Psychon Bull Rev (2018) 25:178–206
DOI 10.3758/s13423-016-1221-4

BRIEF REPORT

The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective

John K. Kruschke[1] · **Torrin M. Liddell**[1]

Published online: 7 February 2017

Abstract In the practice of data analysis, there is a conceptual distinction between hypothesis testing, on the one hand, and estimation with quantified uncertainty on the other. Among frequentists in psychology, a shift of emphasis from hypothesis testing to estimation has been dubbed "the New Statistics" (Cumming, 2014). A second conceptual distinction is between frequentist methods and Bayesian methods. Our main goal in this article is to explain how Bayesian methods achieve the goals of the New Statistics better than frequentist methods. The article reviews frequentist and Bayesian approaches to hypothesis testing and to estimation with confidence or credible intervals. The article also describes Bayesian approaches to meta-analysis, randomized controlled trials, and power analysis.

Keywords Null hypothesis significance testing · Bayesian inference · Bayes factor · Confidence interval · Credible interval · Highest density interval · Region of practical equivalence · Meta-analysis · Power analysis · Effect size · Randomized controlled trial · Equivalence testing

The New Statistics emphasizes a shift of emphasis away from null hypothesis significance testing (NHST) to "estimation based on effect sizes, confidence intervals, and meta-analysis" (Cumming, 2014, p. 7). There are many reasons to eschew NHST, with its seductive lapse to black-and-white thinking about the presence or absence of effects. There are also many reasons to promote instead a cumulative science that incrementally improves estimates of magnitudes and uncertainty. These reasons were recently highlighted in a prominent statement from the American Statistical Association (ASA; Wasserstein & Lazar, 2016) that will be summarized later in this article. Recent decades have also seen repeated calls to shift emphasis away from frequentist methods to Bayesian analysis (e.g., Lindley, 1975).

In this article, we review both of these recommended shifts of emphasis in the practice of data analysis, and we promote their convergence in Bayesian methods for estimation. The goals of the New Statistics are better achieved by Bayesian methods than by frequentist methods. In that sense, we recommend a *Bayesian* New Statistics. Within the domain of Bayesian methods, we have a more nuanced emphasis. Bayesian methods provide a coherent framework for hypothesis testing, so when null hypothesis testing is the crux of the research then Bayesian null hypothesis testing should be carefully used. But we also believe that typical analyses should not routinely stop with hypothesis testing alone. In that sense, we recommend a *New* Bayesian Statistics, that is, Bayesian analyses that also consider estimates of magnitudes and uncertainty, along with meta-analyses.

This article begins with an extensive description of frequentist and Bayesian approaches to null hypothesis testing and estimation with confidence or credible intervals. Subsequently, the article explains Bayesian approaches to meta-analysis, randomized controlled trials, and power analysis. We hope to demonstrate that Bayesian approaches to all these analyses are more direct, more intuitive, and more informative than frequentist approaches.

✉ John K. Kruschke
johnkruschke@gmail.com

[1] Indiana University, Bloomington, USA

Two conceptual distinctions in data analysis

We frame our exposition in the context of the two conceptual distinctions in data analysis that we mentioned earlier, and which are illustrated in Fig. 1. The rows of Fig. 1 mark the distinction between point-value hypothesis tests and estimation of magnitude with uncertainty. The columns of Fig. 1 indicate the distinction between frequentist and Bayesian analysis. We will review the two distinctions in the next sections, but we must first explain what all the distinctions refer to, namely, formal models of data.

Data are described by formal models

In all of the data analyses that we consider, the data are described with formal, mathematical models. The models have meaningful parameters. You can think of a mathematical model as a machine that generates random samples of data in a pattern that depends on the settings of its control knobs. For example, a shower head spews droplets of water (i.e., the data) in a pattern that depends on the angle of the shower head and the setting of the spray nozzle (i.e., the parameters). Different machines can make different patterns of data; for example a lawn sprinkler can make different patterns of water than a bathroom shower. In data analysis, we describe the actually-observed data in terms of a mathematical machine that has its parameters set to values that would generate simulated data that mimic the observed data. When we "fit" a model to data, we are figuring out the settings of the parameters (i.e., the control knobs) that would best mimic the observed data.

For example, suppose we measure the intelligence quotient (IQ) scores of a group of people. Suppose we make a histogram of the scores, and the histogram looks roughly unimodal and symmetric. Therefore we might choose to describe the data in terms of a normal distribution. The normal distribution has two parameters, namely its mean (denoted by Greek letter mu, μ) and its scale or standard deviation (denoted by Greek letter sigma, σ). These two parameters are the control knobs on the machine that generates normally distributed data according to the Gaussian formula. Suppose we find that when we set μ to 100 and we set σ to 15 then the machine generates data that closely mimic that actually observed data. Then we can meaningfully summarize the set of many data values with just two parameter values and the mathematical form of the model.

