complicada y necesaria.

El Factor Bayes (BF)

Antes de continuar con el tema, solamente un apunte breve sobre el denominado Factor Bayes (BF) y su interpretación, ya que su uso en las Ciencias del Comportamiento aun no se ha generalizado, pero poco a poco va apareciendo más en los artículos e informes.

El factor Bayes es un resumen de la evidencia proporcionada por los datos en favor de una teoría científica, representada por un modelo estadístico, como opuesta a otra teoría (Kass y Raftery, 1995). Y tiene la ventaja de que su valor no se ve afectado por tamaños amplios de muestra. Kass y Raftery, 1995 señalan que la estadística frecuencial tiende a rechazar de forma sistemática la hipótesis nula cuando las muestras son grandes, mientras que el Factor Bayes no.

El Factor Bayes informa sobre la probabilidad de la hipótesis nula respecto a la probabilidad de la hipótesis alternativa, dados unos datos, *p*(H|datos). O viceversa: la probabilidad de la hipótesis alternativa respecto a la probabilidad de la hipótesis nula, dados unos datos. Es decir, se trata de una interpretación totalmente diferente al valor *p* de la perspectiva frecuencial que tradicionalmente se aplica en las pruebas de inferencia estadística basadas en la comprobación de la hipótesis nula (NHST) donde nunca se plantea la probabilidad de la hipótesis nula o la hipótesis alternativa. Con el procedimiento clásico NHST lo que se calcula es la $p(\text{datos}|H_0)$, es decir la probabilidad del resultado (o un resultado más extremo), asumiendo que la hipótesis nula es cierta. Y hay que tener muy claro que $p(\text{H}|\text{datos}) \neq p(\text{datos}|H_0)$. Además, esas pruebas clásicas de inferencia estadística frecuencial (NHST) se basan en asumciones fundamentadas en la extracción aleatoria repetida de muestras de una población definida (basadas en la distribución muestral de los estadísticos) y sus principales herramientas para llevara a cabo las inferencias son los valores *p* de probabilidad, el tamaño del efecto y los intervalos de confianza.

En resumen, el valor *p* de la inferencia frecuencial (procedimiento clásico de la hipótesis nula NHST) informa de la probabilidad de unos datos, o datos más extremos, considerando a la hipótesis nula como cierta ($p(\text{datos}|H_0)$), pero no dice nada de la probabilidad que tiene una hipótesis estadística de ser cierta (ni tampoco de ser falsa).

El valor del Factor Bayes, en cambio, se puede expresar como apoyo a la hipótesis nula respecto a la alternativa y se representa como BF_{01} o, en cambio, su valor puede indicar el apoyo de la hipótesis alternativa sobre la hipótesis nula y se representa como BF_{10}.

1. BF01 expresses the likelihood of H0 relative to H1 given the data.
2. BF01 expresses the probability of the data given H0, relative to H1.

El Factor Bayes proporciona un continuo de evidencia para la probabilidad de H_1 sobre la de H_0, o viceversa.

A medida que se incrementa el valor del BF, aumenta la probabilidad de una hipótesis estadística sobre la otra hipótesis. Un valor de BF = 1 significa que los datos apoyan de la misma manera a la hipótesis nula que a la hipótesis alternativa (no hay evidencia que apoye a ninguna hipótesis). Es decir, los datos predicen de la misma manera a ambas hipótesis estadísticas y la evidencia no favorece a un modelo sobre el otro.

La proximación bayesiana a la técnica de comprobación de hipótesis estadísticas fue desarrollada por Jeffreys (1935, 1961). Una de las interpretaciones que más se utilizan del Factor Bayes es la propuesta por Jeffreys (1961) que señala que si FB_{10} (apoyo de la hipótesis alternativa sobre la hipótesis nula) es:

- > 100 entonces hay extrema evidencia para H_1

-Entre 30-100 evidencia muy fuerte para H_1

-Entre 10-30 fuerte evidencia para H_1

-Entre 3-10 hay evidencia moderada para H_1

-Entre 1-3 es evidencia anecdótica para H_1

-Si el BF = 1 entonces no hay evidencia para ninguna de las dos hipótesis estadísticas.

Cuando se calcula FB_{01} (evidencia a favor de la hipótesis nula sobre la hipótesis alternativa) se utilizan los mismos valores anteriormente descritos para llevar a cabo la interpretación de FB (Kass y Raftery, 1995). Así, cuando se trata de la evidencia contra la hipótesis nula (FB_{10}): 1 a 3 es anecdótico, de 3 a 10 es moderada, de 10 a 100 es fuerte evidencia y > 100 es decisiva.

Sin embargo, conviene ser una persona precavida y no caer en la tentación de que esos valores sean la única guía para valorar la calidad de los hallazgos y, de nuevo, se conviertan en un oráculo de la verdad tal y como se ha criticado duramente al valor de punto de corte de alfa = .05 en la perspectiva frecuencial que fue establecido por Fisher de forma arbitaria. Siempre es necesario recurrir a valorar todo el proceso del diseño de investigación y jerarquizar la evidencia o pruebas que aporta

el estudio, actuando siguiendo la perspectiva de la Práctica Basada en la Evidencia. El análisis de todos los elementos que forman parte del proceso del diseño de investigación es la herramienta que facilita la valoración activa de la validez de los resultados, y no solamente los resultados estadísticos.

En definitva, Jeffereys (1961) sitúa el valor de FB = 3 como punto de referencia para considerar que hay evidencia moderada relevante para ser interpretada. Y no es un valor arbitrario, ya que, en general, está vinculado a un resultado estadísticamente significativo ($p < .05$). Sin embargo, hay que anotar que hay debate sobre el valor de referencia para indicar que hay evidencia sustantiva y otros investigadores / revistas optan por un valor de FB = 6 como ocurre en la revista *Cortex* o de FB = 10 tal y como se señala en la revista *Nature Human Behavior for Registered Reports*. Actualmente, y desde el área de la perspectiva de la inferencia clásica frecuencial, también hay debate sobre si seguir manteniendo el valor de alfa = .05 como punto de corte para decidir si hay o no significación estadística, pues hay opiniones de expertos que consideran que debería ser menor, como por ejemplo alfa = .005.

El programa estadístico JASP, gratuito, permite calcular el Factor Bayes de una forma sencilla. Para realizar los cálculos del Factor Bayes se parte de que las dos hipótesis tienen a priori la misma probabilidad (50%) y el BF expresa la probabilidad de los datos bajo los dos modelos, H_0 y H_1. Por ejemplo, un $FB_{10} = 4$ significa que los datos son 4 veces más probables que ocurran bajo la hipótesis alternativa que bajo el modelo de la hipótesis nula Y, por lo tanto, $BF_{01} = ¼$). O, por ejemplo, $BF_{10} = 32$ significa que los datos son 32 veces más probables que ocurran bajo el modelo de la hipótesis alternativa (probabilidad de 32 : 1 a favor de H_1) y, ese dato se corresponde de forma directa con un valor de $BF_{01} = 1/32$ (probabilidad de 1 : 32 a favor de H_0).

Resultados nulos

Continuando con la exposición del contraste de hipótesis desde la perspectiva clásica de inferencia estadística basada en el modelo frecuencial (modelo de comprobación de la significación de la hipótesis nula, NHST), también es importante tener en cuenta que los resultados donde se mantiene la hipótesis nula (conocidos como ‘resultados nulos’ y también como ‘resultados negativos’) deben ser publicados, llevando a cabo una interpretación de su valor de significación

estadística, pero también del tamaño del efecto estimado y de su intervalo de confianza (que permite valorar la incertidumbre en la estimación puntual) junto con la apreciación de todos los elementos metodológicos del diseño que podrían favorecer ese resultado nulo como la baja potencia estadística, el escaso tamaño de la muestra, la estimación a priori excesiva del tamaño del efecto, explicaciones al tamaño del efecto detectado ….

Un problema muy importante que tiene la Ciencia actual es que aproximadamente el 80% - 90% de los resultados que se publican son resultados que han sido estadísticamente significativos, 'olvidándose' en el cajón del archivador aquellos estudios donde se mantiene la hipótesis nula (se trata del problema del denominado 'sesgo de publicación' o 'file drawer'). Como consecuencia de esa situación, es probable que el panorama de resultados que ofrece el mundo científico sea demasiado bueno y no se ajuste a la realidad de los fenómenos. Por ello, hay que luchar desde el mundo académico y profesional para dar valor a los resultados nulos (donde se mantiene la hipótesis nula) que también tienen su papel en la explicación de la realidad detectada e implican reflexionar sobre por qué se producen o qué elementos del diseño han estado amenazados y su valoración requiere, en gran medida, la estimación del tamaño del efecto y su intervalo de confianza. Y desde el modelo de análisis NHST no se pueden interpretar nunca como evidencia de ausencia o evidencia de relación, pues ausencia de evidencia no es evidencia de ausencia. Un resultado nulo solo ofrece información de que el hallazgo no es concluyente, dado que no se ha detectado un efecto o una relación sistemática. Pero no quiere decir que hay evidencia de ausencia de efecto o de relación entre las variables.

En este punto conviene reflexionar sobre la creencia errónea, pero bastante común, que plantea que no encontrar un efecto estadísticamente significativo en un estudio de superioridad indica que los tratamientos tienen el mismo efecto o son equivalentes. Esa idea es absolutamente errónea, pues cuando el resultado señala que el efecto o la diferencia entre las puntuaciones medias de los grupos no es estadísticamente significativo solo se puede concluir que con la información recogida en el estudio no se puede llegar a ningún tipo de afirmación sobre el efecto de la intervención ya que se trata de un resultado no concluyente. Ese resultado no concluyente requiere una nueva valoración en otro estudio donde se planifique

adecuadamente el tamaño de la muestra y el efecto esperado y se replique de nuevo la investigación. Quizás ese efecto nulo (mantener la hipótesis nula) podría tratarse de un problema de escasa potencia estadística para detectar el efecto esperable o estimado, quizás se sobreestimó el efecto esperado, falta de validez de las medidas, medidas no fiables, falta de control experimental .... o quizás realmente el efecto es nulo (Campbell, 1982). Pero la respuesta no la sabemos ya que las pruebas clásicas de inferencia estadística no dan respuesta al porqué de los resultados nulos o resultados que no logran ser estadísticamente significativos.

En definitiva, conviene tener muy presente que un resultado nulo o un resultado negativo (mantener la hipótesis nula en el contraste estadístico) nunca probará que las intervenciones analizadas son equivalentes o similares, ni nunca demostrará que no hay relación entre las variables. Los valores de $p >$ alfa (mantener la hipótesis nula) o los intervalos de confianza que contienen el valor del efecto de efecto nulo (generalmente el valor de 0) dejan abierta la posibilidad de que existan efectos reales así como de que no exista ningún efecto o relación entre las variables. Se trata de un resultado no concluyente y son hallazgos que también contribuyen al conocimiento científico (Amrhein y cols., 2017). Y se pueden considerar hallazgos 'positivos' en el sentido de que dejan abierta la posibilidad de que aún existan efectos importantes o quizás podría ser que no exista el efecto. Por lo tanto, ante un resultado nulo es importante evitar en los informes afirmaciones como 'no hubo efecto', 'no hubo diferencias', 'no existe efecto de inetracción' o 'el tratamiento no influyó'. Este tipo de interpretaciones atribuyen certeza a la hipótesis nula ("no hay efecto") o concluyen que la hipótesis nula se acepta o se ha probado y esa conclusión es totalmente errónea con las pruebas clásicas de inferencia frecuencial basadas en el contraste de la hipótesis nula. En este tipo de pruebas se asume desde el principio que la hipótesis nula es cierta y no se demuestra ni su verdad ni su falsedad. No es un procedimiento que estime la probabilidad de la hipótesis nula o de la hipótesis alternativa. Ese procedimiento clásico solamente ofrece la probabilidad de nuestros datos dado que la hipótesis nula es cierta.

Respecto a los resultados estadísticamente significativos ($p \leq$ alfa, rechazar la hipótesis nula), tampoco es correcta la creencia de que se ha obtenido la certeza absoluta de la presencia de un efecto y menos si es una evidencia de un solo estudio (Oakes, 1986). Ante ese resultado estadísticamente significativo se desconoce si la

hipótesis nula o la hipótesis alternativa es cierta. Ante un resultado $p \leq$ alfa puede haber ocurrido: un efecto existe o que se trata de un resultado muy improbable o extraño dentro del modelo de la hipótesis nula. Obtener un resultado de $p \leq$ alfa puede indicar que la hipótesis nula no es cierta, pero también podría ser cierta dicha hipótesis y el resultado hallado tener una baja probabilidad en su distribución muestral. El valor *p* puede ser interpretado como una medida de grado de sorpresa ante su valor (Greenwald y cols., 1996). Así cuanto más pequeño sea el valor *p* más sorprendente serán los resultados si la hipótesis nula es realmente cierta (Reinhart, 2015). Y, como Fisher (1937) señala, en ningún experimento aislado un resultado estadísticamente signficativo puede ser suficiente para la demostración de un fenómeno natural.

Además, la estimación puntual de un estadístico no mide la incertidumbre del hallazgo. Disponer del intervalo de confianza de esa estimación puntual ayuda a valorar de forma aproximada el grado de incertidumbre de la estimación puntual, pero tampoco ofrece una interpretación de certeza absoluta, pues la 'danza de los intervalos de confianza' (Cumming, 2014) señala que un intervalo de confianza va cambiando de una muestra a otra debido a la variación aleatoria.

Por lo tanto, cuando se valora la calidad de una investigación es importante valorar de forma exhaustiva la calidad del apartado de Método y sus subapartados, es decir, cómo se planificó el proceso del diseño de investigación, cómo se seleccionó la muestra, cómo se midieron las variables o cómo se llevó a cabo la recogida de datos. Y no centrar ese análisis de la calidad del estudio en los resultados de las pruebas de constraste estadístico: efecto estadísticamente significativo / efecto estadísticamente no significativo. Que un resultado estadístico se vincule a $p \leq$ alfa o $p >$ alfa no significa de forma directa que el estudio sea bueno o malo, pues será necesario leer con detalle toda la información del apartado de Método para valorar de forma crítica o activa la calidad del estudio y sus hallazgos. La lectura activa o crítica de un informe o artículo debe ser especialmente exhaustiva cuando se lee el apartado de Método y de ahí la importancia de que los autores y autoras describan de forma transparente y detallada cómo se llevó a cabo el proceso del diseño de investigación y dotar al lector o lectora de toda la información necesaria para que valore y tome decisiones sobre la calidad del estudio y el grado de certeza que se puede atribuir a sus resultados. Y siempre hay que recordar que el hallazgo

de un único estudio no demuestra la certeza del hallazgo ni le otorga generalización a otras muestras y situaciones.

En definitiva, los resultados nulos que se producen al mantener la hipótesis nula tras ejecutar la prueba estadística ($p > .05$) no significa que no hay efecto en la población o que las variables no están relacionadas; solamente significa que, con los resultados obtenidos con la muestra, no se puede llegar a una conclusión sobre la hipótesis teórica propuesta. Se trata de un resultado no concluyente desde el punto de vista de la significación estadística, pero que acompañado del tamaño del efecto y su intervalo de confianza y con la valoración de la calidad del diseño aplicado se pueden plantear ideas para mejorar la próxima investigación y seguir de este modo indagando sobre la naturaleza y características del fenómeno objeto de estudio. Si persiste ese resultado en diferentes estudios sería interesante plantear un análisis de equivalencia (estudio o test de equivalencia) entre las medias para comprobar si el hallazgo es nulo.

Es muy importante dar a conocer los resultados nulos y publicarlos porque su ausencia enturbia la visión que se tiene de un fenómeno e impide seguir estudiando líneas de investigación apoyadas en los esfuerzos que realizaron otros investigadores, pero cuyo trabajo se desconoce, ya que nunca lo publicaron. La implicación de los editores y las editoras es muy importante ya que al final la decisión de publicar o no un manuscrito depende de ellos o ellas.

Por otra parte, si el investigador o investigadora desea comprobar una hipótesis teórica que plantea que dos medias no difieren de forma estadísticamente significativa o que no hay una relación entre dos variables (se plantea una hipótesis de igualdad) entonces su técnica de contraste de hipótesis son las denominadas “pruebas de equivalencia” y no el procedimiento tradicional de contraste estadístico para diseños de superioridad (donde se plantean que una media será superior a otra o que habrán diferencias entre las medias).

El procedimiento tradicional de contraste de hipótesis estadística NHST está dirigido a analizar hipótesis teóricas de superioridad donde se plantea que los grupos difieren, ya que las puntuaciones de un grupo serán superiores a las del otro grupo aunque, como ya se ha comentado, todo el contraste se base en la hipótesis de

nulidad de efectos o ausencia de relación entre las variables (H_0). Pero esa cuestión es una de las paradojas que envuelven al contraste estadística.

Una vez ejecutados los análisis estadísticos, se obtienen unos resultados (evidencia) cuyo grado de validez depende de forma directa de la calidad del proceso de diseño de investigación que se haya planificado y realizado en el estudio. En el diseño debe primar la *maximización de la varianza primaria* que se atribuye a la relación entre las variables que forman la hipótesis (variables independiente y dependiente) y *minimizar la varianza no sistemática o error aleatorio* presente siempre en el modelo. Por supuesto, la calidad o validez de los resultados también depende del control efectivo de la *varianza sistemática secundaria* (variable extraña) cuyo efecto no controlado en el diseño invalidaría las conclusiones del estudio (principio MAX-MIN-CON).

Ecuación estructural del modelo

Para comprobar un modelo teórico es necesario formularlo como un modelo matemático (ecuación estructural) y contrastarlo con datos empíricos, realizando un proceso de modelización estadística. El objetivo de dicha modelización es separar la variación que se atribuye al componente sistemático del modelo, efecto = α, (vinculado a los efectos de la variable de tratamiento; varianza sistemática primaria) de la variación relacionada con el componente aleatorio o término de error del modelo, error = ε (varianza secundaria no sistemática o aleatoria). Incorporando en el modelo, si fuese necesario, las fuente de varianza secundaria sistemática que se han controlado y no forman parte de la hipótesis de investigación: por ejemplo, un factor de bloques o una variable covariada. Por lo tanto, la varianza total de las puntuaciones planteada en el modelo se descompone en:

Varianza total = Varianza sistemática primaria + Varianza secundaria no sistemática o error aleatorio

Y si hay alguna variable de varianza secundaria sistemática que se controla con el diseño (ej. bloqueo), habrá que incorporar su efecto en la ecuación estructural:

Varianza total = Varianza sistemática primaria + *Varianza secundaria sistemática controlada* + Varianza secundaria no sistemática o error aleatorio

En este punto es muy importante definir adecuadamente los componentes del modelo lineal (ecuación estructural) que se pretende ajustar a los datos obtenidos con el estudio. De ahí la importancia de revisar y construir de forma adecuada el modelo teórico, pues el modelo matemático que se formule estará determinado por el planteamiento teórico. Planificar de forma cuidadosa el estudio empírico teniendo en cuenta las conclusiones de la revisión teórica efectuada (revisión del conocimiento previo) es fundamental para plantear de forma correcta las hipótesis del estudio y para planificar el diseño y el análisis de los datos posterior.

El Modelo Lineal General (MLG, representado en la ecuación estructural del diseño) es un modelo estadístico que describe una combinación lineal de los efectos aditivos que forman la puntuación directa (Y) obtenida en la denominada variable dependiente:

Y = constante + efecto de la(s) variable independiente(s) + error

En las Ciencias del Comportamiento la estimación de la variable dependiente no está exenta de error, utilizándose modelos estadísticos o probabilísticos a diferencia de los modelos determinísticos donde se plantean relaciones exactas entre las variables. El término de error (ε) representa el término de varianza secundaria aleatoria o no sistemática, es decir, recoge el efecto de otras variables no incluidas directamente en el modelo y cuyo efecto individual no debe ser ni destacado ni sistemático (debe ser error aleatorio).

Diseño de grupos independientes unifactorial univariado

La ecuación estructural o modelo matemático que plantea el Modelo Lineal General (MLG) que describe a un diseño de grupos independientes (conocido como *diseño entre-grupos* o *diseño entre-sujetos*), con una sola variable independiente (diseño unifactorial) y una sola variable dependiente (diseño univariado) se basa en descomponer la puntuación de la variable dependiente Y en tres partes:

1) *Constante*: es la media general (M).

2) *Efecto* o efectos del tratamiento: se trata de la varianza explicada por el efecto de estar en un grupo u otro (se representa como A, B, C…y se trata de la denominada 'varianza entre-grupos').

3) *Error* (E): se trata de la varianza explicada por el efecto que se produce dentro de cada grupo (efecto intra-celdilla) que debe estar vinculado con las posibles fuentes de variabilidad no sistemática como el error aleatorio vinculado al muestreo o las propias diferencias individuales (se trata de la denominada 'varianza-intragrupo' o varianza intra-celdilla). Es decir, los sujetos que reciben una misma condición de la variable independiente (todos se encuentran por ejemplo en el grupo a_1) tendrán el mismo efecto vinculado a dicha condición (efecto de a_1), pero cada sujeto tendrá un término de error de estimación diferente (de ahí el término de variabilidad o varianza intra-grupo), pues a pesar de recibir el mismo tratamiento a_1 no todos los sujetos obtienen la misma puntuación en la variable dependiente Y. En este punto, hay que tener muy en cuenta que la fuente de varianza o de variabilidad del término de error también podría incluir los efectos de otras variables con efecto sistemático no controladas en el diseño (variables contaminadoras sin control). La calidad o validez de los resultados del estudio depende de que ese término de error solamente incluya variabilidad no sistemática o aleatoria y que el efecto de las variables de varianza sistemática secundaria haya sido controlado de forma efectiva con el diseño de investigación. Si el término de error contiene variables extrañas sistemáticas entonces estará sesgado, ya que su valor será mayor y, por lo tanto, afectará a la validez de conclusión estadística, pues se pierde potencia estadística al efectuar el contraste de hipótesis.

La representación de la ecuación estructural en términos poblacionales de un *diseño entre-grupos unifactorial univariado*, bajo el modelo completo o saturado o de la hipótesis alternativa, es la siguiente:

$$Y = \mu + \alpha + \varepsilon$$

Donde Y representa las puntuaciones en la variables dependiente medida (los datos del estudio), μ es la constante o media total de los datos, α es el efecto de la variable independiente o factor (que tendrá al menos dos condiciones: α_1 y α_2, pero podría tener tres o más condiciones) y ε es el término de error que debe ser aleatorio o no sistemático y que se encuentra en los datos vinculado el error de muestreo y/o diferencias individuales.

En términos de matrices, el modelo o ecuación estructural anterior se representa como:

$$\mathbf{Y = M + A + E}$$

Donde, tal y como ya se ha descrito, Y es la variable dependiente, M es la constante o media general, A es el efecto del tratamiento y E es el error aleatorio. Este tipo de representación de matrices es la que se utiliza en este libro para el desarrollo de las ecuaciones estructurales de cada modelo de diseño de investigación.

Por ejemplo, en un *diseño entre-grupos unifactorial (*A = 2*) univariado* los sujetos tendrán una sola medición o puntuación, ya que formarán parte de una condición u otra (entre-grupos), hay una variable independiente (A) que tiene dos condiciones (a_1 y a_2) y una variable dependiente (Y) y la ecuación estructural plantea que las puntuaciones en la variable dependiente Y de cada uno de los sujetos que forman la muestra puede descomponerse en:

1. la media general del grupo o constante,
2. más el efecto de estar en una determinada condición del factor o variable independiente (un efecto para los sujetos del grupo o condición a_1 y otro efecto para los sujetos del grupo a_2, donde la suma de los efectos siempre será cero, ya que la estimación del Efecto es igual a: A = Ma - M),
3. más el error aleatorio no sistemático, propio de la puntuación de cada sujeto, ya que se vincula a las diferencias individuales o errores de muestreo; (Error E= Y - M - A; la suma de los valores de los errores intra-celdilla o dentro de cada grupo también sumará siempre cero).

Como ya se ha comentado, es un requisito fundamental que el diseño controle con alguna técnica de control las variables extrañas o contaminadoras porque, en caso contrario, su efecto estará en el término de error, aumentando su varianza, y, por lo tanto afectando de forma directa al valor del estadístico de la razón *F* del Análisis de la varianza (ANOVA) que disminuirá al estar contaminado el valor de la varianza que se considera aleatoria por factores de varianza sistemática secundaria (problemas de potencia estadística: amenaza a la validez de conclusión estadística). Como ya se comentará, el denominador de la razón *F* es el valor del término de error

y si este valor aumenta entonces disminuye el numerador (que es el valor del efecto del tramiento) y, por lo tanto, disminuye el valor del estadístico *F*.

Una vez que se ha definido la ecuación estructural o modelo matemático que representa a la hipótesis del estudio y recogidos los datos, se procede con la fase de ejecución del *análisis de los datos*.

Una de las técnicas estadísticas que más se utiliza para analizar el efecto de la variación atribuida al efecto del tratamiento respecto a la que procede del error aleatorio es el denominado Análisis de la Varianza (ANalysis Of VAriance, ANOVA) que utiliza la prueba estadística de la Razón *F* para llevar a cabo el contraste estadístico. Se trata de una prueba paramétrica. Este tipo de modelos o pruebas de ANOVA serán analizadas en el presente libro.

Antes de pasar a la presentación y estudio del Análisis de la Varianza (ANOVA) se van a repasar los conceptos de error de muestreo y distribución muestral de las medias ya que son aspectos importantes para comprender los fundamentos y la lógica que subyace al procedimiento clásico de significación de la hipótesis nula (NHST).

Error de muestreo

Cuando se dispone de un conjunto de datos es fácil pensar que una puntuación individual diferirá de la puntuación media y que todas las puntuaciones individuales diferirán de la media en diferentes cantidades (más alejada de la media o menos alejada de la media) y en diferentes direcciones (mayor a la media o menor a la media). Esas diferencias se pueden cuantificar como varianzas y desviaciones típicas o desviaciones estandards (Varianza total: Σ(Puntuación Y – Media Total)^2, es decir $\Sigma(Y - M)^2$). En definitiva, ya se trate de una puntuación única, un conjunto de puntuaciones o una media, se sabe que su valor diferirá del centro de cualquier distribución a la pertenezca. Se trata de la variabilidad de los datos respecto a la media general.

Además, esas ideas con puntuaciones de una muestra de datos son trasladables al área de las muestras de una población de datos. Es decir, del mismo modo que una puntuación individual difiere de su media, también la media de una muestra diferirá de la media real de una población. Se trata del denominado “error de

muestreo". El error de muestreo señala que si se recogen datos de una muestra, el valor observado medio con esa muestra diferirá (aunque sea ligeramente) del dato real de la media poblacional. Este hecho es natural y esperable. Pero qué ha sucedido si la media muestral es extremadamente diferente de lo que se esperaba según la media de la población. En este caso, cabe preguntarse si algo excepcional ha ocurrido. Y la ejecución de las pruebas de contraste estadístico son la herramienta estadística que puede dar la respuesta a dicho interrogante.

Contraste de hipótesis

En resumen, en el contraste de hipótesis se comienza asumiendo que la hipótesis nula es cierta (efecto cero o nula relación entre las variables). Si se rechaza entonces se podrá aceptar la hipótesis alternativa. Se trata de dos hipótesis estadísticas mutuamente excluyentes (si se mantiene H_0 no se acepta H_1 y si se rechaza H_0 se acepta H_1) y exhaustivas. Es decir, son exhaustivas porque solamente existe esos dos posibles resultados en la decisión estadística: mantener H_0 o rechazar H_0. Y si la H_0 es verdadera entonces la H_1 debe ser falsa. Y viceversa, si la si la H_0 es falsa entonces la H_1 debe ser verdadera. Ante esta situación parecía sencillo tomar la decisión: simplemente se trata de determinar si H_0 es falsa o verdadera. Sin embargo, hay un gran obstáculo: nunca se puede probar si H_0 es falsa o verdadera. Ni tampoco si H_1 es falsa o verdadera. Entoces si no se puede probar si una hipótesis estadística es falsa o verdadera, ¿cómo se pueden detectar los efectos estadísticamente significativos? Esta es la pregunta fundamental del contraste de hipótesis estadísticas y que gran parte de la comunidad científica (investigadores e investigadoras y lectores y lectoras) no comprenden de forma adecuada (MIS TRABJSO).

Lo único que se puede hacer ante esa situación es: si la H_0 no se puede probar entonces se trata de fijar una condición que permita rechazarla. Por ejemplo,si se lanza una moneda 500 veces y se obtiene 495 caras, ¿dudaría de esa moneda?, ¿pensaría que está trucada?, es decir, ¿pensaría que tiene algún sesgo o factor que la inclina de forma sistemática hacia la cara? Claramente se puede pensar que se rechaza la hipótesis nula de efecto cero o moneda con un mismo peso en la cara y la cruz, pues sería lógico pensar que hay un sesgo o factor que la inclina hacia la cara y por ello su mayor frecuencia. Y por qué pensamos así. Porque un resultado

tan raro es improbable que hay sido el resultado del azar y más bien se puede pensar que es un factor de efecto sistemático en la moneda o en el modo como se lanza la moneda. Por supuesto, los resultados improbables (aunque raros o con menor frecuencia) también podría ser efecto del azar y de ahí que siempre haya un error estadístico o una equivocación en la decisión estadística.

Distribución muestral

Entonces en un contraste de hipótesis ¿qué condiciones hay que fijar para poder rechazar H_0? Para dar respuesta a este interrogante y poder tomar la decisión estadística se dispone de la "distribución muestral de un estadístico". Otro concepto fundamental y necesario para entender la decisión estadística de mantener o rechazar H_0. Los valores que forman esa distribución muestral son las condiciones fijadas para mantener o rechazar H_0, valores que serán la referencia para evaluar el resultado del estadístico utilizado en una investigación.

Los datos de las distribuciones muestrales de cada estadístico se encuentran en las conocidas tablas de distribución teórica del estadístico que se encuentran en los libros de estadística y de diseños de investigación. Generalmente, las tablas no proporcionan los valores de probabilidad de las distribuciones muestrales. Esas tablas proporcionan los valores críticos o teóricos (tabulares) que definen la región de rechazo y suelen presentar esos datos con varios niveles de alfa o nivel de significación (tablas de .05, tablas de .01...).

La región de rechazo es aquella parte del área bajo una curva que incluye todos los valores de un test estadístico que conducen al rechazo de la hipótesis nula. Y por ello, cuando se rechaza la hipótesis con un valor concreto del estadístico se dice que se rechaza para ese valor o un valor más extremo ya que la región de rechazo define a todos los resultados que implican el rechazo de la hipótesis nula. Por ejemplo, si con un valor de 7 se rechaza H_0 se puede dedudir que si el valor hubiese sido de 7.1, 7.2 ... 8, 8.1 ... también se rechazaría la hipótesis nula.

Una distribución muestral es una distribución de probabilidad teórica de los posibles valores de algún estadístico que ocurriría si se pudieran obtener de forma aleatoria todas las muestras posibles de una determinado tamaño fijo a partir de una población concreta. Hay una distribución muestral para cada estadístico: media, desviación típica, varianza, proporción, mediana...

Un *estadístico* es un estimador que se utiliza para conocer un parámetro poblacional desconocido. El estadístico es una función de los valores de la muestra y es una variable aleatoria ya que sus valores dependen de la muestra seleccionada. La distribución de probabilidad teórica de un estadístico se conoce como "*distribución muestral del estadístico*".

Cuando se tiene un conjunto de puntuaciones individuales se puede definir su distribución con una puntuación media, por ejemplo. Del mismo modo, si se pueden crear muchas muestras de datos, todas del mismo tamaño, también se puede construir su distribución y calculando la media de cada una de las muestras extraidas se puede construir la denominada "*distribución de medias muestrales*".

Por lo tanto, cuando se está trabajando con variables cuantitativas y se tiene un tamaño suficiente de muchas muestras de datos (todas del mismo tamaño) su distribución acaba adquiriendo la forma de campana de Gauss que se denomina distribución normal. Cuando se dispone de esa distribución muestral ya es posible saber exactamente cuál es la probabilidad de que se obtenga un dato determinado en la variable. Como se puede observar, en la distribución normal los valores más próximos a la media tienen una probabilidad más alta (son más frecuentes o hay más casos) respecto a los valores de los extremos (menos frecuentes o hay menos casos). Concretamente, entre la puntuación de -1.96 y la 1.96 se encuentran el 95% de los valores de la variable.

La distribución muestral del estadístico es el referente para comparar los resultados del estudio con los valores de esa distribución teórica y conocer qué probabilidad tiene ese resultado en esa distribución de efectos nulos donde solo actúa el muestreo aleatorio (el azar) como elemento que crea las diferencias en los valores de frecuencia obtenidos.

Un ejemplo de distribución muestral puede ser la de la media, la de la correlación…, es decir, hay distribuciones muestrales para cualquier estadístico.

Distribución muestral de las medias muestrales

Nos vamos a centrar en la distribución de las medias muestrales para desarrollar el concepto. Conocer la distribución muestral del estadístico es imprescindible para poder llevar a cabo los análisis del proceso de inferencia estadística sobre el

parámetro correspondiente. Por ejemplo, para poder realizar inferencias sobre la media poblacional es necesario conocer la distribución muestral de la media.

Como ya se ha descrito, la *distribución muestral* de un estadístico es la distribución de todos los valores posibles de un estadístico concreto, dado un tamaño fijo de muestra concreto, incluyendo las probabilidades asociadas a cada uno de ellos. Por ejemplo, si se pudiesen obtener todas las muestras de un tamaño determinado de una población y se computase un estadístico (la media, la proporción, la desviación típica...) para cada muestra de dicha distribución (sus valores diferirán de forma natural ya que hay un error de muestreo) entonces la distribución de probabilidad de ese estadístico sería la distribución muestral. Por ejemplo, la distribución muestral de la media incluye todos los valores posibles del estadístico media en las condiciones fijadas de extracción y para un tamaño de muestra concreto o fijo.

La distribución muestral de las medias muestrales tiene la forma de una curva normal, es decir, es una curva en forma de campana con un solo pico y dos colas simétricas. El centro de la distribución muestral de las medias muestrales es la media de las medias muestrales y representa a la media real de la población (μ de las medias muestrales). La amplitud de la distribución muestral se denomina error estándar o error típico (σ de las medias muestrales) y cuantifica el error de muestreo.

Veamos un ejemplo. Si usted decide computar la media de una muestra de 10 elementos, el valor que obtenga no será igual al valor de la media de la población de dicha muestra ya que por azar será o un poco mayor o un poco menor. Si lo que hace es computar una media para un conjunto de muestras de 10 elementos cada una (computa la media cada diez elementos) entonces podrá observar que algunas muestras están más cercas a la media de la población que otras. Imagine que computamos la media una y otra vez con muestras de n = 10 seleccionadas al azar de la población y construimos una distribución de frecuencias de unas 1.000 medias muestrales. Esa distribución de medias es una buena aproximación a la distribución muestral de la media. A medida que el número de muestras se aproxima a infinito, la distribución de frecuencias relativas se aproxima a la distribución muestral del estadístico. Y este hecho ocurre independientemente de como sea la distribución poblacional, pues a medida que se recogen más muestras de datos más se aproxima

la distribución muestral del estadístico a la distribución normal. La estadística ha demostrado que la distribución muestral de medias tiende a ser normal cuando el tamaño de la muestra supera 30. Y los valores que forman esa distribución muestral pueden tener una mayor frecuencia (son más comunes) o una menor frecuencia (son más raros). Esa mayor o menor frecuencia está asociada a su mayor o menor probabilidad y es la clave para llevar a cabo el proceso de decisión estadística ante el resultado empírico del estudio.

Si el resultado empírico del estudio tiene una baja frecuencia en la distribución muestral del estadístico, es decir, es un resultado raro, entonces se puede concluir que hay otros factores que no son el azar (error de muestreo) los responsables de ese resultado extraño o raro, interpretándose que será la variable de intervención o de estudio la que está provocando ese dato. Y esos eventos raros o inusuales solo ocurrirán un 5% de las veces o menos, concluyéndose ante esos eventos raros que se puede rechazar la hipótesis nula como explicación del fenómeno, y se opta por la hipótesis alternativa. Ese punto de corte que sirve para inferir si está actuando el azar o se trata de un efecto sistemático se conoce como el nivel de significación del 5% y es la condición que fija el rechazo de la hipótesis nula. Es decir, es la condición que fija que el resultado es estadísticamente significativo. Se pueden utilizar criterios o condiciones más conservadoras para rechazar H_0, como disminuir el nivel del significación al 1%, hecho que supone buscar eventos aún mucho más raros o poco frecuentes en la distribución muestral del estadístico.

La dispersión que presentan los valores de la distribución muestral (conocido como 'error típico' o error estándar del estadístico) es más pequeña que la dispersión de la variable aleatoria X (conocida como desviación típica).

El error típico o estándar de la distribución muestral se obtiene así,

$$ErrorTípico(\sigma_{muestra}) = \frac{\sigma}{\sqrt{n}}$$

Donde *n* es el tamaño de las muestras con las que se realiza la distribución muestral y σ es la desviación típica de la población.

En la siguiente dirección de Internet se pueden realizar ejercicios de simulación de la distribución muestral de varios estadísticos. Por ejemplo, se puede practicar la

distribución muestral del estadístico de la media utilizando la simulación que se ofrece en la siguiente dirección Web: http://onlinestatbook.com/stat_sim/index.html

En ese programa on-line, una vez el usuario o usuaria se encuentra en la pantalla de la simulación debe especificar dos tipos de información:

1. La distribución de la población de la que se van a tomar las muestras aleatorias (distribución normal o gaussiana, distribución uniforme o rectangular, distribución asimétrica). En el primer histograma aparece la media, la mediana y la desviación típica de la distribución de la población.
2. El tamaño de la muestra y el estadístico que desea utilizar en la simulación.

El *applet* de la simulación devuelve una muestra de la población y grafica la muestra del estadístico. Esto se puede repetir una y otra vez para estimar la distribución muestral del estadístico. En la figura 19 se representa la pantalla de la simulación.

Los datos de la figura 19 indican que se ha seleccionado la distribución normal como población para extraer las muestras. El estadístico que se va a estimar en la distribución es la media. Es decir, la distribución poblacional subyacente es la distribución normal con media y mediana de 16 y desviación típica de 5.

No hay que confundir tamaño de la muestra (de cada muestra individual que se computa) con número de muestras recogidas (réplicas o reposiciones). En la distribución de medias que se ha graficado se han utilizado 1001 muestras (réplicas, 'Reps'), siendo el tamaño de cada muestra de 10 elementos. La distribución muestral de medias obtenida con tamaños de muestra de 10 datos (N = 10) tiene una media de 15.97, una mediana de 16, una desviación típica (error típico) de 1.59, la asimetría es de 0.06 y la curtosis es igual a -0.08.

El programa permite ejecutar al mismo tiempo una segunda distribución muestral donde se puede cambiar el tamaño de las muestras (aparecería en un cuarto histograma).

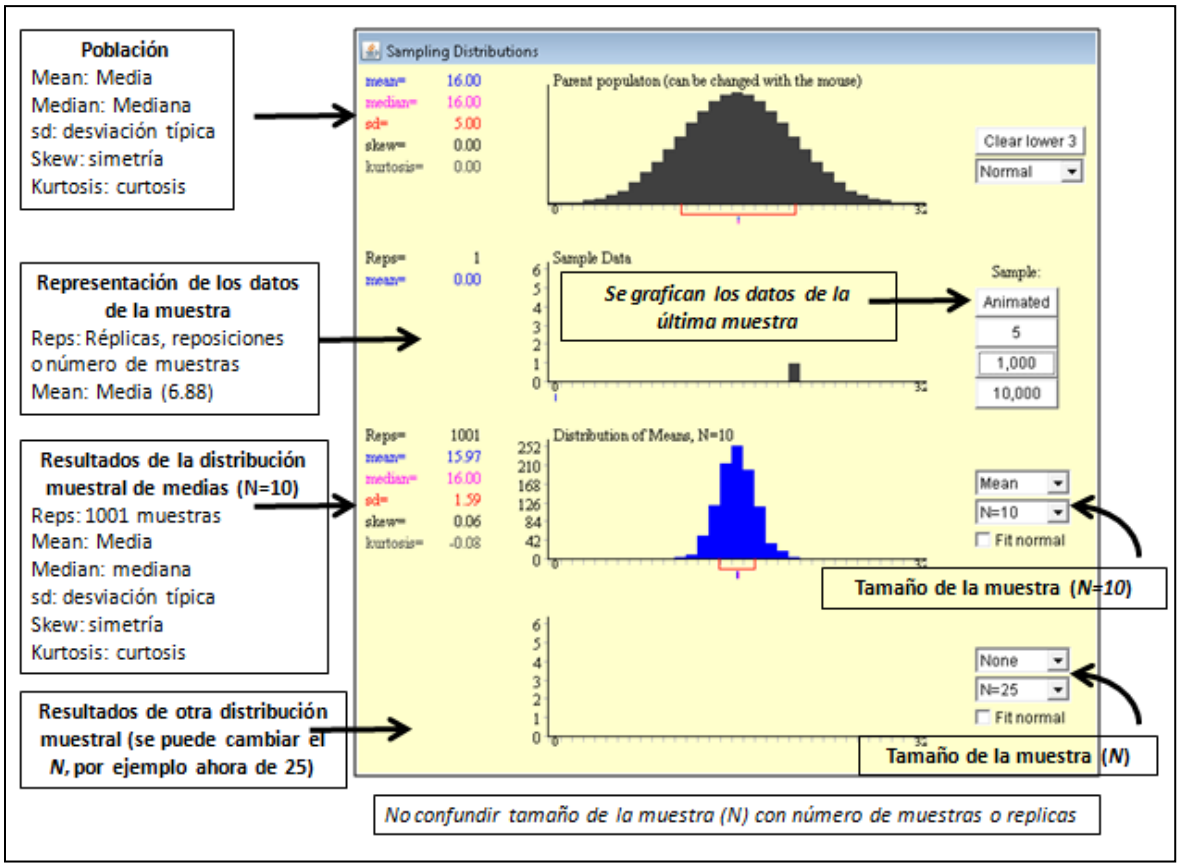


Figura 19. Simulación de la distribución muestral de la media

A continuación se detalla otro ejemplo. Se extraen cinco datos (medias) al azar y la representación de los valores de esos cinco datos es la que aparece en el histograma segundo de la pantalla B, debajo de la distribución de la población ('Sample Data'). Se observa un único valor en el tercer histograma de la pantalla B que corresponde a la media de los cinco datos anteriores (es la media de las medias muestrales). Dicha media de las medias muestrales ha sido calculada con los *N* = 5 elementos (medias) que se observan en el segundo histograma de la pantalla B. Ese primer valor de la distribución muestral de la media es la media de los cinco datos o medias que aparecen en el segundo histograma (Sample Data: datos de la muestra). Se ha obtenido la primera media muestral de la distribución muestral de la media.

Se pueden observar a la izquierda de los histogramos los estadísticos descriptivos de la distribución de frecuencias. En el caso de la distribución poblacional (*parent population*) la media (*mean*) es 16, la mediana (*median*) 16, la desviación típica (*sd*) es de 5. La simetría (*skew*) y la curtosis (*kurtosis*) es igual a cero.

Si se observa la distribución muestral del estadístico de la media (Pantalla B, *Distribution of Means*, $N = 5$) se observan los estadísticos descriptivos de la distribución muestral de medias que hasta ese momento se han calculado. Por lo tanto, a medida que se añaden más medias muestrales a la distribución entonces van cambiando los datos de media, mediana, desviación típica, simetría y apuntamiento de la distribución muestral de la media (Figura 20).

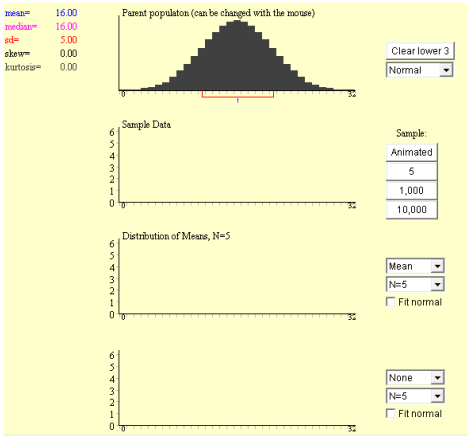
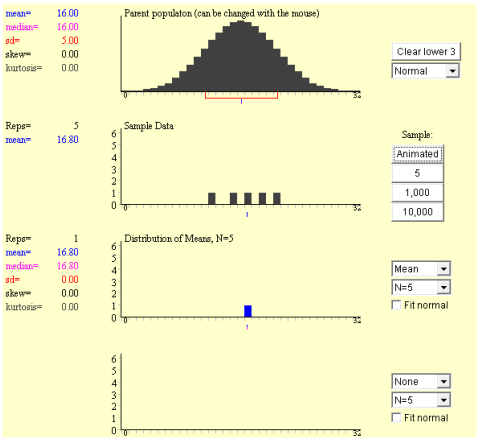
Pantalla A	Pantalla B
	

Figura 20. Simulación de la distribución muestral de la media

Se puede seguir con la simulación seleccionando nuevas muestras de cinco elementos cada vez y se observará cómo se va construyendo la distribución muestral de las medias (tercer histograma). Con ello, poco a poco la distribución muestral de la media adoptará la forma de la curva normal.

También se puede realizar la simulación comparando al mismo tiempo dos distribuciones de medias muestrales con dos tamaños muestrales diferentes por ejemplo $N = 2$ (se presenta en el tercer histograma) y $N = 5$ (se presenta en el cuarto histograma). De este modo se puede reflexionar sobre la importancia del tamaño de la muestra.

Características de la distribución muestral de la media

Si se extraen muestras aleatorias de tamaño *n* de una población infinita que tiene media poblacional μ y varianza σ^2, entonces por el Teorema del Límite Central se sabe que:

1. El valor esperado de la media de las medias muestrales (distribución muestral de la media) es igual a la media de la población: $E(\overline{X}) = \mu$.
2. La varianza de las medias muestrales (distribución muestral de la media) es igual a la varianza poblacional dividido por el tamaño de la muestra, $V(\overline{X}) = \sigma^2 / n$. En consecuencia, la desviación estándar de las medias muestrales (llamada también el error típico o estándar de la distribución muestral de la media), es igual a la deviación estándar poblacional dividida por la raíz cuadrada de *n*: $V(\overline{X}) = \sigma / \sqrt{n}$.

Por lo tanto, gracias a dichas expresiones se pueden relacionar las medias muestral y poblacional mediante una distribución de probabilidad conocida como es la distribución muestral del estadístico. En otras palabras, gracias al conocimiento que se tiene de las distribuciones muestrales se puede conocer la probabilidad asociada a un valor del estadístico obtenido en una muestra de *n* elementos.

En resumen, la media de una distribución muestral es una variable aleatoria que tiene un determinado comportamiento que la estadística ya ha estudiado. Por ejemplo, si la variable que se estudia tiene una distribución normal entonces la media muestral también será normal, pero con una desviación menor. Además, si la variable no tiene una distribución normal (por ejemplo es rectangular), pero la muestra es lo bastante grande entonces la media de la distribución muestral también se aproxima a la distribución normal (Teorema del Límite Central).

La distribución de la media muestral de una población normal es una distribución normal con la misma media poblacional y con una desviación típica conocida como error típico o error estándar. Este hecho permite calcular probabilidades cuando se tiene una muestra de una variable con distribución normal y desviación típica conocida. Cuando no se conoce la desviación típica de la variable (varianza poblacional desconocida), también se pueden hacer cálculos con otros estadísticos como por ejemplo la distribución *t* de Student.

Técnicas de inferencia estadística

En la mayor parte de las ocasiones el proceso de evaluación empírica de las hipótesis se realiza mediante un modelo de decisión probabilística a través de la ejecución de una prueba o test de significación estadística.

La prueba estadística es una fórmula que está basada en la distribución muestral del estimador del parámetro (estadístico) que aparece en la hipótesis estadística. Gracias a dicha prueba se podrá tomar una decisión estadística: mantener o rechazar la hipótesis nula, siempre con un margen de error de equivocación (errores estadísticos).

La inferencia estadística es la parte de la Estadística que incluye los métodos utilizados para tomar decisiones o para obtener conclusiones sobre una característica desconocida de la población a partir de los datos obtenidos con una muestra representativa de la población. Su aplicación requiere el estudio de la teoría de la probabilidad. La técnica del Análisis de la Varianza (ANOVA) permite analizar si la diferencia entre las medias de diferentes condiciones tiene una magnitud más allá de lo que podría pensarse por el error aleatorio natural y propio que se produce al extraer las muestras de la población (error de muestreo).

Análisis de la varianza (ANOVA)

La técnica del Análisis de la Varianza (ANOVA) permite comparar dos o más medias entre sí (medias cuadráticas) para determinar mediante el proceso de contraste de hipótesis estadísticas si la diferencia (o las diferencias) entre las medias de la variable dependiente en los grupos de la variable independiente es estadísticamente significativa. Para ello, la ecuación estructural plantea un modelo matemático o ecuación estructural donde la variable dependiente Y es la variable cuantitativa continua y la o las variables independientes (factores del modelo: A, B, C …) son las variables cualitativas o factores que estarán formados por diferentes condiciones que representarán diferentes efectos (efectos de las condiciones a_1 y a_2: α_1 y α_2; efectos de las condiciones b_1 y b_2: β_1 y β_2; y efectos de las condiciones c_1 y c_2: χ_1 y χ_2 ..., por ejemplo).

La denominada Razón *F* es el estadístico o prueba de contraste que se utiliza en el análisis de la varianza.

Aunque la prueba de la hipótesis se ha formulado para comprobar cierto orden en las medias de las condiciones de las variables independientes, la prueba de la hipótesis del estadístico *F* compara varianzas; de ahí que el nombre que reciba esta prueba sea precisamente el de análisis de la varianza. Como ya se ha comentado, la

prueba estadística es completamente válida en la medida que se cumplan una serie de supuestos fundamentales de las pruebas paramétricas, destacando especialmente que la variable dependiente se mida al menos en escala de intervalo, que se distribuya normalmente y que las distintas observaciones sean independientes y extraídas aleatoriamente, no existiendo entre las condiciones experimentales ninguna otra diferencia que la manipulación de la variable independiente o la relación predicha entre las variables.

En los diseños de grupos independientes o diseño entre-grupos (diseños entre-sujetos), la Razón *F* compara la 'varianza entre los grupos' (efecto de la variable independiente en cada condición o grupo) con la 'varianza intra-grupo' (error aleatorio o varianza aleatoria que se produce dentro del grupo). Si se concluye que la diferencia entre la varianza entre-grupos y la varianza intra-grupos es estadísticamente significativa (el valor *p* del resultado obtenido con la prueba estadística de la razón *F* es $\leq$ alfa (generalmente alfa = .05) dentro del modelo de la distribución de la hipótesis nula) entonces al menos una de las medias es diferente de forma estadísticamente significativa de al menos otra media de la variable dependiente. De ahí el nombre de análisis de la varianza, ya que compara la varianza entre-grupos con la varianza intra-grupo para detectar diferencias entre las medias de los diferentes grupos o condiciones que forman el diseño de la investigación. La fórmula del estadístico de la Razón *F* es la siguiente:

$$F = \frac{\textbf{Varianza del efecto}}{\textbf{Varianza del error}}$$

En resumen, la técnica de ANOVA permite conocer el vínculo de los cambios en la variable dependiente según las condiciones o niveles de las variables independientes. Es decir, permite analizar si las puntuaciones en la variable dependiente difieren o no de forma estadísticamente significativa cuando se comparan los resultados de cada grupo o condición de la variable independiente.

Cuando el diseño tiene más de dos grupos, puede ser que haya una sola diferencia estadísticamente significativa entre dos grupos o puede que hayan más diferencias entre las medias de los grupos, pero el ANOVA no especifica el número de grupos que difieren ni qué grupos difieren, ya que se trata de una prueba 'omnibus'

y solo informa de que hay alguna diferencia estadísticamente significativa entre los grupos.

Por lo tanto, si el diseño tiene más de dos grupos (A > 2, tiene más de una diferencia de medias), será necesario continuar el análisis de los datos con otras técnicas que permitan descubrir entre qué grupos se encuentran las diferencias estadísticamente significativas como las 'pruebas de contraste de hipótesis específicas', también conocidas como pruebas post hoc o pruebas a posteriori (por ejemplo, Tukey, Bonferroni...).

Modelos de la hipótesis nula y la hipótesis alternativa

En la figura 21 se representa la formulación de la ecuación estructural de un *diseño entre-grupos unifactorial univariado* bajo el planteamiento de la hipótesis nula (modelo restringido) y de la hipótesis alternativa (modelo completo o saturado). Se observa en la ecuación estructural del modelo de la hipótesis nula o modelo restringido que la descomposición de la puntuación en la variable dependiente solamente incluye la constante o media general (M, que será la puntuación predicha o puntuación pronostica por el modelo para cada sujeto del estudio) más el término de error:

$$H_0 \equiv Y = M + E$$

En cambio, cuando se trata del modelo completo o modelo de la hipótesis alternativa, la ecuación estructural sí tiene el efecto del tratamiento o el efecto de estar en un grupo u otro (A) junto con la constante o media general y el error:

$$H_1 \equiv Y = M + A + E$$

En el modelo de la hipótesis alternativa la puntuación predicha por el modelo para cada sujeto será la constante más el efecto que haya producido estar en ese grupo o condición (puntuación predicha $\hat{Y} = M + A$). Por ejemplo, para el sujeto 1 que se encuentra en la condición 1 de la variable independiente la puntuación predicha es la constante más el efecto que tenga el nivle o condición 1:

$$\hat{Y}_{Sujeto1} = M + A_1$$

Por lo tanto, los sujetos que se encuentren en el mismo grupo (por ejemplo grupo de a_1) tendrán el mismo efecto de A (efecto de α_1) y la suma de los efectos de todos los grupos siempre será igual a 0 ya que la estimación de los efectos son desviaciones respecto a un valor medio (estimación del efecto de A = M_a - M). La suma del cuadrado de los efectos estimados es igual a la Suma de Cuadrados del efecto de A ($\Sigma(A)^2$).

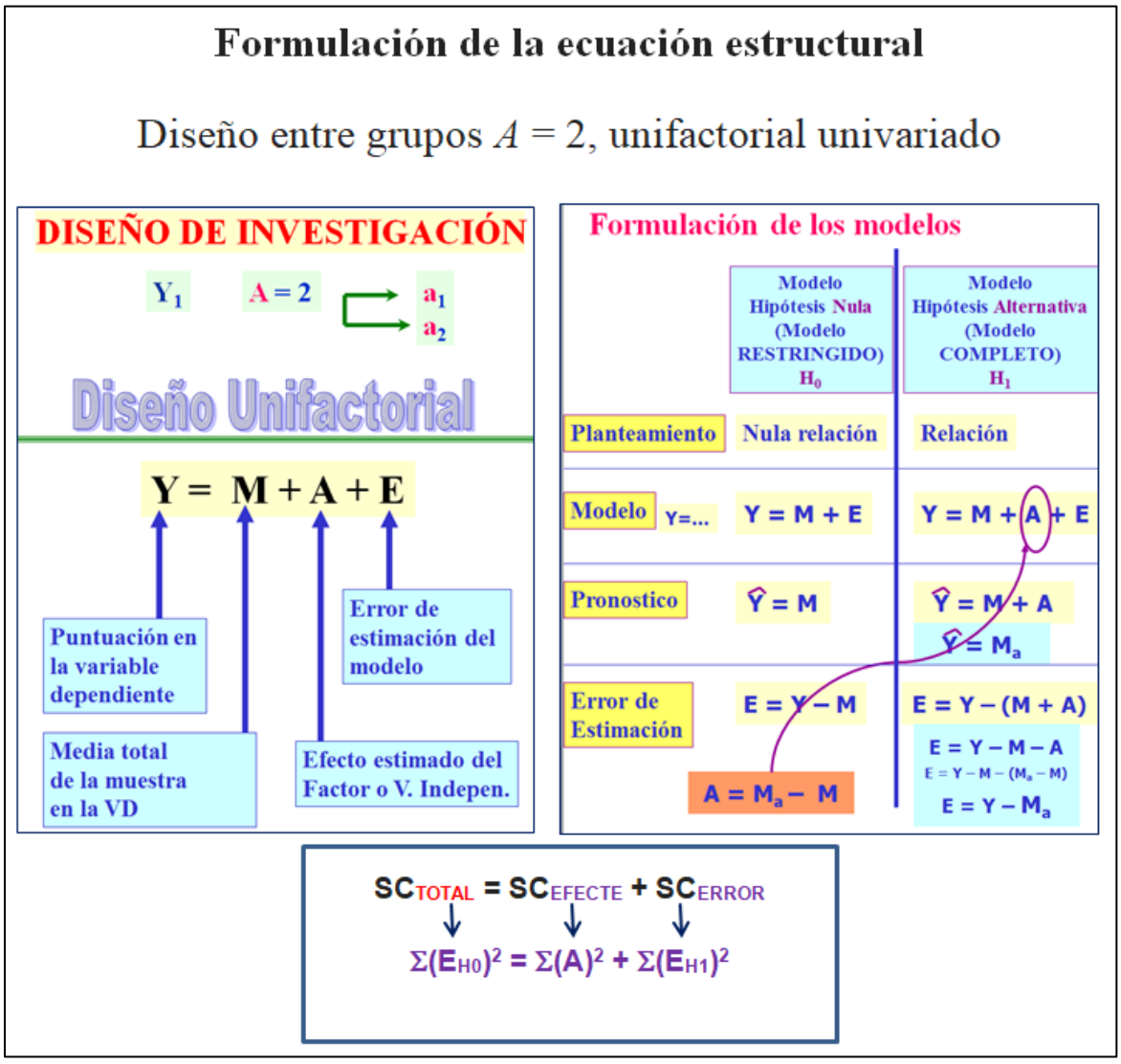


Figura 21. Formulación de los modelos: hipótesis nula / hipótesis alternativa

Cada modelo estadístico plantea un pronostico o una puntuación pronosticada para cada uno de los sujetos que forman la muestra en función del planteamiento que subyace a su hipótesis.

Puntuación pronosticada y error

Modelo de la hipótesis nula. En el caso del modelo de la hipótesis nula que señala que no hay relación entre las variables o el efecto es cero, la puntuación que pronostica para todos los sujetos es la media general o constante (M). Es decir, dado que plantea que no hay efecto de la variable independiente, el pronóstico de las puntuaciones solamente se basa en el valor de la media general de todas las puntuaciones. En este caso, el error de estimación es la diferencia entre la puntuación obtenida en el estudio (dato directo en la variable Y) menos la puntuación pronostica que es la media general, luego $E = Y - M$. La suma del cuadrado de los errores estimados con el modelo de la hipótesis nula es igual a la Suma de Cuadrado Total ($\Sigma(E_{H0})^2$).

Modelo de la hipótesis alternativa. En el modelo de la hipótesis alternativa (plantea que sí hay relación entre las variables o efecto diferente a cero) se introduce en la ecuación estructural la fuente de la varianza del efecto de la variable y, por lo tanto, la puntuación que pronostica para cada sujeto será la constante o media general (M) más el efecto de estar en un determinado grupo del diseño (A), tal y como ya se ha comentado. En este caso, el error de estimación es la diferencia entre la puntuación obtenida en el estudio Y menos la puntuación pronostica por el modelo en su ecuación estructural: $E = Y - M - A$. La suma del cuadrado de los errores estimados con el modelo de la hipótesis alternativa es igual a la Suma de Cuadrado del Error ($\Sigma(E_{H1})^2$).

En resumen, el componente del error de estimación del modelo o ecuación estructural siempre es la puntuación obtenida en el estudio menos la puntuación pronostica por dicho modelo. Recordar que el símbolo $\hat{Y}$ es la puntuación pronosticada en la variable dependiente Y. Así, una vez formulada la ecuación estructural resulta sencillo calcular el error de estimación del modelo como:

$$E = Y - \hat{Y}$$

Tal y como se ha comentado, la Suma de Cuadrados Total es igual a la suma de los cuadrados de los valores del error de estimación del modelo nulo ($\Sigma(E_{H0})^2$). Posteriormente, se detallará la explicación del concepto de sumas de cuadrados.

Además, la variabilidad total o varianza total observada en las puntuaciones Y será igual a la varianza explicada por el efecto de la variable independiente (SC, Suma de Cuadrados del efecto) más la varianza explicada por el término de error (SC, Suma de Cuadrados del error):

$$SC_{TOTAL} = SC_{EFECTE} + SC_{ERROR}$$
$$\Sigma(E_{H0})^2 = \Sigma(A)^2 + \Sigma(E_{H1})^2$$

Para desarrollar la ecuación estructural se puede utilizar la plantilla de la figura 22 donde se especifica cada uno de los pasos necesarios para ejecutar el Análisis de la Varianza (ANOVA) y se explica cómo estimar cada uno de los elementos que forman la descomposición de la ecuación estructural hasta llegar a la tabla del análisis de la varianza o tabla de ANOVA (ver también Anexo 2).

Diseño entre grupos A = 2, unifactorial univariado

Pasos para desarrollar la ecuación estructural del modelo de diseño:

Y = M + A + E

a (condición: a1, a2, a3 ...)	***S*** (sujeto: 1, 2, 3, 4, 5 ...)	**Y** (puntuaciones en la variable dependiente)	**M** (Media general o constante)	***y*** **(Y – M)** E_{H0}	**A** $(M_a - M)$	$\hat{Y}$ **(M + efectes)** Puntuación pronosticada bajo la H1	**E** $(Y - \hat{Y})$ $E_{H1} = Y - M_a$
SC (Suma de Cuadrados)							
gl (grados de libertad)							
MC (Media Cuadrática)							
				SCtotal	**SCentre**		**SCerror**

Figura 22. Plantilla para desarrollar la ecuación estructural

La tabla del ANOVA

La tabla del ANOVA (Análisis de la Varianza) de un diseño entre-sujetos unifactorial univariado está formada por las fuentes de varianza relacionadas con:

1) El efecto de la variable independiente (VI) (fuente de varianza entre-sujetos, A). En esa fuente entre-sujetos se encuentran los factores de varianza sistemática primaria y, si los hubiera, los factores de varianza sistemática secundaria controlada como la variable de bloques.

2) El término de error (S / A: sujetos ligados a un grupo o condición concreta de la variable independiente) o varianza secundaria no sistemática (aleatoria).

3) La variabilidad o varianza total de las puntuaciones que es la suma de los dos componentes anteriores también se suele incluir.

En función del diseño seleccionado para dar respuesta a las hipótesis de investigación (teniendo en cuenta siempre el principio MAX-MIN-CON: maximizar el efecto de la varianza sistemática primaria, minimizar el error y controlar la varianza sistemática secundaria), las fuentes de varianza entre-sujetos pueden ampliarse a 2, 3 factores o más y sus interacciones si el modelo lo especifica (diseño factorial no aditivo) o sin interacciones (diseño de bloques aditivo).

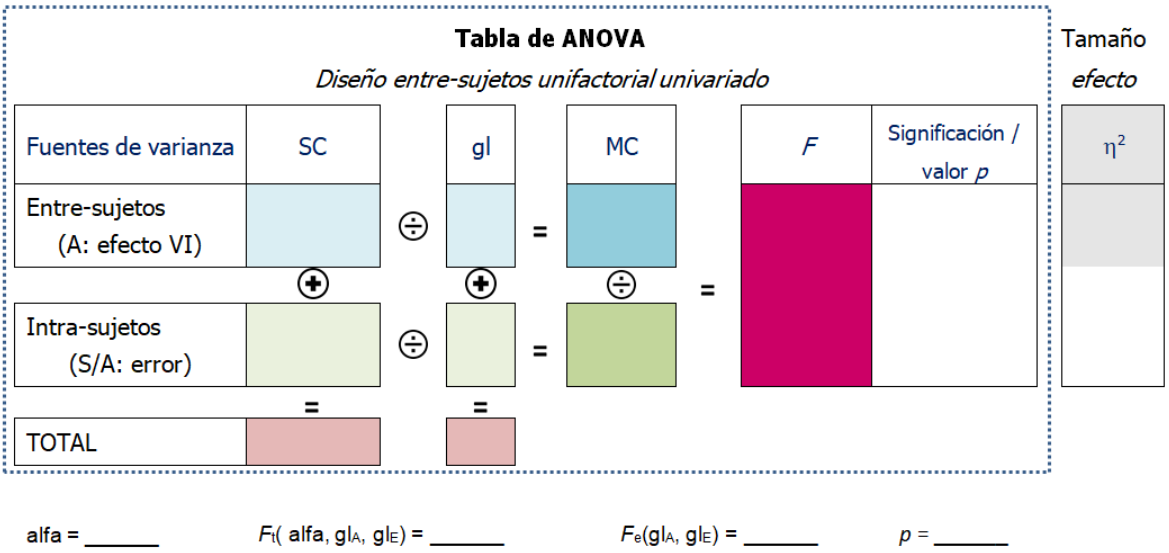
Además, el término de error de la ecuación estructural también puede ser de 2 tipos cuando el diseño es de medidas parcialmente repetidas o mixto ya que en este caso tiene:

1) un término de error para las fuentes de varianza entre-sujetos (S / A: sujetos ligados al factor A) y

2) otro término de error para las fuentes de varianza intra o de medidas repetidas (S *x* B: sujetos cruzados en el factor B) que también se utiliza para las interacciones entre los factores entre-sujetos e intra-sujetos (A *x* B).

Por ejemplo, si el diseño es mixto o de medidas parcialmente repetidas A *x* B, siendo A un factor entre-sujetos y B un factor de medidas repetidas, hay un término de error para la fuente de varianza entre-sujetos (S / A) y otro término de error para la fuente de varianza intra-sujetos (factor A) y para la interacción AB (A *x* B).

En la figura 23 se representa los elementos que componen la Tabla de ANOVA de un diseño entre-sujetos univariado junto con una columna para especificar el tamaño del efecto que se corresponde con los resultados del análisis; concretamente se ha añadido el estadístico denominado eta cuadrado (η^2) o proporción de varianza explicada ($SC_{efecto\ A}$ divido por la SC_{total}).



Nota: **SC**: Suma de Cuadrados, **gl**: grados de libertad, **MC**: Media Cuadrática, ***F***: estadístico de contraste de la Razón *F*, **Significación / valor *p***: el valor de probabilidad asociado al resultado empírico de la Razón $F_{empírica}(gl_A, gl_E)$ (asociado a ese valor concreto o un valor más extremo) de la prueba estadística ejecutada, asumiendo que la hipótesis nula es cierta. En esa última casilla se debe anotar el valor concreto de significación cuando se dispone de un programa estadístico (por ejemplo: $p = .007$ o quizás $p = .255$) y cuando no se dispone del resultado concreto será necesario acudir a las tablas de la Razón *F* y consultar el valor teórico que se corresponde al diseño ejecutado, teniendo en cuenta los grados de libertad 'entre' o del factor y los grados de libertad 'intra' o del error para un valor de alfa concreto ($F_{teórica}$(alfa, gl_A, gl_E). En el caso de consultar las tablas, se debe situar el símbolo de <, >, o = antes del valor de alfa elegido a priori por el investigador o investigadora (por ejemplo si el alfa es .05 entonces se escribirá $p < .05$ o $p > .05$ o $p = .05$; si el alfa es .01 entonces situar la opción que corresponda: $p < .01$ o $p > .01$ o $p = .01$). $\mathbf{SC_{TOTAL}} = SC_{ENTRE\text{-}SUJETOS} + SC_{INTRASUJETOS}$. Y $\mathbf{gl_{TOTALES}} = gl_{ENTRE\text{-}SUJETOS} + gl_{INTRASUJETOS}$. Para realizar de forma manual el contraste estadístico se necesita conocer la siguiente información: valor de alfa, $F_{teórica}$, $F_{empírica}$ y finalmente se obtiene el valor *p* de probabilidad vinculado al resultado empírico obtenido de la Razón *F*, o un valor más extremo, asumiendo que la hipótesis nula es cierta. Si $\boldsymbol{F_{empírica} \geq F_{teórica}}$ entonces $\boldsymbol{p \leq}$ **alfa** y, por lo tanto, se rechaza la hipótesis nula, es decir, hay alguna diferencia entre las medias de los grupos que es estadísticamente significativa. $\boldsymbol{\eta^2}$: Tamaño del efecto vinculado al estadístico eta cuadrado o proporción de varianza explicada.

Figura 23. Tabla de ANOVA y tamaño del efecto

Estimación de los parámetros: término del efecto de A

Una vez el investigador o investigadora identifica el modelo estadístico que representa a la hipótesis de investigación (identificados en una ecuación estructural), a continuación es necesario estimar los parámetros que componen la ecuación estructural.

Los parámetros se pueden definir como 'efectos principales', como 'efectos simples', 'efectos de interacción', 'error' o incluso como 'contrastes concretos'.

La estimación del parámetro del efecto principal del factor A (variable independiente) se define como la media del grupo o condición donde se encuentra el sujeto menos la media poblacional, midiéndose por lo tanto en la escala de puntuaciones de diferencia:

$$\alpha = \mu_{ai} - \mu$$

Es decir,

$$A = M_a - M$$

La suma de los efectos de todas las condiciones de la variable independiente siempre será igual a 0:

$$\Sigma(A) = 0$$

Estimación de los parámetros: término de error

La estimación numérica del parámetro del error mediante el criterio de los 'mínimos cuadráticos' supone obtener un estimador que minimice el componente aleatorio del modelo, ε. Se trata de obtener la minimización de la suma cuadrática de los errores, de manera que la suma de los errores de estimación del modelo debe ser la mínima posible. Es decir, la predicción de los valores de la variable dependiente que realiza el modelo estadístico formulado debe incluir el menor error de estimación posible.

El error o residual estimado $\hat{\varepsilon}_i$ del modelo o ecuación estructural refleja la diferencia entre el valor de la puntuación obtenida en la variable dependiente Y y el valor predicho por el modelo para la puntuación estimada en dicha variable $\hat{Y}$.

$$E = Y - \hat{Y}$$

Y, como ya se ha comentado, la puntuación predicha de la variable dependiente $\hat{Y}$ se estima como:

$$\hat{Y}_i = M + \text{TODOS LOS EFECTOS FORMULADOS EN EL MODELO}$$

Así, en el caso de un diseño con un solo factor o variable independiente (factor A) solo existirá el efecto de A que se descompone en los efectos de cada condición que tenga. La puntuación predicha de la variable dependiente $\hat{Y}$ se estima como:

$$\hat{Y} = M + A$$

Por lo tanto, la estimación del error se efectúa como la puntuación directa obtenida Y menos la suma de la constante y todos los efectos formulados en el modelo):

$$\hat{E} = Y - (M + \text{TODOS LOS EFECTOS FORMULADOS EN EL MODELO})$$

En el caso de un diseño con un solo factor (factor A), el error E se estima como:

$$E = Y - (M + A)$$

Luego,

$$E = Y - M - A$$

Y, en este tipo de diseño entre-grupos unifactorial, se puede comprobar que si $A = M_a - M$, entonces, $Ei = Y - M - (M_a - M)$, y por lo tanto:

$$E = Y - Ma$$

La suma de los errores de todas las puntuaciones siempre será igual a 0:

$$\Sigma(E) = 0$$

Estimación de los parámetros: variabilidad total

La variabilidad total o varianza total de un conjunto de puntuaciones Y se obtiene restando a cada puntuación la media general o constante (M). Así la variabilidad total (*y*) que se observa en un conjunto de datos es igual a:

$$y = Y - M$$

La suma de cada una de las puntuaciones de diferencia de variabilidad total siempre será igual a 0:

$$\Sigma(y) = 0$$

El resumen del plateamiento de la ecuación estructural para un diseño entre-sujetos unifactorial univariado se representa en la figura 24.

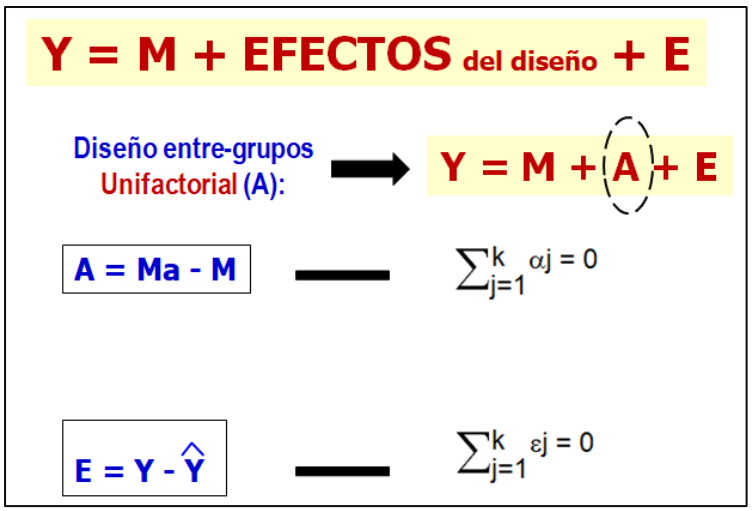


Figura 24. Resumen del planteamiento de la ecuación estructural. Diseño entresujetos

Suma de Cuadrados

Para valorar la parte de variación atribuible al efecto del tratamiento A respecto a la variabilidad aleatoria de los datos (E) hay que considerar todas las observaciones (puntuaciones en la variable dependiente Y) que se han registrado en la investigación. El problema estriba en que como el modelo se expresa en puntuaciones de diferencia, si se suman todos los valores (de todos los sujetos) de algún componente de la ecuación, el resultado siempre será cero tal y como ya se ha señalado anteriormente (pues son desviaciones respecto a la media). Esta medida permite que el usuario compruebe si está realizando bien los cálculos, pues si la suma de los efectos estimados no da 0 o la suma de los errores de los sujetos de la muestra no da 0 o la suma de las puntuaciones de diferencia de la variabilidad total no da 0 entonces es que se ha realizado la estimación de dichos efectos de manera incorrecta.

Por lo tanto, si se suman los efectos del tratamiento ($\hat{\alpha}$) (varianza entre-grupos) de las *N* observaciones de la muestra de un estudio el resultado de la suma siempre será cero $(\sum \hat{\alpha} = 0)$. Del mismo modo ocurre con los valores del componente de error (varianza intra-grupos) que suman cero al sumar todos los errores y las puntuaciones de diferencia *y* de la variabilidad total (varianza total).

Además, respecto al término de error, tambien, se puede comprobar que suman cero al sumar los valores de error de los sujetos que están dentro de la misma condición de A, por ejemplo a_1. Y, también suman cero las puntuaciones de diferencia (*y*) que representan cada puntuación directa en Y menos la constante o media general. Debido a ello, el procedimiento para sumar las puntuaciones de diferencia de cada fuente de variación se realiza elevando al cuadrado estas puntuaciones y posteriormente sumándolas $(\sum(\hat{\alpha})^2)$, $(\sum(\hat{\alpha})^2)$ y $(\sum(\hat{\alpha})^2)$. Con ello se evita que los datos sumen cero y se puede proporcionar un valor de suma de cuadrados.

A la puntuación que resulta de sumar el cuadrado de cada una de las puntuaciones de diferencia de una fuente de variación se le denomina “Suma de Cuadrados” de dicha fuente de variación (Suma de Cuadrados del efecto o variable independiente A, Suma de Cuadrado del error E, Suma de Cuadrados Total T).

En definitiva, la suma de cuadrados es una transformación cuadrática de las puntuaciones de diferencia que mantiene la proporción respecto de las diferencias originales, eliminando el signo de la puntuación y con ello se evita que siempre sume cero. De tal forma que si se extrae la raíz cuadrada del cuadrado, se obtiene la puntuación de diferencia en valores absolutos. El cuadrado permanece invariante ante el signo y aumenta en proporción cuadrática a la distancia que separa la media de la observación. El ANOVA trabaja con las sumas de cuadrados de cada fuente de variación.

Grados de Libertad y Medias Cuadráticas

Como cada una de estas sumas de cuadrados (de los efectos, del error y total) se estima con un número distinto de observaciones independientes, se debe corregir estas diferencias dividiendo cada suma de cuadrados por sus correspondientes “grados de libertad”.

Los grados de libertad indican el número de observaciones cuyos valores son libres de variar, o en otras palabras, el número de observaciones independientes de una fuente de variabilidad menos el número de parámetros estimados al computar dicha variación. Por ejemplo, los grados de libertad de la suma de cuadrados total son $N - 1$ donde N es el número total de observaciones, del factor A son a – 1 donde a es el número de grupos o condiciones, los grados de libertad del error son N – a o si el diseño es ortogonal también se pueden calcular como a (n – 1) donde n es el número de observaciones por grupo o condición.

Al resultado de dividir la suma de cuadrados de una fuente de variación por sus grados de libertad se le denomina “Media Cuadrática” de esa fuente de variación.

Razón *F*

El estadístico *F* mide el número de veces que es mayor la suma de cuadrados del efecto de un determinado tratamiento (efecto de A dividido por sus grados de libertad, MC_A), que la fuente de error del model (error dividido por sus correspondientes grados de libertad, MC_E). La prueba de la hipótesis del estadístico *F* compara varianzas; de ahí que el nombre que reciba esta prueba, tal y como se ha comentado, sea precisamente el de Análisis de la Varianza (ANOVA).

El valor del estadístico de la razón *F* del ANOVA es la razón entre la variación de los datos observada entre los distintos niveles o condiciones de la variable independiente (efecto del factor A) respecto a la varianza que se atribuye al error (efecto de E).

$$F = \frac{\textbf{Varianza del efecto}}{\textbf{Varianza del error}}$$

Por lo tanto, la prueba *F* es la razón entre las medias cuadráticas del efecto A y el error E y siempre que se escriba un valor de *F* hay que informar de los grados de libertad de las dos fuentes de varianza que se han utilizado para obtener el valor de ese estadístico: fuente entre o del efecto y de la fuente intra o del error. Se trata de la razón $F_{empírica}$ del estudio:

$$F(gl\ efecto, gl\ error) = \frac{\frac{SC_A}{gl_A}}{\frac{SC_{ERROR}}{gl_{ERROR}}} = \frac{MC_A}{MC_{ERROR}}$$

SC es la Suma de Cuadrado y MC es la media cuadrática; gl son los grados de libertad de la fuente de varianza.

En un diseño entre-sujetos unifactorial, los grados de libertad del efecto del tratamiento A (varianza 'entre-grupos') se obtiene como número de condiciones que tenga la variable independiente menos 1 ($gl_A = a - 1$).

En un diseño entre-sujetos unifactorial, los grados de libertad del error (varianza 'intra-celdilla' o del error) se obtiene como número total de observaciones menos el número de condiciones que tenga el diseño ($gl_{error} = N - a$). Si el diseño es ortogonal (cada grupo consta del mismo número de observaciones, $n_1 = n_2$) entonces los grados de libertad del error también se pueden calcular como a(n -1).

Decisión estadística

Si los resultados obtenidos con la investigación confirman o no las hipótesis científicas o teóricas se comprueba calculando la probabilidad de obtener el tamaño del efecto detectado con la muestra del estudio en una distribución muestral de una hipótesis estadística que plantea que el efecto es cero (modelo de la hipótesis nula).

Como ya se ha comentado, conviene tener muy presente que con el contraste de hipótesis estadísticas no se comprueba la probabilidad de que la hipótesis nula sea cierta o falsa porque se parte como supuesto o condición que la hipótesis nula es cierta y se conoce su distribución muestral donde se asociada cada valor con un valor de probabilidad bajo dicho supuesto.

Dado que la hipótesis nula generalmente plantea que el efecto es cero en la población entonces todos los valores de su distribución tendrán la misma probabilidad, siendo la distribución muestral uniforme (ver figura 25).

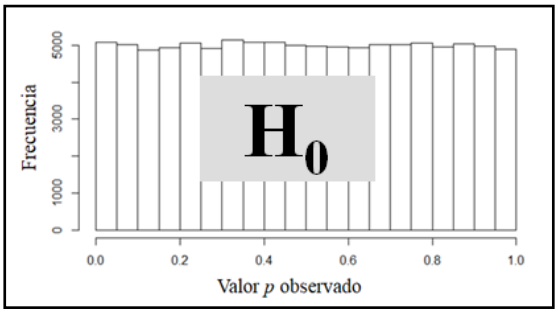


Figura 25. Distribución de los valores *p* de las frecuencias en el modelo de la hipótesis nula

Teniendo en cuenta esa cuestión, el contraste estadístico calcula el valor *p* del resultado del estadístico (se conoce como $F_{empírica}$) aplicado obtenido con los datos de la muestra (o un resultado más extremo), asumiendo que la hipótesis nula es cierta. Y, para tomar la decisión estadística (mantener la hipótesis nula o rechazar la hipótesis nula) compara dicho valor *p* del estadístico con el que tendría en la distribución de la hipótesis nula ($F_{teórica}$), dados los grados de libertad 'entre' o del efecto y los grados de libertad 'intra' o del error y el valor de alfa seleccionado a priori por el investigador o investigadora.

Por lo tanto, para llevar a cabo esa comparación se compara el valor de la $F_{empírica}$ con el valor de la $F_{teórica}$ que se puede encontrar en la tabla de la distribución muestral o teórica del estadístico aplicado que se encuentra en los manuales de estadística y diseño, existiendo diferentes tablas en función del valor del alfa (.05, .01. .025…) donde se consulta según los grados de libertad asociados a la prueba estadística aplicada, necesitando conocer tanto los grados de libertad entre-grupos o del efecto como los grados de libertad del error o intra-grupos (ver figura 26). Por ejemplo, para un alfa de .05, 6 grados de libertad de la fuente de varianza entre-grupos A y 6 grados de libertad de la fuente de varianza del error, el valor de la *F* de tablas, tabular o teórica es de 4.284.

TABLAS Razón *F*

Tabla III (continuación). $F\ (\alpha = 0.050,\ gl_{entre} = columnas,\ gl_{error} = filas)$

gl	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*24*
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882	243.906	249.052
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413	19.454
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.639
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.774
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.527
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.841

Figura 26. Tabla de la razón *F* para un alfa = .05, grados de libertad del efecto (entre) en las columnas y los grados de libertad del error en las filas

En resumen, cuando se consultan las tablas teóricas de la distribución muestral del estadístico (construidas bajo el supuesto de la hipótesis nula) se necesita conocer:

1) el valor de alfa (α) que ha establecido a priori el investigador o investigadora (antes de recoger los datos) como riesgo asumido de error de Tipo I en la fase de planificación del estudio y

2) los grados de libertad (*gl*) que se utilizan en el estadístico o prueba estadística seleccionada por el investigador o investigadora para dar una solución empírica a su hipótesis científica.

Para tomar de forma manual la decisión estadística hay que comparar los valores de la $F_{\text{empírica}}$ y la $F_{\text{teórica}}$. Cuando el valor de la $F_{\text{empírica}} \geq F_{\text{teórica}}$ entonces se rechaza la hipótesis nula, ya que el valor de *p* vinculado al estadístico de la $F_{\text{empírica}}$ será menor que el valor de alfa que se haya utilizado al consultar las tablas de los manuales (generalmente alfa es .05 y, por lo tanto, cuando $p \leq .05$ se considera que el efecto o la relación detectada entre las variables es estadísticamente significativa).

Como ya se ha comentado, en el caso de la razón *F* el estadístico compara dos varianza que tienen diferentes grados de libertad y, por lo tanto, se necesitará conocer los grados de libertad de la fuente de varianza del efecto (fuente de varianza entre-grupos) y los grados de libertad del error o fuente de varianza intra-grupo) para obtener el valor de la $F_{\text{teórica}}$. En el caso de la prueba *t* de Student (solamente se puede utilizar si el diseño tiene solamente 2 grupos, A = 2) para grupos independientes o diseños entre-grupos, únicamente se necesitará conocer los grados del término de error, pues los del tratamiento o efecto siempre serán 1 porque el diseño que utiliza la prueba *t* siempre tendrá dos condiciones o grupos (los grados de libertad del tratamiento o del efecto entre-grupos se obtienen como número de condiciones menos 1; $a - 1$, luego $2 - 1 = 1$). Los grados de libertad del término de error o intra-celdilla se obtienen como $N - a$. Y los grados de libertad totales son $N - 1$.

Errores estadísticos y decisiones correctas

Alfa (error de Tipo I)

El contraste estadístico siempre se realiza asumiendo un determinado riesgo de equivocación al rechazar una hipótesis nula que realmente es cierta. Se trata del "nivel de significación", valor de alfa", α, o "error de Tipo I": probabilidad de rechazar la hipótesis nula siendo realmente cierta. El nivel de significación está establecido por consenso en el 5% (α = .05) si no se indica otro valor en el informe del estudio. Si el investigador o investigadora desea trabajar con un alfa menor y así lo planifica a priori (por ejemplo, (α = .01) será necesario que lo especifique claramente en su informe de investigación. A su probabilidad complementaria se le denomina nivel de confianza (1 – α).

Nivel de confianza (1 – α)

A la probabilidad complementaria del error de Tipo I se le denomina "nivel de confianza" (1 – α) y determina cuál será la probabilidad de que el investigador o investigadora acierte en su decisión cuando la hipótesis nula sea realmente cierta, es decir, cuando se mantiene la hipótesis nula y realmente no existe relación entre las variables independientes y las dependientes especificadas en la hipótesis de investigación. Se trata de una decisión correcta.

Beta (error de Tipo II)

Cuando se rechaza la hipótesis nula, se podría estar cometiendo un error de Tipo II (error beta, β). Se trata de la probabilidad de mantener la hipótesis nula siendo realmente falsa. A su probabilidad complementaria se le denomina potencia estadística (1 – β).

Potencia estadística (1 – beta)

La potencia estadística es la probabilidad de rechazar la hipótesis nula siendo realmente falsa. Es decir, la probabilidad de detectar un efecto si realmente existe en la población. Generalmente se establece que el error de Tipo II debe ser de .2 como mucho, es decir, se considera como aceptable más error de Tipo II ya que sería cuatro veces mayor que el error de Tipo I. En otras palabras, se recomienda que la

potencia estadística se planifique para que sea al menos de .8. Se trata de una decisión correcta.

Por lo tanto, el contraste estadístico se inicia partiendo del supuesto de que no existe relación entre la variables independiente y la dependiente (modelo de la hipótesis nula). El modelo de la hipótesis nula atribuye las diferencias (ya sean las diferencias grandes o sean pequeñas) que pudiesen aparecer entre las puntuaciones de los diferentes grupos al efecto de extracción aleatoria de las muestras (muestreo aleatorio) y a las diferencias individuales no sistemáticas. Es decir, cualquier diferencia, grande o pequeña, detectada entre las condiciones es aleatoria (fruto del azar) porque en la población el efecto es cero.

Cuando se ejecuta la prueba del contraste de hipótesis estadísticas (por ejemplo con la razón *F*, con la *t* de Student, con *r* del coeficiente de correlación…), si se detecta que algún efecto de los postulados en el modelo estadístico no es un efecto nulo o de efecto cero entonces la hipótesis nula se puede rechazar, ya que al menos una de las condiciones experimentales (o alguna posible combinación entre varias de ellas o efecto de interacción) no procede de la misma población.

Cuando se rechaza la hipótesis nula se puede aceptar la hipótesis alternativa que está vinculada con el efecto postulado en la hipótesis científica o teórica si se tiene garantías de la validez de la evidencia aportada por los resultados. Ante el rechazo de la hipótesis nula hay un abanico de posibles hipótesis alternativas, pero si el diseño se planifica de forma adecuada, controlando las amenazas a las valideces, entonces cabe suponer que el efecto obtenido está provocado por las causas propuestas en el modelo teórico del estudio. Si el diseño de investigación es deficiente, evidentemente a pesar de rechazar la hipótesis nula poco se podrá decir de la causa de los efectos hallados o de la relación encontrada entre las variables.

En resumen, una vez se determina el tamaño del efecto experimental con los datos del estudio, se calcula su probabilidad bajo el supuesto de la hipótesis nula. Se puede concluir que el resultado es estadísticamente significativo (valor *p* de probabilidad obtenido con la prueba estadística ≤ valor de alfa fijado a priori por el investigador o investigadora) y rechazar la hipótesis nula. Por el contrario, puede ocurrir que se mantenga la hipótesis nula si el valor de probabilidad asociado con el tamaño del efecto detectado supera el punto de corte de α preestablecido ($p > \alpha$).

Para llevar a cabo dicha decisión estadística se puede emplear el estadístico razón *F* del Análisis de la Varianza cuyo proceso se ha detallado anteriormente.

Hipótesis de causa-efecto (causalidad)

Conviene tener muy claro que solamente el tipo de metodología que se haya empleado en el estudio podrá fundamentar que las hipótesis planteen hipótesis de causalidad donde la variable independiente causa un efecto sobre la variable dependiente. Este tipo de hipótesis de causalidad solamente pueden plantearse cuando se trata de un diseño de investigación que se ha planificado con una metodología experimental (manipulación de la variable independiente y asignación aleatoria del tratamiento). En términos estrictos, sólo cuando se aplica la metodología experimental se puede hablar de causalidad o de efecto causal de la variable independiente sobre la dependiente.

En otros contextos de investigación donde se emplea la metodología cuasi-experimental (manipulación de la variable independiente sin asignación aleatoria del tratamiento) será necesario poner en marcha un control exhaustivo de posibles diferencias previas entre los sujetos que podrían contaminar los resultados por terceras variables que no estaban presentes en la hipótesis del estudio y se podría hablar de cuasi-causal. Con una metodología cuasi-experimental se aspira a conclusiones de causalidad, pero en términos estrictos no podría realizarse. Quizás con diseños complejos que planifiquen y traten de controlar un conjunto de fuentes de varianza sistemática secundaria se podría hablar de “posibles relaciones de causa-efecto”. Por ejemplo, los diseños con puntuaciones de propensión están dirigidos a controlar un amplio número de variables extrañas y trata de aportar conclusiones de causalidad.

Y, por supuesto, si el estudio se lleva a cabo con una metodología no experimental (ni manipulación de la variable independiente, ni asignación aleatoria del tratamiento) nunca se deben plantear las hipótesis como causales ni interpretar los hallazgos como que la variable independiente causa un efecto sobre la variable medida, pues solo se interpretará en términos de covariación entre las variables y magnitud y dirección de la relación entre las variables. La correlación entre las puntuaciones de las variables no implica causalidad.

En definitiva, el tipo de metodología empleada en el estudio determinará el alcance de las inferencias de los resultados del análisis de la varianza que será de causalidad o de asociación entre las variables explicativas (independiente y dependiente) en función de la metodología de investigación empleada en el estudio. La técnica estadística que se aplica no determina la causalidad o las relaciones de causa-efecto entre las variables.

Supuestos estadísticos del modelo paramétrico

La prueba estadística aplicada es completamente válida en la medida que se cumplan los supuestos fundamentales de las pruebas paramétricas: que la variable dependiente se distribuya normalmente, que las varianzas de los grupos sean homogéneas (se puede aplicar la prueba de Levene para comprobar este supuesto donde se requiere que el resultado del contraste estadístico sea $p > .05$) y que las distintas observaciones sean independientes y extraídas aleatoriamente, no existiendo entre las condiciones experimentales ninguna otra diferencia sistemática que la atribuida al efecto de la variable independiente.

Los supuestos de las pruebas paramétricas señalan que debe existir:

-*Normalidad*: los errores (ε_{ij}) sigue una distribución normal. Esto es equivalente a que Y_{ij} sigue una distribución normal.

-*Linealidad*: Esto es equivalente a que $Y_{ij} = \mu + \alpha_i + \beta_j$. Es decir, la relación entre las variables debe ser lineal.

-*Homocedasticidad de las varianzas*: $Var(\varepsilon_{ij}) = \sigma^2$. Esto es equivalente a que $Var(Y_{ij}) = \sigma^2$. Cuando se comparan las puntuaciones de dos o más grupos es necesario que la varianza de dichos grupos sea homogénea. Este supuesto se puede comprobar con la prueba de Levene

-*Independencia*: ε_{ij} son independientes entre sí. Esto es equivalente a que Y_{ij} son independientes entre sí. Es decir, las respuestas dadas a una variable no deben depender de ninguna otra situación.

En general, los estudios de simulación han comprobado que la razón *F* es robusta al incumplimiento de dichos supuestos si la muestra es lo suficientemente grande,

siendo un problema si la muestra es pequeña y si los tamaños de los grupos están muy descompensados (diseños no balanceados).

Autoevaluación: contraste estadístico

A continuación se presenta un ejercicio para repasar los conceptos fundamentales que forman parte del contraste estadístico. Escribe tu respuesta.

Ejercicio 1. Anota la respuesta a cada tipo de probabilidad:

1. La probabilidad de rechazar la H0 cuando es falsa se denomina _____
2. La probabilidad de rechazar la H0 cuando es cierta se denomina: _____
3. La probabilidad de los datos observados en el estudio, siendo la hipótesis nula cierta se denomina: _____
4. La probabilidad de mantener la H0 cuando es cierta se denomina: _____
5. La probabilidad de mantener la H0 cuando es falsa se denomina: _____
6. La probabilidad del resultado de la prueba estadística (o de un resultado más extremo), bajo el supuesto de una distribución que plantea que no hay efecto se denomina: _____
7. La probabilidad de no rechazar la H0, siendo la H1 cierta, se denomina: _____
8. La probabilidad de rechazar una H0 verdadera és: _____

Ejercicio 2. Qué es el valor *p* de probabilidad. Señala Verdadero o Falso:

1. La probabilidad de que la hipótesis nula sea verdadera, dados los datos de la investigación: _____
2. La probabilidad de que la hipótesis alternativa sea verdadera, dados los datos de la investigación: _____

3. La probabilidad de error asociada a la certeza de la hipótesis nula: _____
4. La probabilidad del resultado obtenido en el estudio si la hipótesis nula es falsa: _____
5. La probabilidad del resultado obtenido en el estudio si la hipótesis nula es cierta: _____
6. La probabilidad de la hipótesis nula cuando se compara con el valor de alfa fijado a priori: _____

Capítulo 8. Supuesto: “indefensión aprendida y depresión en ratas”. Diseño entre-sujetos unifactorial (A = 2) univariado

Marcos Pascual-Soler*

Dolores Frías-Navarro**

**ESIC Business & Marketing School, España

*Universidad de Valencia

Índice

Citar el capítulo como:

Pascual-Soler, M. y Frías-Navarro, D. (2021). Supuesto: “indefensión aprendida y depresión en ratas”. Diseño entre-sujetos unifactorial (A = 2) univariado. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

SUPUESTO DE INVESTIGACIÓN. Supongamos que una investigadora desea comprobar si la indefensión aprendida produce déficits depresivos. Diseña una situación experimental donde los sujetos son ocho ratas (N = 8) que deben completar un laberinto de cuyo suelo reciben descargas eléctricas de baja intensidad ininterrumpidamente. La mitad se asigna aleatoriamente a la condición de shock escapable (a_1) y la otra mitad a la condición de shock inescapable (a_2). La tarea experimental consiste en recorrer el laberinto, midiéndose el tiempo (en segundos, Y) que emplean las ratas en su recorrido. La hipótesis experimental mantiene que las ratas del grupo de shock inescapable fracasarán en su aprendizaje, recorriendo el laberinto con lentitud y sin precisión, e incluso en muchas ocasiones no llegaran a completarlo, manifestando un comportamiento depresivo. Sin embargo, las ratas que se encuentran en la condición experimental de shock escapable aprenderán que con su ejecución escapan de la descarga y aumentarán rápidamente la velocidad de carrera con objeto de eliminar la situación aversiva y llegar al habitáculo que les privará de las descargas, a diferencia de la condición de shock inescapable donde perdurará la descarga aunque lleguen a dicho habitáculo. Tras una serie de diez ensayos previos que facilitaron el aprendizaje de la situación o recorrido del laberinto de todas las ratas, los resultados del experimento fueron los que se detallan en la Tabla 7 (ver Anexo 6 con el planteamiento del ejercicio, su solución manual y su solución con los programas SPSS, JASP y JAMOVI).

Tabla 7. Matriz de resultados. Diseño entre-sujetos A = 2

A $\rightarrow$ Shock	**Y** $\rightarrow$ Tiempo
Condición 1: a_1 Escapable	23, 11, 12, 26
Condición 2: a_2 No escapable	39, 38, 23, 28

A continuación se detalla un conjunto de preguntas dirigidas a que los lectores y lectoras autovaloren su grado de comprensión de todos los conceptos expuestos hasta el momento en el libro. Se recomienda realizar dicho ejercicio antes de continuar con el desarrollo del libro. La pregunta 27 y 28 se explicarán posteriormente cuando se presente el concepto de tamaño del efecto.

Autoevaluación del planteamiento del ejercicio y análisis

Contestar a las siguientes preguntas de autoevaluación:

1) Objetivo del estudio

2)Hipótesis del estudio (hipótesis experimental o científica)

2.1. Como constructo

2.2. Operacionalizada en el experimento

3) Variable independiente o factor:

3.1. Como constructo

3.2. Operacionalizada en el experimento

4) Variable dependiente o resultados:

3.1. Como constructo

3.2. Operacionalizada en el experimento

5) Posibles variables extrañas:

4.1. Como constructo y su control

4.2. Operacionalizada en el estudio y su control

6) Plantea otras posibles variables extrañas y su control

7) Metodología del estudio, por qué. Utiliza el contenido del ejercicio para realizar la explicación

8) Tipo de **diseño** del estudio, por qué

9) Plantear las **hipótesis estadísticas**

10) Plantear la **ecuación estructural** del modelo **nulo** o restringido

11) Puntuación pronostica por el modelo **nulo**

12) Plantear la **ecuación estructural** del modelo **alternativo** o completo

13) Puntuación pronostica por el modelo **alternativo**

14) Qué valor tiene ***S***, por qué

15) Qué valor tiene ***N***, por qué

16) Qué valor tiene ***n***, por qué

17) Qué valor tiene la representación del factor **A** y sus condiciones, por qué

18) Qué valores tiene **Y**, por qué

19) Qué valores tiene **M**, por qué

20) Qué valor tienen las **2 medias** de las condiciones de la variable independiente

21) Qué valor tiene la **puntuación de diferencia** de las dos medias

22) Qué valores tienen los efectos de **A** en el modelo

23) Qué valores tiene el error (**E**) del modelo

24) Resolver el **contraste de hipótesis** planteado en el estudio hasta obtener el valor de la Razón *F* con sus grados de libertad

25) Tomar la **decisión estadística**: mantener la hipótesis nula / rechazar la hipótesis nula

26) Calcular el **tamaño del efecto de eta cuadrado**

27) Calcular el **tamaño del efecto de *d* de Cohen y su intervalo de confianza**. Consultar el programa de la Colaboración Campbell

28) Comparar la **relación** directa entre el valor de **eta cuadrado** (η^2) y el valor de la ***d* de Cohen** (tarea de conversión entre los estadísticos)

29) Redacción de resultados utilizando el formato del Manual APA (7ª edición)

Explicación y desarrollo del ejercicio

Las características de la investigación planteada en el ejemplo anterior permiten concluir que se trata de un diseño de investigación con una metodología experimental con dos condiciones de la variable independiente y una variable dependiente medida. La manipulación de la variable independiente (A) y la asignación aleatoria de las unidades de observación (ratas) a una de las condiciones de intervención indica que se trata de un experimento verdadero o realizado con una metodología experimental.

De acuerdo con la hipótesis que se quiere comprobar se formula la ecuación estructural, conocida también como modelo estructural paramétrico, donde se realiza una transformación lineal de los datos separando los dos componentes del modelo (efecto y error) cuya relación lineal determina que los efectos en la variable dependiente Y sean aditivos:

1. el efecto de la variable independiente A (componente fijo o de efecto fijo para cada sujeto de la condición) y
2. el término de error (componente aleatorio vinculado a la puntuación de cada sujeto).

Una vez especificado el modelo del diseño, se pondrá a prueba una serie de hipótesis de nulidad relacionadas con los diferentes efectos atribuidos a las variables independientes de la ecuación estructural; en el ejemplo únicamente hay una variable independiente A con dos condiciones o niveles de tratamiento. Pero podría suceder que el modelo tenga más de una variable independiente, por ejemplo A y B, y entonces se llevarían a cabo contrastes de hipótesis para cada una de las tres fuentes de varianza: el factor A, el factor B y su interacción AB.

Ecuación estructural y efecto de A

La estimación del efecto del shock (variable independiente A) vendrá determinada por la diferencia en tiempo que se observe entre las dos situaciones o condiciones manipuladas (a_1: shock escapable; a_2: shock inescapable). Si la única diferencia constante que ha existido entre las dos condiciones, a lo largo de todo el experimento, es la posibilidad o no de escape del shock, cabría esperar que las medias de los dos grupos fueran distintas y su diferencia estadísticamente significativa. Claro, en la medida que la hipótesis teórica de partida sea cierta y el efecto exista en la población.

El modelo se suele expresar en variaciones en torno a la media. La manera más simple e intuitiva de escribir el modelo estructural consiste en expresar cada puntuación individual de Y como la suma de la media general, más el efecto del grupo o condición en el que se encuentran el sujeto más el término de error específico de cada sujeto. Se trata de la denominada "ecuación estructural" que en términos poblacionales es igual a:

$$Y = \mu + \alpha + \varepsilon$$

Pudiéndose estimar los parámetros del efecto (α) como la media del grupo (μ_a) respecto a la media poblacional (μ), midiéndose, por lo tanto, en la escala de puntuaciones de diferencia:

$$\alpha = \mu_a - \mu$$

Como estos son los valores de la población de ratas y en el ejemplo se trabaja con una muestra de ratas, lo que se obtiene es una estimación de esos parámetros poblacionales desconocidos a partir de las medias muestrales. Es decir, la ecuación estructural se reformula en términos muestrales como:

$$Y = M + A + E$$

En términos muestrales, la estimación del efecto son desviaciones de las medias marginales respecto a la media general:

$$A = Ma - M$$

Donde se deduce que la media de la condición es igual a la constante más el efecto de dicha condición:

$$Ma = M + A$$

En la tabla 8 se han añadido las medias de cada grupo y sus efectos. Así se comprueba que:

$$Ma_1 = 25 + (-7) = 18$$

$$Ma_2 = 25 + 7 = 32$$

Tabla 8. Matriz de resultados. Diseño entre-sujetos A = 2

A → Shock	**Y** → Tiempo	M_a	A
Condición 1: a_1 Escapable	23, 11, 12, 26	18	-7
Condición 2: a_2 No escapable	39, 38, 23, 28	32	7
		M = 25	

Dado que la suma de las desviaciones es igual a cero, como ya se comentó, entonces en un diseño A = 2, el último parámetro de las condiciones del factor (a_2) se puede obtener como diferencia. Aplicando esta fórmula a los datos del supuesto se obtiene que $\alpha_1 = -7$, (18 - 25), por lo tanto como $\sum_{j=1}^{k} A_j = 0$ entonces necesariamente $\alpha_2 = 7$. De este modo, el número de parámetros que se necesita estimar en el modelo son tantos como el número de condiciones experimentales –*a*– menos uno (Riba, 1990). Se trata de los denominados grados de libertad (*gl*). En un diseño de A = 2 los grados de libertad son 1 porque:

$$gl_a = a - 1$$

$$gl_a = 2 - 1 = 1$$

El cambio producido en la escala cuando se trabaja con los efectos de las condiciones, respecto a sus medias, no varía la información, sino que únicamente expresa los datos tomando como punto de referencia la media. Por tanto, si conocemos la media de los datos y la puntuación de diferencia (los efectos), podemos conocer el tiempo que se ha invertido en el recorrido del laberinto. Hay dos

consecuencias fundamentales que se derivan de este primer cambio de escala para el análisis de los resultados:

1) Se calcula el valor de un parámetro, la mediat total $\bar{Y}$, (M), que indica el punto central de todas las puntuaciones.

2) La inclusión de un parámetro implica la pérdida de un grado de libertad en el conjunto de los datos. Esto quiere decir que los *N* valores obtenidos en el experimento se conocen a partir de *N* - 1 puntuaciones de diferencia y del valor del parámetro de media total, M (Box y cols., 1978, página 40). Se trata de los grados de libertad totales vinculados a la variabilidad o varianza total que presentan las puntuaciones de la muestra de datos.

Puntuación pronosticada $\hat{Y}$

Una vez se conocen los parámetros del modelo y se formula la ecuación estructural, se pueden expresar los datos sobre su base y de este modo comprobar la predicción del modelo. En el ejemplo, en cualquiera de las dos situaciones de shock, la puntuación que se pronostica para la velocidad del recorrido resultará de sumar a la constante, o media general M, el efecto correspondiente al nivel del factor donde se encuentra el sujeto. Así, la puntuación pronosticada de cada sujeto es:

$$\hat{Y} = M + A$$

Luego si,

$$\hat{Y} = M + (Ma - M)$$

Entonces en este diseño unifactorial se comprueba que la puntuación pronosticada $\hat{Y}$ es la media de la condición donde se encuentra el participante:

$$\hat{Y} = Ma$$

Por lo tanto, para conocer el valor de la puntuación pronosticada simplemente hay que descomponerla en la constante más los efectos que se plantean en la ecuación estructural.

Error

Y el error (E) cometido en cada observación i será la diferencia que resulte de restar a la puntuación directa o real (Y) del participantes el pronóstico o puntuación pronosticada ($\hat{Y}$) que se obtenga con la ecuación del modelo ($\hat{Y}$):

$$E = Y - \hat{Y}$$

Además, se comprueba que $\sum_{j=1}^{k} E_j = 0$

Por ejemplo, en el ejercicio de las ratas, en la observación cuarta -Y_4- se registraron 26 segundos en el recorrido del laberinto (puntuación directa Y). Como esta unidad experimental se encontraba en la condición de shock escapable (a_1), el pronóstico o la puntuación pronosticada ($\hat{Y}$) para ese sujeto (así como para todos los sujetos que se encuentren en su misma condición o grupo: a_1), será de 18 segundos:

$$\hat{Y}_{a1} = \overline{Y} + A_1 = 25 + (-7) = 18 \text{ segundos}$$

Como realmente esa rata (Y_4) ha recorrido el laberinto en 26 segundos, el error de estimación de esa rata concreta número 4 (cada sujeto o rata tiene su propio error, pues el error es individual) ha sido:

$$E_4 = Y_4 - \hat{Y}_{a1} = 26 - 18 = 8 \text{ segundos}$$

En resumen, la razón última del modelo o ecuación estructural es expresar la hipótesis experimental en términos mensurables para medir el efecto que ha tenido en la variable dependiente los cambios operados en la variable independiente.

Varianza total

Como los datos se expresan en puntuaciones de diferencia se obtiene que la Suma de Cuadrado Total ($\Sigma(Y - M)^2$) es igual a la Suma de Cuadrado del Efecto ($SC_A = \Sigma(\hat{Y} - M)^2$, recordar que en este diseño $\hat{Y} = Ma$, luego $SC_A = \Sigma(Ma - M)^2$) más la Suma de Cuadrados del Error ($\Sigma(Y - \hat{Y})^2$). Es decir:

Suma de Cuadrado Total = Suma de Cuadrado del Efecto + Suma de Cuadrados del error

$$SC_{TOTAL} = SC_{EFECTO\ A} + SC_{ERROR}$$

$$Y - M = (\hat{Y} - M) + (Y - \hat{Y})$$

Además, se comprueba que la suma de las puntuaciones de diferencia (y) ($\Sigma(Y - M)$ relacionada con la varianza total es igual a 0:

$$\sum_{j=1}^{k} y_j = 0$$

En resumen y respecto al supuesto de las ratas, la diferencia de 2 segundos por debajo de la media en el tiempo invertido en el recorrido del laberinto de la rata 4 se corresponde con un descenso de 7 segundos por la posibilidad de escapar del shock al final del recorrido y de 5 por el efecto de otras variables distintas de la manipulada (en este caso se atribuye al error).

En resumen, el modelo o ecuación estructural descompone la variación respecto de la media total en dos partes:

1. Una atribuida al tratamiento o no del shock y
2. Otra a las restantes variables que inciden en la velocidad del recorrido. Siendo estas últimas la fuente de variación de error por proceder de variables que no están contempladas en el modelo teórico.

Suma de Cuadrados

Si se suman los efectos del tratamiento (A) de las *N* observaciones de la muestra el resultado de la suma será cero, tal y como ya se ha explicado. Por ello, es necesario calcular las denominadas Sumas de Cuadrados de cada fuente de variación para obtener un valor de varianza o variabilidad en cada fuente de varianza.

El procedimiento para sumar las puntuaciones de diferencia de cada fuente de variación se realiza elevando al cuadrado estas puntuaciones y sumándolas.

A la puntuación que resulta de sumar el cuadrado de las puntuaciones de diferencia de una fuente de variación se denomina “Suma de Cuadrados” de dicha fuente de variación.

Suma de Cuadrados del Efecto A

Por tanto, la Suma de Cuadrados correspondiente a la manipulación del factor A o explicada por la acción del tratamiento, $\Sigma(Ma - M)^2$, (Suma de Cuadrados del efecto o del tratamiento) en el ejemplo que nos ocupa es igual a 392.

Suma de Cuadrados del Error

La suma de cuadrados correspondiente a las variables no incluidas en el modelo teórico se conoce como Suma de Cuadrados del Error, $(\Sigma(Y - \hat{Y})^2$. En el ejemplo la Suma de Cuadrados residual o del error (varianza intraceldilla) es igual a 356.

Suma de Cuadrados Total

La suma de las puntuaciones de diferencia de los datos respecto a la media general al cuadrado, $(\sum Y - M)^2$, se denomina Suma de Cuadrados Total (la varianza total está representada por el modelo de la hipótesis nula o modelo restringido). En el ejemplo la Suma de Cuadrados total es igual a 748.

Las sumas de cuadrados se descomponen, por lo tanto, en las puntuaciones de diferencia de cada observación. De tal forma que:

$$SC_{TOTAL} = SC_{EFECTO\ A} + SC_{ERROR}$$

La descomposición de la varianza total de las puntuaciones del ejemplo propuesto junto con los resultados de la tabla del ANOVA y el valor del tamaño del efecto de eta cuadrado se puede observar en la imagen siguiente de la pizarra.

$$SC_{TOTAL} = SC_{EFECTO\ DEL\ SHOCK} + SC_{ERROR} = 392 + 356 = 748$$

En la imagen 5 se representa la solución de las Sumas de Cuadrados del ejercicio de las ratas y se completa el Análisis de la Varianza a excepción del cáculo del valor *p* de probabilidad.

