Type S error rates for classical and Bayesian single and multiple comparison procedures

Andrew Gelman[1]
Department of Statistics, Columbia University, New York, USA

Francis Tuerlinckx
Department of Psychology, University of Leuven, Belgium

Summary

In classical statistics, the significance of comparisons (e.g., $\theta_1 - \theta_2$) is calibrated using the Type 1 error rate, relying on the assumption that the true difference is zero, which makes no sense in many applications. We set up a more relevant framework in which a true comparison can be positive or negative, and, based on the data, you can state "$\theta_1 > \theta_2$ with confidence," "$\theta_2 > \theta_1$ with confidence," or "no claim with confidence." We focus on the *Type S* (for sign) error, which occurs when you claim "$\theta_1 > \theta_2$ with confidence" when $\theta_2 > \theta_1$ (or vice-versa). We compute the Type S error rates for classical and Bayesian confidence statements and find that classical Type S error rates can be extremely high (up to 50%). Bayesian confidence statements are conservative, in the sense that claims based on 95% posterior intervals have Type S error rates between 0 and 2.5%. For multiple comparison situations, the conclusions are similar.

[1]Address correspondence to Andrew Gelman, Department of Statistics, Columbia University, New York, NY 10027 USA. We thank David H. Krantz, Frederic Y. Bois, the editor, and two referees for helpful comments. This work was supported in part by the U.S. National Science Foundation grant SBR-9708424 and Young Investigator Award DMS-9796129. The second author is a research assistant for the Fund of Scientific Research - Flanders.

Keywords: Bayesian Inference, Multiple Comparisons, Type 1 Error, Type M Error, Type S Error

1 Introduction

1.1 Type 1 and Type S error rates

Classical comparisons procedures are calibrated based on the Type 1 error, that is, the probability of claiming that $\theta_1 \neq \theta_2$ if, in fact, $\theta_1 = \theta_2$. Thus, for example, if we follow the procedure of claiming that $\theta_1 \neq \theta_2$ if the 95% confidence interval for $\theta_1 - \theta_2$ excludes zero, then our Type 1 error rate should be at most 5%. Based on our experience with data analysis in the social and behavioral sciences, we believe this framework to be generally inappropriate, since we do not believe that $\theta_1 = \theta_2$ is a reasonable possibility for continuous parameters.

We prefer to think in terms of the sign of the comparison. Thus, we identify a 95% interval for $\theta_1 - \theta_2$ that is all-positive as a claim "with confidence" that $\theta_1 > \theta_2$, with an all-negative 95% interval corresponding to the opposite claim and an interval that includes zero results in no confident claim (see, e.g., Tukey, 1960, Harris, 1997, and Rindskopf, 1997). This sign comparison procedure is calibrated using *Type S* (for sign) errors, which correspond to wrongly identifying the sign of a comparison; that is, claiming that $\theta_1 > \theta_2$ when in fact $\theta_2 > \theta_1$. From the perspective of sign comparisons, we believe that the Type S error rate is of more direct interest than the Type 1 error rate. In particular, we shall examine the probability of making a Type S error, conditional on making a comparison with confidence. The Type S error rate will be compared for classical and Bayesian confidence intervals for data from hierarchical normal models.

The structure of the paper is as follows. First, we will continue the introduction by considering the relation between one-sided tests, two-sided tests and confidence intervals. We also explore the relation between single and multiple comparisons and hierarchical models. In Section 2, we lay out the particular hierarchical model that will used throughout the paper, followed by a definition of the classical and Bayesian intervals leading to confidence statements. In Section 3, Type S error rates for single comparisons under the hierarchical model are defined and results are presented. Section 4 presents an evaluation of the Type S error probabilities for multiple comparisons; we consider two classical procedures and one Bayesian procedure. We conclude in Section 5 with a discussion about the relevance of Type S error rates, suggestions for further work and some general recommendations and conclusions.

1.2 One-sided tests, two-sided tests, and the interpretation of confidence intervals

Classical one- and two-sided tests for comparisons test the hypotheses $\theta_j < \theta_k$ and $\theta_j \neq \theta_k$, respectively. Since we are testing the two inequalities $\theta_j < \theta_k$ and $\theta_k < \theta_j$, our procedure of focusing on the sign of the confidence interval corresponds to two simultaneous one-sided tests. This interpretation would be acceptable, but we prefer to think of our "claims with confidence" and Type S errors as arising from the natural interpretation of 95% confidence intervals for comparisons. When the interval for $\theta_j - \theta_k$ includes zero (e.g., $[-1.3, 5.9]$), then it is standard to say that the two parameters are not statistically significantly different. When the interval for $\theta_j - \theta_k$ excludes zero, then it is standard to accept the difference as real and to confidently work with the assumption that the sign of the true difference is as given by the estimated difference.

This procedure—to implicitly make confident claims about the sign of a comparison if the 95% interval for the estimate excludes zero—is standard in applications of linear regression, generalized linear models, and more complicated statistical analyses as well as for simple comparisons of means. Thus, we do not see ourselves as evaluating one-sided or two-sided tests but rather as evaluating the standard statistical procedure—whether classical or Bayesian—based on locating a 95% interval relative to zero.

1.3 Single comparisons, multiple comparisons, and hierarchical models

Statistical theory distinguishes between *single comparisons*, in which error rates are evaluated for each comparison separately, and *multiple comparisons*, in which one evaluates the error rate jointly among a set of comparisons.

