Brazilian Journal of Probability and Statistics
2016, Vol. 30, No. 1, 70–90
DOI: 10.1214/14-BJPS257

Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities

Luis Pericchi[a] **and Carlos Pereira**[b]

[a]*University of Puerto Rico*
[b]*Universidade de São Paulo*

Abstract. Our purpose is to recommend a change in the paradigm of testing by generalizing a very natural idea, originated perhaps in Jeffreys [*Proceedings of the Cambridge Philosophy Society* **31** (1935) 203–222; *The Theory of Probability* (1961) Oxford Univ. Press] and clearly exposed by DeGroot [*Probability and Statistics* (1975) Addison-Wesley], with the aim of developing an approach that is attractive to all schools of statistics, resulting in a procedure better suited to the needs of science. The essential idea is to base testing statistical hypotheses on minimizing a weighted sum of type I and type II error probabilities instead of the prevailing paradigm, which is fixing type I error probability and minimizing type II error probability. For simple vs simple hypotheses, the optimal criterion is to reject the null using the likelihood ratio as the evidence (ordering) statistic, with a *fixed* threshold value instead of a *fixed* tail probability. By defining expected type I and type II error probabilities, we generalize the weighting approach and find that the optimal region is defined by the evidence ratio, that is, a ratio of averaged likelihoods (with respect to a prior measure) and a fixed threshold. This approach yields an optimal theory in complete generality, which the classical theory of testing does not. This can be seen as a Bayesian/non-Bayesian compromise: using a weighted sum of type I and type II error probabilities is Frequentist, but basing the test criterion on a ratio of marginalized likelihoods is Bayesian. We give arguments to push the theory still further, so that the weighting measures (priors) of the likelihoods do not have to be proper and highly informative, but just "well calibrated." That is, priors that give rise to the same evidence (marginal likelihoods) using minimal (smallest) training samples.

The theory that emerges, similar to the theories based on objective Bayesian approaches, is a powerful response to criticisms of the prevailing approach of hypothesis testing. For criticisms see, for example, Ioannidis [*PLoS Medicine* **2** (2005) e124] and Siegfried [*Science News* **177** (2010) 26–29], among many others.

Key words and phrases. Hypothesis testing, significance levels, errors type I and type II, weighting error probabilities.

Received November 2013; accepted August 2014.

1 Changing the paradigm of hypothesis testing and revisiting Bayes factors and likelihood ratios

1.1 Introduction

Classical significance testing, as developed by Neyman and Pearson, is suited to and was designed to perform very specific comparisons, under well designed studies for which the probability of a type I error (false rejection) has been fixed beforehand to some specific value α, and a most powerful statistic is found so that the probability of type II error β is minimized. The sample sizes are chosen so that β is bigger than, or at least of the same order as, α. But the vast majority of studies do not conform to this standard. Even when individual studies conform to the standard, merged studies no longer do. Fixing the probability of type I error, for whatever amount of evidence, as well as fixed tables of p-values, are not justifiable (at least when there is an explicit or implicit alternative hypothesis, as shown in Pereira and Wechsler (1993) and references within), since then the type II error is completely outside the statistician's control, with the possibility that type I error probability may be enormous as compared with type II error probability.

There is a need for an alternative to the prevailing paradigm, which is: (i) Fix type I error probability at α and Minimize type II error probability, or (ii) Calculate p-value and interpret it as the minimum α for which you will reject the null hypothesis, using a fixed table of values like $\{0.1, 0.05, 0.01\}$.

Alternatives to this approach date back to at least Jeffreys (1935, 1961). In our view, Morris DeGroot, in his authoritative book *Probability and Statistics*, 2nd edition (DeGroot, 1975), perhaps the best bridge between schools of statistics ever written, states in a didactic manner, that it is more reasonable to minimize a weighted sum of probabilities of type I and type II errors than to specify a type I error probability and then minimize the probability of type II error. DeGroot proves this, but only in the very restrictive scenario of simple-vs-simple hypotheses. We propose here a very natural generalization for composite hypotheses, by using general weight functions in the parameter space. This was also the position taken by Pereira (1985) and Berger and Pericchi (1996, 2001). Recent proposals for adaptive significance levels are in Varuzza and Pereira (2010) and Perez and Pericchi (2014). The developments in the present paper, provide a general Decision-Theoretic framework to justify alternatives to traditional hypothesis testing. We show, in a parallel manner to DeGroot's proof and Pereira's discussion, that the optimal test statistics are Bayes factors when the weighting functions are priors with mass on the whole parameter space and loss functions that are constant under each hypothesis. When there are areas of indifference (i.e., areas of no practical importance, like "the new drug is less than 10%, say, more effective than the current 'gold standard' "), then loss functions that are equal to zero in the indifference region (i.e., 0 to 10%) achieve the goal of practical significance instead of statistical significance.

Hypothesis testing of precise hypotheses is the most contentious aspect of statistics. There is no agreement between the Bayesian and Frequentist schools of statistics, nor even within those schools. It seems timely to shift to *weighting* paradigms in order to meet the needs of science. We call weighting paradigms those which minimize weighted types of errors, and weight nuisance parameters. We show here that the alternative theory yields general optimal tests, unlike the traditional theory, in which the existence of an optimal test is the exception rather than the rule.

We present only very simple examples, for the sake of clarity of a general argument.

1.2 The disagreement is not about the mathematics, it is about the statistical implementation

To fix ideas let us suppose a simple hypothesis $H_0 : f(\mathbf{x}|\theta_0)$ vs $H_1 : f(\mathbf{x}|\theta_1)$. All schools of Statistics agree that rejection of H_0 should take place when the Likelihood Ratio is small, that is

$$\frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} < \mathrm{r}.$$

The justification for such a rule, comes from different versions of the Fundamental Neyman–Pearson lemma. However, how to choose r is the *crux of the matter*. For example two ways to assign r are the pure Bayesian and the pure Frequentist. In the former, $\mathrm{r} = \frac{(1-P(H_0))\cdot L_0)}{P(H_0)\cdot L_1}$, where $P(H_0)$ is the prior probability of the Null Hypothesis, and L_i the loss for accepting H_i when H_j is true, $i, j = 0, 1; i \neq j$ (see Section 1.7). On the other hand, for the latter r should be chosen indirectly, such that $P(\frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} < \mathrm{r}|H_0) = \alpha$. See, for example, DeGroot (1975, Chapter 8). Thus in the former there is one choice r, while in the latter the choice is α. This seemingly small detail makes an enormous difference. In the sequel, we will propose a mixed way to choose the threshold.

1.3 Why do significance tests work for carefully designed studies, but not otherwise?

Designed studies for testing hypotheses following classical significance testing are based on a careful balance between controlling the probability of type I error and minimizing type II error.