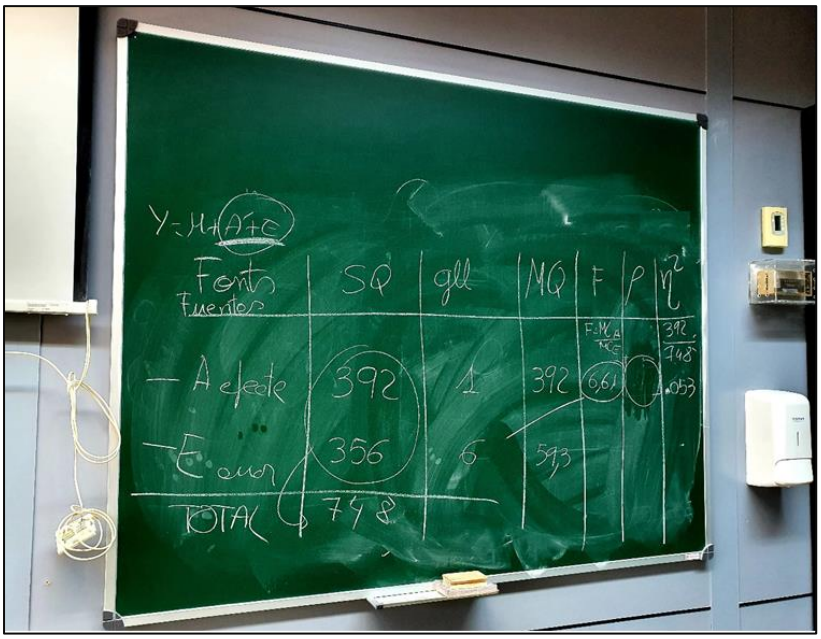


Imagen. Resultados del ANOVA del ejercicio de las ratas e indefensión

Plantilla de aprendizaje

Durante los primeros momentos del aprendizaje del desarrollo de la ecuación estructural de un modelo de diseño se pueden seguir las indicaciones de una plantilla de aprendizaje tal y como se ha señalado anteriormente. Para este ejercicio se puede consultar en el Anexo 2 la plantilla que se ha elaborado para el aprendizaje de un diseño entre-grupos unifactorial univariado.

A continuación se detalla en la figura 27 el inicio del ejercicio utilizando dicha plantilla.

En dicha plantilla también se detallan los conceptos fundamentales para desarrollar el ejercicio. Se recomienda que los lectores y lectoras completen el ejercicio y reflexionen sobre cada uno de los conceptos y pasos realizados para completar la tabla del Análisis de la Varianza, ANOVA.

a *(condición: a1, a2, a3 ...)*	***S*** *(sujeto: 1, 2, 3, 4, 5 ...)*	**Y** *(puntuaciones en la variable dependiente)*	**M** *(Media general o constante)*	***y*** **(Y – M)** E_{H0}	**A** **(M_a – M)**	$\hat{Y}$ **(M + efectes)** Puntuación pronosticada H1	**E** **(Y - $\hat{Y}$)** $E_{H1} = Y\text{-}M_a$
a1 escapable	1	23					
a1 escapable	2	11					
a1 escapable	3	12					
a1 escapable	4	26					
a2 No escapable	5	39					
a2 No escapable	6	38					
a2 No escapable	7	23					
a2 No escapable	8	28					
				SC (Suma de Cuadrados)			
				gl (grados de libertad)			
				MC (Media Cuadrática)			
				SCtotal	**SCentre**		**SCerror**

Figura 27. Inicio del desarrollo de la ecuación estructural y ANOVA

La tabla del Análisis de la varianza, ANOVA, del supuesto de las ratas que permite analizar la relación entre las variables de indefensión (variable independiente o factor del modelo) y síntomas de depresión (variable dependiente o variable medida) se detalla en la figura 28.

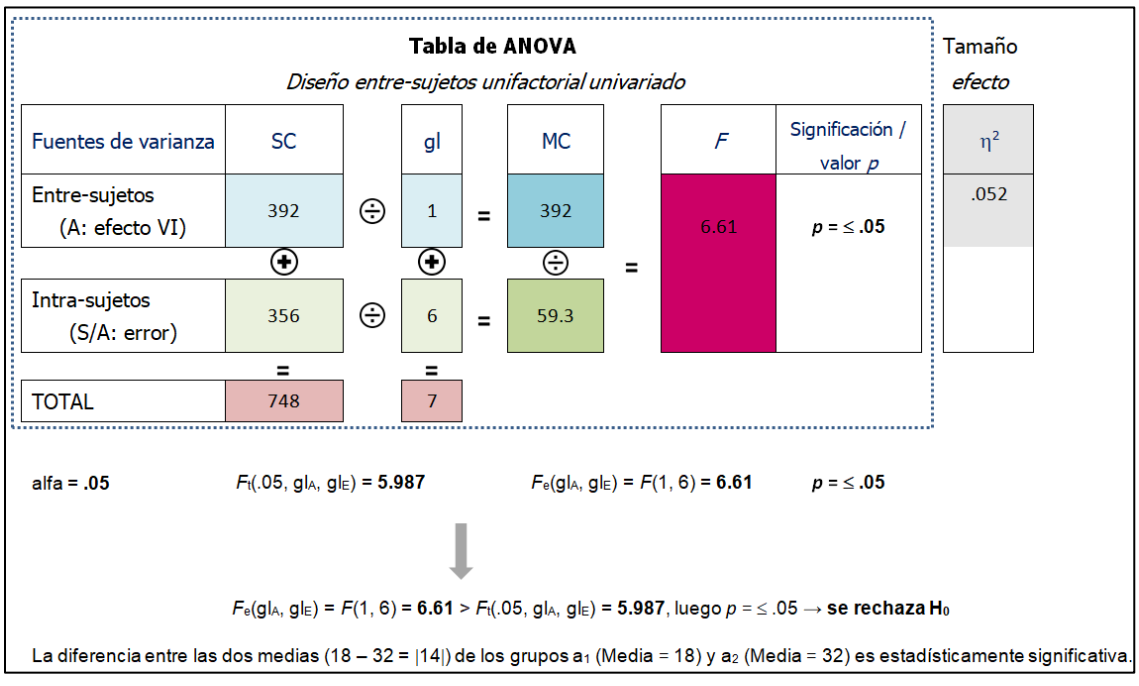


Figura 28. Tabla de ANOVA y tamaño del efecto

Contraste estadístico

Llegados a este punto, solo queda por determinar si el ajuste del modelo a los datos es lo suficientemente considerable como para concluir que es estadísticamente significativo (se trata de ejecutar la decisión estadística) junto con estimar el valor del tamaño del efecto y su intervalo de confianza. La interpretación de los hallazgos se realizará teniendo en cuenta el resultado del valor *p* del contraste estadístico y el valor del tamaño del efecto y la precisión de su intervalo de confianza.

En el ejemplo se pretende comprobar la hipótesis que mantiene que la situación de desamparo aprendido provoca déficits depresivos operacionalizados en conductas de frustración y desinterés (diseño de superioridad). Si los resultados obtenidos confirman o no esta hipótesis se comprueba calculando la probabilidad de obtener el tamaño del efecto detectado en el experimento cuando la hipótesis teórica que sustenta el modelo es falsa, es decir, no hay relación entre la variable independiente (shock) y la variable dependiente (tiempo) (planteamiento de la hipótesis nula). Para esto se necesita conocer el número de resultados que pueden producirse en un determinado experimento (el universo de resultados) y la probabilidad asociada con el ejemplo, para poder determinar si el resultado obtenido es muy improbable o extraño (es estadísticamente significativo) o, por el contrario, está dentro del intervalo de confianza (Arnau, 1981).

Pasos para llevar a cabo el contraste de hipótesis estadísticas

A continuación se detalla un resumen de todos los pasos implicados en el proceso de contraste estadístico.

1. Inicio del contraste: se asume como cierto el modelo de la hipótesis nula. El contraste de hipótesis se inicia partiendo del supuesto de que no existe relación entre las variables independientes y las dependientes, a este enunciado se le denomina modelo de la hipótesis nula. Dicha hipótesis explica que las diferencias que pudiesen aparecer en las puntuaciones estarían provocadas solamente por el efecto del azar atribuido al error de muestreo y/o a las diferencias individuales. Es decir, si el experimento se repitiera extrayendo las infinitas muestras de la variable Y en las

distintas condiciones del factor A y se calculará la media en cada una de las *a* condiciones de la variable, la esperanza matemática de cada $\bar{Y}_{a(i)}$ sería igual al valor de la media poblacional de Y (μ). Por lo tanto, el modelo de la hipótesis nula plantea que el efecto es igual a 0 y la puntuación pronosticada para cada sujeto es la constante o media general, M. El supuesto de la distribución muestral que se comprueba se denomina "prueba de significación de la hipótesis nula" (NHST).

2. Establecer a priori (antes de recoger los datos) el riesgo de equivocación como una probabilidad: probabilidad de detectar un efecto cuando realmente no existe (valor de alfa). En vez de recoger todas las posibles muestras del mismo experimento para determinar los parámetros de la población muestral con exactitud (circunstancia imposible en la realidad), se asume un determinado límite de error que se denomina alfa, α. El valor de α o error de Tipo I limita la probabilidad de equivocarse al rechazar la hipótesis nula (que se considera que realmente es cierta). Como la prueba de la hipótesis se realiza conociendo la distribución muestral del estadístico de la Razón *F* entonces se conoce exactamente la probabilidad acumulada en cada valor del mismo, siendo la probabilidad de error de Tipo I asociada con cada valor de este estadístico (valor *p*) la probabilidad complementaria de dicha probabilidad acumulada. De ahí que consultando dicha dicha tabla se pueda conocer la probabilidad asociada a un valor concreto del estadístico *F* dados los grados del libertad entre o del efecto y los grados de libertad del error y dado un valor de alfa prefijado como riesgo de equivocarse al rechazar la hipótesis nula. Se trata del nivel de significación.

3. El nivel de significación está establecido por consenso en la comunidad científica en el 5% si no se indica otro valor en la fase de planificación del estudio y, de ahí, que no sea necesario especificar dicho valor si el investigador o investigadora va a ejecutar sus contrastes de hipótesis con el máximo de error de Tipo I que se considera apropiado en las Ciencias Sociales y de la Salud. Así, en los artículos o informes donde no se menciona qué valor se ha prefijado para alfa se asume que fue de .05. La probabilidad complementaria al valor de alfa se le denomina nivel de confianza ($1 - \alpha$) y determina cuál será la probabilidad de que el investigador o investigadora acierte en su decisión (mantenga la hipótesis nula) cuando la hipótesis

nula sea cierta dado que no existe relación entre las variables independientes y las dependientes del modelo.

4. Obtener el valor *p* del estadístico calculado. Una vez se determina el tamaño del efecto experimental del estudio (se conoce el valor o resultado de la prueba estadística ejecutada) se calcula su probabilidad bajo el supuesto de la hipótesis nula. Se puede concluir que el resultado es estadísticamente significativo y rechazar la hipótesis nula ($p \leq$ alfa), o por el contrario mantenerla si la probabilidad asociada con el tamaño del efecto detectado supera el punto α preestablecido ($p >$ alfa). En el Análisis de la Varianza el estadístico que se emplea para realizar dicha comprobación es la razón *F* que es la razón entre la variación de los datos observada entre los distintos niveles de un tratamiento (Media Cuadrática del Efecto: Sumas de Cuadrados del efecto o del factor / sus grados de libertad) respecto de la suma de cuadrados del término del error (Media Cuadrática del Error: Sumas de Cuadrados del error / sus grados de libertad).

5. Como cada una de estas sumas de cuadrados se estiman con un número distinto de observaciones independientes, se corrige estas diferencias dividiendo cada suma de cuadrados por sus correspondientes grados de libertad (*gl*). Con ello se estiman las denominadas Medias Cuadráticas de cada fuente de varianza.

6. Los grados de libertad indican el número de observaciones cuyos valores son libres de variar, o en otras palabras, el número de observaciones independientes de una fuente de variabilidad menos el número de parámetros estimados al computar dicha variación.

A continuación, se continúa con el desarrollo del ejercicio de las ratas hasta completar la fase de la decisión estadística: mantener / rechazar la hipótesis nula.

Grados de libertad totales

Los grados de libertad de la suma de cuadrados total son $N - 1$. Es decir, la suma cuadrática de la variación total, $\sum(Y - M)^2$, implica *N* observaciones independientes y se pierde un grado de libertad cuando se estima el parámetro desconocido μ con la media total de las observaciones, M. Entonces como la suma de las desviaciones se realiza de cada puntuación respecto a la media general, $N - 1$ observaciones pueden tomar cualquier valor (tienen grados de libertad), ya que solamente una de

ellas está determinada (no tiene grados de libertad para variar) para poder obtener el valor concreto de la media total de los datos.

Grados de libertad del efecto A

Los grados de libertad de la suma de cuadrados de la fuente de tratamiento es igual $a - 1$, pues la puntuación predicha $\hat{Y}$ en $A = \sum(\hat{Y} - M)^2$, $(A = \sum(Ma - M)^2)$, contiene *a* parámetros a estimar, $\hat{a}_0$, $\hat{a}_1 \ldots \hat{a}_k$ y se sabe que el sumatorio de las desviaciones $\hat{Y} - M$ tiene que ser igual a cero, entonces únicamente $a - 1$ de los parámetros a estimar son libres de variar.

Grados de libertad del error

Los grados de libertad de la suma de cuadrados del error o residual son iguales a $N - a$, ya que la suma de cuadrados del error, $\sum(Y - \hat{Y})^2$, implica *N* observaciones independientes donde la puntuación predicha $\hat{Y}$ supone estimar *a* parámetros independientes dentro de cada condición experimental.

Razón *F*

Los resultados del supuesto 1 de las ratas relacionado con la indefensión aprendida y la sintomatología depresiva señalan que con 1 grado de libertad de la fuente del tratamiento o fuente entre-grupos (varianza entre grupos) A y 6 grados de libertad de la del error (varianza intra-grupos) se obtiene que la Razón *F* es igual a 6.607:

$$F(1, 6) = \frac{\frac{392}{1}}{\frac{356}{6}} = \frac{392}{59.334} = 6.607$$

Decisión estadística (mantener H_0 / rechazar H_0)

Concluyendo, se rechaza el modelo de la hipótesis de nulidad H_0, ya que el valor de $F(1, 6) = 6.607$ (o un valor más extremo) tiene una probabilidad menor del riesgo de error de Tipo I fijado a priori por el investigador o investigadora (en concreto es de $p = .042$).

En otras palabras, el valor empírico de la razón $F_{\text{empírica}}$ es mayor que el valor de la distribución teórica ($F_{\text{teórica}}$) que le corresponde con 1 y 6 grados de libertad para el término del efecto y del error respectivamente ($F_{\text{teórica}}(1, 6) = 5.987$). El valor de la denominada *F* teórica o de tablas $F_t(\alpha = .05, 1, 6)$ es de 5.987, y únicamente indica que el 95% de los valores que pueden obtenerse de *F* en la muestra del experimento, cuando la hipótesis nula sea cierta, serán menores o iguales que 5.987.

En el experimento el valor de la $F_{\text{empírica}}$ es mayor que la $F_{\text{teórica}}$ y, por lo tanto, ese valor, o valores más grandes, forman parte de ese 5% de valores de la distribución de la hipótesis nula que tienen una baja frecuencia. Y dado que tienen una baja frecuencia, el investigador o investigadora decide rechazar el modelo de la hipótesis nula para los resultados de su experimento y aceptar la hipótesis alternativa defendida en su estudio, por supuesto con un riesgo de error estadístico de equivocarse (error de Tipo I) fijado a priori en la fase de planificación del estudio o antes de recoger los datos.

Por último, faltaría la redacción de los resultados de la investigación de las ratas. Se va a realizar dicha redacción con los resultados del Análisis de la Varianza (ANOVA) y los resultados del tamaño del efecto. Concretamente como estadístico del tamaño del efecto se utilizará la eta cuadrado (η^2) o proporción de varianza explicada definida como:

$$R^2 = \eta^2 = \frac{SumadeCuadrados_{TRATAMIENTO\ EFECTO}}{SumadeCuadrados_{TOTAL}}$$

A continuación se ofrece una explicación más detallada del concepto del tamaño del efecto ya que el Manual del APA (7ª edición) y la comunidad científica consideran que cuando se redactan los resultados de una prueba de contraste estadístico siempre hay que añadir el tamaño del efecto y su intervalo de confianza, si es posible. Con ello, el investigador o investigadora debe facilitar una interpretación y redacción de la significación estadística, pero también de la magnitud y dirección de las diferencia halladas junto con la precisión en la estimación de dicha magnitud (intervalo de confianza).

¿Cómo se redactan los resultados de la inferencia estadística?

Cuando se redactan los resultados del contraste de hipótesis es fundamental dar toda la información estadística que rodea al proceso de contraste de hipótesis estadística como el tipo de diseño y las variables implicadas en el análisis que se ejecuta en la ecuación estructural, los descriptivos, al menos, de media y su desviación típica y número de observaciones (*n*) del grupo y el valor de la razón *F* con sus grados de libertad, junto con su valor *p* exacto y la información del tamaño del efecto (en este caso se ha interpretado la proporción de varianza explicada conocida como eta cuadrado, η^2, cuyo calculo es suma de cuadrados del efecto dividido por la suma de cuadrados total) y su intervalo de confianza (en la redacción del supuesto 1 no se ha incluido la información del intervalo de confianza de eta cuadrado).

En la siguiente dirección (https://www.gigacalculator.com/) se pueden llevar a cabo de forma rápida la estimación de los valores de media y desviación típica para que se incluyan en el apartado de redacción de los resultados:

https://www.gigacalculator.com/calculators/standard-deviation-calculator.php

Además, en la siguiente dirección también hay un amplio abanico de herramientas para realizar cáculos estadísticos:

https://www.gigacalculator.com/calculators/statistics/

En la Figura 29 se detallan unos resultados del artículo de Van Dessel y De Houwer (2019) titulado "*Hypnotic suggestions can iduce rapid change in implicit attitudes*". En este momento como lector o lectora de informes de investigación o artículos sería interesante que leyerá esos resultados. Este ejercicio consiste en observar solamente la información que se aporta en los resultados estadísticos. El objetivo es familiarizarse y comprender cómo se redactan los resultados en el apartado de Resultados de un informe de investigación y conocer algunos términos de uso muy frecuente en las Ciencias del Comportamiento así como otros términos que cada vez son más utilizados en la literatura científica.

Los términos metodológicos relacionados con el primer análisis estadístico del artículo hacen referencia a (ver Figura 29):

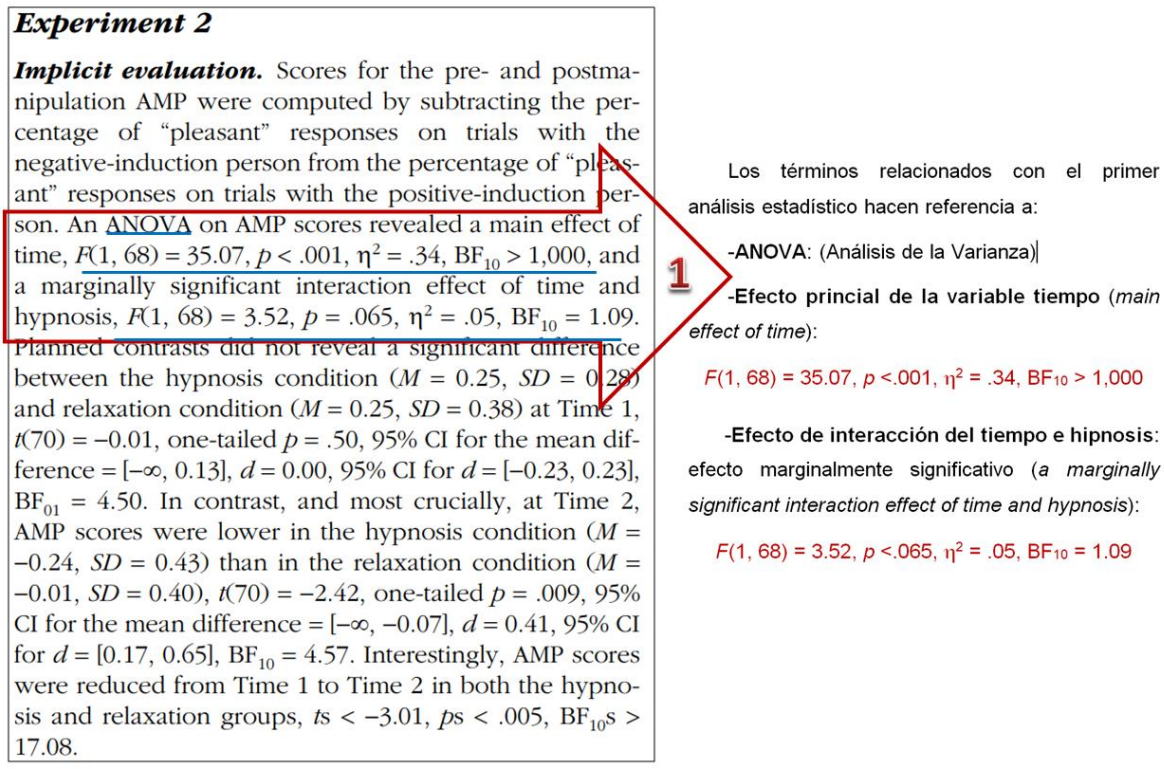
-ANOVA: Análisis de la Varianza

-Efecto princial de la variable tiempo (*main effect of time*):

$$F(1, 68) = 35.07,\ p < .001,\ \eta^2 = .34,\ BF_{10} > 1{,}000$$

-Efecto de interacción del tiempo e hipnosis: efecto marginalmente significativo (*a marginally significant interaction effect of time and hypnosis*):

$$F(1, 68) = 3.52,\ p < .065,\ \eta^2 = .05,\ BF_{10} = 1.09$$



Experiment 2

Implicit evaluation. Scores for the pre- and postmanipulation AMP were computed by subtracting the percentage of "pleasant" responses on trials with the negative-induction person from the percentage of "pleasant" responses on trials with the positive-induction person. An ANOVA on AMP scores revealed a main effect of time, $F(1, 68) = 35.07$, $p < .001$, $\eta^2 = .34$, $BF_{10} > 1{,}000$, and a marginally significant interaction effect of time and hypnosis, $F(1, 68) = 3.52$, $p = .065$, $\eta^2 = .05$, $BF_{10} = 1.09$. Planned contrasts did not reveal a significant difference between the hypnosis condition ($M = 0.25$, $SD = 0.28$) and relaxation condition ($M = 0.25$, $SD = 0.38$) at Time 1, $t(70) = -0.01$, one-tailed $p = .50$, 95% CI for the mean difference = $[-\infty, 0.13]$, $d = 0.00$, 95% CI for $d = [-0.23, 0.23]$, $BF_{01} = 4.50$. In contrast, and most crucially, at Time 2, AMP scores were lower in the hypnosis condition ($M = -0.24$, $SD = 0.43$) than in the relaxation condition ($M = -0.01$, $SD = 0.40$), $t(70) = -2.42$, one-tailed $p = .009$, 95% CI for the mean difference = $[-\infty, -0.07]$, $d = 0.41$, 95% CI for $d = [0.17, 0.65]$, $BF_{10} = 4.57$. Interestingly, AMP scores were reduced from Time 1 to Time 2 in both the hypnosis and relaxation groups, $ts < -3.01$, $ps < .005$, $BF_{10}s > 17.08$.

1

Los términos relacionados con el primer análisis estadístico hacen referencia a:

-**ANOVA**: (Análisis de la Varianza)

-**Efecto princial de la variable tiempo** (*main effect of time*):

$F(1, 68) = 35.07,\ p < .001,\ \eta^2 = .34,\ BF_{10} > 1{,}000$

-**Efecto de interacción del tiempo e hipnosis**: efecto marginalmente significativo (*a marginally significant interaction effect of time and hypnosis*):

$F(1, 68) = 3.52,\ p < .065,\ \eta^2 = .05,\ BF_{10} = 1.09$

Figura 29. Ejercicio de lectura activa de los resultados del informe

A continuación en la figura 30 se detalla la redacción de la evidencia aportada por la prueba estadística de la Razón *F* del Análisis de la varianza (ANOVA) dentro del apartado de Resultados del informe de investigación. Es conveniente fijarse en la redacción de las comas como elemento que separa cada uno de los datos de dicha expresión. Se utiliza la expresión inglesa del punto para separar los decimales y la coma para separar los miles. La redacción debe completar 5 elementos:

1) símbolo del estadístico (*F*)

2) valor de los grados de libertad del numerador de la razón *F* (vinculados al factor o variable independiente)

3) valor de los grados de libertad del denominador (vinculados al término de error)

4) símbolo del valor *p* de probabilidad (*p*)

5) y siempre ese valor *p* debe ir acompañado con su valor concreto. En el caso de que el valor p sea menor a .001 se escribirá < .001. Cuando esta última situación ocurre, los programas estadísticos suelen informar de que *p* = .000, pero siempre en el informe se escribirá *p* < .001 y nunca *p* = .000 porque *p* no es cero sino un valor más pequeño que .001.

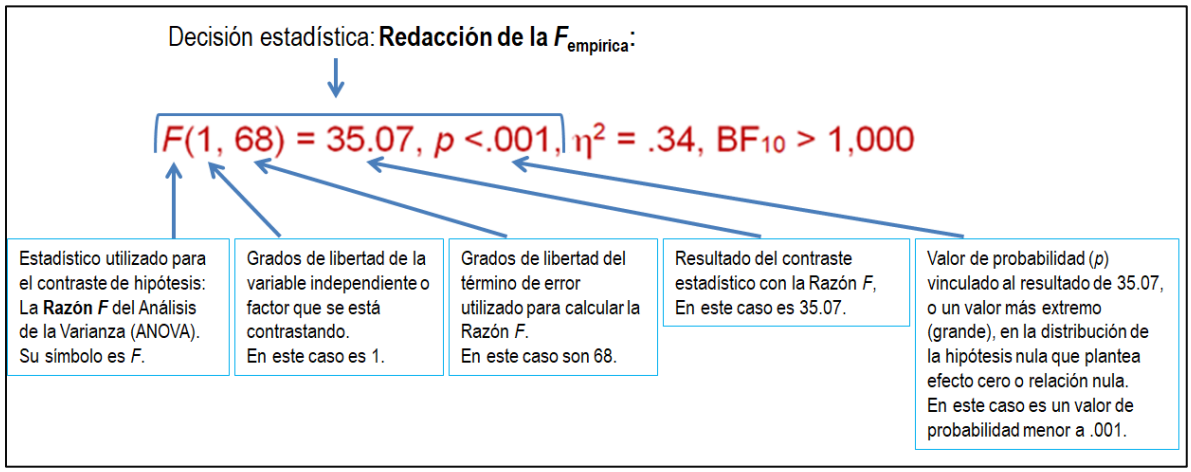


Figura 30. Redacción de la evidencia aportada por la prueba estadística de la Razón *F* de un ANOVA

La redacción del resultado de la inferencia estadística vinculada a la Razón *F* debe completarse con la redacción de la información del tamaño del efecto y su intervalo de confianza, siempre que sea posible. Así, a continuación del valor *p*, y separado por una coma, se escribe el símbolo del estadístico (en el ejemplo es eta cuadrado o proporción de varianza explicada, η^2) y su resultado (ver Figura 31). Si se dispone del intervalo de confianza entonces se seguiría la redacción de la estimación puntual después de una coma. Por ejemplo, 95% IC para η^2 = .04, .11 y se redactaría así en el informe: η^2 = .34, 95% IC .04, .11.

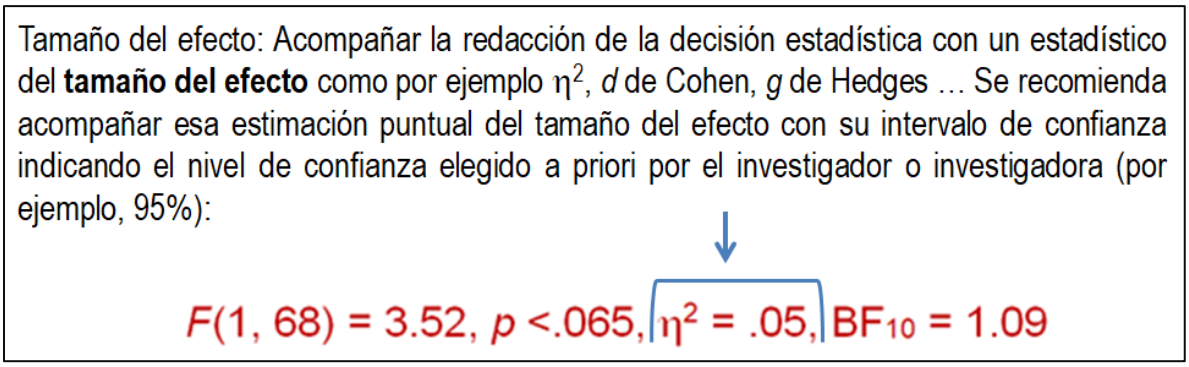


Figura 31. Redacción de la evidencia aportada por el estadístico del tamaño del efecto η^2

Cada vez más la redacción de los resultados se completa con el denominado Factor Bayes (FB). En el siguiente apartado se ofrece una breve explicación de este tipo de análisis que está basado en la estadística bayesiana. Como allí se comenta, se trata de una interpretación totalmente diferente al valor *p* de la perspectiva frecuencial que tradicionalmente se aplica en las pruebas de inferencia estadística basadas en la comprobación de la hipótesis nula (NHST) donde nunca se plantea la probabilidad de la hipótesis nula o la hipótesis alternativa ya que se asume de partida que la hipótesis nula es cierta.

En la inferencia bayesiana se trabaja con información sobre las probabilidades a priori de cada hipótesis estadística y las probabilidades a posteriori una vez conocidos los resultados y, por ello, se puede estimar la probabilidad de la hipótesis nula respecto a la alternativa (BF_{01}) o viceversa, es decir, la probabilidad de la hipótesis alternativa respecto a la nula (BF_{10}), dando así apoyo a una hipótesis estadística u otra. Si el valor del Factor Bayes es igual a 1 entonces se concluye que los datos apoyan por igual a ambas hipótesis estadísticas y la evidencia no favorece a ninguna. Como se observa en la figura 32, en el artículo se escribe que BF_{10} = 1,000, es decir hay una extrema evidencia a favor de la hipótesis alternativa sobre la hipótesis nula.

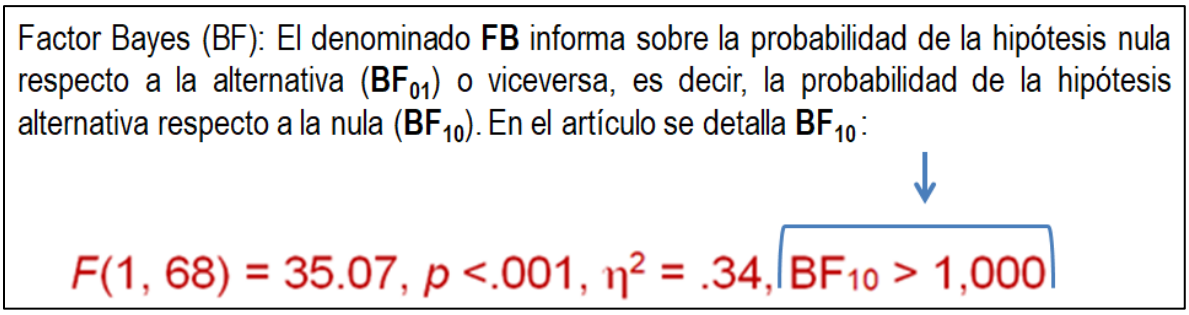


Figura 32. Redacción de la evidencia aportada por el estadístico del Factor Bayes (BF_{10})

En definitiva, la redacción en el artículo de Van Dessel y De Houwer (2019) de la evidencia aportada por los tres estadísticos anteriores (F, η^2 y BF_{10}) señala que hay diferencias estadísticamente significativas entre las medias ("significación estadística", $p \leq$ = alfa), el tamaño del efecto como proporción de varianza explicada es grande ($\eta^2 > .15$) y la probabilidad de la hipótesis alternativa (hay efecto o existe relación entre las variables) es extremadamente superior a la probalidad de la hipótesis nula (no hay efecto o existe una relación nula entre las variables) como explicación de los resultados encontrados en la muestra de participantes.

Como ejercicio de reflexión, puede realizar el mismo análisis que se ha ejecutado anteriormente con el segundo resultado que se ofrece en el artículo (figura 33). Respecto a esta redacción es conveniente anotar que la frase "efecto marginalmente significativo" es totalmente inadecuada y suele ser señal de un sesgo del investigador o investigadora a favor de su hipótesis científica. El resultado o es estadísticamente significativo porque $p \leq$ = alfa o no es estadísticamente significativo porque $p >$ alfa, pero no utilizar esa frase ya que induce a engaño y el lector o lectora podrían creer que al fin y al cabo se puede considerar estadísticamente significativo porque "casi" lo es. No es cierto. Es una falacia.

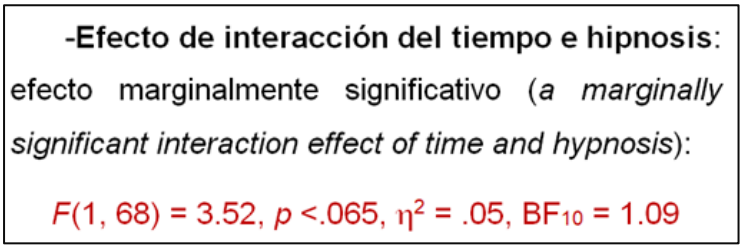


Figura 33. Redacción de la evidencia aportada por el estadístico del Factor Bayes (BF_{10})

En estos momentos el ejercicio anterior es solamente de reflexión y de autoevaluación personal. Pero es importante tener en cuenta que trabajar siguiendo el planteamiento de la Práctica Basada en la Evidencia exige tener los conocimiento de metodología de investigación suficientes para leer de forma crítica o activa los artículos e informes que se publican dentro de su área de trabajo. Además, desde el punto de vista ético, es necesario seguir formándose en las nuevas técnicas de análisis que se van incorporando en los informes o en los nuevos objetivos de los estudios como puede ser leer de forma crítica las revisiones sistemáticas, los trabajos de meta-análisis y de meta-análisis en red, el denominado Factor Bayes, el tamaño del efecto y su intervalo de confianza y comprender adecuadamente los fundamentos de los procedimientos clásicos de análisis como las pruebas de contraste estadístico tipo *t* de Student, razón *F*, coeficiente de correlación o análisis de regresión, todos ellos muy utilizados en las Ciencias Sociales y de la Salud.

Redacción de los resultados del Supuesto: desamparo y depresión en ratas

La redacción de los resultados del *diseño entre-grupos unifactorial (A =2) univariado* del supuesto 1 de investigación siguiendo el formato del Manual del APA sería, por ejemplo, la siguiente:

Los resultados del estudio del efecto de la indefensión aprendida sobre la sintomatología depresiva mediante un diseño entre-grupos unifactorial univariado con dos grupos (grupo 1 = shock eléctrico escapable, grupo 2 = shock eléctrico no escapable) señalan que las ratas que son sometidas a un situación de indefensión (reciben shock eléctrico no escapable o independiente de su conducta) tienen una puntuación media más alta en depresión (Media = 32, *DT* = 7.79, *n* = 4) que las ratas que sí tienen control de la situación del shock y, por lo tanto, no desarrollan la situación de indefensión aprendida (Media = 18, *DT* = 7.62, *n* = 4), siendo la diferencia entre las medias de los dos grupos estadísticamente significativa, con un tamaño del efecto grande, $F(1, 6) = 6.61$, $p = 042$, $\eta^2 = .52$. Por lo tanto, observando las puntuaciones medias de las condiciones experimentales, las ratas del grupo de shock no escapable recorrieron el laberinto con mayor lentitud que las ratas que fueron sometidas a shock escapable.

En definitiva, el planteamiento teórico que el investigador o investigadora desarrolla en su hipótesis determinará la selección de un diseño de investigación concreto cuyo plan de investigación debe seguir la pauta de *maximización* de la

varianza sistemática primaria, *minimización* de la varianza del error y *control* de la varianza sistemática secundaria (principio MAX-MIN-CON) Las características de la investigación y especialmente la posibilidad o no de la asignación aleatoria a las condiciones de tratamiento determinará la modalidad o metodología de investigación y, como consecuenica, el alcance de la interpretación de la relación hallada en términos causales o no.

Redacción de los resultados de un ANOVA entre-grupos, unifactorial A = 2 y univariado

A continuación se detallan unos ejercicios de redacción de los resultados de un análisis de varianza entre-grupos unifactorial univariado siguiendo la normativa del Manual APA. Se redactan como ejemplos de los resultados de las investigaciones.

Redacción 1. Se cumple el supuesto de homogeneidad de las varianzas.

La primera hipótesis de trabajo plantea si existen diferencias estadísticamente significativas entre las respuestas de los padres y las madres que contestan al cuestionario y su perfil tecnológico como usuarios de Internet. El estudio de las posibles diferencias entre los padres y las madres en las puntuaciones obtenidas en la escala de Perfil Tecnológico del padre / madre usuario de Internet señala que los padres obtienen una puntuación media más alta (Media = 13.96, *DT* = 2.37, *n* =563) que la de las madres (Media = 13.19, *DT* = 2.38, *n* =1101), siendo la diferencia entre las medias estadísticamente significativa y el tamaño del efecto pequeño (diseño entre-grupos *A* = 2 entre-grupos, univariado, *F* (1, 1662) = 38.37, $p < .001$, $\eta^2 = .02$). El supuesto de homogeneidad o igualdad de las varianzas de las puntuaciones de los dos grupos (padres y madres) se cumple (Levene *F* (1, 1662) = 0.67, *p* = .41).

En la redacción anterior hay que tener en cuenta que cuando se redactan los resultados siempre debe informarse de los estadísticos descriptivos de media, desviación típica y número de observaciones de cada grupo (*n*). Además, se redacta el resultado de la prueba *F* del ANOVA con toda la información: grados de libertad 'entre' o de la fuente de varianza del efecto, grados de libertad 'intra-celdilla' o de la fuente de varianza del término de error, valor obtenido del estadístico, valor exacto de *p* y valor del tamaño del efecto (eta cuadrado por ejemplo; y si es un diseño con solo dos grupos (A = 2) se puede anotar el valor de la diferencia estandarizada de medias conocida como tamaño del efecto *d* de Cohen). Si los resultados del valor de

p que ofrece el programa estadístico pone .000 nunca se debe anotar .000 en la redacción. Se redactaría $p < .001$ ya que lo que indica .000 es que se trata de un valor muy pequeño, es decir, que es menor a .001.

Redacción 2. Se cumple el supuesto de homogeneidad de las varianzas y se ofrece una tabla de descriptivos.

A continuación se redacta un ejemplo de análisis de la diferencia entre las puntuaciones medias de dos grupos, se cumple el supuesto de hogeneidad de las varianzas, los estadísticos descriptivos se amplían y se ofrecen en una tabla.

Los resultados del análisis de la varianza (ANOVA) entre-grupos unifactorial univariado señala que hay una diferencia estadísticamente significativa entre las medias del grupo de payasos y el grupo de control sin payasos en la variable de ansiedad, siendo el tamaño del efecto muy grande, $F(1, 10) = 7.62$, $p = .02$, $\eta^2 = .43$. Se ha comprobado el supuesto de homogeneidad de las varianzas de los dos grupos, Levene $F(1, 10) = 0.28$, $p = .609$. Por tanto, el grupo de payasos obtiene la puntuación media de ansiedad más baja respecto al grupo de comparación que no recibe ningún tipo de intervención. En términos de diferencia estandarizada de medias, el tamaño del efecto de *d* Cohen = 1.59 (95% IC -2.89 a -0.29), es decir, un tamaño del efecto muy grande con una amplitud muy amplia del intervalo de confianza ya que oscila desde un tamaño del efecto pequeño a un tamaño del efecto muy grande. Los resultados descriptivos se presentan en la Tabla 9.

Tabla 9. Análisis descriptivo de la variable ansiedad

	Grupo de payasos	Grupo de control
n	6	6
Media	4.50	13.50
DT	4.85	6.35
Amplitud	13	17
Mínimo	1	1
Máximo	14	18

En la redacción anterior hay que tener en cuenta que cuando los descriptivos de la variable medida (variable dependiente) se detallan en una tabla ya no deben mencionarse en el texto, pues no hay que repetir la información en el texto y en la

tabla. Y las tablas que ofrece el SPSS tiene que elaborarlas el usuario y nunca copiar y pegar una tabla del SPSS o de otro programa estadístico en la redacción de los resultados. Las tablas con formato del Manual de APA no tienen líneas verticales y solamente se ponen las líneas horizontales en la primera fila en la parte superior e inferior y en la última fila solamente en la parte inferior.

Ejercicio para el lector o lectora

Se recomiendan que los lectores o lectoras lleven a cabo el análisis de forma manual y con un programa estadístico de los datos representados en la Imagen 6.

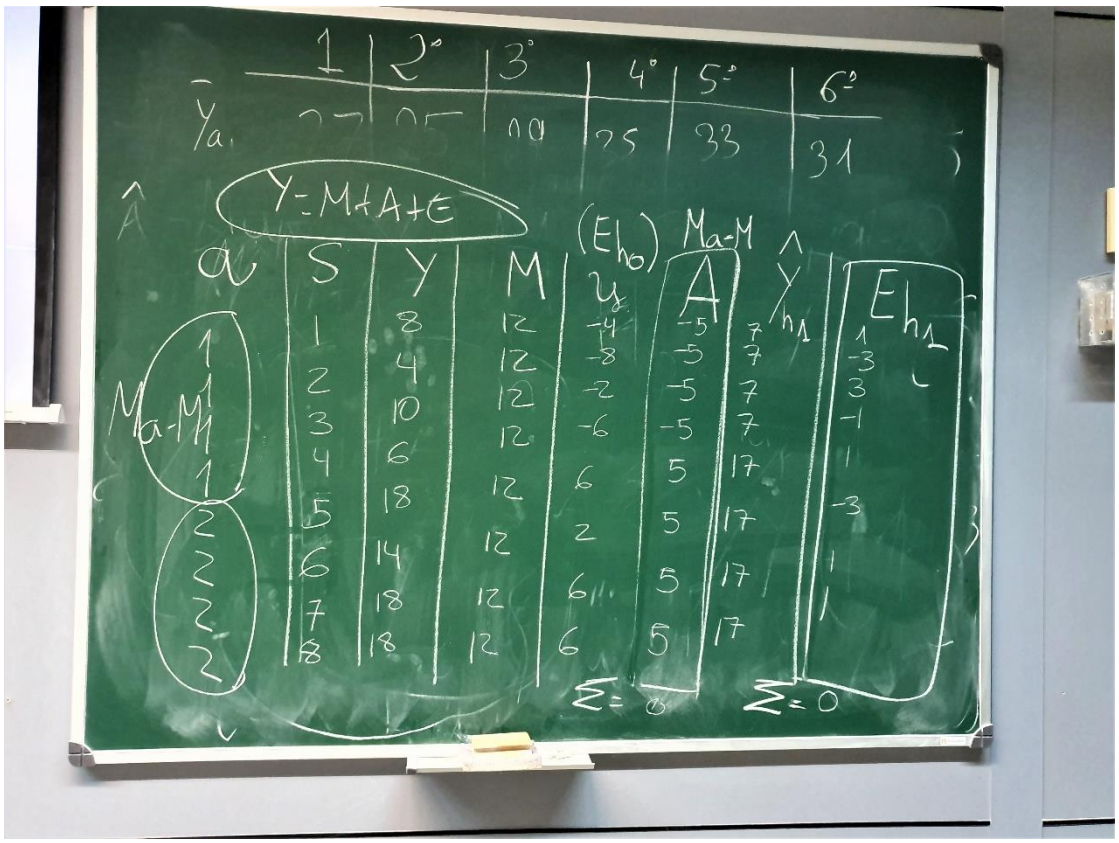


Imagen 6. Descomposición de la ecuación estructural de un conjunto de datos. Desarrollar el ejercicio de forma manual y con un programa estadístico. Imaginar un modelo teórico de esos datos (variables e hipótesis de un estudio) y redactar los resultados siguiendo las recomendaciones del Manual APA (7º edición).

Capítulo 9.Tamaño del efecto

Dolores Frías-Navarro

Universidad de Valencia

Índice

Citar el capítulo como:

Frías-Navarro, D. (2021). Tamaño del efecto. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

Comencemos con unas reflexiones. Qué podemos concluir ante el hallazgo de una diferencia estadísticamente significativa entre las medias de los grupos A (grupo de tratamiento) y B (grupo de control) cuando el valor *p* de probabilidad es igual a .01 ¿Estos resultados determinan que automáticamente el tratamiento A es sustancialmente mejor o con mayor utilidad que el tratamiento B? La respuesta es no. La interpretación de los valores *p* de probabilidad no es suficiente para hacer una inferencia de la significación clínica, sustantiva, social o práctica.

Usted lector o lectora piensa que ¿un resultado estadísticamente significativo con un valor *p* de .002 tiene un efecto bastante mayor que el resultado obtenido con un valor de $p = .045$? ¿Usted considera que la significación estadística otorga importancia al efecto detectado?

Como ya se ha comentado anteriormente, los valores *p* de probabilidad no son un índice del tamaño del efecto, ni tampoco indican la probabilidad de que la hipótesis nula sea verdadera o falsa. Del mismo modo, una falta de significación estadística no significa que la hipótesis nula sea verdadera ni que los efectos de los dos grupos sean equivalentes. Conviene tener simpre presente que ausencia de evidencia no es evidencia de ausencia de efectos (Altman y Bland, 1995).

El valor *p* de probabilidad es un valor que limita el rechazo o no rechazo de la hipótesis nula dentro del procedimiento de significación de la hipótesis nula. De forma arbitraria Sir Ronald A. Fisher (1925) fijó el nivel de alfa en .05 (5% de error de Tipo I). El criterio de $p \leq .05$ se basa en la idea que tuvo Fisher acerca de la razonable confianza que representaba para señalar que un efecto existe. No implica ningún valor mágico, ni detecta importancia del hallazgo. Por ello es totalmente incorrecto concluir que un resultado con $p = .04$ es importante y otro con $p = .06$ no es importante. Seguramente si los tamaños de la muestra fuesen iguales esos efectos serían muy similares. De ahí la importancia de informar siempre de la probabilidad exacta del resultado (por ejemplo $p = .03$ en lugar de $p < .05$).

Un resultado estadísticamente significativo indica que la probabilidad del resultado observado (o más extremo) es menor a .05 ($p \leq .05$), si la hipótesis nula fuese cierta (es decir, si no hubiese efecto). Como consecuencia y dada esa baja probabilidad que el investigador o investigadora asume como riesgo de equivocarse, se rechaza la hipótesis nula y se concluye que probablemente existe un efecto. Pero

este valor de probabilidad que implica el rechazo de la hipótesis nula si el alfa fijado a priori es de .05 no dice nada ni del tamaño del efecto ni de la significación clínica del efecto observado. Por ello, puede ocurrir que un tamaño del efecto pequeño detectado en un estudio con un tamaño muestral grande tenga el mismo valor de *p* que un tamaño del efecto grande de otro estudio cuyo tamaño muestral en cambio es pequeño.

Debemos recordar de nuevo que con un tamaño de muestra suficientemente grande es muy posible alcanzar un resultado estadísticamente significativo aunque el impacto del efecto del tratamiento haya sido mínimo en términos sustantivos (tamaño del efecto trivial). Es cierto que un tamaño del efecto vinculado a un valor *p* de probabilidad de .002 es mayor que el de .045, pero las comparaciones sólo se pueden realizar cuando los tamaños muestrales de los dos estudios son iguales ya que no sería cierto si el tamaño de la muestra es mayor en el estudio que tiene un valor *p* de probabilidad menor (.002). El valor *p* de probabilidad depende del tamaño del efecto y también del tamaño de la muestra:

p = tamaño del efecto *x* tamaño de la muestra

A pesar de las críticas que el procedimiento clásico de significaión de la hipótesis nula NHST ha tenido desde casi el mismo momento que apareció (FRIAS:::)), el procedimiento de significación de la hipótesis nula sigue siendo la técnica habitual de análisis en las Ciencias Sociales y de la Salud y las deficiencias en su interpretación se han tratado de paliar incorporando la estimación del tamaño del efecto junto con los valores *p* de probabilidad.

Qué es el tamaño del efecto

El concepto de tamaño del efecto ha sido ampliamente tratado por Jacob Cohen (1969, 1988) en su clásico libro sobre la potencia estadística. El tamaño del efecto (*effect size*) es el nombre dado a un conjunto de índices cuantitativos que cuantifican la magnitud de la diferencia entre poblaciones o la relación entre las variables. Es decir, cuantifica el efecto de un tratamiento o variable independiente sobre la variable dependiente medida (en los diseños experimentales y cuasi-experimentales) o el grado de asociación o covariación entre las variables (en los diseños no experimentales), siendo su valor independiente del tamaño de la muestra. Y aquí

radica una de sus principales ventajas: el valor del tamaño del efecto no depende del tamaño de la muestra, a diferencia del valor *p* de probabilidad que está directamente relacionado con la muestra. Así, manteniendo constante el efecto, si la muestra se aumenta entonces el valor de *p* disminuye.

El tamaño del efecto es un índice estandarizado (se pueden comparar entre sí los valores de diferentes índices) y estima un parámetro que es independiente del tamaño de la muestra. El tamaño del efecto cuantifica en qué grado el fenómeno estudiado se encuentra en la población, de manera que cuando la hipótesis nula es verdadera el tamaño del efecto es cero.

En definitiva, el tamaño del efecto permite que el resultado pueda ser interpretado en términos de magnitudes, es decir, admite hablar de tamaño del efecto pequeño, mediano o grande. Además, representa el estadístico que se utiliza para los trabajos individuales o primarios en los estudios de meta-análisis para calcular posteriormente el tamaño del efecto medio. La interpretación de la magnitud del efecto depende generalmente del supuesto de que las puntuaciones del grupo experimental y control están normalmente distribuidas y tienen las mismas desviaciones típicas (homogeneidad de varianzas).

El proceso de diseño de investigación implica elaborar y diseminar el conocimiento sobre una temática y los investigadores muchas veces utilizan instrumentos de evaluación diferentes para medir un mismo constructo, impidiendo la comparación directa de los hallazgos. Una de las principales ventajas del uso de los tamaños del efecto es que permite trabajar con índices estandarizados o tipificados, es decir, con una métrica común. De este modo, se permite la comparación de los tamaños del efecto de diferentes estudios que han utilizado métricas distintas en la medida original de las variables dependientes al convertir esos valores en índices del tamaño del efecto. Por ejemplo, la sintomatología depresiva puede ser medida con diferentes auto-informes o con escalas de medida distintas y sin embargo pueden ser directamente comparables sus resultados cuando se computan los tamaños del efecto en cada uno de los estudios ya que gracias a la estandarización de sus valores tienen la misma métrica. Es decir, los valores de los diferentes instrumentos de medida son transformados a una escala común gracias a la estimación de los tamaños del efecto, ya sea vía diferencia estandarizada de

medias, correlaciones o índices de riesgo, pudiendo realizar así comparaciones directas entre los resultados de investigaciones diferentes.

Además, una vez realizada la transformación a una métrica común, como todos los valores del tamaño del efecto son comparables entre sí entonces se puede calcular una media de los tamaños del efecto procedentes de diferentes estudios sobre una misma temática para resumir todos los datos en un único valor tal y como se realiza en los estudios de meta-análisis con el tamaño del efecto medio. Se trata del tamaño del efecto medio obtenido tras efectuar un estudio de revisión sistemática cuantitativa conocido como meta-análisis. En un trabajo de meta-análisis se interpretan e integran los hallazgos de diferentes trabajos de investigación (estudios primarios) y se ofrece un tamaño del efecto medio dentro de un área determinada de investigación.

La información que proporciona el tamaño del efecto puede ser utilizada con tres fines dentro del proceso de diseño de investigación:

1. Para comunicar información sobre el tamaño del efecto individual vinculado a una hipótesis de trabajo y permitir con ello la interpretación del resultado encontrado dentro de un pensamiento meta-analítico.
2. Para permitir los análisis de potencia estadística a priori y estimar el tamaño de la muestra para lograr un nivel aceptable de potencia estadística.
3. Para integrar los resultados de los estudios en un trabajo de meta-análisis o revisión sistemática cuantitativa identificando el tamaño del efecto medio.

En definitiva, el tamaño del efecto se refiere a la magnitud de la relación entre las variables y es un concepto diferente de la significación estadística o del contraste de hipótesis. Es un resultado que debe añadirse a la redacción de los resultados junto al valor *p* de probabilidad si se utilizan pruebas de hipótesis estadísticas en el estudio. Un resultado estadísticamente significativo podría tener un valor de tamaño del efecto pequeño del mismo modo que un resultado estadísticamente no significativo podría tener un valor de tamaño del efecto grande y no detectarse como estadístcamente significativo debido a que se utilizó un tamaño de la muestra pequeño. Hasta ahora,

gran parte de los manuales que explican el contraste estadístico suelen obviar la explicación del tamaño del efecto, del mismo modo que no se incluía en los programas docentes, pero actualmente es necesario conocer su concepto ya que se solicita por las revistas y la comunidad científica como un índice que debe acompañar al valor *p* siempre que sea posible. Actualmente ya se informa del tamaño del efecto en la mayoria de los artículos, sin embargo, la interpretación e implicaciones de sus resultados no llega a integrarse en la discusión de los hallazgos y suelen ofrecerse sus datos, pero se interpreta de forma superficial y con poco valor informativo en la discusión. Quizás, debido a la falta de comprensión del concepto de tamaño del efecto y qué información aporta.

Cómo estimar el tamaño del efecto

Existe un amplio número de índices del tamaño del efecto para cuantificar el efecto de un tratamiento o la relación entre las variables (Kirk, 1996). Las tres principales familias de índices son:

1. La familia de la 'diferencia estandarizada de medias' tipo *d* de Cohen (1988) donde el tamaño del efecto señala el grado de diferencia estandarizada entre dos medias.
2. La familia del 'coeficiente de correlación' *r* donde el tamaño del efecto expresa el grado de asociación entre dos variables. O la proporción de varianza explicada (eta cuadrado, η^2, η^2 parcial, R^2), entre otros índices.
3. La familia de los 'índices de riesgo' para tablas de contingencia 2 *x* 2 con resultados binarios o dicotómicos como el riesgo relativo, la reducción absoluta de riesgo, el número de pacientes que será necesario tratar (NNT) y el denominado *odds ratio* donde se estima la proporción de sujetos que experimentan un determinado resultado.

Los dos índices de tamaño del efecto más utilizados en las publicaciones psicológicas son la *d* de Cohen y la *r* de Pearson. El primero, *d* de Cohen, se utiliza cuando se trabaja con las medias de dos grupos y se calcula como la diferencia entre las medias dividido por la desviación típica común. El segundo, *r* de Pearson, se utiliza cuando los datos se analizan para comprender el grado de asociación entre dos variables o en qué grado una variable se puede predecir a partir de la otra. Los

dos índices de tamaño del efecto se pueden convertir algebraicamente de entre sí, del mismo modo que se pueden realizar conversiones con el resto de índices del tamaño del efecto.

Familia de diferencia estandarizada de medias

Los tres principales índices del tamaño del efecto de la familia de *diferencia estandarizada de medias* son la denominada *d* de Cohen (1988), la *g* de Hedges (1982) y la delta (Δ) de Glass (1976). Su aplicación requiere necesariamente diseños con: 1) dos grupos independientes (A = 2 grupos, grupos independientes, grupo experimental y grupo de comparación o control) y una variable continua medida. 2) O, tambien, diseños con dos muestras relacionadas (A = 2 mediciones o medidas repetidas, grupos dependientes), es decir, cuyos participantes han sido medidos en los dos grupos o condiciones de la investigación.

En los índices del tamaño del efecto de diferencia estandarizada de medias, los tres estadísticos utilizan como numerador la diferencia de medias entre el grupo A y el grupo B, pero difieren en la estimación que hacen de la desviación típica en el denominador del estadístico de tamaño del efecto.

d de Cohen

$$d = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{COMÚN}} \qquad S_{COMÚN} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

,

Donde $\overline{Y}_1$ es la media del grupo 1 (grupo experimental), $\overline{Y}_2$ es la media del grupo 2 (grupo control) y $S_{COMÚN}$ es la desviación típica común de las puntuaciones de los dos grupos.

Y en la fórmula de la desviación típica común, n_1 y n_2 son los tamaños muestrales de los grupos experimental y control respectivamente y S^2_1 y S^2_2 son las varianzas (desviaciones típicas al cuadrado) de los dos grupos mencionados.

Luego, la fórmula de la *d* de Cohen es:

$$d = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}}$$

El valor de tamaño del efecto *d* de Cohen puede oscilar desde 0 hasta infinito, aunque valores más grandes de 1 no suelen ser muy habituales en la literatura de Psicología.

Los valores propuestos por Jacob Cohen (1988) para interpretar la magnitud del efecto detectado entre la diferencia estandarizada de dos puntuaciones medias son los siguientes:

d = 0.2: pequeño

d = 0.5 mediano

d = ≥ 0.8: grande

Se puede realizar una conversión entre todos los índices del tamaño del efecto y posteriormente se ofrecera una tabla con los más comunes en la literatura psicológica junto con la relación entre esos índices respecto a los valores propuestos por Cohen.

Conviene tener muy presente:

1) Un tamaño del efecto pequeño podría ser importante desde el punto de vista sustantivo o clínico y un tamaño del efecto grande podría ser totalmente irrelevante o poco útil. Por ello, hay que contextualizar siempre la magnitud del efecto en el área concreta del fenómeno que se está estudiando.

2) Los valores de Cohen son orientativos cuando el investigador o investigadora desconoce qué magnitud tienen los efectos en su campo y no tiene medio de averiguarlo ni puede plantear el posible efecto en la población. Es decir, no es adecuado utilizar esos 3 valores para identificar siempre pequeño, mediano o grande, pues quizás el valor de *d* = 0.2 podría ser un tamaño del efecto grande en una determinada área de investigación.

En otras palabras, no se deben utilizar los puntos de corte tradicionales como oráculos de la verdad ya que son decisiones que unos científicos o científicas han

marcado utilizando una argumentación general (por ejemplo, .05 para alfa o los tres valores de tamaño del efecto de Jacob Cohen), pero será decisión del investigador o investigadora escoger los valores más adecuados para el fenómeno que está estudiando; decisiones que debe fundamentar y explicar en su artículo o informe de investigación de forma transparente. También, es cierto que en Psicología, en la mayoría de las ocasiones, los tamaños del efecto se sitúan entre *d* = 0.2 y *d* = 0.5. Encontrar tamaños del efecto más allá de *d* = 1 es bastante difícil.

A continuación es conveniente reflexionar sobre los mensajes del profesor Fernando Blanco en Tiwtter acerca de la magnitud o valor de la *d* de Cohen y la importancia de contextualizar los efectos en el área concreta de investigación (Figura 34).

Figura 34. Información de Fernando Blanco obtenida de Tiwtter

g de Hedges (conocida también como d de Hedges y d corregida)

Una cuestión importante relacionada con el diseño de la investigación es el tamaño de la muestra de observaciones (*N*). La estimación del valor del tamaño del efecto puede verse afectada seriamente cuando el tamaño de la muestra es pequeño.

El tamaño del efecto *d* de Cohen es un estimador sesgado del tamaño del efecto δ, estimando tamaños del efecto que son superiores a los reales, especialmente cuando el tamaño de las muestras es pequeño.

El sesgo es especialmente importante cuando el tamaño de la muestra es menor de 20 observaciones (o menor a 10 observaciones por grupo) (Nakagawa y Cuthill, 2007). Hedges y Olkin (1985) proponen la siguiente fórmula para corregir el sesgo (estimador insesgado):

$$d_{corregida} = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{COMUN}}\left(1 - \frac{3}{4(n_1 + n_2) - 9}\right)$$

$$d_{corregida} = d\left(1 - \frac{3}{4(n_1 + n_2) - 9}\right)$$

Donde *n* es el tamaño de los grupos 1 y 2 y *d* es el valor del tamaño del efecto no corregido o sesgado. Se trata de la *d* de Hedges aunque algunos autores lo nombran como la *g* de Hedges para no confundirlo con la *d* de Cohen. Sin embargo, estos cambios en las nominaciones de los estadísticos han provocado una gran confusión entre los investigadores e investigadoras.

Se recomienda utilizar siempre el tamaño del efecto corregido o la *d* corregida, aunque si el tamaño de la muestra es grande entonces la diferencia entre los dos estimadores desaparece.

Delta de Glass

Cuando las varianzas de los grupos difieren de forma destacada (existe heterocedasticidad) se puede utilizar como denominador de la ecuación la desviación típica del grupo control y obtener así el índice denominado *delta* de Glass, Δ, cuya fórmula es la siguiente (Glass, McGaw y Smith, 1981):

$$\Delta = \frac{\overline{Y}_1 - \overline{Y}_2}{S_{CONTROL}}$$

En este caso se asume que la varianza del grupo de control ($S_{control}$) es un estimador más adecuado de la varianza poblacional dado que la varianza del grupo experimental estará afectada por los efectos de la intervención. La varianza del grupo

de control, en cambio, no estará afectada y por lo tanto refleja mejor la desviación típica de la población (Lipsey y Wilson, 2001). La fuerza de esta asunción es directamente proporcional al tamaño del grupo de control. Cuanto mayor el tamaño del grupo de control más probable es que la varianza del grupo de control se parezca a la varianza de la población.

Visualización de la *d* de Cohen y su relación con otros índices

Kristoffer Magnusson (https://twitter.com/krstoffr) ha realizado diversas visualizaciones de diferentes conceptos estadísticos que son muy útiles para comprenderlos de forma adecuada y para su enseñanza, como por ejemplo la relación entre la potencia estadística y las pruebas de significación estadística, los intervalos de confianza, la distribución del valor *p*, la distribución de la *t* de Student, la inferencia bayesiana, las correlaciones, la estimación de máxima verosimilitud, las pruebas de equivalencia y las de no inferioridad y la *d* de Cohen (Figura 35).

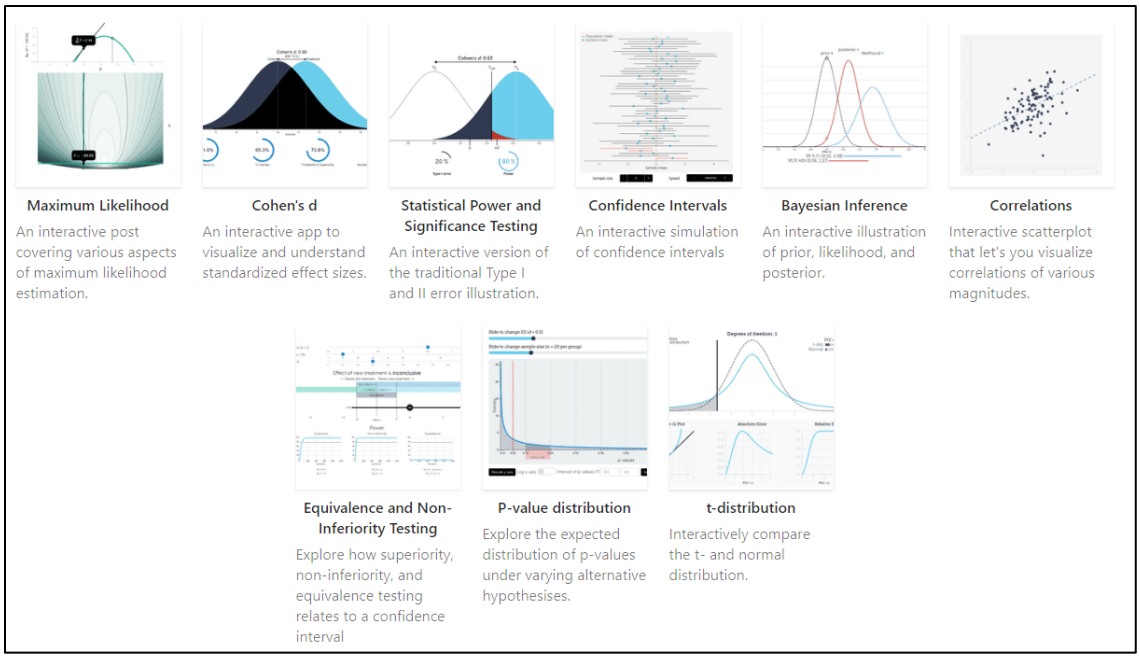


Figura 35. Visualizaciones de diferentes conceptos estadísticos. Disponible en: https://rpsychologist.com/

Para facilitar la interpretación sustantiva del valor de la *d* de Cohen se presenta a continuación la siguiente visualización de su funcionamiento. La herramienta

permite modificar el valor de la desviación típica y ofrece la diferencia no estandarizada entre las puntuaciones medias. En la zona de ajustes ('settings') se pueden cambiar los colores de las gráficas, los nombres de las variables, la media, la desviación típica... Además, se puede descargar fácilmente el gráfico.

Esta herramienta de visualización se encuentra en la siguiente dirección de Internet en su versión en castellano (traducida por Daniel Alcalá López (https://twitter.com/danalclop) y ha sido elaborada por Magnusson tal y como ya se ha comentado, investigador en psicología clínica:

https://rpsychologist.com/es/cohend/

En su versión en inglés se encuentra en la siguiente dirección:

https://rpsychologist.com/cohend/

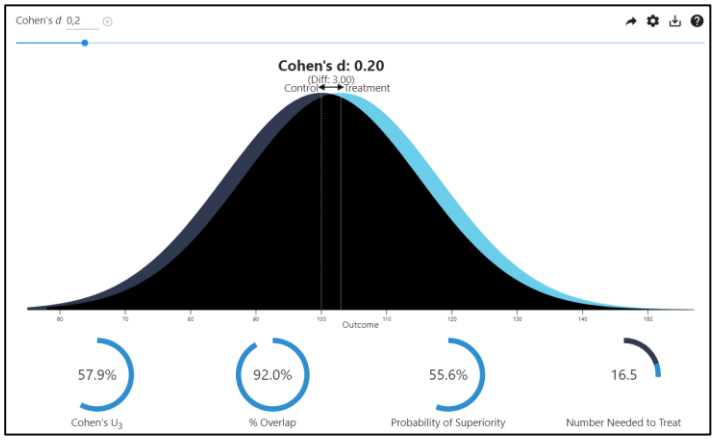
La visualización del valor de la *d* de Cohen (se asume distribución normal e igualdad de varianzas de los dos grupos) ofrece las siguientes relaciones directas entre diferentes índices de tamaño del efecto:

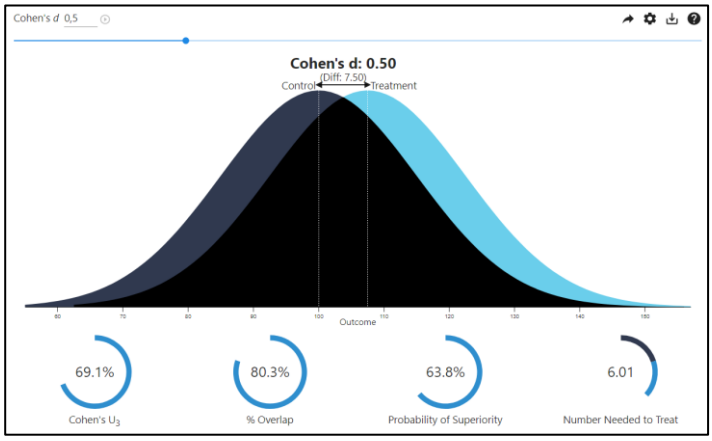
1. Solapamiento visual (visual overlap)
2. U_3 de Cohen (Cohen's U_3)
3. Probabilidad de superioridad (probability of superiority). También conocido como Tamaño del efecto del lenguaje común (common language effect size, CL) y Área Bajo la Curva (Area Under the (ROC) Curve, AUC). Representa la probabilidad de que una persona seleccionada al azar del grupo de tratamiento tenga una puntuación más alta que una persona seleccionada al azar del grupo de control. Si no hay efecto de una intervención ($d = 0$) entonces el valor de CL o probabilidad de superioridad = 50%, es decir, un sujeto elegido al azar del grupo experimental tiene la misma probabilidad de tener una puntuación por encima de la media del grupo de control que un sujeto del grupo de control
4. Porcentaje de solapamiento (percentage of overlap).

5. Número necesario a tratar (Number Needed to Treat, NNT): expresa los beneficios del tratamiento o la prevención de un efecto indeseable y cuantifica el número de casos adicionales que es necesario tratar con la intervención para producir el beneficio o para prevenir el efecto adverso comparado con el grupo de control.

A continuación se ofrece una breve explicación de esos 5 índices de tamaño del efecto para comprender su relación con el valor de la *d* de Cohen.

En la figura 36 se pueden observar los resultados de la visualización utilizando los tres valores clásicos de la *d* de Cohen de tamaño del efecto pequeño *d* = 0.2, mediano *d* = 0.5 y grande *d* = 0.8 junto con su relación con el resto de índices de tamaño del efecto que ofrece la visualización de Magnusson.

Visualización de la diferencia estandarizada de medias (*d*)	Tamaño del efecto
	PEQUEÑO *d* = 0.20 ↓ -U_3 de Cohen = 57.9% -% solapamiento = 92.0% -Probabilidad de superioridad = 55.6% -NNT = 16.5



MEDIANO

d = 0.50

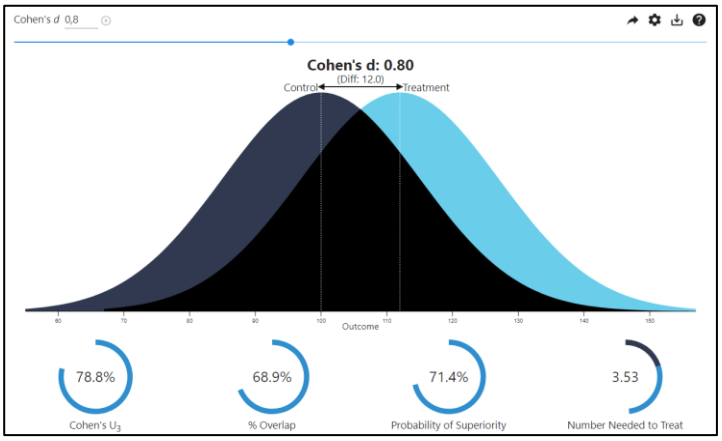
↓

-U_3 de Cohen = 69.1%

-% solapamiento = 80.3.0%

-Probabilidad de superioridad = 63.8%

-NNT = 6.01



GRANDE

d = 0.80

↓

-U_3 de Cohen = 78.8%

-% solapamiento = 68.9%

-Probabilidad de superioridad = 71.4%

-NNT = 3.53

Cuando el valor *d* aumenta:

aumenta U_3 y la probabilidad de superioridad, pero **disminuye** el porcentaje de solapamiento y NNT

Figura 36. Visualización de la diferencia estandarizada de medias (*d*): efecto pequeño, mediano y grande y su relación con otros índices de tamaño del efecto

Un ejemplo para practicar con la visualización. En primer lugar, ajustar en la visualización un valor de *d* = 0.46 y comprobar cómo se relaciona con los 5 índices índices del tamaño del efecto señalados anteriormente. Con una *d* = 0.46, el 68% del grupo de tratamiento estará por encima de la media del grupo de control (U_3 de

Cohen), el 82% de los dos grupos se superpondrán o habrá un solapamiento de sus puntuaciones y hay un 63% de probabilidad de que una persona seleccionada al azar del grupo de tratamiento tenga una puntuación más alta que una persona seleccionada al azar del grupo de control (probabilidad de superioridad). Además, para tener un resultado más favorable en el grupo de tratamiento en comparación con el grupo de control sería necesario tratar a 6.6 personas más o adicionales (se redondea a 7). Esto significa que si 100 personas pasan por el tratamiento, 15.1 personas más tendrán un resultado favorable en comparación con el tratamiento de control. En esta simulación del valor NNT, se asume que el 20% del grupo de control tiene "resultados favorables" (proporción de eventos favorales, *control event rate*, CER), es decir, mejora y este valor se puede cambiar presionando el símbolo de la rueda de ajustes.

Número Necesario a Tratar (NNT): número necesario de sujetos a tratar para observar un efecto beneficioso del tratamiento o para prevenir un efecto indeseable

El índice más utilizado dentro del ámbito de la Práctica Basada en la Evidencia y la estimación de la efectividad de un efecto del tratamiento en un estudio con metodología experimental (ensayo clínico aleatorio, ECA) es el número Necesario a Tratar (NNT) donde se mide la proporción de éxito / beneficio o prevención de un efecto adverso (resultado binario: éxito / no éxito) en un diseño con 2 grupos: grupo experimental y grupo control. Se utiliza una tabla de doble entrada. Cuando se trata de variables continuas, el índice de la diferencia estandarizada de medias *d* de Cohen es el más utilizado para el mismo diseño de investigación con dos grupos, pero cuando la variable medida (variable dependiente) es en este caso continua (Furukawa & Leucht, 2011).

El índice NNT se desarrolló en el contexto de la estadística en Medicina (Altman, 1998; Laupacis y cols., 1988) donde es ampliamente utilizado ya que de una forma sencilla se puede informar a los profesionales de la efectividad de un tratamiento médico, siendo mucho más intuitivo que Odds Ratio o Riesgo Relativo.

En Psicología también es un índice muy útil para expresar de forma más comprensible qué significa el valor de la *d* de Cohen cuando se habla de la eficacia de un tratamiento psicológico frente a un grupo de control activo o un grupo de control de no activo o de la reducción de determinados efectos adversos en la conducta del sujeto. Así, en el ámbito de la Psicología, los eventos o efectos de la terapia se pueden definir en términos de cambios en la conducta, ya se trate de incrementar una conducta o de disminuirla dado su efecto negativo sobre el bienestar de los individuos. Por ejemplo, informar de que el resultado de la intervención tiene un valor de $d = 0.62$ significa poco para los profesionales y las profesionales que no están familiarizados con ese índice, resultando más tangible o intuitivo hablar en términos del número de personas que mejoran con el tratamiento psicológico o de personas que reciben el tratamiento y no desarrollan el evento adverso. Hasta el momento, sin embargo, no suele presentarse en los artículos o informes, pero resultaría muy intuitivo para acompañar la interpretación de la magnitud del efecto mediante la *d* de Cohen.

En el contexto del cambio en la conducta, en un prepint o manuscrito aún no revisado por pares, Gruijters y Peters (2019) proponen cambiar el término por el de Número Necesario para el Cambio (NNC), definido como la cantidad de personas que necesitan ser tratadas con una intervención para lograr el cambio de comportamiento deseado en un sujeto adicional en relación con un grupo de control. Los autores han desarrollado un código R para facilitar las simulaciones de su propuesta y se encuentra disponible en https://github.com/matherion. Los autores destacan que en el ámbito psicológico se trataría de plantear objetivos específicos como por ejemplo promover que los sujetos realicen una actividad física de al menos 45 minutos, al menos tres veces por semana y después puntuar si lo han hecho o no y definir la tasa de respuesta del grupo experimental (EER) como la proporción de personas que han llevado a cabo ese objetivo de la intervención. O, por ejemplo, después de recibir un curso relacionado con los peligros del sexo no seguro, medir la proporción de jóvenes que de forma consistente utilizarían el preservativo. O, en función del nivel de estudios, estudiar la proporción de personas que se han infectado de una enfermedad de transmisión sexual desde de tener sexo sin preservativo frente a las personas que no han recibido (en este ejemplo, no se podrían establecer relaciones de causalidad entre el tratamiento y la respuesta ya que se trata de una

metodología no experimental). Se trata de estimar la frecuencia de esas conductas en un grupo experimental y en un grupo control y comparar sus proporciones. En este caso, el valor NNC informa sobre el número de personas que es necesario tratar para lograr un cambio de conducta favorable en 1 persona adicional en comparación con el grupo de control (más datos sobre la simulación se puede consultar en: https://matherion.github.io/userfriendlyscience/reference/nnc.html#references). Sin embargo, en la literatura psicológica aún no se utiliza mucho este índice ya que las medidas binarias o no son factibles o no son deseables o los datos sobre las tasas de eventos o proporciones no suele ser la medida elegida para medir la variable dependiente y se opta, en mayor medida, por utilizar una variable continua y expresar la magnitud del efecto en términos de desviaciones estándar (*d* de Cohen) o de proporción de varianza explicada (η^2). Y se prefiere, por ejemplo, medir el tiempo que los sujetos invierten en hacer ejercicio y utilizar esos datos como variable dependiente. Conviene tener presente que se pueden realizar conversiones entre los diferentes índices del tamaño del efecto y, por ejemplo, un valor dado de *d* se puede convertir a un valor de NNT tal y como se muestra en la visualización de Magnusson. Otra cuestión es interpretar de forma correcta lo que el valor de NNT quiere mostrar respecto a los objetivos del estudio.

El Número Necesario a Tratar (NNT) expresa de forma sencilla el tamaño del efecto beneficioso de una intervención o la prevención de un efecto indeseable, comparado con los resultados en un grupo de control. Es decir, estima el número de pacientes a los que habría que dar el tratamiento o sobre los que habría que hacer una determinada intervención para conseguir el objetivo deseado (beneficio o prevención en una persona adicional o en una persona más) que no se ha conseguido en los sujetos del grupo control. Con otras palabras, el NNT es el número de sujetos que sería necesario tratar para evitar algún resultado adverso en comparación con el grupo de control.

En resumen, el valor NNT estima cuántas personas deben recibir el tratamiento beneficioso para conseguir que 1 persona adicional mejore (se beneficie) o para prevenir en 1 persona adicional un efecto indeseable (Molina, 2012).

La pregunta que implica es la siguiente: ¿a cuántas personas hay que tratar para obtener un beneficio o para prevenir un evento adverso no deseado respecto a no

recibir ese tratamiento? El valor de estimación puntual NNT debe acompañarse de un intervalo de confianza, generalmente al 95% para interpretar la precisión de dicha estimación.

El computo del valor NNT requiere tener información de la proporción de respuesta del grupo experimental (personas expuestas al tratamiento) y del grupo control (personas no expuestas al tratamiento).

Por lo tanto, el valor del número necesario a tratar (NNT) se define como el número de personas que deben ser tratadas con una intervención o tratamiento concreto (por ejemplo, una determinada dosis de aspirina para las personas que han tenido un infarto de miocardio) para:

1) *producir* que 1 paciente adicional tenga un efecto positivo / favorable atribuido únicamente al efecto del tratamiento (aumentar la supervivencia) o bien,

2) para *evitar* o prevenir que 1 persona desarrolle un resultado negativo (por ejemplo, prevenir la muerte o un efecto secundario)

Es decir, expresa los beneficios de utilizar un tratamiento o una actividad preventiva sobre un grupo de control (Pita-Fernández & López-de-Ullibarri, 1999). Y, por ello, interesa un valor de NNT pequeño, pues si NNT = 1 entonces con un único sujeto tratado ya se demostraría el efecto del tratamiento en un sujeto adicional lo que significa que su efecto sería muy grande y que hubo efecto en todos los sujetos tratados. Esta situación como luego se comentará es casi imposible.

Por ejemplo, si un tratamiento tiene un valor de NNT de 5, significa que 5 personas deben ser tratadas con dicho tratamiento para evitar un resultado negativo en 1 persona adicional o para que 1 persona adicional mejore su dolencia.

Interpretación de los valores NNT

A continuación se detalla la información relacionada con la interpretación del resultado de NNT.

1. Los valores de NNT siempre se redondean al número entero más cercano.

2. Cuanto menor sea el valor de NNT mayor será la magnitud del efecto del tratamiento o más efectiva la actividad de prevención.

3. El valor de NNT debe expresarse junto a su intervalo de confianza para tener información sobre la precisión de su estimación puntual (Altman, 1998).

Por qué es importante incluir el intervalo de confianza de una estimación puntual siempre que se pueda. La respuesta es porque ayuda a interpretar hasta qué grado dicha estimación puntual se aproxima al verdadero valor de un determinado parámetro poblacional en términos de precisión de la estimación en una muestra. Por ejemplo, el intervalo de confianza (IC) del 95% significa que si se repitiera el mismo experimento 100 veces, obteniendo las muestras de la misma población, el IC incluiría el verdadero valor del parámetro poblacional en 95 ocasiones, mientras que el 5 restante no lo incluirían. Nunca se podrá saber si el valor de estimación puntual obtenido en el estudio se encuentra en esos 95 intervalos o se encuentra en los otros 5, pero sí interesa que el IC sea pequeño o estrecho ya que eso significa que es más preciso. Si el IC es muy amplio o grande entonces aporta poca información ya que el valor real del parámetro poblacional puede estar situado en cualquier punto del intervalo existiendo un número mayor de posibilidades.

4. El valor más alto posible de NNT es infinito y se corresponde a la situación donde la proporción de respuesta positiva es la misma en los grupos de tratamiento y control. Cuando el valor de NNT es infinito entonces el tratamiento no es eficaz, pues la Reducción Absoluta del Riesgo sería 0 y el valor de NNT sería infinito.

5. El valor de NNT no puede estar entre 0 y 1 ya que es imposible, por ejemplo, NNT = 0.5, es decir que la mitad de un paciente sea tratado para mejorar 1 paciente. Se redonderaría a 1. Por lo tanto, el valor positivo más bajo de NNT es 1 que sería la situación ideal (diríamos que realmente imposible) ya que las proporciones de respuesta positiva al tratamiento del grupo experimental y el grupo control serían del 100% y del 0%, respectivamente.

6. Un valor de NNT = 1 significa que en todos los pacientes a los que se les da el tratamiento se produce un resultado favorable, y por ello solo con un paciente ya se observa el efecto de la intervención. Es decir, solo es necesario tratar a una persona para obtener el beneficio buscado en otra persona. El valor de 1 es posible, aunque en la realidad es difícil encontrar un tratamiento tan altamente efectivo. Además, conviene recordar que el valor sustantivo de un tamaño del efecto como pequeño, mediano o grande siempre debe ir matizado por el contexto del fenómeno

que se está estudiando, los costes, la seguridad de la intervención o las preferencias del paciente (modelo de la Práctica Basada en la Evidencia). Como se ha comentado, cuanto más reducido es el valor de NNT mayor es el efecto de la magnitud del tratamiento, pero la importancia clínica o sustantiva va más allá de su valor y solo el profesional o la profesional pueden valorar dicha importancia con su juicio clínico. Por ejemplo, dentro del ámbito del estudio de la tasa de suicidios, un valor de tamaño del efecto muy pequeño podría ser relevante como indicador de éxito en la reducción del número de suicidios ya que tendría un impacto importante en la vida de las personas que están en riesgo de suicidio. Desde el ámbito del aprendizaje de un segundo idioma, un tamaño del efecto medio en la evaluación del rendimiento podría ser razonable para adoptar el nuevo método de enseñanza. Por lo tanto, no hay una regla general para especificar qué tamaño del efecto es el relevante ya que depende del propósito del estudio.

7. Respecto a los valores numéricos del intervalo de confianza de NNT, si su límite inferior está por encima de 1 entonces significa que existe una superioridad de una intervención con respecto a la otra o grupo de control. En cambio, si el intervalo de confianza incluye el valor de 1 en su intervalo entonces indica la ausencia de diferencias estadísticamente significativas en la variable medida entre el grupo de tratamiento y el grupo de control.

8. Resultados dañinos: NNH (‘*number needed to harm*’). El valor de NNT también puede utilizarse para estimar la capacidad potencial del tratamiento para producir daño en los pacientes. En este caso, este valor se denomina NNH (‘number needed to harm’), número necesario para hacer daño, y es una medida de la inseguridad de una intervención entendido como el número de personas que se necesitaría atender con un tratamiento específico para producir, o para no evitar, un evento adverso en 1 persona adicional. En este caso, lo que se busca es un valor de NNH alto, ya que indicaría que se necesitaría tratar a muchas personas para que 1 persona adicional tuviese un efecto adverso.

9. No hay un valor absoluto para un NNT que determine si un tratamiento es útil o no. Un valor de NNT = 1 significa que el beneficio se produce en cada paciente que recibe el tratamiento. Teniendo en cuenta que pocos tratamientos tienen eficacia de 100% y pocos controles (placebos o no tratamiento) no tienen ningún efecto, los

tratamientos por muy eficaces que sean, usualmente están en el rango de 2-4, excepto los antibióticos que tiene un valor de NNT de 1.1. Conviene recordar que un valor de tamaño del efecto grande *d* de Cohen = 0.8 se corresponde con un valor NNT = 4, redondeando.

10. Siempre hay que valorar el resultado de NNT dentro del contexto donde se produce el fenómeno y solo comparar los valores de NNT de dos intervenciones si realmente son comparables respecto a su contexto, características del grupos experimental y control y fenómeno estudiado. Por ejemplo, un valor de NNT = 200 podría ser aceptable si supone que tratar a 200 pacientes implica salvar una vida, pero podría ser excesivo si se trata de corregir levemente un problema. Siempre hay que contextualizar y valorar clínicamente su valor.

11. Hay que tener en cuenta que es más fácil demostrar significación estadística al comparar un tratamiento activo con un placebo que al comparar un tratamiento activo con un grupo de comparación que es un grupo de control activo (se trataría de otro tratamiento que tiene su efecto y no de placebo). Debido a ello, el valor de NNT es mayor cuando el *grupo de control es activo* (en los dos grupos hay efecto y se disminuye la distancia entre las proporciones de respuesta del grupo experimental y control). En cambio, si el grupo de control es de placebo (*grupo de control no activo* que se supone que su efecto será próximo a 0) entonces suele haber una mayor distancia entre las proporciones de respuesta positiva entre los dos grupos (efecto mayor) y, por lo tanto, menor será el valor de NNT. Por ejemplo, si la diferencia entre las proporciones de respuesta de los dos grupos es del 13% el valor de NNT = 8 y si la diferencia es del 18% el valor de NNT = 6, redondeando. En definitiva, si el grupo de control es no activo el valor de NNT deberá de ser pequeño ya que dicho grupo carece de efecto, mientras que si se trata de un grupo de comparación activo el valor de NNT deberá de ser mayor ya que dicho grupo tendrá cierto efecto en los sujetos.

Cómo calcular el Número Necesario a Tratar (NNT)

El cálculo del Número Necesario a Tratar (NNT) es sencillo y requiere calcular o disponer de las proporciones de respuesta de mejora o beneficio en el grupo experimental y el grupo control o de comparación que puede ser un grupo de control activo como por ejemplo otro tipo de fármaco o de tratamiento o un grupo control no activo tipo placebo o sin intervención (Chittaranjan, 2015). Hay calculadoras que

facilitan su cálculo aunque manualmente solo requiere sencillos cálculos a partir de los resultados observados en el grupo de intervención (grupo de expuestos al tratamiento) y el grupo de control (grupo de no expuestos o no tratados).

En el siguiente ejemplo se desarrolla el cálculo manual del valor NNT. En un estudio se ha llevado a cabo una intervención para disminuir la sintomatología depresiva en un grupo de 120 pacientes que habían sido asignados al azar al grupo experimental donde recibían un fármaco antidepresivo ($n = 58$) o al grupo de comparación o control que recibían un placebo ($n = 62$). Se trata de un estudio controlado aleatorizado (ECA), es decir, tiene una metodología experimental. Después de 12 semanas, la investigadora comprueba que 32 de los 58 pacientes que habían sido tratados con el fármaco antidepresivo habían mejorado según el criterio clínico de cambio fijado previamente. Por lo tanto, la proporción de respuesta positiva a los antidepresivos fue de 32 /58 = .55, es decir, el 55% de los sujetos del grupo experimental mejoraron en su nivel de sintomatología depresiva. Respecto al grupo de control que recibió un placebo se comprobó que 26 de los 62 pacientes mejoraron en su sintomatología depresiva, siendo la proporción de respuesta al placebo de 26 / 62 = .42, es decir del 42%. Por lo tanto, el tratamiento con el fármaco antidepresivo aumentó la proporción de respuesta basal relacionada con el placebo del 42% al 55%. Es decir, hubo un aumento en la mejora de la depresión en el grupo experimental del 13% respecto a la respuesta del grupo de control (55 – 42 = 13).

Con otras palabras, si el 55% de los pacientes del grupo experimental respondieron positivamente al efecto del fármaco antidepresivo (hubo cambio clínico), el 42% habría respondido al placebo de todos modos (igualmente tendrían un cambio clínico) por lo que la contribución única del fármaco antidepresivo fue solo del 13%. Y esta es la gran peculiaridad del valor NNT: informa del efecto único del tratamiento.

Con esos resultados se puede concluir que el tratamiento con antidepresivos en 100 pacientes da como resultado 13 personas que mejoran por el efecto único del fármaco, es decir, esa es la respuesta al tratamiento por encima de la proporción de respuesta basal relacionada con el efecto placebo del 42%. Como se observa, en los dos grupos, tanto en el experimental como en el control, se produce una respuesta

positiva de beneficio o mejora y el valor de NNT informa del efecto concreto o único que se atribuye al tratamiento que se ha administrado.

La pregunta entonces es: ¿cuántos pacientes necesitarán ser tratados con el fármaco para que haya 1 paciente adicional que mejora por el efecto único de dicho fármaco y no por otras causas?. La respuesta es sencilla: 100 / 13 = 7.7 pacientes deberán de ser tratados con el fármaco antidepresivo para que mejore 1 paciente por el efecto atribuido a dicho fármaco. El NNT se redondea al número entero más cercano, por lo que el valor de NNT en este ejemplo es igual a 8.

Ese resultado de NNT = 8 se puede obtener también calculando la "Reducción Absoluta de Riesgo" (RRA) ya que el valor NNT es el inverso del RRA y se detalla a continuación. Los datos de este ejercicio junto con los resultados de los índices que a continuación se desarrollan se presenta en la Tabla 10.

Grupo	Éxito / beneficio	No éxito / no beneficio	*n*
Experimental (expuestos)	32	26	58
Control (no expuestos)	26	36	62

Tabla 10. Grupo de expuestos y no expuestos al tratamiento y sus proporciones de respuesta positiva (éxito) y negativa (no éxito o beneficio)

Reducción Absoluta de Riesgo

Por lo tanto, para calcular el valor de NNT se necesita conocer la Reducción Absoluta de Riesgo (Absolute Risk Reduction, RRA) que es la resta entre los riesgos del desenlace del grupo de no expuestos (grupo control o grupo de comparación) y menos el riesgo del grupo de expuestos (grupo experimental): se trata de una diferencia de proporciones.

La palabra riesgo se utiliza en el sentido de *riesgo de presentar el desenlace*, desenlace que puede ser:

1) positivo: un beneficio (por ejemplo, mejorar en la depresión) o evitar un efecto adverso (por ejemplo, un ataque de pánico). En este caso interesaría que la proporción de sujetos del grupo experimental fuese mayor que la del grupo de control, por ejemplo, mejora en la conducta desadapativa medida.

2) negativo: puede ser presentar un efecto adverso (por ejemplo, la muerte o un efecto secundario). En este caso interesaría que la proporción de sujetos del grupo experimental fuese menor que la del grupo de control, por ejemplo, menos suicidios o menos síntomas de depresión.

Si se está estudiando un tratamiento que trata de aumentar la tasa de eventos en el grupo experimental (*experimental event rate*, EER) respecto a la tasa del grupo de control (*control event rate*, CER) entonces para facilitar su interpretación con signo positivo se sitúa la proporción del grupo experimenal en primer lugar:

$$NNT = 1 / EER - CER$$

Si el tratamiento tiene como objetivo disminuir una tasa de un evento indeseable o perjudicial entonces se puede cambiar el orden de las proporciones, presentando en primer lugar la proporción del grupo de control, para que el resultado sea positivo:

$$NNT = 1 / CER - EER$$

Donde EER es la tasa de eventos en el grupo experimental y CER es la tasa de eventos del grupo de control. Por ejemplo, si la tasa de respuesta positiva es del 60% en el grupo experimental (EER) y del 30% en el grupo control (CER) entonces el valor de NNT = 3.3 (es decir, 1 / (.6 - .3).

Si el riesgo está expresado como porcentaje, y no como proporción, la fórmula es:

$$NNT = 100 / RRA$$

El denominador de NNT es el valor de la Reducción Absoluta de Riesgo (RRA) y por lo tanto, el valor del número de personas que es necesario tratar para beneficiar un caso adicional o para prevenir el evento en un caso adicional (NNT) es el recíporoco o inverso del valor de RRA (NNT = 1 / RRA). El valor RRA señala la cantidad en la que el tratamiento reduce el riesgo absoluto en el grupo experimental en comparación con las personas que no recibieron dicho tratamiento y que forman parte del grupo de control. El valor de RRA como porcentaje se interpreta como que el tratamiento beneficia o previene el desenlace en un determinado porcentaje en los sujetos del grupo experimental comparado con los sujetos del grupo de control o

comparación. También se puede calcular un intervalo de confianza del valor puntual RRA para obtener información de la precisión de la estimación.

El valor de NNT se puede acompañar de su intervalo de confianza. Si se dispone del intervalo de confianza de la reducción absoluta de riesgo (RRA) entonces su cómputo es sencillo. Si, por ejemplo, en un estudio los límites del intervalo de confianza de RRA son 2.6 y 7.4 entonces se trata de calcular sus valores recíprocos o inversos. El intervalo de confianza del valor NNT sería (95%): 13.5 y 38.5, com límite inferior y superior de la estimación puntual del valor NNT = 20.

Por ejemplo, utilizando los resultados del ejercicio anterior, si el riesgo de beneficiarse del tratamiento es de .55 (55%) en el grupo experimental y de .42 (42%) en el grupo de control, entonces:

$$RRA = .42 - .55 = -.13$$

En términos de proporción, el valor de RRA = -.13. En términos de porcentaje, el 13% es la reducción del riesgo absoluto que tuvo la intervención respecto al grupo control. Significa que el 13% de los sujetos del grupo experimental se benefician de los efectos del tratamiento, en relación con el grupo de control. Por lo tanto, el tratamiento tiene una reducción del riesgo absoluto de depresión del 13% respecto del grupo control.

Si la reducción de riesgo absoluto es grande, se necesitará tratar a un pequeño número de pacientes para que se produzca el beneficio; por el contrario, si la reducción es pequeña, se deben tratar muchos pacientes para observar el beneficio.

Pero, la pregunta clave es ¿cómo se puede saber cuántos pacientes habría que tratar con el tratamiento para conseguir que un paciente más o adicional tenga ese beneficio? ¿Es necesario tratar a muchos pacientes para que haya un beneficio en 1 persona más? La respuesta la ofrece el valor de NNT.

$$NNT = 1 / -.13 = -7.7 \sim -8$$

El valor NNT se redondea y, por lo tanto, se señala que 8 pacientes deben recibir el tratamiento para que 1 persona más se beneficie del tratamiento o se evite un efecto adverso.

En un manuscrito, el lector o lectora debe estar atento al valor de las proporciones que tienen el grupo experimental y el control para observar en qué grupo hubo un mayor beneficio del tratamiento o en qué grupo se actuó de forma más preventiva. Los programas de cálculo suelen mostrar el valor NNT en negativo cuando hay beneficio o prevención ya que lo calculan como proporción de éxito del grupo control menos la proporción de éxito del grupo experimental, pero también hay manuales y programas que cambian el orden de los grupos y evitan el signo negativo si hubo éxito (proporción de éxito del grupo experimental menos la proporción de éxito del grupo control). Por ello, conviene siempre interpretar el valor de las proporciones para interpretar de forma correcta el valor de NNT.

Ejercicios de NNT

A continuación se presentan varios ejemplos.

Si en una investigación se obtiene que la respuesta de mejora para el tratamiento es igual al 60% y para el grupo de comparación es de 49%, qué valor de Número Necesario a Tratar (NNT) tendrían esos resultados. Se observa que 11 pacientes mejoran por el efecto único de la intervención por cada 100 sujetos tratados. Y se concluye que el valor NNT = 9. Es decir, 100 / 11 = 9.09 ~ 9.

En otro estudio se obtiene que el grupo experimental mejora un 15% y el grupo placebo un 10%. Por lo tanto, hay 5 sujetos que mejoran por el efecto único del tratamiento por cada 100 sujetos tratados. Como se observa, hay una mejoría en los dos grupos, pero cuántos sujetos hay que tratar para que 1 paciente adicional mejore por el efecto único del tratamiento. El valor de NNT = 20, es decir, será necesario que 20 pacientes de características similares reciban el tratamiento para conseguir que 1 persona más se cure por el efecto concreto de la intervención que es objeto de estudio que si se hubiese administrado, por ejemplo, un placebo.

Los resultados de una investigación señalan que el porcentaje de remisión de un determinado trastorno psicológico es del 50% con la intervención A y del 20% con la intervención B. Entonces, ¿cuántos pacientes habrá que tratar con la intervención A en lugar de B para encontrar 1 persona más donde remite el trastorno por el efecto de la intervención psicológica A? La respuesta es 4 pacientes, ya que NNT = 4.

También es importante tener en cuenta que el valor de NNT debe ser valorado observando las proporciones de respuesta positiva en el grupo experimental y control, para no caer en engaños. Por ejemplo, ya ha quedado claro que si NNT = 9 entonces 9 sujetos necesitarán recibir el tratamiento para que 1 paciente adicional se beneficie por el efecto único de dicho tratamiento. Pero con el valor de NNT se desconoce cuántos pacientes responden positivamente independientemente del tratamiento (efecto placebo), ni cuántos no responden en absoluto. Para tener esa información es necesario disponer de los datos de las proporciones de respuesta de los grupos experimental y control. Por ejemplo, en un estudio A el porcentaje de respuesta del grupo experimental es del 12% frente al 1% del grupo de control; de ahí que la ventaja del grupo experimental sea del 11% y, por lo tanto, NNT = 9. En otro estudio B también se obtiene un valor de NNT = 9, pero el porcentaje de respuesta del grupo experimental es del 99% frente al 88% del grupo de control, siendo su valor de NNT = 9. A pesar de que los valores de NNT son iguales las situaciones son muy diferentes. En el estudio A casi no hay respuesta positiva en el grupo de placebo mientras que en el grupo de tratamiento hay una moderada respuesta positiva de mejora o beneficio. En cambio, en el estudio B sí hay una alta respuesta positiva en el grupo de placebo y la valoración que se hace ahora de la ganancia que se atribuye a la intervención es pequeña. Los dos estudios tienen el mismo valor de NNT, pero la valoración sustantiva de dicho valor en función de los porcentajes de respuesta de los dos grupos es diferente, de ahí la importancia de valorar no solo la contribución única del tratamiento (NNT) sino también la tasa de respuesta positiva del grupo placebo o grupo de comparación.

Por lo tanto, es importante conocer las proporciones (o los porcentajes) de respuesta positiva o de beneficio en los dos grupos que se comparan cuando se calcula NNT ya que ayuda a distinguir entre un NNT = 10 cuando se calcula a partir de porcentajes de 20% versus 10% en cada grupo de otra situación con el mismo valor de NNT = 10, pero calculado a partir de porcentajes de 80% versus 70%. Se trataría de dos escenarios muy diferentes.

En el video siguiente de Youtube se hace una presentación del concepto de NNT y se detallan varios ejemplos en diferentes situaciones o necesidades de conocimiento del estudio:

https://www.youtube.com/watch?v=vcDiz7z5Pd0 Vídeo

Número Necesario a Dañar (NNH) y valoración de la magnitud de la relación beneficio-riesgo

El NNT puede utilizarse para predecir los daños y, en este caso, se trata del número de personas que hay que tratar para encontrar un efecto adverso o producir daño. Se trataría de un valor NNT negativo cuyo signo indica que el tratamiento tiene un efecto perjudicial o que presenta mayores efectos adversos en el grupo experimental.

El número necesario para dañar o número necesario a tratar para observar un efecto adverso (NNH) indica cuántos pacientes deben recibir un tratamiento particular para que 1 paciente adicional experimente daño o un resultado adverso concreto como respuesta al tratamiento, comparado con lo que sucedería en el grupo de control.

El valor de NNH tendrá una Reducción Absoluta de Riesgo (RRA) con un valor negativo (signo que se ignora al redactar el valor de NNH) dada la fórmula de RRA:

El índice de tamaño del efecto NNT y NNH han sido desarrollados dentro del modelo de Práctica Basada en la Evidencia y en el contexto de los ensayos clínicos para ayudar al clínico en la toma de decisiones en su práctica acerca de la eficacia de un determinado tratamiento o intervención, resultando su interpretación mucho má sencilla que la de otros índices como odds ratio. El juicio clínico del profesional es necesario para valorar la importancia de los valores de NNT y NNH.

La práctica del profesional y su toma de decisiones mejora de forma sustancial si puede valorar la evidencia aportada por los valores NNT y NNH, es decir, valorar si los riesgos son mayores que los beneficios. Así, con esa información, el profesional o la profesional conoce la magnitud del efecto del intervención (NNT) y la magnitud de los efectos adversos (NNH) y puede adoptar una decisión valorando beneficio y riesgo. Por supuesto, como ya se ha comentado, esta valoración la realiza el profesional o la profesional que debe tomar una decisión en función del fenómeno o variable de interés, su experiencia y su juicio clínico para el caso concreto que este valorando en su estudio.

Por lo tanto, los profesionales deben emitir un juicio clínico sobre el valor del beneficio del tratamiento y la gravedad del riesgo. Por ejemplo, una determinada intervención puede tener un valor de NNH = 100 para un riesgo de que ocurra un infarto de miocardio agudo en una persona como consecuencia de la administración del tratamiento. En este caso, la gravedad del infarto probablemente llevaría a la decisión de no prescribir esa intervención aunque solo aparezca el infarto en una persona de cada 100 dada la importancia de sus consecuencias en la salud de los pacientes.

Por ejemplo, si el beneficio es disminuir el dolor de forma leve y el efecto adverso es la muerte de la persona entonces valores de NNT = 4 y NNH = 100 (se necesitaría tratar a 4 pacientes para que 1 paciente se beneficie o le disminuya el dolor levemente y 1 de cada 100 pacientes tendrá un efecto adverso que es la muerte) entonces el balance no sería favorable respecto a instalar el tratamiento ya que una persona de cada 100 moriría frente a disminuir el dolor leve en una persona de cada 4 personas que reciban el tratamiento. La pregunta clave es: qué es más importante evitar el dolor leve o que se produzca una muerte.

Otro ejemplo, si NNT = 40 y NNH = 10 entonces sería necesario tratar a 40 pacientes para evitar que 1 presente un efecto adverso y, por otra parte, 1 de cada 10 presentará un efecto adverso (NNH = 10). En definitiva, la valoración de los valores es una decisión del profesional o la profesional ya que depende del fenómeno objeto de estudio, pero será importante tener en cuenta que es deseable que el valor de NNT sea bajo mientras que el valor de NNH sea elevado.

En resumen, los valores más bajos de NNT y los valores más altos de NNH se asocian con un perfil de tratamiento más favorable. Es decir, cuanto menor sea el valor de NNT mayor será el efecto de la contribución única de la intervención. Así, si NNT = 4 entonces solamente 4 sujetos necesitan ser tratados con la intervención para que 1 sujeto responda al efecto único del tratamiento. En cambio, si NNT = 20 entonces 20 sujetos necesitan recibir el tratamiento para que 1 sujeto responda al efecto único del tratamiento.

En la siguiente figura 37 se detallan los datos presentados en el estudio de Herruzo (2004) donde se realiza una revisión de los resultados aportados por la literatura sobre el tratamiento hormonal del cáncer de mama y el efecto beneficioso

del tamoxifeno en mujeres con tumor estrógeno positivo y ganglios positivos en comparación con un grupo control que no recibe tamoxifeno durante cinco años. En este caso, el número necesario a tratar (NNT) es el número de mujeres que necesitarían ser tratadas con tamoxifeno diario durante cinco años para prevenir 1 caso de cáncer de mama y el número necesario para dañar (NNH) señala cuántas mujeres deben ser tratadas para que 1 mujer desarrolle cáncer endometrial.

TABLE II

Tamoxifen in early breast cancer treatment

Outcome	Years of tamoxifen	NNT (95%CI)
Prevent recurrence	1	18 (13 to 30)
	2	16 (13 to 26)
	5	8 (7 to 10)
Prevent death	1	28 (18 to 66)
	2	30 (21 to 49)
	5	22 (15 to 36)
		NNH (95%CI)
Endometrial cancer	5	97 (68 to 168)

Figura 37. Resultados de NNT y NNH (95% IC) del estudio de Herruzo (2004)

En el artículo, el autor señala que en las mujeres tratadas con tamoxifeno durante 1, 2 ó 5 años, la reducción en el porcentaje de recidiva durante los 10 años siguientes es del 21%, 29% y 47% respectivamente, la reducción en la aparición de tumor contralateral es del 13%, 26% y 47%, respectivamente, y la reducción en la mortalidad es del 12%, 17% y 26%, respectivamente, y concluye el autor que por cada 8 mujeres a las que se les da tamoxifeno por 5 años, se previene la recidiva en 1 mujer (NNT de 8).

Concretamente, en el estudio de Herruzo (2004) se presentan resultados sobre el beneficio del tratamiento con el fármaco tamoxifeno en mujeres tratadas durante 1, 2 ó 5 años y los valores de NNT respecto a la respuesta de prevención de recidiva (‘prevent recurrence’) y la reducción de la mortalidad (‘prevent death’) como efectos beneficiosos y, por otra parte, el riesgo de aparición de cáncer endometrial (‘endomentrial cancer’) como efecto perjudicial durante los 10 años siguientes a la finalización del tratamiento.

Se observa en el estudio de Herruzo (2004) que cuando el tratamiento del tamoxifeno se ha llevado a cabo durante cinco años, el valor de NNT = 8, es decir,

por cada 8 mujeres a las que se les da tamoxifeno durante 5 años, se previene la recidiva o recaída (recurrencia) en 1 mujer (con un intervalo de confianza entre 7 y 10). Si el tamoxifeno se ha recibido durante 2 años entonces el valor de NNT = 16, es decir, por cada 16 mujeres a las que se les da tamoxifeno durante 2 años, se previene la recidiva o recaída (recurrencia) en 1 mujer (con un intervalo de confianza entre 13 y 26).

BENEFICIO:
NNT = 8 → Por cada 8 mujeres a las que se les prescribe el fármaco tamoxifeno durante 5 años, se previene la recidiva en 1 mujer.

Si el tamoxifeno se ha recibido durante 1 año entonces el valor de NNT = 18, es decir, por cada 18 mujeres a las que se les da tamoxifeno durante 1 año, se previene la recidiva o recaída (recurrencia) en 1 mujer (con un intervalo de confianza entre 13 y 30). Por lo tanto, el profesional puede observar que es más eficaz prescribir tamoxifeno durante 5 años para prevenir la recaida del cáncer de mama y, además, la precisión de esa estimación es mayor, pues con 5 años el intervalo de confianza del valor puntual NNT es el más preciso dado que es el más estrecho. En resumen, se puede concluir que el tratamiento de tamoxifeno mejora sustancialmente la supervivencia a 10 años en mujeres con tumor estrógeno positivo y ganglios positivos, reduciendo además las posibilidades de recurrencia del cáncer de mama. Respecto a los riesgos de recibir tamoxifeno se observa que el valor de NNH = 97, es decir, cada 97 pacientes tratadas con tamoxifeno durante 5 años se producirá cáncer endometrial en 1 mujer.

RIESGO:
NNH = 97 → Por cada 97 mujeres a las que se les prescribe el fármaco tamoxifeno durante 5 años, en 1 mujer aparecerá cáncer endometrial.

Otros resultados NNT / NNH en las investigaciones

Con el objetivo de practicar el concepto de Número Necesario a Tratar (NNT) y Número Necesario a Dañar (NNH)se detalla a continuación resultados de investigaciones reales.

En primer lugar, se pueden consultar la Web de ‘The NNT’ (https://www.thennt.com/) donde es posible encontrar el valor NNT en diferentes

investigaciones del ámbito de la Medicina, tanto en sus efectos beneficiosos como en los indeseables. Como ejemplo, se detalla en la figura 38 el efecto de la aspirina para prevenir el infarto de miocardio:

https://www.thennt.com/nnt/aspirin-preventing-first-heart-attack-stroke/

Los resultados de dicho estudio se representan en la Figura 38.

Table 2: Benefits and harms associated with the use of aspirin for primary prevention prevention.*

Outcome		Absolute Risk Difference	Relative Risk (95% CI)	NNT/NNH
All Cause Mortality	Mahmoud et al.[3+]	0.10	0.98 (0.93-1.02)	No benefit
	Zheng et al.[4]	0.13	0.94 (0.88-1.01)	No benefit
Death from cardiovascular causes	Mahmoud et al.[3+]	0.08	0.92 (0.83, 1.01)	No benefit
	Zheng et al.[4]	0.07	0.94 (0.83-1.05)	No benefit
Heart attack	Mahmoud et al.[3+]	0.3	0.82 (0.71–0.94)	NNT: 333
	Zheng et al.[4]	0.28	0.85 (0.73-0.99)	NNT: 361
Ischemic stroke	Mahmoud et al.[3+]	0.10	0.94 (0.86–1.02)	No benefit
	Zheng et al.[4]	0.16	0.81 (0.76-0.87)	NNT: 625
Major bleeding	Mahmoud et al.[3+]	0.40	1.47 (1.31-1.65)	NNH: 250
	Zheng et al.[4]	0.47	1.43 (1.30-1.56)	NNH: 213
Intracranial bleeding	Mahmoud et al.[3+]	0.10	1.33, (1.13– 1.58)	NNH: 1000
	Zheng et al.[4+]	0.11	1.34 (1.14-1.57)	NNH: 909

Abbreviations: NNT: Number Needed To Treat; NNH: Number Needed To Harm; CI: Confidence Interval

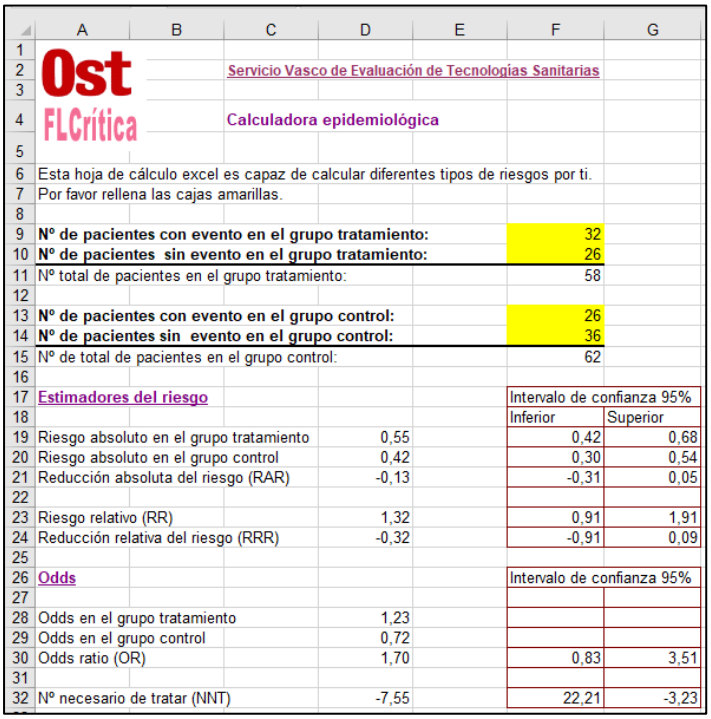
Figura 38. Valor NNT en diferentes investigaciones del ámbito de la Medicina

Programas para calcular NNT

Existen programas que realizan el cálculo de NNT y NNH. En las siguientes direcciones se pueden encontrar algunos de ellos:

-**Calculadora epidemiológica del Servicio Vasco de Evaluación de Tecnologías Sanitarias**. En su página Web se puede descargar un Excel que

permite realizar todos los índices vinculados con el riesgo de presentar un desenlace. Muy recomendable. Los resultados que ofrece se detallan en la Figura 39.



