Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement
2017, Vol. 49, No. 3, 166–182

0008-400X/17/$12.00 http://dx.doi.org/10.1037/cbs0000077

A First Steps Guide to the Transition From Null Hypothesis Significance Testing to More Accurate and Informative Bayesian Analyses

Brian P. O'Connor
University of British Columbia, Okanagan

This article begins with a brief summary of the problems with null hypothesis significance testing (NHST), followed by a short, nontechnical description of perhaps the most useful NHST alternative, Bayesian methods. Simple R commands and output for Bayesian correlations, regressions, and ANOVA are provided. This is followed by examples of how to describe Bayesian analyses in the Methods and Results sections of articles. The focus is on taking the cautious first steps in a transition away from NHST.

Keywords: significance testing, Bayesian statistics, null hypothesis

Supplemental materials: http://dx.doi.org/10.1037/cbs0000077.supp

Most researchers are by now aware that there are serious drawbacks with null hypothesis significance testing (NHST). The various problems have been well described in numerous books and journal articles (e.g., Bakan, 1966; Carver, 1978; Cohen, 1994; Harlow, Mulaik, & Steiger, 1997; Hunter, 1997; Kline, 2004; Nickerson, 2000; Oakes, 1986; Schmidt, 1996; Ziliak & McCloskey, 2008). But NHST is nevertheless still used in the vast majority of research reports in psychology. A perusal of recent issues from most journals and of modern statistics textbooks will quickly reveal how little has changed. The general awareness of problems with NHST and the many articulate descriptions of the problems have had little effect on research practices and on student education. Researchers are now more likely to report effect sizes and confidence intervals than they were in the past, but the *p* values from NHST are still the primary focus of most investigations.

Perhaps the descriptions of the problems with NHST and of the alternatives have been too long or technical. Researchers may also incorrectly believe that software for alternative methods is not available or is not easy to use. The purposes of this article are to provide a short and simple summary of the problems with NHST; to provide a short, nontechnical description of perhaps the most useful NHST alternative, Bayesian methods; and to provide simple R code and descriptions of how to begin using Bayesian methods. The target audience is readers who are new to Bayesian methods and the upcoming generation of psychology researchers.

Problems With NHST

A description of problems with NHST requires first making clear what *p* values do tell us.

> Conventionally, researchers make such decisions by assuming the null hypothesis to be true and, given this assumption, attempting to make inferences based on the probability of obtaining the actual pattern of results observed. Specifically, a statistical test yields the probability of a given result's (or one more extreme) being produced by chance if the null hypothesis is true. If D denotes an outcome or one more extreme and Ho denotes the null hypothesis's being true, then the probability produced by such a statistical test can be expressed as P(D|Ho), that is, the conditional probability of D, given Ho. If this figure is less than a threshold probability or alpha level (typically .05), then chance is concluded to be a sufficiently unlikely explanation of the outcome, and the existence of an effect is held to be supported by the data. (Pollard & Richardson, 1987, p. 159)

> [A] *p* value is the probability of data as extreme or more extreme as that obtained, computed under the presumption of the truth of the null hypothesis. (Maxwell & Delaney, 2004, pp. 47–48)

> [W]hat it tells us is "Given that H_o is true, what is the probability of these (or more extreme) data? (Cohen, 1994, p. 997)

> When we say that a difference is statistically significant at the .05 level, we mean that a difference that large would occur less than 5% of the time if the null were true. (Howell, 2013, p. 94)

> The calculations from a *t* test provide a *p* value, such as $p = .05$; it is a number which tells us the *proportion* of the time that we can expect to find mean differences as large as or larger than the particular sized difference we get when we are sampling from the same population assumed under the null hypothesis. (Carver, 1978, p. 382)

> *p* actually stands for the conditional probability . . . which represents the likelihood of a result or outcomes even more extreme assuming 1) the null hypothesis is exactly true; 2) the sampling method is random sampling; 3) all distributional requirements, such as normality and homoscedasticity, are met; 4) the scores are independent; 5) the scores are also perfectly reliable; and 6) there is no source of error besides sampling or measurement error. In addition to the specific observed result, *p* values reflect outcomes never observed and require many assumptions about those unobserved data. If any of these assumptions are untenable, *p* values may be inaccurate. (Kline, 2013, p. 74)

Correspondence concerning this article should be addressed to Brian P. O'Connor, Department of Psychology, University of British Columbia, Okanagan, IKBSAS, ART 330-1147 Research Road, Kelowna, British Columbia, Canada, V1V 1V7. E-mail: brian.oconnor@ubc.ca

> [A] p value is the conditional probability, given the value of the test statistic $T = t_0$ from the original data, and computed under the assumption that the null is true, that in an independent sample of the same size, a value of T as large or larger than t_0 will be obtained. (This definition was provided by a reviewer of this article.)

Clearly, a *p* value is derived from the distribution of test statistic values that occur when the null hypothesis is true. Imagine the following. Scores for two groups, for example, $N = 20$ for each, are randomly drawn from the same large population (pool) of scores and the *t* test value is computed. Now imagine doing this over and over, millions of times. The distribution of t values from the many comparisons would be the sampling distribution of t-values that occur when $N = 20$ per group and when the null hypothesis is true (remember that the two samples are always drawn from the same population). The formulas that are used in software packages to provide *p* values are mathematical ways of producing the same results. Note that values from a real dataset have not entered the picture at this point. That is, real data were not involved in the production of the sampling distribution of t-values.

Unfortunately, the information that a *p* value does provide is unsatisfying, minimally informative, and prone to misinterpretations. When we obtain a test statistic value that is beyond the 95th percentile, all we can say is that "when the null hypothesis is true, a test statistic value of this magnitude is unlikely on the basis of sampling variability alone." But the *p* value does not "know" anything about our data. It does not know from where our samples were obtained, i.e., it does not know whether or not the samples were drawn from the same population. It does not, and cannot, know if the null hypothesis is true for our dataset. We are left unsatisfied, and the misinterpretations then begin.

A *p* value is not the probability of committing a Type I error in any given study (a common misconception) because we typically do not know whether the null hypothesis is true or not. If we did have such knowledge, there would be no reason to conduct the research. If the null hypothesis is not true, then the probability of committing a Type I error is zero. It is not the *p* value.

The *p* value is not an indication of the proportion of Type I errors that are made across many studies in the literature (another misconception). For example, the claim that the widespread use of $p < .05$ implies that 5% of all published findings are Type I errors could only be true if researchers were always testing true null hypotheses, which is not a reasonable assumption. The .05 *p* value thus provides an unrealistic upper limit on the rate of false rejections of the null hypothesis (Pollard & Richardson, 1987).

The *p* value is not an indication that the findings will replicate. The *p* value is not an indication of effect size or practical importance. The *p* value is not the probability that our results are due to chance (e.g., "$p < .05$, so there is less than 5% probability that my findings are due to chance"). The *p* value is the probability of obtaining a particular test statistic value due to chance (sampling variability) when the null hypothesis is true. But it is not the probability that our particular research results are due to chance because we do not know if the null hypothesis is true in any given study and it (the *p* value) does not know either.

