When are sample means meaningful? The role of modern estimation in psychological science

Clinton P. Davis-Stober

University of Missouri

Jason Dana

Yale University

Jeffrey N. Rouder

University of Missouri

Abstract

Sample means are considered a foundational statistic for understanding experimental psychological data that seemingly can be used in all contexts. We argue here that contrary to common belief, sample means are not always interpretable, and there are not-so-rare cases where they are useless if not misleading. We define the sample mean as uninterpretable if in a measurement environment it is outperformed by a nonsense estimator that does not use the data to estimate the relations among experimental conditions. We consider two such nonsense estimators: one that randomizes the relations among conditions, and another that states that there are no condition effects no matter the data. We show that there are common cases where these nonsense estimators outperform sample means on average, and these may even occur when effects are detected. We argue that in these cases, sample means are uninterpretable and should be avoided. The concept of interpretability highlights regularization in model estimation, such as that used in lasso, ridge, and hierarchical Bayes techniques. We argue that not only is the interpretation of sample means context dependent, but that modern estimators should replace sample means, even for describing data, because they are interpretable in a wider range of contexts.

Progress in science is intimately tied to our ability to accurately measure and judiciously interpret variables of interest. In this paper, we take the usual case where a scientist wishes to measure a variable of interest across a number of experimental conditions. The question we ask is: *under what conditions are differences among sample means interpretable?* At first glance, it may seem that sample means, and their differences, are always interpretable as they are simple statistics commonly used to measure central tendency. Yet, as we will show, the interpretability of sample mean differences is not automatic and depends on the context. We approach the problem of interpretability from a measurement perspective.

One of the key concepts in measurement is that the measurement tool should match the characteristics of the environment and phenomena under study as well as the goals of measurement. A yardstick, for example, is appropriate for measuring the length of straight objects that subtend from about an inch to 36 inches. It is inappropriate for measuring the width of a hair or the distance to the moon. If we wished to measure the width of a hair, we would need a microscope. Of course, the microscope is inappropriate for measuring the length of shoes. Measurement tools have scales and resolutions on which their application is interpretable.

The sample mean, in particular, the differences among sample means, can also be seen as measurement tools. When they are, the question then is about their measurement resolution. This question may seem odd at first. The sample mean is perhaps the most elementary statistic, and researchers use it ubiquitously to describe the to-be-explained phenomena of interest. Moreover, differences among sample means seemingly follow immediately and underlie most common tests. In common usage, sample means and their differences seem directly interpretable without consideration of the environment to which they are applied. Sample means are rarely considered as a measurement instrument and there is little thought about when it might be uninterpretable. Here we argue that researchers need to be cognizant of the sample mean as a measurement tool, and to problematize the relations among it, the phenomena under study, the environment of data collection, and the goals of the research. As we show, some environments are so impoverished that the differences among sample means are uninterpretable and reporting them overstates their fidelity. Moreover, statistical significance of an effect provides no safeguard of interpretability, and, as will be shown, there are environments with typical power for which sample means are uninterpretable.

In this paper, we define minimum thresholds for interpretability of sample means. We argue that these thresholds serve as a fundamental limit. If they are not met, we believe that sample means should not be valued or further analyzed, and should not be the inputs to further statistical tests. We make firm recommendations—when environments are impoverished, researchers should NOT report sample means to readers. Instead, they should report regularized measures of central tendency, a concept we subsequently discuss. These regularized measures are increasingly common in sophisticated data analysis (Gelman & Carlin, 2017) precisely because they are more interpretable in these environments.

Fortunately, statements about interpretability may be assessed before data collection, and in this regard are similar to power. Yet, interpretability is different than power, reflects no inferential considerations, and, as we show, even moderately powered experiments may not meet these thresholds. One consequence is that one could reject null hypotheses using sample means that are uninterpretable by our considerations. Such a state is problematic of course, and we show the conditions under which it occurs.

