The epistemic value of p -values

Marko Kovic^{*1,2,3}

¹*ZIPAR – Zurich Institute of Public Affairs Research, Zurich, Switzerland.*

²*ars cognitionis GmbH, Zurich, Switzerland.*

³*Swiss Skeptics – Association for Critical Thinking, Zurich, Switzerland.*

December 2, 2017

Abstract

How do p -values relate to beliefs? p -values are widely used in data analytic contexts, and the misuse of p -values can have grave consequences. In order to get a better grip on p -values, we need to make epistemological sense of them. I argue that the least inappropriate way of understanding the relation of p -values to beliefs is to regard p -values as a *degree of epistemic luck*, or, in the rare contexts in which the belief in a null model is justified, as a *degree of justifiedness*. When we handle p -values in an epistemically responsible way, there is no need to abandon them. What should be abandoned, however, is the use of p -values as a default tool in data analysis: Instead of relying on default tools, we should justify our data analysis tools each time we perform some data analytic task.

^{*}marko.kovic@ars-cognitionis.ch / marko.kovic@zipar.org / marko.kovic@skeptiker.ch /
+41 76 335 06 17

1 p -values, a ubiquitous tool

When analyzing quantitative data, we are often not only interested “descriptive” properties of the data at hand, but we would like to know how the properties of the data at hand relate to the real world from which this data was sampled. The most common method of performing such “inferential” data analysis is by using so-called p -values. Someone only peripherally interested in science might not be familiar with p -values, but they might have heard of “statistical significance”. Statistical significance is a convention for reporting p -values: When $p \leq 0.05$, then the information at hand is important; if not, then (apparently) not.

p -values are an essential part of the toolkit of frequentist statistics [1], and they are routinely calculated and reported as a *de facto* default inferential heuristic. That is, in and of itself, a good thing: Analyzing data is not easy, and we need tools to help us make sense of it all. In recent years, however, the academic debate about the downsides of data analysis based on p -values has intensified. In the context of the so-called replication crisis, or, more generally, the statistical crisis in science [2], the heavy reliance on, the direct or unwitting manipulation of, as well as the routine misinterpretation of p -values is criticized [3, 4, 5].

p -values are an important tool in data analysis, but one that, it seems, is not easy to handle. When the tool that are p -values is mishandled, the results can be dire, as the statistical crisis in science has taught us. In order for p -values to be used in as rational a manner as possible, we need to have as clear an understanding of p -values as possible. The goal of this paper is to contribute to our understanding of p -values by exploring the concept of p -values from an epistemological perspective.

2 A common misunderstanding of p -values, and why it matters

Describing p -values in a formal manner is relatively simple [6]:

$$\Pr(T(y_{rep}) \geq T(y) \mid H_0)$$

The p -value is the probability that some test statistic T under a hypothetical replication y_{rep} will be equal to or greater than the test statistic T for the

actual data at hand, y , given the null model H_0 . In other words: Given that the null model H_0 is true, the p -value denotes the probability of obtaining the same or more extreme test statistic if we were to replicate the data collection infinitely often. The null model is a specific statistical model that posits that some effect or some difference (of effects) is zero.

p -values are routinely interpreted in ways that deviate from its technical definition. The most common and, I believe, the most problematic misunderstanding of p -values is the belief that the p -value is the probability of the null model [7, 8, 9]. In terms of the relationship of the null model and the observed data, this misunderstanding takes the following form:

$$\Pr(H_0 \mid \text{data})$$

It is understandable why one would be inclined to interpret p -values as the probability of the null model. After all, when calculating a p -value, we are explicitly stating our interest in the null model, and therefore, we intuitively expect the p -value to tell us something about the null model. But it does no such thing: The p -value *cannot* be the probability for the null model because the starting point of the calculation of the p -value is that the null model is true. Interpreting the p -value as the probability of the null model, therefore, *must* be false.

