Bayesian Reanalyses from Summary Statistics: A Guide for Academic Consumers

Alexander Ly
University of Amsterdam
Centrum Wiskunde & Informatica, Amsterdam

Akash Raj
University of Amsterdam

Alexander Etz
UC Irvine

Maarten Marsman
University of Amsterdam

Quentin F. Gronau
University of Amsterdam

Eric-Jan Wagenmakers
University of Amsterdam

Abstract

Across the social sciences, researchers have overwhelmingly used the classical statistical paradigm to draw conclusions from data, often focusing heavily on a single number: p. Recent years, however, have witnessed a surge of interest in an alternative statistical paradigm: Bayesian inference, in which probabilities are attached to parameters and models. We feel it is informative to provide statistical conclusions that go beyond a single number, and –regardless of one's statistical preference– it can be prudent to report the results from both the classical and the Bayesian paradigm. In order to promote a more inclusive and insightful approach to statistical inference we show how the open-source software program `JASP` (`https://jasp-stats.org`) provides a set of comprehensive Bayesian reanalyses from just a few commonly-reported summary statistics such as t and N. These Bayesian reanalyses allow researchers –and also editors, reviewers, readers, and reporters– to quantify evidence on a continuous scale, assess the robustness of that evidence to changes in the prior distribution, and gauge which posterior parameter ranges are more credible than others by examining the posterior distribution of the effect size. The procedure is illustrated using the seminal Festinger and Carlsmith (1959) study on cognitive dissonance.

Classical null hypothesis significance testing (NHST) allows researchers to evaluate scientific propositions in a seemingly straightforward manner: whenever the p-value falls below a threshold α (usually set to .05) researchers feel licensed to reject the null hypothesis that the effect is absent and embrace the alternative hypothesis that the effect is present. For example, in the results section one may encounter conclusions such as "overall classification accuracy was greater than chance", "the analysis revealed a main effect of the manipulation", and "the correlation was significant"; in the discussion section, these statements are abstracted from the standard NHST framework even further, conveying the impression that whenever $p < .05$, the data strongly favor the alternative hypothesis over the null hypothesis of no effect.

The field's mechanistic use of p-values appears to be at odds with the recent warning issued by the *The American Statistical Association* (ASA; Wasserstein & Lazar, 2016, p. 131): "The widespread use of 'statistical significance' (generally interpreted as '$p \leq 0.05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process." Indeed, p-values have been critiqued on numerous grounds (e.g., Greenland et al., 2016; Nickerson, 2000; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016; Wagenmakers, Marsman, et al., 2018). One widely appreciated concern is that p-values do not convey information about the size of the effect or the precision with which that effect is estimated (e.g., Cumming, 2014).

As one prominent alternative to p-value NHST, there is a growing trend for psychologists to employ Bayesian statistics (e.g., Vandekerckhove, Rouder, & Kruschke, 2017; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017; Wagenmakers, Morey, & Lee, 2016). Within the Bayesian framework, prior uncertainties about parameters and models are updated by means of observed data to yield posterior uncertainties (Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2018). For instance, the posterior distribution quantifies our knowledge about a non-zero effect size, which is useful for parameter estimation; on the other hand, the Bayes factor contrasts the predictive adequacy of two competing models, which is useful for hypothesis testing (Etz & Vandekerckhove, 2018). Specifically, the Bayes factor quantifies the degree to which the data are more likely under one model versus another (e.g., Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995; Myung & Pitt, 1997). For example, the Bayes factor can be used to compare a null hypothesis $\mathcal{H}_0$ (i.e., there is no effect) to an alternative hypothesis $\mathcal{H}_1$ (i.e., there is an effect). The Bayes factor BF_{10} for $\mathcal{H}_1$ over $\mathcal{H}_0$ has an intuitive interpretation; $\text{BF}_{10} = 7$ indicates that the observed data are 7 times more likely under the alternative hypothesis $\mathcal{H}_1$ than under the null hypothesis $\mathcal{H}_0$, whereas $\text{BF}_{10} = 0.2$ indicates that the observed data