To motivate the concept of interpretability of sample means, and their differences, we consider Correll, Park, Judd, and Wittenbrink’s (2002) race-based weapons priming task. In this task, participants are shown pictures of actors holding guns or similarly-shaped objects such as cell phones. The task is to decide whether the held object is a gun or not, and participants respond more often and more quickly that the object is a gun when it is held by an African American actor than a White actor. For the sake of this example, let’s assume the effect is both real and small, and that participants respond 40 ms faster when

the gun is held by African American actors than by White actors. Let's also assume each observation has a true standard deviation of 300 ms. The response-time distributions for the African American actor condition (solid line) and White actor condition (dashed line) are displayed in Figure 1A. The difference of 40 ms is quite small compared to the variability. Fortunately, researchers collect multiple observations from each condition. Let's take for example the case where the researcher collects eight observations for each actor condition, for a total of sixteen observations. There are two means of eight observations each, and the distribution of these sample means is plotted in Figure 1A as well. The reduction in variance is obvious.

The environment in this case is defined by the true effect, the true variation in observations, and the sample size. The measurement tool is the difference between the race condition sample means, and this difference serves as an estimate of the true effect. Figure 1B shows the distribution of this measurement, and it is centered at 40 ms, the true effect (black vertical line), but the resolution is quite poor, i.e., the width of this distribution is large. We can ask whether this measurement tool, the difference between the sample means, has sufficient resolution to be interpretable in this environment.

One way of assessing the interpretability of the sample mean difference is to compare it to an alternative measurement tool. The tool we chose is very crude and not useful in general. It returns the grand mean across all 16 observations as the measurement for each race condition, and the measurement of the true effect is identically zero. This estimator is certainly strange. A researcher using this tool will always conclude that the effect is identically zero regardless of the data, and for this reason, we call this measurement tool the *zero estimator*. The zero estimator is shown as the red vertical line in Figure 1B. Note that it is slightly off from the true value of 40 ms.

The zero-estimator is akin to a broken clock. It is correct only when the true effect is identically zero. But in the hypothetical race-priming experiment, the true effect is 40 ms, meaning that the zero estimator is inaccurate in high-resolution environments. We can define the accuracy of the measurement tool as the root-mean-squared error (RMSE) between the measured and the true values. Sample means are known to have favorable properties. No other unbiased estimator yields lower RMSE. Moreover, sample means track the true values well and converge on them rapidly with large sample sizes.

But these statistical facts do not mean that the sample mean outperforms the zero estimator in the above environment. Figure 1C shows the distribution of root-squared error (or absolute error) for the difference between the sample means. Small values are desirable as they indicate less error. Note how the error is quite large across most of the distribution. With just eight observations per condition, the RMS error is about 150 ms. Let's compare this degree of error to that for the zero estimator. The zero estimator has an error of 40 ms — if the true value is 40 ms and the estimate is zero regardless of data, then the error is always the same. This value is indicated with the red vertical line. This error is quite small, about 1/7th of the error from the difference between sample means.

In this environment, the difference between sample means is far less accurate a measurement than the zero estimator even though the true value is nonzero. Considering these sample means, forming conclusions from them, or even reporting them overstates their interpretability. They are fairly uninterpretable, and the lack of interpretability reflects the poor resolution from using small samples to detect small effects.

Figure 1. Estimators and error. **A.** Distributions of observations and sample means across 8 observations for the African American actor condition (solid line) and White actor condition (dashed-line). **B.** Distribution of the measurement tool, the difference in sample means. The vertical line is the zero estimator. **C.** Distribution of the absolute error for the sample mean measurement tool. The root-mean-squared error is 7 times greater for the sample mean tool than the zero estimator.

Sometimes people find it attractive to think of interpretability in terms of statistical power. Interpretability, however, is distinct from power. We need not discuss binary decisions, long term error rates, beliefs, or any other inferential system. In fact, it is a question of whether the measurement itself has the resolution to be interpretable, much as measurements from yardsticks are uninterpretable for measuring the width of a hair. The reason is the same—there needs to be a match between the environment and the instrument.