This common misunderstanding of p -values is worth discussing in an epistemological context. Yes, interpreting p -values as posterior probabilities is plainly incorrect. But the fact that there is a tendency to misinterpret p -values in such a manner is relevant: From an epistemological point of view, the incorrect idea of the p -value as the posterior probability of the null model is rather elegant and it indicates that there is a desire to make epistemic sense of p -values. Having an epistemic understanding of p -values, then, is doubly valuable: An incorrect epistemic interpretation could become less pervasive, and the desire for making epistemic sense of p -values could be satisfied.

3 p -values and beliefs

3.1 A fictional example: Test scores of class A and class B

Imagine two classes in some school, class A and class B, with 30 pupils each. The pupils in both classes had to take a geography test on the same day.

Pupils in both classes did similarly well, as depicted in Figure 1.

Figure 1: Histogram of fictional test scores for classes A and B.

The mean of the test scores for class A is 70.6, and that for class B is 64.8.

On average, class A did a bit better than class B, but Jane Doe, a teacher at the school, is wondering how meaningful this difference is. In order to better understand the average test score difference, Jane Doe decides to calculate a test statistic in order to arrive at a p -value. Test statistic of choice for this is the Welch two sample t-test [10], a widely used procedure for estimating differences in means of two groups.

The test statistic yields a p -value of 0.018. Now, upon having calculated a p -value, the big question is: What, exactly, is Jane Doe, an epistemically rational agent, supposed to believe about the difference in test scores of classes A and classes B?

3.2 The p -value as a posterior, updated belief

As discussed in section 2, the p -value cannot be the posterior probability of the null model. Could it, however, be understood as a posterior degree of belief, or posterior credence [11], in the null model?

In our example, Jane Doe would have started out with the belief that $\Pr(H_0) = 1$. Now, upon calculating a test statistic that allowed to calculate a p -value, Jane decides to update her belief to $\Pr(H_0) = 0.018$. This seems like a rational course of action: Jane Doe has some prior belief, she collects evidence, and then she updates her belief. This kind of interpretation, however, creates two problems.

First, if the p -value should serve as the basis for the description of one's posterior belief in the null model, we need to have some rule for mapping the p -value to the degree of belief in the null model. If we simply take the p -value and use it as the exact quantification of our belief in the null model, we are creating an inconsistency. The p -value cannot be the posterior probability for the null model, but if we use the p -value as a quantification of our posterior belief in the null model, we are making a proposition that is very close to the posterior probability of the null model.

Second, and crucially, the prior belief $\Pr(H_0) = 1$ that we make upon calculating a p -value cannot simply be swapped by a posterior belief. After all, the calculating the p -value is contingent on the explicit belief $\Pr(H_0) = 1$, so using the p -value as the basis of some updated belief $\Pr(H_0) \leq 1$ is problematic. If Jane Doe's prior belief $\Pr(H_0) = 1$ represents the premise for calculating the p -value of 0.018, then Jane Doe's posterior belief cannot be simply $\Pr(H_0) = 0.018$, because that would be a case of circular reasoning [12, 524].

3.3 The p -value as a second doxastic layer

Inferring a posterior, updated belief about the null model with a p -value is problematic. What kind of epistemic options does this leave for Jane Doe? It has been proposed to treat p -values as a *doxastic attitude other than belief*. One specific proposal is the idea of the p -value as a “degree of surprise in the data” [13]. The idea of p -values as a second doxastic layer is an intriguing and elegant option. Jane Doe believes $\Pr(H_0) = 1$, the discrepancy between the data and the model surprises her to the degree of $1 - 0.018 = 0.982$.

If the p -value represents a second doxastic layer, such as the degree of

surprise, then we have circumvented the problems that arise if we try to interpret the p -value as an updated belief. However, this interpretation gives rise to another complication: How are p -values as a second doxastic layer related to beliefs? More specifically: What happens to the belief $\Pr(H_0) = 1$? If the second doxastic layer is supposed to directly inform and update our beliefs, we have simply complicated the process of interpreting p -values as an updated belief described in subsection 3.2 by introducing the second doxastic layer as just one additional step.

