Journal of Experimental Psychology: General
2013, Vol. 142, No. 2, 573–603

0096-3445/13/$12.00 DOI: 10.1037/a0029146

Bayesian Estimation Supersedes the *t* Test

John K. Kruschke
Indiana University, Bloomington

Bayesian estimation for 2 groups provides complete distributions of credible values for the effect size, group means and their difference, standard deviations and their difference, and the normality of the data. The method handles outliers. The decision rule can accept the null value (unlike traditional *t* tests) when certainty in the estimate is high (unlike Bayesian model comparison using Bayes factors). The method also yields precise estimates of statistical power for various research goals. The software and programs are free and run on Macintosh, Windows, and Linux platforms.

Keywords: Bayesian statistics, effect size, robust estimation, Bayes factor, confidence interval

One of the most frequently encountered scientific procedures is a comparison of two groups (e.g., du Prel, Röhrig, Hommel, & Blettner, 2010; Fritz, Morris, & Richler, 2012; Wetzels et al., 2011). Given data from two groups, researchers ask various comparative questions: How much is one group different from another? Can we be reasonably sure that the difference is non-zero? How certain are we about the magnitude of difference? These questions are difficult to answer because data are contaminated by random variability despite researchers' efforts to minimize extraneous influences on the data. Because of "noise" in the data, researchers rely on statistical methods of probabilistic inference to interpret the data. When data are interpreted in terms of meaningful parameters in a mathematical description, such as the difference of mean parameters in two groups, it is Bayesian analysis that provides complete information about the credible parameter values. Bayesian analysis is also more intuitive than traditional methods of null hypothesis significance testing (e.g., Dienes, 2011).

This article introduces an intuitive Bayesian approach to the analysis of data from two groups. The method yields complete distributional information about the means and standard deviations of the groups. In particular, the analysis reveals the relative credibility of every possible difference of means, every possible difference of standard deviations, and all possible effect sizes. From this explicit distribution of credible parameter values, inferences about null values can be made without ever referring to *p* values as in null hypothesis significance testing (NHST). Unlike NHST, the Bayesian method can accept the null value, not only reject it, when certainty in the estimate is high. The new method handles outliers by describing the data as heavy tailed distributions instead of normal distributions, to the extent implied by the data. The new method also implements power analysis in both retrospective and prospective forms.

The analysis is implemented in the widely used and free programming languages R and JAGS and can be run on Macintosh, Linux, and Windows operating systems. Complete installation instructions are provided, along with working examples. The programs can also be flexibly extended to other types of data and analyses. Thus, the software can be used by virtually anyone who has a computer.

The article is divided into two main sections, followed by appendices. The first section introduces the Bayesian analysis and explains its results through examples. The richness of information provided by the Bayesian parameter estimation is emphasized. Bayesian power analysis is also illustrated. The second section contrasts the Bayesian approach with the *t* test from NHST. This section points out not only the relative poverty of information provided by the NHST *t* test but also some of its foundational logical problems. An appendix is provided for readers who are familiar with a different Bayesian approach to testing null hypotheses, which is based on model comparison and uses the Bayes factor as a decision statistic. This appendix suggests that Bayesian model comparison is usually less informative than the approach of Bayesian parameter estimation featured in the first section.

The perils of NHST and the merits of Bayesian data analysis have been expounded with increasing force in recent years (e.g., W. Edwards, Lindman, & Savage, 1963; Kruschke, 2010a, 2010b, 2011c; Lee & Wagenmakers, 2005; Wagenmakers, 2007). Nevertheless, some people have the impression that conclusions from NHST and Bayesian methods tend to agree in simple situations such as comparison of two groups: "Thus, if your primary question of interest can be simply expressed in a form amenable to a *t* test, say, there really is no need to try and apply the full Bayesian machinery to so simple a problem" (Brooks, 2003, p. 2694). This article shows, to the contrary, that Bayesian parameter estimation provides much richer information than the NHST *t* test and that its conclusions can differ from those of the NHST *t* test. Decisions based on Bayesian parameter estimation are better founded than those based on NHST, whether the decisions derived by the two methods agree or not. The conclusion is bold but simple: Bayesian parameter estimation supersedes the NHST *t* test.

This article was published Online First July 9, 2012.

For helpful comments on previous versions of this paper, I thank Michael Masson and especially Wolf Vanpaemel.

Supplementary information can be found at http://www.indiana.edu/~kruschke/BEST/

Correspondence concerning this article should be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th Street, Bloomington, IN 47405-7007. E-mail: kruschke@indiana.edu

The New Approach: Robust Bayesian Estimation

Bayesian Estimation Generally

Bayesian inference is merely the reallocation of credibility across a space of candidate possibilities. For example, suppose a crime is committed, and there are several possible suspects who are mutually unaffiliated. When evidence implicates one suspect, the other suspects are exonerated. This logic of exoneration is merely reallocation of belief, based on data. The complementary reallocation applies when data exonerate some suspects: Suspicion in the remaining suspects increases. Just as the fictional detective Sherlock Holmes said (Doyle, 1890), when you have eliminated the impossible, all that remains, no matter how improbable, must be the truth.

In the context of data analysis, the phenomenon to be explained is a pattern in noisy numerical data. We describe the pattern with a mathematical model such as linear regression, and the parameters in the model, such as the slope in linear regression, describe the magnitude of the trend. The space of possible "suspects" for describing the data is the space of values for the parameters. In Bayesian estimation, we reallocate belief toward the parameter values that are consistent with the data and away from parameter values that are inconsistent with the data.

A Descriptive Model for Two Groups

The first step of most statistical analyses is specifying a descriptive model for the data. The model has parameters that are meaningful to us, and our goal is to estimate the values of the parameters. For example, the traditional t test uses normal distributions to describe the data in each of two groups. The parameters of the normal distributions, namely the means (μ_1 and μ_2) and the standard deviations (σ_1 and σ_2), describe meaningful aspects of the data. In particular, the difference of the mean parameters ($\mu_1 - \mu_2$) describes the magnitude of the difference between central tendencies of the groups, and the difference of the standard-deviation parameters ($\sigma_1 - \sigma_2$) describes the magnitude of the difference between the variabilities of the groups. Our main goals as analysts are to estimate those magnitudes and to assess our uncertainty in those estimates. The Bayesian method provides answers to both goals simultaneously.

I assume that the data are measured on a metric scale (e.g., response time, temperature, weight) for both of two conditions or groups. To describe the distribution of the data, the traditional t test assumes that the data in each group come from a normal distribution (Gosset, 1908). Although the assumption of normality can be convenient for mathematical derivations, the assumption is not necessary when using numerical methods as will be used here, and the assumption is not appropriate when the data contain outliers, as is often the case for real data. A useful way to accommodate outliers is by describing the data with a distribution that has taller tails than the normal distribution. An often-used distribution for this application is the t distribution, treated here as a convenient descriptive distribution of data and not as a sampling distribution from which p values are derived. In other words, I am using the t distribution merely as a convenient way to describe data; I am not using the t distribution to conduct a t test. There is a large literature on the use of the t distribution to describe outliers (e.g., Damgaard, 2007; Jones & Faddy, 2003; Lange, Little, & Taylor, 1989; Meyer & Yu, 2000; Tsionas, 2002; Zhang, Lai, Lu, & Tong, in press). Methods of estimation that accommodate outliers are known as *robust.*

Figure 1 shows examples of the t distribution, superimposed with a normal distribution. The relative height of the tails of the t distribution is governed by a parameter denoted by the Greek letter ν (nu), which can range continuously from 1 to infinity. When ν is small, the t distribution has heavy tails, and when ν is large (e.g., 100), the t distribution is nearly normal. Therefore I will refer to ν as the *normality* parameter in the t distribution. (Traditionally, in the context of sampling distributions, this parameter is referred to as the *degrees of freedom.* Because I will not be using the t distribution in that context, I will not be using that potentially misleading nomenclature.) The t distribution can describe data with outliers by setting ν to a small value, but the t distribution can also describe data that are normal, without outliers, by setting ν to a large value. Just like the normal distribution, the t distribution has a mean parameter μ and a standard deviation parameter σ.

In the present model of the data, I will describe each group's data with a t distribution, with each group having its own mean parameter and standard deviation parameter. Because outliers are usually relatively few in number, I will use the same ν parameter for both groups so that both groups' data can inform the estimate of ν. Thus, my description of the data uses five parameters: the means of the two groups (μ_1 and μ_2), the standard deviations of the two groups (σ_1 and σ_2), and the normality of the data within the groups (ν). I will use Bayesian inference to estimate the five parameters.

As discussed above, Bayesian inference is reallocation of credibility toward parameter values that are consistent with the data. To carry out Bayesian inference, one must start with a distribution of credibility across parameter values that expresses previous knowledge about the parameter values without the newly collected data.

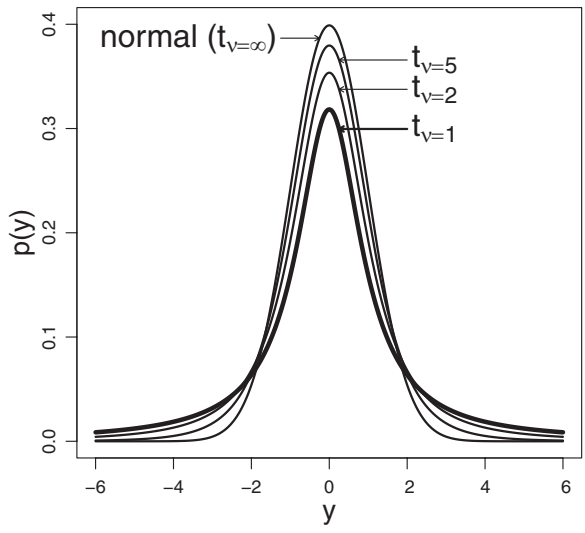

Figure 1. Examples of the t distribution, for different values of the ν parameter. When ν is small, the t distribution has heavier tails than the normal distribution. (For these examples, the mean parameter μ is set to zero, and the standard deviation parameter σ is set to 1.0.)

This allocation is called the *prior distribution.* The prior distribution must be acceptable to the skeptical scientific audience of the analysis. The prior cannot, therefore, trivially presume the desired outcome. The prior can be informed by previous findings if doing so would be appropriate for the purposes and audience of the analysis. Because this article is about general techniques, not a specific application domain, the prior distributions here are made very broad and vague, thereby expressing great prior uncertainty in the values of the parameters. This specification of an uncertain prior implies that the prior has minimal influence on the estimates of the parameters, and even a modest amount of data will overwhelm the prior assumptions when one is doing Bayesian parameter estimation.

Figure 2 depicts the descriptive model along with the prior distribution on its parameters. The *i*th datum from group *j* is denoted as y_{ji} at the bottom of the diagram. The data are described by *t* distributions, depicted in the middle of the figure. The prior distribution is indicated at the top of the figure. In particular, the prior on the mean parameters, μ_1 and μ_2, is assumed to be a very broad normal distribution, depicted in the diagram by an iconic normal shape. To keep the prior distribution broad relative to the arbitrary scale of the data, I have set the standard deviation *S* of the prior on μ to 1,000 times the standard deviation of the pooled data. The mean *M* of the prior on μ is arbitrarily set to the mean of the pooled data; this setting is done merely to keep the prior scaled appropriately relative to the arbitrary scale of the data. Thus, if *y* were a measure of distance, the scale could be nanometers or light-years and the prior would be equally noncommittal. The prior on the standard deviation parameter is also assumed to be noncommittal, expressed as a uniform distribution from a low value *L,* set to one thousandth of the standard deviation of the pooled data, to a high value *H,* set to one thousand times the standard deviation of the pooled data. Finally, the ν parameter has a prior that is exponentially distributed, which spreads prior credibility fairly evenly over nearly normal and heavy tailed data. The exact prior distribution on ν is shown in Appendix A.

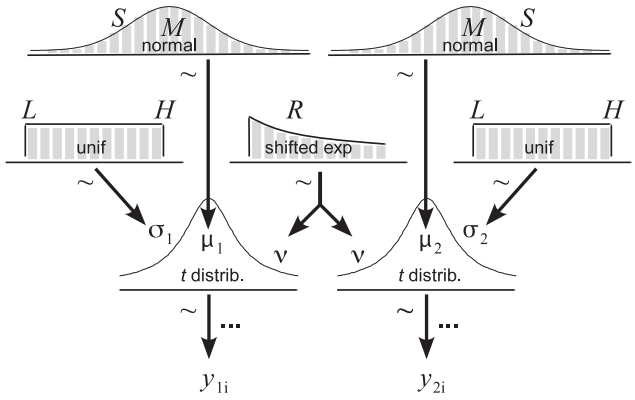

Figure 2. Hierarchical diagram of the descriptive model for robust Bayesian estimation. At the bottom of the diagram, the data from Group 1 are denoted y_{1i} and the data from Group 2 are denoted y_{2i}. The data are assumed to be described by *t* distributions, as indicated by the arrows descending from the *t*-distribution icons to the data. The ~ symbol (tilde) on each arrow indicates that the data are randomly distributed, and the "..." symbol (ellipsis) on the lower arrows indicates that all the y_i are distributed identically and independently. The two groups have different mean parameters (μ_1 and μ_2) and different standard deviation parameters (σ_1 and σ_2), and the ν parameter is shared by both groups, as indicated by the split arrow, for a total of five estimated parameters. The parameters are provided with broad, noncommittal prior distributions, as indicated by the icons in the upper part of the diagram. The prior distributions have histogram bars superimposed on them to suggest their representation by a very large random sample and their correspondence to the histograms of the posterior distributions in Figures 3–5. *S* = standard deviation; *M* = mean; *L* = low value; *H* = high value; *R* = rate; unif = uniform; shifted exp = shifted exponential; distrib. = distribution.

Flexibility: Variations and extensions. The default form of the analysis program uses a noncommittal prior that has minimal impact on the posterior distribution. Users can modify the program to specify other prior distributions if they like, as explained in Appendix B. This flexibility is useful for checking the robustness of the posterior against reasonable changes in the prior. The flexibility is also useful in applications that allow strongly informed priors based on publicly accessible previous research.

The default form of the analysis program uses *t* distributions to describe the shape of the data in each group. Users can modify the program to specify other shapes to describe the data. For example, if the data are skewed, it might be useful to describe the data with a log-normal distribution. Appendix B shows how to do this.

Robust Bayesian estimation can be extended (in the programming languages R and JAGS) to research designs with a single group or with multiple groups. In the case of data from a single group, including the case of a single group of difference scores from repeated measures on the same subjects, a modified model merely estimates μ, σ, and ν of the group. For multiple groups, on the other hand, the model of Figure 2 can be extended in two ways. First, of course, every group is provided with its own μ_j and σ_j parameters but with ν shared by all groups. Second, and importantly, the model can be provided with a higher level distribution across the group means, if desired. This higher level distribution describes the distribution of the μ_j across groups, wherein the overall mean of the groups is estimated, and the between-group variability is estimated. A major benefit of the hierarchical structure is that the estimates of the distinct group means undergo "shrinkage" toward the overall mean, to an extent determined by the actual dispersion across groups. In particular, when several groups have similar means, this similarity informs the higher level distribution to estimate small variability between groups, which, in turn, pulls the estimate of outlying groups toward the majority of the groups. The magnitude of shrinkage is informed by the data: When many groups are similar, there is more shrinkage of outlying groups. Shrinkage of estimates is a natural way to mitigate false alarms when considering multiple comparisons of groups, because shrinkage can restrain chance conspiracies of rogue data. Specification of hierarchical structure can be useful for sharing of information across group estimates, but it is not necessary and is only appropriate to the extent that the top-level distribution is a useful description of variability across groups.

Note that shrinkage is caused by the hierarchical model structure, not by Bayesian estimation. Non-Bayesian methods such as maximum likelihood estimation also show shrinkage in hierarchical models, but Bayesian methods are particularly flexible and allow many complex nonlinear hierarchical models to be easily implemented. For example, the extended model can also place a higher level distribution on the group standard deviations

(Kruschke, 2011b, Section 18.1.1.1), so that every group has its own estimated standard deviation, but the various group estimates mutually inform each other so that some degree of homogeneity of variance can be enforced to the extent that the data suggest it. Complex nonlinear hierarchical models can be very challenging for NHST procedures because of the difficulty of generating sampling distributions for computing *p* values from nested models. Further details regarding so-called hierarchical Bayesian analysis of variance (ANOVA) are provided by Gelman (2005, 2006); Gelman, Hill, and Yajima (2012); and Kruschke (2010a, 2010b, 2011b). Complete programs are provided by Kruschke (2011b; e.g., programs ANOVAonewayJagsSTZ.R and ANOVAtwowayJagsSTZ.R).

Summary of the model for Bayesian estimation. The model describes the data with five parameters: a mean and standard deviation for each group and a normality parameter shared by the groups. The prior allocation of credibility across the five-parameter space is very vague and wide, so that the prior has minimal influence on the estimation, and the data dominate the Bayesian inference. Bayesian estimation will reallocate credibility to parameter values that best accommodate the observed data. The resulting distribution is a joint distribution across the five parameters, thereby revealing combinations of the five parameter values that are credible, given the data.

The Mechanics of Bayesian Estimation

As described earlier, Bayesian inference simply reallocates credibility across the parameter values in the model, given the data. The mathematically correct way to reallocate credibility is provided by a formula called Bayes' rule (Bayes & Price, 1763). It is based on a simple relationship between conditional probabilities, but it has tremendous ramifications when applied to parameters and data. Denote the set of data as *D,* which consists of all the observed y_{ji} from both groups. Bayes' rule derives the probability of the parameter values given the data, in terms of the probability of the data given the parameter values and the prior probabilities of the parameter values. For our descriptive model in Figure 2, Bayes' rule has the following form:

$$\underbrace{p(\mu_1, \sigma_1, \mu_2, \sigma_2, \nu|D)}_{\text{posterior}} = \underbrace{p(D|\mu_1, \sigma_1, \mu_2, \sigma_2, \nu)}_{\text{likelihood}} \times \underbrace{p(\mu_1, \sigma_1, \mu_2, \sigma_2, \nu)}_{\text{prior}} \Big/ \underbrace{p(D)}_{\text{evidence}} \quad (1)$$

In words, Bayes' rule in Equation 1 simply states that the posterior credibility of the combination of values $< \mu_1, \sigma_1, \mu_2, \sigma_2, \nu >$ is the likelihood of that combination times the prior credibility of that combination, divided by the constant $p(D)$. Because it is assumed that the data are independently sampled, the likelihood is the multiplicative product across the data values of the probability density of the *t* distribution in Figure 2. The prior is the product of the five independent parameter distributions in the upper part of Figure 2. The constant $p(D)$ is called the *evidence* or the *marginal likelihood* by various authors. Its value is computed, in principle, by integrating the product of the likelihood and prior over the entire parameter space. The integral is impossible to compute analytically for many models, which was a major impediment to the widespread use of Bayesian methods until the development of modern numerical methods that obviate explicit evaluation of $p(D)$.

