

*Ménage à Trois* Inference Style:  
Unifying Three Hypothesis Testing Doctrines\*

James Abdey<sup>†</sup>

Department of Statistics

London School of Economics and Political Science

**Abstract**

Three prominent ‘schools’ of hypothesis testing exist, propelled by Fisher, Jeffreys and Neyman. Fisher extolled the virtue of the  $p$ -value, whose magnitude signals the strength of evidence in the null hypothesis,  $H_0$ . In contrast, Jeffreys’ approach favours the use of objective posterior probabilities using a Bayesian framework, whilst Neyman resorted to fixed error probabilities, namely the computation of Type I and Type II errors. Here a unified framework of the competing doctrines is offered, using a new conditioning statistic which accommodates the  $p$ -value density under the alternative hypothesis for both simple and composite tests. Critical  $p$ -value curves and surfaces are derived to quickly allow conclusions to be drawn.

**JEL Classification:** Primary C12, C13, C44. Secondary C50.

**Keywords:**  $p$ -values; Significance; Simultaneous testing.

---

\*Financial support is gratefully acknowledged from the Economic and Social Research Council (ESRC) under award PTA-030-2005-00047.

<sup>†</sup>Contact details: Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE. Email: J.S.Abdey@lse.ac.uk

# 1 Introduction

In recent years considerable attention in the literature has focused on the suitability of conventional null hypothesis significance testing (NHST), with a frustrating lack of agreement. For instance in wildlife research, Anderson, Burnham, and Thompson (2000) report on the increasing number of papers criticising the NHST approach, but Thompson's data cite numerous defences as well. A common complaint concerns the misuse of NHST, rather than the procedure itself. Incorrect interpretation of the test conclusions however is hardly justification for an embargo on NHST (as suggested in Schmidt (1996)), but rather simply a matter of researcher training.

This paper seeks to extend previous attempts to provide a methodological unification of the different schools of hypothesis testing (Neyman-Pearson, Fisherian and Bayesian). Each school has its own merits, however each also suffers from limitations which are discussed. Attention focuses on the concept of conditional frequentist testing which has been developed in recent years to help provide unity. New results presented here include a revised conditioning statistic taking into account the behaviour of the  $p$ -value under  $H_1$  by considering its density,  $f_P(p|H_1)$ . As a consequence, new critical  $p$ -value curves and surfaces are constructed to provide a quick-and-easy method for researchers to employ this conditional methodology, with computation limited to obtaining a conventional  $p$ -value and determining certain parameter values which are a product of the specification of  $H_1$ .

The structure of the paper is as follows. Section 2 summarises Bayesian hypothesis testing and considers the inferential conflict between posterior probabilities and  $p$ -values. Section 3 provides a useful review of the different testing doctrines, while section 4 outlines the unifying framework of conditional frequentist testing including a new

alternative conditioning statistic. Section 5 extends this result to construct new critical  $p$ -value curves and surfaces with a discussion of how these should be interpreted. Finally section 6 concludes.

## 2 Bayesian Hypothesis Testing

Prior to the 1920s, statistical inference was foremost Bayesian, following on from the pioneering work of Bayes and Laplace. As a simple illustrative example, consider  $n$  independent Bernoulli trials used to test  $H_0 : \theta = \theta_0$  against  $H_1 : \theta \neq \theta_0$  for  $0 < \theta < 1$ . Prior probabilities for the two hypotheses,  $\Pr(H_i)$ ,  $i = 0, 1$ , are stated as well as the prior density for  $\theta$ ,  $\pi(\theta)$ .<sup>1</sup> Objective Bayesians would use default probabilities of 0.5 and a default prior density  $\pi(\theta) = 1$  over  $0 < \theta < 1$  for a typical Bernoulli problem. A subjective approach would choose probabilities and a density based on personal beliefs or real extraneous information. Once data, say  $x$ , have been collected, the posterior probabilities of the hypotheses can be computed, representing the posterior distribution. For example for the null hypothesis, from Bayes' theorem,

$$\Pr(H_0|x) = \frac{\Pr(H_0)f(x|\theta = \theta_0)}{\Pr(H_0)f(x|\theta = \theta_0) + \Pr(H_1) \int_{\{\theta \neq \theta_0\}} f(x|\theta)\pi(\theta)d\theta}, \quad (1)$$

where  $f(x|\theta)$  is the sampling distribution of  $x$ , given parameter  $\theta$ . Hence, it is necessary to be able to evaluate the integral analytically, or at least approximate it numerically by Monte Carlo methods. Equation (1) then gives a measure of the likelihood of  $H_0$  taking into account  $\pi(\theta)$ . This contrasts and conflicts with a conventional  $p$ -value, an issue returned to later.

An alternative quantity to report is the *Bayes factor* which yields a measure of the odds of  $H_0$  to  $H_1$ , given the data — essentially a weighted likelihood ratio. Formally the

---

<sup>1</sup>One can view this as a weight function permitting calculation of an average likelihood under the alternative hypothesis.

Bayes factor, say  $B_{0,1}$  of  $H_0$  to  $H_1$ , is the posterior odds ratio over the prior odds ratio,

$$B_{0,1} = \frac{\Pr(H_0|x)/\Pr(H_1|x)}{\Pr(H_0)/\Pr(H_1)} \quad (2)$$

$$= \frac{f(x|\theta = \theta_0)}{\int_{\theta \neq \theta_0} f(x|\theta)\pi(\theta)d\theta}. \quad (3)$$

The interpretation of the specification in (3) is as the likelihood of the data under  $H_0$  divided by the average likelihood under  $H_1$ , with the advantage that the Bayes factor is independent of the prior hypothesis probabilities, and so reflects the observed data only. Obviously for the objective approach with  $\Pr(H_0) = \Pr(H_1) = 0.5$ , the Bayes factor is simply the posterior odds ratio. Given (3), the posterior probability of  $H_0$  can alternatively be stated as

$$\Pr(H_0|x) = \left[ 1 + \frac{\Pr(H_1)}{\Pr(H_0)} \cdot \frac{1}{B_{0,1}} \right]^{-1}. \quad (4)$$

Berger and Delampady (1987), for example, derive the Bayes factor for  $\pi(\theta) \sim N(\theta_0, \tau^2)$  in contrast to the Cauchy  $C(\theta_0, \tau^2)$  preferred by Jeffreys (1961), where  $\tau^2$  is a hyperparameter. Berger (1985) provides useful references in defence of objective priors in response to the frequentist criticism of Bayesian techniques requiring a prior specification.<sup>2</sup> Having equal prior probabilities is intuitively acceptable as representing objectivity due to the symmetry of the prior beliefs (despite the fact that just considering Bayes factors removes the need to even consider such probabilities), however there is no clear objective choice for  $\pi(\theta)$ . In Berger and Delampady (1987) it is argued that  $\pi(\theta)$  should be symmetric about  $\theta_0$  for a parameter space spanning the entire real line, and possibly be non-increasing in  $|\theta - \theta_0|$  to avoid bias towards  $\theta \neq \theta_0$ . They note that the functional form of  $\pi(\theta)$  is largely irrelevant, however in the Gaussian versus Cauchy specification, the scale factor  $\tau$  is influential in both Bayes factor and posterior probabilities which means that  $\tau$  must be specified, and for that matter specified

---

<sup>2</sup>Berger and Berry (1988) note the disguised subjectivity within the frequentist ideology.

subjectively since there is no obvious default, objective value. Of particular note however, are the ‘automatic’ Bayesian significance tests of Jeffreys (1961) (specifying a Cauchy  $C(\theta_0, \sigma^2)$  prior) and Smith and Spiegelhalter (1980) (specifying a constant default prior) which, although not completely objective, do yield superior results vis-à-vis  $p$ -values.

Whereas non-Bayesians would be inclined to report a  $p$ -value and perhaps a confidence interval of likely values of the unknown parameter, a Bayesian approach would be to report the posterior  $\Pr(H_0|x)$  with, say, a 95% posterior credible interval for the parameter. So fundamentally, we have two competing statistics for point statistics for empirical conclusions, namely the  $p$ -value and the posterior probability,  $\Pr(H_0|x)$ .<sup>3</sup> Although both seem intuitively appealing, it is possible to encounter an inferential conflict between  $p$ -values and the conditional measures of Bayes factors and posterior probabilities for two-sided tests, such as a small (i.e. significant)  $p$ -value occurring in parallel with a large  $\Pr(H_0|x)$ . When such cases occur,  $p$ -values are very misleading resulting in an irreconcilability between  $p$ -values and posterior probabilities. In terms of quantities to report, such as reporting posterior probabilities for a range of (subjective) prior inputs, see Dickey (1973).

## 2.1 Example of inferential conflicts

Take as an example the interesting experiment investigating the presence of psychokinesis, that is the ability of the mind to influence matter. In 1987, an experiment by Schmidt, Jahn and Radin seemed to prove the existence of this phenomenon. Particles arrived at a quantum gate and the experiment was set up such that the probability of particles veering towards one of two directions was 0.5. Of the 104,900,000 independent Bernoulli trials, there were 53,263,000 successes providing allegedly strong evidence in favour of the paranormal, with the test of  $H_0 : \theta = 0.5$  yielding a  $p$ -value of 0.0003. So does this imply that the X-Files are true? Sadly, no. If the Bayesian approach outlined above is used,

---

<sup>3</sup>The focus here will be on these summary statistics rather than confidence intervals and posterior credible intervals.

then  $\Pr(H_0|x) = 0.94$ , so psychic ability is unlikely. Hence here the  $p$ -value is extremely misleading. Of course in practice we would not expect  $p$  to be exactly equal to  $\Pr(H_0|x)$ , however although  $p < \Pr(H_0|x)$ , the magnitude of the difference is particularly startling. Other examples of such a conflict can be found in Diamond and Forrester (1983). Note the focus here on two-sided tests, i.e. a simple or small interval null hypothesis being tested against  $H_1 : \Omega_\theta \setminus \Theta_0$ , for parameter space  $\Omega_\theta$ .

