https://doi.org/10.25819/ubsi/2961

New Perspectives on Statistical Data Analysis: Challenges and Possibilities of Digitalization for Hypothesis Testing in Quantitative Research

Riko Kelter[1]

Contact: Riko Kelter, Department of Mathematics, University of Siegen, riko.kelter@uni-siegen.de

[1] University of Siegen, Siegen, Germany

Abstract. p-values, the 'gold standard' of statistical validity are not as reliable as many scientists assume. In the last decade, severe problems have been observed regarding the validity of highly reputable research. Additionally, the growing availability of big data challenges the design and statistical analysis of studies and experiments across science. Therefore, it is more important than ever to make the best use of available computational tools, software and possibilities digitalization offers to improve the validity of research results. In this paper, we focus on an essential procedure often carried out in quantitative research, which is directly related to the experienced problems: Statistical hypothesis testing. First, we show that the traditional way of hypothesis testing has severe logical problems. Second, it is shown that due to the increasing availability of computational resources, highly sophisticated methods from the area of computational statistics - namely Bayesian data analysis - can complement and even replace traditional hypothesis testing. Third, we highlight how digitalization helps in making these technologies available to a vast range of researchers in the form of the novel and free software package JASP. Together, this paper shows that considering a change in perspective on statistical data analysis, in particular on hypothesis testing, provides the possibility to improve the transparency and reliability of research in the medical, social and natural sciences.

Keywords: Data Analysis, Mathematical Psychology, Hypothesis Testing, Bayesian Statistics, Statistical Inference

1 Introduction

In 2005, epidemiologist John P. Ioannidis of Stanford University suggested that most published research findings are false (Ioannidis, 2005). Since then, countless papers have explored the situation many scientists face for nearly two decades now (Begley & Ioannidis, 2015; McElreath & Smaldino, 2015). These include problems with the replication of existing study results and the validity of a vast amount of highly reputable research. Entitled as the replication crisis (Baker & Penny, 2016), a string of publications detailing how these problems form has forced scientists to reconsider how research results are evaluated

(Colquhoun, 2017; Ioannidis, 2016). In particular, statistical data analysis has been identified as one major piece in the big puzzle of the replication crisis, causing even scientists with the best intentions into trouble. A large part of the observed problems was already attributed to the "surprisingly slippery nature of the p-value, which is neither as reliable nor as objective as most scientists assume", as (Nuzzo, 2014) notes. While the p-value is often used to identify significant research findings and study results in quantitative research, in a large number of cases it produces false-positive results, that is, states an effect if none is present. This fact is highly problematic, as reducing the number of false-positive results is one of the biggest necessities of contemporary science (McElreath & Smaldino, 2015). The situation even led the American Statistical Association (ASA) to release an official statement in 2016, which stressed that *"by itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."* (American Statistical Association, 2016). What is more, in light of the problems the ASA recommended to supplement or even replace p-values with other approaches which *"emphasize estimation over testing such as (...) Bayesian methods"* (American Statistical Association, 2016) and "alternative measures of evidence such as likelihood ratios or Bayes factors" (American Statistical Association, 2016). Many approaches have been proposed to counteract the problems identified in p-values (Wasserstein et al., 2019). The ideas range from methodological shifts (Kruschke & Liddell, 2018) to simpler options as applying stricter standards for declaring statistical significance (Benjamin et al., 2018). While the ongoing problems are far from being solved, through the debate about statistical significance an increasing number of scientists has become aware that it is necessary to change current practices of data analysis, especially hypothesis testing (McElreath, 2020; Wasserstein et al., 2019). In this paper, we show how Bayesian data analysis can replace traditional p-values, leading to more reliable conclusions. Also, we showcase how digitalization helps to foster transparent and reproducible research by presenting the statistical software JASP. JASP has been developed at the University of Amsterdam and implements a vast range of highly-sophisticated Bayesian statistical methods, making it an attractive candidate to improve the reproducibility of research.

