ARTICLE IN PRESS

Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses☆

M.J. Bayarri[a,†], Daniel J. Benjamin[b], James O. Berger[c,*], Thomas M. Sellke[d]

[a] Universitat de València, Spain
[b] University of Southern California, United States
[c] Duke University, United States
[d] Purdue University, United States

HIGHLIGHTS

- Use odds of a correct rejection of the null hypothesis to incorrect rejection.
- Pre-experimentally, these odds are the power divided by the Type I error.
- Post-experimentally, these odds are the Bayes factor.
- The Bayes factor is shown to be a fully frequentist measure of evidence.
- A useful bound on the Bayes factor is given which depends only on the *p*-value.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Odds
Bayesian
Frequentist
Bayes factors

ABSTRACT

Much of science is (rightly or wrongly) driven by hypothesis testing. Even in situations where the hypothesis testing paradigm is correct, the common practice of basing inferences solely on *p*-values has been under intense criticism for over 50 years. We propose, as an alternative, the use of the odds of a correct rejection of the null hypothesis to incorrect rejection. Both pre-experimental versions (involving the power and Type I error) and post-experimental versions (depending on the actual data) are considered. Implementations are provided that range from depending only on the *p*-value to consideration of full Bayesian analysis. A surprise is that all implementations – even the full Bayesian analysis – have complete frequentist justification. Versions of our proposal can be implemented that require only minor modifications to existing practices yet overcome some of their most severe shortcomings.

1. Introduction

In recent years, many sciences – including experimental psychology – have been embarrassed by a growing number of reports that many findings do not replicate. While a variety of factors contribute to this state of affairs, a major part of the problem is that conventional statistical methods, when applied to standard research designs in psychology and many other sciences, are too likely to reject the null hypothesis and therefore generate an unintentionally high rate of false positives. A number of alternative statistical methods have been proposed, including several in this special issue, and we are sympathetic to many of these proposals. In particular we are highly sympathetic to efforts to wean the scientific community away from an over-reliance on hypothesis testing, with utilization of often-more-relevant estimation and prediction techniques.

Our goal in this paper is more modest in scope: we propose a range of modifications – several *relatively minor* – to existing statistical practice in hypothesis testing that we believe would immediately fix some of the most severe shortcomings of current methodology. The minor modifications would not require any changes in the statistical tests that are commonly used, and would rely only on the most basic statistical concepts and tools, such as significance thresholds, *p*-values, and statistical power. With *p*-values and power calculations in hand (obtained from standard

☆ Authors are listed alphabetically. We were greatly saddened by the death of Susie Bayarri during the preparation of this paper.

* Corresponding author.
E-mail addresses: daniel.benjamin@gmail.com (D.J. Benjamin), berger@stat.duke.edu (J.O. Berger), tsellke@purdue.edu (T.M. Sellke).

† Deceased author.

http://dx.doi.org/10.1016/j.jmp.2015.12.007
0022-2496/

software in the usual way), the additional calculations we recommend can be carried out with a calculator.

In developing and justifying these simple modifications of standard methods, we also discuss additional tools that are available from Bayesian statistics. While these can provide considerable additional benefit in a number of settings, significant improvements in the testing paradigm can be made even without them.

We study the standard setting of precise hypothesis testing.[1] We can observe data $\boldsymbol{x}$ from the density $f(\boldsymbol{x} \mid \theta)$. We consider testing

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0. \tag{1}$$

Our proposed approach to hypothesis testing is based on consideration of the odds of correct rejection of H_0 to incorrect rejection. This 'rejection odds' approach has a dual frequentist/Bayesian interpretation, and it addresses four acknowledged problems with common practices of statistical testing:

1. Failure to incorporate considerations of power into the interpretation of the evidence.
2. Failure to incorporate considerations of prior probability into the design of the experiment.
3. Temptation to misinterpret p-values in ways that lead to overstating the evidence against the null hypothesis and in favor of the alternative hypothesis.
4. Having optional stopping present in the design or running of the experiment, but ignoring the stopping rule in the analysis.

There are a host of other problems involving testing, such as the fact that the size of an effect is often much more important than whether an effect exists, but here we only focus on the testing problem itself. Our proposal – developed throughout the paper and summarized in the conclusion – is that researchers should report what we call the 'pre-experimental rejection ratio' when presenting their experimental design, and researchers should report what we call the 'post-experimental rejection ratio' (or Bayes factor) when presenting their experimental results.

In Section 2, we take a pre-experimental perspective: for a given anticipated effect size and sample size, we discuss the evidentiary impact of statistical significance, and we consider the problem of choosing the significance threshold (the region of results that will lead us to reject H_0). The (pre-experimental) 'rejection ratio' R_{pre}, the ratio of statistical power to significance threshold (i.e., the ratio of the probability of rejecting under H_1 and H_0, respectively), is shown to capture the strength of evidence in the experiment for H_1 over H_0; its use addresses Problem #1 above.