In a more general sense, however, all evaluations of statistical methodology are multiple comparisons problems in the sense that we expect to use a method repeatedly in a variety of situations. The frequency properties of a statistical method are defined with respect to long-term repeated use corresponding to a distribution of true parameter values and data. For single comparisons, we consider the probabilities that claims with confidence are in fact true, on an individual basis; for multiple comparisons, we consider the probability that an ensemble of such claims are true. In either case, we evaluate these probabilities in the context of a model of the distribution of true parameter values θ_j. As we shall see, the variance of this distribution is a key parameter determining the Type S error rates for both classical and Bayesian procedures.

2 Theoretical framework

2.1 Model and definition of replications

Consider data from J independent studies, with n_j observations from each study j. We do not specify a probability model for the individual observations; instead we focus directly on a derived measurement or summary statistic, which we label y_j, from each study. This derived measurement will most likely be the sample mean for sample j, but in general it could be any function of the data for which the following makes sense. We assume the distribution of y_j is normal, that it depends on an unknown parameter θ_j and has variance σ^2,

$$y_j|\theta_j, \sigma \sim \mathrm{N}(\theta_j, \sigma^2). \tag{1}$$

Furthermore, we assume that the population of θ_j's follow a normal distribution,

$$\theta_j|\mu, \tau \sim \mathrm{N}(\mu, \tau^2). \tag{2}$$

Equations (1) and (2) determine the hierarchical normal linear model.

As stated in the introduction, in some situations a researcher wants to make claims of the type "$\theta_j > \theta_k$" with confidence (for some j and k) and it would be interesting to know how such claims can be calibrated using the Type S error rate.

For our error rate calculations, we shall need to know the joint distribution of $\theta_j - \theta_k$ and $y_j - y_k$. Given $\theta_j - \theta_k$ and σ, $y_j - y_k$ follows a normal distribution

$$y_j - y_k|\theta_j, \theta_k, \sigma \sim \mathrm{N}(\theta_j - \theta_k, 2\sigma^2), \tag{3}$$

and the distribution of $\theta_j - \theta_k$ is also normal:

$$\theta_j - \theta_k|\mu, \tau \sim \mathrm{N}(0, 2\tau^2). \tag{4}$$

Because of Equations (3) and (4), the joint distribution of $(y_j - y_k, \theta_j - \theta_k)$ will be bivariate normal. The mean of $\theta_j - \theta_k$ is 0, and the mean of $y_j - y_k$ can be determined as follows:

$$\mathrm{E}(y_j - y_k) = \mathrm{E}(\mathrm{E}(y_j - y_k|\theta_j - \theta_k)) = 0,$$

with variance

$$\begin{aligned} \mathrm{var}(y_j - y_k) &= \mathrm{E}(\mathrm{var}(y_j - y_k|\theta_j - \theta_k)) + \mathrm{var}(\mathrm{E}(y_j - y_k|\theta_j - \theta_k)) \\ &= 2\sigma^2 + 2\tau^2. \end{aligned}$$

The covariance can be established by equating the mean of the distribution of $y_j - y_k$ given $\theta_j - \theta_k$ (i.e., $\theta_j - \theta_k$) to the general formula for the mean of conditional normal distributions

$$\theta_j - \theta_k = \mathrm{E}(y_j - y_k) + \frac{\mathrm{cov}(y_j - y_k, \theta_j - \theta_k)}{\mathrm{var}(\theta_j - \theta_k)}(\theta_j - \theta_k - \mathrm{E}(\theta_j - \theta_k)).$$

Solving $\text{cov}(y_j - y_k, \theta_j - \theta_k)$ from this formula leads to

$$\text{cov}(y_j - y_k, \theta_j - \theta_k) = 2\tau^2.$$

Summarized, the distribution of $(y_j - y_k, \theta_j - \theta_k)$ can be symbolized as follows

$$(y_j - y_k, \theta_j - \theta_k) \sim \text{N}_2(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\beta}$ is the vector containing the means and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $(y_j - y_k, \theta_j - \theta_k)$.

2.2 Classical and Bayesian intervals and confidence statements

In this paper, we perform calculations of the probabilities of Type S errors for classical confidence statements and Bayesian posterior intervals under the proposed hierarchical model. The $100(1-\alpha)\%$ Bayesian posterior interval we consider is a so-called central posterior interval (see, e.g., Gelman et al., 1995) bounded by the posterior $\alpha/2$ and $1-\alpha/2$ quantiles.

Our calculations are both frequentist and Bayesian: frequentist because they evaluate long-term error rates (i.e., under repeated sampling), and Bayesian (or empirical Bayesian, in the sense of Morris, 1983) because the replications average over the population distribution for the θ_j's. As is discussed by Rubin (1984) both classical and Bayesian inferential statements can be considered as data summaries and evaluated using frequentist methods.

We separately consider confidence statements based on classical confidence intervals and Bayesian posterior intervals. For convenience, we work with the conventional 95% intervals; this work generalizes in the obvious way to other probability statements. The classical 95% interval for $\theta_j - \theta_k$ is simply $[(y_j - y_k) \pm 1.96\sqrt{2}\sigma]$, and so a classical claim is made "with confidence" if the absolute difference between the two observed derived measurements exceeds a threshold

$$\text{Classical threshold: } |y_j - y_k| > 1.96\sqrt{2}\sigma. \tag{5}$$

As can be seen in Equation (5), this formula makes no reference to the hierarchical structure of the data.