The posterior distribution is approximated to arbitrarily high accuracy by generating a large representative sample from it, without explicitly computing $p(D)$. A class of algorithms for doing so is called *Markov chain Monte Carlo* (MCMC) methods, and those methods are used here. The MCMC sample, also called a *chain* of values because of the way the values are generated, provides many thousands of combinations of parameter values, $<\mu_1, \sigma_1, \mu_2, \sigma_2, \nu>$. Each combination of values is representative of credible parameter values that simultaneously accommodate the observed data and the prior distribution. The thousands of representative parameter values are summarized graphically by a histogram, as shown in the prior distributions of Figure 2 and in subsequent depictions of posterior distributions. From the MCMC sample, one can easily ascertain any aspect of the credible parameter values in which one might be interested, such as the mean or modal credible value and range of credible values. Importantly, one can also examine the credible difference of means by computing $\mu_1 - \mu_2$ at every combination of representative values, and one can do the same for the difference of standard deviations. Several examples are provided below.

For computing the Bayesian inference, I will use the programming language called R (R Development Core Team, 2011) and the MCMC sampling language called JAGS, accessible from R via a package called rjags (Plummer, 2003). The programs are written in the style of programs in a recent textbook (Kruschke, 2011b). All the software is free. The software is easy to install, and it is easy to run the programs, as explained at http://www.indiana.edu/~kruschke/BEST/, where "BEST" stands for *Bayesian estimation.*

With the software and programs installed, running an analysis is easy. For a complete example, open the file BESTexample.R in R and read the comments in that file.

There are just four simple steps in conducting an analysis. First, one loads the relevant programs into R using the command source ("BEST.R"). Second, the data for the two groups are entered as vectors in R, denoted y1 and y2. Third, the MCMC chain is generated using the command mcmcChain = BESTmcmc(y1,y2). Fourth, the results are plotted, using the command BESTplot(y1,y2,mcmcChain). Examples of results are presented below.

Digression: Technical details of MCMC sampling. The process of MCMC sampling generates a large representative sample of credible parameter values from the posterior distribution. The bigger the sample is, the better it represents the underlying posterior distribution. The program defaults to an MCMC sample size of 100,000. This sample size, also called *chain length,* is adequate for typical applications.

It is important not to confuse the MCMC "sample" of parameter values with the "sample" of empirical data. There is one sample of data, which remains fixed regardless of the MCMC sample size. A longer MCMC chain merely provides a higher resolution representation of the posterior distribution of parameter values, given the fixed data.

Because the MCMC process generates a random sample of credible parameter values, its results will be slightly different on repeated analyses of the same data. These small variations are of no consequence in most applications. If, however, the user requires more stability in the MCMC approximation of the posterior, it is

easy to specify a larger chain length. The analysis program takes proportionally longer time to generate a longer chain. The user is encouraged to use as long a chain as possible.

The goal of the MCMC process is to generate an accurate and reliable representation of the posterior distribution. Unfortunately, MCMC algorithms can suffer from clumpiness (technically called autocorrelation) in the chains that they generate. One way of diluting clumpiness is by thinning the chain, which means using only every *k*th step in the chain, where *k* is an arbitrary number chosen judiciously by the user. Although the thinned chain has less clumpiness, it also is much shorter than the original chain and therefore has less reliable estimates of posterior characteristics. It turns out that in most typical applications, clumpiness can be adequately smoothed simply by running a long chain without thinning, and the long chain produces reliable estimates of the posterior distribution (e.g., Jackman, 2009, p. 263; Link & Eaton, 2012). Therefore the program defaults to no thinning, although the user can thin if desired.

Assessing Null Values

Psychologists and researchers in various other disciplines have been trained to frame research questions in terms of whether or not a null value can be rejected. For example, when investigating two groups, the goal is framed as trying to reject the null hypothesis that the two groups have equal means. In other words, the "null value" for the difference of means is zero, and the goal is to reject that value as implausible.

One problem with framing research this way, with the goal of rejecting a difference of zero, is that theories can be expressed very weakly yet still be confirmed (Meehl, 1967, 1997). For example, a theorist could claim that a drug increases intelligence and have the claim confirmed by any magnitude of increase, however small, that is statistically greater than zero. Strong theories, by contrast, predict particular magnitudes of difference or predict specific forms of relation between variables (e.g., Newtonian mechanics). Scientists who pursue strong theories therefore need to estimate parameter values, not merely reject null values. Bayesian estimation is an excellent tool for pursuing strong theories.

Bayesian estimation can also be used to assess the credibility of a null value. One simply examines the posterior distribution of the credible parameter values and sees where the null value falls. If the null value is far from the most credible values, one rejects it. Examples are provided later.

Bayesian estimation also can accept the null value, not only reject it. The researcher specifies a *region of practical equivalence* (ROPE) around the null value, which encloses those values of the parameter that are deemed to be negligibly different from the null value for practical purposes. The size of the ROPE will depend on the specifics of the application domain. As a generic example, because an effect size of 0.1 is conventionally deemed to be small (Cohen, 1988), a ROPE on effect size might extend from −0.1 to 0.1. When nearly all of the credible values fall within the ROPE, the null value is said to be accepted for practical purposes. Examples are provided later in the article. The use of a ROPE is described further by Kruschke (2011a, 2011b) and in additional settings by Carlin and Louis (2009); Freedman, Lowe, and Macaskill (1984); Hobbs and Carlin (2007); and Spiegelhalter, Freedman, and Parmar (1994). Independently of its use as a decision tool for Bayesian analysis, use of a ROPE has also been suggested as a way to increase the predictive precision of theories (J. R. Edwards & Berry, 2010).

There is a different Bayesian approach to the assessment of null values, which involves comparing a model that expresses the null hypothesis against a model that expresses all possible parameter values. The method emphasizes a statistic called the Bayes factor, which is the overall likelihood of the data for one model relative to the overall likelihood of the data for the other model. In the Bayes-factor approach, parameter estimates are not emphasized. Moreover, the value of the Bayes factor itself can be very sensitive to the choice of prior distribution in the alternative model. Although the Bayes-factor approach can be appropriate for some applications, the parameter-estimation approach usually yields more directly informative results. Interested readers can find more details in Appendix D.

Examples of Robust Bayesian Estimation

I now discuss three examples of robust Bayesian estimation. The first considers two groups of moderate sample sizes, in which there are different means, different standard deviations, and outliers. The second considers two groups of small sample sizes in which the Bayesian analysis concludes that the means are not credibly different. The third considers two groups of large sample sizes in which the Bayesian analysis concludes that the group means are equal for practical purposes. In all three cases, the information provided by the Bayesian analysis is far richer than the information provided by an NHST *t* test, and in all three cases the conclusions differ from those derived from the NHST *t* test. Results from the corresponding NHST *t* tests are discussed later in the article.

Different means and standard deviations with outliers: Figure 3. Consider data from two groups of people who take an IQ test. Group 1 ($N_1 = 47$) consumes a "smart drug," and Group 2 ($N_2 = 42$) is a control group that consumes a placebo. Histograms of the data appear in the upper right panels of Figure 3. (The data for Figure 3 were generated randomly from *t* distributions. The exact data are provided in the example of running the free software at http://www.indiana.edu/~kruschke/BEST/.) The sample mean of Group 1 is 101.91 and the sample mean of Group 2 is 100.36, but there is a lot of variability within groups, and the variances of the two groups also appear to differ. There also appear to be some outliers. Are these groups credibly different?

Robust Bayesian estimation yields rich information about the differences between groups. As explained above, the MCMC method generates a very large number of parameter combinations that are credible, given the data. These combinations of parameter values are representative of the posterior distribution. Figure 3 shows histograms of 100,000 credible parameter-value combinations. It is important to understand that these are histograms of parameter values; they are not histograms of simulated data. The only histograms of data appear in the top right panels of Figure 3 that are labeled with *y* on their abscissas, and these data are fixed at their empirically observed values. All the other histograms display 100,000 parameter values from the posterior distribution, given the single set of actual data. In particular, the five histograms in the left column of Figure 3 show the posteriors corresponding to the five prior histograms in Figure 2. For example, the wide uniform prior on σ_1, shown in the left of Figure 2, becomes the

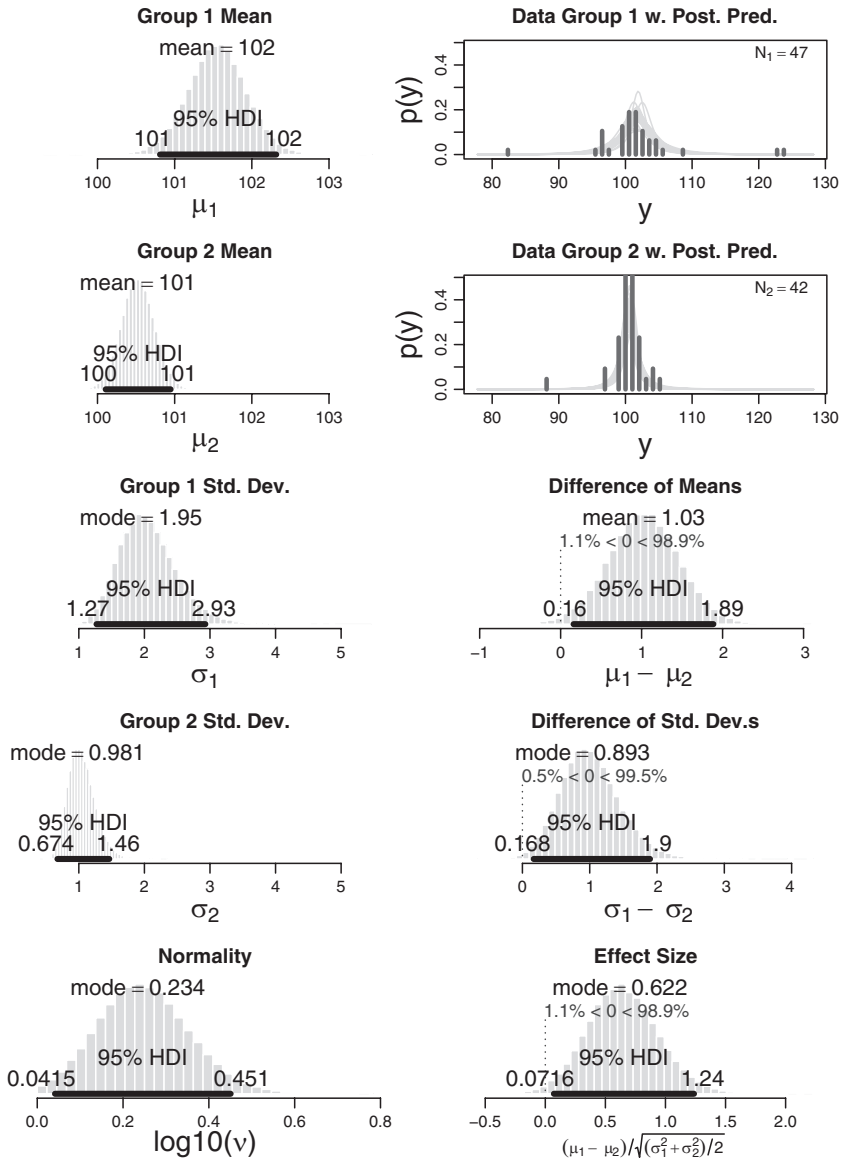

Figure 3. Top right shows histograms of the data in the two groups, with representative examples of posterior predictive (Post. Pred.) distributions superimposed. Left column shows marginals of the five-dimensional posterior distribution, corresponding to the five prior histograms in Figure 2. Lower right shows posterior distribution of differences and effect size. HDI = highest density interval; w. = with; Std. Dev. = standard deviation.

smoothly peaked and relatively narrow posterior distribution in the middle left of Figure 3.

Each histogram is annotated with its central tendency, with the mean used for distributions that are roughly symmetric and the mode used for distributions that are noticeably skewed. Each histogram is also marked with its 95% *highest density interval* (HDI), which is a useful summary of where the bulk of the most credible values falls. By definition, every value inside the HDI has higher probability density than any value outside the HDI, and the total mass of points inside the 95% HDI is 95% of the distribution. The numbers displayed in the plots of Figure 3 are rounded to three significant digits to save display space.

The upper left panel of Figure 3 shows that the mean of the credible values for μ_1 is 101.55 (displayed to three significant digits as 102), with a 95% HDI from 100.81 to 102.32, and the mean of the MCMC chain for μ_2 is 100.52, with a 95% HDI from 100.11 to 100.95. Therefore, the difference $\mu_1 - \mu_2$ is 1.03 on average, as displayed in the middle plot of the right column. One sees that the 95% HDI of the difference of means falls well above zero, and 98.9% of the credible values are greater than zero. Therefore one can conclude that the groups' means are, indeed, credibly different. It is important to understand that the Bayesian analysis yields the complete distribution of credible values, but a separate decision rule converts the posterior distribution to a discrete conclusion about a specific value.

The Bayesian analysis simultaneously shows credible values of the standard deviations for the two groups, with histograms plotted in the left column of Figure 3. The difference of the standard

deviations is shown in the right column, where it can be seen that a difference of zero is not among the 95% most credible differences, and 99.5% of the credible differences are greater than zero. Thus, not only is the mean of the first group credibly larger than the mean of the second group, the standard deviation of the first group is also credibly larger than the standard deviation of the second group. In the context of the smart drug consumed by Group 1, this result means that the drug increased scores on average, but the drug also increased the variability across subjects, indicating that some people may be adversely affected by the drug and others may be quite positively affected. Such an effect on variance has real-world precedents; for example, stress can increase variance across people (Lazarus & Eriksen, 1952).

The lower right panel of Figure 3 shows the distribution of credible effect sizes, given the data. For each credible combination of means and standard deviations, the effect size is computed as $(\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$. The histogram of the 100,000 credible effect sizes has a mode of 0.622, with the shape shown in Figure 3, and a 95% HDI that excludes zero.[1]

The lower left panel of Figure 3 shows credible values of the normality parameter in the *t* distribution. The values are shown on a logarithmic scale, because the shape of the *t* distribution changes noticeably for values of ν near 1 but changes relatively little for $\nu > 30$ or so (see Figure 1). On a base-10 logarithmic scale, $\log_{10}(\nu) = 0$ means $\nu = 1$, $\log_{10}(\nu) = 1$ means $\nu = 10$, and $\log_{10}(\nu) = 2$ means $\nu = 100$. The histogram shows that $\log_{10}(\nu)$ has values close to zero, which means that the credible *t* distributions are large tailed to accommodate outliers in the data.

The upper right panels of Figure 3 show a smattering of credible *t* distributions superimposed on histograms of the data. The curves are produced by selecting several random steps in the MCMC chain and at each step plotting the *t* distribution with parameters $< \mu_1, \sigma_1, \nu >$ on the Group 1 data and plotting the *t* distribution with parameters $< \mu_2, \sigma_2, \nu >$ on the Group 2 data. By visually comparing the data histogram and the typical credible *t* distributions, one can assess whether the model is a reasonably good description of the data. This type of assessment is called a *posterior predictive check* (Gelman, Carlin, Stern, & Rubin, 2004; Gelman & Shalizi, 2012; Gelman & Shalizi, in press; Guttman, 1967; Kruschke, in press; Rubin, 1984). We see from the plots that the credible *t* distributions are a good description of the data. (In fact, the fictitious data were generated from *t* distributions, but for real data one never knows the true generating process in nature.) The posterior predictive check can be useful for identifying cases in which data are strongly multimodal instead of unimodal or strongly skewed instead of symmetric, or with two groups that have very different kurtosis instead of the same kurtosis. In these cases one may seek a more appropriate model. Fortunately, it is easy to modify the programs so they use different models; see Appendix B.

Small sample sizes: Figure 4. Consider a case of small-sample data, with $N_1 = 8$ and $N_2 = 8$, as shown in Figure 4. Although the sample means of the two groups are different, the posterior distribution reveals great uncertainty in the estimate of the difference of means, such that a difference of zero falls within the 95% HDI (middle panel of right column). As is shown later, the traditional NHST *t* test comes to a different conclusion about the difference of means (with $p < .05$). The posterior distribution on the effect size also shows that an effect size of zero falls within the 95% HDI (lowest panel of right column). The posterior distribution on the normality parameter has a mode of $\log_{10}(\nu) = 1.45$, which corresponds to $\nu = 28$ and which can be seen in Figure 1 to be nearly normal. Compared with the prior on ν (see Figure A1 in Appendix A), the posterior on ν has ruled out extremely heavy tails, but otherwise remains very uncertain.

Accepting the null with large sample sizes: Figure 5. As the sample size gets larger, the precision of the parameter estimates also increases, because sampling noise tends to cancel out. If one defines a ROPE around the null value, the precision of the estimate might be fine enough that the 95% HDI falls entirely within the ROPE. If this happens, it means that the 95% most credible values are practically equivalent to the null value. This condition can be a criterion for accepting the null value. Notice that if the ROPE is relatively wide and the 95% HDI is very narrow, the 95% HDI could fall entirely within the ROPE and yet also exclude zero. This is not a contradiction. It simply means that the credible values of the parameter are non-zero, but those non-zero values are so small that they have no practical importance.

Figure 5 shows a case of accepting the null value. The difference between means is nearly zero, but most important, the 95% HDI of the difference falls entirely within a ROPE that extends from -0.1 to 0.1. The same is true of the difference in standard deviations, where, in fact, 100% of the posterior (i.e., all of the 100,000 representative values) falls inside the ROPE. It is important to understand that Bayesian estimation provides a complete distribution over possible differences, but the decision rule is auxiliary and concludes that for practical purposes one accepts the null value.

Bayesian estimation allows one to make this conclusion by virtue of the fact that it provides an explicit posterior distribution on the differences, given the data. Without the explicit posterior distribution, one could not say whether the estimate falls within the ROPE. The decision procedure based on Bayesian estimation allows one to accept the null value only when there is high enough precision in the estimate, which typically can happen only with relatively large sample sizes.

In contrast, the NHST *t* test has no way of accepting the null hypothesis. Even if one were to define a ROPE, the confidence interval from the NHST *t* test does not provide the information one needs. The NHST *t* test and confidence interval are discussed at length in a subsequent section.

Power Analysis for Bayesian Estimation

Researchers can have various goals when analyzing their data. One important goal is to obtain a precise estimate of the descrip-

[1] The effect size is defined here as $(\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$, because I take the perspective that the effect size is merely a re-description of the posterior distribution. In principle, many different data sets could have generated the posterior parameter distribution, and therefore the data should not be used in re-describing the posterior. Nevertheless, some users may prefer to compute an effect size in which the estimates are weighted by the sample sizes in the groups: $\delta = (\mu_1 - \mu_2)/\sqrt{[\sigma_1^2(N_1 - 1) + \sigma_2^2(N_2 - 1)]/(N_1 + N_2 - 2)}$ (Hedges, 1981; Wetzels et al., 2009). This form does not change the sign of the effect size, merely its magnitude, so the proportion of the posterior distribution of the effect size that is greater (or less) than zero remains unaffected.

