Psychological Bulletin
1996, Vol. 119, No. 1, 149-158

0033-2909/96/$3.00

Bayesian Methods for Assessing Importance of Effects

Henry Rouanet
Centre National de la Recherche Scientifique and Université René Descartes

In experimental data analysis when it comes to assessing the importance of effects of interest, 2 situations are commonly met. In Situation 1, asserting largeness is sought: "The effect is large in the population." In Situation 2, asserting smallness is sought: "The effect is small in the population." In both situations, as is well known, conventional significance testing is far from satisfactory. The claim of this article is that Bayesian inference is ideally suited to making adequate inferences. Specifically, Bayesian techniques based on "noninformative" priors provide intuitive interpretations and extensions of familiar significance tests. The use of Bayesian inference for assessing importance is discussed elementarily by comparing 2 treatments, then by addressing hypotheses in complex analysis of variance designs.

When a researcher examines the statistical results of an experiment, two situations are commonly encountered when it comes to assessing the importance of effects of interest.[1]

In Situation 1 (largeness), the descriptive conclusion is informally expressed as "There is an effect"; the underlying idea is that presumably the true effect is (substantially) *large*. Then the researcher looks at the corresponding t or F ratio in hopes of finding a statistically significant result and of concluding "There is evidence of an effect" ($p < .01$, e.g.). Though formally correct, this sentence is open to misinterpretation, however, as the well-known *Ritual Warning 1* states, Statistical significance is not practical significance.

In Situation 2 (smallness), the descriptive conclusion is informally expressed as "There is no effect"; the underlying idea is that presumably the true effect is (trivially) *small*. The researcher then looks at the F ratio in hopes of now finding a statistically nonsignificant result and of concluding "There is no evidence of effect" ($p > .50$, e.g.). Again, though formally correct, this sentence is open to misinterpretation because as *Ritual Warning 2* says, No evidence of effect is not proof of no effect.

Situation 1 is quite common: One wishes to show that there is a marked difference between two experimental conditions or between the performances of two populations. Situation 2 is, however, perhaps not so uncommon: One may wish to show that two teaching methods differ only slightly; that some interaction effect is small enough to be ignored; or that the fit of some model, though imperfect, is acceptable as a first approximation. In the context of model validation, the often-read conclusion "Experimental evidence is consistent with the model"—albeit cautious and in itself legitimate—becomes adventurous, as soon as with ensuing developments the model's validity is taken for granted.

I propose in this article techniques for handling Situations 1 and 2. Two preliminary points are worth mentioning. First, there is some mathematical statistics literature about the subject, going back to Hodges and Lehman (1954), at least. Second, there is an abundant number of related articles in biometrics, and more specifically pharmacokinetics, where trying to assert that two drugs differ only by a slight amount (Situation 2) is known as *testing for the bioequivalence* of the two drugs. For this situation, at least two competing frequentist techniques have been advocated: one based on a noncentral F and another based on two one-sided t tests (for a discussion, see Schuirman, 1987). Such techniques are potentially of general interest and, as it turns out, variants of them have already been addressed to psychologists, namely by Rogers, Howard, and Vessey (1993) and under the banner of the *good-enough principle* by Serlin and Lapsley (1985, 1993).

In the field of biometrics, in addition to frequentist articles, there is a growing Bayesian literature, as reflected for instance by the special issue of *The Statistician* (Smeeton, 1994). Admittedly, the philosophy underlying the Bayesian approach has been a matter of debate among statisticians, but the thrust of this contemporary Bayesian literature is the building of concrete answers to contemporary problems of interest, leaving the "grand debate" between frequentists and Bayesians in the background.

This article finds its proper place within this developing, pragmatic Bayesian literature. Its basic claim is that Bayesian techniques are ideally suited for handling Situations 1 and 2 alike. This article thus offers an introduction to the use of Bayesian techniques in experimental data analysis, in connection with a topic of psychological interest. After all, perhaps the best way to render the Bayesian approach attractive is to show that it offers a sensible and practical answer to a major issue.

The work that has led to this article was initiated years ago with my regretted colleague and friend Dominique Lépine, to whom this article is a posthumous homage.

I warmly thank A. R. Jonckheere and P. Suppes for their encouragements and efficient help.

