*Memorial Sloan-Kettering Cancer Center*

Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

*Year* 2005 *Paper* 1

# The Bayesian two-sample t-test

Mithat Gonen[*] Wesley O. Johnson[†]

Yonggang Lu[‡] Peter H. Westfall[**]

[*] Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org

[†] University of California-Irvine, wjohnson@ics.uci.edu

[‡] Texas Tech University, yonggang.lu@ttu.edu

[**] Texas Tech University, peter.westfall@ttu.edu

# The Bayesian two-sample t-test

Mithat Gonen, Wesley O. Johnson, Yonggang Lu, and Peter H. Westfall

## Abstract

In this article we show how the pooled-variance two-sample t-statistic arises from a Bayesian formulation of the two-sided point null testing problem, with emphasis on teaching. We identify a reasonable and useful prior giving a closed-form Bayes factor that can be written in terms of the distribution of the two-sample t-statistic under the null and alternative hypotheses respectively. This provides a Bayesian motivation for the two-sample t-statistic, which has heretofore been buried as a special case of more complex linear models, or given only roughly via analytic or Monte Carlo approximations. The resulting formulation of the Bayesian test is easy to apply in practice, and also easy to teach in an introductory course that emphasizes Bayesian methods. The priors are easy to use and simple to elicit, and the posterior probabilities are easily computed using available software, in some cases using spreadsheets.

# The Bayesian Two-Sample $t$-Test

**Mithat Gönen, Wesley O. Johnson, Yonggang Lu, and Peter H. Westfall**

SUMMARY. In this article we show how the pooled-variance two-sample $t$-statistic arises from a Bayesian formulation of the two-sided point null testing problem, with emphasis on teaching. We identify a reasonable and useful prior giving a closed-form Bayes factor that can be written in terms of the distribution of the two-sample $t$-statistic under the null and alternative hypotheses respectively. This provides a Bayesian motivation for the two-sample $t$-statistic, which has heretofore been buried as a special case of more complex linear models, or given only roughly via analytic or Monte Carlo approximations. The resulting formulation of the Bayesian test is easy to apply in practice, and also easy to teach in an introductory course that emphasizes Bayesian methods. The priors are easy to use and simple to elicit, and the posterior probabilities are easily computed using available software, in some cases using spreadsheets.

[1] Mithat Gönen is Attending Biostatistician, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10021 (E-mail: gonenm@mskcc.org). Wesley O. Johnson is Professor, Department of Statistics, University of California at Irvine, Irvine, CA 92697 (E-mail: wjohnson@ics.uci.edu). Yonggang Lu is a Ph.D. student and Peter H. Westfall is Professor of Statistics, Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock TX 79409 (E-mail: gentlelu@yahoo.com and peter.westfall@ttu.edu). The author order is alphabetical. The authors are grateful to the referees, the associate editor, and the editor for their suggestions that greatly improved the article.

KEY WORDS: Bayes factor, posterior probability, prior elicitation, teaching Bayesian statistics.

# 1 INTRODUCTION AND THE TEST

The two-sample comparison is a staple in elementary statistics courses. A typical course sequence is as follows: one-sample problems (means and proportions, tests and intervals), two-sample comparisons (differences of means and proportions, tests and intervals), then more advanced topics (ANOVA, regression). Single-sample problems involving the selection of a population reference value for the mean, $\mu_0$, are less interesting than their two-sample counterparts. Most designed experiments involve this latter category, where the samples are experimental and control (drug and placebo in most clinical trials), and interesting applications also exist in virtually all areas of scientific inquiry.

Assuming the data $y_{ir}$ $(i = 1, 2; r = 1, \ldots, n_i)$ are independent and normally distributed with means $\mu_i$ and common variance $\sigma^2$, the pooled-variance two-sample $t$-test is commonly used for testing $H_0 : \mu_1 = \mu_2$ against the two-sided alternative $H_1 : \mu_1 \neq \mu_2$. The test statistic is

$$t = \frac{\overline{y}_1 - \overline{y}_2}{s_p/n_\delta^{1/2}}, \tag{1}$$

where

$$s_p^2 = \{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\}/(n_1 + n_2 - 2)$$

is the pooled variance estimate, $\overline{y}_i$ and $s_i^2$ are the sample mean and sample variance for group $i$, and

$$n_\delta = (n_1^{-1} + n_2^{-1})^{-1},$$

which may be called the "effective sample size" for the two-sample experiment. Letting $\nu = n_1 + n_2 - 2$ denote the degrees of freedom and $t\{1-\alpha/2, \nu\}$ denote the $1-\alpha/2$ quantile of the $T_\nu$ distribution, $H_0$ is rejected in favor of $H_1$ when $|t| \geq t\{1-\alpha/2, \nu\}$; the two-sided $p$-value is obtained as $p = 2 \times P(T \geq |t|)$, where $T$ has the $T_\nu$ distribution. This test has many optimality properties (Lehmann, 1986), it is routinely produced by statistical software, and it is found in most elementary statistics texts.

