Psychonomic Bulletin & Review
2007, 14 (5), 779-804

THEORETICAL AND REVIEW ARTICLES

A practical solution to the pervasive problems of *p* values

ERIC-JAN WAGENMAKERS
University of Amsterdam, Amsterdam, The Netherlands

In the field of psychology, the practice of *p* value null-hypothesis testing is as widespread as ever. Despite this popularity, or perhaps because of it, most psychologists are not aware of the statistical peculiarities of the *p* value procedure. In particular, *p* values are based on data that were never observed, and these hypothetical data are themselves influenced by subjective intentions. Moreover, *p* values do not quantify statistical evidence. This article reviews these *p* value problems and illustrates each problem with concrete examples. The three problems are familiar to statisticians but may be new to psychologists. A practical solution to these *p* value problems is to adopt a model selection perspective and use the Bayesian information criterion (BIC) for statistical inference (Raftery, 1995). The BIC provides an approximation to a Bayesian hypothesis test, does not require the specification of priors, and can be easily calculated from SPSS output.

The primary goal of this article is to promote awareness of the various statistical problems associated with the use of *p* value null-hypothesis significance testing (NHST). Making no claim of completeness, I review three problems with NHST, briefly explaining their causes and consequences (see Karabatsos, 2006). The discussion of each problem is accompanied by concrete examples and references to the statistical literature.

In the psychological literature, the pros and cons of NHST have been, and continue to be, hotly debated (see, e.g., Cohen, 1994; Cortina & Dunlap, 1997; Cumming, 2007; Dixon, 2003; Frick, 1996; Gigerenzer, 1993, 1998; Hagen, 1997; Killeen, 2005a, 2006; M. D. Lee & Wagenmakers, 2005; Loftus, 1996, 2002; Nickerson, 2000; Schmidt, 1996; Trafimow, 2003; Wagenmakers & Grünwald, 2006; Wainer, 1999). The issues that have dominated the NHST discussion in the psychological literature are that (1) NHST tempts the user into confusing the probability of the hypothesis given the data with the probability of the data given the hypothesis; (2) $\alpha = .05$ is an arbitrary criterion for significance; and (3) in real-world applications, the null hypothesis is never exactly true, and will therefore always be rejected as the number of observations grows large.

In the statistical literature, the pros and cons of NHST are also the topic of an ongoing dispute (e.g., Berger & Wolpert, 1988; O'Hagan & Forster, 2004; Royall, 1997; Sellke, Bayarri, & Berger, 2001; Stuart, Ord, & Arnold, 1999).[1] A comparison of these two literatures shows that in psychology, the NHST discussion has focused mostly on problems of interpretation, whereas in statistics, the NHST discussion has focused mostly on problems of formal construction. The statistical perspective on the problems associated with NHST is therefore fundamentally different from the psychological perspective. In this article, the goal is to explain NHST and its problems from a statistical perspective. Many psychologists are oblivious to certain statistical problems associated with NHST, and the examples below show that this ignorance can have important ramifications.

In this article, I will show that an NHST *p* value depends on data that were never observed: The *p* value is a tail-area integral, and this integral is effectively over data that are not observed but only hypothesized. The probability of these hypothesized data depends crucially on the possibly unknown subjective intentions of the researcher who carried out the experiment. If these intentions were to be ignored, a user of NHST could always obtain a significant result through optional stopping (i.e., analyzing the data as they accumulate and stopping the experiment whenever the *p* value reaches some desired significance level). In the context of NHST, it is therefore necessary to know the subjective intention with which an experiment was carried out. This key requirement is unattainable in a practical sense, and arguably undesirable in a philosophical sense. In addition, I will review a proof that the NHST *p* value does not measure statistical evidence. In order for the *p* value to qualify as a measure of statistical evidence, a minimum requirement is that identical *p* values convey identical levels of evidence, irrespective

E.-J. Wagenmakers, ej.wagenmakers@gmail.com

of sample size. This minimum requirement is, however, not met: Comparison to a very general Bayesian analysis shows that p values overestimate the evidence against the null hypothesis, a tendency that increases with the number of observations. This means that $p = .01$ for an experiment with 10 subjects provides more evidence against the null hypothesis than $p = .01$ for an experiment with, say, 300 subjects.

The second goal of this article is to propose a principled yet practical alternative to NHST p values. I will argue that the Bayesian paradigm is the most principled alternative to NHST: Bayesian inference does not depend on the intention with which the data were collected, and the Bayesian hypothesis test allows for a rigorous quantification of statistical evidence. Unfortunately, Bayesian methods can be computationally and mathematically intensive, and this may limit the level of practical application in experimental psychology. The solution proposed here is to use the Bayesian information criterion (BIC; Raftery, 1995). The BIC is still principled, since it approximates a full-fledged Bayesian hypothesis test (Kass & Wasserman, 1995). The main advantage of the BIC is that it is also practical; for instance, it can be easily calculated from SPSS output (see Glover & Dixon, 2004).

Several disclaimers and clarifications are in order. First, this article reviews statistical problems with p values. To bring out these problems clearly, it necessarily focuses on situations for which NHST and Bayesian inference disagree. NHST may yield entirely plausible inferences in many situations, particularly when the data satisfy the interocular trauma test. Of course, when the data hit you right between the eyes, there is little need for a statistical test anyway.

Second, it should be acknowledged that an experienced statistician who analyzes data with thought and carefulness may use NHST and still draw sensible conclusions; the key is not to blindly apply the NHST machinery and base conclusions solely on the resulting p value. Within the context of NHST, sensible conclusions are based on an intuitive weighting of such factors as p values, power, effect size, number of observations, findings from previous related work, guidance from substantive theory, and the performance of other models. Nevertheless, it cannot be denied that the field of experimental psychology has a p value fixation, since for single experiments the p value is the predominant measure used to separate the experimental wheat from the chaff.

Third, the existence of a phenomenon is usually determined not by a single experiment in a single lab, but by independent replication of the phenomenon in other labs. This might lead one to wonder whether the subtleties of statistical inference are of any relevance for the overall progress of the field, since only events that are replicable and pass the interocular trauma test will ultimately survive. I fully agree that independent replication is the ultimate arbiter, one that overrules any statistical considerations based on single experiments. But this argument does not excuse the researcher from trying to calculate a proper measure of evidence for the experiment at hand, and it certainly does not sanction sloppy or inappropriate data analysis for single experiments.

On a related point, Peter Killeen has proposed a new statistic, p_{rep}, that calculates the probability of replicating an effect (Killeen, 2005a, 2005b, 2006). The use of this statistic is officially encouraged by *Psychological Science*, and a decision-theoretic extension of p_{rep} has recently been published in *Psychonomic Bulletin & Review*. Although the general philosophy behind the p_{rep} statistic is certainly worthwhile, and the interpretation of p_{rep} is arguably clearer than that of the NHST p value, the p_{rep} statistic can be obtained from the NHST p value by a simple transformation. As such, p_{rep} inherits all of the p value problems that are discussed in this article.

The outline of this article is as follows. The first section briefly describes the main ideas that underlie NHST. The second section explains why p values depend on data that were never observed. The third section explains why p values depend on possibly unknown subjective intentions. This section also describes the effects of optional stopping. The fourth section suggests that p values may not quantify statistical evidence. This discussion requires an introduction to the Bayesian method of inference (i.e., Bayesian parameter estimation, Bayesian hypothesis testing, and the pros and cons of priors), presented in the fifth section. The sixth section presents a comparison of the Bayesian and NHST methods through a test of the "p postulate" introduced in Section 4, leading to an interim discussion of the problems associated with p values. The remainder of the article is then concerned with finding principled and practical alternatives to p value NHST. The seventh section lays the groundwork for such alternatives, first by outlining desiderata, then by evaluating the extent to which various candidate methodologies satisfy them. The eighth section outlines the BIC as an practical approximation to the principled Bayesian hypothesis test, and the ninth section concludes.

NULL-HYPOTHESIS SIGNIFICANCE TESTING

Throughout this article, the focus is on the version of p value NHST advocated by Sir Ronald Fisher (e.g., Fisher, 1935a, 1935b, 1958). Fisherian NHST considers the extent to which the data contradict a default hypothesis (i.e., the null hypothesis). A Fisherian p value is thought to measure the *strength of evidence* against the null hypothesis. Small p values (e.g., $p = .0001$), for instance, constitute more evidence against the null hypothesis than larger p values (e.g., $p = .04$). Fisherian methodology allows its user to learn about a specific hypothesis from a single experiment.

Neyman and Pearson developed an alternative procedure involving Type I and Type II error rates (e.g., Neyman & Pearson, 1933). Their procedure requires the specification of an alternative hypothesis and pertains to actions rather than evidence, and as such specifically denies that one can learn about a particular hypothesis from conducting an experiment: "We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis" (Neyman & Pearson, 1933, pp. 290–291).

Despite fierce opposition from both Fisher and Neyman, current practice has become an anonymous amalgamation of the two incompatible procedures (Berger, 2003; Christensen, 2005; Goodman, 1993; Hubbard & Bayarri, 2003; Royall, 1997). The focus here is on Fisherian methodology because, arguably, it is philosophically closer to what experimental psychologists believe they are doing when they analyze their data (i.e., learning about their hypotheses).

Before discussing three pervasive problems with Fisherian NHST, it is useful to first describe NHST and establish some notation. For concreteness, assume that I ask you 12 factual true–false questions about NHST. These questions could be, for example:

1. A p value depends on the (possibly unknown) intentions of the researcher who performs the experiment. True or false?
2. A p value smaller than .05 always indicates that the data are more probable under the alternative hypothesis than under the null hypothesis. True or false?

.
.
.

12. Given unlimited time, money, and patience, it is possible to obtain arbitrarily low p values (e.g., $p = .001$ or $p = .0000001$), even if no data are ever discarded and the null hypothesis is exactly true. True or false?

Now assume that you answer 9 of these 12 questions correctly and that the observed data are as follows: $x = \{\text{C, C, C, E, E, C, C, C, C, C, C, E}\}$, where "C" and "E" indicate a correct response and an incorrect response, respectively. Suppose I want to know whether your performance can be explained purely in terms of random guessing.

In Fisherian p value NHST, one proceeds by specifying a test statistic $t(\cdot)$ that summarizes the data x in a single number, $t(x)$. In the example above, the observed data x are reduced to the sum of correct responses, $t(x) = 9$.

The choice of this test statistic is motivated by the statistical model one assumes when analyzing the data. In the present situation, most people would use the binomial model, given by

$$\Pr[t(x) = s \mid \theta] = \binom{n}{s}\theta^s(1-\theta)^{n-s}, \tag{1}$$

where s is the sum of correct responses, n is the number of questions, and θ is a parameter that reflects the probability of answering any one question correctly, $\theta \in [0, 1]$. The model assumes that every question is equally difficult.

The next step is to define a null hypothesis. In this example, the null hypothesis that we set out to test is "random guessing," and in the binomial model this corresponds to $\theta = 1/2$. Plugging in $\theta = 1/2$ and $n = 12$ reduces Equation 1 to

$$\Pr\left[t(x) = s \mid \theta = \frac{1}{2}\right] = \binom{12}{s}\left(\frac{1}{2}\right)^{12}. \tag{2}$$

Under the null hypothesis H_0 (i.e., $\theta = 1/2$), the probability of $t(x) = 9$ is $\Pr[t(x) = 9 \mid \theta = 1/2] \approx .054$. Thus, the observed data do not seem very probable under the null hypothesis of random guessing. Unfortunately, however, the fact that the test statistic for the observed data has a low probability under the null hypothesis does not necessarily constitute any evidence against this null hypothesis. Suppose I asked you 1,000 questions, and you answered exactly half of these correctly; this result should maximally favor the null hypothesis. Yet when we calculate

$$\Pr\left[t(x) = 500 \mid \theta = \frac{1}{2}\right] = \binom{1{,}000}{500}\left(\frac{1}{2}\right)^{1{,}000},$$

we find that this yields about .025, a relatively low probability. More generally, as n increases, the probability of any particular $t(x)$ decreases; thus, conclusions based solely on the observed $t(x)$ are heavily influenced by n.

The solution to this problem is to consider what is known as the *sampling distribution* of the test statistic, given that H_0 holds. To obtain this sampling distribution, assume that H_0 holds exactly, and imagine that the experiment under consideration is repeated very many times under identical circumstances. Denote the entire set of replications $\mathbf{x}^{\text{rep}} \mid H_0 = \{x_1^{\text{rep}} \mid H_0, x_2^{\text{rep}} \mid H_0, \ldots, x_m^{\text{rep}} \mid H_0\}$. Next, for each hypothetical data set generated under H_0, calculate the value of the test statistic as was done for the observed data: $\mathbf{t}(\mathbf{x}^{\text{rep}} \mid H_0) = \{t_1(x_1^{\text{rep}} \mid H_0), t_2(x_2^{\text{rep}} \mid H_0), \ldots, t_m(x_m^{\text{rep}} \mid H_0)\}$. The sampling distribution $\mathbf{t}(\mathbf{x}^{\text{rep}} \mid H_0)$ provides an indication of what values of $t(x)$ can be expected in the event that H_0 is true.

The ubiquitous p value is obtained by considering $\Pr[t(x) \mid H_0]$, as well as that part of the sampling distribution $\mathbf{t}(\mathbf{x}^{\text{rep}} \mid H_0)$ more extreme than the observed $t(x)$. Specifically, the p value is calculated as $p = \Pr[\mathbf{t}(\mathbf{x}^{\text{rep}} \mid H_0) \geq t(x)]$. The entire process is illustrated in Figure 1. Returning to our example, we have already seen that $\Pr[t(x) = 9 \mid \theta = 1/2] \approx .054$. To calculate the p value, we also need the probabilities of values for the test statistic more extreme than the $t(x)$ that was observed. We could follow the general scheme of Figure 1 and obtain the sampling distribution $\mathbf{t}(\mathbf{x}^{\text{rep}} \mid H_0)$ by simulation. However, in simple models, the sampling distribution of popular test statistics is often known in analytic form. Here, of course, we can use Equation 2, from which it follows that $t(10) \approx .016$, $t(11) \approx .003$, and $t(12) \approx .0002$. The one-sided p value, which corresponds to the hypothesis $H_1 : \theta > 1/2$, is given by

$$p_{(\text{1-sided})} = \sum_{x_i=9}^{12} t(x_i) \approx .073.$$

To calculate the two-sided p value, corresponding to the hypothesis $H_1 : \theta \neq 1/2$, the more extreme values at the low end of the sampling distribution also need to be taken into account:

$$p_{(\text{2-sided})} = \sum_{x_i=0}^{3} t(x_i) + \sum_{x_i=9}^{12} t(x_i) \approx .073 + .073 = 0.146.$$

A low p value is considered evidence against H_0. The logic is that a low p value indicates that either the null hypothesis is false or a rare event has occurred.

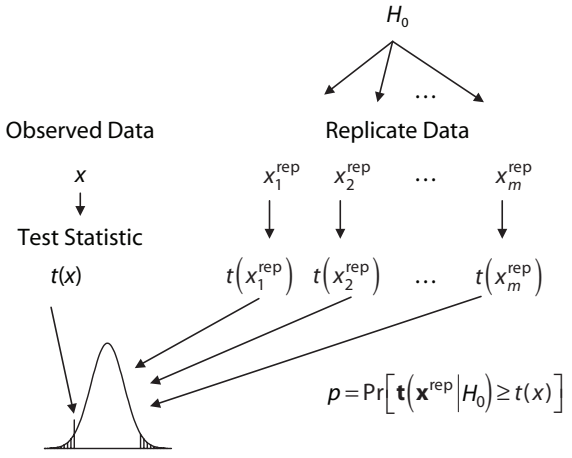

Figure 1. A schematic overview of *p* value statistical null-hypothesis testing. The distribution of a test statistic is constructed from replicated data sets generated under the null hypothesis. The two-sided *p* value is equal to the sum of the shaded areas on either side of the distribution; for these areas, the value of the test statistic for the replicated data sets is at least as extreme as the value of the test statistic for the observed data.

In sum, a *p* value is obtained from the distribution of a test statistic over hypothetical replications (i.e., the sampling distribution). The *p* value is the sum [or, for continuous $t(\cdot)$, the integral] over values of the test statistic that are at least as extreme as the one that is actually observed.

PROBLEM 1
p Values Depend on Data That Were Never Observed

As was explained in the previous section, the *p* value is not just based on the test statistic for the observed data, $t(x) \mid H_0$, but also on hypothetical data that were contemplated yet never observed. These hypothetical data are data expected under H_0, without which it is impossible to construct the sampling distribution of the test statistic $\mathbf{t}(\mathbf{x}^{rep} \mid H_0)$. At first, the concern over dependence on hypothetical data may appear to be a minor quibble. Consider, however, the following examples.