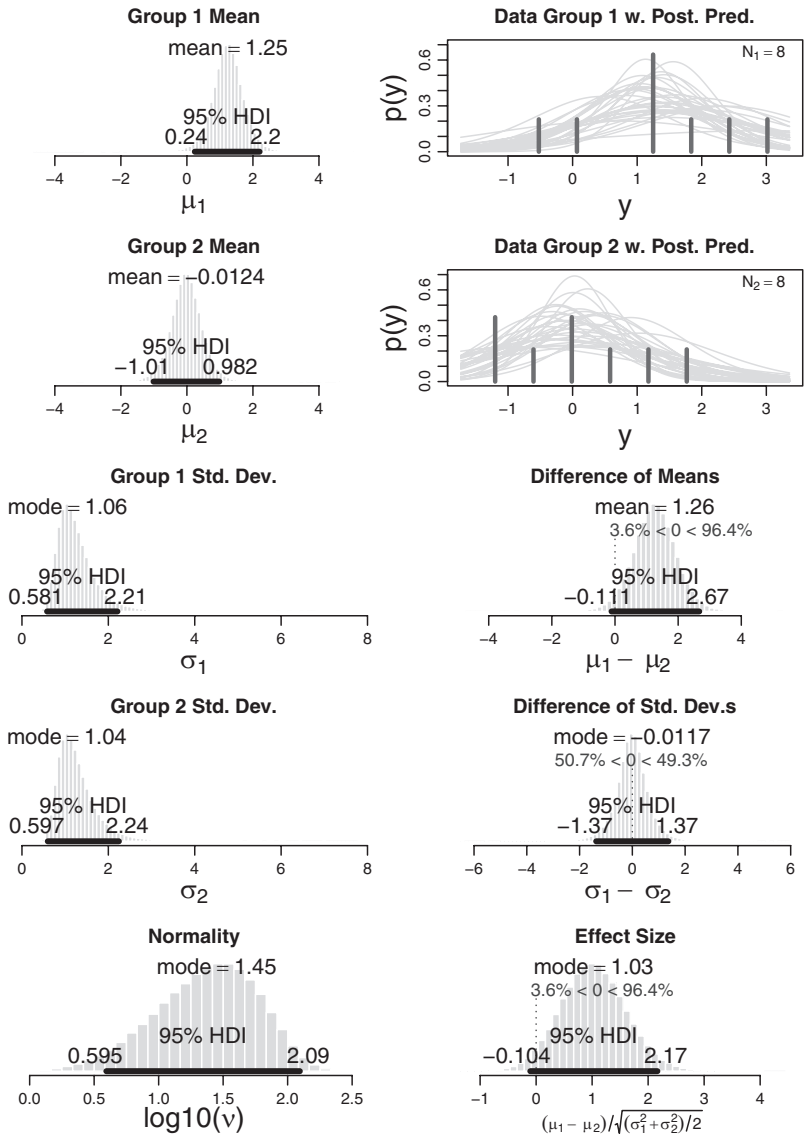

Figure 4. Top right shows histograms of data in the two groups, with representative examples of posterior predictive distributions (Post. Pred.) superimposed. Left column shows marginals of the five-dimensional posterior distribution. Lower right shows posterior distribution of differences and effect size. HDI = highest density interval; w. = with; Std. Dev. = standard deviation.

tive parameters. Success in achieving this goal can be expressed as the width of the 95% HDI being less than some criterial maximum. Other goals regard specific parameter values of interest, such as null values. For example, the analyst can assay whether the 95% HDI falls entirely outside or inside the ROPE and thereby declare the null value to be rejected or accepted. The Bayesian posterior distribution provides complete information to address these goals.

With these various goals for analysis in mind, the analyst may wonder what is the probability of achieving them, if the sampled data were generated by hypothetical parameter values. A traditional case of this issue is NHST power analysis. In NHST, the power of an experiment is the probability of rejecting the null hypothesis if the data were generated from a particular specific alternative effect size. The probability of rejecting the null, for data sampled from a non-zero effect, is less than 100% because of random variation in sampled values, but it is at least 5% because that is the conventionally tolerated false alarm rate from the null hypothesis.

Power can be assessed prospectively or retrospectively. In *retrospective* power analysis, the effect size is estimated from an observed set of data, and then the power is computed for the sample size that was actually used. In *prospective* power analysis, the effect size is hypothesized from analogous previous research or intuition, and the power is computed for a candidate sample size. Prospective power analysis is typically used for sample size determination, because the analyst can plan a sample size that yields the desired power (for the assumed hypothesis). In the context of NHST, many authors have pointed out that prospective power

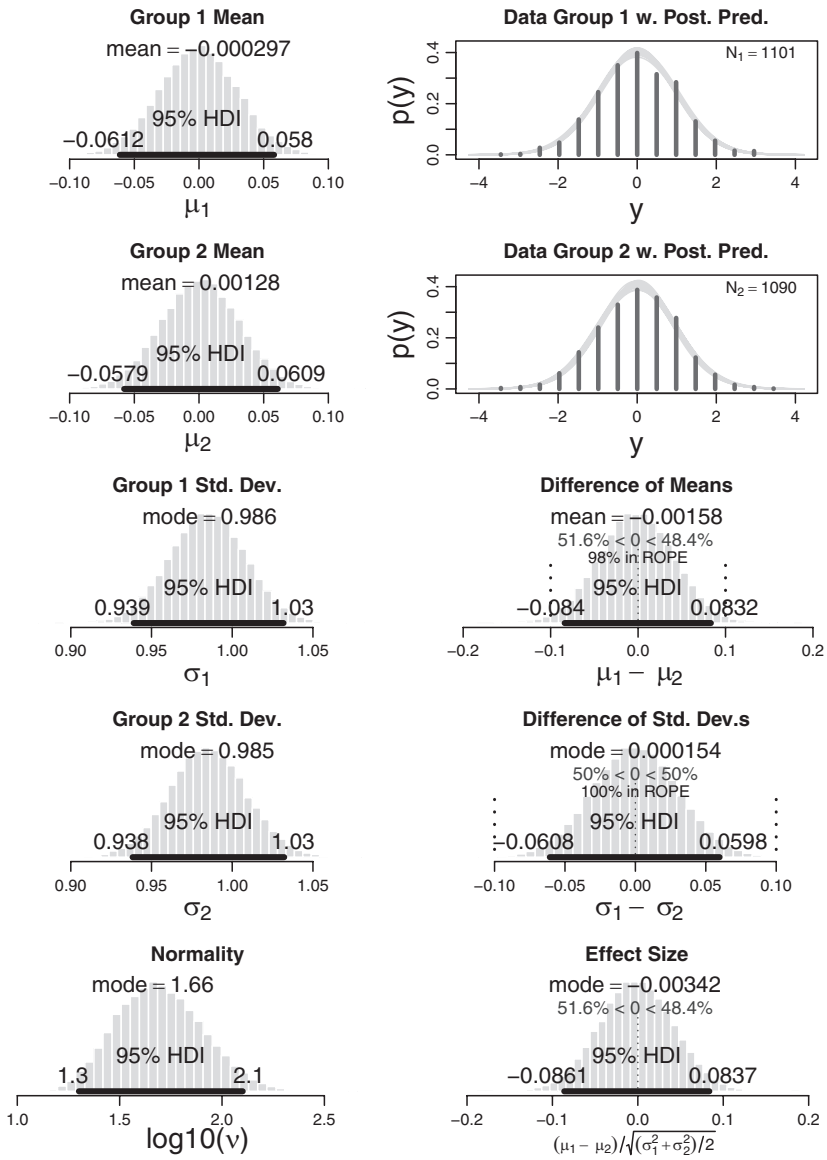

Figure 5. Top right shows histograms of data in the two groups, with representative examples of posterior predictive distributions (Post. Pred.) superimposed. Left column shows marginals of the five-dimensional posterior distribution. Lower right shows posterior distribution of differences and effect size. HDI = highest density interval; w. = with; Std. Dev. = standard deviation; ROPE = region of practical equivalence.

analysis provides useful information, but retrospective power analysis (that uses only the data from a single experiment) provides no additional information that is not already implicit in the *p* value (e.g., Gerard, Smith, & Weerakkody, 1998; Hoenig & Heisey, 2001; Nakagawa & Foster, 2004; O'Keefe, 2007; Steidl, Hayes, & Schauber, 1997; Sun, Pan, & Wang, 2011; Thomas, 1997). Retrospective power analysis can, however, at least make explicit the probability of achieving various goals in the given experiment, even if that information is not useful for additional inference from the given data.

In either retrospective or prospective power analyses, NHST uses a point value for the hypothetical effect size. In Bayesian power analysis, one uses an entire distribution of parameters instead of a single point value for the effect size. Thus, every value of effect size is considered but only to the extent that it is considered to be credible. NHST power analysis can consider various point values, such as the end points of a confidence interval, but the different point values are not weighted by credibility and therefore can yield a huge range of powers. As a consequence, NHST power analysis often yields extremely uncertain results (e.g., Gerard et al., 1998; Miller, 2009; Thomas, 1997), but Bayesian power analysis yields precise estimates of power. A later section describes NHST power analysis in more detail, and the remainder of this section describes Bayesian power analysis.

For Bayesian retrospective power analysis, the distribution of credible parameter values is the posterior distribution from an

observed set of data. At every step in the MCMC chain of the posterior, the analyst uses that step's parameter values to simulate new data, then does a Bayesian analysis of the simulated data, and then checks whether the desired goals are achieved. The process is illustrated in Figure 6, and the caption provides more details. From the many simulations, the proportion of times that each goal is achieved is used to estimate the probability of achieving each goal. The mechanics of the process are explained in detail in Chapter 13 of Kruschke (2011b) and are illustrated with examples in Kruschke (2010a, 2010b). The issue has been explored in technical detail in various domains (e.g., Adcock, 1997; De Santis, 2004, 2007; Joseph, Wolfson, & du Berger, 1995a, 1995b; Wang & Gelfand, 2002; Weiss, 1997).

For prospective power analysis, the same process is executed but starting with hypothetical data instead of actual data. The hypothetical data are designed to represent ideal results for a large experiment that perfectly reflects the hypothesized effect. From the large set of idealized data, a Bayesian analysis reveals the corresponding parameter distribution that is credible. This parameter distribution is then used as the expression of the hypothesis in the left side of Figure 6. A major advantage of this approach is that researchers can usually intuit hypothetical data much more easily than hypothetical parameter values. The researcher merely needs to generate hypothetical data from an idealized experiment, instead of trying to specify abstract distributions of parameters and their trade-offs in high-dimensional space. This approach is also quite general and especially useful for more complex situations involving models with many parameters.

Example of Bayesian prospective power analysis. To facilitate the generation of idealized data for prospective power analysis, a program accompanying this article generates simulated data from two groups. Details are provided in Appendix C. The user specifies the means and standard deviations of the two normally distributed groups and the sample size for each group. The user also specifies the percentage of the simulated data that should come from outlier distributions, which have the same means as the two groups but a larger standard deviation, which is also specified by the user. As an example, suppose the researcher is contemplating the effect of a smart drug on IQ scores. He or she assumes that the control group has a mean of 100 and standard deviation of 15 and the treatment group will have a mean of 108 and standard deviation of 17, with scores normally distributed in each group. Moreover, the researcher hypothesizes that 10% of the data will consist of outliers, simulated as coming from the same group means but with twice as large a standard deviation. An idealized experiment would perfectly realize these hypothetical values in a large sample size, such as 1,000 per group. The sample size expresses the confidence in the hypothesis: The larger the sample size, the higher the confidence. Figure 7 shows an example of such idealized data. The researcher can easily inspect this figure to check that it accurately represents the intended hypothesis.

The idealized data are then submitted to a Bayesian analysis so that the corresponding parameter distribution can be derived. The resulting posterior distribution, shown in Figures 8 and 9, reveals the parameter uncertainty implicit in the idealized data set, for all the parameters including the normality parameter, which might be particularly difficult to specify by prior intuition alone. The posterior distribution also captures *joint* dependencies of credible parameter values, as revealed in Figure 9, where it can be seen that the standard deviation (σ_1 and σ_2) and normality (ν) parameters are correlated with each other. This correlation occurs because higher values of normality, which posit small-tailed data distributions, require larger standard deviations to accommodate the outliers in the data. Although it is fairly easy for a researcher to intuit, generate, and check idealized data as in Figure 7, it is probably considerably more difficult for a researcher to intuit, generate, and check idealized joint parameter values to express the hypothesis as in Figure 9.

Another benefit of this approach is that the researcher's certainty in a hypothesis is expressed in terms of concrete sample size, not in terms of spread in an abstract parameter space. The sample size for the idealized data directly indicates the amount of data in fictitious previous research that provides support for the hypothesis. The example in Figure 7 used 1,000 values in each group to

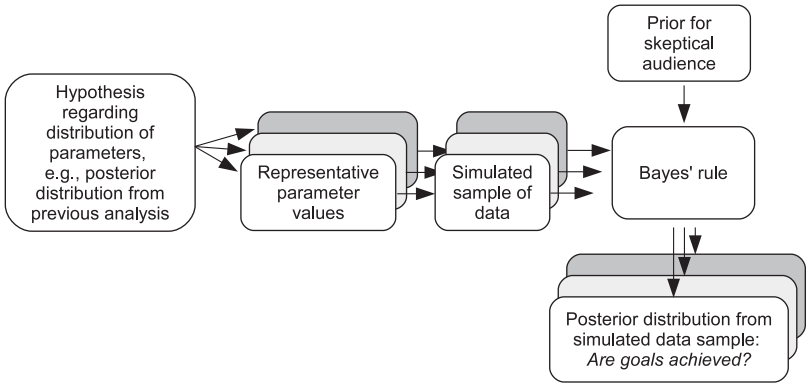

Figure 6. Flowchart for estimating Bayesian power. At the left, a hypothetical distribution of parameter values is used to repeatedly generate representative credible parameter values. For each set of parameter values, a simulated sample of data is generated from the model. Then Bayes' rule is applied to derive the posterior distribution from the simulated data and assay whether the various goals have been achieved for that sample. Across many replications of simulated experiments, the probability of achieving each goal is estimated to arbitrarily high accuracy.

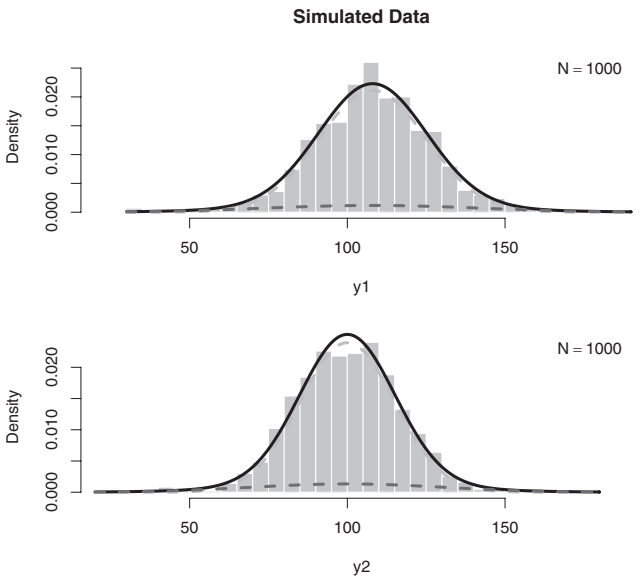

Figure 7. Idealized data used for prospective power analysis. The histograms represent the simulated data values and the curves indicate the generating distributions, with the taller dashed curve representing the main generator, the shorter dashed curve representing the outlier distribution, and the solid curve representing their sum.

represent a fairly large previous study. If the analyst wanted to express even stronger certainty in the hypothesis, a larger sample size could be used for the idealized data. The larger sample size will yield narrower parameter distributions in the Bayesian analysis.

The creation of a hypothetical parameter distribution for prospective power analysis is analogous to eliciting a prior distribution from expert informants. The method advocated here, in which the informant generates hypothetical data from which a parameter distribution is derived, is analogous to the *equivalent prior sample* (EPS) method proposed by Winkler (1967) for eliciting the prior in estimating the bias of a coin. The method advocated here is consistent with the advice of Garthwaite, Kadane, and O'Hagan (2005):

> As a guiding principle, experts should be asked questions about quantities that are meaningful to them. This suggests that questions should generally concern observable quantities rather than unobservable parameters, although questions about proportions and means also might be considered suitable, because psychological research suggests that people can relate to these quantities. . . . Graphical feedback is an important component . . . and it seems to provide a potentially powerful means of improving the quality of assessed distributions. (pp. 689, 692)

A reader might wonder why we go through the effort of creating a parameter distribution to generate simulated data if we already have a way of generating simulated data for the idealized experiment. The answer is that the idealized data are generated from a punctate parameter value without any uncertainty expressed in that parameter value. It is the idealized sample size that expresses the intuitive certainty in the hypothesized parameter value and, subsequently, the Bayesian posterior from the idealized sample that expresses the parameter uncertainty. Thus, the process goes from easily intuited certainty expressed as idealized sample size to less easily intuited uncertainty expressed as a multidimensional parameter distribution. The resulting parameter distribution can be visually examined for consistency with the intended hypothesis.

With the hypothetical parameter distribution now established, I proceed with the power analysis itself. To estimate the probability of achieving the goals, one steps through the credible combinations of parameter values in the MCMC chain. At each step, the chain specifies a combination of parameter values, $<\mu_1, \sigma_1, \mu_2, \sigma_2, \nu>$, which one uses to generate a simulated set of data from the model. For experiments in which data are sampled until a threshold sample size, those sample sizes should be used for the simulated data. For experiments in which data are sampled until a threshold duration elapses, the simulated data should produce random sample sizes to mimic that process (e.g., with a Poisson distribution; Sadiku & Tofighi, 1999). In the analyses presented here (and in the programs available at http://www.indiana.edu/~kruschke/BEST/), I assume fixed sample sizes, merely for simplicity. For each set of simulated data, a Bayesian analysis is conducted to produce a posterior distribution on the parameters given the simulated data. Each of the goals can then be assessed in the posterior distribution. This process repeats for many steps in the MCMC chain from the original analysis. The software commands for executing the power analysis are explained in Appendix C.

The underlying probability of achieving each goal is reflected by the proportion of times that the goal is achieved in the simulated replications. The estimate of the underlying probability is itself a Bayesian estimation from the proportion of successes in the simulated replications. One assumes a noncommittal uniform prior on the power, and therefore the posterior distribution on power is a beta distribution (e.g., Chapter 5 of Kruschke, 2011b). As the number of replications increases, the beta distribution gets narrower and narrower, which is to say that the estimate of power gets more and more certain. The 95% HDI of the posterior beta distribution is used to summarize the uncertainty in the estimated power. Notice that the accuracy of the power estimate is limited in practice only by the number of simulated replications. In principle, the power estimate is precise because it integrates over the entire hypothetical parameter distribution.