The Bayesian interval is based on the posterior distribution of θ. We need to stress that we consider for the moment only the case where the variance σ and the hyperparameters μ and τ are assumed to be known. Equivalently, we could say that there are a large number of studies and that the sample size within each study j are also large, such that μ, τ and σ can be accurately estimated from the data in the "empirical Bayes" sense; see, e.g., Gelman et al. (1995) or Carlin and Louis (1996).

Given the data y and the parameters μ, σ, τ, the θ_j's are independent with the following distributions:

$$\theta_j | y, \sigma, \mu, \tau \sim \text{N}(\hat{\theta}_j, V_j),$$

where

$$\hat{\theta}_j = \frac{\frac{1}{\sigma^2}y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}.$$

The inverse of the posterior variance is simply the sum of the inverse of the inverse of the data variance σ^2 (i.e., the data precision) and the inverse of the prior variance (i.e., the prior precision). By factorizing the formula of the posterior mean, it can be seen easily that the posterior mean is a precision weighted average of the prior and observed derived measure.

The 95% posterior interval for $\theta_j - \theta_k$ is thus

$$[(\hat{\theta}_j - \hat{\theta}_k) \pm 1.96\sqrt{V_j + V_k}] = \left[\frac{\frac{1}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}(y_j - y_k) \pm 1.96\sqrt{\frac{2}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}}\right],$$

and so a Bayesian claim is made "with confidence" if

$$\text{Bayesian threshold: } |y_j - y_k| > 1.96\sqrt{2}\sigma\sqrt{1 + \frac{\sigma^2}{\tau^2}}. \tag{6}$$

This Bayesian threshold is always greater than the classical threshold (5), and thus the Bayesian interval is more "conservative" in the sense of being more likely to include zero. This happens because the posterior mean of $\theta_j - \theta_k$ will be somewhere between the data mean $y_j - y_k$ and the prior mean, which is zero. Hence, the posterior mean will be pulled towards zero (the amount of influence will determined by the ratio of the prior precision to the data precision). Technically, this means that some shrinkage happens.

The only difference between what we call the classical and the Bayesian procedures are in the acknowledgement of the hierarchical structure of the data. In fact, the Bayesian statements have a classical interpretation as random effects or as "predictive" inference (see, e.g., Robinson, 1991) in that they are probability statements about unobserved random variables. Conversely, the classical intervals can be interpreted as Bayesian inferences with τ set to ∞. However, we feel comfortable using the label "classical" for the unshrunken interval centered around $y_j - y_k$ since this is the standard estimate used as a starting point in classical comparisons (e.g., Scheffe, 1959).

3 Type S error rates in the hierarchical model

A Type S error occurs when a claim is made with confidence and with the wrong sign. The plots in Figure 1 illustrate the frequencies of Type S errors for Bayesian and classical statements in 2000 replications in each of three scenarios: $\tau/\sigma = 0.5$, 1, and 2. For each simulation draw, we first sampled the true difference $\theta_j - \theta_k \sim \mathrm{N}(0, 2\tau^2)$ and then sampled the observed difference $y_j - y_k \sim \mathrm{N}(\theta_j - \theta_k, 2\sigma^2)$.

From these plots we can see that, as noted above, the Bayesian thresholds for making a statement with confidence are always higher than the classical