2 Null hypothesis significance testing

In this section, we briefly review the theory behind p-values, which are part of *null hypothesis significance testing* (NHST). Also, we highlight some of the logical fallacies of NHST.

2.1 A brief introduction to NHST

The traditional way of hypothesis testing goes back to the early 20th century. In the approach of (Neyman & Pearson, 1936), who published their highly influential theory in the 1930s, the general format is to test a *null hypothesis* H_0, which makes a statement about a parameter δ against the *alternative hypothesis* H_1. After conducting the experiment and calculating the hypothesis test, the experimenter either has to accept or reject H_0. Due to the nomenclature one often calls this procedure *null hypothesis significance testing* (NHST). A *hypothesis test* can now simply be interpreted as a rule stating for which observed sample values the decision is made to reject H_0. The values for which H_0 will be rejected is called the *rejection region*. To construct a hypothesis test, the so-called *sampling statistic* of the quantity of interest, the parameter δ, is considered. For example, when comparing two normally distributed groups like a treatment and control group in a randomized controlled trial (RCT), often the quantity of interest is the difference in means $\mu_1 - \mu_2$ of both groups. The distribution of the differences in means $\mu_1 - \mu_2$ under the null hypothesis H_0 - that is, the sampling statistic - can be derived theoretically (Held & Sabanés Bové, 2014). After conducting the study and observing the quantity of interest, for example, $\mu_1 - \mu_2 = 3$,

the known distribution of $\mu_1 - \mu_2$ is used to determine how plausible it is to obtain a difference of $\mu_1 - \mu_2 = 3$ or even larger differences. Figure 1 visualizes NHST: On the x-axis, it shows the set of possible results, which in the above example are all possible values of $\mu_1 - \mu_2$. Based on theoretical results, the distribution of this quantity $\mu_1 - \mu_2$ under the

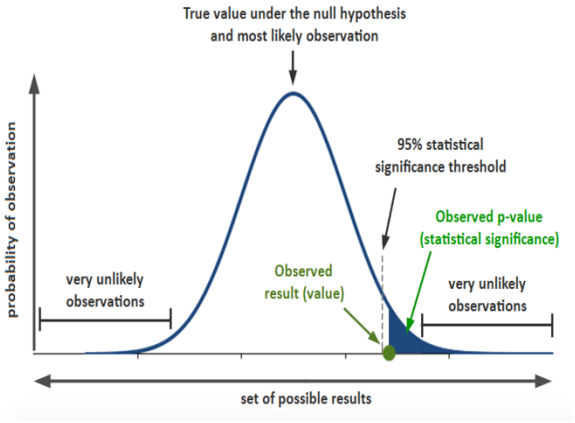

Figure 1. Null hypothesis significance testing

null hypothesis H_0 is well known, shown as the bell-shaped density in figure 1. Under H_0, it is quite unlikely to observe very small values or huge values. In the above example, when observing a result like $\mu_1 - \mu_2 = 3$, the idea of *statistical significance* is to calculate the probability of obtaining a difference equal to or more extreme than the difference observed. This probability is the coloured area right to the observed result in figure 1, and this is exactly the p-value often reported in quantitative research. If the p-value is small, it seems plausible to reject the null hypothesis, because observing such a large difference would be highly unlikely under H_0. *"The p-value is the probability, under the assumption of the null hypothesis H_0, to obtain a result equal to or more extreme than what was actually observed."* (Held & Sabanés Bové, 2014). Over time, the well known 95% statistical significance threshold has manifested itself in science, which was invented by (Fisher, 1925). The threshold is shown as the dashed vertical line in figure 1 and simply states that one should reject the null hypothesis H_0, whenever the p-value is smaller than 0.05. That means one rejects H_0 whenever one would observe a difference equal to the one observed or more extreme with 5% or less probability under the null hypothesis H_0. It is important to note that formally, a continuous quantification of the p-value when using the Neyman-Pearson theory is not allowed. The p-value can only be interpreted as a binary value for the decision against (if $p < 0.05$) or for H_0 (if $p \geq 0.05$).