Suppose I think that a realistic sample size for my study will have $N_1 = N_2 = 50$. Suppose I have several goals for the analysis, selected here more for illustrative purposes than for realism. With regard to the magnitude of $\mu_1 - \mu_2$, suppose I would like to show that its 95% HDI excludes a ROPE of $(-1.5, 1.5)$, which is to say that there is difference between means that is credibly different than 1.5 IQ points. I may also desire a minimum precision on the estimate of $\mu_1 - \mu_2$, such that its 95% HDI has width less than 15 IQ points (i.e., one standard deviation of the normed background population). I may have analogous goals for the difference of standard deviations and for the effect size. The results of running 1,000 simulated experiments are shown in Table 1, where it can be seen that the estimated probability that the 95% HDI on the difference of means excludes a ROPE of $(-1.5, 1.5)$ is only 40.1%. There are fairly tight bounds on this estimate because 1,000 simulated experiments were run. Similarly, the estimated probability that the 95% HDI on the effect size excludes zero is only 54.4%. The prospective power analysis reveals that even if the hypothetical effect of the smart drug were perfectly supported by

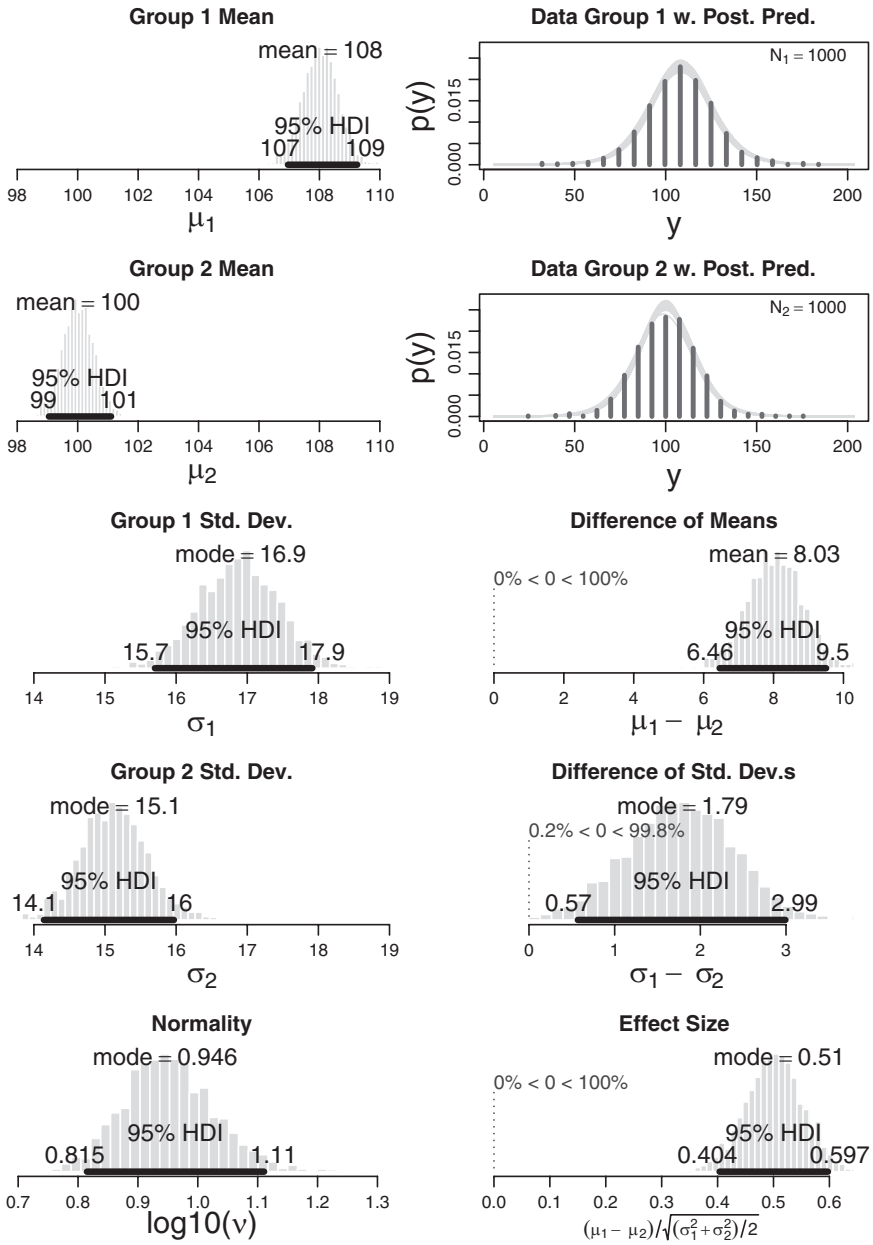

Figure 8. Posterior distribution from idealized data in Figure 7, used for prospective power analysis. (The histograms are a bit choppy because only a short MCMC chain was generated for use in prospective power analysis.) Figure 9 shows pairwise plots of the five parameters. MCMC = Markov chain Monte Carlo; HDI = highest density interval; w. = with; Post. Pred. = posterior predictive; Std. Dev. = standard deviation.

previous research that included 1,000 people in each group, a novel experiment involving only 50 people in each group would be underpowered (assuming that a power of at least 80% would be desired before running an experiment). The researcher could increase power by increasing the novel experiment's sample size or by aspiring to an easier goal (e.g., with a smaller ROPE).

Example of Bayesian retrospective power analysis. Retrospective power analysis proceeds like prospective power analysis, but instead of using idealized data, one uses actually observed data. Recall the example of Figure 3, which involved data from two groups of subjects who took IQ exams. (Although the data were fictitious, suppose that they were real for the present illustration.) The first group ($N_1 = 47$) took a smart drug, and the second group ($N_2 = 42$) was a control. The Bayesian posterior revealed credibly non-zero differences in the means and the standard deviations. The posterior also indicated that the effect size was credibly non-zero. Suppose that the researcher would like to do a retrospective power analysis for these data, and the researcher has several goals, with one being that the 95% HDI on the difference of means is greater than zero. Six goals altogether are

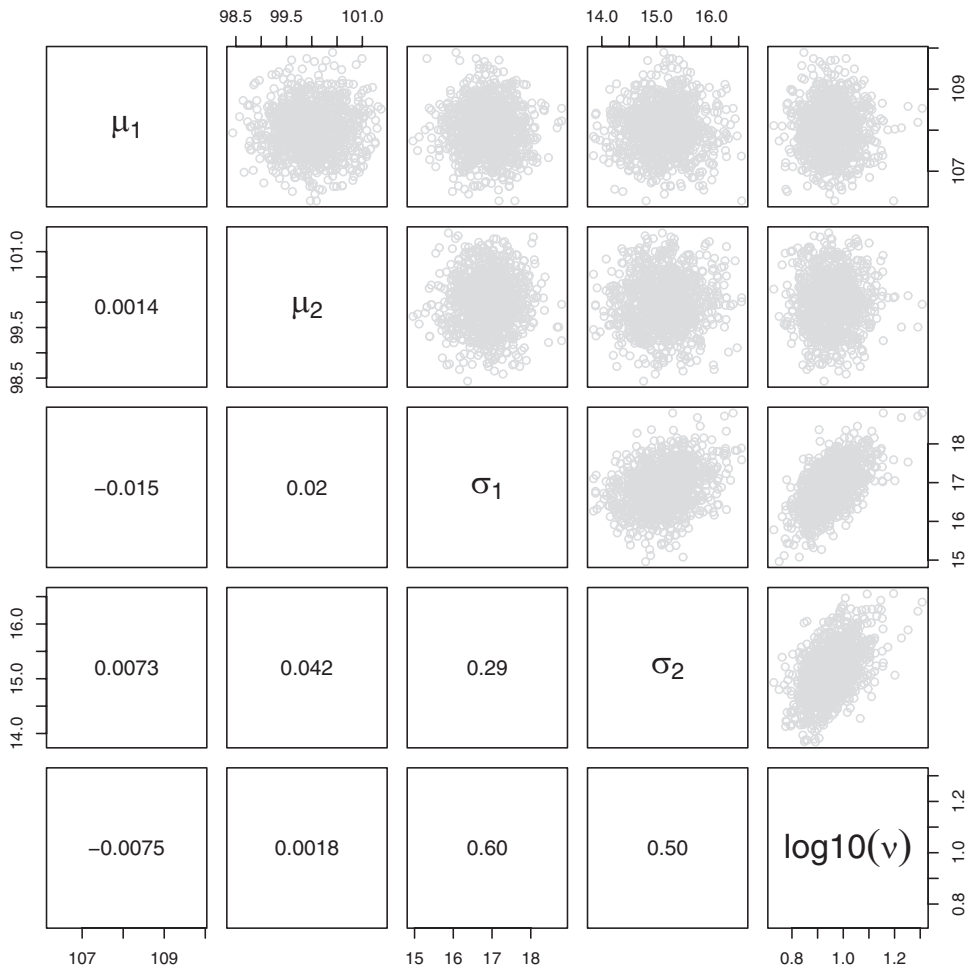

Figure 9. Posterior distribution from idealized data in Figure 7, used for prospective power analysis. The pairwise plots of credible parameter values, shown in the upper right panels, reveal correlations in the standard deviation (σ_1 and σ_2) and normality (ν) parameters. The numerical correlations are shown in the lower left panels. It would be difficult to generate this distribution of parameter values directly from intuition, but it is relatively easy to intuit the idealized data of Figure 7.

indicated in Table 2. The power analysis generated simulated data from 1,000 steps in the MCMC chain (selected evenly from across the entire chain). The analysis revealed that the power for the first goal, regarding the difference of means being greater than zero, was 47.4%, with a 95% HDI on the estimate extending from 44.3% to 50.4%, based on 1,000 simulated replications. If more simulated replications were used, the bounds on the estimated power would be tighter. The power for the other goals is indicated in Table 2. It

Table 1
Bayesian Prospective Power Analysis for Parameter Distribution in Figure 8, Using $N_1 = N_2 = 50$

	Based on 1,000 simulated replications		
Goal	Bayesian power	Lower bound	Upper bound
95% HDI on the difference of means excludes ROPE of (−1.5, 1.5).	40.1%	37.1%	43.1%
95% HDI on the difference of means has width less than 15.0.	72.6%	69.8%	75.3%
95% HDI on the difference of standard deviations is greater than zero.	10.5%	8.6%	12.4%
95% HDI on the difference of standard deviations has width less than 10.0.	15.8%	13.5%	18.0%
95% HDI on the effect size is greater than zero.	54.4%	51.4%	57.5%
95% HDI on the effect size has width less than 1.0.	97.8%	96.9%	98.7%

Note. "Lower bound" and "Upper bound" refer to limits of the 95% HDI on the beta posterior for estimated power, which get closer together as the number of simulated replications increases. HDI = highest density interval; ROPE = region of practical equivalence.

Table 2
Bayesian Retrospective Power Analysis for the Posterior Distribution in Figure 3

Goal	Based on 1,000 simulated replications		
	Bayesian power	Lower bound	Upper bound
95% HDI on the difference of means excludes ROPE of (−0.1, 0.1).	47.4%	44.3%	50.4%
95% HDI on the difference of means has width less than 2.0.	59.5%	56.5%	62.6%
95% HDI on the difference of standard deviations excludes ROPE of (−0.1, 0.1).	62.9%	59.9%	65.9%
95% HDI on the difference of standard deviations has width less than 2.0.	72.9%	70.1%	75.6%
95% HDI on the effect size excludes ROPE of (−0.1, 0.1).	38.2%	35.2%	41.2%
95% HDI on the effect size has width less than 0.2.	0.1%	0.0%	0.3%

Note. "Lower bound" and "Upper bound" refer to limits of the 95% HDI on the beta posterior for estimated power, which get closer together as the number of simulated replications increases. HDI = highest density interval; ROPE = region of practical equivalence.

is worth reiterating that these precise power estimates incorporate the full uncertainty of the parameter estimates and are not based on a single hypothetical parameter value as in NHST power analysis.

Reporting the Results of a Bayesian Analysis

When the results of a robust Bayesian estimation of groups are reported, the posterior distribution is summarized in text. Although there are established conventions for reporting NHST analyses (e.g., American Psychological Association, 2009), there are not yet conventions for reporting Bayesian analyses. General guidelines for reporting a Bayesian analysis are offered by Kruschke (2011b, Chapter 23). Those guidelines were intended to apply to any Bayesian analysis and to when the analyst could not assume much previous knowledge of Bayesian methods in the audience. Hence, the first recommended points were to motivate the use of Bayesian analysis, describe the model and its parameters, justify the prior distribution, and mention the details of the MCMC mechanics. In the present application, all these points are addressed elsewhere in this article. Anyone who uses the accompanying program unaltered can briefly review the points or simply refer to this article. (If the user changes the prior or likelihood function in the program, those changes must be explained, along with assessment of MCMC convergence.) The essential mission of the analyst's report, therefore, is to summarize and interpret the posterior distribution. A summary of each parameter can consist of descriptive values including the central tendency, the 95% HDI, and, if relevant, the percentage of the posterior distribution above or below a landmark value, such as zero, or within a ROPE. These values are displayed graphically in Figure 3 and are output in greater numerical precision on the computer console when the program is run. When making discrete decisions about a null value, the analyst can explicitly define and justify a ROPE, as appropriate, or leave a ROPE unspecified so that readers can use their own.

Summary of Bayesian Estimation

I began with a descriptive model of data from two groups, wherein the parameters were meaningful measures of central tendency, variance, and normality. Bayesian inference reallocates credibility to parameter values that are consistent with the observed data. The posterior distribution across the parameter values gives complete information about which combinations of parameter values are credible. In particular, from the posterior distribution one can assess the credibility of specific values of interest, such as zero difference between means, or zero difference between standard deviations. One can also decide whether credible values of the difference of means are practically equivalent to zero, so that the null value is accepted for practical purposes. (How the Bayesian parameter-estimation approach to assessing null values, described here, differs from the Bayesian model-comparison approach is explained in Appendix D.) The Bayesian posterior distribution also provides complete information about the precision of estimation, which can be summarized by the 95% HDI.

The Bayesian posterior distribution can also be used as a complete hypothesis for assessing power, that is, the probabilities of achieving research goals such as rejecting a null value, accepting a null value, or reaching a desired precision of estimation. The power estimation incorporates all the information in the posterior distribution by integrating across the credible parameter values, using each parameter–value combination to the extent it is credible. Appendix C provides some details for using the software to do power analysis.

The software for computing the Bayesian parameter estimates and power is free and easy to download from http://www.indiana.edu/~kruschke/BEST/. Instructions for its use are provided in the sample programs, and instructions for modifying the programs are provided in Appendix B.

In the next section, I show that the information provided by the NHST *t* test is very impoverished relative to the results of Bayesian estimation.

The NHST *t* Test

In this section I review the traditional *t* test from null hypothesis significance testing (NHST). First I look at the *t* test applied to the three examples presented earlier, to highlight differences between the information and conclusions provided by NHST and Bayesian estimation. Then I turn to general foundational issues that undermine the usefulness of NHST in any application.

Examples of the NHST *t* Test

Recall the data of Figure 3, in which IQ scores of a smart drug group were compared against IQ scores of a control group. The robust Bayesian estimation revealed credible non-zero differences between means and standard deviations of the groups, along with heavy tails (non-normality). A complete posterior distribution on

the effect size was also generated. Bayesian retrospective power analysis (see Table 2) indicated that the 95% HDI on the effect size would be greater than zero on 49% of replications, with tight bounds on the estimate established by 1,000 simulated replications.

I now consider the results of an NHST *t* test applied to the data of Figure 3, which yields $t(87) = 1.62$, $p = .110$, with a 95% confidence interval on the difference of means from −0.361 to 3.477. These results, and the results of all *t* tests reported in this article, use the Welch (1947) modification to degrees of freedom for producing the *p* value and confidence interval, to accommodate unequal variances. According to conventional decision criteria (i.e., $p < .05$), the result implies that the two group means are not significantly different, contrary to the conclusion reached by the Bayesian analysis.

The NHST *t* test tells nothing about the difference between the standard deviations of the two groups, which in the samples are 6.02 and 2.52, respectively. To test the difference of standard deviations, I have to conduct an additional NHST *F* test of the ratio of variances, which yields $F(46, 41) = 5.72$, $p < .001$. However, by conducting a second test on the same data, according to NHST criteria I need to control for the additional possibility of false alarm by using a stricter criterion to reject the null hypothesis in either test. For example, I could apply a Bonferroni correction to the two tests, so that I would require $p < .025$ to declare significance instead of $p < .050$. Notice that the differences between groups remain fixed, but the criterion for declaring significance changes depending on whether or not I intend to test the difference between sample variances. Corrections for multiple comparisons are discussed more below, in the context of showing how *p* values change under various other changes of intention.

Unfortunately, the results from both NHST tests are suspicious, because both tests assume normally distributed data, but the actual data apparently have outliers. In this context of NHST, the appearance of outliers was judged qualitatively. I could run an additional test of normality, but this would incur an additional penalty in setting a stricter criterion for significance in the other tests. The problem with outliers is that conventional *p* values are computed on the basis of sampling distributions drawn from null hypotheses that have normal distributions. Sampling distributions generated from non-normal distributions yield different *p* values (for interactive examples in Excel, see Kruschke, 2005). Although the *t* test for difference of means tends to be fairly robust against violations of normality (e.g., Lumley, Diehr, Emerson, & Chen, 2002, and references cited therein), the *F* test for difference of variances can be strongly affected by violations of normality (e.g., Box, 1953; Pearson, 1931).

A standard way to address violations of distributional assumptions in NHST is to use resampling methods. In resampling, instead of assuming a particular distributional form in the population, one substitutes the data themselves for the population. Under the null hypothesis, the data from the two groups represent the same underlying population, and therefore the data are pooled. A sampling distribution is generated by repeatedly drawing samples of sizes N_1 and N_2 randomly, with replacement, from the pooled population and computing the difference of sample means or difference of sample standard deviations in every replication. Across many replications (i.e., 100,000 for the results reported here), a *p* value is computed as the proportion of randomly generated samples in which the sample difference exceeds the difference in the actual data, multiplied by two for a two-tailed test. With the data from Figure 3, a resampling test of the difference of means yields $p = .116$, which is very close to the result of the conventional *t* test, but a resampling test of the difference of standard deviations yields $p = .072$, which is much larger than the result of the conventional *F* test. The latter also implies that the two standard deviations are not significantly different, contrary to the conventional *F* test (and contrary to the Bayesian estimation). Corrections for multiple comparisons must still be applied when interpreting the *p* values. To recapitulate, even with resampling, which avoids parametric distributional assumptions, both the difference of means and the difference of standard deviations are deemed to be nonsignificant, unlike in Bayesian estimation.