Clearly the (very) large sample size used would easily yield a (very) small  $p$ -value when the sample proportion deviates even slightly from 0.5 due to the standard error. Some departure from  $H_0$  is likely to occur precisely because in any experiment there is likely to be some systematic deviation from the strict  $H_0$  such as a calibration issue in the experimental design as well as the stochastic nature of the experimental particles. Consequently the  $p$ -value will be decreasing in  $n$ , the sample size. Therefore with a sample size of several million, even minor deviations from a strict  $H_0$  will be statistically significant as a result of false positives.

So the Fisherian approach which produces a  $p$ -value, i.e. the probability of the observed outcome or a more extreme one, seems to be flawed. Much better, therefore, to report the likelihoods of all the different hypotheses assessing their strengths conditional on the data, as achieved in Bayesian testing. In essence the hypotheses are all in direct competition with one another<sup>4</sup> and the posterior probabilities allow the researcher to discriminate between them. Should no hypothesis emerge the ‘winner’, i.e. we have inconclusive results, then more data should be collected. Note that in practice it is never possible to be 100% certain in accepting or rejecting a particular hypothesis — an open mind must be maintained since new observations might cause a revision in the posterior probabilities culminating in a previously preferred hypothesis becoming less likely while the less endeared hypothesis might suddenly become in vogue. Initial data may be compatible with the sampling distribution  $f(x|\theta)$ , however the true sampling

---

<sup>4</sup>In model choice problems, it is possible to have several hypotheses each representing a different model.

distribution could be of a completely different functional form, but the data  $x$  might be compatible with both distributions by pure coincidence, unlike new observations (from the true distribution) which may be incompatible with  $f(x|\theta)$ .<sup>5</sup>

A basic deficiency with Fisherian hypothesis testing is that it answers the question ‘Given  $H_0$  is true, what is the probability of these (or more extreme) data?’ i.e.  $\Pr(x|H_0)$ , however what we really want to answer is ‘Given these data, what is the probability that  $H_0$  is true?’ i.e.  $\Pr(H_0|x)$ , that is the conditioning is reversed. The important point is that in general  $\Pr(x|H_0) \neq \Pr(H_0|x)$ .<sup>6</sup> The reason for the considerable disparity between  $\Pr(H_0|x)$  and the  $p$ -value for two-sided tests stems from the conditioning set. The posterior probability takes into account only the data, while the  $p$ -value considers the probability of observing the data *or a more extreme result*. As Jeffreys (1980) commented,

‘I have always considered the arguments for the use of  $P$  absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened.’ (p. 453)

Cohen (1994) provides an entertaining review and critique of the pitfalls of Fisherian and Neyman-Pearson testing citing the ‘*mechanical dichotomous decisions around a sacred .05 criterion*’ (italics in original) and his ‘temptation to call it statistical hypothesis inference testing’, with an eye on its acronymous namesake.

Berger and Sellke (1987) investigate lower bounds on the posterior probability,  $\Pr(H_0|x)$ , for different priors for testing point null hypotheses, and include references to other papers

---

<sup>5</sup>This argument reflects the questions facing any theory, i.e. model assumptions and simplifications should be reasonable and the core theoretical implications should be reflected in the data. However this does not prove that a model is true, rather a completely different mechanism may have generated the data, but the incorrect model provides a good fit by pure coincidence, although the likelihood of this decreases with more data.

<sup>6</sup>Although Fisherian advocates may argue that the definition of a  $p$ -value as  $\Pr(x|H_0)$  is no secret and hence it is foolish to treat a  $p$ -value as measuring  $\Pr(H_0|x)$ , the fact is that most practitioners are non-specialists who confuse the distinction between  $\Pr(x|H_0)$  and  $\Pr(H_0|x)$  — see Diamond and Forrester (1983). Therefore given this conflict for two-sided tests, the reporting of  $p$ -values inevitably leads to a culture of rejecting  $H_0$  too liberally. Such Type I errors are by convention more intolerable than their Type II counterparts.

testing Bayesian point nulls. They report that  $p \ll \Pr(H_0|x)$  where equality can only be achieved provided the prior is heavily biased in favour of  $H_1$ , for example  $\Pr(H_1) = 0.85$ , where the probability mass is symmetrically spread out to most favour  $H_1$ , can achieve a posterior probability of 0.05 for a two-sided  $z$ -statistic of 1.96. Clearly such a biased prior would be unpalatable to most, if not all, however a practitioner wishing to reject the null (if a ‘significant’ result was especially sought) can easily circumvent this perceived bias by just reporting the conventional  $p$ -value, citing ‘standard practice’. Since these lower bounds all exceed the  $p$ -values regardless of prior choice, then it is not possible to dismiss this inferential conflict between  $p$ -values and conditional measures based on the subjective choice of  $\pi(\theta)$ . Edwards, Lindman, and Savage (1963) are considered the first to expose the magnitude of this irreconcilability, such that  $p$ -values are typically at least an order of magnitude less than conditional measures.

### 3 Unifying Bayesians and Frequentists

The previous section highlighted the conflict between classical  $p$ -values and conditional measures, namely Bayesian posterior probabilities and, through (4), Bayes factors. Researchers have sought to bridge the divide between the various schools of testing (Neyman-Pearson, Fisherian and Bayesian). Bayarri and Berger (2004) review achievements in developing a methodological, if not philosophical, union between the opposing camps citing the pedagogical benefits which inevitably result from consistent inference.<sup>7</sup> For a discussion concerning the adverse effects of divided methodologies, see Goodman (1999a) and Goodman (1999b). Synthesising the best of both worlds is naturally appealing.

It should be noted that as far as *estimation* is concerned, frequentist and Bayesian

---

<sup>7</sup>Robinson and Wainer (2001) present a critique of null hypothesis significance testing (NHST) to educate researchers in the art of best practice. They conclude that NHST has its merits but should be treated as an adjunct to other forms, such as Bayesian testing when a probabilistic statement concerning the hypotheses is sought.

approaches typically yield the same, or at least similar, results for common parametric problems involving continuous parameters, allowing the adoption of either frequentist or Bayesian interpretations. Yet despite frequentist estimation being effective, Bayesian tools should be implemented to assess estimator accuracy. Many frequentist methods require asymptotic approximations, and are also used in Bayesian cases (see LeCam (1986) and Schervish (1995) for further details), however unlike frequentist methodologies exact small sample solutions can be obtained for Bayesian procedures, often more easily than asymptotic methods.

### 3.1 Review of Testing Doctrines

Three distinct schools of hypothesis testing exist advocated by Neyman, Fisher and Jeffreys.<sup>8</sup> The trouble arises due to the considerable disagreement in the test results of simple, or small interval, null hypotheses reported by each method.<sup>9</sup> Efron and Gous (2001) consider the differences in the scales of evidence. For a historical review of the different approaches, see Carlson (1976), Savage (1976), Spielman (1978), Hall and Sellinger (1986), Zabell (1992) and Lehmann (1993). For completeness, a brief review of the different techniques and common criticisms of them is now provided.

#### 3.1.1 Neyman-Pearson approach

Both  $H_0$  and  $H_1$  need to be specified (in order to produce error probabilities and to assess test power). A pre-experimental significance level,  $\alpha$ , is chosen which is used to define a critical region,  $C$ , and an appropriate test statistic, say  $T$ , is used. A simple decision rule follows. Reject  $H_0$  when  $t \in C$ <sup>10</sup>, otherwise fail to reject. As appropriate, report Type I or Type II errors,  $\alpha$  and  $\beta$  respectively. This approach is justified by way of the *frequentist*

---

<sup>8</sup>Interestingly, despite the ideological clash concerning testing, the schools reach agreement on estimation and confidence procedures (in terms of the numerical values to report), disagreeing only in the correct interpretation.

<sup>9</sup>Casella and Berger (1987) and Berger and Montera (1999) consider one-sided, i.e. composite null hypotheses with the former highlighting the similarity between results.

<sup>10</sup> $t$  is the observed realisation of the random variable  $T$ .

*principle:*

In *repeated*<sup>11</sup> use of a statistical procedure, the long-run average actual error should not be greater than (and ideally should equal) the long-run average reported error.

An oft-cited criticism with the Neyman-Pearson approach is that the error probabilities are fixed *a priori*, and so fail to adequately reflect variation in test statistic values. Also  $H_1$  must be specified to enable computation of Type II errors and consequently power functions, which rely on parameters which are typically unknown. Classical tests specify  $H_1$  and choose  $n$  to achieve the desired power, but surely choice of the parameter under  $H_1$ ,  $\theta_1$ , is subjective, hence neutralising the critics of Bayesian methods who dispute the use of a prior distribution. A possible remedy is to use a prior distribution for  $\theta_1$  and consider average power with respect to this distribution. Also since the goal is to maximise power subject to the pre-experimental  $\alpha$ , there exists an asymmetric treatment of the hypotheses.

### 3.1.2 Fisher approach

Fisher's approach champions  $p$ -values as reflecting the strength of evidence against  $H_0$ . Unlike the Neyman-Pearson framework, the (subjective) specification of an alternative hypothesis is not required. The original Fisher approach advocated the replication of small studies, and so false negatives were considered costlier (to society) than false positives. The rationale for this is that a significant result would then be tested many more times, resulting in it being discarded if subsequent rejections failed to occur. A false negative on the other hand would be ignored from the start.

Criticisms include violation of the frequentist principle and the very definition of  $p$ -values, i.e. the questionable justification for providing the probability of the data 'or a more extreme value', as remarked upon above. A client is concerned with inference on

---

<sup>11</sup>An important consideration concerns the repeatability of an experiment, however this may not always be feasible, for example a nuclear war.

his/her actual data, not hypothetical data by considering what might have been observed under repeated sampling but did not.