Correspondence concerning this article should be addressed to Henry Rouanet, UFR Math-Info, 45 Rue des Saints Pères, 75270 Paris cedex 06, France. Electronic mail may be sent via Internet to rouanet@math-info.univ-paris5.fr.

[1] In this article, the word *important* is understood throughout in terms of magnitude, not of impact for psychological theories where, as a referee pointed out, asserting the smallness of an effect may be of "theoretical importance."

I first discuss Bayesian procedures for assessing importance in the elementary case of comparing two treatments. I then turn to procedures for addressing hypotheses in the context of complex analysis of variance (ANOVA) designs.

Beyond Effect Sizes

The warnings above are ritual because they are well known, and various recommendations have been made for alleviating misinterpretations of significance testing. A sensible recommendation is to look at *effect sizes* (see, e.g., Tatsuoka, 1993), that is, at the magnitudes of observed effects, estimated by various indicators such as the standardized difference between means or the proportion η^2 of total variance accounted for by a source of variation, and so forth. As good sense suggests (and maximum likelihood confirms), observed effects reflect corresponding population effects, allowing for sampling fluctuations. By looking at effect sizes, one can therefore avoid gross errors. Thus, getting a small observed effect, even though statistically significant, precludes asserting largeness of the corresponding population effect; similarly, getting a large observed effect, even though statistically nonsignificant, precludes asserting smallness. In such cases, however, what positive assertions may be made? Could one contemplate asserting largeness when the effect is large though statistically nonsignificant? Could one contemplate asserting smallness when the effect is small though statistically significant? Such questions are moot ones, because in significance testing it is hard to sort out the influences of effect size and sample size.

It is my contention that in both Situations 1 and 2, looking at observed effects should be done at the start of data analysis and not as an afterthought following significance testing. Description comes first. In Situation 1, the descriptive conclusion should be that the observed effect (taken with its sign of course) is large; in Situation 2, that it is small (regardless of sign). If this is not the case, no corresponding inferential conclusion is reachable. If this is the case, the aim ascribed to inferential procedures should be *to extend descriptive conclusions* to the population, allowing for sampling fluctuations. In other words, in Situation 1, one should try to assert that the population effect is itself large, and in Situation 2, that the population effect is itself small. Naturally, if the observed effect happens to be in an "intermediate zone," there is no hope to assert either largeness or smallness.

For the conceptual discussion, consider the *matched pairs design,* where n subjects undergo two treatments, with a numerical dependent variable (score). Let d_i be the difference of the two scores, or individual effect for Subject i, hence the mean effect

$$\sum \frac{d_i}{n} = \bar{d},$$

henceforth denoted d for simplicity. Let δ denote the corresponding population effect. Let σ^2 denote the variance of individual effects in the population, and

$$s^2 = \sum \frac{(d_i - d)^2}{(n-1)}$$

its unbiased estimate (with $q = n - 1$ df). It is convenient for the purpose of this article to use the *standardized effect* d/s, which seems to have become common practice in behavioral research. Specifically, starting from the idea that $|d|/s = 0.5$ may represent a typical medium-sized effect (see, e.g., Cohen, 1977, 1992), the following conventions are adopted in the sequel: If $|d|/s$ exceeds 0.6, the effect will be deemed large; if $|d|/s$ is below 0.4, it will be deemed small. Needless to say, the foregoing conventions ought to be viewed as rules of thumb. Specifying "good-enough values" depends on the state of the art in the field under study and should be reconsidered in each concrete situation.

Bayesian Inference

Now assume the usual normal sampling model, that is, assume that the individual effects $(d_i)(i = 1, \ldots, n)$ are independently normally distributed $N(\delta, \sigma^2)$ (with mean δ and variance σ^2). Consequently, the sampling distribution of the mean d is $N(\delta, \sigma^2/n)$. To test the null hypothesis H_0: $\delta = 0$, the usual test statistic is $t = \sqrt{n}(d/s)$, which under H_0 is distributed as t_q (elementary Student's t with $q = n - 1$ df). If t_{obs} denotes the observed value of t, H_0 is rejected if given the hypothetical event $\delta = 0$, the probability that t_q exceeds t_{obs} (upper-sided test, e.g.) or that $|t_q|$ exceeds $|t_{obs}|$ (two-sided test) is small enough.