The *p* value is not a probability statement about the truthfulness of the null hypothesis (e.g., "$p < .05$, so there is only a 5% chance that the null hypothesis is true"). A *p* value is based on the assumption that the null hypothesis is true. It cannot be converted into a probability statement about the null hypothesis (Maxwell & Delaney, 2004, p. 48), although our brains seem naturally prone to making such conversions. For example, the probability of a population of firemen generating a person in uniform (high) is not the same as the probability that a sample person in uniform was generated by a population of firemen (low; Pollard & Richardson, 1987). Similarly, the probability of obtaining a particular test statistic value given the null hypothesis (what *p* values do tell us) is not the same as the probability of the null hypothesis, given that a particular test statistic value was obtained. The *p* value is not a probability value for the null hypothesis. NHST does not tell us what we think it does and what we usually most want to know (Masson, 2011): How probable is a hypothesis, given the collected data? We so much want to make such statements about our findings that we often just assume that *p* values provide such information (Cohen, 1994).

There is a tendency to believe that the *p* values are highly accurate. Yet they are typically based on assumptions about the data that are not met, or that are met to varying degrees, which makes them imprecise, mere approximations. NHST is based on the assumption that the data have been randomly sampled from the population, but this is almost never the case. We treat our *p* values as if there are highly accurate anyways.

In NHST there is no way of quantifying the degrees of evidence for competing models. NHST forces us to make an artificial, binary decision about whether or not there is an effect. Most researchers are probably aware that a failure to reach statistical significance does not mean that the null hypothesis is true. But nonsignificant effects are routinely treated as evidence of there being "no effect" (Armstrong, 2007; Falk & Greenbaum, 1995; Oakes, 1986). Discussion sections commonly involve speculations about why there was no effect for one or more variables or for one or more groups of participants. Nonsignificant effects are treated as evidence for the null hypothesis. Ultimately pointless searches for moderator variables, to account for a mixture of significant and nonsignificant effects, may be recommended solely on the basis of *p* values. Failures to replicate and growing piles of apparently conflicting findings may be caused solely by natural sampling variability and by the use of NHST to evaluate raw data (Schmidt, 1996).

Most of our studies involve relatively small samples and thus have modest statistical power. A small-sample study that just happens to provide an accurate estimate of a negligible population effect size is not likely to make it into the peer-reviewed literature because the *p* value will likely be $> .05$, due to the small effect size and low statistical power. However, another study of the same variables and using the same design and the same sample size, that provides an overestimate of the true population effect size on the basis of sampling variability alone, is more likely to make it into the peer-reviewed literature. The pool of published studies ends up overestimating the true effect sizes, simply because of sampling variability, small samples, and NHST. If the variable relationship in question is sufficiently appealing, the research report may be described in an undergraduate textbook and may thus become part of the discipline's popular knowledge base. Reports of lower but accurate effect sizes, which are likely to have nonsignificant *p* values due to our relatively small samples, are required to correct a Type I error, but such findings rarely get published. Rubble is generated and there is a lack of cumulative progress.

Conventional confidence intervals are more informative than *p* values. But they are derived from the same NHST framework, they have the same problems, and they are prone to misinterpretations. A 95% confidence interval indicates that if the study were conducted many times, 95% of the confidence intervals would contain the true population effect size. But this is statement is minimally informative and unsatisfying. There is instead a tendency to claim that a 95% confidence interval means that there is a 95% chance that the confidence interval contains the true population effect size. This interpretation is correct when the intervals are provided by Bayesian analyses (credibility, density, or probability intervals), but the interpretation is incorrect and unwarranted when the analyses are based on NHST.

A variety of data randomization procedures are now readily available and are occasionally used in research reports. Examples include randomization tests, the jackknife, and bootstrapping. These methods can be used to provide *p* values that are more accurate than those provided by the conventional formulas because they are not based on assumptions that our data rarely meet (Edgington, 1995; Efron & Tibshirani, 1993). But the additional *p* value precision does not bypass the many NHST problems because these methods are from the same NHST framework and they have the same interpretational shortcomings.

For over 25 years now there has been a strange, unfortunate dichotomy in how data are evaluated in our discipline. Reviews of the literature, where the degree of support for a hypothesis is more definitively determined, are usually based on meta-analyses. The focus is typically on effect sizes and on the estimate of the population effect sizes. The *p* values and the conclusions based on *p* values in individual studies are often ignored in meta-analyses. (When *p* values are involved, it is usually to assist in the estimation effect sizes that were not provided in original research reports, although see Owen, 2009, for a noteworthy exception.) Yet we continue to make NHST the primary focus when evaluating raw data in our research reports (Schmidt, 1996). Our methods of evaluating raw data need to catch up and become more meaningful and informative.

The American Statistical Association (ASA) recently released a "Statement on Statistical Significance and P-Values" (Wasserstein & Lazar, 2016). Included in the statement are the following declarations:

> While the *p* value can be a useful statistical measure, it is commonly misused and misinterpreted.
>
> *p* values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
>
> Scientific conclusions and business or policy decisions should not be based only on whether a *p* value passes a specific threshold.
>
> A *p* value, or statistical significance, does not measure the size of an effect or the importance of a result.
>
> By itself, a *p* value does not provide a good measure of evidence regarding a model or hypothesis. (Wasserstein & Lazar, 2016, pp. 131–132)

After decades of obsession with NHST, it now seems to be a case of "Oh that, never mind."

Bayesian Analyses

The present description of Bayesian methods is deliberately short and simple. Excellent introductions and tutorials, including discussions of statistical theory and reasoning about evidence, have been provided by Kruschke (2015); Kruschke, Aguinis, and Joo (2012); Lee and Wagenmakers (2013); Wagenmakers, Morey, and Lee (2016), and Zyphur and Oswald (2015). The hope in keeping the present description brief is to give all readers just enough of an overview to be enticed to try Bayesian analyses, without losing anyone with formulas and technical details. It should also become clear that while the analytic methods involved are different from NHST, the parameter values and some portions of the statistical output can be very much the same as in NHST.

In NHST, a test statistic value for an observed dataset is compared to the idealized, mathematically derived distribution of test statistic values that occur when the null hypothesis is true and when all of the test statistic assumptions are met. There are no such comparisons in Bayesian analyses. There are certainly no comparisons with any imaginary, theoretical distribution of values for a hypothesis that is not of direct interest and whose values were not derived from the current data. Instead, Bayesian analyses use input from the researcher to generate a complete distribution for all of the parameters of interest, given the data and the researcher's specifications. The distribution is then carefully examined to reach conclusions about the most likely values for the parameters.

Bayesian analyses begin with the researcher providing an observed dataset along with information that defines the prior probability of the parameters to be estimated, for example, information about the distribution of mean differences, or about the distribution of a regression weight (the specification of prior probabilities is discussed further below). Imagine the prior distribution as a range (which could be a very wide range) of possible values that a parameter of interest could have. The prior distribution provides the initial uncertainties for the parameter values, given current knowledge. These possible values are treated as random or variable across the possible range. The observed data values are treated as fixed or constant. Bayesian methods then estimate the probabilities of all possible combinations of parameter values, given the data and given the constraints in the priors. The output is called a "posterior probability distribution" because it contains the probabilities of the parameter values after observing the data (Zyphur & Oswald, 2015). "Bayesian inference yields a complete posterior distribution over the conjoint parameter space, which indicates the relative credibility of every possible combination of parameter values" (Kruschke, 2010, p. 295). One can also obtain from the posterior distribution information about the correlations of credible parameter values and the standard deviations of the values.