Consider now the example of Figure 4, which involved small samples ($N_1 = N_2 = 8$). An NHST *t* test of the data yields $t(14) = 2.33$, $p = .035$, with 95% confidence interval on the difference of means from 0.099 to 2.399. Notice that the conclusion of significance (i.e., $p < .05$) conflicts with the conclusion from Bayesian estimation in Figure 4, in which the 95% HDI on the difference of means included zero. The Bayesian estimate revealed the full uncertainty in simultaneously estimating five parameters from small samples, and the NHST *t* test relied on a point null hypothesis assuming normal distributions.

Finally, consider the example of Figure 5, which involved large samples ($N_1 = 1{,}101$, $N_2 = 1{,}090$). An NHST *t* test yields $t(2189) = 0.01$, $p = .99$, and a 95% confidence interval on the difference of means from −0.085 to 0.084. According to NHST, we cannot accept the null hypothesis from this result; we can say only that it is highly probable to get a difference of means from the null hypothesis that is greater than the observed difference of means. (In fact, according to NHST, it is so unlikely to get such a small difference of means that we should reject some aspect of the null hypothesis, such as the assumption of independence.) The Bayesian estimation showed that the 95% HDI was completely contained within a ROPE from −0.1 to 0.1, thereby accepting the null value for practical purposes. (NHST has problems when pursuing an analogous decision rule regarding the relation of the confidence interval and a ROPE, as is discussed below.)

In all three cases, NHST and Bayesian estimation came to different conclusions. It would be wrong to ask which conclusion is correct, because for real data we do not know the true state of the world. Indeed, for many studies, we assume in advance that there must be some difference between the conditions, however small, but we go to the effort of obtaining the data in order to assess what magnitude of difference is credible and whether we can credibly claim that the difference is not zero or equivalent to zero for practical purposes. Therefore we should use the analysis method that provides the richest information possible regarding the answer we seek. And that method is Bayesian estimation. Bayesian estimation provides an explicit distribution of credibilities across all possible parameter values in the descriptive model, given the set of actually observed data. From the posterior distribution, Bayesian methods use a decision procedure involving the HDI and ROPE to declare particular values to be credible or not.

In contrast, the method of NHST provides only a point estimate of the parameter values, namely, the parameter values that minimize the squared deviation or maximize the likelihood. The decision process in NHST is based on asking what is the probability of

the data summary statistic (such as t), if the null hypothesis were true. Answering this question provides little direct information about the probability of parameter values given the data, which is what one wants to know. NHST is based on sampling distributions generated by assuming that the null-hypothesis values of the parameters are true. Sampling distributions are also the basis of confidence intervals. As will be shown, for any fixed set of data there are many different p values and many different confidence intervals depending on the sampling intentions of the analyst. For any confidence interval there is no distributional information regarding values between its end points. The poverty of information in the confidence interval also leads to power estimates being very uncertain. In summary, not only is the question asked by NHST not the question one wants answered, but the information provided by NHST is very impoverished. Thus, whether Bayesian analysis indicates credible differences when NHST does not (as for the data of Figure 3), indicates uncertainty when NHST declares significance (as for the data of Figure 4), or indicates accepting the null for practical purposes when NHST cannot (as for the data of Figure 5), it is Bayesian analysis that directly addresses our question and provides richer information.

From the three specific examples reviewed above, I now proceed to general issues that afflict NHST.

A p Value Depends on Sampling Intentions

In the NHST t test, an observed t value is computed from the data, which I will denote t_{obs}. The value of t_{obs} is a direct algebraic function of the data values, which can be computed regardless of any assumptions about where the data came from, just as a mean or standard deviation can be computed for any set of data. However, additional assumptions must be made to compute the p value. The p value is the probability of getting a t value from the null hypothesis, as big or bigger than t_{obs}, if the intended experiment were repeated ad infinitum. The p value indicates the rarity of t_{obs} relative to the space of all possible t values that might have been observed from the intended sampling process if the null hypothesis were true. More formally, the p value is the probability that any t_{null} value generated from the null hypothesis according to the intended sampling process has magnitude greater than or equal to the magnitude of t_{obs}, which is denoted as $p(\text{any } |t_{null}| \geq |t_{obs}|)$.

Importantly, the space of all possible t_{null} values that might have been observed is defined by how the data were intended to be sampled. If the data were intended to be collected until a threshold sample size was achieved, the space of all possible t_{null} values is the set of all t values with that exact sample size. This is the conventional assumption. Many researchers, however, do not intend to collect data with a fixed sample size planned in advance. Instead, they intend to collect data for a certain duration of time, such as 2 weeks, and the actual number of respondents is a random number. In this case, the space of all possible t_{null} values is the set of all t values that could be obtained during that time, which could involve larger or smaller sample sizes. The result is a different space of possible t_{null} values than the conventional assumption of fixed sample size and, hence, a different p value and different confidence interval. The space of possible t_{null} values is also strongly affected by the intention to compare the results with other groups. This is because additional comparisons contribute more possible t_{null} values to the space of all t values that could be obtained, and consequently the p value and confidence interval change. There are many different intentions for generating the space of possible t_{null} values and, hence, many different p values and confidence intervals for a single set of data. This section illustrates this point with several different examples.

The p value for intending to sample until threshold N. The conventional assumption is that the data collector intended to collect data until achieving a specific threshold for sample size. The upper left Panel A of Figure 10 shows the probability of obtaining t values from the null hypothesis when the threshold sample sizes are $N_1 = N_2 = 8$. The x-axis indicates the value of t_{obs}. The y-axis is labeled $p(\text{any } |t_{null}| > |t_{obs}|)$, which is the p value. The plot shows that the value of t_{obs} for which $p = .05$ is $t_{crit} = 2.14$. This is the conventional t_{crit} value reported in many textbooks. For the data of Figure 4, $t_{obs} = 2.33$ exceeds the critical value, and $p < .05$.

The p value for intending to sample from multiple groups. If the data for two groups were collected along with data from other groups and tests were intended for various combinations of groups, the space of all possible t_{null} values is the set of all t_{null} values collected across all tests. Because of this increased space of possible t_{null} values, the relative rarity of t_{obs} changes, and hence the p value for t_{obs} changes. This dependence of the p value on the intention of the experimenter is well recognized by the conventional use of various corrections for different kinds of multiple comparisons (see, e.g., the excellent summary in Maxwell & Delaney, 2004). Notice that the data in the original two groups have not changed when the space of intended comparisons is enlarged, and notice also that the actual data in any of the groups are irrelevant; indeed, the data do not even have to be collected. What matters is the intention that defines the space of possible t_{null} values from the null hypothesis.

Suppose, for example, that we collect data until a fixed sample size is achieved, but for four groups instead of two, and we conduct two independent NHST t tests. The upper middle Panel B of Figure 10 shows the probability of obtaining t_{null} values from the null hypothesis when the sample sizes for all groups are fixed at $N = 8$. The plot shows that the value of t_{obs} for which $p = .05$ is $t_{crit} = 2.5$. In particular, for the data of Figure 4, $t_{obs} = 2.33$ does not exceed the critical value, and $p > .05$. Thus, although the data in Figure 4 have not changed, their p value has changed because it is computed relative to a different space of possibilities.

The p value for intending to sample until threshold duration. If the data were collected until a particular duration was reached (e.g., collecting data until the end of the week), the space of all possible t_{null} values from the null hypothesis is the set of all t_{null} values that could occur with the variety of sample sizes that may have been collected by the end of the duration. This is the way many researchers in the social sciences actually collect data. There is nothing wrong with this procedure, because each datum is completely insulated from the others, and the result of each measurement is unaffected by any measurement that was made before or after. In some respects, sampling for a fixed duration is better insulated from the data than sampling until N reaches a threshold, because sampling for a fixed duration means each measurement is taken without considering any previous measurements at all, but sampling until N reaches a threshold depends on counting the previous measurements. The point, though, is that the space of possible t_{null} values from the null hypothesis is different when

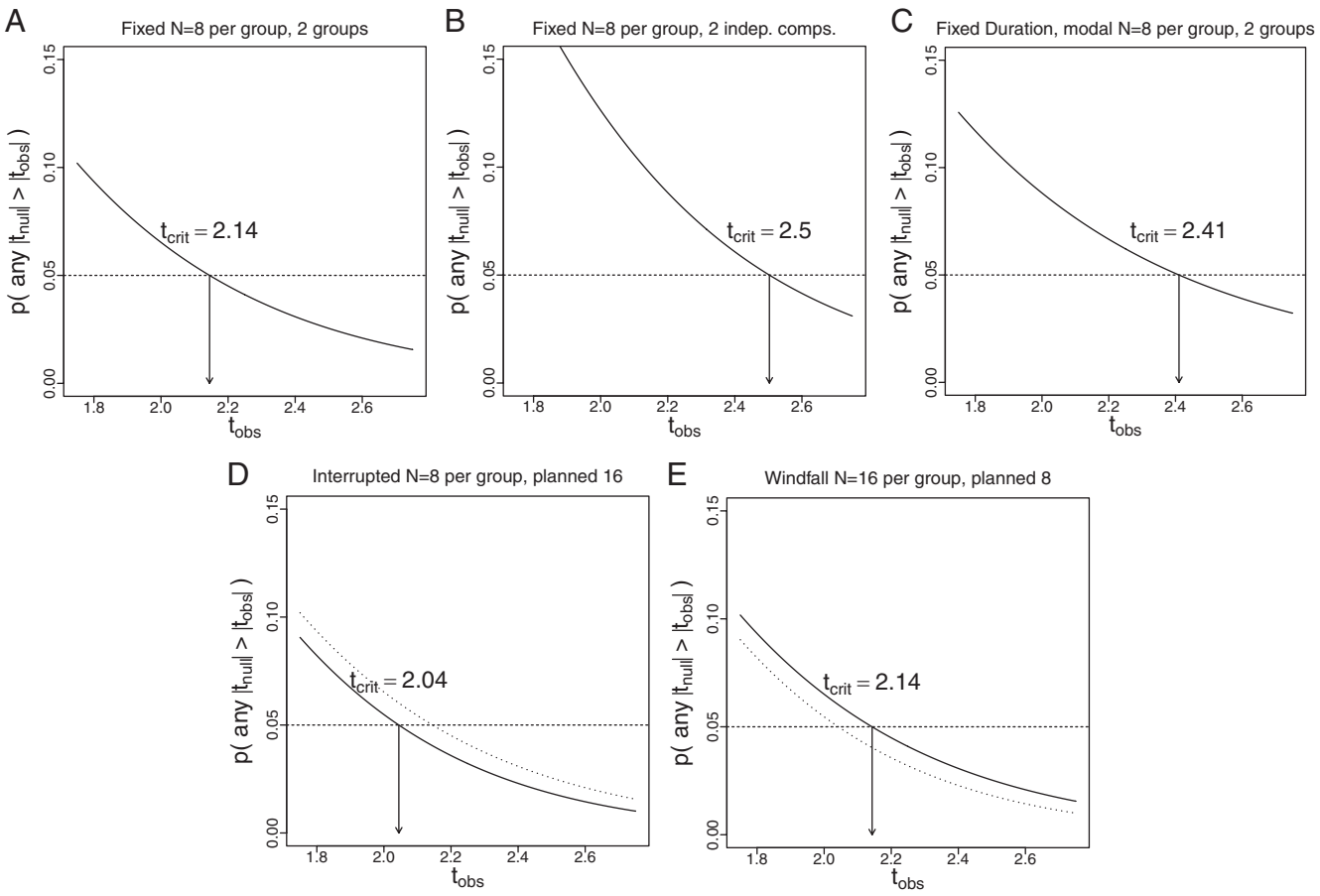

Figure 10. Probability of *t* values sampled from a null hypothesis for different sampling intentions. indep. comps. = independent comparisons; obs = observed; crit = critical.

intending to sample for fixed duration than when intending to sample until threshold *N,* and therefore the *p* value changes.

The upper right Panel C of Figure 10 shows the probability of obtaining t_{null} values from the null hypothesis when the sample sizes are random with a modal value of 8, with a variety of other sample sizes possible. It was assumed that the researcher collected data for 4 hr in a facility that can seat at most 5 subjects per hour. In any given hour, there are usually about four seats filled, occasionally five seats filled, and often fewer than four seats filled. (The exact probabilities determine the quantitative results but are irrelevant to the qualitative argument.) The plot shows that the value of t_{obs} for which $p = .05$ is $t_{crit} = 2.41$. In particular, for the data of Figure 4, $t_{obs} = 2.33$ does not exceed the critical value, and $p > .05$. Thus, although the data in Figure 4 have not changed, their *p* value has changed because it is computed relative to a different space of possibilities.

The *p* value for violated intentions: Interruption or windfall. Because the *p* value is defined by the space of possible t_{null} values generated by the null hypothesis when the intended experiment is repeated, the *p* value should be based on the intended sample size, not merely the actually obtained sample size. This is exactly analogous to basing corrections for multiple tests on the intended space of tests.

Consider the case of interrupted research, in which the researcher intended to collect $N = 16$ per group (say) but was unexpectedly interrupted, perhaps because of falling ill or because of computer failure, and therefore collected only $N = 8$ per group. Most analysts and all statistical software would use $N = 8$ per group to compute a *p* value. This is inappropriate, however, because the space of possible t_{null} values from the null hypothesis should actually be dominated by the intended sampling scheme, not by a rare accidental quirk. Suppose that the probability of the interruption is just 1 in 50 (2%), so that when t_{null} values are generated from the null hypothesis, 98% of the time those t_{null} values should be based on the intended $N = 16$, and only 2% of the time should they be based on the rare occurrence of $N = 8$. The resulting *p* values are shown in the lower left Panel D of Figure 10. Notice that the critical value is much lower than if the analysis had inappropriately assumed a fixed sample size of $N = 8$, indicated by the dotted curve.

Consider instead a case in which there is a windfall of data, perhaps caused by miscommunication so two research assistants collect data instead of only one. That is, the researcher intended to collect $N = 8$ per group, but the miscommunication produced $N = 16$ per group. Most analysts and all statistical software would use $N = 16$ per group to compute a *p* value. This is inappropriate,

however, because the space of possible t_{null} values from the null hypothesis should actually be dominated by the intended sampling scheme, not by a rare accidental quirk. Suppose that the probability of the windfall is just 1 in 50 (2%), so that when t_{null} values are generated from the null hypothesis, 98% of the time those t_{null} values should be based on the intended $N = 8$, and only 2% of the time should they be based on the rare occurrence of $N = 16$. The resulting p values are shown in the lower right Panel E of Figure 10. Notice that the critical value is much higher than if the analysis had inappropriately assumed a fixed sample size of $N = 16$, indicated by the dotted curve.

The p value for intending to sample until threshold t_{obs}. The conventional assumption for sampling data is to continue sampling until N reaches a fixed threshold. Alternatively, sampling could continue until t_{obs} reaches a fixed threshold (e.g., $|t_{obs}| > 3.0$). Notice that I am not conducting a t test with each datum collected and then stopping if I have achieved significance by a conventional fixed-N critical value; the case of sequential testing is considered later. Instead, here I am setting a threshold t value, fixed at 3.0 (say), and I am observing how big N_{obs} gets before exceeding that value. Having exceeded the threshold t value does not indicate significance. Instead, the question is, How big a sample does it take to exceed that value? If there is a real difference between groups, $|t_{obs}|$ should exceed 3.0 after relatively few data values have been collected, but if there is little difference between groups, $|t_{obs}|$ will take a long time to exceed 3.0. If $|t_{obs}|$ exceeds 3.0 with far smaller N_{obs} than would be expected from the null hypothesis, I reject the null hypothesis.

Figure 11 shows the p value for N_{obs}, $p(N_{null} \leq N_{obs})$, when the null hypothesis is true. The p values were computed by Monte Carlo simulation of 200,000 sequences generated from the null hypothesis. For each sequence, data were generated randomly from normal distributions, starting with $N_1 = N_2 = 5$, and alternately increasing N in each group until $t_{obs} > 3.0$ (to a maximum of $N_{obs} = N_1 + N_2 = 52$). Across the 200,000 simulated sequences, the simulation tallied how many stopped at $N_{obs} = 10$, how many stopped at $N_{obs} = 11$, how many stopped at $N_{obs} = 12$, and so on. The p value for N_{obs} is simply the total proportion of sequences that stopped at or before that N_{obs}.

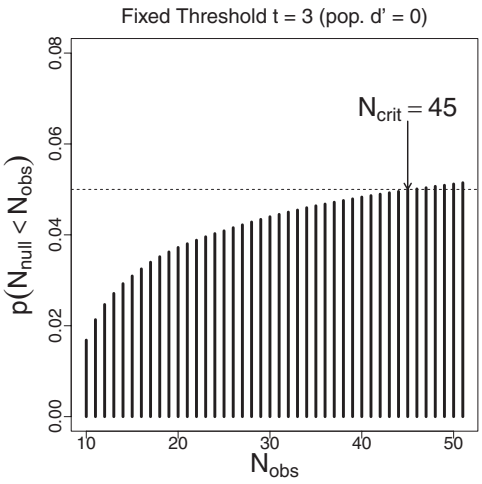

Figure 11. The p value when the data collection continues until when $|t_{obs}| > 3.0$, starting with $N_1 = N_2 = 5$ and increasing N by 1 alternately in each group. The horizontal dashed line indicates $p = .05$. (pop. d'=0) means that population effect size is zero (i.e., the null hypothesis is true). obs = observed; crit = critical.

Figure 11 indicates that N_{crit} is 45, which means that if $|t_{obs}|$ exceeds 3.0 by the time $N_{obs} = 45$, then $p < .05$ and the null hypothesis can be rejected. For example, suppose an experimenter collects data with $N_{obs} = 49$ and $t_{obs} = 3.06$. If the data were collected until $t_{obs} > 3.0$, the p value is not less than .05. But if the data had been collected with a fixed-N sampling scheme, the p value would be less than .05.

Sampling until t_{obs} exceeds a threshold might seem to be unusual, but it could prove to be efficient in cases of large effect sizes because small N_{obs} will be needed to exceed the threshold t. The scheme is exactly analogous to sampling schemes for estimating the bias in a coin, as explained for example by Berger and Berry (1988). To estimate the bias in a coin one could sample until reaching a threshold number of flips and count the number of heads, or one could sample until attaining a threshold number of heads and count the number of flips. The point is that for any given result involving a specific t_{obs} and N_{obs}, there is no unique p value because p is the rarity of the result in the space of possible t_{null} or N_{null} values sampled from the null hypothesis, and that space depends on the intention of the data collector.

Conclusion regarding dependence of p value on sampling intention. This section has emphasized that any specific t_{obs} and N_{obs} has many different p values, depending on the sampling intentions of the data collector. Conventional p values assume that the data collector intended to collect data until N_{obs} reached a preset threshold. Conventional methods also recognize that p values change when the intended comparisons change and therefore prescribe various corrections for various intended comparisons. This section showed examples of p values under other sampling intentions, such as fixed duration, unexpected windfall or interruption, and sampling until threshold t_{obs} instead of sampling until threshold N_{obs}. Again, the point is that any specific t_{obs} and N_{obs} has many different p values, and therefore basing conclusions on "the" p value and "the" significance is a misleading ritual.