While the two-sample $t$-statistic is well understood and widely accepted, it is difficult to find motivation for it in the Bayesian hypothesis testing literature. Recent literature suggesting that we should teach Bayesian methods at the elementary learning stage includes Albert (1997a), Albert and Rossman (2001), Antleman (1997), Berry (1996, 1997) and Bolstad (2004); however, none of these discuss the two-sample $t$-statistic, at least not from the Bayesian formulation of hypothesis testing.

In the general Bayesian formulation of hypothesis testing, one places prior probabilities $\pi_0$ and $\pi_1$ ($\pi_0 + \pi_1 = 1$) on hypotheses $H_0$ and $H_1$, respectively, then updates these values via Bayes theorem to obtain the posterior probabilities

$$P(H_j \,|\, \text{data}) = \frac{\pi_j P(\text{data} \,|\, H_j)}{\pi_0 P(\text{data} \,|\, H_0) + \pi_1 P(\text{data} \,|\, H_1)}, \quad j = 0, 1,$$

where $P(\text{data}|H_j)$ denotes the marginal density of the data under hypothesis $j$. Since the posterior probabilities are sensitive to the priors $\pi_0$ and $\pi_1$, it is often suggested to use the Bayes factor ($BF$) instead:

$$BF = \frac{P(\text{data} \,|\, H_0)}{P(\text{data} \,|\, H_1)}.$$

When $BF > 1$ the data provide evidence for $H_0$, and when $BF < 1$ the data provide evidence for $H_1$ (and against $H_0$). Jeffreys (1961) suggests $BF < 0.1$ provides "strong" evidence against $H_0$ and $BF < 0.01$ provides "decisive" evidence. The posterior probability is simply related to the Bayes factor as

$$P(H_0 \,|\, \text{data}) = \left[1 + \frac{\pi_1}{\pi_0}\frac{1}{BF}\right]^{-1}.$$

Much of the literature on Bayes factors and posterior probabilities is concerned with calculating or approximating (either analytically or via Monte Carlo) the marginal densities

$$P(\text{data} \,|\, H_j) = \int P(\text{data} \,|\, \boldsymbol{\theta}_j, H_j)\Pi_j(\boldsymbol{\theta}_j \,|\, H_j)\, d\boldsymbol{\theta}_j,$$

where $\boldsymbol{\theta}_j$ is the parameter vector under hypothesis $H_j$ and $\Pi_j(\boldsymbol{\theta}_j \,|\, H_j)$ is its prior distribution. Relevant references are Jeffreys (1961), Dickey (1971), Zellner and Siow (1980), Berger and Sellke (1987), Bernardo and Smith (1994), Carlin and Chib (1995), Chib (1995), Kass and Raftery (1995), and Albert, (1997b).

When considering the two-sample case in particular where the hypotheses are $H_0 : \mu_1 = \mu_2 = \mu$, vs. $H_1 : \mu_1 \neq \mu_2$, the parameter vectors are $\boldsymbol{\theta}_0 = (\mu, \sigma^2)$ and $\boldsymbol{\theta}_1 = (\mu_1, \mu_2, \sigma^2)$, and one may consider a variety of priors $\Pi_j(\boldsymbol{\theta}_j \,|\, H_j)$. Such analyses for the Bayesian two-sample $t$ test are found in the literature, but only implicitly as a special cases of more complex regression formulations, or as related to the estimation problem as in Bolstad (2004).

The aim of this paper is to elucidate a reasonable model, prior, and rather simple results that occur in this important special case.

For the two-sample problem with normally distributed, homoscedastic, and independent data, with prior distributions as specified in Section 2, the Bayes factor for testing $H_0 : \mu_1 = \mu_2 = \mu$, vs. $H_1 : \mu_1 \neq \mu_2$ is

$$BF = \frac{T_\nu(t\,|\,0, 1)}{T_\nu(t\,|\,n_\delta^{1/2}\lambda, 1 + n_\delta\sigma_\delta^2)}. \tag{2}$$

Here $t$ is the pooled variance two-sample $t$ statistic (1), $\lambda$ and $\sigma_\delta^2$ denote the prior mean and variance of the standardized effect size $(\mu_1 - \mu_2)/\sigma$ under $H_1$, and $T_\nu(.\,|\,a, b)$ denotes the noncentral $t$ probability density function (pdf) having location $a$, scale $b^{1/2}$, and df $\nu$. Specifically, $T_\nu(.\,|\,a, b)$ is the pdf of the random variable $Y/\sqrt{U/\nu}$, where $Y$ is distributed normally with mean $a$ and variance $b$, and where $U$ has the chi-square distribution with $\nu$ degrees of freedom, independent of $Y$. The mathematical derivation of (2) and further details are available on-line (Gönen et al., 2004). The data enter the $BF$ only through the pooled-variance two-sample $t$-statistic (1), providing a Bayesian motivation for its use. Benefits of having the analytic result (2) are: (i) one can explain Bayesian tests in terms of unconditional (central and non-central $T$) distributions, (ii) it allows simple sensitivity analysis with respect to prior inputs, as we show in Section 4, and (iii) it allows for a simple explanation of "Lindley's Paradox" (Lindley, 1957), which we also illustrate in Section 4.