	A	B	C	D	E	F	G
1							
2	Ost		Servicio Vasco de Evaluación de Tecnologías Sanitarias				
3							
4	FLCrítica		Calculadora epidemiológica				
5							
6	Esta hoja de cálculo excel es capaz de calcular diferentes tipos de riesgos por ti.						
7	Por favor rellena las cajas amarillas.						
8							
9	**Nº de pacientes con evento en el grupo tratamiento:**					32	
10	**Nº de pacientes sin evento en el grupo tratamiento:**					26	
11	Nº total de pacientes en el grupo tratamiento:					58	
12							
13	**Nº de pacientes con evento en el grupo control:**					26	
14	**Nº de pacientes sin evento en el grupo control:**					36	
15	Nº de total de pacientes en el grupo control:					62	
16							
17	Estimadores del riesgo					Intervalo de confianza 95%	
18						Inferior	Superior
19	Riesgo absoluto en el grupo tratamiento			0,55		0,42	0,68
20	Riesgo absoluto en el grupo control			0,42		0,30	0,54
21	Reducción absoluta del riesgo (RAR)			-0,13		-0,31	0,05
22							
23	Riesgo relativo (RR)			1,32		0,91	1,91
24	Reducción relativa del riesgo (RRR)			-0,32		-0,91	0,09
25							
26	Odds					Intervalo de confianza 95%	
27							
28	Odds en el grupo tratamiento			1,23			
29	Odds en el grupo control			0,72			
30	Odds ratio (OR)			1,70		0,83	3,51
31							
32	Nº necesario de tratar (NNT)			-7,55		22,21	-3,23

Figura 39. Calculadora epidemiológica de índices de riesgo. Servicio Vasco de Evaluación de Tecnologías Sanitarias

-**Calculadora desarrollada por el Programa de Habilidades en Lectura Crítica Español (CASPe)** (https://www.redcaspe.org). En su página Web se puede descargar un Excel que permite realizar todos los índices vinculados con el riesgo de presentar un desenlace (https://www.redcaspe.org/herramientas/calculadoras). Concretamente, el valorde NNT se acompaña de su intervalo de confianza. También muy recomendable. Los resultados que ofrece se detallan en la Figura 40.

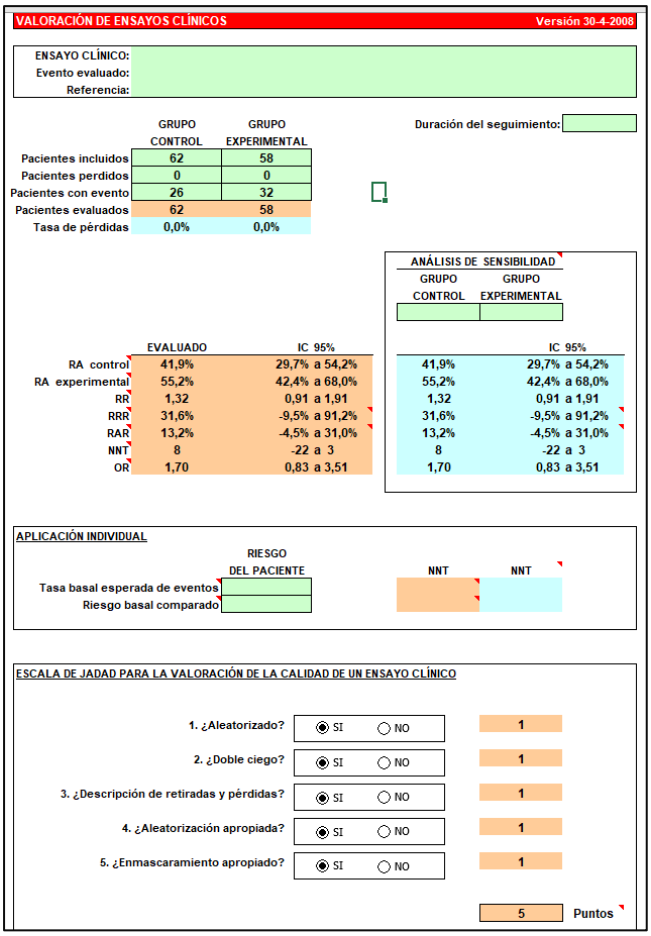


Figura 40. Calculadora epidemiológica de índices de riesgo. Programa de Habilidades en Lectura Crítica Español (CASPe)

-Calculadora desarrollada por la Oficina de Evaluación de Medicamentos del SES y Grupo Evalmed (figura 41):

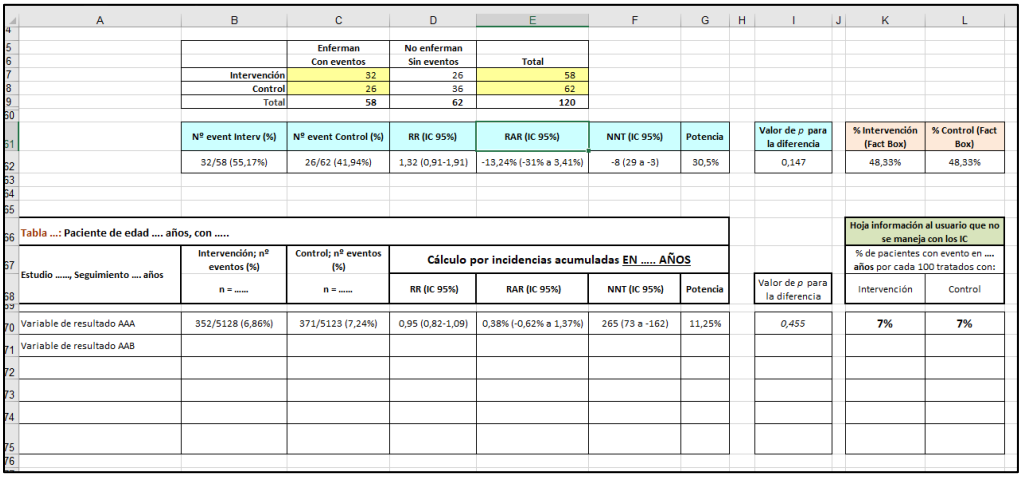
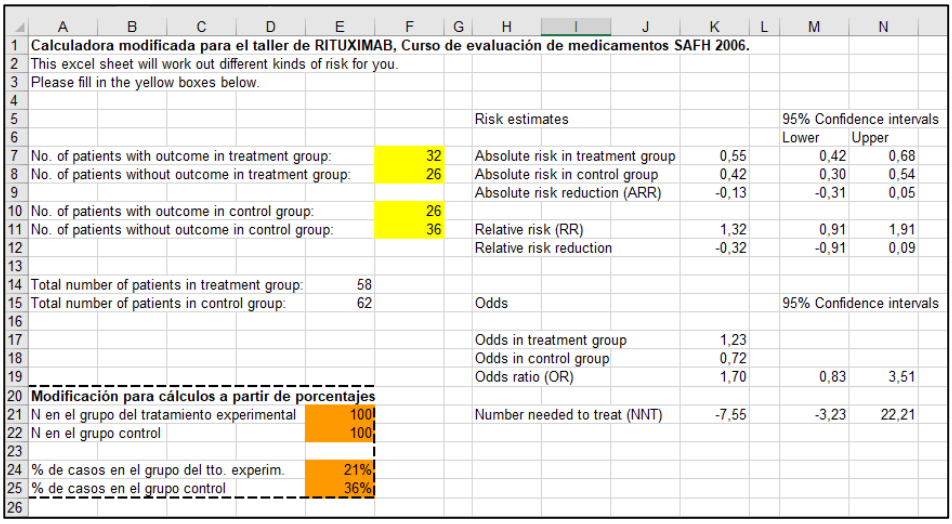


Figura 41. Calculadora epidemiológica de índices de riesgo. Evalmed

-Calculadora SIGN



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**Calculadora modificada para el taller de RITUXIMAB, Curso de evaluación de medicamentos SAFH 2006.**													
2	This excel sheet will work out different kinds of risk for you.													
3	Please fill in the yellow boxes below.													
4														
5								Risk estimates					95% Confidence intervals	
6													Lower	Upper
7	No. of patients with outcome in treatment group:					32		Absolute risk in treatment group			0,55		0,42	0,68
8	No. of patients without outcome in treatment group:					26		Absolute risk in control group			0,42		0,30	0,54
9								Absolute risk reduction (ARR)			-0,13		-0,31	0,05
10	No. of patients with outcome in control group:					26								
11	No. of patients without outcome in control group:					36		Relative risk (RR)			1,32		0,91	1,91
12								Relative risk reduction			-0,32		-0,91	0,09
13														
14	Total number of patients in treatment group:				58									
15	Total number of patients in control group:				62			Odds					95% Confidence intervals	
16														
17								Odds in treatment group			1,23			
18								Odds in control group			0,72			
19								Odds ratio (OR)			1,70		0,83	3,51
20	**Modificación para cálculos a partir de porcentajes**													
21	N en el grupo del tratamiento experimental				100			Number needed to treat (NNT)			-7,55		-3,23	22,21
22	N en el grupo control				100									
23														
24	% de casos en el grupo del tto. experim.				21%									
25	% de casos en el grupo control				36%									
26														

-http://psych.purdue.edu/~gfrancis/EquivalentStatistics/

Specify sample sizes:
n_1: 21 n_2: 22

Statistic	Value		
t-value	2.7533793	Convert to other statistics	
p-value	0.0087505	Convert to other statistics	
Cohen's *d*	0.84	Convert to other statistics	
Lower limit for Cohen's *d* 95% confidence interval	0.2104782	Convert to other statistics	
Upper limit for Cohen's *d* 95% confidence interval	1.4601566	Convert to other statistics	
Hedge's *g*	0.8245399	Convert to other statistics	
Lower limit for Hedges' *g* 95% confidence interval	0.1961178	Convert to other statistics	
Upper limit for Hedges' *g* 95% confidence interval	1.4438262	Convert to other statistics	
Post hoc power (from *d*)	0.7668356	Convert to other statistics	
Post hoc power (from *g*)	0.7514349	Convert to other statistics	
JZS Bayes Factor (alt / null)	5.4625388	Convert to other statistics	(might take a long time)
Λ, log likelihood ratio (full / null)	3.6477427	Convert to other statistics	(cannot be negative or zero)
ΔAIC (null - full)	5.2954854	Convert to other statistics	
ΔAICc (null - full)	4.9801008	Convert to other statistics	
ΔBIC (null - full)	3.5342853	Convert to other statistics	

Capítulo 10. Tamaño del efecto: La proporción de varianza explicada: R^2, η^2 y η^2 parcial

Dolores Frías-Navarro*
Marcos Pascual-Soler**

*Universidad de Valencia
**ESIC Business & Marketing School, España

Índice

Citar el capítulo como:

Frías-Navarro, D. y Pascual-Soler, M. (2021). Tamaño del efecto: La proporción de varianza explicada: R^2, η^2 y η^2 parcial. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

R^2 y η^2

En el análisis de regresión el tamaño del efecto se pueden estimar con el coeficiente de determinación (R^2). El coeficiente de determinación cuantifica la proporción de varianza de la variable respuesta o variable dependiente que es explicada por el efecto de la variable predictora o variable independiente. Su valor es el mismo que el del tamaño del efecto eta cuadrado (η^2) que se estima en el Análisis de la Varianza (ANOVA).

La principal ventaja de estos índices (proporciones) es su fácil interpretación ya que se puede multiplicar por 100 y hablar en términos de porcentaje de varianza explicada por el efecto de la variable independiente. Por ejemplo, si el valor de η^2 es .15 entonces el 15% de las diferencias encontradas entre los dos grupos se atribuye al efecto de la intervención o tratamiento. En términos de Cohen se trataría de un tamaño del efecto grande.

Además, 1 - η^2 es la proporción de varianza no explicada por el efecto del tratamiento y que se atribuye, por lo tanto, a la fuente de varianza del error. En el ejemplo anterior de η^2 = .15, el 85% de la varianza detectada entre las puntuaciones no está explicada por el efecto de la variable independiente y se atribuye a la fuente de la varianza del error. La suma de la proporción de varianza explicada y la no explicada es igual a 1.

Los índices vinculados al análisis de la varianza (ANOVA) como proporción de varianza eta cuadrado (η^2) y eta cuadrado parcial (η_p^2) se pueden obtener facilmente a partir de las sumas de cuadrados de la tabla del ANOVA tal y como se desarrollará a continuación.

Eta Cuadrado: η^2

Cuando se trata de estimar el tamaño del efecto de η^2 de forma total entonces hay que dividir la Suma de Cuadrados del Efecto por la Suma del Cuadrados Total.

El cálculo de eta cuadrado (η^2) o proporción de varianza explicada por el efecto del tratamiento es:

$$R^2 = \eta^2 = \frac{Suma de Cuadrados_{TRATAMIENTO\ o\ EFECTO}}{Suma de Cuadrados_{TOTAL}}$$

Si el diseño únicamente tiene un factor o variable independiente (diseño unifactorial), los valores de eta cuadrado (η^2) y eta cuadrado parcial (η_p^2) coinciden ya que sólo hay una fuente de varianza del efecto o de varianza explicada. Si se trata de un diseño con dos factores o dos variables independientes (A *x* B) (diseño factorial) entonces los valores de eta cuadrado (η^2) y eta cuadrado parcial (η_p^2) ya no coinciden, siendo más altos cuando se utiliza η_p^2 ya que sobreestima el valor del tamaño del efecto.

En el caso de los diseños factoriales es recomedable estimar el tamaño del efecto para los dos efectos principales (A y B) y para el efecto de interacción A *x* B ya que con ello el estadístico indicará la proporción de varianza que explica cada fuente de varianza de forma independiente. Por lo tanto, en un diseño factorial A *x* B habrá que estimar tres tamaños del efecto:

$$\eta^2_{EFECTO A} = \frac{Suma de Cuadrados_{EFECTO A}}{Suma de Cuadrados_{TOTAL}}$$

$$\eta^2_{EFECTO B} = \frac{Suma de Cuadrados_{EFECTO B}}{Suma de Cuadrados_{TOTAL}}$$

$$\eta^2_{EFECTO AB} = \frac{Suma de Cuadrados_{EFECTO AB}}{Suma de Cuadrados_{TOTAL}}$$

Los resultados de eta cuadrado calculado de forma total para la tabla del ANOVA 2 *x* 3 presentada en la figura 30 de la tabla de ANOVA son los siguientes:

$$\eta^2_{EFECTO A} = \frac{12}{58} = .207$$

$$\eta^2_{EFECTO B} = \frac{18}{58} = .310$$

$$\eta^2_{EFECTO AB} = \frac{12}{58} = .207$$

Eta Cuadrado parcial: η_p^2

El estadístico de eta cuadrado parcial (η_p^2) es la proporción de varianza explicada por el efecto (por ejemplo, efecto de A, efecto de B o efecto de interacción AB) más la del error que se puede atribuir a dicho efecto o fuente de varianza.

Cuando se trata de estimar el tamaño del efecto de η^2 parcial entonces hay que dividir la Suma de Cuadrado del Efecto a estimar por la Suma del Cuadrados del Efecto que se está estimado más la Suma de Cuadrados del Error. Es decir, el denominador de la fórmula incluye la suma de cuadrados del efecto que se está considerando más la suma de cuadrados del error.

$$\eta_p^2 = \frac{SumadeCuadrados_{TRATAMIENTO\ EFECTO}}{SumadeCuadrados_{TRATAMIENTO\ EFECTO} + SumadeCuadrdos_{ERROR}}$$

De nuevo, si se utiliza un diseño factorial se pueden estimar tres tamaños del efecto, dos para los efectos principales (A y B) y uno para el efecto de interacción (AB). Por lo tanto, la estimación del valor de η_p^2 para cada una de las tres fuentes de varianza de un diseño factorial A *x* B es la siguiente:

$$\eta_{pEFECTOA}^2 = \frac{SumadeCuadrados_{EFECTOA}}{SumadeCuadrados_{EFECTOA} + SumadeCuadrados_{ERROR}}$$

$$\eta_{pEFECTOB}^2 = \frac{SumadeCuadrados_{EFECTOB}}{SumadeCuadrados_{EFECTOB} + SumadeCuadrados_{ERROR}}$$

$$\eta_{pEFECTOAB}^2 = \frac{SumadeCuadrados_{EFECTOAB}}{SumadeCuadrados_{EFECTOAB} + SumadeCuadrados_{ERROR}}$$

Los resultados de eta cuadrado calculado de forma parcial para la tabla del ANOVA 2 *x* 3 presentada en la figura 30 se presentan a continuación. Estos valores son los que se detallan en el SPSS dado que dicho programa estadístico únicamente calcula la η_p^2.

$$\eta_{EFECTOA}^2 = \frac{12}{12+16} = .429$$

$$\eta_{EFECTOB}^2 = \frac{18}{18+16} = .529$$

$$\eta_{EFECTOAB}^2 = \frac{12}{12+16} = .429$$

Cuando se trabaja con diseños factoriales, Pierce y cols. (2004) subrayan las precauciones que los investigadores deben adoptar cuando interpretan el valor de eta cuadrado parcial. En este caso el valor de eta cuadrado parcial es mayor para una determinada fuente de varianza que el valor de eta cuadrado. Por lo tanto, la η_p^2 sobreestima el tamaño del efecto.

Diseños factoriales: η^2 y $\eta^2_{parcial}$

En los diseños donde se incluye la estimación del efecto de más de una variable independiente o factor ("diseños factoriales"), el tamaño del efecto puede estimarse para cada fuente de varianza de forma total (η^2) o de forma parcial (η_p^2). A continuación se detallan las características de cada estadístico utilizando los datos de la tabla resumen del ANOVA factorial 3 *x* 2 que se detallan en la figura 42 obtenido con el programa SPSS. Se observa que si se suman los valores de eta parcial al cuadrado de las tres fuentes de varianza del modelo de diseño factorial (A, B y AB) su valor supera a 1 ya que comparten información en el cálculo. Sin embargo, si se computa el valor de eta cuadrado la suma de los tres valores nunca superará al valor de 1 ya que se calculan separando la parte de varianza que específicamente explica cada fuente de varianza. Por lo tanto, conviene tener en cuenta que el valor de η_p^2 sobreestima el tamaño del efecto o proporción de varianza explicada.

Pruebas de efectos inter-sujetos

Variable dependiente: Y

Origen	Tipo III de suma de cuadrados	gl	Cuadrático promedio	F	Sig.	Eta parcial al cuadrado
Modelo corregido	42,000[a]	5	8,400	6,300	,004	,724
Interceptación	450,000	1	450,000	337,500	,000	,966
A	12,000	2	6,000	4,500	,035	,429
B	18,000	1	18,000	13,500	,003	,529
A * B	12,000	2	6,000	4,500	,035	,429
Error	16,000	12	1,333			
Total	508,000	18				
Total corregido	58,000	17				

a. R al cuadrado = ,724 (R al cuadrado ajustada = ,609)

Figura 42. Tabla de ANOVA y eta parcial al cuadrado

En definitiva, solamente cuando el diseño tiene una única variable independiente o factor los valores de eta cuadrado y eta cuadrado parcial coinciden. Además, cuando se suman los valores de eta cuadrado parcial de las fuentes de varianza de un diseño factorial se puede obtener un valor superior a 1. Con los datos de la figura 42 se observa que la suma de los valores de η_p^2 es igual a 1.387. Sin embargo, la suma de los valores de eta cuadrado de un determinado diseño factorial nunca puede ser superior a uno ya que se computan con el mismo término de error. Por ello, es

importante redactar en el informe o artículo qué estadístico del tamaño del efecto se ha calculado, recordando que el SPSS siempre computa el valor de η_p^2, sobreestimando el tamaño del efecto. Con los datos de la figura 30 se observa que la suma de los valores de η^2 es igual a .724, y si se suma la varianza atribuida al error o varianza no explicada por los efectos del tratamiento (1 - η^2 = .276) entonces la suma total siempre será igual a 1.

Programas para calcular el tamaño del efecto

A continuación se ofrecen unos datos y se detalla cómo estimar el tamaño del efecto y su intervalo de confianza con diferentes programas estadísticos.

Ejercicio dirigido a calcular el tamaño del efecto con los programas

Calcular el tamaño del efecto *d* de Cohen (diferencia estandarizada entre 2 medias; es un estadístico para diferencias entre pares de medias) respecto a la diferencia de medias de los siguientes dos grupos independientes:

1. Simbología concreta mañana: Media = 6, *DT* = 2.828, *n*=2
2. Simbología concreta noche: Media = 22, *DT* = 14.142, *n*=2

Estadísticos descriptivos

Variable dependiente: VARIABLE3

VARIABLE1	VARIABLE2	Media	Desv. Desviación
Simbología concreta	Mañana	6,00	2,828
	Tarde	15,00	,000
	Noche	22,00	14,142
	Total	14,33	9,647
Simbología matemática	Mañana	14,00	5,657
	Tarde	19,00	9,899
	Noche	44,00	2,828
	Total	25,67	15,306
Total	Mañana	10,00	5,888
	Tarde	17,00	6,164
	Noche	33,00	15,188
	Total	20,00	13,558

Como se puede observar, la diferencia entre las dos medias es de 16 (en valores absolutos), entonces ¿es o no es estadísticamente significativa dicha diferencia de medias? Se ha comprobado que el valor de *p* asociado a la prueba de contraste estadístico es de *p* = .012, luego la diferencia sí es estadísticamente significativa. Pero qué magnitud de tamaño del efecto tiene esa diferencia. Si se calcula se

observa que el valor de *d* = 0.90 (95% IC 0.35, 0.85). El ejercicio incluye la redacción de la diferencia como estadísticamente significativa y el tamaño del efecto como grande en términos de Cohen.

Program estadístico 1

https://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

Con este programa se puede observar que la *d* de Cohen sobrestima mucho el tamaño del efecto de diferencia estandarizada de medias (*d* = -1.568) cuando la *n* es pequeña (en el ejercicio era 2). En ese caso se optaría por el estadístico de diferencia estandarizada de medias denominado *g* de Hedges (*d* = -0.896) que controla el sesgo que se produce cuando la muestra es pequeña. La *g* de Hedges es el estadístico *d* de Cohen ajustado a grupos con tamaño del efecto pequeño.

La delta de Glass (Δ) se aplica cuando se utiliza la desviación típica del grupo de control en la fórmula. A veces, no interesa calcular una desviación típica común con las dos desviaciones de los grupos, sino que se opta por utilizar la del grupo control. Una herramienta de cálculo se encuentra en la figura 43.

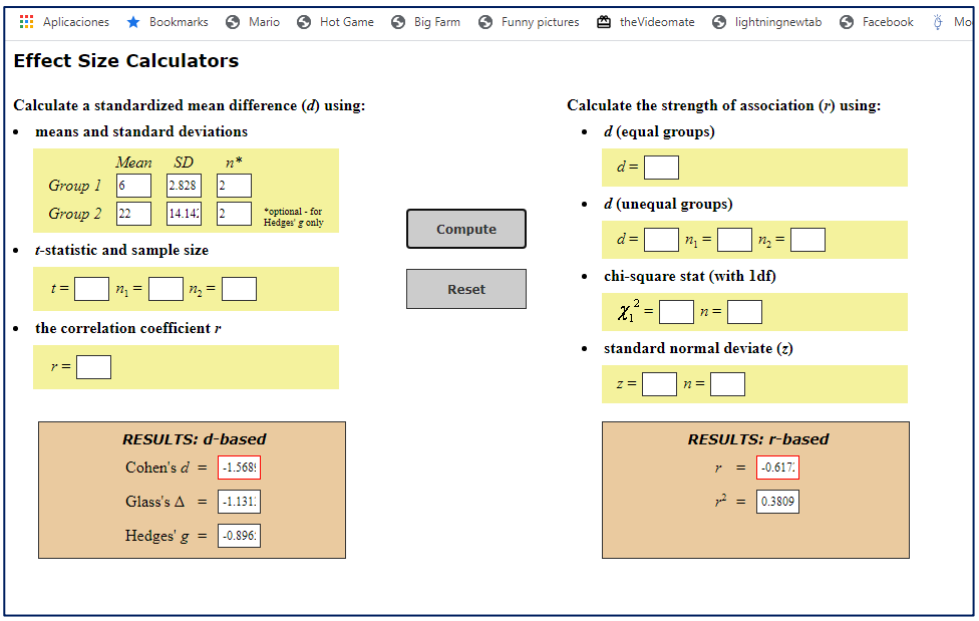


Figura 43. Tamaño del efecto:
https://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

Programa estadístico 2: Colaboración Campbell

https://campbellcollaboration.org/research-resources/effect-size-calculator.html

Aquí se clica sobre MEANS AND STANDARD DEVIATIONS y se utilizan las medias de dos grupos para calcular la *d* de Cohen. Con este programa solo se obtiene la *d* de Cohen (*d* = -1.569) y no estima la g de Hedges. Y hay que poner un punto y no una coma en los decimales cuando se introducen los datos (MUY IMPORTANTE). Este programa ofrece el intervalo de confianza del tamaño del efecto: 95% IC -3.81, 0.67 (figura 44). Si se observan los valores del intervalo de confianza se comprueba que dicho intervalo pasa por el valor de cero (es decir, el 0 es un valor posible del intervalo) indicando que la hipótesis nula se mantiene porque el valor de 0 es un valor del intervalo de confianza del tamaño del efecto.

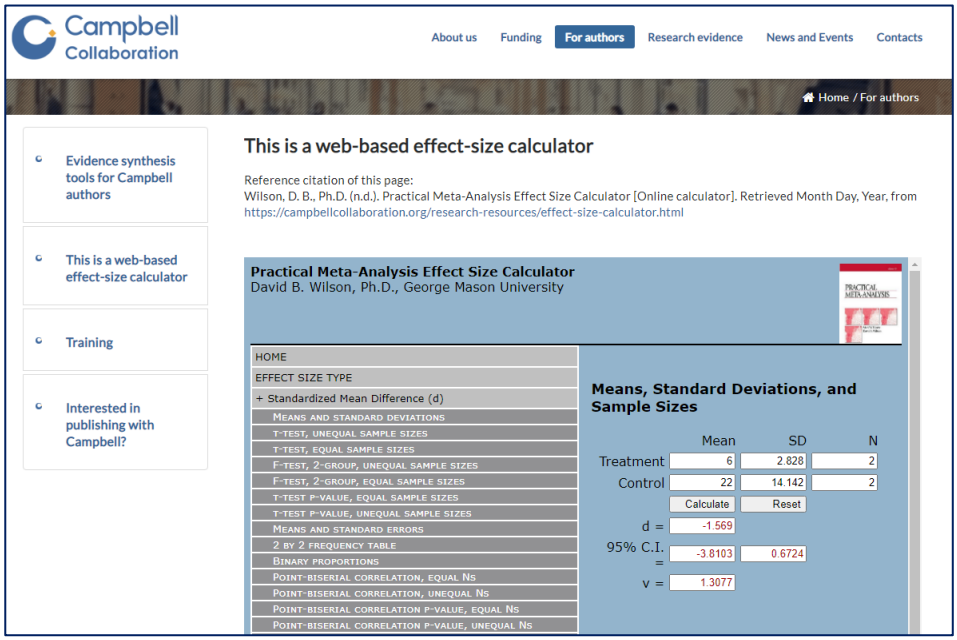


Figura 44. Tamaño del efecto: https://campbellcollaboration.org/research-resources/effect-size-calculator.html

Program estadístico 3: Programa de meta-análisis: Comprehensive meta-analysis

Se obtiene el mismo resultado de *d* de Cohen y su intervalo de confianza (Figura 45):

Model	Study name	Contrastes	Statistics for each study						
			Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	1,000	1,000	-1,57	1,14	1,31	-3,81	0,67	-1,37	0,17
Fixed			-1,57	1,14	1,31	-3,81	0,67	-1,37	0,17

d = -1.57, 95% IC -3.81, 0.67

Se obtiene el mismo resultado de *g* de Hedges y su intervalo de confianza

g de Hedges:

Model	Study name	Contrastes	Statistics for each study						
			Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	1,000	1,000	-0,90	0,65	0,43	-2,18	0,38	-1,37	0,17
Fixed			-0,90	0,65	0,43	-2,18	0,38	-1,37	0,17

g = -0.90, 95% IC -2.18, 0.38

Figura 45. Tamaño del efecto: Comprehensive meta-analysis

Conversión entre diferentes índices del tamaño del efecto

La interpretación del tamaño del efecto como diferencia estandarizada de medias puede ser comparada con el estudio de la correlación, los percentiles, la varianza explicada, el porcentaje de sujetos del grupo de control que se sitúa por debajo de la media del grupo experimental, la probabilidad de que una persona del grupo experimental sea superior a una persona del grupo de control si los dos son elegidos al azar (Common Language Effect Size, LC) y con el porcentaje de solapamiento de las curvas normales de los grupos experimental y control (overlap percent, OL) (ver figura 46).

Cohen (1988) proporciona una imagen mental de los tamaños del efecto gracias a la conversión de sus valores en términos de solapamiento de las curvas normales de los dos grupos (grupo experimental y grupo de control) donde el grado de solapamiento está determinado por el valor de tamaño del efecto *d* (*percentage overlap*, %OL). Cuanto mayor el tamaño del efecto menor es el solapamiento entre las curvas de ambas poblaciones. Es decir, asumiendo que las dos distribuciones de puntuaciones tiene una forma similar entonces las medias de los dos grupos serán diferentes si el solapamiento de sus distribuciones es pequeño. De este modo, el

tamaño del efecto se obtiene midiendo el grado de solapamiento de las puntuaciones de los grupos experimental y control (ver Figura 46).

Interpretación de los valores de tamaño del efecto de d de Cohen

d de Cohen	r Correlación biserial puntual	R^2 Coeficiente determinación (eta cuadrado, η^2) (x100 = % varianza explicada por tratamiento)	% de **Solapamiento** (OL %)*	**U1*** de Cohen (% No solapamiento)	**U2**** de Cohen (%)	**U3***** de Cohen (%)	Probabilidad de superioridad del grupo experimental (efectos +) (**CL**) (**AUC**)*****	**BESD******	**Odds ratio**	***f* de Cohen**
∞	1.000	1.000	0.0	100		100	1.00			
3.0	0.832	0.693	7.2	92.8	0.93	99.9	0.98	0.87	18.86	1.50
2.5	0.781	0.609	10.7	89.3	0.89	99.0	0.96	0.79	13.37	1.25
2.0	0.707	0.500	18.9	81.1	0.84	98.2	0.92	0.68	9.00	1.00
1.8	0.669	0.448	22.6	77.4	0.82	96.4	0.90	0.63	7.56	0.90
1.6	0.625	0.390	26.9	73.1	0.79	94.5	0.87	0.58	6.28	0.80
1.5	0.600	0.360	29.3	70.7	0.77	93.3	0.86	0.55	5.70	0.75
1.4	0.573	0.329	31.9	68.1	0.76	92.2	0.84	0.52	5.16	0.70
1.3	0.545	0.297	34.7	65.3	0.744	90.3	0.82	0.48	4.66	0.65
1.2	0.514	0.265	37.8	62.2	0.73	88.5	0.80	0.45	4.20	0.60
1.1	0.482	0.232	41.1	58.9	0.714	86.4	0.78	0.42	3.77	0.55
1.0	0.447	0.200	44.6	55.4	0.69	84.1	0.76	0.38	3.38	0.50
0.9	0.410	0.168	48.4	51.6	0.67	82.4	0.74	0.35	3.02	0.45
0.8	0.371	0.138	52.6	47.4	0.66	79.1	0.71	0.31	2.69	0.40
0.7	0.330	0.109	57.0	43.0	0.64	76.0	0.69	0.27	2.39	0.35
0.6	0.287	0.083	61.8	38.2	0.62	73.7	0.66	0.24	2.12	0.30
0.5	0.243	0.059	66.6	33.4	0.60	69.1	0.64	0.20	1.88	0.25
0.4	0.196	0.038	72.6	27.4	0.58	66.4	0.61	0.16	1.66	0.20
0.3	0.148	0.022	78.7	21.3	0.56	62.9	0.58	0.12	1.46	0.15
0.2	0.100	0.010	85.3	14.8	0.54	58.2	0.56	0.08	1.29	0.10
0.1	0.050	0.002	92.3	7.7	0.52	54.8	0.53	0.04	1.14	0.05
0.0	0.000	0.000	100.0	0	0.50	50	0.50	0	1.00	0.00

*U1 de Jacob Cohen (1988) es el porcentaje de no solapamiento: U1 = 1-% solapamiento

**U2 de Jacob Cohen (1988) = Percentil: porcentaje del grupo que se situaría por debajo de la media del grupo experimental cuando el tamaño del efecto es positivo

***U3 de Jacob Cohen (1988) = Percentil: porcentaje del grupo experimental que se situaría por debajo de la media del grupo control cuando el tamaño del efecto es positivo.

****CL (AUC): Common Language Effect Size (McGraw y Wong, 1992). Probabilidad de que una persona del grupo experimental sea superior a una persona del grupo control, si los dos son elegidos al azar. Siempre que la media del grupo experimental sea superior a la media del grupo de control. Si la media del grupo experimental es menor que la media del grupo de control (efecto negativo) entonces su probabilidad de exceder es igual a 1 - probabilidad de la tabla.

*****Binomial Effect Size Display (Rosenthal y Rubin, 1982).

- *U1 de Jacob Cohen (1988) es el porcentaje de no solapamiento: U1 = 1-% solapamiento
- **U2 de Jacob Cohen (1988) = Percentil: porcentaje del grupo que se situaría por debajo de la media del grupo experimental cuando el tamaño del efecto es positivo
- ***U3 de Jacob Cohen (1988) = Percentil: porcentaje del grupo experimental que se situaría por debajo de la media del grupo control cuando el tamaño del efecto es positivo.
- ****CL (AUC): Common Language Effect Size (McGraw y Wong, 1992). Probabilidad de que una persona del grupo experimental sea superior a una persona del grupo control, si los dos son elegidos al azar. Siempre que la media del grupo experimental sea superior a la media del grupo de control. Si la media del grupo experimental es menor que la media del grupo de control entonces su probabilidad de exceder es igual a 1 - probabilidad de la tabla.
- *****Binomial Effect Size Display (Rsenthal y Rubin, 1982).

Figura 46. Conversión entre índices del tamaño del efecto

Porcentaje de solapamiento entre las dos distribuciones

Cuanto mayor es el valor de *d* de Cohen menor será el solapamiento entre las distribuciones del grupo experimental y el grupo control o grupo de comparación.

El valor de tamaño del efecto de $d = 0$ indica un completo solapamiento (100%) entre las distribuciones de los dos grupos, siendo por lo tanto sus medias idénticas y las puntuaciones de los grupos indistinguibles (ver Figura 46).

En cambio, si un investigador obtiene un tamaño del efecto de $d = 0{,}1$ entonces el porcentaje de solapamiento entre los dos grupos sería de 92.3% lo que significa que sólo el 7,7% de los sujetos del grupo 1 obtendrían puntuaciones que no son obtenidas por los del grupo 2. En otras palabras, aproximadamente el 92% de los sujetos del grupo 1 obtienen puntuaciones que están dentro de la distribución de las puntuaciones obtenidas por el grupo 2 dado que el porcentaje de solapamiento es del 92.3%.

Si el tamaño del efecto fuese de $d = 3$ entonces existiría sólo un 7.2% de solapamiento entre las dos distribuciones de manera que se podría concluir que el efecto del tratamiento pudo discriminar de forma fiable a los grupos. Es decir, aproximadamente el 92% de los sujetos del grupo 1 obtuvo puntuaciones superiores

a las del grupo de control. En la dirección de Internet http://www.bolderstats.com/jmsl/doc/CohenD.html (Universidad de Colorado, EEUU) se pueden realizar simulaciones gráficas con diferentes tamaños del efecto de diferencia estandarizada de medias *d* de Cohen y el porcentaje de solapamiento entre dos distribuciones normales (*percentage overlap*, %OL) (ver Figura 46). En la dirección http://www.bolderstats.com/jmsl/doc/ se incluyen más simulaciones de conceptos estadísticos.

En la figura 46 se representan algunos ejemplos (*overlap* significa solapamiento de las distribuciones). Si el tamaño del efecto es $d = 0$ entonces el solapamiento de las dos distribuciones es total (100% de solapamiento). A medida que el valor del tamaño del efecto es mayor entonces las dos distribuciones están menos solapadas y va aumentando el porcentaje de puntuaciones del grupo experimental que está por encima de las del grupo de control.

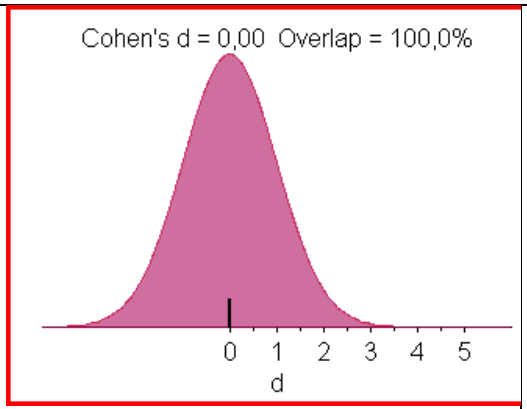
Si se utiliza el valor de *d* como una puntuación *z* entonces se puede realizar una transformación a percentiles y obtener otra interpretación alternativa. Por ejemplo si el valor de *d* es 2 entonces podemos afirmar que la distribución de las puntuaciones del grupo control está por debajo de la media de la distribución de las puntuaciones del grupo experimental en un 97.7%. Es decir, 97.7% es el área bajo la curva normal que corresponde a una puntuación $z = 2$ (ver figura 47). Por lo tanto, la media del grupo experimental se encuentra en el percentil 97.7 del grupo de control. En este caso el porcentaje de no solapamiento entre las distribuciones del grupo experimental y del grupo de control es de 81.1% (porcentaje de solapamiento de 18.9%).

Si el tamaño del efecto fuese cero, (ausencia de efecto dado que las dos medias serían iguales) entonces el porcentaje de casos del grupo de control que se sitúa por debajo del grupo experimental sería del 50%. Del mismo modo, también es la misma proporción de sujetos del grupo experimental que se sitúa por debajo del grupo de control, 50%. En otras palabras, las distribuciones de los dos grupos se superponen totalmente (100% de solapamiento de las dos distribuciones).

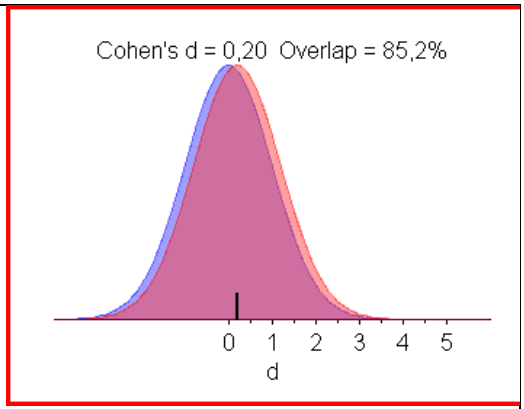
Otra aproximación a la interpretación del tamaño del efecto puede ser realizada con el estadístico conocido como Lenguaje Común o área bajo la curva de ROC (AUC). El tamaño del efecto del Lenguaje Común señala la probabilidad que tiene una puntuación seleccionada aleatoriamente de ser superior de otra puntuación

seleccionada también aleatoriamente de una segunda distribución. En otros términos, el valor de CL señala la probabilidad que tiene un sujeto seleccionado aleatoriamente del grupo experimental de superar la puntuación de otro individuo seleccionado aleatoriamente del grupo de control, asumiendo que las dos distribuciones son normales y las varianzas de cada grupo son homogéneas.

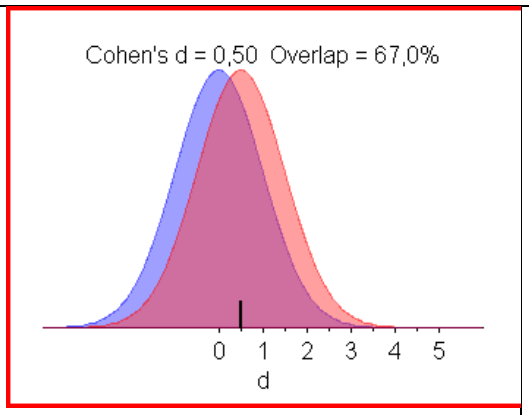
Por ejemplo, un tamaño del efecto grande en términos de Cohen de 0.8 supone que la probabilidad que tiene un sujeto del grupo experimental seleccionado aleatoriamente de superar la puntuación de un sujeto del grupo de control también seleccionado aleatoriamente es del 0.71. Es decir, la probabilidad que hay de obtener una diferencia entre las puntuaciones de los dos grupos mayor de cero es igual a .71. De este modo, el 71% de las veces un sujeto extraído al azar del grupo experimental obtendrá un valor superior que otro sujeto del grupo de control también extraído al azar. Si el tamaño del efecto fuese de *d* = 2 entonces en términos del lenguaje común la probabilidad que hay de obtener una diferencia entre las puntuaciones de los dos grupos es de .92. O lo que es lo mismo, la probabilidad de que la puntuación del sujeto del grupo experimental exceda a la puntuación del sujeto del grupo control es de .92. En el 92% de las ocasiones la puntuación de un sujeto del grupo experimental será superior a la de un sujeto del grupo de control. En la figura 47 se representan diversos valores de la d de Cohen y el % de solapamiento de als distribuciones.



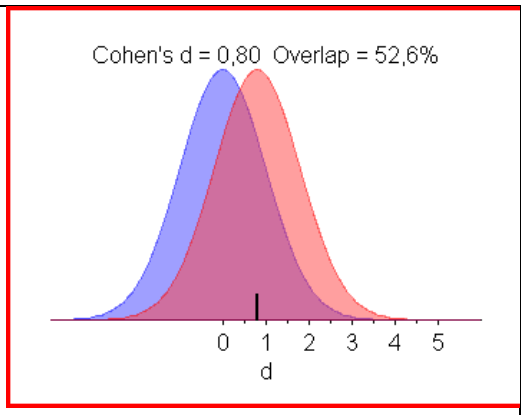
- Las puntuaciones de los grupos son idénticas. Total solapamiento de las distribuciones.



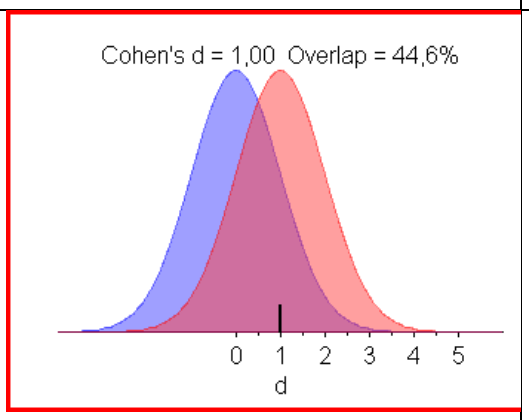
- El 14,8% del grupo experimental está por encima de la media del grupo de control



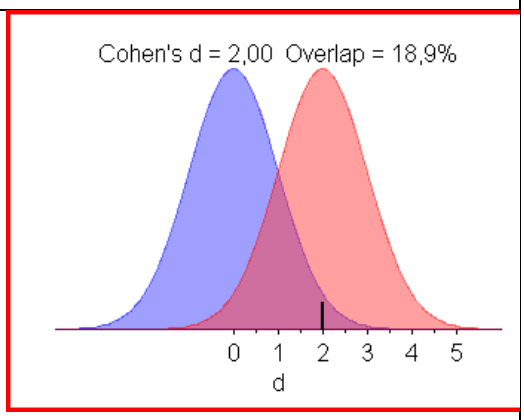
- El 33% del grupo experimental está por encima de la media del grupo de control



- El 47.4% del grupo experimental está por encima de la media del grupo de control



- El 55.4% del grupo experimental está por encima de la media del grupo de control



- El 81.1% del grupo experimental está por encima de la media del grupo de control

Figura 47. La *d* de Cohen y el solapamiento de las dos distribuciones

Redacción de resultados

Cuando se redactan los resultados del contraste de hipótesis es fundamental dar toda la información estadística que rodea al proceso de contraste de hipótesis estadística como el tipo de diseño y las variables implicadas en el análisis que se ejecuta en la ecuación estructural, los descriptivos, al menos, de media y su desviación típica y número de observaciones (*n*) del grupo y el valor de la razón *F* con sus grados de libertad, junto con su valor *p* exacto y la información del tamaño del efecto (en este caso se ha interpretado la proporción de varianza explicada conocida como eta cuadrado, η^2, cuyo calculo es suma de cuadrados del efecto

dividido por la suma de cuadrados total) y su intervalo de confianza (en la redacción del supuesto 1 no se ha incluido la información del intervalo de confianza de eta cuadrado).

En la siguiente dirección (https://www.gigacalculator.com/) se pueden llevar a cabo de forma rápida la estimación de los valores de media y desviación típica para que se incluyan en el apartado de redacción de los resultados:

https://www.gigacalculator.com/calculators/standard-deviation-calculator.php

Además, en la siguiente dirección también hay un amplio abanico de herramientas para realizar cáculos estadísticos:

https://www.gigacalculator.com/calculators/statistics/

Redacción de los resultados del supuesto: desamparo y depresión

La redacción de los resultados del *diseño entre-grupos unifactorial (A =2) univariado* del supuesto 1 de investigación siguiendo el formato del Manual del APA sería, por ejemplo, la siguiente:

Los resultados del estudio del efecto de la indefensión aprendida sobre la sintomatología depresiva mediante un diseño entre-grupos unifactorial univariado con dos grupos (grupo 1 = shock eléctrico escapable, grupo 2 = shock eléctrico no escapable) señalan que las ratas que son sometidas a un situación de indefensión (reciben shock eléctrico no escapable o independiente de su conducta) tienen una puntuación media más alta en depresión (Media = 32, *DT* = 7.79, *n* = 4) que las ratas que sí tienen control de la situación del shock y, por lo tanto, no desarrollan la situación de indefensión aprendida (Media = 18, *DT* = 7.62, *n* = 4), siendo la diferencia entre las medias de los dos grupos estadísticamente significativa, con un tamaño del efecto grande, $F(1, 6) = 6.61$, $p = 042$, $\eta^2 = .52$. Por lo tanto, observando las puntuaciones medias de las condiciones experimentales, las ratas del grupo de shock no escapable recorrieron el laberinto con mayor lentitud que las ratas que fueron sometidas a shock escapable.

Redacción de los resultados de un ANOVA entre-grupos, unifactorial A = 2 y univariado

A continuación se detallan unos ejercicios de redacción de los resultados de un análisis de varianza entre-grupos unifactorial univariado siguiendo la normativa del Manual APA. Se redactan como ejemplos de los resultados de las investigaciones.

Redacción 1. Se cumple el supuesto de homogeneidad de las varianzas.

La primera hipótesis de trabajo plantea si existen diferencias estadísticamente significativas entre las respuestas de los padres y las madres que contestan al cuestionario y su perfil tecnológico como usuarios de Internet. El estudio de las posibles diferencias entre los padres y las madres en las puntuaciones obtenidas en la escala de Perfil Tecnológico del padre / madre usuario de Internet señala que los padres obtienen una puntuación media más alta (Media = 13.96, *DT* = 2.37, *n* =563) que la de las madres (Media = 13.19, *DT* = 2.38, *n* =1101), siendo la diferencia entre las medias estadísticamente significativa y el tamaño del efecto pequeño (diseño entre-grupos *A* = 2 entre-grupos, univariado, *F* (1, 1662) = 38.37, $p < .001$, $\eta^2 = .02$). El supuesto de homogeneidad o igualdad de las varianzas de las puntuaciones de los dos grupos (padres y madres) se cumple (Levene *F* (1, 1662) = 0.67, *p* = .41).

En la redacción anterior hay que tener en cuenta que cuando se redactan los resultados siempre debe informarse de los estadísticos descriptivos de media, desviación típica y número de observaciones de cada grupo (*n*).

Además, se redacta el resultado de la prueba *F* del ANOVA con toda la información: grados de libertad 'entre' o de la fuente de varianza del efecto, grados de libertad 'intra-celdilla' o de la fuente de varianza del término de error, valor obtenido del estadístico, valor exacto de *p* y valor del tamaño del efecto (eta cuadrado por ejemplo; y si es un diseño con solo dos grupos (A = 2) se puede anotar el valor de la diferencia estandarizada de medias conocida como tamaño del efecto *d* de Cohen).

Si los resultados del valor de *p* que ofrece el programa estadístico pone .000 nunca se debe anotar .000 en la redacción. Se redactaría $p < .001$ ya que lo que indica .000 es que se trata de un valor muy pequeño, es decir, que es menor a .001.

Redacción 2. Se cumple el supuesto de homogeneidad de las varianzas y se ofrece una tabla de descriptivos.

A continuación se redacta un ejemplo de análisis de la diferencia entre las puntuaciones medias de dos grupos, se cumple el supuesto de hogeneidad de las varianzas, los estadísticos descriptivos se amplían y se ofrecen en una tabla.

Los resultados del análisis de la varianza (ANOVA) entre-grupos unifactorial univariado señala que hay una diferencia estadísticamente significativa entre las medias del grupo de payasos y el grupo de

control sin payasos en la variable de ansiedad, siendo el tamaño del efecto muy grande, F (1, 10) = 7.62, p = .02, η^2 = .43. Se ha comprobado el supuesto de homogeneidad de las varianzas de los dos grupos, Levene F (1, 10) = 0.28, p = .609. Por tanto, el grupo de payasos obtiene la puntuación media de ansiedad más baja respecto al grupo de comparación que no recibe ningún tipo de intervención. En términos de diferencia estandarizada de medias, el tamaño del efecto de *d* Cohen = 1.59 (95% IC -2.89 a -0.29), es decir, un tamaño del efecto muy grande con una amplitud muy amplia del intervalo de confianza ya que oscila desde un tamaño del efecto pequeño a un tamaño del efecto muy grande. Los resultados descriptivos se presentan en la Tabla 10.

Tabla 10. Análisis descriptivo de la variable ansiedad

	Grupo de payasos	Grupo de control
n	6	6
Media	4.50	13.50
DT	4.85	6.35
Amplitud	13	17
Mínimo	1	1
Máximo	14	18

En la redacción anterior hay que tener en cuenta que cuando los descriptivos de la variable medida (variable dependiente) se detallan en una tabla ya no deben mencionarse en el texto, pues no hay que repetir la información en el texto y en la tabla.

Y las tablas que ofrece el SPSS tiene que elaborarlas el usuario y nunca copiar y pegar una tabla del SPSS o de otro programa estadístico en la redacción de los resultados.

Las tablas con formato del Manual de APA no tienen líneas verticales y solamente se ponen las líneas horizontales en la primera fila en la parte superior e inferior y en la última fila solamente en la parte inferior.

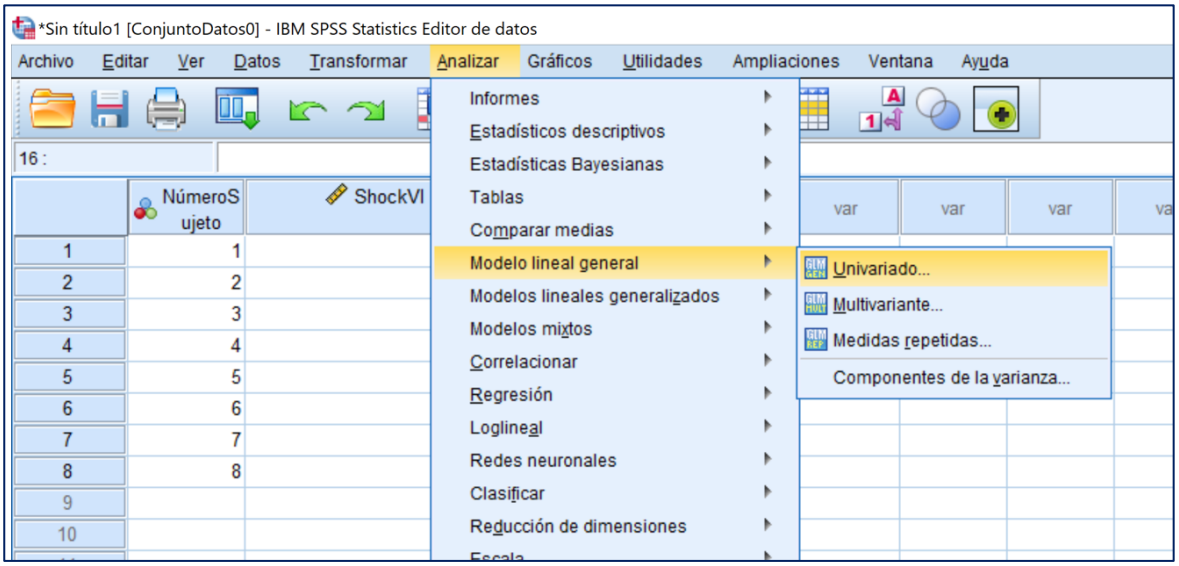
Solución con el SPSS. Diseño entre-sujetos unifactorial univariado

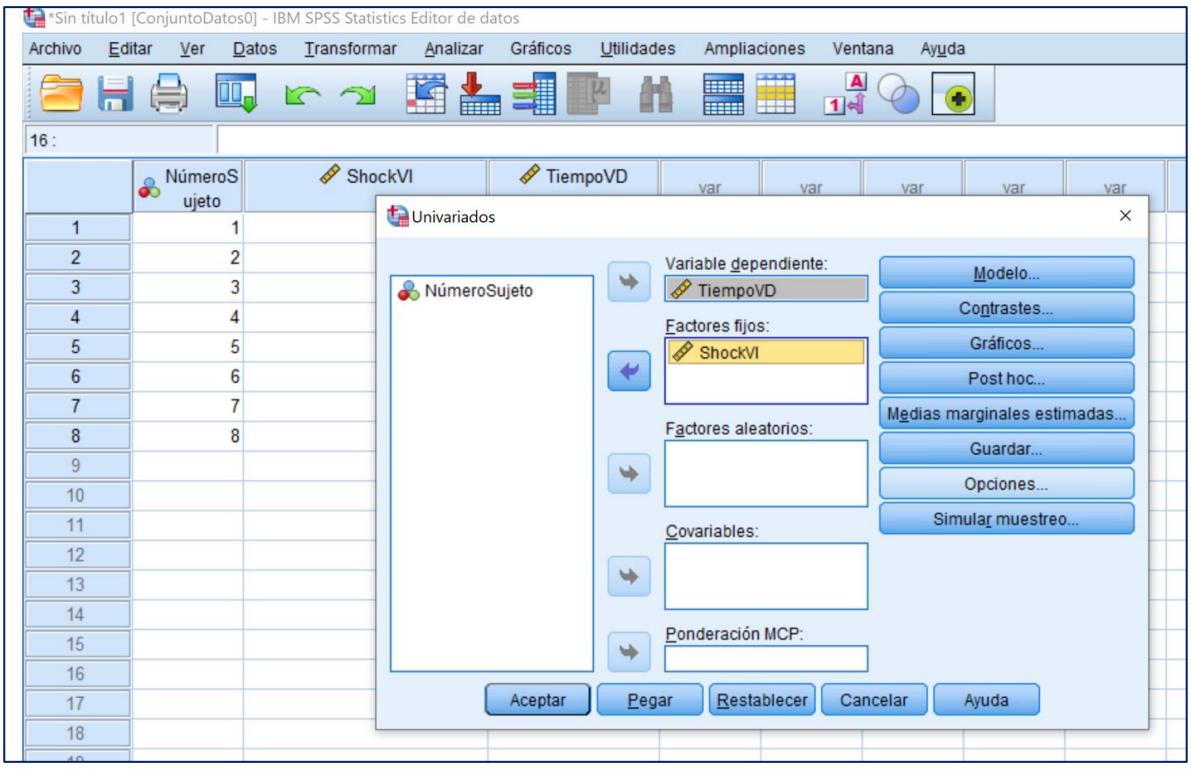
1. Introducción de datos

	NúmeroSujeto	ShockVI	TiempoVD
1	1	1	23
2	2	1	11
3	3	1	12
4	4	1	26
5	5	2	39
6	6	2	38
7	7	2	23
8	8	2	28

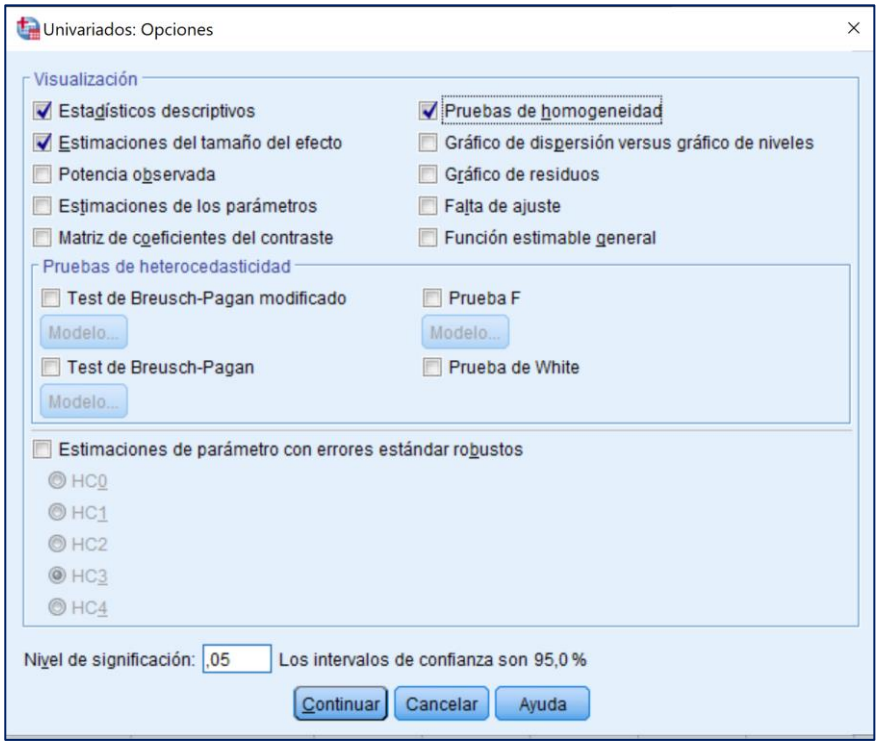
	NúmeroSujeto	ShockVI	TiempoVD
1	1	Escapable	23
2	2	Escapable	11
3	3	Escapable	12
4	4	Escapable	26
5	5	No escapable	39
6	6	No escapable	38
7	7	No escapable	23
8	8	No escapable	28
9			

2. Modelo Lineal General





3. Opciones



RESULTADOS

Análisis univariado de varianza

[ConjuntoDatos0]

Factores inter-sujetos

		Etiqueta de valor	N
ShockVI	1	Escapable	4
	2	No escapable	4

Estadísticos descriptivos

Variable dependiente: TiempoVD

ShockVI	Media	Desv. Desviación	N
Escapable	18,00	7,616	4
No escapable	32,00	7,789	4
Total	25,00	10,337	8

Prueba de igualdad de Levene de varianzas de error[a,b]

		Estadístico de Levene	gl1	gl2	Sig.
TiempoVD	Se basa en la media	,000	1	6	1,000
	Se basa en la mediana	,000	1	6	1,000
	Se basa en la mediana y con gl ajustado	,000	1	4,883	1,000
	Se basa en la media recortada	,000	1	6	1,000

Prueba la hipótesis nula de que la varianza de error de la variable dependiente es igual entre grupos.

a. Variable dependiente: TiempoVD

b. Diseño : Intersección + ShockVI

Pruebas de efectos inter-sujetos

Variable dependiente: TiempoVD

Origen	Tipo III de suma de cuadrados	gl	Media cuadrática	F	Sig.	Eta parcial al cuadrado
Modelo corregido	392,000[a]	1	392,000	6,607	,042	,524
Intersección	5000,000	1	5000,000	84,270	,000	,934
ShockVI	392,000	1	392,000	6,607	,042	,524
Error	356,000	6	59,333			
Total	5748,000	8				
Total corregido	748,000	7				

a. R al cuadrado = ,524 (R al cuadrado ajustada = ,445)

Capítulo 11. Comprobación de hipótesis específicas (diseño entre grupos *A* > 2)

Dolores Frías-Navarro*
Marcos Pascual-Soler**
*Universidad de Valencia
**ESIC Business & Marketing School, España

Índice

Citar el capítulo como:

Frías-Navarro, D. y Pascual-Soler, M. (2021). Comprobación de hipótesis específicas (diseño entre grupos *A* > 2). En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

Hasta ahora se ha presentado la prueba paramétrica para contrastar hipótesis de dos grupos independientes (diseño entre-grupos unifactorial A = 2 univariado) mediante el Análisis de la Varianza (ANOVA). Ahora se va a detallar el contraste de hipótesis específicas, también conocido como pruebas post hoc o comparaciones a posteriori. Se van a presentar diferentes pruebas estadísticas y se van a utilizar dos módulos del SPSS donde se pueden realizar las pruebas de hipótesis específicas en la ventana de 'Analizar' y luego se opta por el 1) módulo de 'Comparar medias' o por el 2) módulo de 'Modelo Lineal General'. Por lo tanto, el análisis de hipótesis específicas se puede realizar con:

ANALIZAR → Comparar medias

O con:

ANALIZAR → Modelo Lineal General

ANALIZAR → Modelo lineal general → univariado

SPSS. ANALIZAR → Comparar medias

El SPSS tiene un grupo de análisis dentro del menú Analizar → Comparar Medias donde se pueden ejecutar las técnicas de contraste de hipótesis de prueba *t* para una muestra, prueba *t* para muestras independientes, prueba *t* para muestras relacionadas y ANOVA de un factor.

Nos vamos a detener en el último apartado: ANOVA de un factor (figura 48). Dentro de este apartado se puede ejecutar un diseño de A = 2 y de A > 2. Es decir, se pueden ejecutar los diseños unifactoriales con un factor de 2 condiciones o con un factor que tenga más de dos condiciones (por ejemplo A = 3).

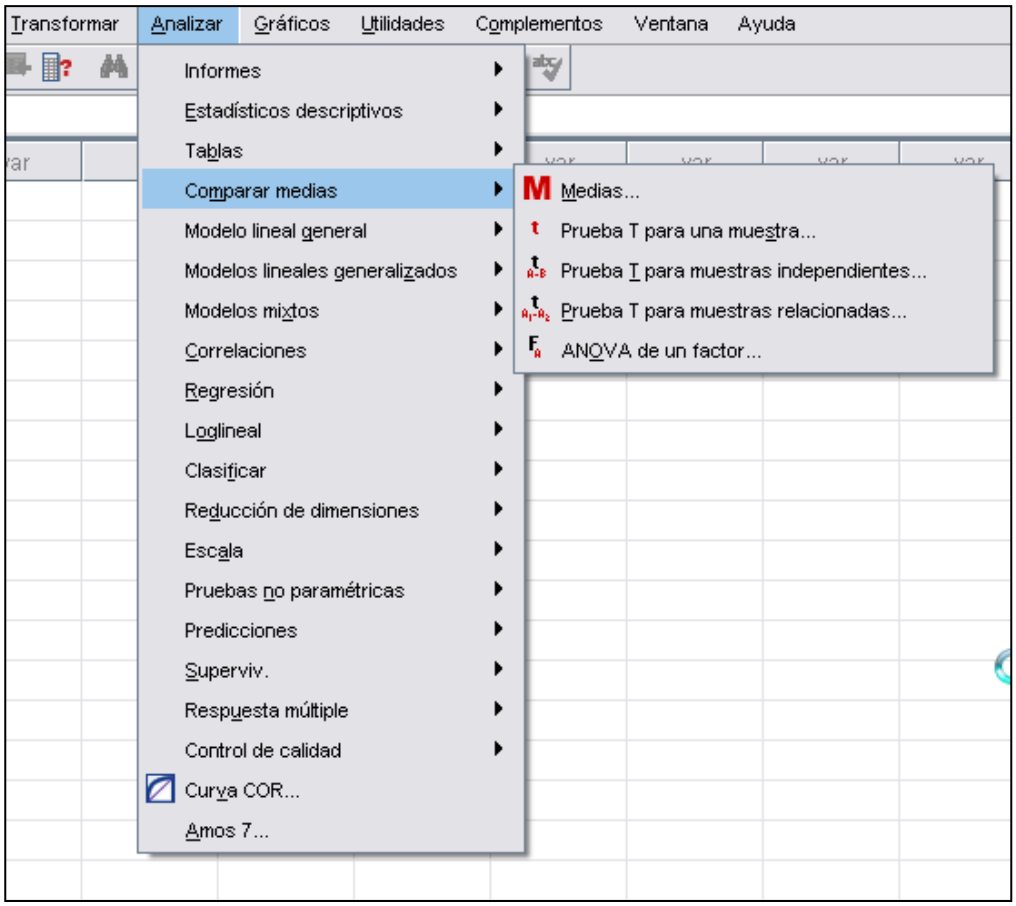


Figura 48. SPSS: Analizar-Comparar medias-ANOVA de un factor

Cuando el diseño tiene un factor o variable independiente con más de dos condiciones será necesario recurrir a la opción de 'post-hoc' (a posteriori) para poder descubrir entre qué pares de medias se encuentran las diferencias (figura 49).

Y para ello es necesario que el profesional (investigador o investigadora) decida qué prueba debe ejecutar a posteriori, teniendo en cuenta si se asume que las varianzas de los grupos son homogéneas ("homocedasticidad de varianzas") o no lo son ("heterocedasticidad de varianzas").

Figura 49. SPSS: Analizar-Comparar medias-ANOVA de un factor

Por lo tanto, cuando la variable independiente tiene más de 2 condiciones, hay que analizar entre qué pares de medias se producen las diferencias estadísticamente significativas y en qué sentido.

Supuesto de investigación

Supongamos que una investigación tiene un diseño entre-grupos unifactorial univariado ortogonal, A = 3 (nota: si en el texto se señala que $n_a = 3$, se interpreta que es un diseño ortogonal y por lo tanto $N = 9$, tres observaciones por 3 condiciones), y tiene los siguientes datos:

-a_1= 12, 8, 10

-a_2= 5, 7, 6

-a_3= 14, 13, 15

El investigador o investigadora desea comprobar si existen diferencias estadísticamente significativas entre esos tres grupos (en general, si hay alguna diferencia entre las medias que es estadísticamente significativa) y también desea conocer entré qué par de medias o pares de medias se encuentran las diferencias que son estadísticamente significativas.

Se lleva a cabo la introducción de datos para este tipo de diseño en el SPSS y, posteriormente, se ejecutan las instrucciones del modelo de diseño que se desea analizar (figura 50).

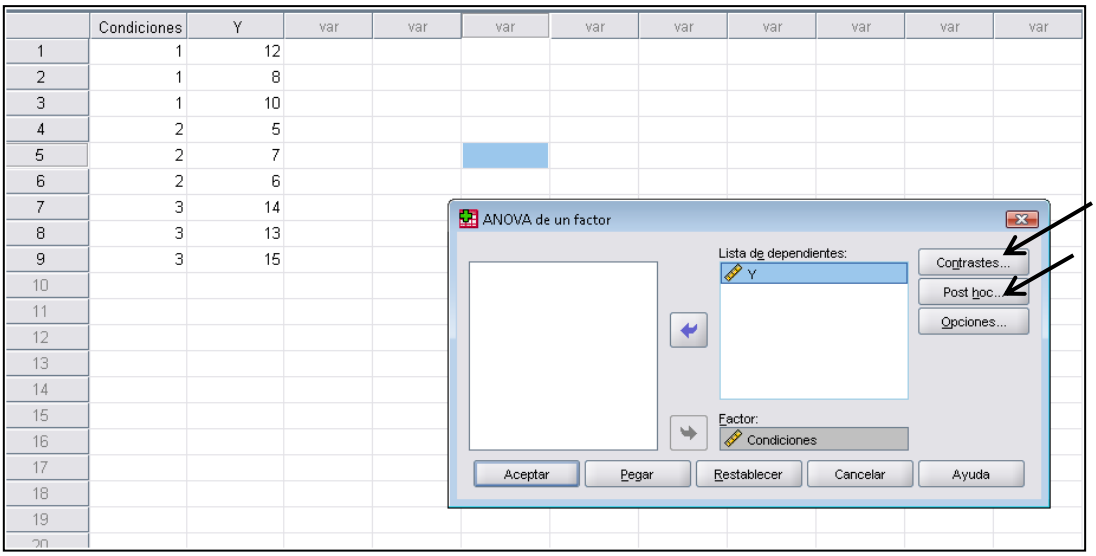


Figura 50. SPSS: Introducción de datos para un diseño entre-grupos y selección de la variable dependiente e independiente del diseño

Si se ejecuta el 'ANOVA de un factor' y en 'Opciones' se selecciona 'descriptivos' y 'prueba de homogeneidad' de las varianzas, el resultado señala la siguiente relación entre las medias de los grupos mediante la razón *F* del Análisis de la Varianza (figura 51). El lector o lectora debe responder a la pregunta de si en el resultado del ANOVA ¿existe *alguna diferencia estadísticamente significativa* entre las medias de las tres condiciones?

A continuación se describe la información que proporciona el ANOVA y las pruebas de hipótesis específicas.

ANOVA de un factor

[Conjunto_de_datos0]

Descriptivos

Y

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	3	10,00	2,000	1,155	5,03	14,97	8	12
2	3	6,00	1,000	,577	3,52	8,48	5	7
3	3	14,00	1,000	,577	11,52	16,48	13	15
Total	9	10,00	3,674	1,225	7,18	12,82	5	15

Prueba de homogeneidad de varianzas

Y

Estadístico de Levene	gl1	gl2	Sig.
,667	2	6	,548

ANOVA

Y

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	96,000	2	48,000	24,000	,001
Intra-grupos	12,000	6	2,000		
Total	108,000	8			

Figura 51. SPSS: resultados del ANOVA y homogeneidad de varianzas

Los resultados de la prueba de homogeneidad de varianzas de Levene señalan que las varianzas de las puntuaciones de las tres condiciones son homogéneas (Levene $F_{(2,\ 6)} = 0.667$, $p = .548$). Por lo tanto se cumple el supuesto de homogeneidad de las varianzas ya que el valor de *p* es mayor al valor de alfa que se sitúa a priori en .05.

Los resultados del modelo de ANOVA de un diseño entre-sujetos unifactorial A = 3 univariado señalan que existen diferencias estadísticamente significativas entre los tres grupos ($F_{(2, 6)} = 24$, $p = .001$).

Pero, cuando A > 2, la pregunta es ¿dónde se encuentran las diferencias que son estadísticamente significativas? ¿Es entre las medias de a1 - a2? ¿Es entre las medias de a1 - a3? O quizás ¿es entre las medias de a3 - a2? Las diferencias siempre se valoran en términos de valores absolutos.

El resultado del ANOVA no permite dar respuesta a esas comparaciones específicas entre los tres grupos ya que solo ofrece información de si globalmente hay alguna diferencia de medias que es estadísticamente significativa (se trata de

una prueba 'omnibus'). Cuando únicamente se trabaja con dos medias (A = 2) estaba claro que si el resultado del ANOVA señalaba que habían diferencias estadísticamente entre los grupos se trataba de la única diferencia posible: a1 - a2 (o lo que es lo mismo, a2 - a1). Pero ahora hay tres posibles diferencias de medias. ¿Cuál de ellas es estadísticamente significativa? ¿Son todas? ¿Es una? ¿Son dos? En las pruebas post-hoc de comparación de medias (también conocidas como pruebas a posteriori y pruebas de hipótesis específicas) se encuentra la solución a esos interrogantes.

Antes de continuar conviene recordar cómo se desarrolla la ecuación estructural para el modelo planteado en el estudio con A=3:

$$Y = M + A + E$$

Las puntuaciones en la variable dependiente Y son igual a:

La constante (media general) + el efecto del tratamiento (Efecto A = Ma - M) + Error (Error = Y - M - Efectos, es decir, Y - M - A en este tipo de diseño).

Además, se puede calcular:

- Los efectos estimados para el diseño planteado son $\alpha_1 = 0$ y $\alpha_2 = -4$ (luego $\alpha 3 = 4$ ya que la suma de los efectos es 0)
- Y los errores son en $a_1 = 2$ y -2, en $a_2 = -1$ y 1 y en $a_3 = 0$ y -1 (luego los errores que faltan son 0 en el grupo 1, en el grupo 2 y 1 también es 0 y en el grupo 3 es 1 ya que los errores deben sumar cero dentro de cada grupo y, por supuesto, también suman cero cuando se suman todos sus valores).

Como ejercicio, los lectores y lectoras pueden completar la tabla de efectos y los errores y desarrollar el modelo de diseño propuesto mediante la descomposición de la ecuación estructural. Se debe comprobar que la suma de los efectos del tratamiento es cero así como la del error. Elevar al cuadrado y sumar (Suma de Cuadrados, SC). Dividir cada suma de cuadrados por sus grados de libertad (gl) (Medias Cuadráticas, MC) y ejecutar la prueba *F* del Análisis de la Varianza (*F* = MCefecto / MCerror). Después, se debe utilizar las tablas de la razón *F* y buscar el valor teórico que corresponde a *F* (.05, 2, 6). A continuación, se comparan los dos

valores de *F* y si el valor de la *F* empírica del estudio es mayor o igual que el valor de la *F* teórica de las tablas entonces se puede rechazar la hipótesis nula (es decir, se rechaza la hipótesis nula cuando $F(2, 6) \geq F(.05, 2, 6)$. Cuando se utilizan las tablas sólo se puede saber si el valor *p* de probabilidad del resultado de la prueba estadística es superior o menor al valor de alfa. Por eso, cuando se utilizan las tablas no se puede dar el valor exacto de probabilidad del resultado del estadístico, pero en nuestro ejemplo sí se puede saber que $p < .05$ pues la *F* empírica es mayor que la *F* teórica. Por lo tanto, se puede completar el proceso de decisión estadística tal y como se efectúa en el contraste de hipótesis.

Cuando se trabaja con un programa estadístico (como por ejemplo SPSS; JASP, JAMOVI, SAS...) ya no hace falta recurrir a las tablas de la distribución del estadístico. El mismo programa añade el valor exacto de probabilidad (valor *p*, a veces se señala como 'significación') junto al resultado de la prueba estadística. El investigador o investigadora tiene que completar el proceso de contraste estadístico comparando el valor de *p* de probabilidad del resultado obtenido (o más extremo) con la prueba estadística con el valor del alfa fijado a priori en su investigación. Si el valor de $p \leq$ alfa entonces se puede rechazar la hipótesis nula. Siempre que se trabaja con un programa estadístico hay que redactar los valores *p* de probabilidad exactos incluso cuando las decisiones suponen mantener la hipótesis nula ($p >$ alfa). No es nada recomendable situar 'ns' para indicar un resultado no estadísticamente significativo. Recordemos que el valor *p* de probabilidad depende del valor del efecto y del tamaño de la muestra. Así, manteniendo constante el efecto, si se aumenta la muestra disminuye el valor de probabilidad. Del mismo modo, si se disminuye el tamaño de la muestra entonces aumenta el valor *p* de probabilidad.

Por lo tanto, cuando el Análisis de la Varianza (ANOVA) ofrece un resultado estadísticamente significativo está indicando que al menos un grupo difiere de los otros grupos. EL ANOVA es un 'test omnibus', es decir, no informa del patrón de diferencias entre las medias, sólo facilita que hay (o no hay) alguna diferencia estadísticamente significativa entre las medias de los grupos, pero sin señalar dónde está la diferencia o las diferencias. Es decir, el ANOVA solamente detecta la presencia o ausencia de un efecto global de la variable independiente sobre la variable dependiente. Pero el investigador o investigadora desea conocer dónde se encuentran las diferencias estadísticamente significativas que detectó de forma

global el ANOVA. Para analizar el patrón de diferencias entre las medias es necesario ejecutar pruebas de hipótesis específicas y las más utilizadas implican comparar pares de medias, de ahí el nombre de comparaciones por parejas (‘pairwise comparisons’). En el SPSS se conocen como ‘pruebas post hoc’.

Siguiendo con el ejercicio, una vez que se ha comprobado que existe un efecto global en el Análisis de la Varianza tal y como se ha detallado en la salida del SPSS ($F_{(2, 6)}$ = 24, p = .001), se puede abordar, a continuación, entre qué pares de medias se encuentran las diferencias estadísticamente significativas. Para ello, se puede elaborar la tabla de diferencias de medias entre los grupos o condiciones (tabla 11).

Tabla 11. Diferencias de medias para un diseño A = 3

Medias	*a*1 Media = 10	*a*2 Media = 6
Media *a*2 = 6	4	-
Media *a*3 = 14	4	8

Con un diseño A = 3 existen 3 diferencias de medias entre los 3 grupos. Para conocer cuántas diferencias de medias simples existente en un diseño se puede aplicar la siguiente fórmula:

$$C = \frac{a\,(a-1)}{2}$$

Donde C es el número de diferencias de medias simples entre los grupos no redundantes y *a* es el número de condiciones que tiene la fuente de varianza que se analiza en el modelo del diseño. En el ejemplo, si a = 3 entonces: C = 3. Es decir, hay 3 comparaciones simples entre los pares de medias, a1-a2, a1-a3 y a2-a3.

Entonces, ¿cuál de las tres diferencias de medias es estadísticamente significativa? Para responder a dicha cuestión se necesita ejecutar una prueba de contraste de hipótesis específicas para comparar las diferencias de medias entre las condiciones experimentales.

Una solución que se podría pensar es ejecutar pruebas *t* de Student dos a dos o ANOVAs para cada par de medias ejecutando en este caso 3 ANOVAs. Pero esta alternativa no se considera válida ya que está sujeta a una grave deficiencia metodológica como es aumentar la probabilidad del error de Tipo I. Es decir, la tasa de error de Tipo I por experimento deja de ser la planteada con el valor del alfa prefijado cuando se planificó el estudio. Para poder realizar esos contrastes estadísticos correctamente es necesario controlar el alfa por comparación. Veamos en primer lugar qué es la Tasa de Error de Tipo I y en segundo lugar cómo controlar dicha tasa de error mediante los procedimientos de pruebas de hipótesis específicas.

Tasa de error de tipo I

La Tasa de Error de Tipo I o alfa por experimento (α_{PE}) es:

$$\alpha_{PE} = 1 - (1 - \alpha_{PC})^{C}$$

Donde alfa por comparación, αPC, es la probabilidad de cometer un error de tipo I en 'una comparación' y el alfa por experimento, αPE, es la probabilidad de cometer al menos un error de Tipo I en un 'conjunto de comparaciones' (C).

Supongamos que en un estudio se van a realizar 4 comparaciones o pruebas de contraste de hipótesis. Si todas las hipótesis nulas fueran ciertas y αPC = .05 entonces la probabilidad de cometer al menos un Error de Tipo I es:

$$\alpha_{PE} = 1 - (1 - .05)^{4} = .1855$$

Ese valor de .1855 se aparta mucho del valor de alfa fijado a priori por el investigador o investigadora en .05. Es decir, se tendría una probabilidad de error de Tipo I de .1855, que es totalmente inadmisible. La probabilidad de rechazar una hipótesis nula siendo cierta es de .1855 (18.5%). ¿Qué ha pasado? La prueba de contraste de hipótesis se ha hecho más liberal (es decir, se rechaza la hipótesis nula con mayor facilidad), aumentando de este modo el error de Tipo I. ¿Cómo se puede corregir ese sesgo que amenaza la validez de conclusión estadística de los resultados del estudio? La opción es ejecutar un procedimiento que haga a la prueba estadística más conservadora para equilibrar ese aumento del error de Tipo I o probabilidad de rechazar la hipótesis nula siendo realmente cierta.

Todos los procedimientos de contraste de hipótesis específicas se basan en hacer el contraste más conservador. Es decir, se reduce el αPC (alfa por comparación) para poder controlar el αPE (alfa por experimento) que no debe superar el valor de .05. En definitiva, la prueba se hace más conservadora.

En este punto el investigador o investigadora debe tomar una decisión de nuevo: tiene que elegir la prueba de contraste de hipótesis específicas que controle correctamente la tasa de error de Tipo I y además que la potencia estadística sea máxima (menor error de Tipo II). Es decir, en un mismo diseño se podría optar por varias pruebas de contraste de hipótesis específicas y el investigador o investigadora debe seleccionar la más adecuada para que la validez de conclusión estadística sea la más óptima. La más adecuada es aquella que controla la tasa de error de tipo I y fija el alfa por experimento en .05 y al mismo tiempo es la prueba que facilita la mayor potencia estadística para de este modo reducir el error de Tipo II.

Por lo tanto, para tomar la decisión el investigador o investigadora debe considerar los siguientes aspectos:

1. El número de comparaciones (*C*) que la hipótesis plantea: pueden ser comparaciones exhaustivas (*a posteriori*) o pueden ser comparaciones planificadas (*a priori*).
2. Si las hipótesis experimentales son simples (plantea diferencias *entre pares de medias*) o complejas (plantea alguna diferencia entre medias que implica *un promedio de medias*).

Comparaciones exhaustivas o a posteriori. El contraste de hipótesis específicas es exhaustivo cuando se realizan todas las comparaciones posibles entre los padres de medias que tiene el diseño de investigación. Por ejemplo, si A = 3 entonces el número de todos los pares posibles de diferencias de medias es igual a 3. Si se analizan las tres diferencias simples de medias entonces se considera que se han realizado comparaciones exhaustivas.

Comparaciones planificadas o a priori. Cuando el número de comparaciones que hay que contrastar es más reducido (no se realizan de forma exhaustiva todas las comparaciones simples), el contraste se denomina contraste planificado o contraste a priori. Por ejemplo, si sólo se desean comparar a1 - a2 y a1 - a3 entonces

ya no es exhaustivo sino planificado dado que por diversas cuestiones teóricas no interesa analizar la comparación entre a2 y a3.

Para poder plantear contrastes a priori es necesario fundamentar esas opciones de contraste en unas hipótesis teóricas que den sentido a la elección de los contrastes o análisis que se quieren realizar. Por ejemplo, en un diseño con un grupo de control y dos grupos de tratamiento (A = 3) podría ser interesantes comparar el grupo de control con uno de los grupos de tratamiento (contraste uno) y el grupo de control con el otro grupo de tratamiento (contraste dos). En este caso el investigador o investigadora ha planteado dos contrastes o pruebas consideradas a priori, es decir, su planteamiento es previo a cualquier tipo de resultado del estudio. Los contrastes se plantean antes de ejecutar cualquier análisis y necesitan de una teoría que los fundamente.

Las **hipótesis son simples** cuando se plantean diferencias entre pares de medias simples, por ejemplo la diferencia entre las medias de los grupos a1 - a2. Se considera que las **hipótesis son complejas** cuando la diferencia de medias implica un promedio de medias, por ejemplo se desea comparar la media de a1 frente a la media compleja de dos grupos. Por ejemplo una media compleja sería (a2 + a3) / 2. Es decir, en el cómputo de la media han intervenido las medias de dos grupos. Al final se contrastan dos medias, pero al menos una de ellas es compleja. Otro ejemplo de hipótesis complejas podría ser la comparación siguiente: la media de (a2 + a3) / 2 respecto a la media (a1 + a3) / 2.

Pruebas de contraste de hipótesis específicas

Existen diferentes pruebas de contraste de hipótesis específicas. En la figura 52 se representan las situaciones en las que se podrían aplicar las *pruebas de hipótesis específicas* más utilizadas:

- **DHS Tukey**: hipótesis exhaustivas y simples.
- **Dunnett**: hipótesis planificadas y simples.
- **Bonferroni**: hipótesis planificadas, exhaustivas, simples y complejas.
- **Scheffé**: hipótesis planificadas, exhaustivas, simples y complejas.

	Planificada	Exhaustiva	Simple	Compleja
DHS Tukey		x	X	
Dunnett	x		X	
Bonferroni	X		X	x
Scheffé	x	x	x	x

Figura 52. Diferentes pruebas de contraste de hipótesis específicas

Como se observa, hay situaciones donde se pueden aplicar varias pruebas de hipótesis específicas, pero dependiendo de la situación hay unas pruebas que tienen más potencia estadística (menor error de Tipo II). Recordar que el investigador o investigadora tiene que seleccionar aquella prueba que además de controlar el error de Tipo I también sea la que tiene más potencia estadística para detectar el efecto (menor error de Tipo II).

A continuación se detallan las situaciones de investigación donde es más apropiada una prueba estadística u otra, dado que controlan la probabilidad del error de Tipo I en el nivel de alfa planteado por el investigador o investigadora a priori y tienen el menor error de Tipo II.

Procedimiento DHS (Honestly Significant Difference) de Tukey

El procedimiento DHS de Tukey es el más potente cuando en el diseño se ejecutan:

1) todas las comparaciones posibles entre las medias (exhaustivo) y además
2) son comparaciones simples

Es decir, es la prueba más adecuada cuando se comparan dos medias simples cada vez y se ejecutan todas las posibles comparaciones simples del diseño.

El procedimiento calcula el denominado ‘rango crítico’ (RC) o valor teórico de la prueba y se compara con cada diferencia empírica de medias simple que tenga el diseño.

Si el valor de la diferencia de medias supera o iguala ese valor de rango crítico entonces se rechaza la hipótesis nula. Es decir, la diferencia entre el par de medias es estadísticamente significativa (p < alfa) si dicha diferencia iguala o supera al valor de rango crítico que ofrece la prueba DHS de Tukey.

Los diferentes procedimientos de contrastes de hipótesis específicas se basan en la estimación de un valor de Rango Crítico obtenido a partir de una distribución concreta. En concreto, la prueba DHS de Tukey está basada en la distribución del rango estudentizado $q_{(\alpha,\ a,\ glError)}$, donde q es el valor de la distribución de rango estudentizado basado en el alfa, el número de medias a contrastar y los grados de libertad del error.

En el ejemplo, el valor del Rango Crítico es igual a 3.543. Por lo tanto, cualquier diferencia de medias que supere o iguale ese valor teórico (en valores absolutos) es estadísticamente significativa según la prueba de Tukey. En el ejemplo las tres diferencias de medias superan el Rango Crítico de Tukey y, por lo tanto, se rechaza la hipótesis nula en los tres contrastes estadísticos de diferencias de medias.

Procedimiento de Dunnett

El procedimiento de Dunnett es el más potente cuando se trata de comparar:

1) la media de un grupo frente al resto de medias que tenga el diseño, es decir, se realizan a-1 comparaciones y además
2) son comparaciones simples.

No es un procedimiento exhaustivo sino que plantea hipótesis planificadas o a priori. También se calcula el rango crítico (RC) de Dunnett y se procede del mismo modo comparándolo con las diferencias de medias empíricas de la investigación.

La prueba de Dunnett utiliza la distribución basada en la comparación con un grupo de control (distribución de Dunnett). En la distribución de Dunnett se obtiene el valor de D en función del alfa, el número de grupos y los grados de libertad del error $D_{(\alpha,\ a,\ glError)}$. De nuevo, cualquier diferencia de medias (en valores absolutos) que iguale o supere dicho valor crítico es estadísticamente significativa.

Por lo tanto, en aquellas situaciones de investigación donde la prueba más adecuada es la de Dunnett (donde se realizan a-1 comparaciones simples) entonces

será necesario indicar en el SPSS qué grupo es el de comparación (categoría de control) para que efectúe las *a*-1 comparaciones respecto a un determinado grupo. Por ejemplo, si la categoría de control se llama 'Primero' entonces eso significa que se va a comparar la media del primer grupo con las medias del resto de condiciones que tenga el diseño (figura 53). En un diseño con A = 3 las comparaciones que efectúa la prueba de Dunnett serían: a_1 - a_2 y a_1 - a_3.

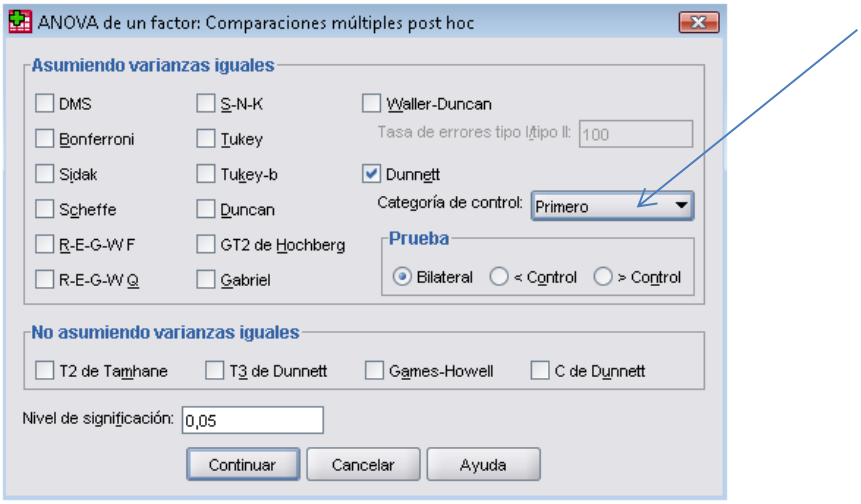


Figura 53. SPSS: selección de la prueba de Dunnett

Suponiendo que el investigador o investigadora desea realizar *a* - 1 comparaciones simples (3 - 1 = 2 contrastes) entonces la prueba de Dunnett ofrece el siguiente resultado en el SPSS (figura 54). Como se observa, sólo aparecen dos contrastes: la media del grupo a2 con la media del grupo a1 (la diferencia de medias es -4) y la media del grupo a3 con la media del grupo a1 (diferencia de medias de 4). Los resultados que ofrece el SPSS para poder tomar la decisión estadística de si se puede o no rechazar la hipótesis nula es el valor de significación (Sig.), es decir, el valor *p* de probabilidad del estadístico de comparación.

Pruebas post hoc

Comparaciones múltiples

Y
t de Dunnett (bilateral)[a]

(I) Condiciones	(J) Condiciones	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
2	1	-4,000*	1,155	,024	-7,31	-,69
3	1	4,000*	1,155	,024	,69	7,31

a. Las pruebas t de Dunnett tratan un grupo como control y lo comparan con todos los demás grupos.

*. La diferencia de medias es significativa al nivel 0.05.

Figura 54. SPSS: resultados de la prueba de Dunnett

Corrección de Bonferroni

La corrección de Bonferroni es el procedimiento más potente siempre que la hipótesis formule el número de comparaciones a priori aunque si C (número de comparaciones) es grande entonces la prueba es poco potente y no se recomienda su uso.

Se puede aplicar con hipótesis simples y con hipótesis complejas.

El procedimiento consiste en aplicar en cada comparación o prueba estadística el siguiente valor de alfa:

Alfa= α_{PE} que se desea en el experimento / Número de comparaciones (C)

$$\alpha_{PC} = \frac{\alpha_{PE}}{C}$$

Por ejemplo si se formulan 4 comparaciones o contrastes a priori, el α_{PE} final se mantendrá en .05 si en cada comparación individual se utiliza un alfa por comparación igual a .0125:

$$\alpha_{PC} = \frac{.05}{4} = .0125$$

Si se ejecuta el procedimiento de Bonferroni se puede comprobar que $\alpha_{PE} = 1 - (1- 0.0125)^4 = .049$. Valor cercano al alfa de.05 fijado a priori por el investigador.

El procedimiento de Bonferroni se basa en la distribución de la prueba *F:* $F_{(\alpha/C,\ 1,\ glError)}$, donde C representa el número de comparaciones y a/C es el alfa que se va a utilizar en cada comparación y por lo tanto la tabla que hay que consultar.

Procedimiento de Scheffé

El procedimiento de Scheffé es válido en cualquier circunstancia de investigación, pero normalmente es la prueba menos potente.

Se calcula el rango crítico (RC) de Scheffé y se procede del mismo modo comparándolo con las diferencias de medias empíricas. Si el valor de la diferencia de medias iguala o supera al valor del Rango Crítico entonces se rechaza la hipótesis nula.

La prueba de Scheffé está basada en la distribución de la prueba *F*: $F_{(\alpha,\ a-1,\ glError)}$.

En la tabla 12 se detalla 'el máximo' número de contrastes (comparaciones) que deberían probarse en una investigación con el procedimiento de Bonferroni. Si el número de comparaciones es mayor entonces el procedimiento de Bonferroni pierde potencia estadística y sería más conveniente optar por la prueba de Scheffé que en esas circunstancias es más potente que Bonferroni.

Tabla 12. Máximo número de contrastes que deberían probarse en un estudio con el procedimiento de Bonferroni

	Número de grupos							
gl_{error}	3	4	5	6	7	8	9	10
5	2	4	8	12	17	24	31	40
6	2	5	9	14	21	30	41	55
7	2	5	10	16	25	37	52	71
8	2	6	11	18	29	44	64	89
9	2	6	12	20	33	51	75	107
10	2	6	12	22	37	58	87	127
12	3	7	13	25	43	70	110	166
14	3	7	14	28	49	82	132	205
16	3	7	15	30	54	93	153	243
18	3	7	16	32	58	103	173	281
20	3	7	17	33	63	112	191	316
30	3	8	18	39	78	147	267	470
40	3	8	20	43	87	170	320	586
50	3	8	20	45	94	187	360	674
60	3	8	21	47	98	199	390	743
70	3	9	21	48	102	209	414	799
80	3	9	21	49	105	217	433	844
90	3	9	22	50	107	223	449	882
100	3	9	22	50	109	228	462	913
110	3	9	22	51	111	232	473	941
120	3	9	22	51	112	236	483	964

Siguiendo con el ejemplo anterior se pueden calcular las diferencias entre los pares de medias simples (ver los datos presentados anteriormente).

¿Qué prueba de contraste de hipótesis específicas será la más adecuada para contrastar todas las diferencias entre los pares de medias? Como ya se ha comentado, la más adecuada es aquella que controla el error de Tipo I en el nivel fijado previamente por el alfa y además es la prueba con mayor potencia estadística (menor error de Tipo II). Por lo tanto, si se trata de un diseño donde se plantean todas las comparaciones de medias dos a dos (comparaciones exhaustivas y simples) entonces la prueba más potente es DHS de Tukey. La prueba DHS (diferencia honestamente significativa) de Tukey es la elección más correcta desde el punto de vista de la validez de conclusión estadística.

A continuación se detalla la salida de resultados que ofrece el programa SPSS (figura 55).

Análisis de varianza univariante

[Conjunto_de_datos1]

Factores inter-sujetos

		Etiqueta del valor	N
A	1	Fármaco A	4
	2	Fármaco B	4
	3	Fármaco C	4

Estadísticos descriptivos

Variable dependiente:Y

A	Media	Desviación típica	N
Fármaco A	26,00	1,826	4
Fármaco B	25,00	3,162	4
Fármaco C	30,00	1,826	4
Total	27,00	3,104	12

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente:Y

F	gl1	gl2	Sig.
1,714	2	9	,234

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intersección + A

Pruebas de los efectos inter-sujetos

Variable dependiente:Y

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Modelo corregido	56,000[a]	2	28,000	5,040	,034	,528
Intersección	8748,000	1	8748,000	1574,640	,000	,994
A	56,000	2	28,000	5,040	,034	,528
Error	50,000	9	5,556			
Total	8854,000	12				
Total corregida	106,000	11				

a. R cuadrado = ,528 (R cuadrado corregida = ,423)

Figura 55. SPSS: resultados del ANOVA

Si se ejecuta el contraste de hipótesis mediante la prueba DHS de Tukey con el SPSS se seleccionará la Opción de Tukey (figura 56).

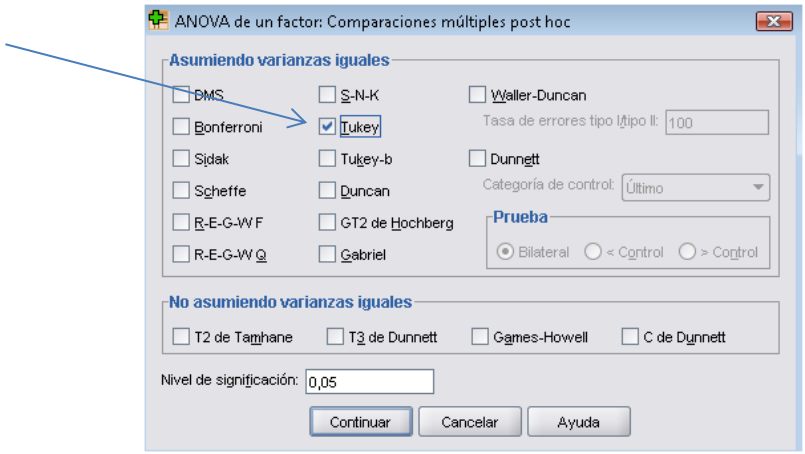


Figura 56. SPSS: resultados del ANOVA

Y el SPSS ofrece los siguientes resultados para la ‘pruebas post hoc’ HSD de Tukey (figura 57):

Pruebas post hoc

Comparaciones múltiples

Y
HSD de Tukey

(I) Condiciones	(J) Condiciones	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
1	2	4,000*	1,155	,031	,46	7,54
	3	-4,000*	1,155	,031	-7,54	-,46
2	1	-4,000*	1,155	,031	-7,54	-,46
	3	-8,000*	1,155	,001	-11,54	-4,46
3	1	4,000*	1,155	,031	,46	7,54
	2	8,000*	1,155	,001	4,46	11,54

*. La diferencia de medias es significativa al nivel 0.05.

Subconjuntos homogéneos

Y

HSD de Tukey[a]

Condiciones	N	Subconjunto para alfa = 0.05		
		1	2	3
2	3	6,00		
1	3		10,00	
3	3			14,00
Sig.		1,000	1,000	1,000

Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa el tamaño muestral de la media armónica = 3,000.

Figura 57. SPSS: resultados de la prueba de Tukey

La interpretación de los subconjuntos homogéneos que ofrece el SPSS es muy útil para visualizar de forma rápida si hay diferencias estadísticamente significativas entre las medias. Aquellas condiciones cuyas medias no difieren de forma estadísticamente significativa aparecen en la misma columna dentro del mismo subconjunto. Por el contrario, cuando dos medias difieren de forma estadísticamente significativa entre sí entonces aparecen en dos subconjuntos diferentes. En la tabla anterior se observa que cada media se encuentra en un subconjunto diferente indicando que las diferencias entre todos los pares de medias son estadísticamente significativas. Hay que revisar siempre la tabla de diferencias de medias, pues la prueba de subconjuntos homogéneos es más conservadora y podría ocurrir que no

se detecten diferencias con el análisis visual de los subconjuntos y sí en la tabla de diferencias de medias de Tukey.

Si se hubiese optado por ejecutar la prueba de Bonferroni o Scheffé los resultados serían los siguientes (figura 58).

Pruebas post hoc

Comparaciones múltiples

Variable dependiente:Y

	(I) Condiciones	(J) Condiciones	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
						Límite inferior	Límite superior
Scheffé	1	2	4,000*	1,155	,037	,30	7,70
		3	-4,000*	1,155	,037	-7,70	-,30
	2	1	-4,000*	1,155	,037	-7,70	-,30
		3	-8,000*	1,155	,001	-11,70	-4,30
	3	1	4,000*	1,155	,037	,30	7,70
		2	8,000*	1,155	,001	4,30	11,70
Bonferroni	1	2	4,000*	1,155	,040	,20	7,80
		3	-4,000*	1,155	,040	-7,80	-,20
	2	1	-4,000*	1,155	,040	-7,80	-,20
		3	-8,000*	1,155	,001	-11,80	-4,20
	3	1	4,000*	1,155	,040	,20	7,80
		2	8,000*	1,155	,001	4,20	11,80

*. La diferencia de medias es significativa al nivel 0.05.

Figura 58. SPSS: resultados de la prueba de Bonferroni y Scheffé

La prueba de Scheffé es más potente (menor error de Tipo II) que Bonferroni ya que si se observa la tabla anterior de 'máximo número de contrastes (comparaciones) que deberían probarse con el procedimiento de Bonferroni' (ver la tabla 12) se puede comprobar que Bonferroni sería más potente si el máximo de comparaciones fuese de dos contrastes para un diseño con tres grupos y seis grados de libertad del error.

De todos modos para esa situación donde se realizan todas las comparaciones posibles dos a dos ya se ha comprobado que la prueba más potente es DHS de Tukey (valores de p = .031 y p = .001). Y esa sería la decisión más correcta para ese diseño y, por lo tanto, la que debería seleccionar el investigador o investigadora para realizar sus hipótesis específicas, asegurando la validez de conclusión estadística.

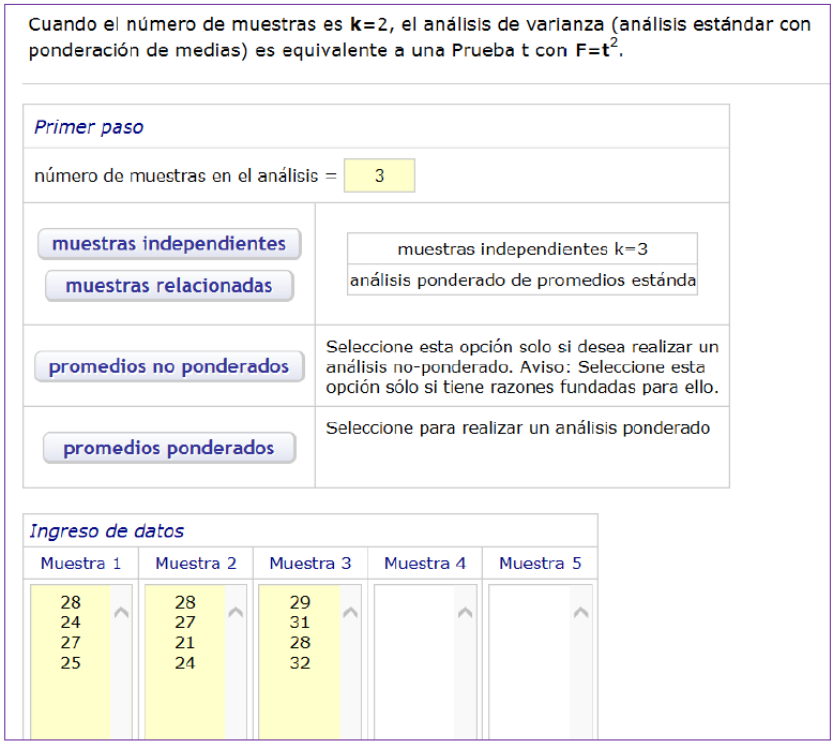
En ningún tipo de prueba de hipótesis específicas informa el SPSS del valor del Rango Crítico del estadístico aplicado sino que directamente informa del valor p de probabilidad vinculado al contraste de diferencia de medias. Otros programas estadísticos sí informan del valor p de probabilidad junto al valor de Rango Crítico

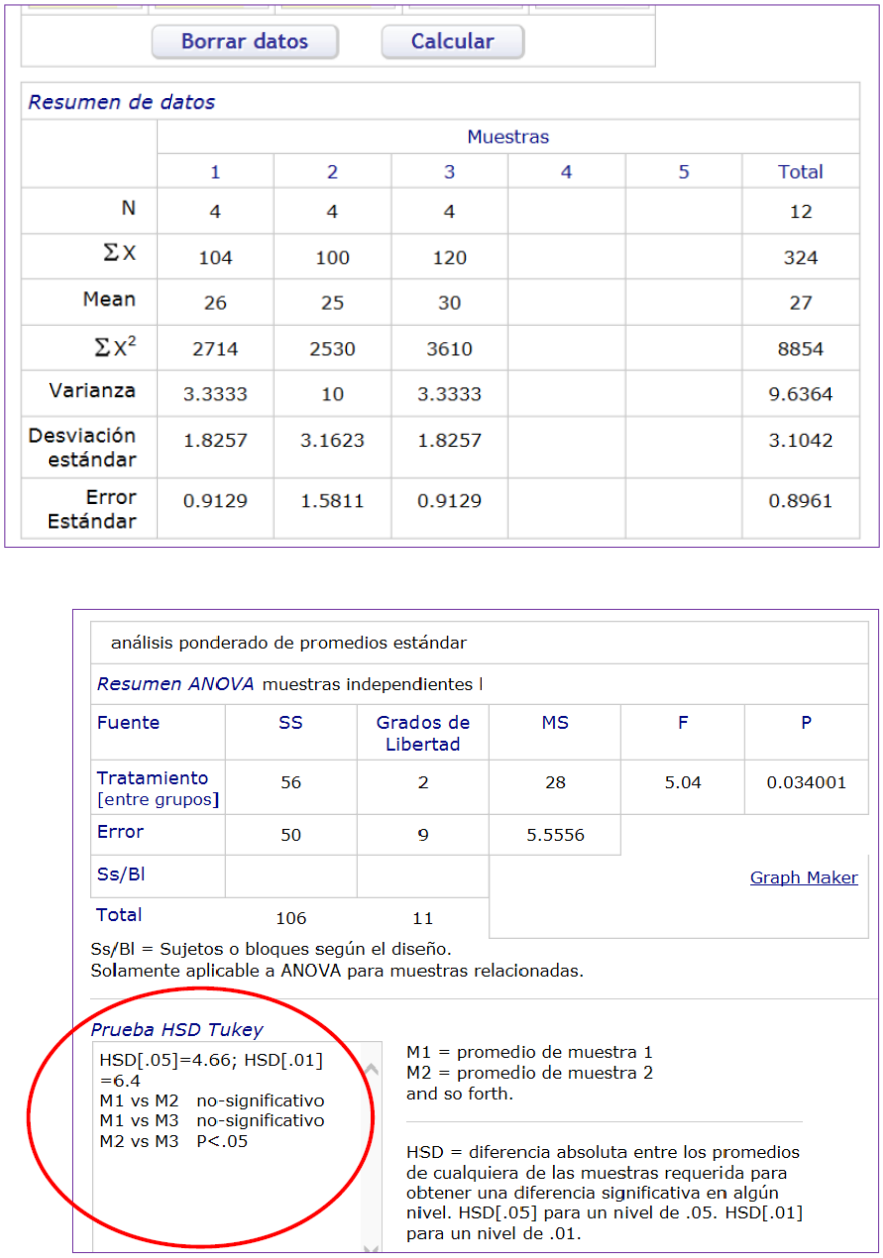
como por ejemplo la aplicación del ANOVA que se encuentra en la Web Vassar Stats (http://vassarstats.net/).

La ejecución de un modelo de diseño entre-sujetos unifactorial A = 3 univariado mediante la Web Vassar Stats se detalla a continuación (figura 59). Se puede observar que sus resultados sí aportan el valor del Rango Crítico de la prueba de Tukey. El lector o lectora puede comprobar el valor del rango crítico para dos valores de alfa: $RC_{.05}$ = 4.66 y $RC_{.01}$ = 6.4. A medida que disminuye el valor *p* de probabilidad, aumenta el valor del Rango Crítico, tal y como sucede con los valores de la *F* teórica o de cualquier estadístico de contraste. Una vez se conoce el valor del Rango Crítico se puede comparar con la diferencia entre las medias. Si la diferencia de medias es ≥ que el valor del Rango Crítico entonces se concluye que la diferencia hallada es estadísticamente significativa. Si la diferencia de medias es < que el valor del Rango Crítico entonces se concluye que la diferencia hallada no es estadísticamente significativa.

ANOVA UNIFACTORIAL A=3 CON VASSAR STATS

http://vassarstats.net/





Borrar datos Calcular

Resumen de datos

	Muestras					
	1	2	3	4	5	Total
N	4	4	4			12
ΣX	104	100	120			324
Mean	26	25	30			27
ΣX^2	2714	2530	3610			8854
Varianza	3.3333	10	3.3333			9.6364
Desviación estándar	1.8257	3.1623	1.8257			3.1042
Error Estándar	0.9129	1.5811	0.9129			0.8961

análisis ponderado de promedios estándar

Resumen ANOVA muestras independientes I

Fuente	SS	Grados de Libertad	MS	F	P
Tratamiento [entre grupos]	56	2	28	5.04	0.034001
Error	50	9	5.5556		
Ss/Bl					
Total	106	11			

Graph Maker

Ss/Bl = Sujetos o bloques según el diseño.
Solamente aplicable a ANOVA para muestras relacionadas.

Prueba HSD Tukey

HSD[.05]=4.66; HSD[.01] =6.4
M1 vs M2 no-significativo
M1 vs M3 no-significativo
M2 vs M3 P<.05

M1 = promedio de muestra 1
M2 = promedio de muestra 2
and so forth.

HSD = diferencia absoluta entre los promedios de cualquiera de las muestras requerida para obtener una diferencia significativa en algún nivel. HSD[.05] para un nivel de .05. HSD[.01] para un nivel de .01.

Figura 59. Aplicación on line de Vassar Stats: diseño entre-grupos A = 3

SPSS. ANALIZAR A = 2 → Modelo Lineal General → univariado

En la figura 60 se reproducen los resultados de un estudio tal y como se obtienen con el programa SPSS junto con la ventana de datos que se ha elaborado para introducirlos en la base cuando se trata de un *diseño entre-grupos unifactorial univariado* (en la ventana de datos poner las etiquetas al factor A: situar 1 para shock escapable y 2 para shock inescapable).

Para llevar a cabo el análisis con el SPSS se seleccionan las siguientes instrucciones una vez nos encontramos siguiendo el siguiente recorrido en las ventanas del programa:

ANALIZAR → Modelo lineal general → univariado

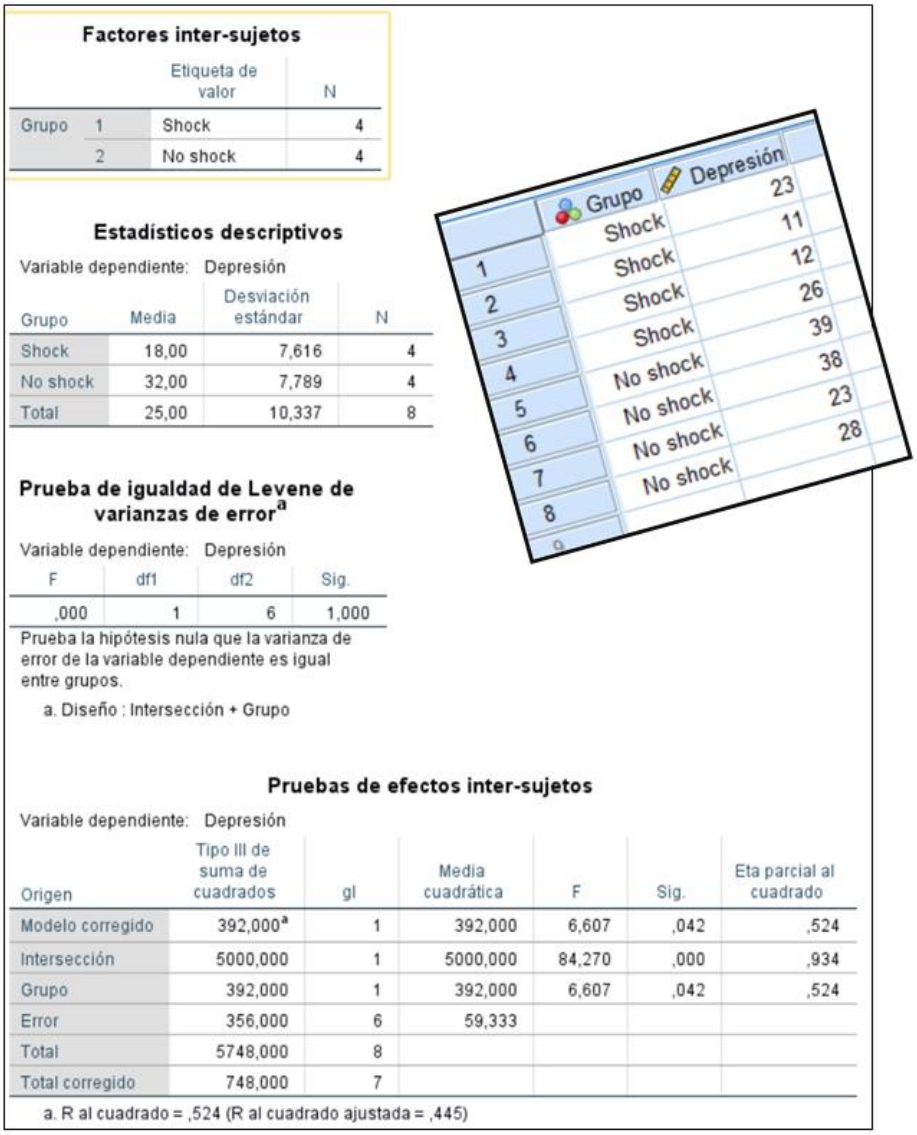
A continuación se sitúan en la base las variables en su lugar correcto:

-variable dependiente Y

-factores fijos: variable independiente A

Y también se selecciona, a la derecha, en la ventana de 'Opciones' los apartados de: estadísticos descriptivos, estimaciones del tamaño del efecto y pruebas de homogeneidad.