Example 1. Hypothetical events affect the *p* value (Barnard, 1947; Berger & Wolpert, 1988, p. 106; D. R. Cox, 1958, p. 368). Assume a variable x that can take on six discrete values, $x \in \{1, 2, \ldots, 6\}$. For convenience, take x to be the test statistic, $t(x) = x$. Furthermore, assume that we observe $x = 5$ and wish to calculate a one-sided *p* value. Table 1 shows that under the sampling distribution given by $f(x)$, the *p* value is calculated as $p = f(x = 5) + f(x = 6) = .04$. Under the sampling distribution given by $g(x)$, however, the same calculation yields a different *p* value: $p = g(x = 5) + g(x = 6) = .06$. This discrepancy occurs because $f(x)$ and $g(x)$ assign a different probability to the event $x = 6$. Note, however, that $x = 6$ has not been observed and is a purely hypothetical event. The only datum that has actually been observed is $x = 5$. The observed $x = 5$ is equally likely to occur under $f(x)$ and $g(x)$—isn't it odd that our inference should differ? Sir Harold Jeffreys summarized the situation as follows: "*What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred*. This seems a remarkable procedure" (Jeffreys, 1961, p. 385). The following example further elaborates this key point.

Example 2. The volt-meter (Pratt, 1962). A famous example of the effect of hypothetical events on NHST was given by Pratt (1962) in his discussion of Birnbaum (1962).[2] Rather than summarize or paraphrase Pratt's well-written example (see Berger & Wolpert, 1988, pp. 91–92; Stuart et al., 1999, p. 431), it is given here in its entirety:

> An engineer draws a random sample of electron tubes and measures the plate voltage under certain conditions with a very accurate volt-meter, accurate enough so that measurement error is negligible compared with the variability of the tubes. A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary normal analysis, giving a confidence interval for the true mean. Later he visits the engineer's laboratory, and notices that the volt meter used reads only as far as 100, so the population appears to be "censored." This necessitates a new analysis, if the statistician is orthodox. However, the engineer says he has another meter, equally accurate and reading to 1000 volts, which he would have used if any voltage had been over 100. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all. But the next day the engineer telephones and says: "I just discovered my high-range volt-meter was not working the day I did the experiment you analyzed for me." The statistician ascertains that the engineer would not have held up the experiment until the meter was fixed, and informs him that a new analysis will be required. The engineer is astounded. He says: "But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking me about my oscilloscope."
>
> I agree with the engineer. If the sample has voltages under 100, it doesn't matter whether the upper limit of the meter is 100, 1000, or 1 million. The sample provides the same information in any case. And

Table 1
Two Different Sampling Distributions, $f(x)$ and $g(x)$, Lead to Two Different *p* Values for $x = 5$

Distribution	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$x = 6$
$f(x) \mid H_0$	.04	.30	.31	.31	.03	.01
$g(x) \mid H_0$	.04	.30	.30	.30	.03	.03

Note—$x = 5$ is equally likely under $f(x)$ and $g(x)$. See the text for further details.

this is true whether the end-product of the analysis is an evidential interpretation, a working conclusion, a decision, or an action. (Pratt, 1962, pp. 314–315)

More specifically, consider the case of a sample $x_1, x_2, \ldots, x_n$ from a variable X that is exponentially distributed, with scale parameter θ. Under normal circumstances, the expected value of X, $E(X)$, equals θ, so that the sample mean $\bar{X}$ is an *unbiased* estimator of θ. However, when the measurement apparatus cannot record observations greater than some value c, the data are said to be censored, and as a result $\bar{X}$ is now a *biased* estimator of θ. Because values higher than c are set equal to c, $\bar{X}$ will underestimate the true value of θ. For this censored setup, the expected value of X is equal to

$$E(X) = \theta[1 - \exp(-c/\theta)]. \quad (3)$$

Note that $E(X)$ is now a proportion of θ: When $c \to \infty$ (i.e., effectively no censoring), $E(X) \to \theta$, and when $c \to 0$ (i.e., complete censoring), $E(X) \to 0$.

Because Equation 3 leads to unbiased results, a traditional statistician may well recommend its use for the estimation of θ, *even when no observations are actually censored* (Stuart et al., 1999, p. 431; for a critique of the concept of "unbiased estimators," see Jaynes, 2003, pp. 511–518). When replicate data sets are generated according to some null hypothesis, such as $\theta = 1$, the resulting sampling distribution depends on whether or not the data are censored; hence, censoring also affects the p value. Note that censoring has an effect on two-sided p values regardless of whether the observed data were actually censored, since censoring affects the shape of the sampling distribution.

In the statistical literature, the fact that p values depend on data that were never observed is considered a violation of the conditionality principle (see, e.g., Berger & Berry, 1988b; Berger & Wolpert, 1988; D. R. Cox, 1958; A. W. F. Edwards, 1992; Stuart et al., 1999). This principle is illustrated in the next example.

Example 3. The conditionality principle (see Berger & Wolpert, 1988, p. 6; Cornfield, 1969; D. R. Cox, 1958). Two psychologists, Mark and René, decide to collaborate to perform a single lexical decision experiment. Mark insists on 40 trials per condition, whereas René firmly believes that 20 trials are sufficient. After some fruitless deliberation, they decide to toss a coin for it. If the coin lands heads, Mark's experiment will be carried out, whereas René's experiment will be carried out if the coin lands tails. The coin is tossed and lands heads. Mark's experiment is carried out. Now, should subsequent statistical inference in this situation depend on the fact that René's experiment might have been carried out, had the coin landed tails?

The conditionality principle states that statistical conclusions should only be based on data that have actually been observed. Hence, the fact that René's experiment might have been carried out, but was not, is entirely irrelevant. In other words, the conditionality principle states that inference should proceed conditional on the observed data.[3]

In contrast, the situation is not so clear for NHST. As was illustrated earlier, the sampling distribution of the test statistic is constructed by considering hypothetical replications of the experiment. Does this include the coin toss? Indeed, in a hypothetical replication, the coin might have landed tails, and suddenly René's never-performed experiment has become relevant to statistical inference. This would be awkward, since most rational people would agree that experiments that were never carried out can safely be ignored. A detailed discussion of the conditionality principle can be found in Berger and Wolpert (1988).[4]

PROBLEM 2
p Values Depend on Possibly Unknown Subjective Intentions

As illustrated above, the p value depends on data that were never observed. These hypothetical data in turn depend on the sampling plan of the researcher. In particular, the same data may yield quite different p values, depending on the intention with which the experiment was carried out. The following examples underscore the problem.

Example 4. Binomial versus negative binomial sampling (e.g., Berger & Berry, 1988b; Berger & Wolpert, 1988; Bernardo & Smith, 1994; Cornfield, 1966; Howson & Urbach, 2006; Lindley, 1993; Lindley & Phillips, 1976; O'Hagan & Forster, 2004). Consider again our hypothetical questionnaire of 12 factual true–false questions about NHST. Recall that you answered 9 of the 12 questions correctly and that $p_{(2\text{-sided})} \approx 0.146$. However, now I tell you that I did not decide in advance to ask you 12 questions. In fact, it was my intention to keep asking you questions until you made your third mistake, and this just happened to take 12 questions. This procedure is known as *negative binomial sampling* (Haldane, 1945). Under this procedure, the probability of n, the total number of observations until the final mistake, is given by

$$\Pr[t(x) = n \mid \theta] = \binom{n-1}{f-1}\theta^s(1-\theta)^{n-s}, \quad (4)$$

where $f = n - s$ is the criterion number of mistakes; in this example, $f = 3$.

By comparing this formula with Equation 1 for binomial sampling, one may appreciate the fact that the part of the equation that involves θ is the same in both cases—namely, $\theta^s(1-\theta)^{n-s}$. The main difference between the binomial and negative binomial sampling plans is the number of possible permutations of correct and incorrect decisions: In the negative binomial sampling plan, the result for the final question has to be a mistake. Another difference is that for negative binomial sampling, the dependent measure is the total number of trials, a quantity that is fixed in advance in the case of binomial sampling.

Note that the observed data, x = (C, C, C, E, E, C, C, C, C, C, E), are consistent with both the binomial and the negative binomial sampling plans—hence, nothing in the data per se informs you about my sampling intention. Nevertheless, this intention does affect the p value. This happens solely because the sampling plan affects the hypothetical data expected under the null hypothesis. For instance, under the negative binomial sampling plan, a hy-

pothetical data set may have $n = 3$, as in x^{rep} = (E, E, E), or $n = 15$, as in x^{rep} = (E, E, C, C, C, C, C, C, C, C, C, C, C, C, E). My intention influences the sampling distribution of the test statistic, and with it, the p value. In the case of sampling until the third mistake, the more extreme results under H_0 consist of the hypothetical replicate experiments that take more than 12 trials to produce the third mistake. Thus, using Equation 4, the p value under the negative binomial sampling plan is given by

$$\sum_{n=12}^{\infty} \binom{n-1}{2} \left(\frac{1}{2}\right)^n \approx .033.$$

As in the previous example, the data are precisely the same, but the p value differs.

The present example is well known and clearly demonstrates that p values depend on the sampling plan. It thus follows that the p value is undefined when the sampling plan is unknown. The following example illustrates this notion in a way that is familiar to every experimental psychologist.

Example 5. The importance of hypothetical actions for imaginary data (see, e.g., Berger & Berry, 1988a; Berger & Wolpert, 1988, p. 74.1). Consider the following scenario: Amy performs a lexical decision experiment to test whether words immediately preceded by an emotional prime word (e.g., *cancer* or *love*) are categorized as quickly as words that are preceded by a neutral prime word (e.g., *poster* or *rice*). Amy's experiment involves 20 subjects. A standard null-hypothesis test on the mean response times for the two conditions yields $p = .045$, which leads Amy to conclude that the emotional valence of prime words affects response time for the target words. Amy is obviously pleased with the result and is about to submit a paper reporting the data. At this point, you ask Amy a seemingly harmless question: "What would you have done if the effect had not been significant after 20 subjects?"

Among the many possible answers to this question are the following:

1. "I don't know."
2. "I can't remember right now, but we did discuss this during a lab meeting last week."
3. "I would not have tested any more subjects."
4. "I would have tested 20 more subjects, and then I would have stopped the experiment for sure."
5. "That depends on the extent to which the effect would not have been significant. If p was greater than .25, I would not have bothered testing additional subjects, but if p was less than .25, I would have tested about 10 additional subjects and then would have stopped the experiment."
6. "That depends on whether my paper on retrieval inhibition gets accepted for *Psychonomic Bulletin & Review*. I expect the action letter soon. Only if this paper gets accepted would I have the time to test 20 additional subjects in the lexical decision experiment."

After the previous examples, it should not come as a surprise that the p value depends in part on Amy's answer, since her answer reveals the experimental sampling plan. In particular, after Answer 1 the p value is undefined, and after Answer 6 the p value depends on a yet-to-be-written action letter for a different manuscript. Only under Answer 3 does the p value remain unaffected. It is awkward that the conclusions of NHST depend critically on events that have yet to happen—events that, moreover, are completely uninformative with respect to the observed data.

Thus, in order to calculate a p value for data obtained in the past, it is necessary that you look into the future and consider your course of action for all sorts of eventualities that may occur. This requirement is very general; in order to calculate a p value, you need to know what you would have done had the data turned out differently. This includes what you would have done if the data had contained anomalous responses (e.g., fast guesses or slow outliers); what you would have done if the data had clearly been nonnormally distributed; what you would have done if the data had shown an effect of practice or fatigue; and in general, what you would have done if the data had violated any aspect of your statistical model (Hill, 1985).

Thus, p values can only be computed once the sampling plan is fully known and specified in advance. In scientific practice, few people are keenly aware of their intentions, particularly with respect to what to do when the data turn out not to be significant after the first inspection. Still fewer people would adjust their p values on the basis of their intended sampling plan. Moreover, it can be difficult to precisely quantify a sampling plan. For instance, a reviewer may insist that you test additional participants. A different reviewer might hold a different opinion. What exactly is the sampling distribution here? The problem of knowing the sampling plan is even more prominent when NHST is applied to data that present themselves in the real world (e.g., court cases or economic and social phenomena), for which no experimenter was present to guide the data collection process.

It should be stressed that NHST practitioners are not at fault when they adjust their p value on the basis of their intentions (i.e., the experimental sampling plan). When one uses the NHST methodology, this adjustment is in fact mandatory. The next example shows why.

Example 6. Sampling to a foregone conclusion (i.e., optional stopping) (see, e.g., Anscombe, 1954; Berger & Berry, 1988a; Kadane, Schervish, & Seidenfeld, 1996). It is generally understood that in the NHST framework, every null hypothesis that is not exactly true will eventually be rejected as the number of observations grows large. Much less appreciated is the fact that, even when a null hypothesis *is* exactly true, it can always be rejected, at any desired significance level that is greater than 0 (e.g., $\alpha = .05$ or $\alpha = .00001$). The method to achieve this is to calculate a p value after every new observation or set of observations comes in, and to stop the experiment as soon as the p value first drops below α. Feller (1940) discussed this sampling strategy with respect to experiments that test for extrasensory perception.

Specifically, suppose we have data $x_1, x_2, \ldots, x_n$ that are normally distributed with standard deviation $\sigma = 1$ and unknown mean μ. The null hypothesis is $\mu = 0$, and the alternative hypothesis is $\mu \neq 0$. The test statistic Z is then given by $Z = \bar{x}\sqrt{n}$, where $\bar{x}$ is the sample mean. When the null hypothesis is true, the sampling distribution of Z is the standard normal [i.e., $N(0, 1)$]. For a fixed-sample-size design, the p value is given by $p(H_0 : \mu = 0) = 2[1 - \Phi(|Z|)]$, where Φ is the standard normal cumulative distribution function.

Now suppose that the researcher does not fix the sample size in advance, but rather sets out to obtain a significant p value by using the following stopping rule: "Continue testing additional subjects until $|Z| > k$, then stop the experiment and report the result." It can be shown that this strategy will always be successful, in that the experiment will always stop, and $|Z|$ will then be greater than k (Feller, 1970). When the resultant data are subsequently analyzed as if the researcher had fixed the sample size in advance, the researcher is guaranteed to obtain a "significant" result and reject the null hypothesis (see Armitage, McPherson, & Rowe, 1969).

To appreciate the practical ramifications of optional stopping, consider successive tests for the mean of normally distributed data with known variance. Suppose the data become available in batches of equal size, and a test is conducted on the accumulating data after each new batch of data arrives. Table 2 shows that the probability that at least one of these tests is significant at the .05 level increases with the number of tests (Jennison & Turnbull, 1990; McCarroll, Crays, & Dunlap, 1992; Pocock, 1983; Strube, 2006). After only four "sneak peaks" at the data, this probability has risen to about .13.

This example forcefully demonstrates that *within the context of NHST*, it is crucial to take the sampling plan of the researcher into account; if the sampling plan is ignored, the researcher is able to always reject the null hypothesis, even if it is true. This example is sometimes used to argue that any statistical framework should somehow take the sampling plan into account. Some people feel that "optional stopping" amounts to cheating, and that no statistical inference is immune to such a devious sampling strategy. This feeling is, however, contradicted by a mathematical analysis (see, e.g., Berger & Berry, 1988a; W. Edwards, Lindman, & Savage, 1963; Kadane et al., 1996; Royall, 1997; for a summary, see an online appendix on my personal Web site, users.fmg.uva.nl/ewagenmakers/).

Table 2
The Effect of Optional Stopping on Statistical Inference Through *p* Values (cf. Jennison & Turnbull, 1990, Table 1; Pocock, 1983, Table 10.1)

Number of Tests K	Pr(Signif | H_0)
1	.05
2	.08
3	.11
4	.13
5	.14
10	.19
20	.25
50	.32
∞	1.00

Note—Pr(Signif | H_0) indicates the probability that at least one of K tests is significant, given that H_0 is true.

It is not clear what percentage of p values reported in experimental psychology have been contaminated by some form of optional stopping. There is simply no information in Results sections that allows one to assess the extent to which optional stopping has occurred. I have noticed, however, that large-sample-size experiments often produce small effects. Perhaps researchers have a good a priori idea about the size of the experimental effects that they are looking for, and thus assign more subjects to experiments with smaller expected effect sizes. Alternatively, researchers could be chasing small effects by increasing the number of subjects until "the pattern of results is clear." We will never know.

The foregoing example should not be misinterpreted. There is nothing wrong with gathering more data, examining these data, and then deciding whether or not to stop collecting new data (see the online appendix cited above). The data constitute evidence; gathering more evidence is generally helpful. It makes perfect sense to continue an experiment until the pattern of results is clear. As stated by W. Edwards et al. (1963), "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" (p. 193). What the previous example shows is that within the NHST paradigm, the researcher may not use this sensible and flexible approach to conducting an experiment. The practical consequences are substantial.