In summary, frequentist hypothesis testing can be seen as a procedure targeted at the long-term type I error control.

2.2 Problems with NHST and p-values

NHST may seem reasonable at first. Nevertheless, there are some severe logical fallacies which we want to pinpoint here. These problems question the usefulness of NHST for practical research and call for other options. First, there are two types of errors which need to be considered: If the null hypothesis H_0 is true, but the hypothesis test incorrectly decides to reject H_0, then the test has made a *type I error*. If the null hypothesis H_0 is false, but the hypothesis test incorrectly accepts H_0, then the test has made a *type II error*. Table 1 gives an overview:

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision	Type I error
	H_1	Type II error	Correct decision

Table 1. Type I and II errors in hypothesis tests

Formally, every method of statistical testing can make these two types of errors. Nevertheless, NHST was developed to control the type I error while simultaneously minimizing the type II error (Neyman & Pearson, 1936).

2.3 Type I error control is not always appropriate and is not bullet-proof

The preference for type I error control is highly questionable in applied research. Consider a diagnostic test for a disease which uses a blood

sample to calculate the concentration of specific antibodies. Suppose one knows the number of antibodies follows a particular distribution in healthy individuals. Suppose also, that very large (or small) values which pass the 95% significance threshold indicate the presence of an autoimmune disease. When applying NHST and testing patients, the testing procedure will minimize the type I error. A type I error happens if a patient is told that she has the disease, but the patient is healthy. The consequence of a type I error is mild: Further diagnostics will show that the result was a false-positive one, and the caused costs are small. Consider now a type II error: A patient who has the disease will be sent home with a false-negative result. The condition will progress until the patient makes a second test and is diagnosed and treated correctly. The damage done is considerable: The disease has progressed, causing subsequent treatments to be more expensive next to the fact that the patient suffers unnecessarily. The example shows that type II error control is preferable to type I error control in some prevalent settings, making the usefulness of NHST questionable.

Additionally, countless papers have demonstrated that the type I error control guaranteed by NHST is often not attained. This leads to an uncomfortable situation in which a long-term type I error rate of 5%, in reality, equals a skyrocketing 36% false-positive rate (Colquhoun, 2014).

2.4 Falsification or confirmation?

The second problem of NHST is rooted in the philosophy of science itself. Due to space limitations, we cannot offer a full account here. However, falsification only makes sense when the goal is to narrow down a substantial number of research hypotheses. In other cases, researchers are more interested in *confirming* research hypotheses. Whether this refers to showing the effectiveness of a new drug, the efficacy of psychological interventions, or the improved performance of a new computational algorithm, scientists often need to confirm that a hypothesis is indeed correct (or at least the most suitable of a set of candidates). Additionally, scientists often need to rephrase research hypotheses to make them rejectable via falsification. For example, if the goal is to show that a drug for lowering blood pressure works, falsification forces scientists to formulate the hypothesis as $H_0: \mu_1 = \mu_2$, where μ_1 and μ_2 are the group means of the treatment and control group in the study. The actual goal is to *reject* this hypothesis H_0 to *confirm* that the drug works as expected. When discarding H_0, scientists still do not know how large the difference between μ_1 and μ_2 is, which is of much more interest than only stating that the difference is non-zero. After all, the difference could be negligibly small, although significant, making the research results scientifically less or even entirely irrelevant.

2.5 Dependence on the researcher's intentions

The third point is the most problematic: The findings and interpretation of NHST depend on the researcher's intentions. For example, it plays a crucial role if the number of participants in a study is fixed in advance, or if researchers sample participants until time or money runs out (Kruschke & Liddell, 2018). This situation causes unnecessary strain on financial and personal resources and makes the interpretation of results obtained via NHST difficult. When reporting such findings, researchers can unintendedly invalidate all their work by violating their sampling plan. Also, this opens the door to misuse of statistics by reporting a different sampling plan after the actual study has been conducted only to obtain a significant result. This practice is often called 'p-hacking' and is observed widely by now, which is worrisome (Ioannidis, 2019). Also, NHST violates the *likelihood principle* (LP), which is one of the most critical proven results in mathematical statistics (Berger & Wolpert, 1988).