The continuous posterior distribution is sometimes called a "probability density function" and there is usually much focus on the "highest density interval" (HDI) within this function/distribution.

> Points inside an HDI have higher probability density (credibility) than points outside the HDI, and the points inside the 95% HDI include 95% of the distribution. Thus, the 95% HDI includes the most credible values of the parameter. The 95% HDI is useful both as a summary of the distribution and as a decision tool. Specifically, the 95% HDI can be used to help decide which parameter values should be deemed not credible, that is, rejected. . . . One simple decision rule is that any

> value outside the 95% HDI is rejected. In particular, if we want to decide whether the regression coefficients are nonzero, we consider whether zero is included in the 95% HDI. (Kruschke et al., 2012, p. 730)

The HDI (a credibility interval) is thus the range of parameter estimates that captures 95% (or 99%, or whatever percentage is preferred) of the posterior probability distribution. Statements such as, "There is a 95% chance that the parameter value falls between ___ and ___", are possible that is, are meaningful and warranted. NHST provides no such information about the probability of a parameter. A kind of NHST conclusion about the data can nevertheless be derived. The null can be rejected as improbable if the 95% HDI does not contain a null value.

Upon examining the HDI, the mean, median or mode of the distribution may be selected as the best estimate of a parameter. The mode will often be the more logical choice because the width of an HDI is an indication of the accuracy of the estimate. Posterior distributions that are more peaked provide better estimates, that is, have more narrow credibility intervals. Larger sample sizes usually result in more peaked posterior distributions.

Bayesian methods typically result in a change in beliefs about a coefficient from before to after the data collection and analyses. If a broad, uninformative prior was used, the change in beliefs could be as follows: "Before analysing the data I had no idea what values the coefficient might take. I conservatively assumed that the coefficient was zero. Now that I have run the analyses on my data, I believe that the likely value for the coefficient is between ___ and ___, and the most likely value is ___." Beliefs shift in the direction of the evidence.

The Prior Distribution

There are two broad kinds of prior distributions: (a) informative priors, from previous empirical findings or from theoretical predictions, and (b) diffuse, noncommittal, noninformative priors. Informative priors (e.g., an effect size from a previous study or from a meta-analysis) are not used to skew the results in one's preferred direction. Bayesians consider it inappropriate (even foolish) not to use informed or empirical priors if such information is available. It would be a missed opportunity that could bias the results (Efron, 2013). Not using available information would be analogous to police detectives refusing to jointly consider all of the available clues when evaluating any single piece of information about a crime. Informed or empirical priors assist in the development of cumulative knowledge. They are especially helpful in making firmer inferences possible and avoiding Type II errors in small sample research.

Uninformative priors are commonly used when researchers want their priors to have no influence on the posterior distribution. An example of an uninformative prior would be specifying that the range of possible values for a correlation coefficient in a particular study is -1 to $+1$, even though a meta-analysis might suggest a narrower range. Here is a good description of the common reasoning:

> When a study is exploratory, there may be little to no prior knowledge that can be used for estimation. Similarly, prior knowledge may be diffuse because of contradictory findings or competing theories, leading to prior distributions that are also diffuse. Alternatively, researchers may decide to eliminate the importance of priors in the estimation process to rely as much as possible on the likelihood (i.e., rely primarily on the data). In these cases, it is common to specify prior probabilities that allow the data (the likelihood) to dominate, such as by specifying a uniform (flat) prior distribution. This distribution specifies that, before any data are collected, no parameter values are more probable than others. A distribution that serves a similar purpose is a prior distribution with a huge variance, such as a normal distribution for an effect β with a mean $\mu_\beta = 0$ and variance $\sigma^2_\beta = 10^{10}$. This variance makes the prior probability distribution of the parameter values nearly flat, which is the default setting in some statistics programs (see Asparouhov & Muthén, 2010b; B. Muthén, 2010). Such distributions are so uninformative that they allow the data to dominate the estimation of posteriors through the likelihood (Gill, 2008; Kass & Wasserman, 1996). . . . Using uninformative priors in this manner has a long history (e.g., Bayes, 1763; Keynes, 1921/2008; Laplace, 1825/1995). (Zyphur & Oswald, 2015, p. 398)

The expectation is that uninformative priors will result in unbiased parameter estimates that mimic those from traditional analytic methods while permitting intuitive Bayesian statements to be made about the probabilities of the parameters.

Emerging evidence, however, indicates that uninformative priors can sometimes have unintended influences on the results and produce misleading findings (Baskurt & Evans, 2013; Seaman, Seaman, & Stamey, 2012). The use of diffuse priors is apparently not always so harmless after all. The potential problems are more likely to occur in smaller data sets and when effect sizes are weak, or when there are very large prior estimates for variances. The checking methods described by Kruschke (2015), Seaman et al. (2012), and Kamary and Robert (2014) can be used to assess the influence of one's priors on the posterior distribution. "When there is contention about the prior, it can be most convincing simply to conduct the analysis with different priors and demonstrate that the essential conclusions from the posterior do not change" (Kruschke, 2015, p. 724). Ghosh (2011) provided a review of approaches to specifying uninformative priors for different types of data distributions.

The process of selecting informed priors for the analysis, possibly on the basis of conflicting previous research or researcher objectives, is called "elicitation" (see Bocker, 2011, for an overview). An R package named BayesPref for elicitation is available from https://github.com/jlepird/BayesPref. Efron (2013) stated that he recently completed his term as editor of an applied statistics journal, that approximately one quarter of the papers used Bayesian methods, and that "almost all of these were based on uninformative priors" (p. 1178). Efron was concerned that the failure to use informed priors may eventually "bust" the current wave of enthusiasm for Bayesian methods. Researchers and journal editorial boards need to become comfortable with incorporating existing knowledge into data analyses, which will likely be no small leap for those who are more comfortable with NHST.

The MCMC

Although Bayes's rule for estimating posterior probabilities was devised in 1763, Bayesian analyses for most applications could not be implemented until recently due to the lack of computers and mathematical algorithms. The Markov Chain Monte Carlo (MCMC) is the modern computational breakthrough that has made all sorts of Bayesian analyses possible.