The condition  $p < \alpha$  acts as a screen for potentially useful innovations. The original idea of  $p$ -value testing in the context of a continuing series of experiments is intuitively sensible. Originally, inference was performed as follows:  $p < 0.05$  identified an effect,  $p > 0.2$  indicated no effect or one too small to be discovered in an experiment of the current size, inbetween these cases a revision to the experiment would be proposed. In practice most studies are ‘single-shot’ studies with no replications and any  $p > 0.05$  is automatically ignored. However such one-off studies can be combined to form meta-analyses such as the Cochrane Collaboration. Also, because of the potentially high-cost consequences of rejection errors, it is unlikely that in practice high-stakes decisions would be based on a single study.

### 3.1.3 Jeffreys approach

Jeffreys favoured an alternative hypothesis which allows the Bayes factor, as per (3), to be specified. Inference can then be based on a balance-of-probabilities basis, whereby we reject  $H_0$  if  $B_{0,1} \leq 1$  and fail to reject if  $B_{0,1} > 1$ . (Recall the Bayes factor is a likelihood ratio, hence values sub-unity suggest that  $H_0$  is less likely, hence its rejection.) In addition, objective posterior error probabilities are reported. If equal prior probabilities are used, i.e.  $\Pr(H_i|x) = 0.5$ ,  $i = 0, 1$ , then the posterior probabilities are,

$$\Pr(H_0|x) = \frac{B_{0,1}}{1 + B_{0,1}} = \alpha(B_{0,1}) \quad (5)$$

$$\Pr(H_1|x) = \frac{1}{1 + B_{0,1}} = \beta(B_{0,1}). \quad (6)$$

Intuitively, a fully accurate subjective prior distribution should result in optimal inferential decision-making. However to be *fully* accurate requires all prior beliefs to be incorporated which in principle means an infinite number of assessments, i.e.

$F_\theta(\theta = k) \forall k \in \Theta \subseteq \mathbb{R}$ , for distribution function  $F_\theta$  and parameter space  $\Theta$ , need to be reflected in  $\pi(\theta)$ . Partially-elicited priors, for example  $\pi(\theta)$  reflecting particular quartiles or moments, are problematic due to the omitted prior beliefs concerning the remainder of the distribution. Therefore it is appropriate to work with a class of prior distributions,  $\Gamma$ , encompassing the residual uncertainties. Hence use of an objective prior is prudent. However, is it really possible to have a completely impartial prior distribution? Adoption of different types of prior distributions may not achieve such impartiality, for example use of conjugate priors, but yield analytical convenience and tractability.

For a collective review of the different approaches, see Berger (2003).

## 4 Conditional Frequentist Testing

Although frequentist and Bayesian methods yield similar results in terms of estimation, this is not so for testing as characterised by the conflict between  $p$ -values and conditional measures. The problem with conventional frequentist testing is a lack of suitable conditioning. Berger, Brown, and Wolpert (1994) offer a helpful unification focusing on simple hypothesis tests following in the footsteps of Kiefer (1977) and also Brownie and Kiefer (1977) who propose the conditional confidence approach.<sup>12</sup>

The goal of conditional frequentist testing (CFT) is to obtain agreement over the numerical values to report when performing hypothesis tests, if not agreement in terms of interpretation (i.e. a methodological unification rather than a philosophical one), similar to estimation and confidence procedures.

The unconditional error probabilities  $\alpha$  and  $\beta$  in the Neyman-Pearson world suffer from their inflexibility, i.e. Type I and Type II errors fail to distinguish between test statistic values on (or just inside) the critical region boundary and those values deep within it. To remedy this deficiency, Berger, Brown, and Wolpert (1994) recommend reporting the conditional error probabilities given in (5) and (6). Since  $\alpha(B_{0,1})$  and  $\beta(B_{0,1})$  are functions

---

<sup>12</sup>Alternative approaches have been suggested, such as Hwang, Casella, Robert, Wells, and Farrell (1992).

of the Bayes factor, the Bayesian influence has now been incorporated into the Neyman-Pearson framework. Birnbaum (1961) referred to these as ‘intrinsic significance levels’ providing a likelihoodist interpretation.

So the basic conditional test can be summarised as follows<sup>13</sup> for critical value  $c$ ,

- If  $B_{0,1} \leq c$ , reject  $H_0$  and report conditional error probability  $\alpha(B_{0,1})$ .
- If  $B_{0,1} > c$ , do not reject  $H_0$  and report conditional error probability  $\beta(B_{0,1})$ .

If the Bayes factor is evaluated on a balance-of-probabilities basis, then set  $c = 1$ . From a decision-theoretic perspective, consider  $H_0 : \theta \in \Theta_0$  and  $H_1 : \theta \in \Theta_1 \equiv \Omega_\theta \setminus \Theta_0$ , for parameter space  $\Omega_\theta$ . Consider the ‘0-1’ loss function,  $L$ , on action space  $\mathcal{A} = \{a_0, a_1\}$  where  $a_0 =$  do not reject  $H_0$  and  $a_1 =$  reject  $H_0$ . Then,

$$L(\theta, a_0) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, \\ 1 & \text{if } \theta \in \Theta_1, \end{cases} \quad L(\theta, a_1) = \begin{cases} 1 & \text{if } \theta \in \Theta_0, \\ 0 & \text{if } \theta \in \Theta_1. \end{cases} \quad (7)$$

The Bayes decision rule is not to reject  $H_0$ , provided

$$\Pr(\theta \in \Theta_0|x) > \Pr(\theta \in \Theta_1|x). \quad (8)$$

Berger, Brown, and Wolpert (1994) do consider a ‘no decision’ region for inconclusive values of  $B_{0,1}$ , i.e.  $\epsilon < B_{0,1} < \nu$  for arbitrary constants  $\epsilon$  and  $\nu$ , typically such that  $\epsilon < 1 < \nu$ . However for simplicity it is easier to partition  $B_{0,1} \in \mathbb{R}^+$  into solely ‘reject’ and ‘not reject’ sets and report a large conditional error probability if  $\epsilon < B_{0,1} < \nu$ . Readers can then readily interpret the conclusiveness of the test result based on this information themselves.

Up to this point, CFT unifies and fully satisfies frequentist, likelihoodist and Bayesian principles. It remains to explicitly incorporate  $p$ -values in order to offer a plausible

---

<sup>13</sup>Berger, Brown, and Wolpert (1994), report that CFT can also be used for sequential testing noting that the Bayes factor is not affected by the chosen stopping rule.

methodological unification of hypothesis testing.

#### 4.1 Conditioning

Reid (1995) and Bjørnstad (1996) discuss conditioning, although the following basic example from Berger and Wolpert (1988) nicely highlights the benefits of conditional frequentism. Consider the observations  $X_i$ ,  $i = 1, 2$ , such that

$$X_i = \begin{cases} \theta + 1 & \text{with probability } \frac{1}{2}, \\ \theta - 1 & \text{with probability } \frac{1}{2}. \end{cases} \quad (9)$$

Define a confidence set,  $C(X_1, X_2)$ , for the unknown parameter  $\theta$  as

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 2 & \text{if } X_1 = X_2, \end{cases} \quad (10)$$

which yields unconditional frequentist coverage of 0.75. However this can be considerably improved if we condition on the observed data. The sample mean provides the precise value of  $\theta$  when  $x_1 \neq x_2$  with probability 1, yet if  $x_1 = x_2$ , then all we know is that this observed value is  $\theta + 1$  or  $\theta - 1$ . If we define a *conditioning statistic*,  $S = |X_1 - X_2|$ , then  $S$  can only take two values, i.e. 0 or 2. Hence conditional on  $S$ ,

$$\Pr_{\theta}(\theta \in C(X_1, X_2) | S = 0) = \frac{1}{2} \quad (11)$$

$$\Pr_{\theta}(\theta \in C(X_1, X_2) | S = 2) = 1. \quad (12)$$

Such conditioning still satisfies frequentist as well as Bayesian ideology through the conditional error probabilities (5) and (6) which can also be viewed as

$$\alpha(B_{0,1}) = \alpha(s) = \Pr(\text{Type I error} | S(X) = s) = \Pr_0(\text{Reject } H_0 | S(X) = s)$$

$$\beta(B_{0,1}) = \beta(s) = \Pr(\text{Type II error} | S(X) = s) = \Pr_1(\text{Not reject } H_0 | S(X) = s).$$

All that is required now is to introduce  $p$ -values into the methodology. Wolpert (1995) and Sellke, Bayarri, and Berger (2001) consider the conditioning statistic for simple hypotheses

$$S = \max\{p_0, p_1\}, \quad (13)$$

where  $p_i$  is the  $p$ -value when testing hypothesis  $H_i$  against  $H_{i^c}$ ,  $i = 0, 1$ , where  $c$  denotes the complement. It follows that the decision rule should be

$$\begin{aligned} \text{If } p_0 \leq p_1 & : \text{ Reject } H_0, \text{ report } \alpha(s) \\ \text{If } p_0 > p_1 & : \text{ Do not reject } H_0, \text{ report } \beta(s). \end{aligned}$$

Of course, any strictly increasing function  $\psi(p_i)$  would yield the same decision, hence the importance of the use of  $p$ -values in (13) is less than their interpretation as a measure of evidence in support of a hypothesis.