It is important to recognize that NHST cannot be salvaged by attempting to fix or set the sampling intention explicitly in advance. For example, consider two researchers who are interested in the effects of a smart drug on IQ. They collect data from identical conditions. The first researcher obtains the data shown in Figure 3. The second researcher happens to obtain identical data (or at least data with identical t_{obs} and N_{obs}). Should the conclusions of the researchers be the same? Common sense, and scientific acumen, suggests that the conclusions should indeed be the same because the data are the same. But NHST says no, the conclusions should be different, because, it turns out, the first researcher collected the data with the explicit intention to stop at the end of the week and compare with another group of data to be collected the next week, and the second researcher collected the data with the explicit intention to stop when a threshold sample size of $N_1 = N_2 = 50$ was achieved but had an unexpected interruption. Bayesian analysis, on the other hand, considers only the actual data obtained, not the space of possible but unobtained data that may have been sampled from the null hypothesis.

NHST Has 100% False Alarm Rate in Sequential Testing

Under NHST, sequential testing of data generated from the null hypothesis will eventually lead to a false alarm. With infinite patience, there is 100% probability of falsely rejecting the null. This is known as "sampling to reach a foregone conclusion" (e.g., Anscombe, 1954). To illustrate this phenomenon, a computer simulation generated random values from a normal distribution with mean zero and standard deviation one, assigning each sequential value alternately to one or the other of two groups, and at each step conducting a two-group *t* test assuming the current sample sizes were fixed in advance. Each simulated sequence began with $N1 = N2 = 3$. If at any step the *t* test indicated $p < .05$, the sequence was stopped and the total $N (= N1 + N2)$ was recorded. Otherwise, another random value was sampled from the zero-mean normal and included in the smaller group, and a new *t* test was conducted. For purposes of illustration, each sequence was limited to a maximum sample size of $N = 5{,}000$. The simulation ran 1,000 sequences.

The results are displayed in the left panel of Figure 12, which shows the proportion of the 1,000 sequences that had (falsely) rejected the null with a sample size of *N* or less. As can be seen, by the time $N = 5,000$, nearly 60% of the sequences had falsely rejected the null. The increase is false alarm proportion is essentially linear on $\log(N)$, and rises to 100% as *N* grows arbitrarily large. Intuitively, this 100% false alarm rate occurs because NHST can only reject the null and therefore must do so eventually.

Bayesian decision making, using the HDI and ROPE, does not suffer a 100% false alarm rate in sequential testing. Instead, the false alarm rate asymptotes at a much lower level, depending on the choice of ROPE. For illustration, again a computer simulation generated random values from a normal distribution with mean of zero and standard deviation of one, assigning each sequential value alternately to one or the other of two groups but at each step conducting a Bayesian analysis and checking whether the 95% HDI completely excluded or was contained within a ROPE from −0.15 to 0.15.

The right panel of Figure 12 shows the results. The false alarm rate rose to an asymptotic level of 8.7% at a sample size of about 200 per group (400 total). Once the sample size got to approximately 300 per group (600 total), the 95% HDI became small enough to fall completely inside the ROPE when the sample means happened to be nearly equal. When the sample size got to approximately 1,850 per group (3,700 total), the 95% HDI essentially always fell within the ROPE, correctly accepting the null hypothesis. The qualitative behavior exhibited in the right panel of Figure 12 is quite general, with the quantitative detail depending on the width of the ROPE. When the ROPE is wider, the asymptotic false alarm rate is lower, and a smaller sample size is required for the 95% HDI to fall inside the ROPE.

Confidence Intervals Provide No Confidence

The previous sections focused on *p* values and false alarm rates in NHST. This section focuses on confidence intervals.

A confidence interval depends on sampling intentions. There are various equivalent definitions of the confidence interval, but they all are based on sampling distributions. The most general and coherent definition is this:

> A 95% confidence interval on a parameter is the range of parameter values that would not be rejected at $p = .05$ by the observed set of data.

In the case of the NHST *t* test, instead of checking merely whether the null hypothesis, $\mu_1 - \mu_2 = 0$, can be rejected at $p = .05$, one checks whether every other value of $\mu_1 - \mu_2$ can be rejected at $p = .05$. The range of unrejected values is the 95% confidence interval. This definition is general because it applies to any model and any stopping rule. And the definition is coherent because it makes explicit that the confidence interval is directly linked to the *p* value.

A crucial implication of the definition is this: When the sampling intention changes, the *p* value changes and so does the confidence interval. There is a different 95% confidence interval

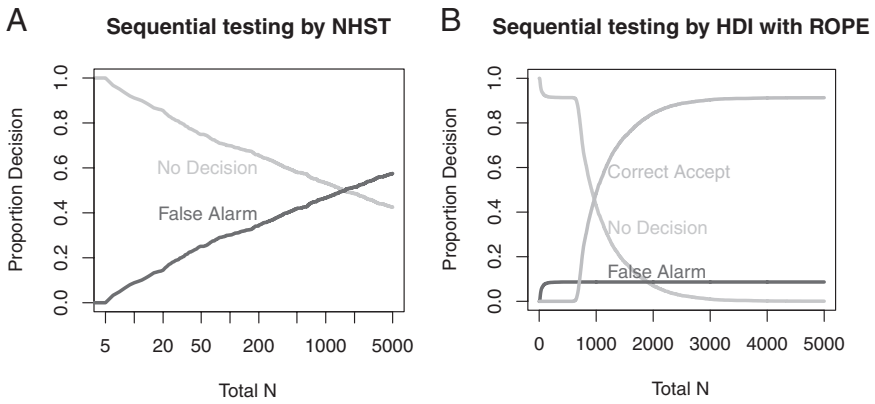

Figure 12. Proportion of decisions when data are sequentially sampled from the null hypothesis and testing is conducted with every datum. Left panel shows that the probability of false alarm in NHST continually rises with sample size. Right panel shows that the probability of false alarm in Bayesian analysis asymptotes at a relatively small value. NHST = null hypothesis significance testing; HDI = highest density interval; ROPE = region of practical equivalence.

for every different sampling intention, which includes different comparison intentions. Standard software packages for NHST typically implement changes in confidence intervals only for a subset of multiple-comparison intentions in ANOVA, but the software should also implement changes in confidence intervals for other sorts of multiple tests and for sampling intentions other than threshold sample size.

A confidence interval carries no distributional information. A confidence interval provides no information regarding which values within its bounds are more or less credible. In particular, a confidence interval on the difference of means does not indicate that a difference of means in its middle is more credible than a difference of means at its end points.

It is tempting to imbue the confidence interval with distributional information that is not actually present. As an example of imposing distributional information on a confidence interval, consider a plot of the *p* value (for a particular sampling intention) as a function of the parameter value (e.g., Poole, 1987; Sullivan & Foster, 1990). Such a graph captures the intuition that some sort of probability is higher in the middle of the interval than near its ends. Unfortunately, the plot of $p(\text{any } |t_{hyp}| > |t_{obs}|)$ as a function of hypothesized $\mu_1 - \mu_2$ is not a probability distribution at all; for instance, it does not integrate to one, as probability distributions must. Moreover, the *p* value is not the probability of the hypothesized parameter difference conditional on the data, which is provided only by the Bayesian posterior distribution. More sophisticated forms of the approach construct actual probability distributions over the parameter space, such that different areas under the distribution correspond to confidence levels (e.g., Schweder & Hjort, 2002; Singh, Xie, & Strawderman, 2007). These *confidence distributions* can correspond exactly with the Bayesian posterior distribution when using a particular form of noninformative prior (Schweder & Hjort, 2002, pp. 329–330). But unlike Bayesian posterior distributions, confidence distributions change when the sampling intention changes, just as *p* values and confidence intervals change.

Another way to imbue a confidence interval with a distributional interpretation is by superimposing a sampling distribution upon it. In particular, take the sampling distribution of the difference of sample means from the null hypothesis, denoted $p(\bar{y}_1 - \bar{y}_2|\mu_1 - \mu_2 = 0)$, re-center it on the observed difference of sample means, and then superimpose that distribution on the parameter axis, $\mu_1 - \mu_2$ (Cumming, 2007; Cumming & Fidler, 2009). Unfortunately, this approach already assumes that $\mu_1 - \mu_2$ has a specific, fixed value to generate the sampling distribution; hence, the result cannot represent a probability distribution over other candidate values of $\mu_1 - \mu_2$. Moreover, this approach is possible only because the parameter value μ and the sample estimator $\bar{y}$ happen to be on commensurate scales, so the sampling distribution of $\bar{y}_1 - \bar{y}_2$ can be superimposed on the parameter difference $\mu_1 - \mu_2$ despite their different meanings. As an example in which the sampling distribution of an estimator is quite different than the underlying parameter, consider estimating the probability of left handedness in a population. The parameter is a value on a continuous scale from zero to one, and the confidence interval on the parameter is a continuous subinterval. But the sample estimator is the proportion of left handers out of the sample size *N,* and the sampling distribution is a binomial distribution on discrete values 0/*N,* 1/*N,* 2/*N,* . . ., *N*/*N.* There is no way to re-center the discrete sampling distribution of the observed proportion to produce a continuous distribution on the parameter scale.

In summary, the classic confidence interval has no distributional information about the parameter values. A value in the middle of a confidence interval cannot be said to be more or less credible than a parameter value at the limits of the confidence interval. Superimposing a sampling distribution, from a fixed parameter value, onto the parameter scale says nothing about the probability of parameter values and is not generally possible. Recent elaborations of the confidence-interval concept into confidence distributions are dependent on the sampling intention of the data collector. Only the Bayesian posterior distribution explicitly indicates the probability of parameter values without being dependent on sampling intentions.

ROPE method cannot be used to accept null value in NHST. Because an NHST confidence interval (CI) has some properties analogous to the Bayesian posterior HDI, it may be tempting to try to adopt the use of the ROPE in NHST. Thus, we might want to accept a null hypothesis in NHST if a 95% CI falls completely inside the ROPE. This approach goes by the name of *equivalence testing* in NHST (e.g., Rogers, Howard, & Vessey, 1993; Westlake, 1976, 1981). Unfortunately, the approach fails because the meaning of the CI is not the same as the HDI. In a Bayesian approach, the 95% HDI actually includes the 95% of parameter values that are most credible. Therefore, when the 95% HDI falls within the ROPE, we can conclude that 95% of the credible parameter values are practically equivalent to the null value. But a 95% CI from NHST says nothing directly about the credibility of parameter values. Crucially, even if a 95% CI falls within the ROPE, a change of intention will change the CI and the CI may no longer fall within the ROPE. For example, if the two groups being compared are intended to be compared to other groups, then the 95% CI is much wider and may no longer fall inside the ROPE.

Summary regarding NHST confidence interval. In summary, a confidence interval provides very little information. Its end points can vary dramatically depending on the sampling intention of the data collector because the end points of a confidence interval are defined by *p* values, which depend on sampling intentions. Moreover, there is no distributional information regarding points within a confidence interval, and we cannot say that a parameter value in the middle of a confidence interval is more probable than a parameter value at the end of a confidence interval. One consequence of this dearth of information is that the confidence interval cannot be used with a ROPE to decide to accept the null hypothesis. Another consequence is that power estimates are extremely uncertain, as is shown next.

In NHST, Power Is Extremely Uncertain

In NHST, power is computed by assuming a punctate value for the effect size, even though there is uncertainty in the effect size. In retrospective power analysis, the range of uncertainty in the NHST power estimate is indicated by computing power at the limits of the 95% confidence interval on the effect size. Unfortunately, this typically results in a huge range on the power estimate, rendering it virtually useless, as many authors have pointed out (e.g., Gerard et al., 1998; Miller, 2009; Nakagawa & Foster, 2004; O'Keefe, 2007; Steidl et al., 1997; Sun et al., 2011; Thomas, 1997). As an example, recall the data in Figure 3. A traditional

two-group t test yielded $t(87) = 1.62$, $p = .110$, with a 95% confidence interval on the difference of means from -0.361 to 3.477. (Because these data have outliers, the traditional t test is not applicable to these data, as discussed earlier in the article, but this issue is tangential to the points made here about the poverty of information in an NHST power analysis.) At the point estimate of the effect size, the power is 35.0%. But at the limits of the 95% confidence interval on the effect size, the power is 5.0% and 94.0%, which spans almost the full possible range of power. Thus, NHST power analysis tells us almost nothing about the power of the experiment. Consider instead the large-sample ($N > 1,000$) data of Figure 5, which showed essentially no difference between sample means. The NHST power at the point estimate of the effect size is 5.0% (i.e., the false alarm rate for the null hypothesis). But at the limits of the confidence interval on the effect size, the NHST power is 49.5% and 50.6% (for effects of opposite signs). The reason that there is such a high probability of rejecting the null, even at the small limits of the confidence interval, is that a large sample size can detect a small effect. Thus, even with a huge sample size, NHST estimates of retrospective power can be very uncertain. These uncertain power estimates by NHST contrast with the precise estimates provided by the Bayesian approach.

In prospective power analysis, frequentists can try different hypothetical parameter values, but because the hypothetical values are not from a probability distribution, they are not integrated into a single power estimate. The Bayesian approach to power, illustrated in Figure 6, is awkward to implement in a frequentist framework. The approach requires that the hypothesis is expressed as a probability distribution over parameters (shown in the leftmost box of Figure 6), which is shunned in frequentist ontology. Perhaps more important, even if a frequentist admits a hypothesis expressed as a probability distribution over parameter values, it is difficult to imagine where the distribution would come from, especially for complex multidimensional parameter spaces, if it were not generated as a Bayesian posterior distribution. Finally, even if the approach were adapted, with NHST conducted on simulated data instead of Bayesian analysis, there would still be the inherent fickleness of the resulting p values and confidence intervals. In other words, the simulated data could be generated by one sampling intention, but the NHST could assume many different sampling intentions (because the data bear no signature of the sampling intention), and many different powers could be computed for the same hypothetical effect size.

Conclusion

In the examples presented above, which contrasted results from Bayesian estimation (BEST) and the NHST t test, the advantage of BEST was not solely from model flexibility in Bayesian software. The main advantage was in Bayesian estimation per se, which yields an explicit posterior distribution over parameters unaffected by sampling intentions. Recall that BEST revealed far richer information than the NHST t test even when parametric modeling assumptions were removed from NHST by using resampling. A crucial argument against NHST is completely general and does not rely on any particular illustrative model, namely, that in NHST p values and CIs are based on sampling distributions, and sampling distributions are based on sampling intentions, and different sampling intentions change the interpretation of data even though the intentions have no effect on the data.

Some people may have an impression that Bayesian estimation merely substitutes an assumption about a prior distribution in place of an assumption about a sampling intention in NHST, and therefore both methods are equally dubious. This perspective is not correct, because the assumptions of Bayesian estimation are epistemologically appropriate, and the assumptions of NHST are not.

In NHST, the sampling intentions of the data collector (which determine the p value and CI) are unknowable and, more important, irrelevant. The intentions are unknowable in the sense that true intentions can be subconscious and hidden from one's self, covert and hidden from one's peers, or overt but changing through time in response to a dynamic environment. The intentions are especially unknowable to the data themselves, which are collected in such a way as to be insulated from the experimenter's intentions. More important, because the data were not influenced by the intentions, the intentions are irrelevant to one's interpretation of the data. There is no reason to base statistical significance on whether the experimenter intended to stop collecting data when $N = 47$ or when the clock reached 5:00 p.m.

On the other hand, in Bayesian estimation the prior distribution is both explicitly presented and relevant. The prior cannot be chosen capriciously or covertly to predetermine a desired conclusion. Instead, the prior must be justified to a skeptical audience. When there is lack of prior knowledge, the prior distribution explicitly expresses the uncertainty, and modest amounts of data will overwhelm the prior. When there is disagreement about appropriate priors, different priors can be used and the resulting posterior distributions can be examined and checked for differences in conclusions. When there is strong prior information, it can be a serious blunder not to use it. For example, consider random drug or disease tests. Suppose a person selected at random from a population is tested for an illicit drug, and the test result is positive. What is the probability that the person actually used the drug? If the prior probability of drug use in the population is small (and the test is realistically imperfect), then the posterior probability that the person used the drug is also surprisingly small. (e.g., Berry, 2006; Kruschke, 2011b, p. 71). The proper interpretation of the data (i.e., the test result) depended on the Bayesian incorporation of prior knowledge. Thus, the prior distribution in Bayesian estimation is both explicitly justified and epistemologically relevant.

Some people may wonder which approach, Bayesian or NHST, is more often correct. This question has limited applicability because in real research we never know the ground truth; all we have is a sample of data. If we knew the ground truth, against which to compare the conclusion from the statistical analysis, we would not bother to collect the data. If the question of correctness is instead asked of some hypothetical data generator, the assessment is confined to that particular distribution of simulated data, which likely has only limited resemblance to real-world data encountered across research paradigms. Therefore, instead of asking which method is more often correct in some hypothetical world of simulated data, the relevant question is asking which method provides the richest, most informative, and meaningful results for any set of data. The answer is always Bayesian estimation.

Beyond the general points about the relative richness of information provided by Bayesian estimation, there are also many practical advantages to Bayesian estimation over NHST. The software for Bayesian estimation (i.e., JAGS/BUGS) is very flexible and can accommodate realistic data situations that cause difficulties for NHST. For example, the Bayesian software can incorporate non-normal data distributions, censored data, unequal variances, unbalanced sample sizes, nonlinear models, and multiple layers of hierarchical structure in a straightforward unified framework. NHST can have great difficulties with those situations because it can be problematic to derive sampling distributions for *p* values (even when assuming a fixed *N* sampling intention).

Summary

Some people have the impression that conclusions from NHST and Bayesian methods tend to agree in simple situations such as comparison of two groups: "Thus, if your primary question of interest can be simply expressed in a form amenable to a *t* test, say, there really is no need to try and apply the full Bayesian machinery to so simple a problem" (Brooks, 2003, p. 2694). The present article has shown, to the contrary, that Bayesian estimation always provides much richer information than the NHST *t* test and sometimes comes to different decisions.

Bayesian estimation provides rich and complete information about the joint distribution of credible parameter values, including the means, standard deviations, effect size, and normality. Bayesian estimation can accept the null value by using a decision procedure involving the HDI and ROPE. Bayesian estimation provides precise power analysis for multiple goals of research.

The NHST *t* test, on the other hand, has many foundational problems. The *p* values on which it bases decisions are ill defined, as are confidence intervals because they are inherently linked to *p* values. Confidence intervals (CIs) carry no distributional information and therefore render power to be virtually unknowable because of its uncertainty. And NHST has no way to accept the null hypothesis, because a CI changes when the sampling intention changes, and a CI does not have the meaning of an HDI.