Consider, for instance, a hypothetical experiment on inhibitory control in children with ADHD. In this experiment, Hilde has decided in advance to test 40 children with ADHD and 40 control children. She examines the data after 20 children in each group have been tested and discovers that the results quite convincingly demonstrate the pattern she hoped to find: Children with ADHD have longer stop-signal response times than control children. Unfortunately for Hilde, she cannot stop the experiment and claim a significant result, since she would be guilty of optional stopping. She has to continue the experiment, wasting not just her own time and money, but also the time of the children who still need to undergo testing, as well as the time of the children's parents.

In certain situations, it is not only wasteful, but in fact *unethical* to refrain from monitoring the data as they come in, and to continue the experiment regardless of how convincing the interim data may be. This especially holds true for clinical trials, where one seeks to minimize the number of patients that have to undergo an inferior treatment. The NHST framework has been extended in several ways to deal with such situations, but the resulting procedures have also not been without controversy, as the next example illustrates.

Example 7. Sequential procedures in clinical trials (see, e.g., Anscombe, 1954, 1963; Berger & Berry, 1988a; Berger & Mortera, 1999; Cornfield, 1966; W. Edwards et al., 1963; Howson & Urbach, 2006; Royall, 1997, pp. 97–107; Ware, 1989). A team of medical doctors sets

out to study the effect of an experimental Method A on the treatment of a dangerous infectious disease. The conventional treatment is Method B. Obviously, it is important to stop the experiment as soon as it is apparent that the new drug is either superior or inferior (see Ware, 1989). Procedures that involve the online monitoring of data are known as *interim analyses*, *sequential analyses*, or *repeated significance tests*. For the reasons outlined above, NHST dogma outlaws data from sequential designs from being analyzed as if the number of participants was fixed in advance. In order to analyze such data within the framework of NHST, several methods have been proposed (e.g., Armitage, 1960; Friedman, Furberg, & DeMets, 1998; Siegmund, 1985; for concise reviews, see Jennison & Turnbull, 1990, and Ludbrook, 2003).

For concreteness, the focus here is on a matched-pairs *restricted* sequential procedure proposed by Armitage (1957). This procedure is applied to binomial data as follows (see Figure 2). Pairs of participants are assigned to Treatment A and Treatment B. The quantity to be monitored is the number of pairs for which Method A leads to better results than Method B. Thus, when Treatment A works better for one member of the pair than Treatment B does for the other member, a "counter" is increased by 1. When Treatment B works better than Treatment A, the counter is decreased by 1, and when both treatments are equally effective the counter is not updated.

The characteristic feature of the restricted sequential procedure is that one can make one of three decisions. First, as soon as the counter exceeds the upper threshold shown in Figure 2, one stops the experiment and concludes that Treatment A is better than Treatment B. Second, as soon as the counter exceeds the lower threshold, one stops the experiment and concludes that Treatment B is better than Treatment A. Third, if neither threshold has been crossed after a preset number of participants have been tested, one stops the experiment and declares that the data are inconclusive. In Figure 2, the maximum number of participants is set at $66 \times 2 = 132$. When, say, the counter is still close to 0 after about 45 pairs of subjects have been tested, it may be impossible to reach either of the two horizontally slanted boundaries before exceeding the maximum number of participants. This explains why the vertical threshold is wedge-shaped instead of vertically straight.

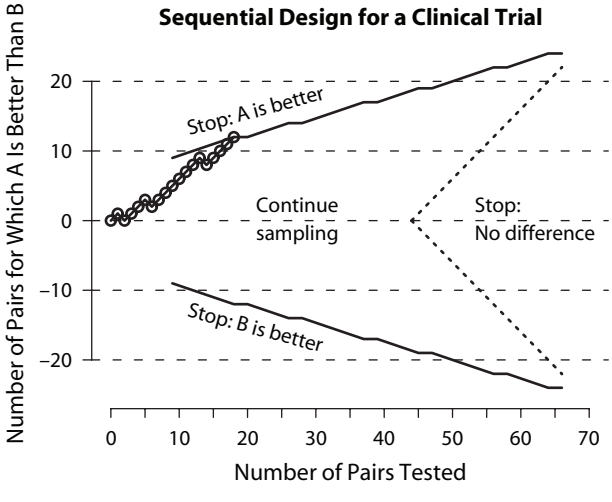

Figure 2. Example of a matched-pairs restricted sequential procedure (Armitage, 1957). Data from Freireich et al. (1963) are represented by open circles. Because the data are discrete, the upper and lower boundaries are not perfectly straight. See the text for details.

The intercept and slope for the upper and lower boundaries are determined by considering the Type I error rate α and the power $1 - \beta$ against a specific alternative. For the design in Figure 2, the probability of crossing either the upper or the lower boundary under the null hypothesis is $\alpha = .05$. When the alternative hypothesis is true, and the proportion of pairs in which one treatment is better than the other equals $\theta = .75$, the power to reject the null hypothesis is $1 - \beta = .95$. From this information, one can calculate the intercept to be ± 6.62 and the slope to be $\pm 0.2619n$, where n is the number of pairs tested (for details, see Armitage, 1957). The data shown in Figure 2 are from an experiment by Freireich et al. (1963). In this experiment, the upper threshold was reached after observing 18 pairs, with a corresponding Type I error rate of .05 (see also Berger & Berry, 1988a).

The motivation for using a restricted procedure is that it imposes an upper limit on the number of participants. In unrestricted sequential procedures, in contrast, there can be considerable uncertainty as to the number of participants required to reach a decision. Despite its elegance, the sequential procedure has several shortcomings (see also Jennison & Turnbull, 1990). First, it is unclear what can be concluded when none of the three boundaries has yet been reached. A related concern is what should be concluded when the trial is stopped, but then some additional observations come in from patients whose treatment had already started at the time that the experiment was terminated (see Freireich et al., 1963).

Much more important, however, is that sequential procedures may lead to very different conclusions than fixed-sample procedures—and for exactly the same set of data. For instance, when the experiment in Figure 2 is stopped after 18 trials, the restricted sequential procedure yields a Type I error probability of .05, whereas the fixed-sample-size p value equals .008. For the design in Figure 2, data that cross the upper or lower threshold are associated with a sequential $\alpha = .05$ Type I error rate, whereas, for the same data, the fixed-sample p values are lower by about a factor of 10.

Thus, the NHST methodology demands a heavy price for merely entertaining the idea of stopping an experiment in the entirely fictional case that the data might have turned out differently than they did. That is, "the *identical* data would have been obtained whether the experimenter had been following the sequential design or a fixed sample size design. The drastically differing measures of conclusiveness are thus due solely to thoughts about other possibilities that were in the experimenter's mind" (Berger &

Berry, 1988a, pp. 41–43). Anscombe (1963, p. 381) summarized the state of affairs as follows:

> "Sequential analysis" is a hoax. . . . So long as all observations are fairly reported, the sequential stopping rule that may or may not have been followed is irrelevant. The experimenter should feel entirely uninhibited about continuing or discontinuing his trial, changing his mind about the stopping rule in the middle, etc., because the interpretation of the observations will be based on what was observed, and not on what might have been observed but wasn't.

In sum, *p* values depend on the sampling plan of the researcher. Within the context of NHST, this is necessary, for it prevents the researcher from biasing the *p* value through optional stopping (see Example 6). The problem is that the sampling plan of the researcher reflects a subjective process that concerns hypothetical actions for imaginary events. For instance, Example 5 shows that the sampling plan may involve events that are completely unrelated to the data that had been observed.

From a practical perspective, many researchers probably ignore any sampling plan dependence and compute *p* values as if the size of the data set was fixed in advance. Since sampling plans refer in part to future events that could have occurred but did not, there is no way to check whether a reported *p* value is biased or not. For staunch supporters of NHST, this must be cause for great concern.

PROBLEM 3
p Values Do Not Quantify Statistical Evidence

In the Fisherian framework of statistical hypothesis testing, a *p* value is meant to indicate "the strength of the evidence against the hypothesis" (Fisher, 1958, p. 80); the lower the *p* value, the stronger the evidence against the null hypothesis (see also Hubbard & Bayarri, 2003). Some authors have given explicit guidelines with respect to the evidential interpretation of the *p* value. For instance, Burdette and Gehan (1970; see also Wasserman, 2004, p. 157) associated specific ranges of *p* values with varying levels of evidence: A *p* value greater than .1 yields "little or no real evidence against the null hypothesis"; a *p* value less than .1 but greater than .05 implies "suggestive evidence against the null hypothesis"; a *p* value less than .05 but greater than .01 yields "moderate evidence against the null hypothesis"; and a *p* value less than .01 constitutes "very strong evidence against the null hypothesis" (Burdette & Gehan, 1970, p. 9).

The evidential interpretation of the *p* value is an important motivation for its widespread use. If the *p* value is inconsistent with the concept of statistical evidence, there is little reason for the field of psychology to use the *p* value as a tool for separating the experimental wheat from the chaff. In this section, I review several arguments against the interpretation of *p* values as statistical evidence (see, e.g., Berger & Delampady, 1987; Berger & Sellke, 1987; Cornfield, 1966; Royall, 1997; Schervish, 1996; Sellke et al., 2001).

In order to proceed with our argument against the evidential interpretation of the *p* value, it seems that we first need to resolve a thorny philosophical issue and define precisely what is meant by "statistical evidence" (for a discussion, see, e.g., Berger & Sellke, 1987; Birnbaum, 1962, 1977; Dawid, 1984; De Finetti, 1974; Jaynes, 2003; Jeffreys, 1961; Royall, 1997; Savage, 1954). The most common and well-worked-out definition is the Bayesian definition, which will be dealt with in some detail below. For the moment, the focus is on a simple rule, a violation of which clearly disqualifies the *p* value as a measure of statistical evidence. This rule states that identical *p* values provide identical evidence against the null hypothesis. Henceforth, this rule will be referred to as the *p postulate* (see Cornfield, 1966).

Example 8. The *p* postulate: Same *p* value, same evidence? Consider two experiments in which interest centers on the effect of lexical inhibition from orthographically similar words (i.e., "neighbors"). Experiment S finds that $p = .032$ after 11 participants are tested, and Experiment L finds $p = .032$ after 98 participants are tested. Do the two experiments provide equally strong evidence against the null hypothesis? If not, which experiment is the more convincing?

When the *p* value is kept constant, Rosenthal and Gaito (1963) found that the confidence with which a group of psychologists were willing to reject the null hypothesis increased with sample size (see also Nelson, Rosenthal, & Rosnow, 1986). Thus, psychologists tend to think that Experiment L provides more evidence against the null hypothesis than does Experiment S. The psychologists' reasoning may be that a significant result is less likely to be due to chance fluctuations when the number of observations is large than when it is small. Consistent with the psychologists' intuition, an article co-authored by 10 reputable statisticians maintained that "A given *p* value in a large trial is usually stronger evidence that the treatments really differ than the same *p* value in a small trial of the same treatments would be" (Peto et al., 1976, p. 593, as cited in Royall, 1997, p. 71).

Nevertheless, Fisher himself was of the opinion that the *p* postulate is correct: "It is not true . . . that valid conclusions cannot be drawn from small samples; if accurate methods are used in calculating the probability [the *p* value], we thereby make full allowance for the size of the sample, and should be influenced in our judgement only by the value of probability indicated" (Fisher, 1934, p. 182, as cited in Royall, 1997, p. 70). Thus, Fisher apparently argues that Experiments L and S provide equal evidence against the null hypothesis.

Finally, several researchers have argued that when the *p* values are the same, studies with small sample size actually provide *more* evidence against the null hypothesis than do studies with large sample size (see, e.g., Bakan, 1966; Lindley & Scott, 1984; Nelson et al., 1986). Note that the *p* value is influenced both by effect size and by sample size. Hence, when Experiments L and S have different sample sizes but yield the same *p* value, it must be the case that Experiment L deals with a smaller effect size than does Experiment S. Because Experiment L has

more power to detect a difference than does Experiment S, the fact that they yield the same *p* value suggests that the effect is less pronounced in Experiment L. This reasoning suggests that Experiment S provides more evidence against the null hypothesis than does Experiment L.

In sum, there is considerable disagreement as to whether the *p* postulate holds true. Among those who believe the *p* postulate is false, some believe that studies with small sample size are less convincing than those with large sample size, and others believe the exact opposite. As will shortly become apparent, a Bayesian analysis strongly suggests that the *p* postulate is false: When two experiments have different sample sizes but yield the same *p* value, the experiment with the smallest sample size is the one that provides the strongest evidence against the null hypothesis.

BAYESIAN INFERENCE

This section features a brief review of the Bayesian perspective on statistical inference. A straightforward Bayesian method is then used to cast doubt on the *p* postulate. This section also sets the stage for an explanation of the BIC as a practical alternative to the *p* value. For more detailed introductions to Bayesian inference, see, for instance, Berger (1985), Bernardo and Smith (1994), W. Edwards et al. (1963), Gill (2002), Jaynes (2003), P. M. Lee (1989), and O'Hagan and Forster (2004).

Bayesians Versus Frequentists

The statistical world has always been dominated by two superpowers: the *Bayesians* and the *frequentists* (see, e.g., Bayarri & Berger, 2004; Berger, 2003; Christensen, 2005; R. T. Cox, 1946; Efron, 2005; Lindley & Phillips, 1976; Neyman, 1977). Bayesians use probability distributions to quantify uncertainty or degree of belief. Incoming data then reduce uncertainty or update belief according to the laws of probability theory. Bayesian inference is a method of inference that is *coherent* (i.e., internally consistent; see Bernardo & Smith, 1994; Lindley, 1972). A Bayesian feels free to assign probability to all kinds of events—for instance, the event that a fair coin will land tails in the next throw or that the Dutch soccer team will win the 2010 World Cup.

Frequentists believe that probability should be conceived of as a limiting frequency. That is, the probability that a fair coin lands tails is 1/2 because this is the proportion of times a fair coin would land heads if it were tossed very many times. In order to assign probability to an event, a frequentist has to be able to repeat an experiment very many times under exactly the same conditions. More generally, a frequentist feels comfortable assigning probability to events that are associated with "aleatory uncertainty" (i.e., uncertainty due to randomness). These are events associated with phenomena such as coin tossing and card drawing. On the other hand, a frequentist may refuse to assign probability to events associated with "epistemic uncertainty" (i.e., uncertainty due to lack of knowledge, which may differ from one person to another). The event that the Dutch soccer team wins the 2010 World Cup is one associated with epistemic uncertainty—for example, the Dutch trainer has more knowledge about the plausibility of this event than I do (for a discussion, see R. T. Cox, 1946; Fine, 1973; Galavotti, 2005; O'Hagan, 2004).

Over the past 150 years or so, the balance of power in the field of statistics has shifted from the Bayesians, such as Pierre-Simon Laplace, to frequentists such as Jerzy Neyman. Recently, the balance of power has started to shift again, as Bayesian methods have made an inspired comeback. Figure 3 illustrates the popularity of the Bayesian paradigm in the field of statistics; it plots the proportion of articles with the letter string "Bayes" in the title or abstract of articles published in statistics' premier journal, the *Journal of the American Statistical Association* (JASA). Figure 3 shows that the interest in Bayesian methods is steadily increasing. This increase is arguably due mainly to pragmatic reasons: The Bayesian program is conceptually straightforward and can be easily extended to more complicated situations, such as those that require (possibly nonlinear) hierarchical models (see, e.g., Rouder & Lu, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005) or order-constrained inference (Klugkist, Laudy, & Hoijtink, 2005). The catch is that, as we will see later, practical

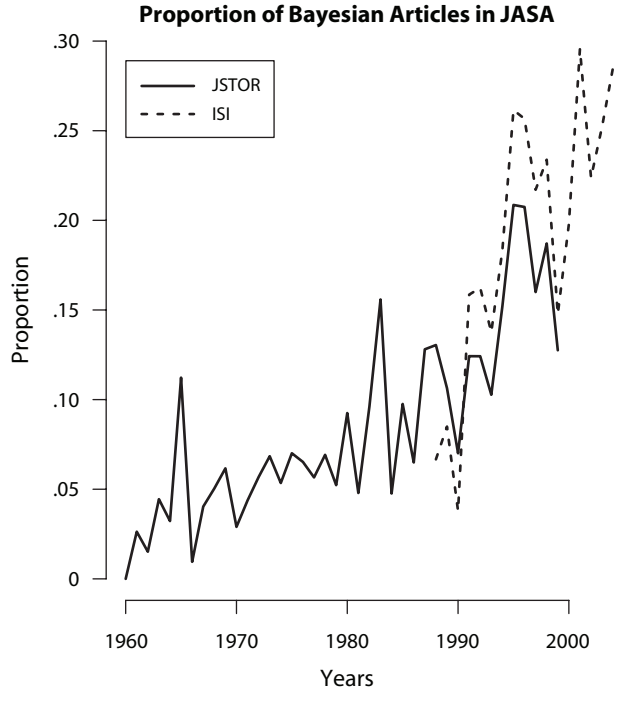

Figure 3. The proportions of Bayesian articles in the *Journal of the American Statistical Association* (JASA) from 1960 to 2005, as determined by the proportions of articles with the letter string "Bayes" in title or abstract. This method of estimation will obviously not detect articles that use Bayesian methodology but do not explicitly acknowledge such use in the title or abstract. Hence, the present estimates are biased downward, and may be thought of as a lower bound. JSTOR (www.jstor.org) is a not-for-profit scholarly journal archive that has a 5-year moving window for JASA. The ISI Web of Science (isiknowledge.com) journal archive does not have a moving window for JASA, but its database only goes back to 1988. ISI estimates are generally somewhat higher than JSTOR estimates because ISI also searches article keywords.

implementation of Bayesian methods requires integration over the parameter space. For many problems, the values of such integrals are not available as analytic expressions, and one has to resort to computer-intensive numerical approximation methods. The increasing development and feasibility of such computer-intensive methods (i.e., Markov chain Monte Carlo [MCMC] methods; Gilks, Richardson, & Spiegelhalter, 1996; Robert & Casella, 1999) has greatly fueled the Bayesian fire.