All of the approaches to data analysis that we consider in this article assume that the data are described by mathematical models with meaningful parameters. Mathematical models of data are profoundly useful for a variety of reasons. In particular, mathematical models are useful for *describing* data because people who are familiar with the behavior of the model can grasp the form of the data from only a few parameter values. Mathematical models are useful for *making inferences* about data because the formal logic of mathematics allows us to derive specific properties of parametric descriptions that would not be obvious from the data alone. For example, in a Bayesian framework we could derive the probability that the mean μ falls within the range 99 to 101, given the observed data.

The conceptual distinctions in Fig. 1 indicate different sorts of analyses for a given model of data. The distinctions in Fig. 1 apply to any particular model of data, regardless of its complexity. The model could have its parameter values examined by a hypothesis test, or the model could have its parameter values estimated with uncertainty (i.e., the row distinction). The model could be addressed using frequentist methods or with Bayesian methods (i.e., the column distinction). To understand why various people recommend emphasizing some approaches over others, it is important to understand the different information provided by the different analyses. Therefore we tour the four cells of Fig. 1, starting with the top-left cell.

	Frequentist	**Bayesian**
Hypothesis test	*p* value (null hypothesis significance test)	Bayes factor
Estimation with uncertainty	maximum likelihood estimate with confidence interval (The "New Statistics")	posterior distribution with highest density interval

Fig. 1 Two conceptual distinctions in the practice of data analysis. *Rows* show point-value hypothesis testing versus estimating magnitude with uncertainty. *Columns* show frequentist versus Bayesian methods. *Cells* indicate the typical information provided by each approach

Frequentist hypothesis test: Null hypothesis significance test (NHST)

The top-left cell of Fig. 1 corresponds to a frequentist hypothesis test. In this article we use "frequentist" loosely to refer to the constellation of methods founded on sampling distributions of imaginary data. Sampling distributions are at the core of p values and confidence intervals, as explained later. Bayesian analysis, by contrast, is not based on sampling distributions.

For example, a traditional t-test involves computing a p value, and if $p < .05$ then the null hypothesis is rejected. The test proceeds as follows. We collect some data and compute a summary statistic called t. For a single group of data, $t = (\bar{y} - \mu_0)/(s/\sqrt{N})$ where $\bar{y}$ is the sample mean, μ_0 is the null-hypothesis mean, s is the sample standard deviation,

and N is the sample size. For our purposes, the exact formula for the t statistic is not important; what matters is the idea that the t statistic summarizes the data. Frequentists want to know how likely it would be to get a value of t at least this extreme if the null hypothesis were true and if we were to collect data the same way we collect data in the actual research. When this probability is small enough, frequentists decide to reject the null hypothesis. Therefore, the probability is the rate of committing a false alarm (also known as a "Type I error"), and the goal of the decision threshold (usually set at 0.05) is to limit false alarms to that frequency.

Figure 2 illustrates the idea of a p value. The summary value of the observed data is indicated as the concrete block labeled "Actual Outcome." The analyst, illustrated as the face in the lower left, has a hypothesis in mind, such as a null hypothesis of zero effect. We consider repeatedly randomly sampling from the null hypothesis, every time generating a sample in the same way that the actual data were sampled, and for every simulated sample we compute a summary statistic like the one we computed for our actual sample. The resulting distribution of randomly generated summary values is illustrated as the cloud in Fig. 2. The center of the cloud is where most simulated samples fall, and the periphery of the cloud represents extreme values that occur more rarely by chance. Intuitively, if the actual outcome falls in the fringe of the cloud, then the actual outcome is not very likely to have occurred by the hypothesis, and we reject the hypothesis. The proportion of the cloud that is as extreme as or more extreme than the actual outcome is the p value, shown at the right of Fig. 2.

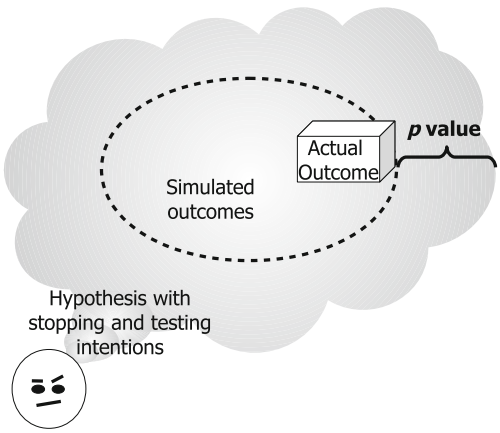

Fig. 2 *Definition of a p value.* A summary of the observed data is indicated as the concrete block labeled "Actual Outcome." The analyst, illustrated as the face in the lower left, has a hypothesis in mind, which generates a distribution of possible outcomes when simulated data are sampled according to the analyst's stopping and testing intentions. The sampling distribution is illustrated as the cloud. The proportion of the cloud that is as or more extreme than the actual outcome is the p value, shown on the *right*. Different stopping and testing intentions generate different clouds of possibilities, hence different p values