At this point, recall the principle of the Bayesian approach, referring for a detailed presentation to standard Bayesian textbooks such as those by Lindley (1965), Box and Tiao (1973), Press (1989), or Lee (1989). A *prior distribution,* expressing one's uncertainty about parameters independently from the data, is postulated. This prior is combined with the sampling distribution and data using the classical Bayes' theorem and yields a *posterior distribution,* which expresses the uncertainty about the parameters, conditionally on data.

Noninformative Priors

What distribution should be taken as a prior? According to one Bayesian conception, the prior should incorporate all available information or opinion about the parameters. Early attempts at introducing Bayesian inference in psychology, such as those by Edwards, Lindman, and Savage (1963), were strongly committed to this conception. There is, however, another conception, more in the spirit of this article, where priors are chosen to express a "state of ignorance" about the parameters, with the following motivation: If the prior expresses ignorance about parameters, the posterior expresses the evidence brought by the data. In current Bayesian terminology, such priors are called *noninformative.*[2]

Posterior Distributions

In the above elementary situation, it can be shown, assuming the noninformative prior on (δ, σ), that given $d = d_{obs}$ (the observed value of d) and $s = s_{obs}$ (the observed value of s), the posterior distribution of δ is such that $\sqrt{n}(\delta - d_{obs})/(s_{obs})$ is distributed

[2] Technically, in this article, noninformative priors derived from classical Jeffreys' rule are used (see Box & Tiao, 1973, pp. 41–42).

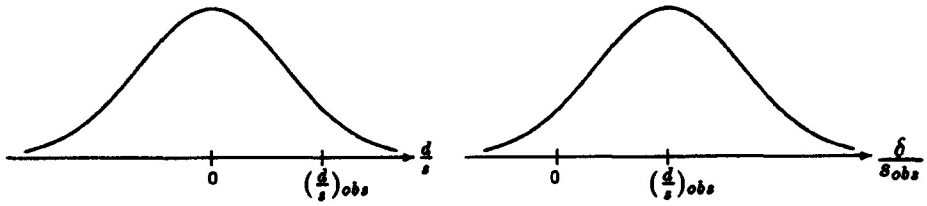

Figure 1. Sampling distribution under H_0 (left) and posterior distribution (right).

as an elementary t_q variable (Box & Tiao, 1973, p. 97). In other words, the distribution of δ is a *scaled* t_q with mean d and scale s^2_{obs}/n (taking $d = 0$ and $s^2_{obs}/n = 1$, one gets back to the elementary t). Using the notation of scaled t_q (standard in Bayesian statistics), the basic result about the posterior distribution of δ is written (where $\sim$ is read "is distributed as"):

$$\delta \sim t_q\left(d_{obs}, \frac{s^2_{obs}}{n}\right).$$

The connection between sampling and posterior distributions is apparent if one scales both the observed effect d_{obs} and the population effect δ by the observed value s_{obs}. Using $(d/s)_{obs}$ to denote the observed standardized effect d_{obs}/s_{obs}, one gets the posterior distribution $\delta/s_{obs} \sim t_q[(d/s)_{obs}, 1/n]$ by shifting the sampling distribution under H_0 by the amount $(d/s)_{obs}$ (see Figure 1).

At this point, the methodology basic to this article should emerge. With the observed effect taken as an estimate of the population effect, the Bayesian approach provides a distribution around this estimate, and this distribution is used to draw inferences about how large or small the effect may be in the population. It is clear how this methodology contrasts with that of significance testing. Instead of asking whether a population in which the null hypothesis were true could have produced the observed effect (significance testing), one asks what a population that has produced the observed effect may look like (Bayesian inference).

Owing to the way the distribution of δ is expressed in terms of s^2_{obs}, to assess the importance of effect in the population, the simplest choice is to take $|\delta|/s_{obs}$ as an index of importance. This will be done in the sequel to this article.[3]

Asserting Largeness or Smallness

The conclusions that can be drawn from a posterior distribution are direct extensions of the descriptive conclusions taken from the observed effects. If the bulk of the distribution lies in the region of large effect values, the probability is high that the population effect is large, so largeness will be asserted; if it lies in the region of small effect values, the probability is high that the population effect is small, so smallness will be asserted.

