Commentary

A Bayesian Perspective on Hypothesis Testing

A Comment on Killeen (2005)

Eric-Jan Wagenmakers[1] and Peter Grünwald[2]

[1]University of Amsterdam, Amsterdam, The Netherlands, and [2]Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands

In a recent article, Killeen (2005a) proposed an alternative to traditional null-hypothesis significance testing (NHST). This alternative test is based on the statistic p_{rep}, which is the probability of replicating an effect. We share Killeen's skepticism with respect to null-hypothesis testing, and we sympathize with the proposed conceptual shift toward issues such as replicability. One of the problems associated with NHST is that p values are prone to misinterpretation (cf. Nickerson, 2000, pp. 246–263). Another problem with NHST is that it can provide highly misleading evidence against the null hypothesis (Killeen, 2005a, p. 345): NHST can lead one to reject the null hypothesis when there is really not enough evidence to do so.

Killeen's p_{rep} statistic successfully addresses the problem of misinterpretation, and this is a major contribution (cf. Cumming, 2005; Doros & Geier, 2005; Killeen, 2005b; Macdonald, 2005). However, the p_{rep} statistic does not remedy the second, more fundamental NHST problem mentioned by Killeen. Here we perform the standard analysis to show that p_{rep} can provide misleading evidence against the null hypothesis (cf. Berger & Sellke, 1987; Edwards, Lindman, & Savage, 1963). This analysis demonstrates the discrepancy between Bayesian hypothesis testing and p_{rep}, and highlights the necessity of considering the plausibility of both the null hypothesis and the alternative hypothesis.

Consider an experiment in taste perception in which a participant has to determine which of two beverage samples contains sugar. After n trials, with s successes (i.e., correct decisions) and f failures, we wish to choose between two hypotheses: H_0 (i.e., random guessing) and H_1 (i.e., gustatory discriminability). For inference, we use the binomial model, in which the likelihood $L(\theta)$ is proportional to $\theta^s(1-\theta)^f$, where θ denotes the probability of a correct decision on any one trial.

Address correspondence to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands, e-mail: ewagenmakers@fmg.uva.nl.

A *Bayesian* hypothesis test (Jeffreys, 1961) proceeds by contrasting two quantities: the probability of the observed data D given H_0 (i.e., $\theta = \frac{1}{2}$) and the probability of the observed data D given H_1 (i.e., $\theta \neq \frac{1}{2}$). The ratio $B_{01} = p(D|H_0)/p(D|H_1)$ is the *Bayes factor*, and it quantifies the evidence that the data provide for H_0 vis-à-vis H_1. Assuming equal prior plausibility for the models, the posterior probability for H_0 is given by $B_{01}/(1 + B_{01})$.

In the taste perception experiment, $p(D|H_0) = \frac{1}{2}^n$. The quantity $p(D|H_1)$ is more difficult to calculate, because it depends on our prior beliefs about θ. Specifically, when prior knowledge of θ is given by a prior distribution $p(\theta)$, one obtains $p(D|H_1)$ by integrating $L(\theta)$ over all possible values of θ, weighted by the prior distribution $p(\theta)$: $p(D|H_1) = \int_0^1 L(\theta)p(\theta)d\theta$. We consider two classes of priors.

SCENARIO 1: THE CRYSTAL-BALL POINT PRIOR

To calculate the Bayes factor that maximally favors H_1, we let the data determine our prior beliefs about θ. That is, $p(D|H_1)$ is maximal when we choose a "crystal-ball point prior" that assigns all prior probability to the single value of θ under which the observed data have maximum probability (Edwards et al., 1963). This analysis favors H_1 quite unfairly. In a theoretical analysis of the taste perception experiment, we varied the number of observations n from 50 to 10,000 in increments of 50 and calculated for each n the number of successes required to obtain a classical p value that just reaches significance at the ubiquitous .05 level.

The lower set of functions in Figure 1a shows that the quantity $1 - p_{rep}$ is a constant .08, which indicates a .92 chance of replicating the effect. In contrast, the minimum posterior probability for H_0 is a relatively constant .128. Hence, an analysis that quite unfairly favors H_1 still cannot make H_1 more than 6.8 times as likely as H_0 when $p = .05$ and $p_{rep} = .92$. This key result shows that both classical p values and p_{rep} can overestimate the evidence against H_0. The posterior probability for H_0 increases further when we consider more realistic priors.

