Bayes factors for research workers

Ly, A.

Chapter 7

Four Requirements for an Acceptable Research Programme

Abstract

In a recent article for *Basic and Applied Social Psychology*, Witte and Zenker (2016) proposed a research strategy that rests on the sequential evaluation of a point-alternative hypothesis. At first a large study is used to determine a "specific theoretical effect size" and then, in a series of follow-up studies, this estimated effect size is contrasted against an effect size of zero. The authors deem this strategy "free of various deficits that beset dominant strategies (e.g., meta-analysis, Bayes factor analysis)" and argue that its broad adoption constitutes "one way in which the confidence crisis may be overcome".

We disagree with their research strategy as it does not go far enough. One should avoid hindsight bias and acknowledge uncertainty that comes with scientific learning. The four requirements given here provide the context in which Bayes factors can help empirical scientists learn from data.

Keywords: Crisis of confidence, exploratory versus confirmatory research, scientific learning.

We agree with Witte and Zenker (2016) that it can be useful to test an alternative hypothesis that is constructed, in part or in whole, from earlier data (e.g., Verhagen and Wagenmakers, 2014; Wagenmakers et al., 2016c). We also agree that it can be informative to take into account a sequence of studies as it unfolds over time (e.g., Scheibehenne et al., 2016). In this comment, however, we focus mainly on areas of disagreement, which centre on what we believe to be mistakes and omissions. First we address the mistakes and discuss how, in our opinion, Witte

This chapter is published as Marsman, M., Ly, A., & Wagenmakers, E.–J. (2016). Four requirements for an acceptable research programme. *Basic and Applied Social Psychology, 38*(6), 308–312. doi: http://dx.doi.org/10.1080/01973533.2016.1221349.

and Zenker (2016) fell prey to two common fallacies: the power fallacy and the fallacy of the transposed conditional. Even for experienced scholars, these fallacies may be difficult to recognise. Second, we address the omissions and discuss four requirements for an acceptable research programme.

7.1 The power fallacy

On repeated occasions, Witte and Zenker (2016) lament the lack of statistical power while at the same time boasting about the strength of statistical evidence. This confused interpretation of the data can be overcome by recognising that power and evidence are inherently different concepts. Before we start, let's take for granted that the desired test is between $\mathcal{H}_0 : \delta = 0$ versus a point-alternative $\mathcal{H}_1 : \delta = 0.30$.

Now power is a pre-data concept, a metric constructed by averaging across all possible data sets that could be obtained in the envisioned experiment. A priori and *on average* –with respect to all possible data sets– experiments designed with low power are unlikely to yield a significant outcome given that $\mathcal{H}_1$ is true. In contrast, evidence is a post-data concept. In this specific scenario it is given by the likelihood ratio, the relative probability of the observed data under the competing hypotheses. The likelihood ratio considers only the data that have in fact been obtained.

As discussed elsewhere in detail, after the data have been observed, data that could have been observed –but were not– are evidentially irrelevant (e.g., Berger and Wolpert, 1988; Bayarri et al., 2016; Wagenmakers et al., 2015a; Wagenmakers et al., 2017c). Basically, our pre-data state of knowledge has changed by the observation of the data, and after the data have arrived our post-data state of knowledge is all that ought to matter.

When the pre-data concept of power is erroneously used for post-data purposes –such as inference and the quantification of evidence–, this entails a deliberate loss of important information, namely the actual outcomes of the experiment.

7.2 The fallacy of the transposed conditional

Witte and Zenker (2016) correctly point out that the Bayes factor is the probability of the data under $\mathcal{H}_0$ versus $\mathcal{H}_1$ (Wagenmakers et al., 2016b). They also acknowledge that the Bayes factor and the likelihood ratio are "quantitatively" equivalent whenever the hypotheses are both simple (i.e., consisting of a single specified point value for effect size). However, Witte and Zenker (2016) argue that by simply changing the nomenclature[1] –from Bayes factors to likelihood ratios– allows them to interpret the likelihood ratio as the relative plausibility of the hypotheses. So even though what is calculated is the relative probability of the data given the hypotheses, the result is interpreted as the relative probability of the hypotheses given the data. By doing so Witte and Zenker (2016) commit the fallacy of the transposed conditional.

[1] "What's in a name? That which we call a rose by any other name would smell as sweet" – Juliet, Act 2 Scene 2

Unfortunately, in statistical inference there is no such thing as a free lunch (Rouder et al., 2016a). Any time one wishes to assign probabilities to parameters or models, one is automatically committed to the Bayesian framework (Ly et al., 2016a, 2016b). Specifically, the only way to obtain a posterior probability is by using the data to update a prior probability. Bayes factors quantify the extent to which the data change the prior model odds to posterior model odds, and as such they can be considered the relative evidence that the data provide for the models under consideration. The Bayes factor is therefore only one ingredient for inference. The other ingredient is the prior model odds. One is licensed to interpreted Bayes factors (or likelihood ratios, for simple models) as posterior odds, but only when the prior odds equals 1, and *not* when the prior odds is ignored.