Una vez se ejecutan esas instrucciones, los resultados que aporta el programa relativos al análisis de la varianza solictado se detallan en la figura 41. Esos resultados del ANOVA y sus estadísticos deben ser redactados en el informe de investigación tal y como se ha detallado anteriormente.



Factores inter-sujetos

		Etiqueta de valor	N
Grupo	1	Shock	4
	2	No shock	4

	Grupo	Depresión
1	Shock	23
2	Shock	11
3	Shock	12
4	Shock	26
5	No shock	39
6	No shock	38
7	No shock	23
8	No shock	28

Estadísticos descriptivos

Variable dependiente: Depresión

Grupo	Media	Desviación estándar	N
Shock	18,00	7,616	4
No shock	32,00	7,789	4
Total	25,00	10,337	8

Prueba de igualdad de Levene de varianzas de error[a]

Variable dependiente: Depresión

F	df1	df2	Sig.
,000	1	6	1,000

Prueba la hipótesis nula que la varianza de error de la variable dependiente es igual entre grupos.

a. Diseño : Intersección + Grupo

Pruebas de efectos inter-sujetos

Variable dependiente: Depresión

Origen	Tipo III de suma de cuadrados	gl	Media cuadrática	F	Sig.	Eta parcial al cuadrado
Modelo corregido	392,000[a]	1	392,000	6,607	,042	,524
Intersección	5000,000	1	5000,000	84,270	,000	,934
Grupo	392,000	1	392,000	6,607	,042	,524
Error	356,000	6	59,333			
Total	5748,000	8				
Total corregido	748,000	7				

a. R al cuadrado = ,524 (R al cuadrado ajustada = ,445)

Figura 60. Datos introducidos en el SPSS y resultados del ANOVA

SPSS: ANOVA de un factor para muestras o grupos independientes

A continuación se presenta un conjunto de datos para un diseño entre grupos unifactorial univariado A=2 analizado con la prueba de contraste de hipótesis de ANOVA (Análisis de la Varianza) mediante el programa estadístico SPSS. Los datos corresponden a un diseño entre-sujetos o entre-grupos con una única variable independiente o factor (A). El diseño 'entre-sujetos' señala que cada sujeto solamente es medido en una ocasión ya sea expuesto a la condición a_1 del factor o variable independiente o a la condición a_2 de dicho factor. También se conoce como un

'diseño de muestras independientes'. Además es un diseño con una sola variable dependiente o variable medida (Y) y por ello es un diseño 'univariado'. La introducción de los datos de un diseño entre-grupos unifactorial univariado en el programa de SPSS se realiza tal y como se detalla en la figura 61.

	Condición	Y	var
1	1	13	
2	1	5	
3	1	11	
4	1	11	
5	2	27	
6	2	21	
7	2	32	
8	2	24	
9			

Figura 61. Datos introducidos en el SPSS y resultados del ANOVA

A continuación se puede ejecutar el análisis de la prueba estadística de ANOVA para un factor (figura 62). En el SPSS se selecciona la ventana de:

ANALIZAR → comparar medias → ANOVA de un factor

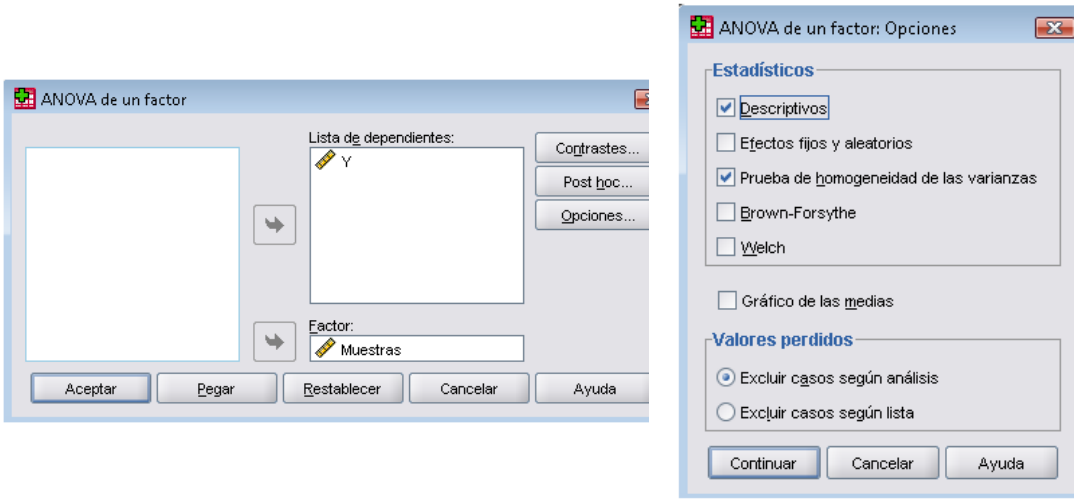


Figura 62. Datos introducidos en el SPSS y resultados del ANOVA

Dentro de 'ANOVA de un factor' seleccionamos Opciones y clicamos sobre Estadísticos: Descriptivos y Prueba de homogeneidad de las varianzas de los dos

grupos. Se aconseja realizar el análisis con la ventana de Modelo Lineal General ya que en ese apartado se encuentran todos los modelos de diseño.

Los resultados del ANOVA de un diseño entre-grupos A= 2 univariado son los siguientes (figura 63):

ANOVA de un factor

[Conjunto_de_datos1]

Descriptivos

Y

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	4	10,00	3,464	1,732	4,49	15,51	5	13
2	4	26,00	4,690	2,345	18,54	33,46	21	32
Total	8	18,00	9,366	3,311	10,17	25,83	5	32

Prueba de homogeneidad de varianzas

Y

Estadístico de Levene	gl1	gl2	Sig.
,429	1	6	,537

ANOVA

Y

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	512,000	1	512,000	30,118	,002
Intra-grupos	102,000	6	17,000		
Total	614,000	7			

Figura 63. Datos introducidos en el SPSS y resultados del ANOVA

Los resultados señalan, por un parte,que sí existe homogeneidad de las varianzas (Levene $F_{(1, 6)} = 0.429$, $p = .537$). Y, por otra parte, la prueba de contraste estadístico del Análisis de la Varianza (ANOVA) señala que la diferencia entre las medias del grupo 1 (Media = 10, DT=3.46) y del grupo 2 (Media = 26, DT=4.69) es estadísticamente significativa ($F_{(1, 6)}= 30, 118$, $p = .002$). Por lo tanto, la media del grupo 2 es mayor que la media del grupo 1, siendo la diferencia entre las medias estadísticamente significativa.

Se puede establecer una relación directa entre el valor del estadístico de contraste *t* de Student y el valor del estadístico de contraste *F* del Análisis de la Varianza dado que $t = F^2$ y $F = \sqrt{t}$.

Análisis con el programa JASP

A continuación, en la figura 64 se detalla la base de datos que se utiliza con el programa gratuito JASP y el resultado que se obtiene. La introducción de datos no se realiza en el programa JASP y por ello se necesitan otros programas. La introducción de los datos se puede realizar con Excel o con SPSS y luego se llama a ese archivo de datos desde el programa JASP. Para ello, se abre el programa JASP y se selecciona en la izquierda donde están situadas tres líneas verticales y se busca en el ordenador al fichero que se desea abrir:

Open → Computer

Así, se busca en el ordenador el fichero de datos que se va a utilizar para el contraste de hipótesis estadísticas y se selecciona para que se abra en el programa JASP.

Cuando ya se dispone en la pantalla del ordenador los datos que son objeto de análisis, se selecciona la ventana que tiene la opción de ANOVA y se van seleccionando las casillas que se corresponden con el diseño que se va a ejecutar: *diseño entre-grupos unifactorial univariado*. Es decir, al abrir la ventana se selecciona 'ANOVA' y se sitúan las variables dependiente ('Depresión') e independiente (factor fijo) ('Grupo') (figura 64).

Además, se seleccionan las opciones de:

-1) 'Assumption Checks: Homogeneity tests' para analizar si las varianzas de los dos grupos son homogéneas.

2) Marginal Means: Descriptive statistics, Estimates of effect size y η^2.

El estadístico η^2 es el tamaño del efecto conocido como eta cuadrado o proporción de la varianza explicada por el efecto del tratamiento o Grupo. Así, se calcula como:

η^2 = Suma de Cuadrados del efecto / Suma de Cuadrados Total.

Su valor complementario, 1 - η^2, es la proporción de varianza no explicada por dicho efecto y que se corresponde con la varianza explicada por el término de error en la ecuación estructural planteada en este diseño de investigación:

1 - η^2 = Suma de Cuadrados del error / Suma de Cuadrados Total.

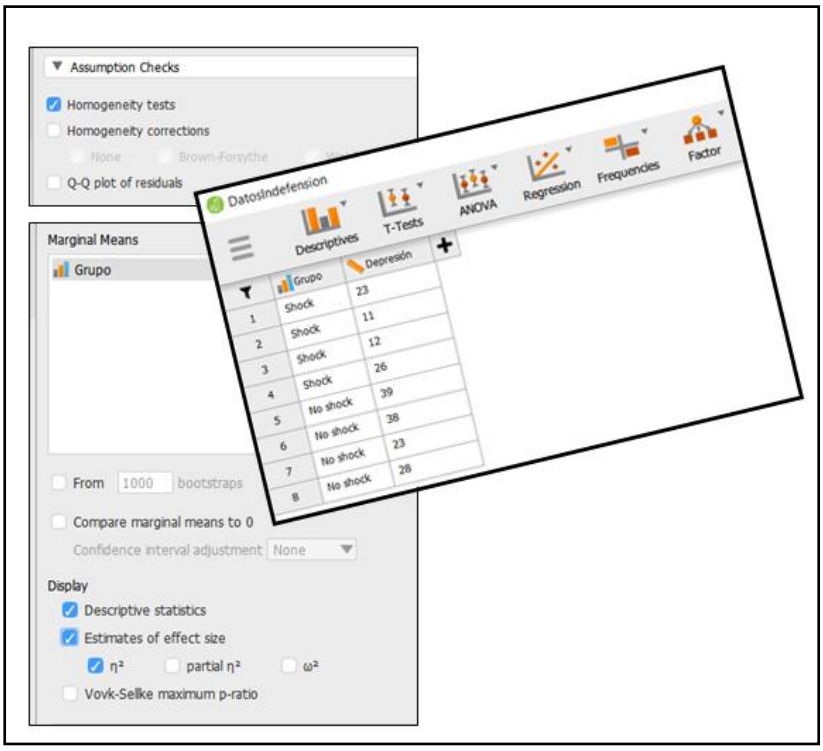
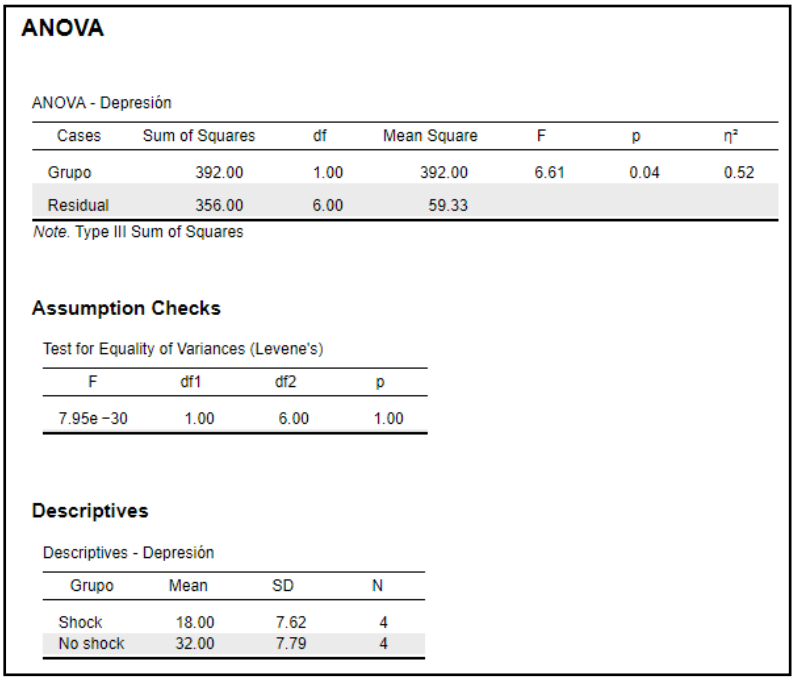


Figura 64. Datos introducidos en JASP y resultados del ANOVA

Los resultados de las instrucciones dadas al programa JASP se representan en la figura 65.



ANOVA

ANOVA - Depresión

Cases	Sum of Squares	df	Mean Square	F	p	η^2
Grupo	392.00	1.00	392.00	6.61	0.04	0.52
Residual	356.00	6.00	59.33			

Note. Type III Sum of Squares

Assumption Checks

Test for Equality of Variances (Levene's)

F	df1	df2	p
7.95e −30	1.00	6.00	1.00

Descriptives

Descriptives - Depresión

Grupo	Mean	SD	N
Shock	18.00	7.62	4
No shock	32.00	7.79	4

Figura 65. Resultados del ANOVA realizado con JASP

Una de las ventajas que tiene el programa JASP (además de ser un programa gratuito que se descarga de su página web (https://jasp-stats.org/) y que constantemente se está actualizando porque va incorporando más posibilidades de análisis) es que permite estimar el denominado Factor Bayes, FB. El Factor Bayes (BF) estima la probabilidad de una hipótesis estadística sobre la probabilidad de la otra hipótesis como ya se comentó anteriormente. En este caso, el usuario selecciona:

ANOVA → Bayesian ANOVA

Una vez seleccionado al ANOVA bayesiano, se sitúa la variable dependiente y la variable independiente en su lugar correcto y se marca la opción de BF10 si se desea estimar la probabilidad de la hipótesis alternativa respecto a la probabilidd de la hipótesis nula (figura 66). Se podría marcar la opción de BF01 y, entonces, el programa ofrece el resultado de la la probabilidad de la hipótesis nula respecto a la probabilidd de la hipótesis aletrnativa.

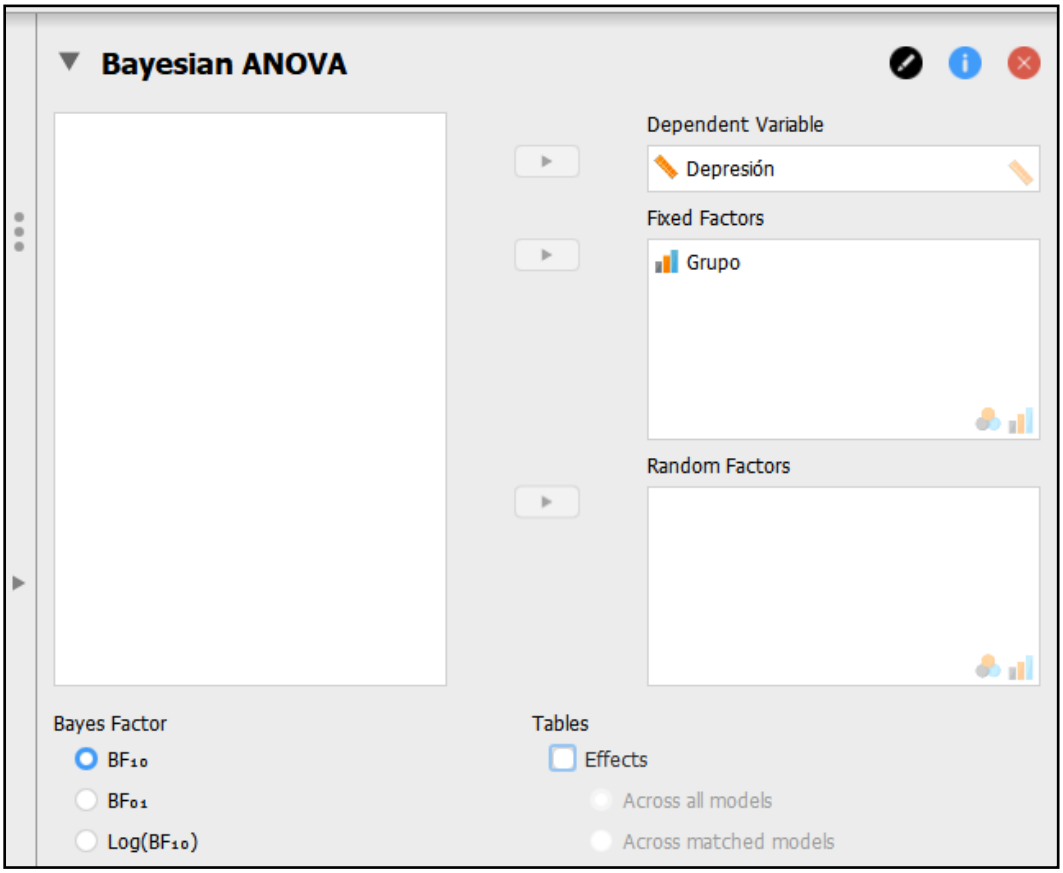


Figura 66. Factor Bayes con JASP

A continuación el programa JASP ofrece los resultados que se detallan en la figura 67. El valor del Factor Bayes es: BF10 = 2.11, es decir, la hipótesis alternativa es 2.11 más probable que la hipótesis nula.

Results

Bayesian ANOVA

Model Comparison

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Null model	0.50	0.32	0.47	1.00	
Grupo	0.50	0.68	2.11	2.11	1.04e −4

Figura 67. Factor Bayes con JASP

Capítulo 12. Potencia estadística y tamaño de la muestra

Dolores Frías-Navarro

Universidad de Valencia

Índice

Citar el capítulo como:

Frías-Navarro, D. (2021). Potencia estadística y tamaño de la muestra. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

A continuación, en primer lugar, se resumen y se amplían los conceptos que están relacionados con la potencia estadística ("statistical power") y la planificación del tamaño de la muestra. La potencia estadística es un concepto fundamental en la fase de planificación del diseño de la investigación, ya que es necesario un mínimo de potencia para garantizar la captación de un efecto en la muestra cuando se lleva a cabo el contraste de hipótesis estadísticas, si es que realmente dicho efecto existe en la población (validez de conclusion estadística). Y la probabilidad de detectar ese efecto está vinculado a un tamaño de muestra apropiado. En segundo lugar, se detallan los aspectos metodológicos que están relacionados con la potencia estadística como su valor mínimo, su importancia, su relación con el tamaño del efecto, *N* y alfa y los pasos necesarios para llevar a cabo un análisis de potencia estadística dirigido a planificar el tamaño de la muestra de un estudio con el programa G*Power, entre otras cuestiones.

Qué es la potencia estadística

El análisis de la potencia estadística se utiliza para diseñar estudios que tengan una probabilidad deseada de observar un resultado estadísticamente significativo, asumiendo que hay un efecto verdadero de una determinada magnitud que se concreta en el diseño en un valor concreto. Por supuesto, ese análisis de potencia a priori solo será preciso cuando la estimación del tamaño del efecto sea precisa. Conviene tener en cuenta que los investigadores y las investigadoras utilizan un estimador del tamaño del efecto del parámetro poblacional que es desconocido y que, probablemente, tendrá cierto sesgo y variabilidad.

Cuatro parámetros están implicados en un análisis de potencia estadística a priori ("potencia estadística a priori"): alfa, potencia estadística deseada, *N* (número de observaciones necesarias en el diseño) y tamaño del efecto esperado.

En varias ocasiones ya se ha comentado que el objetivo de las pruebas clásicas de inferencia estadística (NHST) es determinar si se puede o no rechazar la hipótesis nula (H_0). Por supuesto, la decisión que se adopte estará sujeta a una probabilidad de error estadístico, ya sea el error de Tipo I (rechazar H_0 cuando es cierta) o el error de Tipo II (mantener H_0 cuando es falsa).

El *error de Tipo I* es la probabilidad de equivocarse al rechazar la hipótesis nula cuando es realmente verdadera (error alfa, α, falso positivo). Es decir, es la probabilidad de encontrar un resultado estadísticamente significativo cuando la hipótesis nula es cierta. Generalmente, el máximo de error de Tipo I se sitúa en un valor de alfa = .05.

El *error de Tipo II* es la probabilidad de no rechazar la hipótesis nula o mantenerla cuando realmente es falsa (error beta, β, falso negativo). Es decir, es la probabilidad de no detectar un resultado estadísticamente significativo cuando la hipótesis nula es falsa. Generalmente, el máximo de error de Tipo II se sitúa en un valor de beta = .20.

Las decisiones correctas cuando se interpreta el resultado de la prueba estadística son el *nivel de confianza* o probabilidad de mantener la hipótesis nula cuando es cierta (1 – alfa) y la *potencia estadística* (1 – beta). El valor del nivel de confianza es el complementario del error alfa y, por ello, generalmente, se plantea un nivel de confianza de al menos .95.

La potencia de la prueba estadística es la probabilidad de rechazar con éxito la hipótesis nula cuanda es falsa y es igual a 1 – beta. Con otras palabras, la potencia estadística es la probabilidad de rechazar la hipótesis nula de no efecto cuando el efecto verdadero es no nulo en la población. Su valor es el complementario del error beta y, por ello, generalmente se plantea una potencia estadística de al menos .80. Jacob Cohen (1962,1988, 1990) es uno de los primeros autores que destacaron la importancia de la potencia estadística en el contraste de hipótesis junto con el tamaño del efecto.

La potencia estadística de una prueba de contraste de hipótesis ya ejecutada ("potencia estadística a posteriori") depende del tamaño del efecto real obtenido en el estudio, del tamaño de la muestra utilizado en el estudio y del nivel de alfa que se utilizó.

La potencia estadística que realmente es importante cuando se diseña una investigación es la potencia estadística a priori ya que guiará al investigador o investigadora para diseñar las condiciones que permitan detectar el efecto o la relación entre las variables planteada en la hipótesis del estudio, si realmente hay un efecto. En cambio, la potencia estadística a posteriori solo informa de la potencia que se tuvo para ejecutar la prueba estadística y no permite avanzar hacia una solución

si la potencia fue escasa. Por supuesto, cuando se mantiene la hipótesis nula siempre la potencia estadística de esa prueba de contraste (potencia estadística a posteriori) no fue suficiente para detectar el efecto, si había un efecto.

El objetivo de un análisis de potencia estadística a priori es controlar el error de Tipo II o limitar la probabilidad de observar un resultado no estadísticamente significativo cuando sí existe un efecto de un tamaño específico en la población. Este tipo de potencia está unida al diseño de la investigación y es fundamental para garantizar que la prueba estadística se ejecuta con la sensibilidad adecuada para detectar un efecto, si realmente existe. Con este tipo de actuaciones se fortalece la validez de conclusión estadística: capacidad para detectar el efecto, si realmente existe.

Conviene recordar que cuando se rechaza la hipótesis nula, si se ha cometido un error estadístico solamente podría ser el error de Tipo I, pues con el contraste estadístico el efecto se detectó y otra cosa es que no exista en la población y por lo tanto se habría cometido el error Tipo I. Por ello, en esa situación de rechazo de la hipótesis nula sería imposible cometer un error de Tipo II.

En cambio, si la decisión estadística es mantener la hipótesis nula entonces el único error estadístico que podría cometerse es el error de Tipo II, es decir, que realmente existe el efecto en la población y no se ha detectado y en este caso sería imposible cometer error de Tipo I ya que se mantiene la hipótesis nula.

Es muy importante conocer que el análisis de la potencia estadística a priori está relacionado con tres elementos implicados en el diseño de la investigación:

1. Tamaño de la muestra o número de observaciones (*N*) (generalmente es el parámetro que se quiere planificar con el estudio de potencia).
2. Tamaño del efecto esperado (fijado a priori por el investigador o investigadora y, por lo tanto, es el tamaño del efecto que se espera detectar con el estudio).
3. Criterio adoptado de decisión estadística (valor de alfa) (fijado a priori por el investigador o investigadora como máximo de error de Tipo I).

Si el investigador o investigadora conoce o planifica tres de los cuatro elementos señalados (potencia, tamaño de la muestra, tamaño del efecto y valor de alfa) entonces el valor del cuarto elemento se puede deducir. Es decir, si conoce por

ejemplo el tamaño de la muestra, el tamaño del efecto y se fija el valor de alfa entonces el valor de la potencia estadística para esa situación está determinado por los tres elementos anteriores. También si se conoce el tamaño del efecto, el grado de potencia estadística y se fija el valor de alfa entonces el tamaño de la muestra está determinado por los tres elementos anteriores. Por ello, el valor del tamaño de la muestra de un estudio se puede planificar con la información de los otros tres elementos utilizando por ejemplo el programa G*Power. Y esta cuestión es la que se va a tratar con detalle en este capítulo para presentar cómo planificar el tamaño de la muestra (*N*) de un estudio.

Valores de la potencia estadística (1 - beta)

De forma consensuada por la comunidad científica, el valor mínimo de potencia estadística que se considera adecuado para mantener en un nivel óptimo la validez de conclusión estadística del contraste estadístico es de .80 (es decir, beta = .20). Como se puede observar, el error de Tipo II o error beta es mayor que el que se considera para el error de Tipo I que se ha fijado como mucho en .05, es decir, el costo del error de Tipo I es más serio. Así, el costo de un error de Tipo I es cuatro veces mayor que el de un error de Tipo II (por lo tanto, la ratio β / α = 4).

Se considera más peligroso rechazar una hipótesis nula que realmente es cierta (decir que hay un efecto cuando realmente no lo hay) que mantener una hipótesis nula falsa (decir que no hay efecto cuando realmente sí lo hay). Conviene tener en cuenta que cuando se rechaza la H_0 se concluye que existe un efecto sistemático y supone un nuevo conocimiento o evidencia sobre el fenómeno, pero cuando se mantiene la H_0 se trata de un resultado no concluyente que debe seguir investigándose para indagar sobre su naturaleza, no cerrando posibles hipótesis de trabajo ni dirigiendo la investigación hacia alguna posición concreta. En cambio, el rechazo de la H_0 implica que se ha detectado un efecto sistemático (que debería ser replicado para otorgarle fiabilidad y consistencia), y ya supone una explicación a la naturaleza del fenómeno que es objeto de estudio. De ahí que la comunidad científica sea más conservadora con el error de Tipo I, y solamente rechazar H_0 cuando la probabilidad del resultado (o un resultado más extremo) sea tan pequeña que el efecto resulta extraño en la distribución de la hipótesis nula, con un margen de error de como mucho alfa = .05, es decir se rechazará solamente 5 hipótesis de cada 100

de forma errónea porque realmente no son prueba de ningún efecto sistemático. Utilizando la analogía del proceso de un juicio, es más grave declarar a alguien culpable cuando realmente no lo es, conduciéndolo a una condena, que declarar a alguien no culpable cuando realmente sí lo es.

Es importante tener en cuenta que el valor de alfa lo determina el investigador o la investigadora a priori y es fijo, es decir, no variará a lo largo del análisis. Si desea trabajar con un alfa de .01, y así lo planifica, ese valor no va a variar y no se debe variar, por supuesto, a posteriori cuando ya se conocen los resultados del estudio (mala conducta del investigador o investigadora).

En cambio, no ocurre lo mismo con el valor de la potencia estadística a priori (se trata de la potencia 'deseada' o esperada) que es un valor que se plantea en la fase de planificación del diseño del estudio, pero que puede variar con los datos de la muestra ya que quizás el efecto detectado ha sido menor al planificado o quizás ha sido mayor al planificado. Si esto sucede entonces la calidad del diseño quedará afectada ya que la prueba estadística no se habría ejecutado con el valor de potencia planificado (amenaza a la validez de conclusión estadística).

En resumen, el valor de la potencia estadística con la que se quiere ejecutar la prueba estadística se determina a priori por el investigador o investigadora, pero se planifica como un valor deseable en la fase de planificación del diseño para planificar, a su vez, el tamaño de la muestra con la que se realizará el estudio, esperando siempre que en ese estudio concreto se obtenga un determinado valor de tamaño del efecto tal y como se pronosticó y que tampoco es un valor fijo que controla el investigador o investigadora como valor que no cambiará. Entonces, solamente una vez ejecutado el análisis de los datos se podrá saber si el tamaño del efecto esperado o pronosticado es el que se ha obtenido en el estudio con una determinada muestra de participantes y, por lo tanto, se ha ejecutado ese análisis con la potencia estadística deseada por el investigador o investigadora y planificada a priori, siempre y cuando se utilizó el tamaño de la muestra adecuado para los tres parámetros o elementos planificados (alfa, beta y tamaño del efecto). Una vez ejecutado el análisis se puede calcular la potencia estadística a posteriori con la que se ha llevado a cabo realmente ese análisis estadístico, que puede ser la potencia planificada a priori o no. Sin embargo, como ya se ha comentado, la información que aporta la potencia

estadística a posteriori no es especialmente relevante ya que si se rechazó la hipótesis nula hubo suficiente potencia estadística y si no se rechazó hubo poca potencia estadística, si el efecto realmente existe en la población.

La importancia de la potencia estadística

En el diseño de la investigación es muy importante ejecutar los análisis con un nivel adecuado de potencia estadística por varios motivos:

1. Para equilibrar los falsos positivos (α) con falsos negativos (β). Es decir, puede ser adecuado utilizar diferentes valores valores de β, dependiendo del equilibrio deseado entre los errores de Tipo I y Tipo II, pues si un valor sube el otro baja y viceversa. Es decir, si se controla más el error de Tipo I con un valor más conservador o un valor más pequeño de alfa (por ejemplo se utiliza $\alpha = .01$) entonces se pierde potencia estadística (aumenta beta). Y si se toma la decisión estadística con un valor de alfa más liberal o un valor más grande de alfa (por ejemplo se utiliza $\alpha = .05$) entonces aumenta la potencia estadística (disminuye beta).
2. También es importante para aumentar la precisión de la estimación puntual del estadístico. Es decir, para disminuir el intervalo de confianza de dicha estimación. Para un valor de alfa concreto y un tamaño del efecto dado, la potencia estadística aumenta si se amplia la muestra con más datos y con ello aumenta la precisión de la estimación puntual. Con ello, las inferencias estadísticas que se realizan son, en general, más correctas o precisas (Maxwell y cols., 2008; Ioannidis, 2005; Sterne y Smith, 2001).
3. Además, cuando los resultados son más precisos y las inferencias estadísticas más correctas es más probable que los resultados se repliquen (Asendorpf y cols., 2013; Maxwell, 2004) y esta cuestión es fundamental para la acumulación correcta del conocimiento científico. Los estudios con baja potencia estadística para detectar el tamaño del efecto que se considera relevante o importante son más difíciles de replicar ya que la mayoria de ellos son falsos positivos (Button y cols., 2013; Maxwell, 2004).
4. Además, un beneficio importante es que el investigador o investigadora hace consciente, en su tarea de investigación de planificación del diseño, el hecho

de que encontrar un efecto que se considera relevante o importante también es una cuestión que debe tener un cierto grado de probabilidad de detectarlo con el diseño. Y con ello, debe plantearse que si planifica un diseño con una probabilidad insuficiente (por ejemplo, bajo tamaño de muestra) debería adoptar medidas para incrementar esa probabilidad de observar un efecto estadísticamente significativo cuando analice los datos. En definitiva, el análisis de la potencia a priori permite diseñar un estudio donde se incrementa la probabilidad de alcanzar una decisión correcta cuando se rechaza la hipótesis nula.

Durante décadas se ha comprobado que los investigadores e investigadoras no han llevado a cabo sus estudios con la suficiente potencia estadística (Bakker, van Dijk, & Wicherts, 2012; Hallahan, & Rosenthal, 1996; Perugini, y cols., 2014; Rossi, 1990; Sedlmeier, & Gigerenzer, 1989). Si a esta situación se añade el problema del sesgo de publicación (se publica un mayor número de estudios que muestran efectos estadísticamente significativos que los que obtienen resultados nulos) entonces se puede pensar que la literatura está llena de falsos positivos, explicando en cierto modo los problemas de replicación en la Ciencia y vinculando esa crisis de la replicación con la baja potencia estadística de los trabajos publicados (Frías-Navarro y cols, 2019). Por ello, disponer de suficiente potencia estadística para llevar a cabo el contraste de hipótesis es esencial para comprobar si hubo un efecto o una relación entre las variables y poder rechazar la hipótesis nula. Si la potencia estadística es baja entonces no se podrá detectar el efecto, no porque no haya efecto sino porque la técnica no ha sido lo suficientemente poderosa para detectarlo, teniendo que mantener H_0 aunque realmente es una hipótesis falsa (se estaría cometiendo error de Tipo II; falso negativo).

Actualmente ya no está justificada la falta de planificación de la potencia estadística por motivos de dificultad en el cálculo (generalmente, antes se consultaban las tablas para estimar la potencia teniendo en cuenta los parámetros de alfa, tamaño del efecto y *N*; Cohen, 1988; Tang, 1938), pues se han desarrollado programas que ejecutan de forma sencilla dicha estimación y, además, son gratuitos como G*Power que es uno de los más utilizados.

En resumen, la potencia estadística es importante porque aumenta directamente la probabilidad de encontrar un efecto si existe, porque contribuye indirectamente a reducir la tasa general de errores de inferencia de datos (O'Brien & Castelloe, 2007), hace consciente la necesidad de estimar o planificar el valor del tamaño del efecto esperado o deseado en un estudio concreto y porque favorece la replicación de los hallazgos. Por lo tanto, es muy conveniente que el investigador o investigadora dedique un tiempo a la planificación de la potencia estadística y los elementos estadísticos que la rodean, pues con ello aumenta las posibilidades de acertar en su decisión estadística y si existe el efecto esperado, pueda llegar a ser detectado.

Cómo aumentar la potencia estadística

La respuesta es sencilla: dado un nivel concreto de alfa, la potencia estadística aumenta cuando se incrementa el tamaño de la muestra y el tamaño del efecto, (Perugini y cols., 2018).

Manteniendo constante el valor del alfa y el tamaño del efecto, a mayor muestra de datos mayor es la potencia estadística. Y manteniendo constante el tamaño de la muestra y el tamaño del efecto, a mayor valor de alfa mayor es la potencia estadística. Y manteniendo constante el tamaño de la muestra y el valor de alfa, a mayor valor del tamaño del efecto mayor es la potencia estadística.

Por lo tanto, manteniendo constante dos de los elementos (por ejemplo, alfa y tamaño del efecto), la potencia estadística aumenta cuando aumenta el tercer elemento (en el ejemplo, el tamaño de la muestra). También se conoce que la potencia estadística aumenta cuando aumenta el tamaño del efecto. Y, además, la potencia estadística aumenta si se utilizan valores de alfa o criterios de decisión más indulgentes (por ejemplo, α = .06 en lugar de α = .05 o α = .05 en lugar de α = .01).

Por lo tanto, la potencia estadística aumenta cuando:

1. Aumenta el tamaño de la muestra
2. Aumenta el tamaño del efecto
3. Aumenta el valor de alfa

Potencia estadística y tamaño del efecto

La potencia de una prueba estadística es la probabilidad de rechazar la hipótesis cuando es realmente falsa. Por lo tanto, cuando se habla de potencia estadística se hace referencia a detectar un efecto que realmente existe (decisión correcta). En otras palabras, es la capacidad que tiene la prueba estadística para detectar una diferencia entre las medias o una asociación entre las variables de una determinada magnitud o tamaño de efecto.

La potencia estadística está directamente vinculada con el concepto del tamaño del efecto: potencia a priori y potencia a posteriori

La primera idea que hay que tener clara cuando se habla de potencia estadística es que el valor de la potencia estadística planificado a priori (el valor que se planifica antes de ejecutar el análisis de datos, denominada "potencia a priori") solo sera preciso cuando el valor del tamaño del efecto estimado sea también preciso. Es decir, si cuando se ejecuta el análisis de datos se observa que el valor del tamaño del efecto obtenido se ha alejado de su valor estimado entonces este hecho afectará directamente al valor de la potencia estadística con la que se ha ejecutado dicho análisis (potencia estadística a posteriori). Si el tamaño del efecto disminuye entonces disminuirá la potencia estadística o si no se logró el tamaño de muestra que se había planificado también disminuirá la poetencia estadística a posteriori. Esa potencia estadística real con la que se ha ejecutado la prueba de contraste estadístico es la denominada "potencia estadística a posteriori".

Por lo tanto, es muy importante que el investigador o investigadora valore con detalle el valor del tamaño del efecto que pronostica que encontrará en su estudio para ajustar adecuadamente el tamaño de la muestra según el valor de la potencia estadística deseada a priori y el valor de alfa elegido a priori. Por supuesto, también es necesario diseñar una investigación que permita alcanzar ese valor de tamaño del efecto estimado para garantizar que la potencia estadística se mantendrá en el valor que se ha planificado (cumplir el principio MAX-MIN-CON).

Si se cumplen los pronosticos de las estimaciones de tamaño del efecto y de potencia estadística solo se podrá saber una vez ejecutado el análisis de los datos

del estudio. Si el tamaño del efecto resulta que ha sido menor del pronosticado entonces la potencia estadística disminuirá.

En resumen, el diseño de la investigación debe planificarse para que la prueba estadística tenga una alta potencia estadística, es decir, para que sea sensible para detectar un efecto en la muestra. Se trata de la potencia estadística entendida como sensibilidad de la prueba estadística. El papel de la magnitud del tamaño del efecto es fundamental para estimar el tamaño de la muestra y la potencia estadística de la prueba de análisis de datos. El valor del tamaño del efecto poblacional es, por definición, desconocido, siendo necesario que el investigador o investigadora estime su valor para poder planificar el tamaño de la muestra en función de un valor de potencia estadística (al menos de .80) y un valor de alfa (como mucho de .05).

Tal y como se ha señalado, la potencia estadística depende del tamaño del efecto que suele ser un parámetro desconocido. En el capítulo del tamaño del efecto ya se han comentado las características de diferentes índices del tamaño del efecto. Los diferentes índices del tamaño del efecto pueden ser comparados entre sí a través de la conversión al mismo índice del tamaño del efecto. Esta es la actuación que se lleva a cabo en los estudios de meta-análisis para calcular el tamaño del efecto medio. Por ejemplo, todos los índices se pueden convertir a la diferencia estandarizada de medias *d* de Cohen o quizás se pueden convertir todos los índices al coeficiente de correlación. Es una decisión del investigador o investigadora basada, generalmente, en presentar los hallazgos del tamaño del efecto medio de la forma más comprensiva para los lectores y lectoras.

Un repaso de los valores de los índices de tamaño del efecto que más se utilizan y la conversión a los valores de *d* de Cohen de tamaño del efecto pequeño, mediano y grande se detalla en la tabla 13. Conviene recordar la importancia de contextualizar los efectos en un ámbito concreto de investigación y valorar su interpretación sustantiva o clínica como pequeño, mediano o grande en ese contexto concreto.

Tabla 13. Diferentes índices del tamaño del efecto y sus valores

	Pequeño	Mediano	Grande
d	**0.20**	**0.50**	**0.80**
η^2 (R^2)	.01	.06	.14
r	.10	.24	.37

r biserial puntual	.10	.30	.50
ANOVA: *f*	.10	.25	.40
AUC	.56	.64	.71
Regresion: f^2	.02	.15	.35
Asociación Tabla 2 *x* 2 OR	1.5	3.5	9
Asociación Ji Cuadrado W o Phi	.10	.30	.50

Tipo de análisis de potencia estadística

Con G*Power se pueden llevar a cabo cinco tipos de análisis de potencia estadística. En cada tipo se definen los valores de 3 parámetros y se busca el valor del cuarto parámetro (figura 68):

1) **Potencia a priori**: se computa el valor de *N*, dados los valores de alfa, potencia y tamaño del efecto.

2) **Análisis de alfa y beta**: se computan los valores de alfa y beta (nivel de significación y potencia estadística), dados los valores de la ratio β / α, *N* y tamaño del efecto.

3) **Potencia a posteriori**: se computa el valor de la potencia estadística, dados los valores de alfa, *N* y tamaño del efecto.

4) **Análisis del criterio**: se computa el valor alfa, dados los valores de potencia estadística, *N* y tamaño del efecto.

5) **Análisis de sensibilidad**: se computa el valor del tamaño del efecto, dados los valores de alfa, potencia estadística y *N*.

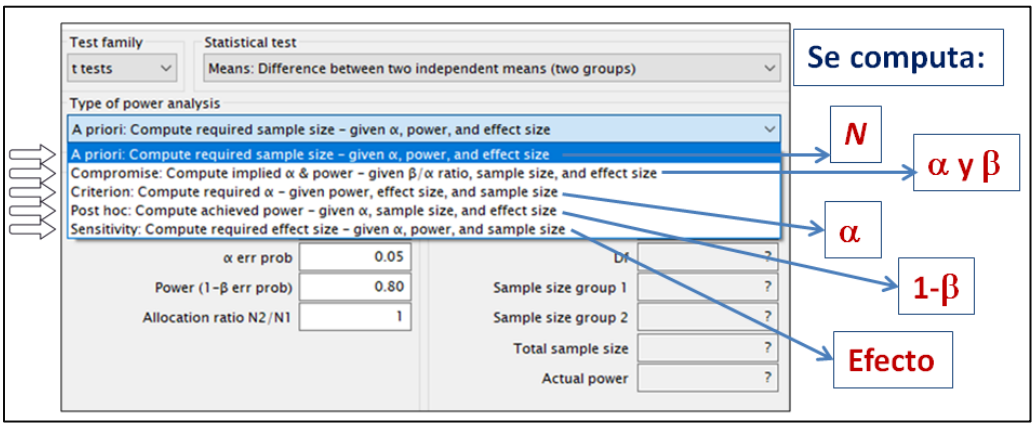


Figura 68. Tipo de análisis de potencia estadística con G*Power

Generalmente, el *análisis de potencia a priori* es el que más se utiliza ya que su objetivo es planificar el tamaño de la muestra del estudio (*N*) y esta cuestión es esencial en el diseño de investigación. Se recomienda que todos los artículos y las propuestas de proyectos de investigación presenten un análisis de potencia estadística a priori como medida que trata de controlar la validez de conclusión estadística de los resultados, con el objetivo de que en el estudio se utilice el tamaño de muestra necesario para captar un efecto esperado de un valor determinado, si realmente existe en la población.

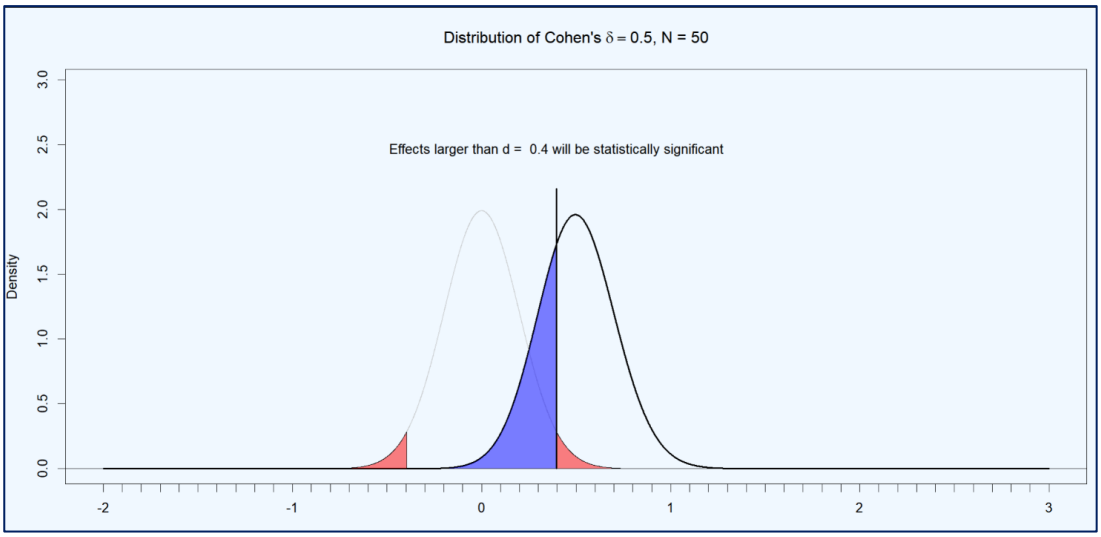
El *análisis de potencia estadística a posteriori* (cuando ya se ha realizado el análisis de los datos) se utiliza para comprobar con qué grado de potencia realmente se trabajo en el estudio, una vez que ya se conoce el efecto real que se obtuvo en dicho estudio. Si hay discrepancias entre el valor del efecto real y el del efecto esperado entonces la planificación del estudio de potencia a priori no habrá sido exitosa respecto al tamaño muestral planificado.

El *análisis de los valores de alfa y beta* en función de la ratio beta / alfa, no suele ser usual entre los investigadores e investigadoras, en general. Tampoco el análisis de criterio suele ser usual, pues son valores que fija el investigador a priori.

El *análisis de sensibilidad* sí suele utilizarse cuando se trata de conocer el efecto mínimo detectable (“minimal detectable effect”, MDE):

1) Se utiliza cuando ya se ha ejecutado el estudio, sobre todo si el resultado fue mantener la hipótesis nula, y se desea saber qué efecto mínimo hubiese sido necesario para rechazar la hipótesis nula.

2) También se utiliza cuando se dispone de pocos sujetos en el estudio. Por ejemplo, si la investigadora solo dispone de 20 participantes sería relevante conocer que valor del tamaño del efecto podría detectarse con ese tamaño de muestra para rechazar la hipótesis nula. Es decir, si solo tiene 20 sujetos en su estudio, la cuestión relevante ya no sería cuántos sujetos necesita sino saber qué efecto mínimo será detectable con ese tamaño de muestra disponible.



Cómo hacer un análisis de potencia estadística a priori

En un experimento concreto, una investigadora ha valorado que un tamaño del efecto de $d = 0.5$ es la mejor estimación del tamaño del efecto más pequeño que sería de interés en su estudio. Su objetivo es comparar las medias de 2 grupos con un diseño entre-sujetos univariado (A = 2), utilizando un $\alpha = .05$. La preguna fundamental que debe realizarse cuando planifica su estudio es la siguiente: ¿cuántos participantes serán necesarios en el estudio si se desea realizar un contraste estadístico con un 80% de potencia estadística o de probabilidad de detectar ese efecto estimado, $d = 0.5$, si realmente existe? La respuesta se presenta a lo largo de este apartado.

Primero: seleccionar el valor del error de Tipo I o alfa para los contrastes estadísticos

Segundo: definir el valor de tamaño del efecto esperado

Como se puede estar pensando, si la investigadora desconoce esa estimación del tamaño del efecto que espera encontrar en su estudio entonces no es posible plantearse cuántos datos necesita en su estudio (*N*). Y aquí radica la principal dificultad del estudio de potencia estadística: es necesario plantear a priori el tamaño del efecto que se desea detectar (tamaño del efecto deseado).

El tamaño del efecto poblacional por definición es desconocido y el investigador o investigadora deberá plantear un tamaño del efecto estimado o el tamaño del efecto

que espera detectar con su análisis estadístico y que representa al efecto o la relación entre las variables del estudio.

Por lo tanto, el investigador o investigadora debe decidir qué valor del tamaño del efecto espera en su campo concreto de trabajo y plantea en su hipótesis de investigación. Y hay que destacar que se trata del tamaño del efecto en un contexto concreto de trabajo ya que un tamaño del efecto de un determinado valor (por ejemplo, d = 0.4) puede ser muy pequeño en un área de trabajo y, en cambio, puede ser un valor muy grande en otra área de trabajo. De ahí la importancia de que los investigadores y las investigadoras indagen y reflexionen sobre el valor del tamaño del efecto que se espera encontrar en su tema objeto de estudio durante la fase de conocimiento previo del proceso de diseño de la investigación. Por lo tanto, el valor del tamaño del efecto se debe contextualizar en un área concreta de trabajo.

Y, además, hay que tener en cuenta que la importancia clínica o sustantiva de ese valor del tamaño del efecto para el fenómeno estudiado se la dará el juicio clínico del profesional, pues un tamaño del efecto pequeño podría ser muy importante en esa área concreta de estudio. Por ejemplo, salvar una vida podría ser muy importante aunque el tamaño del efecto de un fármaco fuese de d = 0.01.

En definitiva, un tema es la significación estadística (valor p), otro la significación o valor del tamaño del efecto y otro la significación sustantiva o clínica (juicio del profesional).

Si finalmente el investigador o investigadora no tiene ningún conocimiento sobre ese posible tamaño del efecto esperable entonces, quizás, sería mejor no seguir con el estudio porque sería como andar con una venda en los ojos sin tener ningún tipo de control sobre si sufriré un accidente o no. Por lo tanto, es fundamental que el investigador o investigadora se plantee el tamaño del efecto esperado en su estudio y la potencia deseada y lleve a cabo la planificación del tamaño de la muestra (N) para finalmente ejecutar el diseño de la investigación con las mejores condiciones posibles, tratando con ello de garantizar la validez de conclusión estadística de sus hallazgos.

Entonces, cómo estimar qué valor del tamaño del efecto se espera en un estudio. A continuación se detallan diversas actuaciones:

1. **Escenario 1.** La respuesta se puede obtener consultando un trabajo de meta-análisis sobre el área concreta de estudio que es de interés para el investigador o investigadora. Ese meta-análisis proporciona un tamaño del efecto medio y su intervalo de confianza y podría ser la información que se necesita para plantear como tamaño del efecto esperado en el estudio que está planificando el investigador o la investigadora. Es conveniente estudiar en ese meta-análisis si hubo o no sesgo de publicación ya que si fuera así ese tamaño del efecto podría estar sobrevalorado. Esta es una opción recomendable siempre y cuando el estudio de meta-análisis se realizó de forma correcta y no hubo sesgo de publicación.

2. **Escenario 2.** El investigador o investigadora podría utilizar el tamaño del efecto de un estudio individual publicado sobre la misma temática como valor poblacional desconocido. Esta alternativa no se considera, en general, óptima debido al sesgo de publicación tal y como se ha comentado anteriormente (se publican en mayor medida los resultados estadísticamente significativos y el tamaño del efecto podría estar sobrevalorado). Señalan Anderson y cols., (2017) que, generalmente, este escenario da como resultado la planificación de estudios con baja potencia estadística. Este escenario podría mejorar si se dispone de más de un estudio individual y se estima el tamaño del efecto medio ponderando por el tamaño de la muestra. No se trataría de un meta-análisis, pero sería un poco más preciso que solo un estudio. Como es lógico pensar, cuando se utiliza como estimación el valor del tamaño del efecto de un trabajo publicado (ya sea un meta-análisis o un estudio individual) su calidad o certeza dependerá de si el tamaño del efecto de la muestra de dicho estudio es una estimación precisa del valor poblacional. Hecho sobre el que siempre existirá una incertidumbre, sobre todo si el estudio se realizó con baja potencia estadística.

3. **Escenario** 3. El investigador o investigadora podría valorar el "tamaño del efecto mínimo" que sería de interés o importante en su estudio (útiles o teóricamente significativos) y utilizarlo como tamaño del efecto esperado, independientemente del valor real del tamaño del efecto (Lakens y cols., 2018).

4. **Escenario 4.** Si no se dispone de más información, se podría optar por utilizar uno de los valores o puntos de corte clásicos establecidos por Cohen: $d = 0.2$ tamaño del efecto pequeño, $d = 0.5$ tamaño del efecto mediano y $d \geq 0.8$ tamaño del efecto

grande. Se podría utilizar el tamaño del efecto medio de 0.5 como una estimación razonable de un tamaño del efecto útil y planificar el tamaño de la muestra en función de ese valor para un valor de alfa concreto y una determinada potencia estadística. Conviene resaltar que está decisión es recomedable solo cuando no hay respuesta concreta que justifique el uso de un determinado valor del tamaño del efecto esperado, pues como ya se ha comentando es importante conextualizar los efectos y quizás un tamaño de d = 0.8 podría ser trivial en alguna área de investigación o un valor de d = 0.2 podría ser muy grande en otro contexto de investigación.

Estudios recientes señalan que la mayoría de los tamaños del efecto en Psicología están en torno a d = 0.4 (Camerer y cols., 2018; Open Science Collaboration, 2015**;** Bosco, Aguinis, Singh, Field, & Pierce, 2015; Gignac, & Szodorai, 2016; Stanley, Carter, & Doucouliagos, 2018). Sin embargo, ese valor podría ser muy grande en algunas áreas de investigación y de ahí la importancia de la valoración teórica por parte del investigador o investigadora. Por ejemplo, en el estudio de replicación de Klein y cols. (2018) de 28 hallazgos publicados clásicos y contemporáneos, el tamaño medio del efecto fue solo de d = 0.15 (en comparación con d = 0.6 de los estudios originales).

Por lo tanto, y si no hay información que contextualice mejor el valor esperado del tamaño del efecto, el valor de d = 0.4 parece ser la estimación más razonable para buscar un tamaño del efecto mínimo, útil o teóricamente significativo en el ámbito de la Psicología. Ese valor de d = 0.4 se corresponde con un coeficiente de correlación de r = .2 como ya se ha comentado. Cuando se trata de un diseño entre-grupos, ese valor de tamaño del efecto d = 0.4 significa que tiene un 61% de posibilidades de encontrar la diferencia esperada si selecciona aleatoriamente a un participante de cada grupo (AUC).

En definitiva, solamente utilizar un tamaño del efecto de d = .4 como la estimación más razonable para buscar un efecto no despreciable, útil o teóricamente significativo si no se dispone de más evidencia sólida sobre el tamaño del efecto esperado.

Una reflexión que cada vez está cobrando más fuerza en la comunidad científica está relacionada con la presencia de los tamaños del efecto pequeños. En el ámbito psicológico, los efectos reales suelen ser pequeños. Probablemente los fenómenos

psicológicos están determinados por una multitud de causas y probablemente cada causa individual tenga un efecto pequeño. Por ello, los tamaños del efecto grandes es poco probable que sean reales. Como Funder y Ozer (2019) señalan, en el contexto de la investigación psicológica, un tamaño del efecto muy grande (como por ejemplo, r = .40 o mayor) es probable que esté sobreestimado ya que raramente ocurren o se replican. Por ello, señalan los autores, vale la pena tomar en serio los tamaños del efecto pequeños y, además, son más creíbles. Concluyen los autores señalando que los tamaños del efecto pequeños obtenidos con estudios con N grandes son los que probablemente mejor reflejen el estado del fenómeno estudiado. Además, conviene reflexionar sobre el hecho de que, aunque el tamaño del efecto sea pequeño cuantitativamente, si el efecto aparece de forma sistemática en diferentes situaciones y ambientes, probablemente algún tipo de información valiosa nos está transmitiendo.

Por ejemplo, se han encontrado tamaños del efecto pequeños en términos de los valores establecidos por Cohen (1988), pero muy importantes en cuestiones de salud como la correlación entre la aspirina y la prevención de un infarto (r = .03) (Rosnow & Rosenthal, 2003; Rosenthal, 1990; Steering Committee of the Physician's Health Study Research Group, 1988), la ingesta de calcio y la masa ósea en mujeres con premenopusia (r = .08) (Meyer y cols., 2001) y la ingesta de ibuprofeno y el alivio del dolor (r = .14) (Meyer y cols., 2001).

5. **Escenario 5.** Se podría pensar en realizar un estudio piloto a pequeña escala para obtener un valor del tamaño del efecto como estimador del futuro tamaño del efecto, pero los autores opinan que "los estudios piloto son casi inútiles para estimar los tamaños del efecto" (Brysbaert, 2019). Esos estudios pilotos son buenos para demostrar que un estudio o una técnica es viable, pero no ofrecen información fiable para estimar el tamaño del efecto (Albers & Lakens, 2018; Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006). Y es más, señala Brysbaert (2019), ese estudio piloto podría engañar al investigador o investigadora y conducirle por un camino falso si se detecta un efecto estadísticamente significativo en las pruebas piloto (falso positivo) y se utiliza como la única razón para embarcarse en un proyecto de investigación. En definitiva, no se recomienda llevar a cabo un estudio piloto para argumentar el valor de un tamaño del efecto esperado ya que no son estimaciones

válidas, pues son diseños con pocas observaciones y, por lo tanto, seguramente con escasa potencia estadística.

Programa de potencia estadística: G*POWER

El programa G*Power (http://www.gpower.hhu.de/) es uno de los más utilizados para llevar a cabo estudios de potencia estadística y del tamaño del efecto (Erdfelder, Faul, & Buchner, 1996; Faul, Erdfelder, Lang, & Buchner, 2007; Faul y cols., 2009). Su uso es sencillo cuando se comprenden los elementos que forman parte en el proceso de inferencia estadística así como las relaciones entre ellos: alfa, tamaño del efecto, potencia estadística y tamaño de la muestra. Su manual se puede descargar gratuitamente de:

https://oir.umh.es/files/2020/04/GPowerManual.pdf

*G*Power* y potencia estadística a priori

Siguiendo con el ejemplo anterior de la investigadora que ha planteado un tamaño del efecto esperado de $d = 0.5$ en su estudio, se va a desarrollar a continuación el cálculo del tamaño de la muestra (*N*) utilizando un alfa de .05 y fijando una potencia estadística deseada de .80. El objetivo es comparar las medias de 2 grupos con un diseño entre-sujetos univariado (A = 2) y comprobar si la diferencia entre ellas es estadísticamente significativa. La investigadora ha revisado la literatura (conocimiento previo) y el resultado de un trabajo de meta-análisis le permite plantear ese valor de tamaño del efecto como el esperable en su estudio.

La investigadora lleva a cabo un estudio de la potencia estadística a priori en la fase de planificación del diseño de investigación para estimar el tamaño mínimo de observaciones que necesariamente debe recoger (*N*) si desea encontrar un hallazgo que sea estadísticamente significativo con al menos una probabilidad del 80% si realmente el efecto de 0.5 pronosticado existe. Plantea su estudio como un diseño ortogonal o balanceado, es decir, $n_1 = n_2$ tal y como se observa al señalar que "Allocation ratio N2 / N1 = 1" (figura 69). El programa ofrece la posibilidad de ejecutar un diseño no ortogonal o no balanceado especificando la ratio que se planifica para el tamaño de los dos grupos

Una vez descargado el programa gratuito G*Power, versión 3.1.9.7., y abierto el programa, la primera acción es seleccionar el estadístico de aquella prueba de

contraste de hipótesis que se utilizará en el estudio en el menú. En la figura 67 se detallan los pasos para seleccionar el procedimiento de análisis de la potencia estadística deseado y los criterios estadísticos que ha adoptado la investigadora para llevar a cabo su análisis de potencia estadística.

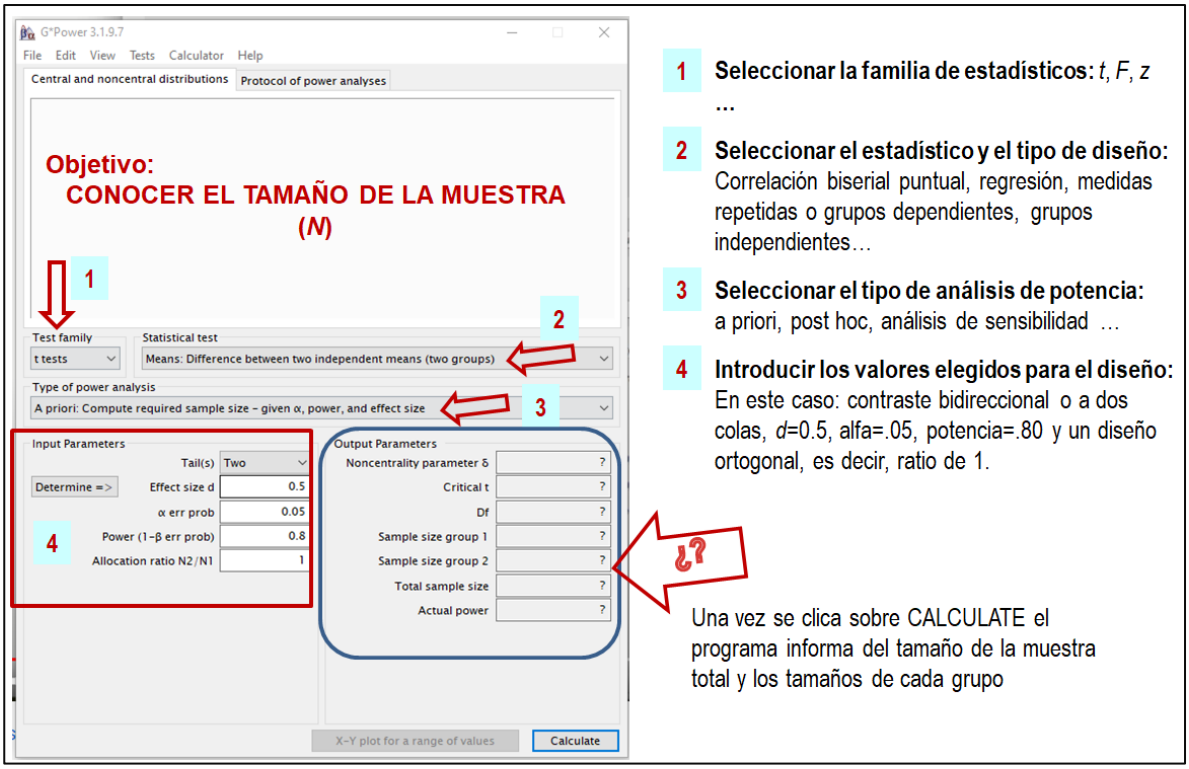


Figura 69. G*Power. Selección de los parámetros

Los resultados que ofrece el programa G*Power respecto al estudio planteado anteriormente son los siguientes (ver figura 70):

1. El diseño necesita un tamaño de *N* = 128 observaciones
2. Cada grupo (A = 2) debe tener, por lo tanto, 64 observaciones

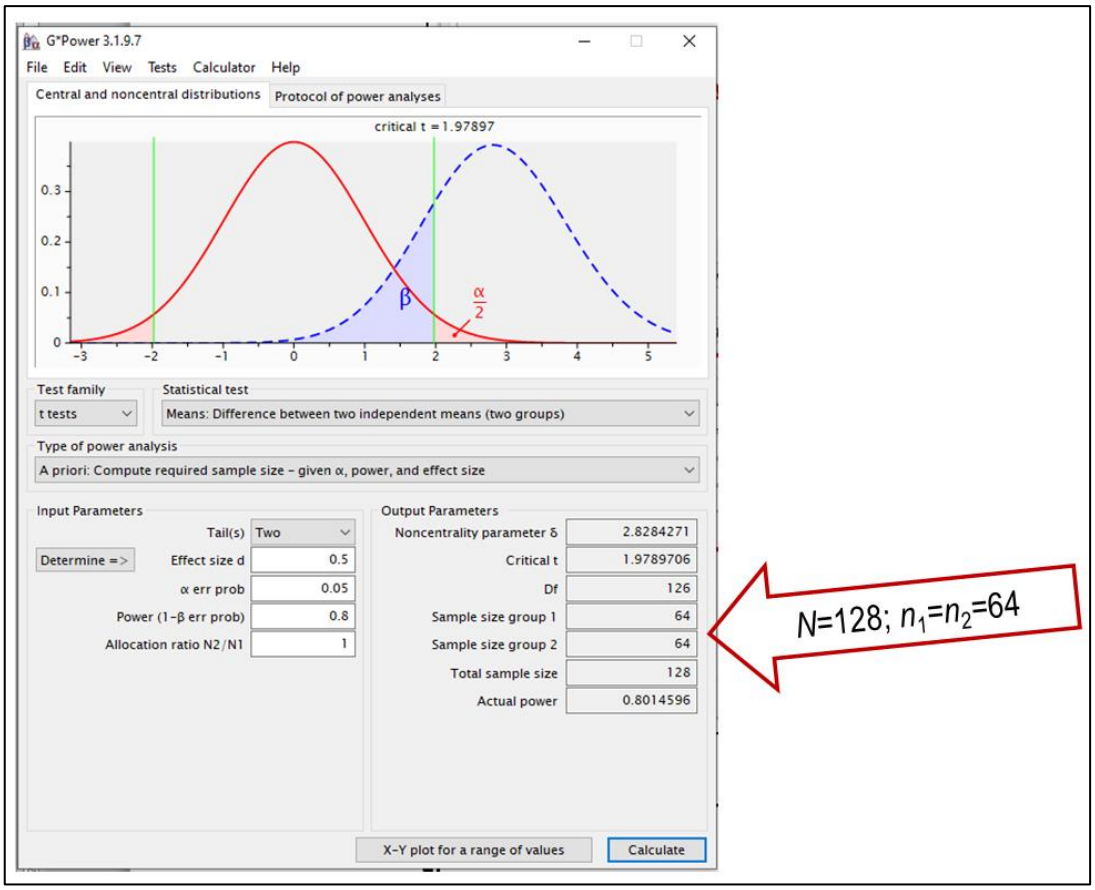


Figura 70. Resultados del tamaño de muestra necesario para una potencia de .80 utilizando un diseño ortogonal con 2 grupos independientes

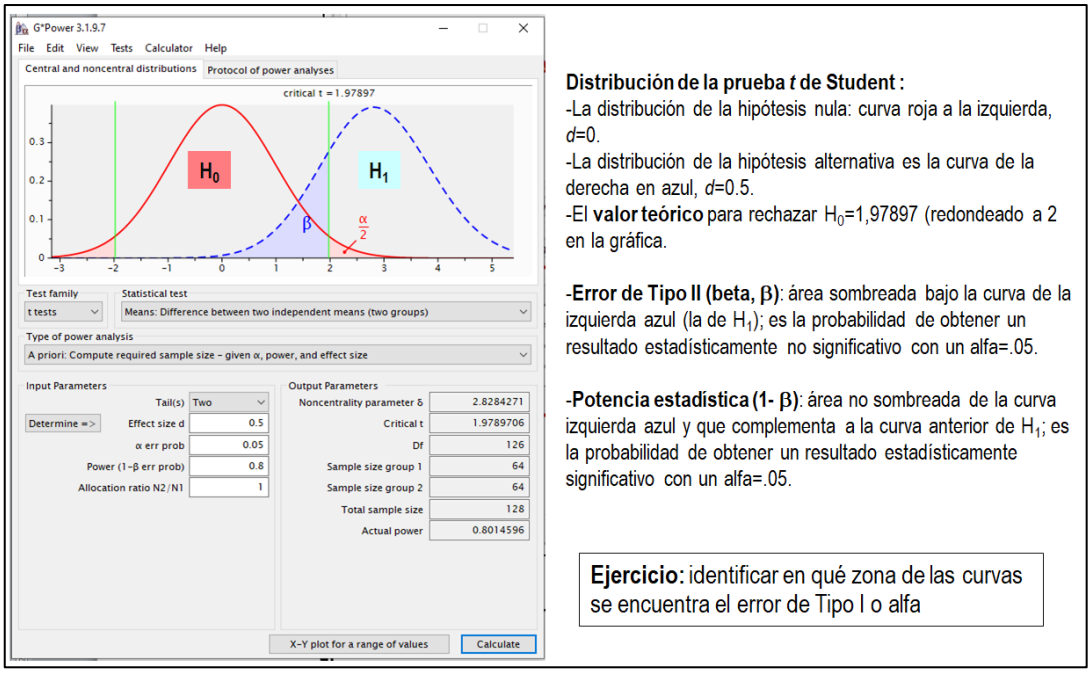
Por lo tanto, si se seleccionan 128 observaciones (64 en cada grupo) de una población donde el tamaño del efecto *d* de Cohen es de 0.5, en el 80% de las ocasiones se espera que la prueba *t* de Student señalará que la diferencia es estadísticamente significativa, utilizando un alfa de 0.05.

A continuación se detalla la información que ofrece el programa G*Power en dos secciones:

1) Curvas. Describiendo la información que proporcionan las dos curvas de las gráficas (distribución central de H_0 y distribución no central de H_1)

2) Parámetros. Describiendo todos los elementos que aporta el programa relacionados con los parámetros.

Curvas de distribución: distribución central de H_0 y distribución no central de H_1



Distribución de la prueba *t* de Student :
-La distribución de la hipótesis nula: curva roja a la izquierda, *d*=0.
-La distribución de la hipótesis alternativa es la curva de la derecha en azul, *d*=0.5.
-El **valor teórico** para rechazar H_0=1,97897 (redondeado a 2 en la gráfica.

-Error de Tipo II (beta, β): área sombreada bajo la curva de la izquierda azul (la de H_1); es la probabilidad de obtener un resultado estadísticamente no significativo con un alfa=.05.

-Potencia estadística (1- β): área no sombreada de la curva izquierda azul y que complementa a la curva anterior de H_1; es la probabilidad de obtener un resultado estadísticamente significativo con un alfa=.05.

Ejercicio: identificar en qué zona de las curvas se encuentra el error de Tipo I o alfa

Figura 71. Información proporciona por el programa G*Power

Información sobre los parámetros

En la figura 72 se detalla los resultados que aporta el programa G*Power (“output parameters”) al ejecutar ‘Calculate’.

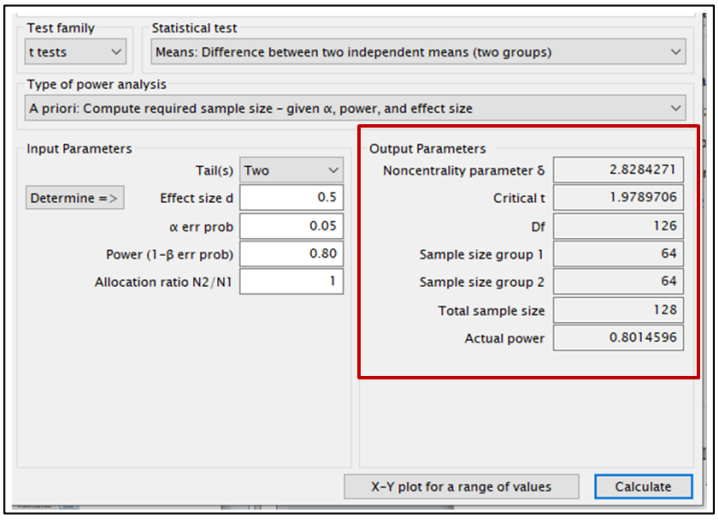


Figura 72. Información proporciona por el programa G*Power

El programa G*Power no ofrece solamente información sobre el tamaño de muestra necesario para los parámetros especificados de alfa, potencia y tamaño del efecto sino que, además, aporta otros datos que pueden ser relevantes para llevar a cabo más cálculos como el intervalo de confianza del tamaño del efecto. La información completa que proporciona el programa es la siguiente

1. Párametro de no centralidad (δ, "noncentrality parameter") (Liu & Raudenbush, 2004). La distribución de la H_0 es la "distribución central" de la prueba *t* de Student (grupos independientes: $t(N - 2, 0)$ y la "distribución no central" de la prueba *t* de Student es la distribución de H_1 $t(N - 2, \delta)$, siendo $N = n_1 + n_2$ y δ el parámetro de no centralidad que está relacionado con el tamaño del efecto y ponderado por el tamaño de los grupos:

$$\delta = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

El parámetro de no centralidad describe el grado de diferencia entre la H_0 y la H_1, es decir, describe la separación entre el centro de la distribución de la hipótesis alternativa y el centro de la distribución de la hipótesis nula. La distribución muestral que se asume como cierta en el contraste de hipótesis es la de la hipótesis nula, es decir, la distribución central de H_0 (centrada alrededor del valor 0). La distribución no central es la distribución muestral vinculada a la hipótesis alternativa, H_1, y depende de los grados de libertad y del valor del parámetro de no centralidad. La hipótesis alternativa está asociada a la presencia de efectos diferentes a cero y pueden ser muchas ya que dependen del tamaño del efecto y, por ello, hay una gran cantidad de posibles distribuciones no centrales.

2. El valor teórico de la prueba t de Student ("Critical *t*").
3. Los grados de libertad ("*Df*")
4. Tamaño de la muestra del grupo 1
5. Tamaño de la muestra del grupo 2
6. Tamaño total de la muestra
7. Potencia real

*Tamaño del efecto con G*Power*

El programa G*Power también permite calcular el valor del tamaño del efecto *d* de Cohen cuando se está trabajando con el estadístico de la *t* de Student. Se trata de presionar sobre 'Determine', en el lado izquierdo de la etiqueta del tamaño del efecto ('effect size d'). Se abre una ventana para calcular el tamaño del efecto a partir de las medias y las desviaciones típicas de los dos grupos (figura 73). Conviene recordar que la prueba *t* asume que las varianzas de los dos grupos son iguales (homocedasticidad de varianzas), aunque es una prueba relativamente robusta al incumplimiento de dicho supuesto si el tamaño de los grupos es el mismo ($n_1 = n_2$).

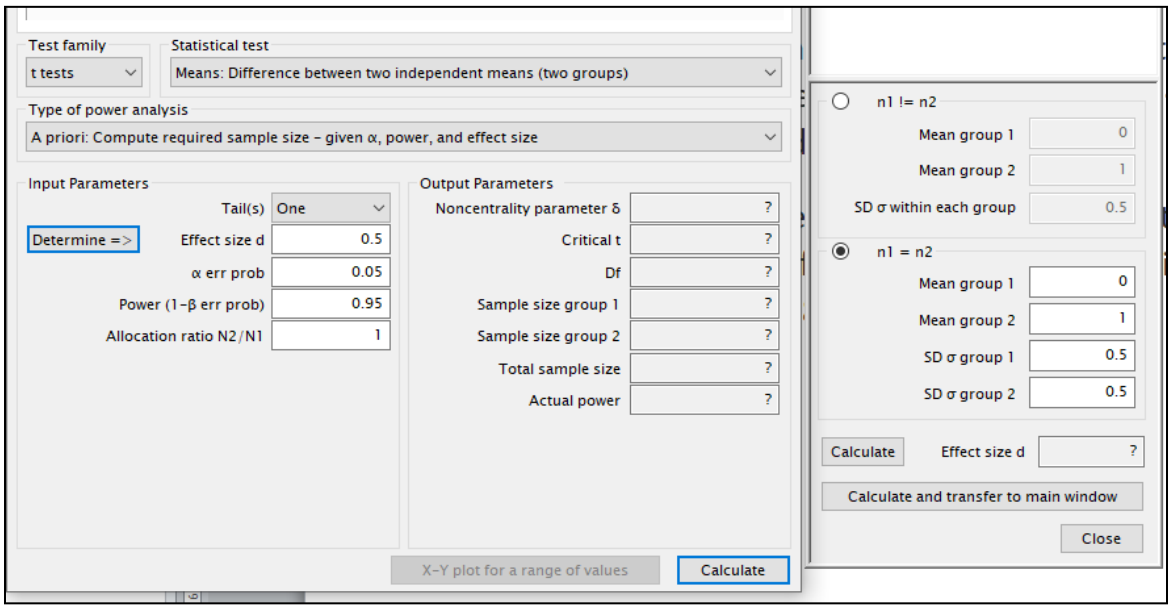


Figura 73. Estimación del tamaño del efecto *d* de Cohen

Capítulo 13. Diseño factorial: dos variables independientes, A x B

Dolores Frías-Navarro*
Marcos Pascual-Soler**
*Universidad de Valencia
**ESIC Business & Marketing School, España

Índice

Citar el capítulo como:

Frías-Navarro, D. y Pascual-Soler, M. (2021). Diseño factorial: dos variables independientes, A x B. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

La comprensión de los fenómenos psicológicos supone en muchas ocasiones analizar el efecto conjunto de varias variables dado que sólo su interacción puede explicar la ocurrencia o presencia de dichos fenómenos o conducta. En estos casos el investigador o investigadora tendrá que trabajar con los diseños denominados *factoriales*. El diseño factorial incluye el análisis simultáneo del efecto de dos o más variables independientes (designados *factores*) en una misma ecuación estructural.

La aplicación de este tipo de diseños permite analizar al mismo tiempo los efectos de varios factores sobre las respuestas de los sujetos y, además, y ahí radica su principal característica, permite investigar, si se planeta en la ecuación estructural, la influencia que pueden tener en las observaciones registradas o datos (variable dependiente) las *interacciones* entre dos o más variables independientes. Es decir, analiza los efectos debidos a que dos o más factores estén actuando simultáneamente sobre la variable dependiente. Este tipo de diseños son especialmente adecuados cuando el investigador o investigadora presupone o predice una relación entre dos o más variables independientes sobre la variable dependiente medida.

La inclusión de los factores dentro de un diseño se realiza siempre en función de consideraciones teóricas sobre posibles vínculos entre las variables independientes que son fuente de varianza sistemática primaria o sobre posibles fuentes de contaminación (varianza sistemática secundaria) que se desean controlar con el diseño de la investigación incluyendo dicha fuente en la ecuación estructural del modelo de diseño que se va a analizar (por ejemplo, los diseños de bloques).

Como ya se ha comentado, las variables que explican los cambios que se producen en la variable dependiente son múltiples, y a veces es necesario analizar el efecto conjunto de dos o más factores para poder comprender convenientemente los cambios que se ocasionan en la variable dependiente.

Hasta ahora, en el libro solamente se han considerado ejemplos donde el investigador o investigadora ha sometido a contraste una sola variable independiente (un factor, diseños unifactoriales univariados), que podría tener dos condiciones, tres, cuatro o más. Mediante la aplicación de un diseño factorial (por ejemplo, A *x* B, A *x* B *x* C, A *x* B *x* C *x* D …), en cambio, se pretende analizar el efecto simultáneo de dos o más variables independientes (dos o más factores, se trata de diseños

factoriales) sobre una variable dependiente medida (diseño univariado) o sobre más de una variable dependiente (diseño multivariado) (ver figura 74).

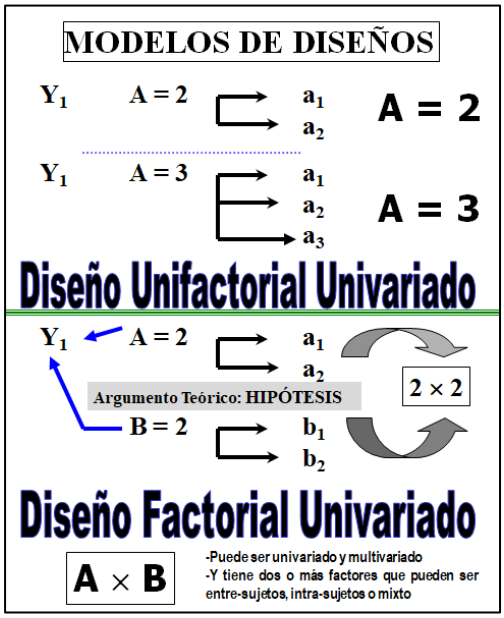


Figura 74. Diseño unifactorial y diseño factorial

El *diseño factorial completo* se caracteriza por requerir tantos grupos de tratamiento como posibles combinaciones entre los niveles de las variables independientes. En un diseño factorial completamente cruzado, las observaciones se realizan bajo todas las situaciones resultantes de combinar los niveles de un factor con los niveles de otro factor o bien con los niveles que resultan de la combinación de varios factores.

En el *diseño factorial incompleto* algunas de las combinaciones entre las condiciones de los factores están ausentes, no formándose todos los grupos que se derivan de la estructura de cruzamiento entre las variables manipuladas (por ejemplo, los diseños anidados o los diseños de cuadrado latino y diseños grecolatinos). Como consecuencia, el investigador o investigadora no puede estimar algunos de los efectos factoriales.

Los diseños factoriales están codificados en función del número de *factores* y del número de *niveles* de cada variable independiente. Por ejemplo, el diseño factorial más simple es el 2 *x* 2, es decir, el compuesto por dos variables independientes (o factores) con dos niveles cada una de ellas. Si se analizan dos factores, A con tres niveles o condiciones y B con cuatro, se tratará de un diseño factorial 3 *x* 4. Cuando

los niveles de los factores son los mismos también suelen codificarse mediante una potencia, como por ejemplo 2^3, diseño factorial de tres *factores* con dos niveles cada uno de ellos (representa lo mismo que la expresión 2 *x* 2 *x* 2). El número de grupos o condiciones experimentales resultantes de la aplicación de diseños factoriales será igual al producto de los niveles de los factores entre sí.

El *plan factorial* se expresa a partir del número de factores y número de niveles que tiene cada factor. Un diseño factorial completo 2^2 está compuesto por cuatro condiciones experimentales mientras que un diseño factorial completo 2^3 está formado por ocho grupos, representándose así todas las posibles combinaciones de los niveles de los factores entre sí.

Por lo tanto, para definir un diseño factorial se numeran los factores que se analizan y el número de condiciones experimentales (‘celdillas de interacción’) que configuran cada uno de estos factores. El número de condiciones experimentales resultantes de la aplicación de un determinado diseño factorial (las denominadas ‘celdillas de interacción’) será igual al producto de los niveles de los factores entre sí. En un diseño factorial 2 *x* 2, existirán cuatro condiciones experimentales: a_1b_1, a_1b_2, a_2b_1 y a_2b_2, resultado de combinar los dos niveles del factor A con los dos de B (ver figura 75).

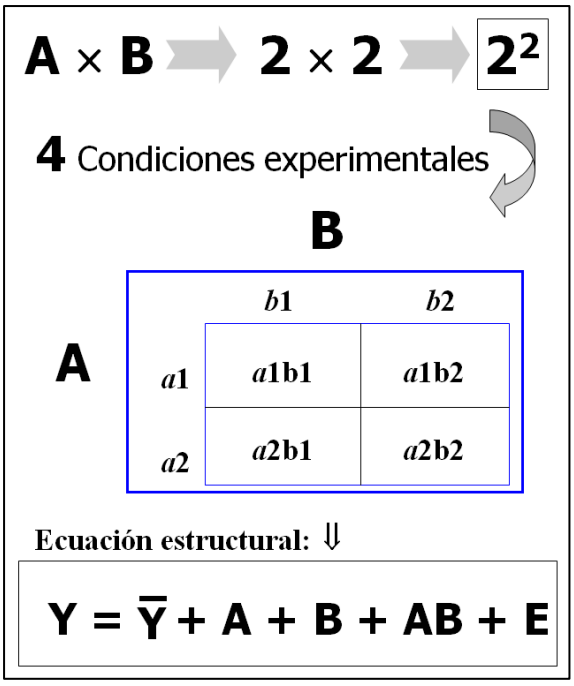


Figura 75. Tabla de datos de la interacción en un diseño 2 *x* 2

Para realizar el cálculo de los efectos que corresponden a cada fuente de varianza será necesario estimar las medias de los efectos principales (A y B) para cada una de sus condiciones o grupos (a1, a2 y b1, b2) y las medias del efecto de interacción (A x B) de cada una de las celdillas de interacción (a1b1, a1b2, a2b1, a2b2), junto con la media general o total (constante). Para ello, en cada ocasión, habrá que seleccionar el *n* que se corresponde con la media que se desea estimar en cada fuente de varianza. Por ejemplo, si se desea estimar la media de a1, hay que utilizar como numerador el número de observaciones que hay en a1. Y si se desea estimar la media de a1b1 hay que utilizar como numerador el número de observaciones que hay en esa interacción. En la figura 76 se describen todas las puntuaciones medias que hay que estimar en un diseño factorial A *x* B y, posteriormente, dichas medias se utilizarán para estimar los efectos de cada fuente de varianza que tiene el modelo completo o modelo de la hipótesis alternativa.

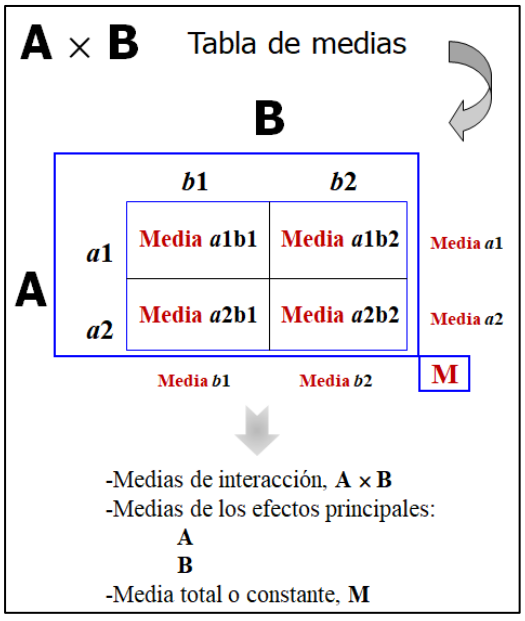


Figura 76. Medias que es necesario calcular en un diseño A *x* B

Ventajas del diseño factorial

La estrategia factorial tiene dos ventajas principales respecto a la unifactorial:

1) Aumenta la potencia estadística de la prueba de la razón *F*. Al analizar conjuntamente el efecto de dos o más variables independientes, si realmente estos factores están relacionados con la variable dependiente, el término residual será menor en el diseño factorial que si se analiza cada una de estas variables

unifactorialmente. Por lo tanto, el diseño factorial es, en principio, más potente (la potencia estadística es mayor, 1 - beta) que varios diseños unifactoriales, puesto que cada uno de los factores (si realmente están relacionados con la variable dependiente) reduce el componente residual del modelo factorial o término de error del modelo (denominador de la Razón *F*). Esa reducción del término de error se produce en el denominador del estadístico de la Razón *F* y por ello aumenta el valor de dicho estadístico, es decir, aumenta la varianza explicada o varianza atribuida a la fuente del efecto o tratamiento (numerador de la Razón *F*).

2) Plantea hipótesis teóricas de interacción entre las variables independientes. La segunda ventaja, ya no es únicamente cuantitativa en el sentido de aumentar la potencia estadística, sino que puede incluso variar la interpretación que se hace de la relación entre los factores y la variable dependiente. Si la relación que los factores tienen con la variable dependiente varía al manipularlos conjuntamente, quiere decir que el efecto de un factor depende de que el otro asuma ciertos valores. Es decir, lo importante no es valorar el efecto de cada factor de forma independiente sino explorar y valorar el efecto de interacción que se produce entre las condiciones de los factores sobre la puntuación obtenida en la variable dependiente. Cuando ocurre este fenómeno se produce un efecto en el modelo que se conoce con el término de *efecto de interacción*. Por el contrario, cuando en un diseño factorial se hace referencia al efecto de un factor individual (es decir, se analiza su efecto de forma independiente del resto de factores) se denomina *efecto principal* del factor en cuestión. Así, por ejemplo, en un diseño factorial A *x* B hay dos efectos principales (el de A y el de B) y un efecto de interacción (el de A *x* B). Cuando se aplica un diseño factorial, al investigador o investigadora realmente le interesa estudiar ese efecto de interacción (estará reflejado en la hipótesis del estudio), pero necesita estimar los dos efectos principales para poder estimar el efecto de interacción y por ello las tres fuentes de varianza son sometidas a contraste estadístico. Si el efecto de interacción no es estadísticamente significativo entonces se procederá con la interpretación de los efectos principales para comprobar si alguno de ellos o todos alcanzan un nivel de significación estadísticamente significativo ($p \leq$ alfa). En la redacción de los resultados se redactará el efecto de interacción como no estadísticamente significativo (por lo tanto, no se habrá comprobado la hipótesis del estudio que señalaba un posible efecto de interacción entre los factores) y, a continuación, se

centrará la redacción en el análisis de los efectos principales y si alguno de ellos tiene más de dos condiciones o niveles hay que recordar que será necesario ejecutar una prueba de hipótesis específicas o prueba post-hoc.

En definitiva, la complejidad de la conducta humana obliga a pensar que está determinada por múltiples causas conexas entre sí. Es por ello que el diseño factorial, que permite descubrir y analizar el sentido de las interacciones entre variables determinantes de la conducta, sea el diseño por excelencia y el de más amplio uso en el contexto de la Psicología experimental básica y aplicada.

Tipo de interacción entre los factores

El patrón interaccional es diverso en función de la relación que se detecte entre las variables o factores del modelo. Cuando existe interacción, las líneas se cruzan o convergen en algún punto, mientras que cuando no se produce el efecto de interacción, las líneas se mantienen paralelas, ya que la distancia entre las medias es constante. La representación gráfica de las puntuaciones medias de la interacción puede estar indicando un patrón de *interacción ordinal*, *interacción no ordinal* o *cruzada*, *interacción mixta* e *interacción no lineal* (véase Figura 77).

Una *interacción es ordinal* cuando el orden de superioridad de un factor sobre el otro se mantiene o es constante aunque el efecto cuantitativo puede variar. La interacción ordinal es *positiva* si se observa un crecimiento en el grupo mayor y una disminución en el menor y *negativa* si se produce un acercamiento entre los factores.

Un efecto de interacción se denomina *no ordinal* o *interacción cruzada* cuando el orden de superioridad entre los factores se cambia, no manteniéndose constante.

La *interacción mixta* es aquel efecto de interacción que siendo no ordinal presenta una tendencia clara hacia el cambio de orden, dando lugar a una interacción ordinal.

La *interacción no lineal se* caracteriza por la falta de vínculo lineal entre los factores.

La interacción ordinal y la no ordinal son las dos piezas claves del fenómeno de la interacción en los diseños factoriales.

La distinción entre los tipos de interacción es importante. Como Lubin (1961) señala, supongamos que una investigadora está interesado en analizar el efecto de dos drogas y dos tipos de terapia en relación al éxito de la intervención efectuada (diseño entre-sujetos 2 *x* 2). Si la relación gráfica entre las dos variables no es paralela, pero nunca llegan a cruzarse las líneas se tratará de un efecto de *interacción ordinal*. En este caso, para cada intervención realizada el efecto de la droga tendrá el mismo orden de superioridad o inferioridad en relación al factor terapia. Así, si la droga A_1 tiene un efecto mayor sobre la terapia B_1 también lo tendrá sobre la terapia B_2. La droga A_2 tendrá un efecto menor en ambos tipos de terapia. En este tipo de interacción ordinal las líneas representadas nunca se cruzarán.

Si se produce una interacción *no ordinal* o *cruzada* tiene que producirse un cruce entre dos o más factores. En este caso, los efectos de las drogas no mantienen el mismo orden para cada terapia. Puede ocurrir que la droga A_1 tenga un efecto mayor sobre la terapia B_1 pero menor sobre B_2 en comparación con la droga A_2.

En definitiva, cuando existe un efecto estadísticamente significativo de interacción, la diferencia de efectos entre combinaciones factoriales no es constante.

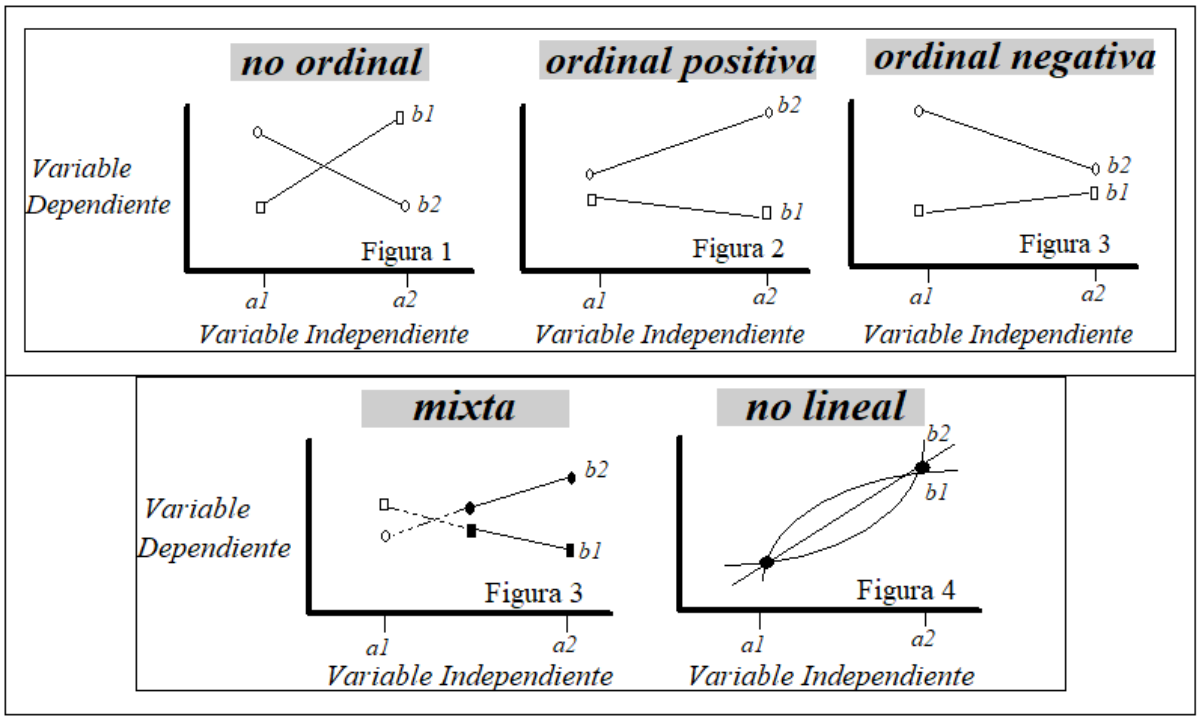


Figura 77. Representación del efecto de interacción

Modelo aditivo y no aditivo. Ecuación estructural

Para analizar el efecto conjunto o el efecto de interacción de dos factores A y B (se trata del diseño factorial más sencillo) sobre una única variable dependiente Y (diseño univariado) se define el *modelo completo* como un '*modelo de efectos no aditivos*' (modelo no aditivo) mediante la siguiente ecuación estructural:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB} + \mathbf{E}$$

Y = valores de la variable dependiente

M = media de la variable dependiente

A = efecto principal del primer factor, A

B = efecto principal del segundo factor, B

AB = efecto de interacción A × B

E = error de estimación del modelo

Podría ocurrir que en un diseño factorial se asumiese la no existencia del efecto de interacción entre los dos factores, ya que se estaría prediciendo la variable dependiente a partir de los efectos principales e independientes de los dos factores. En este caso se dice que se trata de un 'modelo de *efectos aditivos*' (sin el efecto de interacción entre los factores) porque los efectos del tratamiento y los de los bloques sobre la variable dependiente son aditivos. Así, cuando se habla de aditividad de los efectos se refiere a que no hay interacción entre la variable de tratamiento y la de bloques; en otras palabras, la relación entre los efectos o condiciones del tratamiento es la misma en cada uno de los bloques.

La ecuación estructural de un diseño de bloques es la siguiente:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{E}$$

Y = valores de la variable dependiente

M = media de la variable dependiente

A = efecto principal del primer factor, A

B = efecto principal del segundo factor, B

E = error de estimación del modelo

El modelo de efectos aditivos se desarrollará posteriormente cuando se detalle el modelo de bloques con dos factores (un factor de tratamiento y un factor de bloqueo) cuya ecuación estructural no incluye el efecto de interacción ya que se plantea un diseño sin efecto de interacción entre las condiciones de los factores. Esta asunción de ausencia del efecto de interacción se conoce como 'aditividad del modelo' (efecto de interacción con *p* > alfa; primer supuesto del diseño de bloques) y debe comprobarse ante de pasar a ejecutar el modelo factorial aditivo junto con comprobar también que el factor de bloqueo sí tiene un efecto estadísticamente significativo (efecto de la variable de bloqueo con *p* < alfa; segundo supuesto del diseño de bloques). Si en un diseño de bloques (modelo aditivo) se detectará un efecto de interacción entre los factores estadísticamente significativo entonces se habría producido un desajuste con la hipótesis teórica planteada ya que la hipótesis del diseño de bloques no hace referencia a un efecto de interacción entre los factores principales del modelo, pues parte de un modelo aditivo.

En cambio, cuando el modelo incluye el término de interacción, que por definición es un componente multiplicativo, se dice que se está en presencia de un modelo de *efectos no aditivos*. En ambos casos, los dos modelos son aditivos en la medida que constan de componentes que sumados entre sí predicen la variable dependiente; pero sólo el modelo no aditivo contiene un término multiplicativo (efecto de interacción), entendido como un componente aditivo más del modelo de predicción (figura 78).

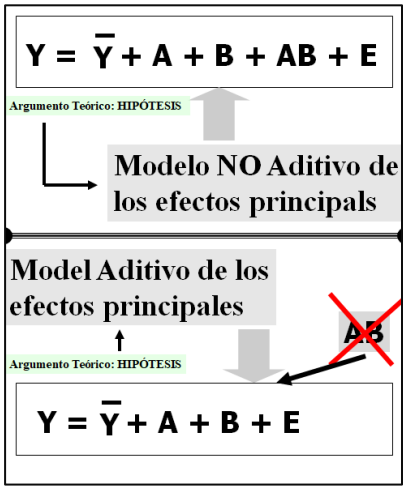


Figura 78. Ecuación estructural: modelo no aditivo y modelo aditivo

Modelo no aditivo. Efecto de interacción

Cuando se trata de un modelo no aditivo (es decir, el modelo que contiene el efecto de interacción entre las condiciones de los factores, A x B), el punto crucial de la explicación de los resultados vendrá determinado (siempre en primer lugar) por el estudio de la significación estadística del componente de interacción entre los factores (A *x* B). Un efecto de interacción entre los factores señala que el efecto de una condición de la variable independiente (A), por ejemplo, A1, sobre la variable dependiente (Y) no es el mismo en todos los niveles de la otra variable independiente (B), es decir, no tiene el mismo efecto en B1 que en B2. Por lo tanto, el efecto que tiene una condición de una de las variables independientes sobre la variable dependiente depende del nivel o condición que tenga la otra variable independiente (figura 79).

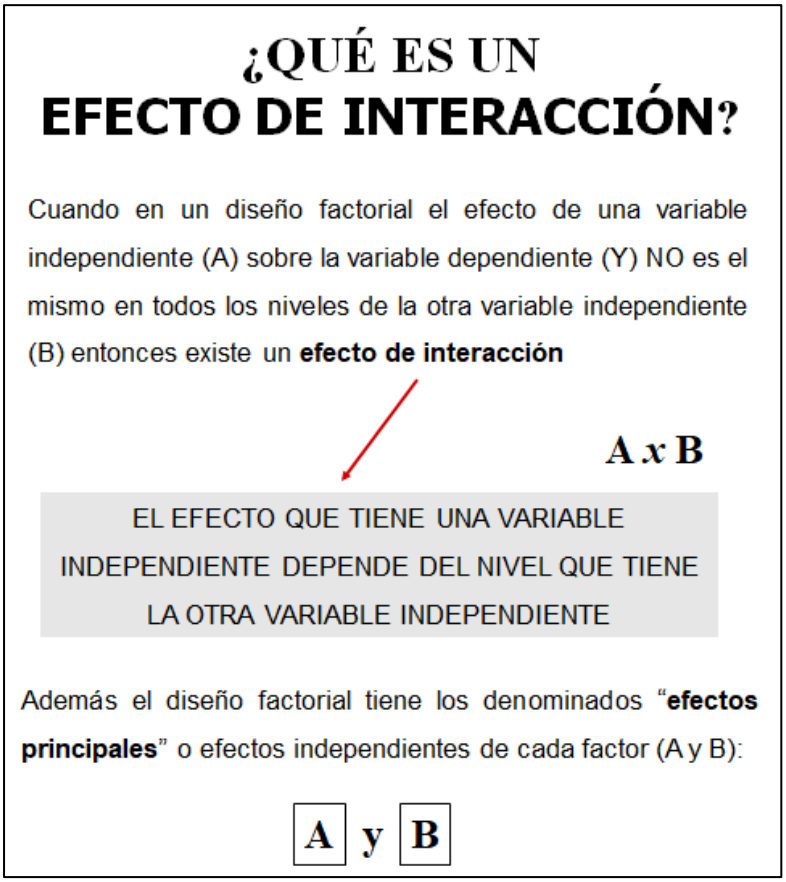


Figura 79. Efecto de interacción y efectos principales

Cuando se formula un modelo no aditivo, la hipótesis teórica plantea un efecto de interacción entre las variables independientes sobre la variable dependiente y, por

ello, es la fuente de varianza que debe interpretarse en primer lugar cuando ya se dispone de la tabla resumen de ANOVA.

Por lo tanto, cuando se plantea un diseño factorial no aditivo lo primero que hay que comprobar en la tabla de ANOVA es si el término de interacción es estadísticamente significativo y sí lo es hay que centrar la interpretación en el análisis y reflexión de dicho efecto. Efectivamente, no tendría sentido centrar la explicación desde el efecto principal de la variable A (o de la variable B) cuando el efecto detectado en la variable dependiente depende de ciertas condiciones definidas por los valores de la otra variable (efecto de interacción). Además, conviene recordar que si el efecto de interacción es estadísticamente significativo será necesario continuar con el análisis de las diferencias entre los pares de medias (pruebas de hipótesis específicas o pruebas post-hoc) o también se podría optar por aplicar un análisis de los efectos simples.

La ecuación estructural de un modelo de diseño factorial con interacción (modelo no aditivo) se representa junto con la estimación de los efectos principales, interacción y el error en la Figura 53. Como se observa, la estimación del efecto de interacción consiste en restar de la media de la condición de interacción la media general y los efectos principales de los factores implicados en la interacción que se está analizando. Así, el efecto de la fuente de varianza de interacción entre los factores A y B se estima como:

$$\mathbf{AB} = \mathbf{M_{ab}} - \mathbf{M} - \mathbf{A} - \mathbf{B}$$

Supongamos que se desea estimar el efecto de interacción de la celdilla a_1b_1 entonces habrá que calcular dicho efecto como:

$$\mathbf{A_1B_1} = \mathbf{M_{a1b1}} - \mathbf{M} - \mathbf{A_1} - \mathbf{B_1}$$

Es decir, la estimación del efecto de interacción de la celdilla a_1b_1 es igual a la media de la interacción a_1b_1 menos la media general o constante, menos el efecto principal de la condición a_1 y menos el efecto principal de la condición b_1. En este punto un error que se suele cometer al hacer la estimación del efecto de interacción es utilizar las puntuaciones medias de las condiciones a_1 y b_1 y no los efectos estimados de a_1 y b_1. Por lo tanto, precaución cuando se calcula el efecto de

interacción y recordar que se restan al final los efectos principales de las condiciones de cada factor que están implicadas en el efecto de interacción que se desea estimar.

En la figura 80 se observa que el término de error del modelo de diseño planteado siempre es la puntuación directa obtenida menos la puntuación predicha por el modelo formulado.

Como se está presentando el diseño entre-grupos o entre-sujetos, se ha anotado que el error es S / A *x* B, es decir, el error hace referencia a aquella parte de la variabilidad que se atribuye al error aleatorio, a las diferencias individuales de los participantes. En el diseño entre-grupos los sujetos están anidados (representado por la línea vertical, S /) en una cierta combinación de las condiciones del factor A y el factor B. De ahí que se represente el error como S / A *x* B: sujetos anidados o ligados a una combinación concreta de los factores del diseño entre-grupos.

Teniendo en cuenta que la puntuación predicha siempre es la media general más los efectos que plantea el modelo de diseño en su ecuación estructural entonces en este modelo no aditivo A *x* B, el error es igual a la puntuación obtenida en Y, menos la media general o constante, menos los efectos de A, de B y menos el efecto de la interacción A x B. Conviene recordar que la puntuación predicha siempre es la media general o constante más los efectos que se plantean en la ecuación estructural (en este caso concreto son A, B y AB).

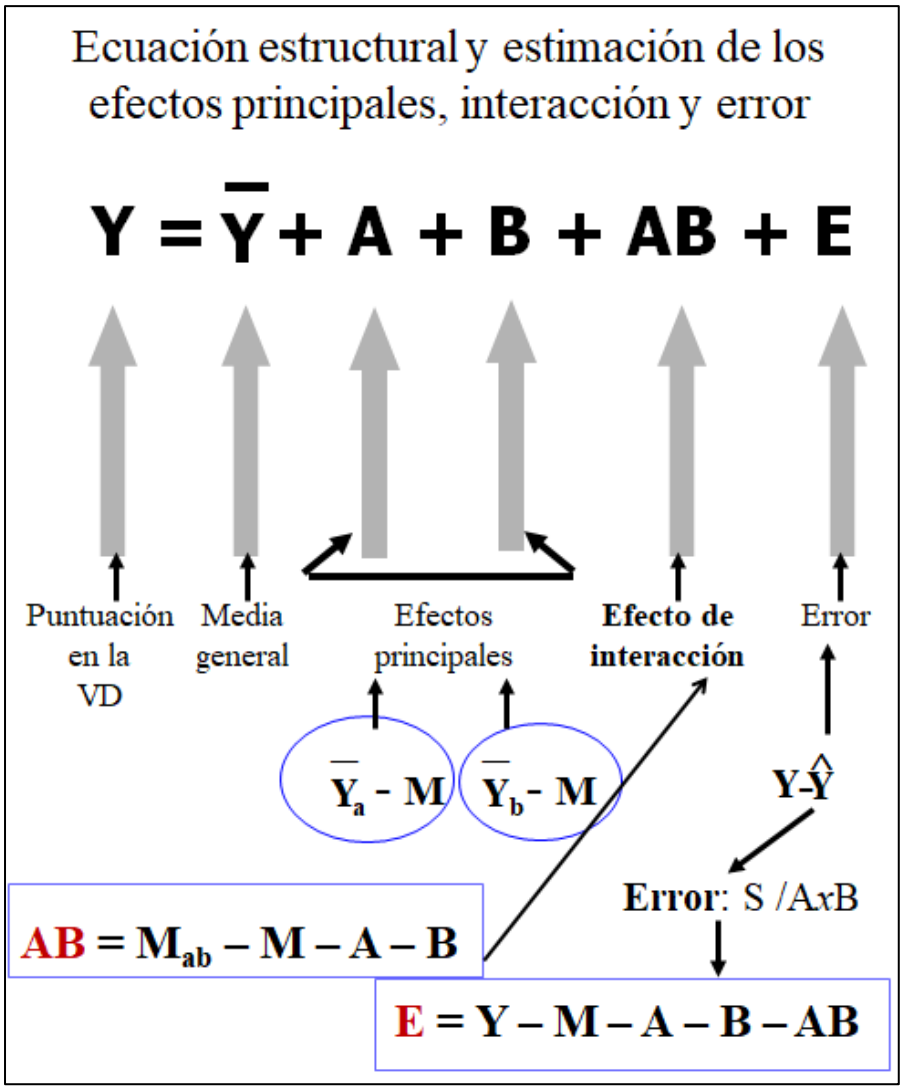


Figura 80. Ecuación estructural del diseño A *x* B y estimación de los efectos

A continuación se presenta un esquema con los conceptos básicos del modelo factorial no aditivo y aditivo tal y como ya se realizó anteriormente en el modelo de diseño unifactorial (ver Figura 81). Cada uno de los elementos que definen a los dos modelos (completo no aditivo y completo aditivo) se explica posteriormente con un ejercicio que plantea un supuesto de investigación.

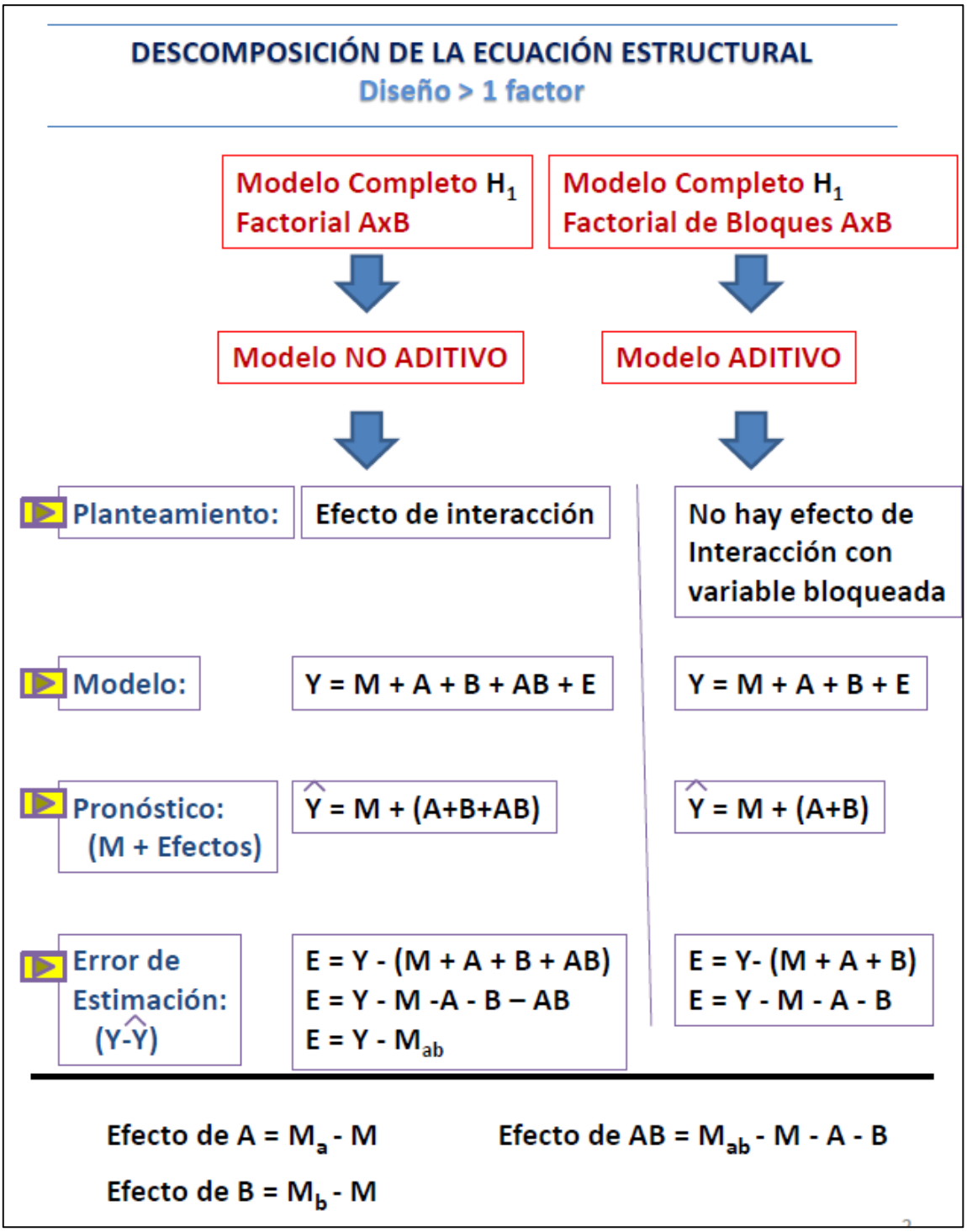


Figura 81. Descomposición de la ecuación estructural y modelos: aditivo y no aditivo

Desarrollo de un supuesto de investigación (A *x* B)

A continuación se plantea un supuesto de investigación con una hipótesis teórica de interacción entre dos variables independientes.