Consider the following example motivated by DeGroot (1975, Section 8.2).

Example 1. Suppose we have Normal data with scale $\sigma_0 = 3$, and we are interested in comparing two possible means:

$$H_0 : \theta = \theta_0 = -1 \quad \text{vs} \quad H_1 : \theta = \theta_1 = 1.$$

It is desired to design an experiment (the observations are costly) so that the probability of an error of type I (False Rejection of H_0) is 0.05 and the probability of a type II error (False Acceptance of H_0) is 0.1. Application of the Classical Neyman–Pearson lemma yields an optimal criterion based on the likelihood ratio:

$$\text{Reject } H_0 \text{ if } \frac{\bar{x}-(-1)}{\sigma_0/\sqrt{n}} \geq C_\alpha .$$

Now the constant C_α is chosen as to have a type I error probability equal 0.05, that is:

$$\Pr\left(\frac{\bar{X}-(-1)}{\sigma_0/\sqrt{n}} \geq C_\alpha|H_0\right)=\alpha,$$

which is immediately recognized as the familiar $C_{\alpha=0.05}=1.645$. Turning to the type II requirement then,

$$\beta=\Pr\left(\frac{\bar{X}-(-1)}{\sigma_0/\sqrt{n}} < 1.645|H_1\right)=0.1,$$

which gives $n=19.25$, so we settle for $n=20$ as our designed experiment, resulting in $\beta=0.091$. This implies that H_0 is rejected if $\bar{x}>0.1$. Notice that in this situation $\alpha/\beta=0.55$, or a ratio of about 1 to 2 between the probabilities of type I and type II errors.

1.4 The conundrum of "an approach bothered by good information"

Example 1 (Continued). After you give the researchers your design, they come back to you and proudly give you $n=100$ data, since it cost the same to produce $n=100$ as $n=20$, a situation that is not that unusual. The researchers are very satisfied with their prolific experiment, but the statistician is disturbed. As usual, type I error probability is kept fixed and equal to $\alpha=0.05$, but... what is the new type II error probability? The statistician makes a calculation and obtains that the new type II error probability is $\beta=0.00000026$, or $\alpha/\beta=195{,}217$, quite different from 1 to 2 as designed. In fact, the rejection region becomes $\bar{x}>-0.51$. Thus, if, for example, the observed sample mean is $\bar{x}=-0.5$, the null hypothesis $H_0:\theta=-1$ is rejected in favor of an alternative much further away from the observed value. This leads to a conundrum: why is more information a bad thing for the traditional approach to hypothesis testing? Incidentally, in the perhaps more frequent situation in which information is lost, the type I and type II error probabilities can be unbalanced in the opposite direction. For example, if the real sample size delivered by the researchers were $n=10$, then type II error probability would be inflated to $\beta=0.32$ if α were kept fixed, yielding a ratio of $0.05/0.32=0.156$, noticeably lower than the ratio of 0.55 in the designed experiment.

Example 1 illustrates why significance testing is inadequate for measuring the evidence in favor of or against a hypothesis in general. This is motivation to go back to the essentials.

Among others, DeGroot (1975) argues that instead of fixing type I error probability (or computing a p-value with a scale held fixed) and minimizing type II error probability, a better hypothesis testing paradigm is to choose the test to minimize a weighted sum of the error probabilities, namely

$$\mathrm{Min}_\delta\big[a \cdot \alpha_{\theta_0}(\delta) + b \cdot \beta_{\theta_1}(\delta)\big], \tag{1}$$

where δ denotes the test: $\delta(\mathbf{x}) = I_{\mathrm{R}}(\mathbf{x})$, where R is the rejection region of H_0, $I_S(y)$ is the indicator function, equal to 1 if $y \in S$ and 0 otherwise. Notice the apparently slight difference between the expression denoted by (1) and the traditional approach of significance testing at a fixed significance level α_0:

$$\begin{gathered}\text{Restricting to those tests } \delta \text{ on which type I error:}\\ \alpha_{\theta_0}(\delta) \le \alpha_0, \qquad \mathrm{Min}_\delta\, \beta_{\theta_1}(\delta).\end{gathered} \tag{2}$$

This difference is far-reaching, as we will see. In the following, the superiority of the weighting approach becomes distinctly apparent.

Example 1 (Continued). For simple-vs-simple hypotheses DeGroot proves that the optimal test is, for a and b defined in (1):

$$\text{Reject } H_0 \text{ if } \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} < \frac{b}{a}. \tag{3}$$

This approach can handle effectively *any* sample size, as long as we are prepared to select a and b, or more precisely b/a. This has been perhaps the most important contention, the reason given not to embrace a more balanced and sensible combination of the two types of error. The point we make here is that the choice of b/a has already been made! To see this, we go back to the design situation in which the sample size was chosen to be $n = 20$. Now for $\alpha = 0.05$ and $\beta = 0.091$, the weighted rejection region (3) is equivalent, after some algebra, to

$$\text{Reject if } \exp\left(\frac{n}{\sigma_0^2}(\theta_1 - \theta_0)\big[(\theta_0 + \theta_1)/2 - \bar{x}\big]\right) < \frac{b}{a}, \tag{4}$$

and since the rejection region is $\bar{x} > 0.1034$, that is equivalent in our particular example to $-\frac{3}{2\sqrt{20}} \cdot \log(\frac{b}{a}) + \sqrt{20}/3 = 1.645$, obtaining $\frac{b}{a} = 0.63$. Thus if we set $a = 1$, the implicit value of b is 0.63. Now the weighted approach leads to a criterion that makes sense for any sample size, n

$$\text{The Optimal Rejection Region } \mathrm{R} \text{ is: } \bar{x} > \frac{9}{2 \cdot n} \times 0.46,$$

with a cutoff point that is always positive but approaches zero, as is intuitively reasonable for $\theta_0 = -1$ and $\theta_1 = 1$, and $a > b$. Furthermore, now the ratio between $\alpha = 1 - \Phi(3 \cdot 0.23/\sqrt(n) + \sqrt{n}/3)$ and $\beta = \Phi(3 \cdot 0.23/\sqrt(n) - \sqrt{n}/3)$, as a function of n, is extremely stable, ranging from 0.55 at $n = 20$ to 0.61 at $n = 100$. Thus, we have found that an α of 0.05 for $n = 20$, is equivalent to an α of 0.00033 for $n = 100$. This shows the extent to which the usual method of leaving α unchanged, whatever the information available, is unbalanced. Notice that not even changing to $\alpha = 0.01$ would have been an effective remedy for $n = 100$, since the equivalent α is about thirty times smaller.