To appreciate the importance of the prior odds, consider the competing models $\mathcal{H}_1$: "people have extra-sensory perception (ESP)" versus $\mathcal{H}_0$: "people do not have ESP". Few researchers would seriously entertain equal prior odds in this case. Moreover, suppose the likelihood ratio for an ESP experiment yielded a factor of 30 in favour of ESP; do we conclude from this that the ESP hypothesis is 30 times more likely than the null hypothesis? Of course we do not, and if the authors methodology were to sanction this inference (which it does not), then this would be a compelling argument against their methodology instead of a compelling argument for ESP. Extraordinary claims require extraordinary evidence, and in order to assess the posterior plausibility of ESP one needs to combine the evidence from the data (i.e., the Bayes factor) with the prior plausibility of the ESP phenomenon (Wagenmakers et al., 2015b).

7.3 Requirements of a research programme

A research programme that can cure the current "crisis of confidence" (Pashler and Wagenmakers, 2012) needs to be more ambitious than the approach proposed by Witte and Zenker (2016). Below we outline four key requirements and point to the relevant literature.

7.3.1 I. Preregistration

Philosophers, psychologists, physicians, and physicists have long argued that empirical research needs to respect the distinction between work that is exploratory or hypothesis-generating and work that is confirmatory or hypothesis-testing, and that this needs to be done by preregistering the analysis plan in all of its details (e.g., Barber, 1976; Chambers, 2013; Feynman, 1998; Goldacre, 2009; Peirce, 1878,8; Wagenmakers et al., 2012).

These theoretical arguments have garnered empirical support in the sense that preregistered replications rarely support the original effects (e.g., Nosek and Lakens, 2014; Open Science Collaboration, 2012). Without preregistration, researchers can easily and unwittingly fall prey to hindsight bias and confirmation bias. In our opinion, any research programme that does not include preregistration is seriously incomplete.

7.3.2 II. Transparency

In reproducible research, transparency is essential. Indeed, one can argue that preregistration falls under the general heading of transparency as well. Here we use transparency to refer to open materials, open data, and open analysis code. Recent initiatives such as TOP (Transparency and Openness Promotion, Nosek et al., 2015), PRO (The Peer Reviewers' Openness Initiative, Morey et al., 2016), and the Center for Open Science badges for good academic behaviour (Kidwell et al., 2016) aim and change the dominant culture so that openness becomes the norm, not the exception.

In our own work, we have developed the open-source statistical software program `JASP` (`jasp-stats.org`; JASP Team, 2017). In `JASP`, users can save data, analysis input, analysis output, and analysis annotations in a single .jasp file.[2] When this file is uploaded to the Open Science Framework, the OSF `JASP` previewer allows anybody with an online browser to inspect the annotated output, even without having `JASP` installed.

7.3.3 III. Comprehensive knowledge updating

A mature research programme allows knowledge to be updated as new data come in (Scheibehenne et al., 2016). This requirement is fulfilled by Witte and Zenker (2016), but only in part: what is updated is the likelihood ratio, but not the value of the parameter. In other words, based on the initial study, Witte and Zenker (2016) committed themselves 100% to the single point estimate $\delta = 0.30$. This violates what Lindley termed "Cromwell's rule". Cromwell famously told the Church of Scotland "I beseech you, in the bowels of Christ, think it possible you may be mistaken". Cromwell's rule states that one should not categorically rule out anything, for this makes it impossible to learn. As explained by Lindley (1985), "So leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved."

Occasionally there are reasons to violate Cromwell's rule. For instance, one may wish to evaluate the relative adequacy of the predictions from a theoretically meaningful hypothesis – perhaps a general law or invariance (Rouder et al., 2009), or perhaps a physical law involving gravity or the speed of light. In the current example, however, the point estimate of 0.30 is devoid of theoretical content; the effect size could differ from one context to the next, or it could be lower or higher. The original data set suggested $\delta = 0.30$, but what if a second, much larger data set[3] had suggested $\delta = 0.10$? This value is still consistent with the general theory of there being an effect, only it is a little smaller than suggested in the original study. The likelihood ratio would have favoured $\mathcal{H}_0$, but at the same time it would be obvious that $\mathcal{H}_0$ is false. This is the equivalent of the Lindley's astronaut scenario.

[2] The analysis output may also be saved separately.

[3] For concreteness and to avoid ambiguity, let's say one thousand times as large.

The correct way to update knowledge is to update both the plausibility of competing models and the plausibility of the parameters within those models.[4] This implies that we also need priors on the parameters within the models. The introduction of these priors have led to much debate in the statistical community at first, as they were perceived as highly subjective. However, it has since been mathematically proved that the influence of the prior on the posterior washes out easily with enough data (e.g., Bickel and Kleijn, 2012; Kleijn and van der Vaart, 2012; van der Vaart, 1998) for the regular models typically used in the psychological sciences. As such, rather than using a point estimate of the parameter from a first data set as a point alternative hypothesis, we propose to use the posterior of the effect size instead. By using the posterior as a prior in the next study, we incorporate all the relevant information from the first data set for inference in a next experiment. Hence, subjectivity simply refers to the incorporation of previously collected data rather than an opinion. This method of extracting information from one study to another is further explored in Ly et al. (2017b), Verhagen and Wagenmakers (2014), and Wagenmakers et al. (2016c), and by doing so we adhere to the laws of probability. Hence, our proposition of using Bayesian methods leads to a principled method of learning. Moreover, it automatically gives us posteriors that can be readily used to quantify the uncertainty of our inference.