Formally, a p value can be defined as follows. For a set of actual data, let $T(D_{actual})$ be a descriptive summary value of the data, such as a t statistic. Suppose that the actual data were sampled according to certain stopping and testing intentions denoted I. Then the p value is defined as

$$p \text{ value} \equiv p\left(T(D_{simulated}) \succeq T(D_{actual}) \,\middle|\, \mu, I\right) \quad (1)$$

where $T(D_{simulated})$ are the descriptive summaries of simulated data sampled from a hypothetical population characterized by parameter value μ according to the same stopping and testing intentions, I, as the actual data. In Eq. 1, the relation "$\succeq$" means "is at least as extreme as, relative to the expected value of $T(D_{simulated})$." Usually, a p value refers to the null hypothesis, in which case the model-parameter μ is set to zero or some other value that represents no effect. When the resulting p value is used to decide whether or not to reject the null-hypothesis μ_0, the procedure is called a null-hypothesis significance test, abbreviated as NHST.

As a concrete example of NHST, we consider the simplest sort of data: dichotomous outcomes such as correct/wrong, agree/disagree, left/right, female/male, and so on. For example, we might be interested in knowing the probability that people agree with a particular policy statement. Suppose that of 18 randomly selected people, 14 agree with the statement. In the sample, the proportion of agreement is $14/18 \approx 0.778$, which apparently differs from the value 0.50 that represents the "null" value of ambivalence or equal preference. Formally, we denote the underlying probability of agreement as θ, and the null hypothesis as $\theta_{null} = 0.50$. The actual sample size is denoted $N = 18$, and the number who agree is denoted $z = 14$, as shown at the top of Fig. 3. (Please note that z refers to the number of people who agree, not to any sort of standardized score.)

To compute a p value for the data in Fig. 3, we first must declare the stopping and testing intentions. We make the conventional assumption that the stopping intention is to sample until $N = 18$ and that this is the only test of the data we intend to make. Then we create a sampling distribution of the summary statistic z/N (that is, we create the cloud of imaginary possibilities in Fig. 2). From this sampling distribution, we determine the probability that simulated z/N would be as or more extreme than the actual z/N. The resulting p value is indicated in the top-left cell of Fig. 3. (The displayed p value is the *two-tailed* p value, which considers the probability that simulated z/N would be more extreme than actual z/N in either direction away from θ_{null}, because either direction could reject the null hypothesis.) In this case, because p is less than the conventional threshold for tolerable false alarms (i.e., $p < .05$), the null hypothesis is rejected.

It is important to understand that the p value would be different if the stopping or testing intentions were different.

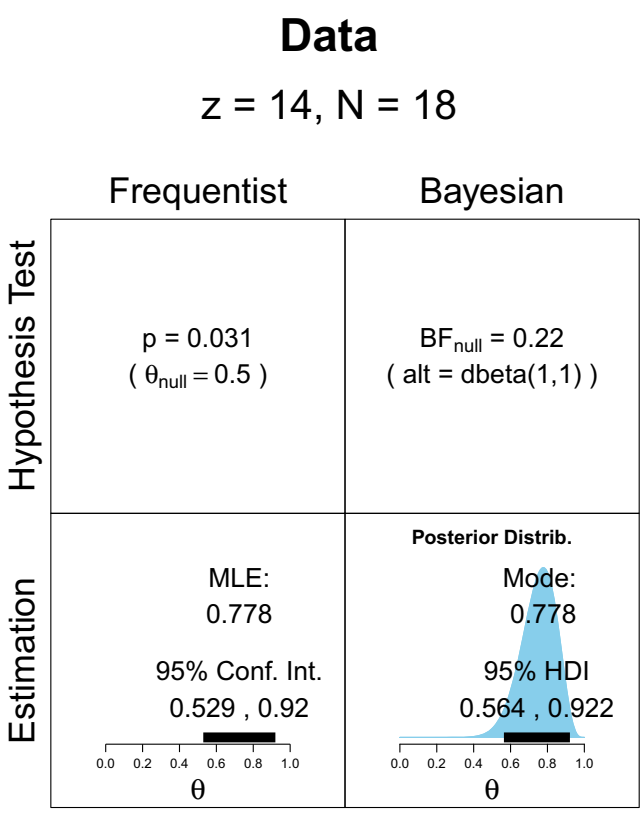

Fig. 3 Dichotomous data, with z occurrences in N trials shown at the *top*. The 2 × 2 table shows results from analyses corresponding to Fig. 1. Details of each cell are discussed in various sections of the article. (*BF* Bayes factor, *MLE* maximum likelihood estimate, *Conf. Int.* confidence interval. *HDI* highest density interval)