MCMC experts tend to explain the procedure using pages of formulas. The following two extended quotes are the best simple language descriptions that could be found. They together provide a good bird's eye view of what is involved. The first is from a statistics blog:

> Imagine you want to find a better strategy to beat your friends at the board game Monopoly. Simplify the stuff that matters in the game to the question: which properties do people land on most? The answer depends on the structure of the board, the rules of the game and the throws of two dice. One way to answer the question is this. Just follow a single piece around the board as you throw the dice and follow the rules. Count how many times you land on each property. Eventually, you will build up a good picture of which properties get the most business. This should help you win more often. What you have done is a MCMC analysis. The board defines the rules. Where you land next only depends on where you are now, not where you have been before and the specific probabilities are determined by the distribution of throws of two dice. MCMC is the application of this idea to mathematical or physical systems like what tomorrow's weather will be or where a pollen grain being randomly buffeted by gas molecules will end up. (matt_black, 2011, https://stats.stackexchange.com/questions/165/how-would-you-explain-markov-chain-monte-carlo-mcmc-to-a-layperson)

The next quote is an example of the common tendency in the Bayesian literature to describe the MCMC process as a directed random walk:

> A Markov chain can be thought of as a directed random walk through the parameter space that describes all the possible values of the parameter of interest. A directed random walk implies that although the next value drawn in the Markov chain is random, some values are more likely to be drawn than others; a well-constructed Markov chain will sample from these more likely regions of the sample space. The sampling of parameter values proportionally to their probability (which is determined by the information in the data and, in the case of a Bayesian analysis, by any informative prior information the user provides) allows the user to reconstruct the parameter's entire distribution. (Hamra, MacLehose, & Richardson, 2013, p. 628)

Although the parameter space may potentially be very large, especially when noninformative priors are used, the chain spends more of its time "walking around" the more important regions. The samples that are drawn mimic samples from the true posterior distribution. The end result is analogous to a topological map or a heat map of high and low density regions. Another good, but longer description of the MCMC procedure was provided by van de Schoot et al. (2014, supplementary materials). A program named MCMCRobot provides graphical illustrations of the MCMC process (Lewis, 2001). It is available for free for Windows computers and for iOS devices (iPhone and iPad) through iTunes. (True!)

Model-Checking

There are three general methods of checking the results from Bayesian analyses. One method focuses on the MCMC chains. Graphic analyses can reveal whether the MCMC sampling has gone awry or become stuck (a more elaborate description will be provided in the section on illustrative data analyses below). The MCMC can also be run multiple times. One hopes to find that the differences between the chains are equivalent to the average differences within chains. A second method is to rerun the analyses using different priors, looking convergent evidence in the findings. A third method, called a posterior predictive check, is to generate simulated data using the selected, credible parameters (e.g., the regression equation) from the posterior distribution. A model fits when the generated, predicted data greatly resembles or mimics the actual data. It is also possible for a model to fit well and for the results to conflict with those of previous research (results that could have been used for informed priors). See the above references on Bayesian methods, and Gelman et al. (2013), for more extended discussions of model-checking and for dealing with conflicts with prior data.

Some Benefits of Bayesian Analyses

Bayesian methods are not restricted by sample size. They do not require equal numbers of cases per groups. There are no computational problems when group sizes are unequal, and they do not require decisions about which sums of squares to use when group sizes are unequal. Bayesian methods are less adversely affected by missing data, and Bayesian estimation is less affected by subjective biases than is multiple imputation (Yalch, 2016). Bayesian analyses do not break down when predictors are strongly correlated. Bayesian methods do not generate prediction errors in moderated regression when group sizes are unequal. When the data have outliers or non-normal distributions, one merely specifies that the analyses be conducted using an appropriate non-normal distribution. In Bayesian chi-square tests there is no requirement that the expected cell frequency values be greater than or equal to five. See Kruschke (2015) for more details.

Our natural but incorrect interpretation of NHST confidence intervals is correct for Bayesian credibility (i.e., density or probability) intervals. A Bayesian 95% credibility interval provides the 95% most credible values for a parameter. Bayesian credibility intervals also contain distributional information about the parameter values, whereas NHST confidence intervals do not.

In Bayesian methods there is potentially less concern with correcting one's decision about a parameter, or about a group comparison, on the basis of the total number of parameters that are estimated or on the basis of the total number of group comparisons that are conducted. Two broad perspectives on this issue can be found in the Bayesian literature. One view is that stringency adjustments do not have to be made when a researcher decides to conduct further analyses. Bayesian results are determined only by the data and by the structure of the model and not by the researcher's explicit or implicit intentions to conduct a specific number of tests (Kruschke et al., 2012, p. 744). When multiple tests are conducted, NHST confidence intervals grow wider, while the Bayesian credibility intervals remain unchanged because the multidimensional posterior distribution remains unchanged when it is examined from different perspectives. A second view is that the multiple comparison issue is an open question that requires investigation and that it is potentially resolvable. Gelman, Hill, and Yajima (2012) described how a Bayesian multilevel model approach can cause multiple comparison problems to "disappear entirely." However, the benefits occur when well-informed priors are used but not necessarily when broad priors are used (see http://andrewgelman.com/2016/08/22/bayesian-inference-completely-solves-the-multiple-comparisons-problem/). There is also a small, separate literature on Bayesian approaches to NHST multiple comparisons (e.g., Berry & Hochberg, 1999).

Bayesian methods can be used to make statements about the likelihood of the null hypothesis. The strengths of the evidence for the null and alternative hypotheses can be quantified and compared. The "Bayes factor" is most commonly used for this purpose. It is the probability of the data under one hypothesis compared to the probability of the data under the other hypothesis (see Wetzels et al., 2011, for an overview). The Bayes factor is interpreted as an odds ratio. For example, a Bayes factor of 3 for the alternative hypothesis indicates that the data are three times more likely to have occurred under the alternative than under the null hypothesis. Jeffreys (1961) provided conventions for comparing Bayes factor values to the conventional NHST interpretations of *p* values. Bayes factors above 3 or below 0.33 are considered "substantial." See Wetzels et al. (2011, Table 1, p. 293) for the full list of interpretive guidelines.

The null can be accepted, and not merely rejected, depending on the data (Kruschke, 2011). The focus is on determining whether the null value is a credible value in the posterior distribution:

> This . . . involves establishing a region of practical equivalence (ROPE) around the value of interest. For example, if we are interested in the null value (i.e., zero) for a particular regression coefficient, we establish slope values that are equivalent to zero for practical purposes in the particular application. Suppose we specify that slopes between −0.05 and 0.05 are practically equivalent to zero. We would decide to reject the null value if the 95% HDI falls completely outside the ROPE, because none of the 95% most credible values is practically equivalent to the null value. Moreover, we would decide to accept the null value if the 95% HDI falls completely inside the ROPE, because all of the 95% most credible values are practically equivalent to the null value. The 95% HDI gets narrower as the sample size gets larger. (Kruschke et al., 2012, pp. 730–731)

A decision would be withheld when the ROPE does not completely contain the HDI because the data are not sufficient for a decision to be made in this case. See Kruschke (2015, Chapter 12) for an extended discussion of the reasoning and possibilities.

In NHST, the probability that the obtained results will be replicated is unknowable because NHST does not provide a distribution of credible parameter values. In Bayesian analyses, both power and replication probability can be computed using the posterior distribution. The methods work for any model regardless of complexity or distributional assumptions (Kruschke et al., 2012).

R Code and Illustrative Data Analyses

R code for Bayesian correlations, regressions, and a one-way ANOVA are provided in the Appendix, available online as supplemental material. Due to space constraints, only the results of the correlation and regression analyses are provided in the Appendix. For each procedure, commands are provided that install the required R packages, that generate artificial data for the analyses, that run the conventional (non-Bayesian) analyses on the data, and that run Bayesian analyses on the data. Note that the commands for running the Bayesian analyses are very brief in every case. Copy and paste the commands into the R console, and the analyses will run without any further input from the user. (Sometimes copying and pasting from a pdf can change some characters and add new ones. An R file with the commands was uploaded as a supplementary material file when the manuscript was submitted to CJBS. The file, along with complete program output, is available from https://people.ok.ubc.ca/brioconn/bayes/bayes.html) To run the analyses on your own data, simply read in a data file and substitute your variable names into the commands.