In this example, the environment is defined by the true effect, the true variability in the data, and the sample size. In practice, two of the three elements are unknown, at least to arbitrary precision. In fact, the goal is to measure these true values, and knowing them would obviate the goal. Fortunately, we do not need to define these true values to assess interpretability. Instead, one could specify a range or the order-of-magnitude of possible effect values. We ask that if effects ranged on a specified scale—say 10s or 100s or milliseconds—then, what are the sample sizes needed to insure that sample means are more accurate than the zero estimator.

In a later section we introduce a new, crude estimator that is in some cases better than the zero estimator but still too insensitive for general use. This estimator measures some important features of the data, such as the scale of possible effects, but randomizes the relations between groups or conditions. We consider sample mean differences interpretable in those environments in which it is more accurate than this new random estimator.

Before we introduce this new random estimator, we need to review some advances in modern statistical theory. These advances, which broadly go under the name of regularization in estimation, serve two purposes here. First, they help us develop the random

estimator alternative that we use to define interpretability in a measurement environment. Second, they become the basis for our tangible recommendations. We recommend that these regularization estimators be adopted broadly especially in impoverished environments.

Stein’s Paradox and Shrinkage Estimation

Although it is common to think of the sample mean as a natural or obvious choice for measurement, the seminal work of Stein, 1956 showed that the situation is more complex. Stein argued that the sample mean was not, in general, admissible, and better accuracy could be obtained with biased measurements. The result was controversial at first though widely influential today and foundational in analysis under the moniker of *regularization* (e.g., lasso regression, ridge regression, Bayesian shrinkage, see Lehmann & Casella, 1998). Here, we review Stein’s result and its consequences for understanding the measurement properties of the sample mean.

Consider a researcher who wishes to know whether reading speed of words is affected by the color these words are displayed in. The researcher chooses colors of red, green, yellow and blue, and presents words, one-at-a-time, to participants. A seemingly natural statistic to report is the sample mean for each color and the differences among them. Note that each sample mean is an isolated, separate measure. The sample mean for red words is not informed by the responses to blue words. Stein’s paradox reveals that this approach is not ideal. If one knew from the other three colors that response times tend to be on the scale of half a second or so, then this information should be used in measuring the response to red words. That is, it should be used to calibrate the measurement instrument itself.

Figure 2 shows an example. The top row shows true condition response time means as a function of color. Red words are read more quickly than green words and so on. The data are shown next, and they are noisy draws from the respective true values. Below them are sample means, which are less perturbed from the true means than the data themselves. The bottom row shows Bayesian hierarchical-model estimates (Rouder, Morey, Verhagen, Swagman & Wagenmakers, 2017) which are different than the sample means. The hierarchical model uses the combined scale of observations to “shrink” estimates toward the grand mean. These shrunk estimates here are more accurate than the sample means. This advantage holds generally—estimators that pool scale information across conditions are more accurate than corresponding estimators that do not.

This shrinkage property has motivated much of the work in modern estimation. Whereas pooling information always improves accuracy, the amount to pool is necessarily a function of prior assumptions. Different methods, say empirical Bayes, hierarchical Bayes, lasso, and ridge regression, vary in these prior assumptions. Even so, all are based on the above insight.

A surprising consequence of this work, known as Stein’s paradox, is that any information about the scale, no matter how small, helps improve accuracy. This argument for scaling measurements may be taken to an extreme, and doing so helps build intuition. Consider the following popular example where the goal is to measure the true mean of three populations: the weight of all hogs in Montana, the per-capita tea consumption in Taiwan, and the height of all redwood trees in California. These values could be estimated separately by getting a sample of hogs, a sample of Taiwanese households, and a sample of redwood trees and using each sample mean as an estimate of the corresponding population