## 4.2 Alternative Conditioning Statistic, $S$

However the conditioning statistic in (13) requires two separate hypothesis tests:  $H_0$  v.  $H_1$  and  $H_1$  v.  $H_0$  in order to obtain  $p_0$  and  $p_1$  respectively. A new alternative to this approach presented here is to make use of the second-order  $p$ -value,  $p'$ , as detailed in the appendix. The advantage of doing this is that  $p'$  can be computed directly from  $p_0 = p$  as per (41) for the general case, which is easily applicable to specific test statistic distributions. Hence the proposed variant of (13) is

$$S = \max\{p, p'\}. \quad (14)$$

where  $p$  is the conventional  $p$ -value obtained from testing  $H_0$  against  $H_1$ , and  $p'$  is the corresponding second-order  $p$ -value. The  $p$ -value density under  $H_1$ ,  $f_P(p|H_1)$  is

$$f_P(p|H_1) = \frac{\partial}{\partial p} F_P(p|H_1) = \frac{g_{X_n}(F_{X_n}^{-1}(1-p))}{f_{X_n}(F_{X_n}^{-1}(1-p))}, \quad (15)$$

and is therefore readily computable for various test statistic distributions, so there should be no additional computational burden of obtaining  $p'$  as opposed to  $p_1$ . Indeed, since most common hypothesis tests involve Gaussian and  $t$ -distributed test statistics,  $f_P(p|H_1)$  is already known — see Hung, O'Neill, Bauer, and Köhne (1997).

### 4.3 Use of Conditional Error Probabilities in $S$

The conditional error probabilities,  $\alpha(s) = \alpha(B_{0,1})$  and  $\beta(s) = \beta(B_{0,1})$ , sum to one as evident from (5) and (6). Therefore it could be said that a conditioning statistic  $S$  with  $p$ -value arguments is redundant, as decision making could also be based on the conditional error probabilities instead which require the computation of the Bayes factor.

An intuitive decision rule would be to conclude in favour of the hypothesis which minimises the reported conditional error probability, i.e.  $\alpha(B_{0,1})$  if  $H_0$  is rejected, or  $\beta(B_{0,1})$  otherwise. This leads to the following conditioning statistic,

$$S = \min\{\alpha(B_{0,1}), \beta(B_{0,1})\}. \quad (16)$$

So the corresponding decision rule would be,

$$\text{If } \alpha(B_{0,1}) < \beta(B_{0,1}) \quad : \quad \text{Reject } H_0, \text{ report } \alpha(B_{0,1}) \quad (17)$$

$$\text{If } \alpha(B_{0,1}) \geq \beta(B_{0,1}) \quad : \quad \text{Do not reject } H_0, \text{ report } \beta(B_{0,1}). \quad (18)$$

## 5 Critical $p$ -value curves and surfaces

Although the concept of conditional frequentist testing is appealing, for a particular methodology to be widely employed in practice it is necessary to have a simple implementation along qualitative lines, i.e. a simple-to-understand reject or not reject rule. At this point, the reader is encouraged to consult the appendix for background information regarding  $p$ -value distributions and notation used below.

This section develops a new concept of critical  $p$ -value curves and surfaces which can be constructed for simple and composite hypotheses respectively. The idea is to find the  $p$ -value which yields equality between the rejection regions under  $f_P(p|H_0)$  and  $f_P(p|H_1)$ ,  $p$  and  $p'$  respectively, for a particular effect size  $\delta$ . Hence if  $p = p'$ , the researcher should conclude that each hypothesis is equally likely to be true, perhaps resulting in a randomisation or further testing. However, should  $p \neq p'$  we should invoke the conditioning statistic in (14), but for ease the appropriate conclusion can be quickly established from the curve/surface. The following examples illustrate.

### 5.1 Simple hypotheses with standard Gaussian-distributed test statistics

Hung, O'Neill, Bauer, and Köhne (1997) show that for Gaussian distributed test statistics testing simple hypotheses,

$$f_P(p|H_1) = g_\delta(p) = \frac{\phi(Z_p - \sqrt{n}\delta)}{\phi(Z_p)}, \quad 0 < p < 1, \quad (19)$$

where  $\delta = \mu/\sigma$  denotes the effect size,  $n$  the sample size,  $\phi$  is the standard Gaussian density and  $Z_p$  its  $(1 - p)^{th}$  percentile.

For  $\sqrt{n}\delta > 0$ , Figure 1 illustrates the typical shape of  $f_P(p|H_1)$ . Under  $H_1$ , the concentration of the probability density near zero provides the rationale for sufficiently small  $p$ -values to warrant rejection of  $H_0$  (lower 'tail' of  $f_P(p|H_0)$ ) and sufficiently large  $p$ -values to warrant rejection of  $H_1$  (upper tail of  $f_P(p|H_1)$ ). To obtain the associated

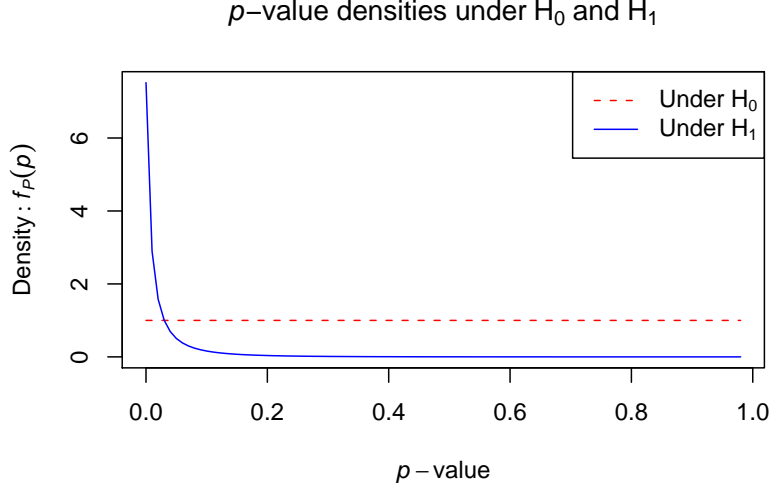


Figure 1:  $P$ -value densities under both  $H_0 : \mu = 0$  (uniform distribution) and  $H_1 : \mu = 2$  (ratio of two Gaussian densities as per (19)) such that  $\sqrt{n}\delta = 3.5$ .

critical  $p$ -value curve, we solve<sup>14</sup>

$$p = p' \tag{22}$$

$$= \Phi(Z_p - \sqrt{n}\delta), \tag{23}$$

which rearranges to

$$\sqrt{n}\delta = Z_p - \Phi^{-1}(p). \tag{24}$$

For  $\sqrt{n}\delta < 0$ ,  $f_P(p|H_1)$  is left-skewed, as per Figure 2.<sup>15</sup> Consequently sufficiently large

<sup>14</sup>Applying the result of Hung, O'Neill, Bauer, and Köhne (1997), and given in (19), the distribution function of the  $p$ -value density under  $H_1$  is

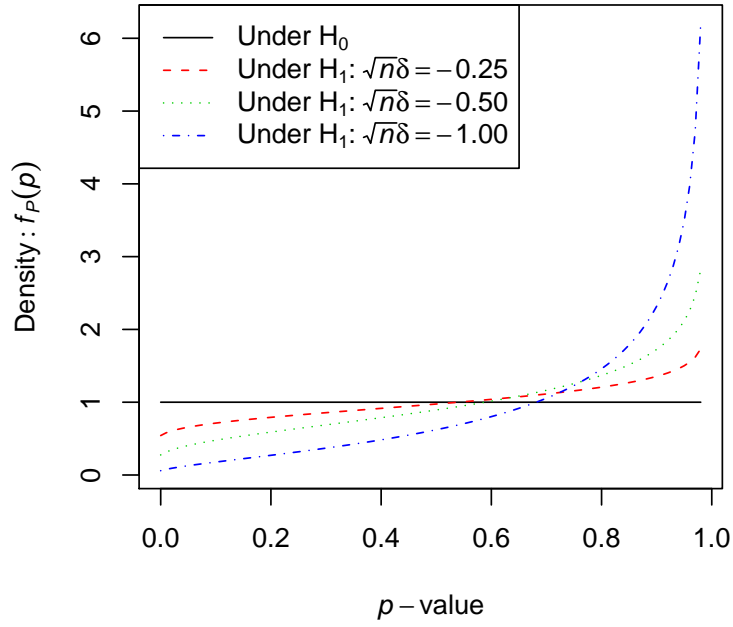
$$F_P(p|H_1) = \int_0^p \frac{\phi(Z_x - \sqrt{n}\delta)}{\phi(Z_x)} dx = 1 - \Phi(Z_p - \sqrt{n}\delta), \tag{20}$$

where recall  $Z_p$  is the  $(1-p)^{th}$  percentile of the standard Gaussian distribution, and  $\delta = k/\sigma$ . Given the definition of the  $p'$ -value in (41) for generic test statistic distributions, for the upper-tailed  $z$  test,

$$\begin{aligned} P' = \Pr(P > p|H_1) &= 1 - F_P(p|H_1) \\ &= \Phi(Z_p - \sqrt{n}\delta). \end{aligned} \tag{21}$$

<sup>15</sup>Note Figure 2 depicts  $t$ -distributed test statistics, but the standard Gaussian case is just the limiting distribution in terms of degrees of freedom.

### (i) Density functions



### (ii) Distribution functions

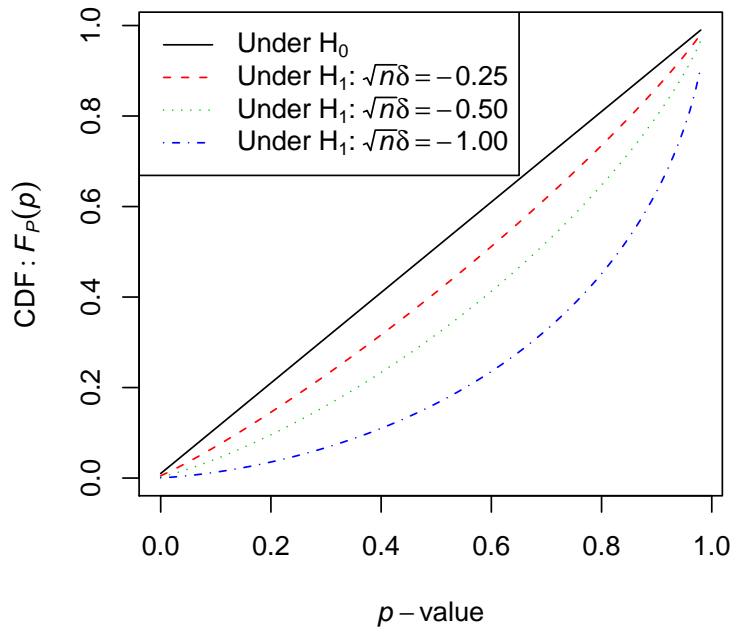


Figure 2:  $P$ -value density and distribution functions for simple forms of  $H_1$  where the effect size function  $\sqrt{n}\delta$  takes negative values.