Note: The previous analysis provides an interesting method to "decrease α with n." Notice that from the formula for α above, using Mill's ratio, we get the following simple approximation:

$$\alpha_n \sim 1 - \Phi(\sqrt{n}/\sigma_0) \approx \frac{\phi(\sqrt{n}/\sigma_0)}{\sqrt{n}/\sigma_0}, \tag{5}$$

giving clear guidance on how to decrease the scale of p-values with increasing sample size. Notice how fast the p-values ought to decrease with the sample size to give a comparable amount of surprise against the model. In Section 7, we will see that the rate of decrease is different (much slower) for more complex tests.

1.5 The Lindley Paradox is not necessarily a difference between Bayesian and non-Bayesian, but between fixed significance levels and minimizing the weighted sum of error probabilities

Lindley's Paradox (Lindley, 1957) has been understood as the increasing divergence (as information accumulates) between the evidence measures of Classical hypothesis testing and Bayesian testing. There is also a divergence between Weighting's and Classical testing, even when there are no prior densities.

To see this, we go back to the motivating example of a simple hypothesis against a simple alternative as the simplest setting in which it becomes clear that the discrepancy is due to differences in what is to be minimized. If one relinquishes fixed significance levels and adopts the weighting approach of minimizing the weighted sum of error probabilities, then there is a one-to-one relationship with Bayesian posterior model probabilities (as it is with testing based on the likelihood ratio).

To see this, recall that in the approach recommended by DeGroot, the minimization condition is given by equation (1).

Example 1 (Continued). In this example, the optimal rejection region given by the weighting approach can be written as

$$\text{Reject if } \frac{\bar{x}+1}{\sigma_0/\sqrt{n}} \geq \frac{2.07}{\sigma_0\sqrt{n}} + \frac{\sqrt{n}}{\sigma_0}, \tag{6}$$

to be compared to the traditional rejection rule with fixed significance levels

$$\text{Reject if } \frac{\bar{x}+1}{\sigma_0/\sqrt{n}} \geq 1.645. \tag{7}$$

This is the divergence, or Lindley's Paradox, but between two Frequentist rejection rules. Notice the striking difference in behavior of the two right-hand sides: the fact is that under the weighting criterion as n grows the type I error probability goes to zero (as does that of type II error, and so consistency is achieved), but in the traditional approach the type I error probability is kept fixed and consistency fails (there is a positive probability of false rejection no matter how large n is).

On the other hand, a general one-to-one relationship can be established between the probability of the null hypothesis and the criteria obtained by minimizing the weighted sum of error probabilities. For any simple-vs-simple comparison, if π_0 and π_1 are, respectively, the prior probabilities of the null and the alternative, then Bayes's theorem yields as the posterior probability of the null

$$P(H_0|x) = \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)} = \left[1 + \frac{\pi_1 f(x|\theta_1)}{\pi_0 f(x|\theta_0)}\right]^{-1}. \tag{8}$$

Therefore, if the ratio b/a is interpreted as π_1/π_0 (assuming equal losses $L_0 = L_1$), then the rejection region obtained by the weighting method is equivalent to the region in which rejection of the null occurs if $P(H_0|x) < 0.5$. There is no divergence between schools of statistical inference here; on the contrary, there is a perfect correspondence.

1.6 A more general setting

Suppose now that we are testing the following two general hypotheses:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1. \tag{9}$$

We define, in the Neyman–Pearson tradition, type I and type II error probabilities for the test δ at the parameter value θ as

$$\alpha_\theta(\delta) = \Pr(\text{Rejecting } H_0|\theta \in \Theta_0), \tag{10}$$

$$\beta_\theta(\delta) = \Pr(\text{Accepting } H_0|\theta \in \Theta_1). \tag{11}$$

Definition. The weighted (or expected) type I and type II error probabilities are defined respectively, as:

$$\alpha(\delta) = \int_{\Theta_0} \alpha_\theta(\delta)\pi_0(\theta)\, d\theta \tag{12}$$

and

$$\beta(\delta) = \int_{\Theta_1} \beta_\theta(\delta)\pi_1(\theta)\, d\theta, \tag{13}$$

where $\pi_j(\theta) \geq 0$ are such that $\int_{\Theta_j} \pi_j(\theta) = 1$, $j \in \{0, 1\}$ (this condition will be relaxed in Section 1.8). Why the expectation? It is important to notice that the error probabilities depend on θ, which is obviously unknown. How to deal with this?

In Berger (2008), it is stated that: "There are many situations in which it has been argued that a Frequentist (statistician) should use an average coverage criterion; see Bayarri and Berger (2004) for examples and references." Similarly, we argue here that both Bayesians and Frequentists should average error probabilities. Among the main reasons we have: (i) averaging error probabilities permits a completely general theory of optimality, as we will see; (ii) averaging is natural (from a probabilistic point of view, it is the marginal error probability) and flexible in the choice of weight functions; (iii) the methodologies for assessing weighings have advanced (see, e.g., Pereira (1985), or Berger and Pericchi (1996)); and (iv) it is a natural mix of Frequentist error probabilities and Bayesian averaging.

1.6.1 *Interpretations of the weight* (*prior*) *measures*. We pause here to discuss interpretations, because there are multiple possible interpretations of the weight measures $\pi_j(\theta)$, $j = 0, 1$:

1. Prior Measures: The most obvious interpretation of these measures is that they are the assumed prior densities of the parameter values conditional on each hypothesis, which is the natural interpretation under a Bayesian framework. Notice that this interpretation does not necessarily lead to a subjective approach. If a general method for generating conventional priors, like the *Intrinsic Prior* method, is used, then this can be considered an objective approach. There is room for other conventional priors.
2. Regions of Statistical Importance: In order to state "statistical importance" rather than "statistical significance," the weight function can be combined with a loss structure to define "indifference regions" on which the difference between the null and alternative is of no practical importance. See the examples.

It turns out that under the prior measure interpretation, we obtain a Frequentist Decision Theory justification of Bayes Factors. For the second interpretation, the decisions are based on posterior probabilities of sets that actually embody rules based on "statistical importance."

Note: It is tempting, because it is so simple, to use weight functions that are point masses in the null and the point where statistical importance starts. These are point masses signalling specific points for which the error probabilities ought to be controlled by design. For example, if for a particular value of $\theta_1 \in \Theta_1$, where there is "practical significance," such as a novel medical treatment improvement of 20%, then the weight function may be set as a point mass on 20% improvement (this typically would work only for monotone likelihood ratio families). We consider these point masses (i) for simplicity and (ii) to compare to frequentist solutions of the problem of "too much power," that is, when there is statistical significance but not practical significance (Bickel and Doksum (1977)). However, this simple solution is not based on the most reasonable prior.