3 Bayesian data analysis as an alternative

In this section, we review the theory behind Bayesian data analysis, which is an often proposed alternative to NHST. We highlight how some of the logical fallacies of NHST are avoided by considering the Bayesian approach.

3.1 Bayesian parameter estimation

It is helpful first to introduce the general idea behind Bayesian parameter estimation to get familiar with the conventional notation. Bayesian parameter estimation centres on the posterior distribution $p(\theta|x)$ of the unknown parameter θ after observing the experimental data x, which are assumed to follow a specific probability density $p(x|\theta)$, the likelihood function. The posterior distribution reflects the relative plausibility of different parameter values after the available prior knowledge $p(\theta)$ has been updated employing the data x via the likelihood function $p(x|\theta)$. Correctly, one starts by assigning the model parameters θ a prior distribution $p(\theta)$. The information in the observed data x is then used to update this prior information to the posterior distribution, where parameter values which yielded good predictions of the observed data x get a boost in plausibility.

$$p(\theta) \cdot p(x|\theta) \propto p(\theta|x) \quad (1)$$

The $\propto$ symbol in equation (1) means 'proportional to'. As modern sampling algorithms like Markov-Chain-Monte-Carlo (MCMC) which produce the posterior distribution numerically only need a function proportional to the posterior, it suffices to write the posterior in this way (McElreath, 2020). Analytic derivations of the posterior distribution are possible only for simple statistical models so that most realistic models require the use of MCMC algorithms (Robert & Casella, 2004). The posterior distribution can be summarized using point or interval estimates, like the posterior mean or median, or credible intervals. Credible intervals include a fixed percentage - for example, 95% - of the posteriors probability mass, and thereby make it possible to state in what range of parameter values the true parameter θ lies with a given probability. Note that this interpretation is often applied to frequentist confidence intervals, which is false.

3.2 Bayesian hypothesis testing

The structured approach to Bayesian hypothesis testing uses the Bayes factor (Jeffreys, 1961). The Bayes factor quantifies the relative predictive performance of two rival hypotheses. It can be interpreted as the degree to which the data demand a change in beliefs towards one of both hypotheses under consideration.

$$\frac{P(H_1)}{P(H_0)} \cdot \frac{p(x|H_1)}{p(x|H_0)} = \frac{P(H_1|x)}{P(H_0|x)} \quad (2)$$

The first term in equation (2) is the prior odds, which is the relative plausibility of the two hypotheses before observing any data. The second quantity is the Bayes factor $BF_{10}(x)$, which indicates the evidence provided by the data x observed. The third term, the posterior odds, indicates the relative plausibility of both hypotheses after having seen the data and is calculated as the product of the prior odds and the Bayes factor. The subscript in the Bayes factor $BF_{10}(x)$ indicates which hypothesis is supported by the observed data: $BF_{10}(x)$ is the Bayes factor in favour of H_1, and $BF_{01}(x)$ is the Bayes factor in favour of H_0. Algebraic rearrangements show that $BF_{01}(x) = 1/BF_{10}(x)$. Large values of $BF_{10}(x)$ signal more support for H_1 and the Bayes factor ranges from zero to ∞. A Bayes factor of 1 indicates that H_0 and H_1 both predict the observed data x equally well.

4 An example of digitalization in statistical data analysis: JASP

In this section, we show how Bayesian data analysis can be conducted by using the open-source statistical software package JASP (JASP Team, 2019). Through the advancing

digitalization and availability of more powerful computing resources, Bayesian methods are available to researchers in the form of software like JASP today without the need to code complicated programming. We use an example from medical science to show that more valuable information is obtained when considering Bayesian hypothesis testing. Also, we show how the reporting of research results is digitalized and made more transparent through JASP. A typical question arising in medical research is used as a scaffold to showcase the usefulness of Bayesian hypothesis testing: Do two groups (pre-treatment, after-treatment) differ on an observed metric variable, and if so, how large is the effect size between both groups? Usually, NHST compares the means μ_1 and μ_2 of the same population at two different time points via Student's paired-samples t-test to reject the null hypothesis via the use of p-values (Kelter, 2020a).