How much a researcher should believe in H_1 over H_0 depends not only on the rejection ratio but also on the prior odds, the relative prior probability of H_1 to H_0. The 'pre-experimental rejection odds,' which is the overall odds in favor of H_1 implied by rejecting H_0, is the product of the rejection ratio and the prior odds. When the prior odds in favor of H_1 are low, the rejection ratio need to be greater in order for the experiment to be equally convincing. This line of reasoning, which addresses Problem #2, implies that researchers should adopt more stringent significance thresholds (and generally use larger sample sizes) when demonstrating surprising, counterintuitive effects. The logic underlying the pre-experimental odds suggests that the standard approach in many sciences (including experimental psychology) – accepting H_1 whenever H_0 is rejected at a conventional 0.05 significance threshold – can lead to especially misleading conclusions when power is low or the prior odds is low.

In Section 3, we turn to a post-experimental perspective: once the experimental analysis is completed, how strong is the evidence implied by the observed data? The analog of the pre-experimental odds is the 'post-experimental odds': the prior odds times the Bayes factor. The Bayes factor is the ratio of the likelihood of the observed data under H_1 to its likelihood under H_0; for consistency in notation (and because of a surprising frequentist interpretation that is observed for this ratio), we will often refer to the Bayes factor as the 'post-experimental rejection ratio,' R_{post}.

Common misinterpretations of the observed p-value (Problem #3) are that it somehow reflects the error probability in rejecting H_0 (see Berger, 2003; Berger, Brown, & Wolpert, 1994) or the related notion that it reflects the likelihood of the observed data under H_0. Both are very wrong. For example, it is sometimes incorrectly said that $p = 0.05$ means that there was only a 5% chance of observing the data under H_0. (The correct statement is that $p = 0.05$ means that there was only a 5% chance of observing a test statistic as extreme or more extreme as its observed value under H_0 – but this correct statement is not very useful because we want to know how strong the evidence is, given that we actually observed the value of the test statistic that we did.) Given this misinterpretation, many researchers dramatically overestimate the strength of the experimental evidence for H_1 provided by a p-value. The Bayes factor has a straightforward interpretation as the strength of the evidence in favor of H_1 relative to H_0, and thus its use can avoid the misinterpretations that arise from reliance on the p-value.

The Bayes factor approach has been resisted by many scientists because of two perceived obstacles. First, determination of Bayes factors can be difficult. Second, many are uneasy about the subjective components of Bayesian inference, and view the familiar frequentist justification of inference to be much more comforting. The first issue is addressed in Section 3.2, where we discuss the 'Bayes factor bound' $1/[-ep \log p]$ (from Sellke, Bayarri, & Berger, 2001 and Vovk, 1993). This bound is the *largest* Bayes factor in favor of H_1 that is possible (under reasonable assumptions). The Bayes factor bound can thus be interpreted as a best-case scenario for the strength of the evidence in favor of H_1 that can arise from a given p-value. Even though it favors H_1 amongst all (reasonable) Bayesian procedures, it leads to far more conservative conclusions than the usual misinterpretation of p-values; for example, a p-value of 0.05 only represents *at most* 2.5 : 1 evidence in favor of H_1. The 'post-experimental odds bound' can then be calculated as the Bayes factor bound times the prior odds.

In Section 3.3, we address the frequentist concerns about the Bayes factor. In fact, we show that in our setting, using the Bayes factor is actually a fully frequentist procedure – and, indeed, we argue that it is actually a much better frequentist procedure than that based on the p-value or on the pre-experimental rejection ratio. Our result that the Bayes factor has a frequentist justification is novel to this paper, and it is surprising because the Bayes factor depends on the prior distribution for the effect size under H_1. We point out the resolution to this apparent puzzle: the prior distribution's role is to prioritize where to maximize power, while the procedure always maintains frequentist error control for the rejection ratio that is analogous to Type I frequentist error control.

[1] By *precise hypothesis testing*, we mean that H_0 is a lower dimensional subspace of H_1, as in (1). In particular, the major problem with p-values that is highlighted in this paper is muted if the hypotheses are, say, $H_0 : \theta < 0$ versus $H_1 : \theta > 0$. As an example, suppose θ denotes the difference in mean treatment effects for cancer treatments A and B:

- Scenario 1: Treatment A = standard chemotherapy and Treatment B = standard chemotherapy + steroids. This is a scenario of precise hypothesis testing, because steroids could be essentially ineffective against cancer, so that θ could quite plausibly be essentially zero.
- Scenario 2: Treatment A = standard chemotherapy and Treatment B = a new radiation therapy. In this case there is no reason to think that θ could be zero, and it would be more appropriate to test $H_0 : \theta < 0$ versus $H_1 : \theta > 0$.

See Berger and Mortera (1999) for discussion of these issues.

Please cite this article in press as: Bayarri, M. J., et al., Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. Journal of Mathematical Psychology (2016), http://dx.doi.org/10.1016/j.jmp.2015.12.007