I suspect that many psychologists are not aware that the Bayesian paradigm is quite popular in the field of statistics, for frequentist statistics is the version that is taught to psychology students in undergraduate classes and used by experimental psychologists to analyze their data. Figure 3 suggests that a large gap indeed exists between statistics theory, as studied in JASA, and statistical practice, as displayed in the standard journals for experimental psychology. I estimate that the proportion of articles in, say, the *Journal of Experimental Psychology: Learning, Memory, and Cognition* that use Bayesian methods for data analysis is approximately 0.

Bayesian Parameter Estimation

For simplicity and consistency, this section continues our binomial example. To reiterate, I asked you 12 factual true–false questions, and you answered 9 of them correctly. We adopt the binomial model, in which the probability of 9 correct responses out of 12 questions is given by

$$\Pr(s = 9 \mid \theta, n = 12) = \binom{12}{9} \theta^9 (1-\theta)^3,$$

where $\theta \in [0, 1]$ indicates the probability of answering any one question correctly. Let D denote the observed data (i.e., $s = 9$ out of $n = 12$). Thus, we have already obtained $\Pr(D \mid \theta)$—that is, the probability of the data given θ. The object of our inference, however, is $\Pr(\theta \mid D)$, the *posterior distribution* of θ given the data.

Probability theory gives the relation between $\Pr(\theta \mid D)$ and $\Pr(D \mid \theta)$:

$$\Pr(\theta \mid D) = \frac{\Pr(D \mid \theta)\Pr(\theta)}{\Pr(D)}. \quad (5)$$

This simple equation is known as *Bayes's theorem*. The initial state of knowledge about θ is indicated by the *prior distribution* $\Pr(\theta)$. This prior distribution is updated by means of the likelihood $\Pr(D \mid \theta)$ to yield the posterior distribution $\Pr(\theta \mid D)$. Note that, in contrast to the method of *p* values, the Bayesian method is conditioned on the data D that have actually been observed, and does not make any reference to imaginary data that could have been observed but were not.

When additional data D_2 come in, the posterior distribution for θ is obtained as in Equation 5, save that all quantities are conditioned on the presence of the old data D—that is, $\Pr(\theta \mid D_2, D) = \Pr(D_2 \mid \theta, D)\Pr(\theta \mid D)/\Pr(D_2 \mid D)$. When D_2 is independent of D, so that $\Pr(D_2 \mid \theta, D) = \Pr(D_2 \mid \theta)$, the posterior simplifies to $\Pr(\theta \mid D_2, D) = \Pr(D_2 \mid \theta)\Pr(\theta \mid D)/\Pr(D_2)$. Comparison with Equation 5 shows that sequential updating in the Bayesian paradigm takes a particularly simple form: The posterior distribution $\Pr(\theta \mid D)$ after observation of the first batch of data D becomes the prior distribution for the next batch of data D_2, and this "prior" distribution gets updated through the likelihood for the new data D_2.

In Equation 5, the normalizing constant $\Pr(D)$ does not depend on θ. In fact, $\Pr(D)$ is calculated by integrating out θ—that is, $\Pr(D) = \int \Pr(D \mid \theta)\Pr(\theta)\, d\theta$. $\Pr(D)$ is also known as the *marginal probability* or the *prior predictive probability* of the data. For many purposes related to parameter estimation, $\Pr(D)$ can safely be ignored, and we can write Equation 5 as

$$\Pr(\theta \mid D) \propto \Pr(D \mid \theta)\Pr(\theta), \quad (6)$$

where $\propto$ stands for "is proportional to."

Let's apply the foregoing to our true–false example. First, we need to specify $\Pr(\theta)$, the prior distribution for the binomial parameter that gives the probability of answering any one question correctly. It seems reasonable, a priori, to expect at least some knowledge of the topic to be present, suggesting a prior with most of its mass on $\theta > 1/2$. On the other hand, a closer look at the questions reveals that they were probably specially selected to elicit an incorrect answer. In this situation, it is reasonable to assume as little about θ as possible. A standard choice for such an "uninformative" prior distribution is the uniform distribution shown in the top panel of Figure 4 (for a more principled argument as to why the uniform is an appropriate uninformative distribution, see Jaynes, 2003; Lee &

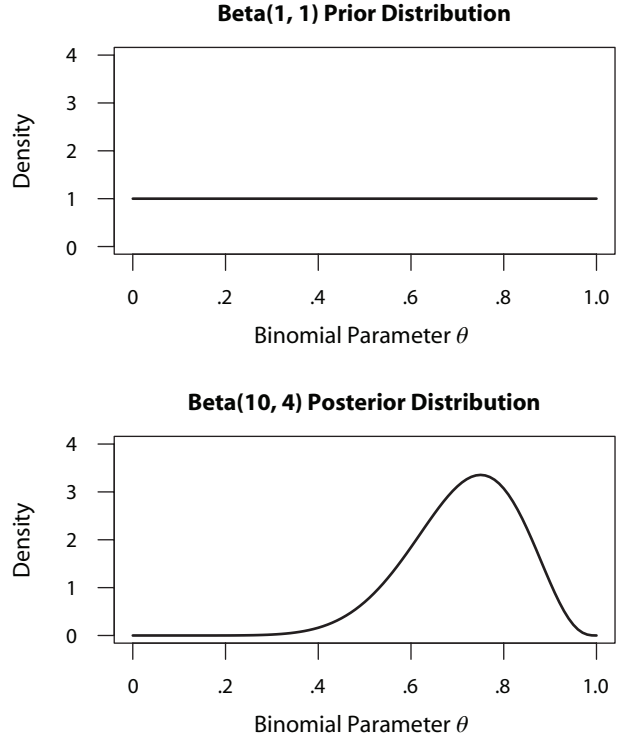

Figure 4. Bayesian parameter estimation for the binomial model. See the text for details.

Wagenmakers, 2005). For this flat prior, $\theta \sim \text{Uniform}(0, 1)$, and it conveniently drops out of subsequent calculations.

The binomial likelihood of the data is

$$\Pr(s \mid \theta, n) = \binom{n}{s} \theta^s (1-\theta)^{n-s}.$$

In this particularly simple problem, it is possible to obtain the marginal probability of the data, $\Pr(D)$, in analytic form. Appendix A shows that $\Pr(D) = 1/(n+1)$. Now we have all the information required to calculate $\Pr(\theta \mid D)$. After plugging in the required information, Equation 5 yields $\Pr(\theta \mid s, n) = (n+1)\Pr(s \mid \theta, n)$. For the hypothetical data from the true–false example,

$$\Pr(\theta \mid s = 9, n = 12) = 13 \binom{12}{9} \theta^9 (1-\theta)^3.$$

This posterior distribution is shown in the bottom panel of Figure 4. After the data have been observed, values of $\theta < .5$ are less plausible than they were under the prior distribution, and values of $\theta \in (.5, .9)$ have become more plausible than they were under the prior. Once we have the posterior distribution, the estimation problem is solved, and we can proceed to report the entire distribution or useful summary measures (see Lee & Wagenmakers, 2005). For instance, the mean of the distribution is $5/7 \approx .71$, and the 95% Bayesian confidence interval for θ is (.462, .909).

In this particular example, there exists an easier method to obtain the posterior distribution. This is how it works: The Beta probability density function, $\text{Beta}(\theta \mid \alpha, \beta)$, has two parameters that determine the distribution of θ (for details, see Appendix A). The uniform distribution happens to correspond to a Beta distribution with $\alpha = 1$ and $\beta = 1$. When a Beta prior is updated through a binomial likelihood function, the resulting posterior distribution is still a Beta distribution, albeit with parameters $\alpha + s$ and $\beta + n - s$. Thus, the posterior we calculated earlier,

$$\Pr(\theta \mid s = 9, n = 12) = 13 \binom{12}{9} \theta^9 (1-\theta)^3,$$

is in fact a $\text{Beta}(\theta \mid 10, 4)$ distribution.

Bayesian Hypothesis Testing

One of the most obvious features of a Bayesian hypothesis test is that it is comparative in nature. A Bayesian hypothesis test always involves at least two different models. Consequently, from a Bayesian perspective, the fact that the null hypothesis is unlikely is not sufficient reason to reject it—the data may be even more unlikely under the alternative hypothesis. After observing the data, the posterior odds in favor of the null hypothesis H_0 versus the alternative hypothesis H_1 are given by

$$\frac{\Pr(H_0 \mid D)}{\Pr(H_1 \mid D)} = \frac{\Pr(D \mid H_0)}{\Pr(D \mid H_1)} \cdot \frac{\Pr(H_0)}{\Pr(H_1)}. \quad (7)$$

This equation states that the posterior odds are equal to the ratio of prior predictive probabilities times the prior odds. Often the focus of interest is on the change in odds from prior to posterior, brought about by the data. This change is indicated by the ratio of prior predictive probabilities, $\Pr(D \mid H_0)/\Pr(D \mid H_1)$, a quantity known as the *Bayes factor* (see, e.g., Jeffreys, 1961). The Bayes factor, or the log of the Bayes factor, is often interpreted as the weight of evidence coming from the data (e.g., Good, 1985). Thus, a Bayes-factor hypothesis test prefers the hypothesis under which the observed data are most likely (for details, see Bernardo & Smith, 1994, chap. 6; Gill, 2002, chap. 7; Kass & Raftery, 1995).

Some of the details involved can best be illustrated by once more turning to our true–false example (see also Nickerson, 2000; Wagenmakers & Grünwald, 2006; Wasserman, 2000). In a Bayesian hypothesis test, the interest might center on whether your performance (i.e., 9 correct answers out of 12 questions) is consistent with an explanation in terms of random guessing. Thus, H_0, the null hypothesis of random guessing, posits that the success rate parameter in the binomial model is fixed at $\theta = 1/2$. For H_1, the alternative hypothesis, θ is a free parameter, $\theta \neq 1/2$. As discussed above, in this example a reasonable prior is the uninformative uniform prior. The Bayes factor is then given by

$$\begin{aligned} BF_{01} &= \frac{\Pr(D \mid H_0)}{\Pr(D \mid H_1)} = \frac{\Pr(D \mid \theta = 1/2)}{\Pr[D \mid \theta \sim \text{Beta}(1,1)]} \\ &= \frac{\binom{12}{9}\left(\frac{1}{2}\right)^{12}}{\int_0^1 \Pr(D \mid \theta)\Pr(\theta)\, d\theta} \\ &= (n+1)\binom{12}{9}\left(\frac{1}{2}\right)^{12} \approx 0.70. \end{aligned} \quad (8)$$

BF_{01} is smaller than 1, indicating that the data are more likely under H_1 than they are under H_0. Specifically, the data are $1/0.7 \approx 1.4$ times more likely under H_1 than they are under H_0. If we believe H_0 and H_1 to be equally likely a priori [i.e., $\Pr(H_0) = \Pr(H_1)$], one can calculate the posterior probability of H_0 as $\Pr(H_0 \mid D) \approx 0.7/(1 + 0.7) \approx .41$. The posterior probability of the alternative hypothesis is the complement, $\Pr(H_1 \mid D) = 1 - \Pr(H_0 \mid D)$.

From the foregoing analysis, a Bayesian would happily conclude that the data are about 1.4 times as likely to have occurred under H_1 than under H_0, and be done with it. For many experimental psychologists, however, such a conclusion may be unsatisfactory; these researchers desire to see the Bayes factor reduced to a dichotomous decision: Either H_0 is true or H_0 is false. Statistically, this desire is entirely unfounded. Just like the ubiquitous .05 criterion in NHST, any criterion on the Bayes factor will be somewhat arbitrary. Moreover, any summary of the Bayes factor will lose information. Nevertheless, people apparently find it difficult to deal with continuous levels of evidence. It should also be acknowledged that it is useful to be able to summarize the quantitative results from a Bayesian hypothesis test in words.

In order to accommodate the need for a verbal description of the Bayes factor, Jeffreys (1961) proposed a division into four distinct categories. Raftery (1995) proposed

Table 3
Interpretation of the Bayes Factor in Terms of Evidence (cf. Raftery, 1995, Table 6)

Bayes Factor BF_{01}	$\Pr(H_0 \mid D)$	Evidence
1–3	.50–.75	weak
3–20	.75–.95	positive
20–150	.95–.99	strong
>150	>.99	very strong

Note—$\Pr(H_0 \mid D)$ is the posterior probability for H_0, given that $\Pr(H_0) = \Pr(H_1) = 1/2$.

some minor adjustments of these rules of thumb. Table 3 shows the Raftery (1995) classification scheme. The first column shows the Bayes factor, the second column shows the associated posterior probability when it is assumed that both H_0 and H_1 are a priori equally plausible, and the third column shows the verbal labels for the evidence at hand. Consistent with intuition, the result that the data from the true–false example are 1.4 times as likely under H_1 than they are under H_0 constitutes only "weak evidence" in favor of H_1.

Equation 8 shows that the prior predictive probability of the data for H_1 is given by the average of the likelihood for the observed data over all possible values of θ. Generally, the prior predictive probability is a weighted average, the weights being determined by the prior distribution $\Pr(\theta)$. In the case of a uniform Beta(1, 1) prior, the weighted average reduces to an equally weighted average. Non-Bayesians cannot average over θ, since they are reluctant or unwilling to commit to a prior distribution. As a result, non-Bayesians often determine the maximum likelihood—that is, $\Pr(D \mid \hat{\theta})$, where $\hat{\theta}$ is the value of θ under which the observed results are most likely. One of the advantages of averaging instead of maximizing over θ is that averaging automatically incorporates a penalty for model complexity. That is, a model in which θ is free to take on any value in [0, 1] is more complex than the model that fixes θ at 1/2. By averaging the likelihood over θ, values of θ that turn out to be very implausible in light of the observed data (e.g., all θs $< .4$ in our true–false example) will lead to a relative decrease of the prior predictive probability of the data (Myung & Pitt, 1997).

"That Wretched Prior . . ."

This short summary of the Bayesian paradigm would be incomplete without a consideration of the prior $\Pr(\theta)$. Priors do not enjoy a good reputation, and some researchers apparently believe that by opening a Pandora's box of priors, the Bayesian statistician can bias the results at will. This section illustrates how priors are specified and why priors help rather than hurt statistical inference. More details and references can be found in a second online appendix on my personal Web site (users.fmg.uva.nl/ewagenmakers/). The reader who is eager to know why the *p* postulate is false can safely skip to the next section.

Priors can be determined by two different methods. The first method is known as "subjective." A "subjective Bayesian" argues that all inference is necessarily relative to a particular state of knowledge. For a subjective Bayesian, the prior simply quantifies a personal degree of belief that is to be adjusted by the data (see, e.g., Lindley, 2004). The second method is known as "objective" (Kass & Wasserman, 1996). An "objective Bayesian" specifies priors according to certain predetermined rules. Given a specific rule, the outcome of statistical inference is independent of the person who performs the analysis. Examples of objective priors include the unit information priors (i.e., priors that carry as much information as a single observation; Kass & Wasserman, 1995), priors that are invariant under transformations (Jeffreys, 1961), and priors that maximize entropy (Jaynes, 1968). Objective priors are generally vague or uninformative—that is, thinly spread out over the range for which they are defined.