1.7 A general optimality result

Define the weighted likelihoods, which we may call the *evidence measures* for the data $\mathbf{y}$ under each hypothesis, as

$$\varpi_0(\mathbf{y}) = \int_{\Theta_0} f(\mathbf{y}|\theta)\pi_0(\theta)\,d\theta, \tag{14}$$

and

$$\varpi_1(\mathbf{y}) = \int_{\Theta_1} f(\mathbf{y}|\theta)\pi_1(\theta)\,d\theta. \tag{15}$$

Lemma 1. *It is desired to find a test function δ that minimizes, for specified $a > 0$ and $b > 0$:*

$$\text{SERRORS}(\delta) = a \cdot \alpha(\delta) + b \cdot \beta(\delta). \tag{16}$$

The test δ^* is defined as

$$\textit{accept } H_0 \textit{ if } \frac{\varpi_0(\mathbf{y})}{\varpi_1(\mathbf{y})} > \frac{b}{a}, \tag{17}$$

$$\textit{accept } H_1 \textit{ if } \frac{\varpi_0(\mathbf{y})}{\varpi_1(\mathbf{y})} < \frac{b}{a}, \tag{18}$$

and accept any if $a \cdot \varpi_0(\mathbf{y}) = b \cdot \varpi_1(\mathbf{y})$. Then for any other test function δ:

$$\text{SERRORS}(\delta^*) = a \cdot \alpha(\delta^*) + b \cdot \beta(\delta^*) \le \text{SERRORS}(\delta). \tag{19}$$

In words, rejecting the null when the ratio of evidences is smaller than b/a is globally optimal.

Proof. Denote by R the rejection region of the test δ, that is, those data points on which H_0 is rejected. Then, under the mild assumptions of Fubini's theorem that allows interchanging the order of the integrals, for any test function δ,

$$\begin{aligned}
&a\alpha(\delta) + b\beta(\delta)\\
&\quad = a\int_{\Theta_0}\left[\int_{\mathrm{R}} f(\mathbf{y}|\theta)\,d\mathbf{y}\right]\pi_0(\theta)\,d\theta + b\int_{\Theta_1}\left[\int_{\mathrm{R}^C} f(\mathbf{y}|\theta)\,d\mathbf{y}\right]\pi_1(\theta)\,d\theta\\
&\quad = a\int_{\Theta_0}\int_{\mathrm{R}} f(\mathbf{y}|\theta)\pi_0(\theta)\,d\mathbf{y}\,d\theta + b\int_{\Theta_1}\int_{\mathrm{R}^C} f(\mathbf{y}|\theta)\pi_1(\theta)\,d\mathbf{y}\,d\theta\\
&\quad = a\int_{\Theta_0}\int_{\mathrm{R}} f(\mathbf{y}|\theta)\pi_0(\theta)\,d\mathbf{y}\,d\theta + b\left[1 - \int_{\Theta_1}\int_{\mathrm{R}} f(\mathbf{y}|\theta)\pi_1(\theta)\,d\mathbf{y}\,d\theta\right]
\end{aligned} \tag{20}$$

$$= b + \int_{\mathrm{R}} \left[a \int_{\Theta_0} f(\mathbf{y}|\theta)\pi_0(\theta)\,d\theta - b \int_{\Theta_1} f(\mathbf{y}|\theta)\pi_1(\theta)\,d\theta \right] d\mathbf{y}$$

$$= b + \int_{\mathrm{R}} \left[a\varpi_0(\mathbf{y}) - b\varpi_1(\mathbf{y}) \right] d\mathbf{y}.$$

The result follows from application of the definition of δ^* in expressions (17) and (18), since every point on which $a \cdot \varpi_0(\mathbf{y}) - b \cdot \varpi_1(\mathbf{y}) < 0$ is in R, but no point on which $a \cdot \varpi_0(\mathbf{y}) - b \cdot \varpi_1(\mathbf{y}) > 0$ is included. Therefore, δ^* minimizes the last term in the sum, and the first does not depend on the test. The result has been established. □

Regarding the assessment of the constants a and b, notice that it suffices to specify the ratio $\mathrm{r} := a/b$. This can be done in multiple ways: (i) by finding the *implicit a and b* of a carefully designed experiment, as in Example 1; (ii) by using a conventional table of ratio of evidences, like the Jeffreys evidence ratio scale table, and Kass and Raftery modification of it (see Tables 4 and 5 in the Appendix); or (iii) using a ratio of prior probabilities of H_0 times the loss incurred by false rejection of H_0 over the product of the prior probability of H_1 and the loss incurred by false acceptance of H_0. In symbols, calling L_0 the loss for false rejection of H_1 and L_1 the loss for false rejection of H_0:

$$\mathrm{r} = \frac{b}{a} = \frac{P(H_1) \cdot L_0}{P(H_0) \cdot L_1}.$$

To see this, notice that the risk function can be written as $R(\theta, \delta) = L_1 \alpha_\theta(\delta)$ if $\theta \in \Theta_0$, and as $R(\theta, \delta) = L_0 \beta_\theta(\delta)$ if $\theta \in \Theta_1$. Assuming a priori that the probability of the null hypothesis is $P(H_0)$, then the average (Bayesian) risk, taking expectations with respect to $(P(H_0), \pi_0)$ and $((1 - p(H_0)), \pi_1)$, we get the averaged risk

$$r(\delta) = P(H_0) \cdot L_1 \cdot \alpha(\delta) + \bigl(1 - P(H_0)\bigr) \cdot L_0 \cdot \beta(\delta), \tag{21}$$

and we see that the correspondence with expression (16) is: $a \mapsto P(H_0) \cdot L_1$ and $b \mapsto (1 - P(H_0)) \cdot L_0$, assuming that the loss is constant on each of the hypotheses.

The Rejection Region $\mathtt{R}$ in (18) takes two different shapes under interpretations 1 and 2.

- For interpretation 1, $\mathtt{R}$ is defined by

$$\frac{\varpi_0(\mathbf{y})}{\varpi_1(\mathbf{y})} < \frac{b}{a}. \tag{22}$$

  That is, the null hypothesis is rejected if the Bayes factor of H_0 over H_1 is small enough.
- For interpretation 2, $\mathtt{R}$ is defined as the region in which

$$\frac{\Pr(H_0 \cup H_0^* | \mathbf{y})}{\Pr(H_1 | \mathbf{y})} = \frac{\Pr(H_1^C | \mathbf{y})}{\Pr(H_1 | \mathbf{y})} < \frac{b}{a}, \tag{23}$$

where H_0 is the null hypothesis, H_0^* the indifference region, where it is not worthwhile to abandon the null because the gain from doing so is insufficient, and H_1 the alternative of practical significance. This assumes that the loss of rejecting H_0 under H_0 and H_0^* is the same, and that the loss for accepting H_0, both under H_0 and H_0^*, is zero.