Figure 2. The benefits of shrinkage. True values, data, sample means, and a Bayesian hierarchical shrinkage estimates are shown. The shrinkage estimates are more accurate in this case as well as on average across repeated samples from the true values. The Bayesian model is described in Rouder, Morey, Verhagen, Swagman, and Wagenmakers, 2017.

mean. Stein's paradox shows that the total error of these three estimates is expected to be reduced if, rather than using sample means to estimate each individual population mean, scale information from all three samples is pooled. To be clear, pooling only helps in overall estimation - hog weight information does not reduce the error of the redwood tree heights estimate. Rather, pooling with leave some individual estimates worse off (say hog weight) with other individual estimates better off, and the gains outweigh the losses on balance. In other words, the most accurate estimator for the three properties considered *simultaneously* is not the same as the most accurate estimator for each quantity individually. In experimental psychology, we are interested in the relations among conditions and having the most accurate overall estimator is desirable.

Regularization and Interpretability

We are now in a position to use shrinkage estimation to assess whether sample means are interpretable. To do so, we ask if we can construct a ridiculous measurement tool, similar in spirit to the zero estimator, that outperforms sample means in common environments. If so, we consider sample means uninterpretable in these environments.

Here, we use randomness rather than the zero value. We will randomly assign measurement values to each condition so that the relations among them, say which is largest, is completely random and made without any influence of data. This notion of using a random benchmark to calibrate the performance of a method is standard in classification and machine learning, and is even preceded in psychology (Dawes, 1979; Wilks, 1938). The innovation here is to use random estimators that have modern shrinkage properties. Although the relations are random, the data are used to determine the scale or calibration of the values. This approach of using a smart random estimator as a benchmark is also preceded in psychology (Davis-Stober & Dana, 2014). The argument for interpretability we present here is more general than our previous work. The random estimator presented in the next section, though related to the class considered in Davis-Stober and Dana (2014), is both novel and ideal for assessing the interpretability of sample means.

Here is an example of how smart random estimators work: The shrinkage estimates in Figure 2 for the colors red, green, blue, and yellow are the values of 617 ms, 636 ms, 638 ms, and 712 ms. One way we can construct a random estimator with these calibrations is to randomly assign each value to a color. For example, we may assign the colors red, green, blue and yellow the values 638 ms, 636 ms, 712 ms, 617 ms. The key property of this new random estimator is that it does exactly the opposite of what psychologists value. It randomly scrambles the exact information of interest, the relations among the conditions. For example, one is unable with this tool to truly assess whether red text is read more quickly than green text. It preserves the characteristic of no intrinsic interest, the scale of measurement, say that differences are about 40 ms. Such information, while useful in fine tuning estimates, is of no real relevance in practice to most psychologists.

We consider the sample means to be interpretable if they are more accurate than random estimators that scramble the relations among conditions. To characterize interpretability we develop a smart random estimator in the next section. This estimator leverages modern advances in shrinkage and uses the scale afforded in the data to calibrate the random relations. The argument is that if this smart random estimator, which randomizes critical information, can outperform sample means, then sample means are uninterpretable in the measurement environment. The critical question is whether there are such uninterpretable environments. The answer is yes, and the environments are, unfortunately, not so rare in psychological science.

A Random Estimator For Condition Means

The key step in a random estimator is that the relations among the condition means is random, and so, we start with random numbers. A bit of notation is helpful, and we let p denote the number of conditions. Let $a_1, a_2, \dots, a_p$ be a sequence of independent draws from a continuous uniform distribution over the interval $[-1, 1]$. We assign a_1 to the first condition, a_2 to the second condition, and so on. Because each of a_i 's are random, the

relations among the conditions are random and, in particular, are completely independent of any structure in the data.