Calculation of (2) requires evaluation of the noncentral $T$ pdf with general scale parameter. Many software packages provide the pdf of the noncentral $t$ having scale parameter 1.0, and a simple modification is needed for the general case: $T_\nu(t\,|\,a, b) = T_v(t/b^{1/2}\,|\,a/b^{1/2}, 1)/b^{1/2}$. Thus, for example, using the statistics freeware package `R` (http://www.r-project.org/), the Bayes factor can be computed as

$$BF = \text{dt(t,n1+n2-2)/(dt(t/sqrt(postv),n1+n2-2,nc)/sqrt(postv))},$$

where 't' is the value of the two-sample $t$-statistic, postv $= 1 + n_\delta \sigma_\delta^2$ and nc$=$ $n_\delta^{1/2}\lambda/(1+n_\delta\sigma_\delta^2)^{1/2}$. The noncentral $t$ density is also available in commercial packages including SAS, SPSS, and *Mathematica*, and it may be obtained using specialized programs or add-ins with other packages as well. For the case where the prior mean $\lambda$ of the effect size is assumed to be zero, the Bayes factor requires only the central $T$ pdf and is calculated more simply (e.g., using a spreadsheet) as

$$BF = \left[\frac{1 + t^2/\nu}{1 + t^2/\{\nu(1+n_\delta\sigma_\delta^2)\}}\right]^{-(\nu+1)/2} (1 + n_\delta\sigma_\delta^2)^{1/2}.$$

Assessment of priors is discussed generically in Section 2, and Section 3 discusses prior selection in a specific context involving clinical trials. Section 4 presents an analysis of a data set comparing blood pressure drop in patients receiving either calcium supplements or placebo, along with a sensitivity analysis, and Section 5 concludes.

# 2 PRIOR DISTRIBUTION AND ASSESSMENT

Let $N(y \,|\, a, b)$ denote the pdf of a normally distributed random variable with mean $a$ and variance $b$, and as usual, $Y \sim N(a,b)$ means that $Y$ has pdf $N(y \,|\, a, b)$. The assumption for the two-sample $t$-test is that the data are conditionally independent with $Y_{ir}|\{\mu_i, \sigma^2\} \sim N(\mu_i, \sigma^2)$. The goal is to test the null hypothesis $H_0 : \delta = \mu_1 - \mu_2 = 0$ against the two-sided alternative $H_1 : \delta \neq 0$.

In order to obtain the usual two-sample $t$ statistic, prior knowledge is modeled for $\delta/\sigma$ rather than for $\delta$. Let $\mu = (\mu_1 + \mu_2)/2$, and reparameterize $(\mu_1, \mu_2, \sigma^2)$ to $(\mu, \delta, \sigma^2)$. The prior for $\delta/\sigma$ is specified as

$$\delta/\sigma \,|\, \{\mu, \sigma^2, \delta/\sigma \neq 0\} \sim N(\lambda, \sigma_\delta^2).$$

For Jeffreys (1961), dependence of the prior for $\delta$ on the value of $\sigma$ is implicit in his assertion "from conditions of similarity, it [the mean] must depend on $\sigma$, since there is nothing in the problem except $\sigma$ to give a scale for [the mean]." This dependence is also found in Dickey (1971), Zellner and Siow (1980) and Berger et al. (1997).

The standardized effect size $\delta/\sigma$ is a familiar dimensionless quantity, easily modeled *a priori*. Cohen (1988) reports that $|\delta/\sigma|$ values of 0.20, 0.50, and 0.80 are "small," "medium," and "large," respectively, based on a survey of studies reported in the social sciences literature. These benchmarks can be used to check whether the specifications of hyperparameters $\lambda$ and $\sigma_\delta^2$ are reasonable; a simple check based on $\lambda \pm 3\sigma_\delta$ can determine whether the prior allows unreasonably large effect sizes.

The remaining parameters $(\mu, \sigma^2)$ are assigned a standard non-informative prior, no matter whether $\delta = 0$ or $\delta \neq 0$. While non-informative priors are attractive in the sense of minimizing prior inputs, they also ensure that the Bayes factor depends on the data only through the two-sample $t$ statistic. One can verify numerically that two different data sets having identical $t$

statistics and sample sizes can yield different Bayes factors when the prior for $(\mu, \sigma^2)$ is informative.