For example, if data collection were stopped because time ran out instead of because N reached 18, then the sample size would be a random number and the sampling distribution would be different, hence the p value would be different. When N is a random value, the sampling distribution is a probabilistic mixture of different fixed-N sampling distributions, and the resulting mixture is (in general) not the same as any one of the fixed-N distributions. Moreover, if a second survey question were being asked, then we would have to consider the probability that z/N from *either* question could be as or more extreme than the proportion observed in the first question, and hence the p value would be different. Corrections for multiple tests are discussed at the end of our tour through the four cells of Fig. 1. Complete numerical examples are provided in Kruschke (2015).

Notice that the only information delivered by NHST itself is the p value, from which the user decides whether or not to reject the hypothesized value of θ_{null}. This dearth of information can easily let the user slip into fallacious "black and white" thinking: believing either that the null is true and correct (which does not follow from $p > .05$) or that some meaningfully large effect is true and correct (which does not follow from $p < .05$). Moreover, the p value by itself indicates nothing about the estimated magnitude of the parameters, nor how uncertain the estimate is.

These issues and others are reviewed again later in the article, when we describe the reasons for a shift in emphasis from hypothesis testing to estimation. But for now we continue our expository tour through the four cells of Figs. 1 and 3, moving to the lower-left cell: frequentist estimation.

Frequentist estimation and confidence interval

In some situations we might be interested only in whether or not the null value is an acceptable description of the data. But in many and perhaps most situations, we would like to know the magnitude of the trend or effect, and to have some sense of how uncertain we are about the magnitude. For example, in the case of agreement with a policy statement, we might not be satisfied with only deciding whether or not the population is ambivalent, but we would want to know the magnitude of agreement and its range of uncertainty.

In frequentist methods, the best estimate of a parameter in a descriptive mathematical model is usually obtained as the *maximum likelihood estimate*, abbreviated as MLE. A special case of the MLE is the least-squares estimate (LSE), which may be familiar to readers who have previously studied analysis of variance (ANOVA) or linear regression. The MLE is the value of a parameter that makes the data most probable within the context of the descriptive mathematical model. For example, if we are describing dichotomous data with a simple model in which the probability of occurrence is denoted by the parameter θ, then the MLE of θ is z/N. As another example, if we are describing a set of metric numerical values (such as heights of people) by a mathematical normal distribution, then the MLE of the mean parameter μ is the arithmetic mean of the sample.

The uncertainty of the estimated parameter value is represented, in frequentist methods, by the *confidence interval*. The most general definition of a confidence interval, which applies to all models and situations, is the following (e.g., Cox, 2006, p. 40): The 95 % confidence interval of a parameter contains all the values of the parameter that would not be rejected by $p < .05$. NHST asks whether or not the null value of the parameter would be rejected. The confidence interval merely asks which other values of the parameter would not be rejected. Clearly the MLE of the parameter would not be rejected, but how far away from the MLE can we go before we reject the parameter value?

Formally, the frequentist 95 % confidence interval (CI) is the range of values for a parameter μ such that the corresponding p value (defined in Eq. 1) is greater than or equal to 0.05:

μ is in the 95% CI if and only if

$$p\left(T(D_{simulated}) \succeq T(D_{actual}) \,\middle|\, \mu, I\right) \geq 0.05 \qquad (2)$$

where $D_{simulated}$ are sampled from the hypothesized value of μ according to the same stopping and testing intentions I as the actual data.

An example of a CI is shown in the lower-left cell of Fig. 3. The horizontal black bar marks the 95 % CI, which is the range of parameter values that would not be rejected by $p < .05$. In other words, any value for the parameter θ *outside* the CI *would* be rejected by $p < .05$.

The CI reported in Fig. 3 was computed using the conventionally assumed intentions: data collection was stopped at a predetermined sample size N and there were no other tests being conducted. If there were a different stopping intention or a different testing intention, then the CI would be different. Because a confidence interval is defined in terms of p values, and p values depend on sampling and testing intentions, it follows that researchers with identical data but different stopping or testing intentions will have different confidence intervals. Just as any set of data has many different p values depending on the stopping and testing intentions, any set of data has many different confidence intervals. Complete numerical examples are provided in Kruschke (2015).