AL and EJW are supported by the starting grant "Bayes or Bust" awarded by the European Research Council (Grant #283876) and grant 016.Vici.170.083 from the Netherlands Organisation for Scientific Research (NWO). AE was supported by grant #1534472 from NSF's Methods, Measurements, and Statistics panel, as well as the National Science Foundation Graduate Research Fellowship Program #DGE1321846. QFG was supported by grant #406.16.528 from NWO. We thank the editor, Michael Masson, Sang Ho Lee, Jay Myung, and two reviewers for their constructive suggestions for improvement. We are also grateful to Raoul Grasman for his contributions to `JASP`. Centrum Wiskunde & Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands. Correspondence concerning this article may be addressed to Alexander Ly, University of Amsterdam, Psychological Methods Department, Postbus 15906, 1001 NK Amsterdam, the Netherlands. Email address: a.ly@uva.nl.

are 5 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$. In general, the Bayes factor returns a non-negative number, and the higher (lower) the value of BF_{10}, the more (less) evidence the data provide for $\mathcal{H}_1$ over $\mathcal{H}_0$. When the two hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$ are equally plausible a priori, then a Bayes factor of $\text{BF}_{10} = 6$ indicates that the posterior plausibility for the alternative hypothesis $\mathcal{H}_1$ is 86% (i.e., 6/7), leaving 14% (i.e., 1/7) posterior probability for the null hypothesis $\mathcal{H}_0$.

A discussion on the merits and demerits of the different statistical paradigms is beyond the scope of this paper. We agree with the ASA's recommendation to go beyond p, and that it is prudent to adopt an inclusive statistical approach. For when the results of different statistical paradigms point in the same direction, this bolsters one's confidence in the conclusions, but when the results are in blatant contradiction, this weakens one's confidence.

In the spirit of promoting a more inclusive statistical approach, our primary goal is to demonstrate the ease with which published classical results can be subjected to a Bayesian reanalysis using the recently developed "Summary Stats" module in `JASP` (JASP Team, 2018). Depending on the analysis at hand, this module takes as input commonly-reported statistics such as t, r, and R^2 together with sample size N, and returns a comprehensive Bayesian assessment.[1] Importantly, this Bayesian assessment can be executed in the absence of the raw data. This is essential when the data are no longer available or when they cannot be shared; but even when the raw data are publicly available, the reanalysis with the "Summary Stats" module is much more efficient – reviewers, readers, and reporters can obtain a comprehensive Bayesian assessment almost instantaneously. We believe that the richness of a Bayesian report contrasts favorably with a report of just the summary statistics themselves. We illustrate this claim using a seminal study published more than half a century ago.

The Festinger & Carlsmith (1959) Cognitive Dissonance Study

In a landmark publication,[2] Festinger and Carlsmith (1959, hereafter FC) outlined a theory to account for *cognitive dissonance*, a phenomenon they described as follows: "If a person is induced to do or say something which is contrary to his private opinion, there will be a tendency for him to change his opinion so as to bring it into correspondence with what he has done or said" (p. 209). Earlier experiments on cognitive dissonance (e.g., Kelman, 1953) induced participants to make a statement contrary to their personal opinion for a chance to gain a reward. It was hypothesized that for greater rewards there would be a greater change to the opinion, but the data showed the reverse: the smaller the reward, the greater change in opinion. FC proposed a theory that could account for this behavioral pattern, which they subsequently put to the test in an ingenious experiment.

FC's experiment included control, high reward, and low reward conditions, each with twenty participants. All participants performed a boring task for one hour, after which they were asked to take a survey and answer questions about, among other things, their enjoyment of the study. Where the conditions differ is what happens after completing the

[1] The website `https://web.archive.org/web/20170212075534/http://pcl.missouri.edu/bayesfactor`, designed and maintained by Jeff Rouder, exploits the same idea, but focuses exclusively on the Bayes factor.