To summarize, the prior is as follows:

$$\Pi(\delta\,/\sigma|\,\mu, \sigma^2, \delta \neq 0) = N(\delta/\sigma\,|\,\lambda, \sigma_\delta^2),$$

with the nuisance parameters assigned the improper prior

$$\Pi(\mu, \sigma^2) \propto 1/\sigma^2.$$

Finally, the prior is completed by specifying the probability that $H_0$ is true:

$$\pi_0 = P(\delta = 0),$$

where $\pi_0$ is often taken to be 1/2 as an "objective" value (Berger and Sellke, 1987). However, $\pi_0$ can be simply assigned by the experimenter to reflect prior belief in the null; it can be assigned to differentially penalize more complex models (Jeffreys, 1961, p. 246); it can be assessed from multiple comparisons considerations (Jeffreys, 1961, p. 253; Westfall et al., 1997); and it can be estimated using empirical Bayes methods (Efron et al. 2001). The next section provides a case study for prior assessment.

It should be mentioned prominently that Jeffreys, who pioneered the Bayesian testing paradigm, derived a Bayesian test for $H_0 : \mu_1 = \mu_2$ that is also a function of the two-sample $t$-statistic (1). However, his test (Jeffreys, 1961, Section 5.41) uses an unusually complex prior that partitions the simple alternative $H_1 : \mu_1 \neq \mu_2$ into three disjoint events depending upon a hyperparameter $\mu$: $H_{11} : \mu_2 = \mu \neq \mu_1$, $H_{12} : \mu_1 = \mu \neq \mu_2$, and $H_{13} : \{(\mu_1 \neq \mu_2)$ and neither equals $\mu\}$. Jeffreys further suggests prior probabilities in the ratio $1 : 1/4 : 1/4 : 1/8$ for $H_0$, $H_{11}$, $H_{12}$, and $H_{13}$ respectively, adding another level of avoidable complexity. An additional concern with Jeffreys' two-sample $t$ test is that it does not accommodate prior information about the alternative hypothesis.

## 3 A CASE STUDY: CLINICAL TRIALS

This section provides a case study in clinical trials to suggest how priors can be specified. Prior information to suggest the expected effect size (i.e., the value of $\lambda$) is routinely used for sample size calculations. In clinical trials, the outcome is considered positive if it is significant in the correct tail using a standard two-sided test with Type I error probability $\alpha = 0.05$. The large-sample sample size calculation formula for two-sample tests is given by

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\delta/\sigma)^2}$$

where $n = n_1 = n_2 = 2n_\delta$ is the sample size per group and $\beta$ is the Type II error probability. The analyst must specify $\delta/\sigma$. In a study powered at $100(1-\beta)\% = 80\%$, the analyst will have used

$$\delta/\sigma = \frac{z_{1-\alpha/2} + z_{1-\beta}}{n_\delta^{1/2}},$$

or $\delta/\sigma = (1.96 + 0.84)/n_\delta^{1/2} = 2.80/n_\delta^{1/2}$ as an anticipated standardized effect size. For example, if $n = 100$, then the analyst anticipated $\delta/\sigma = 2.80/50^{1/2} = 0.396$ ("small" to "medium" in the terminology of Cohen).

The value $\sigma_\delta$ can be expressed as a function of the prior probability that the effect is in the wrong direction. For example, if $\lambda = 0.396$ and one thinks $P(\delta < 0 \,|\, \delta \neq 0) = 0.10$, then one obtains $\sigma_\delta = 0.309$ using normal distribution calculations. More generally, if $\lambda = 2.80/n_\delta^{1/2}$, then $\sigma_\delta = 2.19/n_\delta^{1/2}$, again assuming $P(\delta < 0 \,|\, \delta \neq 0) = 0.10$. These calculations involved the choice of zero for the tenth percentile of the prior on $\delta/\sigma$; other percentiles could have been selected as well. Yet another calibration would involve selection of $\sigma_\delta$ based on a prior assumed value for $P(\delta/\sigma > 2\lambda \,|\, \delta \neq 0)$. It would be useful to try several such values to ensure consistency.

The remaining parameter to specify is $\pi_0 = P(H_0)$. Observing that it is unethical to randomize patients when the outcome is certain, the quantities $P(\delta \leq 0\,)$ and $P(\delta > 0\,)$ should be roughly comparable. One may set $\pi_0 = 0.5$, which, in conjunction with $P(\delta < 0 \,|\, \delta \neq 0) = 0.10$, yields $P(\delta \leq 0\,) = 0.5 + \; 0.10(0.5) = 0.55$. Alternatively, one may first set $P(\delta \leq 0\,) = 0.5$, which, in conjunction with $P(\delta < 0 \,|\, \delta \neq 0) = 0.10$, implies $\pi_0 = 0.444$.