Example 1

Let $(d/s)_{obs} = 0.9$ (large observed effect) with $n = 25$. The posterior distribution is $\delta/s_{obs} \sim t_{24}(0.9, 1/25)$ (see Figure 2). Most of the distribution clearly lies on the region of large effects, which means that largeness can be asserted.

To get a compact assertion stating that the probability is high that the effect is large, one may set up (a) a *credibility level* γ (> 0.50) and (b) a *limit for largeness* l_{lar}; one may then attempt to show that $P(\delta/s_{obs} > l_{lar}) > \gamma$. For example, letting $\gamma = .90$ and $l_{lar} = 0.6$, one finds from the t_{24} distribution that $P(\delta/s_{obs} > 0.6) = .927$,[4] and because $0.927 > 0.90$, one may assert that the effect in the population exceeds 0.6 with a probability higher than .90. Alternatively, one finds the value 0.64 such that $P(\delta/s_{obs} > 0.64) = .90$ (see Figure 2), and because $0.64 > 0.6$, one reaches the same conclusion. In short, for $\gamma = .90$ and $l_{lar} = 0.6$, largeness of effect is asserted.

It may be remarked that assessing importance is a more demanding task than significance, and experience suggests that taking $\gamma = .90$ may be a reasonable (not mandatory) convention (rather than, e.g., taking the complementary values of familiar α levels).

Example 2

Let $(d/s)_{obs} = 0.1$ (small observed effect) with $n = 25$. The posterior distribution is $\delta/s_{obs} \sim t_{24}(0.1, 1/25)$ (see Figure 3). Here most of the distribution lies in the region of small values.

To assert smallness, again take $\gamma = .90$ as a credibility level and $l_{sma} = 0.4$ as a *limit for smallness*. Then one can find $P(|\delta|/s_{obs} < 0.4) = .917$, and because $.917 > .90$, one may assert that the effect in the population is (in absolute value) smaller than 0.4 with a probability higher than .90. Alternatively, one finds $P(|\delta|/s_{obs} < 0.38) = .90$ (see Figure 3), and because 0.38 is less than 0.40, one reaches the same conclusion, that $P(|\delta|/s_{obs} <$

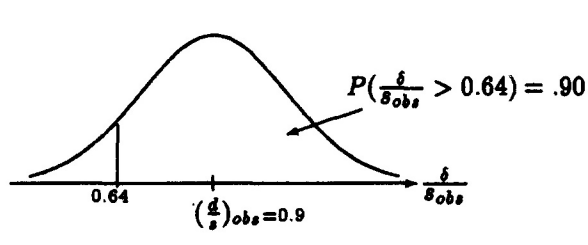

Figure 2. Example 1: $P(\delta/s_{obs} > 0.6) > .90$. Largeness of effect is asserted (for $\gamma = .90$ and $l_{lar} = 0.6$).

[3] If one took δ/σ as an index, Bayesian inference would still be feasible but technically more complicated: One would replace the t_q distribution by a distribution that I call the L'_q distribution, which is closely linked to the classical noncentral t_q distribution (Rouanet & Lecoutre, 1983).

[4] Accurate t distributions are nowadays obtainable from any standard software and even calculators.

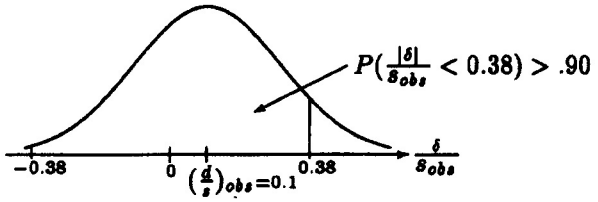

Figure 3. Example 2: $P(|\delta|/s_{obs} < 0.4) > .90$. Smallness of effect is asserted (for $\gamma = .90$ and $l_{sma} = 0.4$).

$0.4) > .90$. In short, for $\gamma = .90$ and $l_{sma} = 0.4$, smallness of effect is asserted.

Taking $\gamma = .90$, largeness can actually be asserted in Example 1 for any l_{lar} less than 0.64; smallness can actually be asserted in Example 2 for any l_{sma} greater than 0.38. Therefore, in the process of data analysis, instead of setting rigid upper or lower limits, it is often more convenient for the psychologist to first set γ, for example, $\gamma = .90$, and then to figure out whether the corresponding value can be deemed large (in Situation 1) or small (in Situation 2).