Supuesto de investigación. Una investigadora desea comprobar si la memoria está relacionada con la motivación y el estrés percibido por los individuos. Para poder contrastar dicho supuesto, diseña una investigación en la que manipula simultáneamente las variables de *Estrés* y *Motivación* (*bajo* y *alto*, en los dos factores).

Antes de la prueba de recuerdo, la mitad de los participantes son sometidos a un cuestionario de preguntas matemáticas: en un caso las preguntas tienen una solución (estrés bajo) y en el otro grupo las preguntas no tienen una solución (y los sujetos son 'ciegos' a esta circunstancia, es decir, no saben que son preguntas sin una solución) y provoca que aumente el estrés de quienes se encuentran en esta condición ya que insisten e insisten en buscar una solución, pero no la hayan (estrés alto). Además, a la mitad de los sujetos se les ha notificado que recibirán una recompensa económica sustanciosa si resuelven adecuadamente la prueba de recuerdo de un listado de 20 palabras que posteriormente se les entregará (se trata de motivarles; motivación alta) y a la otra mitad no se les ha dicho nada (motivación baja). El diseño utiliza dos participantes en cada grupo experimental, asignandolos aleatoriamente a las condiciones experimentales. Los ocho participantes son varones seleccionados aleatoriamente de una facultad de economía de la Universidad de Navarra.

Una vez que los participantes han sido sometidos a una de las condiciones de interacción del diseño (diseño entre-sujetos), los resultados en el número de errores cometidos al recordar el listado son los que se detallan en la tabla 14.

Tabla 14. Resultados del experimento de errores y medias

		(A) Estrés		
		a_1 *bajo*	a_2 *alto*	*Medias B*
(*B*)	b_1	13	15	13
	baja	11	13	
		Media a1b1: 12	*Media* a2b1: 14	
Motivación	b_2	1	11	7
	alta	3	13	
		Media a1b2: 2	*Media* a2b2: 12	
	Medias A	7	13	10

Una vez leído el supuesto de investigación es importante que el lector o lectora reflexione sobre el tipo de hipótesis de investigación planteada, el tipo de diseño de investigación que es apropiado para analizar dicha hipótesis, las variables dependientes e independientes que se utilizan y qué constructos representan y cómo se operacionalizan en el diseño, qué variables se han controlado y qué técnicas de control se han aplicado en el estudio (si las hay), la metodología que se utiliza, el número de observaciones total, el número de observaciones por condición o grupo intraceldilla (¿el diseño es ortogonal o no?) y el número de observaciones que hay en cada factor, qué valor de alfa se fija a priori como probabilidad del error de tipo I.

Si en el texto del supuesto no se dice nada sobre el valor de alfa se asume que es .05 ya que si es otro valor, como por ejemplo alfa = .01, sería necesario que el investigador o investigadora lo especificase en el texto del supuesto de la investigación de manera explícita. Del mismo modo, si en el texto del supuesto de investigación no se menciona nada sobre si el contraste de hipótesis es unidireccional o bidireccional entonces se asume que lo ha realizado de forma bilateral o a dos colas. Si fuese de una cola o unidireccional (la hipótesis marca claramente la dirección de la diferencia entre las medias) entonces es necesario que el investigador o investigadora lo resalte de forma explícita cuando describe el diseño de su estudio. A veces se plantean hipótesis bidireccionales y el investigador o investigadora marca la dirección de la diferencia según el conocimiento previo que ha revisado, pero desea aplicar hipótesis bidireccionales ya que no renuncia a observar también qué ocurre si la dirección de la diferencia es la contraria. En este caso, la elección del análisis es llevar a cabo el contraste de hipótesis estadísticas observando las dos colas de la distribución de la hipótesis nula ya que la diferencia podría ir en una dirección (grupo primero mayor que el segundo) o al revés (grupo segundo mayor que el primero). En general, en Psicología, en la mayoría de las ocasiones, se lleva a cabo el análisis con pruebas bilaterales y los investigadores o investigadoras no suelen mencionarlo en sus artículos ya que se asume que si no se dice nada esa es la forma como han actuado cuando han tomado la decisión estadística.

El primer paso para poder realizar el análisis de la varianza es conocer las puntuaciones medias de cada una de las condiciones de los efectos principales y las medias de interacción de las celdillas de interacción que se producen al cruzar las

condiciones de los factores del diseño, situando además la media general o constante (tabla 15). En la tabla anterior ya se habían incluido, pero se detallan ahora para recordar la importancia de calcular las medias como primer paso para estimar posteriormente cada uno de los efectos que tiene el modelo. En la tabla de medias se puede observar la dirección de las medias, permitiendo observar si van en la dirección propuesta por el modelo teórico revisado. Así se comprueba que los sujetos que han recibido la condición de estrés bajo y motivación alta son los que menos errores producen (por lo tanto, tienen un nivel mayor de memoria). De momento las puntuaciones siguen el planteamiento de la hipótesis teórica, pero falta comprobar si las medias difieren de forma estadísticamente significativa, y si difieren todos los pares de medias o solamente algunos de ellos. Para ello es necesario ejecutar un análisis de varianza con el proceso de decisión estadística.

Tabla 15. Tabla de medias de errores cometidos

		(A) Estrés		
		a_1 *bajo*	a_2 *alto*	*Medias B*
(*B*)	b_1 *baja*	12	14	13
Motivación	b_2 *alta*	2	12	7
	Medias A	7	13	10

Por lo tanto, la ecuación estructural que habrá que descomponer para resolver el supuesto de investigación planteado es la siguiente:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB} + \mathbf{E}$$

Efectos principales

Se pueden estimar los efectos principales del *bajo* y *alto* nivel de estrés (factor principal A) sobre el número de errores cometidos al recordar el listado a través de:

$$\mathbf{A} = \mathbf{M}_a - \mathbf{M} = \begin{pmatrix} 7 \\ 7 \\ 7 \\ 7 \\ 13 \\ 13 \\ 13 \\ 13 \end{pmatrix} - \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} = \begin{pmatrix} -3 \\ -3 \\ -3 \\ -3 \\ 3 \\ 3 \\ 3 \\ 3 \end{pmatrix} \rightarrow \begin{pmatrix} a_1 \\ a_1 \\ a_1 \\ a_1 \\ a_2 \\ a_2 \\ a_2 \\ a_2 \end{pmatrix}$$

El efecto de $\hat{\alpha}_1$ es de -3 puntos y el de $\hat{\alpha}_2$ de 3 puntos. Se comprueba que la suma de los efectos es cero. Por tanto, en la muestra se aprecia que el grupo sometido a un *estrés alto* aumenta su promedio de errores al recordar el listado en 3 puntos respecto de la media general, mientras que los que están en la condición de *bajo estrés* reducen tres puntos su puntuación promedio en el conteo de los errores cometidos al recordar el listado.

Los grados de libertad correspondientes al término del efecto principal A serán igual a 1 ($gl_A = a - 1 = 2 - 1 = 1$).

Se puede observar igualmente cómo afecta la motivación (factor principal B) a la cantidad de errores cometidos al recordar las palabras del listado:

$$\mathbf{B} = \mathbf{M}_b - \mathbf{M} = \begin{pmatrix} 13 \\ 13 \\ 7 \\ 7 \\ 13 \\ 13 \\ 7 \\ 7 \end{pmatrix} - \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ -3 \\ -3 \\ 3 \\ 3 \\ -3 \\ -3 \end{pmatrix} \rightarrow \begin{pmatrix} b_1 \\ b_1 \\ b_2 \\ b_2 \\ b_1 \\ b_1 \\ b_2 \\ b_2 \end{pmatrix}$$

Los cuatro casos en que la motivación es *baja* (b_1), aumentan los errores en el recuerdo de las palabras del listado en promedio 3 puntos ($\hat{\beta}_1 = 3$) respecto a la media general de errores de recuerdo cometidos. Si la motivación es *alta*, entonces disminuyen los errores en el recuerdo del listado en otros 3 puntos ($\hat{\beta}_2 = -3$) sobre el promedio general. Se observa de nuevo que la suma de los efectos del factor B es cero.

Los grados de libertad correspondientes al término del efecto principal B serán igual a 1 ($gl_B = b - 1 = 2 - 1 = 1$).

Efectos de interacción en el modelo no aditivo

Para estimar el efecto de interacción AB se resta de la media de cada celdilla de interacción la media general o constante y los efectos principales de los factores que se han definido en la ecuación estructural; en este diseño son A y B.

Por lo tanto:

AB = $\mathbf{M}_{ab}$ – (M + EFECTOS PLANTEADOS EN EL MODELO O ECUACIÓN ESTRUCTURAL)

Es decir,

$$\mathbf{AB} = \mathbf{M}_{ab} - \mathbf{(M + A + B)}$$

De lo que se deduce directamente que:

$$\mathbf{AB} = \mathbf{M}_{ab} - \mathbf{M} - \mathbf{A} - \mathbf{B}$$

Para obtener los efectos de interacción a partir de los datos del supuesto que se está desarrollando:

$$\mathbf{AB} = \mathbf{M}_{ab} - \mathbf{M} - \mathbf{A} - \mathbf{B} =$$

$$= \begin{pmatrix} 12 \\ 12 \\ 2 \\ 2 \\ 14 \\ 14 \\ 12 \\ 12 \end{pmatrix} - \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} - \begin{pmatrix} -3 \\ -3 \\ -3 \\ -3 \\ 3 \\ 3 \\ 3 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3 \\ -3 \\ -3 \\ 3 \\ 3 \\ -3 \\ -3 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ -2 \\ -2 \\ -2 \\ -2 \\ 2 \\ 2 \end{pmatrix} \rightarrow \begin{pmatrix} ab_{11} \\ ab_{11} \\ ab_{12} \\ ab_{12} \\ ab_{21} \\ ab_{21} \\ ab_{22} \\ ab_{22} \end{pmatrix}$$

Los grados de libertad correspondientes al término de interacción se calculan multiplicando los grados de libertad de los efectos principales implicados; en este caso los grados de libertad serán igual a 1 ($gl_{AB} = gl_A\ gl_B = 1 \times 1 = 1$).

Por lo tanto, ahora ya se puede completar la tabla de efectos de un diseño factorial A *x* B (tabla 16). Los efectos que están en gris están determinados (se pueden deducir ya que todos los efectos deben sumar 0) dados los grados de libertad que tiene cada fuente de varianza.

Tabla 16. Tabla de efectos

		(A) Estrés		
		α_1 bajo	α_2 alto	*Efectos B*
(*B*)	β_1 *baja*	$\alpha_1 \beta_1$ 2	$\alpha_2 \beta_1$ -2	β_1 3
Motivación	β_2 *alta*	$\alpha_1 \beta_2$ -2	$\alpha_2 \beta_2$ 2	β_2 -3
	Efectos A	α_1 -3	α_2 3	M = 10

Siguiendo con el análisis del efecto de interacción, ahora habrá que comprobar si estas diferencias entre las celdillas de interacción son o no estadísticamente significativas. Para ello se procede con la estimación de la fuente de varianza del error y, posteriormente, se completa el análisis de la varianza (ANOVA) con el estadístico de la razón *F*. En este diseño con dos efectos principales y un efecto de interacción será necesario calcular 3 razones *F* con sus correspondientes valores de *p* de probabilidad del efecto detectado, asumiendo que la hipótesis nula es cierta; es decir, una prueba *F* para cada fuente de varianza y se lleva a cabo la decisión dicotómica de mantener o rechazar la hipótesis nula en cada uno de los tres casos.

Error de estimación

El error de estimación que se comete está definido por la diferencia entre la la puntuación obtenida en el estudio y la puntuación pronosticada por el modelo de diseño planteado en el estudio.

Respecto a la puntuación pronosticada, en un diseño A *x* B la estimación del valor de la variable dependiente se hace a partir de la suma de la media general **M** y los efectos de **A**, **B** y **AB** que se han planteado en la ecuación estructural. Por lo tanto, $\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB}$.

$$\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB}$$

Se puede analizar cuál será el pronóstico de cada una de las cuatro situaciones experimentales que tiene el diseño A *x* B del estudio planteado en el supuesto de investigación:

Estrés *bajo* / Motivación *baja* (a_1b_1):

$$\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB} = \bar{Y} + \hat{\alpha}_1 + \hat{\beta}_1 + \hat{\alpha\beta}_{11} =$$

$$= 10 + -3 + 3 + 2 = 12$$

Estrés *bajo* / Motivación *alta* (a_1b_2):

$$\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB} = \bar{Y} + \hat{\alpha}_1 + \hat{\beta}_2 + \hat{\alpha\beta}_{12} =$$

$$= 10 + -3 + -3 + -2 = 2$$

Estrés *alto* / Motivación *baja* (a_2b_1):

$$\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB} = \bar{Y} + \hat{\alpha}_2 + \hat{\beta}_1 + \hat{\alpha\beta}_{21} =$$

$$= 10 + 3 + 3 + -2 = 14$$

Estrés *alto* / Motivación *alta* (a_2b_2):

$$\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB} = \bar{Y} + \hat{\alpha}_2 + \hat{\beta}_2 + \hat{\alpha\beta}_{22} =$$

$$= 10 + 3 + -3 + 2 = 12$$

Se observa que cuando el modelo de diseño es un diseño entre-sujetos factorial no aditivo A *x* B, la puntuación pronosticada para cada uno de los cuatro efectos de interacción es la media de interacción de cada grupo. Sin embargo, conviene tener presente que eso sucede solamente en este diseño, pues si el diseño se plantea como aditivo entonces ya no se produce esa coincidencia tal y como se detallará después.

Siguiendo con los datos del supuesto de investigación, se observa que los participantes del grupo de estrés bajo y motivación alta (a_1b_2) son los que tienen menos errores al recordar el listado de las palabras. Y cuando el estrés es alto y la motivación es baja se produce el mayor número de errores al recordar el listado de las palabras.

Por lo tanto, el error de estimación que se comete para cada una de las observaciones que se han recogido en el estudio es el siguiente:

$$\mathbf{E} = \mathbf{Y} - \mathbf{M} - \mathbf{A} - \mathbf{B} - \mathrm{A}\mathbf{B} =$$

$$= \begin{pmatrix} 13 \\ 11 \\ 1 \\ 3 \\ 15 \\ 13 \\ 11 \\ 13 \end{pmatrix} - \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} - \begin{pmatrix} -3 \\ -3 \\ -3 \\ -3 \\ 3 \\ 3 \\ 3 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3 \\ -3 \\ -3 \\ -3 \\ -3 \\ 3 \\ 3 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \\ -2 \\ -2 \\ -2 \\ -2 \\ 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} S1ab_{11} \\ S2ab_{11} \\ S3ab_{12} \\ S4ab_{12} \\ S5ab_{21} \\ S6ab_{21} \\ S7ab_{22} \\ S8ab_{22} \end{pmatrix}$$

Se observa de nuevo que la suma de todos los errores de estimación es igual a cero. También se puede comprobar que dentro de una celdilla concreta la suma de los errores estimados también debe dar cero. Es decir, si hay dos participantes que están en la misma condición de tratamiento (por ejemplo, a_1b_1), por qué sus puntuaciones directas difieren (puntuaciones de 13 y 11) si tienen la misma media general, el mismo efecto de A (a_1) y el mismo efecto de B (b_1), pues por las diferencias individuales y/o el error aleatorio y de ahí que el valor del término de error es propio y vinculado a cada puntuación directa. Y el sumatorio de los errores dentro de cada celdilla (errores intra-celdilla) debe ser igual a cero si el diseño es ortogonal. Es decir, si se trata de un diseño ortogonal ($n_{a1b1} = n_{a1b2} = n_{a2b1} = n_{a2b2}$) entonces los grados de libertad del error se pueden calcular como $(n - 1)ab$, siendo n el número de observaciones dentro de cada celdilla de interacción y ab el resultado de la multiplicación de A *x* B. Si el diseño no es ortogonal entonces los grados de libertad del error se pueden calcular como $N - ab$, siendo N el número total de observaciones. Las dos formulas de los grados de libertad ofrecen el mismo resultado, por supuesto.

En resumen, el desarrollo de la ecuación estructural (modelo no aditivo) para cada una de las observaciones del estudio es la siguiente:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB} + \mathbf{E}$$

$$\begin{pmatrix} 13 \\ 11 \\ 1 \\ 3 \\ 15 \\ 13 \\ 11 \\ 13 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{pmatrix} + \begin{pmatrix} -3 \\ -3 \\ -3 \\ -3 \\ 3 \\ 3 \\ 3 \\ 3 \end{pmatrix} + \begin{pmatrix} 3 \\ 3 \\ -3 \\ -3 \\ -3 \\ -3 \\ 3 \\ 3 \end{pmatrix} + \begin{pmatrix} 2 \\ 2 \\ -2 \\ -2 \\ -2 \\ -2 \\ 2 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$

Sumas de cuadrados

Una vez ya se dispone de la información vinculada a la descomposición de la ecuación estructural en cada uno de sus componenetes, ya se procede con el cálculo de las sumas de cuadrados (SC) de los efectos de cada fuente de varianza planteada en el modelo de diseño de investigación: A, B, AB y E.

Suma de cuadrados del efecto de A, nivel de estrés:

$$\mathbf{SC}_A = \mathbf{A}'\,\mathbf{A} = \begin{pmatrix} -3 & -3 & -3 & -3 & 3 & 3 & 3 & 3 \end{pmatrix} \begin{pmatrix} -3 \\ -3 \\ -3 \\ -3 \\ 3 \\ 3 \\ 3 \\ 3 \end{pmatrix} = 72$$

Suma de cuadrados del efecto de B, nivel de motivación:

$$\mathbf{SC}_B = \mathbf{B}'\,\mathbf{B} = \begin{pmatrix} 3 & 3 & -3 & -3 & 3 & 3 & -3 & -3 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \\ -3 \\ -3 \\ 3 \\ 3 \\ -3 \\ -3 \end{pmatrix} = 72$$

Suma de cuadrados del efecto de interacción AB, nivel de estrés por nivel de motivación:

$$\mathbf{SC}_{AB} = \mathbf{AB}'\,\mathbf{AB} = \begin{pmatrix} 2 & 2 & -2 & -2 & -2 & -2 & 2 & 2 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \\ -2 \\ -2 \\ -2 \\ -2 \\ 2 \\ 2 \end{pmatrix} = 32$$

Suma de cuadrados de la fuente de varianza del error, E:

$$\mathbf{SC}_E = \mathbf{E}'\,\mathbf{E} = \begin{pmatrix} 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} = 8$$

A continuación, se aplica la prueba de la hipótesis para determinar si la variabilidad correspondiente al término de interacción permite desechar el modelo aditivo y mantener la validez de un modelo con término de interacción o modelo no aditivo (tabla 17). Se trata de la prueba de aditividad del modelo. Previamente se dividen las Sumas de Cuadrados (SC) por su grados de libertad (gl) y se obtienen las denominadas Medias Cuadráticas (MC) y ya se puede calcular el valor del estadístico de la razón $F = MC_{efecto} / MC_{error}$ para cada una de las fuentes de varianza entre-grupos o del efecto (A, B y A *x* B) planteadas en el modelo de diseño del estudio.

Puede comprobarse en la Tabla 17 que existen diferencias estadísticamente significativas ($p < .05$) para el término de interacción entre los dos factores principales del modelo de diseño planteado (efecto A *x* B). Por tanto, el modelo que se deriva del análisis estadístico no es el de efectos aditivos sino el modelo con efectos de interacción o modelo no aditivo tal y como planteaba la hipótesis teórica del estudio.

Tabla 17. Diseño factorial 2 × 2 entre estrés y motivación. Modelo con interacción

Fuente	*SC*	*gl*	*MC*	*F*	*p*	η^2
A	72.00	1	72.00	36.0	< .05	.391
B	72.00	1	72.00	36.0	< .05	.391
A *x* B	32.00	1	32.00	16.0	< .05	.174
Error	8.00	4	2.00			
Total	184.000	7				

El modelo que mejor se ajusta a los datos supone que la cantidad de recuerdo de las palabras dependerá no sólo del mero efecto independiente del alto o bajo nivel de estrés y motivación; también estará relacionada con el efecto particular que cada combinación de motivación y estrés alto o bajo ocasiona a los participantes.

El cálculo del valor *p* de probabilidad de cada fuente de varianza se ha realizado consultando las tablas de la Razón *F* (.05, 1, 4) = 7.709. Sería necesario buscar en las tablas el valor teórico de *F* que corresponde a cada fuente de varianza, pero como en este diseño en los tres casos los grados de libertad son 1 para el efecto y 4 para el error pues con un valor teórico es suficiente ya que se utiliza el mismo en los tres casos. Se observa que los valores de las tres *F* empíricas (F_A1, 4) = 36, F_B1, 4, F_{AB}1, 4) = 16) son mayores a 7.709, luego en los tres casos se rechaza la hipótesis nula. Sin embargo, se procederá con la interpretación teórica del efecto de interacción ya que las puntuaciones en la variable dependiente dependen de las condiciones del factor de A y del factor de B conjuntamente.

La mejor ayuda para interpretar el efecto de interacción consiste en realizar una representación gráfica del mismo, en la que se sitúa en el eje de las abscisas los niveles de uno de los factores y en las ordenadas las puntuaciones en la variable dependiente de cada condición experimental (ver Figura 82). Se observa en la gráfica un efecto de interacción entre los niveles o condiciones de las variables independientes (interacción ordinal negativa). Se comprueba que los participantes que se someten a la situación de estrés bajo y con alta motivación son los que menos errores producen al recordar el listado de las palabras. Se podrían intercambiar los ejes y realizar otra representación gráfica. Las dos gráficas son adecuadas y el investigador o investigadora elegirá para su informe aquella gráfica que considere que explica o representa mejor el efecto de interacción de los factores.

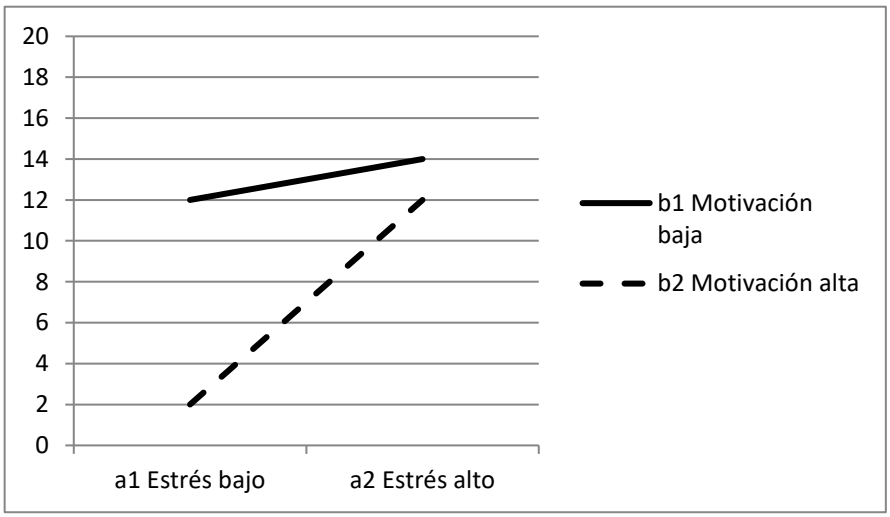


Figura 82. Representación gráfica del efecto de interacción AB

Contraste de hipótesis específicas

También en el diseño factorial, al igual que cuando se aplican diseños unifactoriales, si se rechaza la hipótesis nula hay que aplicar contrastes específicos para determinar entre qué medias se producen diferencias estadísticamente significativas, controlando la *tasa de error de Tipo I por experimento*. Las comparaciones en el diseño factorial se aplican de forma análoga al diseño unifactorial.

Como no se ha concretado ninguna hipótesis específica en el enunciado del supuesto de investigación y todas las comparaciones que se realizan incumben únicamente a las existentes entre dos grupos parece oportuno aplicar la prueba de Tukey ya que es la que tiene la máxima potencia estadística en este tipo de situaciones y controla de forma adecuada la tasa de error de tipo I.

$$|\overline{Y}_g - \overline{Y}_h| \geq \frac{q_{(\alpha,\, a\, b,\, gl_{error})}}{\sqrt{2}} \sqrt{MC_{error} \sum_{i=1,\, j=1}^{a\ b} \frac{c_{ij}^2}{n_{ij}}}$$

Se consulta la tabla de Tukey del rango estudentizado y se observa que para *q*(alfa, ab, g_{lerror}: *q*(.05, 4, 4) el valor de *q* es 5.757 (figura 83). Posteriormente se aplica la formula anterior de Tukey y se obtiene que el valor del Rango Crítico para comprobar si la diferencia entre los pares de medias es o no es estadísticamente significativa es igual a 82.

Estadístico del rango estandarizado

Tabla *Comparaciones múltiples del Test de Tukey (**valores q**)*

gl		número de grupos													
error	α	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.05	17.97	36.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07	50.6	52.0	53.2	54.3	55.4
	.01	90.03	135.0	164.3	185.6	202.2	215.8	227.2	237.0	245.6	253	260	266	272	277
2	.05	6.085	8.331	9.798	10.880	11.732	12.436	13.026	13.539	13.989	14.392	14.746	15.071	15.370	15.646
	.01	14.04	19.02	22.29	24.720	26.63	28.20	29.53	30.68	31.69	32.59	33.44	34.11	34.82	35.39
3	.05	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462	9.714	9.941	10.149	10.340	10.516
	.01	8.261	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69	17.13	17.51	17.86	18.19	18.51
4	.05	3.927	5.040	**5.757**	6.287	6.707	7.053	7.347	7.602	7.826	8.026	8.208	8.371	8.522	8.662
	.01	6.512	8.120	9.173	9.958	10.584	11.097	11.542	11.928	12.268	12.579	12.837	13.082	13.314	13.527

Figura 83. Consulta en la tabla de Tukey del valor de *q*(alfa, ab, gl_{error})

Si se ajusta la prueba de Tukey a los datos del supuesto de investigación se puede comprobar que:

$$|\overline{Y}_g - \overline{Y}_h| \geq \frac{q_{(0.05,\, 4,\, 4)}}{\sqrt{2}} \sqrt{2\left(\frac{1^2}{2} + \frac{-1^2}{2} + \frac{0^2}{2} + \frac{0^2}{2}\right)} \Rightarrow$$

$$\Rightarrow \frac{5.757}{\sqrt{2}}\sqrt{2} = 5.757$$

En este diseño el número total de comparaciones posibles simples dos a dos no redundantes entre las medias de las condiciones es igual a seis ya que:

$$C = \frac{ab(ab-1)}{2} = \frac{4(3)}{2}6$$

Se puede realizar una tabla de diferencias de medias para comprobar qué diferencia supera o iguala al valor del Rango Crítico de Tukey de 5.757 (tabla 18). Se observa que tres diferencias de medias de interacción superan dicho valor.

Tabla 18. Tabla de diferencia de medias entre las condiciones de interacción AB

Medias	a1b1 Media=12	a1b2 Media=2	a2b1 Media=14
a1b2: media 2	**10**	-	-
a2b1: media 14	2	**12**	-
a2b2: media 12	0	**10**	2

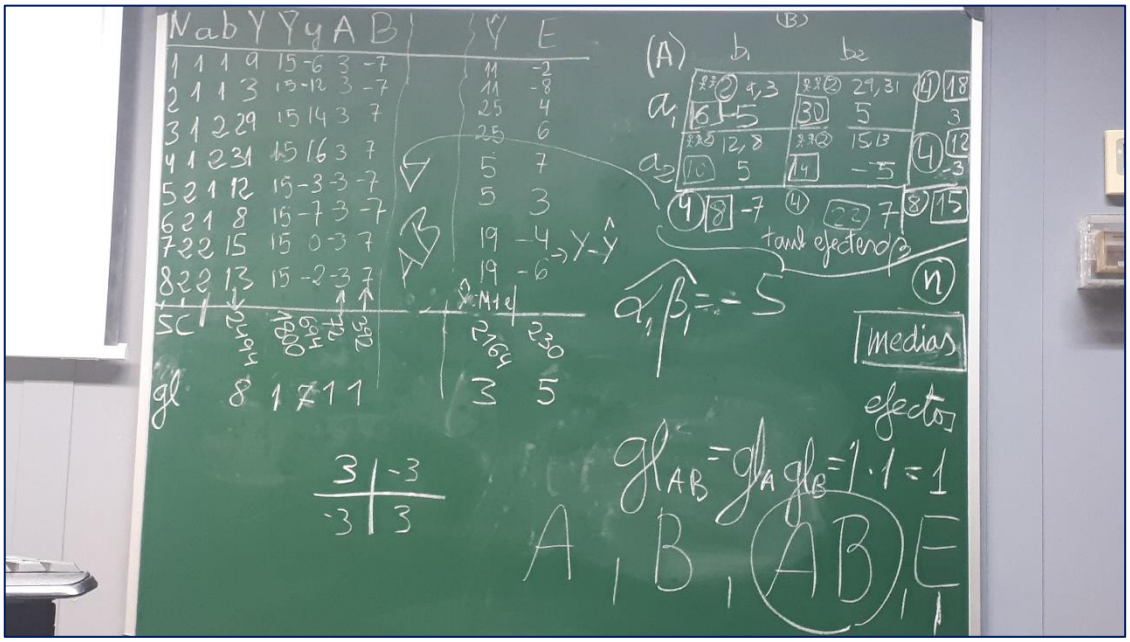
Analizada la distancia o la diferencia que existe entre las seis comparaciones simples, consideradas dos a dos, se puede comprobar que únicamente existen diferencias estadísticamente significativas entre el grupo con *bajo* estrés y *alta* motivación (Media de a_1b_2 = 2) respecto de las puntuaciones medias de las otras tres situaciones experimentales. Es decir, el valor crítico de Tukey de 5.757 únicamente se supera o se iguala cuando se compara la diferencia de medias de la condición que tiene la media de a_1b_2 (grupo de estrés bajo y motivación alta) respecto al resto de las medias de los otros tres grupos (a_1b_2 - a_1b_1, a_1b_2 - a_2b_1 y a_1b_2 - a_2b_2). El resto de diferencias de medias entre las medias de las condiciones no son estadísticamente significativas (a_2b_1 - a_1b_1, a_2b_1 – a_1b_2 y a_2b_1 - a_2b_2) porque el valor del Rango Crítico de la prueba de Tukey supera al valor de las diferencias empíricas entre las medias de los grupos o condiciones de los efectos de interacción.

Como conclusión final de los análisis realizados, el investigador o investigadora tendría que concluir que la memoria es mayor cuando los sujetos se encuentran en una situación de estrés bajo y alta motivación para recordar. Así, el efecto en la mejora de la mejora no depende del efecto del estrés y de la motivación por separado

o de forma independiente sino que el efecto de la interacción de ambas situaciones es fundamental para entender por qué la memoria es mayor.

Ejercicio de diseño factorial

Desarrollar el siguiente ejercicio de forma manual y con un programa estadístico. Redactar los resultados utilizando el formato del Manual del APA (7º edición).



Diseño factorial: SPSS, JASP, JAMOVI

La tabla del ANOVA del diseño factorial anterior va a ser ahora completada con el valor de *p* exacto que les corresponde a cada fuente de varianza del tratamiento o varianza entre-grupos del modelo de diseño especificado (A, B y AB). Dicho valor de *p* se calcula asumiendo que la hipótesis nula es cierta (es decir, efecto o relación cero o diferencia entre las medias igual a cero) y por ello se consulta qué valor de probabilidad tendría el estadístico o resultado de $F\,(1,\ 4) = 16$ obtenido en el experimento (o un valor más extremo) en la distribución de la hipótesis nula que es la distribución muestral conocida del estadístico de la Razón $F\,(.05,\ 1,\ 4)$. Este valor *p* del resultado (calculado a partir de la información de la distribución muestral de la hipótesis nula) se puede obtener de forma directa con un programa estadístico como el SPSS, JASP o JAMOVI o con alguna aplicación como por ejemplo la que se ha desarrollado en la siguiente página Web: http://davidmlane.com/hyperstat/F_table.html, detallada por el profesor David Lane.

Por ejemplo, el valor de *p* que se corresponde con los resultados del efecto de la fuente de varianza de interacción AB (*F* (1, 4) =16) es de .01613 (figura 84), que si se redonde en el tercer decimal sería de *p* = .016. El lector o lectora puede comprobar que el valor de *p* de las fuentes de varianza A (*F* (1, 4) =16) y B (*F* (1, 4) =16) es .00388 en los dos casos ya que tienen el mismo valor en la *F* empírica y los grados de libertad entre y del error también son los mismos.

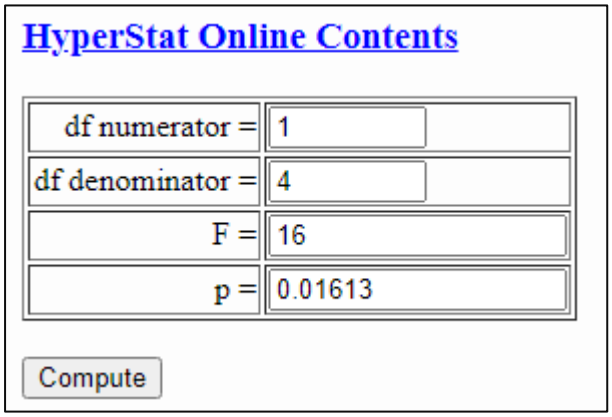


Figura 84. Consulta en la tabla de Tukey del valor de *q*(alfa, ab, gl_{error})

Ahora ya se puede completar la tabla del ANOVA con los valores exactos del valor *p* de probabilidad que se corresponden con los resultados de las tres fuentes de varianza relacionadas con los efectos principales de las variables independientes y su interacción (tabla 19). En la redacción de los informes siempre hay que detallar el valor *p* exacto de cada fuente de varianza.

Tabla 19. Diseño factorial 2 × 2 entre estrés y motivación. Modelo con interacción

Fuente	*SC*	*gl*	*MC*	*F*	*p*	η^2
A	72.00	1	72.00	36.0	.004	.391
B	72.00	1	72.00	36.0	.004	.391
A *x* B	32.00	1	32.00	16.0	.016	.174
Error	8.00	4	2.00			
Total	184.000	7				

Otra forma de obtener el valor de *p* exacto de cada una de las fuentes de varianza que se plantean en el diseño de una determinada investigación es llevar a cabo el análisis con un programa estadístico. A continuación se detallará la ejecución de los análisis con los programas SPSS, JASP y JAMOVI.

SPSS

Una vez introducidos los datos en el SPSS tal y como se detalla en la figura 85, se procede al análisis con el diseño factorial acudiendo a la ventana del programa:

Análisis → Modelo Lineal General → Univariante

	Aestrés	Bmotivación	Yerrores
1	1	1	13
2	1	1	11
3	1	2	1
4	1	2	3
5	2	1	15
6	2	1	13
7	2	2	11
8	2	2	13
9			

Figura 85. Base de datos de un diseño factorial A *x* B con SPSS

El proceso de acceso al diseño elegido en el SPSS es entonces el especificado en la figura 86.

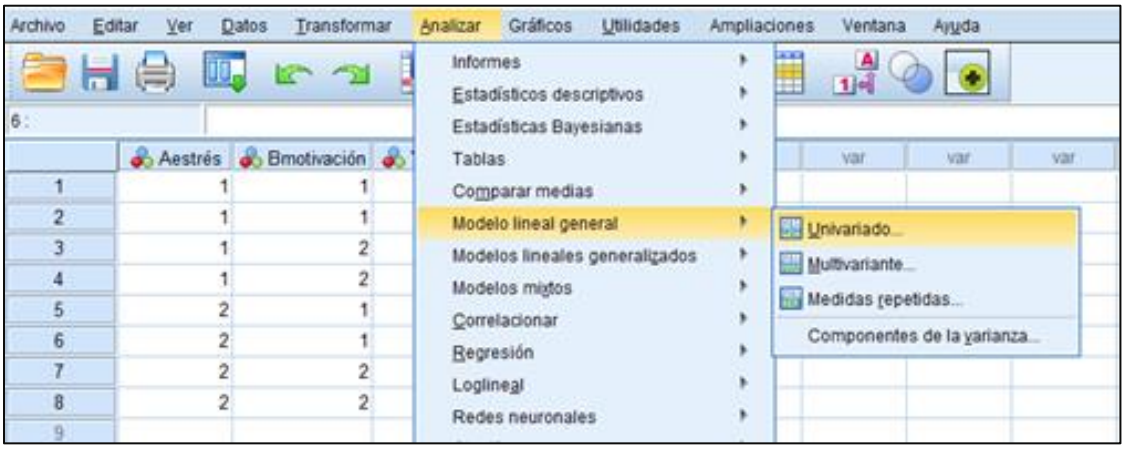


Figura 86. Acceso al diseño univariado con el SPSS

A continuación se sitúa cada una de las variables que se desea estudiar en el lugar correspondiente para llevar a cabo el análisis y en el apartado de Opciones se seleccionan estadísticos y estimaciones del tamaño del efecto y también se selecciona el apartado de Medias marginales estimadas para que el output o salida

de resultados contenga las medias marginales de los efectos principales A y B y las medias de las celdillas de interacción AB (figura 87).

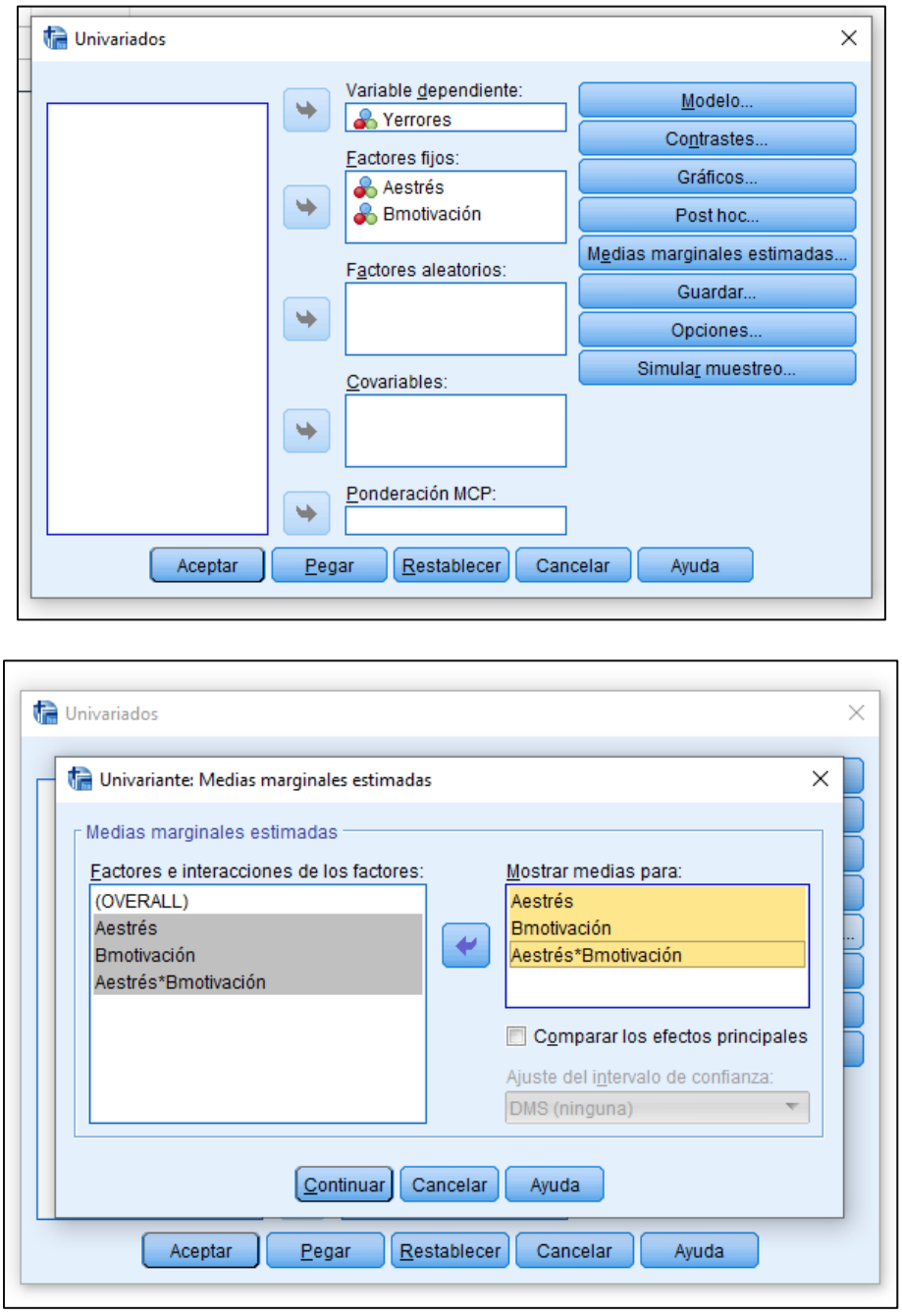


Figura 87. Acomodación de las variables junto con opciones complementarias de análisis y pantalla de las medias marginales estimadas con SPSS

El SPSS ofrece los siguientes resultados. En primer lugar, detalla el número de observaciones que hay en cada condición de los efectos principales A y B (Factores inter-sujetos). En segundo lugar, incluye los estadísticos descriptivos de la fuente de varianza de la interacción: media desviación típica y número de observaciones

(Estadísticos descriptivos). En tercer lugar, se presenta la tabla del ANOVA que incluye el valor de la proporción de varianza explicada (η_p^2) ya que se le solicitó en Opciones junto con los estadísticos descriptivos anteriores (figura 88). El lector o lectora puede comprobar que el programa SPSS ofrece los resultados de la eta cuadrado parcial y por ello no coincide con los valores de η^2 que se han calculado en las tablas de ANOVA anteriores. Para una explicación detallada se puede consultar el apartado dedicado al tamaño del efecto que se ha presentado anteriormente y al apartado anterior donde se valoran los tamaños del efecto de eta cuadrado y eta cuadrado parcial en un diseño factorial. Se recomienda utilizar los valores de η^2 para elaborar el informe de resultados ya que la η_p^2 sobreestima el valor del tamaño del efecto, especialmente si el tamaño de la muestra es pequeño.

Factores inter-sujetos

		N
Aestrés	1	4
	2	4
Bmotivación	1	4
	2	4

Estadísticos descriptivos

Variable dependiente: Yerrores

Aestrés	Bmotivación	Media	Desv. Desviación	N
1	1	12,00	1,414	2
	2	2,00	1,414	2
	Total	7,00	5,888	4
2	1	14,00	1,414	2
	2	12,00	1,414	2
	Total	13,00	1,633	4
Total	1	13,00	1,633	4
	2	7,00	5,888	4
	Total	10,00	5,127	8

Pruebas de efectos inter-sujetos

Variable dependiente: Yerrores

Origen	Tipo III de suma de cuadrados	gl	Media cuadrática	F	Sig.	Eta parcial al cuadrado
Modelo corregido	176,000[a]	3	58,667	29,333	,003	,957
Intersección	800,000	1	800,000	400,000	,000	,990
Aestrés	72,000	1	72,000	36,000	,004	,900
Bmotivación	72,000	1	72,000	36,000	,004	,900
Aestrés * Bmotivación	32,000	1	32,000	16,000	,016	,800
Error	8,000	4	2,000			
Total	984,000	8				
Total corregido	184,000	7				

a. R al cuadrado = ,957 (R al cuadrado ajustada = ,924)

Figura 88. Resultados del ANOVA entre-grupos A x B con SPSS

En quinto lugar, el programa SPSS detalla las medias marginales de los efectos principales A y B y de la interacción ya que se le solicitó al programa SPSS cuando se seleccionó el apartado de Medias marginales estimadas (figura 89). El valor de 1 indica 'bajo' y el valor de 2 indica 'alto' en la condición del factor o variable independiente ya que así se codificaron las variables en la base de datos.

Medias marginales estimadas

1. Aestrés

Variable dependiente: Yerrores

Aestrés	Media	Desv. Error	Intervalo de confianza al 95%	
			Límite inferior	Límite superior
1	7,000	,707	5,037	8,963
2	13,000	,707	11,037	14,963

2. Bmotivación

Variable dependiente: Yerrores

Bmotivación	Media	Desv. Error	Intervalo de confianza al 95%	
			Límite inferior	Límite superior
1	13,000	,707	11,037	14,963
2	7,000	,707	5,037	8,963

3. Aestrés * Bmotivación

Variable dependiente: Yerrores

Aestrés	Bmotivación	Media	Desv. Error	Intervalo de confianza al 95%	
				Límite inferior	Límite superior
1	1	12,000	1,000	9,224	14,776
	2	2,000	1,000	-,776	4,776
2	1	14,000	1,000	11,224	16,776
	2	12,000	1,000	9,224	14,776

Figura 89. Resultados del ANOVA entre-grupos A *x* B con SPSS

El programa SPSS no calcula la prueba de Tukey (u otras pruebas de hipótesis específicas) para el efecto de la interacción de dos o más factores. Solamente calcula esas pruebas de contraste de hipótesis específicas para los efectos principales del modelo de diseño. Por lo tanto, si se utiliza el programa SPSS para analizar un diseño factorial y se desea conocer qué pares de medias simples del efecto de interacción difieren entre sí de forma estadísticamente significativa será necesario calcular el valor del Rango Crítico, teórico o tabular de Tukey y después comparar si la diferencia de medias empírica obtenida en el estudio supera o iguala dicha diferencia para poder concluir que esa diferencia tiene un valor de *p* menor o igual al valor del alfa.

Si es así, se puede concluir que esa diferencia entre las dos medias es estadísticamente significativa.

Otra solución para continuar con el estudio del efecto de interacción es introducir los datos como si fuese un diseño unifactorial con cuatro condiciones (A = 4), utilizando cada efecto de interacción como si fuese una condición del diseño (por ejemplo, A1,B1 se corresponde con la condición A1 del diseño unifactorial A = 4) y llevar a cabo un diseño unifactorial solicitando la prueba de comparaciones múltiples en el programa SPSS. En la figura siguiente (figura 90) se puede observar la introducción de datos para estudiar el efecto de interacción en un diseño AB con el programa SPSS. Se ejecuta ese diseño unifactorial y directamente se observan los resultados de la prueba de hipótesis específicas (por ejemplo, la prueba HSD de Tukey), sin mirar el ANOVA que carece de sentido para el supuesto de investigación.

	Medida	GrupoVI
1	13	1
2	11	1
3	1	2
4	3	2
5	15	3
6	13	3
7	11	4
8	13	4

Con los valores de las etiquetas:

	Medida	GrupoVI
1	13	A1B1
2	11	A1B1
3	1	A1B2
4	3	A1B2
5	15	A2B1
6	13	A2B1
7	11	A2B2
8	13	A2B2

Figura 90. Base de un diseño A = 4 cuyo objetivo es estudiar el efecto de interacción de un diseño 2 *x* 2 con el SPSS

Se ejecuta el diseño unifactorial A = 4 y en Opciones se solicita por ejemplo la prueba de Tukey para el factor o variable independiente de GrupoVI que tiene cuatro condiciones. Los resultados se detallan en la Figura 91. Se observa que las diferencias entre los pares de medias que son estadísticamente significativas se encuentran entreb a_1b_1 respecto a las condiciones de a_1b_2, a_2b_1 y a_2b_2. Son las mismas diferencias que las que se detectan con el uso del Rango Crítico de la prueba de Tukey. Recordar que si se opta por realizar ese diseño unifactorial solamente se debe interpretar los resultados de la prueba de Tukey y hacer caso omiso de los resultados del ANOVA.

Pruebas post hoc

GrupoVI

Comparaciones múltiples

Variable dependiente: Medida

HSD Tukey

(I) GrupoVI	(J) GrupoVI	Diferencia de medias (I-J)	Desv. Error	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
A1B1	A1B2	10,00*	1,414	,007	4,24	15,76
	A2B1	-2,00	1,414	,553	-7,76	3,76
	A2B2	,00	1,414	1,000	-5,76	5,76
A1B2	A1B1	-10,00*	1,414	,007	-15,76	-4,24
	A2B1	-12,00*	1,414	,004	-17,76	-6,24
	A2B2	-10,00*	1,414	,007	-15,76	-4,24
A2B1	A1B1	2,00	1,414	,553	-3,76	7,76
	A1B2	12,00*	1,414	,004	6,24	17,76
	A2B2	2,00	1,414	,553	-3,76	7,76
A2B2	A1B1	,00	1,414	1,000	-5,76	5,76
	A1B2	10,00*	1,414	,007	4,24	15,76
	A2B1	-2,00	1,414	,553	-7,76	3,76

Se basa en las medias observadas.

El término de error es la media cuadrática(Error) = 2,000.

*. La diferencia de medias es significativa en el nivel ,05.

Figura 91. Resultados de la prueba HSD de Tukey con el diseño A = 4 cuyo objetivo es estudiar el efecto de interacción de un diseño 2 *x* 2

JASP

Desde el programa gratuito JASP se puede llamar a la base de datos que se haya creado con Excel o con SPSS, por ejemplo, y ejecutar posteriormente el diseño entre-sujetos univariado 2 x 2 (figura 92).

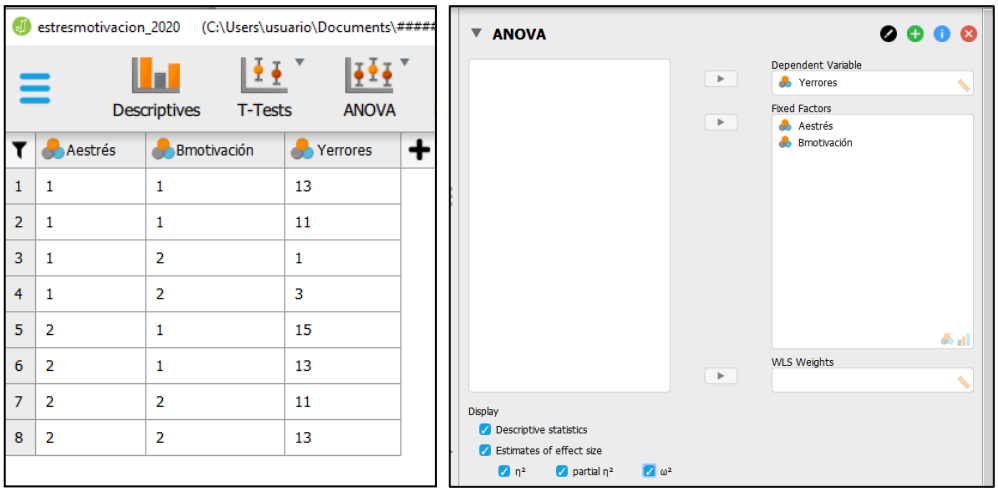


Figura 92. Base de datos y comodación de las variables junto con opciones complementarias de análisis con JASP

En el programa JASP se pueden solicitar más estadísticos del tamaño del efecto. Concretamente, se ha solicitado que se calculen los valores de η^2 y η_p^2 y también el de omega cuadrado (ω^2). El investigador o investigadora debe interpretar el que considere más conveniente para detallar en su informe de investigación y, por lo tanto, informar de solo un estadístico del tamaño del efecto.

Los resultados que ofrece el programa JASP se detallan en la figura 93.

Results

ANOVA

ANOVA - Yerrores

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
Aestrés	72.00	1	72.00	36.00	3.88e −3	0.39	0.90	0.38
Bmotivación	72.00	1	72.00	36.00	3.88e −3	0.39	0.90	0.38
Aestrés ✻ Bmotivación	32.00	1	32.00	16.00	0.02	0.17	0.80	0.16
Residuals	8.00	4	2.00					

Note. Type III Sum of Squares

Descriptives

Descriptives - Yerrores

Aestrés	Bmotivación	Mean	SD	N
1	1	12.00	1.41	2
	2	2.00	1.41	2
2	1	14.00	1.41	2
	2	12.00	1.41	2

Figura 93. Resultados del ANOVA entre-grupos A x B con JASP

En el apartado que tiene el programa JASP de medias marginales se puede solicitar que calcule las medias marginales de los efectos principales A y B y las medias de los efectos de interacción AB (figura 94).

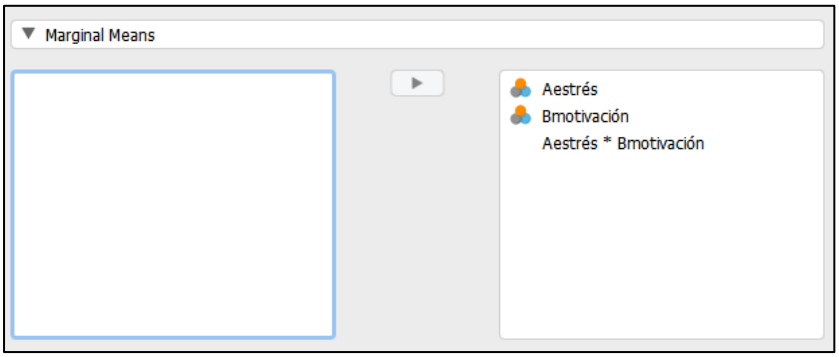


Figura 94. Seleccionar el análisis de las medias marginales con JASP

El resultado que detalla el programa JASP de las medias marginales de A, B y AB se puede observar en la figura 95. El programa JASP ofrece las medias y el error estándar (SE, Standard Error) que no hay que confundir con la desviación típica. Como ofrece el intervalo de las medias detalla el SE ya que el intervalo se construye con el error estándar. Cuando se realice la redacción de los resultados se utilizarán siempre los descriptivos de medias (M), desviación típica (DT) y número de observaciones (N).

Marginal Means

Marginal Means - Aestrés

		95% CI for Mean Difference		
Aestrés	Marginal Mean	Lower	Upper	SE
1	7.00	5.04	8.96	0.71
2	13.00	11.04	14.96	0.71

Marginal Means - Bmotivación

		95% CI for Mean Difference		
Bmotivación	Marginal Mean	Lower	Upper	SE
1	13.00	11.04	14.96	0.71
2	7.00	5.04	8.96	0.71

Marginal Means - Aestrés ✻ Bmotivación

			95% CI for Mean Difference		
Aestrés	Bmotivación	Marginal Mean	Lower	Upper	SE
1	1	12.00	9.22	14.78	1.00
2		14.00	11.22	16.78	1.00
1	2	2.00	−0.78	4.78	1.00
2		12.00	9.22	14.78	1.00

Figura 95. Medias marginales con JASP

Una de las ventajas del programa JASp es que sí realiza un análisis de hipótesis específicas para el efecto de interacción AB del diseño factorial. Para ejecutar los estadísticos de hipótesis específicas se selecciona Post Hoc Tests y se activa la fuente de varianza de la interacción (Aestrés * Bmotivación), por ejemplo. El programa JASP ofrece directamente las 6 diferencias de medias que hay que analizar y no 12 como el SPSS ya que son redundantes y duplica las diferencias de medias. Por lo tanto, conviene recordar que si se utiliza el SPSS hay que tener en cuenta que dichas diferencias de medias están duplicadas, pues solo hay que interpretar seis diferencias simples entre los pares de medias.

En la figura 96 se detalla cómo se solita la prueba post hoc para el efecto de interacción AB.

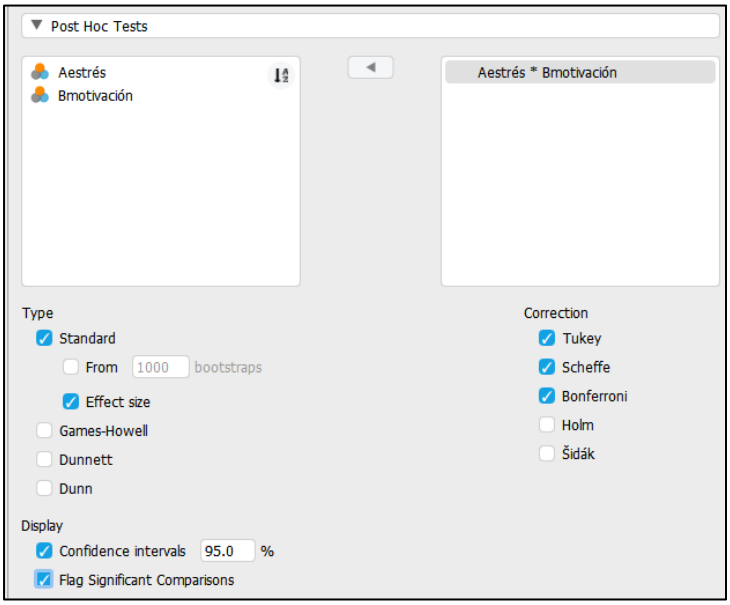


Figura 96. Selección del contraste de hipótesis específicas con JASP

El programa JASP ofrece un conjunto de resultados sobre las diferencias de medias entre todos los pares de medias y acaba con las columnas de los estadísticos de contraste de hipótesis específicas (figura 97). El investigador o investigadora debe seleccionar aquella prueba que sea la más adecuada para el modelo de diseño que está utilizando en su investigación, teniendo en cuenta que debe controlar de forma adecuada la tasa de error de Tipo I y debe ser la más potente (la prueba que tiene el menor error de Tipo II) para maximizar la validez de conclusión estadística. Un inconveniente del programa JASP es que no estima el valor del tamaño del efecto (*d* de Cohen) para cada diferencia entre un par de medias simples, siendo necesario

recurrir a otro programa para calcularlo como la página Web de la Colaboration Campbell, o se puede utilizar el programa JAMOVI que sí calcula dicho valor de tamaño del efecto con el estadístico de diferencia estandarizada de medias *d* de Cohen.

Post Hoc Tests

Standard

Post Hoc Comparisons - Aestrés ✻ Bmotivación

			95% CI for Mean Difference						
		Mean Difference	Lower	Upper	SE	t	p_{tukey}	$p_{scheffe}$	p_{bonf}
1, 1	2, 1	−2.00	−7.76	3.76	1.41	−1.41	0.55	0.62	1.00
	1, 2	10.00	4.24	15.76	1.41	7.07	7.28e −3**	0.01*	0.01*
	2, 2	−4.44e −15	−5.76	5.76	1.41	−3.14e −15	1.00	1.00	1.00
2, 1	1, 2	12.00	6.24	17.76	1.41	8.49	3.68e −3**	5.10e −3**	6.35e −3**
	2, 2	2.00	−3.76	7.76	1.41	1.41	0.55	0.62	1.00
1, 2	2, 2	−10.00	−15.76	−4.24	1.41	−7.07	7.28e −3**	0.01*	0.01*

* p < .05, ** p < .01

Note. P-value and confidence intervals adjusted for comparing a family of 4 estimates (confidence intervals corrected using the tukey method).

Figura 97. Resultados del contraste de hipótesis específicas con JASP

JAMOVI

El programa gratuito JAMOVi puede abrir los ficheros de SPSS, JASP o Excel, por ejemplo y la presentación de su base de datos es muy semejante (figura 98).

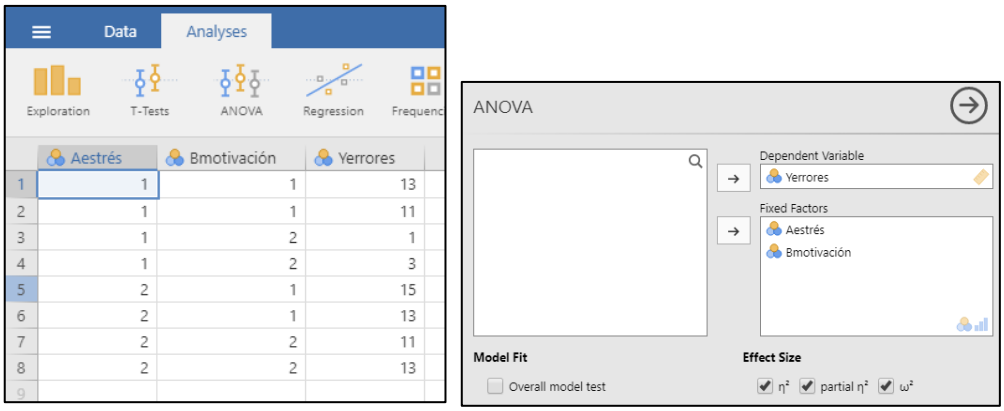
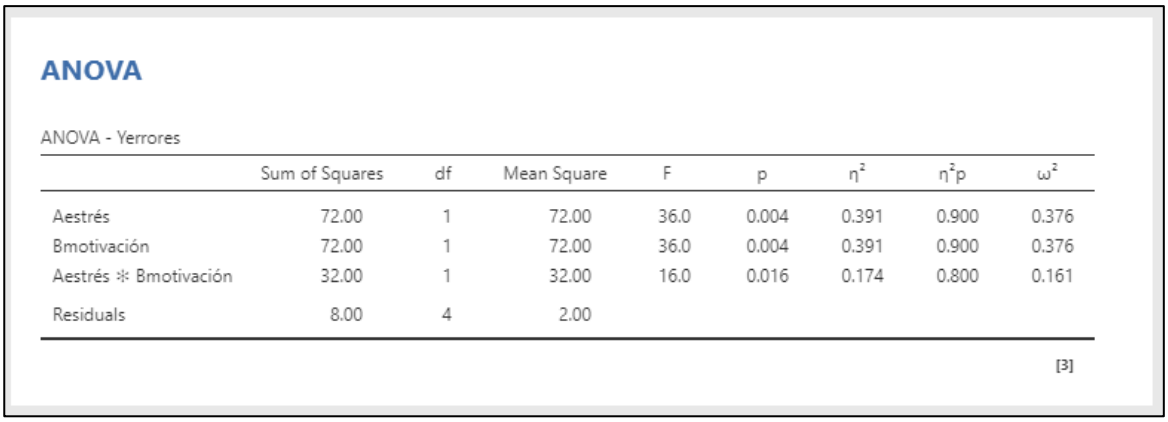


Figura 98. Base de datos y comodación de las variables junto con opciones complementarias de análisis con JAMOVI

Los resultados con el programa JAMOVI del ANOVA junto con los estadísticos del tamaño del efecto solicitados se detallan en la figura 99.



ANOVA

ANOVA - Yerrores

	Sum of Squares	df	Mean Square	F	p	η^2	η^2p	ω^2
Aestrés	72.00	1	72.00	36.0	0.004	0.391	0.900	0.376
Bmotivación	72.00	1	72.00	36.0	0.004	0.391	0.900	0.376
Aestrés ✻ Bmotivación	32.00	1	32.00	16.0	0.016	0.174	0.800	0.161
Residuals	8.00	4	2.00					

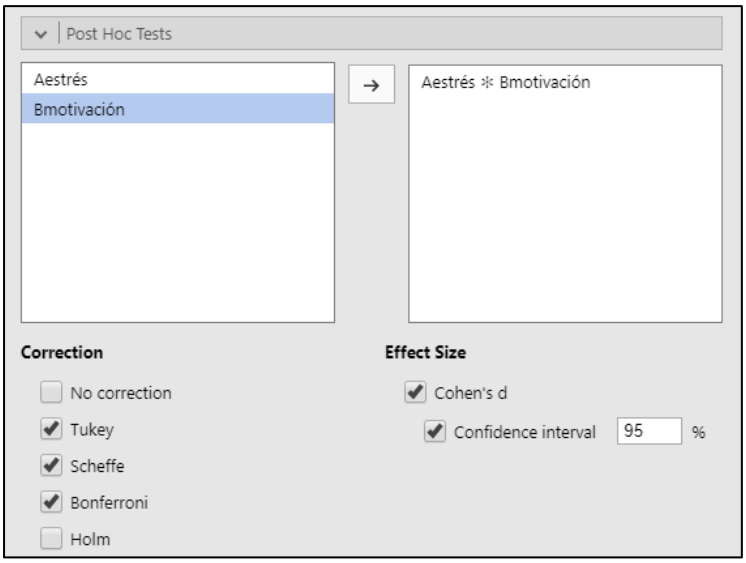
[3]

Figura 99. Resultados del ANOVA entre-grupos A x B con JAMOVI

Como se observa en la figura 100, el programa JAMOVI puede efectuar las pruebas post o estadísticos de contraste de hipótesis específicas y, además, ofrece para cada diferencia entre un par de medias el valor del tamaño del efecto *d* de Cohen denominado diferencia estandarizada de medias.

Por lo tanto, de los tres programas estadísticos analizados, solamente el programa JAMOVI calcula los valores del tamaño del efecto para el efecto de interacción AB, es decir, calcula una diferencia de medias estandarizada *d* de Cohen para cada uno de los pares de medias que tiene el diseño del estudio junto con su intervalo de confianza.

Conviene tener presente que esos valores tan altos de *d* de Cohen son irreales en la investigación real y son propios de un diseño cuyos datos se han programado para la docencia. Siguiendo la propuesta de Jacob Cohen, un tamaño del efecto de diferencia estandarizada de media de $d = 0.2$ es pequeño, $d = 0.5$ es medio y un valor de $d = 0.8$ sería grande. Y, gran parte de las terapias psicológicas tienen en la realidad un tamaño del efecto en torno a $d = 0.5$, es decir, un tamaño del efecto mediano.



Post Hoc Tests

Post Hoc Comparisons - Aestrés ✻ Bmotivación

Comparison												95% Confidence Interval	
Aestrés	Bmotivación		Aestrés	Bmotivación	Mean Difference	SE	df	t	P_{tukey}	$P_{scheffe}$	$P_{bonferroni}$	Cohen's d	Lower	Upper
1	1	-	1	2	10.00	1.41	4.00	7.07	0.007	0.010	0.013	7.07	−0.405	14.547
		-	2	1	−2.00	1.41	4.00	−1.41	0.553	0.615	1.000	−1.41	−4.518	1.690
		-	2	2	−4.44e−15	1.41	4.00	−3.14e−15	1.000	1.000	1.000	0.00	−2.776	2.776
	2	-	2	1	−12.00	1.41	4.00	−8.49	0.004	0.005	0.006	8.49	−0.295	17.265
		-	2	2	−10.00	1.41	4.00	−7.07	0.007	0.010	0.013	−7.07	−14.547	0.405
2	1	-	2	2	2.00	1.41	4.00	1.41	0.553	0.615	1.000	1.41	−1.690	4.518

Note. Comparisons are based on estimated marginal means

Figura 100. Resultados de las pruebas de hipótesis específicas para el efecto de interacción AB y valores de *d* de Cohen con JAMOVI

Capítulo 14. Diseño de bloques

Dolores Frías-Navarro

Universidad de Valencia

Índice

Citar el capítulo como:

Frías-Navarro, D. (2021). Diseño de bloques. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

En ocasiones puede ocurrir que se introduzca una variable en el diseño de investigación para controlar su efecto y reducir con ello la varianza del error aleatorio. Se trata de una variable que es fuente de varianza sistemática secundaria (mantiene una relación estadísticamente significativa con la variable dependiente) y se desea controlar su efecto a través del diseño que se utiliza para analizar los datos del estudio (control de la varianza sistemática secundaria). Conviene recordar, que la hipótesis del estudio está formada por las variables explicativas, es decir la variable independiente (fuente de varianza sistemática primaria) y la variable dependiente. Sin embargo, pueden haber más variables en el estudio como, por ejemplo las variables de bloqueo cuya función es de control.

El diseño de bloques supone utilizar la técnica de control de la constancia a través de la introducción de la variable de bloqueo en la ecuación estructural. Se trata de un factor o variable independiente que está formado por condiciones o grupos y estadísticamente es indistinguible de una variable independiente fuente de varianza sistemática primaria. Sus efectos se estiman del mismo modo. Y, para determinar si un factor del diseño tiene la función de bloque o se trata de una variable independiente objeto de estudio teórico es necesario conocer el enunciado de la hipótesis de investigación. En la hipótesis de investigación únicamente se detallan las variables explicativas, es decir, las variables del efecto o varianza sistemática primaria (variable independiente) y la variable medida (variable dependiente). La variable de bloqueo se explica en el texto, pero cuando se hace referencia al diseño del estudio y a su planificación ya que ahí se debe detallar que se planificó el control de la variable de bloqueo y por ello se crearon los grupos que forman las condiciones de dicha variable de bloqueo.

Las variables de bloqueo se introducen en la ecuación estructural del diseño como fuente de varianza que no forma parte de la hipótesis sustantiva del estudio, pero gracias a ello se controla su efecto y se reduce el término de error del modelo. En todos los diseños de investigación se desea que el error aleatorio sea el mínimo posible (minimizar el error). Cuando se introduce una variable de bloqueo en la ecuación estructural se está controlando su efecto y se está reduciendo el término de error.

Cuando el investigador o investigadora considera que puede haber una variable que no tiene una distribución homogénea entre los participantes y podría ser causa de heterogeneidad entre ellos (se trata de grupos no homogéneos), puede optar por incluirla en el diseño como un factor o variable independiente y actuar como variable de bloqueo. Esa variable de bloqueo permite crear grupos o condiciones de la variable independiente objeto de estudio (fuente de varianza sistemática primaria) donde en gran medida se garantice que hay una distribución homogénea de la variable de bloqueo en cada condición de la variable independiente, controlando con ello su efecto (ya que los grupos serán homogéneos en la variable de bloqueo) y, al mismo, tiempo aumentando la potencia estadística ya que se reduce el término de error del modelo. Este diseño se conoce como diseño de bloques.

El diseño de bloques se utiliza especialmente cuando el estudio tiene una metodología cuasi-experimental y la ausencia de asignación aleatoria del tratamiento podría suponer que hay alguna variable extraña que habría que controlar y una técnica útil puede ser bloquearla (diseño de bloques no aleatorizados). Sin embargo, también en los estudios con metodología experimental se puede utilizar una variable de bloqueo, pero antes de la asignación aleatoria del tratamiento se han configurado los bloques con sus condiciones o grupos y, posteriormente, se podrían asignar de forma aleatoria las condiciones del tratamiento a cada bloque de forma independiente (diseño de bloques aleatorizados). Es decir, diseños aleatorizados por bloques.

En resumen y repasando cuestiones que ya se han tratado anteriormente en el libro, en un diseño completamente aleatorizado (*metodología experimental sin restricciones en la aleatorización*), los tratamientos (condiciones de la o las variables independientes) se asignan al azar a los participantes sin restricciones. Pero en un diseño con una metodología experimental con restricciones hay un paso previo donde se crean los bloques en función de las características propias de los participantes (se estratifica a los participantes en alguna variable no aleatoria) y, posteriormente, se asigna al azar el tratamiento a cada uno de los bloques creados (*metodología experimental con restricciones en la aleatorización* o *diseños parcialmente aleatorios*). Cada bloque estará formado por unidades experimentales que forman un grupo homogéneo y gracias a ello se reduce el residual o error del modelo.

Cuando en una investigación hay que bloquear dos variables extrañas (dos variables de bloqueo) y se puede suponer justificadamente -por razonamientos teóricos o por el aval de otros estudios- que no interactúan las variables extrañas entre ellas ni con la independiente, es posible reducir los costes del experimento aplicando un diseño de cuadrado latino.

Conviene recordar que para considerar un diseño con dos o más variables independientes como experimental al menos una de estas debe ser manipulada y asignada aleatoriamente. La otra variable no debe ser manipulada o asignada necesariamente.

Se recomienda a los lectores y lectoras que repasen el apartado de asignación aleatoria que se ha detallado anteriormente en el libro y valoren de nuevo el ejemplo de la glucosa en sangre en la línea base que se medía para que actuase como una variable de bloqueo: glucosa alta, glucosa media y glucosa baja. Se creaban esos tres grupos de sujetos y, posteriormente se asignaba al azar a cada uno de los bloques formados, y de forma independiente, las dos condiciones de la variable independiente que eran objeto de estudio: un fármaco para el tratamiento de la diabetes o un fármaco placebo. Por lo tanto, se controla que en el grupo de la condición de fármaco para el tratamiento de la diabetes hay participantes de las tres de categorías de nivel de glucosa en la línea base y también se controla que ocurra lo mismo en el grupo de fármaco placebo. De este modo se garantiza la distribución homogénea de los distintos niveles de glucosa de línea base en el grupo experimental y en el grupo de control. Se trata de un *diseño de bloques con restricciones en la aleatorización* porque no todas las variables o factores del diseño se han asignado al azar (la glucosa previa en sangre es una variable asignada y por lo tanto no es posible aleatorizarla).

Ecuación estructural del diseño de bloques

La ecuación estructural de los diseños con variables de bloqueo no difiere de la del diseño factorial, pero sí impone que el modelo sea necesariamente aditivo. De todas maneras, si el modelo que plantea la hipótesis del estudio es aditivo, también hay que estimar el efecto de la interacción en la ecuación estructural para comprobar que esta fuente de variación no es estadísticamente significativa (primer supuesto del diseño con bloques).

El diseño diseño de bloques más sencillo es factorial y tiene dos factores: factor A vinculado a la variable independiente (factor de 'varianza sistemática primaria') y el Factor B vinculado a la variable de bloqueo que se controla (factor de 'varianza sistemática secundaria' controlada), configurando un diseño entre grupos de bloques A *x* B que está definido como un modelo de efectos aditivos en la siguiente ecuación estructural:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{E}$$

Y = valores de la variable dependiente

M = media de la variable dependiente

A = efecto principal del primer factor, A

B = efecto principal del segundo factor, B

E = error de estimación del modelo

Supuestos del diseño de bloques

Los dos supuestos del modelo de bloques son: 1) el modelo factorial es aditivo y 2) la variable de bloqueo está relacionada con la variable dependiente.

1- Comprobar que no hay efecto de interacción estadísticamente significativo entre los factores A (variable de tratamiento) y B (variable de bloqueo).

El primer supuesto de un diseño con variables de bloqueo señala que la variable independiente o variable de tratamiento y la de bloqueo no deben interaccionar; es decir, la relación entre la variable de bloqueo y la variable de tratamiento no afecta a la expresiópn de la variable dependiente. Si la variable de bloqueo interacciona con la independiente, la de bloqueo no podría ser extraña a los objetivos del estudio sino que debería ser otra variable independiente de tratamiento, y la hipótesis debería explicar la naturaleza de la interacción. Para comprobar este principio se aplica una ANOVA donde se analiza, exclusivamente, si el efecto de la interacción es estadísticamente significativo. Si no se cumpliera este supuesto (valor *p* de la fuente de varianza de la interacción menor a .05), la justificación teórica que avalara el modelo quedaría desconfirmada por los datos. En cambio, si el modelo factorial es aditivo (el efecto de interacción no es estadísticamente significativo, *p* > alfa)

entonces se podrá continuar con el análisis del diseño de bloques. Una vez tenemos los resultados del modelo no aditivo con interacción no es necesario volver a construir todo el desarrollo de la ecuación ecuación estructural, pues el resultado de las operaciones se puede conocer sin tener que hacer los cálculos. Simplemente hay que añadir a la suma de cuadrados del error y a sus grados de libertad los valores de la fuente de varianza de la interacción.

Por lo tanto, el diseño de bloques se corresponde con una ecuación estructural definida por un modelo aditivo, es decir, no tiene efecto de interacción entre los dos factores. Esta ausencia de efecto de interacción debe comprobarse con los datos del estudio a través de lo que se denomina 'prueba de la aditividad' (el efecto de interacción no es estadísticamente significativo, $p >$ alfa, primer supuesto del diseño de bloques). Si se detectará un efecto de interacción estadísticamente significativo entonces se habría producido un desajuste con la hipótesis teórica planteada ya que dicha hipótesis no debe hacer referencia a un efecto de interacción entre los factores principales del modelo ya que se optó por este modelo aditivo cuando se planificó el estudio. Es decir, si se detecta un efecto de interacción, entonces la hipótesis de la investigación resultaría automáticamente invalidada.

2- Comprobar que en el modelo aditivo la variable dependiente está relacionada con la de bloqueo (la fuente de varianza de la variable bloqueada es estadísticamente significativa cuando se plantea el modelo aditivo de los efectos).

Y una vez comprobado que el modelo es aditivo, se analizan los datos con el modelo aditivo y se somete al contraste estadístico mediante el ANOVA. Y antes de interpretar el efecto de la variable independiente o factor de la hipótesis, se debe comprobar que el factor de bloqueo sí es una variable con un efecto estadísticamente significativo ($p \leq$ alfa, segundo supuesto del diseño de bloques) sobre la variable dependiente. Si se comprueba que el efecto de la variable bloqueada no es estadísticamente significativo entonces su varianza podría formar parte de la varianza de error y afectará poco a la prueba de hipótesis del factor A o factor de varianza sistemática primaria; en ese caso se trataría de una variable no extraña sino una variable irrelevante como muchas otras variables. Los datos deberían analizarse en un modelo sin factor de bloqueo.

Si se cumple con los dos supuestos se habrá conseguido reducir una parte de la varianza del error y la prueba de la hipótesis se podrá realizar con un término residual menor que si no se hubiera controlado la variable extraña en cuestión. Obviamente, no se hacen pruebas entre las medias de los bloques, ya que este análisis no tiene interés teórico.

Hay que tener en cuenta que la variable extraña se bloquea porque se supone que está relaciona de forma estadísticamente significativa con la variable dependiente y si no se controla su efecto entonces el término de error se incrementará. Si el investigador o investigadora sospecha que la asignación aleatoria no es una técnica de control suficiente para controlar esa fuente de varianza extraña (metodología experimental) o si se conoce su efecto y no hay asignación aleatoria (metodología no experimental) se puede optar por incluir dicha variable en la ecuación estructural como un factor y controlar su efecto ya que se mantiene constante en todos los grupos de tratamiento. Además, cuando se controla su efecto, aumenta la potencia estadística de la prueba estadística del factor de varianza sistemáticas (disminuye el error y, por lo tanto, el numerador de la prueba *F*)

Estimación de los efectos

La estimación de las fuentes de varianza del diseño de bloques se realizan exactamente igual que en todos los diseños.

El efecto de la manipulación del primer factor, A, en la variable dependiente se estima a partir de las diferencias entre las medias de los grupos sometidos a los distintos niveles de *A* menos la media general, por el mismo procedimiento que se utiliza para estimar un efecto principal:

$$\mathbf{A} = \mathbf{M}_a - \mathbf{M}$$

El efecto de la manipulación del segundo factor, B, se estima a partir de las diferencias entre las medias de los grupos sometidos a los distintos niveles de B menos la media general:

$$\mathbf{B} = \mathbf{M}_b - \mathbf{M}$$

Tanto **A** como **B** estiman los denominados *efectos principales* de los factores, y ambos efectos tienen resultados aditivos, esto es, podemos estimar **A**

independientemente de **B**, y **B** independientemente de **A**. Por tanto el modelo aditivo de un diseño factorial con dos factores es:

$$\mathbf{Y} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{E}$$

Este modelo no contempla, por tanto, el efecto de interacción en la expresión de su ecuación estructural.

El lector o lectora puede comprobar la diferencia de pronóstico que se deriva del modelo aditivo y del modelo no-aditivo tal y como se ha detallado anteriormente en el diseño factorial. A continuación se detalla la predicción que se hace del primer sujeto a partir del modelo aditivo:

$$\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B} = \bar{Y} + \hat{\alpha}_1 + \hat{\beta}_1 = 10 + -3 + -3 = 4$$

La predicción del mismo sujeto bajo el modelo no-aditivo:

$$\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B} + \mathbf{AB} = \bar{Y} + \hat{\alpha}_1 + \hat{\beta}_1 + \hat{\alpha\beta}_{11} = 10 + -3 + -3 + -2 = 2$$

Lógicamente la predicción de los dos modelos no será la misma, ni tampoco serán iguales los errores de predicción cometidos por ambos modelos.

Por lo tanto, en el modelo aditivo, el error de estimación que se comete está definido por la diferencia entre la puntuación pronosticada y la puntuación obtenida. Como la estimación del valor de la variable dependiente se hace a partir de los efectos de **A** y **B** entonces la puntuación pronostica es:

$$\hat{\mathbf{Y}} = \mathbf{M} + \mathbf{A} + \mathbf{B}$$

El error de estimación a partir del modelo aditivo será:

$$\mathbf{E}_{H_1} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - (\mathbf{M} + \mathbf{A} + \mathbf{B})$$

Es decir:

$$\mathbf{E}_{H_1} = \mathbf{Y} - \mathbf{M} - \mathbf{A} - \mathbf{B}$$

Para comprobar si los efectos estimados en el experimento son estadísticamente significativos hay que aplicar la prueba de la hipótesis. Para ello, se eleva al cuadrado cada uno de los efectos y se obtienen las Sumas de Cuadrados que deben ser ajustadas por sus grados de libertad (SC / gl) para obtener las Medias Cuadráticas.

4. Los grados de libertad de los *efectos principales* se obtiene restando un punto del número de condiciones de cada factor:

-los grados de libertad de *A* son *a* – 1 y

- los grados de libertad de *B* son *b* – 1

Los grados de libertad del término residual o error se calculan descontando de los grados de libertad totales los de los efectos principales: $gl_{total} - gl_A - gl_B$. Los grados de lbertad totales son *N* – 1.

Supuestos de investigación: diseño de bloques 2 x 2 univariado

SUPUESTO DE DISEÑO DE BLOQUES 1. En la práctica totalidad de las investigaciones y manuales de hipnosis se hace referencia a la relación entre la hipnosis y el recuerdo por la posibilidad de evocar mediante los trances hipnóticos sucesos y vivencias que se encuentran postergados en el olvido por el paso del tiempo, o para encontrar vínculos con procesos traumáticos. La finalidad terapéutica de utilizar la relajación, sofrología o hipnosis es la de recuperar la información a la que el sujeto parece no tener acceso en un estado de activación normal. Pero la misma desconexión con el mundo exterior que se consigue en estos estados parece entorpecer los procesos de aprendizaje. Se ha comprobado que el aprendizaje (medido por el recuerdo después del trance) se dificulta cuando se aprendió en estado de hipnosis. Un investigador o investigadora pretende replicar los resultados de las investigaciones, pero con el fin de aumentar la potencia del diseño bloquea la capacidad de los sujetos para recordar listas de palabras con tres condiciones: baja, media y alta. El proceso de selección de la muestra se realiza aleatoriamente a partir del total de alumnos matriculados en la asignatura de Pensamiento. Todos los alumnos completaron una prueba para medir la capacidad para recordar listas de palabras; a partir de su puntuación en esta prueba se dividió a los participantes utilizando los percentiles 25, 50 y 75 en los tres bloques. Para cada bloque se extrajeron aleatoriamente cuatro alumnos, asignando mediante el azar la mitad a cada condición de la variable independiente. Se hipotetiza que el aprendizaje producido en el trance hipnótico será siempre menor, no interactuando este factor con la capacidad de retentiva de los sujetos. Los resultados se detallan en la tabla

20. Y a continuación se debe elaborar el análisis para completar la tabla de ANOVA y redactar los resultados.

Tabla 20. Puntuaciones en el experimento

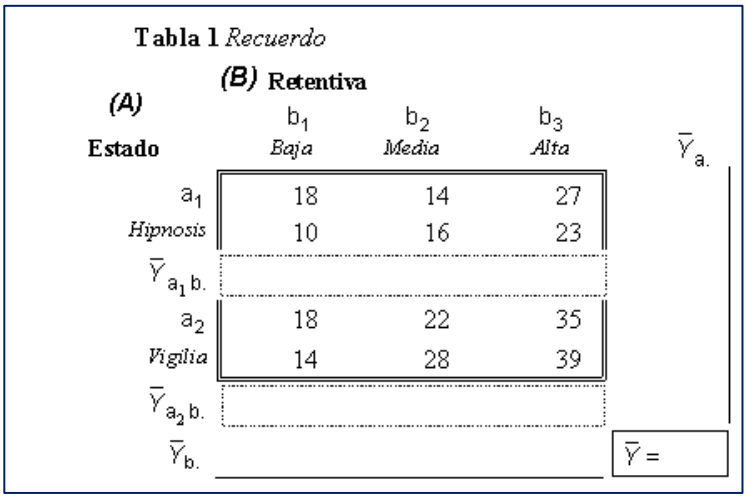
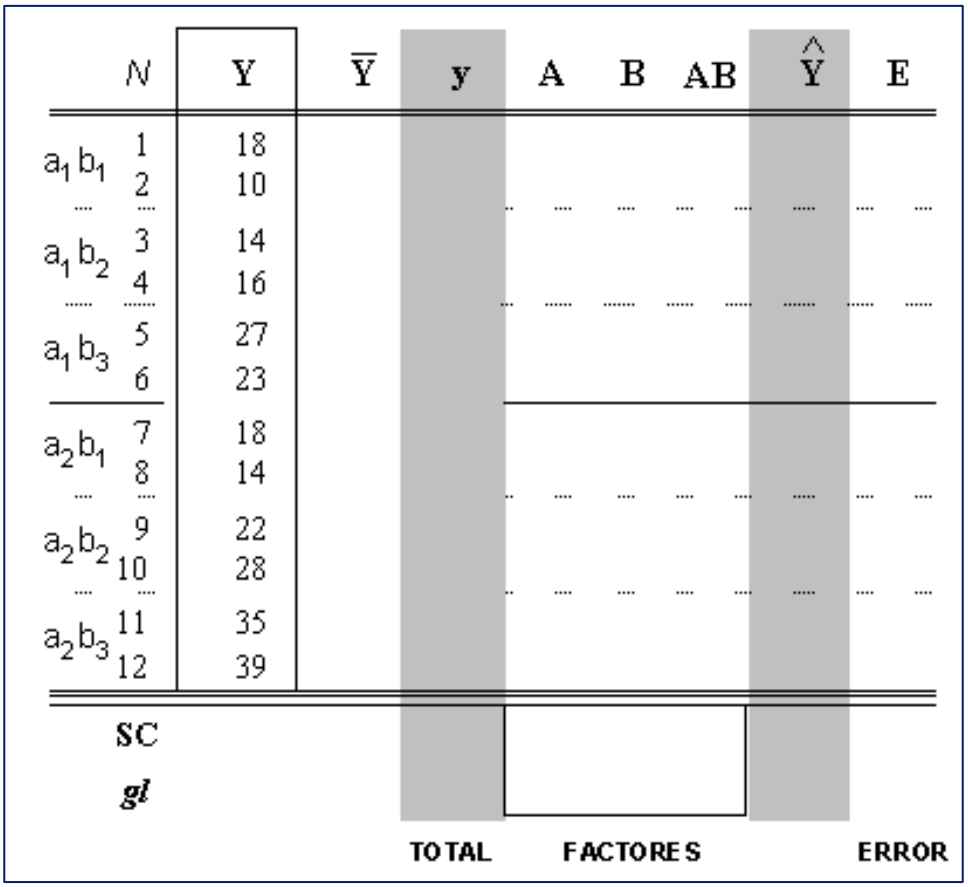


Tabla 1 *Recuerdo*

(A) Estado	(B) Retentiva b_1 *Baja*	b_2 *Media*	b_3 *Alta*	$\bar{Y}_{a.}$
a_1 *Hipnosis*	18 10	14 16	27 23	
$\bar{Y}_{a_1 b.}$				
a_2 *Vigilia*	18 14	22 28	35 39	
$\bar{Y}_{a_2 b.}$				
$\bar{Y}_{b.}$				$\bar{Y}$ =

Ejercicios:

1. Completa la ecuación estructural y calcula las Sumas de Cuadrados.



	N	Y	$\bar{Y}$	y	A	B	AB	$\hat{Y}$	E
$a_1 b_1$	1	18							
	2	10							
...	...	...		...	...	...	...	...	...
$a_1 b_2$	3	14							
	4	16							
.....	.....	.....		.....	.....	.....	.....	.....	.....
$a_1 b_3$	5	27							
	6	23							
$a_2 b_1$	7	18							
	8	14							
...	...	...		...	...	...	...	...	...
$a_2 b_2$	9	22							
	10	28							
...	...	...		...	...	...	...	...	...
$a_2 b_3$	11	35							
	12	39							
SC									
gl									
				TOTAL		**FACTORES**			**ERROR**

2. Comprueba si se cumple el supuesto de no interacción (¿es un modelo no aditivo o es un modelo aditivo?).

Tabla 3 *Diseño factorial 2 × 3 con interacción*

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	$\hat{\eta}^2$
A					0.050	
B					0.050	
A × B					0.050	
Error						
Total				F_{tablas}	=	
				F_{tablas}	=	

3. Completa la tabla resumen del Análisis de la Varianza (modelo aditivo: diseño de bloques)

Tabla 4 *Diseño factorial 2 × 3*

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	$\hat{\eta}^2$
A					0.050	
B					0.050	
Error						
Total				F_{tablas}	=	
				F_{tablas}	=	

4. Redacció del resultats.

_____
_____
_____
_____
_____

5. Si no se hubiera realizado el bloqueo de la variable, ¿qué conclusión se habría obtenido en el estudio?

Tabla 5 *Diseño unifactorial*

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	$\hat{\eta}^2$
A					0.050	
Error						
Total				F_{tablas}	=	

Supuesto de diseño de bloques

Supuesto de investigación. Diversos autores han analizado la relación entre la ansiedad y la ejecución, encontrando que es curvilínea; de tal manera que la ejecución es óptima con un nivel de arousal moderado o intermedio, y los niveles de arousal demasiado bajos o demasiado altos conducen a ejecuciones pobres. Un psicólogo deportivo ha diseñado un nuevo procedimiento que prepara psicológicamente a los futbolistas para el lanzamiento de los penalties. Emplea unas técnicas de autocontrol que incluyen la visualización de la situación, la concentración mental y la relajación. En un estudio previo comprobó que los deportistas de alta competición mejoraban la ejecución, pero algunos estudios posteriores no confirmaban los satisfactorios resultados iniciales. El psicólogo consideraba que su procedimiento había fracasado en algunas ocasiones porque el futbolista se había distendido excesivamente por los ejercicios de relajación. Para paliar esta deficiencia incluye en su programa de entrenamiento un ejercicio cognitivo de autosugestión que incrementa el nivel de arousal. El cambio introducido en su programa lo somete a prueba con 12 futbolistas extraídos aleatoriamente de cada división, cuatro juegan en la primera división, cuatro en la segunda A y otros tantos, en la segunda B. Una vez concluido el programa inicial, dos de cada categoría, seleccionados al azar, practican el ejercicio cognitivo que incrementa el arousal. Todos los deportistas, después del entrenamiento realizan una tanda de 30 lanzamientos al mismo portero. La hipótesis es que los futbolistas entrenados para aumentar el arousal conseguirán

más goles. En el diseño se bloquea la calidad técnica de los jugadores mediante la categoría en la que milita el futbolista, considerando que este factor no interactúa con el tratamiento, mejorando la ejecución de los futbolistas sin que importe su categoría.

Por lo tanto, el supuesto 2 de bloqueo plantea un diseño de bloques con la variable de entrenamiento en un ejercicio cognitivo de autosugestión (arousal medio) frente al grupo de arousal bajo que no recibe el programa de intervención, representando esta fuente de varianza la denominada varianza sistemática primaria. Además, se controla con el diseño una posible fuente de varianza sistemática secundaria que es la técnica deportiva de cada jugador operacionalizada por su pertenencia a un determinado grupo o categoría futbolística. Los datos se representan en la tabla 21 y la primera tarea consiste en calcular las medias de cada fuente de varianza.

Tabla 21. Puntuaciones en el experimento

Tabla 1 *Número de goles*

(A) Arousal		(B) Categoría b_1 *2ª B*	b_2 *2ª A*	b_3 *1ª*	$\bar{Y}_{a.}$
a_1		8	11	21	
Bajo		14	7	17	
$\bar{Y}_{a_1 b.}$					
a_2		14	16	25	
Medio		8	22	29	
$\bar{Y}_{a_2 b.}$					
$\bar{Y}_{b.}$					$\bar{Y}$ =

Como se trata de un diseño de bloques que plantea un modelo de efectos aditivos, lo primero que hay que hacer es comprobar el supuesto de aditividad. Es decir, comprobar que efectivamente el modelo es de efectos aditivos y no hay un efecto de interacción que es multiplicativo.

Para ello es necesario obtener, en primer lugar, los resultados de la tabla del ANOVA de un modelo no aditivo, es decir, con efecto de interacción AB y comprobar que esa fuente de interacción no es estadísticamente significativa (p > .05).

Para comprobar el supuesto de aditividad de los efectos habrá que trabajar estimando los efectos de las fuentes de varianza A, B, AB y error.

Así, se comienza cn el cálculo de los grados de libertad de cada fuente de varianza (tabla 22) y después se estiman tantos efectos como grados de libertad tenga la fuente de varianza (tabla 23) y el resto se completa hasta que sumen 0 los efectos de cada fuente de varianza (tabla 24).

Tabla 22. Calcular los grados de libertad de cada fuente de varianza

$$
\begin{aligned}
totales \equiv\ gl_T &= N-1 = \quad -1 = \\
entre\ A \equiv\ gl_A &= a-1 = \quad -1 = \\
entre\ B \equiv\ gl_B &= b-1 = \quad -1 = \\
interacción\ AB \equiv\ gl_{AB} &= gl_A\ gl_B = \quad \cdot \quad = \\
residual \equiv\ gl_W &= gl_T-(gl_A+gl_B+gl_{AB}) = \\
&= \quad -(\quad +\quad +\quad) =
\end{aligned}
$$

Teniendo en cuenta el resultado de los grados de libertad de cada fuente se varianza se pasa a completar la tabla de efectos, estimando solamente el número de efectos necesarios que se han especificado en los grados de libertad de cada fuente de varianza (tabla 23).

Tabla 23. Efectos que necesariamente hay que estimar según los grados de libertad de cada fuente de varianza

a_1: *Arousal: Bajo*

$\rightarrow \hat{\alpha}_1 = \bar{Y}_{a_1} - \bar{Y} = \quad - \quad =$

b_1: *Categoría: 2ª B*

$\rightarrow \hat{\beta}_1 = \bar{Y}_{b_1} - \bar{Y} = \quad - \quad =$

b_2: *Categoría: 2ª A*

$\rightarrow \hat{\beta}_2 = \bar{Y}_{b_2} - \bar{Y} = \quad - \quad =$

a_1b_1: $\rightarrow \widehat{\alpha_1\beta_1} = \bar{Y}_{a_1b_1} - \bar{Y} - \hat{\alpha}_1 - \hat{\beta}_1 =$

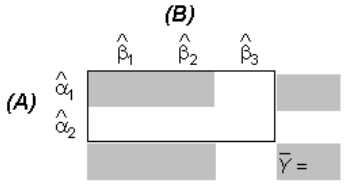
$= \quad - \quad - \quad - \quad =$

a_1b_2: $\rightarrow \widehat{\alpha_1\beta_2} = \bar{Y}_{a_1b_2} - \bar{Y} - \hat{\alpha}_1 - \hat{\beta}_2 =$

$= \quad - \quad - \quad - \quad =$

Posteriormente ya se puede completar toda la tabla de efectos sabiendo que la suma de los efectos de cada fuente de varianza especificada en el modelo de diseño debe ser igual a cero (tabla 24). Las celdillas son gris señalan que esos efectos necesariamente hay que estimarlos y las celdillas en blanco se obtienen como valor complementario hata que todos los efectos sumen cero.

Tabla 24. Completar la tabla de efectos



A continuación ya se procede con el desarrollo de la ecuación estructural del modelo factorial no aditivo (tabla 25) ya que, como se ha comentado anteriormente, el primer paso es comprobar el supuesto de aditividad de los efectos y por eso se realiza este modelo para observar si en la tabla del ANOVA el efecto de interacción AB es o no estadísticamente significativo (tabla 26). Si no es un efecto estadísticamente significativo ($p > .05$) entonces se habrá comprobado que se cumple el supuesto de aditividad y se prosigue con la comprobación del supuesto 2 que hace referencia a que la variable de bloqueo sí debe ser una fuente de varianza cuyo efecto es estadísticamente significativo ($p \leq .05$) (tabla 27).

Tabla 25. Desarrollar la ecuación estructural

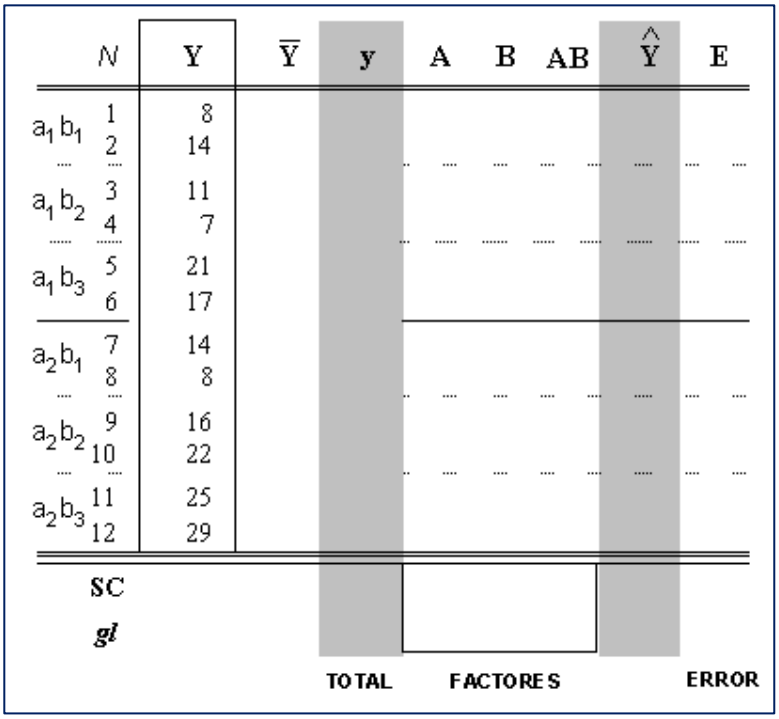


Tabla 26. Realizar la tabla del ANOVA del modelo no aditivo y comprobar el supuesto de aditividad

Tabla 3 *Diseño factorial 2 × 3 con interacción*

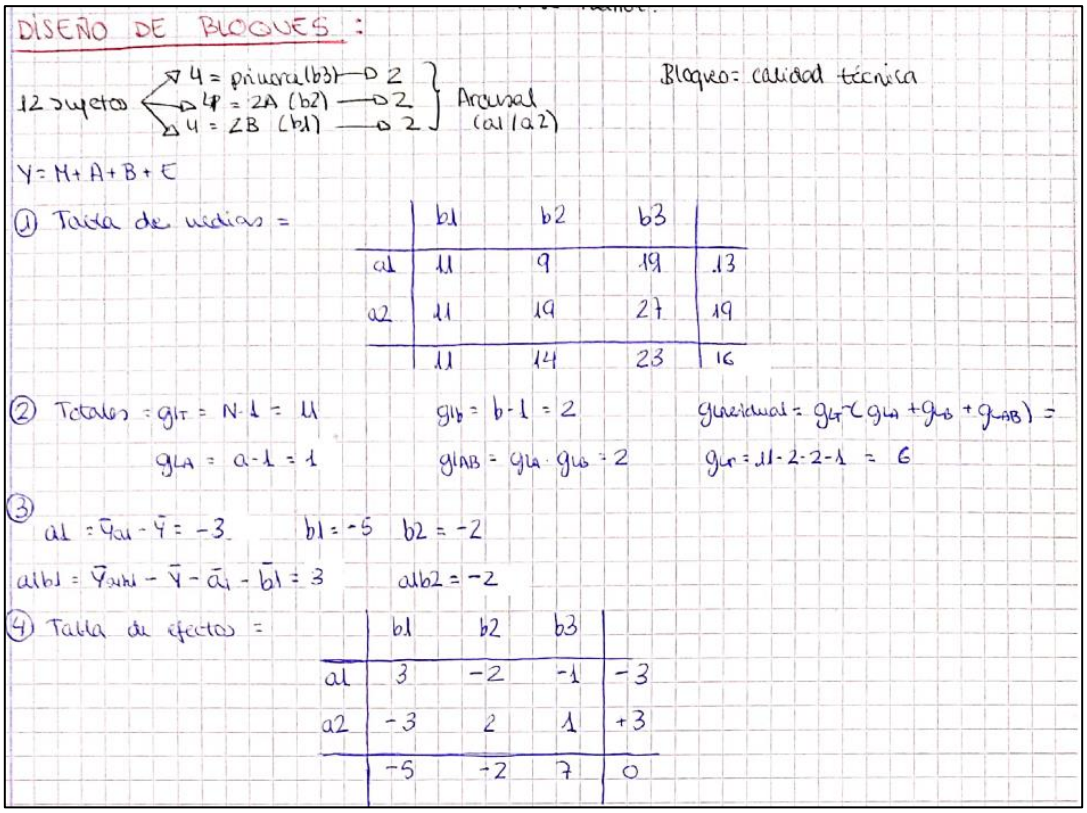
Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	$\hat{\eta}^2$
A					0.050	
B					0.050	
A × B					0.050	
Error						
Total				F_{tablas}	=	
				F_{tablas}	=	

A continuación se detalla en ejercicio tal y como ha sido elaborado por un alumna del grupo VM de la materia de Diseños de Investigación en Psicología de la Facultad de Psicología (Universidad de Valencia). El ejercicio se ha desarrollado hasta completa la tabla del ANOVA del modelo no aditivo para comprobar si se cumple el efecto de adtividad.

Arousal y goles, bloqueo de la técnica futbolística

Trabajo presentado por una alumna del grupo VM de Diseños de Investigación.

Muchas gracias por tu aportación a la clase (24/11/2020)



DISEÑO DE BLOQUES:

12 sujetos: 4 = primera (b3) → 2; 4 = 2A (b2) → 2; 4 = 2B (b1) → 2 } Arousal (a1 / a2)

Bloqueo: calidad técnica

Y = M + A + B + E

① Tabla de medias =

	b1	b2	b3	
a1	11	9	19	13
a2	11	19	27	19
	11	14	23	16

② Totales: $gl_T = N-1 = 11$ $gl_b = b-1 = 2$ $gl_{residual} = gl_T - (gl_A + gl_b + gl_{AB}) =$

$gl_A = a-1 = 1$ $gl_{AB} = gl_A \cdot gl_b = 2$ $gl_E = 11-2-2-1 = 6$

③ $a1 = \bar{Y}_{a1} - \bar{Y} = -3$ $b1 = -5$ $b2 = -2$

$a1b1 = \bar{Y}_{a1b1} - \bar{Y} - \bar{a}_1 - \bar{b}_1 = 3$ $a1b2 = -2$

④ Tabla de efectos =

	b1	b2	b3	
a1	3	-2	-1	-3
a2	-3	2	1	+3
	-5	+2	7	0

⑤

Y	$\bar{Y}$	+ A	+ B	+ E	+ AB
8	16	-3	-5	-3	3
14	:	-3	-5	+3	3
11	:	-3	-2	+2	-2
7	:	-3	-2	-2	-2
21	:	-3	7	+2	-1
17	:	-3	7	-2	-1
14	:	+3	-5	+3	-3
8	:	+3	-5	-3	-3
16	:	+3	-2	-1	2
22	:	+3	-2	+3	2
25	:	+3	7	-2	1
29	:	+3	7	+2	1
SC		108	312	78	56
gl		1	2		

$SC_T = 554$

⑥

	SC	gl	MC	F	p	η^2
A	108	1	108	8'3	<	0'19
B	312	2	156	12	<	0'56
A·B	56	2	28	2'15	>	0'1
E	78	6	13	—		—

$F(1;6) = 5'987$

$F(2;6) = 5'143$

No hay interacción

Como repaso de conceptos:

-Error = Y - Ypronosticada

-Y pronosticada = M + efectos del modelo

-Efectes del modelo no aditivo = A, B, AB

-$Y_{sujeto1}$pronosticada = 16 + (-3, -5, 3)=11

-Recordar que en diseño factorial no aditivo: la media es la puntuación pronosticada. Luego, el error sería: $E_{sujeto1}$ = Y- media)= 8 – 11 = -3.

-Es decir, el error es igual: $E_{sujeto1}$ = Y – Ypronosticada 8 –(16-3-5-3) = 11.

-Por lo tanto, $E_{sujeto1}$ = 8 - 11= -3

A continuación hay que ejecutar el modelo de efectos aditivos (Y=M+A+B+E), comprobar el segundo supuesto y si se cumple pasar a interpretar la hipótesis del estudio. Redactar los resultados

Una vez se ha comprobado que el efecto de interacción AB no es estadísticamente significativo y, por lo tanto, se cumple el supuesto de aditividad de los efectos de los factores del modelo, se pasa ya a plantear el modelo aditivo y se recalculan los resultados de la tabla de ANOVA (tabla 27).

Y hasta aquí la presentación de la alumna.

Tabla 27. Realizar la tabla del ANOVA del modelo aditivo y comprobar el supuesto del efecto del bloqueo. Interpretar el efecto de la varianza sistemática primaria

Tabla 4 *Diseño factorial 2 × 3*

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	$\hat{\eta}^2$
A					0.050	
B					0.050	
Error						
Total				F_{tablas}	=	
				F_{tablas}	=	

Por último y como ejercicio docente, se puede comprobar qué hubiese sucedido si el investigador o investigadora no bloquea la variable que es fuente de varianza sistemática secundaria y hubiese llevado a cabo un diseño unifactorial con la varriable A únicamente (tabla 28).

Tabla 28. ¿Cuál habría sido la conclusión en el caso de no haber realizado el bloqueo?

Tabla 5 *Diseño unifactorial*

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	$\hat{\eta}^2$
A					0.050	
Error						
Total				F_{tablas}	=	

Solución del Supuesto con el SPSS

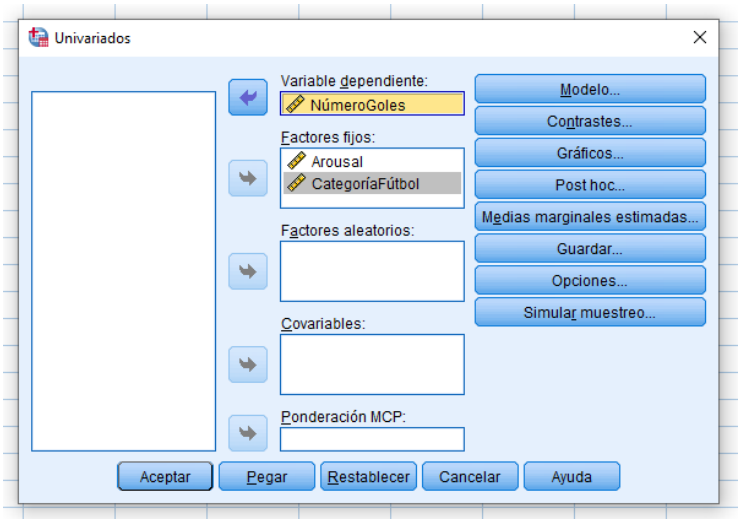
En primer lugar, se introducen los datos en el SPSS para obtener toda la información vinculada con el diseño factorial de la investigación.

	Arousal	CategoríaFútbol	NúmeroGoles
1	Bajo	2ª B	8
2	Bajo	2ª B	14
3	Bajo	2ª A	11
4	Bajo	2ª A	7
5	Bajo	1ª	21
6	Bajo	1ª	17
7	Medio	2ª B	14
8	Medio	2ª B	8
9	Medio	2ª A	16
10	Medio	2ª A	22
11	Medio	1ª	25
12	Medio	1ª	29

→

	Arousal	CategoríaFútbol	NúmeroGoles
1	1	1	8
2	1	1	14
3	1	2	11
4	1	2	7
5	1	3	21
6	1	3	17
7	2	1	14
8	2	1	8
9	2	2	16
10	2	2	22
11	2	3	25
12	2	3	29

Se ejecuta con el SPSS y se obtienen las siguientes tablas relacionadas con los estadísticos descriptivos y el análisis de la varianza (ANOVA).



Análisis univariado de varianza

[ConjuntoDatos0]

Factores inter-sujetos

		Etiqueta de valor	N
Arousal	1	Bajo	6
	2	Medio	6
CategoríaFútbol	1	2ª B	4
	2	2ª A	4
	3	1ª	4

Estadísticos descriptivos

Variable dependiente: NúmeroGoles

Arousal	CategoríaFútbol	Media	Desv. Desviación	N
Bajo	2ª B	11,00	4,243	2
	2ª A	9,00	2,828	2
	1ª	19,00	2,828	2
	Total	13,00	5,404	6
Medio	2ª B	11,00	4,243	2
	2ª A	19,00	4,243	2
	1ª	27,00	2,828	2
	Total	19,00	7,746	6
Total	2ª B	11,00	3,464	4
	2ª A	14,00	6,481	4
	1ª	23,00	5,164	4
	Total	16,00	7,097	12

¿Se cumple el supuesto de aditividad de los efectos del modelo?