If historical (meta-analysis) data are available on rejection rates, one can check whether the prior specification is consistent with historical data by calculating the proportion of nulls that would be expected to be rejected. Since (for large sample sizes) the $t$-statistic is approximately distributed as $N(0, 1)$ when $\delta = 0$, and approximately (marginally) distributed as $N(n_\delta^{1/2}\lambda, 1 + n_\delta \sigma_\delta^2)$ when $\delta \neq 0$, the proportion of rejected nulls (upper-tailed, $\alpha = 0.025$) is expected to be

$$\pi_0(0.025) + \pi_1 \left[ 1 - \Phi\left( \frac{1.96 - n_\delta^{1/2}\lambda}{\sqrt{1 + n_\delta \sigma_\delta^2}} \right) \right].$$

Using, as suggested above, $\lambda = 2.80/n_\delta^{1/2}$, and $\sigma_\delta = 2.19/n_\delta^{1/2}$, this expression yields 33.1% rejections when $\pi_0 = 0.5$ and 36.5% when $\pi_0 = 0.444$. For comparison, Lee and Zelen (2000) surveyed the oncology literature for a variety of diseases and found that only 28.7% of the randomized trials reported rejection of the null hypothesis. Hence the choice of $\pi_0 = 0.5$, along with $(\lambda, \sigma_\delta) = (2.80/n_\delta^{1/2}, 2.19/n_\delta^{1/2})$, yields a model that is roughly consistent with results of randomized trials, at least in oncology.

# 4 AN EXAMPLE

The Data and Story Library (DASL; the website is http://lib.stat.cmu.edu/DASL) provides data sets that illustrate the use of basic statistical methods. Under the "Pooled t test" method one finds the "Calcium and Blood Pressure Story," which contains a subset of the data shown in Lyle et al. (1987). As posted on the DASL website, the data consist of blood pressure measurements on a subgroup of 21 African-American subjects, 10 who have taken calcium supplements and 11 who have taken placebo. The primary analysis variable is the blood pressure difference ("Begin" minus "End"). Summary statistics are as follows:

| Group | n | mean | StdDev |
|---|---|---|---|
| Calcium | 10 | 5.0000 | 8.7433 |
| Placebo | 11 | -0.2727 | 5.9007 |

Here, $s_p = 7.385$, $n_\delta = 5.238$, and $t = 1.634$; the positive $t$-value suggests calcium is beneficial for reducing blood pressure. The two-sided frequentist $p$-value, from the $T_{19}$ distribution, is $p = 0.1187$.

To perform the Bayesian test, priors must be specified. The previous section provided a case study to suggest particular values based on frequentist power considerations; however, this particular study was not powered for the African-American subgroup and those results do not apply. For the purposes of discussion, we will be as generic as possible in our initial specification and then provide sensitivity analysis.

While not experts in the subject matter, we might suppose that, if there is an effect, that the direction is completely uncertain, and set $\lambda = 0$. Further, we might assume that a standardized effect size greater than 1 is unlikely; setting $\sigma_\delta = 1/3$ seems reasonable as this would imply $P(|\delta/\sigma| > 1 \,|\, H_1) = 0.003$. We now compute the Bayes factor: $BF = 0.791$, suggesting that the data support $H_1 : \mu_1 \neq \mu_2$ better than $H_0 : \mu_1 = \mu_2$. If we wish to calculate posterior probabilities, then we need the prior probabilities as well; generically we may set $\pi_0 = 0.5$. With these settings we have $P(H_0 \,|\, \text{data}) = 0.442$. While it is true that the null hypothesis that calcium has no effect is less likely after seeing the data, the results are not compelling.

Figure 1 shows a sensitivity analysis of the posterior probability $P(H_0 \,|\, \text{data})$ with respect to $\lambda$, for $\sigma_\delta = 0.01, 0.33, 0.67$, and $1.00$, assuming the prior probability is $\pi_0 = 0.5$. There is not reasonable evidence against $H_0$ no matter which combinations of the prior values $\lambda$ and $\sigma_\delta$ are chosen. Smaller posterior probabilities occur for $\lambda$ near the sample estimate $(\overline{y}_1 - \overline{y}_2)/s_p = 0.714$ and for $\sigma_\delta = 0.01$, but even these are not small enough to rule out $H_0$. The graph shows large differences in the posterior probability for different $\lambda$; e.g., if $\lambda$ is near $-1$ (meaning that, if there is a difference, then calcium is expected to be much worse than placebo for reducing blood pressure), the positive $t$-statistic $t = 1.634$ provides much more evidence for $H_0$ than for $H_1$. While this lack of sensitivity may be troubling, one can question whether such values of $\lambda$ would have been reasonable choices; after all, presumably the goal of the study was to assess whether calcium causes greater reductions in blood pressure, and therefore non-negative values of $\lambda$ might have been more plausible *a priori*.

Figure 2 shows the special case where $\lambda = 0$ and $\sigma_\delta$ is varied over a wider range. Here the minimum posterior probability is $P(H_0 \,|\, \text{data}) = 0.423$, much larger than the frequentist $p$-value ($p = 0.1187$). This graph highlights the central point of Berger and Sellke (1987); namely that $P(H_0 \,|\, \text{data})$ is typically much higher than the frequentist $p$-value. For comparison, the posterior probability that results when $t = 2.093$, for which the frequentist two-sided $p$-value is exactly 0.05, is also displayed in the graph as a dotted line. The curve corresponding to $t = 2.093$ ($p = 0.05$) illustrates Berger and Sellke's (perhaps surprising) conclusion that $H_0$ will be true in at least 30% of studies for which the $p$-value is observed to be in a small neighborhood of 0.05 (assuming that $H_0$ is true, *a priori*, in 50% of all studies considered, and assuming that the prior effect sizes for the non-null studies come from a symmetric unimodal distribution centered at 0).