Role of Sample Size

Sample size controls the dispersion of posterior distributions: The greater the sample size, the more concentrated the distribution is. In Situation 1, for a given $(d/s)_{obs} > l_{lar}$, the probability $P(\delta/s_{obs} > l_{lar})$ is an increasing function of n. For instance, with $(d/s)_{obs} = 0.9$, one finds for $n = 100$, $P(\delta/s_{obs} > 0.60) = .998$; for $n = 400$, $P(\delta/s_{obs} > 0.60) = 1$.

Similarly, in Situation 2, given $|d/s|_{obs} < l_{sma}$, the probability $P(|\delta|/s_{obs} < l_{sma})$ is again an increasing function of n. For instance, with $(d/s)_{obs} = 0.1$, one finds for $n = 100$, $P(|\delta|/s_{obs} < 0.4) = .998$; for $n = 400$, $P(|\delta|/s_{obs} < 0.4) = 1$. Thus, in Situation 2 as well as in Situation 1, the more observations one has, the better position one is in to enforce a conclusion, a highly desirable property from a statistical standpoint. Large samples can do no harm.[5]

Choosing an Appropriate Sample Size

Bayesian calculations may be performed prior to gathering data. Thus, in Situation 1, one may calculate in advance that for a sample size of $n \geq 20$, the standardized observed effect should be at least 0.9 to assert largeness (with the aforementioned conventions); for $n \geq 25$, a value of 0.87 would suffice, and so forth. Such calculations prove helpful for choosing a sample size, according to plausible values for the effect to be observed.

In Situation 2, because $d_{obs} = 0$ is obviously the observed value most favoring smallness, an absolute minimal sample size can be determined. If for $d_{obs} = 0$, one wants to have $P(|\delta|/s_{obs} < 0.4) > .90$, n must be such that $P(|t_{n-1}| < 0.4\sqrt{n}) > .90$, which leads to $n \geq 19$. Stated otherwise, if one hopes to assert smallness, at least 19 subjects are needed in any case.

Reinterpreting Significance Levels

Posterior distributions based on noninformative priors provide reinterpretations of frequentist procedures in terms of probabilities about parameters (Box & Tiao, 1973, p. 102; Lewis, 1993; Lindley, 1965). In the elementary case, the reinterpretation readily follows from the relation between sampling and posterior distributions. If p denotes the two-sided, observed level (p value), then $(1 - p)/2$ is simply the probability that the effect δ has the same sign as the observed effect d_{obs}, and $1 - p$ is the probability that δ lies between 0 and $2d_{obs}$ (see Figure 4).

Now apply those reinterpretations to Situations 1 and 2. In Situation 1, it seems natural to reinterpret the directional level. Thus, in Example 1, the observed value of the t statistic is $t_{obs} = \sqrt{25} \times 0.9 = 4.5$, hence $p/2 = .00007$, and therefore $P(\delta > 0) = .99993$ (see Figure 5). With a very high credibility, one may assert that the effect is positive. Clearly, this does not amount to asserting that with a high credibility, the effect is large. Thus, Warning 1 interpreted in Bayesian terms says: A high probability that δ has the sign of d_{obs} does not entail a high probability that δ is large. Finding a statistically significant effect is perhaps a first step, but not a sufficient one, toward asserting largeness.

In Situation 2, it seems natural to reinterpret the two-sided level. Thus, in Example 2, one has $t_{obs} = \sqrt{25} \times 0.1 = 0.5$, hence $p = .622$. That is, one may state that $P(\delta/s_{obs} \in [0, 0.2]) = .378$ (see Figure 6). Now this in no sense points toward asserting smallness. If one considers the limit case where $p = 1$ (a "perfectly statistically nonsignificant" result), the reinterpretation tells no more than $P(\delta \in [0, 0]) = 0$, a trivial statement. Thus, Warning 2 interpreted in Bayesian terms says: Not having a high probability that δ is of the sign of d_{obs} does not entail a high probability that δ is small. In plain words, finding a statistically nonsignificant result is not at all a step toward asserting smallness.