From a pragmatic perspective, the discussion of subjective versus objective priors would be moot if it could be shown that the specific shape of the prior did not greatly affect inference (see Dickey, 1973). Consider Bayesian inference for the mean μ of a normal distribution. For parameter estimation, one can specify a prior $\Pr(\mu)$ that is very uninformative (e.g., spread out across the entire real line). The data will quickly overwhelm the prior, and parameter estimation is hence relatively robust to the specific choice of prior. In contrast, the Bayes factor for a two-sided hypothesis test is sensitive to the shape of the prior (Lindley, 1957; Shafer, 1982). This is not surprising; if we increase the interval along which μ is allowed to vary according to H_1, we effectively increase the complexity of H_1. The inclusion of unlikely values for μ decreases the average likelihood for the observed data. For a subjective Bayesian, this is not really an issue, since $\Pr(\mu)$ reflects a prior belief. For an objective Bayesian, hypothesis testing constitutes a bigger challenge: On the one hand, an objective prior needs to be vague. On the other hand, a prior that is too vague can increase the complexity of H_1 to such an extent that H_1 will always have low posterior probability, regardless of the observed data. Several objective Bayesian procedures have been developed that try to address this dilemma, such as the local Bayes factor (Smith & Spiegelhalter, 1980), the intrinsic Bayes factor (Berger & Mortera, 1999; Berger & Pericchi, 1996), the partial Bayes factor (O'Hagan, 1997), and the fractional Bayes factor (O'Hagan, 1997) (for a summary, see Gill, 2002, chap. 7).

For many Bayesians, the presence of priors is an asset rather than a nuisance. First of all, priors ensure that different sources of information are appropriately combined, such as when the posterior after observation of a batch of data D_1 becomes the prior for the observation of a new batch of data D_2. In general, inference without priors can be shown to be internally inconsistent or *incoherent* (see, e.g., Bernardo & Smith, 1994; Cornfield, 1969; R. T. Cox, 1946; D'Agostini, 1999; De Finetti, 1974; Jaynes, 2003; Jeffreys, 1961; Lindley, 1982). A second advantage of priors is that they can prevent one from making extreme and implausible inferences; priors may "shrink" the extreme estimates toward more plausible values (Box & Tiao, 1973, pp. 19–20; Lindley & Phillips, 1976; Rouder et al., 2005). A third advantage of specifying priors is that it allows one to focus on parameters of interest by eliminating so-called *nuisance parameters* through the law of total

probability (i.e., integrating out the nuisance parameters). A fourth advantage of priors is that they might reveal the true uncertainty in an inference problem. For instance, Berger (1985, p. 125) argues that "when different reasonable priors yield substantially different answers, can it be right to state that there is a single answer? Would it not be better to admit that there is scientific uncertainty, with the conclusion depending on prior beliefs?"

These considerations suggest that inferential procedures that are incapable of taking prior knowledge into account are incoherent (Lindley, 1977), may waste useful information, and may lead to implausible estimates. Similar considerations led Jaynes (2003, p. 373) to state that "If one fails to specify the prior information, a problem of inference is just as ill-posed as if one had failed to specify the data."

A BAYESIAN TEST OF THE *p* POSTULATE

We are now in a position to compare the Bayesian hypothesis test to the *p* value methodology. Such comparisons are not uncommon (see, e.g., Berger & Sellke, 1987; Dickey, 1977; W. Edwards et al., 1963; Lindley, 1957; Nickerson, 2000; Wagenmakers & Grünwald, 2006), but their results have far-reaching consequences: Using a Bayes-factor approach, one can undermine the *p* postulate by showing that *p* values overestimate the evidence against the null hypothesis.[5]

For concreteness, assume that Ken participates in an experiment on taste perception. On each trial of the experiment, Ken is presented with two glasses of beer, one glass containing Budweiser and the other containing Heineken. Ken is instructed to identify the glass that contains Budweiser. Our hypothetical participant enjoys his beer, so we can effectively collect an infinite amount of trials. On each trial, assuming conditional independence, Ken has probability θ of making the right decision. For inference, we will again use the binomial model. For both the Bayesian hypothesis tests and the frequentist hypothesis test, the null hypothesis is that performance can be explained by random guessing—that is, $H_0 : \theta = 1/2$. Both the Bayesian and the frequentist hypothesis tests are two-sided.

In the Bayesian analysis of this particular situation, it is assumed that the null hypothesis H_0 and the alternative hypothesis H_1 are a priori equally likely. That is, it is assumed that Ken is a priori equally likely to be able to discriminate Budweiser from Heineken as he is not. Via Equation 7, the calculation of the Bayes factor then easily yields the posterior probability of the null hypothesis: $\Pr(H_0 \mid D) = BF_{01}/(1 + BF_{01})$. To calculate the prior predictive probability of H_1: $\theta \neq 1/2$, we need to average the likelihood weighted by the prior distribution for θ:

$$\Pr\left(D \mid H_1\right) = \int_0^1 \Pr\left(D \mid \theta\right)\Pr\left(\theta\right)\, d\theta.$$

As mentioned earlier, the prior distribution can be determined in various ways. Figure 5 shows two possible priors. The uniform Beta(1, 1) prior is an objective prior that is often chosen in order to "let the data speak for themselves." The peaked Beta(6, 3) prior is a subjective prior that reflects my personal belief that it is possible for an experienced beer drinker to distinguish Budweiser from Heineken, although performance probably will not be anywhere near perfection.

A third prior is the "oracle point prior." This is not a fair prior, since it is completely determined by the observed data. For instance, if Ken successfully identifies Budweiser in *s* out of *n* trials, the oracle prior will concentrate all its prior mass on the estimate of θ that makes the observed data as likely as possible—that is, $\hat{\theta} = s/n$. Because it is determined by the data, and because it is a single point instead of a distribution, the oracle prior is totally implausible and would never be considered in actual practice. The oracle prior is useful here because it sets a lower bound for the posterior probability of the null hypothesis (W. Edwards et al., 1963). In other words, the oracle prior maximizes the posterior probability of the alternative hypothesis; thus, whatever prior one decides to use, the probability of the null hypothesis has to be at least as high as it is under the oracle prior.

To test the *p* postulate that equal *p* values indicate equal evidence against the null hypothesis, the data were constructed as follows. The number of observations *n* was varied from 50 to 10,000 in increments of 50, and for each of these *n*s I determined the number of successful decisions *s* that would result in an NHST *p* value that is barely significant at the .05 level, so that for these data *p* is effectively fixed at .05. For instance, if $n = 400$, the number of successes has to be $s = 220$ to result in a *p* value of about .05. When $n = 10{,}000$, *s* has to be 5,098. Next, for the data that were constructed to have the same *p* value of .05, I calculated the Bayes factor $\Pr(D \mid H_0)/\Pr(D \mid H_1)$ using the objective Beta(1, 1) prior, the subjective Beta(6, 3) prior, and the unrealistic oracle prior. From the Bayes factor, I then computed posterior probabilities for the null hypothesis.

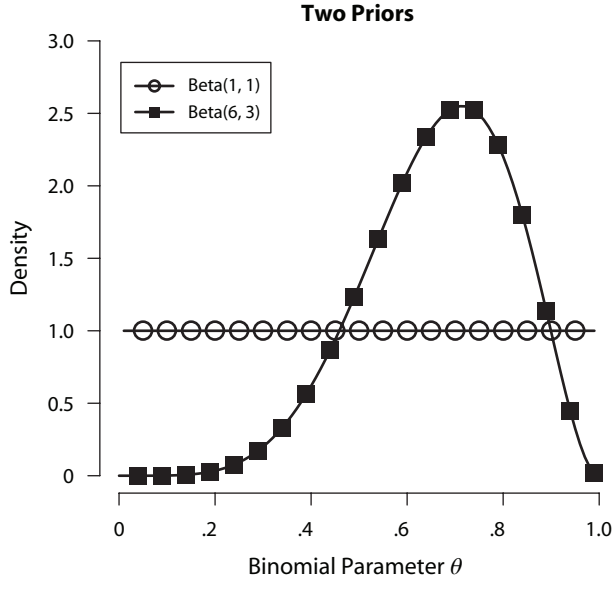

Figure 5. Distributions for a flat objective prior and an asymmetric subjective prior.

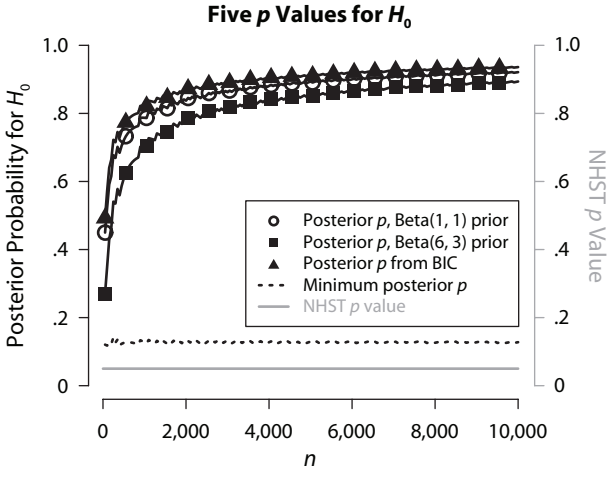

Figure 6. Four Bayesian posterior probabilities are contrasted with the classical *p* value, as a function of sample size. Note that the data are constructed to be only just significant at the .05 level (i.e., $p \approx .05$). This means that as *n* increases, the proportion of correct judgments *s* has to decrease. The upper three posterior probabilities illustrate that for realistic priors, the posterior probability of the null hypothesis strongly depends on the number of observations. As *n* goes to infinity, the probability of the null hypothesis goes to 1. This conflicts with the conclusions from NHST (i.e., "$p = .05$, reject the null hypothesis"), which is shown by the constant lowest line. For most values of *n*, the fact that $p = .05$ constitutes evidence in support of the null hypothesis rather than against it.

Figure 6 shows the results. By construction, the NHST *p* value is a constant .05. The lower bound on the posterior probability for H_0, obtained by using the oracle prior, is a fairly constant .128. This is an important result, for it shows that—for data that are just significant at the .05 level—even the utmost generosity to the alternative hypothesis cannot make it more than about 6.8 times as likely as the null hypothesis. Thus, an oracle prior that is maximally biased against the null hypothesis will generate "positive evidence" for the alternative analysis (see Table 3), but this evidence is not as strong as the NHST "1 in 20" that $p = .05$ may suggest. The posterior probability for H_0 increases when we abandon the oracle prior and consider more realistic priors.

Both the uniform "objective" prior and the peaked "subjective" prior yield posterior probabilities for H_0 that are dramatically different from the constant NHST *p* value of .05. The posterior probabilities for H_0 are not constant, but rather increase with *n*. For instance, under the uniform prior, the Bayes factor in favor of the null hypothesis is about 2.17 when $s = 220$ and $n = 400$, and about 11.69 when $s = 5{,}098$ and $n = 10{,}000$. In addition, for data that would prompt rejection of the null hypothesis by NHST methodology, the posterior probability for H_0 may be considerably higher than the posterior probability for H_1.

Figure 6 also shows the posterior probability for H_0 computed from the BIC approximation to the Bayes factor. In a later section, I will illustrate how the BIC can be calculated easily from SPSS output and how the raw BIC values can be transformed to posterior probabilities. For now, the only important regularity to note is that all three posterior probabilities (i.e., objective, subjective, and BIC) show the same general pattern of results.

The results above are not an artifact of the particular priors used. You are free to choose any prior you like, as long as it is continuous and strictly positive on $\theta \in [0, 1]$. For any such prior, the posterior probability of H_0 will converge to 1 as *n* increases (Berger & Sellke, 1987), provided of course that the NHST *p* value remains constant (see also Lindley, 1957; Shafer, 1982). The results from Figure 6 therefore have considerable generality.

At this point, the reader may wonder whether it is appropriate to compare NHST *p* values to posterior probabilities, since these two probabilities refer to different concepts. I believe the comparison is insightful for at least two reasons. First, it highlights that *p* values should not be misinterpreted as posterior probabilities: A *p* value smaller than .05 does not mean that the alternative hypothesis is more likely than the null hypothesis, even if both hypotheses are equally likely a priori. Second, the comparison shows that if *n* increases and the *p* value remains constant, the posterior probability for H_0 goes to 1 for any plausible prior. In other words, there is no plausible prior for which the posterior probability of the null hypothesis is monotonically related to the NHST *p* value as the number of observations increases.

It should be acknowledged that the interpretation of the results hinges on the assumption that there exists a prior for which the Bayesian analysis will give a proper measure of statistical evidence, or at least a measure that is monotonically related to statistical evidence. Thus, it is possible to dismiss the implications of the previous analysis by arguing that a Bayesian analysis cannot yield a proper measure of evidence, whatever the shape of the prior. Such an argument conflicts with demonstrations that the only coherent measure of statistical evidence is Bayesian (see, e.g., Jaynes, 2003).

In conclusion, NHST *p* values may overestimate the evidence against the null hypothesis; for instance, a data set that does not discredit H_0 in comparison with H_1 may nonetheless be associated with a *p* value lower than .05, thus prompting a rejection of H_0. This tendency is compounded when sample size is large (for practical consequences in medicine and public policy, see Diamond & Forrester, 1983). I believe these results demonstrate that the *p* postulate is false; equal *p* values do not provide equal evidence against the null hypothesis. Specifically, for fixed *p* values, the data provide more evidence against H_0 when the number of observations is small than when it is large. This means that the NHST *p* value is not a proper measure of statistical evidence.

One important reason for the difference between Bayesian posterior probabilities and frequentist *p* values is that the Bayesian approach is comparative and the NHST procedure is not. That is, in the NHST paradigm, H_0 is rejected if the data—and more extreme data that could have been observed but were not—are very unlikely under H_0. Therefore, the NHST procedure is oblivious to the very real possibility that although the data may be unlikely under H_0, they are even less likely under H_1.

In the field of cognitive modeling, the advantages of a comparative approach are duly recognized. For instance, Busemeyer and Stout (2002, p. 260) state that "It is meaningless to evaluate a model in isolation, and the only way to build confidence in a model is to compare it with reasonable competitors." In my opinion, the distinction between cognitive models and statistical models is one of purpose rather than method, so the quotation above would apply to statistical inference as well. Many statisticians have criticized the selective focus on the null hypothesis (e.g., Hacking, 1965). Jeffreys (1961, p. 390) states the problem as follows:

> Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place? If there is no clearly stated alternative, and the null hypothesis is rejected, we are simply left without any rule at all, whereas the null hypothesis, though not satisfactory, may at any rate show some sort of correspondence with the facts.

Interim Conclusion

The use of NHST is tainted by statistical and practical difficulties. The methodology requires knowledge of the intentions of the researcher performing the experiment. These intentions refer to hypothetical events, and can therefore not be subjected to scientific scrutiny. It is my strong personal belief that the wide majority of experiments in the field of psychology violate at least one of the major premises of NHST. One common violation of NHST logic is to test additional subjects when the effect is not significant after a first examination, and then to analyze the data as if the number of subjects was fixed in advance. Another common violation is to take sneak peeks at the data as they accumulate and to stop the experiment when they look convincing. In both cases, the reported p value will be biased downward.

This abuse of NHST could be attributed either to dishonesty or to ignorance. Personally, I would make a case for ignorance. An observation by Ludbrook (2003) supports this conjecture: "Whenever I chastise experimentalists for conducting interim analyses and suggest that they are acting unstatistically, if not unethically, they react with surprise." Abuse by ignorance is certainly more sympathetic than abuse by bad intentions. Nevertheless, ignorance of NHST is hardly a valid excuse for violating its core premises (see Anscombe, 1963). The most positive interpretation of the widespread abuse is that researchers are guided by the Bayesian intuition that they need not concern themselves with subjective intentions and hypothetical events, since only the data that have actually been observed are relevant to statistical inference. Although this intuition is, in my opinion, correct as a general guideline, in the framework of NHST it is completely off.

In sum, NHST imposes upon the user a straitjacket of restrictions and requirements. In return, the user may obtain an indication of goodness of fit that depends on the oftentimes unknown intention of the researcher performing the experiment. Moreover, the p value ignores the alternative hypothesis and therefore fails to quantify statistical evidence. The same p value also does not always carry the same weight; a Bayesian analysis confirmed that for a fixed p value, a small study provides more evidence against the null hypothesis than does a large study. Furthermore, data exist for which the null hypothesis is rejected on the basis of the NHST p value, whereas a Bayesian analysis of the same data finds the null hypothesis to be much more plausible than the alternative hypothesis. In a fixed-sample-size design, a Bayesian analysis that is maximally biased against the null hypothesis does not cast as much doubt on the null hypothesis as does the NHST p value (see Figure 6).

One may well wonder why p values have been used so extensively, given their obvious drawbacks. Several reasons may be identified, but one of the most important ones is surely that many of the p value criticisms have not been accompanied by a concrete proposal for a feasible and easy-to-use alternative procedure. This has perhaps fueled the sentiment that although p values may have their drawbacks, alternative procedures are complicated and arbitrary. Nickerson (2000, p. 290) graphically summarized the situation as follows: "NHST surely has warts, but so do all the alternatives."