4.1 A Bayesian paired-samples t-test

The dataset used is from (Moore et al., 2012) and provides the number of disruptive behaviours by dementia patients during two different phases of the lunar cycle. The hypothesis tested is H_0: "The average number of disruptive behaviours in patients with dementia does not differ between full moon and other days" against the alternative H_1 of a differing average number of disruptive behaviours.

	t	df	p	Mean Difference
Moon - other	6.452	14	< .001	2.433

Table 2. Paired-samples t-test results for the dementia dataset obtained from NHST

Table 2 shows the results of the paired-samples t-test, indicating with $p < .001$ that H_0 can be rejected. Note that this is not what researchers want to know: The desired answer is which hypothesis is more probable after observing the data, which is precisely quantified by the posterior odds $P(H_1|x)/P(H_0|x)$. Note also that the Bayes factor BF_{10} is a crucial ingredient in the posterior odds because the posterior odds are the product of the Bayes factor and the prior odds. A large BF_{10} therefore necessitates a change in beliefs towards H_1. Assumption checks include a Shapiro-Wilk test on normality, which is not significant at $p = .148$.

	BF_{10}	Error %
Moon - other	1521.058	5.014e-7

Table 3. Bayesian paired-samples t-test results for the dementia dataset

Now, the Bayesian paired-samples t-test shown in table 3 in contrast yields $BF_{10} = 1521.058$, indicating extreme evidence for H_1. JASP also produces a plot of the prior and posterior distribution of the effect size δ, which is of interest in most medical research settings. Figure 2 shows the prior and posterior

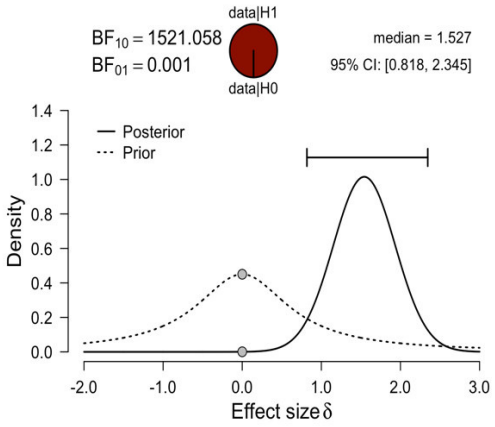

Figure 2. Bayesian data analysis of the dementia dataset of (Moore et al., 2012): Prior-posterior plot of the effect size

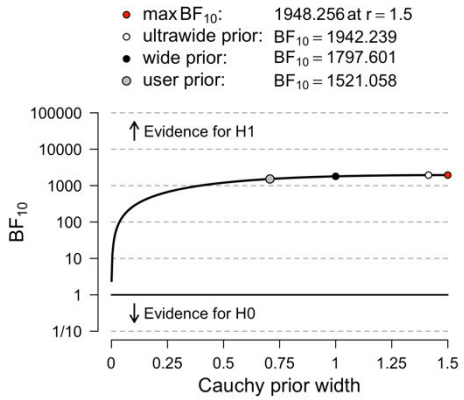

Figure 3. Bayesian data analysis of the dementia dataset of (Moore et al., 2012): Robustness analysis for the Bayes factor using varying Cauchy prior widths