Appendix D explains that Bayesian estimation is typically also more informative than the Bayesian model-comparison approach, which involves the Bayes factor. The Bayes factor can be extremely sensitive to the choice of alternative-hypothesis prior distribution. The Bayes factor hides the uncertainty in the parameter estimation, even concluding substantial evidence for the null hypothesis when there is great uncertainty.

The software for Bayesian parameter estimation is free, easy to use, and extendable to complex designs and models (as explained in the section that described the model and in Appendix B). The programs can be downloaded from http://www.indiana.edu/~kruschke/BEST/, where instructions for software installation are also provided. An extensive introduction to the methods used in those programs is available in a textbook (Kruschke, 2011b).

All of these facts point to the conclusion that Bayesian parameter estimation supersedes the NHST *t* test.

References

Adcock, C. J. (1997). Sample size determination: A review. *Statistician, 46,* 261–283. doi:10.1111/1467-9884.00082

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anscombe, F. (1954). Fixed sample size analysis of sequential observations. *Biometrics, 10,* 89–100. doi:10.2307/3001665

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions, 53,* 370–418. doi:10.1098/rstl.1763.0053

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76,* 159–165.

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery, 5,* 27–36. doi:10.1038/nrd1927

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40,* 318–335. doi:10.2307/2333350

Brooks, S. P. (2003). Bayesian computation: A statistical revolution. *Philosophical Transactions of the Royal Society of London: Series A. Mathematical, Physical and Engineering Sciences, 361,* 2681–2697.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics, 29,* 89–93. doi:10.1111/j.1467-9639.2007.00267.x

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology, 217,* 15–26. doi:10.1027/0044-3409.217.1.15

Damgaard, L. H. (2007). Technical note: How to use WinBUGS to draw inferences in animal models. *Journal of Animal Science, 85,* 1363–1368. doi:10.2527/jas.2006-543

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference, 124,* 121–144. doi:10.1016/S0378-3758(03)00198-8

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170,* 95–113. doi:10.1111/j.1467-985X.2006.00438.x

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Annals of Mathematical Statistics, 41,* 214–226. doi:10.1214/aoms/1177697203

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* Basingstoke, England: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. doi:10.1177/1745691611406920

Doyle, A. C. (1890). *The sign of four.* London, England: Spencer Blackett.

du Prel, J.-B., Röhrig, B., Hommel, G., & Blettner, M. (2010). Choosing statistical tests: Part 12 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International, 107,* 343–348. doi:10:3238/arztebl.2010.0343

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods, 13,* 668–689. doi:10.1177/1094428110380467

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242. doi:10.1037/h0044139

Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics, 40,* 575–586. doi:10.2307/2530902

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141,* 2–18. doi:10.1037/a0024338

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods

for eliciting probability distributions. *Journal of the American Statistical Association, 100,* 680–701. doi:10.1198/016214505000000105

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *Annals of Statistics, 33,* 1–53. doi:10.1214/009053604000001048

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1,* 515–533.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5,* 189–211. doi:10.1080/19345747.2011.618213

Gelman, A., & Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology.* Advance online publication. doi:10.1111/j.2044-8317.2011.02037.x

Gelman, A., & Shalizi, C. R. (in press). Philosophy and the practice of Bayesian statistics in the social sciences. In H. Kincaid (Ed.), *The Oxford handbook of philosophy of social science.* Oxford, England: Oxford University Press.

Gerard, P. D., Smith, D. R., & Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management, 62,* 801–807. doi:10.2307/3802357

Gosset, W. S. (1908). The probable error of a mean. *Biometrika, 6,* 1–25.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 29,* 83–100.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128. doi:10.2307/1164588

Hobbs, B. P., & Carlin, B. P. (2007). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics, 18,* 54–80. doi:10.1080/10543400701668266

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician, 55,* 19–24. doi:10.1198/000313001300339897

Jackman, S. (2009). *Bayesian analysis for the social sciences.* New York, NY: Wiley.

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

Jones, M. C., & Faddy, M. J. (2003). A skew extension of the *t*-distribution, with applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65,* 159–174. doi:10.1111/1467-9868.00378

Joseph, L., Wolfson, D. B., & du Berger, R. (1995a). Sample size calculations for binomial proportions via highest posterior density intervals. *Statistician, 44,* 143–154. doi:10.2307/2348439

Joseph, L., Wolfson, D. B., & du Berger, R. (1995b). Some comments on Bayesian sample size determination. *Statistician, 44,* 167–171. doi:10.2307/2348442

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Kruschke, J. K. (2005). *Excel workbooks that generate sampling distributions from arbitrary populations for any sample statistic: An overview and invitation.* Retrieved from http://www.indiana.edu/~jkkteach/ExcelSampler/ UsersIntroduction_21_07_2005.pdf

Kruschke, J. K. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science, 1,* 658–676. doi:10.1002/wcs.72

Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14,* 293–300. doi:10.1016/j.tics.2010.05.001

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312.

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS.* Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K. (2011c). Introduction to special section on Bayesian data analysis. *Perspectives on Psychological Science, 6,* 272–273. doi:10.1177/1745691611406926

Kruschke, J. K. (in press). The posterior predictive check can and should be Bayesian. *British Journal of Mathematical and Statistical Psychology.*

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association, 84,* 881–896.

Lazarus, R. S., & Eriksen, C. W. (1952). Effects of failure stress upon skilled performance. *Journal of Experimental Psychology, 43,* 100–105. doi:10.1037/h0056614

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112,* 662–668. doi:10.1037/0033-295X.112.3.662

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3,* 112–115. doi:10.1111/j.2041-210X.2011.00131.x

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52,* 362–375. doi:10.1016/j.jmp.2008.03.002

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23,* 151–169. doi:10.1146/annurev.publhealth.23.100901.140546

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115. doi:10.1086/288135

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests* (pp. 395–425). Mahwah, NJ: Erlbaum.

Meyer, R., & Yu, J. (2000). BUGS for a Bayesian analysis of stochastic volatility models. *Econometrics Journal, 3,* 198–215. doi:10.1111/1368-423X.00046

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review, 16,* 617–640. doi:10.3758/PBR.16.4.617

Nakagawa, S., & Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica, 7,* 103–108. doi:10.1007/s10211-004-0095-z

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS.* Hoboken, NJ: Wiley.

O'Keefe, D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures, 1,* 291–299. doi:10.1080/19312450701641375

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika, 23,* 114–133. doi:10.2307/2333631

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.* Retrieved from http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Poole, C. (1987). Beyond the confidence interval. *American Journal of Public Health, 77,* 195–199. doi:10.2105/AJPH.77.2.195

R Development Core Team. (2011). *R: A language and environment for statistical computing.* Retrieved from http://www.R-project.org

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565. doi:10.1037/0033-2909.113.3.553

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review, 12,* 195–223.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G.

(2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi:10.3758/PBR.16.2.225

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12,* 1151–1172. doi:10.1214/aos/1176346785

Sadiku, M. N. O., & Tofighi, M. R. (1999). A tutorial on simulation of queueing models. *International Journal of Electrical Engineering Education, 36,* 102–120.

Schweder, T., & Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics, 29,* 309–332. doi:10.1111/1467-9469.00285

Singh, K., Xie, M., & Strawderman, W. E. (2007). Confidence distribution (CD)—distribution estimator of a parameter. In R. Liu, W. Strawderman, & C.-H. Zhang (Eds.), *Complex datasets and inverse problems* (Vol. 54, pp. 132–150). Beachwood, OH: Institute of Mathematical Statistics.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 157,* 357–416.

Steidl, R. J., Hayes, J. P., & Schauber, E. (1997). Statistical power analysis in wildlife research. *Journal of Wildlife Management, 61,* 270–279. doi:10.2307/3802582

Sullivan, K. M., & Foster, D. A. (1990). Use of the confidence interval function. *Epidemiology, 1,* 39–42. doi:10.1097/00001648-199001000-00009

Sun, S., Pan, W., & Wang, L. L. (2011). Rethinking observed power: Concept, practice and implications. *Methodology, 7,* 81–87.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology, 11,* 276–280. doi:10.1046/j.1523-1739.1997.96102.x

Tsionas, E. G. (2002). Bayesian inference in the noncentral Student-*t* model. *Journal of Computational and Graphical Statistics, 11,* 208–221. doi:10.1198/106186002317375695

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54,* 491–498. doi:10.1016/j.jmp.2010.07.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. doi:10.3758/BF03194105

Wang, F., & Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science, 17,* 193–208. doi:10.1214/ss/1030550861

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Statistician, 46,* 185–191. doi:10.1111/1467-9884.00075

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34,* 28–35.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32,* 741–744. doi:10.2307/2529259

Westlake, W. J. (1981). Response to bioequivalence testing—a need to rethink. *Biometrics, 37,* 591–593.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J., Iverson, G., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6,* 291–298. doi:10.1177/1745691611406923

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review, 16,* 752–760. doi:10.3758/PBR.16.4.752

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association, 62,* 776–800.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the first international meeting* (pp. 585–603). Valencia, Spain: University of Valencia Press.

Zhang, Z., Lai, K., Lu, Z., & Tong, X. (in press). Bayesian inference and application of robust growth curve models using Student's *t* distribution. *Structural Equation Modeling.*

Appendix A

The Prior Distribution on ν

Figure A1 shows the prior distribution on the normality parameter, ν. Mathematically, it is $p(\nu|\lambda) = (1/\lambda) \exp[-(\nu - 1)/\lambda]$ for $\nu \geq 1$ and $\lambda = 29$. This prior was selected because it balances nearly normal distributions ($\nu > 30$) with heavy tailed distributions ($\nu < 30$). This prior distribution was chosen instead of several others that were considered, including various uniform distributions, various shifted gamma distributions, and various shifted and folded *t* distributions. It is easy to change this prior if the user desires, as described in Appendix B.

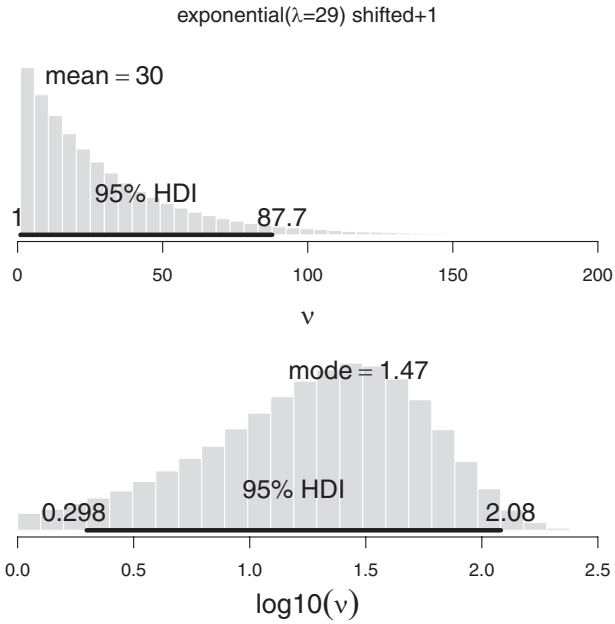

Figure A1. The prior distribution on the normality parameter ν. The upper panel shows the distribution on ν, as graphed iconically in the middle of Figure 2. The lower panel shows the same distribution on $\log_{10}(\nu)$ for easy comparison with the posterior distributions shown in Figure 3. HDI = highest density interval.

(*Appendices continue*)

Appendix B

Modifying the Program for Other Priors or Likelihoods

This appendix explains how to modify the Bayesian estimation program BEST.R to use other prior distributions or likelihood functions. This Appendix assumes that the reader is familiar with the basic operation of the programs available from http://www.indiana.edu/~kruschke/BEST/. Because of space constraints, I must assume that the reader is familiar with the basic structure of JAGS/BUGS programs, as explained, for example, by Kruschke (2011b, Chapter 7).

The model of Figure 2 is expressed in JAGS as

```
model {
  for ( i in 1:Ntotal ) {
    y[i] ~ dt( mu[x[i]], tau[x[i]], nu )
  }
  for ( j in 1:2 ) {
    mu[j] ~ dnorm( muM, muP )
    tau[j] <- 1/pow( sigma[j], 2 )
    sigma[j] ~ dunif( sigmaLow, sigmaHigh )
  }
  nu <- nuMinusOne + 1
  nuMinusOne ~ dexp(1/29)
}
```

where x[i] is the group membership (1 or 2) of the *i*th datum. The values for the constants in some of the priors are provided by the data statement later in the program:

```
muM = mean(y),
muP = 0.000001 * 1/sd(y)^2,
sigmaLow = sd(y)/1000,
sigmaHigh = sd(y) * 1000
```

where y is the vector of pooled data. The second line above says that the precision on the prior for μ_j, namely muP, is 0.000001 times the precision in the pooled data. The third line above says that the lowest value considered for σ_j is the standard deviation of the pooled data divided by 1,000. The fourth line above says that the highest value considered for σ_j is the standard deviation of the pooled data times 1,000. The prior on ν is set in the model specification above, in the line nuMinusOne ~ dexp(1/29). The value 1/29 makes the mean of the exponential to be 29.

If the user has strong previous information about the plausible values of the means and standard deviations, that information can be used to set appropriate constants in the prior. It is important to understand that the prior should be set to be agreeable to a skeptical audience.

For example, it could be that Group 2 is a control group drawn from the general population and Group 1 is a novel treatment. In this case one might have strong prior information about the control group but not about the novel treatment. In the case of IQ scores, it is known that the mean of the general population is 100 and the standard deviation is 15. But one's particular control group may deviate somewhat from the general population. Therefore one might want to change the prior specification in the model to

```
# Group 1 mean is uncertain:
mu[1] ~ dnorm(muM, muP)
tau[1] <- 1/pow(sigma[1], 2)
# Group 1 SD is uncertain:
sigma[1] ~ dunif(sigmaLow, sigmaHigh)
# Group 2 mean is nearly 100:
mu[2] ~ dnorm(100, 0.25)
tau[2] <- 1/pow(sigma[2], 2)
# Group 2 SD is between 10 and 20:
sigma[2] ~ dunif(10, 20)
```

In actual research analysis, the user would have to strongly justify the choice of informed prior, to convince a skeptical audience that the analysis is cogent and useful.

Changing the likelihood function for the data is also straightforward in JAGS. For example, suppose that one wanted to describe the data with a log-normal distribution instead of with a *t* distribution. Then the model specification could be as follows:

```
model {
  for ( i in 1:Ntotal ) {
    # log-normal likelihood:
    y[i] ~ dlnorm( log(mu[x[i]]), tau[x[i]] )
  }
  for ( j in 1:2 ) {
    mu[j] ~ dnorm( muM, muP )
    tau[j] <- 1/pow( sigma[j], 2 )
    sigma[j] ~ dunif( sigmaLow, sigmaHigh )
  }
}
```

Because the log-normal function has no normality parameter ν, that parameter is removed from the prior specification and from the rest of the program.

JAGS has many different distributions that can be used to define likelihood functions. For example, for modeling skewed distributions such as response times, a Weibull distribution could be used (Rouder, Lu, Speckman, Sun, & Jiang, 2005). If the analyst desires a likelihood function other than one already built into JAGS, the Poisson zeros trick can be used to specify virtually any likelihood function (e.g., Ntzoufras, 2009, p. 276). It is also straightforward to model censored data in JAGS; search my blog (http://doingbayesiandataanalysisogspot.com/) with the term "censor" for a detailed example.

(*Appendices continue*)

Appendix C

Doing Power Analysis

Examples for doing power analyses are provided in the program BESTexamplePower.R. Excerpts from that program are presented here.

For doing prospective power analysis, the user first creates an idealized data set that expresses the hypothesis. The function makeData creates the data in the following code:

```
prospectData = makeData(
  mu1 = 108, # mean of group 1
  sd1 = 17, # standard deviation of group 1
  mu2 = 100, # mean of group 2
  sd2 = 15, # standard deviation of group 2
  nPerGrp = 1000, # sample size in each group
  pcntOut = 10, # percent from outlier distrib.
  sdOutMult = 2.0, # SD multiplier of outlier dist.
  rnd.seed = NULL # seed for random number )
# Rename for convenience below:
y1pro = prospectData$y1
y2pro = prospectData$y2
```

(The resulting data are displayed in Figure 7.) Then the idealized data are submitted to a Bayesian data analysis. Only a short MCMC chain is created because it will be used for simulating experiments, not for creating a high-resolution representation of a posterior distribution from real data.

```
mcmcChainPro = BESTmcmc(y1pro, y2pro,
  numSavedSteps = 2000)
BESTplot(y1pro, y2pro, mcmcChainPro,
  pairsPlot = TRUE)
```

(The resulting posterior is displayed in Figures 8 and 9.) Then the power is estimated with the function BESTpower, as follows:

```
N1plan = N2plan = 50 # specify planned sample size
powerPro = BESTpower(
  # MCMC chain for the hypothetical parameters:
  mcmcChainPro,
  # sample sizes for the proposed experiment:
  N1 = N1plan, N2 = N2plan,
  # number of simulated experiments to run:
  nRep = 1000,
  # number of MCMC steps in each simulated run:
  mcmcLength = 10000,
  # number of simulated posteriors to display:
  showFirstNrep = 5,
  # ROPE on the difference of means:
  ROPEm = c(−1.5, 1.5),
  # ROPE on the difference of standard dev's:
  ROPEsd = c(−0.0,0.0),
  # ROPE on the effect size:
  ROPEeff = c(−0.0,0.0),
  # maximum 95% HDI width on the diff. of means:
  maxHDIWm = 15.0,
  # maximum 95% HDI width on the diff. of SD's:
  maxHDIWsd = 10.0,
  # maximum 95% HDI width on the effect size:
  maxHDIWeff = 1.0,
  # file name for saving results:
  saveName = "BESTexampleProPower.Rdata"
)
```

Retrospective power analysis uses the same call to the function BESTpower, but it uses an MCMC chain from a previous real data analysis instead of from a hypothetical data analysis and uses the actual sample sizes in the experiment rather than planned sample sizes.

(*Appendices continue*)

Appendix D

The Bayes-Factor Approach to Null Hypothesis Testing

The main body of this article explains Bayesian estimation of parameters in a descriptive model for data from two groups. From the complete posterior distribution on those parameters, one could make discrete decisions about the credibility of particular values of interest, such as null values. The Bayesian estimation approach provides rich information about the magnitude of the difference between means, difference between standard deviations, effect size, and normality.

If the researcher is not interested in estimating effect size or other aspects of the groups but instead is focused on rejecting or accepting a specific value relative to a distribution of alternative values, then there is another Bayesian approach to consider. This approach is called Bayesian model comparison, and it involves a statistic called the *Bayes factor*. I have previously discussed this topic with different examples (Kruschke, 2011a, 2011b, Chapter 12). Here I focus on the specific case of testing the difference of means between two groups.