The default priors (diffuse/uninformative) were used for all three illustrative data analyses. The default priors do not have to be specified in the command syntax. This makes running the analyses seem very similar to conventional NHST analyses, with no complicated, uncertain extra steps. The supplementary material file contains commands for running the analyses using different and quite varying priors, for the purpose of assessing the stability of the results.

The MCMCglmm package was used for the Bayesian analyses because the installation and commands are so simple and brief. The R programs provided by Kruschke (2015) require just a bit more effort to install and they require minor changes to his example files (e.g., to read in your own data and to specify your design). But they provide more design options, more detailed output, and there are more possibilities for examining the posterior distribution and for running additional comparisons. It is also easier to specify informed priors and the supporting documentation is more user friendly.

Some aspects of the MCMC sampling process are controlled by the user. For example, note the final three arguments on the MCMCglmm command for the regression analyses that involve predicting variable "dv" from variables "A" and "B":

```
MCMCglmm(dv ~  A + B, data = dataset1, family
= 'gaussian', nitt = 13000, burnin = 3000, thin = 10)
```

The *nitt* argument instructs the procedure to take 13,000 samples. The *burnin* argument defines the burn-in period. The initial samples generated by (random) starting points are less trustworthy and could distort the results. The *burnin* value of 3,000 instructs the procedure to ignore the first 3,000 values that are produced when constructing the posterior distribution. The *thin* argument is used to ensure that the successive values that are used in the resulting output are clearly distinct. In other words, values on the *thin* argument are adjusted to reduce "autocorrelation." The thin value of 10 instructs the procedure to keep every 10th value. The end result of these particular settings is 1,000 sample values for each parameter. The above values for the three arguments are the default values.

The *family = 'Gaussian'* argument on the MCMCglmm command instructs the procedure to conduct the regression assuming that the dependent variable has a normal distribution. When the dependent variable is dichotomous, as in alive versus dead, the *family* argument could be changed to '*categorical*' in order to conduct a logistic regression. When the dependent variable is a count, as in number of criminal offences, the *family* argument could be changed to '*poisson*' for a Poisson regression.

Bayesian analyses typically provide graphic diagnostic information about the MCMC sampling process. The parameter estimates that are drawn across the MCMC iterations are plotted to determine whether or not the generated samples are providing us with a trustworthy posterior distribution. The Trace plots in the left column of Figure 1 are the regression parameter estimates for

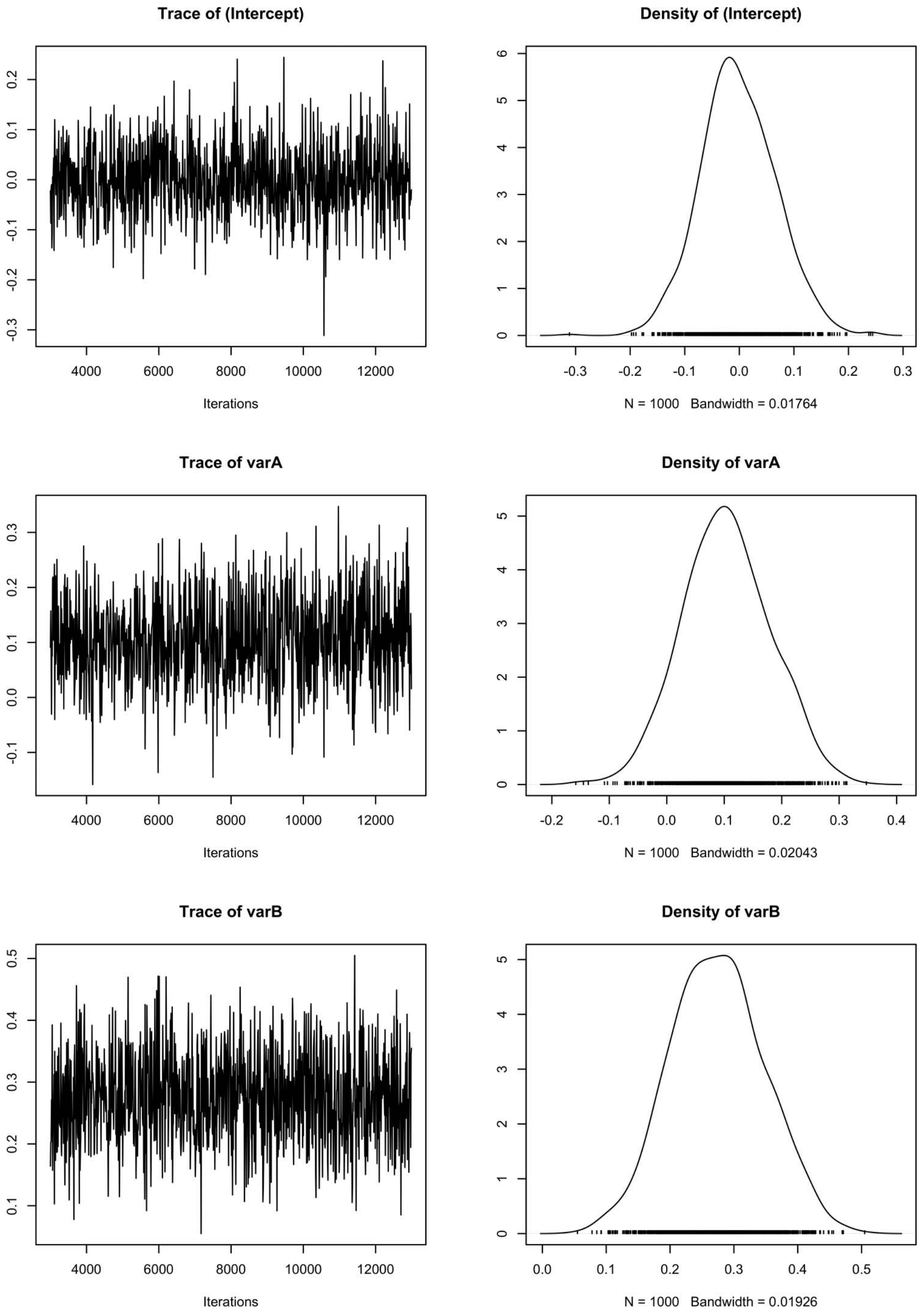

Figure 1. Trace & Density plots for the regression data.

iterations 3001 to 12,991, ignoring the first 3,000 burn-in samples. The series of values is called a "chain". What we want to see is consistent variation around a minimally changing average value. One sign of autocorrelation would be if the chain had sections with more distinct, snake-like, upward or downward bends, indicating insufficient "stationarity." This is not happening in the Trace plots in Figure 1, which are considered stable and desirable, despite the occasional blips. Another indication that autocorrelation is not a problem is the autocorrelation coefficients that are provided for various lags by the program, which are clearly low (see the regression output in the Appendix). The indications are that the posterior would likely not be changed by increasing the number of iterations. Had there been a problem with autocorrelation or signs of a poorly formed chain, the number of iterations and the thinning interval could be increased, because autocorrelation decreases over time. The Density plots in Figure 1 are the posterior distributions of the parameter values across the MCMC sampling process.