We suggest that (23) is more reasonable than rejecting the null when, say,

$$\frac{\Pr(H_0|\mathbf{y})}{\Pr(H_1|\mathbf{y})} < \frac{1}{3},$$

an approach popular in medical statistics, since it may happen that, for example, $P(H_0|\mathbf{y}) = 0.1\varepsilon$ and $P(H_1|\mathbf{y}) = 0.9\varepsilon$, and ε can be minute, like $\varepsilon = 0.001$. If both posterior probabilities are minute, one should *not* abandon H_0 in favor of H_1.

Finally, for the simple and simplistic point masses at θ_0 and θ_1, the optimal rule becomes

$$\frac{f(\mathbf{y}|\theta_0)}{f(\mathbf{y}|\theta_1)} < \frac{b}{a}. \tag{24}$$

1.8 Relaxing the assumptions: "Well calibrated" priors

In the proof of Lemma 1, it was assumed that the weights were proper, that is $\int_{\Theta_j} \pi_j(\theta)\, d\theta = 1$. This may be seen as too heavy an assumption. Fortunately, the assumption can be relaxed, at least for weights that are improper but *well calibrated*. See Pericchi (2005) for a discussion of well calibrated priors. We give two illustrations of well calibrated priors:

1. Illustration 1: Let us consider the priors Jeffreys suggested for the Normal mean testing problem: $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ and the variances are unknown. The Jeffreys priors for this problem are:

   $$\pi_0^J(\sigma_0) = \frac{1}{\sigma_0}$$

   and

   $$\pi_1^J(\mu, \sigma_1) = \frac{1}{\sigma_1} \cdot \frac{1}{\pi\sigma_1(1 + \mu^2/\sigma_1^2)}.$$

   Notice that the priors are not proper. We define a training sample of minimal size $x(l)$ as a sub-sample of $\mathbf{x}$ such that both $\pi_0^J(\sigma_0|x(l))$ and $\pi_1^J(\mu, \sigma_1|x(l))$ are proper, that is, the priors integrate to 1, but any sub-sample of $x(l)$ will not. In this illustration case the minimal training size is one, that is $x(l) = x_l$. It turns out that Jeffreys priors are well calibrated (Pericchi (2005)), that is, for any x_l,

$$\int f(x_l|\sigma_0)\pi_0^J(\sigma_0)\,d\sigma_0 = \int f(x_l|\mu,\sigma_1)\pi_1^J(\mu,\sigma_1)\,d\mu\,d\sigma_1,$$

or $m_0(x_l) = m_1(x_l)$.

2. Illustration 2: Suppose one wishes to compare a Normal model with a Cauchy model, both with location μ and scale σ unknown. For location-scale models, the objective prior is usually chosen to be $\pi(\mu,\sigma) = 1/\sigma$. It turns out (see Berger et al. (1998)) that for any location-scale likelihood, the minimal training sample size is 2, that is, $x(l) = (x_{l_1}, x_{l_2})$. It follows that

$$\int \frac{1}{\sigma^3}\cdot f\left(\frac{x_{l_1}-\mu}{\sigma}\right) f\left(\frac{x_{l_2}-\mu}{\sigma}\right) d\mu\,d\sigma = \frac{1}{2|x_{l_2}-x_{l_1}|},$$

that is, the marginal for any two different data points is *the same* for *any* location-scale family, and so, if the prior is $1/\sigma$, is well calibrated between any location-scale family for the minimal training sample of two observations.

Corollary 1. *For priors that do not integrate to* 1, *but are well calibrated, Lemma* 1 *still holds.*

Proof. Take an arbitrary minimal training sample $x(l)$, so that the remaining sample is denoted by $x(-l)$. Now use the priors $\pi_0(\theta_0|x(l))$ and $\pi_1(\theta_1|x(l))$, and the corresponding likelihoods $f_0(x(-l)|\theta_0)$ and $f_1(x(-l)|\theta_1)$ in Lemma 1. Assuming we have a sample bigger than the minimal training sample, then Lemma 1 follows with the priors and likelihoods above. Now the result follows from the following identity, for well calibrated priors:

$$\frac{\int f_0(x(-l)|\theta_0)\pi_0(\theta_0|x(l))\,d\theta_0}{\int f_1(x(-l)|\theta_1)\pi_1(\theta_1|x(l))\,d\theta_1} = \frac{\int f_0(x|\theta_0)\pi_0(\theta_0)\,d\theta_0}{\int f_1(x|\theta_1)\pi_1(\theta_1)\,d\theta_1}.$$

This corollary substantially expands the applicability of Lemma 1 and highlights the usefulness of well calibrated priors. □

2 Two-sided alternatives

Example 2. Consider a univariate normal distribution with known variance σ_0^2 and the following hypotheses about the value of the mean θ:

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta \neq \theta_0.$$

2.1 Bayesian intrinsic prior

One "objective Bayesian" approach is the intrinsic prior approach. For this example, it turns out that the intrinsic prior is (Pericchi (2005), Berger and Pericchi

(1996)): $\pi_1(\theta) = N(\theta|\theta_0, 2\sigma_0^2)$ and π_0 is a Dirac delta at point θ_0. Calculation yields that the optimal test δ^* is

$$\text{Reject } H_0 \text{ if } \frac{f(\bar{y}|\theta_0)}{\varpi_1(\bar{y})} = \frac{N(\bar{y}|\theta_0, \sigma_0^2/n)}{N(\bar{y}|\theta_0, \sigma_0^2(2+1/n))} < \text{r}, \tag{25}$$

where n is the sample size.

2.2 Practical significance

Assume that the test is meant to detect a difference if $\theta_1 = \theta_0 \pm \Delta$. Then the simplistic prior weight is a Dirac delta centered at the two points $\theta_0 \pm \Delta$ with weight equal to $1/2$ on each point. The test now becomes

Reject H_0 if

$$\frac{f(\bar{y}|\theta_0)}{\varpi_1^*(\bar{y})} = \frac{N(\bar{y}|\theta_0, \sigma_0^2/n)}{(1/2)N(\bar{y}|\theta_0 - \Delta, \sigma_0^2/n) + (1/2)N(\bar{y}|\theta_0 + \Delta, \sigma_0^2/n)} < \text{r}, \tag{26}$$

which can be written as

$$\text{Reject } H_0 \text{ if } \exp\Big(-\frac{n\Delta}{2\sigma_0^2}[\Delta - 2(\bar{y} - \theta_0)]\Big) + \exp\Big(-\frac{n\Delta}{2\sigma_0^2}[\Delta + 2(\bar{y} - \theta_0)]\Big) > \frac{2}{\text{r}}.$$

This is a reasonable criterion that can be compared to the usual significance test, which is extremely biased against H_0 for sample sizes larger than those that were carefully chosen to achieve specified type I and type II error probabilities.