[2] Cited over 3,540 times according to Google Scholar, February 2, 2018.

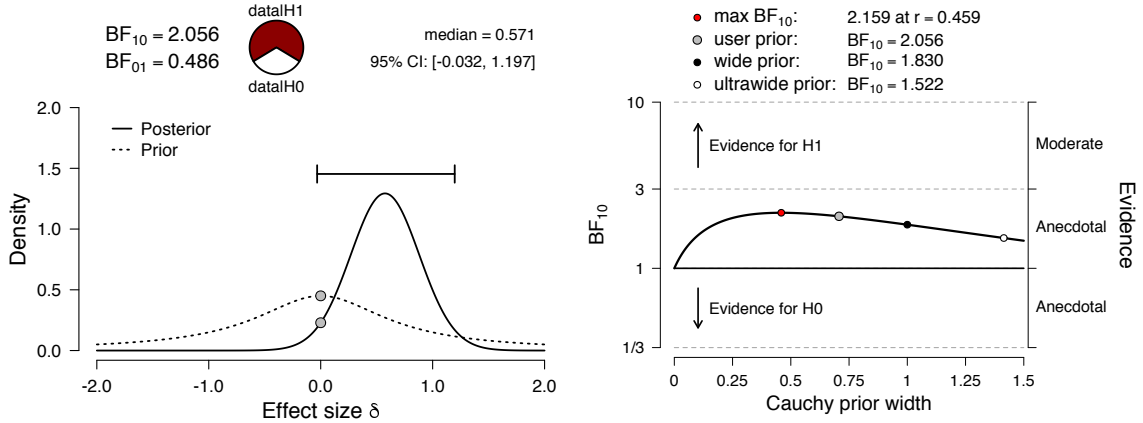

Figure 1. A comprehensive Bayesian reanalysis of the seminal study by Festinger and Carlsmith (1959), obtained by entering $t = 2.22$ and $N_1 = N_2 = 20$ into the JASP Summary Stats module, see text for details. Figures from JASP.

boring task, but before completing the survey. In the reward conditions, participants were asked to interact with a confederate by telling them that the experiment was interesting and fun; for this they received either twenty dollars (high reward) or one dollar (low reward). In the control condition participants went straight to the post-interview and did not interact with the confederate. According to FC, the crucial test of their theory lies in comparing the post-interview enjoyment ratings from the low versus high reward conditions, where the low reward condition is predicted to have higher enjoyment ratings. In line with their theory's prediction, FC found a higher (sample) mean enjoyment rating in the low reward group than in the high reward group, $t(38) = 2.22$, $p = .032$, and this was taken as support for their theoretical position. No effect size estimate is reported in the original paper, but this can be easily computed from the t-value and the group sizes, giving a Cohen's d of $d = 0.702$.

Bayesian Reanalysis

We wish to conduct a Bayesian reanalysis of the FC result, but the raw data from this study are no longer available. However, the Summary Stats module in JASP affords a comprehensive Bayesian reanalysis using only the test statistic reported in the original paper.[3] Inputting the reported t-value and sample sizes for the two groups yields the results shown in Fig. 1.

In the left panel, the dotted line represents the default prior distribution for the population effect size δ under $\mathcal{H}_1$: a zero-centered *Cauchy* distribution (i.e., a t-distribution with one degree of freedom; Jeffreys, 1948; Ly, Verhagen, & Wagenmakers, 2016a, 2016b), here with a default scale of $\gamma = 0.707$ (e.g., Morey & Rouder, 2015). Thus, under $\mathcal{H}_1$ –that is, presuming the effect is present– the expectation is that the effect is most likely to be small, although the possibility that it is large is not ruled out.