The utility of p -values as a second doxastic layer might lie in their ability to serve as a heuristic for switching to and updating our beliefs in a different epistemic context. If p -values quantify our degree of surprise, then we might need to update our belief $\Pr(H_0) = 1$ according to that surprise, but do so *outside* of the current data analytic context. In our example, Jane Doe might conclude that she is highly surprised by the data; so much so that she decides to update her belief $\Pr(H_0) = 1$. Within the narrow epistemic context of the data analysis, her belief $\Pr(H_0) = 1$ remains, but the p -value as a degree of surprise is a second doxastic layer that also serves as a window into a broader, real-world epistemic context beyond the narrow confines of the data analysis.

Unfortunately, however, this approach leaves us once again with the open question of how to map p -values to our beliefs. More precisely, we have to, first, map p -values to our degree of surprise, and then, we have to map our degree of surprise to our broad, real-world posterior belief. If we do this by simply repurposing the exact value of the p -value twice in a row, we are taking an unjustified shortcut.

3.4 The p -value as an *reductio ad absurdum* argument

One tricky aspect of p -values is the adoption of the belief that $\Pr(H_0) = 1$. What if there was a way to interpret p -values without really adopting that belief? This could potentially be feasible if p -values represented a form of epistemic thought experiment: A *reductio ad absurdum* argument.

In our example, Jane Doe could interpret the p -value she obtained by turning it into a *reductio ad absurdum* argument: If the test scores of both classes did not differ at all, the p -value should be exactly 1. The actual p -value is much lower, only 0.018, so the two classes probably do really differ. Understanding p -values in this manner seems to be an elegant epistemic solution insofar as one seems to be fairly unburdened from doxastic attitudes – you can make the case against the null model without really believing in it.

Unfortunately, the interpretation of p -values as a *reductio ad absurdum* argument creates a number of epistemic inconsistencies. *First*, and rather obviously, viewing p -values as a *reductio* thought experiment still relies on some prior belief; not saying it out loud does not make it go away. Within the *reductio* interpretation, the prior belief is not $\Pr(H_0) = 1$, but simply $\Pr(H_0) = 0$. *Second*, and once again, there is the open question of mapping the p -value to whatever kind of posterior belief one wants to adopt within the *reductio* thought experiment interpretation. The posterior belief, again, cannot simply be $\Pr(H_0) = p$. *Third*, and most importantly, the *reductio* argument does not work logically. The p -value cannot be used as any kind of evidence against the null model, because the p -value is premised on the truth of the null model, regardless of the numerical value of the p -value. In other words: p -values fail as a *reductio* argument because they are not a consequence of the premise $\Pr(H_0) = 1$, but actually a part of the premise. This last point can be illuminated with a simple example. If we want to determine whether a stopwatch is broken, we could measure the exact same action (like dropping a feather in a vacuum tube) multiple times and simply compare the results: If the stopwatch were broken, we should get different results. This is a valid *reductio* argument. However, if we change this setup and introduce the condition that any test result is premised on the stopwatch not being broken, then we do not have a valid *reductio* argument any longer.

3.5 The p -value as the degree of epistemic luck

The attempts so far to interpret p -values in an epistemically consistent way have not worked very well. What other epistemic routes can Jane Doe explore? Jane Doe's prior belief is $\Pr(H_0) = 1$. Upon calculating the p -value, Jane Doe still holds the same belief. However, a new piece of information, the p -value, has been obtained. The p -value indicates, technically, how well the data fit the null model. In her current epistemic state, Jane Doe can interpret the p -value as a quantification of how *lucky* she got: Jane Doe believes $\Pr(H_0) = 1$, but that belief is not very deliberate. Jane Doe has simply used a data analytic tool, the p -value, that posits $\Pr(H_0) = 1$. In this setting, the p -value indicates how well the unjustified belief $\Pr(H_0) = 1$ corresponds to the data.