$p$ -values warrant rejection of  $H_0$  (upper ‘tail’ test of  $H_0$ ), while sufficiently small  $p$ -values warrant rejection of  $H_1$  (lower tail test of  $f_P(p|H_1)$ ). This translates into<sup>16</sup>

$$1 - p = 1 - \Phi(Z_p - \sqrt{n}\delta) \quad (26)$$

which still rearranges to (24). Figure 3 plots the corresponding critical  $p$ -value curve. So all a researcher needs to do is obtain the conventional  $p$ -value from the test statistic in the usual way and determine  $\sqrt{n}\delta$  using  $H_1$ . Using this information all that is required is to determine where the observed co-ordinates  $(p, \sqrt{n}\delta)$  fall in relation to the critical  $p$ -value curve. If the point is on the curve itself, then  $p = p'$  and so the hypotheses are equally plausible, however if the point departs from the curve, then it is appropriate to reject or not reject  $H_0$  as indicated. Note the  $x$ -axis represents the divide between the decision rule. Formally, we have

$$\text{For } \sqrt{n}\delta > 0 : \begin{cases} \text{Reject } H_0, \text{ report } \alpha(s) & \text{if } \sqrt{n}\delta < Z_p - \Phi^{-1}(p) \\ \text{Do not reject } H_0, \text{ report } \beta(s) & \text{if } \sqrt{n}\delta > Z_p - \Phi^{-1}(p). \end{cases} \quad (27)$$

$$\text{For } \sqrt{n}\delta < 0 : \begin{cases} \text{Reject } H_0, \text{ report } \alpha(s) & \text{if } \sqrt{n}\delta > Z_p - \Phi^{-1}(p) \\ \text{Do not reject } H_0, \text{ report } \beta(s) & \text{if } \sqrt{n}\delta < Z_p - \Phi^{-1}(p). \end{cases} \quad (28)$$

Of course, having a graphical depiction of the critical  $p$ -value curve removes the need for any formal computation along the lines of (27) or (28), making implementation of the methodology fast and simplistic.<sup>17</sup>

---

<sup>16</sup>A lower-tail test is performed when testing  $H_1$ , since sufficiently small  $p$ -values warrant rejection of  $H_1$ . Therefore it is necessary to restate the  $p'$ -value for negative effect sizes as the complement of the  $p'$ -value defined in (21). Hence,

$$(P')^c = 1 - P' = \Pr(P \leq p|H_1) = F_P(p|H_1) = 1 - \Phi(Z_p - \sqrt{n}\delta). \quad (25)$$

<sup>17</sup>Obviously the conditional error probabilities,  $\alpha(s)$  and  $\beta(s)$  will need to be calculated, however the critical  $p$ -value curve itself is indicative of the level of significance of a particular  $p$ -value.

Critical  $p$ -value curve for Gaussian-distributed test statistics with simple hypotheses

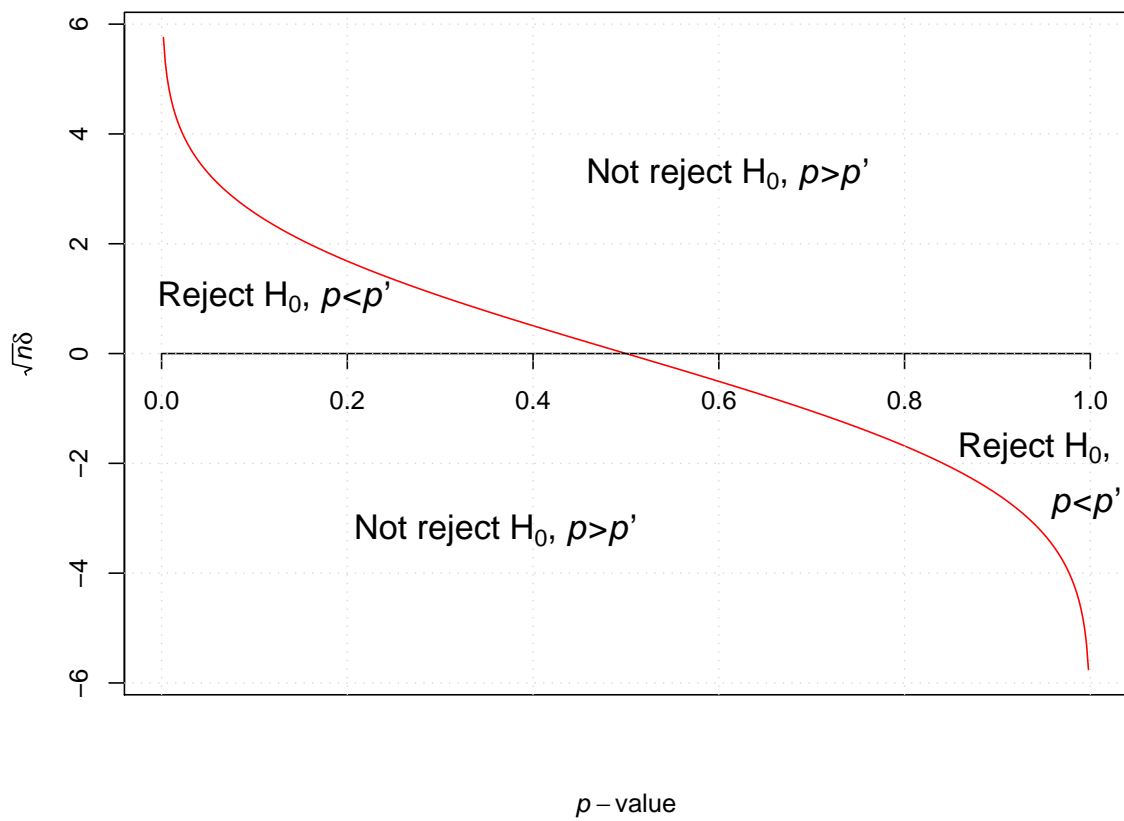


Figure 3: Critical  $p$ -value curve for standard Gaussian-distributed test statistics with simple forms for  $H_0$  and  $H_1$ .

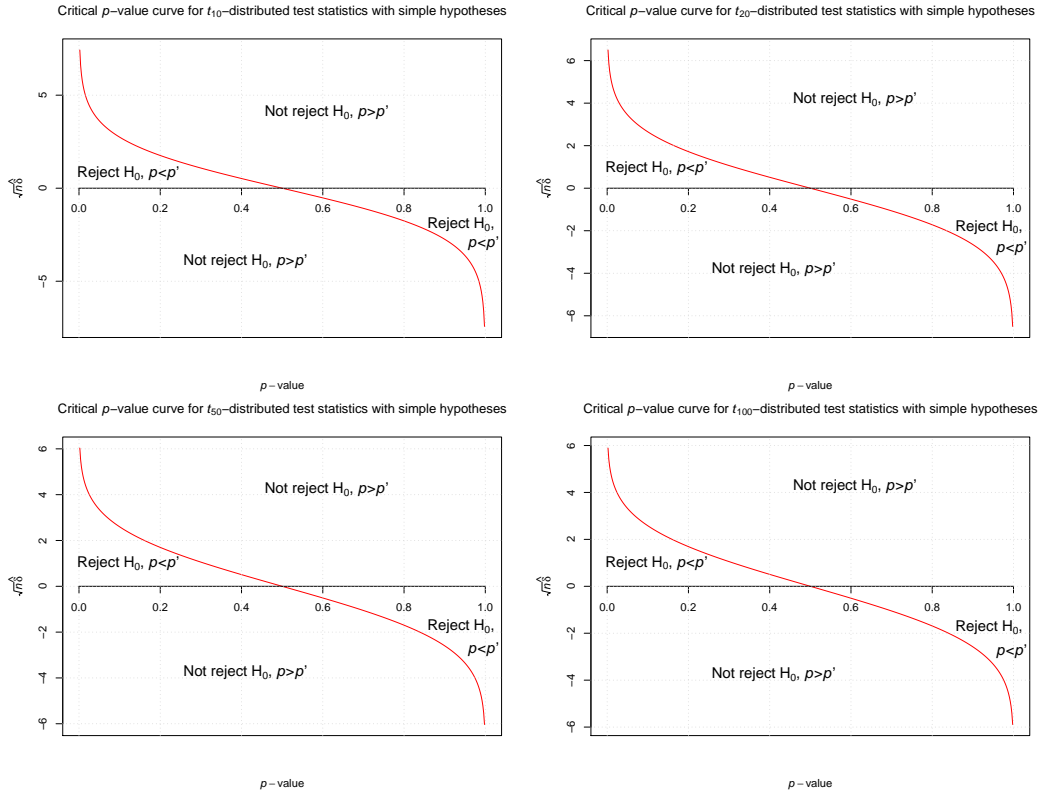


Figure 4: Critical  $p$ -value curve for  $t$ -distributed test statistics with 10, 20, 50 and 100 degrees of freedom with simple forms for  $H_0$  and  $H_1$ .

## 5.2 Simple hypotheses with $t$ -distributed test statistics

Critical  $p$ -value curves and surfaces involve setting  $p = p'$ .  $f_P(p|H_0)$  is known to be uniform, while  $f_P(p|H_1)$  depends on the test statistic distribution. In the previous subsection, standard Gaussian test statistics were considered, however curves can be constructed for a variety of distributions. In Figure 4 critical  $p$ -value curves are presented for  $t$ -distributed test statistics for various degrees of freedom for simple forms of  $H_1$ . As can be seen there is little change in the position of the curve as the degrees of freedom are adjusted, with Figure 3 representing the limiting case.