Null Hypothesis Model and Alternative Hypothesis Model

In the model-comparison approach to null value assessment, one model expresses the null hypothesis, wherein the only available value for the difference of means is zero. This model effectively puts a spike-shaped prior distribution on the difference of means, such that the prior probability of non-zero differences of means is zero, and the prior probability density of zero difference of means is infinity. The second model expresses a particular alternative hypothesis, wherein there is a complete spectrum of available values for the difference of means, with a specific prior probability distribution on those values. The model comparison therefore contrasts a model that requires the difference to be zero against a model that allows many non-zero differences with particular prior credibility.

It is important to emphasize that this method compares the null hypothesis, expressed as a spike-shaped prior, against a particular shape of an alternative broad prior, for which there is no unique definition. The results of the model comparison do not provide an absolute preference for or against the null hypothesis; instead, the results indicate only the relative preference for or against the null with respect to the particular shape of alternative prior. There are typically a variety of alternative-hypothesis priors, and the relative preference for the null hypothesis can change dramatically depending on the choice of alternative-hypothesis prior (e.g., Dienes, 2008; Kruschke, 2011a; Liu & Aitkin, 2008; Vanpaemel, 2010). This sensitivity to the choice of alternative-hypothesis prior is one key reason to advise caution when using the Bayes-factor method.

The Bayes-factor method produces merely the relative credibilities of the two priors as descriptions of the data, without (necessarily) producing an explicit posterior distribution on the parameter values. Although the Bayes factor can be very sensitive to the choice of alternative hypothesis prior, the posterior distribution on the parameter values (as provided by BEST, for example) is typically very robust against reasonable changes in the prior when there are realistic numbers of data points. Thus, Bayesian estimation, with its explicit parameter distribution, not only is more informative than Bayesian model comparison but is also more robust.

Bayesian Model Comparison and the Bayes Factor

This section describes the mathematics of Bayesian model comparison, which are necessary for summarizing two approaches in the recent literature. Applying Bayes' rule (cf. Equation 1 in the main body of the article) to each model, we have the posterior probability of models M_1 and M_2 given by

$$\begin{aligned} p(M_1|D) &= p(D|M_1)\, p(M_1) \Big/ \sum_m p(D|M_m)\, p(M_m) \\ p(M_2|D) &= p(D|M_2)\, p(M_2) \Big/ \sum_m p(D|M_m)\, p(M_m) \end{aligned} \tag{D1}$$

where M_{m} denotes model m. Hence,

$$\frac{p(M_1|D)}{p(M_2|D)} = \underbrace{\frac{p(D|M_1)}{p(D|M_2)}}_{\mathrm{BF}} \frac{p(M_1)}{p(M_2)} \tag{D2}$$

where the ratio marked BF is the Bayes factor. Equation D2 shows that the BF converts the prior odds of the models, $p(M_1)/p(M_2)$, to the posterior odds of the models, $p(M_1|D)/p(M_2|D)$.

As the BF increases greater than 1.0, the evidence increases in favor of model M_1 over model M_2. The convention for interpreting the magnitude of the BF is that there is "substantial" evidence for model M_1 when the BF exceeds 3.0 and, equivalently, "substantial" evidence for model M_2 when the BF is less than 1/3 (Jeffreys, 1961; Kass & Raftery, 1995; Wetzels et al., 2011).

The terms of the Bayes factor are the marginal likelihoods within each model, which are the probability of the data given the parameter values in the model weighted by the prior probability of the parameter values:

$$\begin{aligned} p(D|M_m) = \int\int\int\int & d\mu_1\, d\mu_2\, d\sigma_1\, d\sigma_2 \\ & \times \underbrace{p(D|\mu_1,\ \mu_2, \sigma_1,\ \sigma_2, M_m)}_{\text{likelihood}} \\ & \times \underbrace{p(\mu_1, \mu_2, \sigma_1, \sigma_2|M_m)}_{\text{prior}} \end{aligned} \tag{D3}$$

(*Appendices continue*)

Equation 3 includes only the means and standard deviations of the groups and not a normality parameter ν, because the models in the literature assume normally distributed data. In the present application we have two models, namely the null hypothesis H_{null} and the alternative hypothesis H_{alt}, which have identical likelihood functions (i.e., normal density functions) and differ only in their prior distributions. In the null model, the prior $p(\mu_1, \mu_2, \sigma_1, \sigma_2|H_{null})$ puts zero probability on any parameter combination with $\mu_1 \neq \mu_2$. In the alternative model, the prior $p(\mu_1, \mu_2, \sigma_1, \sigma_2|H_{alt})$ spreads credibility over the full spectrum of combinations of μ_1 and μ_2.

The recent literature in psychological sciences includes at least two versions of Bayesian model comparison applied to two groups. One approach solved equations for the BF analytically for a particular choice of prior distribution. A subsequent approach used an MCMC solution for a more general model. The two approaches are now summarized.

Analytical Approach to Bayes Factor

An analytical approach was described by Rouder, Speckman, Sun, Morey, and Iverson (2009). Their descriptive model uses normal distributions of equal variance in the two groups. Hence, there are only three parameters to describe the groups, instead of five parameters as in BEST. The common standard deviation for the two groups is denoted σ, the overall mean across groups is denoted μ, and the difference between groups is denoted α, with $\mu_1 = \mu + \alpha/2$ and $\mu_2 = \mu - \alpha/2$. The model is reparameterized in terms of the effect size: $\delta = \alpha/\sigma$. The model for the null hypothesis assumes $\delta = 0$. If this prior distribution on δ were graphed, it would be zero for all values of δ except for an infinitely tall spike at $\delta = 0$. The model for the alternative hypothesis assumes that the prior on the effect size δ is a Cauchy(0,1) distribution, which is equivalent to a t distribution with mean of zero, standard deviation of one, and $\nu = 1$. In both the null and the alternative hypotheses, the prior on the precision $1/\sigma^2$ is assumed to be a gamma distribution with shape and rate parameters set to 0.5 (equivalent to a chi-square distribution with 1 degree of freedom), and the prior on μ is assumed to be an improper uniform distribution of infinite width. These assumptions for the prior distributions follow precedents of Jeffreys (1961) and Zellner and Siow (1980), and the approach is therefore called the JZS prior by Rouder et al. (2009).

Dienes (2008, 2011) provided another analytical solution, with a corresponding online calculator. A strength of the approach is that it allows a variety of forms for the alternative hypothesis, including normal distributions with non-zero means. A weakness is that it assumes a single value for the standard deviation of the groups, rather than a prior distribution with uncertainty. For succinctness, its specifics are not further discussed in this article.

MCMC Approach to Bayes Factor

Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009) used a model with separate standard deviations for the two groups and evaluated the BF using MCMC methods instead of analytical mathematics. The model assumes that the data in both groups are normally distributed. The grand mean μ is given a Cauchy(0,1) prior instead of the improper uniform distribution used by Rouder et al. (2009). The standard deviations, σ_1 and σ_2, each have folded Cauchy$^+$(0,1) priors instead of the gamma(0.5,0.5) distribution used by Rouder et al. In the alternative hypothesis, the effect size δ has a Cauchy(0,1) prior, the same as used by Rouder et al. The group means are $\mu_1 = \mu + \alpha/2$ and $\mu_2 = \mu - \alpha/2$, where $\alpha = \delta\sigma_{pooled}$ and $\sigma_{pooled} = \sqrt{(\sigma_1^2(N_1 - 1) + \sigma_2^2(N_2 - 1)/(N_1 + N_2 - 2)}$ (Hedges, 1981).

Instead of deriving the BF analytically, Wetzels et al. (2009) used MCMC methods to obtain a posterior distribution on the parameters in the alternative model and then adapted the Savage–Dickey (SD) method to approximate the Bayes factor (Dickey & Lientz, 1970). The SD method assumes that the prior on the variance in the alternative model, at the null value, equals the prior on the variance in the null model: $p(\sigma^2|H_{alt}, \delta = 0) = p(\sigma^2|H_{null})$. From this it follows that the likelihood of the data in the null model equals the likelihood of the data in the alternative model at the null value: $p(D|H_{null}) = p(D|H_{alt}, \delta = 0)$. Therefore, the Bayes factor can be determined by considering the posterior and prior of the alternative hypothesis alone, because the Bayes factor is just the ratio of the probability density at $\delta = 0$ in the posterior relative to the probability density at $\delta = 0$ in the prior: $BF = p(\delta = 0|H_{alt}, D)/p(\delta = 0|H_{alt})$. The posterior density $p(\delta = 0|H_{alt}, D)$ is approximated by fitting a smooth function to the MCMC sample, and the prior density $p(\delta = 0|H_{alt})$ is computed from the mathematical specification of the prior.

The SD method can be intuitively related to the ROPE in parameter estimation. Suppose we have a ROPE on the difference of means, perhaps from -0.1 to 0.1 as in Figure 5. The Bayes factor can be thought of as the ratio of (a) the proportion of the posterior within the ROPE relative to (b) the proportion of the prior within the ROPE. This ratio is essentially what the SD method computes when the ROPE is infinitesimally narrow. As is shown by example later, the proportion of the parameter distribution inside the ROPE may increase by a substantial factor but still leave the posterior proportion inside the ROPE very small.

Wetzels et al. (2009) showed that their approach closely mimicked the analytical results of Rouder et al. (2009) when the model was restricted to have equal variances in the two groups. The approach of Wetzels et al. is more general, of course, because the model allows different standard deviations in the two groups. Wetzels et al. also explored applications in which there is a directional hypothesis regarding the group means, expressed as a prior on the effect size that puts zero prior probability on negative effect sizes.

In principle, the implementation by Wetzels et al. (2009) could be modified to use t distributions to describe the groups instead of normal distributions. This would make the model similar to the one used in BEST except for the choice of prior distribution and reparameterization in terms of effect size. Convergence of BEST and the approach of Wetzels et al. could also be achieved by

(*Appendices continue*)

implementing the Savage–Dickey BF in BEST. The model-comparison and parameter-estimation approaches can also be combined as different levels in a hierarchical model (Kruschke, 2011a, 2011b, Chapter 12). Although the two approaches can be merged, there is a crucial difference in emphasis: The model-comparison approach emphasizes the Bayes factor, whereas the parameter-estimation approach emphasizes the explicit posterior distribution on the parameter values.

Examples of Applying the Bayes Factor Method

Consider the data of Figure 3, which involved fictitious IQ scores of a "smart drug" group and a placebo group. Recall that the parameter-estimation approach (using t distributions to describe the data) revealed a credible non-zero difference between means, a credible non-zero difference between standard deviations, and explicit distributions on the effect size and all parameters, as well as a depiction of model distributions superimposed on the data (i.e., a posterior predictive check).

The model-comparison approaches summarized above (Rouder et al., 2009; Wetzels et al., 2009) should not be applied to these data because they are not normally distributed. Of course, we can only visually guess the non-normality without a model of it, so we might apply the model-comparison methods anyway and see what conclusions they produce. The analytical BF method of Rouder et al. (2009) yields a BF of 1.82 in favor of the null hypothesis regarding the difference of means. The SD/MCMC method of Wetzels et al. (2009) yields a BF of 2.20 in favor of the null hypothesis. These BFs are not substantial evidence in favor of the null, but notice they lean the opposite way of the conclusion from the parameter estimation in Figure 3. The reason for the opposite-leaning conclusion is that the outliers in the data can be accommodated in the models only by large values of the standard deviation and, hence, small values of effect size. But notice that the BF by itself reveals no information about the magnitude of the difference between means, nor about the standard deviation(s), nor whether the data respect the assumptions of the model.

A problem with the Bayes-factor approach to null hypothesis assessment is that the null hypothesis can be strongly preferred even with very few data and very large uncertainty in the estimate of the difference of means. For example, consider two groups with $N_1 = N_2 = 9$, with data randomly sampled from normal distributions and scaled so that the sample means are 0 and the standard deviations are 1. The results of Bayesian parameter estimation in Figure D1 show that the most credible difference of means is essentially 0 and the most credible difference of standard deviations is essentially 0, but the explicit posterior distribution also reveals huge uncertainty in the estimates of the differences because the sample size is small. The 95% HDI on the difference of means goes from -1.27 to 1.23 (with only 14% of the posterior falling in the ROPE extending from -0.1 to 0.1), the 95% HDI on the difference of standard deviations extends from -1.15 to 1.17 (with only 18% of the posterior falling in the ROPE extending from -0.1 to 0.1), and the 95% HDI on the effect size extends from -0.943 to 0.952. Thus, Bayesian estimation shows that the most credible difference is essentially zero, but there is huge uncertainty in the estimate (and only a small proportion of the posterior distribution within a ROPE). The analytical BF method of Rouder et al. (2009) yields a BF of 3.11 in favor of the null hypothesis regarding the difference of means. The SD/MCMC method of Wetzels et al. (2009) yields an even larger BF of 4.11 in favor of the null hypothesis that the effect size is zero. Thus, the model-comparison method concludes that there is substantial evidence ($BF > 3$) in favor of the null hypothesis. But this conclusion seems unwarranted when there is so much uncertainty in the parameter values as revealed by parameter estimation. The Bayes factor hides crucial information about parameter uncertainty.

Summary: Model Comparison Versus Parameter Estimation

In summary, the BF from model comparison by itself provides no information about the magnitudes of the parameters, such as the effect size. Only explicit posterior distributions from parameter estimation yield that information. The BF by itself can be misleading, for example in cases when the null hypothesis is favored despite huge uncertainty in the magnitude of the effect size.

Both the analytical and MCMC approaches to model comparison provide a BF only for the difference of means and no BF for the difference of standard deviations. In principle, the analytical or MCMC approach could be extended to produce a BF for the difference of standard deviations, but this awaits future development. By contrast, the parameter-estimation approach as implemented in BEST includes the posterior difference of standard deviations as a natural aspect of its output.

Neither the analytical nor the MCMC approach to model comparison provides any power analysis. In principle, the MCMC chain from the approach of Wetzels et al. (2009) could be used to compute power in the manner described in Figure 6. But power computations are already implemented in the BEST software.

The two model-comparison approaches summarized here used mathematically motivated "automatic" prior distributions that were meant to be relatively neutral and generically applicable. Wetzels et al. (2011) showed that the decisions from the automatic BF correlate with decisions from the NHST t test. Nevertheless, the BF can be highly sensitive to the choice of alternative-hypothesis prior distribution (e.g., Dienes, 2008, 2011; Kruschke, 2011a; Liu & Aitkin, 2008; Vanpaemel, 2010), even to such an extent that the BF can change from substantially favoring the null hypothesis to substantially favoring the alternative hypothesis or vice versa. For the model-comparison approach to be meaningful, therefore, the prior in the alternative hypothesis must be a meaningful representation of a viable alternative hypothesis, and this is not automatically true.

(*Appendices continue*)

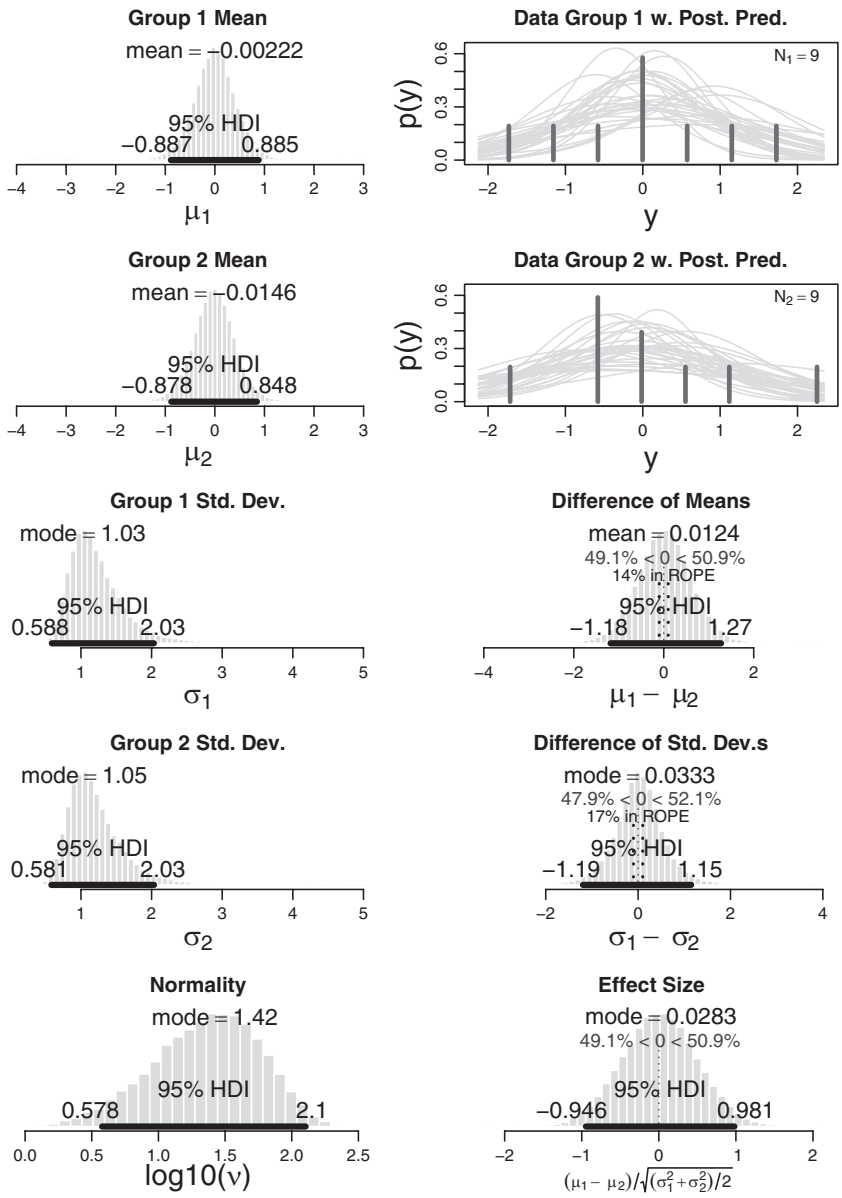

Figure D1. For small data sets, parameter estimation reveals huge uncertainty in the estimated difference of means, difference of standard deviations, and effect size. Model comparison, on the other hand, concludes that there is substantial evidence in favor of the null hypothesis despite this uncertainty, with $BF = 3.11$ when using the method of Rouder et al. (2009) and $BF = 4.11$ when using the method of Wetzels et al. (2009). BF = Bayes factor; HDI = highest density interval; w. = with; Post. Pred. = posterior predictive; Std. Dev. = standard deviation; ROPE = region of practical equivalence.

When the conclusion from model comparison, using the Bayes factor, differs from the conclusion from parameter estimation, using the relation of the HDI and ROPE, which should be used? The two approaches ask different questions, and there may be times when the model comparison is the specific answer being sought. But in general, parameter estimation yields richer information about the magnitudes of the meaningful parameters and their uncertainties, along with a natural method for computing power.

Received June 13, 2011
Revision received May 30, 2012
Accepted May 30, 2012 ■