7.3.4 IV. Acknowledging uncertainty

In our experience, researchers strongly desire unambiguous yes/no answers, even when these are unavailable due to the stochastic nature of the data. Paradoxically, the noisier the data, the stronger this desire seems to become.

The decision-making framework of null hypothesis significance testing (NHST) offers some certainty: if $p < .05$, we may "reject the null hypothesis". This is fulfilling, because by making a decision we have swept all of the existing uncertainty under the rug. There is no more need to debate the outcome any longer, the researcher may feel, because we were sanctioned to make a Decision to Reject the Null Hypothesis. After the Gordian knot has been cut, it is futile to argue about other possible decisions that could have been made. This way, NHST offers an illusion of certainty, and with it the protection again critique and self-doubt.

Unfortunately there are several problems with the decision-making framework of null hypothesis significance testing. The list is endless, but here we highlight the following concerns:

1. Utilities are ignored. If the purpose of statistical inference in academia is to make decisions, then one needs to specify utilities or loss functions associated to the potential outcomes (e.g., Lindley, 1985). Without utilities there can be no sensible decision making.

2. Scientists often do not make decisions. One of our favourite quotations is from Rozeboom (1960, p. 420): "The null-hypothesis significance test treats

[4] Point hypotheses are a good approximation to posterior distributions that are highly peaked, but in the case of Witte and Zenker (2016) we see no compelling reason in this case to violate Cromwell's rule and update knowledge only partially.

acceptance or rejection of a hypothesis as though these were *decisions* one makes on the basis of the experimental data—i.e., that we elect to adopt one belief, rather than another, as a result of an experimental outcome. *But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested.*"

3. The p-value from the framework of null hypothesis significance testing –upon which the Decision to Reject the Null Hypothesis is based– is "violently biased against the null hypothesis." (Edwards, 1965, p. 400; see also Berger and Delampady, 1987; Edwards et al., 1963; Johnson, 2013; Marsman and Wagenmakers, 2017; Sellke et al., 2001; Wetzels et al., 2011). For these and other reasons we sympathise with the p-value ban in *Basic and Applied Social Psychology* (Trafimow and Marks, 2015).

Instead of using ad-hoc decision rules for seeking certainty where there is none, it is better to acknowledge and quantify uncertainty. If a Bayes factor indicates that the data are 4 times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$, this does not mean that $\mathcal{H}_0$ has been refuted, or that $\mathcal{H}_1$ is true. Authors should make claims that are in accordance with the strength of evidence in the data – often, this means that the claims should be more modest. In turn, editors and reviewers should reward such modesty, not punish it.

7.4 Concluding comments

We proposed four requirements for an acceptable research programme, which we believe to be at odds with Witte and Zenker's (2016) proposal. Specifically, their proposal fails to acknowledge uncertainty and does not result in coherent knowledge updates. This is because Witte and Zenker (2016) sweep the prior model probabilities under the rug and violate the laws of probability by using an intermediate estimate as a point alternative hypothesis. Moreover, by using a point alternative hypothesis, Witte and Zenker (2016) ignore the uncertainty with which the alternative was specified.

For comprehensive knowledge updating, that is, statistical learning, we have to adhere to the laws of probability, the same way the motion of stars has to obey the laws of physics. Our advocacy for Bayesian methods in psychology is, in essence, a call to adopt a principled method of learning. This call is neither new nor controversial, as Bayesian methods have been adopted in fields such as econometrics, statistics and computer science with great success.

The reward for adopting Bayesian methods in psychology is substantial: not only do our conclusions adhere to the laws of probability, but we also obtain automatic uncertainty quantification in terms of posterior distributions. These posteriors provide a full summary of the previous data sets and can be transformed into so-called posterior predictives which give an indication of how our previous findings generalise to new experiments (Liu and Aitkin, 2008). The posterior predictive as a measure of replicability will be better in predicting future

outcomes compared to Witte and Zenker's (2016) approach as was shown in Wagenmakers et al. (2006). Their loss in performance is due to their commitment to a single point alternative hypothesis, thus, their disregard of the uncertainty in their intermediate step and, therefore, their violation of the laws of probability.

The proposed four requirements for an acceptable research programme are relatively straightforward to execute, but they imply that researchers acknowledge and counteract fundamental human biases and desires. Implementing the programme therefore requires a change in academic culture. Academic culture is difficult to change, but the past five years have demonstrated that it can be done. Driven by the combined efforts from researchers, journals, funders, and institutes (especially the Center for Open Science), there has been a dramatic and positive reorientation of academic values. The caterpillar known as psychological science has finally started its metamorphosis, and only the future will show whether the butterfly is willing to learn from the data that were actually observed.