Next, we use data to scale the a_i 's. As described in the previous section, this scaling term will leverage shrinkage properties to put these random estimates on a scale appropriate to the data. We denote this scaling coefficient as b , and it is a single number. The random estimator for each condition mean, denoted $\hat{\mu}_i^{re}$, is

$$\hat{\mu}_i^{re} = \bar{G} + b \times a_i, \quad (1)$$

where $\bar{G}$ is the observed grand mean across conditions. Here, we see the random estimator is a simple scaling of random numbers. The value of b , discussed subsequently, can be positive or negative, depending on the data. In either case, the order among condition estimates is random, that is, it reflects the random relations among the random numbers $a_1, \dots, a_p$.

Remaining is the construction of b . We use a least-squares argument. Scalar b is constructed to insure that $\hat{\mu}^{re}$ minimizes squared error subject to the above random assignment constraints. The resultant is

$$b = \left(\frac{p - \sqrt{p(p-1)}}{\sqrt{\sum_{i=1}^p a_i^2}} \sum_{i=1}^p a_i \alpha_i \right),$$

where $\alpha_i = \bar{y}_i - \bar{G}$, with $\bar{y}_i$ defined to be the sample mean for the i^{th} condition. While the scaling term b makes full use of the data, which is ‘smart’ from an estimation perspective, this scaling parameter can only do one of two things to the ordering of the random a_i 's. It can either: (1) leave the random ordering of the a_i 's unchanged (if $b > 0$) or (2) it can completely reverse the ordering (if $b < 0$). Hence, the order of estimates is random¹.

Whether the random estimator incurs less expected squared error than sample means depends upon the environment, i.e., sample size and effect size. As a simple example of how to calculate our random estimator, consider the hypothetical data in the previous example on naming words as a function of display color. The sample means for this set of data are equal to: $\bar{y}_1 = 593$ ms, $\bar{y}_2 = 626$ ms, $\bar{y}_3 = 630$ ms, and $\bar{y}_4 = 756$ ms, for Conditions 1-4 respectively. We refer to this pattern of means as the observed pattern of treatment effects. Now, let's consider what the random estimator would yield as estimates of the population means for these data. First, we draw four uniform random numbers from the $[-1, 1]$ interval that correspond to the four conditions. For Condition 1, we obtain -.190, for Condition 2, we obtain -.973, for Condition 3, we obtain .823, and for Condition 4 we obtain .600. Note that once a random number has been drawn for a condition, we cannot shuffle or re-assign it to another condition after the fact. We calculate the scale factor of b as 21.5 ms, which portends quite a bit of shrinkage. Using this value, the random estimators are $\mu_1^{re} = 647$ ms, $\mu_2^{re} = 630$ ms, $\mu_3^{re} = 669$ ms, $\mu_4^{re} = 664$ ms. Recall that the order of the sample means was: Condition 1 < Condition 2 < Condition 3 < Condition 4. In contrast, the random estimator now assigns the order: Condition 2 < Condition 1 < Condition 4 < Condition 3. This is the same random order that was produced by the random a_i draws. Note also that

¹Since the b term can reverse the order of the random estimates, the $p = 2$ case does not randomize order. For this case, the zero estimator comparison may be more appropriate, although, as we later discuss, the zero and random estimators yield similar sample size recommendations.

the relative distance between the random estimates has also been randomized; the estimates for Conditions 3 and 4 are quite close – in stark contrast to the sample means, where $\bar{y}_4$ is relatively farther from $\bar{y}_3$.

This illustration was based on a single sample. Of course, the stronger argument is what may be expected across all samples for an environment. Perhaps the most straightforward approach to generalization is to simply simulate many samples from the same environment. As a simulation example, we will define a particular environment and compare the accuracy of sample means to the random estimator directly. As before, consider four treatment conditions under a balanced design. Let the four population means be equal to: $\mu_1 = 600$ ms, $\mu_2 = 620$ ms, $\mu_3 = 640$ ms, and $\mu_4 = 660$ ms. For this simulation we will assume that the dependent variable being measured is normally distribution with a standard deviation of 100 ms. This combination of means and variance gives an overall effect size of $f^2 = .05$, which is a “small” effect according to the Cohen conventions (Cohen, 1988). A single repetition will consist of the following steps: (1) generate ten samples ($n = 10$) from each of the four distributions, (2) calculate the four sample means and our random estimates, and (3) calculate the squared error from each set of estimates and the four population means. By repeating steps (1)-(3) many times and averaging the squared error for each estimation method we can obtain an estimate of average squared error. To arrive at a good estimate, we carried out 10,000 repetitions.