TOWARD AN ALTERNATIVE TO NHST

Desiderata for Principled and Practical Alternatives to *p* Values

Several methods for inference are able to replace p values. The list of alternatives includes Bayesian procedures, Bayesian–frequentist compromises (see, e.g., Berger, 2003; Berger, Boukai, & Wang, 1997; Berger, Brown, & Wolpert, 1994; Good, 1983), Akaike's information criterion (AIC; e.g., Akaike, 1974; Burnham & Anderson, 2002), cross-validation (e.g., Browne, 2000; Geisser, 1975; Stone, 1974), bootstrap methods (e.g., Efron & Tibshirani, 1997), prequential methods (e.g., Dawid, 1984; Wagenmakers, Grünwald, & Steyvers, 2006), and methods based on the principle of minimum description length (MDL; e.g., Grünwald, 2000; Grünwald, Myung, & Pitt, 2005; Pitt, Myung, & Zhang, 2002; Rissanen, 2001). These alternative methods are all *model selection methods*, in that the explicit or implicit goal is to compare different models and select the best one (for applications of model selection in the field of psychology, see two special issues of the *Journal of Mathematical Psychology*: Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006). Thus, model selection methods do not assess the adequacy of H_0 in isolation. Rather, the adequacy of H_0 is compared with the adequacy of an alternative model, H_1, automatically avoiding the negative consequences that arise when the focus is solely on H_0.

All model selection methods require a quantification of model adequacy. This concept is well defined. An ideal model extracts from a given data set only those features that are replicable. This ideal model will therefore yield the best prediction for unseen data from the same source. When a model has too few parameters, it is unable to de-

scribe all the replicable features in a data set. Hence, the model *underfits* the data and yields suboptimal predictions. When a model has too many parameters, it is too powerful and captures not only the replicable features in the data, but also captures idiosyncratic random fluctuations. Hence, the model *overfits* the data, and since its parameter estimates are contaminated by noise, predictive performance will suffer. Thus, the universal yardstick for selecting between competing models is predictive performance.

In experimental psychology, model selection procedures are mostly used to adjudicate between nonnested complicated nonlinear models of human cognition. There is no reason, however, why these procedures could not be applied to run-of-the-mill statistical inference problems involving nested linear models such as ANOVA. Consider, for instance, normally distributed data with a known variance and unknown mean, $D \sim N(\mu, \sigma^2 = 1)$. Under the null hypothesis, μ is fixed at 0, whereas under the alternative hypothesis, μ is a free parameter. It is clear that, on account of its extra free parameter, H_1 will always provide a better fit to the data than will H_0. Thus, goodness of fit alone cannot be used to select between H_0 and H_1. Instead, model selection methods can be used to assess whether the decrease in model parsimony is warranted by the associated increase in goodness of fit. The tradeoff between parsimony and goodness of fit is quantified by the criterion of minimizing prediction error (Myung, 2000; Myung, Navarro, & Pitt, 2006; Myung & Pitt, 1997; Wagenmakers et al., 2006). Thus, the problem of hypothesis testing can be profitably viewed as a problem of model selection (M. D. Lee & Pope, 2006): When the prediction error associated with H_1 is lower than that associated with H_0, the data support H_1 over H_0. From now on, the terms *hypothesis testing* and *model selection* will be used interchangeably.

Before choosing among the many attractive model selection alternatives to *p* value hypothesis testing, it is useful to consider a number of desirable properties that an alternative method for inference should possess. A list of theoretical desiderata that address the limitations of NHST *p* values discussed earlier should arguably contain at least the following elements:

1. Ideally, a statistical procedure should depend only on data that were actually observed. Data that could have been observed but were not are irrelevant for the situation at hand.
2. Ideally, the results of a statistical procedure should not depend on the unknown intentions of the researcher.
3. Ideally, a statistical procedure should provide a measure of evidence that takes into account both the null hypothesis and the alternative hypothesis.

A consequence of the third desideratum is that an ideal procedure should be able to quantify evidential support in favor of the null hypothesis. This requirement may seem self-evident, but note that Fisherian *p* values are not designed to quantify support in favor of the null hypothesis. A *p* value indicates the evidence against the null hypothesis. It is not possible to observe the data and corroborate the null hypothesis; one can only fail to reject it. Hence, the null hypothesis exists only in a state of suspended disbelief. The APA Task Force on Statistical Inference underscored this point by issuing the warning "Never use the unfortunate expression 'accept the null hypothesis'" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599).

What is unfortunate here is not the expression, but rather the fact that Fisherian *p* values are incapable of providing support for the null hypothesis. This sets Fisherian *p* values apart from all of the model selection methods mentioned above. The practical implications are substantial. For instance, experiments may be designed in order to test theories that predict no difference between experimental conditions. A colleague or reviewer may later note that such an experiment is flawed from the outset, since it hinges on the acceptance of the null hypothesis, something that is not possible in the Fisherian NHST framework. Such a state of affairs not only impedes theoretical progress, but also has serious practical ramifications. Consider, for instance, the statistical problem of determining whether the glass found on the coat of a suspect in a criminal case is the same as that from a broken window (see Shafer, 1982). From the characteristics of the pieces of glass (i.e., the refractive index), the prosecution wants to claim that the glass on the coat matches that of the window; that is, the prosecution is interested in supporting the null hypothesis that the refractive indices do not differ. Of course, within the framework of Fisherian NHST, one may calculate all kinds of measures to make the problematic conclusion "the null hypothesis is true" more digestible, for instance by calculating power, reporting effect sizes, and so on, but such calculations do not produce what one wants to know—namely, the evidence that the data provide in favor of the null hypothesis.

To the theoretical desiderata above one might then add desiderata of a more pragmatic nature:

4. Ideally, a statistical procedure should be very easy to implement.
5. Ideally, a statistical procedure should be "objective," in the sense that different researchers, who have the same models and are confronted with the same data, will draw the same conclusions.

These pragmatic desiderata are motivated by the fact that the majority of experimental psychologists have only had very limited training in mathematical statistics. This is why methods for statistical inference that require much computer programming and independent mathematical thought are unlikely to gain much popularity in the field.

Selecting a Method for Model Selection

Subjective and objective Bayesian model selection methods satisfy theoretical Desiderata 1–3. Consistent with Desideratum 1, the Bayesian hypothesis test is conditioned on the observed data only (see Equation 7). Consistent with Desideratum 2, the intention of the researcher

performing the experiment (i.e., the stopping rule) is irrelevant for Bayesian inference. This is shown mathematically in the online appendix on my Web site. Finally, from Equations 7 and 8 it is evident that the Bayesian hypothesis test compares the plausibility of H_0 versus H_1 and quantifies the result by an odds ratio (see Table 3). This is consistent with Desideratum 3.

With respect to other methods of model selection, the MDL principle integrates over the sample space and therefore violates Desiderata 1 and 2 (see Myung et al., 2006; Wallace & Dowe, 1999).[6] Depending on the implementation, the prequential methodology is either identical to Bayesian model selection or asymptotically equivalent to it (see Wagenmakers et al., 2006). Bootstrap and cross-validation methods are promising alternatives, but do not generally quantify evidence in terms of probability. The AIC follows the *p* postulate, in that it assumes that the weight of statistical evidence is not influenced by the size of the data set. For instance, application of AIC to the data from Figure 6 produces a line that is almost horizontally straight at .284. A Bayesian analysis showed that for any plausible prior, the *p* postulate is false: For a fixed *p* value, the evidence in favor of H_0 increases with the number of observations.

Unfortunately, subjective and objective Bayesian methods do not satisfy pragmatic Desideratum 4 (i.e., ease of application), and only the objective Bayesian method satisfies Desideratum 5 (i.e., objectivity). The true–false example discussed earlier is easy enough, but it is one of a limited set of hypothesis testing problems for which analytic expressions have been worked out. For other problems, integration over the parameter space has to be achieved using MCMC methods (see Kass & Raftery, 1995; Raftery, 1996). For someone with limited quantitative training, mastering a method such as Bayesian hypothesis testing using MCMC may take years. Contrast this with the "point-and-click" NHST approach implemented in computer packages such as SPSS. These programs have completely automatized the inference process, requiring nothing of the user except the input of the data and the very general specification of a model. Using the point-and-click approach, most experimental psychologists are able to arrive at a statistical conclusion in a matter of minutes.

Thus, we are faced with a dilemma. On the one hand, the Bayesian methodology successfully addresses the limitations of the *p* value methodology (see Desiderata 1–3, and the earlier discussion on Bayesian methodology). On the other hand, a commitment to the Bayesian hypothesis tests requires an investment of time and energy that most experimental psychologists are unwilling to make. The solution to the dilemma is to forgo the full-fledged objective Bayesian hypothesis test and settle instead for an accurate and easy-to-calculate approximation to the Bayesian hypothesis test.

BAYESIAN HYPOTHESIS TESTING USING THE BIC APPROXIMATION

Some experimental psychologists are already familiar with the BIC (for foundations, see, e.g., Hannan, 1980; Kass, 1993; Kass & Raftery, 1995; Kass & Wasserman, 1995; Pauler, 1998; Raftery, 1995; Schwarz, 1978; Smith & Spiegelhalter, 1980; Wasserman, 2000; for applications in structural equation modeling, see Raftery, 1993; for critical discussion, see, e.g., Burnham & Anderson, 2002, as well as Firth & Kuha, 1999 [a special issue of *Sociological Methods & Research*]; Gelman & Rubin, 1999; Raftery, 1999; Weakliem, 1999; Winship, 1999; Xie, 1999). However, most experimental psychologists use the BIC only to compare nonnested models, and then carry out the comparison in a very rough qualitative fashion. That is, when Model A has a lower BIC value than Model B, Model A is simply preferred over Model B. Unfortunately, the extent of this preference is almost never quantified on a scale that researchers can easily intuit (see Wagenmakers & Farrell, 2004).

What is often not realized is that the difference between BIC values transforms to an approximation of the Bayes factor; hence, the BIC can be used for nested and nonnested models alike. The latter feature is nicely illustrated by Vickers, Lee, Dry, and Hughes (2003), who conducted BIC-style statistical inference to approximate the Bayes factor for a set of six nested and nonnested models. Furthermore, assuming the models under consideration are equally plausible a priori, a comparison of their BIC values easily yields an approximation of their posterior probabilities. For instance, in the case of comparing a null hypothesis (e.g., $\mu = 0$) to an alternative hypothesis (e.g., $\mu \neq 0$), one option is to report that, say, $\mathrm{BIC}(H_0) = 343.46$ and $\mathrm{BIC}(H_1) = 341.58$, and conclude that H_1 is "better" than H_0. Another option is to transform these values to posterior probabilities and to report that $\Pr_{\mathrm{BIC}}(H_0 \mid D) \approx .28$, and consequently that $\Pr_{\mathrm{BIC}}(H_1 \mid D) \approx .72$. The latter method of presenting the statistical conclusions is much more insightful.

Appendix B shows how the BIC approximation may be derived (for more details, see two excellent articles by Raftery: 1995, 1999). Here, only the end result of this derivation is given. The BIC for model H_i is defined as

$$\mathrm{BIC}(H_i) = -2\log L_i + k_i \log n, \tag{9}$$

where n is the number of observations, k_i is the number of free parameters of model H_i, and L_i is the maximum likelihood for model H_i—that is, $L_i = c\Pr(D \mid \hat{\theta}, H_i)$, with c an arbitrary constant (A.W. F. Edwards, 1992). The BIC approximation to the prior predictive probability $\Pr(D \mid H_i)$ may then be obtained by the following simple transformation: $\Pr_{\mathrm{BIC}}(D \mid H_i) = \exp[-\mathrm{BIC}(H_i)/2]$. In the case of two models, H_0 and H_1, the Bayes factor is defined as the ratio of the prior predictive probabilities; hence, the BIC approximation of the Bayes factor is given by

$$BF_{01} \approx \frac{\Pr_{\mathrm{BIC}}\left(D \mid H_0\right)}{\Pr_{\mathrm{BIC}}\left(D \mid H_1\right)} = \exp\left(\Delta\mathrm{BIC}_{10}/2\right), \tag{10}$$

where $\Delta\mathrm{BIC}_{10} = \mathrm{BIC}(H_1) - \mathrm{BIC}(H_0)$. For instance, if data from an experiment yielded $\mathrm{BIC}(H_0) = 1{,}211.0$ and $\mathrm{BIC}(H_1) = 1{,}216.4$, the Bayes factor in favor of H_0 would be $\exp(5.4/2) \approx 14.9$. With equal priors on the models, this would amount to a posterior probability of H_0 of

14.9/15.9 ≈ .94. According to Table 3, this would constitute "positive" evidence for H_0. Note that in the Fisherian tradition of *p* value hypothesis testing, one can never reach this conclusion: One can fail to reject the null hypothesis, but the null hypothesis can never be accepted.

Now consider a similar experiment that finds $\mathrm{BIC}(H_0) = 1{,}532.4$ and $\mathrm{BIC}(H_1) = 1{,}534.2$. For this experiment, the Bayes factor in favor of H_0 equals $\exp(1.8/2) \approx 2.5$. The Bayes factors from these two experiments can be combined into an overall Bayes factor by simple multiplication: $BF_{01}(\text{total}) = 14.9 \times 2.5 = 37.25$. This corresponds to a posterior probability of H_0 of 37.25/38.25 ≈ .97, which according to Table 3 constitutes "strong" evidence for H_0.

In the case of *k* models, each of which is a priori equally plausible, the posterior probability of a particular model H_i is obtained by the following transformation (see, e.g., Wasserman, 2000):

$$\Pr_{\mathrm{BIC}}\left(H_i \mid D\right) = \frac{\exp\left[-\frac{1}{2}\mathrm{BIC}\left(H_i\right)\right]}{\sum_{j=0}^{k-1}\exp\left[-\frac{1}{2}\mathrm{BIC}\left(H_j\right)\right]}. \qquad (11)$$

In the case of two models, H_0 and H_1, the posterior probability of H_0 is given by

$$\Pr_{\mathrm{BIC}}\left(H_0 \mid D\right) = \frac{1}{1+\exp\left(-\frac{1}{2}\Delta\mathrm{BIC}_{10}\right)}. \qquad (12)$$

This means that the posterior probability of a null hypothesis is a logistic or sigmoid function of half the BIC difference between the null hypothesis H_0 and the alternative hypothesis H_1.

The sigmoid function is shown in Figure 7, from which it is apparent that BIC differences in excess of 12 or −12 constitute very strong evidence. For instance, when $\Delta\mathrm{BIC}_{10} = 12$, the posterior probability for H_0 is close to 1—that is, $\Pr_{\mathrm{BIC}}(H_0 \mid \Delta\mathrm{BIC}_{10} = 12) = 1/[1 + \exp(-6)] \approx .998$. Note that the absolute values of the BIC are irrelevant—only the differences in BICs carry evidential weight. Recall also that the function shown in Figure 7 incorporates the effects of sample size and the number of free parameters, and that it holds regardless of whether H_0 is nested within H_1 or not.

As discussed previously, the Bayes factor is sensitive to the shape of the prior distribution. The equations that involve BIC do not, however, appear to specify any prior, and this may lead one to wonder how the BIC can then be used to approximate the Bayes factor. As is explained in Appendix B, one prior that is consistent with the BIC approximation is the "unit information prior." This prior contains as much information as does a single observation (Raftery, 1999). In the following discussion, I assume that the prior that the BIC implicitly assumes is the unit information prior.

Advantages and Disadvantages of the BIC Approximation

The BIC approximation of the Bayesian hypothesis test is attractive for two reasons. First, the BIC approximation does not require the researcher to specify his or her own prior distribution. This ensures that the BIC is "objective," in the sense that different researchers, confronted with the same data, will draw the same statistical inference from the same set of models. The objectivity inherent in the BIC is especially appealing to those who feel that the specification of priors is subjective and therefore inherently unscientific. However, as pointed out by Xie (1999, p. 429), the "drawbacks of the BIC are precisely where its virtues lie," and it has been argued that the unit information prior that the BIC implicitly assumes is often too wide. A prior that is too wide decreases the prior predictive probability of the alternative hypothesis, and therefore makes the null hypothesis appear more plausible than it actually is. In other words, it has been argued that more information needs to be injected in the prior distribution. In response, Raftery (1999) pointed out that when the BIC does not support H_1, whereas subjective Bayesian methods do support H_1, the difference must be due to the specific prior that is assumed. The BIC can therefore be viewed as providing an objective baseline reference for automatic Bayesian hypothesis testing. In sum, the drawback of the BIC is that it does not incorporate substantive information into its implicit prior distribution; the virtue of the BIC is that the specification of the prior distribution is completely automatic.