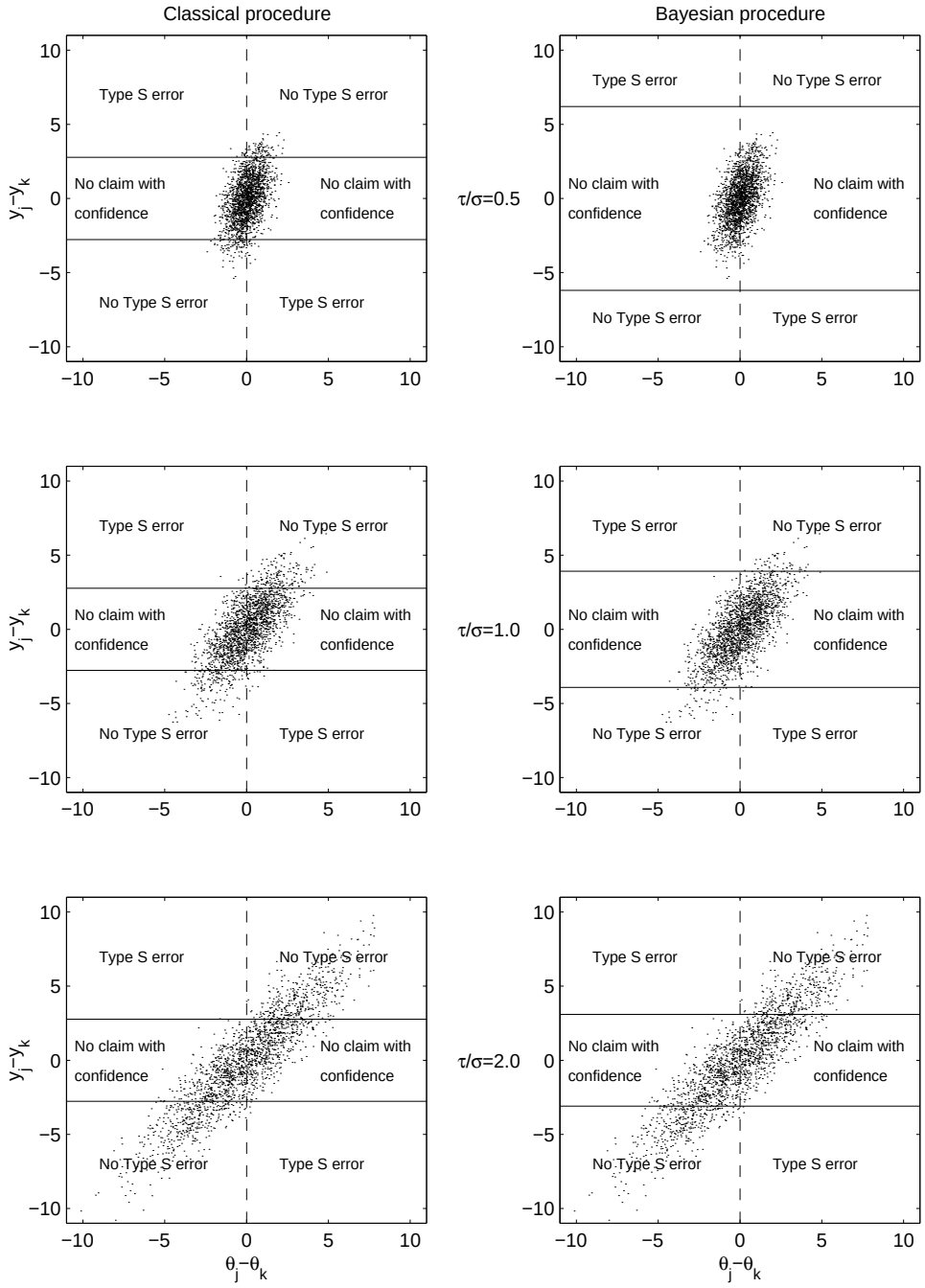

Figure 1: Illustration of claims with confidence and Type S errors for Bayesian and classical comparisons. Scatterplots show the long-run frequency properties using 2000 simulations from the hierarchical model with $\sigma = 1$ and variance ratio τ/σ set to 0.5, 1, and 2.

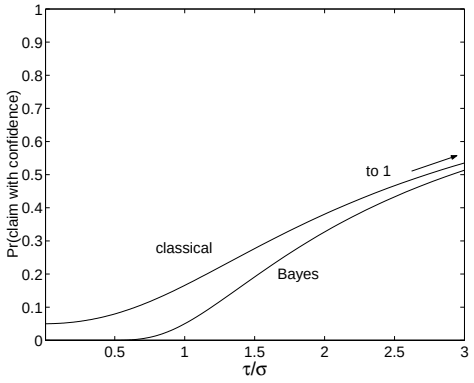

Figure 2: Probability of making a claim with confidence for classical and Bayesian comparisons: long-run frequencies are shown as a function of the variance ratio τ/σ.

threshold—in fact, for $\tau/\sigma = 0.5$, the Bayesian threshold is set so high that even in 2000 replications we would not expect to find any Bayesian claims with confidence. The difference between the two procedures decreases for higher values of τ/σ. (Figure 2 displays the frequency of Bayesian and classical claims with confidence—i.e., the "power" of the comparisons procedures—as a function of τ/σ.) The plots in Figure 1 also illustrate a familiar tradeoff: as the threshold for making statements with confidence is lowered, the rate of Type S errors increases. In this simple example, the Bayesian and classical rules for whether to make a statement "with confidence" differ only in their thresholds. The two procedures have the same receiver operating characteristic (ROC) curves (i.e., the same tradeoff relation between Pr(claiming that $\theta_j > \theta_k | \theta_j > \theta_k$) and Pr(claiming that $\theta_j > \theta_k | \theta_j < \theta_k$)) but because of the different thresholds they give quite different results for any fixed confidence level, as we shall illustrate for 95% intervals.

We examine the rate of Type S errors as a proportion of the statements made with confidence. We believe that this *conditional* probability is the appropriate error rate to consider, since our primary concern is to understand the frequency properties of claims with confidence derived from signs of 95% intervals. The conditional probability of a Type S error is,

$$\Pr(\text{Type S error} \mid \text{claim made with confidence}) = \Pr(\text{sign}(\theta_j - \theta_k) \neq \text{sign}(y_j - y_k) \mid |y_j - y_k| > T),$$

where T is a generic symbol for the threshold (which can be classical or Bayesian). To compute this probability, we have to compute a volume under

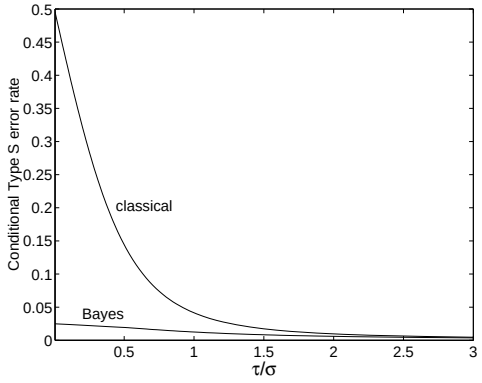

Figure 3: Conditional probability of Type S error, conditional on making a claim with confidence for classical and Bayesian comparisons, as a function of τ/σ.