Figure 3 shows the mean-squared error in estimation for each repetition, and the big green dot shows the mean over repetitions. The mean error is 30% larger for the sample mean than for the random estimator, and, moreover, 62% of the samples show larger error for the sample mean than for the random estimator. Given these results, it is hard to take the orderings implied by sample means seriously when they are more error prone than random values. In this sense, they are uninterpretable.

Planning and Evaluating Designs

The main advantage of zero- and random-estimators is that they provide a benchmark for using sample means to interpret the ordering of conditions. If these nonsense estimators with no or randomized orders are expected to do a better job of accounting for the structure in data than sample means, then sample means should not be interpreted. For simple designs, say one-way, between-group ANOVA designs, it is possible to derive expressions for the minimum sample size per condition for which sample means are indeed more accurate than the zero- or random-estimator.

The key input to these expressions for minimum sample size is a minimum Cohen noncentrality effect-size measure one wishes to resolve, f^2 . This measure, along with sample size and number of conditions, defines the environment. The noncentrality measure is

$$f^2 = \frac{\frac{1}{p} \sum_{i=1}^p (\mu_i - \mu^*)^2}{\sigma^2},$$

where μ^* is the true grand mean given by $\mu^* = (\sum_i \mu_i)/p$ and p is the number of conditions. For calibration purposes, Cohen recommends using f^2 values of .01, .0625, and .16 as emblematic of small, medium, and large effects, respectively. These values for f^2 are analogs to the values of .2, .5, and .8, respectively, for the better known d measure. The f^2 measure

Figure 3. Mean-square error for the sample mean and the random estimator for 10 observations in each of four conditions. Each point shows the result from a simulation repetition, and there are 10,000 such repetitions. Contours are bivariate kernel density estimates. The large point is the mean of MSEs across repetitions. This mean for sample means is 40% larger than the mean for the random estimator, and 63% of samples show larger MSE for the sample mean.

is also related to R^2 , the proportion of variance accounted for by the condition structure as

$$f^2 = \frac{R^2}{1 - R^2}.$$

Expressing a minimum sample size for sample mean interpretability works as follows: First the researcher chooses a value of f^2 to be resolved in an experiment. Then, the minimum sample size per condition needed to insure that on average sample means are more accurate than the random estimator is

$$n_r = \frac{\sqrt{p(p-1)}}{pf^2}. \quad (2)$$

A proof is provided in the appendix. As an example, if there are 5 conditions and the minimum f^2 to be resolved is .05, then the value of n_r , the minimum sample size per condition, is 17.9, which rounds up to 18. (Strictly speaking, the minimum sample size is the smallest integer greater than or equal to n_r .) Similarly, the expression for n_0 , the

minimum sample size so that the sample mean is more accurate than the zero estimator is

$$n_0 = \frac{(p-1)}{pf^2}. \quad (3)$$

A bit of algebra reveals that

$$n_0 = n_r \times \sqrt{\frac{p-1}{p}}.$$

The square root term is less than 1.0 indicating that $n_r > n_0$. For this reason, we focus on the larger size, n_r . Figure 4A shows the dependency of n_r on f^2 for experiments with 2, 4, 6, and 8 conditions. As can be seen, the minimum sample sizes per condition for interpretability can be quite large, say around 80 observations per condition to resolve small effects. In designs with smaller sample sizes, random estimators will outperform sample means on average. Nobody wants this.