At this point, the main contribution of the Bayesian approach in Situation 2 becomes apparent: It distinguishes those situations where, owing to insufficient sample size, the conclusion in favor of some approximate null hypothesis is not obtainable and those situations where the sample size is big enough for a genuine smallness conclusion to be reached. In this connection, consider the following situation. Suppose that not only d_{obs} but $2d_{obs}$ is small and that the result is significant, that is, p is low, and consequently $1 - p$ is high. Then, from the reinterpretation of the two-sided level, it follows that the probability is high that δ is between 0 and $2d_{obs}$, hence the probability is high that δ is small. As a conclusion, when the observed effect is very small, getting a statistically significant result—far from ruling out the conclusion of a small effect—is actually a cue for it. This is what I call the *negligibility paradox* (Rouanet, in press).

Bayesian Extensions of ANOVA

Psychological experiments often involve complex designs, and there is no "established" Bayesian ANOVA, covering numerous common designs, comparable to frequentist ANOVA textbooks like, for example, Winer's (1979).[6] My colleagues

[5] Large samples may not be necessary, of course. Moreover, as pointed out by a referee, it is a strength of the Bayesian approach that it enables an investigator to monitor the accumulation of data and search for an early conclusion, when practical, ethical, or both considerations make it desirable (as reflected in the biometric literature).

[6] The excellent book by Box and Tiao (1973), in spite of its mathematical sophistication, far from covers most common designs.

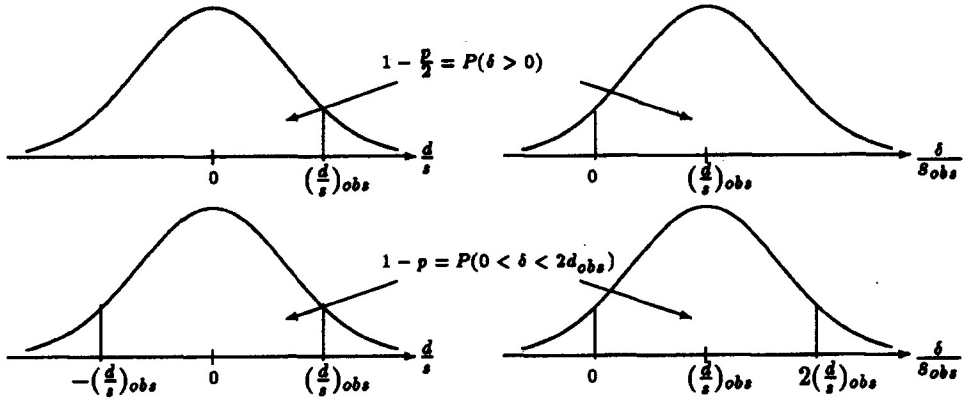

Figure 4. Sampling distribution under H_0 (left) and posterior distribution (right). Reinterpretations of observed levels: directional $p/2$ (top) and two-sided p (bottom). Figures for $d_{obs} > 0$.

and I have long been working on filling in this gap by devising methods and software that include Bayesian procedures in addition to significance tests. A good part of my work and that of my colleagues has dealt with Bayesian ANOVA (see Hoc, 1983; Lecoutre, 1984; Lépine & Rouanet, 1975; Rouanet & Lépine, 1977). Another part has been concerned with categorized data (see Bernard, 1991, in press). Worked examples of real data sets are also found in Rouanet, Lépine, and Pelnard-Considère (1976); Rouanet, Lépine, and Holender (1978); Bernard, Blancheteau, and Rouanet (1985); Rouanet, Bernard, and Le Roux (1990); and Rouanet, Bernard, Lecoutre, and Le Roux (in press). Especially relevant to this article is the reference of Rouanet and Lecoutre (1983), which discusses specific inference (see *Bayesian Distributions*) on the same concrete example.

The principle of Bayesian ANOVA that underlies this article can be simply stated. For any sampling model, the corresponding noninformative Bayesian inference is obtained by assuming a noninformative prior on the parameters of the model. Therefore, for any source of variation for which there is a valid F test, there is a corresponding valid Bayesian procedure based on the same two sums of squares (SS) as the F statistic.

The material in the remainder of this article is organized as follows. In this section, I state the basic theoretical results for one then several degrees of freedom (df); then I turn to assumptions and computing considerations. In the next sections, I present a concrete numerical example in detail.