a bivariate normal distribution:

$$\frac{\int_{-\infty}^{0}\int_{T}^{\infty} \mathrm{N}_2(0,\mathbf{\Sigma})d(y_j-y_k)d(\theta_j-\theta_k) + \int_{0}^{\infty}\int_{-\infty}^{-T} \mathrm{N}_2(0,\mathbf{\Sigma})d(y_j-y_k)d(\theta_j-\theta_k)}{\int_{-\infty}^{\infty}\int_{T}^{\infty} \mathrm{N}_2(0,\mathbf{\Sigma})d(y_j-y_k)d(\theta_j-\theta_k) + \int_{-\infty}^{\infty}\int_{-\infty}^{-T} \mathrm{N}_2(0,\mathbf{\Sigma})d(y_j-y_k)d(\theta_j-\theta_k)},$$

with $\mathbf{\Sigma}$, the 2×2 variance matrix of $(y_j - y_k, \theta_j - \theta_k)$, as derived in Section 2.1.

Maghsoodloo and Huang (1995) present an algorithm for approximating the bivariate normal integrals by transforming the random variables such that only cumulative distribution functions of univariate normal distributions are necessary in the calculation.

Figure 3 displays the conditional Type S error rates for the Bayesian and classical procedures as a function of τ/σ. It is no surprise that the Bayesian error rates are lower since the Bayesian threshold (6) is more stringent than the classical rule (5). What may be surprising, at first, is that neither procedure comes close to an error rate of 5%.

Consider the classical procedure first. For τ/σ near 0, the classical procedure is set up to have a Type 1 error rate of 5%, and thus to make confident claims 5% of the time (for $\tau/\sigma = 0$) or slightly more than 5% of the time (for τ/σ near 0). However, with σ so much greater than τ, the patterns in the data are mostly noise, and, in particular, even if $(y_j - y_k) > 1.96\sqrt{2}\sigma$, the true difference $\theta_1 - \theta_2$ is just about equally likely to be negative as positive (see the upper left plot in Figure 1). Thus, the conditional Type S error rate is close to 50%. Conversely, when τ/σ is large enough, essentially all of the comparisons are statistically significantly different from zero, and the Type

S error rate approaches 0 (not 5%).

Now consider the Bayesian procedure. Because we are assuming the model is known, Bayesian posterior distributions have direct long-run frequency interpretations. Consider a comparison $\theta_j - \theta_k$ that is made with confidence (for convenience, suppose the claim is that $\theta_j > \theta_k$). For such a comparison, the expected conditional Type S error rate is simply the posterior probability $\Pr(\theta_k > \theta_j | y)$, which we know must be less than 2.5% (a claim that $\theta_j > \theta_k$ with confidence means that the central 95% interval excludes 0, so less than 2.5% of the posterior distribution is in the range $\theta_j - \theta_k < 0$). Under the model, the Bayesian Type S error rate is thus bounded above by half the nominal error rate. Figure 3 shows that, for small values of τ/σ, this bound is approximately achieved, but for large values of τ/σ, as with the classical procedure, most of the posterior intervals are far from zero, and so the Type S error rate approaches 0.

How should we interpret these results? From our perspective, the usual Type 1 error rate is not particularly useful here. In particular, if τ/σ is near zero, the classical statements, when they are made with confidence, are wrong nearly half the time. Thus, in the very setting where Type 1 errors are relevant—when the null hypothesis is approximately true—we believe that the Type S error rate is more relevant than the Type 1 error rate for the key question: what is the long-run reliability of a set of statistical claims?

4 Multiple comparisons

By considering error rates, one is implicitly considering a long sequence of comparisons; this is in fact our fundamental justification for our above analysis using a hierarchical model. The distribution of the θ_j's is tautologically defined as the set of θ_j's that will appear in the long run, and this is the distribution that should be averaged over in evaluating error rates.

It is natural at this point to consider multiple comparisons procedures, which are constructed to control the probability of making at least one error in a given set of comparisons. For various multiple comparisons procedures, we study the *comparisonwise* Type S error rate (that is, among all the claims made with confidence, the proportion that are of the wrong sign) and the *experimentwise* type S error rate (that is, among the times that the set of claims made with confidence is nonempty, the proportion of such nonempty sets that contain at least one claim that is of the wrong sign).

Multiple comparisons is an extensive topic (see, e.g., Kirk, 1995), and we do not attempt a complete treatment here; rather, we illustrate the general properties of Type S errors for multiple comparisons in a relatively simple situation. We consider the canonical example of J parameters, $\theta_1, \ldots, \theta_J$, with interest in the $J(J-1)/2$ pairwise comparisons, $\theta_j - \theta_k$.

4.1 Classical multiple comparisons procedures

A classical multiple comparisons procedure requires, if the true θ_j's are all equal, that the probability of making *no* claims with confidence be at least 95%. This can be achieved in a variety of ways, including the approach of widening the confidence intervals for all pairs $\theta_j - \theta_k$ by an amount determined by J so that the experimentwise Type 1 error rate is 5%. Here, we consider two standard classical procedures: Tukey's Honestly Significant Difference test (HSD) and Wholly Significant Difference test (WSD), which are both range tests, in which all the pairwise differences are compared with a critical value under the distribution of the studentized range under the null hypothesis that the J groups are identical. Typically one starts with the largest difference, working down until no significant result is found anymore.