In the left panel, the solid line is the posterior distribution for effect size, that is,

[3]The Summary Stats module is activated via the + icon next to the "Common" tab at the top of the JASP window.

the knowledge about effect size obtained after updating the prior distribution using the observed data, and assuming that $\mathcal{H}_1$ holds. This posterior distribution has a median of 0.571,[4] and a relatively wide 95% credible interval that ranges from −0.032 to 1.197. In other words, 95% of the posterior mass lies in the interval from −0.032 to 1.197; clearly, the effect has not been estimated with much precision. More generally, by computing the area under the posterior distribution between $\delta = a$ and $\delta = b$, one can assess how plausible it is that the population effect size δ falls in the interval from a to b after the data have been observed (e.g., Wagenmakers, Love, et al., 2018; Wagenmakers et al., 2016). For instance, by comparing the area under the posterior distribution to the right of zero against that to the left of zero, we quantify how much more likely it is that the effect is positive rather than negative, under $\mathcal{H}_1$ – that is, under the presumption that the effect is present.

In general, the posterior distribution quantifies all that we know about the population effect size δ, given that $\mathcal{H}_1$ holds and the effect exists. The latter point is worth emphasizing since it has been argued that one may perform a Bayesian null hypothesis test by judging whether the 95% credible interval overlaps with zero. Despite its beguiling simplicity, such a procedure is incorrect (Berger, 2006; Jeffreys, 1961; Wagenmakers et al., 2017), because it begs the question – the extent to which a null hypothesis is plausible cannot be assessed when this hypothesis has been ruled out in advance (i.e., under the continuous prior distribution (i.e., the Cauchy prior) under $\mathcal{H}_1$, the probability of any single point such as $p(\delta = 0)$ equals zero).

In order to perform a Bayesian hypothesis test, one needs to compare the predictive performance of the null hypothesis $\mathcal{H}_0$ against that of the alternative hypothesis $\mathcal{H}_1$. The result of this comparison is known as the Bayes factor, and the left panel of Fig. 1 reveals that it equals 2.056 – that is, the observed FC data are only about twice as likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$. Bayesian statistician Harold Jeffreys deemed this level of evidence "not worth more than a bare mention" (Jeffreys, 1961, p. 432). The proportion wheel on top visualizes the strength of the evidence.[5] The Bayes factor quantifies relative predictive performance, and the predictive performance from $\mathcal{H}_1$ is determined in part by the prior distribution. Under a default prior specification, it is natural to wonder how robust the conclusions are to plausible changes in the prior distribution. To address this issue, the Summary Stats module allows one to select the option "Bayes factor robustness check". The right panel of Fig. 1 shows the result and the depicts the value of the Bayes factor BF_{10} on the y-axis as a function of the scale γ of the Cauchy prior distribution on the x-axis. The values on the x-axis range from $\gamma = 0$ (when $\mathcal{H}_1$ reduces to $\mathcal{H}_0$ and the Bayes factor is 1 regardless of the data) to $\gamma = 1.5$. Across this entire range, the Bayes factor in favor of $\mathcal{H}_1$ over $\mathcal{H}_0$ never exceeds 3; in fact, the maximum Bayes factor in favor of $\mathcal{H}_1$ equals 2.159, obtained when the width γ is set to 0.459.

So far, we assumed that the prior for effect size is always centered at zero. However, the Bayesian framework can be extended to include informed prior distributions that incorporate context-specific expectations and need not be centered around zero (Gronau, Ly, & Wagenmakers, 2017). As in the reanalysis with default priors, the reanalysis with informed priors is a function solely of the summary statistics. To illustrate the use of an informed prior, we once again reanalyze the result of FC study, but this time with two different priors

[4] Note that the prior distribution of δ has shrunk the sample value of $d = 0.702$ toward zero.

[5] See also `https://osf.io/3acm7/`.