The Bayesian ANOVA using MCMCglmm provides the regression coefficients for the dummy-coded group variable. The intercept (41.01) is the mean of Group1. The coefficient for GroupGroup2 (18.56) represents the mean difference between Group1 and Group2. The coefficient for GroupGroup3 (2.75) represents the mean difference between Group1 and Group3. The 95% HDI credibility interval for the Group 1 versus 2 comparison does not include zero (38.21 to 44.11), whereas the interval for the Group 1 versus 3 comparison does include zero (−2.16 to 7.67). The Bayes factors for the full model, for the Group 1 versus 2 comparison, and for the Group 2 versus 3 comparison were very high, indicating "extreme evidence" in favor of the alternative over the null hypothesis. However, the Bayes factor for the Group 1 versus 3 comparison was much lower, at 0.59, indicating weaker, "anecdotal evidence" in favor of the alternative hypothesis.

The output for the correlation analyses and for the regression analyses indicates that the results from the conventional NHST and Bayesian analyses were highly similar. The 95% credibility intervals from the HDIs were almost identical to the NHST confidence intervals. Statistical significance conclusions based on the 95% HDIs are identical to the NHST statistical significance conclusions. NHST and Bayesian analyses do not always yield such similar findings. They likely yielded highly similar results in these cases because of the use of clean random normal data. But the fact that NHST and Bayesian results can be highly similar does not justify sticking with NHST. All of the problems and interpretational difficulties with the NHST results remain, as do all of the benefits of the Bayesian methods.

How Bayesian Analyses Can Be Described in a Method Section

The following text provides a model of how Bayesian methods for the regression data (see the Appendix) can be described to readers in the Analytic Methods subsection of a Methods section of a journal article.

We conducted Bayesian analyses to obtain the most credible estimates of the parameter values and to assess the statistical significance of the regression coefficients. Conventional, parametric significance tests do not indicate the probability of Type I errors for a dataset, and they do not indicate whether obtained results are due to chance or whether obtained results will replicate (Pollard & Richardson, 1987; Schmidt, 1996). The results of Bayesian analyses are more credible (Gill, 2015; Kruschke, 2015). Bayesian methods use only the observed data and not *p* values from hypothetical unobserved distributions. The posterior distribution for a parameter (e.g., for a regression coefficient) is accurately approximated by a very large representative random sample of parameter values drawn from the posterior distribution. The large sample is used to determine the quantiles of the parameter distribution and its shape. One primary output from Bayesian analyses is a "highest density interval," or HDI. Points inside an HDI have higher probability density (credibility) than points outside. Points inside the HDI include 95% of the distribution and thus are the most credible values of the parameter. The 95% HDI can be used to decide which parameter values should be deemed not credible (e.g., any value outside the 95% HDI). For the present study, we considered any regression coefficient with a 95% HDI that did not include zero to be "statistically significant." We used the MCMCglmm package in R to run the Bayesian analyses. We used the default noncommittal priors, which potentially have minimal influence on the posterior. The Markov Chain Monte Carlo (MCMC) chains were initialized at maximum likelihood values and burned in for 3,000 steps. Ten thousand further iterations were conducted and a thinning interval of 10 was used, resulting in a total of 1,000 steps that were saved.

Three follow-up analyses were then conducted to assess the stability and adequacy of the posterior. First, the trace and density plots were examined for signs of autocorrelation. Second, the analyses were run again, five times, using dramatically varying values for the priors. Specifically, we used the following prior values for the regression weights: −1, −.5, 0, .5, and 1 (recall that the variables were standardized). In each case, we used the MCMCglmm default, noncommittal values for the degree of belief (variance) parameters. We were hoping for consistency in the findings across the different prior specifications. Third, we conducted a posterior predictive check. We generated 200 simulated data sets using the most credible parameters from the posterior distribution. Plots were then examined to determine whether the generated, predicted data resembled or mimicked the actual data.

How Findings From Bayesian Analyses Can Be Described in a Results Section

The following text provides a model of how Bayesian methods for the regression data (see the Appendix) can be described to readers in the Results section of a journal article.

The linear regression analysis was conducted using varA and varB as predictors of varDV. The trace and density plots are provided in Figure 1. There was very little autocorrelation in the well-mixed chains. The resulting MCMC sample was therefore considered highly representative of the underlying posterior distribution. The regression weight for varB was .278, with a 95% HDI that ranged from .128 to .417. The regression weight was therefore considered nonzero and statistically significant. In contrast, the regression weight for varA was .098, with a 95% HDI that ranged from −.045 to .247. The credibility interval includes zero, and so the parameter estimate for varA was considered not significant.

The regression weights and the credibility intervals were stable across the analyses using different priors. Variability in the findings occurred at only the second or third decimal places of the coefficient estimates across the analyses using different priors. Density plots of the posteriors for the varB regression weight are provided in Figure 2. The posterior predictive check analyses confirmed that predicted data from the posterior distribution greatly resembled the actual data.

Software for Conducting Bayesian Analyses

NHST would almost certainly not have become our predominant analytic method had computers and algorithms for Bayesian analyses been available in the early 1900s. Fisher and other developers of NHST were aware of the shortcomings. But there were no feasible alternatives at that point and they were proud of the formula-based sampling distributions for test statistics that they used instead. Efficient algorithms and software for Bayesian analyses are now readily available. There are numerous, free R packages for conducting Bayesian analyses on the R CRAN web site and on GitHub. R code, along

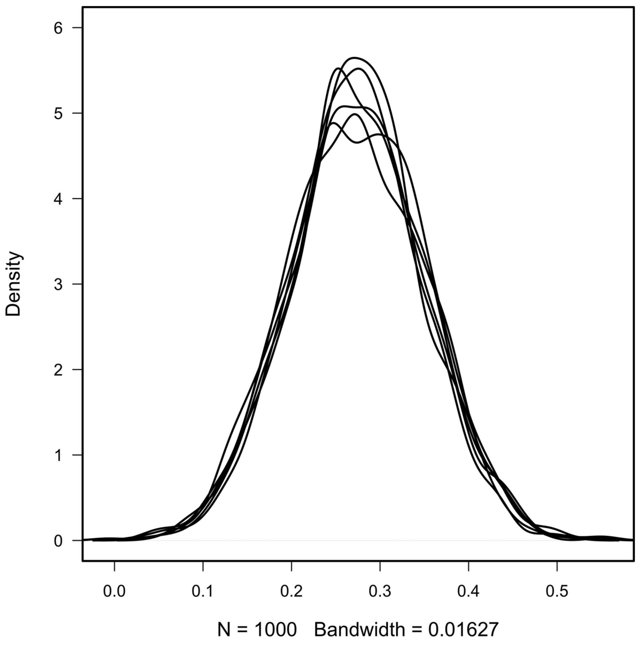

Figure 2. Density plots of posteriors for varB produced using differing priors.

with excellent tutorials, have been provided by Kruschke (2015) and by Lee and Wagenmakers (2013). There is an online calculator that provides Bayes factors for entered data (e.g., http://pcl.missouri.edu/bayesfactor). There are also MPlus, Matlab, and some SAS routines for Bayesian analyses. Experienced Bayesians often use programs named BUGS, STAN, and JAGS.