Confidence intervals have no distributional information Notice that a confidence interval has no distributional information. The limits of the confidence interval merely define the range of parameter values that would not be rejected by $p < 0.05$. There is no direct sense by which parameter values in the middle of the confidence interval are more probable than values at the ends of the confidence interval. This absence of distributional information is why the confidence interval in Fig. 3 is drawn as a flat line.

It is easy to imagine a probability distribution superimposed over the confidence interval, such that parameter values in the middle are more probable than parameter values at the ends. But this is a Bayesian interpretation of the interval and is not what the frequentist confidence interval actually provides. Bayesian intervals will be described later in the article. Some frequentists have discussed functions on confidence intervals that resemble probability distributions. For example, a plot of the p value as a function of the parameter value (i.e., a plot of the p value in Eq. 1 as a function of μ) will resemble a probability distribution (e.g., Poole, 1987; Sullivan & Foster, 1990). But the p value curve is not a probability distribution, and it does not indicate the probability of the parameter value. Some analysts have suggested normalizing the p value curve (e.g., Schweder & Hjort, 2002; Singh, Xie, & Strawderman, 2007), but the meaning of such a distribution is remote and the result is still sensitive to stopping and testing intentions. Cumming (e.g., Cumming, 2007, 2014; Cumming & Fidler, 2009) has discussed superimposing sampling distributions or relative likelihood curves on the CI, but neither of these approaches provides the probability of the parameter value, given the data. For more details, see Kruschke (2013, p. 592) or Kruschke (2015, pp. 323–324).

Summary of frequentist approach In summary, frequentist approaches rely on sampling distributions, illustrated by the cloud of imaginary possibilities in Fig. 2. The sampling distribution is defined by the stopping and testing intentions of the researcher. Thus, for any fixed set of actual data, different stopping or testing intentions yield different p values and confidence intervals. Moreover, confidence intervals have no distributional information.

At this point in the article, we are in the midst of explaining the information provided by the various types of analyses, laid out in the conceptual framework of Fig. 1 and the numerical examples in Fig. 3. So far, we have described frequentist approaches to hypothesis testing and estimation with uncertainty, corresponding to the left column of Figs. 1 and 3. We next explain Bayesian approaches to estimation and hypothesis testing in the right column of Figs. 1 and 3. After that, we will explore various arguments for shifting emphasis away from hypothesis testing to estimation with uncertainty, and away from frequentist to Bayesian methods. Ultimately, we will discuss meta-analysis, randomized controlled trials, and power analysis.

Bayesian estimation and highest density interval

To explain how the goals of the New Statistics are achieved through Bayesian estimation, we must first explain the information provided by a Bayesian analysis. We have limited space here and can therefore provide only a cursory overview. Other introductory resources include a companion article in this issue (Kruschke & Liddell, 2015), Chapter 2 of Kruschke (2015, available free online), and other articles (e.g., Kruschke, 2013; Kruschke, Aguinis, & Joo, 2012).

Bayesian analysis is re-allocation of credibility across possibilities. Sherlock Holmes was doing Bayesian reasoning

when he famously said, "when you have eliminated the impossible, whatever remains, however improbable, must be the truth" (Doyle, 1890). Holmes started with various degrees of suspicion about each suspect, then collected new evidence, then re-allocated degrees of suspicion to the suspects who were not eliminated by the evidence. Even if the prior suspicion for a suspect may have been low, when the other suspects are eliminated then the ultimate suspicion for the remaining suspect must be high. A tacit premise of Holmes' statement is that the truth is among the considered possibilities. A more accurate rephrasing would be, "when you have depreciated the most improbable, whatever remains is the least improbable of the options under consideration." Unfortunately, that nuanced phrasing sounds more like Hamlet than Holmes. Nevertheless, the logic still involves reallocation of credibility across possibilities. We start with a prior degree of belief in each possibility, then we collect some data and re-allocate credibility across the possibilities, resulting in a posterior degree of belief in each possibility.

In data analysis, the possibilities are parameter values in a descriptive model of data. Just as Holmes started his investigation with a space of possible explanations for evidence, we start our analysis with a space of possible parameter values in a descriptive model of data. (Frequentist approaches also start with an assumed descriptive model.) The degree of belief in each parameter value, without considering the data, is called the prior distribution, and the degree of belief in the parameter values after taking the data into account is called the posterior distribution. The exact reallocation of credibility across parameter values is provided by the mathematics of Bayes' rule, the details of which are not necessary to describe here. Instead, we provide a numerical example.