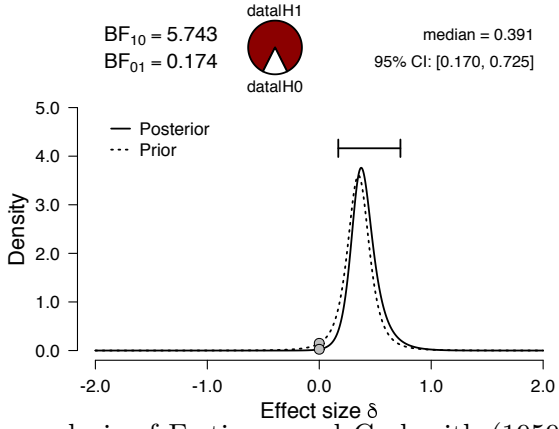

Figure 2. A Bayesian reanalysis of Festinger and Carlsmith (1959), obtained by entering $t = 2.22$ and $N_1 = N_2 = 20$ into the `JASP` Summary Stats module and changing the prior setting to a t-distribution with location 0.35, scale 0.102, and degrees of freedom 3. Figure from `JASP`.

that are centered away from zero. The first informed prior was elicited from Dr. Suzanne Oosterwijk, a social psychologist at the University of Amsterdam. This "Oosterwijk prior" (Gronau et al., 2017) was elicited in the context of a specific effect, but we believe it is plausible more generally for effects in experimental psychology whose presence needs to be ascertained by a statistical analysis. The Oosterwijk prior is a t-distribution with location 0.350, scale 0.102, and 3 degrees of freedom (dotted line in Fig. 2); it assigns most mass to effect sizes from 0.1 to about 0.6. Because this prior is highly informative and the number of observations in the FC study is fairly small, the change from prior to posterior (solid line in Fig. 2) is modest. The Bayes factor BF_{10} indicates that the data are 5.743 times more likely under the informed alternative than under the null hypothesis. Under equal prior model probabilities, this reanalysis leaves $\mathcal{H}_0$ a non-negligible posterior probability of 14.8% (i.e., 1/6.743).

One might suspect that it is possible to find as much evidence for there being an effect, that is, $\mathcal{H}_1$, as desired for a given data set just by changing the prior distribution. This, however, is not true. The second informed prior that we consider is an "oracle" prior that assigns all of its mass to a single point: the observed effect size $d = 0.702$. This "prior" can be obtained in JASP by choosing a normal distribution prior with mean of 0.702 and standard deviation of 0. Note that as a prior it is unrealistic, since in practice it is impossible to know the observed effect size before conducting the experiment. However, it showcases the maximum evidence possible in favor of the alternative hypothesis (Edwards, Lindman, & Savage, 1963). Using this oracle prior, we obtain a Bayes factor of 10.45 in favor of the alternative hypothesis $\mathcal{H}_1$ over the null hypothesis $\mathcal{H}_0$. Hence, even if researchers are blatantly cheating by assigning all prior mass to the observed effect size, the data are only 10.45 times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$; under equal prior model probabilities, such an extreme form of cheating still leaves $\mathcal{H}_0$ a posterior probability of 8.7% (i.e., 1/11.45). In this particular scenario we find that a seminal result, significant with a p-value of .032, does not yield compelling evidence against $\mathcal{H}_0$ when assessed from a default or an informed

Bayesian perspective.[6] We wish to stress that the strength of evidence provided by a Bayes factor can be best appreciated by considering the raw numbers, perhaps visualized as a proportion wheel (see Fig. 1 and 2); the classification scheme proposed by Jeffreys provides a useful, but rough guideline that should not take precedence over a more careful assessment of the strength of evidence.

In sum, the Bayesian reanalyses shown in Figs. 1 and 2 are easily obtained in JASP and paint an inferential picture more complete than the one provided by the statement "$t(38) = 2.22$, $p = .032$".