Pruebas de efectos inter-sujetos

Variable dependiente: NúmeroGoles

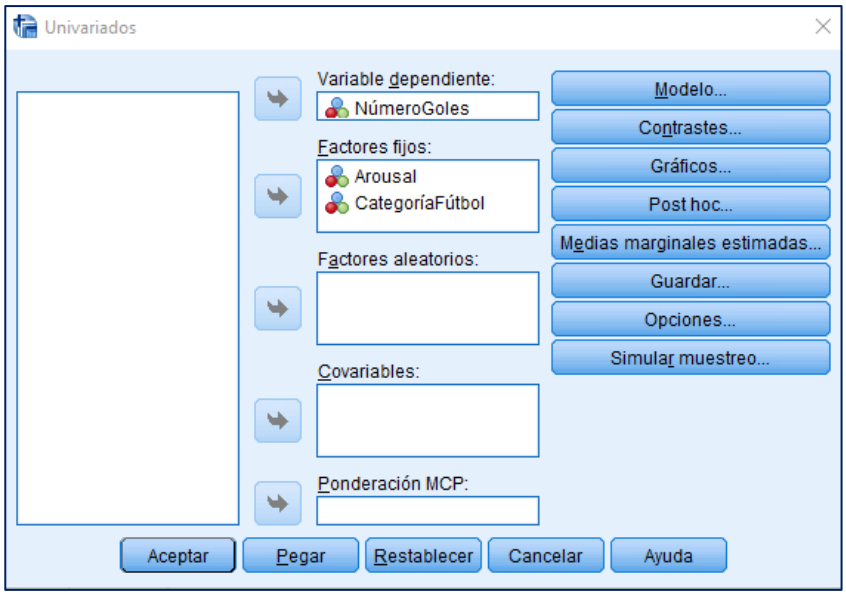
Origen	Tipo III de suma de cuadrados	gl	Media cuadrática	F	Sig.	Eta parcial al cuadrado
Modelo corregido	476,000[a]	5	95,200	7,323	,015	,859
Intersección	3072,000	1	3072,000	236,308	,000	,975
Arousal	108,000	1	108,000	8,308	,028	,581
CategoríaFútbol	312,000	2	156,000	12,000	,008	,800
Arousal * CategoríaFútbol	56,000	2	28,000	2,154	,197	,418
Error	78,000	6	13,000			
Total	3626,000	12				
Total corregido	554,000	11				

a. R al cuadrado = ,859 (R al cuadrado ajustada = ,742)

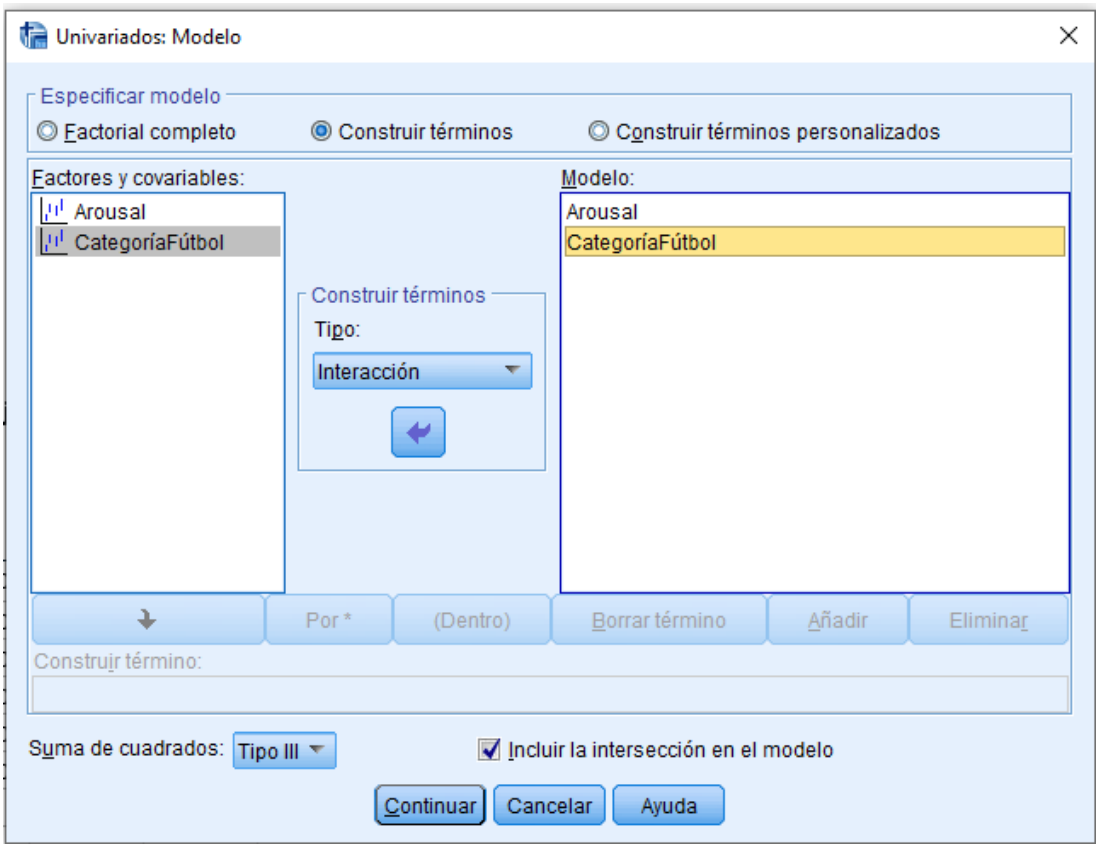
Sí se cumple el supuesto de adtividad de los efectos señaldos en el modelo o ecuación estructural y, posteriormente se procede con el planteamiento del modelo de efectos aditivos de los factores. Se ejecuta ese diseño factorial con efectos aditivos y se comprueba que la variable de bloqueo es estadísticamente significativa y si lo es ya se pasa a la interpretación de la relación entre las variables explicativas que han sido redactadas en la hipótesis de la investigación (variable independiente, que es fuente de varianza sistemática primaria, y la variable dependiente).

Para **ejecutar el diseño de bloques de forma manual** se procede sumando la información de suma de cuadrados y grados de libertad de la fuente de la interacción al término de error. Y se recalcula de nuevo la información de la fuente de varianza del error. Las sumas de cuadrados y grados de libertad de los efectos principales A y B no se modifican. A continuación se recalculan todos los valores de las razones *F*, tanto la del factor A como la del B, ya que se ha modificado el valor del término de error. Y, finalmente se realiza el contraste estadístico comparando con los valores de la *F* teórica teniendo en cuenta los nuevos grados de libertad del error.

Para **ejecutar el diseño de bloques con el SPSS** es necesario forzar el modelo aditivo en el apartado de Modelo dentro de la ventana de modelo lineal general y univariado.



Y se construye un modelo de diseño factorial sin efecto de interacción (modelo aditivo). Se puede observar que en el apartado de Modelo solamente aparecen los efectos principales de A y B.



Se ejecuta ese modelo aditivo y ahora se obtiene el análisis de la varianza del diseño de bloques.

Pruebas de efectos inter-sujetos

Variable dependiente: NúmeroGoles

Origen	Tipo III de suma de cuadrados	gl	Media cuadrática	F	Sig.	Eta parcial al cuadrado
Modelo corregido	420,000[a]	3	140,000	8,358	,008	,758
Intersección	3072,000	1	3072,000	183,403	,000	,958
Arousal	108,000	1	108,000	6,448	,035	,446
CategoríaFútbol	312,000	2	156,000	9,313	,008	,700
Error	134,000	8	16,750			
Total	3626,000	12				
Total corregido	554,000	11				

a. R al cuadrado = ,758 (R al cuadrado ajustada = ,667)

A partir de aquí ya se pueden redactar los resultados del diseño de bloques.

Redacción de los resultados del diseño de bloques

A continuación se redactan los resultados del diseño de bloques aplicado al supuesto 2 de bloqueo y se ha dejado en blanco los resultados de los ANOVA, tamaño del efecto y estadísticos descriptivos para que los lectores y lectoras los completen acudiendo a las tablas anteriores del SPSS donde se encuentra toda la información.

Los resultados del diseño entre-grupos de bloques 2 x 3 (arousal: bajo / medio y categoría en el fútbol: 2ª B, 2ª A y 1ª) univariado respecto a la variable de número de goles marcados en una tanda de treinta lanzamientos de penalti señalan que los datos se ajustan a un modelo aditivo sin interacción entre la variable de arousal y la variable de bloqueo de categoría futbolística, F(___,___) = ___, p = ___, η^2 = ___ y, además, la variable de bloqueo de de categoría futbolística es estadísticamente significativa (F(___,___) = ___, p = ___, η^2 = ___), siendo adecuada el control de su variabilidd en el diseño. El análisis del nivel de arousal permite concluir que hay un efecto estadísticamente significativo de la variable de arousal, F(___,___) = ___, p = ___, η^2 = ___. En concreto, los jugadores que marcan más goles son aquellos que tienen un nivel de arousal medio (M = ___, DT = ___, n = ___) en comparación con los jugadores que tienen un nivel bajo de arousal (M = ___, DT = ___, n = ___). En definitiva, se ha comprobado que tener un nivel medio de arousal mejora la ejecución en el fútbol ya que se marcan más goles en una situación de penalti respecto a los jugadores que tienen un nivel bajo de arousal. Por ello, se recomienda que los jugadores reciban un entrenamiento mediante ejercicios cognitivos que incrementen el arousal hasta un nivel medio.

A continuación se redacta el supuesto con todos los resultados del análisis del diseño entre-grupos de bloques univariado. Conviene tener en cuenta que se ha completado con la eta cuadrado y no con la eta cuadrado parcial ya que ésta sobreestima el valor de proporción explicada.

Los resultados del diseño entre-grupos de bloques 2 x 3 (arousal: bajo / medio y categoría en el fútbol (bloqueo): 2ª B, 2ª A y 1ª) univariado respecto a la variable de número de goles marcados en una tanda de treinta lanzamientos de penalti, señalan que los datos se ajustan a un modelo aditivo sin interacción entre la variable de arousal y la variable de bloqueo de categoría futbolística, F(2, 6) = 2.2, p = .197, η^2 = .10 y, además, la variable de bloqueo de categoría futbolística es estadísticamente

significativa ($F(2, 8) = 9.1$, $p = .008$, $\eta^2 = .56$), siendo adecuada el control de su variabilidd en el diseño. El análisis del nivel de arousal permite concluir que hay un efecto estadísticamente significativo de la variable de arousal, $F(1, 8) = 6.4$, $p = .035$, $\eta^2 = .20$. En concreto, los jugadores que marcan más goles son aquellos que tienen un nivel de arousal medio ($M = 19$, $DT = 7.7$, $n = 6$) en compración con los jugadores que tienen un nivel bajo de arousal ($M = 13$, $DT = 5.4$, $n = 6$). En definitiva, se ha comprobado que tener un nivel medio de arousal mejora la ejecución en el fútbol ya que se marcan más goles en una situación de penalti respecto a los jugadores que tienen un nivel bajo de arousal. Por ello, se recomienda que los jugadores reciban un entrenamiento mediante ejercicios cognitivos que incrementen el arousal hasta un nivel medio.

Ejercicio 1 de diseño de bloques

SUPUESTO. DISEÑO DE BLOQUES. EJECUCIÓN CON EL SPSS

Un profesional está analizando si existen diferencias entre administrar una o dos dosis de un preparado farmacéutico. La terapia farmacológica está diseñada específicamente para el tratamiento de la *ansiedad generalizada*. La muestra experimental está compuesta por pacientes aquejados de esta psicopatología pero que una parte de los mismos tienen un diagnóstico grave y otra parte un diagnóstico moderado. El investigador decide bloquear el grado de gravedad de la enfermedad para mejorar la potencia del diseño. Determine si se cumple el supuesto de aditividad habiendo obtenido los siguientes resultados.

	b_1 *moderado*	b_2 *grave*
	3	14
a_1 *2 dosis*	2	10
	1	12
	13	18
a_2 *1 dosis*	8	14
	9	16

1º. Para comprobar el supuesto de aditividad hay que comprobar si hay un efecto de interacción significativo entre el *tratamiento* y la *gravedad* de la *ansiedad generalizada*. EJECUTAR EL MODELO CON EL SPSS.

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	η^2
A	108	1	108	27	< 0.05	0.314
B	192	1	192	48	< 0.05	0.558
A × *B*	12	1	12	3	> 0.05	0.035
Error	32	8	4			
Total	344	11				

Como $F_{A \times B\ 1,\ 8} = 3$ mantenemos la hipótesis nula ($p > 0.05$). Por tanto, no hay efectos de interacción y se cumple el supuesto de aditividad. Los efectos del tratamiento y la gravedad de la psicopatología son por tanto independientes, si no fuera así entonces estaríamos afirmando que en función de que la enfermedad se encontrase en un punto grave o moderado el tratamiento tendría efectos diferentes.

2º Comprobar que si existe relación entre la variable dependiente y la bloqueada

Como ya calculamos en el ejercicio anterior las sumas de cuadrados correspondientes a cada fuente de variación y no encontramos diferencias significativas en el término de interacción, eliminamos el componente de interacción de la ecuación estructural y aplicamos la prueba de la hipótesis de nuevo. EJECUTAR EL MODELO CON EL SPSS.

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	η^2
A	108	1	108.000	22.091	< 0.05	0.314
B	192	1	192.000	39.272	< 0.05	0.558
Error	44	9	4.889			
Total	344	11				

Sí que existe relación entre la variable bloqueada y la dependiente ($F_{B\ 1,\ 9} = 39.27$; $p < 0.05$), haber bloqueado el grado de la enfermedad ha supuesto descontar 192 de la suma de cuadrados del error. Si no se hubiese rechazado la hipótesis nula en el término de bloqueo habría supuesto que la variable dependiente y la bloqueada no tienen relación, no estando justificado aplicar el diseño de bloques.

Cabe por tanto deducir que existen diferencias en la variable dependiente en función de si se administra una o dos dosis del preparado. En el caso de que se administre *una* única dosis la media observada es 13 y si se administran *dos*, la ansiedad observada disminuye a 7.

Cuál habría sido el resultado si no hubiese bloqueado el investigador el efecto de la variable *gravedad*.

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	η^2
A	108	1	108.0	4.576	> 0.05	0.314
Error	236	10	23.6			
Total	344	11				

No se habría demostrado que existiese diferencia entre administrar una o dos dosis del fármaco ($F_{A\ 1,\ 10} = 4.58$; $p > 0.05$), pese a que la estimación del tamaño del efecto es la misma que en el diseño de bloques ($\eta^2 = 0.314$).

Ejercicio 2. SPSS. A x B : modelo aditivo

EJECUTAR CON EL SPSS. DISEÑO DE BLOQUES

1º COMPROBAR EL SUPUESTO DE ADITIVIDAD DE LOS FACTORES

Pruebas de los efectos inter-sujetos

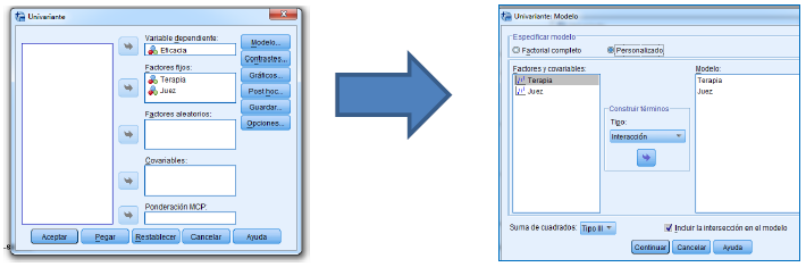
Variable dependiente:Eficacia

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Modelo corregido	246,000[a]	5	49,200	14,057	,000	,854
Intersección	1800,000	1	1800,000	514,286	,000	,977
Terapia	162,000	1	162,000	46,286	,000	,794
Juez	84,000	2	42,000	12,000	,001	,667
Terapia * Juez	,000	2	,000	,000	1,000	,000
Error	42,000	12	3,500			
Total	2088,000	18				
Total corregida	288,000	17				

a. R cuadrado = ,854 (R cuadrado corregida = ,793)

2º REDEFINIR EL MODELO: 1º COMPROBAR EL EFECTO ESTADÍSTICAMENTE SIGNIFICATIVO DE LA VARIABLE DE BLOQUEO Y 2º INTERPRETAR EL EFECTO DE LA VARIABLE DE TRATAMIENTO CON ESE DISEÑO DE BLOQUES

ANALIZAR----MODELO LINEAL GENERAL----UNIVARIANTE----MODELO



ejecutar

Pruebas de los efectos inter-sujetos

Variable dependiente:Eficacia

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Modelo corregido	246,000[a]	3	82,000	27,333	,000	,854
Intersección	1800,000	1	1800,000	600,000	,000	,977
Terapia	162,000	1	162,000	54,000	,000	,794
Juez	84,000	2	42,000	14,000	,000	,667
Error	42,000	14	3,000			
Total	2088,000	18				
Total corregida	288,000	17				

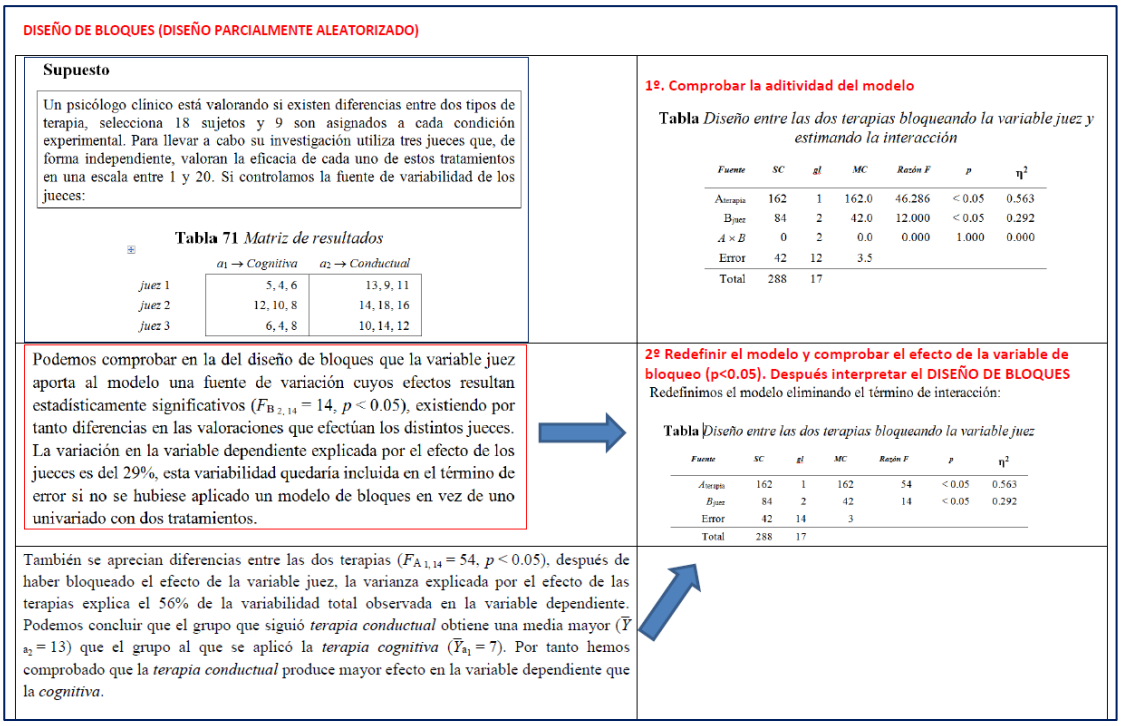
a. R cuadrado = ,854 (R cuadrado corregida = ,823)

Medias marginales estimadas

1. Terapia

Variable dependiente:Eficacia

Terapia	Media	Error típ.	Intervalo de confianza 95%	
			Límite inferior	Límite superior
1	7,000	,577	5,762	8,238
2	13,000	,577	11,762	14,238



DISEÑO DE BLOQUES (DISEÑO PARCIALMENTE ALEATORIZADO)

Supuesto

Un psicólogo clínico está valorando si existen diferencias entre dos tipos de terapia, selecciona 18 sujetos y 9 son asignados a cada condición experimental. Para llevar a cabo su investigación utiliza tres jueces que, de forma independiente, valoran la eficacia de cada uno de estos tratamientos en una escala entre 1 y 20. Si controlamos la fuente de variabilidad de los jueces:

Tabla 71 *Matriz de resultados*

	$a_1 \rightarrow$ *Cognitiva*	$a_2 \rightarrow$ *Conductual*
juez 1	5, 4, 6	13, 9, 11
juez 2	12, 10, 8	14, 18, 16
juez 3	6, 4, 8	10, 14, 12

Podemos comprobar en la del diseño de bloques que la variable juez aporta al modelo una fuente de variación cuyos efectos resultan estadísticamente significativos ($F_{B\,2,\,14} = 14$, $p < 0.05$), existiendo por tanto diferencias en las valoraciones que efectúan los distintos jueces. La variación en la variable dependiente explicada por el efecto de los jueces es del 29%, esta variabilidad quedaría incluida en el término de error si no se hubiese aplicado un modelo de bloques en vez de uno univariado con dos tratamientos.

También se aprecian diferencias entre las dos terapias ($F_{A\,1,\,14} = 54$, $p < 0.05$), después de haber bloqueado el efecto de la variable juez, la varianza explicada por el efecto de las terapias explica el 56% de la variabilidad total observada en la variable dependiente. Podemos concluir que el grupo que siguió *terapia conductual* obtiene una media mayor ($\bar{Y}_{a_2} = 13$) que el grupo al que se aplicó la *terapia cognitiva* ($\bar{Y}_{a_1} = 7$). Por tanto hemos comprobado que la *terapia conductual* produce mayor efecto en la variable dependiente que la *cognitiva*.

1º. Comprobar la aditividad del modelo

Tabla *Diseño entre las dos terapias bloqueando la variable juez y estimando la interacción*

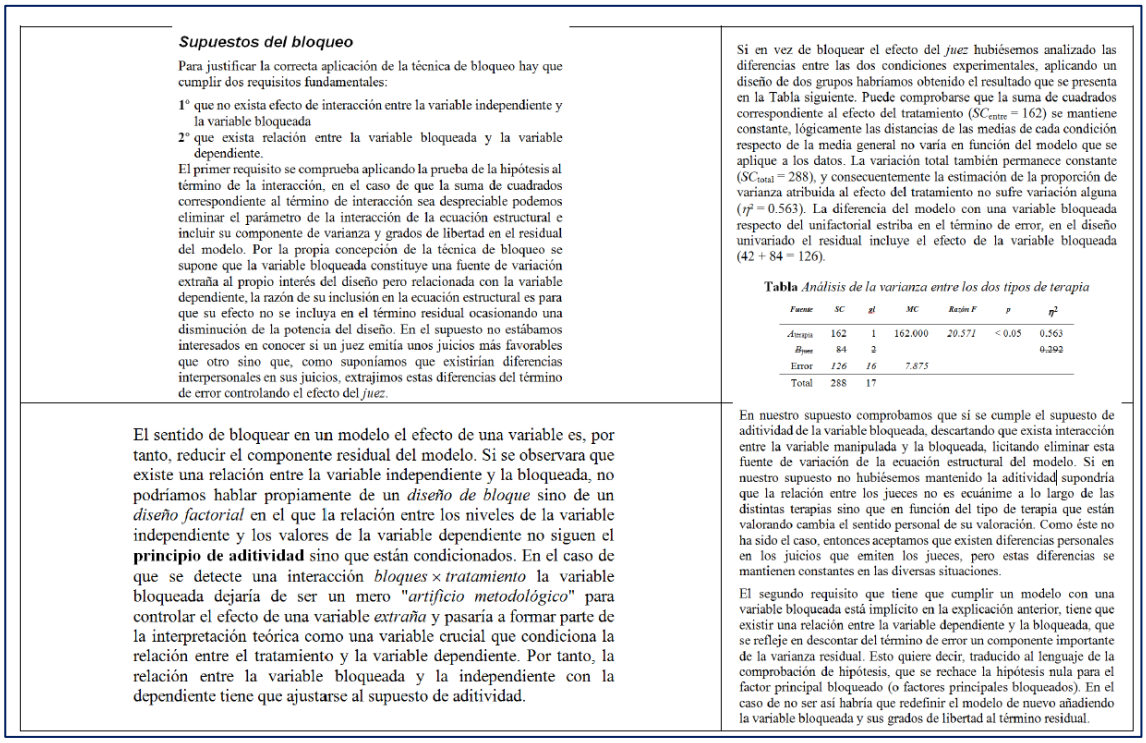
Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	η^2
$A_{terapia}$	162	1	162.0	46.286	< 0.05	0.563
B_{juez}	84	2	42.0	12.000	< 0.05	0.292
$A \times B$	0	2	0.0	0.000	1.000	0.000
Error	42	12	3.5			
Total	288	17				

2º Redefinir el modelo y comprobar el efecto de la variable de bloqueo (p<0.05). Después interpretar el DISEÑO DE BLOQUES

Redefinimos el modelo eliminando el término de interacción:

Tabla *Diseño entre las dos terapias bloqueando la variable juez*

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	η^2
$A_{terapia}$	162	1	162	54	< 0.05	0.563
B_{juez}	84	2	42	14	< 0.05	0.292
Error	42	14	3			
Total	288	17				



Supuestos del bloqueo

Para justificar la correcta aplicación de la técnica de bloqueo hay que cumplir dos requisitos fundamentales:

1º que no exista efecto de interacción entre la variable independiente y la variable bloqueada

2º que exista relación entre la variable bloqueada y la variable dependiente.

El primer requisito se comprueba aplicando la prueba de la hipótesis al término de la interacción, en el caso de que la suma de cuadrados correspondiente al término de interacción sea despreciable podemos eliminar el parámetro de la interacción de la ecuación estructural e incluir su componente de varianza y grados de libertad en el residual del modelo. Por la propia concepción de la técnica de bloqueo se supone que la variable bloqueada constituye una fuente de variación extraña al propio interés del diseño pero relacionada con la variable dependiente, la razón de su inclusión en la ecuación estructural es para que su efecto no se incluya en el término residual ocasionando una disminución de la potencia del diseño. En el supuesto no estábamos interesados en conocer si un juez emitía unos juicios más favorables que otro sino que, como suponíamos que existirían diferencias interpersonales en sus juicios, extrajimos estas diferencias del término de error controlando el efecto del *juez*.

El sentido de bloquear en un modelo el efecto de una variable es, por tanto, reducir el componente residual del modelo. Si se observara que existe una relación entre la variable independiente y la bloqueada, no podríamos hablar propiamente de un *diseño de bloque* sino de un *diseño factorial* en el que la relación entre los niveles de la variable independiente y los valores de la variable dependiente no siguen el **principio de aditividad** sino que están condicionados. En el caso de que se detecte una interacción *bloques* × *tratamiento* la variable bloqueada dejaría de ser un mero "*artificio metodológico*" para controlar el efecto de una variable *extraña* y pasaría a formar parte de la interpretación teórica como una variable crucial que condiciona la relación entre el tratamiento y la variable dependiente. Por tanto, la relación entre la variable bloqueada y la independiente con la dependiente tiene que ajustarse al supuesto de aditividad.

Si en vez de bloquear el efecto del *juez* hubiésemos analizado las diferencias entre las dos condiciones experimentales, aplicando un diseño de dos grupos habríamos obtenido el resultado que se presenta en la Tabla siguiente. Puede comprobarse que la suma de cuadrados correspondiente al efecto del tratamiento ($SC_{entre} = 162$) se mantiene constante, lógicamente las distancias de las medias de cada condición respecto de la media general no varía en función del modelo que se aplique a los datos. La variación total también permanece constante ($SC_{total} = 288$), y consecuentemente la estimación de la proporción de varianza atribuida al efecto del tratamiento no sufre variación alguna ($\eta^2 = 0.563$). La diferencia del modelo con una variable bloqueada respecto del unifactorial estriba en el término de error, en el diseño univariado el residual incluye el efecto de la variable bloqueada ($42 + 84 = 126$).

Tabla *Análisis de la varianza entre los dos tipos de terapia*

Fuente	*SC*	*gl*	*MC*	*Razón F*	*p*	η^2
$A_{terapia}$	162	1	162.000	*20.571*	< 0.05	0.563
B_{juez}	84	2				0.292
Error	*126*	*16*	*7.875*			
Total	288	17				

En nuestro supuesto comprobamos que sí se cumple el supuesto de aditividad de la variable bloqueada, descartando que exista interacción entre la variable manipulada y la bloqueada, licitando eliminar esta fuente de variación de la ecuación estructural del modelo. Si en nuestro supuesto no hubiésemos mantenido la aditividad supondría que la relación entre los jueces no es ecuánime a lo largo de las distintas terapias sino que en función del tipo de terapia que están valorando cambia el sentido personal de su valoración. Como éste no ha sido el caso, entonces aceptamos que existen diferencias personales en los juicios que emiten los jueces, pero estas diferencias se mantienen constantes en las diversas situaciones.

El segundo requisito que tiene que cumplir un modelo con una variable bloqueada está implícito en la explicación anterior, tiene que existir una relación entre la variable dependiente y la bloqueada, que se refleje en descontar del término de error un componente importante de la varianza residual. Esto quiere decir, traducido al lenguaje de la comprobación de hipótesis, que se rechace la hipótesis nula para el factor principal bloqueado (o factores principales bloqueados). En el caso de no ser así habría que redefinir el modelo de nuevo añadiendo la variable bloqueada y sus grados de libertad al término residual.

Redactar los resultados

Els resultats del disseny entre-grups de blocs 2 x 3 (teràpia x jutge) assenyalen que les dades s'ajusten a un model additiu sense interacció entre la variable de tractament i la variable de bloqueig $F(2,12) = 0$, $p = 1$, $\eta^2 = 0$) i, a més, la variable de bloqueig de jutge és estadísticament significativa, $F(2,14) = 14$, $p < .001$, $\eta^2 = .67$),

sent adequada el control de la seua variabilitat en el disseny. L'anàlisi de l'eficàcia de la teràpia permet concloure que hi ha un efecte estadísticament significatiu de la teràpia $F(1,14) = 54$, $p < .001$, $\eta^2 = .80$). En concret, l'eficàcia de la teràpia és més gran quan els pacients reben la teràpia conductual ($M = 13$, $DT = 0.58$, $n = 9$) en comparació amb la teràpia cognitiva ($M = 7$, $DT = 0.58$, $n = 9$).

Referencias

Altman, D. (1994). The scandal of poor medical research. *British Medical Journal*, 308, 283-284.

Altman, D. G. & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal*, 311, 485.

Altman, D. G. (1998). Confidence intervals for the number needed to treat*. British Medical Journal*, *317*, 1309-1312.

American Psychological Association (1953). *Ethical standards of psychologists*. Washington, DC: American Psychological Association.

American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th Ed.). Washington, DC: American Psychological Association.

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060-1073.

American Psychological Association (2010). *Publication Manual of the American Psychological Association* (6th Ed.). Washington, DC: American Psychological Association.

American Psychological Association (2020). *Publication Manual of the American Psychological Association* (7th Ed.). Washington, DC: American Psychological Association.

Amrhein V, Korner-Nievergelt F, Roth T. 2017. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJournal*, *5*, e3544 https://doi.org/10.7717/peerj.3544

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*, 1547-1562. Doi: 10.1177/0956797617723724

Antonelli, M., & Sandroni, C. (2013). Hydroxyethyl starch for intravenous volume replacement: More harm than benefit. *Journal of the American Medical Association*, *309*(7), 723-724.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in Psychology Why do we need them? What might they be? *American Psychologist*, *63*, 839-851.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal Article Reporting Standards for quantitative research in Psychology: The APA Publications and Communications Board Task Force report. *American Psychologist*, *73*, 3-25.

Ariza, L. M. (2020). A la caza de fraudes en la ciencia. *El País Semanal, 15 marzo 2020*. https://elpais.com/elpais/2020/03/10/eps/1583855907_030021.html

Arnau, J. (1981). *Diseños experimentales en psicología y educación. Vol.* I. México D.F.: Trillas (Preedición).

Ato, M. (1991). *Investigación en Ciencias del Comportamiento. I: Fundamentos*. Barcelona: PPU.

Badenes-Ribera, L., Frias-Navarro, D., Pascual-Soler, M., & Monterde-i-Bort, H. (2016). Knowledge level of effect size statistics, confidence intervals and meta-analysis in Spanish academic psychologists. *Psicothema*, *28*(4), 448-456. http://doi.org/10.7334/psicothema2016.24

Badenes-Ribera, L., Frías-Navarro, Iotti, N. O., Bonilla-Campos, A., & Longobardi, C. (2018). Perceived statistical knowledge level and self-reported statistical practice among academic psychologists. *Frontiers in Psychology*, 9, 1-12.

Bakker, M., van Dijk, A., Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi:10.1177/1745691612459060

Belia, S., Fidler, F., Wiiliams, J., & Cumming, G. (2005). Researcher misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389-396.

Bouter, L. (2008). *Knowledge as public property: The societal relevance of scientific research. Higher Education Management and Policy*. Disponible en: https://research.vu.nl/en/publications/knowledge-as-public-property-the-societal-relevance-of-scientific

Bouter, L. (2019). *What is research integrity? Filmed at the Amsterdam Scholarly Summit, 2 July 2019*. Disponible en: http://editorresources.taylorandfrancis.com/wp-content/uploads/2019/07/What-is-research-integrity-Transcript.pdf

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building* (1st ed.). New York, NY: Wiley & Sons.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (2005). *Statistics for experimenters: design, innovation and discovering* (2nd ed.). New York, NY:Wiley & Sons.

Breen, K. J. (2016). Research misconduct: Time for a re-think?. *Internal Medicine Journal*, *46*(6), 728-733. doi:10.1111/imj.13075.

Campbell JP. 1982. Editorial: some remarks from the outgoing editor. *Journal of Applied Psychology*, 67, 691-700.

Chida, Y., Hamer, M., Wardle, J., & Steptoe, A. (2008). Do stress-related psychosocial factors contribute to cancer incidence and survival? *Nature Clinical Practice Oncology*, *5*, 466-475. https://doi.org/10.1038/ncponc1134

Chittaranjan, A. (2015). The Numbers Needed to Treat and Harm (NNT, NNH) statistics: What they tell us and what they do not. *Journal of Clinical Psychiatry*, 76(3):e330–e333. Doi:10.4088/JCP.15f09870).

Chow, S. H., & Liu J. P. (2004). *Design and analysis of clinical trials* (2nd ed.). Hoboken, NJ: Wiley.

Ciapponi, A. (2018). AMSTAR-2: herramienta de evaluación crítica de revisiones sistemáticas de estudios de intervenciones de salud. *Evidencia, Actualizacion En La práctica Ambulatoria*, *21*(1). Recuperado a partir de http://www.evidencia.org/index.php/Evidencia/article/view/6834

Código Deontológico del Psicólogo (2010). Colegio Oficial de Psicólogos de Madrid. Disponible en: http://www.copmadrid.org/webcopm/recursos/codigodeontologicojunio2010.pdf

Código Europeo de Conducta para la Integridad en la Investigación. Edición revisada (2018). *ALLEA-All European Academies, Berlín 2018*. Disponible en: https://www.allea.org/wp-content/uploads/2018/01/SP_ALLEA_Codigo_Europeo_de_Conducta_para_la_Integridad_en_la_Investigacion.pdf

Cohen, J. (1988). *Statistical power analysisfor the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Consejo Superior de Investigaciónes Cientíicas (2021). *Código de Buenas Prácticas Científicas del CSIC*. Editorial CSIC. https://www.icmm.csic.es/es/comision-igualdad/img/Codigo-de-Buenas-Practicas.pdf

Cook, R. J., & Sackett, D. L. (1995). The Number Needed to Treat - a Clinically Useful Measure of Treatment Effect. *British Medical Journal*, *310*(6977), 452-454.

Cumming, G. & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-575.

Cumming, G. & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*, 7-29.

David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *American Statistician*, *49*, 121-133. doi: 10.2307/2684625

De Lecuona, I. (2020). La integridad científica en las instituciones de educación superior en el siglo XXI. *Dilemata, Revista Internacional de Éticas Aplicadas*, *31*, 95-107.

delMas, R.C. & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*, 55-82.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*, 917-928.

Duyx, B., Urlings, M. J. E., Swaen, G. M. H. M., Bouter, L. M., y Zeegers, M. P. (2017). Scientific citations favor positive results: a systematic review and meta-analysis. *Journal of Clinical Epidemiology*, *88*, 92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge UK: Cambridge University Press.

Estrada-Pérez, C. y Jaimes-Barragán, F. (2013). Ronda clínica y epidemiológica Estudios de superioridad vs. Estudios de no inferioridad. *Iatreia*, *26*, 232-237. https://revistas.udea.edu.co/index.php/iatreia/article/view/14516

Evans, R. B., Sexton, V. S., & Cadwallader, T. C. (Eds.). (1992). *The American Psychological Association: A historical perspective. American Psychological Association*. https://doi.org/10.1037/10111-000

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*, 83-96.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *Plos One*, *4*(5), e5738. https://doi.org/10.1371/ journal.pone.0005738

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Fisher, R. A. (1937), The Design of Experiments(2nd ed.), Edinburgh: Oliver and Boyd.

Frías-Navarro, D. & Pascual-Llobell, J. (2003). Psicología clínica basada en pruebas: efecto del tratamiento. *Papeles del Psicólogo*, *85*, 11-18.

Frías-Navarro, D. & Pascual-Soler, M. (2011). Prácticas del análisis factorial exploratorio en la investigación sobre conducta del consumidor y marketing: una visión general y algunas recomendaciones. *Suma Psicológica*, *19*, 47-58.

Frías-Navarro, D. (2011). *Técnica estadística y diseño de investigación*. Valencia: Palmero Ediciones.

Frías-Navarro, D. (2020). *Apuntes de consistencia interna de las puntuaciones de un instrumento de medida*. Universidad de Valencia. España. Disponible en: https://www.uv.es/friasnav/AlfaCronbach.pdf

Frías-Navarro, D., & Pascual, J. (2003). Psicología clínica basada en pruebas: efecto del tratamiento. *Papeles del Psicólogo*, *85*, 11-18.

Frías-Navarro, D., Pascual, J., & García, J. F. (2000). La hipótesis nula y la significación práctica. *Revista de la Asociación Española de Metodología de las Ciencias del Comportamiento*, *Volumen especial*, 181-185.

Frías-Navarro, D., Pascual, J., & García, J. F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicothema*, *12*, 236-240.

Frías-Navarro, D., Pascual, J., & García, J. F. (2002). Concepto y método de la Psicología Basada en la Evidencia. *Interpsiquis* 2002.

Frías-Navarro, D., Pascual-Llobell, J., Monterde-i-Bort, H., & García-Pérez, F. (2007). Tests de equivalencia. Actas IX Congreso de Metodología de las Ciencias Sociales y de la Salud. Granada: Ediciones Sider.

Frías-Navarro, D., Pascual-Llobell, J., Pascual-Soler, M., Perez-Gonzalez, J., & Berríos-Riquelme, J. (2020). Replication crisis or an opportunity to improve scientific production? *European Journal of Education*, 55(4), 618-631. http://doi.org/ 10.1111/ejed.12417

Frías-Navarro, D., Pascual-Soler, M., Berrios-Riquelme, J., Gomez-Frias, R., & Caamaño-Rocha, L. (2021). COVID-19. Effect of moral messages to persuade the population to stay at home in Spain, Chile, and Colombia. *Journal of Spanish Psychology*. Published online by Cambridge University Press: 13 August 2021.

Frías-Navarro, D., Pascual-Soler, M., Monterde-i-Bort, H., Perezgonzalez, J. & Pascual-Llobell, J. (2021). Opinions of Spanish scientists on the science and behavior of the researcher. *Journal of Spanish Psychology*. Published online by Cambridge University Press: 05 February 2021.

Funder, D. C. & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*, 156-168. doi: 10.1177/2515245919847202 www.psychologicalscience.org/AMMPS

Furukawa TA, Leucht S (2011) How to Obtain NNT from Cohen's d: Comparison of Two Methods. *PLoS One*, *6*(4): e19070. doi:10.1371/ journal.pone.0019070

Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. Cambridge: Cambridge University Press.

García de la Banda, G., Martínez-Abascal, M.A., Riesco, M., & Pérez G. (2004). La respuesta de cortisol ante un examen y su relación con otros acontecimientos estresantes y con algunas características de personalidad. *Psicothema*, *16*, 294-298.

García, J. F., Frías-Navarro, D., & Pascual, J. (2006). *Los diseños de la investigación experimental. Comprobación de las hipótesis*. Valencia: editorial CSV.

Gisberta, J. P. & Bonfill, X (2004). ¿Cómo realizar, evaluar y utilizar revisiones sistemáticas y metaanálisis? *Gastroenterología y Hepatología*, *27*, 129-149.

Glass, G. V, Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumption underlying the fixed effects analysis of variance and covariance. *Review of Educational Research*, *42*, 237-288.

Glasziou, P. P., Sanders, D., & Hoffmann, T. (2020). Waste in covid-19 research. *British Medical Journal*, 1847.

Glezerson, B. A. y Bryson, G. L. (2019). Focus on peer review: An online peer review course by Nature Masterclasses. *Canadian Journal of Anesthesia*, *66*, 348-349. https://doi.org/10.1007/s12630-018-1254-4

Gómez-Benito, J. e Hidalgo, M. D. (2015). *La validez en los tests, escalas y cuestionarios. Meta base de recursos educativos de la UAEM (Universidad Autónoma del Estado de Morelos)*. http://metabase.uaem.mx//handle/123456789/1014

González de Dios, J., González-Muñoz, M., Alonso-Arroyo, A., y Benavent, A. (2014). Comunicación científica (XVIII). Conocimientos básicos para leer (y escribir) un artículo científico (5): listas de comprobación de documentos. *Acta Pediátrica Española*, *72*(11), e389-e392.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*:337-350

Gruijters, S, L K. & Peters, G-J. Y. (2019). Introducing the Numbers Needed for Change (NNC): A practical measure of effect size for intervention research. *Preprint*. DOI: 10.31234/osf.io/2bau7

Grupo Europeo de Ética de la Ciencia y las Nuevas Tecnologías (2018). *Declaración sobre la elaboración de un código de conducta para la integridad en la investigación en proyectos financiados por la Comisión Europe*. Oficina de Publicaciones de la Unión Europea.

Hallahan, M., & Rosenthal, R. (1996). Statistical power: concepts, procedures, and appli- cations. *Behavioral Research Theory*, *34*(5/6), 489-499.

Hancock, G.R. & Mueller, R.O. (2010). *The Reviewer's guide to quantitative methods in the Social Sciences*. New York, NY: Taylor & Francis.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*, 315-339. doi:10.2307/1165127

Herruzo, I. (2004). Tratamiento hormonal del cáncer de mama. *Oncología*, *27*, 427-434.

Hoekstra, R., Finch, S., Kiers H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, *13*, 1033-1037.

Hofmann, B., Lone, B. J., Brandt, E. M., Gert, H., Niklas, J., & Søren, H. (2020). Research integrity among PhD students at the Faculty of Medicine: A comparison of three scandinavian universities. *Journal of Empirical Research on Human Research Ethics*,*15*(4), 320-329. doi: 10.1177/1556264620929230

Holtfreter, K., Reisig, M. D., Pratt, T. C., & Mays, R. D. (2020). The perceived causes of research misconduct among faculty members in the natural, social, and applied sciences. *Studies in Higher Education*, *45*, 2162-2174. Doi: 10.1080/03075079.2019.1593352

Huck, S. W. (2007). Reform in statistical education. *Psychology in the School*, *44*, 527-533.

Imai, K., King, G., & Stuart. E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, *171*, 481-502.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine*, *2*(8), https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2018) Meta-research: Why research on research matters. *PLoS Biology*, *16*(3): e2005468, 1-6.

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biology*, *13*(10): e1002264, 1-7.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, *31*, 203-222.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University press.

Jones, B., Jarvis, P., Lewis, J. A., & Ebbutt, A. F. (1996). Trials to assess equivalence: the importance of rigorous methods. *BMJ*. 1996;313:36-9.

Kakuk, P. (2009). The legacy of the Hwang case: Research misconduct in biosciences. *Science and Engineering Ethics*, *15*(4), 545–562. https://doi.org/10.1007/s11948- 009-9121-x

Kass, R. E., Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York, NY:Holt, Rinehart and Winston.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350-386.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Kirk, R.E. (1995). *Experimental design: procedures for the behavioral sciences*. (3rd Ed.). Belmont, CA: Brooks/Cole Publishing.

Kish, L. (1975). *Diseño estadístico para la investigación* (versión en castellano, 1995). Madrid: Centro de Investigaciones Sociológicas (CIS), Colección Monografías Nº 146.

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An Assessment of Clinically Useful Measures of the Consequences of Treatment. *New England Journal of Medicine*, *318*(26), 1728– 1733. http://doi.org/10.1056/NEJM198806303182605

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal Article Reporting Standards for qualitative primary, qualitative meta-analytic, and mixed methods research in Psychology: The APA Publications and Communications Board Task Force report. *American Psychologist*, *73*, 26-46.

Liu, X. & Stephen Raudenbush, S. (2004). A note on the noncentrality parameter and effect size estimates for the F Test in ANOVA. *Journal of Educational and Behavioral Statistics*, 29, 251-255. https://www.jstor.org/stable/3701269

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "F" test. *Review of Educational Research*, *66*, 579-619.

Mariño-Hernández, E. L. (2016). Fraudes científicos y otras malas prácticas. *Farma Journal*, *1*, 183-187.

Martínez-Franco, M., Nirta-Pérez, A. R., y Donado-Gómez, J. H. (2021). Tipos de ensayos clínicos con asignación aleatoria publicados en PubMed durante 40 años. *Acta Médica Colombiana*, *46*, 1-8. https://doi.org/10.36104/amc.2021.1884

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature*, *435*(7043), 737–738.

Maxwell, S. E. & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd Ed.). Mahwah, NJ: Erlbaum.

Maxwell, S. E., Delaney, H.D., & Kelley (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). New York, NY: Routledge

Mayor, J. & Pérez, J. (1989). Psicología o psicologías? Un problema de identidad. En J. Mayor y J. L. Pinillos (Eds.), *Tratado de psicología general, vol. 1: Historia, teoría y método*. (pp. 3-69). Madrid: Alhambra Universidad.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., … Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128-165. doi:10.1037/0003-066x.56.2.128

Monterde i Bort, H., Pascual, J., & Frías-Navarro, D. (2006). Errores de interpretación de los métodos estadísticos: importancia y recomendaciones. *Psicothema*, *18*, 848-856.

Mood, A. M. & Graybill, F. A. (1972). *Introducción a la teoría de la estadística*. Madrid: Aguilar ediciones.

Nesselroade, J. R. & Cattell, R. B. (1988) (Eds.). *Handbook of multivariate experimental Psychology*. (2nd ed. rev.). New York, NY: Plenum Press.

Onwuegbuzie, A. J. (2001). Common methodological, analytical, and interpretational errors in published educational studies: An analysis of the 1998 volume of the British Journal of Educational Psychology. *Educational Research Quarterly*, *26*, 11-22.

Pascual-Llobell, Frías-Navarro, D., & Monterde-i-Bort, H. (2004). Tratamientos psicológicos con apoyo empírico y práctica basada en la evidencia. *Papeles del Psicólogo*, 87.

Pascual-Llobell, J., Frías-Navarro, D., & García, J. F. (1996). *Manual de Psicología Experimental*. Barcelona: Ariel.

Pascual-Llobell, J., Frías-Navarro, D., & García, J. F. (2000). El procedimiento de significación estadística (NHST): su trayectoria y actualidad. *Revista de Historia de la Psicología*, *21*, 9-26.

Pascual-Llobell, J., Frías-Navarro, D., & García-Pérez, J. F. (1996). *Manual de Psicología Experimental*. Barcelona: Ariel.

Pascual-Llobell, J., Frías-Navarro, D., & García-Pérez, J. F. (2000). El procedimiento de significación estadística (NHST): su trayectoria y actualidad. *Revista de Historia de la Psicología*, *21*, 9-26.

Pascual-Llobell, J., Frías-Navarro, D., & Monterde-i-Bort, H. (2004). Tratamientos psicológicos con apoyo empírico y práctica basada en la evidencia. *Papeles del Psicólogo*, *87*, 1-8.

Pelosi, A. J. (2019). Personality and fatal diseases: Revisiting a scientific scandal. *Journal of Health Psychology*, *24*(4), 421-439. https://doi.org/10.1177/1359105318822045

Perugini, M., Gallucci, M., Constantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*, 319-332. doi:10.1177/1745691614528519

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., Evans, S. J. W. (2006). CONSORT Group for the. Reporting of Noninferiority and Equivalence Randomized Trials. *JAMA*, *295*(10):1152.

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., Evans, S. J. W. (2012). Reporting of noninferiority and equivalence randomized trials (for the CONSORT Group). *Journal American Medical Association*, *308*(24):2594-604.

Pita-Fernández, S. & López-de-Ullibarri, I. (1999). Número necesario de pacientes a tratar para reducir un evento. *Cuadernos de Atención Primaria*, *6*, 96-98.

Pupovac, V., & Fanelli, D. (2015). Scientists admitting to plagiarism: A meta-analysis of surveys. *Science and Engineering Ethics*, *21*(5), 1331–1352.

Reinhart, A. (2015). *Statistics done wrong*. San Francisco: No Starch Press.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, *45*, 775-777. doi:10.1037/0003-066x.45.6.775

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews. *Annual Review of Psychology*, *52*, 59-82. doi:10.1146/annurev.psych.52.1.59

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists*. Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *57*, 221-237. doi:10.1037/h0087427

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.

Rossi, J. S. (1990). Statistical power of psychological research: what have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646-657.

Sackett, D. L., Strauss, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM* (2nd Ed.). Edinburgh: Churchill Livingstone.

Sánchez-Meca J., Alacid-de-Pascual I., López-Pina J. A., de la Cruz Sánchez-Jiménez, J. (2016). Meta-análisis de generalización de la fiabilidad del inventario de obsesiones de Leyton versión para niños auto-aplicada. *Revista Española de Salud Pública*, *90*, e1-e14.

Sánchez-Meca, J. & Botella, J. (2010). Revisiones sistemáticas y meta-análisis: Herramientas para la práctica profesional. *Papeles del Psicólogo*, *31*, 7-17.

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Mª Blázquez-Rincón, D., López-Ibáñez, C., & López-Nicolás, R. (2021). Improving the Reporting Quality of Reliability Generalization Meta-analyses: The REGEMA Checklist. *Research Synthesis Methods, f*irst published: 20 March 2021 https://doi.org/10.1002/jrsm.1487

Santiago, M. I., Hervada, X., Naveira, G., Silva, L. C., Fariñas, H., Bacallao, J., & Mújica, O. J. (2010). El programa epidat: usos y perspectivas. *Revista Panamericana de Salud Pública*, *27*, 80-82.

Schwab, S. y Held, L. (2020). Science after Covid-19: faster, better, stronger? *Significance*, *4*, 8-9.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309-316.

Shamoo, A. E. & Resnik, D. B. (2015). *Responsible conduct of research* (3rd ed.). New York, NY: Oxford University Press.

Shaughnessy, J. J. (2009). Evaluating and understanding articles about treatment. *American Family Physician*, *79*, 668-670.

Skidmore, S. T. & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavioral Research*, *45*, 536-546. doi: 10.3758/s13428-012-0257-2

Smaldino, P.E. & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*, 160384, 1-17.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine*, *318*, 262-264. doi: 10.1056/nejm198801283180431

Tang, P. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. *Statistics Research Memorandum*, *2*, 126-149. Author

Thompson, B. (1996) AERA Editorial Policies regarding Statistical Significance Testing: Three Suggested Reforms. *Educational Researcher*, *25*, 26-30.

Tudela, J. & Aznar, J. (2013). ¿Publicar o morir? El fraude en la investigación y las publicaciones científicas. *Persona y Bioética*, *17*, 12-27.

Vacha-Haase, T., Nilsson, J.E., Reetz, D.R., Lance, T.S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology*, *10*, 413-425.

Viladrich, C., Angulo-Brunet, A., y Doval, E. (2017). Un viaje alrededor de alfa y omega para estimar la fiabilidad de consistencia interna. *Anales de Psicología*, *33*, 755-782. http://dx.doi.org/10.6018/analesps.33.3.268401

Volker, M. A. (2006). Reporting effect sizes in school psychology research. *Psychology in the Schools*, *43*, 653-672.

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

BLOQUE 3.

Análisis con JASP: Bayes y NHST

Capítulo 15. Análisis paralelos frecuentista-bayesianos con JASP

Jose D. Perezgonzalez
Nicholas Vincent
Universidad Massey (Nueva Zelanda)

Índice

Citar el capítulo como:

Perezgonzalez, J. D. y Vincent, N.. (2021). Análisis paralelos frecuentista-bayesianos con JASP. En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

Como dice el refrán, ‘Cada maestrillo tiene su librillo’, y en la Filosofía de la Estadística dos de esos librillos son la filosofía frecuencial y la filosofía bayesiana, dos perspectivas distintas a la hora de analizar datos y de generar inferencias a partir de los mismos. La filosofía frecuencial—o frecuentista—es la que prevalece en psicología (ej., las pruebas de comprobación de la significación de la hipótesis nula), si bien la filosofía bayesiana ha hecho grandes avances últimamente, sobre todo después de la famosa ‘crisis de la replicación en la investigación psicológica’ (ej., Open Science Collaboration, 2012; Klein y col., 2014). Esta última, la perspectiva bayesiana, ha sido tradicionalmente relegada a un segundo plano porque necesita estimar las probabilidades iniciales de las hipótesis con las que trabaja, lo que a menudo conlleva estimaciones subjetivas. Por su parte, la primera perspectiva se ha ganado un lugar preferencial en la investigación científica porque se asocia tanto con la objetividad investigativa como con la replicación empírica de los resultados. Aquí no vamos a discutir los beneficios y perjuicios de cada perspectiva. Lo que vamos a hacer es llevar a cabo un análisis paralelo sobre los mismos datos usando ambas, ayudados por el programa estadístico JASP, un programa que permite dicho análisis paralelo de forma clara y sencilla (Perezgonzalez y Frías-Navarro, 2018).

Este capítulo también va a ser algo distinto al resto de capítulos del libro, en el sentido de que va a ser más bien “pedante”, valga la palabra. Con el propósito de prevenir confusiones estadísticas frecuentes y, a la vez, favorecer inferencias más fidedignas a partir de los análisis estadísticos que haremos, vamos a ser puntillosos tanto con nuestros conceptos como con la interpretación de los resultados (sin intención de proponer dicha conducta como modelo para el lector). Por ejemplo, no vamos a usar el concepto de ‘pruebas de comprobación de la significación de la hipótesis nula’ por dos razones. La primera, porque es un concepto que se usa para referirse, al menos, a tres perspectivas frecuentistas distintas: la de Fisher, la de Neyman y Pearson, y a la mezcla de ambas en un potaje un tanto insalubre (ej., Gigerenzer, 2004; Perezgonzalez, 2015a). La segunda razón es porque lo que esas perspectivas ponen a prueba no es la significación de las hipótesis mismas sino la de los datos empíricos como ocurrencias potenciales bajo dichas hipótesis. Otros conceptos que evitaremos usar a lo largo de este capítulo son el nivel alfa (preferimos el concepto de ‘nivel de significación’ en su lugar), errores de Tipo II y poder estadístico (ambos son constructos de Neyman y Pearson, ej., 1928; por tanto,

dichos estadísticos no se pueden estimar bajo una perspectiva fisheriana), y pruebas de hipótesis (ya que tanto la perspectiva fisheriana como la jeffreysiana solo ponen a prueba los datos, no las hipótesis mismas) (léase también Perezgonzalez, 2014).

JASP

JASP es un programa estadístico de código abierto (open source), hoy en día en su versión 0.13.1, y bastante avanzado a la hora de llevar a cabo análisis estadísticos tanto frecuenciales como bayesianos. JASP se puede descargar gratuitamente a través de la página web https://jasp-stats.org. Esa página también proporciona acceso a varios tutoriales de uso.

JASP es fácil e intuitivo de usar, bastante flexible, y permite explorar varias opciones analíticas por medio de un menú de comandos visuales. JASP proporciona los resultados correspondientes de manera inmediata, en la misma pantalla donde se encuentra el menú de comandos. Por tanto, cambiar la selección de opciones automáticamente regenera los resultados a simple vista; y deseleccionar una opción previa elimina los resultados de la pantalla, lo cual ayuda a mantener una visualización pulcra y ordenada de los resultados retenidos. Copiar tablas y figuras de JASP a otro programa (ej., a un procesador de texto) también es una tarea sencilla, y las tablas y figuras son copiadas con el formato recomendado por la Asociación de Psicólogos Americanos (ej., APA, 2010), lo cual ahorra tiempo y preocupaciones a la hora de preparar un artículo de investigación. Sin embargo, siempre es conveniente retocarlas para obtener una presentación precisa.

Con JASP se pueden hacer análisis exploratorios de los datos, análisis inferenciales de corte frecuencial y análisis inferenciales de corte bayesiano. Los análisis exploratorios incluyen descriptivos, análisis de componentes principales y análisis factoriales exploratorios. Los análisis inferenciales de corte frecuencial están restringidos a la filosofía de análisis de datos de Fisher, basada en la comprobación de la significación estadística de los datos (ej., Fisher, 1954). Por su parte, los análisis inferenciales de corte bayesiano están restringidos a la filosofía de comparación de modelos de Jeffreys, también conocida como el factor bayesiano (ej., Jeffreys, 1961). Por lo tanto, cuando en este capítulo hablamos de análisis frecuenciales, en realidad nos estamos refiriendo solo a la comprobación de la probabilidad de los datos bajo una hipótesis nula de no efecto, según la filosofía de Fisher. Y cuando hablamos de

análisis bayesianos, en realidad nos estamos refiriendo a la probabilidad relativa de un modelo sobre el otro, según la filosofía de Jeffreys.

La ventaja del JASP es que permite llevar a cabo un análisis frecuencial – bayesiano en paralelo y de manera bastante simple (Perezgonzalez y Frías-Navarro, 2018). Dicho análisis paralelo permite realizar inferencias más precisas y sin malentendidos filosóficos. La perspectiva fisheriana permite analizar los datos con la ayuda de modelos probabilísticos teóricos, con la intención de someter dichos datos a una prueba severa y así prevenir errores inferenciales (ej., Mayo y Spanos, 2010). Sin embargo, si bien la perspectiva fisheriana puede ser adecuada a la hora de rechazar una hipótesis nula de no efecto, poco puede decir a la hora de interpretar un resultado probable bajo la hipótesis nula—algo que se podría solventar con la perspectiva de Neyman y Pearson (ej., 1928), excepto que JASP todavía no incluye esta última perspectiva.

La perspectiva jeffreysiana permite comparar los mismos datos empíricos usando dos modelos probabilísticos distintos, dando mayor peso relativo al modelo bajo el cual la probabilidad de los datos es mayor. Dichos modelos son más extremos que los usados en la perspectiva frecuencial. Por ejemplo, el modelo de la llamada hipótesis nula es, en realidad, un modelo de hipótesis 'cero', ya que toda la distribución de probabilidades queda colapsada en ese punto ('cero' tiene probabilidad '1'; el resto de valores tiene probabilidad '0'; Kruschke, 2011). El modelo alternativo, por su parte, se basa en otra distribución extrema, tal como la distribución *Cauchy* que, si bien es simétrica y está centrada en 'cero', es muy platykúrtica (parecida a una distribución *t* con un grado de libertad). Bajo dicha distribución *Cauchy*, el efecto 'cero' tiene una probabilidad menor, y los efectos distintos de 'cero' una probabilidad mayor, que bajo el modelo 'cero'. Ambos modelos no solo son extremos en comparación con ellos mismos sino también en comparación con los modelos más "normales" usados por los frecuentistas, lo que le da a las pruebas de factor bayesiano su valor particular. Esto es así porque al comparar la probabilidad de los datos bajo ambos modelos a la vez, es posible determinar qué modelo destaca, si el 'cero' o el alternativo (este índice es el factor bayesiano—Bayes Factor, BF, en inglés). Consecuentemente, se puede concluir tanto en contra como a favor de la hipótesis nula, algo que los fisherianos no pueden hacer.

La duda que surge inmediatamente es si no sería mucho más fácil y rápido el ignorar los análisis frecuenciales y usar solo los jeffreysianos. Eso es lo que han propuesto algunos autores últimamente (ej., Wagenmakers y col., 2017; Rouder y col., 2009). El problema con la propuesta es esa dependencia subjetiva que los bayesianos tienen a la hora de establecer la probabilidad previa de sus constructos que, en el caso jeffreysiano, vemos en la selección de modelos extremos. Es decir, se ha comprobado empíricamente que la mayoría de variables de investigación se distribuyen de una manera cercana a la curva normal frecuencial (un fenómeno que tiene su propia ley estadística, la ley fuerte de los grandes números). Por tanto, es difícil de justificar objetivamente el uso de modelos extremos tales como el 'todo-o-nada' del modelo 'cero' y el 'todo-vale' del modelo alternativo, aun cuando pudieran justificarse subjetivamente.

Sin embargo, hemos de recordar que las perspectivas frecuenciales y bayesianas buscan distintas maneras de aprender de los datos. La perspectiva frecuencial busca aprender a través del control de errores estadísticos (Mayo y Spanos, 2010) con el objetivo de rechazar hipótesis que puedan ser falsas. El uso de modelos probabilísticos más ajustados a la realidad es una estrategia adecuada, incluso cuando nos limite lo que podamos aprender: aprendemos algo acerca del colectivo, no acerca de la muestra, y puede que seamos capaces de rechazar una hipótesis como falsa, pero nunca de aceptarla como verdadera. No obstante, una vez que hayamos pasado ese obstáculo (es decir, una vez que hayamos completado el análisis frecuencial), la perspectiva bayesiana nos permite aprender más de nuestra muestra y más profundamente. El uso de modelos extremos no solo no es problemático, sino que dichos modelos complementan nuestro aprendizaje y, por tanto, son bienvenidos. Por ejemplo, un test frecuencial puede que descubra la baja probabilidad de un resultado bajo la hipótesis nula pero ni nos informa de dónde localizar la hipótesis alternativa, ni siquiera de si dicha hipótesis alternativa existe (¡puede que sea un error!). La perspectiva bayesiana nos puede ayudar, ya que proporciona una distribución posterior de posibles hipótesis alternativas, junto con un grado de certeza en la estimación de dichas alternativas. Por otra parte, un resultado frecuencial razonablemente probable bajo la hipótesis nula es incapaz de asegurar la veracidad de la misma. La perspectiva bayesiana ayuda, ya que nos informa del peso relativo del modelo 'cero' y, si la evidencia así lo justifica, de si podemos concluir

que la hipótesis de un efecto ‘cero’ es más probable y, consecuentemente, probablemente cierta.

Por tanto, es este análisis paralelo desde las perspectivas fisheriana y jeffreysiana el que vamos a desarrollar en este capítulo, ofreciendo un ejemplo práctico que sirva de tutorial didáctico. Los resultados aquí obtenidos han sido extraídos de una investigación empírica llevada a cabo por Vincent (2018) como parte de su tesis de maestría (master) en aviación en la Universidad Massey (Nueva Zelanda). El objetivo didáctico es el de demostrar la estrategia de análisis de datos paralelo discutida anteriormente, por medio del uso del programa estadístico JASP. Aunque escribiremos una pequeña introducción metodológica para centrar al lector, el enfoque es en el análisis de datos tal y como aparecería en la sección correspondiente de un manuscrito de investigación; por tanto, ignoramos aquí otras secciones tales como el marco teórico, la discusión de los datos y la conclusión. Además, incluso la sección de resultados la adaptaremos para ajustarla al objetivo didáctico de un tutorial, con notas a pie de página intercaladas allí donde hicieran falta.

Análisis de datos en paralelo con JASP

La estrategia llevada a cabo para el análisis de datos es, por tanto, el objetivo principal de este tutorial. Dicha estrategia implica tres tipos de análisis: análisis exploratorio de datos, pruebas de significación estadística y análisis de factor bayesiano.

Análisis exploratorio de datos, según Tukey

Este tipo de análisis persigue la intención inicial de Tukey (1977) de aprender de los datos lo más posible sin necesidad de depender de la estadística inferencial. JASP tiene algo de capacidad para permitir dicho análisis exploratorio. Nosotros nos limitaremos a unos pocos estadísticos, tales como el tamaño de los grupos (casos válidos), medidas de centralidad y varianza, e interpretación de los datos en el contexto de las escalas en las que se han medido.

También incluiremos como parte del análisis exploratorio estadísticos tanto descriptivos como inferenciales que ayuden a describir los grupos de manera más fidedigna, tales como el tamaño estandarizado del efecto e intervalos de credibilidad.

Pruebas de significación estadística, según Fisher

Los análisis inferenciales de corte frecuentista siguen el uso convencional de las pruebas de significación estadística de los datos de Fisher (ej., 1954), contrastando los datos empíricos con una hipótesis nula de no diferencia entre grupos (H_0). Dicha hipótesis funciona no tanto como hipótesis de investigación (es decir, demostrar que la diferencia entre grupos es 'cero' no es la razón principal de la prueba; Perezgonzalez, 2015a) sino que más bien funciona como diana para comprobar la validez de nuestro estudio. Como Mayo bien explica (ej., 1996), una prueba de significación estadística permite someter nuestros datos a un test severo, un test que capturará, por ejemplo, 95% de las diferencias entre grupos cercanas a 'cero' (bien sea mirando a una cola bien a las dos, dependiendo de la prueba)—dicho de otra manera, un test tan severo que pocos datos lo superarán si la diferencia entre grupos fuera realmente 'cero' o cercana a 'cero'.

Como es inherente en una prueba de significación estadística de Fisher, no se especifica una hipótesis estadística alternativa a la nula, aunque se sugiera *post hoc* cuando se obtiene un resultado significativo. Por lo tanto, a falta de hipótesis alternativa, no podemos controlar ni la probabilidad del error de Tipo II ni el poder estadístico de la prueba. Y a falta de control del poder estadístico y del error Tipo II, tampoco podemos aceptar la hipótesis nula en aquellos casos en el que el resultado no sea significativo.

No obstante, si bien no es posible controlar el poder estadístico, sí que es posible controlar la sensitividad de la prueba (sensitiveness). Estos análisis de sensitividad no necesitan ni hipótesis alternativas ni errores Tipo II, solo una decisión acerca del tamaño mínimo del efecto que es de interés (Perezgonzalez, 2016, 2017a). Si bien las razones para seleccionar dicho efecto mínimo deberían ser prácticas (ej., coste de implementación, precisión de las medidas, etc.), razones teóricas son también posibles, especialmente cuando las primeras son difíciles de establecer. Digamos, por ejemplo, que decidimos calcular la sensitividad de las pruebas estadísticas en base a un efecto estándar mínimo de tamaño medio. El subsecuente tamaño muestral aseguraría que efectos de tamaño medio (o mayores) aparecerán como significativos en las correspondientes pruebas estadísticas, mientras que efectos menores, que ya habíamos decidido no eran de importancia práctica o teórica,

aparecerán como estadísticamente no significativos. Por tanto, los análisis de sensitividad complementan las pruebas de Fisher, ya que no necesitan ni de una hipótesis alternativa que no se expresa ni de un error Tipo II que no puede estimarse.

Análisis de factor bayesiano, según Jeffreys

Una vez las pruebas frecuentistas han sido concluidas, seguiremos el análisis de datos con las pruebas de factor bayesiano (Jeffreys, 1961), que ayudarán a calcular tanto la probabilidad relativa de una hipótesis nula en caso de obtener un resultado no significativo, como la probabilidad relativa de una hipótesis alternativa en caso de obtener un resultado significativo.

La prueba de factor bayesiano simplemente contrasta los datos empíricos bajo dos modelos distintos a la par:

El modelo nulo base usado en JASP es un modelo de punto 'cero' (M_0), que establece que el efecto en la población es exactamente 'cero'.

El modelo alternativo (M_1) base en JASP es una distribución *Cauchy*, que es una distribución platikúrtica, si bien simétrica y centrada también en 'cero'[3].

El factor bayesiano es la proporción obtenida al dividir la probabilidad de los datos bajo un modelo por la probabilidad de los datos bajo el otro modelo. Y en esa proporción es donde radica el valor del factor bayesiano.

Por ejemplo, una vez que obtenemos un resultado estadísticamente significativo, el factor bayesiano nos permite responder la pregunta: ¿cuán probable es el modelo alternativo—de que el efecto en la población es real—comparado con el modelo nulo—de que el efecto no existe?

No solo eso, sino que también nos permite responder una pregunta que el análisis frecuentista no puede: una vez que obtenemos un resultado no significativo, ¿cuán probable es el valor 'cero'? Dicho de otra manera, el factor bayesiano nos permite decidir si aceptar la hipótesis nula, algo que las pruebas de significación de Fisher no pueden hacer.

[3] JASP 0.13.1 tiene otras distribuciones que se pueden usar como modelo alternativo, tales como la distribución normal y la distribución *t*.

Antecedentes metodológicos

El proyecto de investigación original tenía como objetivo el comprobar si el seguir un procedimiento de escaneo visual sistemático ayudaría a incrementar la atención situacional de los pilotos de aviación cuando éstos no se encontraban operando el avión de manera directa (es decir, cuando volaban con piloto automático).

La muestra fue una muestra conveniente: un total de 46 estudiantes de pilotaje de aviación general, recibiendo instrucción técnica en distintas escuelas de aviación en la isla norte de Nueva Zelanda. La participación de los pilotos en el estudio de investigación fue voluntaria. La muestra total fue calculada a partir de un análisis de sensitividad, usando como criterio el obtener un efecto mínimo de tamaño medio (Cohen d = 0.50) para la ocurrencia más interesante, que fue la de una prueba direccional en favor de la efectividad de la intervención experimental, medida como la diferencia de medias entre grupos independientes a un nivel de significación estadística convencional del 5%.

El diseño fue experimental con grupo control. Los participantes fueron divididos entre los dos grupos de manera aleatoria pero pareada, por medio del tiro de una moneda: para asegurar que ambos grupos fueran de tamaño similar, cuando un participante era aleatoriamente puesto en uno de los grupos, el siguiente era automáticamente puesto en el grupo contrario.

Ambos grupos se enfrentaron a condiciones experimentales similares: un vuelo simulado con el simulador de vuelo Microsoft X y jugado en un ordenador portátil. La simulación duró 10 minutos y estaba basada en un avión Cessna 172 Skyhawk, ya en el aire a 3.500 pies de altura, volando entre dos puntos específicos en el mapa, con dirección 000 en la brújula y en condiciones de vuelo visual pero con piloto automático. Un cambio de dirección y otro de altura ocurrían a intervalos fijos durante el vuelo, así como tres fallos instrumentales sin consecuencia inmediata. Como el vuelo se efectuaba de manera automática, los participantes no tenían ningún rol activo en el vuelo, y solo tenían que sentarse y mantener el escaneado típico de un vuelo de tipo visual. A ninguno de los participantes se les informó ni de que el vuelo experimentaría cambios de dirección y altitud, ni de que existía la posibilidad de que el avión experimentara fallos técnicos. A los participantes también se les proporcionó una tableta Apple iPad con información aeronáutica del aeropuerto de destino, y

papel para anotaciones. Se les señaló que podían tomar notas y que podían usar el reloj de cabina para mantener el tiempo, si hiciera menester.

A los participantes del grupo experimental también se les dio instrucciones de seguir un procedimiento de escaneo basado en tres puntos y que podía completarse en 10-15 segundos: i. escaneo del cielo; ii. comprobar la alineación de la brújula y el indicador de dirección [direccional giroscópico]; iii. comprobar la presión y la temperatura. También se les dijo que llevaran a cabo el escaneo cada dos minutos, aproximadamente (siguiendo el reloj de cabina), y se puso una copia del listado de puntos en un lugar visualmente accesible, para que no tuvieran que memorizarlo. El procedimiento de escaneo no fue preparado con la intención explícita de facilitar que los participantes percibieran los fallos técnicos de la simulación (es decir, el escaneo no apuntaba hacia esos fallos), si bien la expectativa general de la investigación era que el procedimiento incrementaría el nivel de atención general y lo mantendría alto durante toda la simulación y, como consecuencia, que los pilotos se encontrarían en una mejor posición de percibir esos fallos. Los participantes del grupo control no recibieron instrucciones de escaneo alguno. Por tanto, la única diferencia experimental entre ambos grupos fue el uso o no de dicho procedimiento.

Una vez que el vuelo había concluido, los participantes rellenaron un cuestionario de diez preguntas sobre la simulación, incluyendo sobre la información aeronáutica presentada en el iPad. Mientras rellenaban el cuestionario, ninguno de los participantes tuvo acceso ni al iPad ni al ordenador portátil, y solo podían responder bien de memoria bien en referencia a las anotaciones que hubieran tomado durante la simulación de vuelo.

Tutorial y resultados

1. Comprobación y limpieza de datos[4]

La comprobación de los datos antes de adentrarnos de lleno en el análisis estadístico propiamente dicho nos permite una evaluación rápida de la integridad de los mismos (Tabachnick y Fidell, 2001). Las funciones descriptivas de JASP vienen muy bien al caso. Para llevar a cabo esta comprobación, todas las variables deben analizarse a la vez, corregirse si es necesario (ej., si hay casos perdidos, casos extremos, etc.), y cualquier corrección anotarse. Si bien este paso ha de ser exhaustivo, no todos los resultados tienen

[4] Estos resultados no fueron incluidos en la investigación original.

que comunicarse: solo aquellos más informativos para el lector y que den una buena cuenta de la integridad final de la base de datos.
***JASP 0.13.1**: Abrimos JASP, cargamos la base de datos correspondiente, seleccionamos la lengüeta 'Descriptives' y elegimos la opción "Descriptive Statistics". Una vez que se abra la pantalla de comandos, seleccionamos todas las variables de interés, que moveremos a la casilla correspondiente (también es posible comprobar la integridad de los datos a nivel de subgrupo, moviendo la variable que identifica a dichos subgrupos a la casilla 'Split'). JASP crea una tabla básica de resultados descriptivos que incluye casos válidos, caso perdidos, y valores mínimos y máximos. La opción 'Frequency tables (nominal and ordinal variables)' permite comprobar la distribución de frecuencias, porcentajes y valores perdidos de las variables nominales y ordinales. En la opción 'Plots' elegimos "Distribution plots" y "Boxplots". En la opción 'Statistics' elegimos otros estadísticos de interés—"Median", "Skewness" y "Kurtosis"—cuyos resultados son añadidos automáticamente a la tabla anterior.*

Los resultados obtenidos tras la comprobación de la integridad de los datos pueden verse en la Tabla 1[5]. Los casos válidos (también los casos perdidos, que no se presentan en la tabla, pero que pueden calcularse por diferencia), así como los valores mínimos y máximos, no presentan resultados sospechosos. Las medias y las medianas tienen resultados parecidos, lo cual indica cierto grado de normalidad en lo que respecta a las medidas de centralidad. Las puntuaciones 'z' para la asimetría y la kurtosis (obtenidas al dividir sus estadísticos por el error estándar de los mismos, según Tabachnick y Fidell, 2001), no muestran discrepancias serias bajo la curva normal para las variables Static SA (Atención situacional estática) y Active SA (Atención situacional activa); sin embargo sí muestran una discrepancia extrema de la kurtosis de Timing SA (Atención situacional temporal), y una discrepancia extrema tanto de la kurtosis como de la asimetría de Continual SA (Atención situacional continua).

Static SA y Active SA aparecen como relativamente simétricas en sus distribuciones de frecuencias y diagramas de cajas y bigotes (figuras que no se presentan en los resultados). Static SA tiene varios casos extremos (dos casos extremos en cada uno de los valores '7', '2' y '1' de su escala de medida). Como dichos casos extremos aparecen a ambos lados de la distribución—lo que mantiene el equilibrio de la media—y no parecen afectar a la normalidad de la variable, hemos decidido dejar la variable tal como está, sin corregirla.

[5] La investigación original tenía muchas más variables. Aquí simplemente presentamos aquellas que vamos a usar en este tutorial.

Table 1 | Descriptive Statistics[6]

	Static SA	Active SA	Timing SA	Continual SA
Valid	46	46	33	46
Minimum	1.000	0.000	1.000	2.000
Maximum	7.000	2.000	3.000	4.000
Mean	4.304	0.957	2.606	3.558
Median	4.000	1.000	3.000	3.670
Std. Deviation	1.314	0.698	0.715	0.429
Z Skewness	-0.300	0.169	-3.804	-4.546
Z Kurtosis	0.118	-1.259	1.108	5.224

Timing SA tiene una asimetría destacada hacia sus valores bajos, y cinco participantes aparecen como casos extremos (aquellos con valores de '1.5' y '1'). Sin embargo, la naturaleza de la variable misma, que solo contaba el tiempo que se tardaba en percibir los fallos técnicos en la cabina, también implica que el no percibir fallos significaba contar dichos casos como perdidos (la variable tiene 13 casos perdidos). Por lo tanto, tanto la distribución asimétrica como la presencia de casos extremos no es sorprendente.

Las limitaciones de Timing SA empeoran cuando miramos a Continual SA, ya que es una variable creada al combinar Active SA y Timing SA, e incluye todas las ocurrencias, tanto percibidas como no (por tanto, cuenta los fallos no percibidos) así como el tiempo que se tardó en percibirlas (el anclaje '4' sirve para contar las ocurrencias no percibidas y, por tanto, sin tiempo). Continual SA tiene asimetría y kurtosis extremas, no casos perdidos, pero sí algunos casos extremos en los valores más pequeños de su escala ('2.5' y '2').

Como vamos a llevar a cabo un análisis frecuentista – bayesiano en paralelo, y dado que la normalidad de las variables no preocupa cuando se trata de la estadística bayesiana, hemos decidido no cambiar ninguna de las dos variables anteriores. En su lugar, llevaremos a cabo pruebas frecuentistas no paramétricas a la hora de analizar tanto Timing SA como Continual SA.

[6] Si bien la tabla fue copiada directamente de JASP, hemos eliminado información redundante, reorganizado la información presentada, y substituido los estadísticos de la asimetría y la kurtosis por sus puntuaciones 'z' correspondientes.

2. Atención Situacional Estática (Static SA)

Static SA fue evaluada por medio de siete preguntas, cada una de ellas acerca de algún aspecto aeronáutico que un piloto debería recordar durante el vuelo (ej., la altitud inicial del vuelo, las condiciones meteorológicas, e información acerca del aeropuerto de destino) pero que no formaban parte de las tareas de operación dentro de la cabina. La respuesta dada a cada pregunta fue evaluada como una dicotomía—es decir, como correcta o no—y todas ellas sumadas en un solo componente. Por tanto, la escala final de Static SA abarcaría las puntuaciones comprendidas entre un mínimo de '0' (si ninguna respuesta fuera correcta) y un máximo de '7' (si todas las respuestas lo fueran).

2.1. Static SA; análisis exploratorio de datos

Análisis descriptivos básicos adecuados para nuestro análisis exploratorio se pueden obtener a través de las opciones de análisis bayesianos. Nosotros estamos interesados en estadísticos a nivel de grupo, tales como los descriptivos y los intervalos de credibilidad (preferimos los intervalos de credibilidad bayesianos a los intervalos de confianza frecuentistas simplemente porque los primeros reflejan mejor una interpretación descriptiva de las distribuciones muestrales—es decir, una medida de centralidad y una distribución que incorpora el 95% de las estimaciones más probables dentro del intervalo. En cualquier caso, JASP solo proporciona intervalos de confianza para la diferencia entre grupos, lo que conllevaría calcular dichos intervalos a mano para los grupos en sí).

***JASP 0.13.1**: Seleccionamos la lengüeta 'T-Tests' y elegimos la opción "Bayesian Independent Samples T-Test". Un vez dentro de la pantalla de comandos, seleccionamos 'Static SA' como variable dependiente y 'Group'[7] como variable de grupo. En 'Additional Statistics' elegimos "Descriptives"—que genera los estadísticos descriptivos para cada grupo—y en 'Plots' elegimos "Descriptives plots" con "Credible interval" al 95%—que genera la gráfica correspondiente pero también añade los intervalos de credibilidad a la tabla de estadísticos descriptivos.*

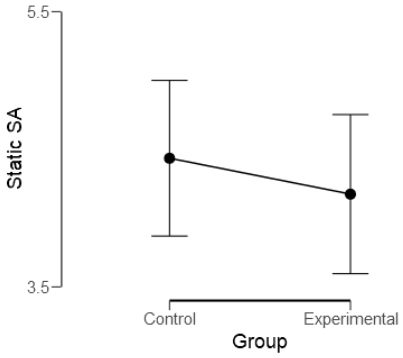
Los resultados obtenidos son similares para cada grupo, los cuales muestran poca diferencia tanto en centralidad como en variabilidad (Tabla 2). El grupo experimental tiene una media algo más baja (M = 4.17)—que también se ve en el

[7] Esta variable fue llamada 'Checklist or no' en la investigación original.

intervalo de credibilidad algo descentrado (95% BCI[8] [3.60, 4.75])—que el grupo control (M = 4.44; 95% BCI [3.87, 5.00]).

Table 2 | Static SA; Group Descriptives

				95% BCI	
Group	**N**	**Mean**	**SD**	**Lower**	**Upper**
Control	23	4.435	1.308	3.869	5.001
Experimental	23	4.174	1.337	3.596	4.752



En el contexto de la escala de medida usada, ambos grupos obtienen un desempeño moderadamente positivo, recordando, de media, cuatro ítemes del total de siete presentados (es decir, un 57% de ellos).

2.2. Static SA; prueba de significación estadística

Si bien las pruebas de significación estadística de Fisher contrastan los datos empíricos con las expectativas derivadas de las hipótesis nulas correspondientes, JASP simplemente enumera las hipótesis de investigación. En nuestro caso, la expectativa general es que la intervención no tendría ningún efecto en Static SA, lo que conlleva una prueba no direccional (de dos colas). JASP asigna los grupos de manera automática, así que hay que referirse a los estadísticos descriptivos a la hora de interpretar los resultados correctamente. En cuanto a estadísticos descriptivos e inferenciales toca, nos interesan los más informativos, que son los tamaños de efecto estandarizados midiendo la diferencia entre grupos, así como las pruebas estadísticas mismas.

***JASP 0.13.1**: Volvemos a la lengüeta 'T-Tests' y elegimos la opción "Independent Samples T-Test". Seleccionamos 'Static SA' como variable dependiente y 'Group' como variable de grupo (antes de llevar a cabo la prueba podríamos comprobar que los requisitos de normalidad e igualdad de varianza se mantienen; como ya hemos comprobado dichos requisitos anteriormente, simplemente pasamos por alto este paso). Seleccionamos "Student"*

[8] BCI es el acrónimo que usamos para referirnos al intervalo de credibilidad bayesiano (Bayesian credible interval), mientras que usamos FCI para referirnos al intervalo de confianza frecuentista (Frequentist confidence interval). El estudio original presentó FCIs en lugar de BCIs.

como prueba estadística, "Group 1 ≠ Group 2" como hipótesis de investigación[9] *y, en 'Additional Statistics', "Effect size" con "Confidence Interval" al 95%.*

Dado que Static SA evalúa información que un piloto debería recordar pero que no requería una atención constante en la cabina, no esperábamos que esta variable fuera afectada de manera alguna por la intervención experimental. Por lo tanto, la Tabla 3 presenta estadísticos no direccionales, calculados a dos colas.

Los resultados muestran que la diferencia media del efecto estandarizado entre grupos fue una Cohen *d* = 0.20, una diferencia que no es rara en psicología y que, de hecho, sería considerada pequeña pero no despreciable. Sin embargo, ya que habíamos decidido con anterioridad que solo estábamos interesados en efectos de tamaño medio (o mayores), ese efecto lo consideramos trivial en nuestro contexto experimental.

Table 3 | Static SA; Independent Samples T-Test

				95% FCI for Cohen's d	
t	**df**	**p**	**Cohen's d**	**Lower**	**Upper**
0.669	44	0.507	0.197	-0.383	0.776

El intervalo de confianza[10] para la diferencia del efecto estandarizado cubre el rango de valores entre -0.38 y 0.78[11]; por lo tanto, nuestros datos tienen una probabilidad razonable bajo cualquiera de las hipótesis en ese intervalo, las cuales no podemos rechazar, incluyendo bajo la hipótesis nula (o valor 'cero'), una

[9] La hipótesis nula sería, por tanto, H_0: el grupo Experimental no tendrá un desempeño significativamente diferente al del grupo Control, dada una medida adecuada de la significación (o Group 1 = Group 2).

[10] Los intervalos de confianza frecuentistas (FCI) son relevantes aquí, ya que las pruebas de significación también son frecuentistas. Siendo no más la segunda cara de la misma moneda inferencial (Perezgonzalez, 2015b), un FCI nos indica aquellos efectos poblacionales que pueden rechazarse con un margen de error del 5%, es decir aquellos que no contienen la media muestral dentro de su intervalo respectivo (Perezgonzalez, 2017b). Por lo tanto, un FCI no debe interpretarse de la misma manera que interpretamos un BCI: ni como una distribución de efectos potenciales centrados en la media muestral, ni como efectos poblacionales cuya probabilidad disminuye cuanto más nos movamos hacia las colas de dicha distribución (es decir, no podemos decir que hay un 95% de probabilidad de que el efecto poblacional se encuentre en el intervalo).

[11] Este rango fue erróneamente descrito como [-0.53, 1.00] en el estudio original.

probabilidad bajo esta última que podemos precisar por medio de la prueba *t* (p = 0.51[12]).

De estos resultados aprendemos que los datos observados tienen una probabilidad razonable bajo la hipótesis nula de no efecto (de hecho, es un resultado que se espera que ocurra cerca de un 50% de las veces cuando no hay efecto real[13]).

Por lo tanto, no podemos rechazar la hipótesis substantiva nula de que la intervención no afecta a la atención situacional estática[14]—ni tampoco podemos afirmar dicha falta de efecto (es decir, tampoco podemos decir que la hipótesis nula se mantiene o que es real[15]).

2.3. Static SA; análisis de factor bayesiano

El factor bayesiano de Jeffreys pone a prueba los datos bajo dos modelos distintos, modelos que JASP también presenta como hipótesis de investigación y a los que etiqueta como H_0 y H_1[16]. En cuanto a la estadística bayesiana se refiere, estamos interesados principalmente en el factor bayesiano, que nos informa del peso de la evidencia empírica a favor de un modelo o del otro.

***JASP 0.13.1**: Regresamos a la lengüeta 'T-Tests' y elegimos la opción "Bayesian Independent Samples T-Test". Luego seleccionamos 'Static SA' como variable dependiente y 'Group' como variable de grupo. Seguidamente seleccionamos "Group 1 ≠ Group 2" como hipótesis de investigación. Como no hay una manera clara de elegir el 'Bayes Factor' apropiado, simplemente seleccionamos "BF_{10}" al azar, comprobamos los resultados, vemos que el factor bayesiano es menor de '0' (BF_{10} = 0.351), y seleccionamos "BF_{01}" en su lugar (ya que ambos estadísticos son intercambiables—el subíndice solo indica qué modelo se usa en el numerador y cuál en el denominador—el procedimiento convencional es el de usar el factor*

[12] Que un efecto pequeño no alcance significación estadística no es sorprendente, ya que la prueba misma no tiene una sensitividad adecuada para efectos de ese tamaño, sino para efectos de tamaño medio (d = 0.5). De hecho, podríamos decir que la prueba de significación misma es redundante una vez que conocemos el efecto estandarizado obtenido.

[13] Ésta es la única información que el valor *p* provee: la probabilidad de que resultados similares a los obtenidos ocurran bajo la H_0 estadística de no efecto (Perezgonzalez, 2015c).

[14] El rechazar la H_0 substantiva depende no del valor *p*, por sí mismo, sino del *Modus Tollens* definido por el nivel de significación elegido. El *Modus Tollens* es el silogismo que permite rechazar la hipótesis cada vez que logramos contradecirla: si H_0 es verdadera, los datos serán no significativos (al nivel elegido). Por tanto, resultados no significativos, tales como los obtenidos aquí, no contradicen la hipótesis (más bien la afirman), lo que conlleva a que el *Modus Tollens* no pueda completarse y, por tanto, a que no pueda concluirse nada de manera realmente lógica (Perezgonzalez, 2017c).

[15] Dado que la H_0 funciona más como diana para comprobar nuestros datos que como hipótesis real, no existe un silogismo *Modus Ponens* con el que poder afirmarla. Por lo tanto, resultados no significativos como los aquí obtenidos tampoco pueden usarse para completar este otro silogismo. En todo caso, un test severo en contra de la hipótesis nula requiere poder estadístico (Mayo, 1996), el cual no podemos controlar cuando usamos las pruebas de Fisher.

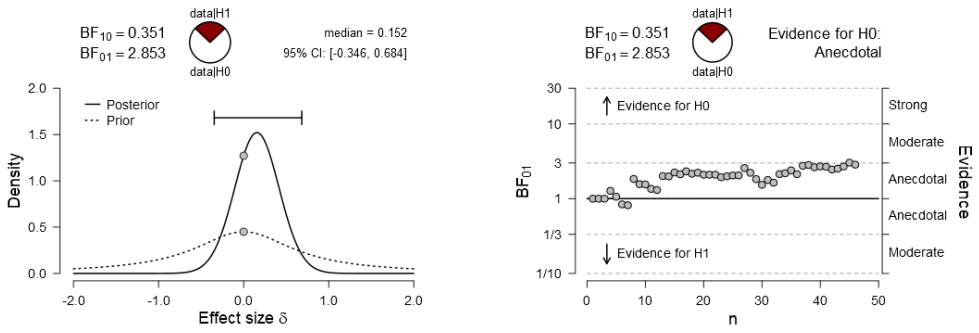
[16] En realidad, el factor bayesiano pone a prueba los datos bajo dos modelos, no bajo dos hipótesis. Para prevenir confusiones con un análisis inferencial bayesiano proprio, en el texto no haremos referencia a hipótesis (H_0 o H_1) sino a modelos (M_0 y M_1).

bayesiano más fácil de entender, que es aquel que resulta en un valor mayor de '1', e interpretar los resultados de manera consistente con dicha selección). También elegimos dos 'Plots': "Prior and posterior" con "Additional info", y "Sequential analysis".

El análisis de factor bayesiano muestra que el modelo 'cero' (M_0[17]) es casi tres veces más probable, relativamente hablando, que el modelo alternativo (BF_{01} = 2.85; Tabla 4).

La distribución posterior muestra que, aun cuando el efecto aparece algo desviado del centro de la distribución, 'cero' todavía alcanza una gran probabilidad de ocurrencia en dicha distribución posterior.

Table 4 | Static SA; Bayesian Independent Samples T-Test

BF_{01}	
2.853	

Con estos resultados aprendemos que la evidencia proporciona un peso anecdótico (tirando a moderado) a favor de la conclusión de que la intervención experimental no tuvo efecto alguno en Static SA, tal y como esperábamos[18].

[17] También reservamos la etiqueta 'nula' para referirnos a la hipótesis nula de Fisher (H_0) y usamos 'cero' para referirnos al modelo nulo bayesiano (M_0), ya que lo que se modela con este último es una distribución de punto 'cero'.

[18] La inferencia bayesiana permite concluir a favor del modelo 'cero' si la evidencia así lo sugiere. Por ejemplo, el factor bayesiano aquí encontrado le da al modelo 'cero' un mayor peso proporcional que al alternativo, y esa proporción se puede convertir en una probabilidad con la fórmula P = BF/(1+BF). Por tanto, el modelo 'cero' tiene una probabilidad del 74% de ser el modelo correcto (comparado con una probabilidad, por diferencia, del 26% del modelo alternativo).

3. Atención Situacional Activa (Active SA)

Active SA evaluaba los resultados de acciones que requerían observación frecuente, si bien no necesariamente de forma constante. Este nivel de supervisión fue puesto a prueba a través de tres fallos en la consola de vuelo, sin consecuencias inmediatas, y fue evaluado por medio de tres preguntas en el cuestionario, cada una en referencia a si el piloto podía identificar correctamente cada uno de los fallos. Cada respuesta fue evaluada como una dicotomía—es decir, como correcta o no—y todas ellas fueron sumadas en un solo componente. Por tanto, la escala final de Active SA abarcaría las puntuaciones comprendidas entre un mínimo de '0' (si ningún fallo fuera identificado correctamente) y un máximo de '3' (si todos los fallos así lo fueran).

3.1. Active SA; análisis exploratorio de datos

Un 95% BCI cubre el 95% del total de estimaciones plausibles del parámetro poblacional que se está infiriendo. Por lo tanto, podemos concluir que, con un grado de certeza o de confianza del 95%, el parámetro poblacional se encuentra dentro del intervalo calculado. Dicho de otra manera, podemos decir que hay una probabilidad del 95% de que el parámetro ahí se encuentre.

JASP 0.13.1*: Regresamos a la lengüeta 'T-Tests' y elegimos la opción "Bayesian Independent Samples T-Test". Luego seleccionamos 'Active SA' como variable dependiente y 'Group' como variable de grupo, además de los descriptivos seleccionados anteriormente: "Descriptives", "Descriptives plots" y 95% "Credible interval".*

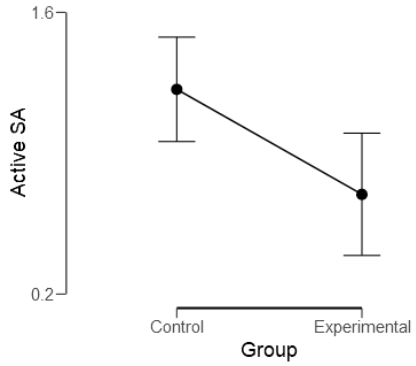
Los resultados obtenidos muestran que el grupo control tiene un mejor desempeño que el grupo experimental, obteniendo una media más alta en el componente (M = 1.22; 95% BCI [0.96, 1.48]) que el grupo experimental (M = 0.70; 95% BCI [0.39, 1.00]). La diferencia en desempeño es más llamativa en la figura (incluida en la Tabla 5), en la que vemos que el BCI del grupo control se sitúa claramente más alto en la escala que el del grupo experimental.

No obstante, cuando tomamos en cuenta el contexto de la escala de medida misma, vemos que ambos grupos tienen un desempeño realmente pobre, ya que perciben, como media, solo uno de los tres fallos (y consultando la Tabla 1 vemos que el máximo número de fallos percibidos fue de dos fallos, mientras que en ambos grupos también hubo pilotos que no percibieron fallo alguno). El grupo control

también obtuvo un desempeño mayor que el grupo experimental, algo que va en contra de las expectativas iniciales para con esta variable.

Table 5 | Active SA; Group Descriptives

				95% BCI	
Group	**N**	**Mean**	**SD**	**Lower**	**Upper**
Control	23	1.217	0.600	0.958	1.477
Experimental	23	0.696	0.703	0.392	1.000



3.2. Active SA; prueba de significación estadística

La expectativa inicial era que la intervención tendría un efecto positivo sobre Active SA, lo que implica un test direccional (de una cola). Dado que JASP asigna los grupos de manera automática, hemos de seleccionar cuál de las dos hipótesis direccionales que provee es la adecuada. Por tanto, antes de lanzarnos de cabeza a interpretar los resultados, es necesario comprobar que la hipótesis direccional seleccionada es la correcta (JASP provee una nota a pie de tabla que viene bien para ello[19]).

***JASP 0.13.1**: Volvemos a la lengüeta 'T-Tests' y elegimos la opción "Independent Samples T-Test". Luego seleccionamos 'Active SA' como variable dependiente y 'Group' como variable de grupo. También seleccionamos la prueba "Student". Para determinar cuál de las dos hipótesis direccionales es la apropiada, elegimos al azar "Group 1 > Group 2", leemos la nota a pie de tabla en los resultados, nos percatamos de que no es la hipótesis correcta, y cambiamos a "Group 1 < Group 2"[20]. Terminamos seleccionando los mismos 'Additional Statistics' que usamos con anterioridad: "Effect size" con "Confidence Interval" al 95%.*

Dado que Active SA medía ítemes ligados a operaciones activas en cabina, se esperaba que dichos ítemes fueran percibidos más fácilmente cuando el nivel de

[19] Esto conlleva el riesgo de que un investigador, una vez observados los resultados de la prueba, cambie su hipótesis por aquella que provea los resultados más deseados (ej., aquella con resultados significativos).

[20] La hipótesis nula sería, por tanto, H_0: el grupo Experimental no mostrará un desempeño significativamente mejor al del grupo Control, dada una medida adecuada de la significación (o Group 1 ≥ Group 2).

escaneo visual en cabina se incrementara. Por tanto, la tabla 6 muestra resultados direccionales (de una cola).

Los resultados muestran que la diferencia media del efecto entre grupos es una Cohen *d* = 0.80[21], típicamente considerada en psicología un efecto de tamaño grande. El intervalo de confianza, de una cola, cubre el rango de valores desde el infinito negativo hasta *d* = 1.30[22] como límite superior, por tanto conteniendo el valor 'cero' de la hipótesis nula dentro del intervalo. La probabilidad razonable de los resultados bajo la hipótesis nula también se observa en la prueba de significación (p = 0.995).

Table 6 | Active SA; Independent Samples T-Test

				95% FCI for Cohen's d	
t	**df**	**p**	**Cohen's d**	**Lower**	**Upper**
2.708	44	0.995	0.799	-∞	1.299

Por tanto, aprendemos que los resultados son bastante probables bajo la hipótesis nula, la cual, dada la naturaleza direccional de la inferencia, no podemos rechazar. Tampoco podemos concluir a favor de dicha hipótesis nula y, por tanto, no podemos sugerir la existencia de un efecto negativo del procedimiento de escaneo en la atención activa, como parecen indicar los datos descriptivos (aun cuando dicha posibilidad de efecto negativo aparece como parte de la hipótesis nula).

3.3. Active SA; análisis de factor bayesiano

Como dijimos anteriormente, el factor bayesiano de Jeffreys pone a prueba la probabilidad de los datos bajo dos modelos distintos, no la probabilidad de las hipótesis mismas. De hecho, el factor bayesiano asume que dichas hipótesis tienen la misma probabilidad previa, del 50% cada una, lo cual reduce su proporción a '1' [$p(H_0) / p(H_1) = 1$) y así se elimina su influencia en la fórmula bayesiana. Esto es una forma sutil de esquivar el problema de la probabilidad de las hipótesis previas, ya que 'muerto el perro se acabó la rabia'. Por tanto, en el factor bayesiano, la probabilidad posterior depende exclusivamente de la evidencia que proviene de los datos empíricos.

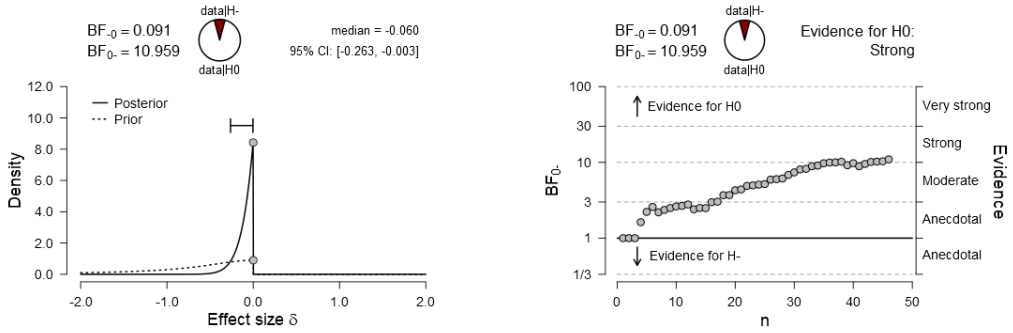
[21] La *d* de Cohen parece un tanto contra-intuitiva. Pero vale recordar que el estadístico señala una diferencia entre grupos pero a favor del grupo control (que no es la diferencia direccional esperada).

[22] Estos resultados fueron erróneamente descritos como [-∞, 0.10] en el estudio original.

***JASP 0.13.1**: Volvemos a la lengüeta 'T-Tests' y elegimos la opción "Bayesian Independent Samples T-Test". Luego seleccionamos 'Active SA' como variable dependiente y 'Group' como variable de grupo. También seleccionamos "Group 1 < Group 2" como hipótesis de investigación y el 'Bayes Factor' correspondiente—que resulta ser "BF_{01}"—así como las gráficas usadas anteriormente: "Prior and posterior" con "Additional info", y "Sequential analysis".*

El factor bayesiano muestra que el modelo 'cero' (M_0) tiene una probabilidad relativa casi once veces mayor que el modelo alternativo (BF_{01} = 10.96; Tabla 7). De hecho, la distribución posterior (de una cola) muestra que el efecto muestral es prácticamente 'cero' (mediana = -0.06; 95% BCI [-0.263, -0.003]). Por tanto, la evidencia empírica a favor del modelo 'cero'—que sostiene que el procedimiento de vigilancia no tendría ningún efecto positivo sobre Active SA (sin por ello descartar un efecto negativo)—es fuerte.

Table 7 | Active SA; Bayesian Independent Samples T-Test

BF_{01}	
10.959	

3.4. Active SA; análisis de dos colas

Esta sección del tutorial muestra análisis de datos no planeados, motivados por la obtención de resultados no esperados. Un aspecto positivo de JASP (al menos en esta versión) es que permite un acceso directo a cualquier pantalla de comandos previos a través de los resultados que generaron, lo cual ahorra tiempo y esfuerzo: solo hay que activar la tabla correspondiente. No obstante, hay que percatarse de que cualquier cambio en la pantalla de comandos actualiza los resultados de la tabla de manera automática (es decir, no genera nuevas tablas).

***JASP 0.13.1**: En 'Results', activamos la tabla "Independent Samples T-Test"—que activa, a*

su vez, la pantalla de comandos del t-test frecuentista de una cola—y elegimos la hipótesis de investigación no direccional "Group 1 ≠ Group 2". La tabla de resultados se actualiza y muestra los resultados de dos colas de la prueba. Luego activamos la tabla "Bayesian Independent Samples T-Test"—que activa la pantalla de comandos del test bayesiano de una cola—y también seleccionamos la hipótesis no direccional "Group 1 ≠ Group 2". Esto actualiza tanto la tabla de resultados como los gráficos. Comprobamos si el 'Bayes Factor' usado anteriormente es todavía adecuado—que, en este caso, no lo es—y seleccionamos "BF_{10}" en su lugar.

Antes de seguir analizando las variables restantes, consideramos relevante el explorar más detenidamente esa discordancia entre las expectativas iniciales y los resultados empíricos obtenidos, especialmente en vista de que los últimos parecen ir en la dirección opuesta a la esperada. Por tanto, aquí re-analizamos los datos usando pruebas a dos colas[23].

La tabla 8 presenta los resultados del test de significación. Vemos que el intervalo de confianza cubre los efectos estandarizados desde el 0.19 hasta el 1.40, indicando que efectos de tamaño pequeño a muy grande son estimaciones que caen dentro del intervalo (de hecho la mitad de las estimaciones son mayores que Cohen *d* = 0.80[24]). El intervalo de confianza no contiene el valor 'cero', hipótesis que podemos rechazar con un margen de error bajo, también afirmado por la prueba de Fisher (p = 0.01).

Table 8 | Active SA; Independent Samples T-Test

				95% FCI for Cohen's d	
t	**df**	**p**	**Cohen's d**	**Lower**	**Upper**
2.708	44	0.010	0.799	0.193	1.396

Note. Two-tailed analyses

Como fue mostrado en la tabla 5, la dirección del efecto favorece al grupo control. El tamaño del efecto, más bien grande, el intervalo de confianza, que no incluye efectos tan pequeños que los calificaríamos de despreciables en psicología, además de la baja probabilidad de los resultados, llevarían a rechazar la hipótesis nula de no

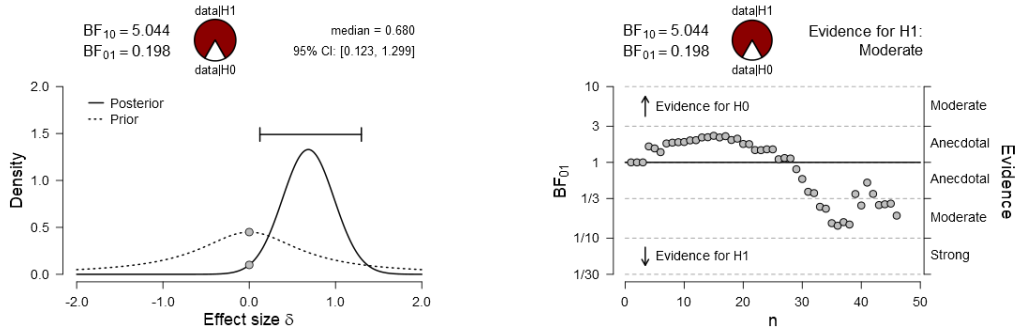
[23] Dado el carácter *post hoc* de esta exploración de datos, preferimos usar pruebas a dos colas, que son más conservadoras.

[24] Este estadístico fue erróneamente descrito como *d* = 0.65 en el estudio original.

efecto[25]. En suma, los resultados sugieren que el procedimiento de escaneo ha funcionado como un obstáculo tanto en cuanto a Active SA se refiere.

La tabla 9 presenta los resultados bayesianos[26]. El análisis de factor bayesiano muestra que el modelo alternativo (M_1) es cinco veces más probable que el modelo 'cero' (BF_{10} = 5.04[27]). De hecho, la distribución posterior de dos colas muestra la mediana del efecto centrada en Cohen *d* = 0.68, con un intervalo de credibilidad (95% BCI [0.12, 1.30]) que no contiene 'cero' y que, por lo tanto, le da una baja probabilidad al mismo. La evidencia empírica a favor del modelo alternativo es moderada, lo que parece soportar la conclusión de que el efecto de la intervención en Active SA fue más bien pernicioso.

Table 9 | Active SA; Bayesian Independent Samples T-Test

BF_{10}	
5.044	

Note. Two-tailed analyses

[25] Como comentábamos anteriormente, el *Modus Tollens* es el silogismo que permite rechazar la hipótesis cada vez que logramos contradecirla: si H_0 es verdadera, los datos serán no significativos (al nivel elegido). Los resultados significativos aquí obtenidos contradicen el consecuente del silogismo, lo que conlleva la contradicción del antecedente (la hipótesis) de una manera lógica válida. Por tanto, podemos concluir que la H_0 substantiva no es correcta, y la podemos rechazar (Perezgonzalez, 2017c).

[26] En el estudio original no se llevaron a cabo pruebas bayesianas en esta sección.

[27] Este factor bayesiano le da al modelo alternativo una probabilidad del 83% de ser el modelo correcto (comparado con una probabilidad del 17%, por diferencia, de que el modelo 'cero' sea el correcto; P = BF/[1+BF]). Dicho de otra manera, la probabilidad posterior de la hipótesis alternativa se ha incrementado un 33% y la de la hipótesis 'cero' ha disminuido un 33%, comparadas con la suposición inicial de que ambas tenían una probabilidad del 50%.

4. Atención Situacional Temporal (Timing SA)

Timing SA evalúa el tiempo que le llevó a los pilotos el percibir cualquiera de los tres fallos una vez ocurridos, en minutos (durante la simulación, los pilotos podían tomar notas, incluyendo del tiempo en referencia al reloj de cabina). Un fallo identificado correctamente en el intervalo de un minuto recibió una puntuación de '1'; en el intervalo de dos minutos, de '2'; en un intervalo mayor, de '3'; y se le daba un valor perdido si no daba un intervalo de tiempo o no identificaba el fallo correctamente, independientemente de cualquier tiempo dado[28]. Las puntuaciones de los tres ítemes fueron promediadas y tomadas como parte de un solo componente. Por tanto, la escala final de Timing SA abarcaba las puntuaciones comprendidas entre un mínimo de '1' (en el caso de que los tres fallos fueran identificados en el intervalo de un minuto cada uno) y un máximo de '3' (en el caso de que los tres fallos fueran identificados con un retraso mayor de dos minutos desde su ocurrencia).

4.1. Timing SA; análisis exploratorio de datos

Un BCI descriptivo es calculado asumiendo el principio de equi-probabilidad de las hipótesis; su centralidad es la de la muestra y cubre, por ejemplo, el 95% de la distribución posterior más cercana a la media. Por tanto, dicha distribución posterior se puede interpretar de una manera directa, como el 95% de estimaciones que son creíbles para el parámetro poblacional, cada una con una probabilidad decreciente cuanto más nos alejamos de la media. Aun así, estamos más interesados en el intervalo en sí, el BCI, que en estimaciones particulares dentro de dicho intervalo.

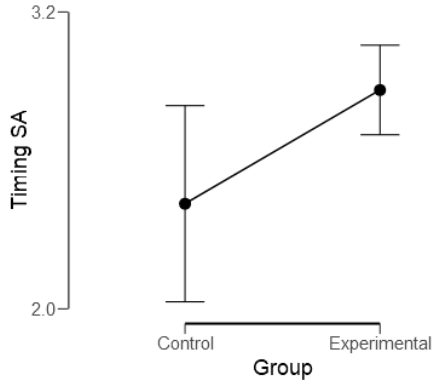
JASP 0.13.1*: Seleccionamos la lengüeta 'T-Tests' y elegimos la opción "Bayesian Independent Samples T-Test". Luego seleccionamos 'Timing SA' como variable dependiente y 'Group' como variable de grupo. También seleccionamos los mismos descriptivos que hicimos anteriormente: "Descriptives", "Descriptives plots" y 95% "Credible interval".*

Los resultados de Timing SA (tabla 10) muestran que el grupo experimental obtuvo un peor desempeño que el grupo control, con un retraso medio más largo (M = 2.89; 95% BCI [2.70, 3.06]) que este último (M = 2.43; 95% BCI [2.03, 2.82]).

[28] Esta valoración aparece invertida en el estudio original, donde '1' cuenta como el intervalo de tres minutos o más, y '3' como el intervalo de un minuto. Esta falta de consistencia cognitiva, aun cuando se describió correctamente en el estudio, produjo errores de interpretación de los datos y motivó el llevar a cabo pruebas estadísticas en la dirección equivocada. Por lo tanto, los resultados aquí discutidos difieren sensiblemente de los descritos en el estudio original.

Table 10 | Timing SA; Group Descriptives

				95% BCI	
Group	**N**	**Mean**	**SD**	**Lower**	**Upper**
Control	20	2.425	0.847	2.028	2.822
Experimental	13	2.885	0.300	2.704	3.066



Sin embargo, teniendo en cuenta el contexto temporal de la escala de medida, vemos que ambos grupos tuvieron un desempeño más bien pobre, percibiendo cualquiera de los fallos con un retraso medio en el intervalo de los dos minutos. Como muestra la figura, el desempeño del grupo control es sensiblemente mejor, pero también muestra una variabilidad mucho más larga, y ese desempeño medio va en contra de lo esperado.

También de interés es el número de pilotos que percibieron fallos, donde observamos que 20 de los pilotos en el grupo control observaron al menos uno de los fallos, pero solo lo hicieron 13 de los pilotos en el grupo experimental.

4.2. Timing SA; prueba de significación estadística

Timing SA es una variable que no cumple los requisitos básicos de un test paramétrico, sobre todo en lo que respecta a la distribución normal de la variable. JASP provee una prueba no paramétrica que podemos utilizar como alternativa: la prueba U de Mann-Whitney.

***JASP 0.13.1**: Volvemos a la lengüeta 'T-Tests' y elegimos la opción "Independent Samples T-Test". Luego seleccionamos 'Timing SA' como variable dependiente y 'Group' como variable de grupo (bajo 'Assumption Checks' podríamos seleccionar "Normality" y "Equality of variances" para comprobar si los datos cumplen los requisitos paramétricos de la prueba t. Como ya lo comprobamos anteriormente, ignoramos esas opciones aquí). Seleccionamos tanto "Mann-Whitney" como "Student" bajo 'Tests' (en el contexto global de esta investigación, la prueba U de Mann-Whitney es la que nos dirá si los resultados son significativos o no, pero la prueba t de Student provee estadísticos de tamaño del efecto que son más fáciles de comparar con las otras variables ya analizadas). A la hora de decidir cuál es la hipótesis de investigación adecuada, elegimos "Group 1 > Group 2" al azar, que, una vez que leemos la nota a pie de tabla en los resultados, resulta ser la opción correcta.*

También seleccionamos 'Additional Statistics': "Effect size" con "Confidence Interval" al 95%.

Ya que Timing SA está relacionada con Active SA, también esperábamos que fuera afectada de manera positiva por la intervención experimental. Por tanto, la tabla 11 muestra resultados direccionales (de una cola). La prueba principal para interpretar la significación de los resultados es la prueba *U* no paramétrica de Mann-Withney, si bien también damos los resultados de la prueba *t* de Student para facilitar la comparación de los resultados con las variables restantes de este estudio.

Los resultados tienen una alta probabilidad bajo la hipótesis nula (U = 98.50, p = 0.934). El tamaño del efecto es una correlación de rango biserial media (r = -0.24). Quizás más comparable es el efecto Cohen *d*, también medio (d = -0.67, 95% FCI [-1.26, ∞]).

Table 11 | Timing SA; Independent Samples T-Test

					95% FCI for Effect Size	
Test	**Statistic**	**df**	**p**	**Effect Size**	**Lower**	**Upper**
Mann-Whitney	98.500		0.934[a]	-0.242	-0.530	∞
Student	-1.873	31	0.965[a]	-0.667	-1.264	∞

Notes. For the Student's t-test, effect size is given by Cohen's *d;* for the Mann-Whitney's test, effect size is given by the rank biserial correlation.

[a] Levene's test is significant ($p < .05$), suggesting a violation of the equal variance assumption

Aprendemos, por tanto, que los resultados son altamente probables y que, dada la naturaleza direccional de la inferencia, no podemos rechazar la hipótesis nula de que o bien no hay efecto o bien el efecto es negativo. Aun así, tampoco podemos concluir a favor del efecto negativo mostrado (es decir, tampoco podemos aceptar la hipótesis nula).