The kind of luck that the p -value indicates is sometimes referred to as *epistemic luck*. Epistemic luck is an epistemological constellation in which our beliefs happen to be true, even though we have no justification for those beliefs [14]. A famous case of epistemic luck is the Gettier problem [15]. If

Jane Doe takes this route, then she is, first, believing without justification that $\Pr(H_0) = 1$. Then, she calculates the p -value and realizes that she has not gotten very epistemically lucky, since the data fits her irrational belief rather poorly, with a p -value of merely 0.018.

Interpreting p -values as a quantification of epistemic luck is, of course, somewhat provocative. After all, when we engage in data analysis, we are doing so very deliberately. The notion of epistemic luck in the context of data analysis might sound like an accusation that researchers are not thinking carefully enough about what they are doing. Of course, I am not making such an accusation. The notion of p -values as a quantification of epistemic luck simply describes a common context of data analysis: By calculating p -values, we are expressing a belief in the null model, and in most cases, this belief is unjustified since the belief in the null model is not the result of some careful prior deliberation, but simply a premise of the data analytic tool we need to use.

But data analytic contexts in which we hold a *justified* belief in the null model do exist. For example, when we are measuring the speed of light in North America and in South America, our belief that the difference in measurements will be zero is justified. In such cases, the p -value is not a quantification of epistemic luck, because we are not simply getting lucky if the data correspond to the null model – we have good reasons for believing $\Pr(H_0) = 1$. When we are calculating p -values in such a context, p -values do not represent the degree of epistemic luck. Instead, they represent the quantification of the *degree of justifiedness* of the belief $\Pr(H_0) = 1$.

p -values as the degree of epistemic luck (or the degree of justifiedness) can also be described as the degree of *mind-to-world fit* [16]. If the degree of Jane Doe’s epistemic luck is 0.018, then the degree to which her mind fits the world is also 0.018. Describing p -values in terms of direction of fit is not necessary and adds no new information compared to the epistemic luck interpretation, but it can elucidate the notion of degree of epistemic luck. After all, as rational agents, we should be interested how well the things in our mind (belief that $\Pr(H_0) = 1$) fit the world (data).

4 Discussion: The complicated epistemic nature of p -values

In the preceding section, I have tried to explore different possible ways in which p -values might relate to our beliefs. The least inconsistent of those ways, I believe, is the notion of p -values as the degree of epistemic luck (Or, in the contexts in which a belief in the null model is justified, the degree of the justifiedness of the belief in the null model.). Such an epistemic view of p -values is not exactly intuitive, but it is, I believe, simple enough so as to be useful.

How useful is this proposed interpretation of p -values? Having a clearer understanding of how p -values relate to our beliefs does not address other epistemological issues of p -values, such as the nature of frequentist probability [17] on which p -values are premised. However, having a clearer picture of the epistemological status of p -values contributes both to epistemic and to instrumental rationality. Epistemically, of course, having a framework of how p -values relate to beliefs is helpful for holding accurate beliefs. Instrumentally, an epistemological framework for p -values is helpful for deciding whether p -values are the appropriate tool for some data analytic task.

4.1 What about statistical significance?

In section 1, I mention statistical significance as a concept that might be better known than p -values themselves. Statistical significance as practiced today refers to what is sometimes described as null hypothesis significance testing (NHST), a way of using p -values that is a combination of the Fisherian with the Neyman-Pearson approach [18, 19]. The aim of NHST is commendable: We do not only want to analyze data, but we also want to make decisions based on the results of the data analysis. NHST helps in this regard by defining thresholds of statistical significance, such as, probably most frequently, $p \leq 0.05$. Working within the NHST framework has at least two benefits. *First*, from a purely proceduralist view, the fact that the NHST framework is widely applied means that it provides a common data analytic language among researchers of all stripes; a sort of statistical Esperanto. *Second*, the NHST framework is a very convenient decision-making tool. Defining p -value thresholds might be mostly arbitrary, but using a threshold as a decision-making heuristic can be an effective incentive for making decisions

rather than not making them.