Concluding Comments

The Summary Stats module in JASP unlocks a comprehensive Bayesian experience from a few commonly-reported summary statistics. Here we illustrated use of the module for the case of an independent-samples t-test, but the Summary Stats module can also be used for inferences concerning paired-samples t-tests, correlation coefficients, binomial proportions, and linear regression models. An entire literature filled with classical statistics is now open for a straightforward Bayesian reanalysis.

In addition to being able to look back on the existing literature, we can also look forward. For instance, editors and reviewers may request that authors include a Bayesian analysis alongside their classical results–which can be accomplished with JASP in mere seconds. For a specific data set, the classical and Bayesian results may disagree (e.g., Wetzels et al., 2011). We believe that such a discrepancy is cause for additional reflection, because it suggests that the data are perhaps not as informative as one would have otherwise believed. As reviewers and editors we can request that authors acknowledge this uncertainty and be transparent about the conflicting accounts of the data.

An additional advantage of a Bayesian analysis is that one can use the data efficiently to inform follow-up studies by taking the posterior distribution from the current study as a prior distribution for further studies. This allows one to compute the so-called replication Bayes factors (Ly, Etz, Marsman, & Wagenmakers, in press; Verhagen & Wagenmakers, 2014), which quantify the additional evidence brought forth by the new data.

Note however, even when the summary statistics are "sufficient" (i.e., they capture all relevant information; e.g., Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017) on general grounds it is still beneficial to have access to the raw data. The raw data can be used to confirm that the statistical model is appropriate, the desirability of which is vividly displayed by Anscombe's quartet (e.g., Anscombe, 1973; Matejka & Fitzmaurice, 2017). Anscombe's quartet, shown here in Fig. 3, consists of four scatter plots; in each scatter plot, the summary statistics (i.e., sample size, mean, variance, and Pearson correlation) for the X and Y variables are identical, and so are the Bayes factors – nevertheless, for three of the four scatter plots the inference in terms of the strength of a linear association based on the observed Pearson's correlation efficient r is meaningless.[7]

In closing, the kind of Bayesian reanalyses outlined here provides an opportunity to expand summary statistics to statements about posterior distributions and Bayes factors.

[6] For a further discussion of the FC results, see https://web.archive.org/web/20170901132636/https://mattiheino.com/2016/11/13/legacy-of-psychology/.

[7] See also Alberto Cairo's Anscombosaurus at https://web.archive.org/web/20170901133148/http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html.