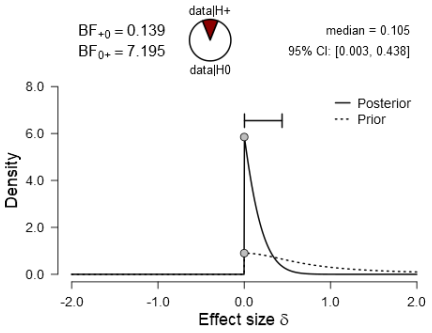
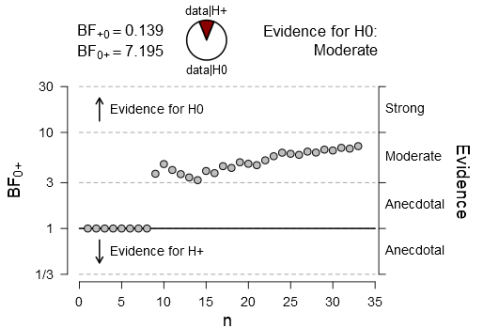
4.3. Timing SA; análisis de factor bayesiano

La inferencia bayesiana—incluyendo el factor bayes de Jeffreys—tiene como punto de partida la muestra observada, que se toma como una realización ya cumplida y no como una muestra tomada al azar de entre todas las muestras posibles. Como la muestra es una realización ya cumplida, es lo que es, y así se tiene que tomar. Por tanto, no hay necesidad de comprobar si cumple o no con requisitos ideales para la prueba que se use para analizarla. Por tanto, los mismos análisis bayesianos valen independientemente de la normalidad de la muestra o de la homogeneidad de su varianza (en el caso bayesiano, y completando el refrán, 'es la montaña la que ha de ir a Mahoma').

JASP 0.13.1*: Volvemos a la lengüeta 'T-Tests' y elegimos la opción "Bayesian Independent Samples T-Test". Luego seleccionamos 'Timing SA' como variable dependiente y 'Group' como variable de grupo. También seleccionamos "Group 1 > Group 2" como hipótesis de investigación y el 'Bayes Factor' apropiado—que es "BF_{01}"—así como los mismos gráficos de antes: "Prior and posterior" con "Additional info", y "Sequential analysis".*

El análisis de factor bayesiano muestra que el modelo 'cero'(M_0) es siete veces más probable que el modelo alternativo (BF_{01} = 7.20; tabla 12). El efecto en la distribución posterior de una cola es pequeño (mediana = 0.105; 95% BCI [0.003, 0.438]) y la evidencia es moderadamente fuerte a favor del modelo 'cero'—es decir, de que no existe efecto alguno (sin por ello descartar la posibilidad de un efecto negativo).

Table 12 | Timing SA; Bayesian Independent Samples T-Test

BF_{01}		
7.195		

4.4. Timing SA; análisis de dos colas

Como apuntábamos anteriormente, JASP es bastante flexible a la hora de llevar a cabo análisis de datos improvisados. Sin embargo, eso conlleva el riesgo de que se abuse bien en busca de la significación estadística bien en busca de un soporte positivo bayesiano. Cautela y responsabilidad debida han de ir mano a mano con dicha flexibilidad.

JASP 0.13.1*: Activamos la tabla "Independent Samples T-Test" en los resultados direccionales obtenidos anteriormente y elegimos la hipótesis de investigación no direccional "Group 1 ≠ Group 2". Luego hacemos lo mismo con la tabla "Bayesian Independent Samples T-Test", comprobamos si el 'Bayes Factor' anterior es todavía el adecuado—que no lo es—y seleccionamos "BF_{10}" en su lugar.*

Como hicimos anteriormente, nos adentramos un poco más en los datos con la idea de aprender qué pasa con esa falta de ajuste entre las expectativas iniciales y los resultados empíricos. La tabla 13 resume los resultados de la prueba de significación a dos colas. Observamos que los resultados no son improbables bajo la hipótesis nula (U = 98.50, p = 0.145) y que no tenemos un argumento razonable para rechazar tal hipótesis de no existencia de efecto negativo[29] (de hecho, si concluyéramos en favor de tal rechazo, tenemos una probabilidad del 15% de errar con tal conclusión). El tamaño del efecto es moderadamente negativo, pero aun así su intervalo de confianza no logra rechazar '0' con poco error (r = -0.24; Cohen's *d* = -0.67, 95% FCI [-1.38, 0.06]).

Table 13 | Timing SA; Independent Samples T-Test

					95% FCI for Effect Size	
Test	**Statistic**	**df**	**p**	**Effect Size**	**Lower**	**Upper**
Mann-Whitney	98.500		0.145 [a]	-0.242	-0.576	0.161
Student	-1.873	31	0.071 [a]	-0.667	-1.380	0.056

Notes. For the Student's t-test, effect size is given by Cohen's *d*; for the Mann-Whitney's test, effect size is given by the rank biserial correlation.

[a] Levene's test is significant (p < .05), suggesting a violation of the equal variance assumption.

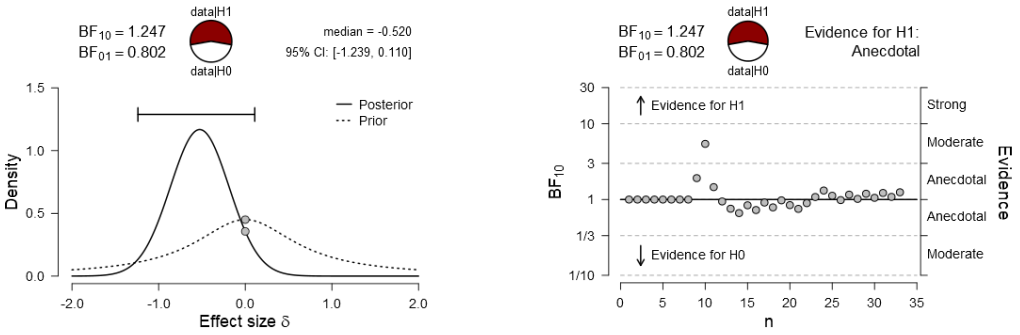
Two-tailed analyses.

Presentamos los resultados bayesianos en la tabla 14, que muestran un poco de apoyo, más bien anecdótico, en favor del modelo alternativo (BF_{10} = 1.25) y en la dirección de un efecto negativo. Dicha evidencia es bastante endeble como para proclamar con confianza que el procedimiento de escaneo tuvo un efecto pernicioso en la velocidad media con la que los pilotos se percataron de aquellos fallos de cabina que lograron percibir[30].

Table 14 | Timing SA; Bayesian Independent Samples T-Test

[29] Si bien la prueba es no direccional, ya hemos puesto a prueba la hipótesis nula de no efecto anteriormente. Por tanto, podemos centrarnos aquí en la posibilidad no explorada de que el efecto fuera negativo de manera estadísticamente significativa, si bien todavía usando una prueba a dos colas, que es más conservadora.

[30] Valga anotar que, si bien la inferencia jeffreysiana desaconseja el interpretar resultados anecdóticos y moderados, las etiquetas mismas pueden llevar a motivar dichas proclamaciones de apoyo poco justificadas. Como la estadística bayesiana no trabaja con probabilidades de error, los análisis frecuentistas en paralelo pueden ayudar a proporcionar una interpretación más moderada de esos resultados bayesianos (en nuestro caso, que concluir que la evidencia empírica soporta la influencia perniciosa de la intervención todavía conlleva una probabilidad razonable, del 15%, de ser una conclusión errónea).

BF_{10}	
1.247	

Note. Two-tailed analyses.

5. Atención Situacional Continua (Continual SA)[31]

Continual SA es una medida *post hoc*, creada al combinar Active SA y Timing SA con la intención de proveer una mejor idea de la seguridad en cabina. Si bien Timing SA nos da una idea del desempeño de aquellos pilotos que percibieron algún fallo, el peor desempeño está limitado, al menos, a la percepción de uno de los fallos. Sin embargo, podemos concebir un desempeño aún peor desde el punto de vista de sus consecuencias para la seguridad: el de los pilotos que no percibieron fallo alguno. Por tanto, Continual SA cubre el rango total de desempeño, evaluando tanto la percepción de fallos (o su falta de percepción) como el retraso, en minutos, que se tardara en percatarse de dichos fallos. Continual SA es un promedio de las puntuaciones recibidas por cada fallo. La escala de Continual SA (y, de manera concurrente, las puntuaciones individuales) abarcaría desde un valor mínimo de '1' (si los tres fallos fueran percibidos en el minuto en que ocurrió) hasta un valor máximo de '4' (si ninguno de los fallos fuera así percibido), con '3' siendo el valor obtenido si los tres fallos fueran percibidos con un retraso mayor a dos minutos. Por tanto, el desempeño obtiene una mayor puntuación promedio cuanto más negativamente afecta a la atención situacional de manera global (incluyendo el retraso en percatarse de todos los fallos, no solo de aquellos que fueron percibidos).

[31] Continual SA fue calculada de manera distinta en este tutorial, con el fin de capturar el desempeño individual de manera más sistemática. Por lo tanto, estos resultados difieren substancialmente de los descritos en la investigación original.

5.1. Continual SA; análisis exploratorio de datos

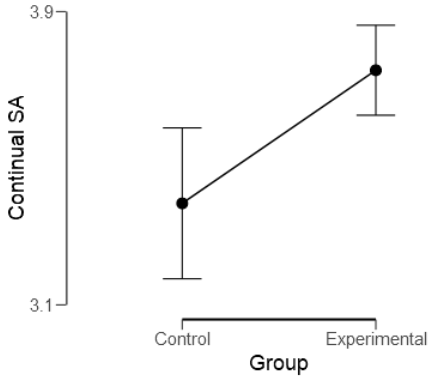
Dado que un BCI descriptivo se calcula asumiendo la equi-probabilidad inicial de las hipótesis, los resultados son bastantes similares a los de un intervalo de confianza frecuentista, aun cuando la interpretación de ambos difiera, filosóficamente hablando. Aquí preferimos BCIs porque la interpretación de los mismos es más coherente con la imagen de una distribución de frecuencias (que es una interpretación errónea para con los FCIs, como hace, por ejemplo, Cumming, 2012[32]).

***JASP 0.13.1**: Seleccionamos la lengüeta 'T-Tests' y elegimos la opción "Bayesian Independent Samples T-Test". Luego seleccionamos 'Continual SA' como variable dependiente y 'Group' como variable de grupo, así como los descriptivos seleccionados anteriormente: "Descriptives", "Descriptives plots" y 95% "Credible interval".*

Los resultados de Continual SA (tabla 15) muestran que el grupo control tuvo un desempeño mejor que el grupo experimental, con un retraso medio menor (M = 3.38; 95% BCI [3.17, 3.58]) que este último (M = 3.74; 95% BCI [3.62, 3.86]).

Table 15 | Continual SA; Group Descriptives

				95% BCI	
Group	**N**	**Mean**	**SD**	**Lower**	**Upper**
Control	23	3.377	0.476	3.171	3.583
Experimental	23	3.740	0.284	3.617	3.863



[32] Cumming genera su gráfico de 'ojos de gato' para representar un FCI a partir de la distribución muestral, pero lo utiliza para describir la muestra original, incluyendo su centralidad (ej., su media) y cobertura (ej., un intervalo del 95%). Sin embargo, un FCI es inferencial (por lo tanto describe la distribución muestral, no la muestra observada), le da la misma probabilidad a cualquier estimación (es decir, la media es tan probable como cualquier otra estimación, lo que también significa que el dibujar tanto la media como la distribución muestral es no sólo irrelevante sino que lleva a confusión), y simplemente representa un intervalo que cubre un porcentaje de la distribución muestral más cercano a la media como intervalo inferencial (es decir, un rango de hipótesis alternativas que no pueden rechazarse sin cometer un error razonable, mayor del 5% de error inferencial). Un BCI, por su parte, sí es una distribución frecuencial (a posteriori) y, por tanto, puede representarse como tal: con una medida de centralidad, una distribución, e invitando la conclusión de que, con un grado de confianza o certeza del 95%, el parámetro se encuentra en el intervalo, y probablemente más cerca de la media que de las colas.

En el contexto de medida usado, sin embargo, vemos que ambos grupos tuvieron un desempeño relativamente pobre, puntuando muy cerca del valor máximo de la escala (máximo = '4'). Como puede verse en la figura, el grupo control no solo obtuvo mejores puntuaciones, sino que su desempeño también fue en la dirección opuesta a la esperada.

5.2. Continual SA; prueba de significación estadística

Continual SA no cumple los requisitos básicos del test paramétrico y, por tanto, será analizada de forma no paramétrica.

***JASP 0.13.1**: Volvemos a la lengüeta 'T-Tests' y elegimos la opción "Independent Samples T-Test". Luego seleccionamos 'Continual SA' como variable dependiente y 'Group' como variable de grupo. También elegimos "Mann-Whitney" y "Student" como 'Tests'. Para determinar la hipótesis de investigación adecuada, seleccionamos al azar "Group 1 > Group 2", que resulta ser la opción correcta, según la nota a pie de tabla en los resultados. También seleccionamos los mismos 'Additional Statistics' que antes: "Effect size" y su "Confidence Interval" al 95%.*

Dado que Continual SA fue creada a partir de Active SA, se esperaba que la intervención tuviera un efecto positivo sobre la variable. Por lo tanto, la tabla 16 muestra resultados direccionales (de una cola). La prueba principal es la *U* de Mann-Whitney, si bien también añadimos los resultados *t* de Student.

Podemos observar que los resultados son altamente probables bajo la hipótesis nula (U = 131.50, p = 0.999), con un efecto de correlación de rango biserial alta (r = -0.50; Cohen's d = -0.93, 95% FCI [-1.43, ∞]).

Table 16 | Continual SA; Independent Samples T-Test

					95% FCI for Effect Size	
Test	**Statistic**	**df**	**p**	**Effect Size**	**Lower**	**Upper**
Mann-Whitney	131.500		0.999	-0.503	-0.684	∞
Student	-3.142	44	0.999	-0.927	-1.433	∞

Note. For the Student's t-test, effect size is given by Cohen's *d*; for the Mann-Whitney's test, effect size is given by the rank biserial correlation.

Con ello aprendemos que los resultados son probables y que, dada la naturaleza direccional de la prueba, no podemos rechazar la hipótesis nula, pero tampoco

podemos aceptar que la intervención tuvo una influencia negativa, como parecen sugerir los datos.

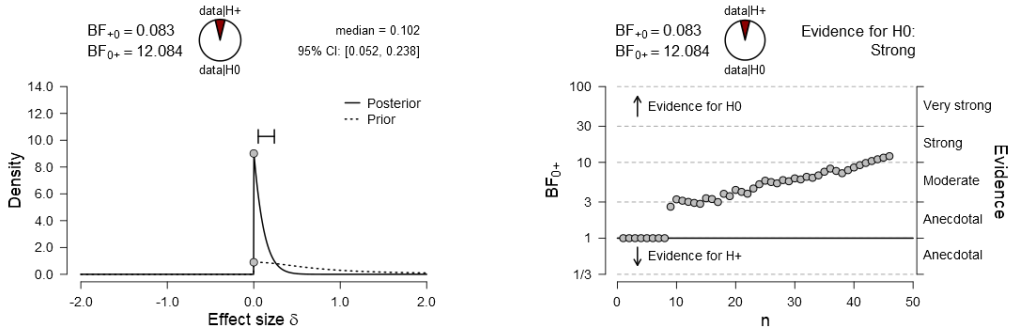
5.3. Continual SA; análisis de factor bayesiano

Dado que solo los datos empíricos tienen peso en la distribución posterior jeffreysiana, el grado de soporte de un modelo sobre el otro puede igualmente traducirse como un grado de soporte de una hipótesis sobre la otra. Cabe recabar, sin embargo, que ya que el factor bayes no trabaja realmente con la probabilidad previa de las hipótesis, siempre existe el riesgo de concluir a favor de una hipótesis de manera errónea (salvo en el caso en que la probabilidad previa real de dichas hipótesis sea del 50%, como se asume).

JASP 0.13.1*: Regresamos a la lengüeta 'T-Tests' y elegimos la opción "Bayesian Independent Samples T-Test". Luego seleccionamos 'Continual SA' como variable dependiente y 'Group' como variable de grupo. También seleccionamos "Group 1 > Group 2" como hipótesis de investigación, el 'Bayes Factor' apropiado—que resulta ser el "BF_{01}"—y los mismos gráficos de antes: "Prior and posterior" con "Additional info", y "Sequential analysis".*

El análisis de factor bayesiano muestra que el modelo 'cero' (M_0) es doce veces más probable que el modelo alternativo (BF_{01} = 12.08; tabla 17). La distribución posterior de una cola identifica el efecto como uno de poco tamaño (mediana = 0.10, 95% BCI [0.052, 0.238]), la evidencia a favor del modelo 'cero' siendo moderada a alta, lo que lleva a la conclusión de que la intervención no tuvo un efecto positivo sobre la atención situacional tal y como fue evaluada por medio de Continual SA.

Table 17 | Continual SA; Bayesian Independent Samples T-Test

BF_{01}	
12.084	

5.4. Continual SA; análisis de dos colas

Una forma de interpretar el factor bayesiano que permite prevenir errores de inferencia es la de describirlo como el cambio proporcional que los datos le dan a la hipótesis previa que favorecen. Es decir, puede que no sepamos con certeza la probabilidad previa de la hipótesis, pero sí podemos concluir algo acerca del peso relativo con el que los datos observados favorecen dicha hipótesis.

***JASP 0.13.1**: En 'Results' activamos la tabla "Independent Samples T-Test" y seleccionamos la hipótesis de investigación no direccional "Group 1 ≠ Group 2". Luego activamos la tabla "Bayesian Independent Samples T-Test" e igualmente seleccionamos la hipótesis "Group 1 ≠ Group 2", comprobamos si el 'Bayes Factor' anterior es apropiado—que no lo es—y lo cambiamos a "BF_{10}".*

Exploramos más profundamente el desajuste entre las expectativas iniciales y los resultados obtenidos, usando análisis a dos colas.

La tabla 18 muestra los resultados de las pruebas de significación, que indican que los resultados tienen una baja probabilidad bajo la hipótesis nula (U = 131.5, p = 0.002), la cual podemos rechazar. De hecho, el tamaño del efecto es largo (correlación de rango biserial = -0.50; Cohen's d = -0.93, 95% FCI [-1.53, -0.31]), y el intervalo de confianza sugiere que pueden rechazarse tanto efectos positivos como efectos negativos muy pequeños (así como efectos negativos muy grandes).

Table 18 | Continual SA; Independent Samples T-Test

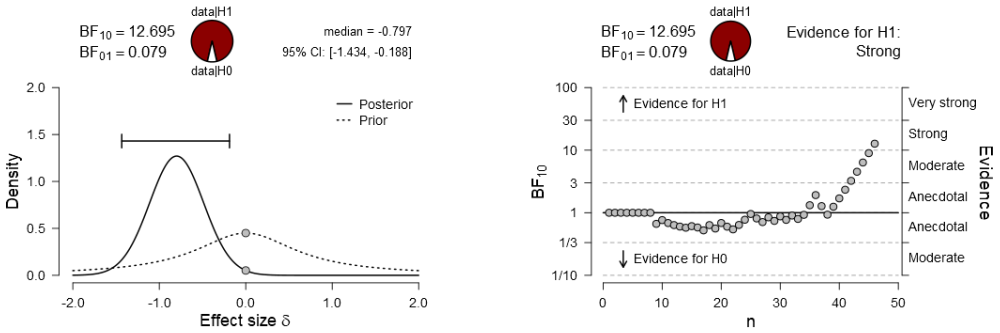
					95% FCI for Effect Size	
Test	**Statistic**	**df**	**p**	**Effect Size**	**Lower**	**Upper**
Mann-Whitney	131.500		0.002	-0.503	-0.712	-0.213
Student	-3.142	44	0.003	-0.927	-1.531	-0.312

Notes. For the Student's t-test, effect size is given by Cohen's *d*; for the Mann-Whitney's test, effect size is given by the rank biserial correlation.
Two-tailed analyses.

La tabla 19 muestra los resultados bayesianos. Aquí también observamos que el modelo alternativo (M_1) es casi trece veces más probable que el modelo 'cero' (BF_{10} = 12.70), con la mediana del efecto centrada en un efecto largo (Cohen d = -0.80, 95% BCI [-1.43, -0.19]) en la distribución posterior. La evidencia a favor de dicho modelo alternativo es relativamente fuerte, lo que soporta la conclusión de que la

intervención realmente parece haber tenido un efecto pernicioso sobre la atención situacional.

Table 19 | Continual SA; Bayesian Independent Samples T-Test

BF_{10}	
12.695	

Note. Two-tailed analyses

Notas finales

En la sección anterior hemos tenido la oportunidad de comprobar tanto la flexibilidad de uso de JASP como la capacidad que ofrece de aprender más de nuestros datos. En suma, hemos demostrado tanto análisis exploratorios como fisherianos y jeffreysianos; tanto análisis planificados como análisis *ad hoc* motivados por la inconsistencia encontrada entre hipótesis y observaciones; tanto análisis direccionales (de una cola) como no direccionales (de dos colas); tanto análisis paramétricos como no-paramétricos; y hemos obtenido tanto resultados significativos como no significativos.

En el último caso, también hemos tenido la oportunidad de comprobar cómo una aproximación jeffreysiana provee una medida del poder relativo de la evidencia a favor o bien del modelo 'cero' o bien del modelo alternativo. Tuvimos incluso la oportunidad de demostrar el caso donde ambas perspectivas pueden contradecirse; es decir, donde un resultado frecuentista no-significativo obtiene un factor bayesiano que soporta el modelo alternativo, aunque fuera de forma anecdótica. Dicho caso también ha servido para argumentar a favor de la idea de Mayo (2017), de que las pruebas frequentistas bien pudieran servir para calibrar los resultados bayesianos con el fin de prevenir conclusiones potencialmente erróneas.

Durante este viaje que aquí termina hemos hecho unas cuantas llamadas de atención por medio de notas a pie de página, llamadas de atención que bien merecen una consideración algo más larga. Por ejemplo, muchas de las llamadas de atención han sido en referencia a la discrepancia entre los resultados descritos en la tesis original y los aquí descritos. Algunas de esas discrepancias (notas 9, 20, 22) tuvieron su causa bien en errores de copiado manual de los resultados bien en desidia (ej., la discrepancia evidente entre expectativa y resultados llevó, en varias ocasiones, a descubrir discrepancias entre la base de datos y los datos originales; pero una vez corregidas, solo los estadísticos con que se estaban trabajando fueron actualizados, en lugar de la tabla al completo). Otras discrepancias tuvieron como causa principal la divergencia en la manera de codificar algunas variables (notas 26, 29), típicamente porque se ignoraron las recomendaciones de codificarlas como se hizo en este tutorial.

Esas discrepancias nos sirven como clave para señalar que ni JASP ni los análisis bayesianos—lo cual se extiende también a cualquier otro programa y análisis estadístico—tienen poder alguno para prevenir dichos errores metodológicos. Es responsabilidad de los investigadores el estar atentos a la posibilidad de cometer errores tanto a nivel metodológico como analítico. Como dice el refrán, ‘de tal palo, tal astilla’, un refrán que nos sirve para enfatizar que la calidad de los resultados depende de la integridad de los datos, la cual, a la hora de determinar la calidad final de cualquier investigación, da prioridad a la integridad metodológica frente a la tecnología estadística.

En nuestro tutorial también hemos tenido la oportunidad de comprobar la simplicidad de comandos de JASP y la velocidad inmediata con la que genera resultados. A menudo hemos usado esa capacidad para asegurarnos de que estábamos trabajando con la hipótesis de investigación correcta (ej., nota 18). Sin embargo, como indicamos en la nota 17, dicha flexibilidad analítica también conlleva el riesgo de decidirnos por aquellas hipótesis que resulten en los resultados deseados o más propicios. JASP tampoco provee protección alguna contra dicha conducta, una conducta que requiere una integridad ética que va más allá de las características técnicas del programa estadístico.

Un tercer punto a discutir es el recordatorio casi constante de que los resultados no concordaban con las hipótesis de investigación, en su mayor parte. Si bien se recuerda, la sensitividad general de la investigación fue estimada de acuerdo a la hipótesis más interesante: la de que la intervención experimental tendría un efecto positivo sensible en la atención situacional. Obviamente, dicha sensitividad contenía en sí misma la posibilidad de efectos positivos de poco tamaño, incluyendo 'cero' (nota 10). Sin embargo, en ningún momento concebimos la posibilidad de que el efecto fuera negativo; mucho menos que lo fuera de manera estadísticamente significativa. (Tenemos ahora una posible explicación que bien pudiera dar sentido a dichas discrepancias. Sin embargo, hasta ponerla a prueba en un nuevo estudio, no sabemos a ciencia cierta en qué medida los resultados aquí obtenidos describen una situación real.)

Y lo dicho arriba nos sirve como punto de partida para discutir dichas expectativas en el contexto de los datos obtenidos, expectativas que un análisis bayesiano propio trataría como parte de la inferencia: es decir, integraría la probabilidad previa de las hipótesis con la probabilidad de los datos bajo dichas hipótesis. Sin embargo, como ya dijimos anteriormente (notas 14, 16, 25), un análisis de factor bayesiano da un rodeo para evitar enfrentarse a dicha integración y solo provee el peso de la evidencia derivada de los datos observados. Por lo tanto, vale la pena recordar aquí que ni los jeffreysianos ni JASP estiman la probabilidad previa de las hipótesis, por tanto tampoco pueden informarnos acerca de la probabilidad posterior de dichas hipótesis una vez que los datos han sido analizados.

Relacionado con lo anterior emerge un cuarto punto de discusión, que es si un análisis paralelo frecuentista – bayesiano realmente añade conocimiento substancial, excepto en aquellas instancias más bien excepcionales como la señalada en la nota 28. Es decir, ¿es dicha propuesta simplemente un juego de pedantería verbal con poca repercusión práctica, si es que tiene alguna, en las inferencias cotidianas del investigador aplicado?

Antes de responder la pregunta, veamos la tabla 20, que presenta las inferencias que razonablemente pudieran haberse hecho basándonos en los análisis realizados. Vemos que tanto las inferencias basadas en las pruebas fisherianas (columnas *p* y *Decision*) como las basadas en las jeffreysianas (columnas *BF* y *Evidence*) son

similares, excepto en el caso V—esto es, el caso excepcional señalado en la nota 28. No solo eso, sino que gracias a que hicimos un análisis de sensitividad, podríamos ignorar las pruebas frecuentistas y basar nuestras inferencias en los datos exploratorios, ya que el conocer el tamaño del efecto también nos permitiría llegar a conclusiones razonables (véase nota 10). Eso es lo que se muestra en la columna 'Cohen's *d*', que motivaría inferencias similares a aquellas alcanzadas con los resultados frecuentistas (teniendo en cuenta que los efectos de los casos II, IV y VI van en la dirección opuesta a la de la hipótesis de investigación planteada). Además, también podríamos añadir los análisis de severidad de Mayo (columna *SEV*) que, si bien son de base frecuentista, las inferencias realizadas a partir de los mismos concordarían bien con las que tomaría un jeffreysiano, incluyendo la diferencia en el peso que se le daría a la evidencia anecdótica de los casos I y V.

Table 20 | Reasonable conclusions based on frequentist and Bayesian results

Case	Cohen's d	p	Decision	SEV	BF	Evidence
I (2t)	0.20	0.507	H_0	0.75	BF_{01} = 2.85	M_0 = anecdotal
II (1t)	0.80	0.995	H_0	0.99	BF_{01} = 10.96	M_0 = strong
III (2t)	0.80	0.010	noH_0	0.99	BF_{10} = 5.04	M_1 = moderate
IV (1t)	-0.67	0.965	H_0	0.99	BF_{01} = 7.20	M_0 = moderate
V (2t)	-0.67	0.071	H_0	0.51	BF_{10} = 1.25	M_1 = anecdotal
VI (1t)	-0.93	0.999	H_0	0.99	BF_{01} = 12.08	M_0 = strong
VII (2t)	-0.93	0.003	noH_0	0.99	BF_{10} = 12.70	M_1 = strong

Note. For comparability purposes, all *p*-values are those of *t*-tests. SEV = Mayo's severity tests.

Por lo tanto, parece que cualquiera de esas perspectivas es suficiente en sí misma y que todas son redundantes. Sin embargo, hemos de recordar que cada una de ellas analiza los datos desde perspectivas diferentes y que cada una permite que aprendamos algo distinto. De hecho, no solo no son intercambiables dichas perspectivas, sino que tampoco pueden usarse para corroborar los resultados de las restantes. Por ejemplo, ya hemos visto cómo incluso un análisis estadístico significativo puede que tenga un peso relativamente minúsculo cuando interpretamos los datos en el contexto de la escala en que la variable fue medida. Por otra parte,

en la tabla 20 se puede observar claramente que un resultado significativo conlleva un peso distinto como evidencia bayesiana, la cual generalmente evita interpretar resultados de valor anecdótico o moderado. El peso de dicha evidencia bayesiana es incluso más palpable cuando traducimos el factor bayesiano a una probabilidad, como mostramos en las notas 16 y 25. Si bien por razones meramente prácticas es posible que usando cualquiera de esas alternativas lleguemos a conclusiones similares, eso no significa que las conclusiones que derivemos de dicha coincidencia estén garantizadas por el análisis llevado a cabo. Valga el siguiente caso ilustrativo: tanto un médico como un homeópata pueden coincidir en un tratamiento particular; sin embargo, esa coincidencia ni hace científicos los principios homeopáticos ni equipara la base del conocimiento de ambos terapeutas. Es decir, 'no todo lo que brilla es oro' y asumir que lo es conlleva el riesgo de tomar decisiones pseudocientíficas—en este caso, no basadas en los análisis llevados a cabo—cuando, con muy poco esfuerzo, el análisis apropiado estaba disponible. El análisis paralelo con JASP, si bien no es perfecto, sí que es una oportunidad excepcional de no confundir lo que una perspectiva analítica provee con lo que provee la otra, lo cual nos permite prevenir errores conceptuales e inferenciales a la hora de generalizar nuestros resultados.

A pesar de lo dicho, siempre podríamos concluir que, al final de cuentas, todo queda en un caso de 'tanto vale, vale tanto': si llegamos a la misma conclusión, entonces no es tan importante qué herramienta se use. O, dicho de otra manera, el usar los análisis apropiados puede que sea suficiente pero no necesario para alcanzar una conclusión válida. Es por ello que avanzamos un argumento más: el de quién es el beneficiario final de la investigación. Si dichos beneficiarios son investigadores, entonces es posible pasarles a ellos la responsabilidad '*caveat emptor*'; es decir, de leer entre líneas y concluir por sí mismos, no sea que 'les den gato por liebre'. Sin embargo, si el beneficiario final es la sociedad, en general (véase, por ejemplo, Stengers, 2010), entonces recae en nosotros la responsabilidad de proveer conclusiones claras y sin disimulos. Por ejemplo, es bastante típico ver exhortaciones del proceso de evolución darwiniana como ley biológica que, sin embargo, usan ejemplos que son característicos de la evolución lamarkiana. La excusa de que el que sabe, entiende, es una excusa pobre para justificar tal uso, ya

que el mismo ni enseña al que no sabe ni protege, al que sabe, de confundirse. El mismo argumento vale para los análisis estadísticos.

El último punto a tratar se relaciona con el anterior, pero es más bien filosófico (ej., Mayo, 1996; Mayo y Spanos, 2010). Como señalamos en las notas 12, 13 y 23, la perspectiva frecuentista se encuentra anclada en la epistemología lógica basada en el argumento por contradicción de Popper. El *Modus Tollens* propio de dicha perspectiva depende del nivel de significación elegido para poner a prueba la hipótesis nula substantiva con la que se trabaja y, si llega el caso, rechazarla. Igualmente, dicha perspectiva tiene poco interés en un argumento por afirmación y en el silogismo lógico de un *Modus Ponens* que pueda demostrar la veracidad de dicha hipótesis nula. Es decir, el interés es en rechazar una hipótesis que pueda ser falsa, no en retenerla o demostrarla.

La perspectiva bayesiana, por su parte, se encuentra anclada en la epistemología lógica basada en el argumento por afirmación de Carnap, concluyendo que la hipótesis (o el modelo) más probable es aquella favorecida por la evidencia encontrada. El interés es en determinar qué hipótesis es la más probable considerando la evidencia que tenemos.

La diferencia entre ambas perspectivas puede observarse mejor en otro contexto. La perspectiva bayesiana decide entre inocente o culpable según el peso de la evidencia—si bien es posible poner límites para que dicha evidencia solo cuente a partir de un mínimo dado, lo cual llevaría a no decidir entre inocente o culpable a menos que el peso de la evidencia demuestre lo contrario. Por su parte, la perspectiva frecuentista solo decide cuándo se es culpable más allá de una duda razonable.

Referencias

APA (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: APA.

Cumming, G. (2012). Understanding the New Statistics. Effect sizes, confidence intervals, and meta-analysis. New York, NY: Routledge.

Fisher, R. A. (1954). Statistical methods for research workers (12th ed.). Edinburgh, U.K.: Oliver and Boyd.

Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33, 587–606. doi:10.1016/j.socec.2004.09.033

Jeffreys, H. (1961). Theory of probability (3rd ed.). Oxford, N.Y.: Clarendon Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability. A “Many Labs” replication project. Social Psychology, 45, 142-152. doi:10.1027/1864-9335/a000178

Kruschke, J. K. (2011). Doing Bayesian data analysis. A tutorial with R and BUGS. Oxford, UK: Academic Press.

Mayo, D. (2017). *New venues for the statistics wars* [Web log post]. Retrieved from https://errorstatistics.com/2017/10/05/new-venues-for-the-statistics-wars

Mayo, D. G. (1996). Error and the growth of experimental knowledge. Chicago, IL: The University of Chicago Press.

Mayo, D. G., and Spanos, A. (eds.). (2010). Error and inference. New York, NY: Cambridge University Press.

Neyman, J., and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. Biometrika, 20A, 175–240. doi: 10.2307/2331945

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. Perspectives in Psychological Science, 7, 657–660. doi:10.1177/1745691612462588

Perezgonzalez, J. D. (2014). A reconceptualization of significance testing. Theory & Psychology, 24, 852–859. doi:10.1177/0959354314546157

Perezgonzalez, J. D. (2015a). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology, 6, 223. doi:10.3389/fpsyg.2015.00223

Perezgonzalez, J. D. (2015b). P-values as percentiles. Frontiers in Psychology, 6, 34. doi:10.3389/fpsyg.2015.00034

Perezgonzalez, J. D. (2015c). Confidence intervals and tests are two sides of the same research question. Frontiers in Psychology, 6, 341. doi:10.3389/fpsyg.2015.00341

Perezgonzalez, J. D. (2016). Statistical sensitiveness for science. Arxiv. Retrieved from https://arxiv.org/abs/1604.01844.

Perezgonzalez, J. D. (2017a). Statistical sensitiveness for the behavioural sciences. PsyArxiv. doi:10.17605/OSF.IO/Y969T. Retrieved from https://psyarxiv.com/qd3gu.

Perezgonzalez, J. D. (2017b). The fallacy of placing confidence in confidence intervals – A commentary. PsyArxiv. doi:10.31234/osf.io/kvxc4. Retrieved from https://psyarxiv.com/kvxc4.

Perezgonzalez, J. D. (2017c). Commentary: The need for Bayesian hypothesis testing in psychological science. Frontiers in Psychology, 8, 1434. doi:10.3389/fpsyg.2017.01434

Perezgonzalez, J. D., and Frías-Navarro, M. D. (2018). Retract $p < 0.005$ and propose using JASP, instead [version 2]. F1000Research, 6, 2122. doi:10.12688/f1000research.13389.2

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin Review, 16, 225–237. doi:10.3758/PBR.16.2.225

Stengers, I. (2018). Another science is possible. A manifesto for slow science. Cambridge (U.K.): Polity Press.

Tabachnick, B. G., and Fidell, L. S. (2001). Using multivariate statistics (4th ed.). Boston, MA: Allyn & Bacon.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley Publishers.

Vincent, N. (2018). Situational awareness of pilots in the cruise (Master's thesis, Massey University, New Zealand).

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., . . . Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld and I. D. Waldman (Eds.), Psychological science under scrutiny: recent challenges and proposed solutions (pp. 123–138). Chichester, UK: John Wiley & Sons. doi:10.1002/9781119095910.ch8

Bloque 4. Elaborar un informe

Capítulo 16. Versión Revisada de la Escala de Arraigo de Inmigrantes Latinoamericanos en España (redacción de un informe de investigación)

José Berríos-Riquelme
Manuel Martín-Fernández
Viviana Vargas-Salinas
Carla Vidal-Figueroa
Cristóbal Pulido-Iparraguirre

*Universidad de Tarapacá, Chile
**Universidad Autónoma de Madrid, España
***Universidad de Valencia, España
****Universidad de Concepción, Chile

Citar el capítulo como:

Berríos-Riquelme, J., Martín-Fernández, M., Vargas-Salinas, V., Vidal-Figueroa, C., y Pulido-Iparraguirre, C. (2021). Versión Revisada de la Escala de Arraigo de Inmigrantes Latinoamericanos en España (redacción de un informe de investigación). En D. Frías-Navarro y M. Pascual-Soler (Eds.), *Diseño de la investigación, análisis y redacción de los resultados*. Universidad de Valencia. España.

Versión Revisada de la Escala de Arraigo de Inmigrantes Latinoamericanos en España

José Berríos-Riquelme

Trabajador Social y doctor en Movilidad Humana
Universidad de Tarapacá - Chile
jberrios@uta.cl

Manuel Martín-Fernández

Psicólogo y doctor en Psicología
Universidad Autónoma de Madrid – España
manuel.martinfernandez@uam.es

Viviana Vargas-Salinas

Psicóloga y doctora en Psicología
Universidad de Valencia - España
viviana.vargas@uv.es

Carla Vidal-Figueroa

Trabajadora Social y doctora en Ciencias Sociales
Universidad de Concepción - Chile
carvidal@udec.cl

Cristóbal Pulido-Iparraguirre

Psicólogo y doctor en Estudios Latinoamericanos
Universidad de Tarapacá - Chile
cpulido@academicos.uta.cl

Resumen

El arraigo en el contexto migratorio se entiende como el grado de apego que tienen los inmigrantes con la sociedad y cultura receptora. Considerando la importancia de esta variable, son escasos los instrumentos que evalúan este constructo para la población latinoamericana. Considerando esto, el objetivo del presente trabajo es evaluar la estructura factorial y las evidencias de validez de la escala de arraigo en inmigrantes latinoamericanos en España. Se realizó un estudio con dos fases en el que participaron un total de 710 inmigrantes latinoamericanos. Para probar la estructura factorial de la escala se utilizaron análisis factoriales exploratorios, análisis paralelo y análisis factorial confirmatorio. Para valorar la validez en relación a otras variables, se utilizaron escalas de emociones percibidas y la variable contacto intergrupal. Los resultados de ambas fases indican que la escala presenta una estructura bifactorial de dos dimensiones: arraigo ecológico y arraigo cultural. Las medidas de ajuste y consistencia interna fueron buenas. Se concluye que la versión revisada de la escala de arraigo es una medida válida y fiable para ser utilizada con población latinoamericana en España. Esta herramienta es breve, rápida y fácil de usar en el campo de la intervención psicosocial con población inmigrante.

Palabras claves: propiedades psicométricas, bifactor, análisis factorial, inmigración latinoamericana, políticas públicas.

Abstract

Rooting in the migratory context is understood as the degree of attachment that immigrants have with the receiving society and culture. Few instruments evaluate this construct for the Latin American population considering the importance of this variable. The present study aimed to evaluate the factorial structure and the evidence of the rooting scale's validity in Latin American immigrants in Spain, considering this aspect. A two-phase study was conducted in which a total of 710 Latin American immigrants participated. Exploratory factor analysis, parallel analysis, and confirmatory factor analysis were used to test the scale's factorial structure. Scales of perceived emotions and the intergroup contact variable were used to assess validity concerning other variables. The results of both phases indicate that the scale presents a bifactor structure with two dimensions: ecological attachment and cultural attachment. The measures of fit and internal consistency were good. It is concluded that the revised version of the rooting scale is a valid and reliable measure to be used with the Latin American population in Spain. This tool is brief, quick, and easy to use in the field of psychosocial intervention with the immigrant population.

Keywords: psychometric properties, bifactor, factor analysis, Latin American immigration, public policies.

Introducción

La inmigración se ha convertido en un fenómeno a nivel mundial durante las últimas tres décadas. Las grandes cantidades de personas que se desplazan de una sociedad a otra, lo hacen para escapar de contextos sociales, económicos y políticos que no les garantizan el cumplimiento de su proyecto de vida. En la actualidad, se estima que en el mundo hay 272 millones de migrantes internacionales, una cifra que equivale al 3.5% de la población mundial (Naciones Unidas, 2019).

En este contexto global, la sociedad española pasó de ser un país emisor de personas a convertirse en un foco receptor de inmigrantes (Berríos, 2017). De acuerdo a datos entregados por el Instituto Nacional de Estadística (INE, 2019), el total de la población extranjera de España es de 5.036.878, las cuales representan el 11.4% de la población total del país. Del total de extranjeros residentes, el número que pertenece a países no comunitarios es de 3.204.338, donde los conglomerados más numerosos son los procedentes de Latinoamérica con (1.254.004) y África (1.122.409).

Distintas razones podrían explicar el alto número de personas procedentes de estas regiones. En el caso del colectivo africano, es posible mencionar la proximidad geográfica y su interés por utilizar la península ibérica como puerta de acceso al mundo occidental. La explicación de la población latinoamericana recae en el pasado histórico que conecta a ambas regiones y las características culturales compartidas (Bayona et al., 2018). De acuerdo a datos del INE (2019), las principales nacionalidades de inmigrantes latinoamericanos en España son Colombia (206.719), Venezuela (137.776), Ecuador (131.814), Bolivia (95.717) y Brasil (90.304).

A pesar de contar con un pasado común y lazos culturales, la realidad del proceso migratorio y sus consecuencias psicosociales en la calidad de vida de estas personas siguen siendo un desafío para los inmigrantes y para el país de acogida. En este sentido, las variables

relacionadas al bienestar emocional y subjetivo de los inmigrantes suelen estar asociadas al nivel de arraigo que tienen (Berríos, 2017).

El arraigo es una variable psicosocial que se entiende como el proceso de conformación de vínculos afectivos de los inmigrantes con la sociedad de acogida, donde se lleva a cabo una reorganización simbólica del espacio y de las relaciones intergrupales que se establecen en el nuevo país. En este sentido, el arraigo permite conocer cómo los inmigrantes vinculan sus proyectos migratorios con la formación de sus hogares y familias, así como las pautas de empleo sobre las que se organiza la supervivencia individual y familiar (Sampedro y Camarero, 2016).

El arraigo tendría un papel fundamental en los procesos cognitivos, afectivos e identitarios que viven las personas en la sociedad de acogida, propiciando el bienestar emocional y favoreciendo la calidad de las relaciones intergrupales. Asimismo, el arraigo como apego significativo tendría un papel predominante para calmar la angustia emocional que viven las personas inmigrantes por dejar su vida pasada (Morgan, 2010).

El arraigo de los inmigrantes en la sociedad de acogida es fundamental para entender los elementos relacionados con la adaptación social de estas personas, puesto que sus costumbres, formas de relacionarse y de expresarse estarán determinadas por el apego que tienen hacia el nuevo lugar donde se integran. De este modo, el estudio del arraigo ayuda a entender cómo la vinculación con el nuevo entorno favorece las relaciones interétnicas positivas y contribuye a formar el sentimiento de pertenencia hacia la nueva sociedad (Torrente et al., 2011).

El arraigo de los inmigrantes en ciencias sociales

El estudio del arraigo ha sido abordado principalmente desde la antropología y la sociología para discutir cómo las personas y los lugares se relacionan con la cultura. Estas disciplinas han enfatizado aspectos como la raza, la etnicidad, las minorías políticas, el racismo,

la diáspora, la migración y la identidad (Gustafson, 2001). Al mismo tiempo existen acercamientos teóricos desde diversas disciplinas que entienden el arraigo como apego al lugar, ejemplo de aquello son las definiciones aportadas desde la psicología social, la antropología, la sociología comunitaria, la psicología ambiental, la socio-demografía y la geografía, entre otras (López-Mosquera, y Sánchez, 2013; Raymond y et al., 2010; Scannell y Gifford, 2010).

Considerando estas cuestiones se pueden encontrar distintas definiciones de arraigo relacionado con el apego al lugar. El diccionario de la lengua española entiende el arraigo como el "echar o criar raíces" o "establecerse de manera permanente en un lugar, vinculándose a personas y cosas". Además, es posible mencionar la acepción de Kyle et al. (2004), para quienes el arraigo son los "vínculos que los humanos comparten con lugares específicos" (p. 65). Por otra parte, Milligan (1998) define el arraigo como el "vínculo emocional formado por un individuo a un lugar físico al que ha dado significado a través de la interacción" (p. 2).

Desde el punto de vista académico, una de las definiciones más utilizada es la de Quezada (2007), quien entiende el arraigo como "el proceso y efecto a través del cual se establece una relación particular con el territorio, en la que metafóricamente se echan raíces en él por diversas situaciones, creando lazos que mantienen algún tipo de atadura con el lugar" (p. 43). En esta misma línea, el arraigo ha sido entendido como la experiencia de un vínculo afectivo de largo término a un área geográfica en particular y el sentido atribuido a este vínculo (Morgan, 2010). Por otro lado, Torrente et al. (2011) conceptualizaron el arraigo haciendo referencia cuando una persona está:

> Insertada social y culturalmente, que entiende y comparte su nueva sociedad, que ha desarrollado las destrezas necesarias para vivir en esa nueva estructura social y cultural, que se siente unida a su entorno, tanto al más lejano como al más próximo e inmediato, tanto en términos de relaciones (amigo, vecinos, compañeros de trabajo) como en términos ambientales (su hogar o su barrio). (p. 844)

Desde este punto de vista, Acebo (1996) señala que el arraigo más que ser un proceso de inserción en la sociedad, es una construcción simbólica de vínculos y de relaciones sociales que se articulan dentro de una ciudad-territorio. Como complemento de esta idea y desde la teoría de apego al lugar, el arraigo es el significado simbólico que se le otorga al vínculo afectivo que la persona tiene con el territorio que habita (Morgan, 2010; Vidal y Pol, 2005). Es decir, el arraigo se articula por la adquisición de vínculos significativos a nivel social y territorial. De esta manera, el arraigarse se considera como una fuente de apego que suscita vínculos emocionales y lazos con la comunidad (Gustafson, 2001).

Como se aprecia, los autores que tratan el tema del apego al lugar reconocen, explícita o implícitamente, que las personas desarrollan sentimientos de arraigo hacia lugares como barrios y ciudades (Hidalgo y Hernández, 2001). Las definiciones del término comparten la idea que el arraigo es un proceso que se forma por la manera en que los inmigrantes se apropian del espacio y de las costumbres de la sociedad de acogida. Esta diversidad de definiciones indica la preocupación por encontrar un cuerpo teórico que permita explicar el fenómeno adecuadamente (Dvorak, et al., 2013; Jennings y Krannich, 2013; Raymond et al., 2010; Scannell y Gifford, 2010).

Los estudios del arraigo en el área de las migraciones se han destacado por analizar el vínculo entre el arraigo y los índices de adaptación del inmigrante en la sociedad de acogida (Sochos y Diniz, 2012). Además, se han realizado investigaciones acerca de la relación entre el arraigo y la decisión de no migrar (Barcus y Brunn, 2010) y el apego al lugar del inmigrante (Ng, 1998). Dentro de la literatura española, el arraigo es una temática de interés en el área de las migraciones, encontrando alcances del concepto desde distintas perspectivas, tales como una norma jurídica (Carrera, 2006; Franch, et al., 2011), como sinónimo de integración (Rinken, 2006) y como apego al lugar a través de vínculos (Torrente et al., 2011).

Medición del arraigo

En España, Torrente et al. (2011) propusieron una medida para estudiar el arraigo en población latinoamericana. Mediante una primera fase de corte cualitativo, realizaron un focus group para indagar sobre las temáticas comunes de la población objeto de estudio. La información obtenida de las sesiones se utilizó para elaborar los ítems de la escala que daban cuenta del vínculo que tenían los inmigrantes latinoamericanos con España. Luego seleccionaron los ítems que pasarían a formar parte de la escala de tipo Likert de 5 puntos, donde las personas debían indicar el grado en que se sentían vinculados en temas relacionados al ambiente físico, social y cultural, desde *nada en absoluto* (1), hasta *mucho* (5).

Posteriormente, los autores de la escala utilizaron una muestra de 648 inmigrantes latinoamericanos para la evaluación psicométrica del instrumento, donde 356 (54.9%) eran hombres y 292 (45.1%) fueron mujeres. Sus edades comprendían entre los 20 y 63 años. La muestra se estratificó por sexo y por edad, la cual tuvo tres categorías (entre 20 y 34 años, entre 35 y 49 años y desde 50 o más años). Los criterios utilizados para depurar la escala antes de realizar el análisis factorial fueron los siguientes: se eliminaron los ítems que mostraron menor relación con la correlación ítem-escala.

Con los ítems restantes se ejecutó un análisis factorial de componentes principales con rotación varimax de acuerdo a los criterios estipulados por Kaiser (1960), donde la carga en cada factor debía ser mayor .50. La consistencia interna de la escala de 16 ítems fue de .90 y el análisis factorial arrojó tres factores. El primer factor explicó el 29.45% de la varianza y tuvo una consistencia interna de .89, denominado "Arraigo Cultural", el cual contiene ítems relacionados con la forma en que se organizan los inmigrantes latinoamericanos en España, tales como "el sistema político y administrativo de España", "el sistema educativo español", "la forma de hablar y expresar sentimientos de los españoles". La correlación ítem-factor fluctuó entre .57 y .76. El segundo factor se definió como Arraigo Ecológico y explicó el

16.83% de la varianza y obtuvo una consistencia interna de .79. Esta dimensión está compuesta por ítems que aluden al vínculo del inmigrante con su entorno cercano, tales como "la ciudad, el pueblo o el barrio donde vive". La correlación ítem-factor fue de .41 a .71. Por último, el factor denominado "Arraigo social y laboral "explicó el 14.71% de la varianza y obtuvo una consistencia interna de .73. Sus ítems explicitan las relaciones sociales o laborales de arraigo, como por ejemplo "los compañeros de trabajo que tiene", "las amistades que tiene", entre otros. La correlación ítem factor varió entre .47 y .58.

Tras analizar el proceso seguido para la construcción de esta escala, es posible afirmar que tiene varias ventajas, destacando un diseño específico para ser utilizado con población inmigrante latinoamericana en España y que considera el arraigo como un fenómeno multidimensional. Sin embargo, también es posible mencionar algunas limitaciones, como que gran parte de su sustento teórico es desde una perspectiva socio-jurídica del arraigo y no desde un enfoque psicosocial. Además, la rotación ortogonal escogida en el modelo original asume que las dimensiones del constructo no están relacionadas. Por último, no se encontraron evidencias suficientes de su estructura factorial y de su validez convergente, motivo por el que no ha avanzado desde su fase inicial.

Considerando lo anterior, la escala manifiesta propiedades relacionadas a su proceso de creación que requieren un análisis más profundo. Por ejemplo, en el estudio inicial uno de los factores encontrados se denominó "Arraigo Social y Laboral", lo que genera algunas contradicciones si se considera que casi un tercio de la muestra, exactamente 187 personas (29.4%), no estaban trabajando y 279 personas (43.9%) tenían un trabajo eventual. Entonces ¿cómo es posible que posean un vínculo con un trabajo que no tienen o que realizan de forma esporádica? Por otro lado, surgen interrogantes si se analiza el procedimiento utilizado en el análisis factorial que son los que vienen por defecto en el software SPSS: componentes principales, rotación varimax y la regla K1 de Kaiser para retener factores. Este procedimiento

tiende a sobredimensionar el número real de factores presentes en una escala, extrayendo un número de variables independientes entre sí (p.ej., componentes) que explican el mayor porcentaje de varianza de los datos. Por lo que de esta manera y de acuerdo con Schmitt (2011) en que la mayoría de los factores psicológicos están correlacionados, al emplear este procedimiento se producen estructuras factoriales poco realistas.

Por estos motivos, el objetivo del presente estudio es evaluar la estructura factorial y las evidencias de validez de la escala de arraigo en inmigrantes latinoamericanos en España.

Método

Se utilizó una metodología no experimental con un diseño transversal. La selección de la muestra se realizó mediante un muestreo no probabilístico (muestreo por conveniencia) en la ciudad de Valencia (España). Para recoger la información se aplicó un cuestionario auto administrado, el que solicitó información de variables socio-demográficas y un conjunto de escalas que forman parte de una investigación más amplia. Los criterios de inclusión fueron ser inmigrante proveniente de Latinoamérica, llevar más de un año en España y ser mayor de edad según la legislación española.

Participantes

El primer grupo de participantes estaba compuesto por 298 inmigrantes latinoamericanos mayores de edad y que llevaban más de un año residiendo en España (dos criterios de inclusión). Los hombres fueron 114 (38,3%) y las mujeres 184 (61,7%). La edad de estas personas comprende entre los 18 y los 64 años de edad (M_{edad} = 36.84, DT_{edad} = 11.34). La mayoría de ellos provienen de Chile (50%), Ecuador (18%) y Bolivia (18%). En cuanto a su situación legal en el país, 275 participantes (93,3%) señalaron encontrarse en situación regular, mientras que 21 personas (7%) declararon estar en situación irregular y dos personas dejaron en blanco esta pregunta (0,7%). En lo que respecta a la solicitud de ciudadanía, 168 personas

(56,4%) declararon haberla solicitado, 127 (42,6%) no la han solicitado y tres personas (1%) no contestaron la pregunta.

El segundo grupo de participantes estaba constituido por 412 inmigrantes latinoamericanos (mujeres N = 248; hombres N = 162) que cumplían los mismos criterios de inclusión que la muestra anterior. Presentaron una media de edad de 37.3% (*DT* = 11.4). Según nacionalidad el mayor porcentaje correspondió a las personas inmigrantes provenientes de Colombia (40.5%) seguido por las personas provenientes de Ecuador (23.4%) y de Bolivia (14.6%). Con respecto a su situación legal de residencia, 361 personas (90%) tenían una situación regularizada y 40 (10%) irregular (11 personas no contestaron a esta pregunta). En cuanto a la solicitud de ciudadanía un 57.7% (N = 232) había solicitado la ciudadanía, un 42.3% (N = 170) no había realizado el trámite y un 2.4% (N = 10) no contestaron a esta pregunta.

Instrumentos

Escala de Arraigo de Inmigrantes Latinoamericanos (Torrente et al., 2011). Consta de 16 ítems que se dividen en tres dimensiones que evalúan el grado de arraigo de los participantes: arraigo ecológico (p.ej., *¿Hasta qué punto se siente vinculado con la ciudad en la que vive?),* arraigo social y laboral (p.ej., *¿Hasta qué punto se siente vinculado con los compañeros de trabajo que tiene?*), y arraigo cultural (p.ej., *¿Hasta qué punto se siente vinculado con la forma de pensar de los españoles?*). El formato de respuesta fue una escala tipo Likert de 5 categorías (1 = *Nada*; 5 = *Mucho*). En el estudio original la consistencia interna de la escala fue buena (α de Cronbach = .90).

Contacto con población española. Para medir la variable se crearon tres ítems: 1) ¿hay alguna persona española dentro de tus amistades más cercanas? Cuyas opciones de respuesta fueron *No* (1) y *Sí* (2); 2) ¿Compartes con alguna persona española alguna actividad que no se relacione con tu trabajo o estudios?: *Nunca* (1), *Algunas veces* (2) y *Siempre* (3); 3) y ¿Cómo valoras tu relación con los españoles? *No afectiva* (1), *Normal* (2) y *Afectiva* (3). Al sumar las

puntuaciones de los respectivos ítems se puede clasificar a los sujetos en función del grado de contacto que tienen con la población española. Los valores hasta 6 puntos indican un contacto moderado con españoles, mientras que las puntuaciones entre 7 y 8, señalan un contacto elevado.

Escala de Creencias sobre las Emociones que el Exogrupo percibe del propio Endogrupo de Inmigrantes (Emociones Auto-percibidas, EMOAUTO, Frías-Navarro, 2015). Esta escala tiene como propósito evaluar las creencias que los inmigrantes tienen respecto a las emociones que creen que sienten los españoles (exogrupo) hacia ellos (endogrupo). La escala plantea: "Valora cada palabra en función del grado de emoción que crees causar en los españoles. Por ejemplo, con la palabra Agradecimiento, si piensas que los españoles sienten agradecimiento hacia los inmigrantes, puedes marcar las opciones medio, bastante o mucho. En cambio, si crees que no sienten agradecimiento hacia ti, puedes marcar nada, poco o medio".

Las emociones estudiadas pueden separarse en una dimensión de emociones positivas y en otra dimensión de emociones negativas. Dentro de las emociones negativas se encuentra creer que el exogrupo (los nacionales) manifiesta rechazo, desconfianza, inseguridad, lástima, odio, indiferencia y miedo hacia el propio endogrupo del inmigrante; la consistencia interna de la escala fue buena según el valor alfa de Cronbach .845 (95% IC .82 a .87). Las emociones positivas estudiadas fueron simpatía, atracción, admiración, respeto y solidaridad por parte del exogrupo; la consistencia interna de la escala es buena según el valor del alfa de Cronbach .765 (95% IC .72 a .81). Para ambos tipos de emociones las opciones de respuesta tuvieron un formato tipo Likert donde las puntuaciones oscilaban entre *Nada* (1), hasta *mucho* (5). A mayor puntuación obtenida, mayor es el grado de la emoción percibida.

Procedimiento

Los instrumentos fueron aplicados en dos fases y muestras distintas con un intervalo de tiempo de 3 años. La recolección de datos se realizó en asociaciones de inmigrantes

latinoamericanos de forma individual y en grupos de no más de cinco personas en la ciudad de Valencia, España. Se resguardó la confidencialidad de todos los datos y se contó con el consentimiento de todos los participantes.

Análisis estadísticos

El análisis de los datos fue realizado en dos fases siguiendo los criterios descritos por Goetz et al. (2013) consistentes en: (1) estudiar las propiedades psicométricas de la escala (p.ej., fiabilidad y validez) en una muestra diferente y seleccionar aquellos ítems más relevantes atendiendo a la validez de constructo, la validez de contenido y la consistencia interna; (2) seleccionar los ítems de acuerdo al modelo conceptual propuesto por Torrente et al. (2011); (3) seleccionar los ítems más relevantes para medir el constructo de acuerdo a un panel de expertos; (4) estudiar las propiedades psicométricas de la versión reducida del instrumento comparándola con la versión original y (5) validar la escala reducida en una nueva muestra independiente. Además, un panel de tres expertos evaluó la relevancia de cada uno de los ítems en una escala Likert (1 = *muy poco relevante para medir el constructo*; 5 = *muy relevante para medir el constructo*; referencias del VB).

En la primera fase de este estudio se abordaron las cuatro primeras recomendaciones para construir la versión reducida de la escala de arraigo. Mientras que en la segunda fase se realizaron los análisis de las propiedades psicométricas y la validación de la versión reducida en una nueva muestra.

Fase 1. En esta etapa se analizaron las propiedades psicométricas de la escala de arraigo. Primero se realizó un análisis descriptivo de todos los ítems, se obtuvieron las correlaciones entre cada ítem y el resto del test y se calculó la consistencia interna de la escala mediante el α de Cronbach. Para estudiar la validez de constructo del instrumento se llevó a cabo un análisis factorial exploratorio (AFE). Se comenzó con un análisis paralelo utilizando la matriz de correlaciones policóricas (Horn, 1965). Este procedimiento permite determinar el número

mínimo de factores para explicar la variabilidad de las respuestas de la escala al comparar los autovalores de los datos empíricos con los autovalores de un conjunto de datos simulados (Garrido et al., 2013).

Posteriormente se compararon dos modelos factoriales atendiendo al análisis paralelo y a los resultados encontrados por Torrente et al. (2011). Dada la naturaleza ordinal de los ítems, se utilizó el método de estimación mínimos cuadrados ponderados con medias y varianzas ajustadas (*WLSMV*) con la rotación oblicua OBLIMIN, ya que es un método más robusto con este tipo de datos (Asparouhov y Muthén, 2010). Para evaluar el ajuste de los modelos se utilizaron los estadísticos *CFI*, *TLI*, *RMSEA* y *SRMR*. Valores superiores a .90 - .95 para el *CFI* y el *TLI* son considerados como un indicador de buen ajuste (Hu y Bentler, 1998; 1999). Respecto a RMSEA, los valores inferiores a .08, a .06 y a .01 se considera como un indicador de ajuste moderado, bueno o excelente, respectivamente (MacCallum et al., 1996). Por último, valores inferiores a .08 del *SRMR* se consideraron como un buen ajuste del modelo.

Fase 2. En primer lugar, se realizó un análisis descriptivo de los ítems, obteniendo su media, desviación típica, rango, y estadísticos de asimetría y kurtosis. Se obtuvieron asimismo las correlaciones entre cada ítem y el resto del test.

En segundo lugar, se llevó a cabo un análisis factorial confirmatorio (AFC) comparando tres modelos factoriales utilizando mínimos cuadrados ponderados con medias y varianzas ajustadas (*WLSMV*) como método de estimación. El primer modelo fue un modelo unifactorial, en el que todos los ítems saturaron en un único factor (p.ej., arraigo general). El segundo modelo plantea una estructura de dos factores correlacionados, donde los ítems de arraigo ecológico saturaron en un factor, y los ítems de arraigo cultural en otro. El tercer modelo, por su parte, plantea una estructura bifactorial, donde todos los ítems saturarían en un factor no-específico común (p.ej., Arraigo General), y en un único factor específico (p.ej., Arraigo Ecológico o Arraigo Cultural). En este último modelo los factores son ortogonales, de manera que el factor

general explicaría la varianza común de todos ítems, mientras la proporción de varianza debida a los factores específicos quedaría aislada y explicada por los factores de arraigo ecológico y arraigo cultural.

Para evaluar el ajuste de los modelos se emplearon los estadísticos *CFI*, *TLI*, *RMSEA* y *SRMR*. Valores superiores a .90 - .95 para el *CFI* y el *TLI* fueron considerados como un indicador de buen ajuste (Hu y Bentler, 1998; 1999), mientras que valores inferiores a .08, a .06 y a .01 del RMSEA se considera como un indicador de ajuste moderado, bueno o excelente, respectivamente (MacCallum et al., 1996).

En tercer lugar, para poner a prueba las evidencias de validez basadas en la relación con otras variables, se obtuvieron las correlaciones de Pearson entre las puntuaciones factoriales de Arraigo con las Emociones positivas y Emociones Negativas. Todos los análisis se realizaron con el paquete estadístico R (R Core Team, 2016), con excepción del AFE que se realizó con Mplus 7.1 (Múthen y Múthen, 2010).

Resultados

Fase 1. Análisis descriptivo de los ítems y fiabilidad. La mayoría de los ítems presentaron una media centrada a 3.5 y con una desviación típica ligeramente superior a 1 (ver Tabla 1). Los ítems mostraron una ligera asimetría, por lo que la mayoría de los participantes tendió a seleccionar las categorías de frecuencia intermedias (i.,e., *Medio*, *Bastante*). Las correlaciones entre los ítems y el resto del test resultaron en general elevadas con valores superiores a .50. La consistencia interna de la escala de arraigo en la muestra fue excelente según el valor alfa de Cronbach $\alpha = .91$.

Tabla 1. *Estadísticos descriptivos de los ítems (N = 298).*

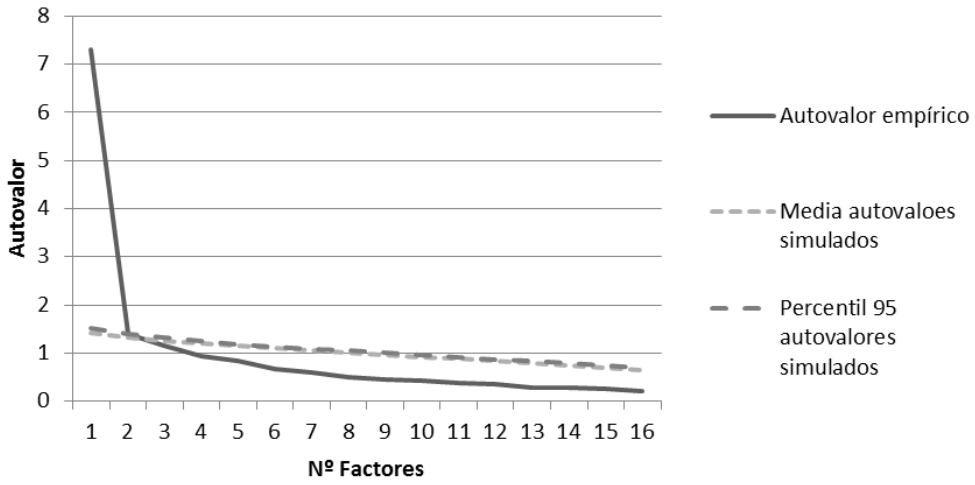
ar	M	DT	Min	Max	Asimetría		Kurtosis		*r* item-test
ar1	3.82	1.06	1	5	-0.69	(.06)	-0.16	(.06)	.66

ar	M	DT	Min	Max	Asimetría		Kurtosis		*r* item-test
ar2	3.70	1.08	1	5	-0.51	(.06)	-0.39	(.06)	.65
ar3	3.81	1.07	1	5	-0.68	(.06)	-0.14	(.06)	.59
ar4	3.37	1.22	1	5	-0.38	(.07)	-0.74	(.07)	.55
ar5	3.42	1.10	1	5	-0.39	(.06)	-0.50	(.06)	.68
ar6	3.19	1.24	1	5	-0.33	(.07)	-0.87	(.07)	.62
ar7	3.59	1.15	1	5	-0.57	(.07)	-0.39	(.07)	.61
ar8	4.08	1.01	1	5	-0.93	(.06)	0.05	(.06)	.58
ar9	3.48	1.19	1	5	-0.44	(.07)	-0.63	(.07)	.50
ar10	3.32	1.15	1	5	-0.37	(.07)	-0.56	(.07)	.49
ar11	2.71	1.20	1	5	0.07	(.07)	-0.89	(.07)	.48
ar12	3.49	1.20	1	5	-0.53	(.07)	-0.50	(.07)	.51
ar13	3.24	1.03	1	5	-0.17	(.06)	-0.40	(.06)	.69
ar14	3.09	1.01	1	5	-0.05	(.06)	-0.33	(.06)	.70
ar15	3.01	1.07	1	5	-0.06	(.06)	-0.49	(.06)	.59
ar16	3.32	1.15	1	5	-0.33	(.07)	-0.61	(.07)	.56

Nota: ar: ítem de la escala de arraigo. M: media; DT: desviación típica; Min: mínimo; *Max*: máximo; *Asimetría* y *Kurtosis*: entre paréntesis se muestra el error estándar de estos estadísticos*;* $r_{item\text{-}test}$: correlación entre el ítem y el resto del test.

Análisis factorial exploratorio. Primeramente, se llevó a cabo un análisis paralelo para determinar el número mínimo de dimensiones latentes necesarias (p.ej., factores). Los resultados de este análisis mostraron que los autovalores de las soluciones factoriales de uno y dos factores eran superiores a los autovalores esperados en una matriz de datos generados aleatoriamente a través de una matriz de correlaciones entre los ítems similares (ver figura 1). Según este análisis convendría extraer dos factores.

Figura 1. Análisis paralelo.



Torrente y cols. (2011) encontraron un modelo de tres factores en la versión original de la escala, por ello, se decidió estimar dos modelos factoriales y comparar su ajuste. El primero era un modelo de dos factores de acuerdo a los resultados del análisis paralelo y el segundo un modelo de tres factores. Tal y como se muestra en la Tabla 2, ambos modelos mostraron un buen ajuste relativo, con valores del *CFI* y el *TLI* superiores a .90. Asimismo, los valores del *SRMR* fueron aceptables para los dos modelos. No obstante, los valores del *RMSEA* resultaron inadecuados para ambos modelos, especialmente para el modelo de dos factores, cuyo intervalo de confianza ni siquiera alcanza el punto de corte de ajuste moderado (p.ej., *RMSEA* $\leq$ .08). El intervalo de confianza del *RMSEA* del modelo de tres factores, por su parte, incluye valores de ajuste moderado o, incluso, valores muy próximos al punto de corte de buen ajuste. Por este motivo se decidió retener la solución factorial de tres factores para la escala.

Tabla 2. Índices de ajuste del análisis factorial exploratorio.

	χ2	gl	CCFI	TTLI	RMSEA	RMSEA05	RMSEA95	SRMR
Modelo 2 Factores	441.11	89	0.93	0.90	0.115	0.105	0.126	0.063
Modelo 3 Factores	247.84	75	0.96	0.93	0.088	0.062	0.100	0.047

Nota: *CFI*: comparative fir index; *TLI*: Tucker-Lewis index; *RMSEA*: Root Mean Squared Error of Approximation; *SRMR*: Standardized Root Mean Residual.

La estructura factorial del modelo de tres factores resultó similar a la obtenida por Torrente y cols. (2011) en la validación de la escala original (ver Tabla 3), donde también se obtuvieron tres factores: *Arraigo Ecológico*, *Arraigo Cultural* y *Arraigo Social*. La mayoría de los ítems mostraron saturaciones factoriales muy elevadas (superiores a .50) en un único factor, aunque algunos de los ítems del factor Arraigo Ecológico y Arraigo Social presentaron saturaciones cruzadas en más de un factor (superiores a .30). Lo anterior indica que algunos de los ítems no son unidimensionales, aspecto que debe considerarse a la hora de calcular las puntuaciones de los participantes en los diferentes factores de arraigo.

Tabla 3. Modelo de tres factores

	Factores		
	Ecológico	Cultural	Social
ar1	**.86**	.00	.01
ar2	**.92**	-.12	.08
ar3	**.60**	.08	.14
ar4	**.66**	.23	-.21
ar5	**.54**	**.39**	-.09
ar6	**.49**	.16	.19
ar7	**.37**	.11	**.44**
ar8	**.33**	**.34**	.16
ar9	.15	.11	**.52**
ar10	.02	.20	**.55**
ar11	.03	.23	**.50**
ar12	**.31**	.04	**.44**
ar13	-.03	**.78**	.21
ar14	.03	**.75**	.18
ar15	-.03	**.84**	-.04
ar16	.16	**.70**	-.16
Correlaciones:			
Ecológico	α = .87	.57	.36
Cultural		α = .84	.43
Social			α = .75

Nota: ar: ítem de la escala de arraigo; α de Cronbach en la diagonal de la matríz de correlaciones.

Los tres factores correlacionaron positivamente, de manera que aquellos participantes con mayores puntuaciones en uno de los factores de arraigo, tienden a presentar puntuaciones de arraigo elevadas en los otros factores. La consistencia interna de los ítems que componen cada factor también resultó adecuada, obteniendo valores del α de Cronbach superiores a .80 para los factores de arraigo ecológico y cultural y de .75 para el factor de arraigo social.

Validez en relación a otras variables. Una vez determinada la estructura factorial de la escala de arraigo, se correlacionaron sus factores con la variable de contacto sostenido con la población española y la escala de emociones (Tabla 4). Se encontró una estrecha relación entre los factores de arraigo y las emociones positivas y el contacto con los españoles, de manera que aquellas personas con mayores puntuaciones en la escala de emociones positivas percibidas y de contacto con los españoles, mostraron a su vez mayores niveles en los tres factores de arraigo. Por el contrario, se encontró una relación inversa entre los tres factores de arraigo y las emociones negativas, de modo que aquellas personas con menores niveles de arraigo tienden a presentar mayores puntuaciones en la escala de emociones negativas percibidas.

Tabla 4. Correlaciones entre los factores de la escala y las variables de validez convergente

	Emociones Positivas	Emociones Negativas	Contacto
Ecológico	.44*	-.28*	.37*
Social	.37*	-.20*	.31*
Cultural	.42*	-.24*	.36*

Nota: *p < .001

Escala Breve de Arraigo (EBA). En consonancia a los resultados descritos y para articular la Escala Breve de Arraigo (EBA), se consideraron los factores teóricos a los que pertenece cada ítem de la versión original y los factores empíricos obtenidos en los que saturaron en el presente estudio. En segundo lugar, se consideró la saturación factorial más alta de cada ítem considerando que esta indica la relación de cada ítem con su factor. Así, aquellos

con una saturación factorial por debajo de .30 se clasificaron como ítems con una relación baja con su factor, aquellos con saturaciones factoriales entre .30 y .60 con una relación media, y aquellos con saturaciones factoriales iguales o superiores a .60 con una relación alta. En tercer lugar, se tuvo en cuenta si los ítems presentaban saturaciones cruzadas. Por último, un panel de expertos evaluó la relevancia de cada ítem para medir el arraigo ecológico, social y cultural de los inmigrantes en una escala Likert de 5 puntos. Aquellos ítems con puntuaciones medias por debajo de la categoría 4 (p.ej., *relevante para medir el constructo*), fueron considerados como "no relevantes" (ver Tabla 5).

Tabla 5. Criteruis oara reducir la Escala de Arraigo

	Factor Teórico (Torrente y cols., 2011)	Factor Empírico (Estudio 1)	Saturación factorial	Saturación cruzada	Evaluación Expertos
ar1	Ecológico	Ecológico	Alta		
ar2	Ecológico	Ecológico	Alta		
ar3	Ecológico	Ecológico	Alta		
ar4	Ecológico	Ecológico	Alta		
ar5	Cultural	Ecológico	Media	Sí	
ar6	Social	Ecológico	Media		No relevante
ar7	Social	Social	Media	Sí	
ar8	Social	Cultural	Media	Sí	
ar9	Cultural	Social	Media		No relevante
ar10	Cultural	Social	Media		No relevante
ar11	Cultural	Social	Media		No relevante
ar12	Social	Ecológico	Media	Sí	
ar13	Cultural	Cultural	Alta		
ar14	Cultural	Cultural	Alta		
ar15	Cultural	Cultural	Alta		
ar16	Cultural	Cultural	Alta		

Nota: ar: ítem de la escala de arraigo

Para componer la versión breve de la escala se seleccionaron aquellos ítems que: (1) saturaban en el mismo factor en el estudio de validación de la escala original y en el presente estudio; (2) presentaban saturaciones factoriales medias o altas en su factor; (3) no presentaban saturaciones factoriales cruzadas; y (4) no fueron clasificados como *no relevantes* por el panel

de expertos. Por lo tanto, para componer la escala breve se seleccionaron los cuatro primeros y los cuatro últimos ítems de la escala original, prescindiendo de los ítems de arraigo social (ver Tabla 6).

Tabla 6. Ïtems finales de la Escala Breve de Arraigo (EBA)

Ítem	Dimensión
1.La ciudad o pueblo en el que vive	Arraigo Ecológico
2.El barrio en el que vive	
3.La casa o piso en el que vive	
4.Las fiestas locales	
5.La forma de comportarse de los españoles	Arraigo Cultural
6.La forma de pensar de los españoles	
7. La forma de hablar y expresar sentimientos de los españoles	
8.La forma de divertirse de los españoles	

Esta versión breve de la Escala de Arraigo cuenta con una buena consistencia interna (α de Cronbach = .87). Además, los ítems de arraigo ecológico de la escala breve mostraron una fuerte correlación con el factor de arraigo ecológico de la escala original, $r = .95$ ($p < .001$); al igual que los ítems de arraigo cultural con el factor de arraigo cultural, $r = .97$, ($p < .001$). Esto implica que las personas con mayores niveles en los factores de arraigo ecológico y cultural de la escala original, obtienen también mayores puntuaciones en estos mismos factores de la versión reducida de la escala

Fase 2. Análisis descriptivo de los ítems y fiabilidad. Todos los ítems mostraron una media centrada en torno a 3.5 con una desviación típica cercana a uno (ver Tabla 7). Esto indica que la mayoría de los participantes se decantaron por las categorías centrales, con una leve inclinación hacia las categorías superiores (p.ej., *Bastante, Mucho*). Esto se ve reflejado en los estadísticos de asimetría, ligeramente negativos. Las correlaciones entre los ítems y el resto del test fueron altas, con valores entre .50 y -.60, indicando una fuerte relación entre cada ítem y el resto de la escala. La consistencia interna fue elevada, con un α de Cronbach = .84 y un ω de

McDonald = .89 para el total de la escala (con α = .81 y .82, y ω = .84 y ω = .84 para las subescalas de arraigo ecológico y arraigo cultural, respectivamente).

Tabla 7. Estadísticos descriptivos de los ítems (N = = 395)

	M	*DT*	*Min*	*Max*	*Asimetría*	*Kurtosis*	*r item-test*
ar1	3.65	1.01	1	5	-0.51 (.05)	-0.20 (.05)	0.67
ar2	3.50	1.05	1	5	-0.25 (.05)	-0.69 (.05)	0.54
ar3	3.75	0.97	1	5	-0.60 (.05)	0.04 (.05)	0.60
ar4	3.27	1.13	1	5	-0.20 (.06)	-0.70 (.06)	0.61
ar13	3.07	0.95	1	5	-0.10 (.05)	-0.06 (.05)	0.61
ar14	3.07	0.93	1	5	-0.08 (.05)	0.02 (.05)	0.67
ar15	3.01	1.05	1	6	0.02 (.05)	-0.37 (.05)	0.61
ar16	3.21	1.13	1	5	-0.20 (.06)	-0.67 (.06)	0.59

Nota: ar: ítem de la escala de arraigo. M: media; DT: desviación típica; Min: mínimo; *Max*: máximo; *Asimetría* y *Kurtosis*: entre paréntesis se muestra el error estándar de estos estadísticos*;* $r_{item\text{-}test}$: correlación entre el ítem y el resto del test.

Análisis factorial confirmatorio. Para determinar la estructura factorial de la escala de arraigo reducida se pusieron a prueba tres modelos: un modelo unifactorial, un modelo de dos factores correlacionados diferenciando entre arraigo ecológico y arraigo cultural, y un modelo bifactorial, con un factor general de arraigo y dos factores específicos de arraigo ecológico y arraigo cultural. Los tres modelos convergieron satisfactoriamente.

Con respecto al ajuste de los modelos (ver Tabla 8), los índices de ajuste relativo (p.ej., *CFI* y *TLI*) mostraron un ajuste adecuado para todos ellos, con valores por encima de .95, con la excepción del *TLI* para el modelo unifactorial, que se queda cerca del punto de corte propuesto por Hu y Bentler (1999). Sin embargo, los valores del *RMSEA* de los modelos de uno y dos factores resultaron bastante inadecuados, con valores muy encima de un ajuste mediocre. El único modelo con valores del *RMSEA* adecuados fue el modelo bifactorial. Por este motivo, decidimos quedarnos con la solución bifactorial como la estructura latente de la escala, ya que éste fue el único modelo que mostró un buen ajuste en los tres índices considerados.

Tabla 8. Índices de ajuste del AFC

Model	$\chi 2$	*gl*	CFI	TLI	RMSEA [95% CI]	
Unifactorial	298.08	20	0.959	0.942	0.188	[.169 - .207]
Dos factores	136.99	19	0.982	0.974	0.126	[.106 - .146]
Bifactorial	28.04	12	0.998	0.994	0.058	[.030 - .087]

Nota: *CFI*: comparative fir index; *TLI*: Tucker-Lewis index; *RMSEA*: Root Mean Squared Error of Approximation; *SRMR*: Standardized Root Mean Residual

Todos los ítems mostraron saturaciones elevadas en el factor de arraigo general, con valores superiores a .55 en todos ellos (ver Tabla 9). Por su parte, las saturaciones de los ítems en los factores específicos fueron más moderadas, con uno de los ítems presentando una saturación no significativa en el factor de arraigo ecológico (p.ej., "las fiestas locales").

Tabla 9. Saturaciones factoriales del modelo bifactorial

	Arraigo General	Arraigo Ecológico	Arraigo Cultural
ar1	.73 (.04)	.44 (.07)	
ar2	.56 (.06)	.57 (.07)	
ar3	.62 (.06)	.62 (.07)	
ar4	.81 (.05)	-.09 (.13)	
ar13	.60 (.05)		.54 (.06)
ar14	.68 (.05)		.44 (.08)
ar15	.57 (.05)		.58 (.06)
ar16	.66 (.05)		.27 (.08)

Nota: ar: ítem de la escala de arraigo. Entre paréntesis se muestra el error de estimación de las saturaciones factriales estandarizadas.

Este modelo asume que los elementos comunes entre arraigo ecológico y arraigo general quedan absorbidos por un factor general común, quedando los elementos no comunes de ambos tipos de arraigo en dos factores específicos bien diferenciados. De la varianza común explicada por el modelo, el factor de arraigo general es responsable del 82.64%, mientras que el factor específico de arraigo cultural explica el 10.24% y el factor de arraigo ecológico el 7.11% restante. De los cuatro ítems de arraigo cultural, el factor de arraigo general explica el 65.09% de la variabilidad en las respuestas de los participantes, mientras que, de los cuatro ítems de arraigo ecológico, el factor de arraigo general es responsable del 75.80% de la varianza.

Evidencias de validez basadas en relación a otras variables. Se utilizaron las puntuaciones factoriales del modelo bifactor de arraigo para poner a prueba su validez en relación a otras variables. Para esto se utilizó la escala de emociones positivas y negativas, encontrando al igual que en el modelo original de una manera estadísticamente significativa, que a mayor nivel de arraigo, mayores serán las emociones positivas percibidas y menores serán las emociones negativas percibidas (ver Tabla 10).

Tabla 10. Correlaciones entre la Escala Breve de Arraigo y otras variables

	Arraigo General	Arraigo Ecológico	Arraigo Cultural
Emociones positivas	0.40***	0.11*	0.18***
Emociones negativas	-0.22***	-0.18***	-0.08

Nota: *: $p < .05$; *** $p < .01$

Discusión

El objetivo del presente trabajo fue evaluar la estructura factorial y las evidencias de validez de la escala de arraigo en inmigrantes latinoamericanos en España. Los hallazgos permiten aseverar que la versión revisada se ajusta mejor a la propuesta teórica del arraigo de inmigrantes con dos dimensiones.

Para proveer evidencia sobre la validez, se examinó la estructura factorial de la escala con medidas exploratorias y confirmatorias a través de dos fases con muestras diferentes. Los resultados demostraron que había ítems con cargas cruzadas que distorsionaban la escala, por lo que se siguieron las recomendaciones de Goetz et al. (2013) para revisar el instrumento. El producto de este proceso sugirió la reducción de los ítems, por lo que se propuso una versión revisada de la escala que se denominó Escala Breve de Arraigo (EBA). La EBA está compuesta de 8 ítems que muestran dos factores claramente identificables: Arraigo Ecológico y Arraigo Cultural. Al mismo tiempo se examinó la validez de la escala en relación a otras variables que teóricamente están asociadas con el arraigo.

Los resultados permiten aseverar que la EBA presenta una estructura bifacotiral estable de dos dimensiones. La primera dimensión de la EBA es el arraigo ecológico, el que da cuenta de los sentimientos de apego al lugar como un espacio físico (Torrente et al., 2011), en el que la persona se vincula con la sociedad estructurando y reestructurando los aspectos simbólicos en la sociedad de acogida. La segunda dimensión es el arraigo cultural, donde se enmarca la forma en que el lugar influye en el desarrollo de conductas y hábitos que originan sentimientos de apego (Torrente et al., 2011). Estas dimensiones son consistentes con la amplia evidencia que plantea al arraigo como una estructura de dos factores (Bricker y Kerstetter, 2000; Brown y Raymond, 2007; Dvorak et al., 2013; Hammitt et al., 2006; Jorgensen y Stedman, 2006; Kyle et al., 2005; Pretyy et al., 2003; Raymond et al., 2010; Riger y Lavrakas, 1981; Taylor, Gottfredson, y Browser, 1985; Williams y Vaske, 2003).

Si bien los resultados son prometedores, es necesario plantear algunas limitaciones que tuvo el estudio. Lo primero es su naturaleza transversal, por lo que no se puede asumir asociaciones de causalidad entre las variables estudiadas. Otro aspecto es la muestra no probabilística, por lo que los resultados deben ser analizados con cautela. Pese a lo anterior, es pertinente mencionar que, dada la dificultad de identificar a todas las unidades de análisis de esta población, conseguir una muestra probabilística con estas características es una tarea sumamente complicada, si no prácticamente imposible (D'Ancona, 2005, 2009).

Para futuras investigaciones se sugiere cumplir con el último paso propuesto por Goetz et al. (2013) y validar la EBA con una nueva muestra. Continuar con el estudio del arraigo y su relación con otras variables psicosociales es importante en el actual contexto español, debido a que el aislamiento y la exclusión derivada del desarraigo de esta población incidiría en enfermedades como ansiedad y depresión (García, et al., 2009). En contrapartida, abordar este fenómeno permitiría trabajar con el desarrollo de un sentido de pertenencia que se evidenciaría

en medidas del bienestar psicológico como la identidad y la autoestima (Lewikca, 2005; Scannell y Gifford, 2010).

El arraigo en el ámbito de las migraciones debe considerarse en el contexto donde se desarrollan los vínculos con el territorio, por el hecho de que incluye variables derivadas de la interacción social con la población autóctona, como las estrategias de aculturación, el prejuicio percibido, entre otras. Proseguir con esta línea de investigación será provechoso en la indagación de procesos emocionales y conductuales que podrían ser mediados o moderados por el arraigo.

Referencias

Acebo, E. 1996. Sociología del arraigo. Una lectura crítica de la teoría de la ciudad. Argentina: Editorial Claridad.

Asparouhov, T. y Muthén, B. 2010. Weighted least squares estimation with missing data. Disponible en: http://www.statmodel.com/download/GstrucMissingRevision.pdf

Barcus, H. y Brunn, S. 2010. Place elasticity: Exploring a new conceptualization of mobility and place attachment in rural America. Geografiska Annaler: Series B, Human Geography, *92*(4), 281-295. https://doi.org/10.1111/j.1468-0467.2010.00353.x

Bayona, J. Pujadas, I. y Avila, R. 2018. Europa como nuevo destino de las migraciones latinoamericanas y caribeñas. Revista Bibliográfica de Geografía y Ciencias Sociales. Vol. XXIII, nº 1.242. Barcelona: Universidad de Barcelona,

Berríos, J. 2017. Construcción de una escala de prejuicio percibido por inmigrantes latinoamericanos en España. Tesis Doctoral. Valencia: Universidad de Valencia.

Bricker, K., y Kerstetter, D. 2000. Level of specialization and place attachment: An exploratory study of whitewater recreationists. Leisure Sciences. 22(4), 233–257. https://doi.org/10.1080/01490409950202285

Brown, G. y Raymond, C. 2007. The relationship between place attachment and landscape values: Toward mapping place attachment. Applied Geography. 27(2), 89-111. https://doi.org/10.1016/j.apgeog.2006.11.002

Carrera, S. 2006. Programas de integración para inmigrantes: una perspectiva comparada en la Unión Europea. Migraciones, (20), 37-73. https://revistas.comillas.edu/index.php/revistamigraciones/article/view/2910

D'Ancona, M. 2005. La exteriorización de la xenofobia (The Exteriorization of Xenophobia). Reis, (112), 197-230. https://doi.org/10.2307/40184716

D'Ancona, M. 2009. Filias y fobias ante la imagen poliédrica cambiante de la inmigración: claves en la comprensión del racismo y la xenofobia. Revista del Ministerio de Trabajo e Inmigración, (80), 39-60. http://www.caritasvitoria.org/datos/documentos/filiasyfobias.pdf

Dvorak, R., Borrie, W., y Watson, A. 2013. Personal wilderness relationships: building on a transactional approach. Environmental Management, *52*(6), 1518-1532. https://doi.org/10.1007/s00267-013-0185-7

Franch, X., Domingo, A., y Sabater, A. 2011. Perspectiva municipal del arraigo en la provincia de Barcelona, 2006-2009. Documents d' Anàlisi Geogràfica, *57*(3), 517-547. https://doi.org/10.5565/rev/dag.252

Frías-Navarro, D. 2015. Escala de Creencias sobre las Emociones que el Exogrupo percibe del propio Endogrupo de Inmigrantes *(*Emociones Auto-percibidas, EMOAUTO). Universidad de Valencia. España.

García, A., Jiménez, B., y Redondo, A. 2009. La inmigración latinoamericana en España en el siglo XXI. Investigaciones geográficas, (70), 55-70. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-46112009000300004&lng=es&nrm=iso>

Garrido, L., Abad, F., y Ponsoda, V. 2013. A new look at Horn's parallel analysis with ordinal variables. Psychological Methods, 18(4), 454. https://doi.org/10.1037/a0030005

Goetz, C., Coste, J., Lemetayer, F., Rat, A., Montel, S., Recchia, S., Debouverie, M., Pouchot, J., Spitz, E., y Guillemin, F. 2013. Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. Journal of Clinical Epidemiology, *66*(7), 710-718. https://doi.org/10.1016/j.jclinepi.2012.12.015

Gustafson, P. 2001. Roots and routes exploring the Relationship between place attachment and mobility. Environment and behavior, *33*(5), 667-686. https://doi.org/10.1177/00139160121973188

Hammitt, W., Backlund E. y Bixler, R. 2006. Place Bonding for Recreation Places: Conceptual and Empirical Development. Leisure Studies, 25(1), 17-41. https://doi.org/10.1080/02614360500098100

Hidalgo, M. y Hernández, B. 2001. Place attachment: Conceptual and empirical questions. Journal of Environmental Psychology, 21(3), 273-281. https://doi.org/10.1006/jevp.2001.0221

Horn, J. 1965. A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179-185. https://doi.org/10.1007/BF02289447

Hu, L. y Bentler, P. 1998. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, *3*(4), 424. https://doi.org/10.1037//1082-989X.3.4.424

Hu, L. y Bentler, P. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Instituto Nacional de Estadística. 2019. Estadística del padrón continuo a 10 de Junio de 2020. Recuperado desde Instituto Nacional de Estadística Disponible en: https://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p04/provi/l0/&file=0ccaa002.px&L=0

Naciones Unidas. 2019. International Migrant Stock 2019. DAES de las Naciones Unidas, División de Población. Nueva York. Disponible en https://www.un.org/en/development/desa/population/migration/data/estimates2/estimatesmaps.asp?0t0

Jennings, B. y Krannich, R. 2013. A multidimensional exploration of the foundations of community attachment among seasonal and year-round residents. Rural Sociology, *78*(4), 498-527. https://doi.org/10.1111/ruso.12019

Jorgensen, B., y Stedman, R. 2006. A comparative analysis of predictors of sense of place dimensions: Attachment to, dependence on, and identification with lakeshore properties. Journal of environmental management, 79(3), 316-327. https://doi.org/10.1016/j.jenvman.2005.08.003

Kaiser, H. 1960. The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141–151. https://doi.org/10.1177%2F001316446002000116

Kyle, G., Mowen, A., y Tarrant, M. 2004. Linking place preferences with place meaning: An examination of the relationship between place motivation and place attachment. Journal of Environmental Psychology, *24*, 439-454. https://doi.org/10.1016/j.jenvp.2004.11.001

Kyle, G. Graefe, A. y Manning, R. 2005. Testing the Dimensionality of Place Attachment in Recreational Settings." Environment and Behavior 37(2), 153–77. https://doi.org/10.1177/0013916504269654

Lewicka, M. 2005. Ways to make people active: The role of place attachment, cultural capital, and neighborhood ties. Journal of environmental psychology, 25(4), 381-395. https://doi.org/10.1016/j.jenvp.2005.10.004

López-Mosquera, N. y Sánchez, M. 2013. Direct and indirect effects of received benefits and place attachment in willingness to pay and loyalty in suburban natural areas. Journal of Environmental Psychology, *34*, 27-35. https://doi.org/10.1016/j.jenvp.2012.11.004

Maccallum, R., Browne, M., y Sugawara, H. 1996. Power analysis and determination of sample size for covariance structure modeling. Psychological Methods, *1*, 130-149. https://doi.org/10.1037//1082-989X.1.2.130

Milligan, M. 1998. Interactional past and potential: The social construction of place attachment. Symbolic interaction, *21*(1), 1-33. https://doi.org/10.1525/si.1998.21.1.1

Morgan, P. 2010. Towards a developmental theory of place attachment. Journal of Environmental Psychology, *30*(1), 11-22. https://doi.org/10.1016/j.jenvp.2009.07.001

Muthén, L. y Muthén, B. 2010. Mplus user's guide. Los Angeles, CA: Muthén & Muthén.

Ng, C. 1998. Canada as a new place: the immigrant's experience. Journal of Environmental Psychology, *18*(1), 55-67. https://doi.org/10.1006/jevp.1997.0065

Pretty, G., Chipuer, H., y Bramston, P. 2003. Sense of place amongst adolescents and adults in two rural Australian towns: The discriminating features of place attachment, sense of community and place dependence in relation to place identity. Journal of environmental psychology, 23(3), 273-287. https://doi.org/10.1016/S0272-4944(02)00079-8

Quezada, M. 2007. Migración, arraigo y apropiación del espacio en la recomposición de identidades socioterritoriales. Cultura y representaciones Sociales, *2*(3), 35-67. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-81102007000200003&lng=es&tlng=es.

R Core Team. 2016. R: A language and environment for statistical computing. R foundation for Statistical Computing. https://www.R-project.org/

Raymond, C., Brown, G., y Weber, D. 2010. The measurement of place attachment: Personal, community, and environmental connections. Journal of Environmental Psychology, *30*(4), 422-434. https://doi.org/10.1016/j.jenvp.2010.08.002

Riger, S., y Lavrakas, P. 1981. Community ties: Patterns of attachment and social interaction in urban neighborhoods. American Journal of Community Psycholgy 9, 55–66 https://doi.org/10.1007/BF00896360

Rinken, S. 2006. ¿Vivir transnacional? Envío de remesas versus arraigo en la sociedad de acogida: el caso de Andalucía. Migraciones, *20*, 173-199. https://revistas.comillas.edu/index.php/revistamigraciones/article/view/2914

Sampedro, R. y Camarero, L. 2016. Inmigrantes, estrategias familiares y arraigo: las lecciones de la crisis en las áreas rurales. Revista Migraciones. Num. 40 Pags. 3-31. https://doi.org/1014422/mig.i40y2016.008

Scannell, L. y Gifford, R. 2010. Defining place attachment: A tripartite organizing framework. Journal of Environmental Psychology, *30*(1), 1-10. https://doi.org/10.1016/j.jenvp.2009.09.006

Schmitt, T. 2011. Current methodological considerations in exploratory and confirmatory factor analysis. Journal of Psychoeducational Assessment, *29*(4), 304-321. https://doi.org/10,1177/0734282911406653

Sochos, A. y Diniz, M. 2012. The role of attachment in immigrant sociocultural adaptation and psychological distress. Journal of Community & Applied Social Psychology, *22*(1), 75-91. https://doi.org/10.1002/casp.1102

Taylor, R., Gottfredson, S. y Brower, S. 1985. Attachment to place: Discriminant validity, and impacts of disorder and diversity. American Journal Community Psychology 13, 525–542. https://doi.org/10.1007/BF00923265

Torrente, G., Ruiz-Hernández, J., Ramírez, M., y Rodríguez, A. 2011. Construcción de una escala para medir el arraigo en inmigrantes latinoamericanos. Anales de Psicología, *27*(3), 843-851. http://www.redalyc.org/articulo.oa?id=16720048032

Vidal, T. y Pol, E. 2005. La apropiación del espacio: una propuesta teórica para comprender la vinculación entre las personas y los lugares. Anuario de Psicología, *36*(3), 281-298.

Williams, D. y Vaske, J. 2003. The Measurement of Place Attachment: Validity and Generalizability of a Psychometric Approach. Forest Science, 49(6), 830–840, https://doi.org/10.1093/forestscience/49.6.830

BLOQUE 5. ANEXOS

Anexo 1. Breve explicación de conceptos fundamentales de diseño de investigación

Diseño de entre-grupos o entre-sujetos

Diseño de entre-grupos o entre-sujetos (between subjects or between groups design)

Se trata de un diseño de investigación donde el investigador:

1) asigna un determinado grupo o condición que tiene la variable independiente manipulada o factor (por ejemplo, el tratamiento nuevo o el tratamiento tradicional) a los participantes del estudio. Si la asignación de la condición es aleatoria se trata de un estudio con una metodología experimental y si no es aleatoria se trata de una metodología cuasi-experimental, o bien

2) selecciona a los participantes de determinados grupos independientes o diferentes que ya están formados ('grupos intactos') en función de sus características o propiedades (por ejemplo, vivir en zona rural o urbana, ser hombre o ser mujer…). Se trata de un diseño entre-grupos realizado con una metodología no experimental.

Diseño factorial

Diseño factorial (factorial design)

El diseño factorial permite analizar el efecto simultáneo de dos o más variables independientes (dos o más factores) sobre una variable dependiente medida (diseño univariado) o sobre un conjunto de variables dependientes (diseño multivariado). Por lo tanto, permite estudiar el efecto de interacción de dos o más variables independientes. Para definir un diseño factorial se numeran los factores que se analizan y el número de condiciones experimentales ('celdillas de interacción') que configuran cada uno de estos factores. Por ejemplo, el diseño factorial más simple es el 2 *x* 2, es decir, el compuesto por dos *variables independientes* o factores con dos *niveles* cada una de ellas. Si se analizan dos factores, A con tres niveles o condiciones y B con cuatro, se tratará de un diseño factorial 3 *x* 4. Cuando todos los factores tienen el mismo número de niveles suelen codificarse mediante una potencia: 2^3 es un diseño factorial de tres factores con dos niveles cada uno de ellos (representa lo mismo que 2 *x* 2 *x* 2).

Diseño de medidas repetidas o diseño intra-sujetos

Diseño de medidas repetidas o diseño intra-sujetos (repeated measures designo or within-subjects design)

Se trata de un diseño de investigación que implica que todos los participantes son medidos en los diferentes niveles o condiciones del factor (también conocido como variable independiente).

En ocasiones, estos diseños suponen que los participantes se miden en diferentes momentos temporales (por ejemplo, pre-test y post-test e incluso seguimiento) o pueden suponer que los participantes reciben los diferentes tratamientos que se prueban en el estudio (por ejemplo, reciben el fármaco denominado Z y también el fármaco denominado W y sus efectos se miden en una determinada variable).

Diseño mixto o de medidas parcialmente repetidas

Diseño mixto o de medidas parcialmente repetidas (mixed design or design with both between and within-subjects factors)

Se trata de un diseño que al menos tiene un factor entre-grupos (factor A, por ejemplo) y un factor de medidas repetidas (factor B, por ejemplo).

En la Imagen 7 se representan las ecuaciones de un diseño entre-sujetos (parte izquierda de la imagen) con su correspondiente término de error representado por S / A dado que los sujetos están anidado a ligados a una única condición de la variable independiente o factor.

En la parte derecha se representa la ecuación estructural de un diseño de medidas repetidas siendo el término de error S x A ya que todos los sujetos pasan o se cruzan (obtienen datos) por todas las condiciones o grupos de la variable independiente, de ahí que el sujeto 1 se encuentre en A_1 (S_1 x A_1) y también en A_2 (S_1 x A_2).

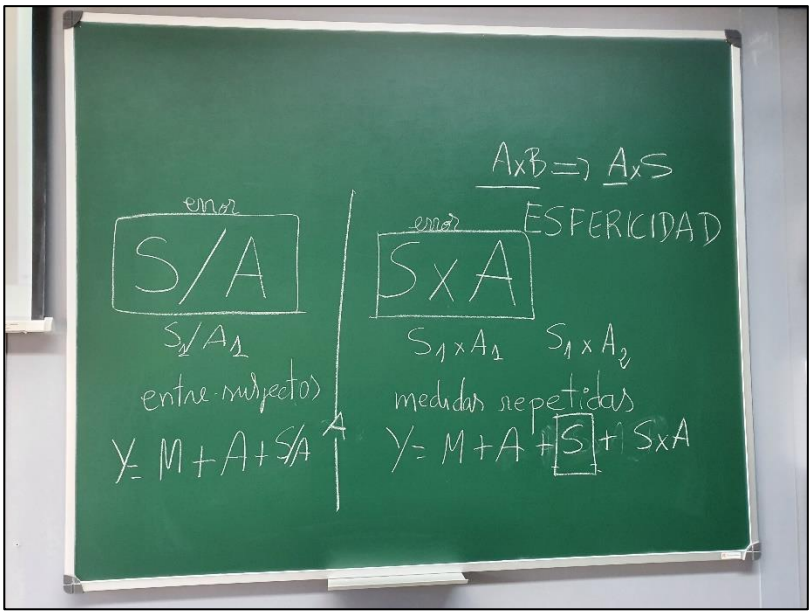


Imagen 7. Términos de error S / A y S x A

El diseño mixto A *x* B tiene dos términos de error:

1) un error formado por los sujetos (S) anidados en una condición del factor entre-grupos A (Error = S / A) para calcular el denominador de la razón *F* de las fuentes de varianza denominadas *entre-grupos* (factor A entre-grupos). Sería el término de error que se utiliza para calcular la razón *F* del factor A.

2) un error formado por la interacción entre el factor B de medidas repetidas y el factor S de Sujetos que a su vez está anidado en el factor entre-sujetos A (Error = B *x* S / A) para las fuentes de varianza de medidas repetidas y para las interacciones que la variable de medidas repetidas tenga con los factores entre-grupos.). Sería el término de error que se utiliza para calcular la razón *F* del factor B y de la interacción A *x* B.

Diseño de bloques

Diseño de bloques (block design)

Se trata de un diseño con al menos una variable bloqueda (variable categorizada en condiciones). Cuando se estudia la relación entre una variable independiente y la variable dependiente es frecuente que aparezcan otras variables o factores que también influyen y que deben ser controladas en el diseño de investigación. A estas variables se las denomina ‘variables bloque’ o ‘variables de bloqueo’, y se

caracterizan por 1) no ser el motivo del estudio (es decir no están formuladas en la hipótesis de investigación, no son variables explicativas) sino que forman parte del estudio dado que tienen un efecto estadísticamente significativo sobre las puntuaciones de la variable dependiente y 2) porque se asume que no tienen interacción con el factor o variable independiente del estudio que forma parte de la hipótesis de investigación); es decir, se debe cumplir el supuesto de que el efecto de la interacción entre la variable independiente de la hipótesis y la variable de boqueo no es estadísticamente significativo.

El modelo es de “bloques aleatorizados completos” cuando en cada bloque se presentan todos los posibles tratamientos del factor objeto de estudio y, además, dentro de cada bloque se asignan los tratamientos de forma aleatoria. Podría ser de bloques no aleatorizados si no hay asignación aleatoria del tratamiento a cada bloque. En ocasiones también puede ocurrir que no se pueden asignar todos los tratamientos sobre cada bloque, de modo que se tienen los diseños por bloques aleatorizados (o no aleatorizados) incompletos.

En el diseño completamente aleatorizado se asignan los tratamientos al azar a los grupos, sin restricción alguna, mientras que en el diseño en bloques aleatorizados primero se forman niveles o condiciones de los bloques (las denominadas parcelas) y a continuación se asignan los tratamientos (las condiciones) a las parcelas de cada bloque. Por lo tanto, se trata de un diseño con aleatorización restringida, ya que las unidades experimentales (sujetos) son primero clasificadas en grupos homogéneos (bloques), y después los tratamientos o condiciones son asignados aleatoriamente dentro de los bloques.

Diseño con variables covariadas, ANCOVA

Diseño con variables covariadas, ANCOVA (Analysis of Covariance, ANCOVA)

Se trata de un diseño que al menos tiene una variable covariada (variable medida cuantitativa continua) que se introduce en el diseño como una fuente de varianza más, pero dicha variable no forma parte de la hipótesis de investigación. La variable covariada se puede utilizar para controlar una variable extraña, para reducir el error o para producir ambas cosas.

Se conoce como ANCOVA cuando el diseño solo tiene una variable dependiente y se trata de un MANCOVA cuando el diseño tiene más de una variable covariada. Se trata de un diseño análogo al diseño de bloques y la diferencia está en que en el ANCOVA los participantes se aparean matemáticamente en lugar de hacerlo físicamente formando bloques con participantes que tienen características similares en la variable de bloqueo (diseño de bloques) (Huitema, 1980; Reichardt, 1979).

En el ANCOVA se realiza un ajuste matemático de la variabilidad observada en los datos del estudio a través de las puntuaciones que se obtienen en la variable continua covariada.

Diseño de grupo equivalente o diseño con grupo de control equivalente

Diseño de grupo equivalente o diseño con grupo de control equivalente (equivalente group design)

Se trata de un diseño elaborado con una metodología experimental. Este diseño permite comparar grupos de participantes o unidades experimentales que son asignados a los diferentes tratamientos o intervenciones de una manera aleatoria (por azar). De ahí que se denomine diseño de 'grupos equivalentes', ya que se utiliza la técnica de la aleatorización para mantener controladas por azar las posibles diferencias entre los grupos antes de introducir el tratamiento o las diferentes condiciones de intervención. Los sujetos de los grupos son homogéneos o equivalentes en las variables observadas y no observadas antes de introducir la variable manipulada o variable independiente. La principal técnica de control que se utiliza es la aletorización en la asignación del tratamiento, pero si hay dudas de la calidad de esta medida se pueden aplicar también otras técnicas de control para controlar posibles variables extrañas conocidas por el investigador.

Diseño de grupo no equivalente o diseño con grupo de control no equivalente

Diseño de grupo no equivalente o diseño con grupo de control no equivalente (nonequivalente group design)

Se trata de un diseño con una metodología cuasi-experimental. Este diseño permite comparar grupos de participantes o unidades experimentales que son asignados a los diferentes tratamientos o intervenciones de una manera no aleatoria. De ahí que se denomine diseño de 'grupos no equivalentes'. Los sujetos de los

grupos no son homogéneos o equivalentes en las variables observadas y no observadas antes de introducir la variable manipulada o variable independiente y, por lo tanto, requiere planificar técnicas de control dirigidas a controlar posibles variables extrañas que el investigador o investigadora considere relevantes. Por ejemplo, la eliminación, el bloqueo o la medida de una variable covariada.

Análisis de la Varianza (ANOVA)

Análisis de la Varianza (ANOVA), (Analysis of Variance)

Sir Ronal Fisher utilizó por primera vez el término de Análisis de la varianza (ANOVA) en 1918 (David, 1995). Fisher consideraba el ANOVA como una técnica o procedimiento para analizar las diferencias en el rendimiento de los cultivos en función de las parcelas agrícolas donde se llevaba a cabo el cultivo que diferirían por el tipo de abono o tratamiento que recibían a lo largo de su crecimiento (Gamst, Meyers, y Guarino, 2008).

El ANOVA es una técnica estadística paramétrica que analiza las diferencias entre las puntuaciones medias cuando el diseño utiliza una única variable dependiente que se mide en dos o más grupos o variables independientes o en dos o más mediciones repetidas a los mismos sujetos. Se trata de una técnica de análisis muy popular en las Ciencias Sociales y de la Salud. Es la técnica de análisis de inferencia estadística que se utiliza mayoritariamente en los diseño de investigación univaridos entre-grupos e intra-grupos. Por ejemplo, Keselman y cols. (1998) señalan que en el 93.3% de los artículos que aplican diseños entre-grupos univariados utilizan la técnica del ANOVA para obtener una solución a sus hipótesis.

Como todas las técnicas paramétricas, antes de ejecutar el ANOVA es necesario comprobar los supuestos relacionados con las puntuaciones en la variable de resultados o variable dependiente: independencia de las puntuaciones de los diferentes grupos, normalidad de las puntuaciones y homogeneidad de las varianzas de los grupos junto con la medida realizada al menos en escala de intervalo de la variable dependiente. En muy pocas ocasiones los investigadores señalan en sus artículos el cumplimiento de tales supuestos (Skidmore y Thompson, 2013). Sin embargo, hay que tener en cuenta que los estudios de simulación señalan que la razón *F* es robusta a leves desviaciones de la normalidad (Harwell, Rubinstein, Haye, y Olds, 1992) y que la razón *F* es relativamente insensible a la violación de la

asunción de normaldiad cuando el diseño tiene grupos con el mismo tamaño de observaciones (Glass, Peckham, y Sanders, 1972; Lix, Keselman, y Keselman, 1996).

Análisis de la Varianza (MANOVA)

El investigador o investigadora también podría plantear un diseño con dos variables dependientes (o más) y analizar los datos con un diseño de multivariado de la varianza (MANOVA). Cuando en una investigación al científico le interesa registrar y analizar de forma conjunta más de una variable dependiente para observar el efecto estimado del tratamiento se dice que está aplicando un diseño *multivariado*; el caso más simple de diseño multivariado es aquel que cuenta con una única variable independiente y al menos dos variables dependientes. De manera análoga a lo que ocurre con el diseño factorial respecto del diseño simple, cuando se aplica un diseño multivariado los resultados obtenidos después del análisis estadístico pueden ser distintos de los que se conseguirían tras aplicar múltiples diseños univariados.

En el diseño multivariado no sólo se analiza la relación existente entre las variables independientes y dependientes sino que se tiene en cuenta la relación entre las variables dependientes, relación que se analiza bajo los dos supuestos, modelo restringido y modelo completo; el planteamiento y el proceso analítico es análogo al que se sigue cuando se aplica el ANOVA. Para consultar el diseño multivariado y su análisis se puede consultar García y cols.(1999).

