Bayes factors for research workers

Ly, A.

Publication date
2018
Document Version
Other version
License
Other

Chapter 8

Bayesian Reanalyses from Summary Statistics: A Guide for Academic Consumers

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` (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. The procedure is illustrated using the seminal Festinger and Carlsmith (1959) study on cognitive dissonance.

Keywords: Bayes factor, data visualisation, effect size, hypothesis testing, p-value.

This chapter is submitted for publication and also available as PsyArXiv preprint: `https://osf.io/7dzmk` as: Ly, A., Raj, A., Marsman, M., Etz, A., & Wagenmakers, E.–J. (2017). Bayesian reanalyses from summary statistics: A guide for academic consumers.

8.1 Introduction

Classical null hypothesis statistical 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 favour 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 and 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., Nickerson, 2000; Rouder et al., 2016a; Wagenmakers et al., 2017b). 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, recent years have seen an increased interest in Bayesian inference (Vandekerckhove et al., 2017; Wagenmakers et al., 2016b), a paradigm in which prior uncertainty about parameters and models is updated by means of observed data to yield posterior uncertainty. Specifically, the posterior distribution quantifies the information about the effect size under the alternative hypothesis, whereas the Bayes factor quantifies the predictive adequacy of the null hypothesis as compared to the predictive adequacy of an exactly-specified alternative hypothesis (e.g., Etz and Wagenmakers, 2017; Jeffreys, 1961; Kass and Raftery, 1995; Myung and Pitt, 1997).

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 will weaken 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, 2017). 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.

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

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 analysis presented here is much more efficient – reviewers, readers, and reporters can obtain a comprehensive Bayesian assessment almost instantaneously. We believe that the richness of the Bayesian report contrasts favourably with a report of just the summary statistics themselves. We illustrate this claim using a seminal study published more than half a century ago.

8.2 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). Early experiments on cognitive dissonance (e.g., Kelman, 1953) induced participants to make a statement contrary to their personal opinion for the chance to gain a reward. It was hypothesised 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 behavioural 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 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 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 is reported in the original paper but this can be easily computed from the t-value and degrees of freedom, giving $d = 0.720$.

[2] Cited over 3,300 times according to Google Scholar, May 19, 2017.

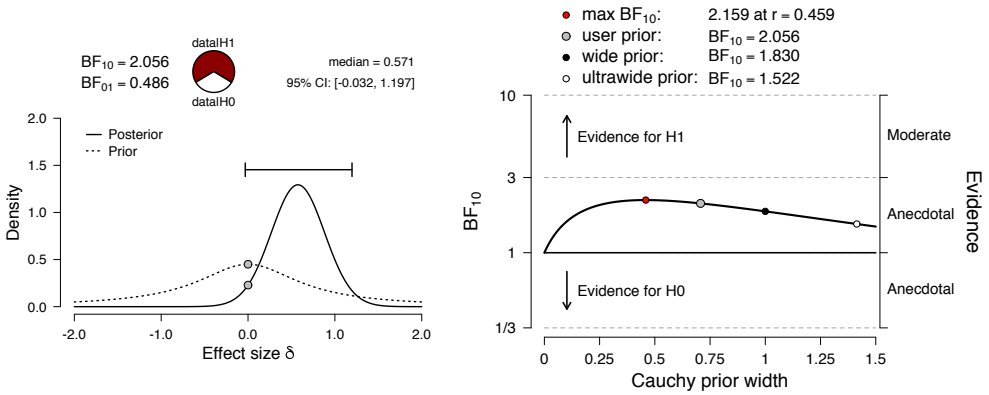

Figure 8.1: A comprehensive Bayesian reanalysis of the seminal study by Festinger and Carlsmith (1959), obtained by entering $t = 2.22$ and $N_1 = N_2 = 20$ into the JASP Summary Stats module. See text for details.

8.3 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. 8.1.

In the left panel, the dotted line represents the prior distribution for effect size under $\mathcal{H}_1$: a zero-centred *Cauchy* distribution (i.e., a t-distribution with one degree of freedom; Jeffreys, 1948; Ly et al., 2016a, 2016b), here with interquartile range set to a default value of $r = 0.707$ (e.g., Morey and Rouder, 2015; for a larger family of informed prior distributions see Gronau et al., 2017a). Thus, under $\mathcal{H}_1$ –that is, assuming 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, 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% central 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 δ falls in the interval from a to b (e.g., Wagenmakers et al., 2016b; Wagenmakers et al., 2017a). 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

[3]The Summary Stats module is activated via the options menu located in the top right corner of the JASP window.

[4]Note that the prior distribution has shrunk the sample value of $d = 0.720$ toward zero.

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 effect size δ, given that $\mathcal{H}_1$ holds and the effect exists. The latter point is worth emphasising 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), 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 assumed by $\mathcal{H}_1$, 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. 8.1 reveals that it equals 2.056 – that is, the observed FC data are only about twice as likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$. Arch-Bayesian Harold Jeffreys deemed this level of evidence "not worth more than a bare mention" (Jeffreys, 1961, p. 432). The proportion wheel on top visualises 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. 8.1 shows the result: the Bayes factor as a function of the interquartile range r of the Cauchy prior distribution. The values range from $r = 0$ (when $\mathcal{H}_1$ reduces to $\mathcal{H}_0$ and the Bayes factor is 1 regardless of the data) to $r = 1.5$. Across this entire range, the Bayes factor never exceeds 3; in fact, the maximum Bayes factor in favour of $\mathcal{H}_1$ equals 2.159, obtained when the width r is set to 0.459.

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 Bayesian perspective.[6] Even though the evidence against $\mathcal{H}_0$ is relatively inconclusive, the posterior distribution can nevertheless be used as a prior in further studies, and allows one to compute the so-called replication Bayes factor (Ly et al., 2017b; Verhagen and Wagenmakers, 2014).

In sum, the Bayesian reanalyses shown in Fig. 8.1 are easily obtained in `JASP` and paint an inferential picture more complete than the one provided by the statement "$t(38) = 2.22$, $p = .032$".

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

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

8.4 Concluding comments

The Summary Stats module in JASP unlocks a comprehensive Bayesian experience from a few commonly-reported summary statistics. Here we illustrated the module for the case of an independent-samples t-test, but the Summary Stats module can also be used for inference 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.

Two remarks are in order. First, even when the summary statistics are "sufficient" (i.e., they capture all relevant information) 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 and Fitzmaurice, 2017).[7]

Second, the Bayesian analyses discussed above are "objective" or "uninformative" in the sense that under $\mathcal{H}_1$, the prior distributions for effect size are centred around zero, the value specified by $\mathcal{H}_0$. However, the Bayesian framework can be extended to include *informed* prior distributions – these distributions incorporate context-specific expectations and need not be centred around zero (Gronau et al., 2017a). We plan to add the extensions to informed priors to JASP in the near future. Just like the reanalysis with objective priors, the reanalysis with informed priors is a function solely of the summary statistics.

In closing, the Bayesian reanalyses outlined here provide an opportunity to expand summary statistics to statements about posterior distributions and Bayes factors. This expansion affords (1) an additional inferential perspective that supplements the classical perspective; (2) a reanalysis 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 this reanalysis will spur a more graded assessment of statistical evidence and a reporting of statistical outcome measures that is both comprehensive and inclusive.

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