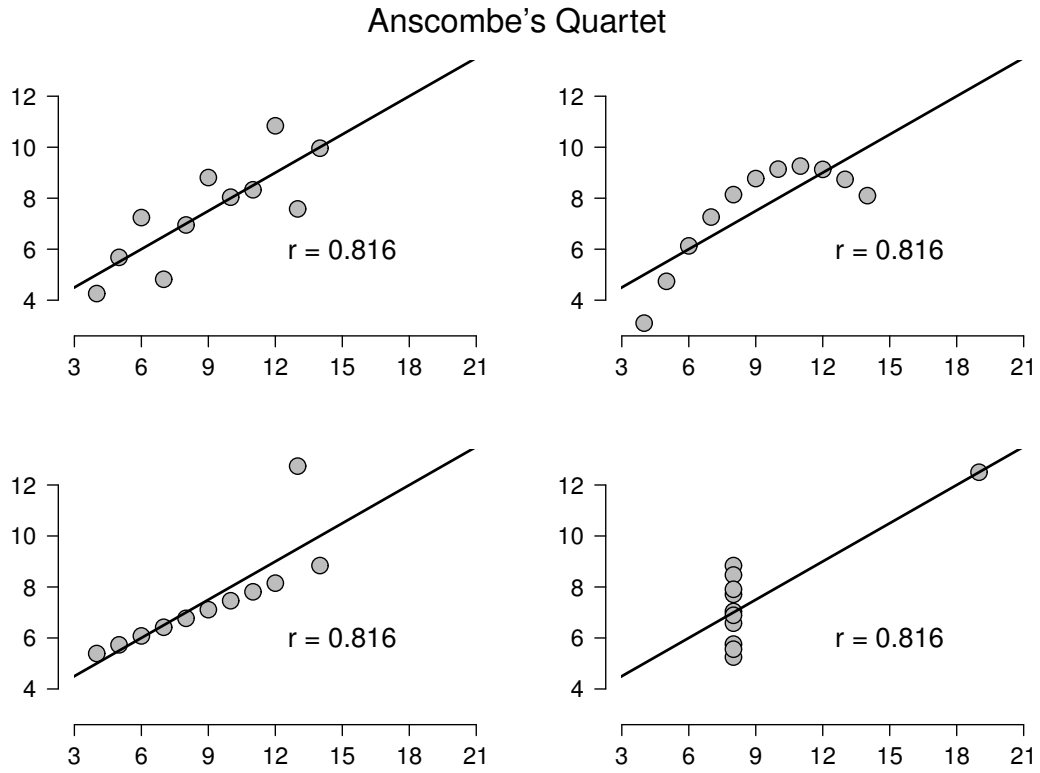

Figure 3. Anscombe quartet (Anscombe, 1973). In each panel, the summary statistics are identical: The X variable has a sample mean of $\bar{x} = 9$ and a sample variance of $s_x^2 = 11$, the Y variable has a sample mean of $\bar{y} = 7.50$ and a sample variance of $s_y^2 \approx 4.125$, and the Pearson correlation coefficient is $r = 0.816$.

Such an expansion affords (1) an additional inferential perspective that supplements the classical perspective, (2) reanalyses of published findings without requiring the raw data, and (3) a highly efficient method for editors, reviewers, readers, and reporters to gauge whether the conclusions from a different statistical paradigm contradict or confirm the classical conclusions. We hope that the Summary Stats Module will spur more nuanced assessments of statistical evidence and reporting of statistical outcome measures that are both comprehensive and inclusive.

Author Contributions

EJW generated the idea for the method. AL and AR implemented the module in `JASP`. AL and MM wrote the first draft of the manuscript, which was further developed by AL and AE, and all authors critically edited it. QFG performed and described the informed analyses. All authors approved the final submitted version of the manuscript.

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219–234.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34. doi: 10.3758/s13423-017-1262-3

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*(2), 313–329.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, *58*(2), 203–210.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337–350.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2017). Informed Bayesian *t*-tests. *arXiv preprint arXiv:1704.02479*.

JASP Team. (2018). *JASP (Version 0.9.0.1)[Computer software].* Retrieved from `https://jasp-stats.org/`

Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kelman, H. C. (1953). Attitude change as a function of response restriction. *Human Relations*, *6*(3), 185–214.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (in press). Replication Bayes factors from evidence updating. *Behavior Research Methods*. doi: 10.3758/s13428-018-1092-x

Ly, A., Marsman, M., Verhagen, A. J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, *80*, 40–55. doi: https://doi.org/10.1016/j.jmp.2017.05.006

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. doi: http://dx.doi.org/10.1016/j.jmp.2015.06.004

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55. doi: http://dx.doi.org/10.1016/j.jmp.2016.01.003

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *CHI 2017 Conference Proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems*. Retrieved from `https://www.autodeskresearch.com/publications/samestats`

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.11-1.* Comprehensive R Archive Network. Retrieved from `http://cran.r-project.org/web/packages/BayesFactor/index.html`

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing

controversy. *Psychological Methods*, *5*, 241–301.

Rouder, J. N., Morey, R. D., Verhagen, A. J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2017). Beyond the new statistics: Bayesian inference for psychology [special issue]. *In preparation for Psychonomic Bulletin & Review*.

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. doi: https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi: https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123–138). Chichester, England: John Wiley and Sons. doi: http://dx.doi.org/10.1002/9781119095910.ch8

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298.