For comparison, Figure 4B shows the minimum sample size per condition to maintain 50% power (at the 5% Type I level) in a one-way ANOVA F -test. As can be seen, these minimum sizes are larger than that needed for interpretability, indicating that interpretability is a low bar. Also evident is that the number of conditions affects minimum sample size differently for interpretability than for power. The minimum sample size increases with condition size for interpretability but decreases for power (indeed, increasing the number of conditions leads to a more stable estimate of within-condition variance). As a result, with several conditions, it may be possible to have high powered designs that have uninterpretable sample means. In such cases, it is hard to see how the ensuing inferences are sound even with high power. Overall, increasing the number of conditions (while maintaining sample size per condition) makes it easier to detect differences, but it makes it harder measure the ordering relationships accurately.

The above minimum sample size expressions are general in that they hold no matter how the dependent variable is distributed. The sole assumption is the familiar one of homogeneity where variance is assumed constant across conditions.

General Discussion

We have shown here that sample means are measurement instruments, and like physical measurements, are interpretable in specific environments. There are low-resolution environments where sample means are outperformed by smart random estimators that scramble the relations among conditions. This performance gain happens because the random estimators use the scale of data to inform the estimates. As a result, there is much shrinkage which helps eliminates noise. Sample means, in contrast, do not use the overall scale of data and offer less noise control.

Sample Means Are Not Ideal

Modern estimation theory stresses shrinkage estimators over sample means, and examples include Bayesian hierarchical estimators, lasso estimators, and ridge estimators. These estimators make use of condition relational information, like the sample mean, and scale information, like the random estimator. Because they use information wisely, these

Figure 4. **A.** Minimum sample size (per condition) for interpretability of the sample mean as a function of effect size f^2 and the number of conditions. One of the more noteworthy aspects is that the needed sample size per condition increases with the number of conditions. **B.** Minimum sample size (per condition) for a power of 50% at the 5% level. The needed sample size per condition decreases with the number of conditions. **C.** The power at the 5% level for the minimum sample size for sample mean interpretability. Surprisingly, commonly powered designs often yield uninterpretable sample means.

modern estimators outperform sample means and random estimators in all measurement environments. They are better measurement tools.

One of our main conclusions is that researchers should avoid reporting sample means in favor of more sophisticated estimates. Two issues remain: First, and perhaps of lesser importance, is that modern shrinkage estimates require some prior assumptions. In our view, these assumptions are benign and default settings are useful. The second issue is more consequential. Researchers are comfortable with sample means. And why shouldn't they be? We have all been taught the arithmetic average since grade school. The sample mean is the most privileged measure of central tendency, and it takes center stage over its less glorious siblings, median and mode. Yet, the performance of the random estimators in certain environments, and the superiority of modern estimators more generally, serve to undermine this privileged and natural status. In our view, sample means, sample medians, and modes should not be taught as descriptives. Instead, they should be taught within the context of estimation where statistics serve not to describe but to estimate unknown parameters. Different estimators have different desiderata pertaining to bias, efficiency and loss. Although this approach is more complicated and difficult than putting the sample mean on a pedestal, it seems necessary for clear scientific thought.

How often are sample means uninterpretable?

One might ask: how often do psychologists utilize sample means that are uninterpretable? One way to answer this question is to examine statistical power values for sample sizes that are at the threshold for sample mean interpretability. Figure 4C shows the statistical power (at the 5% Type-I level) for a one-way ANOVA for n_r , the minimum sample size per condition for sample-mean interpretability. It is clear from this graph that sample means are uninterpretable for non-trivial power values, especially when there are many conditions.

How well powered are studies in the literature? Although this is a difficult question, there are several model-based meta-analyses that estimate power across different areas of psychology. Average statistical power in many areas of psychology is at or below .5, with average power from studies with small effects hovering around .17 (Rossi, 1990; Sedlmeier & Gigerenzer, 1989, see also Maxwell, 2004, and Tressoldi, 2012). A recent study by Szucs and Ioannidis (2016) examined more than 100,000 statistical records from about 10,000 cognitive neuroscience and psychology papers published within the past five years. They found that the median power to detect small effect sizes was .12 - which clearly equates to an uninterpretable sample mean environment - see Figure 4C. Sample means should not be used in such environments. Instead, it is imperative for researchers to use more sophisticated methods that incorporate modern inference when investigating such effects.