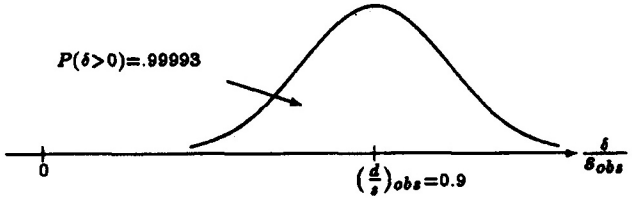

Figure 5. Example 1. Reinterpretation of significance level: $p/2$ = .00007, $1 - p/2$ = .99993 = $P(\delta > 0)$.

Basic Results

1-df source of variation: Inference on effect. A 1-df effect is a signed effect and can be represented by a contrast among cells. Carrying over the notations of the elementary case, I denote δ as the population effect, d_{obs} as the corresponding observed effect, and s^2_{obs} as the observed variance (corrected, with q df) based on n units. Then the posterior distribution of the population effect, standardized by the observed (corrected) standard deviation, reads like in the elementary case,

$$\frac{\delta}{s_{obs}} \sim t_q\left[\left(\frac{d}{s}\right)_{obs}, \left(\frac{1}{n}\right)\right].$$

As a consequence, for any 1-df source of variation, Bayesian inference involves only the elementary t distribution. The practical interest of this result cannot be overemphasized, owing to the paramount importance of 1-df comparisons in experimental data analysis.

It may be remarked that the index of importance that has been taken in this article is scale invariant. Consequently, for assessing the importance of a 1-df effect of interest, the scale of the contrast selected to represent the effect is irrelevant. In practice, of course, analyses will more naturally be conducted in terms of "meaningfully scaled" contrasts, such as differences between means for main effects, differences of differences for interaction, and so forth.

Several-df source of variation: Inference on importance of effect. As classically done in ANOVA (see, e.g., Scheffé, 1959), the magnitude of an effect with m df ($m \geq 1$; m denotes a number of df) may be defined as a variance (corrected, with m df). From now on I use ANOVA notations: σ^2_{eff} denotes the magnitude of effect in the population; s^2_{eff}, the corresponding observed magnitude; if MS_{eff} denotes the mean square of effect, and n_{eff}, the coefficient of σ^2_{eff} in $E(MS_{eff})$, one gets the following equation: $s^2_{eff} = MS_{eff}/n_{eff}$. Similarly, for the corresponding error term, I define s^2_{err} (with q df) and n_{err} with $s^2_{err} = MS_{err}/n_{err}$. Last, let $\hat{n} = n_{eff}/n_{err}$, with $F_{obs} = \hat{n}s^2_{eff}/s^2_{err}$. Then the posterior distribution of the population variance, standardized by the ob-

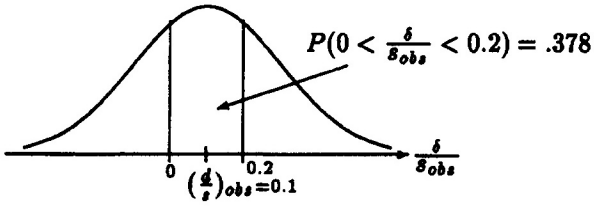

Figure 6. Example 2. Reinterpretation of significance level: $p = .622$, $1 - p = .378 = P(0 < \delta/s_{obs} < 0.2)$.

served error variance, reads, as a direct generalization of Rouanet and Lecoutre (1983, p. 262),

$$\frac{\sigma^2_{eff}}{s^2_{err}} \sim \frac{1}{\hat{n}m}\psi^2_{m,q}(mF_{obs}).$$

The distribution $\psi^2_{m,q}(mF_{obs})$ (psi-square distribution with *df* m and q and eccentricity parameter mF_{obs}) is the Bayesian noncentral extension of the classical F distribution (see Rouanet & Lecoutre, 1983; Schervish, 1992); its numerical evaluation has been investigated in Lecoutre, Guigues, and Poitevineau (1992).

As a consequence, all computations may be performed from the relevant terms that appear in the ANOVA table, using a single computer program implementing the ψ^2 distribution.