While the posterior probability $P(H_0 \,|\, \text{data})$ does not appear to be overly sensitive to the prior inputs $\lambda$ and $\sigma_\delta$ (provided a sensible range of inputs is considered), it is clearly much more sensitive to the prior probability $\pi_0$. For example, when $(\lambda, \sigma_\delta) = (0, 1/3)$, the posterior probabilities are determined as follows:

| | | | | | |
|---|---|---|---|---|---|
| Prior Probability $\pi_0$: | 0.100 | 0.250 | 0.500 | 0.750 | 0.900 |
| Posterior Probability $P(H_0 \,|\, \text{data})$: | 0.081 | 0.209 | 0.442 | 0.704 | 0.877 |

The posterior is sensitive to the prior as expected, but what is more interesting is that *these* data barely modify one's prior belief about $H_0$.

As a concluding note, it is simple to discuss "Lindley's Paradox" (Lindley, 1957), using 2). Lindley had noticed that data from large sample sizes that are "highly significant" from a frequentist standpoint can support $H_0$ better than $H_1$. Imagine, in the case above, that $t = 3.00$, highly significant by any measure. From the frequentist standpoint, the result would be considered even more significant for larger values of $n_1$ and $n_2$. On the other hand, $t = 3.00$ becomes less likely under $H_1$ for extremely large $n_\delta$: the denominator of (2) decreases (since the variance $1 + n_\delta \sigma^2$ increases) while the numerator remains fixed. Figure 3 shows the effect of increasing $n_\delta$ (assuming $n_1 = n_2$) on the posterior probability of $H_0$ when $t = 3.00$, showing a minimum posterior probability of 0.055 at $n_\delta = 81.5$ ($n_1 = n_2 = 163$), and increasing to 1.0 thereafter for larger $n_\delta$. This seeming “paradox” is not really a paradox at all, since the frequentist statistical significance with large $n_\delta$ is a result of a large sample amplification of a very small effect size.

## 5 CONCLUSION

The two-sample comparison is one of the most important problems in statistics. From the teaching standpoint, two-sample testing problems are usually much more interesting and relevant than single-sample problems. However, it is difficult to find the Bayesian two-sample $t$-test explicitly in the literature. We present a simple, relatively easy-to-elicit prior for which the Bayes factor for the two-sample comparison of means is a function of the usual two-sample $t$-statistic, thus providing a Bayesian motivation for this statistic. Because the analytic result itself is easy to teach and compute, and because it facilitates discussions of Bayesian concepts such as prior selection and Lindley’s Paradox, we recommend that this test be incorporated routinely when teaching elementary statistics from a Bayesian perspective.

# REFERENCES

Albert, J. (1997a), "Teaching Bayes Rule: A Data-Oriented Approach," *The American Statistician*, 51, 247–253.

Albert, J. (1997b), "Bayesian Testing and Estimation of Association in a Two-Way Contingency Table," *Journal of the American Statistical Association*, 92, 685–693.

Albert, J., and Rossman, A. (2001), *Workshop Statistics: Discovery with Data, A Bayesian Approach*, Emeryville, CA: Key College.

Antleman, G. (1997), *Elementary Bayesian Statistics*, Cheltenham: Edward Elgar Publishing.

Berger, J. O., Boukai, B., and Wang, Y. (1997), "Unified Frequentist and Bayesian Testing of a Precise Hypothesis," *Statistical Science*, 12**,** 133–160.

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of $P$ Values and Evidence," *Journal of the American Statistical Association*, 82**,** 112–122.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Berry, D. A. (1996), *Statistics: A Bayesian Perspective*, Belmont, CA: Wadsworth.

Berry, D. A. (1997), "Teaching Elementary Bayesian Statistics with Real Applications in Science," *The American Statistician*, 51, 241–246.

Bolstad, W. M. (2004), *Introduction to Bayesian Statistics*, Hoboken, New Jersey: Wiley.

Carlin, B. P., and Chib, S. (1995), “Bayesian Model Choice via Markov Chain Monte Carlo Methods,” *Journal of the Royal Statistical Society,* Ser. B, 57, 473–484.

Chib, S. (1995), “Marginal Likelihood from the Gibbs Output,” *Journal of the American Statistical Association*, 90**,** 1313–1321.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), New York: Academic Press.

Dickey, J. M. (1971), “The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters,” *Annals of Mathematical Statistics*, 42, 204–223.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), “Empirical Bayes Analysis of a Microarray Experiment,” *Journal of the American Statistical Association*, 96, 1151–1160.

Gönen, M., Westfall, P., H., Johnson, W. O., and Lu, Y. (2004), “The Two-Sample $t$ Test: A Bayesian Perspective,” Unpublished manuscript, http://www.ba.ttu.edu/isqs/westfall/Bayes2samplet.pdf.