For  $p = p'$  in this environment where  $\hat{\delta} = k/\hat{\sigma}$  (case of unknown, hence estimated,

variance), we seek<sup>18</sup>

$$p = F_T(t_{p,\nu} - \sqrt{n}\hat{\delta}; \nu) \quad (31)$$

for  $\sqrt{n}\hat{\delta} > 0$ , with compliments for negative  $\sqrt{n}\hat{\delta}$  comparable with (26) which still yields (31) upon rearrangement, analogous to the standard Gaussian case above.

Formally, the decision rule can be stated as

$$\text{For } \sqrt{n}\hat{\delta} > 0 : \begin{cases} \text{Reject } H_0, \text{ report } \alpha(s) & : \sqrt{n}\hat{\delta} < t_{p,\nu} - T_\nu^{-1}(p) \\ \text{Do not reject } H_0, \text{ report } \beta(s) & : \sqrt{n}\hat{\delta} > t_{p,\nu} - T_\nu^{-1}(p). \end{cases} \quad (32)$$

$$\text{For } \sqrt{n}\hat{\delta} < 0 : \begin{cases} \text{Reject } H_0, \text{ report } \alpha(s) & : \sqrt{n}\hat{\delta} > t_{p,\nu} - T_\nu^{-1}(p) \\ \text{Do not reject } H_0, \text{ report } \beta(s) & : \sqrt{n}\hat{\delta} < t_{p,\nu} - T_\nu^{-1}(p). \end{cases} \quad (33)$$

Note  $t_{p,\nu}$  is the  $(1-p)^{th}$  percentile of a Student's  $t$  variable on  $\nu$  degrees of freedom (analogous to  $Z_p$  in the standard Gaussian case) and  $T_\nu^{-1}$  is the quantile function for such a distribution (analogous to  $\Phi^{-1}$ ).

### 5.3 Critical $p$ -value surfaces for composite alternative hypotheses

The critical  $p$ -value curves presented above are ideal for testing simple hypotheses. However when testing a null, say, of  $H_0 : \mu = 0$ , very often we are interested in composite forms of  $H_1$ , for example  $H_1 : \mu \neq 0$ . Hung, O'Neill, Bauer, and Köhne (1997) derive the respective  $p$ -value density and distribution functions. For the Gaussian

---

<sup>18</sup>Using (40), the  $t$ -distribution equivalents of (20) and (21) become respectively,

$$F_P(p|H_1) = \int_0^p \frac{f_T(t_x - \sqrt{n}\hat{\delta}; \nu)}{f_T(t_x; \nu)} dx = 1 - F_T(t_{p,\nu} - \sqrt{n}\hat{\delta}; \nu) \quad (29)$$

and

$$P' = \Pr(P > p|H_1) = F_T(t_{p,\nu} - \sqrt{n}\hat{\delta}; \nu). \quad (30)$$

case such that  $\delta \sim N(\xi, \omega^2)$  with sample size  $n$ ,

$$f_P(p|\mathbf{H}_1) = g_{\xi, \omega, n}(p) = \left[ \omega(n + \omega^{-2})^{1/2} \right]^{-1} \times \exp \left\{ -\frac{1}{2} \left[ (\xi/\omega)^2 - (\sqrt{n}Z_p + \xi/\omega^2)^2 / (n + \omega^{-2}) \right] \right\} \quad (34)$$

$$F_P(p|\mathbf{H}_1) = G_{\xi, \omega, n}(p) = 1 - \Phi \left\{ (Z_p - \sqrt{n}\xi) / (\omega^2 n + 1)^{1/2} \right\}, \quad (35)$$

where (34) and (35) denote the density and distribution functions, under  $\mathbf{H}_1$ , respectively.

This construction introduces two new parameters, namely  $\xi$  and  $\omega^2$ . In order to provide a graphical depiction for when  $p = p'$ , it is possible to construct a critical  $p$ -value surface by controlling for one of these additional parameters.  $\xi$  will be chosen for this purpose.

As Figure 5 demonstrates, when the domain of  $\delta$  under  $\mathbf{H}_1$  encompasses both positive and negative values (as is the case for the Gaussian distribution),  $f_P(p|\mathbf{H}_1)$  has significant density concentration around both 0 and 1, achieving a minimum in the vicinity of 0.5. In order to accommodate these features it is necessary to re-state the subsets of the respective densities which define the  $p$  and  $p'$  regions.

Given we seek to reject  $\mathbf{H}_0$  when the observed  $p$ -value is sufficiently unlikely vis-à-vis  $\mathbf{H}_1$ , it is necessary to associate extremely small and extremely large  $p$ -values with this region, due to the distribution of  $f_P(p|\mathbf{H}_1)$ . Similarly,  $p'$  will be associated with low probabilities of  $p$  under  $f_P(p|\mathbf{H}_1)$  vis-à-vis  $f_P(p|\mathbf{H}_0)$ . Such a region exists around  $p \approx 0.5$ .

### 5.3.1 Surface for $\xi = 0$

Controlling for the mean parameter  $\xi$ , by setting it to zero the distribution of  $\delta$  under  $\mathbf{H}_1$  is symmetric about zero. To obtain equality between  $p$  and  $p'$ , it is necessary to solve the following,

$$2p = F_P(1 - p|\mathbf{H}_1) - F_P(p|\mathbf{H}_1). \quad (36)$$

The left-hand-side value of  $2p$  represents the rejection region  $p$  being comprised of two equal-sized tails (each of area  $p$ ), the lower covering  $[0, p]$  and the upper  $[1 - p, 1]$  with

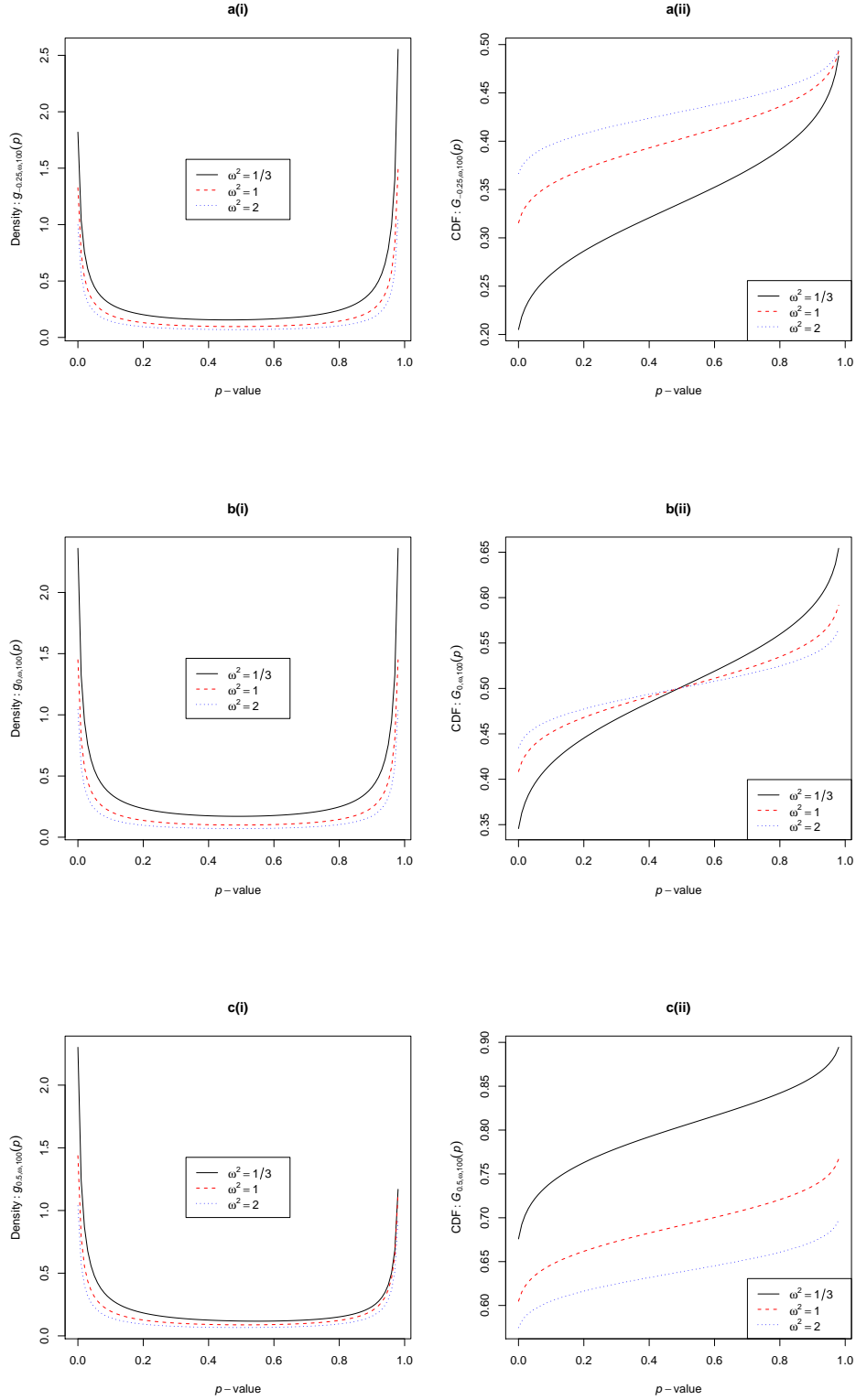


Figure 5:  $P$ -value density and distribution functions for composite forms of  $H_1$  where the effect size parameter  $\delta \sim N(\zeta, \omega^2)$  as per (34) and (35). a(i) =  $g_{-0.25, \omega, 100}(p)$ , a(ii) =  $G_{-0.25, \omega, 100}(p)$ , b(i) =  $g_{0, \omega, 100}(p)$ , b(ii) =  $G_{0, \omega, 100}(p)$ , c(i) =  $g_{0.5, \omega, 100}(p)$  and c(ii) =  $G_{0.5, \omega, 100}(p)$ .

total area of  $2p$ . Meanwhile the right-hand-side specifies  $p'$ , that is the probability of being between  $p$  and  $1-p$  under  $H_1$ , whose area can be computed from the distribution function under  $H_1$ .