Consider again the scenario of the previous section, in which respondents are asked whether they agree or disagree with a particular policy statement. We model the data as if each respondent is a random representative of the underlying probability of agreement, denoted by the value of the parameter θ. The parameter θ can take on values anywhere from $\theta = 0$ to $\theta = 1$. From a Bayesian perspective, our goal is to re-allocate credibility across the possible values of θ, away from values of θ that are inconsistent with the data.

We start with a prior distribution over the candidate values of θ, which for simplicity of illustration we take to be the (essentially) uniform distribution shown in the upper panel of Fig. 4. Then we take into account the data, namely the number of agreements (z) out of the number of respondents (N). Bayes' rule re-allocates credibility to values of θ that are reasonably consistent with the observed proportion of agreement, z/N, as shown in the lower panel of Fig. 4. Notice that small values of θ are not very consistent with the high proportion of agreements in the data, so those small values of θ have low posterior credibility. The modal value

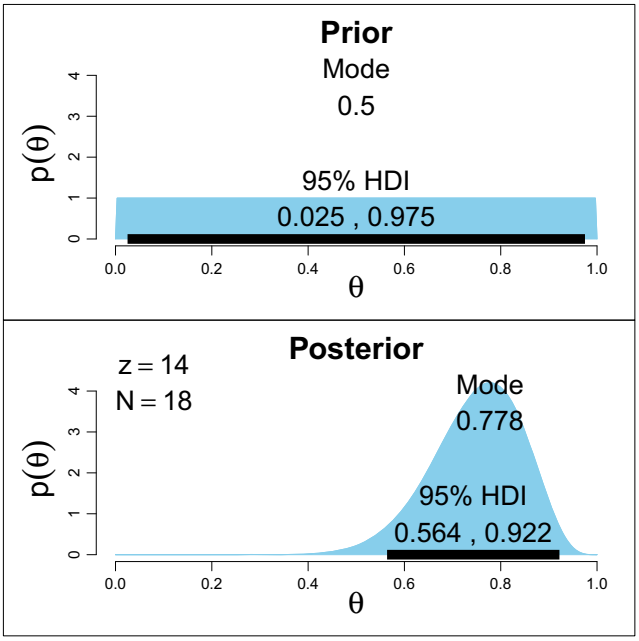

Fig. 4 Bayesian inference is re-allocation of credibility across candidate parameter values. Distributions show the probability density of parameter value θ. The *upper panel* shows the prior distribution. In this case, the prior distribution is essentially uniform, but peaked very slightly so it has a well-defined mode and HDI for purposes of illustration. The *lower panel* shows the posterior distribution, given the data z, N. The posterior distribution here is the same as that shown in the *lower-right cell* of Fig. 3 (*HDI* highest density interval)

of the posterior distribution shows the most credible value of the parameter, given the data.

The uncertainty of the estimate is explicitly indicated by the spread of the posterior distribution. When there is great uncertainty (e.g., because of having a small set of data) then the posterior distribution is spread over a broad range of parameter values, but when there is great certainty (e.g., because of having a large set of data) then the posterior distribution is spread over a narrow range of parameter values. A convenient way to summarize the uncertainty is with the 95 % *highest density interval* (HDI), which contains the parameter values of highest probability and that span the 95 % most probable values. Any parameter value inside the HDI has higher probability density (i.e., higher credibility) than any parameter value outside the HDI. The parameter values inside the 95 % HDI are the 95 % most credible values. Figure 4 marks the HDI's with horizontal black lines at the bottom of the distributions. The HDI is merely a summary statistic and can be applied to any probability distribution. Notice that the 95 % HDI in the prior distribution is wider than the 95 % HDI in the posterior distribution. This reduction of uncertainty reflects greater precision of estimation as more data are included.

It is important to understand that the distributions in Fig. 4 are probabilities of parameter values. The distributions in