Jeffreys, H. (1961), *Theory of Probability*, Oxford: Oxford University Press.

Kass, R. E., and Raftery, A. E. (1995), “Bayes Factors,” *Journal of the American Statistical Association*, 90**,** 773–795.

Lehmann, E.L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York:Wiley.

Lee, S. J., and Zelen, M. (2000), “Clinical Trials and Sample Size Considerations: Another Perspective,” *Statistical Science*, 15, 95–110.

Lindley, D. V. (1957), “A Statistical Paradox,” *Biometrika*, 44, 187–192.

Lyle, R. M., Melby, C. L., Hyner, G. C., Edmondson, J. W., Miller., J. Z., and Weinberger, M. H. (1987), “Blood Pressure and Metabolic Effects of

Calcium Supplementation in Normotensive White and Black Men," *Journal of the American Medical Association*, 257, 1772–1776.

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997), "A Bayesian Perspective on the Bonferroni Adjustment," *Biometrika*, 84, 419–427.

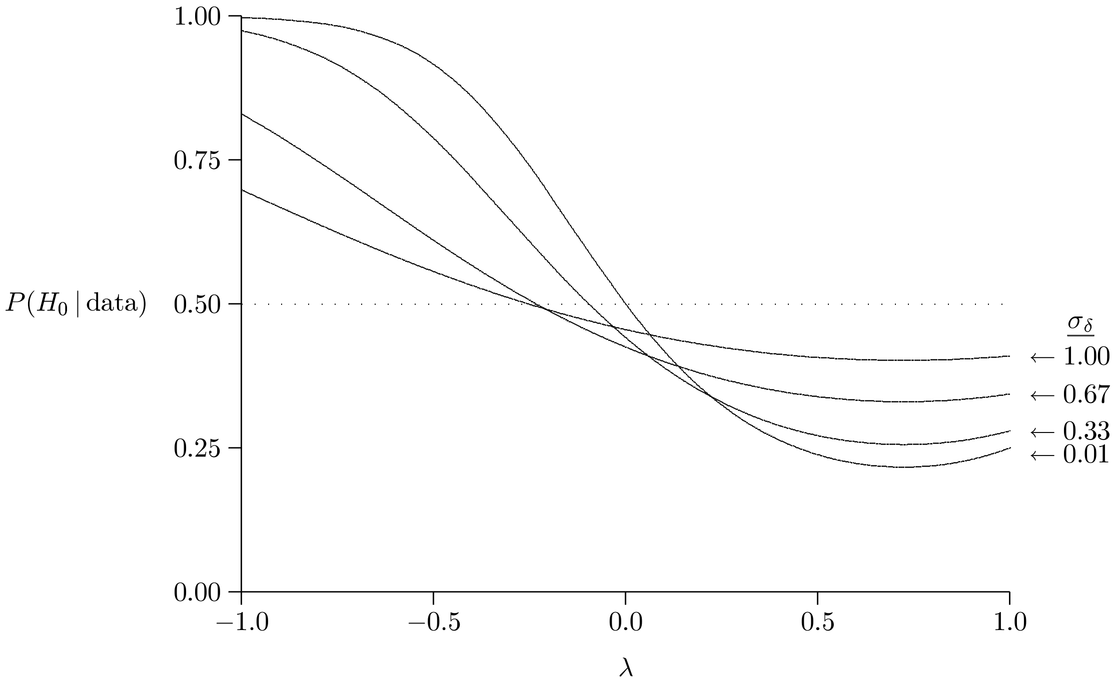
Zellner, A., and Siow, A.(1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, Valencia: University Press, pp. 585–603.