Software availability is no longer an excuse. Imagine how Fisher and colleagues from his era would react if they were to now wake from the dead and discover that we are still using NHST despite each of us having tremendous computing power for much better alternatives. Here is a suggestion for newcomers that will hopefully not be offensive to both Bayesians and NHST devotees. Use your familiar, non-Bayesian (e.g., SPSS) software to explore a dataset and to generate parameter estimates. Then switch to Bayesian software and output to reach more informative and definitive conclusions before writing your report.

A Joint Bayesian-Frequentist Data Scenario That Could Facilitate Progress

Bayesian methods do not solve all statistical problems. Datasets can have idiosyncrasies that will present challenges to both Bayesian newcomers and veterans. Bayesian experts, like most statisticians, may disagree on the best approaches to a data analysis. Statisticians are typically averse to "mechanical" or "mindless" approaches to data analysis. Gigerenzer and Marewski (2015), for one, cautions against hoping for a universal approach to emerge. He claims that Bayesian methods should be viewed as but one tool in the statistics toolbox and that informed judgment is the key. These realities could make otherwise interested users hesitate. But reverting to familiar, mechanical NHST methods is not the solution.

There is a simple scenario that, should it occur with your dataset, might help you side-step some of the hurdles that might otherwise come your way. If the statistical coefficients from the Bayesian and from the traditional ("frequentist") NHST analyses are equivalent, if the Bayesian credibility intervals are essentially the same as the NHST confidence intervals, and if the decision regarding a significant effect is the same when using the Bayesian HDI and ROPE as it is under NHST, then doubts about the findings should be reduced. You will be able to say that the data were analyzed both ways and the findings were the same. Bayesian statements about the credibility of the coefficients would be permitted, while the kinds of problematic NHST statements described at the outset of this paper could be avoided.

This simple scenario occurred in all three of the illustrative data analyses that were described above. The scenario will certainly not always exist. But it would perhaps be the safest and easiest scenario for making one's first efforts in moving forward with findings from Bayesian analyses.

In empirical science, one can never be completely certain that a statistical model is correct. Our data and our judgments can be misleading, even when Bayesian methods are used and when the results converge with those from traditional methods. Convergent findings might nevertheless inspire more confidence than would be the case when a research report is based on significance testing alone.

Journal Editorial Policy Encouragement Would Be Helpful

Given that software is now readily available for Bayesian analyses, perhaps the biggest damper on the use of Bayesian and other non-NHST methods, apart from the lack of education and exposure, is the lack of explicit encouragement for such methods in journal editorial policies. Researchers will remain reluctant to learn and use non-NHST methods as long as they believe that it is not necessary to do so, and as long as they believe that their chances of getting published may be reduced if they do not use NHST. Simple statements in journal editorial policies could be very influential. A single sentence could suffice: "Authors are encouraged, but are not required, to submit research reports that involve rigorous, non-NHST methods of data analysis." This would make it clear to researchers that the editorial board will not be biased against their work if they do not use NHST methods. It might also tempt researchers to use alternative methods in order to make their work more state-of-the-art. A cautious reaction to this might be that there are presently no definitive, universally recognized alternatives to NHST and that statisticians do not always agree on which Bayesian method is best. So we should stay the course. Bayesian methods are far less problematic and are far less prone to misinterpretation than NHST. There is no justification for us to continue generating conflicting, confusing findings in our evaluations of raw data.

References

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23,* 321–327. http://dx.doi.org/10.1016/j.ijforecast.2007.03.004

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437. http://dx.doi.org/10.1037/h0020412

Baskurt, Z., & Evans, M. (2013). Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Analysis, 8,* 569–590. http://dx.doi.org/10.1214/13-BA824

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference, 82,* 215–227. http://dx.doi.org/10.1016/S0378-3758(99)00044-0

Bocker, K. (2011). *Bayesian methods and expert elicitation*. London, England: Risk Books.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399. http://dx.doi.org/10.17763/haer.48.3.t490261645281841

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Edgington, E. S. (1995). *Randomization tests*. New York, NY: Marcel-Dekker.

Efron, B. (2013). Bayes' theorem in the 21st century. *Science, 340*(7, June), 1177–1178.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall. http://dx.doi.org/10.1007/978-1-4899-4541-9

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75–98. http://dx.doi.org/10.1177/0959354395051004

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall-CRC.

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. http://dx.doi.org/10.1080/19345747.2011.618213

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statistical Science, 26,* 187–202. http://dx.doi.org/10.1214/10-STS338

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41,* 421–440. http://dx.doi.org/10.1177/0149206314547522

Gill, J. (2015). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: Chapman & Hall/CRC.

Hamra, G., MacLehose, R., & Richardson, D. (2013). Markov chain Monte Carlo: An introduction for epidemiologists. *International Journal of Epidemiology, 42,* 627–634. http://dx.doi.org/10.1093/ije/dyt043

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Howell, D. C. (2013). *Statistical methods for psychology*. Belmont, CA: Wadsworth, Cengage Learning.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3–7. http://dx.doi.org/10.1111/j.1467-9280.1997.tb00534.x

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

Kamary, K., & Robert, C. P. (2014). Reflecting about selecting noninformative priors. *Journal of Applied and Computational Mathematics, 3,* 1–7.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10693-000

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

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. http://dx.doi.org/10.1177/1745691611406925

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15,* 722–752. http://dx.doi.org/10.1177/1094428112457829

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139087759

Lewis, P. O. (2001). Phylogenetic systematics turns over a new leaf. *Trends in Ecology & Evolution, 16,* 30–37. http://dx.doi.org/10.1016/S0169-5347(00)02025-5

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. http://dx.doi.org/10.3758/s13428-010-0049-5

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

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301. http://dx.doi.org/10.1037/1082-989X.5.2.241

Oakes, M. L. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York, NY: Wiley.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *Annals of Statistics, 37*(6B), 3867–3892. http://dx.doi.org/10.1214/09-AOS697

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102,* 159–163. http://dx.doi.org/10.1037/0033-2909.102.1.159

Schmidt, F. L. (1996). Significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129. http://dx.doi.org/10.1037/1082-989X.1.2.115

Seaman, J., III, Seaman, J., Jr., & Stamey, J. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician, 66,* 77–84. http://dx.doi.org/10.1080/00031305.2012.695938

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development, 85,* 842–860. http://dx.doi.org/10.1111/cdev.12169

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25,* 169–176. http://dx.doi.org/10.1177/0963721416643289

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70,* 129–133. http://dx.doi.org/10.1080/00031305.2016.1154108

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

Yalch, M. M. (2016). Applying Bayesian statistics to the study of psychological trauma: A suggestion for future research. *Psychological Trauma: Theory, Research, Practice and Policy, 8,* 249–257. http://dx.doi.org/10.1037/tra0000096

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: University of Michigan Press.