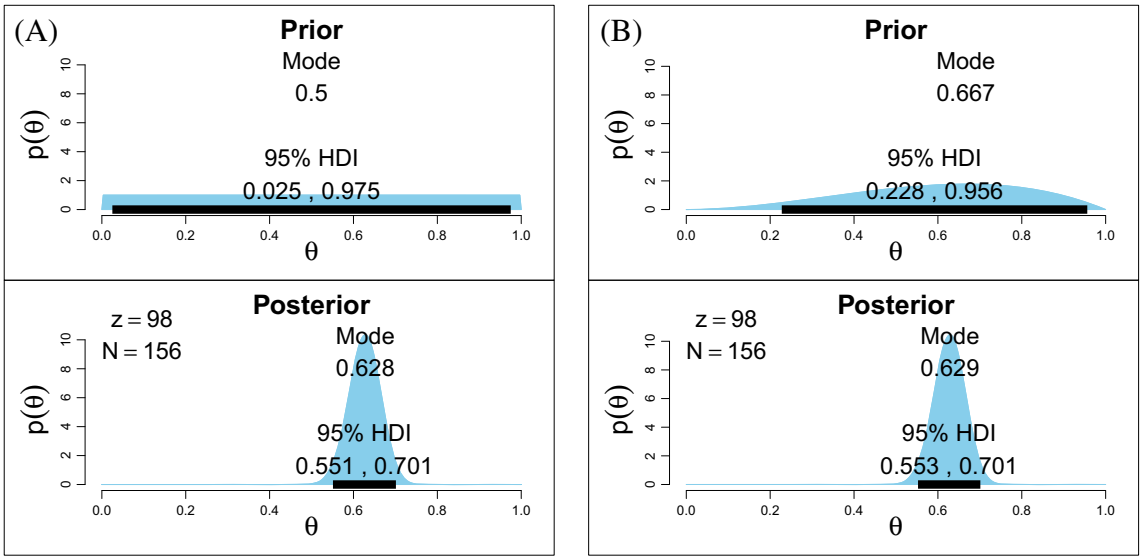

Fig. 5 Diffuse priors have little influence on the posterior for moderate amounts of data. Panel **A** shows results for a uniform prior. Panel **B** shows results for a diffuse but slightly informed prior. Notice that the posterior modes and HDI's are nearly the same for both prior distributions (*HDI* highest density interval)

Fig. 4 are *not* sampling distributions of data, and have nothing to do with the sampling distributions for p values referred to in Fig. 2. The Bayesian posterior distribution in Fig. 4 illustrates the probability of parameter values, given the observed data.

If the prior distribution is broad and diffuse, it has little influence on the posterior distribution when there are realistic (i.e., not tiny) amounts of data. Figure 5 shows that the choice of prior distribution has virtually no effect on the posterior distribution when the prior is relatively flat across the parameter values. Panels A and B of Fig. 5 show two different prior distributions. In panel A, an essentially uniform prior is used. In panel B, the prior distribution is what would result from Bayesian updating of a uniform proto-prior with a small representative pilot in which $z = 2$ and $N = 3$. The resulting posterior distributions in Panels A and B are virtually indistinguishable.

Please now look back to Fig. 3. Compare the information in the lower-right cell, which indicates the information delivered by Bayesian estimation, with the information in the lower-left cell, which indicates the information delivered by frequentist estimation. The most obvious graphical difference is that the Bayesian estimate includes an explicit probability distribution on the parameter values. It is worth re-emphasizing that the posterior distribution on the parameter explicitly indicates (i) the best estimate as the modal value and (ii) the uncertainty as the HDI. The posterior distribution also gives complete information about the credibilities of all the parameter values.

Although the frequentist CI and Bayesian HDI have similar numerical limits in the present example (Fig. 3), the frequentist CI is highly sensitive to the stopping and testing intentions. By contrast, the Bayesian posterior distribution does not change when the stopping or testing intentions change because there is no sampling distribution involved in Bayesian estimation. On the other hand, the Bayesian posterior distribution can be affected by strongly informed prior knowledge, while the frequentist CI is not affected by prior knowledge (because the sampling distribution is not affected by prior knowledge). It is rational that parameter estimation *should* take into account prior knowledge. On the contrary, it is questionable whether parameter estimation should be based on which other tests were intended or whether data collection stopped at fixed duration or fixed sample size. These issues will be discussed in more depth later in the article.

Assessing null values using intervals From the posterior distribution on the parameter, we can assess the extent to which particular values of interest, such as null values, are among the most credible values. People who are familiar with frequentist NHST and confidence intervals, whereby a parameter value is rejected if it falls outside a 95 % confidence interval, may be tempted to apply analogous logic to Bayesian posterior distributions and reject a parameter value if it falls outside a posterior 95 % HDI. Two problems arise from this candidate decision rule. First, it can only reject a parameter value and never accept it. Second, with optional stopping (i.e., gradually accumulating data and repeatedly testing) the decision rule will eventually always reject a null value even when it is true.

To avoid those problems, we adopt a different decision rule, somewhat analogous to frequentist *equivalence testing* (e.g., Rogers, Howard, & Vessey, 1993; Westlake,

1976, 1981). The procedure requires establishing a *region of practical equivalence* (ROPE) around the null value that expresses a range of parameter values that are equivalent to the null value for current practical purposes. For example, if measuring the IQ of people in a treatment group, we might say that any group mean in the interval from 98.5 to 101.5 is practically equivalent to the general population mean of 100 (recall that the standard deviation of IQ scores in the general population is 15, hence the ROPE in this case extends plus or minus 0.1 standard deviations from the general population mean). ROPE's are routinely established in clinical studies to test equivalence or non-inferiority, where care must be taken in high-stakes applications (e.g., Lesaffre, 2008).