Even though NHST has clear benefits, it also has two major drawbacks. *First*, partitioning the p -value, a continuous measure, into categories that demarcate thresholds creates numerous problems in practice [20, 21]. NHST thresholds can be a positive incentive in that they encourage decision-making, but in practice, they are also negative incentives that contribute to problems such as the publication bias and, ultimately, the replication crisis. Proposals to lower the p -value thresholds within the NHST framework [22] are unlikely to alleviate the practical problems of NHST [23].

Second, and just as importantly: The concept of statistical significance creates additional epistemic problems¹. The first epistemic problem that NHST creates has to do with the idea of rejection of the null model. By calculating a test statistic that is conditional on the truth of the null model, we are supposed to either accept or reject the null model. From an epistemological point of view, it is unclear what rejecting the null model means. In the context of the discussion in this paper, rejecting the null model could be interpreted as the adoption of the belief that the null model is false, based on a low level of epistemic luck we have achieved in the data analysis. That, however, catapults us right back to the beginning of the problem, since p -values cannot be the posterior probability of the null model, and they cannot plausibly be our posterior, updated belief about the null model.

The second epistemic problem that arises within the NHST framework is that the rejection of the null model is coupled to the adoption of the belief in some alternative model. The desire for that kind of epistemic leap is understandable – after all, we probably mostly do not really care all that much about the null model, but about some other model. However, adopting a belief in some alternative model after rejecting a null model is somewhat puzzling. Within the NHST framework, one can specify power, which is the probability of correctly rejecting the null model. But, obviously, power cannot tell us anything about the data analysis at hand, since it relies on frequentist probability. Furthermore, even if we somehow decide to reject the null model, that rejection does not in any way pertain to some alternative model but only to the null model. Making this kind of epistemic leap is intuitively understandable: We want to knock down the null model and support some

¹For the purpose of this argument, it is not important whether we are talking about a traditional Neyman-Pearson hypothesis test involving a critical value or a test statistic involving p -values.

alternative model. But wanting this kind of epistemic leap does not make it rational.

4.2 Should we abandon p -values?

Using p -values in applied work is trivially easy, and it is a common heuristic in large parts of academia. p -values, therefore, have obvious practical benefits. But, as the replication crisis has shown us, their misuse can have grave consequences. Should we therefore try to abandon p -values, or at least ignore them until they fade into obscurity [24]?

There are good reasons to believe that abandoning p -values might not be the data analytic panacea that perhaps we might want to believe it is [25]. For example, a common argument posits that p -values should be abandoned in favor of, or gradually supplanted by, so-called confidence intervals [26], but confidence intervals open a whole other can of epistemic worms [27]. The question, I believe, is not whether we should categorically ban p -values or not, but rather whether p -values are the right epistemic tool for the job. As an epistemic heuristic, or an epistemic crutch, they are probably useful in many contexts, and in some contexts, the belief in the null model might actually be justified.

In any number of contexts, however, p -values might not be all that useful, neither epistemically nor instrumentally. In such contexts, tools other than p -values should be considered and, if appropriate, applied. For example, Bayesian data analysis might be easier to stomach epistemically in certain data analytic contexts [28, 29]. These “contexts” I talk of are, unfortunately, not purely objective in nature or definable *a priori* – ideally, some justification for the data analytic procedure of choice should be provided in every single data analytic instance.

References

- [1] J. Neyman. Frequentist probability and frequentist statistics. *Synthese*, 36(1):97–131, September 1977.
- [2] Andrew Gelman and Eric Loken. The Statistical Crisis in Science. *American Scientist*, 102(6):460, 2014. bibtex: gelman2014statistical.
- [3] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366, November 2011.
- [4] Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, 13(3):e1002106, March 2015.
- [5] John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Med*, 2(8):e124, August 2005.
- [6] Andrew Gelman. P Values and Statistical Practice. *Epidemiology*, 24(1):69–72, 2013.
- [7] Steven Goodman. A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology*, 45(3):135–140, July 2008.
- [8] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350, April 2016.
- [9] Ronald L. Wasserstein and Nicole A. Lazar. The ASA’s Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, April 2016.
- [10] B. L. Welch. The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika*, 29(3/4):350–362, 1938.
- [11] Richard Pettigrew. Accuracy and the credence-belief connection. *Philosopher’s Imprint*, 15(16), July 2015.