A second advantage of the BIC approximation is that it is particularly easy to compute. For some models, popular statistical computer programs already provide the raw BIC numbers, so that in order to perform an approximate Bayesian hypothesis test, one only needs to transform these

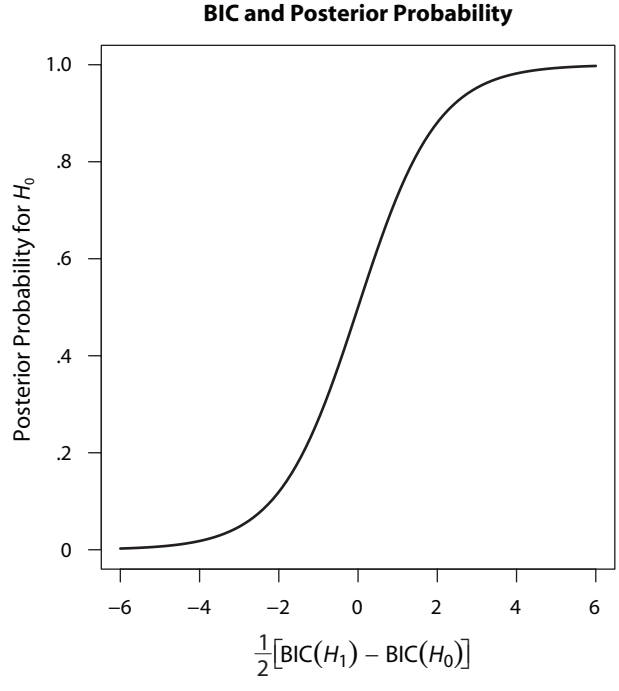

Figure 7. The posterior probability of two a priori equally plausible models is a sigmoid function of half their difference in BIC. See the text for details.

numbers to posterior probabilities (see Equation 12). For other models, such as the ones used in standard ANOVA, the BIC can be readily computed from the sums of squared errors (see Glover & Dixon, 2004; Raftery, 1995); this transformation will be illustrated in Example 9.

The BIC also has a few important limitations. One such limitation is that its approximation ignores the functional form of the model parameters, focusing exclusively on the number of free parameters. Jay Myung, Mark Pitt, and coworkers have shown on multiple occasions that models with the same number of parameters may differ in complexity (see, e.g., M. D. Lee, 2002; Myung, 2000; Myung & Pitt, 1997; Pitt et al., 2002; see also Djurić, 1998; Wagenmakers, 2003; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). For example, Stevens's law of psychophysics can handle both decelerating and accelerating functions, whereas Fechner's law can only account for decelerating functions (see Pitt et al., 2002). Thus, Stevens's law is a priori more flexible, despite the fact that it has just as many parameters as Fechner's law. A full-fledged Bayesian analysis is sensitive to the functional form of the parameters because it averages the likelihood across the entire parameter space. Thus, when the data follow a decelerating function, the prior predictive probability of Stevens's law suffers from the fact that part of its parameter space is dedicated to accelerating functions; therefore, in that part of parameter space, the observed data are very unlikely indeed. Although the issue of functional form is important, it is much more important in complicated nonlinear models than it is in standard linear statistical models such as linear regression and ANOVA.

A final note of caution concerns the number of observations n in Equation 9. In some situations, this number may be difficult to determine (for a discussion and examples, see Raftery, 1995, p. 135). For instance, consider an experiment that has 10 subjects, each of whom contributes 20 observations to each of two conditions. In such a hierarchical or multilevel design, it is not quite clear what n should be. In this case, the standard choice is to take n to be the number of subjects.

The Relation Between BIC and ANOVA

The wide majority of statistical analyses in experimental psychology concern the normal linear model. In the case of linear regression with normal errors, the BIC for model H_i can be written (Raftery, 1995) as

$$\mathrm{BIC}(H_i) = n\log(1 - R_i^2) + k_i \log n, \qquad (13)$$

where $1 - R_i^2$ is the proportion of the variance that model H_i fails to explain. Because ANOVA is a special case of regression, the same equation also holds for ANOVA. The proportion of unexplained variance is readily obtained from tables of sums of squares, as $1 - R_i^2 = \mathrm{SSE}_i/\mathrm{SS}_{\mathrm{total}}$, where SSE_i is the sum of squared errors for model H_i.

When comparing a null hypothesis H_0 with an alternative hypothesis H_1, the common $\mathrm{SS}_{\mathrm{total}}$ factor cancels in the likelihood ratio, and the difference in BIC values is given by

$$\Delta\mathrm{BIC}_{10} = n\log\left(\frac{\mathrm{SSE}_1}{\mathrm{SSE}_0}\right) + \left(k_1 - k_0\right)\log n. \qquad (14)$$

Example 9. How to calculate the BIC from SPSS output (Glover & Dixon, 2004). In a recent article, Glover and Dixon demonstrated with several examples how the output of standard statistical software packages such as SPSS may be used to compare the maximum-likelihood fit of H_0 against that of H_1. To illustrate how easily SPSS output can be transformed to BIC-based estimation of posterior probabilities, I will discuss just one example from the Glover and Dixon article.

Consider a hypothetical recall experiment that has 40 participants in a 2×2 design. The two manipulations concern the amount of previous study ("few" versus "many" study trials) and the concreteness of the word stimuli that need to be recalled ("abstract" versus "concrete"). The hypothesis of interest concerns a possible interaction between the two factors. For instance, it may be expected that recall for both abstract and concrete words increases with additional study trials, but that this increase is more pronounced for abstract words. This hypothesis would be contrasted with a null hypothesis in which the two factors do not interact.

Figure 8 shows the means for the synthesized data, as well as the fits of the interaction model (left panel) and the additive model (right panel). The error bars reflect the standard errors calculated from the residual variation. The interaction model fits the means of the data exactly, whereas the additive model does not. However, the additive model is simpler than the interaction model, and the key question is whether the simplicity of the additive model can compensate for its lack of fit.

Table 4 shows an ANOVA table for the synthesized data set. The sum of squares that are not explained by the interaction model is the error sum of squares (i.e., 470.1). The sum of squares that are not explained by the additive

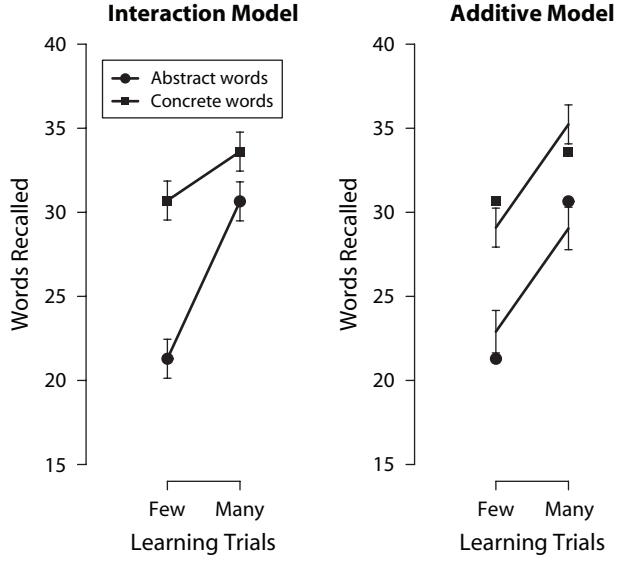

Figure 8. Comparison of an interaction model (left panel) and an additive model (right panel) in accounting for the effects of word type in a hypothetical data set (see Glover & Dixon, 2004, p. 798, Figure 3).

Table 4
ANOVA Table for the Synthesized Data From Glover and Dixon (2004, Table 2)

Source	*df*	Sums of Squares	Mean Square	*F*	*p*
Trials (T)	1	383.1	383.1	29.34	≤.0001
Concreteness (C)	1	376.9	376.9	28.86	≤.0001
T × C	1	104.1	104.1	7.97	.0077
Error	36	470.1	13.1		
Total	39	1,334.2			

model is the error sum of squares plus the sum of squares associated with the interaction effect (i.e., 470.1 + 104.1 = 574.2). Thus, $SSE_1 = 470.1$, $SSE_0 = 574.2$, $n = 40$, and $k_1 - k_0 = 1$, since the interaction model has one parameter more than the additive model. Plugging this information into Equation 14 yields

$$\Delta BIC_{10} = 40 \log\left(\frac{470.1}{574.2}\right) + \log 40 \approx -4.31.$$

The fact that this number is negative indicates that the BIC is higher for H_0 than it is for H_1; hence, according to the BIC, the interaction model is to be preferred over the additive model. The extent of the preference can be quantified precisely by plugging in the −4.31 number into Equation 12. Assuming equal prior plausibility of the interaction model and the additive model, Equation 12 yields $Pr_{BIC}(H_0 \mid D) = 1/[1 + \exp(2.16)] \approx .10$. This means that the posterior probability for the interaction model is about .90, which according to Table 3 constitutes "positive" evidence.

CONCLUDING COMMENTS

The field of experimental psychology uses *p* values as the main vehicle for statistical inference. The statistical literature, however, reveals a number of fundamental problems with *p* values. The primary goal of this article was to review three such problems with *p* values: The *p* value is calculated from imaginary data, is based on subjective intentions, and does not quantify statistical evidence. Many researchers are not aware of these problems, and the many examples in this article demonstrate how such ignorance can have serious practical consequences.

The solution to the problem of statistical inference in psychology is to switch from the *p* value methodology to a model selection methodology. Model selection methods assess which of several models, for instance H_0 and H_1, provides the best explanation of the data. In psychology, model selection methods are generally employed only for nonlinear and nonnested models, but there is no reason why these methods should not be applied to nested models. Unfortunately, model selection methods are generally somewhat more involved than *p* values, and for many experimental psychologists this will take away some of their initial attraction. Fortunately, it is possible to have one's inferential cake and eat it too. Although the objective Bayesian hypothesis test is not trivial to implement, its BIC approximation can be easily calculated from standard output (e.g., SPSS output; see Glover & Dixon, 2004). Under equal priors, a simple transformation then yields the posterior probability of H_0 (see Equation 12). Most importantly, in contrast to *p* values, the Bayesian hypothesis test does not depend on imaginary data, is insensitive to subjective intentions, and does quantify statistical evidence. It should also be noted that it is straightforward to equip Bayesian models with utility functions (see, e.g., Berger, 1985; Bernardo & Smith, 1994) so that—if desired—one can make the discrete decision that has the highest expected utility (see Killeen, 2006).

In the psychological literature on visual perception, memory, reasoning, and categorization, Bayesian methods are often used because they provide an optimal standard against which to compare human performance. I suggest that in addition to modeling an optimal decision maker through the laws of probability calculus, one can actually become an optimal decision maker and apply probability calculus to problems of statistical inference. Instead of reporting what nobody wants to know, namely "the probability of encountering a value of a test statistic that is as least as extreme as the one that is actually observed, given that the null hypothesis is true," psychologists can easily report what everybody wants to know: the strength of the evidence that the observed data provide for and against the hypotheses under consideration.

AUTHOR NOTE

This research was supported by a Veni Grant from the Dutch Organization for Scientific Research (NWO). I thank Scott Brown, Peter Dixon, Simon Farrell, Raoul Grasman, Geoff Iverson, Michael Lee, Martijn Meeter, Jay Myung, Jeroen Raaijmakers, Jeff Rouder, and Rich Shiffrin for helpful comments on earlier drafts of this article. Mark Steyvers convinced me that this article would be seriously incomplete without a consideration of practical alternatives to the *p* value methodology. I am grateful to Scott Glover and Peter Dixon for providing the synthetic data shown in Figure 8. Correspondence concerning this article may be addressed to E.-J. Wagenmakers, University of Amsterdam, Department of Psychology, Methodology Unit, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands (e-mail: ej.wagenmakers@gmail.com; URL: users.fmg.uva.nl/ewagenmakers/).

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

ANSCOMBE, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, **10**, 89-100.

ANSCOMBE, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association*, **58**, 365-383.

ARMITAGE, P. (1957). Restricted sequential procedures. *Biometrika*, **44**, 9-26.

ARMITAGE, P. (1960). *Sequential medical trials*. Springfield, IL: Thomas.

ARMITAGE, P., MCPHERSON, C. K., & ROWE, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A*, **132**, 235-244.

BAKAN, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, **66**, 423-437.

BARNARD, G. A. (1947). The meaning of a significance level. *Biometrika*, **34**, 179-182.

BASU, D. (1964). Recovery of ancillary information. *Sankhya: Series A*, **26**, 3-16.

BAYARRI, M.-J., & BERGER, J. O. (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.

BERGER, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, **18**, 1-32.

BERGER, J. O., & BERRY, D. A. (1988a). The relevance of stopping rules in statistical inference. In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics IV* (Vol. 1, pp. 29-72). New York: Springer.

BERGER, J. O., & BERRY, D. A. (1988b). Statistical analysis and the illusion of objectivity. *American Scientist*, **76**, 159-165.

BERGER, J. O., BOUKAI, B., & WANG, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science*, **12**, 133-160.

BERGER, J. O., BROWN, L., & WOLPERT, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *Annals of Statistics*, **22**, 1787-1807.

BERGER, J. O., & DELAMPADY, M. (1987). Testing precise hypotheses. *Statistical Science*, **2**, 317-352.

BERGER, J. O., & MORTERA, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, **94**, 542-554.

BERGER, J. O., & PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

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

BERGER, J. O., & WOLPERT, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward, CA: Institute of Mathematical Statistics.

BERNARDO, J. M., & SMITH, A. F. M. (1994). *Bayesian theory*. Chichester, U.K.: Wiley.

BIRNBAUM, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, **53**, 259-326.

BIRNBAUM, A. (1977). The Neyman–Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley–Savage argument for Bayesian theory. *Synthese*, **36**, 19-49.

BOX, G. E. P., & TIAO, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

BROWNE, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108-132.

BURDETTE, W. J., & GEHAN, E. A. (1970). *Planning and analysis of clinical studies*. Springfield, IL: Thomas.

BURNHAM, K. P., & ANDERSON, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

BUSEMEYER, J. R., & STOUT, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, **14**, 253-262.

CHRISTENSEN, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *American Statistician*, **59**, 121-126.

COHEN, J. (1994). The earth is round ($p < .05$). *American Psychologist*, **49**, 997-1003.

CORNFIELD, J. (1966). Sequential trials, sequential analysis, and the likelihood principle. *American Statistician*, **20**, 18-23.

CORNFIELD, J. (1969). The Bayesian outlook and its application. *Biometrics*, **25**, 617-657.

CORTINA, J. M., & DUNLAP, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, **2**, 161-172.

COX, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, **29**, 357-372.

COX, D. R. (1971). The choice between alternative ancillary statistics. *Journal of the Royal Statistical Society: Series B*, **33**, 251-255.

COX, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, **14**, 1-13.

CUMMING, G. (2007). *Replication and* p *values:* p *values predict the future vaguely, but confidence intervals do better*. Manuscript submitted for publication.

D'AGOSTINI, G. (1999). Teaching statistics in the physics curriculum: Unifying and clarifying role of subjective probability. *American Journal of Physics*, **67**, 1260-1268.

DAWID, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A*, **147**, 278-292.

DE FINETTI, B. (1974). *Theory of probability: A critical introductory treatment* (Vols. 1 & 2; A. Machí & A. Smith, Trans.). London: Wiley.

DIAMOND, G. A., & FORRESTER, J. S. (1983). Clinical trials and statistical verdicts: Probable grounds for appeal. *Annals of Internal Medicine*, **98**, 385-394.

DICKEY, J. M. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *Journal of the Royal Statistical Society: Series B*, **35**, 285-305.

DICKEY, J. M. (1977). Is the tail area useful as an approximate Bayes factor? *Journal of the American Statistical Association*, **72**, 138-142.

DIXON, P. (2003). The p value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, **57**, 189-202.

DJURIĆ, P. M. (1998). Asymptotic MAP criteria for model selection. *IEEE Transactions on Signal Processing*, **46**, 2726-2735.

EDWARDS, A. W. F. (1992). *Likelihood*. Baltimore: Johns Hopkins University Press.

EDWARDS, W., LINDMAN, H., & SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193-242.

EFRON, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, **100**, 1-5.

EFRON, B., & TIBSHIRANI, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548-560.

FELLER, W. (1940). Statistical aspects of ESP. *Journal of Parapsychology*, **4**, 271-298.

FELLER, W. (1970). *An introduction to probability theory and its applications: Vol. 1* (2nd ed.). New York: Wiley.

FINE, T. L. (1973). *Theories of probability: An examination of foundations*. New York: Academic Press.

FIRTH, D., & KUHA, J. (1999). Comments on "A critique of the Bayesian information criterion for model selection." *Sociological Methods & Research*, **27**, 398-402.

FISHER, R. A. (1934). *Statistical methods for research workers* (5th ed.). London: Oliver & Boyd.

FISHER, R. A. (1935a). *The design of experiments*. Edinburgh: Oliver & Boyd.

FISHER, R. A. (1935b). The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society*, **98**, 39-82.

FISHER, R. A. (1958). *Statistical methods for research workers* (13th ed.). New York: Hafner.

FREIREICH, E. J., GEHAN, E., FREI, E., III, SCHROEDER, L. R., WOLMAN, I. J., ANBARI, R., ET AL. (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood*, **21**, 699-716.

FRICK, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, **1**, 379-390.