For the HSD procedure the reference set is always the original one of size J, with thus the same critical value for each pairwise comparison. The HSD procedure is known to be the most conservative classical procedure for pairwise comparisons (Klockars and Sax, 1986). In contrast, for the WSD procedure the critical value for a pairwise comparison depends on the remaining number of group measures, because a significant test leads to the exclusion of one of the members of the significant pair. The WSD procedure is less conservative than HSD since the critical value for pairs in later testing stages (after already some significant values are determined) is smaller.

4.2 A Bayesian multiple comparisons procedure

In contrast, we define a Bayesian multiple comparison procedure (as in Pruzek, 1997) as a set of "statements with confidence" of the form $\theta_j > \theta_k$, such that the posterior probability is 95% that all these statements are true. (Of course, it is possible that such a procedure will result in *no* statements made with confidence, which occurs if none of the parameters are statistically distinguishable from each other, in this Bayesian sense.)

Our Bayesian procedure is, like the WSD method, a multistage testing procedure. To apply it, posterior simulation draws of the vector of parameters θ has to be computed (which is straightforward in our case since the parameters μ, σ, τ are fixed in our simulations and, given these, the θ_j parameters can be sampled directly from their Gaussian posterior distribution). We then consider the statements of the form "$\theta_j > \theta_k$" one at a time, starting with the comparison that has the highest posterior probability (based on the simulations), then continuing in decreasing probability order. If none of the comparisons has at least a 95% posterior probability of being the correct sign, then we can make no statements with confidence. Otherwise, we include as many of the statements as possible, stopping when the *joint* posterior probability of all of them being true is less than 95%.

	5 groups			10 groups			15 groups		
τ/σ	Bayes	HSD	WSD	Bayes	HSD	WSD	Bayes	HSD	WSD
0.5	.001	.107	.108	.003	.128	.128	.008	.140	.140
1.0	.279	.295	.298	.626	.420	.420	.824	.519	.520
2.0	.868	.742	.744	.993	.926	.926	1.000	.973	.973

Table 1: The probability of making at least one statement with confidence for three multiple comparisons procedures—Bayesian, classical Honestly Significant Difference, and classical Wholly Significant Difference—as a function of the number of studies and the between-study standard deviation. Computations for each procedure and each value of τ/σ are based on 10000 simulations from the hierarchical normal model.

	5 groups			10 groups			15 groups		
τ/σ	Bayes	HSD	WSD	Bayes	HSD	WSD	Bayes	HSD	WSD
0.5	.000	.106	.123	.000	.100	.110	.012	.081	.085
1.0	.017	.024	.032	.018	.016	.018	.021	.014	.018
2.0	.014	.006	.010	.021	.004	.007	.025	.004	.007

Table 2: The computed *experimentwise* Type S error rate—that is, the number of simulations in which at least at least one Type S error was made, divided by the number of simulations in which at least one comparison was made with confidence—for three multiple comparisons procedures. See caption of Table 1 for more information.

4.3 Results

As before, we evaluate the Type S error rates of the classical and Bayesian procedures in the context of the hierarchical normal model with known hyperparameters. For each dataset y simulated under the model, we separately perform classical and Bayesian multiple comparisons procedures, each of which yields a (possibly empty) set of statements of the form "$\theta_j > \theta_k$ with confidence at the 95% level."

We compute both these error rates based on the 10,000 simulations, with $\sigma = 1$ and for each of the nine combinations of the variance ratio $\tau/\sigma = 0.5$, 1.0, 2.0, and the number of studies $J = 5$, 10, 15 (with the number of comparisons equal to $J(J-1)/2 = 15$, 45, 105). The results of the Bayesian procedure are based on 5000 posterior draws. Results appear in Tables 1–3. Table 1 displays the probability that at least one statement with confidence is made among all possible pairwise comparisons. Table 2 shows the probability of making at least one Type S error in a set of comparisons, conditional on at

	5 groups			10 groups			15 groups		
τ/σ	Bayes	HSD	WSD	Bayes	HSD	WSD	Bayes	HSD	WSD
0.5	.000	.083	.087	.000	.068	.071	.004	.054	.054
1.0	.013	.015	.017	.011	.006	.007	.008	.005	.005
2.0	.006	.003	.003	.003	.001	.001	.009	.000	.001

Table 3: The computed *comparisonwise* Type S error rate—that is, the number of comparisons in which a Type S error was made, divided by the number of comparisons made with confidence—for three multiple comparisons procedures. See caption of Table 1 for more information.

least one statement with confidence being made. (Thus, the denominator for the ratio computed in Table 2 is the numerator of the ratio for Table 1.) Table 3 summarizes the comparisonwise Type S error rates—that is, the proportion of confident claims that have the wrong sign. These values are lower than the probabilities in Table 2, which makes sense since the probability in Table 2 of at least one error is primarily determined by the weakest comparisons made with confidence (that is, the comparisons that are on the border of statistical significance), whereas the probability in Table 3 averages over all comparisons made with confidence, weak and strong.

These tables reveal that the Bayesian procedure is neither uniformly more nor less conservative than the classical methods. For small values of τ/σ, the Bayesian multiple comparison procedure is more conservative—it leads to fewer claims with confidence and lower Type S errors as a fraction of claims made with confidence. For τ/σ near zero, the classical procedures yield relatively high Type S error rates, as in the single comparisons setting.