- [12] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK ; New York, NY, June 2003.
- [13] MJ Bayarri and James O Berger. Quantifying surprise in the data and model verification. *Bayesian statistics*, 6:53–82, 1999. bibtex: bayarri1999quantifying.
- [14] Duncan Pritchard. Epistemic Luck. *Journal of Philosophical Research*, 29:191–220, July 2004.
- [15] Edmund L. Gettier. Is Justified True Belief Knowledge? *Analysis*, 23(6):121–123, 1963.
- [16] I. L. Humberstone. Direction of Fit. *Mind*, 101(401):59–83, 1992.
- [17] Alan Hájek. Fifteen Arguments Against Hypothetical Frequentism. *Erkenntnis*, 70(2):211–235, 2009.
- [18] Joseph Lee Rodgers. The Epistemology of Mathematical and Statistical Modeling: A Quiet Methodological Revolution. *The American Psychologist*, 65(1):1–12, January 2010.
- [19] Jose D. Perezgonzalez. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, 6, March 2015.
- [20] Valentin Amrhein, Fränzi Korner-Nievergelt, and Tobias Roth. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. *PeerJ*, 5:e3544, July 2017.
- [21] Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. Abandon Statistical Significance. *arXiv:1709.07588 [stat]*, September 2017.
- [22] Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard

- Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, page 1, September 2017.
- [23] Harry Crane. Why "Redefining Statistical Significance" Will Not Improve Reproducibility and Could Make the Replication Crisis Worse. *arXiv:1711.07801 [stat]*, November 2017. arXiv: 1711.07801.
 - [24] Geoff Cumming. The New Statistics: Why and How. *Psychological Science*, 25(1):7–29, January 2014.
 - [25] Victoria Savalei and Elizabeth Dunn. Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*, 6, 2015.
 - [26] Jacob Cohen. The earth is round ($p < .05$). *American Psychologist*, 49(12):997–1003, 1994.
 - [27] Richard D. Morey, Rink Hoekstra, Jeffrey N. Rouder, Michael D. Lee, and Eric-Jan Wagenmakers. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1):103–123, February 2016.
 - [28] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Taylor & Francis Ltd, Boca Raton, 3 rev ed. edition, November 2013.
 - [29] John K. Kruschke. What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14(7):293–300, July 2010.
 - [30] R Core Team. R: A Language and Environment for Statistical Computing, 2015.

A Reproducing the test score example in R

The test score example in the text can be reproduced within the statistical environment R [30] with the following commands:

```
1 # Necessary packages
  library(ggplot2)
3 library(reshape2)
  library(dplyr)
5
  # Creating the fake data
7 set.seed(42)
  A <- rnorm(30, mean=70, sd=8)
9
  set.seed(21)
11 B <- rnorm(30, mean=65, sd=8)
13
  A <- data.frame(A)
  B <- data.frame(B)
15
  ab <- cbind.data.frame(A$A, B$B)
17 names(ab) <- c("A$A" = "A", "B$B"="B")
19
  ab.long <- melt(ab)
  ab.long <- rename(ab.long, Class = variable)
21
  # Creating the chart
23 ggplot(ab.long, aes(x=value, fill=Class)) +
  geom_histogram(binwidth=2) +
25 xlab("Test scores") +
  ylab("") +
27 scale_x_continuous(limits=c(0,90), breaks=seq(0,90, by=10))
  +
  facet_wrap(~ Class, ncol = 1)
29
  # Performing the Welch t-test
31 t.test(ab$A, ab$B)
```