1 df again. The foregoing result becomes

$$\frac{\sigma^2_{eff}}{s^2_{err}} \sim \frac{1}{\hat{n}}\psi^2_{1,q}(t^2_{obs}).$$

It can be shown that the ratio s_{eff}/s_{err} is equal (up to the sign) to the ratio $(d/s)_{obs}$ defined before, that is, $s_{eff}/s_{err} = |d/s|_{obs}$, similarly, that $\sigma_{eff}/s_{err} = |\delta|/s_{obs}$, and lastly that $\hat{n}$ here is equal to n. Provided the sign of d_{eff} is taken into account, the study of a 1-*df* effect can therefore be also conducted from the ANOVA table.

Assumptions on Sampling Models

The assumptions for the noninformative Bayesian inference are the assumptions of the corresponding sampling model. Several error terms may be available for a given source of variation, depending on the degree of restrictiveness in the assumptions made on the model. In this connection, recall the assumptions in repeated measurement designs (see Rouanet & Lépine, 1970). For each within-subject source of variation, one can take as an error term either a term that is specific to this source or under *circularity* (sometimes called *sphericity*) assumptions, a term that is common to several sources. Different choices lead to different F ratios and consequently to different posterior distributions.[7]

The extension to m *df* sources of variation (with $m > 1$) is straightforward, assuming circularity assumptions leads to both valid F ratios and posterior distributions of σ^2_{eff} based on the $\psi^2_{m,q}$ distribution (see Rouanet & Lecoutre, 1983).

Computing Considerations

If one knows how to build the ANOVA table, one knows how to get its Bayesian extensions. All 1-*df* analyses can be carried out, similar to elementary comparisons between means, with a calculator equipped with the t distribution. For any number of *df*, a program such as Programme d'Inférence Fiducio–Bayésienne (PIF; Lecoutre & Poitevineau, 1991), which implements the ψ^2 distribution, may be used. For each source of variation, this program takes as input the effect and error SS with their *df* and constructs the posterior distribution, from which it furnishes on request all probabilities relevant for asserting either largeness or smallness as the case may be. More extensive software packages have also been written: Lecoutre and Poitevineau (1991) and Bernard and Poitevineau (1986). An assistant software for assessing importance in the context of categorical data is currently being written (Le Roux, Durand, & Walfard, 1995).

Bayesian Inference in Practice

The practice of Bayesian inference for assessing importance in the context of ANOVA designs is best conveyed by means of an example of moderate complexity. On this example, I derive Bayesian distributions for the effects of interest. In the next section, I use these distributions to assess importance.

Reaction Time Data

Consider the following experiment (after Rouanet & Lecoutre, 1983) devised to investigate the model of additive stages in reaction times (RT). Four conditions were defined by crossing the following two within-subject factors: Signal Frequency (Factor A) with two levels, frequent ($a1$) and rare ($a2$); and Foreperiod Duration (Factor B), also with two levels, short ($b1$) and long ($b2$). Because the psychological model of additive stages entails the absence of interaction between the experimental Factors A and B, the main research hypothesis was that the interaction between Signal Frequency and Foreperiod Duration, denoted $A.B$, was (about) null.[8]

There was also a between-subject Factor G, classifying the 12 subjects into three groups of 4 subjects each. The data treated here are RTs (averaged over trials) for subjects and conditions as shown in Table 1.

This table also shows the 12 individual main A effects [e.g., $(435 + 473)/2 - (387 + 416)/2 = 52.5$] and the 12 individual interaction $A.B$ effects (e.g., $-387 + 435 + 416 - 473 = -9$). The ANOVA table is presented in Table 2.

Bayesian Distributions

First example: Main effect of factor A (1 df). The source of variation denoted A in the ANOVA table has 1 *df*. In what follows, the effect is defined by the contrast $(-1/2, +1/2, -1/2, +1/2)$

[7] In the reaction time (RT) data set, specific errors terms have been taken leading to "F ratios," as defined in Rouanet and Lépine (1970). Taking common error terms, leading to "F' ratios" would lead to virtually the same conclusions (as may be checked from the data).

[8] The motivation for denoting $A.B$ the interaction is that I reserve the notation $A \times B$ for the source of variation associated with the Cartesian product of A and B. Investigating the source $A \times B$ here would mean examining the overall differences between the four levels defined by the crossing of Factors A and B.