For large values of τ/σ, the pattern is reversed: the Bayesian procedure yields more frequent, but less reliable statements. However, this is not such a bad thing, considering that, for moderate and high values of τ, both the classical and Bayesian procedures yield very low Type S error rates.

5 Discussion

5.1 The relevance of Type S error rates

Frequentist error rates and Bayesian hierarchical models are complementary: both are based on an ensemble of comparisons (over time or within a group) that are a priori exchangeable (in the sense that, before seeing any data, no comparison is treated any differently than any other, and all are weighted equally in the error rate).

It is standard in applied statistics, both classical and Bayesian, for comparisons to be made with confidence when interval estimates exclude zero.

When the true variation is comparable to or smaller than the estimation uncertainty, we have found that Bayesian procedures tend to be more conservative, in the sense of being less likely than classical methods to make claims with confidence. This is somewhat ironic, considering that the bias in shrinkage estimates sometimes leads them to be viewed with suspicion compared to classical procedures. The key here is that the situations in which one is worried about Type 1 errors are exactly those settings where shrinkage is most effective and most clearly motivated by the data.

For multiple comparisons problems, the conservatism of the Bayesian procedure is even more apparent. Perhaps this is one reason why multiple comparisons issues are typically ignored in Bayesian inference (see, e.g., Gelman et al., 1995, and Carlin and Louis, 1996): the multiple comparisons problem is serious when τ/σ is small, and in such settings the ordinary Bayesian inferences (without "multiple comparisons" adjustments) shrink so much that typically few or no claims are made with confidence. (See the Bayesian curve in Figure 2 for single comparisons in the range $\tau/\sigma < 0.5$.) When τ/σ is large, Bayesian classical non-multiple comparison procedures become identical, and Type S errors approach 0 in both cases.

Related findings (for single comparisons) were reported by Berger and Sellke (1987) and Berger and Delampandy (1987), comparing classical p-values to a Bayesian point null hypothesis (that is, a prior distribution for $\theta_j - \theta_k$ with a point mass at zero). Casella and Berger (1987) found smaller discrepancies between Bayesian inference with a uniform prior distribution and classical one-sided tests (which is similar to our Type S error problem except that we are considering comparisons in both directions). The point and diffuse null hypotheses correspond in our model to $\tau/\sigma = 0$ and ∞, respectively. Our analysis connects these two extremes using the hierarchical modeling framework, with the pleasing result that our findings do not depend on the assumption of a null hypothesis being true.

We emphasize that the only way that our Bayesian procedure differs from the classical method is in using the hierarchical structure of the data, which is implicitly possible given the repeated-sampling frequency interpretation. As always in these hierarchical settings, the methods will differ in practice only if τ/σ is not too large; that is, if the variance between studies is not much larger than the estimation variance for the individual studies.

5.2 Further work

Our results can be generalized in several ways. As stated above, we see the empirical Bayes hierarchical model as the natural mathematical framework for studying error rates of statistical comparisons. However, the normal model we have used is only one possible family of distributions. It would be natural to also consider longer-tailed families such as t distributions, which would yield a higher rate of statements with confidence even for small values of τ. (We are content to leave our Gaussian error model unchanged, since

the central limit theorem makes it appropriate for data y_j that are means, regression coefficients, or any other estimates with approximate normal distributions.)

Another direction towards realism is to allow the hyperparameters in the model to be estimated, rather than assumed known, for the Bayesian inference. We examined this with a simulation in which the Bayes method, rather than "knowing" the population parameters μ and τ, averages over them using the standard hierarchical Bayes method with uniform prior distribution (which reduces to the classical procedure if $J < 3$). For simplicity, we still assume σ is known in this simulation; in general, as long as sufficient data are available within the J studies, estimating σ is not a problem. We ran three small simulation studies with $J = 15$ and $\tau/\sigma = 0.5$, 1.0, 2.0, with 1000 simulations of parameters and data for each condition. For each simulation, a sample from the posterior distribution of size 2000 was drawn (see Gelman et al., 1995, Sections 5.4–5.5), yielding a (possibly empty) set of claims with confidence corresponding to all the comparisons for which the 95% simulation-based posterior intervals excluded zero. We then computed the Type S error rate as the proportion of these claims that had the wrong sign, averaging over the 1000 simulations.

We found that, as τ/σ increases from 0.5 to 1.0 to 2.0, the probability of making at least one statement with confidence increases also from 0.070 to 0.453 to 0.980. As a comparison, if τ is assumed known, the theoretical probabilities of making at least one statement with confidence are 0.008, 0.824 and 1.000 (see Table 1). Thus in general, the uncertainty about τ makes the Bayesian testing procedure less likely to make a claim with confidence (except for small values of τ/σ). Next, we examined the experimentwise Type S error rates (conditional on making at least one claim with confidence): for $\tau = 0.5$, 1.0, and 2.0, the Bayesian procedure with unknown τ has experimentwise error rates of .114, .020, and .022. Again we can compare to Type S error rates for the Bayesian procedure when τ is known, which are 0.012, 0.021 and 0.025 (see Table 2). That is, when τ is uncertain, the claims made with confidence are less reliable, especially for small values of τ. This makes sense if our prior distribution on τ is uniform on $(0, \infty)$, which tends to favor large values of τ, pushing in the direction of $\tau = \infty$, which corresponds to the classical inferences that have high Type S error rates when the true τ is small.