Zyphur, M., & Oswald, F. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management, 41,* 390–420. http://dx.doi.org/10.1177/0149206313501200

(*Appendix follows*)

Appendix

R Commands and Output for Correlation, Regression, and ANOVA

R Commands for all three procedures:

```
# installing and loading packages
install.packages("MCMCglmm");    library(MCMCglmm)
install.packages("BayesFactor"); library(BayesFactor)


# generating scores for the Correlation and Regression analyses
# fix the random number generator seed in order to produce the
# same data on every run
set.seed(123)
N = 200
X = rnorm(N, mean = 0, sd = 1)
Y = rnorm(N, mean = 0, sd = 1)
Z = rnorm(N, mean = 0, sd = 1)
pc.cr = princomp( cbind(X,Y,Z) )
fscores = pc.cr$scores
# imposing correlations between the factor scores
hyp = '
 1  .4  .2
.4   1  .3
.2  .3   1'
rtarget = data.matrix( read.table(text=hyp))
dataset1 = fscores %*% chol(rtarget)
dataset1 = data.frame(scale(dataset1)) # standardizing
colnames(dataset1) = c('varA','varB','varDV')

round(cor(dataset1),2) # conventional Pearson correlations
```

(*Appendix continues*)

```
# CORRELATION

# conventional Pearson correlation & NHST for varB & varDV
cor.test(dataset1$varB,dataset1$varDV)

# correlation through MCMCglmm
# the variables are already standardized
model1 = MCMCglmm(varDV ~ varB, dat=dataset1,
                  nitt=13000, burnin=3000, thin=10)
summary(model1)
autocorr(model1$VCV)
plot(model1$Sol)


# REGRESSION

# conventional multiple linear regression with NHST
fit = lm(varDV ~ varA + varB, data=dataset1)
summary(fit)
confint(fit, level=0.95)
anova(fit)


# Bayesian REGRESSION using MCMCglmm
model2 = MCMCglmm(varDV ~ varA + varB, data=dataset1,
                  family='gaussian', nitt=13000, burnin=3000, thin=10)
summary(model2)
autocorr(model2$VCV)
plot(model2$Sol)
```

(*Appendix continues*)

```
# ANOVA

# generating data for oneway ANOVA, 3 groups
# fix the random number generator seed in order to produce the
# same data on every run
set.seed(123)
X = rnorm(20,40,7)
Y = rnorm(20,60,7)
Z = rnorm(20,43,7)
dataset2 = data.frame(c(X,Y,Z)); colnames(dataset2) = c('DV')
dataset2$Group = gl(3,20, labels = c("Group1", "Group2", "Group3"))


# descriptives & plot
tapply(dataset2$DV,dataset2$Group, summary) # display the group means
tapply(dataset2$DV,dataset2$Group, sd) # display the group SDs
plot(DV ~ Group, data=dataset2)


# conventional ANOVA & multiple comparisons
results = aov(DV ~ Group, data=dataset2)
summary(results)
pairwise.t.test(dataset2$DV,dataset2$Group, p.adjust='bonferroni')
TukeyHSD(results,conf.level=.95)


# Bayesian ANOVA using MCMCglmm
model3 = MCMCglmm(DV ~ Group, data=dataset2, verbose=F,
                  nitt=13000, burnin=3000, thin=10)
summary(model3)
autocorr(model3$VCV)
plot(model3$Sol)
```

(*Appendix continues*)

```
# Bayes factors
# Bayes factor for the model
anovaBF(DV ~ Group, iterations = 10000, data=dataset2)
# Bayes factor for Groups 1 vs 2
anovaBF(DV ~ Group, iterations = 10000,
        data=subset(dataset2, Group != 'Group3'))
# Bayes factor for Groups 1 vs 3
anovaBF(DV ~ Group, iterations = 10000,
        data=subset(dataset2, Group != 'Group2'))
# Bayes factor for Groups 2 vs 3
anovaBF(DV ~ Group, iterations = 10000,
        data=subset(dataset2, Group != 'Group1'))
```

Selected Output from the Above Correlation Commands:

```
 > # conventional Pearson correlation & NHST for varB & varDV
 > cor.test(dataset1$varB,dataset1$varDV)

 Pearson's product-moment correlation

data:  dataset1$varB and dataset1$varDV
t = 4.6897, df = 198, p-value = 5.091e-06

95 percent confidence interval:
 0.1855820 0.4358053

sample estimates:
     cor
0.316182
```

(*Appendix continues*)

```
> # correlation through MCMCglmm
> # the variables are already standardized
> model1 = MCMCglmm(varDV ~ varB, dat=dataset1)

> summary(model1)

 Iterations = 3001:12991
 Thinning interval  = 10
 Sample size  = 1000



             post.mean   l-95% CI   u-95% CI eff.samp  pMCMC
(Intercept)  0.0007881 -0.1459576  0.1214560    720.4  0.942
varB         0.3126429  0.1746920  0.4318655   1127.6 <0.001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> autocorr(model1$VCV)

              units
Lag 0    1.00000000
Lag 10   0.02728845
Lag 50   0.01037955
Lag 100  0.06381656
Lag 500 -0.03396188
```

Selected Output from the Above Regression Commands:

```
> # conventional multiple linear regression with NHST
> fit = lm(varDV ~ varA + varB, data=dataset1)
> summary(fit)
```

(*Appendix continues*)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.783e-18  6.709e-02   0.000 1.000000
varA        1.032e-01  7.393e-02   1.395 0.164487
varB        2.734e-01  7.393e-02   3.698 0.000282 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9488 on 197 degrees of freedom
Multiple R-squared:  0.1088,   Adjusted R-squared:  0.09973
F-statistic: 12.02 on 2 and 197 DF,  p-value: 1.185e-05

> confint(fit, level=0.95)
                  2.5 %    97.5 %
(Intercept) -0.13231085 0.1323109
varA        -0.04263971 0.2489561
varB         0.12756225 0.4191581

> anova(fit)
Analysis of Variance Table

Response: varDV
           Df  Sum Sq Mean Sq F value    Pr(>F)
varA        1   9.339  9.3390  10.373 0.0014955 **
varB        1  12.308 12.3080  13.671 0.0002822 ***
Residuals 197 177.353  0.9003
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(*Appendix continues*)

```
> # Bayesian REGRESSION using MCMCglmm
> model2 = MCMCglmm(varDV ~ varA + varB, data=dataset1,
 family='gaussian', nitt=13000, thin=10, burnin=3000)
> summary(model2)

 Iterations = 3001:12991
 Thinning interval  = 10
 Sample size  = 1000


            post.mean  l-95% CI  u-95% CI eff.samp  pMCMC
(Intercept) -0.001926 -0.124576  0.128453     1107  0.994
varA         0.098120 -0.044695  0.247153     1038  0.198
varB         0.277865  0.127949  0.416467     1000 <0.001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> autocorr(model2$VCV)

              units
Lag 0    1.00000000
Lag 10   0.01371913
Lag 50  -0.06528034
Lag 100  0.01686856
Lag 500  0.02007962
```

Received August 8, 2016
Revision received March 29, 2017
Accepted May 2, 2017 ■