plot of the effect size δ as well as the produced BF_{10}. The posterior of the effect size δ precisely estimates which effect size is the most probable after observing the data x. Note that the traditional paired-samples t-test did not yield any information about the effect size. Although it was significant, it did not state whether the observed effect is small, medium or large. The prior-posterior plot shows how the prior probability mass is reallocated to the posterior after observing the data and shows that with 95% probability, the true effect size δ is in $[0.818, 2.345]$ and the posterior median is 1.527, which indicates a large effect (Cohen, 1988). Another benefit is given by the robustness check shown in figure 3: Different prior distribution widths are used for the effect size δ and the Bayes factor BF_{10} is computed. Specifically, the prior width of the Cauchy prior $C(0, \gamma)$ on the effect size δ is increased gradually, showing how the prior shape influences the resulting Bayes factor BF_{10}. Figure 3 shows that even when changing the prior from the user prior, which equals a medium $C(0, \sqrt{2}/2)$ prior, to a wide $C(0,1)$ or even ultrawide $C(0, \sqrt{2})$ prior, the Bayes factor for H_1 stays above 1000. Therefore, the influence of the prior is negligible here, and only an insignificant amount of subjectivity goes into the analysis.

5 Discussion

The two examples above highlighted how Bayesian data analysis, including hypothesis testing via the Bayes factor, is efficiently conducted with JASP. Next to the ease-of-use, there are multiple benefits when considering the Bayesian way of hypothesis testing: (1) Testing statistical hypothesis with the Bayesian approach is following the likelihood principle. (2) It does not matter if one fixes the sample size of the study or experiment in advance or samples until time or money runs out. This fact is particularly important from a practical perspective. (3) In contrast to NHST, Bayesian data analysis can *confirm* research hypotheses under consideration. (4) The computational requirements to conduct Bayesian data analyses have been reduced significantly in the last years, making the approach available to a wide range of users. In combination with attractive software options like JASP, digitalization has therefore opened up new possibilities for researchers to improve the reliability and transparency of research results.

Still, there remain some challenges and limitations: The computational effort is larger when conducting Bayesian data analyses, which is caused by the substantial numerical calculations required for producing the posterior distribution. This is, in particular, true for complex and high-dimensional models (Kelter, 2020b). Still, for most standard models like linear regression or Student's two-sample t-test, there exist either analytic solutions or the computational effort is moderate, which leads to a seamless experience when using JASP. Note, that the flexibility of extending and adapting statistical models to one own's needs is also a big benefit of the Bayesian approach, and for a brief introduction, see (Kelter, 2020b) or (Kelter, 2020c; McElreath, 2020).

Another problem is concerned with keeping the influence of the prior selection as minimal as possible. While prior elicitation is an important topic in the literature (Held & Sabanés Bové,

2014; Kruschke & Liddell, 2018), the robustness checks available in JASP prevent cherry-picking the most suitable prior for obtaining the desired conclusions from raw research data.

Two aspects of particular importance not mentioned so far remain: First, all analyses conducted in JASP can be saved in a `.jasp` workflow file, which includes all data, analyses and results obtained. This possibility enables researchers from other laboratories to recreate reported analyses. Second, rich visualizations like the ones presented in the two examples above can easily be exported in JASP, which improves the digitalization of research. Third, JASP has built-in support for the Open Science Foundation (Open Science Foundation, 2020), which gives scientists the possibility to make their data, code and material available to others digitally.

6 Conclusion

Digitalization poses various challenges and opens new possibilities for scientists. In this paper, we focussed on the essential procedure of statistical hypothesis testing often carried out in quantitative research. First, we showed that the traditional way of hypothesis testing, NHST, has severe logical problems. Second, it was shown that due to the increasing availability of computational resources, Bayesian data analysis could complement and even replace NHST. Third, we highlighted how digitalization helps in incorporation of these methods into work. A brief presentation of the free statistical software JASP showed how easily Bayesian hypothesis testing is conducted, and a vast range of researchers should be able to benefit from considering the Bayesian approach. Interested readers should also take a look at the R software packages `bayest` (Kelter, 2019), which provides a convenient implementation of Bayesian t-tests in R. A review of how to improve the reproducibility in medical research by employing Bayesian posterior indices is given by (Kelter, 2020a). In summary, this paper highlighted the emerging possibilities digitalization has created for scientists from the medical, social and natural sciences when it comes to statistical hypothesis testing. Considering a change in perspective towards Bayesian hypothesis testing should, therefore, foster transparent, reproducible research across science.