Our results are also relevant for problems more complicated than simple exchangeable models. If the studies differ in known ways, and these known differences can be expressed as linear predictors or as unequal variances, then τ^2 and σ^2 can be interpreted as residual variances in a hierarchical regression model.

Another direction of further work is to study the multiple comparisons procedures more carefully. In particular, one can examine Type S error rates of various subsets of comparisons that lie between the extremes of compar-

isonwise and experimentwise errors. This might be relevant for practical uses of multiple comparison procedures to distinguish reliably between subsets of an ordered list of parameter estimates.

Finally, one can consider the rates of other kinds of error. For example, a problem that arises in many studies is the overestimation of true effects, a bias that can occur because, conditional on an estimate being statistically significant, it is by necessity far from zero. We can define the Type M (for magnitude) error as the overestimation of a true difference, and then the conditional Type M error rate is the probability that the stated 95% interval is larger than the true parameter value, conditional on a claim being made with confidence. Other definitions of error rates would perhaps be relevant for other concerns, with the idea being that different statistical criteria can be relevant in different settings.

5.3 Recommendations

We do not claim that Type S errors are the only concern, or even the most important concern, in statistical comparisons. We do believe, however, that in a situation in which one is making repeated or multiple comparisons with a concern about possible errors, that the conditional Type S error rate is more relevant than the usual Type 1 error rate in evaluating the long-term frequency properties of statistical inferences for continuous parameters.

In particular, the dramatically high Type S error rates of classical procedures when τ is near zero suggests that, for the purposes of inference (as opposed to testing), control over Type 1 errors—in either the single or multiple comparisons case—does not necessarily mean that statistically significant claims are reliable. This casts some doubt on those statistical procedures if in some cases half of the classical comparisons are false in the sense of having the wrong sign.

The purpose of this paper, however, is not to make a claim that Bayesian intervals are better or worse than classical intervals (such a comparison is best made by comparing precision and empirical coverage of the intervals, as in Gelman and Little, 1997). Rather, we demonstrate that the Type S error rate—which we argue is the relevant error rate for statistical analyses in the social and behavioral sciences—differs dramatically from the Type 1 error rate, which is usually used to calibrate interval estimation procedures. This difference occurs for both classical and Bayesian procedures, and even in the best-case scenario in which the assumed model is true.

We conclude by addressing the following question: how can we be so dismissive of the Type 1 error, given that it is such a popular concept in statistics? Our answer is that Type 1 error calculations are important for understanding p-values for testing a null hypothesis, both in classical and Bayesian settings (Meng, 1994, Robins et al., 1998). However, for evaluating the frequency properties of statistical inferences about comparisons, we see Type S errors as the more relevant concept. (Comparisons are special be-

cause, for them, zero is a natural point of comparison. For inference about quantities of interest without this symmetry, other sorts of errors—for example, the probability of claiming that an effect is large when it is in fact small, or vice-versa—might be more relevant.)

In any case, our recommendation in constructing inferential procedures is *not* to try to recalibrate to a fixed Type S error rate, but rather to recognize that, even in the context of the "alternative hypothesis" that $\theta_j \neq \theta_k$, error rates can be empirically and theoretically evaluated.

References

Berger, J. O., and Delampandy, M. (1987). Testing precise hypotheses (with discussion). *Statistical Science, 2*, 317–352.

Berger, J. O., and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P-values and evidence (with discussion). *Journal of the American Statistical Association, 82*, 112–139.

Carlin, B. P., and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

Casella, G., and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *Journal of the American Statistical Association, 82*, 106–111.

Gelman, A. (1996). Discussion of "Hierarchical generalized linear models," by Y. Lee and J. A. Nelder. *Journal of the Royal Statistical Society B*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Gelman, A., and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology, 23*, 127–135.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In *What if there were no Significance Tests?*, ed. L. L. Harlow, S. A. Mulaik, and J. H. Steiger, 145–174. Mahwah, N.J.: Lawrence Erlbaum Associates.

Klockars, A.J., and Sax, G. (1986). *Multiple Comparisons*. Newbury Park: Sage.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, third edition. Brooks/Cole.

Maghsoodloo, S., and Huang, C. L. (1995) Computing probability integrals of a bivariate normal distribution. *Interstat*. `http://interstat.stat.vt.edu/`

Meng, X. L. (1994). Posterior predictive p-values. *Annals of Statistics, 22*, 1142–1160.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Association, 78*, 47–65.

Pruzek, R. M. (1997). An introduction to Bayesian inference and its applications. In *What if there were no Significance Tests?*, ed. L. L. Harlow, S. A. Mulaik, and J. H. Steiger, 287–318. Mahwah, N.J.: Lawrence Erlbaum Associates.

Rindskopf, D. M. (1997). Testing "small," not null, hypotheses: classical and Bayesian approaches. In *What if there were no Significance Tests?*, ed. L. L. Harlow, S. A. Mulaik, and J. H. Steiger, 319–332. Mahwah, N.J.: Lawrence Erlbaum Associates.

Robins, J. M., van der Vaart, A., and Ventura, V. (1998). The asymptotic distribution of p-values in composite null models. Technical report.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science, 6*, 15–51.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12*, 1151–1172.

Scheffe, H. (1959). *The Analysis of Variance.* New York: Wiley.

Tukey, J. W. (1960). Conclusions vs. decisions. *Technometrics, 2*, 423–433.