Valor *p*

Valor p (*p*-value)

En el procedimiento clásico de comprobación de la significación de la hipótesis nula (NHST), el valor *p* calculado del estadístico computado en el estudio realizado (probabilidad calculada del valor del estadístico, asumiendo que la hipótesis nula es cierta) es: la probabilidad del resultado del hallazgo observado, o un resultado más extremo, cuando la hipótesis nula (H_0) sobre la cuestión que se estudia es verdadera (*The p value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true*). Con otras palabras, el valor *p* calculado es la probabilidad de los resultados de la muestra (no de la población), o más extremos, asumiendo que la muestra procede de una

población donde la hipótesis nula es verdadera. Los valores de los parámetros poblacionales que se presuponen para calcular el valor de *p* pueden ser cualquiera y no siempre $H_0 = 0$ (nil-hypothesis). Sin embargo, la mayoría de los paquetes estadísticos (ejemplo, SPSS, JASP, JAMOVI, STATS…) (y consecuentemente, la mayoría de los investigadores) asumen que H_0 de no diferencias entre las puntuaciones (diferencia de cero entre las puntuaciones) es cierto en la población (Cohen, 1994; Thompson, 1996). Generalmente, se considera que la significación estadística o valor *p* del resultado de la muestra es estadísticamente significativo cuando $p \leq$ alfa, es decir se concluye que hubo un efecto sistemático. Conviene tener muy presente, ya que es un error frecuente, que la significación o valor *p* no informa de la significación práctica o utilidad del hallazgo (solo el juicio del experto ofrece esa valoración), ni de la significación del tamaño del efecto (esta información requiere computar un estadístico del tamaño del efecto) (Kirk, 1996). Por lo tanto, el valor *p* de probabilidad es la probabilidad de obtener el efecto observado en la muestra de la investigación (o más grande) en una distribución de la hipótesis nula que se ha construido sobre el supuesto de efecto cero en la población o relación cero entre las variables. En la distribución muestral del estadístico (construida sobre el supuesto de efecto cero en la población) hay valores muy frecuentes y valores muy poco frecuentes, pero todos se construyen asumiendo que la hipótesis nula es cierta o efecto cero. Es decir, tanto las grandes diferencias como las pequeñas diferencias entre los grupos son producto de la extracción aleatoria de las muestras (azar). Las pequeñas diferencias tienen una probabilidad más baja (son menos frecuentes) y las grandes diferencias tienen una probabilidad má alta (son más frecuentes), pero conviene tener muy presente que todas las diferencias que se producen son efectos del azar o por el proceso de extracción aleatoria de las muestras. Por otro lado, se encuentra el concepto de valor *p* de probabilidad asociado a un resultado que se ha obtenido con una prueba estadística y unos datos de una muestra de participantes. Ese valor *p* de probabilidad puede ser muy pequeño, indicando que el efecto observado es muy poco probable en la distribución de la hipótesis nula, pero eso no quiere decir que no se haya podido producir por azar. Del mismo modo, un valor de *p* grande indica que el efecto observado es muy probable en la distribución de la hipótesis, pero eso no quiere decir que necesariamente se ha producido por azar. Cuando se calculan los valores de *p* se asume como un supuesto que el azar es la

causa de las diferencias, pero unos valores de *p* son más compatibles con el modelo de la hipótesis nula (los que tienen valor de *p* grandes) y otros valores de p se consideran que son poco compatibles con un modelo que plantea efecto cero (los que tienen valores de *p* pequeños). El valor *p* del resultado se calcula consultado la distribución de la hipótesis nula de efecto cero. Para decidir que es un valor pequeño de *p* o que es un valor alto de p se utiliza la regla de .05 (alfa): menor o igual a .05 es un valor de *p* pequeño (se concluye que hay un efecto / relación estadísticamente significativo y con ello se justifica el rechazo de la hipótesis nula) y mayor a .05 es un valor alto de *p* (se concluye el efecto / relación no es estadísticamente significativo y se concluye que se mantiene la hipótesis nula). No se demuestra ni la verdad ni la falsedad de la hipótesis nula, se asume que cierta. De ahí, que no debe interpretarse un resultado nulo (se mantiene la hipótesis nula) como ausencia de diferencias o ausencia de efecto (no se demuestra la verdad de la hipótesis nula), ya que solo se sabe que no hay suficiente evidencia de que exista un efecto y, por lo tanto, se trata de un resultado no concluyente. Del mismo modo, un valor de *p* pequeño no implica de forma necesaria que el efecto o la relación entre las variables es importante o es beneficiosa. El valor de *p* solo valora si el efecto es cero y no informa de la magnitud del efecto ni de su importancia, ni tampoco, por supuesto de la utilidad del hallazgo. Podría suceder que en un estudio con una *N* muy grande se detecte un efecto de una magnitud trivial como estadísticamente significativo. De ahí la importancia de acompañar el valor p con un estadístico del tamaño del efecto y su intervalo de confianza. El intervalo de confianza del tamaño del efecto nos indica la precisión de la estimación puntual y si, por ejemplo, abarca desde efectos pequeños hasta efectos grandes, pues en este caso indicaría la escasa precisión de la estimación puntual. El valor de .05 es arbitrario y es el valor máximo de alfa que se acepta en las Ciencias de la Salud y las Ciencias Sociales por consenso de la comunidad científica. Hay que resaltar que en el cómputo del estadístico y su valor *p* intervienen el efecto estimado y el tamaño de la muestra (es decir, la precisión del efecto estimado). A medida que el tamaño de la muestra aumenta, el rango de efectos posibles que podrían ocurrir por azar se reduce. De ahí que la significación estadística (el valor *p*) de un efecto de una magnitud concreta sea mayor (valor *p* más pequeño) en un estudio con una muestra grande que en un estudio con una muestra pequeña, siendo en ambos casos el mismo tamaño del efecto. En definitiva, manteniendo constante el tamaño del

efecto, si aumenta *N* baja el valor de *p* de probabilidad vinculado al estadístico que se aplicado en la prueba de inferencia estadística.

Tamaño del efecto

Tamaño del efecto (effect size)

El tamaño del efecto es un estadístico que informa de la magnitud o la fuerza del hallazgo y la dirección de las diferencias entre las puntuaciones de los grupos o de la relación entre las variables. Un tamaño del efecto puede ser una diferencia entre dos medias, un porcentaje o una correlación (Vacha-Hasse y Thompson, 2004). El tamaño del efecto proporciona información sobre cómo de grande es la diferencia entre las puntuaciones de los grupos (ejemplo, *d* de Cohen) o cómo de fuerte es la relación entre las variables (por ejemplo, *r* del coeficiente de correlación) o cuanta varianza explica la variable independiente respecto a la varianza total (eta cuadro, η^2). La información que ofrece el tamaño del efecto no se puede obtener únicamente interpretando el valor *p* obtenido en el estudio (Volker, 2006) porque el valor *p* no es un estadístico del tamaño del efecto. Además, no existe siempre una relación directa entre un valor *p* y la magnitud del efecto, pues un valor *p* pequeño puede relacionarse con un efecto bajo, medio o alto. Del mismo modo que valor *p* grande puede relacioanrse con un efecto bajo, medio o alto Tampoco existe siempre una relación directa entre la magnitud del efecto y su valor práctico o clínico. Dependiendo de las circunstancias, un efecto de menor magnitud puede ser más importante que un efecto de mayor magnitud (Durlak, 2009).

Error de Tipo I

Error de Tipo I (Type I error)

El error de Tipo I es la probabilidad de rechazar la hipótesis nula cuando realmente es cierta. Se conoce como criterio de significación o alfa. En un contraste de hipótesis es el rechazo incorrecto de la hipótesis nula. Es decir, la hipótesis nula es falsa y se mantiene tras finalizar el contraste estadístico. Generalmente, el criterio de significación o alfa máximo se sitúa en .05, de tal manera que la hipótesis nula se rechaza cuando $p \leq .05$ y se mantiene cuando $p > .05$.

Nivel de confianza

Nivel de confianza

Es el valor complementario al valor de alfa (1 - alfa). Se trata de una decisión correcta. El nivel de confianza es la probabilidad de mantener la hipótesis nula siendo realmente cierta. Es decir, la hipótesis nula es cierta y se mantiene correctamente tras finalizar el contraste estadístico.

Error de Tipo II

Error de Tipo II (Type II error)

La probabilidad de mantener la hipótesis nula cuando realmente es falsa. Se conoce como error beta. En un contraste de hipótesis se produce cuando se mantiene de forma incorrecta la hipótesis nula. Es decir, la hipótesis nula es cierta y se rechaza tras finalizar el contraste estadístico.

Potencia estadística

Potencia estadística (statistical power)

La probabilidad de rechazar correctamente la hipótesis nula (Cohen 1988). Es decir, la hipótesis nula es falsa y se rechaza correctamente tras finalizar el contraste estadístico. Se conoce como potencia estadística del estadístico y representa 1 - error beta.

La potencia estadística depende: 1) del tamaño de la muestra, 2) del tamaño del efecto y 3) del criterio de significación o nivel de confianza, alfa fijado a priori.

En un análisis de potencia a priori (planificando la potencia estadística del diseño del estudio que se va a realizar), el investigador o investigadora puede calcular el tamaño de muestra necesario (número de observaciones necesarias) para un determinado efecto estimado de un valor concreto y con un nivel de significación o alfa concreto fijado a priori y una potencia estadística deseada (generalmente, al menos de .8) (Cohen, 1988). En general, el valor mínimo recomendado de potencia estadística (1- beta) de .8 se basa en la idea de que con un criterio de significación o alfa de .05, la ratio del error de Tipo II respecto al error de Tipo I (.20 / .05 = 4) es cuatro veces más serio concluir que hay efecto cuando realmente no existe en la

población (error de Tipo I) que concluir que no hay efecto cuando realmente sí existe el efecto en la población (error de Tipo II).

Revisión sistemática y meta-análisis

Revisión sistemática y meta-análisis

La revisión sistemática es un método de síntesis de la evidencia o resultados aportados en la literatura sobre una determinada temática común. Requiere un protocolo de revisión con criterios de inclusión y exclusión y una revisión exhaustiva, objetiva y replicable de las publicaciones e informes que existan sobre el constructo objeto de estudio. Generalmente se utilizan bases de datos que deben ser descritas junto con las palabras clave y los criterios de búsqueda en la redacción del informe o artículo. La revisión sistemática debe realizar una valoración crítica de la calidad de los hallazgos de los estudios primarios que forman parte de dicha revisión. Para llevar a cabo dicha valoración existen diferentes listados de comprobación en función de la metodología del estudio. El estudio de meta-análisis se realiza después de la revisión sistemática e implica resumir de forma cuantitativa la evidencia o los resultados con un estadístico del tamaño del efecto medio y su intervalo de confianza. Implica calcular los tamaños del efecto del conjunto de trabajos que forman la revisión sistemática utilizando el mismo estadístico de tamaño del efecto para llevar a cabo finalmente el cómputo del tamaño del efecto medio y su intervalo de confianza. El estudio de meta-análisis incluye analizar la posible heterogeneidad de los estudios y si está presente llevar a cabo un modelo de análisis de efectos aleatorios junto con un estudio del tamaño del efecto a través de las diferentes variables moderadoras que se propongan como relevantes teóricamente por el investigador. También es importante que en trabajo de meta-análisis se valore el posible sesgo de publicación ya que su presencia sobrestima el tamaño del efecto medio.

Anexo 2. Plantilla de aprendizaje: descomposición de la ecuación estructural. Diseño entre-grupos

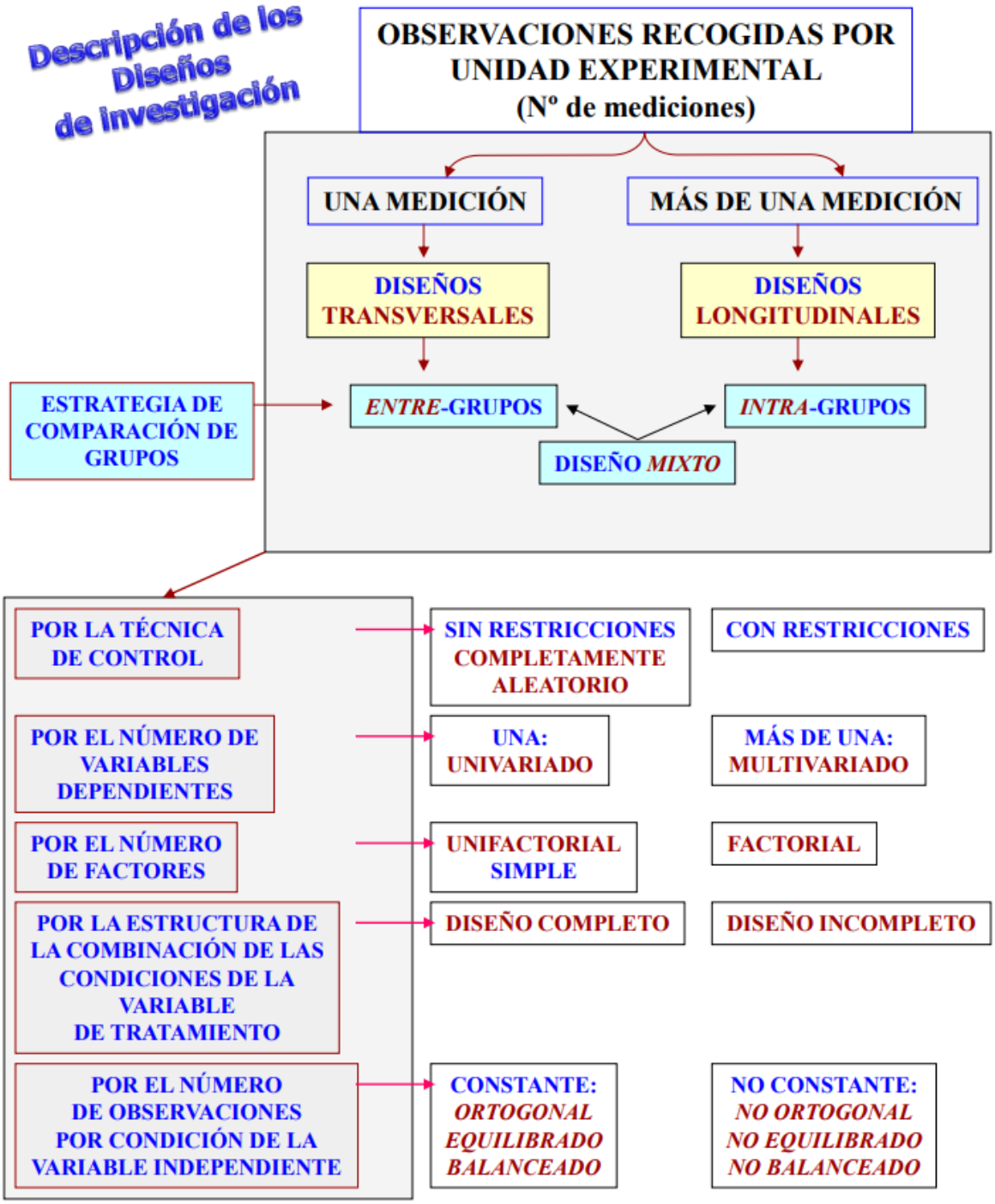
Diseño entre grups A = 2, unifactorial univariado

Pasos para desarrollar la ecuación estructural del modelo de diseño:

Y = M + A + E

a (condición: a1, a2, a3 …)	*S* (sujeto: 1, 2, 3, 4, 5 …)	Y (puntuaciones en la variable dependiente)	M (Media general o constante)	*y* (Y – M) E_{H0}	A (M_a – M)	$\hat{Y}$ (M + efectes) Puntuación pronosticada bajo la H1	E (Y - $\hat{Y}$) $E_{H1} = Y - M_a$
		SC (Suma de Cuadrados) *gl* (grados de libertad) *MC* (Media Cuadrática)		SCtotal	SCentre		SCerror

Anexo 3. Descripción de los diseños de investigación



Anexo 4. Tablas estadísticas

Cálculo on-line

-**Calcular el valor *p* de la Razón *F*** conociendo los grados de libertad del numerador (fuente 'entre' o del efecto), los grados de libertad del denominador ('intra-celdilla' o del término de erro) y el valor de la *F* empírica**:** http://davidmlane.com/hyperstat/F_table.html

Tablas estadísticas. Distribución *F*

1) Distribución *F*: alfa = .05

Tabla I. **Distribución *F* (α = .05, gl_{entre} = columnas, gl_{error} = filas)**

gl	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*24*
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882	243.906	249.052
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413	19.454
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.639
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.774
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.527
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.841
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.410
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.284	3.115
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	2.900
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.737
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.609
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.505
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.420
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.349
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.288
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.235
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.190
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.150
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.114
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.082
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.250	2.054
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.028
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.204	2.005
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.183	1.984
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	1.964
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.148	1.946
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.132	1.930
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.118	1.915
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.104	1.901
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	1.887
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199	2.153	2.080	1.875
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142	2.070	1.864
33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179	2.133	2.060	1.853
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123	2.050	1.843
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114	2.041	1.833
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.003	1.793
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.952	1.737
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.917	1.700
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969	1.893	1.674
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951	1.875	1.654
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938	1.861	1.639
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927	1.850	1.627
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.834	1.608
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850	1.772	1.539
1000	3.851	3.005	2.614	2.381	2.223	2.108	2.019	1.948	1.889	1.840	1.762	1.528
5000	3.843	2.998	2.607	2.374	2.216	2.100	2.011	1.940	1.882	1.833	1.754	1.519

2) Distribución *F*: alfa = .01

Tabla II. **Distribución** *F* (α **= .01**, gl_{entre} = *columnas*, gl_{error} = *filas)*

gl	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*24*
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.98	5928.35	5981.08	6022.47	6055.84	6106.32	6234.63
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.416	99.458
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.598
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	13.929
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.888	9.466
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.313
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.074
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.279
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.729
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.327
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.021
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	3.780
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.587
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.427
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.294
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.181
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.084
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	2.999
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	2.925
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	2.859
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	2.801
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.749
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.702
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.659
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.620
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.585
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.552
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.522
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.495
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.469
31	7.530	5.362	4.484	3.993	3.675	3.449	3.281	3.149	3.043	2.955	2.820	2.445
32	7.499	5.336	4.459	3.969	3.652	3.427	3.258	3.127	3.021	2.934	2.798	2.423
33	7.471	5.312	4.437	3.948	3.630	3.406	3.238	3.106	3.000	2.913	2.777	2.402
34	7.444	5.289	4.416	3.927	3.611	3.386	3.218	3.087	2.981	2.894	2.758	2.383
35	7.419	5.268	4.396	3.908	3.592	3.368	3.200	3.069	2.963	2.876	2.740	2.364
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.288
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698	2.562	2.183
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.115
70	7.011	4.922	4.074	3.600	3.291	3.071	2.906	2.777	2.672	2.585	2.450	2.067
80	6.963	4.881	4.036	3.563	3.255	3.036	2.871	2.742	2.637	2.551	2.415	2.032
90	6.925	4.849	4.007	3.535	3.228	3.009	2.845	2.715	2.611	2.524	2.389	2.004
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503	2.368	1.983
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.336	1.950
500	6.686	4.648	3.821	3.357	3.054	2.838	2.675	2.547	2.443	2.356	2.220	1.829
1000	6.660	4.626	3.801	3.338	3.036	2.820	2.657	2.529	2.425	2.339	2.203	1.810
5000	6.640	4.609	3.786	3.323	3.021	2.806	2.643	2.515	2.411	2.324	2.188	1.795

3) Distribución *F*: alfa = .001

Tabla III. **Distribución *F* (α = .001, gl_{entre} = columnas, gl_{error} = filas)**

gl	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*24*
1	405285.	500001.	540379.	562500.	576408.	585970.	592873.	598153.	602284.	605621.	610669.	624005.
2	998.500	999.000	999.167	999.250	999.300	999.333	999.357	999.375	999.389	999.400	999.417	999.458
3	167.029	148.500	141.108	137.100	134.580	132.847	131.583	130.619	129.860	129.247	128.316	125.935
4	74.137	61.246	56.177	53.436	51.712	50.525	49.658	48.996	48.475	48.053	47.412	45.766
5	47.181	37.122	33.202	31.085	29.752	28.834	28.163	27.649	27.244	26.917	26.418	25.133
6	35.507	27.000	23.703	21.924	20.803	20.030	19.463	19.030	18.688	18.411	17.989	16.897
7	29.245	21.689	18.772	17.198	16.206	15.521	15.019	14.634	14.330	14.083	13.707	12.732
8	25.415	18.494	15.829	14.392	13.485	12.858	12.398	12.046	11.767	11.540	11.194	10.295
9	22.857	16.387	13.902	12.560	11.714	11.128	10.698	10.368	10.107	9.894	9.570	8.724
10	21.040	14.905	12.553	11.283	10.481	9.926	9.517	9.204	8.956	8.754	8.445	7.638
11	19.687	13.812	11.561	10.346	9.578	9.047	8.655	8.355	8.116	7.922	7.626	6.847
12	18.643	12.974	10.804	9.633	8.892	8.379	8.001	7.710	7.480	7.292	7.005	6.249
13	17.815	12.313	10.209	9.073	8.354	7.856	7.489	7.206	6.982	6.799	6.519	5.781
14	17.143	11.779	9.729	8.622	7.922	7.436	7.077	6.802	6.583	6.404	6.130	5.407
15	16.587	11.339	9.335	8.253	7.567	7.092	6.741	6.471	6.256	6.081	5.812	5.101
16	16.120	10.971	9.006	7.944	7.272	6.805	6.460	6.195	5.984	5.812	5.547	4.846
17	15.722	10.658	8.727	7.683	7.022	6.562	6.223	5.962	5.754	5.584	5.324	4.631
18	15.379	10.390	8.487	7.459	6.808	6.355	6.021	5.763	5.558	5.390	5.132	4.447
19	15.081	10.157	8.280	7.265	6.622	6.175	5.845	5.590	5.388	5.222	4.967	4.288
20	14.819	9.953	8.098	7.096	6.461	6.019	5.692	5.440	5.239	5.075	4.823	4.149
21	14.587	9.772	7.938	6.947	6.318	5.881	5.557	5.308	5.109	4.946	4.696	4.027
22	14.380	9.612	7.796	6.814	6.191	5.758	5.438	5.190	4.993	4.832	4.583	3.919
23	14.195	9.469	7.669	6.696	6.078	5.649	5.331	5.085	4.890	4.730	4.483	3.822
24	14.028	9.339	7.554	6.589	5.977	5.550	5.235	4.991	4.797	4.638	4.393	3.735
25	13.877	9.223	7.451	6.493	5.885	5.462	5.148	4.906	4.713	4.555	4.312	3.657
26	13.739	9.116	7.357	6.406	5.802	5.381	5.070	4.829	4.637	4.480	4.238	3.586
27	13.613	9.019	7.272	6.326	5.726	5.308	4.998	4.759	4.568	4.412	4.171	3.521
28	13.498	8.931	7.193	6.253	5.656	5.241	4.933	4.695	4.505	4.349	4.109	3.462
29	13.391	8.849	7.121	6.186	5.593	5.179	4.873	4.636	4.447	4.292	4.053	3.407
30	13.293	8.773	7.054	6.125	5.534	5.122	4.817	4.581	4.393	4.239	4.001	3.357
31	13.202	8.704	6.993	6.067	5.480	5.070	4.766	4.531	4.344	4.190	3.953	3.311
32	13.117	8.639	6.936	6.014	5.429	5.021	4.719	4.485	4.298	4.145	3.908	3.268
33	13.039	8.579	6.883	5.965	5.382	4.976	4.675	4.441	4.255	4.102	3.867	3.228
34	12.965	8.522	6.833	5.919	5.339	4.934	4.633	4.401	4.215	4.063	3.828	3.191
35	12.896	8.470	6.787	5.876	5.298	4.894	4.595	4.363	4.178	4.027	3.792	3.156
40	12.609	8.251	6.595	5.698	5.128	4.731	4.436	4.207	4.024	3.874	3.642	3.011
50	12.222	7.956	6.336	5.459	4.901	4.512	4.222	3.998	3.818	3.671	3.443	2.817
60	11.973	7.768	6.171	5.307	4.757	4.372	4.086	3.865	3.687	3.541	3.315	2.694
70	11.799	7.637	6.057	5.201	4.656	4.275	3.992	3.773	3.596	3.452	3.227	2.608
80	11.671	7.540	5.972	5.123	4.582	4.204	3.923	3.705	3.530	3.386	3.162	2.545
90	11.573	7.466	5.908	5.064	4.526	4.150	3.870	3.653	3.479	3.336	3.113	2.497
100	11.495	7.408	5.857	5.017	4.482	4.107	3.829	3.612	3.439	3.296	3.074	2.458
120	11.380	7.321	5.781	4.947	4.416	4.044	3.767	3.552	3.379	3.237	3.016	2.402
500	10.957	7.004	5.506	4.693	4.176	3.813	3.542	3.332	3.163	3.023	2.806	2.195
1000	10.892	6.956	5.464	4.655	4.139	3.778	3.508	3.299	3.130	2.991	2.774	2.164
5000	10.840	6.917	5.430	4.624	4.110	3.750	3.481	3.272	3.104	2.965	2.749	2.139

4) Distribución *t* de Student

Distribución t de Student

Tabla II. *($T \leq tp$, gl)*

gl	*0.750*	*0.900*	*0.950*	*0.975*	*0.980*	*0.990*	*0.995*	*0.999*
1	1.0000	3.0777	6.3138	12.7062	15.8945	31.8205	63.6567	318.3088
2	0.8165	1.8856	2.9200	4.3027	4.8487	6.9646	9.9248	22.3271
3	0.7649	1.6377	2.3534	3.1824	3.4819	4.5407	5.8409	10.2145
4	0.7407	1.5332	2.1318	2.7764	2.9985	3.7469	4.6041	7.1732
5	0.7267	1.4759	2.0150	2.5706	2.7565	3.3649	4.0321	5.8934
6	0.7176	1.4398	1.9432	2.4469	2.6122	3.1427	3.7074	5.2076
7	0.7111	1.4149	1.8946	2.3646	2.5168	2.9980	3.4995	4.7853
8	0.7064	1.3968	1.8595	2.3060	2.4490	2.8965	3.3554	4.5008
9	0.7027	1.3830	1.8331	2.2622	2.3984	2.8214	3.2498	4.2968
10	0.6998	1.3722	1.8125	2.2281	2.3593	2.7638	3.1693	4.1437
11	0.6974	1.3634	1.7959	2.2010	2.3281	2.7181	3.1058	4.0247
12	0.6955	1.3562	1.7823	2.1788	2.3027	2.6810	3.0545	3.9296
13	0.6938	1.3502	1.7709	2.1604	2.2816	2.6503	3.0123	3.8520
14	0.6924	1.3450	1.7613	2.1448	2.2638	2.6245	2.9768	3.7874
15	0.6912	1.3406	1.7531	2.1314	2.2485	2.6025	2.9467	3.7328
16	0.6901	1.3368	1.7459	2.1199	2.2354	2.5835	2.9208	3.6862
17	0.6892	1.3334	1.7396	2.1098	2.2238	2.5669	2.8982	3.6458
18	0.6884	1.3304	1.7341	2.1009	2.2137	2.5524	2.8784	3.6105
19	0.6876	1.3277	1.7291	2.0930	2.2047	2.5395	2.8609	3.5794
20	0.6870	1.3253	1.7247	2.0860	2.1967	2.5280	2.8453	3.5518
21	0.6864	1.3232	1.7207	2.0796	2.1894	2.5176	2.8314	3.5272
22	0.6858	1.3212	1.7171	2.0739	2.1829	2.5083	2.8188	3.5050
23	0.6853	1.3195	1.7139	2.0687	2.1770	2.4999	2.8073	3.4850
24	0.6848	1.3178	1.7109	2.0639	2.1715	2.4922	2.7969	3.4668
25	0.6844	1.3163	1.7081	2.0595	2.1666	2.4851	2.7874	3.4502
26	0.6840	1.3150	1.7056	2.0555	2.1620	2.4786	2.7787	3.4350
27	0.6837	1.3137	1.7033	2.0518	2.1578	2.4727	2.7707	3.4210
28	0.6834	1.3125	1.7011	2.0484	2.1539	2.4671	2.7633	3.4082
29	0.6830	1.3114	1.6991	2.0452	2.1503	2.4620	2.7564	3.3962
30	0.6828	1.3104	1.6973	2.0423	2.1470	2.4573	2.7500	3.3852
31	0.6825	1.3095	1.6955	2.0395	2.1438	2.4528	2.7440	3.3749
32	0.6822	1.3086	1.6939	2.0369	2.1409	2.4487	2.7385	3.3653
33	0.6820	1.3077	1.6924	2.0345	2.1382	2.4448	2.7333	3.3563
34	0.6818	1.3070	1.6909	2.0322	2.1356	2.4411	2.7284	3.3479
35	0.6816	1.3062	1.6896	2.0301	2.1332	2.4377	2.7238	3.3400
36	0.6814	1.3055	1.6883	2.0281	2.1309	2.4345	2.7195	3.3326
37	0.6812	1.3049	1.6871	2.0262	2.1287	2.4314	2.7154	3.3256
38	0.6810	1.3042	1.6860	2.0244	2.1267	2.4286	2.7116	3.3190
39	0.6808	1.3036	1.6849	2.0227	2.1247	2.4258	2.7079	3.3128
40	0.6807	1.3031	1.6839	2.0211	2.1229	2.4233	2.7045	3.3069
45	0.6800	1.3006	1.6794	2.0141	2.1150	2.4121	2.6896	3.2815
50	0.6794	1.2987	1.6759	2.0086	2.1087	2.4033	2.6778	3.2614
60	0.6786	1.2958	1.6706	2.0003	2.0994	2.3901	2.6603	3.2317
100	0.6770	1.2901	1.6602	1.9840	2.0809	2.3642	2.6259	3.1737
120	0.6765	1.2886	1.6577	1.9799	2.0763	2.3578	2.6174	3.1595
500	0.6750	1.2832	1.6479	1.9647	2.0591	2.3338	2.5857	3.1066
1000	0.6747	1.2824	1.6464	1.9623	2.0564	2.3301	2.5808	3.0984
5000	0.6745	1.2817	1.6452	1.9604	2.0543	2.3271	2.5768	3.0919

Anexo 5. Comparación entre diferentes pruebas: tamaño del efecto pequeño, mediano y grande

Cálculo on-line

-Calcular el tamaño del efecto *d* de Cohen y su intervalo de confianza: https://campbellcollaboration.org/research-resources/effect-size-calculator.html

Tabla 1. Pruebas estadísticas y valores de los tamaños del efecto

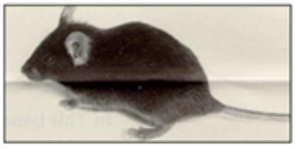
Prueba	Tamaño del efecto	Pequeño	Mediano	Grande
Diferencia de medias estandarizada, *d* de Cohen	d	0.20	0.50	0.80
Correlación de Pearson	r	0.10	0.30	0.50
Correlación de Pearson al cuadrado	r^2	0.01	0.09	0.25
Correlación biserial-puntual	r_{bp}	0.10	0.24	0.37
Correlación biserial-puntual al cuadrado	r^2_{bp}	0.01	0.06	0.14
Eta Cuadrado	η^2	0.01	0.06	0.14
Coeficiente de determinación	R^2	0.01	0.06	0.14
Omega Cuadrado	ω^2	0.01	0.06	0.14
f de Cohen del ANOVA unifactorial	f	0.10	0.25	0.40
f^2 del análisis de regresión (más de dos grupos)	f^2	0.02	0.15	0.35
Ji Cuadrado	w	0.10	0.30	0.50
Odds Ratio	OR	1.29	1.88	2.69
Porcentaje de solapamiento	OL %	85.3	66.6	52.6
Porcentaje de no solapamiento (1-% de solapamiento) (U1 de Cohen)	U1	14.7%	33.4%	47.4%
Common Language Effect size	CL (AUC)	0.56	0.64	0.71

Anexo 6. Solución al ejercicio del supuesto de indefensión aprendida y déficits depresivos. De formal manual, SPSS y JASP

Ejercicio 1. Para casa. Tendrá una evaluación continua

Plantear las características que tiene esta investigación

SUPUESTO DE INVESTIGACIÓN. Supongamos que un investigador desea comprobar si la indefensión aprendida produce déficits depresivos. Diseña una situación experimental donde los sujetos son ocho ratas ($N = 8$) que deben completar un laberinto de cuyo suelo reciben descargas eléctricas de baja intensidad ininterrumpidamente. La mitad se asigna aleatoriamente a la condición de shock *escapable* (a_1) y la otra mitad a la condición de shock *inescapable* (a_2). La tarea experimental consiste en recorrer el laberinto, midiéndose el tiempo (en segundos) que emplean (Y_i) en su recorrido. La hipótesis experimental mantiene que las ratas del grupo de shock inescapable fracasarán en su aprendizaje, recorriendo el laberinto con lentitud y sin precisión, e incluso en muchas ocasiones no llegaran a completarlo, manifestando un comportamiento depresivo. Sin embargo, las ratas que se encuentran en la condición experimental de shock escapable aprenderán que con su ejecución escapan de la descarga y aumentarán rápidamente la velocidad de carrera con objeto de eliminar la situación aversiva y llegar al habitáculo que les privará de las descargas, a diferencia de la condición de shock inescapable donde perdurará la descarga. Tras una serie de diez ensayos previos que facilitaron el aprendizaje de la situación, los resultados del experimento fueron los que se detallan en la Tabla 1.



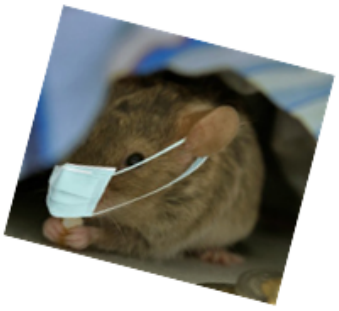


Tabla 1. Matriz de resultados

$A \rightarrow$ Shock	$Y \rightarrow$ Tiempo
a_1 Escapable	23, 11, 12, 26
a_2 No escapable	39, 38, 23, 28

Preguntas:

1-Hipótesis del estudio

2-Variable Independiente:
-Constructo
-Operacionalizada

3-Variable Dependiente:
-Constructo
-Operacionalizada

4-*N*
5-*n*
6-A
7-Y
8-E
9-Metodología del estudio
10-Tipo de diseño del estudio
11-*Plantear* la Ecuación estructural
12-Análisis: *plantear* hipótesis estadísticas

El ejercicio 1 se resolverá una vez se haya realizado la autoevaluación por parte del alumnado

13. Resuelve el contraste de hipótesis hasta obtener el valor de la Razón *F* del ANOVA

14. Decisión dicotómica

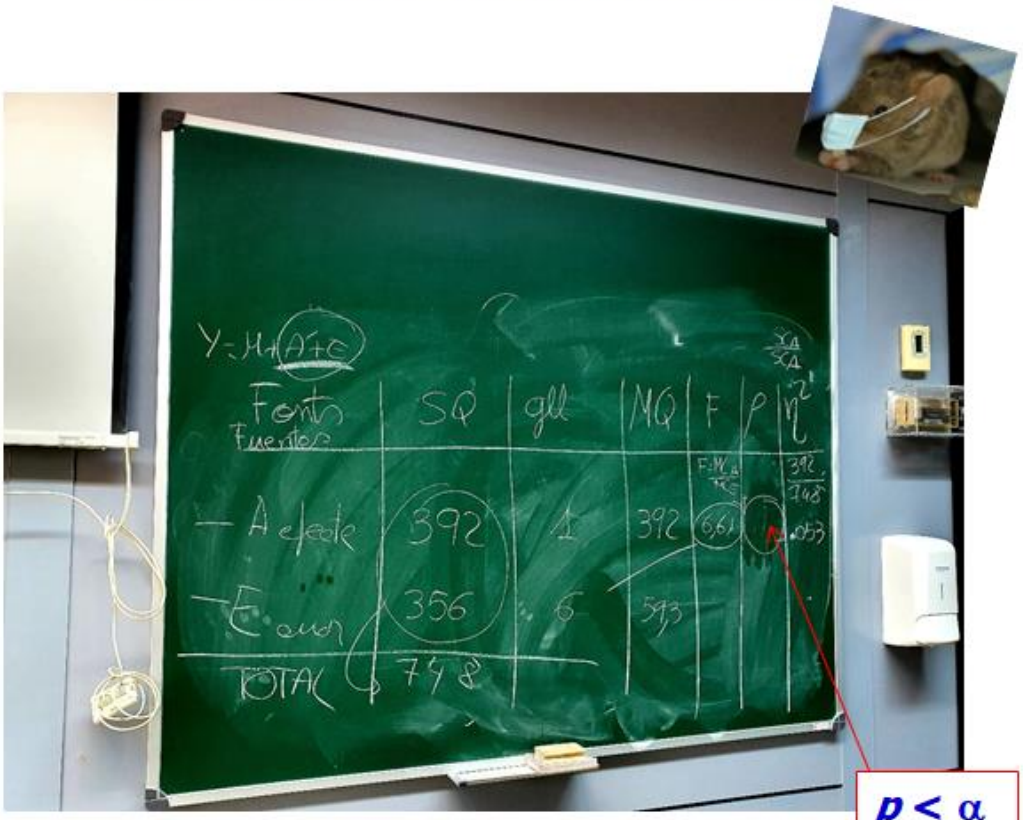
15. Redacción de resultados (formato APA)

Solución manual del Análisis de la Varianza (ANOVA)

La ecuación estructural de este ejercicio se ha desarrollado de forma manual en el libro.

Ejercicio 1. Para casa. Tendrá una evaluación continua

Plantear las características que tiene esta investigación



$p < \alpha$
$p < .05$

Solución consultando las tablas:

-*F* teórica, tabular o de tablas: *F*(.05, 1, 6) = 5,987

-*F* empírica, obtenida en el estudio: *F*(1, 6) = 6,61

Y comparando los valores de la F empírica y la F teórica puede ocurrir una de las dos situaciones siguientes:

F empírica $\geq$ F teórica *... entonces* ***se rechaza la H_0***

O puede ocurrir que:

F empírica $<$ F teórica *... entonces* ***se mantiene la H_0***

Ejercicio 1. Para casa. Tendrá una evaluación continua

Plantear las características que tiene esta investigación

En el ejercicio se puede observar que:

F* empírica** $\geq$ ***F* teórica** *... entonces* ***se rechaza la H_0

6,61 $>$ *5,987* *... entonces* ***se rechaza la H_0***

Por lo tanto, se deduce que → $p \leq \alpha$; $p \leq .05$

Utilizando las tablas del estadístico *F* no se puede obtener el valor exacto de probabilidad que está asociado al resultado obtenido en el análisis (en concreto el valor de *p* vinculado a 6,61). Solamente se puede saber si es menor o igual al valor de alfa (error de Tipo I) o si es mayor.

En el ejercicio se observa que el valor de la *F* empírica es mayor y, por lo tanto, se concluye que su valor *p* de probabilidad en la distribución de la hipótesis nula es menor al valor de alfa prefijado a priori en .05 (son las tablas estadísticas que se han consultado para alfa = .05).

Conclusión: dado que su valor de *p* es menor al calor de alfa se concluye que el resultado de la *F* empírica (6,61) no es compatible con el modelo de la hipótesis nula y se rechaza H_0. Como consecuencia, se acepta la hipótesis alternativa (H_1) y se concluye que la diferencia entre las dos medias de los grupos de shock es estadísticamente significativa. Se miran las medias de los grupos y se observa qué media es la más alta y la más baja y se concluye señalando que los sujetos con la media más alta son los que presentan más déficits depresivos.

A continuación se detalla cómo llevar a cabo la redacción de los resultados en un informe de investigación.

Ejercicio 1. Para casa. Tendrá una evaluación continua

Plantear las características que tiene esta investigación

REDACCIÓN DE LOS RESULTADOS

La redacción de los resultados del *diseño entre-grupos univariado* en un informe de investigación siguiendo el formato del Manual del APA sería, por ejemplo, la siguiente:

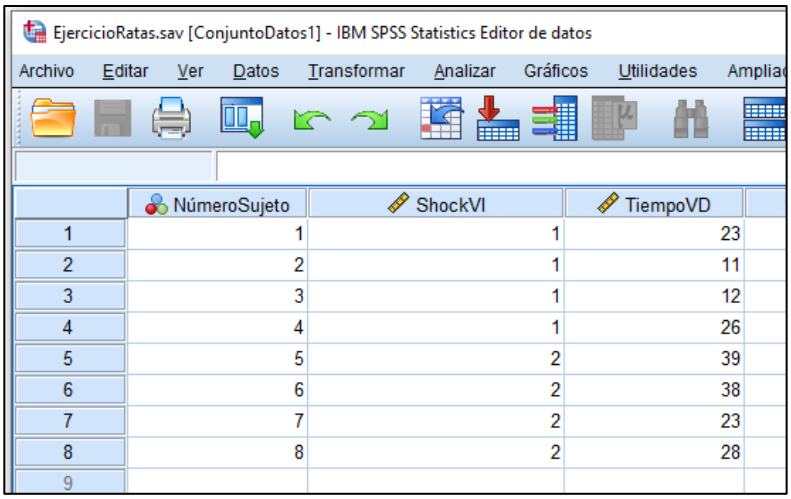
> Los resultados del estudio del efecto de la indefensión aprendida sobre la sintomatología depresiva mediante un diseño entre-grupos unifactorial univariado con dos grupos (grupo 1 = shock eléctrico escapable, grupo 2 = shock eléctrico no escapable) señalan que las ratas que son sometidas a un situación de indefensión (reciben shock eléctrico no escapable o independiente de su conducta) tienen una puntuación media más alta en depresión (Media = 32, *DT* = 7.79, *n* = 4) que las ratas que sí tienen control de la situación del shock y, por lo tanto, no desarrollan la situación de indefensión aprendida (Media = 18, *DT* = 7.62, *n* = 4), siendo la diferencia entre las medias de los dos grupos estadísticamente significativa con un tamaño del efecto grande, $F(1, 6) = 6.61$, $p = 042$, $\eta^2 = .52$. Por lo tanto, observando las puntuaciones medias de las condiciones experimentales, las ratas del grupo de shock no escapable recorrieron el laberinto con mayor lentitud que las ratas que fueron sometidas a shock escapable.

A esta redacción se le puede añadir el resultado de la prueba *F* de Levene que analiza si las varianzas de los dos grupos son homogéneas o no. Uno de los supuestos de las pruebas paramétricas es que las varianzas de los dos dos grupos (se compara la variabilidad que hay entre las puntuaciones de un grupo con la variabilidad que hay en el otro grupo) no debe ser estadísticamente diferente. Así, la redacción anterior se puede completar con la siguiente oración:

"Se ha comprobado que las varianzas de los dos grupos no difieren de forma estadísticamente significativa (Levene *F* (1, 6) = 0, *p* = 1)".

Solución con el programa SPSS

1º INTRODUCCIÓN DE DATOS. Diseño entre-grupos: variable independiente y variable dependiente.

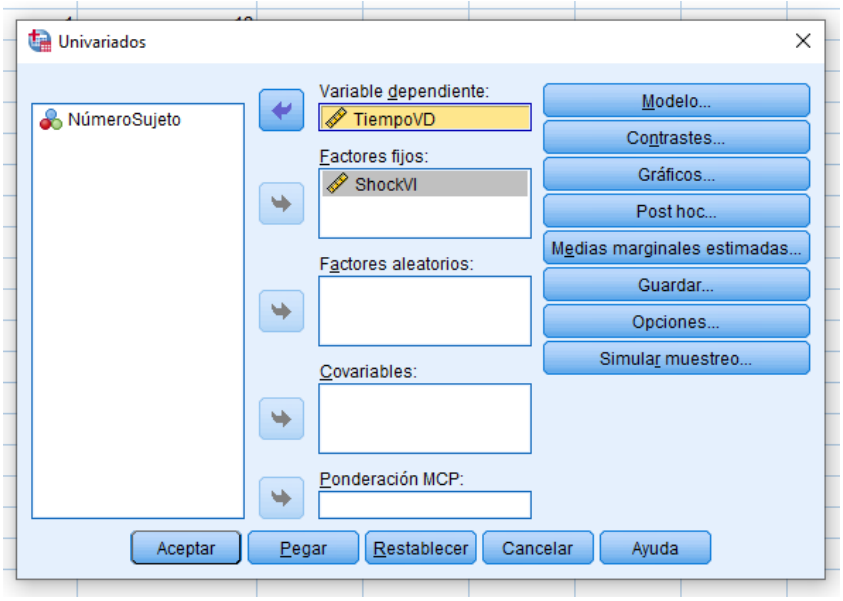


	NúmeroSujeto	ShockVI	TiempoVD
1	1	1	23
2	2	1	11
3	3	1	12
4	4	1	26
5	5	2	39
6	6	2	38
7	7	2	23
8	8	2	28
9			

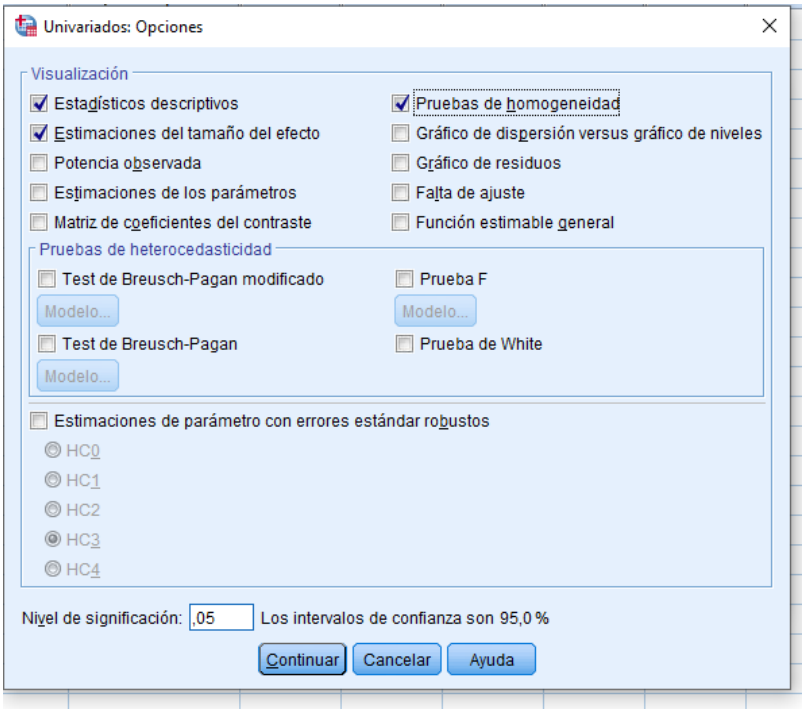
2º PASOS PARA CALCULAR EL ANOVA CON EL SPSS. **PRIMERO**: en la ventana de ANALIZAR seleccionar Modelo Lineal General y Univariado:

ANALIZAR → MODELO LINEAL GENERAL → UNIVARIADO

3º **SEGUNDO**: pasar cada variable del modelo que se va a contrastar a su apartado: variable dependiente o variable independiente (factor fijo):



3º **TERCERO**: abrir OPCIONES y seleccionar: estadísticos descriptivos, estimaciones del tamaño del efecto y pruebas de homogeneidad.



Se selecciona Continuar y posteriormente se Acepta y se obtiene en la ventana de Resultados los datos finales del ANOVA.

Análisis univariado de varianza

Factores inter-sujetos

		Etiqueta de valor	N
ShockVI	1	Escapable	4
	2	No escapable	4

Estadísticos descriptivos

Variable dependiente: TiempoVD

ShockVI	Media	Desv. Desviación	N
Escapable	18,00	7,616	4
No escapable	32,00	7,789	4
Total	25,00	10,337	8

Prueba de igualdad de Levene de varianzas de error[a,b]

		Estadístico de Levene	gl1	gl2	Sig.
TiempoVD	Se basa en la media	,000	1	6	1,000
	Se basa en la mediana	,000	1	6	1,000
	Se basa en la mediana y con gl ajustado	,000	1	4,883	1,000
	Se basa en la media recortada	,000	1	6	1,000

Prueba la hipótesis nula de que la varianza de error de la variable dependiente es igual entre grupos.

a. Variable dependiente: TiempoVD

b. Diseño : Intersección + ShockVI

Pruebas de efectos inter-sujetos

Variable dependiente: TiempoVD

Origen	Tipo III de suma de cuadrados	gl	Media cuadrática	F	Sig.	Eta parcial al cuadrado
Modelo corregido	392,000[a]	1	392,000	6,607	,042	,524
Intersección	5000,000	1	5000,000	84,270	,000	,934
ShockVI	392,000	1	392,000	6,607	,042	,524
Error	356,000	6	59,333			
Total	5748,000	8				
Total corregido	748,000	7				

a. R al cuadrado = ,524 (R al cuadrado ajustada = ,445)

A partir de aquí, ya se dispone de toda la información para pasar a redactar los resultados.

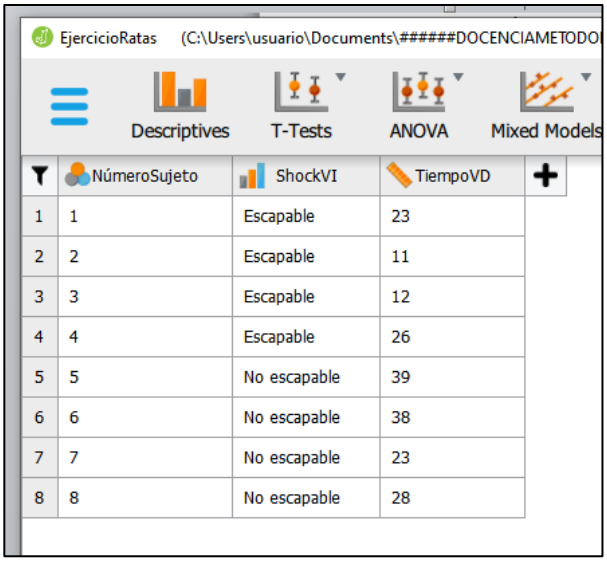
Solución con el programa JASP

El programa estadístico JASP es libre y gratuito. Se puede acceder a su página web y descargarlo en el propio ordenador en la siguiente dirección:

https://jasp-stats.org/

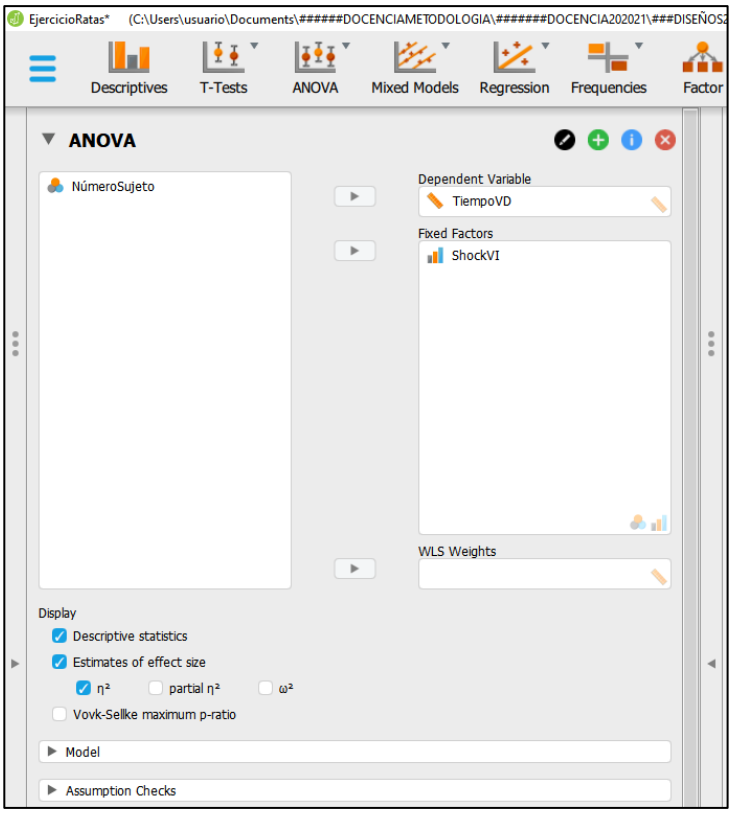
Cuando se abre el programa JASP, a la izquierda de la pantalla hay tres líneas horizontales y desde ahí hay que llamar a la base de datos que se ha preparado previamente en Excel o en SPSS, por ejemplo.

Una vez ya se dispone de la base de datos abierta en JASP ya se puede proceder con el análisis que se crea conveniente.

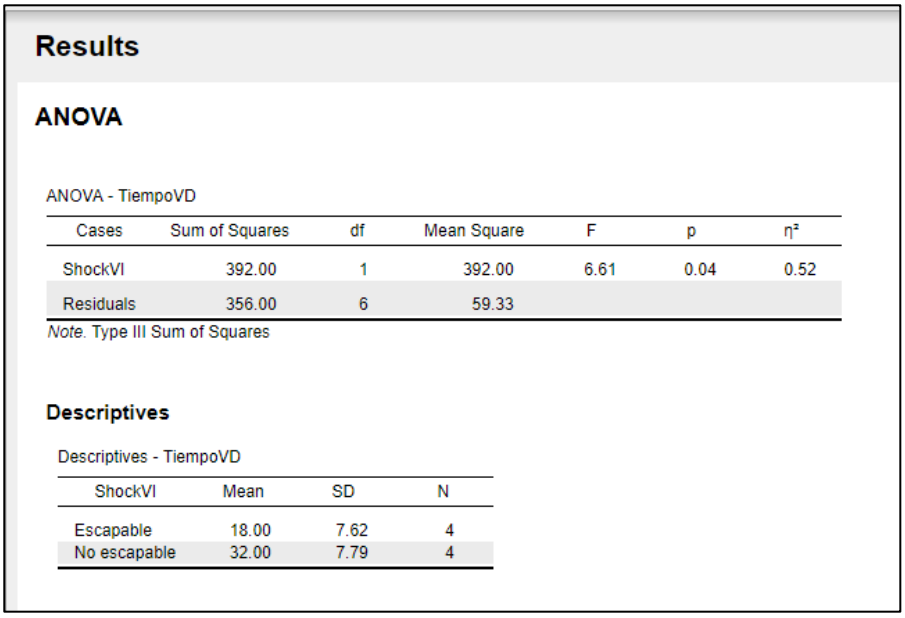


	NúmeroSujeto	ShockVI	TiempoVD
1	1	Escapable	23
2	2	Escapable	11
3	3	Escapable	12
4	4	Escapable	26
5	5	No escapable	39
6	6	No escapable	38
7	7	No escapable	23
8	8	No escapable	28

En concreto, el análisis del ejercicio requiere acceder a la ventana de ANOVA y seleccionar **Classical** → **ANOVA**. Una vez dentro del apartado de ANOVA se procede con situar cada variable del modelo en su lugar correspondiente y, además, se selecciona la estimación del tamaño del efecto de eta cuadrado y los descriptivos.



De forma paralela, a la derecha de la pantalla del ordenador van apareciendo los análisis que se han seleccionado en la parte izquierda de la pantalla.



Results

ANOVA

ANOVA - TiempoVD

Cases	Sum of Squares	df	Mean Square	F	p	η²
ShockVI	392.00	1	392.00	6.61	0.04	0.52
Residuals	356.00	6	59.33			

Note. Type III Sum of Squares

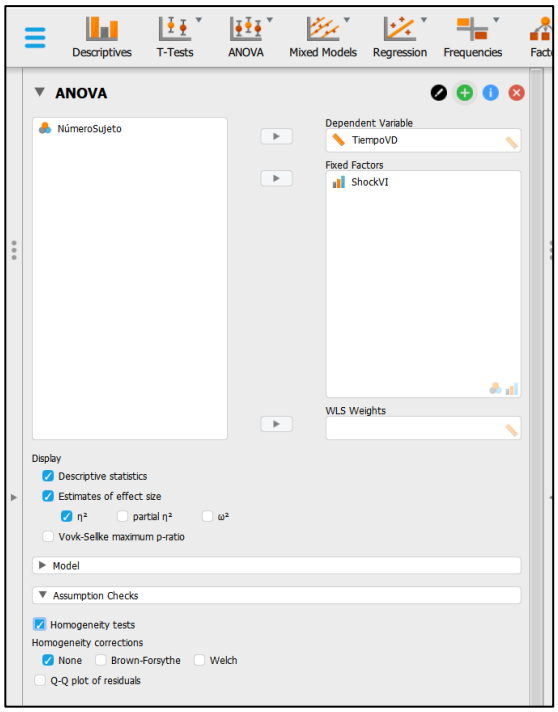
Descriptives

Descriptives - TiempoVD

ShockVI	Mean	SD	N
Escapable	18.00	7.62	4
No escapable	32.00	7.79	4

A partir de aquí ya se dispone de toda la información para redactar los resultados del diseño entre-grupos unifactorial univariado.

También se puede seleccionar la prueba de homogeneidad de las varianzas en Assumptions Checks → Homogenity tests.



Y cuando se selecciona Homogenity tests el programa JASP añade los resultados de la prueba de Levene que contrasta si las varianzas de los grupos implicados en el análisis de la varianza son o no diferentes de forma estadísticamente significativa. En el caso del ejercicio, JASP ofrece el siguiente resultado cuando se seleccionan todos los apartados mencionados:

Results

ANOVA

ANOVA - TiempoVD

Cases	Sum of Squares	df	Mean Square	F	p	η²
ShockVI	392.00	1	392.00	6.61	0.04	0.52
Residuals	356.00	6	59.33			

Note. Type III Sum of Squares

Descriptives

Descriptives - TiempoVD

ShockVI	Mean	SD	N
Escapable	18.00	7.62	4
No escapable	32.00	7.79	4

Assumption Checks

Test for Equality of Variances (Levene's)

F	df1	df2	p
7.95e −30	1.00	6.00	1.00

Anexo 7. Autoevaluaciones

A continuación se ha elaborado un cuestionario de autoevaluación cuyas respuestas se encuentran en la web: https://www.uv.es/friasnav en el apartado que se ha desarrollado para el libro:

Frías-Navarro, D. y Pascual-Soler, M. (Eds.) (2020*). Diseño de la investigación, análisis y presentación de resultados*. Universidad de Valencia. España.

Instrucciones

Para realizar la autoevaluación hay que tener en cuenta que solamente se puede dar una respuesta a cada pregunta. Es decir, no se pueden señalar dos o más opciones como respuestas a la pregunta. Si se considera que hay dos respuestas verdaderas entonces hay que seleccionar la opción de “No se puede responder a la pregunta”, ya que solamente se puede seleccionar una opción de respuesta. Si todas las respuestas son falsas y existe la opción de “No se puede responder a la pregunta” entonces dicha opción será la correcta. Del mismo modo, si todas las respuestas son verdaderas y existe la opción de “No se puede responder a la pregunta” entonces dicha opción será la correcta.

Las respuestas correctas a las autoevaluaciones se encuentran en un documento que se depositará en la página web docente una vez que el alumando haya realizado las autoevaluaciones. Por lo tanto, se irán corrigiendo parte de las preguntas en clase y la otra parte se dará la opción correcta en la página web docente.

FORMULARIO QUE SE NECESITA EN LOS EJERCICIOS:

MC = SC/gl

F = MCefecte / MCerror

1) Autoevaluación 1. Cuestionario

1. Qué quiere decir maximizar la varianza sistemàtica primaria:

A. Aumentar el valor de alfa que se fija a priori (> .05) para maximizar la probabilidad de detectar el efecto que se desea estudiar.
B. Diseñar un estudio con un número amplio de observaciones (> 100) para maximizar el efecto que se desea estudiar.
C. Planificar en el diseño las condiciones necesarias que maximicen el efecto que se desea estudiar.
D. Disminuir el valor de alfa que se fija a priori (> .05) para maximizar la probabilidad de detectar el efecto que se desea estudiar.

2. En un diseño de investigación, el problema de la tercera variable se refiere a:

A. La validez de conclusión estadística.
B. La validez interna.
C. La validez externa.
D. La validez de constructo.

3. Los resultados de un diseño tienen validez interna cuando la variabilidad de las puntuaciones en la variable dependiente es:

A. Sistemática y provocada por el efecto de la variable independiente.
B. Sistemática, aleatoria y provocada por el efecto de la variable medida.
C. No sistemática y constante y producida por el experimentador.
D. No se puede responder a la pregunta.

4. Una investigadora desea estudiar la cantidad de sudor que segregan los pacientes fóbicos según el grado de cercanía del estímulo que desencadena el miedo. Por lo tanto, en el diseño:

A. La variable dependiente será el tipo de estímulo que desencadena la fobia y la independiente la intensidad de la secreción.
B. La variable independiente será la intensidad de la secreción y la dependiente la cercanía del estímulo que desencadene la fobia.
C. La variable independiente será la cercanía del estímulo que desencadene la fobia y la variable dependiente la intensidad de la secreción.

D. La variable dependiente será el grado de cercanía del estímulo y la independiente la intensidad de la secreción.

5. El proceso de búsqueda y localización de la información bibliográfica será más válido cuando:

A. El problema de investigación (necesidad de conocimiento) se plantea una vez examinados los resultados del estudio.
B. El problema de investigación (necesidad de conocimiento) no se define previamente a dicho proceso de búsqueda.
C. El problema de investigación (necesidad de conocimiento) se define de forma concreta previamente a dicho proceso de búsqueda.
D. El problema de investigación (necesidad de conocimiento) se define de forma amplia para así poder captar más referencias bibliográficas.

6. En el denominado principio MAXMINCON se señala que hay que controlar:

A. La asignación aleatoria de las unidades experimentales a los grupos o condiciones de medida.
B. La varianza no sistemàtica (aleatoria).
C. La varianza sistemática secundaria.
D. La varianza sistemática primaria.

7. La validez interna de un diseño está relacionada con:

A. La selección aleatoria de la muestra.
B. La generalización de los resultados.
C. La aplicación correcta de las pruebas de inferencia estadística.
D. La minimización del sesgo y el incremento en la precisión de la estimación del efecto.

8. El efecto del tratamiento está vinculado con:

A. La varianza sistemática primaria.
B. La varianza sistemática secundaria.
C. La varianza sistemática terciaria.
D. La varianza sistemática aleatoria.

9. En un diseño de investigación la manipulación de la variable significa que:

A. La investigadora controla las condiciones de la variable independiente.
B. La investigadora manipula las puntuaciones en la variable dependiente.
C. La investigadora actúa sin ética y/o integridad.
D. La investigadora manipula las opiniones de los participantes para facilitar que cooperen en el desarrollo del estudio.

10. Las variables extrañas en un modelo de diseño de investigación son:

A. Son una fuente de varianza sistemática primaria.
B. Variables dependientes que no forman parte de la hipotesis.
C. Variables independientes que no forman parte de la hipotesis.
D. Son una fuente de varianza sistemática aleatoria.

2) Autoevaluación 2. Alzheimer

SUPÒSIT D'INVESTIGACIÓ 1. En els últims anys s'ha comprovat que les persones majors i fins i tot aquelles diagnosticades de malaltia d'Alzheimer en fase lleu o moderada, encara que de forma limitada, són capaços d'aprendre (Calero, 2000). Les bases biològiques d'aquesta capacitat d'aprenentatge procedeix de l'àmplia evidència empírica sobre la capacitat de les neurones lesionades per regenerar-se i establir noves connexions (Goldman, 1995). S'ha demostrat la plasticitat del sistema nerviós o neuroplasticitat en el cervell de l'ancià, fins i tot quan la demència és lleu, però no passa el mateix en les fases greus de la malaltia ja que hi ha una gran pèrdua de neurones i falta de connexions sinàptiques (Kass, 1995). En el present estudi s'ha dut a terme un disseny on es mesurava a l'inici la línia base del deteriorament cognitiu (es diagnosticava el deteriorament cognitiu), després es realitzava un entrenament en tasques cognitives (psicoestimulació cognitiva) i posteriorment (al cap de sis mesos) es tornava a mesurar els subjectes en aquesta variable per observar si s'havia produït millora en el potencial d'aprenentatge. El deteriorament cognitiu es mesurava amb el Mini-Mental State Examination (MMSE) on a major puntuació major deteriorament.

En l'estudi van participar 8 pacients d'Alzheimer en fase lleu de deteriorament cognitiu (edat mitjana: 75.3 anys, DT: 6.4; 6 homes i 2 dones) diagnosticats per la

Unitat de Diagnòstic de l'hospital. Quatre d'ells van participar en un programa de psicoestimulació cognitiva (PPC) i la resta no participaven en cap programa (grup control). L'assignació dels subjectes als grups experimental i control es va deure a raons clíniques ja que aquells subjectes que tenien un pitjor estat mental a la línia base i un major deteriorament en memòria aspecte-espacial, àudio verbal i fluïdesa verbal pertanyien al grup experimental. Tots els pacients (grup experimental i control) portaven un tractament estable més de quatre mesos amb inhibidors d'acetilcolinesterasa abans de començar l'estudi i van continuar amb el mateix tractament durant tot el procés experimental. Tots els contrastos estadístics són bilaterals (alfa = .05).

11. Selecciona la respuesta correcta:

A. La investigadora controla las condiciones de la variable independiente.
B. La investigadora manipula las puntuaciones en la variable dependiente.
C. La investigadora actúa sin ética y/o integridad.
D. La investigadora manipula las opiniones de los participantes para facilitar que cooperen en el desarrollo del estudio.

12. La metodologia del disseny de la investigació és:

A. Experimental.
B. Quasi-experimental.
C. No Experimental.
D. Qualitativa.

13. La o les variables independents en el 'Disseny Tipus A' són:

A. El grau ansietat del pacient en la fase de línia base (lleu / no lleu).
B. La variable Grup de subjectes (experimental/control).
C. El sexe i el grup de subjectes.
D. Les fases de línia base i la de post-intervenció.

14. Els efectes del entrenament cognitiu:

A. Es mesuren en la línia base.
B. Representen a la variable dependent del disseny.
C. És una variable estranya que es necessari controlar.
D. Totes las opcions anteriors són verdaderes

15. Una variable estranya que s' ha controlat per constància:

A. La fase de la malatia de Alzheimer.

B. El sexe dels pacients.
C. Tractament farmacològic prescrit.
D. Les opcions A i C són verdaderes.

3) Autoevaluación 3. Cuestionario

16. Teniendo en cuenta los siguientes resultados, se puede concluir que:

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente:NOPpost

F	gl1	gl2	Sig.
1,495	5	302	,191

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intersección + NOPpre + Parenting + Clima + Parenting * Clima

A. Las diferencias entre las puntuaciones medias de un diseño de cinco grupos no son estadísticamente significativas.
B. Se cumple el supuesto de homocedasticidad o igualdad de las varianzas de los seis grupos del diseño.
C. Existe heterocedasticidad o falta de homogeneidad entre las varianzas de los cinco grupos del diseño.
D. Las tres opciones anteriores son falsas.

17. Cuando el investigador o investigadora no conoce qué variable extraña podría contaminar los resultados la mejor opción es:

A. La técnica de la constancia de la variable extraña.
B. La técnica de la eliminación de la variable extraña.
C. La técnica de la asignación aleatoria de la variable extraña.
D. La técnica del bloqueo de la variable extraña.

18. Teniendo en cuenta los resultados siguientes, se puede concluir que:

Comparaciones múltiples

NOPpost
DHS de Tukey

(I)Clima	(J)Clima	Diferencia de medias (I-J)	Error típ.	Sig.	Intervalo de confianza 95%	
					Límite inferior	Límite superior
Luminosidad	Oscuridad	-1,86*	,741	,034	-3,60	-,11
	Tiniebla	-,84	,698	,455	-2,48	,81
Oscuridad	Luminosidad	1,86*	,741	,034	,11	3,60
	Tiniebla	1,02	,676	,288	-,57	2,61
Tiniebla	Luminosidad	,84	,698	,455	-,81	2,48
	Oscuridad	-1,02	,676	,288	-2,61	,57

Subconjuntos homogéneos

NOPpost

DHS de Tukey[a,b,c]

Clima	N	Subconjunto	
		1	2
Luminosidad	86	11,47	
Tiniebla	126	12,30	12,30
Oscuridad	96		13,32
Sig.		,463	,318

A. No se logra rechazar la hipótesis nula en ninguna diferencia de medias.
B. Las diferencias estadísticamente significativas solamente se encuentran entre luminosidad y tiniebla, siendo mayor la puntuación en la condición de tiniebla.
C. Las diferencias estadísticamente significativas solamente se encuentran entre luminosidad y oscuridad, siendo menor la puntuación en la condición de oscuridad.
D. Las diferencias estadísticamente significativas solamente se encuentran entre luminosidad y oscuridad, siendo mayor la puntuación en la condición de oscuridad.

19. El error de tipo I se puede cometer cuando:

A. Los resultados obtenidos no son estadísticamente significativos.
B. El valor p de probabilidad es menor a 0.05.
C. Se mantiene la hipótesis nula.
D. El valor de alfa es menor a 0.05.

20. Cometer Error de Tipo II es imposible cuando:

A. Los resultados obtenidos no son estadísticamente significativos.
B. El valor p de probabilidad es menor a 0.05.
C. Se mantiene la hipótesis nula.
D. El valor de alfa es menor a 0.05.

21. Supongamos que una investigadora lleva a cabo un estudio y los resultados del efecto del tratamiento A del análisis de varianza que ofrece el SPSS señalan que $F(5, 123) = 43$, $p = .000$. Qué conclusión habrá obtenido el investigador o investigadora con su análisis:

A. No se puede rechazar la hipótesis nula porque $p = 0$.
B. Hay diferencias estadísticamente significativas entre los grupos, pero no se sabe entre qué pares de medias.
C. Se mantiene la hipótesis nula.
D. Las opciones A y C son correctas.

22. $F(5, 123) = 43$: ¿cuántos grupos tiene el diseño entre-grupos unifactorial?

A. 3.
B. 4.
C. 5.
D. 6.

23. Si el diseño entre-grupos unifactorial: $F(5, 123) = 43$, ¿cuántas comparaciones de medias dos a dos simples no redundantes se podrían realizar?

A. 3.
B. 15.
C. 16.
D. 18.

24. El diseño entre-grupos unifactorial: $F(5, 123) = 43$. Si la investigadora decide comparar el último grupo con el resto y realiza solamente cinco comparaciones simples, qué procedimiento de hipótesis específicas sería el más adecuado desde el punto de la validez de conclusión estadística:

A. Bonferroni.
B. Tukey.
C. Dunnett.
D. Las opciones B y C serían correctas.

25. La técnica de control de la asignación aleatoria implica:

A. Aumentar la variabilidad de los grupos antes de administrar el tratamiento.
B. Aumentar el efecto del tratamiento aplicado.
C. Disminuye el valor del tamaño del efecto y el valor p de probabilidad.
D. Disminuir la probabilidad de diferencias previas al tratamiento entre los grupos del diseño de investigación.

26. ¿Qué definición de valor *p* de probabilidad es la adecuada?

A. El valor de alfa o más extremo asumiendo que la hipótesis nula es verdadera.
B. La probabilidad del resultado obtenido o más extremo si la hipótesis nula es verdadera.
C. La probabilidad del resultado obtenido o más extremo si la hipótesis nula es falsa.
D. La probabilidad de que la hipótesis nula sea falsa dados unos datos o datos más extremos.

27. La potencia estadística se refiere a la:

A. Probabilidad de mantener la hipótesis nula siendo realmente falsa.
B. Probabilidad de detectar una diferencia cuando realmente existe.
C. Probabilidad de que la hipótesis nula sea verdadera.
D. Las opciones B y C son correctas.

4) Autoevaluación 4. Estado emocional

SUPÒSIT D'INVESTIGACIÓ 2. En el treball de Beneyto i García (2012) es pretén contrastar si un tret positiu de personalitat com és l'optimisme podria minimitzar el biaix en el record de paraules després d'induir una emoció determinada. En un nou estudi elaborat per la Universitat de València, es va induir un determinat estat emocional a 12 adolescents homes que es van seleccionar aleatòriament de un Centre Juvenil d' Oci (distribuïts aleatòriament i de forma equilibrada en tres grups en funció de si l'estat emocional induït era positiu (grup 1), negatiu (grup 2) o neutre (grup 3).

En l'experiment, en primer lloc, tots els subjectes visionaven una mateixa llista de 60 paraules amb valència neutra, instant-los a recordar aquestes paraules. Tots els participants van rebre les mateixes instruccions: "aneu a participar en un experiment de memòria. Quan siguis preparat se't presentarà una sèrie de paraules sobre les que versarà la prova". Quan el subjecte ho indicava es donava pas a la presentació amb la llista de paraules.

Un cop finalitzat el visionat de les paraules, se li indicava que anava a veure una pel·lícula curta i que intentés ficar-se al màxim en l'escena. Depenent de la condició experimental a la qual estigués adscrit es va visionar el curt corresponent. Finalitzat el visionat de la pel·lícula, se'ls va lliurar un full en blanc indicant-los que disposaven

de 3 minuts per anotar totes les paraules que recordessin de la presentació que prèviament havien vist.

Per a la inducció emocional es van emprar tres curts de pel·lícula extretes de l'estudi de Rottenberg, Ray i Gross (2007). Per induir l'emoció positiva es va emprar el curt "Quan Harry va trobar la Sally", que presenta una situació còmica en una cafeteria amb una durada de 2,357 minuts. Per a la emoció negativa es va emprar el curt "Crida Llibertat", el qual presenta una matança de tall racista i té una durada de 2,361 minuts. Finalment, per l'emoció neutra es va emprar el curt "Sticks", que és una presentació semblant a un estalvi de pantalla d'ordinador on van apareixent uns bastons a la pantalla; aquesta pel·lícula no provoca cap emoció i va tenir una durada de 2,355 minuts. Per mesurar el nivell d'optimisme es va utilitzar l'instrument LOT-R (Scheier, Carver i Bridges, 1994) en la versió adaptada i validada al castellà de Fernández i Bermudes (1999). Per obtenir els grups experimentals es va dividir als participants en dos grups en virtut de la puntuació obtinguda en el qüestionari LOT-R: alt optimisme amb una mitjana igual o superior a 22 (grup b1) i baix optimisme amb una mitjana inferior a 22 (grup b2). La hipòtesi d'investigació planteja que els subjectes que tenen un estat emocionat positiu recordaran més paraules però, aquest efecte estarà moderat per un grau d'optimisme positiu. Els resultats trobats indiquen una tendència dels més optimistes a recordar major nombre de paraules però si l'estat emocional en què es troben també és positiu ja que si és negatiu ja no es diferencien dels subjectes menys optimistes.

28. Hi ha manipulació de la o les variables independents?

A. Hi ha dos variables independents manipulades.
B. Les pel·lícules s'han utilitzat per manipular la variable independent.
C. La divisió del subjectes en nivell d'optimisme és una manipulació de la variable independent.
D. Les tres opcions anteriors són correctes.

29. Hi ha assignació aleatòria?

A. Sí, perquè els participants es van seleccionar aleatòriament
B. Sí, perquè els participants s'han distribuït aleatòriament.
C. Sí, perquè els participants s'han distribuït de forma equilibrada.
D. Les opcions B i C són verdaderes.

30. Quin tipus de metodologia s'ha utilitzat?

A. Experimental.
B. Observacional.
C. Quasi-experimental.
D. No experimental.

31. La variable dependent és:

A. El nombre de paraules recordades.
B. El grau d'optimisme.
C. La duració de la pel·lícula.
D. La inducció emocional.

32. Les variables independents són:

A. El nombre de paraules recordades i la duració de la pel·lícula.
B. El grau d'optimisme i el nombre de paraules recordades.
C. La duració de la pel·lícula i el sexe.
D. La inducció emocional i el grau d'optimisme.

33. Respecte a la metodologia, es tracta d'un disseny:

A. Totalment aleatori.
B. No aleatori.
C. Parcialment aleatori.
D. No es pot saber amb les dades del supòsit.

34. Què variable o variables s'han controlat per constància

A. La valència de les paraules de la llista.
B. La durada de la pel·lícula.
C. El sexe.
D. Totes les opcions anteriors son verdaderes.

35. Segons l'hipòtesi, quin grup recordarà més paraules:

A. Grup a1b1.
B. Grup a1b2.
C. Grup a2b1.
D. Grup a2b2.

5) Autoevaluación 5. Trastorno obsesivo-compulsivo

SUPÒSIT D'INVESTIGACIÓ 3. El trastorn obsessiu-compulsiu (TOC) és un trastorn heterogeni pel que fa al contingut de les obsessions i compulsions. El rentat, una de les compulsions més freqüents, s'ha associat a diferents variables com: la por al contagi físic o mental, la por a l'afecte negatiu i a la pèrdua de control, un elevat

perfeccionisme o l'evitació de la sensació de brutícia o de inacabat. Un grup d'investigadors tenen el propòsit d'estudiar la simptomatologia psicopatològica i la seua resposta al llarg de tres fases d'intervenció cognitiva per tractar el TOC: fase de pretest, fase primera de avaluació i fase de post-test. L'estudi compta amb 20 subjectes, homes, i el tractament té una durada de cinc mesos. També es mesura l'edat. Quant més alta és la puntuació, més psicopatologia tenen els subjectes. El valor de alfa en aquesta investigació es de α = .01.

36. **Assenyala les variables dependents i independent, i possibles variables extranyes**
37. **Assenyala la metodologia, *N*, *n*, *S*, tipus de disseny**

6) Autoevaluación 6. Prejuicio manifiesto y sutil

SUPÒSIT D'INVESTIGACIÓ 4. En el treball de Molero, Navas i Morales (2001) se assenyalen els següents comentaris: "els conceptes de "prejudici manifest" i "prejudici subtil" de Pettigrew i Meertens (1995) impliquen "l'exclusió social del grup objecte de prejudici encara que a través de diferents vies. El prejudici manifest ho fa a través del rebuig directe i sense pal·liatius dels membres de l'exogrup per considerar-los "biològicament inferiors". El prejudici subtil condueix a un rebuig indirecte que es justifica per la defensa dels valors tradicionals que els immigrants qüestionen o no comparteixen, i l'exageració de les diferències culturals entre la societat d'acollida i la d'arribada, entre "nosaltres" i "ells ". Tot això porta a la negació d'emocions positives cap als membres de l'exogrup. Per això la persona amb prejudici subtil no té, o almenys no expressa, emocions negatives cap als immigrants, però és incapaç també de manifestar emocions positives cap a ells. Cal assenyalar que la persona que té prejudici subtil cap a un determinat grup no és conscient del seu prejudici i de les conductes discriminatòries que aquest prejudici pot arribar a produir" (p. 20).

Un grup d'investigadors esta interessat pel tema del prejudici subtil i desitgen analitzar la seua relació amb les emociones positives. Els investigadors construeixen la tipologia de prejudici (igualitari, subtil i fanàtic) i plantegen analitzar la seua relació amb les puntuacions d'una escala de emocions positives.

La mostra esta formada per 12 estudiants de 4º de l' ESO, seleccionats aleatòriament d'un institut, 4 homes i 4 dones. La seua hipòtesi d'investigació planteja que hi haurà un efecte d'interacció les variables.

38. Si un del factors és el sexe (més prejudici en els homes), qui serà el segon factor?.

A. El tipus de emocions: positives o negatives.
B. La tipologia de prejudici.
C. El curs acadèmic.
D. La intensitat de l'emoció positiva manifestada.

39. La variable dependent és:

A. El tipus de emocions: positives o negatives.
B. La tipologia de prejudici.
C. El curs acadèmic.
D. La intensitat de l'emoció positiva manifestada.

40. La metodologia de la investigació és:

A. Experimental amb restriccions en la aleatorització.
B. Experimental.
C. Quasi-experimental.
D. No experimental.

7) Autoevaluación 7. Efecto de los payasos de hospital

SUPÒSIT D'INVESTIGACIÓ 4. Un equip d'investigació desitja estudiar l'efecte de la intervenció amb pallassos en la avantsala d'operacions sobre l'ansietat dels xiquets respecte a un grup de control que no rep pre-medicació sedant ni cap altre tipus de teràpia i un grup que rep un ansiolític ('midazolam'). Els xiquets s'assignen a la condició d'investigació en funció del criteri del metge i la severitat de la seua malaltia.

Seleccionen a 12 pacients entre 16 anys i s'utilitza un disseny ortogonal o equilibrat.

La simptomatologia de ansietat es mesura amb un instrument tipus termòmetre on el xiquet ha de valorar de 0 a 22 el seu nivell de ansietat. Tots els participants són homes.

La hipòtesi de treball planteja que el nivell d'ansietat dels xiquets del grup experimental que rep la intervenció dels pallassos d'hospital serà diferent al dels xiquets de la resta dels grups.

41. La variable dependent és:

A. Grups d'intervenció.
B. Simptomatologia depressiva.
C. Simptomatologia ansiosa.

42. Hi ha alguna variable estranya pertorbadora?

A. El sexe.
B. La gravetat de la malaltia.
C. La edat.

43. Hi ha alguna variable estranya controlada?

A. El sexe.
B. La gravetat de la malaltia.
C. El metge.

44. La metodologia de la investigació és:

A. Experimental.
B. Quasi-experimental.
C. No experimental.

45. Quines tècniques de control s'han aplicat en el disseny:

A. Aleatorització.
B. Constància.
C. Les opcions de resposta A i B són correctes.

8) Autoevaluación 8. Efecto de los fármacos

SUPUESTO DE INVESTIGACIÓN 5. Una investigadora pretende comprobar la eficacia de un fármaco Z para incrementar la extroversión de los individuos. 12 voluntarios son asignados aleatoriamente a una de las tres condiciones experimentales siguientes. Se supone que el fármaco Z producirá mejores resultados que cualquiera de las otras dos sustancias que actualmente están comercializadas. El objetivo del estudio es comparar el efecto del fármaco Z respecto al fármaco A y el fármaco B. Los resultados del experimento son los siguientes.

Condición	Puntuaciones
Fármaco A	26, 22, 25, 23
Fármaco B	26, 25, 19, 22
Fármaco Z	35, 30, 32, 35

1. Qué procedimiento de análisis procede dada la hipótesis de investigación.
2. Atendiendo únicamente a la tendencia de las medias, ¿se observa el planteamiento de la hipótesis sustantiva?
3. Ejecutar el análisis de la varianza (ANOVA) y aplicar la prueba de contraste de hipótesis específica que sea más adecuada.
4. Realizar un informe de los hallazgos.

SUPUESTO DE INVESTIGACIÓN 6. Supongamos ahora que el investigador o investigadora hubiese planteado analizar el efecto del fármaco Z respecto a la media de los fármacos A y B. ¿Qué procedimiento de contraste de hipótesis específicas hubiese sido más adecuado?

9) Autoevaluación 9. El sueño (1)

SUPUESTO DE INVESTIGACIÓN 7. El sueño es muy importante en el desarrollo infantil y favorece la maduración. Los patrones electroencefalográficos (EEG) durante el sueño presentan cambios que, si no se tienen en cuenta, pueden confudirse con actividad clínica paroxística o alteraciones bruscas del trazado, aunque son fisiológicos, especialmente en los niños. En general, el EEG es uno de los estudios que suelen realizarse en los niños con problemas de aprendizaje ya que algunas investigaciones señalan que existe actividad paroxística (*hipersincronía hipnagógica*) en los niños con este problema. Una investigadora está interesado por el estudio de la relación entre la hipersincronía hipnagógica y las dificultades del aprendizaje. En un primer momento de su investigación planteó el siguiente estudio. Considerando que existe una relación lineal directa entre las variables objeto de estudio, de tal manera que mayor hipersincronía mayores son los problemas de aprendizaje, selecciona nueve niños de cinco años. Los nueve sujetos habían sido diagnósticados en sus Centros Escolares con un problema de trastornos del aprendizaje antes de comenzar el estudio. Aleatoriamente tres de ellos no recibieron

ningún fármaco actuando como grupo de control (a_1), otros tres recibieron asistencia psicológica para su trastorno escolar (a_2) y el resto de niños recibió un fármaco que provocaba la relajación neuronal (a_3). Una vez finalizado el tratamiento la investigadora midió el nivel electroencefalográfico computando el número de alteraciones bruscas del trazado o cambios paroxísticos producidos. Su hipótesis señala que los niños que no son sometidos a ningún tipo de intervención tendrán un mayor número de alteraciones en el EEG, siendo el número menor cuando reciben tratamiento psicológico. Los resultados de parte del análisis de la varianza y de la estimación de los efectos se presentan a continuación.

Ejercicios: completa los resultados, ejecuta el ANOVA y redacta los resultados del informe de investigación.

Fuente	SC	gl	MC	Razón F	p
Entre					0.05
Error	12				
Total					

Tabla de efectos

α_1 Fármaco	0
α_2 Tratamiento psicológico	-4
α_3 Grupo control	
	M = 10

<u>CUESTIONARIO DEL SUPUESTO DE INVESTIGACIÓN 7</u>: responder desde la pregunta 45 a la pregunta 53:

46. La metodología de la investigación se considera:

A. Experimental porque se manipula la variable independiente y hay asignación aleatoria de las unidades experimentales a los grupos de la variable independiente.
B. Experimental porque se manipula la variable independiente y hay asignación aleatoria en el orden de la administración de los grupos de la variable independiente.

C. Cuasi-experimental porque aunque existe asignación aleatoria a los grupos de tratamiento, sin embargo no se manipula la variable independiente de tratamiento.
D. No se puede responder a la pregunta.

47. La validez externa del estudio anterior queda garantizada ya que:

A. Aleatoriamente los sujetos son asignados a las condiciones experimentales.
B. Aleatoriamente son seleccionados los niveles de la variable independiente de tratamiento.
C. Aleatoriamente los sujetos son seleccionados del conjunto de la población.
D. No se puede responder a la pregunta.

48. La variable dependiente (o variables dependientes) es:

A. Número de alteraciones paroxísticas.
B. Número de trastornos que presenta el niño.
C. La edad del niño.
D. d) No se puede responder a la pregunta.

49. La variable independiente (o variables independientes) es:

A. La cantidad de alteraciones paroxísticas que presenta el niño (ninguna, entre 3 y 4, más de cuatro).
B. El grupo de intervención al que es sometido el niño (control, tratamiento psicológico, fármaco).
C. La edad del niño (cinco años, más de cinco, menos de cinco).
D. No se puede responder a la pregunta.

50. La Suma de Cuadrados del Tratamiento es:

A. 96.
B. 116.
C. 506.
D. No se puede responder a la pregunta.

51. En esta primera fase de su investigación la decisión estadística permite concluir que:

A. Existen diferencias estadísticamente significativas ya que $F(2, 8) = 38$, $p < 0.05$.
B. Existen diferencias estadísticamente significativas ya que $F(2, 6) = 43.8$, $p < 0.05$.
C. Existen diferencias estadísticamente significativas ya que $F(2, 6) = 24$, $p < 0.05$.
D. No se puede responder a la pregunta.

52. La investigadora decide calcular una prueba de comparación de medias que le permita analizar el número total posible de comparaciones simples entre pares de medias. Cuántas comparaciones simples son posibles analizar sin ser redundantes:

A. 3.
B. 4.
C. 2.
D. d) No se puede responder a la pregunta.

53. Si la investigadora decide realizar comparaciones de medias, la prueba estadística más adecuada estará guiada por:

A. Controlar la tasa de error de Tipo I y ser la más potente.
B. Controlar la tasa de exceso y ser exacta.
C. Controlar la tasa de error de Tipo I y trabajar con el menor Error de Tipo II.
D. No se puede responder a la pregunta.

54. Si la investigadora hubiese planteado la comparación de la puntuación media del grupo control frente a las puntuaciones medias de todos los demás grupos de tratamiento, realizando comparaciones simples que hubiesen estado definidas a priori, la prueba de comparación de medidas más adecuada sería:

A. La prueba de Dunnett siempre que se realicen 2 comparaciones de medias.
B. La prueba de Dunnett siempre que se realicen 3 comparaciones de medias.
C. La prueba de Tukey siempre que se realicen más de 4 comparaciones de medias.
D. No se puede responder a la pregunta.

10) Autoevaluación 10. El sueño (2)

SUPÒSIT D'INVESTIGACIÓ. El somni és molt important en el desenvolupament infantil i afavoreix la maduració. Els patrons electroencefalogràfics (EEG) durant el son presenten canvis que, si no es tenen en compte, poden confondre amb activitat clínica paroxística o alteracions brusques del traçat, encara que són canvis fisiològics, especialment en els xiquets. En general, l'EEG és un dels estudis que solen realitzar-se en els xiquets amb problemes d'aprenentatge ja que algunes investigacions assenyalen que hi ha activitat paroxística (hipersincronía hipnagógica) en els xiquets amb aquest problema. Un investigador està interessat per l'estudi de

la relació entre la hipersincronía hipnagógica i les dificultats de l'aprenentatge. En un primer moment de la seua investigació va plantejar el següent estudi. Considerant que hi ha una relació lineal directa entre les variables objecte d'estudi, de tal manera que major hipersincronía majors són els problemes d'aprenentatge, selecciona nou xiquets de cinc anys. Els nou subjectes havien estat diagnosticats en els seus Centres Escolars amb un problema de trastorns de l'aprenentatge abans de començar l'estudi. Aleatòriament tres d'ells no van rebre cap fàrmac actuant com a grup de control (a1), altres tres van rebre assistència psicològica per a la seua trastorn escolar (a2), i la resta de xiquets va rebre un fàrmac que provocava la relaxació neuronal però amb moltes efectes secundàries de salut (a3). Un cop finalitzat el tractament l'investigador va mesurar el nivell electroencefalogràfic computant el nombre d'alteracions brusques del traçat o canvis paroxístics produïts. La seua hipòtesi assenyala que els xiquets que no són sotmesos a cap tipus d'intervenció tindran un major nombre d'alteracions en l'EEG, sent el nombre menor quan reben tractament psicològic. L'investigador opta per fer un anàlisi d'ANOVA en primer lloc. Els valors del efectes són: $\alpha 1 = 2$ i $\alpha 2 = -2$.

	Grupo	Medida
1	1	9
2	1	6
3	1	9
4	2	4
5	2	4
6	2	4
7	3	6
8	3	4
9	3	8

Estadísticos descriptivo

Variable dependiente: Medida

Grupo	Media	Desviación estándar
control	8,00	1,732
terapia	4,00	,000
fàrmac	6,00	2,000
Total	6,00	2,179

Comparaciones múltiples

Variable dependiente: Medida

HSD Tukey

(I) Grupo	(J) Grupo	Diferencia de medias (I-J)	Error estándar	Sig.
control	terapia	4,00*	1,247	,042
	fàrmac	2,00	1,247	,315
terapia	control	-4,00*	1,247	,042
	fàrmac	-2,00	1,247	,315
fàrmac	control	-2,00	1,247	,315
	terapia	2,00	1,247	,315

55. La variable dependent és:

A. Nombre d'alteracions paroxístiques.

B. Nombre d'alteracions brusques del traçat.
C. Les opcions A i B són correctes.

56. La puntuació pronosticada pel model de la hipòtesis alternativa per al subjecte amb puntuació de 4 (nº S =5):

A. 0.
B. 2.
C. 4.

57. Suma de Quadrats de l'error és:

A. 14.
B. 12.
C. 13.

58. La decisió estadística condueix a (posa els graus de llibertat de la prova F i el valor de p respecte al valor de alfa):

A. Mantenir la hipòtesi nul·la, F(,) = 5.14, *p* _____.
B. Rebutjar la hipòtesi nul·la, F ,) = 5.14, *p* _____.
C. Rebutjar la hipòtesi nul·la, F(,) = 5.57, *p* _____.

59. Redacta resultats complets del supòsit segons el format d' un informe d'investigació tipus APA i utilitzant el contingut del supòsit: Si et falta alguna dada d'alguna prova, deixa-ho en blanc però posa el seu símbol.

_____
_____
_____
_____
_____
_____
_____

60. Com a professional, segons els resultats, quina opció d'intervenció recomanaries i per què:

_____
_____

11) Autoevaluación 10. Conocimiento abstracto

SUPUESTO DE INVESTIGACIÓN 8. SUPUESTO DE INVESTIGACIÓN. El conocimiento abstracto, como por ejemplo el matemático, es muy difícil de adquirir y

sobre todo de aplicar en situaciones nuevas. Existe la creencia de que el aprendizaje con ejemplos concretos de la vida diaria facilita la adquisición del conocimiento. Sin embargo, los resultados de las investigaciones señalan que el aprendizaje de conceptos matemáticos no mejora con el uso de "buenos ejemplos concretos" (Kaminski, Sloutsky, y Heckler, 2008). Un grupo de investigadores decide llevar a cabo un estudio para analizar cómo los estudiantes universitarios aprenden un simple concepto matemático bajo dos condiciones. En la condición 1 (a_1) la enseñanza se realiza con simbología matemática y en la condición 2 (a_2) se realiza con ejemplos cotidianos concretos (simbología concreta). Los investigadores miden los errores de los alumnos en una prueba matemática donde el máximo de error es veinte. Seleccionan aleatoriamente a 8 alumnos de primero de Psicología de la Universidad de Alicante y reciben de forma aleatoria un tipo de aprendizaje u otro. Los datos del estudio y sus estadísticos descriptivos son los siguientes y la media cuadrática del error es 18.667.

	Variable1	variable2
1	1	7
2	1	7
3	1	8
4	1	10
5	2	14
6	2	7
7	2	19
8	2	20

Variable1	Media	Desviación estándar	N
1	8,00	1,414	4
2	15,00	5,944	4
Total	11,50	5,477	8

61. La variable dependiente es:

A. Grado de aprendizaje del concepto matemático.
B. Cantidad de errores cometidos en la prueba.
C. Las opciones A y B son correctas.

62. La puntuación pronosticada por el modelo de la hipótesis nula para el sujeto que tiene una puntuación de 20 es:

A. 15.
B. 11.5.
C. 0.

63. Realitza el desenvolupament de l'equació estructural que planteja la hipòtesi:

Y = M + A + E

64. La puntuación pronosticada por el modelo de la hipótesis alternativa para el sujeto que tiene una puntuación de 20 es:

A. 15.
B. 11.5.
C. 0.

65. Los valores estimados de los efectos son:

A. $\alpha_1 = 3.5$, $\alpha_2 = -3.5$.
B. $\alpha_1 = 0$, $\alpha_2 = -3.5$.
C. $\alpha_1 = -3.5$, $\alpha_2 = 3.5$.

66. Completa la tabla del ANOVA. Será necesario conocer la siguiente información para tomar la decisión estadística: mantener o rechazar la hipótesis nula:

- *alfa* = _____
- $F_t\,(\alpha, gl_A, gl_E) \equiv F(\ ,\ ,\ )$ = _____
- $F_e\,(gl_A, gl_E) \equiv F(\ ,\ )$ = _____
- *p* = Si el valor de *p* se obtiene manualmente con las tablas del estadístico habrá que optar por situar "$p \leq$ alfa" o por "$p >$ alfa". Por lo tanto, una vez llevado a cabo el contraste, el valor de *p* _____ alfa.

Tabla Anova. **Y = M + A + E**

Fuente de varianza	SC	gl	MC	F	p	η^2
Efecto:					☐.05	

Error:						
Total						

67. Suma de Cuadrados del error es:

A. 112.00.
B. 149.34.
C. 37.33.

68. Estima la proporción de varianza explicada: η2 = _____ E interpreta ese valor:

_____

69. Realiza el ejercicio con el programa SPSS (o con el programa JASP) y señala qué valor exacto de *p* está asociado al estadístico obtenido con los datos del ejercicio en la distribución muestral de la hipótesis nula (H0): F (glA, glE). F (,) = _____, p = _____

70. La decisión estadística conduce a:

A. Rechazar la hipótesis nula, $F(1, 6) = 5.99$.
B. Mantener la hipótesis nula, $F(1, 8) = 3.57$.
C. Mantener la hipótesis nula, $F(1, 6) = 5.25$.

71. En este diseño podria darse un error de Tipo:

A. Error de Tipo I.
B. Error de Tipo II.
C. Desconocemos con los datos que tenemos que tipo de error se ha podido cometer .

72. La metodología es: _____Por qué? Utiliza para la explicación el contenido del supuesto.

_____

_____

73. Según los resultados del ANOVA se puede concluir que:

A. Kaminski, Sloutsky, y Heckler (2008) no tenían razón en su postura científica.
B. El aprendizaje con ejemplos concretos de la vida diaria facilita la adquisición de conocimientos.
C. No se detecta un efecto estadísticamente significativo de la variable tipo de aprendizaje.

74. Redacta los resultats del supuesto 1 utlizando el formato de un informe de investigación tipo Manual del APA y utilizando el contenido del supuesto:

_____
_____
_____
_____
_____
_____

12) Autoevaluación 11. Cuestionario

75. Dentro del contraste estadístico, valor p es ... (Anotar Verdadero V / Falso F):

A. La probabilidad de que la decisión estadística sea correcta. ____
B. La probabilidad del azar. ____
C. La probabilidd de que la hipótesis nula sea verdadera, dados los datos de la investigación. ____
D. La probabilidd de que la hipótesis alternativa sea verdadera, dados los datos de la investigación. ____
E. La probabilidad de error que puede haber en la prueba estadística cuando se calcula el estadístico del contraste: ____
F. La probabilidad del resultado obtenido en el estudio, si la hipótesis nula es falsa. ____
G. La probabilidad del resultado obtenido en el estudio, si la hipótesis nula es cierta. ____
H. La probabilidad de que una hipótesis sea rechazada o mantenida en el contraste estadístico. ____
I. La probabilidad que tiene el resultado obtenido en el estudio, o un dato más extremo, si la hipótesis nula es cierta. ____
J. La probabilidad de la significación de la hipótesis nula cuando se compara con un alfa. ____
K. La probabilidad de que el tamaño del efecto sea importante. ____
L. La probabilidad de que el tamaño del efecto sea útil. ____

76. Contesta:

A. La probabilitat de rebutjar la H0 quan es falsa s'anomena: _____

B. La probabilitat de rebutjar la H0 quan es certa s'anomena: _____ La probabilitat de les dades observades en l'estudi, sent la hipòtesi nul·la certa és: _____

D. La probabilitat de mantenir la H0 quan es certa s'anomena: _____ La probabilitat de mantenir la H0 quan es falsa s'anomena: _____ La probabilitat del resultat de la prova estadística (o de un resultat més extrem), baix el supòsit d'una distribució on no hi ha efecte s'anomena: _____ La probabilitat de no rebutjar la H0, sent la H1 certa s'anomena: _____ La probabilitat de rebutjar una H0 verdadera és: _____ **¿Qué es el valor p de probabilidad en un contraste de hipótesis estadísticas?**

_____

78. ¿Qué es alfa en un contraste de hipótesis estadísticas? ¿Y su complementario?

_____

79. ¿Qué es beta en un contraste de hipótesis estadísticas? ¿Y su complementario?

_____

80. Segon les normes APA, es recomana que els resultats estadísticament no significatius han d'escriure amb:

A. El valor *p* exacte.
B. El valor *p* com > o < al valor de alfa que s'utilitze en l'estudi (per exemple, $p < .05$).
C. Anotar les sigles '*ns*' ('*no significance*') quan és no estadísticament significatiu.

81. La significació substantiva dels resultats fa referència a:

A. La presencia de diferències que són estadísticament significatives.
B. La significació indicada pel interval de confiança de la grandària de l'efecte.
C. La utilitat clínica.

82. La potència estadística:

A. És la probabilitat que té el resultat d' una prova estadística per rebutjar una hipòtesi nul·la falsa.
B. És la probabilitat que té el resultat d' una prova estadística per rebutjar una hipòtesi alternativa verdadera.
C. És la probabilitat que té el resultat d' una prova estadística per rebutjar una hipòtesi nul·la que realment és verdadera.

83. El error de Tipus II:

A. Podria cometre si la hipòtesi alternativa s'accepta.
B. Podria cometre quan s'accepta la hipòtesi nul·la.
C. Només es pot cometre quan la hipòtesis nul·la es manté.

84. En un disseny unifactorial, el resultat de la prova F(4, 221) és 101.25. Quantes observacions té el disseny:

A. 222.
B. 226.
C. 884.

85. El resultat de la prova F(4, 221) és 101.25, p = .065. La prova de Tukey seria la més adequada si es fan:

A. Es necessari fer 10 comparacions simples.
B. Es necessari fer 15 comparacions simples.
C. No cal fer cap comparació múltiple.

86. Per què el valor p no és la probabilitat de la hipòtesi nul·la

_____

_____

_____

87. Cuanto menor es el valor de *p*:

A. Menos probable es el resultado en el modelo de la hipòtesis nula.
B. Más probable es el resultado en el modelo de la hipòtesis nula.
C. Menos probable es el resultado en el modelo de la hipòtesis alternativa.

88. Un resultat estadísticament no significatiu assenyala que:

A. El resultat no és important.
B. Podria ser un resultat important.
C. La hipòtesi nul·la es certa.

89. El resultat de la prova F(4, 10) és 3.02. Podem rebutjar la hipòtesi nul·la:

A. No.
B. Sí.
C. Depèn de la validesa de l'estudi (del grau d'evidència del resultat).

90. Quin tipus d'hipòtesis són les més adequades perquè la prova de Tukey sigui la més idònia?:

A. Hipòtesis simples, complexes i exhaustives.
B. Hipòtesis simples i exhaustives.
C. Tukey sempre és la més adequada quan el disseny és unifactorial.

91. En una investigació s'obté un valor de p < .001, què és pot concloure?

A. La grandària de l'efecte ha segut gran.
B. L'efecte detectat és important.
C. Probablement les diferències detectades no es deuen a l'atzar.

13) Autoevaluación 12. Edad, consumo de alcohol (1)

SUPUESTO DE INVESTIGACIÓN 9. Durant l' adolescència el cervell és especialment vulnerable als efectes de l'alcohol. El consum d'hora d' alcohol pot augmentar el risc de simptomatologia psicopatològica. Pocs estudis han analitzat la relació entre consum d' alcohol i simptomatologia psicopatològica en adolescents en la població general. L' objectiu d'aquest estudi és determinar l' associació entre edat d 'inici del consum d' alcohol, sexe i símptomes psicopatològics en estudiants de batxillerat. Els símptomes es van mesurar amb la sub-escala d' Índex de Malestar anomenada SCL-90-R (Symptom Check List-Revised; Derogatis, 1983) que mesura els símptomes psicopatològics manifestats durant l'última setmana. La hipòtesi d'investigació assenyala un patró de simptomatologia diferent de manera que l' edat de començament del consum d'alcohol està relacionada amb l' Índex de Malestar però depèn del sexe de l' individu. Es va seleccionar una mostra aleatòria de tres aules d' un institut de 2º de Batxiller i es van reclutar 24 individus. Després d' excloure els estudiants majors de 18 anys i aquells que no havien begut mai (no hi havia alcohòlics diagnosticats tampoc), la mostra es va compondre finalment de 18 participants (9 van ser dones: b1). Es va registrar l'edat de començament del consum d' alcohol i posteriorment es van crear tres grups: abans dels 12 anys a1, entre 13 i 15 anys a2 i més de 15 anys (factor A). Els resultats van ser els següents

Sexe	Consum	Malestar
Dona	Abans12	83
Dona	Abans12	86
Dona	Abans12	86
Dona	Entre 12 i 15	76
Dona	Entre 12 i 15	74
Dona	Entre 12 i 15	75
Dona	Més de 15	64
Dona	Més de 15	65
Dona	Més de 15	66
Home	Abans12	56
Home	Abans12	55
Home	Abans12	54
Home	Entre 12 i 15	42
Home	Entre 12 i 15	42
Home	Entre 12 i 15	42
Home	Més de 15	34
Home	Més de 15	35
Home	Més de 15	36

Sexe	Consum	Media	Desviación estándar
Dona	Abans12	85,00	1,732
	Entre 12 i 15	75,00	1,000
	Més de 15	65,00	1,000
	Total	75,00	8,732
Home	Abans12	55,00	1,000
	Entre 12 i 15	42,00	,000
	Més de 15	35,00	1,000
	Total	44,00	8,818

92. Hi ha alguna variable estranya pertorbadora controlada per constància?

A. El sexe.
B. Consumir alcohol.
C. Les opcions A i B son falses.

93. Quin tipus de variable és l'edat de començament del consum d'alcohol en el disseny?

A. Variable independent assignada.
B. Variable pertorbadora.
C. Variable estranya controlada.

94. La metodologia de la investigació és:

A. Experimental.
B. Quasi-experimental.
C. No experimental.

Explica per què utilitzant el contingut del supòsit (no parles en general):

_____
_____
_____
_____
_____
_____
_____

95. Desenvolupa la equació estructural:

Y=_____I posa la taula d'ANOVA↓

96. La puntuació pronosticada pel model de la hipòtesis alternativa per al subjecte que té una puntuació de 83 és:

A. 85.
B. 59.5.
C. 25.5.

97. Els graus de llibertat de la interacció són:

A. 4.
B. 6.
C. 2.

98. Els valors del efecte d'interacció d' homes i consum entre 12 i 15 és:

A. $\alpha_2\ \beta_2 = -1$.
B. $\alpha_2\ \beta_2 = -2$.
C. $\alpha_2\ \beta_2 = -0.5$.

99. La Mitjana Quadràtica de l' error és:

A. 7
B. 3.5.
C. 1.17.

100. La decisió estadística de la hipòtesi d'investigació condueix a:

A. Rebutjar la hipòtesi nul·la, $F(2, 12) = 22.5$, $p < .05$
B. Mantenir la hipòtesi nul·la, $F(2, 12) = 3.85$, $p > .05$.
C. Rebutjar la hipòtesi nul·la, $F(2, 12) = 3.85$, $p < .05$.

101. El resultat de la prova de hipòtesi de la interacció AB assenyala que:

A. La hipòtesi nul·la era falsa.
B. La hipòtesi nul·la era certa.
C. La probabilitat del resultat obtingut era major a .05.

102. En aquest disseny podria donar-se un error de Tipus:

A. Error de Tipus I.
B. Error de Tipus II.
C. Error de Tipus Gamma.

103. El valor de la grandària de l' efecte de interacció de 'eta quadrat parcial' és:

A. .29.
B. .32.
C. .39.

104. Segons els resultats del ANOVA podem concloure que:

A. Hi ha un efecte d'interacció estadísticament significatiu.
B. El sexe i la edat del començament del consum d'alcohol estan vinculats amb el símptomes de psicopatologia com factors principals.
C. La edat del començament del consum d'alcohol està vinculada amb el grau d'alcoholèmia, sent el sexe una variable de bloqueig.

105. El valor *p* de probabilitat de l' efecte del Consum informa de:

A. La probabilitat de la hipòtesi nul·la.
B. La probabilitat de la hipòtesi alternativa.
C. Les respostes A i B són falses.

106. Respecte a l'efecte del Sexe es pot concloure que:

A. La hipòtesi nul·la és falsa.
B. La hipòtesi alternativa és verdadera.
C. Les respostes A i B són falses.

107. S'ha comprovat que les dades del disseny corresponen a un model:

A. Niat.
B. Additiu.
C. No additiu.

108. En quina font de variància podrien haver problemes de potència estadística:

A. Efecte principal A.
B. Efecte principal B.

C. Efecte d'interacció.

109. Els resultats de les proves d'hipòtesi específiques per a la variable Edat de començament del Consum són els següents. Quina prova s'ha aplicat:

(I) Consum	(J) Consum	Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Abans12	Entre 12 i 15	11,50*	,624	,000	9,84	13,16
	Més de 15	20,00*	,624	,000	18,34	21,66
Entre 12 i 15	Abans12	-11,50*	,624	,000	-13,16	-9,84
	Més de 15	8,50*	,624	,000	6,84	10,16
Més de 15	Abans12	-20,00*	,624	,000	-21,66	-18,34
	Entre 12 i 15	-8,50*	,624	,000	-10,16	-6,84

Se basa en las medias observadas.

A. Tukey
B. Bonferroni.
C. Dunnett.

110. Redacta els resultats del supòsit 1 segons el format d' un informe d'investigació:

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

CONCLUSIÓ

FINAL:_____

_____

SUPUESTO DE INVESTIGACIÓN 10 (Continuación del supuesto anterior). Anem a suposar que "el Factor B del disseny anterior es un factor de bloqueig i que no hi ha efecte d'interacció estadísticament significatiu (p > ,05)". La resta del disseny té les mateixes característiques i dades.

111. La hipòtesi de la investigació quina seria ara?

A. L'edat del començament del consum d'alcohol està vinculada amb psicopatologia.
B. Les dones consumen més alcohol que les homes.
C. Quan més edat té el subjecte més consum hi ha d'alcohol.

112. Ara la metodologia de la investigació és:

A. Experimental.
B. Quasi-experimental.
C. No experimental.

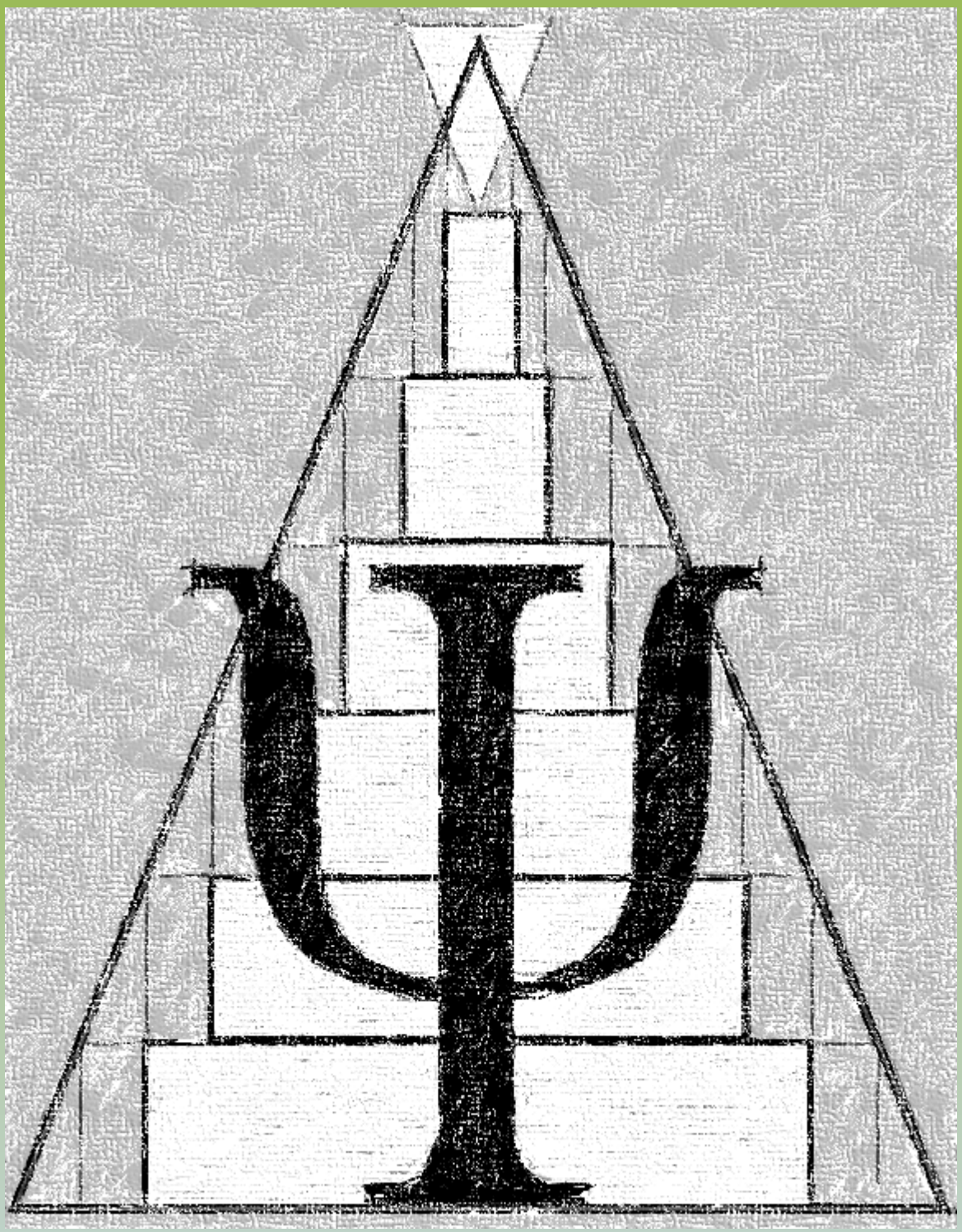
113. El resultat de l'ANOVA de la hipòtesi d'investigació del supòsit d'ara és el següent:

A. $F(2, 14) = 367.9$, $p < .05$.
B. $F(2, 14) = 388.6$, $p < .05$.
C. $F(2, 12) = 315.3$, $p < .05$.

114. Realitza l'exercici amb el SPSS i aporta el valor *p* exacte de cada font de variància de l'ANOVA:

Valor *p* de _____ = _____
Valor *p* de _____ = _____
Valor *p* de _____ = _____

Frías-Navarro, D. y Pascual-Soler, M. (Eds.) (2021). ***Diseño de la investigación, análisis y redacción de los resultados*****. Universidad de Valencia. España. https://doi.org/10.17605/osf.io/hetw2**



Edición 2ª, septiembre de 2021
DOI 10.17605/OSF.IO/KNGTP