SCENARIO 2: REALISTIC PRIORS

Figure 1b shows three beta priors that reflect different amounts of prior substantive knowledge or personal belief about θ. Such prior belief about θ may originate in part from knowledge about the amount of sugar used, or from knowledge of previous outcomes for similar experiments. The upper set of functions in Figure 1a shows the corresponding posterior probabilities for H_0 (cf. O'Hagan & Forster, 2004, p. 5). These functions show that the posterior probability is sensitive to the prior distribution. This is because the Bayes factor penalizes vague hypotheses that could potentially explain a wide range of results. More important here is that Figure 1a shows that the evidence for H_0 increases with n.[1]

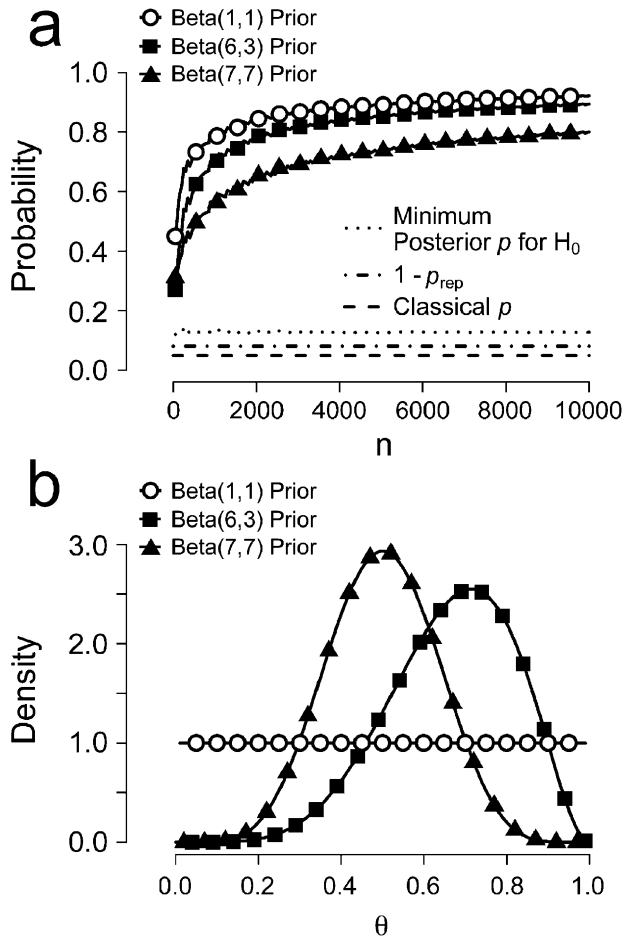

Fig. 1. Discrepancy between Bayesian hypothesis testing and null-hypothesis significance testing (NHST) or p_{rep}. The graph in (a) shows four Bayesian posterior probabilities, the classical p value, and $1 - p_{rep}$ as a function of sample size. Note that the data are constructed to be just significant at the .05 level (i.e., $p = .05$ and $p_{rep} = .92$). The upper set of functions illustrates that for any realistic prior, the posterior probability of the null hypothesis strongly depends on the number of observations. As n goes to infinity, the probability of the null hypothesis goes to 1. This conflicts dramatically with the conclusions from NHST (i.e., "$p = .05$, reject the null hypothesis") and Killeen's (2005a) p_{rep} (i.e., "$p_{rep} = .92$, the effect is highly replicable"), which are shown by the two lowest functions. For most values of n, the fact that p equals .05 (i.e., $p_{rep} = .92$) constitutes evidence in support of the null hypothesis, rather than evidence against it. The graph in (b) shows the three realistic priors that were used to compute the posterior probability of the null hypothesis in (a).

In fact, for any prior $p(\theta)$ that is continuous and strictly positive on $\theta \in [0, 1]$, the posterior probability of H_0 converges to 1 as n increases in Figure 1a (Berger & Sellke, 1987). Thus, regardless of the specific prior used, p_{rep} may indicate that the effect is highly replicable, whereas the Bayesian hypothesis test may strongly favor H_0. This discrepancy occurs because the Bayesian analysis explicitly takes the alternative hypothesis into account.

Bayesian hypothesis tests are often criticized because of their dependence on prior distributions. Yet in our example, no matter what prior is used, the Bayesian test provides substantially less evidence against H_0 than either p values or p_{rep}. One may argue about the pros and cons of priors, but one cannot argue with numbers: Over the past 5 years, at least 30% of the articles in the *Journal of the American Statistical Association* have concerned Bayesian methods.[2] It is our subjective belief that Bayesian methods will prove useful not only for statisticians, but also for psychologists.

REFERENCES

Berger, J.O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*, 112–139.

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, *16*, 1002–1004.

Doros, G., & Geier, A.B. (2005). Probability of replication revisited: Comment on "An alternative to null-hypothesis significance tests." *Psychological Science*, *16*, 1005–1006.

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

Jeffreys, H. (1961). *Theory of probability*. Oxford, England: Oxford University Press.

Killeen, P.R. (2005a). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*, 345–353.

Killeen, P.R. (2005b). Replicability, confidence, and priors. *Psychological Science*, *16*, 1009–1012.

Macdonald, R.R. (2005). Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen (2005). *Psychological Science*, *16*, 1007–1008.

Nickerson, R.S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics: Vol. 2B. Bayesian inference* (2nd ed.). London: Arnold.

(Received 9/14/05; Revision accepted 11/9/05; Final materials received 12/8/05)

[1]We thank Peter Killeen for suggesting this analysis.

[2]To arrive at this percentage, we determined the proportion of articles with "Bayes" or "Bayesian" in the title or abstract.