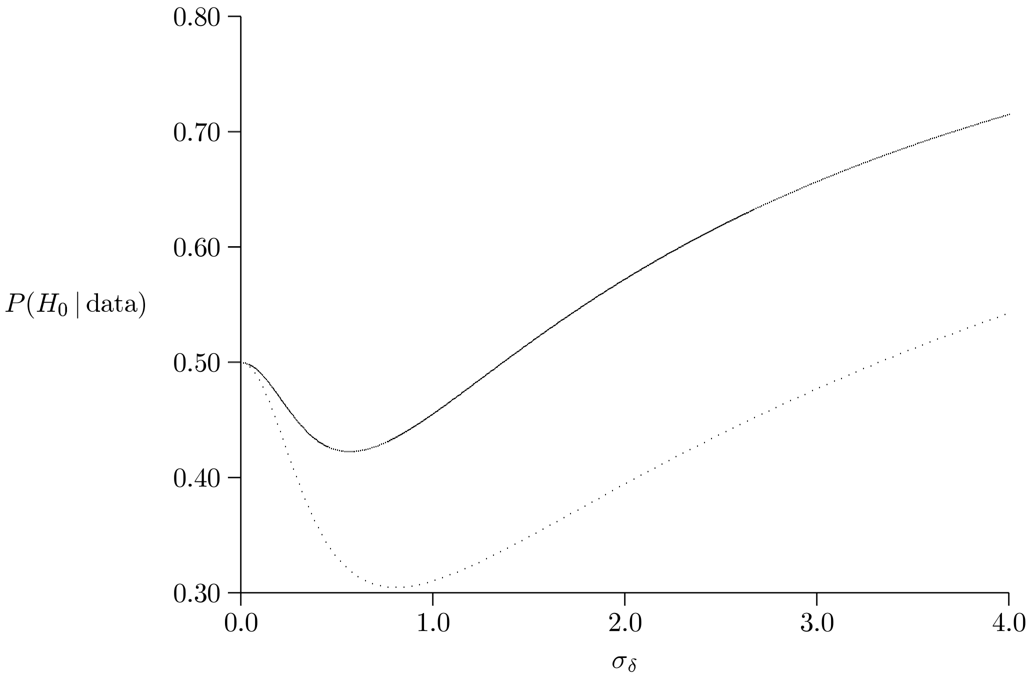
# Figure Legends

**Figure 1:** Posterior probabilities of $H_0$ as a function of $\lambda$, when $\pi_0 = 0.5$, and $\sigma_\delta = 0.01, 0.33, 0.67$, and 1.00 (solid lines). The prior probability $\pi_0 = 0.5$ is also shown (dotted line).

**Figure 2:** Posterior probability of $H_0$ as a function of $\sigma_\delta$, when $\lambda = 0$ and $\pi_0 = 0.5$, both for the observed data (solid line) where the $p$-value is $p = 0.1187$, and for hypothetical data with $p = 0.05$ (dotted line). The minimum posterior probability for the case where $p = 0.05$ is $P(H_0 \,|\, \text{data}) = 0.305$, illustrating Berger and Sellke's "irreconcilability" of frequentist $p$-values with posterior probabilities.

**Figure 3:** Posterior probability of $H_0$ as a function of $n_\delta$ when $\pi_0 = 0.5$ and $(\lambda,\, \sigma_\delta) = (0, 1/3)$ and $t = 3.00$, illustrating Lindley's paradox.


1.00
0.75
0.50
0.25
0.00
$P(H_0\,|\,\text{data})$
−1.0
−0.5
0.0
0.5
1.0
$\lambda$
$\sigma_\delta$
← 1.00
← 0.67
← 0.33
← 0.01


0.80
0.70
0.60
$P(H_0 \,|\, \mathrm{data})$
0.50
0.40
0.30
0.0
1.0
2.0
3.0
4.0
$\sigma_\delta$

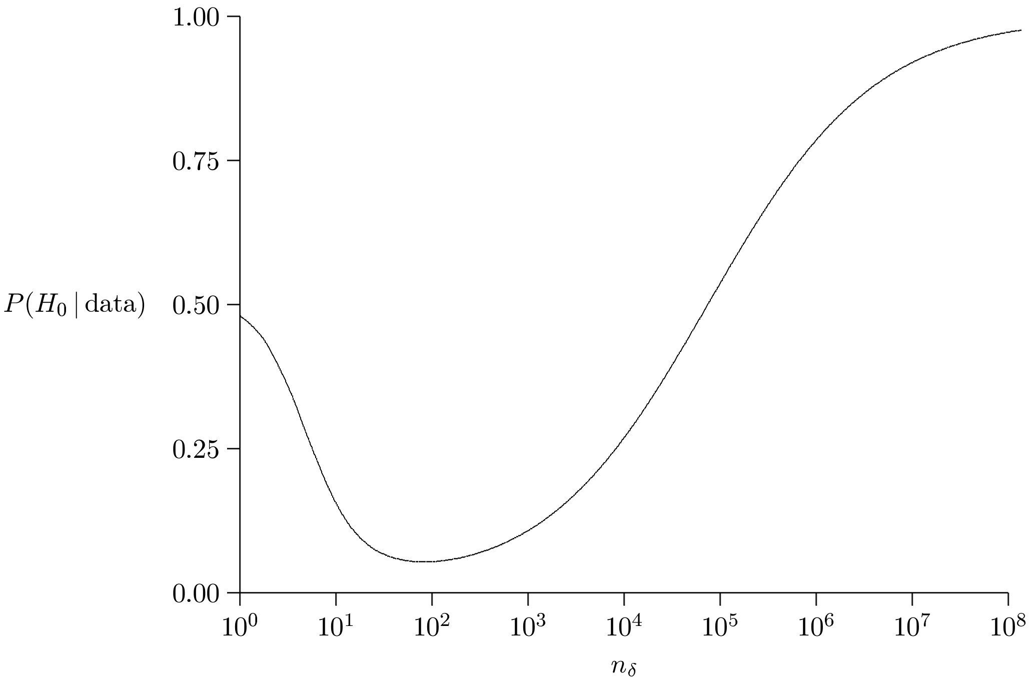

1.00
0.75
$P(H_0 \,|\, \text{data})$
0.50
0.25
0.00
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
$10^6$
$10^7$
$10^8$
$n_\delta$

Figure 3.