7 References

American Statistical Association. (2016). American Statistical Association Press Release for the Statement on Statistical Significance and P-Values. https://doi.org/10.1080/00031305.2016.1154108.Vt2XIOaE2MN

Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? Nature, 533(7604), 452–454. https://doi.org/10.1038/533452A

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. In Circulation Research (Vol. 116, Issue 1, pp. 116–126). https://doi.org/10.1161/CIRCRESAHA.114.303819

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. Nature Human Behaviour, 2(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berger, J. O., & Wolpert, R. L. (1988). The Likelihood Principle. In S. S. Gupta (Ed.), Lecture Notes - Monograph Series. Institute of Mathematical Statistics. http://www.jstor.org/stable/4355509

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2 edition). Routledge.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, 1(3), 140216–140216. https://doi.org/10.1098/rsos.140216

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. Royal Society Open Science, 4(12). https://doi.org/10.1098/rsos.171085

Fisher, R. A. (1925). Statistical Methods for Research Workers (O. and Boyd (ed.)). Oliver and Boyd, Hafner Publishing Company.

Held, L., & Sabanés Bové, D. (2014). Applied Statistical Inference. Springer-Verlag Berlin

Heidelberg. https://doi.org/10.1007/978-3-642-37887-4

Ioannidis, J. P. A. (2005). Why most published research findings are false. In PLoS Medicine (Vol. 2, Issue 8, pp. 0696–0701). Public Library of Science. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2016). Why Most Clinical Research Is Not Useful. PLoS Medicine, 13(6). https://doi.org/10.1371/journal.pmed.1002049

Ioannidis, J. P. A. (2019). What Have We (Not) Learnt from Millions of Scientific Papers with p-Values? The American Statistician, 73, 20–25. https://doi.org/10.1080/00031305.2018.1447512

JASP Team. (2019). JASP (0.11.1). https://jasp-stats.org/

Jeffreys, H. (1961). Theory of Probability. In Oxford Classic Texts in the Physical Sciences (3rd ed.). Oxford University Press.

Kelter, R. (2019). bayest - Bayesian t-test (R package version 1.1). Comprehensive R Archive Network. https://cran.r-project.org/web/packages/bayest/index.html

Kelter, R. (2020a). Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Medical Research Methodology, (in press). https://doi.org/https://doi.org/10.1186/s12874-020-00968-2

Kelter, R. (2020b). Bayesian survival analysis in STAN for improved measuring of uncertainty in parameter estimates. Measurement: Interdisciplinary Research and Perspectives, (in press). https://doi.org/10.1080/15366367.2019.1689761

Kelter, R. (2020c). Statistical Rethinking: A Bayesian Course with examples in R and STAN (2nd ed.). Measurement: Interdisciplinary Research and Perspectives, (in press). https://doi.org/10.1080/15366367.2020.1742561

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics : Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin and Review, 25, 178–206. https://doi.org/10.3758/s13423-016-1221-4

McElreath, R. (2020). Statistical rethinking: A Bayesian course with examples in R and Stan. CRC Press. https://doi.org/10.1201/9781315372495

McElreath, R., & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. PLoS ONE, 10(8), 1–16. https://doi.org/10.1371/journal.pone.0136088

Moore, D. S., McCabe, G. P., & Craig, B. A. (2012). Introduction to the practice of statistics (9th ed.). W. H. Freeman.

Neyman, J., & Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses. Statistical Research Memoirs, 1, 1–37. https://psycnet.apa.org/record/1936-05541-001

Nuzzo, R. (2014). Statistical errors: P values, the “gold standard” of statistical validity, are not as reliable as many scientists assume. Nature, 506(7487), 150–152. https://doi.org/10.1136/bmj.1.6053.66

Open Science Foundation. (2020). OSF - Open Science Foundation. https://osf.io/

Robert, C., & Casella, G. (2004). Monte Carlo statistical methods. Springer.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond “p<0.05.” The American Statistician, 73(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913