A more careful analysis of indifference regions using the same Intrinsic prior as above leads us to the following rejection region:

$$\frac{\Pr(H_{0,\Delta}|\mathbf{y})}{1 - \Pr(H_{0\Delta}|\mathbf{y})} < \frac{b}{a}, \tag{27}$$

where $H_{0,\Delta} = [\theta_0 - \Delta, \theta_0 + \Delta]$. This criterion is extremely simple, based on the ratio of two normal probabilities, and takes specific account of the indifference region. It can be verified that the tests (25), (26) and (27) are all consistent as the sample size n grows. That is, *both* type I and type II errors go to zero as the sample size grows.

The alternative testing paradigm enjoys several desirable properties, some of which we describe here.

3 Hypothesis testing under the new paradigm obeys the Likelihood Principle

One of the usual criticisms of significance testing is that it does not obey the Likelihood Principle, a principle that is of importance not only to Bayesians, given that it was actually enunciated and defended by eminent non-Bayesians. Loosely speaking, the Likelihood Principle establishes that if two likelihoods are proportional to each other, the information about the parameter vector θ is the same. The following example is eloquent.

3.1 The Lindley and Phillips (1976) example revisited

Example 3. It is desired to test whether a coin is balanced, because it is suspected that it is more prone to "heads."

$$H_0 : \theta = 1/2, \quad \text{vs} \quad H_1 : \theta > 1/2.$$

It is known that the number of Heads is $S = 9$ and the number of tails is $n - S = 3$. It is desired to conduct a test with $\alpha = 0.05$. However, significance testing (in its two current versions, based on p-values or fixed significance) cannot decide whether to accept or reject the null hypothesis. How was the sample size determined? Was it fixed beforehand? Or it was decided that the experiment would stop at the third occurrence of "tails"? The results are *not* the same for these two situations.

Suppose $n = 12$ was decided beforehand. In that case, we have a binomial likelihood:

$$f_{\mathrm{B}}(S|\theta) = \frac{12!}{9!3!}\theta^9(1-\theta)^3 = 220\theta^9(1-\theta)^3. \tag{28}$$

However, if the experiment was stopped at the third occurrence of "tails," then we have a negative binomial experiment, with likelihood function

$$f_{\mathrm{NB}}(S|\theta) = \frac{11!}{9!2!}\theta^9(1-\theta)^3 = 55\theta^9(1-\theta)^3. \tag{29}$$

That is, we have two proportional likelihood functions, so according to the Likelihood Principle, we should have the same inference. However, the observed p-values differ:

$$\alpha_{\mathrm{B}} = \Pr(S \geq 9|\theta = 0.5, \text{Binom}) = \sum_{S=9}^{12} f_{\mathrm{B}}(S|\theta = 0.5) = 0.073,$$

while

$$\alpha_{\mathrm{NB}} = \Pr(S \geq 9|\theta = 0.5, \text{NegBinom}) = \sum_{S=9}^{\infty} f_{\mathrm{NB}}(S|\theta = 0.5) = 0.0327.$$

Therefore, the result is considered statistically significant in the second scenario but not in the first.

Examples like these seem to have convinced many that frequentist hypothesis testing is *bound* to violate the Likelihood Principle. The good news, which we would guess is surprising to many, is that the violation of the Likelihood Principle can be avoided by using the weighting method, minimizing a weighted sum of (averaged) type I and type II error probabilities.

Corollary 2. *Testing by minimizing a weighted sum of errors automatically obeys the Likelihood Principle.*

Proof. From Lemma 1, the optimal test is

$$\text{Reject } H_0 \text{ if } \frac{\varpi_0(\mathbf{y})}{\varpi_1(\mathbf{y})} = \frac{\int_{\Theta_0} f(\mathbf{y}|\theta)\pi_0(\theta)\,d\theta}{\int_{\Theta_1} f(\mathbf{y}|\theta)\pi_1(\theta)\,d\theta} < \frac{b}{a},$$

and in the ratio on the left-hand side, the constant in the likelihood cancels out. □

Example 3 (Continued). In this example we assume the uniform prior on (0.5, 1): $\pi(\theta) = 2 \times 1_{(0.5,1)}(\theta)$. We assume this prior for simplicity, and although we do not think it is "optimal" in any sense, it is not unreasonable and does not influence the outcome heavily. The evidence ratio is easily found numerically.

$$\frac{f(S|\theta = 0.5)}{\int_{0.5}^{1} f(S|\theta)\cdot 2\,d\theta} = \frac{(1/2)^{12}}{(1 - \text{pbeta}(0.5|10,4) \times \text{Beta}(10,4) \times 2)} = 0.366,$$

where $\text{pbeta}(x|a,b)$ is a probability of obtaining a value between zero and x when drawing from a beta distribution with parameters a and b, and $\text{Beta}(a,b) = \frac{\Gamma(a)\cdot\Gamma(b)}{\Gamma(a+b)}$, $a > 0$, $b > 0$, is the beta function with parameters a and b.

Thus, according to the Jeffreys table of evidence ratios (see Table 4 in the Appendix), the ratio is less than 1 but greater that $1/\sqrt{10} = 0.32$, so there is mild evidence against H_0, which agrees with the modified table by Kass and Raftery (see Table 5 in the Appendix).

Procedures that depend on ratios of probabilities rather than tail probabilities are more realistic and more flexible.

4 When statistical significance meets practical significance

One of the most criticized points of the current significance testing approach is the lack of correspondence between practical significance and statistical significance. One such example is found in Freeman (1993).

4.1 Freeman's example

Example 4. Consider four hypothetical studies in which equal numbers of patients are given treatments A and B and are asked which of the two they prefer. The results are given in Table 1.

An objective (and proper) prior weight function for the parameter θ is the Jeffreys prior

$$\pi^J(\theta) = \frac{1}{\pi}\theta^{-1/2}(1-\theta)^{-1/2} \qquad \text{for } 0 < \theta < 1.$$

Table 1 *Freeman's example*

Number of patients receiving A and B	Number of patients preferring A : B	Percentage preferring A	Two-sided p-value	Evidence ratio
20	15 : 5	75.00	0.04	0.42
200	115 : 86	57.50	0.04	1.85
2000	1046 : 954	52.30	0.04	6.75
2,000,000	1,001,445 : 998,555	50.07	0.04	219.66

Computation yields the evidence ratio

$$\frac{f(s|\theta = 1/2)}{\varpi(s)} = \frac{\pi \cdot 0.5^N}{\text{Beta}(s + 1/2, N - s + 1/2)}.$$

The results are shown in the fifth column of Table 1, and are consistent with the conclusions put forward by Freeman on intuitive grounds: the first trial is too small to permit reliable conclusions, while the last trial would be considered evidence *for*, rather than against, equivalence, because from any practical perspective, the two treatments are equivalent. In fact, the ratio gives, according to Table 4 in the Appendix, "decisive," or "grade 5" evidence in favor of the null hypothesis for a sample of two million patients, or "very strong" in the modified Table 5.