Consequently the critical  $p$ -value surface comprises a floor and a ceiling due to the dual-nature of unlikely  $p$ -values under  $H_0$ , namely the lower and upper tails. Figure 6 presents these. With respect to the floor (ceiling), as the  $p$ -value decreases,  $p$  decreases ( $1-p$  increases), while  $p'$  increases, hence for  $p$ -values below the floor (above the ceiling)  $H_0$  should be rejected, while it should not be rejected between the floor and ceiling. Formally,

$$\begin{aligned} \text{Reject } H_0, \text{ report } \alpha(s) & \quad \text{if} \quad \frac{1}{2} \left( \Phi \left( \frac{Z_p - \sqrt{n}\xi}{(\omega^2 n + 1)^{1/2}} \right) - \Phi \left( \frac{Z_{1-p} - \sqrt{n}\xi}{(\omega^2 n + 1)^{1/2}} \right) \right) > p \\ \text{Do not reject } H_0, \text{ report } \beta(s) & \quad \text{if} \quad \frac{1}{2} \left( \Phi \left( \frac{Z_p - \sqrt{n}\xi}{(\omega^2 n + 1)^{1/2}} \right) - \Phi \left( \frac{Z_{1-p} - \sqrt{n}\xi}{(\omega^2 n + 1)^{1/2}} \right) \right) < p. \end{aligned}$$

### 5.3.2 Surface for non-zero $\xi$

For  $\xi \neq 0$ ,  $f_P(p|H_1)$  is no longer symmetric about  $p = 0.5$ , as demonstrated in Figure 5. Therefore the critical  $p$ -value floor/ceiling surfaces cannot simply be obtained using  $p$  and  $1-p$ , as the density  $f_P(p|H_1)$  has different weights around 0 and 1. Therefore to equate  $p$  and  $p'$ , we must solve

$$a + (1 - b) = F_p(b|H_1) - F_p(a|H_1), \text{ s.t. } a < m < b, \quad (37)$$

where  $m = \arg_p f_p(p|H_1)$ . Note  $\xi = 0$  is simply a special case of (37) with  $a = p$  and  $b = 1 - p$ . Sample floors and ceilings for  $\xi = 0.5$  and  $\xi = -1$  are given in Figure 7. Note in particular the behaviour of the ceiling for small values of  $\omega^2$ . Also, recall each point on the surface corresponds to a different distribution  $f_P(p|H_1)$  due to the specification of changing values for the  $\omega^2$  parameter. Interpretation of the surfaces in terms of when to reject  $H_0$  is analogous to the  $\xi = 0$  case.

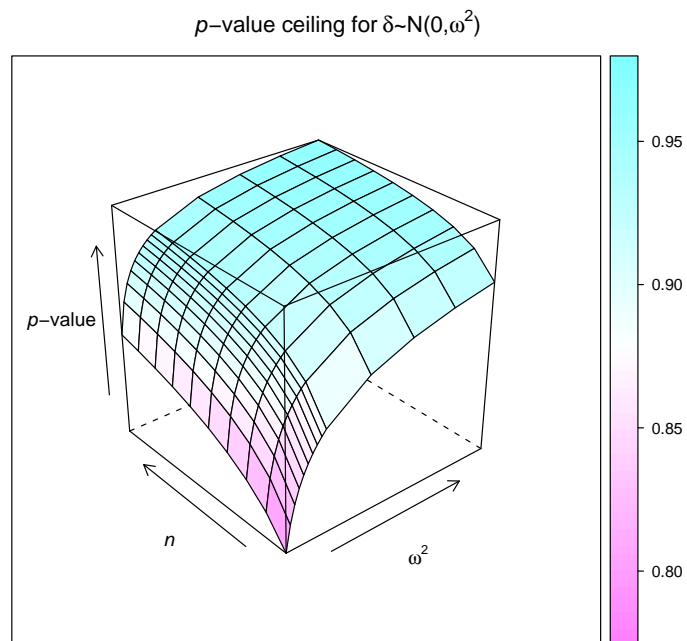
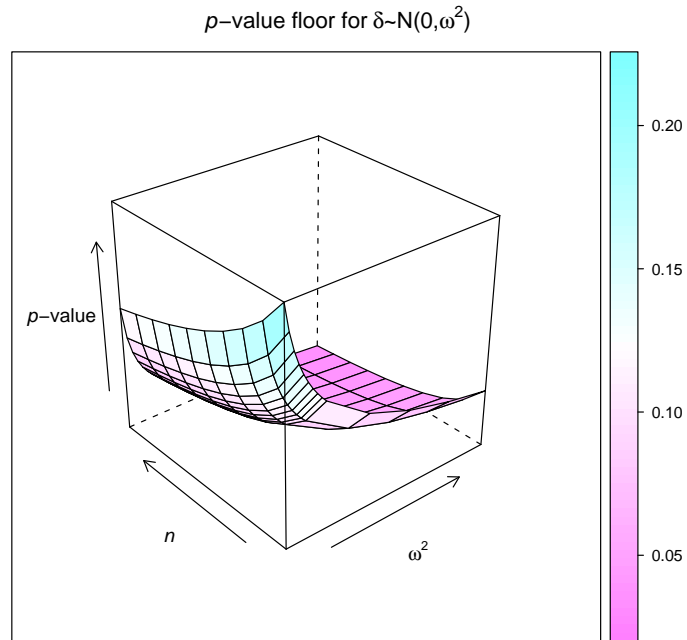


Figure 6: Critical  $p$ -value surface floor and ceiling for  $\delta \sim N(0, \omega^2)$  under  $H_1$ .

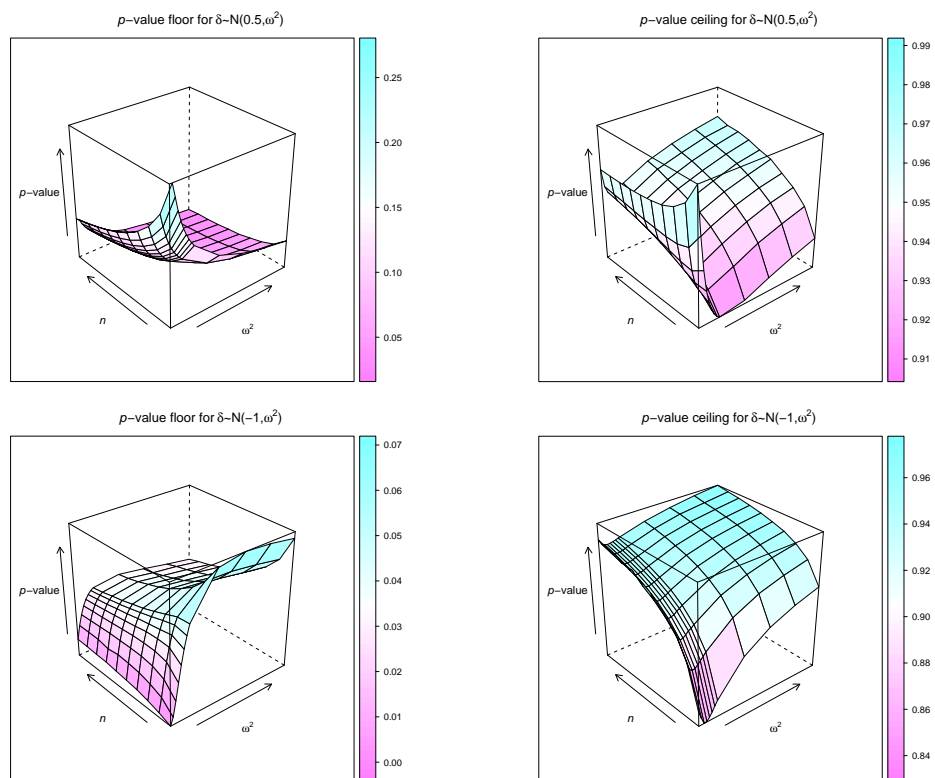


Figure 7: Critical  $p$ -value floors and ceilings for  $\delta \sim N(0.5, \omega^2)$  and  $\delta \sim N(-1, \omega^2)$  respectively under  $H_1$ .

## 6 Conclusions

This paper has endeavoured to extend the methodological unification of the Neyman-Pearson, Fisherian and Bayesian schools of hypothesis testing. Although each doctrine has its merits, each also carries limitations, as discussed. To date, the concept of conditional frequentist testing has offered a plausible unification, however results in this paper extend this methodology by explicitly considering the behaviour of the  $p$ -value under  $H_1$  through  $f_P(p|H_1)$ .

A variant of an oft-cited conditioning statistic has been proposed making use of the so-called second-order  $p$ -value,  $p'$ , which is obtainable from the original  $p$ -value. By taking the maximum of  $p$  and  $p'$ , the researcher can conclude in favour of the most plausible hypothesis, *conditional on the observed data*. By reporting the conditional error probability in conjunction with the reject/not reject decision, the end-user of the test result can decide the strength of the conclusion themselves.

In order to offer a simple-to-use framework of this methodology, the research above also presents new critical  $p$ -value curves and surfaces. By graphically displaying, for a range of parametric specifications relating to  $H_1$ , the  $p$ -value which results in equality between the  $p$  and  $p'$  rejection regions under  $H_0$  and  $H_1$  respectively, it is possible to quickly identify the correct decision in relation to whether to reject  $H_0$  while taking into account the specification of  $H_1$  and sample size.

Applying this methodology allows informed decision making, by accommodating the plausibility of *both* hypotheses for a given set of data. Just as in conventional hypothesis testing, inferential errors can occur but these are reflected in the more useful conditional error probabilities. Collectively, this approach helps in the quest for the holy grail of a unified inferential framework universally accepted by proponents of the various testing schools.