FRIEDMAN, L. M., FURBERG, C. D., & DEMETS, D. L. (1998). *Fundamentals of clinical trials* (3rd ed.). New York: Springer.

GALAVOTTI, M. C. (2005). *A philosophical introduction to probability*. Stanford: CSLI Publications.

GEISSER, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320-328.

GELMAN, A., & RUBIN, D. B. (1999). Evaluating and using statistical methods in the social sciences. *Sociological Methods & Research*, **27**, 403-410.

GIGERENZER, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

GIGERENZER, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral & Brain Sciences*, **21**, 199-200.

GILKS, W. R., RICHARDSON, S., & SPIEGELHALTER, D. J. (EDS.) (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.

GILL, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: CRC Press.

GLOVER, S., & DIXON, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, **11**, 791-806.

GOOD, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.

GOOD, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2: Proceedings of the Second Valencia International Meeting, September 6/10, 1983* (pp. 249-269). Amsterdam: North-Holland.

GOODMAN, S. N. (1993). *p* values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, **137**, 485-496.

GRÜNWALD, P. [D.] (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133-152.

GRÜNWALD, P. D., MYUNG, I. J., & PITT, M. A. (EDS.) (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

HACKING, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.

HAGEN, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, **52**, 15-24.

HALDANE, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, **33**, 222-225.

HANNAN, E. J. (1980). The estimation of the order of an ARMA process. *Annals of Statistics*, **8**, 1071-1081.

HELLAND, I. S. (1995). Simple counterexamples against the conditionality principle. *American Statistician*, **49**, 351-356.

HILL, B. M. (1985). Some subjective Bayesian considerations in the selection of models. *Econometric Reviews*, **4**, 191-246.

HOWSON, C., & URBACH, P. (2005). *Scientific reasoning: The Bayesian approach* (3rd. ed.). Chicago: Open Court.

HUBBARD, R., & BAYARRI, M.-J. (2003). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *American Statistician*, **57**, 171-182.

JAYNES, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science & Cybernetics*, **4**, 227-241.

JAYNES, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

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

JENNISON, C., & TURNBULL, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science*, **5**, 299-317.

KADANE, J. B., SCHERVISH, M. J., & SEIDENFELD, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, **91**, 1228-1235.

KARABATSOS, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*, **50**, 123-148.

KASS, R. E. (1993). Bayes factors in practice. *Statistician*, **42**, 551-560.

KASS, R. E., & RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 377-395.

KASS, R. E., & WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928-934.

KASS, R. E., & WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343-1370.

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

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

KILLEEN, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, **13**, 549-562.

KLUGKIST, I., LAUDY, O., & HOIJTINK, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, **10**, 477-493.

LEE, M. D. (2002). Generating additive clustering models with limited stochastic complexity. *Journal of Classification*, **19**, 69-85.

LEE, M. D., & POPE, K. J. (2006). Model selection for the rate problem: A comparison of significance testing, Bayesian, and minimum description length statistical inference. *Journal of Mathematical Psychology*, **50**, 193-202.

LEE, M. D., & WAGENMAKERS, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, **112**, 662-668.

LEE, P. M. (1989). *Bayesian statistics: An introduction*. New York: Oxford University Press.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187-192.

LINDLEY, D. V. (1972). *Bayesian statistics: A review*. Philadelphia: Society for Industrial & Applied Mathematics.

LINDLEY, D. V. (1977). The distinction between inference and decision. *Synthese*, **36**, 51-58.

LINDLEY, D. V. (1982). Scoring rules and the inevitability of probability. *International Statistical Review*, **50**, 1-26.

LINDLEY, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, **15**, 22-25.

LINDLEY, D. V. (2004). That wretched prior. *Significance*, **1**, 85-87.

LINDLEY, D. V., & PHILLIPS, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *American Statistician*, **30**, 112-119.

LINDLEY, D. V., & SCOTT, W. F. (1984). *New Cambridge elementary statistical tables*. Cambridge: Cambridge University Press.

LOFTUS, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, **5**, 161-171.

LOFTUS, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In H. Pashler (Ed. in Chief) & J. Wixted (Vol. Ed.), *Stevens' Handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 339-390). New York: Wiley.

LUDBROOK, J. (2003). Interim analyses of data as they accumulate in laboratory experimentation. *BMC Medical Research Methodology*, **3**, 15.

MCCARROLL, D., CRAYS, N., & DUNLAP, W. P. (1992). Sequential ANOVAs and Type I error rates. *Educational & Psychological Measurement*, **52**, 387-393.

MYUNG, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190-204.

MYUNG, I. J., FORSTER, M. R., & BROWNE, M. W. (EDS.) (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, **44**(1).

MYUNG, I. J., NAVARRO, D. J., & PITT, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, **50**, 167-179.

MYUNG, I. J., & PITT, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

NELSON, N., ROSENTHAL, R., & ROSNOW, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, **41**, 1299-1301.

NEYMAN, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, **36**, 97-131.

NEYMAN, J., & PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society: Series A*, **231**, 289-337.

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

O'HAGAN, A. (1997). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B*, **57**, 99-138.

O'HAGAN, A. (2004). Dicing with the unknown. *Significance*, **1**, 132-133.

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

PAULER, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika*, **85**, 13-27.

PETO, R., PIKE, M. C., ARMITAGE, P., BRESLOW, N. E., COX, D. R., HOWARD, S. V., ET AL. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design. *British Journal of Cancer*, **34**, 585-612.

PITT, M. A., MYUNG, I. J., & ZHANG, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.

POCOCK, S. J. (1983). *Clinical trials: A practical approach*. New York: Wiley.

PRATT, J. W. (1961). [Review of Lehmann, E. L., Testing statistical hypotheses]. *Journal of the American Statistical Association*, **56**, 163-167.

PRATT, J. W. (1962). On the foundations of statistical inference: Discussion. *Journal of the American Statistical Association*, **57**, 314-315.

RAFTERY, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 163-180). Newbury Park, CA: Sage.

RAFTERY, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111-196). Cambridge, MA: Blackwell.

RAFTERY, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 163-187). Boca Raton, FL: Chapman & Hall/CRC.

RAFTERY, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research*, **27**, 411-427.

RISSANEN, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, **47**, 1712-1717.

ROBERT, C. P., & CASELLA, G. (1999). *Monte Carlo statistical methods*. New York: Springer.

ROSENTHAL, R., & GAITO, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, **55**, 33-38.

ROUDER, J. N., & LU, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, **12**, 573-604.

ROUDER, J. N., LU, J., SPECKMAN, P., SUN, D., & JIANG, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, **12**, 195-223.

ROYALL, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.

SAVAGE, L. J. (1954). *The foundations of statistics*. New York: Wiley.

SCHERVISH, M. J. (1996). *P* values: What they are and what they are not. *American Statistician*, **50**, 203-206.

SCHMIDT, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, **1**, 115-129.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

SELLKE, T., BAYARRI, M.-J., & BERGER, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *American Statistician*, **55**, 62-71.

SHAFER, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, **77**, 325-351.

SIEGMUND, D. (1985). *Sequential analysis: Tests and confidence intervals*. New York: Springer.

SMITH, A. F. M., & SPIEGELHALTER, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society: Series B*, **42**, 213-220.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society: Series B*, **36**, 111-147.

STRUBE, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods*, **38**, 24-27.

STUART, A., ORD, J. K., & ARNOLD, S. (1999). *Kendall's advanced theory of statistics: Vol. 2A. Classical inference and the linear model* (6th ed.). London: Arnold.

TRAFIMOW, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, **110**, 526-535.

VICKERS, D., LEE, M. D., DRY, M., & HUGHES, P. (2003). The roles of the convex hull and the number of potential intersections in performance on visually presented traveling salesperson problems. *Memory & Cognition*, **31**, 1094-1104.

WAGENMAKERS, E.-J. (2003). How many parameters does it take to fit an elephant? [Book review]. *Journal of Mathematical Psychology*, **47**, 580-586.

WAGENMAKERS, E.-J., & FARRELL, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, **11**, 192-196.

WAGENMAKERS, E.-J., & GRÜNWALD, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science*, **17**, 641-642.

WAGENMAKERS, E.-J., GRÜNWALD, P., & STEYVERS, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, **50**, 149-166.

WAGENMAKERS, E.-J., RATCLIFF, R., GOMEZ, P., & IVERSON, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, **48**, 28-50.

WAGENMAKERS, E.-J., & WALDORP, L. (EDS.) (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, **50**(2).

WAINER, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, **4**, 212-213.

WALLACE, C. S., & DOWE, D. L. (1999). Refinements of MDL and MML coding. *Computer Journal*, **42**, 330-337.

WARE, J. H. (1989). Investigating therapies of potentially great benefit: ECMO. *Statistical Science*, **4**, 298-340.

WASSERMAN, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92-107.

WASSERMAN, L. (2004). *All of statistics: A concise course in statistical inference*. New York: Springer.

WEAKLIEM, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, **27**, 359-397.

WILKINSON, L., & THE TASK FORCE ON STATISTICAL INFERENCE (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, **54**, 594-604.

WINSHIP, C. (1999). Editor's introduction to the special issue on the Bayesian information criterion. *Sociological Methods & Research*, **27**, 355-358.

XIE, Y. (1999). The tension between generality and accuracy. *Sociological Methods & Research*, **27**, 428-435.

NOTES

1. A list of 402 articles and books criticizing the use of NHST can be found at biology.uark.edu/coop/Courses/thompson5.html.

2. For related critiques, see Jeffreys (1961), Pratt (1961, p. 166), and Royall (1997, p. 22).

3. Despite its intuitive appeal, the conditionality principle is not universally accepted; see Helland (1995) and the ensuing discussion.

4. NHST generally deals with conditioning by invoking "relevant subsets" or "ancillary statistics," concepts whose statistical problems are discussed in Basu (1964), Cornfield (1969, p. 619), D. R. Cox (1971), and A. W. F. Edwards (1992, p. 175).

5. This overestimation occurs for the sampling plan with fixed sample size. For the sequential sampling plan, *p* values *underestimate* the evidence against the null hypothesis (Berger & Berry, 1988a, p. 70).

6. Specifically, the normalized maximum likelihood instantiation of MDL divides the maximum likelihood for the observed data by the sum of the maximum likelihoods for all data sets that were not observed but that could have been observed in replicate experiments.

APPENDIX A
Some Bayesian Results for the Binomial Distribution

1. Before proceeding, recall the following extremely useful result:

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{A1}$$

where $\Gamma(\cdot)$ is the gamma function, which is a generalization of the factorial function. For positive integers, $\Gamma(x) = (x-1)!$. More generally, the gamma function is given by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\, dt.$$

Thus, $\Gamma(4) = 3! = 6$.

Using Equation A1 above, we can easily calculate the marginal probability of the data, $\Pr(D)$, for a binomial model with uniform prior distribution $\theta \sim \text{Uniform}(0, 1)$. Note that the prior distribution immediately drops out of the equation, so we have

$$\begin{aligned}\Pr(D) &= \int_0^1 \Pr(D \mid \theta)\Pr(\theta)\, d\theta \\ &= \int_0^1 \Pr(D \mid \theta)\, d\theta.\end{aligned}$$

The derivation then goes

$$\begin{aligned}\Pr(D) &= \int_0^1 \Pr(D \mid \theta)\, d\theta \\ &= \binom{n}{s}\int_0^1 \theta^s(1-\theta)^{n-s}\, d\theta \\ &= \frac{n!}{s!(n-s)!}\frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)} \\ &= \frac{\Gamma(n+1)}{\Gamma(n+2)} \\ &= \frac{1}{n+1}.\end{aligned} \tag{A2}$$

2. The Beta distribution is given by

$$\text{Beta}(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{A3}$$

When the prior is any Beta distribution $\text{Beta}(\theta \mid \alpha, \beta)$, its updating through the binomial likelihood will result in a posterior distribution that is still of the Beta form, albeit with different parameters: $\Pr(\theta \mid D) = \text{Beta}(\theta \mid \alpha + s, \beta + n - s)$. Priors that are of the same form as posteriors are said to be *conjugate*.

3. Now assume that we have two models: H_0 that assigns all probability mass to a single value of θ (i.e., θ_0), and H_1 that allows θ to vary freely. Under H_1, we again assume a uniform prior $\theta \sim \text{Uniform}(0, 1)$. The Bayes factor BF_{01} that gives the ratio of averaged likelihoods in favor of H_0 is then given by

$$\begin{aligned}BF_{01} &= \frac{\Pr(D \mid H_0)}{\Pr(D \mid H_1)} \\ &= \frac{\Pr(D \mid \theta_0)}{\int_0^1 \Pr(\theta)\Pr(D \mid \theta)\, d\theta} \\ &= \frac{\binom{n}{s}\theta_0^s(1-\theta_0)^{n-s}}{\binom{n}{s}\int_0^1 \theta^s(1-\theta)^{n-s}\, d\theta} \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)}\theta_0^s(1-\theta_0)^{n-s}.\end{aligned} \tag{A4}$$

APPENDIX B
The BIC Approximation to the Bayes Factor

The prior predictive probability of a given model or hypothesis i is given by

$$\Pr(D \mid H_i) = \int \Pr(D \mid \boldsymbol{\theta}_i, H_i) \Pr(\boldsymbol{\theta}_i \mid H_i)\ d\boldsymbol{\theta}_i, \tag{B1}$$

where D denotes the observed data and $\boldsymbol{\theta}_i$ denotes the parameter vector associated with model H_i. Using a Taylor series expansion about the posterior mode for $\boldsymbol{\theta}_i$ and the Laplace method for integrals, one can derive the following approximation (see Raftery, 1995, pp. 130–133):

$$\log \Pr(D \mid H_i) = \log \Pr(D \mid \hat{\boldsymbol{\theta}}_i, H_i) - (k_i/2)\log n + O(1). \tag{B2}$$

In this equation, $\hat{\boldsymbol{\theta}}_i$ denotes the maximum-likelihood estimate for $\boldsymbol{\theta}_i$, k_i denotes the number of parameters in model H_i, and n is the number of observations. The $O(1)$ term indicates that the error of approximation does not go to 0 as $n \rightarrow \infty$. However, the error of approximation will go to 0 as a proportion of $\log \Pr(D \mid H_i)$.

For certain classes of priors, the error of approximation reduces to $O(n^{-1/2})$. One such prior is the popular "Jeffrey's prior," with a specific choice for the arbitrary constant that precedes it (Wasserman, 2000, p. 99). Another prior that leads to approximation of order $O(n^{-1/2})$ is the unit information prior. This prior contains the same amount of information, on average, as does a single observation. Consider the case of $x^n = (x_1, x_2, \ldots, x_n) \sim N(\mu, 1)$ and a test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. The unit information prior for μ is then normal, with the mean given by the mean of the data, and the standard deviation equal to 1 (Raftery, 1999, pp. 415–416). Thus, for certain reasonable "noninformative" priors, the prior predictive probability of the data is

$$\log \Pr(D \mid H_i) = \log \Pr(D \mid \hat{\boldsymbol{\theta}}_i, H_i) - (k_i/2)\log n + O(n^{-1/2}), \tag{B3}$$

which means that the error of approximation goes to 0 as $n \rightarrow \infty$. As is evident from Equation B3, the approximation to the prior predictive probability is based on a component that quantifies the goodness of fit [i.e., the maximum likelihood $\Pr(D \mid \hat{\boldsymbol{\theta}})$] and a component that penalizes for model complexity [i.e., $(k/2)\log n$]. The penalty term depends not only on the number of free parameters, but also on the sample size n.

The BIC is obtained by multiplying Equation B3 by −2, yielding

$$\mathrm{BIC}(H_i) = -2\log L_i + k_i \log n, \tag{B4}$$

where L_i is the maximum likelihood for model H_i—that is, $L_i = c\Pr(D \mid \hat{\boldsymbol{\theta}}, H_i)$, with c an arbitrary constant (A. W. F. Edwards, 1992). The BIC approximation to the prior predictive probability $\Pr(D \mid H_i)$ is obtained by the reverse transformation: $\Pr_{\mathrm{BIC}}(D \mid H_i) = \exp[-\mathrm{BIC}(H_i)/2]$.

Recall from the discussion following Equation 7 that the Bayes factor is the ratio of prior predictive probabilities—that is, the probability of the data under H_0 divided by the probability of the data under H_1. It follows that in the case of two models H_0 and H_1, the BIC approximation of the Bayes factor is given by

$$BF_{01} \approx \frac{\Pr_{\mathrm{BIC}}\left(D \mid H_0\right)}{\Pr_{\mathrm{BIC}}\left(D \mid H_1\right)} = \exp\left(\Delta\mathrm{BIC}_{10}/2\right), \tag{B5}$$

where $\Delta\mathrm{BIC}_{10} = \mathrm{BIC}(H_1) - \mathrm{BIC}(H_0)$.

(Manuscript received June 28, 2006;
revision accepted for publication February 13, 2007.)