5 Is there extrasensory perception (ESP), or are there just extremely large numbers?

In one of their books, Wonnacott and Wonnacott declared: "Do you want to reject a hypothesis? Take a large enough sample!"

5.1 ESP example

The so-called "extrasensory experiment" found in Good (1992) is an excellent example of how p-values are increasingly misleading with extremely large samples.

Example 5 (Extrasensory perception—ESP or not ESP?). Here the question is whether a "gifted" individual can change the proportion of 0's and 1's emitted with "perfectly" balanced proportions. The null hypothesis is "no change in the proportion" against the alternative hypothesis "some change." That is, $H_0 : \theta = 1/2$ vs $H_1 : \theta \neq 1/2$. We have a huge sample: $N = 104{,}490{,}000$; Successes: $S = 52{,}263{,}471$; Ratio: $S/N = 0.5001768$.

The p-value against the null is minute: pval $= 0.0003$, leading to a compelling rejection of H_0.

On the other hand, there exists an objective (proper) prior that can be used as a weight function here. Specifically, the Jeffreys prior $\pi^J(\theta) = \frac{1}{\pi \times \sqrt{\theta(1-\theta)}}$. Then

Table 2 *Table for different values of* Δ

Δ	0.0002	0.0003	0.0004	0.0005
r	2.15	169	397,877	51,369,319,698

the Bayes Factor, or *evidence ratio*, is

$$\begin{aligned} B_{H_0,H_1} &= \frac{f(\text{data}|\theta=1/2)}{\int f(\text{data}|\theta)\pi^J(\theta)\,d\theta} \\ &= \frac{\pi\cdot(1/2)^N}{\text{Beta}(S+0.5,N-S+0.5)} \\ &= B_{H_0,H_1} = \exp(2.93) = 18.7. \end{aligned}$$

This is strong support for the null hypothesis. The Bayes Factor is equal to 12 for the uniform prior, still strong support for the null.

Taking the second route, setting the a priori points of *practical significance*, we may agree that a value of θ that is, say, Δ above or below the value given in the null hypothesis (here 0.5) can be acceptable as practical significance.

The criterion now reads

$$\frac{\Pr(0.5-\Delta<\theta<0.5+\Delta|\text{data})}{(1-\Pr(0.5-\Delta<\theta<0.5+\Delta|\text{data}))} < \mathrm{r} = \frac{b}{a}. \tag{30}$$

In Table 2, we present the values of r for values of Δ running from 0.0002 (or 0.02%) to 0.0005 (or 0.05%) in increments of 0.0001 (or 0.01%). Even for the smallest Δ considered, the ratio of likelihoods is greater than 1, and for $\Delta = 0.0005$, the ratio indicates compelling evidence against H_1. This is in sharp contrast to the p-value of 0.0003 we obtained above for the same example.

6 A general inequality: The discrepancy between tests at fixed significance levels and tests that minimize a weighted sum of error probabilities is general

Even though the discrepancies between tests at fixed significance levels and those based on minimizing a weighted sum of error probabilities have been illustrated here with specific examples, this phenomenon is more general, as shown in the following result (see also Birnbaum (1969); Dempster (1997), for related results).

Lemma 2. *For the optimal test* δ^* *of Lemma* 1, *it turns out that*:

$$\frac{\alpha(\delta^*)}{1-\beta(\delta^*)} \le \frac{b}{a}.$$

Proof. First, notice that the rejection region for the test δ^* can be written as: $\mathtt{R} = \{\mathbf{y} : \frac{\varpi_1(y)\cdot b}{\varpi_0(y)\cdot a} \geq 1\}$. Denote by $\mathtt{S}$ the set $\mathtt{S} \subset \mathtt{R}$ where $\varpi_0(y) > 0$. Then

$$\begin{aligned}\alpha(\delta^*) &= \int_{\mathtt{R}}\int_{\Theta_0} f_0(y|\theta)\pi_0(\theta)\,d\theta\,dy = \int_{\mathtt{S}} \varpi_0(y)\,dy \\ &\leq \int_{\mathtt{S}} \frac{\varpi_1(y)b}{\varpi_0(y)a}\varpi_0(y)\,dy \\ &= \frac{b}{a}\int_{\mathtt{S}} \varpi_1(y)\,dy \leq \frac{b}{a}\int_{\mathtt{R}} \varpi_1(y)\,dy = \frac{b}{a}(1-\beta(\delta^*)).\end{aligned}$$

□

Corollary 3.

$$\alpha(\delta^*) \leq \frac{b}{a}.$$

Thus, for example, if $b/a = 20$, $\varpi_1(y^*)/\varpi_0(y^*)$ is considered equivalent to $\alpha(\delta^*) = 0.05$, and if the power is 0.8, then

$$\frac{\varpi_1(y^*)}{\varpi_0(y^*)} \leq \frac{1-\beta}{\alpha} = \frac{0.8}{0.05} = 16,$$

and by Corollary 3, we have $\frac{\varpi_1(y^*)}{\varpi_0(y^*)} \leq 20$, so weighted tests rejects less often.

7 A formula for decreasing the p-value (or α) as the sample size increases

The previous analysis of Example 2 and other related examples suggests an interesting method to "decrease α with n" in such a way to give an *"asymptotically equivalent"* result, as in (25). In other words, can we find a formula for the level of the test as a function of the sample size such that it produces approximately the same decisions as rejecting the null when the data obey (25)? In Perez and Pericchi (2014), the following asymptotic approximation is obtained in the simplest case of one parameter, as in Example 2. It is called the *square root* $n \times \log(n)$ *formula*:

$$\alpha(n) = \frac{\alpha * \sqrt{n_0 \times (\log(n_0) + \chi^2_\alpha(1))}}{\sqrt{n \times (\log(n) + \chi^2_\alpha(1))}}, \tag{31}$$

where n_0 is the sample size of a well designed experiment (as in Example 1 $n_0 = 20$), α the initial significance level designed for n_0 and $\chi^2_\alpha(1)$ is the chi-squared quantile with one degree of freedom.