Appendix

Mean squared error is defined as $E[\sum_{i=1}^p (\hat{\mu}_i - \mu_i)^2]$, where $\hat{\mu}_i$ is an estimate of μ_i . It is straightforward, using basic properties of the expectation operator, to show that

$$E[\sum_{i=1}^p (\hat{\mu}_i^{re} - \mu_i)^2] = \frac{\sigma^2(p + (p-1)(p - \sqrt{p^2 - p})^2)}{pn} + 2f^2\sigma^2(p-1)(p - \sqrt{p^2 - p}).$$

The mean squared error for sample means is well known to be

$$E\left[\sum_{i=1}^p (\bar{y}_i - \mu_i)^2\right] = \frac{p\sigma^2}{n}.$$

We prove the result by solving the inequality $E[\sum_{i=1}^p (\hat{\mu}_i^{rce} - \mu_i)^2] < E[\sum_{i=1}^p (\bar{y}_i - \mu_i)^2]$ in terms of n ,

$$\begin{aligned} & E\left[\sum_{i=1}^p (\hat{\mu}_i^{rce} - \mu_i)^2\right] < E\left[\sum_{i=1}^p (\bar{y}_i - \mu_i)^2\right], \\ \iff & \frac{\sigma^2(p + (p-1)(p - \sqrt{p^2 - p})^2)}{pn} + 2f^2\sigma^2(p-1)(p - \sqrt{p^2 - p}) < \frac{p\sigma^2}{n}, \\ \iff & 2f^2\sigma^2(p-1)(p - \sqrt{p^2 - p}) < \frac{\sigma^2(p^2 - p - (p-1)(p - \sqrt{p^2 - p})^2)}{pn}, \\ \iff & 2f^2\sigma^2(p-1)(p - \sqrt{p^2 - p}) < \frac{2\sigma^2(p-1)(\sqrt{p^2 - p} - (p-1))}{n}, \\ \iff & n < \frac{\sqrt{p^2 - p} - (p-1)}{(p - \sqrt{p^2 - p})f^2}, \\ \iff & n < \frac{\sqrt{p(p-1)}}{pf^2}, \end{aligned}$$

which completes the proof.

References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). Hillsdale, NJ: Erlbaum.
- Correll, J., Park, B., Judd, C., & Wittenbrink, C. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83(6), 1314–1329.
- Davis-Stober, C. P. & Dana, J. (2014). Comparing the accuracy of experimental estimates to guessing: a new perspective on replication and the “Crisis of Confidence” in psychology. *Behavioral Research Methods*, 46, 1–14. Retrieved from <http://dx.doi.org/10.3758/s13428-013-0342-1>
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571.
- Gelman, A. & Carlin, J. (2017). *Some natural solutions to the p-value communication problem—and why they won't work*.
- Lehmann, E. L. & Casella, G. (1998). *Theory of point estimation*, 2nd edition. New York: Springer.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, 9, 147–163.
- Rossi, J. S. (1990). Statistical power of psychological research: what have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, 58(5), 646–656. Retrieved from <http://search.ebscohost.com.proxy-ub.rug.nl/login.aspx?direct=true&db=psyh&AN=1991-03239-001&site=ehost-live&scope=site>

- Rouder, J. N., Morey, R. D., Verhagan, A. J., Swagman, A. R., & Wagenmakers, E.-J. (in press). Bayesian analysis of factorial designs. *Psychological Methods*.
- Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distributions. In *Proceedings of the third berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 197–206).
- Szucs, D. & Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *bioRxiv*.
- Tressoldi, P. E. (2012). Replication unreliability in psychology: elusive phenomena or "elusive" statistical power? *Frontiers in Psychology*, 3.
- Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 3(1), 23–40.