## A Appendix: $P$ -value distributions and second order $p$ -values

Suppose  $X_n$  is used to test  $H_0 : \theta = a$  against  $H_1 : \theta = b$ ,  $a < b$ . Let  $X_n$  have the left-continuous distribution function  $F_{X_n}(x_n) = \Pr(X_n \leq x_n)$  under  $H_0$ , for realised  $x_n$ . The  $p$ -value statistic, i.e. significance probability, is then a random variable,  $P$ , calculated for continuous<sup>19</sup> test statistics as<sup>20</sup>

$$P = \Pr(X_n > x_n) = 1 - F_{X_n}(x_n) = \bar{F}_{X_n}(x_n), \quad (38)$$

hence  $p$ -values are one-to-one transformations of the random variable  $X_n$ , so are themselves random variables.<sup>21</sup> Pearson (1938) refers to this as the ‘probability integral transformation’ of the sample data. Advantages of the  $p$ -value include its simplicity, i.e. a single real number restricted to the unit interval, and also its universal application across test statistics with *any* distribution under  $H_0$  due to the transformation in (38). Consequently the unit interval  $[0, 1]$  is a common scale for comparison allowing meta-analyses to be performed.

Given either hypothesis could be true,  $p$ -values will have different distributions accordingly. As shown in many studies, under a non-composite null this is a uniform distribution for *any* continuous test statistic. Let  $P$  be the  $p$ -value random variable with

---

<sup>19</sup>For discrete distributions a correction may be required — see Cox (1977).

<sup>20</sup>This yields *exact*  $p$ -values, as opposed to *approximate*  $p$ -values which are computed by using an approximation of  $F_{X_n}$ , for example when  $F_{X_n}$  is unknown.

<sup>21</sup>As shown in (38),  $p$ -values can be viewed in terms of the survival function,  $\bar{F}_{X_n}$ .

$F_P(p|\mathbf{H}_0)$  being the corresponding distribution function under  $\mathbf{H}_0$ , then using (38)

$$\begin{aligned}
F_P(p|\mathbf{H}_0) &= \Pr(P \leq p|\mathbf{H}_0) \\
&= \Pr(1 - F_{X_n}(x_n) \leq p|\mathbf{H}_0) \\
&= 1 - F_{X_n}(F_{X_n}^{-1}(1 - p)) \\
&= 1 - (1 - p) \\
&= p
\end{aligned} \tag{39}$$

for  $p \in [0, 1]$ . Hence this gives a density function,  $f_p(p|\mathbf{H}_0) = 1$ , consistent with a uniform density. It should be noted that this density is independent of the test statistic distribution, sample size and effect size. Therefore under  $\mathbf{H}_0$  it is impossible to distinguish  $p$ -values obtained from small and large samples as well as between tests engineered to have high power and those less powerful.

Denoting the left-continuous distribution function of  $X_n$  under the alternative hypothesis by  $G_{X_n}(x_n)$ , then the  $p$ -value distribution function under  $\mathbf{H}_1$ ,  $F_P(p|\mathbf{H}_1)$ , becomes

$$\begin{aligned}
F_P(p|\mathbf{H}_1) &= \Pr(P \leq p|\mathbf{H}_1) \\
&= \Pr(1 - F_{X_n}(x_n) \leq p|\mathbf{H}_1) \\
&= 1 - G_{X_n}(F_{X_n}^{-1}(1 - p)),
\end{aligned} \tag{40}$$

which clearly depends on the test statistic's distribution under both hypotheses.

Given two  $p$ -values both greater than  $\alpha$ , say 0.4 and 0.8, although both 'lend support' to  $\mathbf{H}_0$  under classical hypothesis testing, does it mean that the larger value offers stronger evidence in favour of  $\mathbf{H}_0$ ? This is an important issue, since most empirical studies are only concerned about whether the null can be rejected and so ignore test power and the consequences of  $\mathbf{H}_1$ . To answer this, Donahue (1999) considers not only how far the data

fall from the null hypothesis, but also how far the data fall from a *specific* alternative hypothesis. To achieve this, reporting two summary statistics, the original  $p$ -value and a ‘second-order’  $p$ -value,  $p'$ , defined below, is required. Both act as quasi-*post hoc* risk levels indicating certain inferential decision errors.

For the general case covering all possible test statistic distributions,

$$\begin{aligned}
 P' &= \Pr(P > p | H_1) \\
 &= 1 - \Pr(P \leq p | H_1) \\
 &= G_{X_n}(F_{X_n}^{-1}(1 - p)),
 \end{aligned}
 \tag{41}$$

from which a rejection region (for  $H_1$ ) can be obtained, for a given significance level, say  $\gamma = 0.05$ . Consequently, instead of basing inference solely on  $H_0$  as is frequently the case, assessments on the merits of both hypotheses can be presented namely:

- i.  $p$  provides a summary statistic measuring the deviation of the data from  $H_0$
- ii.  $p'$  provides a summary statistic measuring the deviation of the data from  $H_1$ .

## References

- Anderson, D. R., K. P. Burnham, and W. L. Thompson (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management* 6, 912–923.
- Bayarri, M. J. and J. O. Berger (2004). The interplay of Bayesian and Frequentist analysis. *Statistical Science* 19, 58–80.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer Verlag.
- Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing (with discussion)? *Statistical Science* 18, 1–32.

- Berger, J. O. and D. Berry (1988). Statistical analysis and the illusion of objectivity. *American Scientist* 76, 159–165.
- Berger, J. O., L. D. Brown, and R. L. Wolpert (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics* 22, 1787–1807.
- Berger, J. O. and M. Delampady (1987). Testing precise hypotheses. *Statistical Science* 2, 317–335.
- Berger, J. O. and J. Montera (1999). Default Bayes factors for non-nested hypothesis testing. *Journal of the American Statistical Association* 94, 542–554.
- Berger, J. O. and T. Sellke (1987). Testing a point null hypothesis: The irreconcilability of p-values and evidence. *Journal of the American Statistical Association* 82, 112–122.
- Berger, J. O. and R. L. Wolpert (1988). *The Likelihood Principle*. Hayward, California: Institute of Mathematical Statistics.
- Birnbaum, A. (1961). On the foundations of statistical inference: binary experiments. *The Annals of Mathematical Statistics* 32, 414–435.
- Bjørnstad, J. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association* 91, 791–806.
- Brownie, C. and J. Kiefer (1977). The ideas of conditional confidence in the simplest setting. *Communications in Statistics - Theory and Methods* 6, 691–751.
- Carlson, R. (1976). The logic of tests of significance (with discussion). *Philosophy of Science* 43, 116–128.
- Casella, G. and R. L. Berger (1987). Reconciling Bayesian and Frequentist evidence in the one-sided testing problem (with discussion). *Journal of the American Statistical Association* 82, 106–111.

- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist* 49, 997–1003.
- Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics* 4, 49–70.
- Diamond, G. A. and J. S. Forrester (1983). Clinical trials and statistical verdicts: probable grounds for appeal. *Annals of Internal Medicine* 98, 385–394.
- Dickey, J. M. (1973). Scientific reporting. *Journal of the Royal Statistical Society* 35, 285–305.
- Donahue, R. M. J. (1999). A note on information seldom reported via the P value. *The American Statistician* 53, 303–306.
- Edwards, W., H. Lindman, and L. J. Savage (1963). Bayesian statistical inference for psychological research. *Psychological Review* 70, 193–242.
- Efron, B. and A. Gous (2001). Scales of evidence for model selection: Fisher versus Jeffreys (with discussion). In P. Lahiri (Ed.), *Model Selection*. Beachwood.
- Goodman, S. (1999a). Toward evidence-based medical statistics 1: the  $p$ -value fallacy. *Annals of Internal Medicine* 130, 995–1004.
- Goodman, S. (1999b). Toward evidence-based medical statistics 2: the Bayes factor. *Annals of Internal Medicine* 130, 1005–1013.
- Hall, P. and B. Sellinger (1986). Statistical significance: balancing evidence against doubt. *Australian Journal of Statistics* 28, 354–370.
- Hung, H. M. J., R. T. O’Neill, P. Bauer, and K. Köhne (1997). The behavior of the  $p$ -value when the alternative hypothesis is true. *Biometrics* 53, 11–22.
- Hwang, J. T., G. Casella, C. Robert, M. T. Wells, and R. Farrell (1992). Estimation of accuracy in testing. *The Annals of Statistics* 20, 490–509.
- Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press.

- Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics*, pp. 451–454. Amsterdam: North-Holland.
- Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *Journal of the American Statistical Association* 72, 789–827.
- LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer Verlag.
- Lehmann, E. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association* 88, 1242–1249.
- Pearson, E. S. (1938). The probability transformation for testing goodness of fit and combining independent tests of significance. *Biometrika* 30, 134–148.
- Reid, N. (1995). The roles of conditioning in inference. *Statistical Science* 10, 138–157.
- Robinson, D. H. and H. Wainer (2001). On the past and future of null hypothesis significance testing. Technical report, Research Report RR-01-24, Educational Testing Service, Princeton.
- Savage, L. J. (1976). On rereading R. A. Fisher. *The Annals of Statistics* 4, 441–500.
- Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1, 115–129.
- Sellke, T., M. J. Bayarri, and J. O. Berger (2001). Calibration of  $p$ -values for testing precise null hypotheses. *The American Statistician* 55, 62–71.
- Smith, A. F. M. and D. J. Spiegelhalter (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society* 42, 213–220.
- Spielman, S. (1978). Statistical dogma and the logic of statistical testing. *Philosophy of Science* 45, 120–135.

Wolpert, R. L. (1995). Testing simple hypotheses. In H. H. Bock and W. Polasek (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization*, Volume 7. Heidelberg: Springer Verlag.

Zabell, S. (1992). R. A. Fisher and the fiducial argument. *Statistical Science* 7, 369–387.