Versions of this approximate rule have appeared in Cox and Hinkley (1974) and in Good (1992), both with unfortunate typographical errors. The *square root of* $n \times \log(n)$ *formula* above, gives clear guidance on how to decrease the scale of p-values with the sample size. The value n_0 is the "origin" of the "planned" experiment (e.g., $n_0 = 20$ in Example 1). See Table 3 for specific values.

Table 3 *Table of adaptive significance level*

Sample size	α
20	0.05
50	0.03
100	0.02
250	0.012
500	0.008
1000	0.006

7.1 Table for decreasing the significance level

Assume $n_0 = 20$ and $\alpha = 0.05$.

Notice that the *equivalent* α in Example 1 (simple-vs-simple hypotheses) given by equation (5) and the $\sqrt{n \cdot (\log(n) + \chi^2_\alpha(1))}$ formula (for two-sided hypotheses) have different speeds, (5) being much faster than the "square root of $n\log(n)$" formula.

8 Conclusions

What emerges in the implementation of the *weighted* approach to testing statistical hypotheses is a practical implementation of the ideas of decision theory, with a bridge between Bayesian and Frequentist philosophies. This implementation, we argue along with DeGroot, is superior to the two implementations dominant in practice: (i) The use of p-values with fixed cut points, like the ubiquitous α-set $\{0.1, 0.05, 0.01\}$; and (ii) the use of fixed type I error probabilities in the α-set, and then choosing a criterion to minimize type II error.

By doing (i) or (ii), a statistician is in danger of having a minute effective type II error probability and a relatively enormous type I error probability. Furthermore, fixing the type I error probability leads to inconsistency "by design": *no matter how informative the experiment is, one forces the method to have a type I error probability no smaller than one of the numbers in the α-set.* In contrast, minimizing the weighted sum or error probabilities, a method that is more balanced between the two error types emerges, and consistency flows as an automatic consequence: *by minimizing the sum of error probabilities as evidence grows, one is letting both error probabilities converge to zero, so the method is consistent.* As virtues of the approach we have a general theory of optimal testing that obeys the Likelihood Principle, reconciles the disagreement between schools of statistics, and is more in line with the demands of the scientific method.

Finally, to achieve the benefits of the general theory, is not necessary to assume fully proper priors: well calibrated improper priors suffices.

Table 4 *Jeffreys scale of evidence*

Grade 0	$r \geq 1$	Null supported
Grade 1	$1 > r > 10^{-1/2}$	Mild evidence against H_0
Grade 2	$10^{-1/2} > r > 10^{-1}$	Substantial evidence against H_0
Grade 4	$10^{-1} > r > 10^{-3/2}$	Strong evidence against H_0
Grade 5	$10^{-3/2} > r > 10^{-2}$	Very strong evidence against H_0
Grade 6	$10^{-2} > r$	Decisive evidence against H_0

Table 5 *Kass–Raftery scale of evidence*

$-2\log_e(r)$	$\frac{1}{r}$	Evidence against the null H_0
0 to 2	1 to 3	Mild
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
Bigger than 10	Bigger than 150	Very strong

Appendix: Jeffreys table of evidence ratios and, Kass and Raftery's (1995) modification of it, respectively

Here $r = b/a$. See Tables 4 and 5.

Acknowledgments

We are grateful to the Brazilian institutions FAPESP and CNPq that have supported visits of the authors to each other. Luis Pericchi was also supported by NIH Grant: P20-RR016470. The following colleagues read the paper and gave us useful input: P. M. Andrade, M. Gannon, A. Grieves, K. Guo, M. E. Perez and S. Wechsler.

References

Bayarri, M. J. and Berger, J. O. (2004). The interplay between Bayesian and frequentist analysis. *Statist. Sci.* **19**, 58–80. MR2082147

Berger, J. O. (2008). A comparison of testing methodologies. In *The Proceedings of PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, *June 2007*, *CERN 2008-001*, 8–19. Geneve: CERN.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109–122. MR1394065

Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankyā Ser. A* **60**, 307–321. MR1718789

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian model selection. Introduction and comparisons. In *Model Selection* (P. Lahiri, ed.). *Lecture Notes Monogr. Ser.* **68**, 135–207. Beachwood, OH: IMS. MR2000753

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco, CA: Holden-Day, Inc. MR0443141

Birnbaum, A. (1969). Concepts of Statistical Evidence. In *Philosophy, Science and Methods: Essays in Honor of Ernest Nagel* (S. Morgenbesser, P. Suppes and M. White, eds.). New York: St. Martin's Press.

Cox, D. R. and Hinkley, D. V. (1974). *Concepts of Statistical Evidence*. *Theoretical Statistics*. London: Chapman and Hall. MR0370837

DeGroot, M. (1975). *Probability and Statistics*, 2nd ed. Reading, MA: Addison-Wesley. MR0373075

Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Stat. Comput.* **7**, 247–252. MR0408052

Freeman, P. (1993). The role of p-values in analyzing trial results. *Stat. Med.* **12**, 1443–1452.

Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *J. Amer. Statist. Assoc.* **87**, 597–606. MR1185188

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine* **2**, e124. MR2216666

Jeffreys, H. (1935). Some test of significance, treated by the theory of probability. *Math. Proc. Cambridge Philos. Soc.* **31**, 203–222.

Jeffreys, H. (1961). *The Theory of Probability*, 3rd ed. Oxford: Clarendon Press. MR0187257

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 791.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192. MR0087273

Lindley, D. V. and Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.* **30**, 112–119. MR0445656

Pereira, C. A. B. (1985). Testing hypotheses of diferent dimensions: Bayesian view and classical interpretation (in Portuguese). Professor thesis, The Institute of Mathematics and Statistics, Univ. de São Paulo.

Pereira, C. A. B. and Wechsler, S. (1993). On the concept of p-value. *Braz. J. Probab. Stat.* **7**, 159–177. MR1323121

Perez, M. E. and Pericchi, L. R. (2014). Changing statistical significance as the amount of information changes: the adaptive α significance level. *Statist. Probab. Lett.* **85**, 20–24. MR3157877

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. *Handbook of Statist.* **25**, 115–149. MR2490524

Siegfried, T. (2010). Odds are, it's wrong science fails to face the shortcomings of statistics. *Science News* **177**, 26–29.

Varuzza, L. and Pereira, C. A. B. (2010). Significance test for comparing digital gene expression profiles: Partial likelihood application. *Chil. J. Stat.* **1**, 91–102. MR2756086

Department of Mathematics and Center
for Biostatistics and Bioinformatics
University of Puerto Rico
Rio Piedras, San Juan
Puerto Rico
E-mail: luis.pericchi@upr.edu

Instituto de Matemática e Estatística
Universidade de São Paulo
São Paulo, SP
Brasil
E-mail: cpereira@ime.usp.br