At the risk of slipping into black-and-white thinking, if we must make a dichotomous decision about the null value, we can use the following decision rule: If the 95 % HDI falls entirely outside the ROPE then we decide to reject the ROPE'd value (not the entire ROPE'd interval), and if the 95 % HDI falls entirely inside the ROPE then we decide to accept the ROPE'd value for practical purposes, and otherwise we remain undecided. The decision rule follows directly from the meanings of the intervals: When the 95 % HDI falls outside the ROPE, it literally means that the 95 % most credible values of the parameter are all *not* practically equivalent to the null value. When the 95 % HDI falls inside the ROPE, it literally means that all the 95 % most credible values of the parameter *are* practically equivalent to the null value. Notice that the statements made in this context do not use the terms "null hypothesis" or "hypothesis testing" which are reserved for a different approach described later.

For example, in our ongoing example about the probability of agreement with a policy statement, suppose we are interested in whether or not the population can be said to be ambivalent (i.e., $\theta = 0.50$), and for practical purposes of decision making we define a ROPE from $\theta = 0.45$ to $\theta = 0.55$. From the posterior distribution in Figs. 3 and 4, we can make statements as follows: "The 95 % most credible values are all *not* practically equivalent to the null value (i.e., the 95 % HDI excludes the ROPE)", and, "there is only 2.4 % probability that θ is practically equivalent to the null value (i.e., 2.4 % of the posterior distribution falls within the ROPE)". The area of the posterior distribution inside a ROPE is easily computed but is not displayed in Figs. 3 and 4 (but is displayed later in Fig. 11). A different Bayesian decision rule for equivalence examines only the posterior probability mass inside the ROPE, without regard to the HDI (e.g., Wellek, 2010) but we prefer to treat the probability density as a meaningful quantity that better handles skewed distributions (Kruschke, 2015, Section 12.1.2.2, p. 342).

As discussed by Kruschke (2015, p. 337), use of a ROPE is also motivated from the broader perspective of scientific method. Serlin and Lapsley (1985, 1993) pointed out that using a ROPE to affirm a predicted value is essential for scientific progress, and is a solution to Meehl's paradox (e.g., Meehl, 1967, 1978, 1997). ROPE's go by different names in the literature, including "interval of clinical equivalence," "range of equivalence," "equivalence interval," "indifference zone," "smallest effect size of interest," and "good-enough belt" (e.g. Carlin & Louis, 2009; Freedman, Lowe, & Macaskill, 1984; Hobbs & Carlin, 2008; Lakens, 2014; Schuirmann, 1987; Serlin & Lapsley, 1985, 1993; Spiegelhalter, Freedman, & Parmar, 1994).

We bring up the HDI+ROPE decision method here only to make sure that readers do not assume, by analogy to confidence intervals, that a null value can be rejected if it falls outside the 95 % HDI. Instead, the decision rule uses a ROPE around the null value. The ROPE is crucial to allow a decision to accept a null value, and to make the decision rule technically consistent: As N increases, the decision rule converges to the correct decision (i.e., either practically equivalent to the null or not). This decision rule is described in more detail by Kruschke (2011a, 2015, Ch. 12). However, dichotomous decision making is not meant to be the goal here, and the emphasis is on the full information provided by the continuous posterior distribution.

Highest density interval vs. confidence interval Here we reiterate some essential differences between a Bayesian HDI and a frequentist CI. An important quality of the posterior 95 % HDI is that it really does indicate the 95 % most probable values of the parameter, given the data. The posterior distribution depends only on the actually observed data (and the prior), and does not depend on the stopping or testing intentions of the analyst. The frequentist CI is often misinterpreted as if it were a posterior distribution, because what analysts intuitively want from their analysis is the Bayesian posterior distribution, as we discuss more later.

The posterior 95 % HDI refers explicitly to an actual probability distribution over the parameter values, such that parameter values in the middle of the HDI tend to have higher credibility than parameter values at the limits of the HDI. The posterior distribution shows the exact shape of the distribution. A frequentist CI, on the other hand, does not refer to a probability distribution over parameter values and carries no distributional information, as was discussed earlier in the article.

The posterior mode and HDI do not change if the stopping or testing intentions change. By contrast, the frequentist p value and CI are defined in terms of simulated data generated by the stopping and testing intentions, and therefore the p value and CI change when stopping or testing intentions change (as was explained in a previous section). In a Bayesian analysis, it is straight forward to indicate the 95 % HDI around a modal parameter estimate. In frequentist

analyses, on the other hand, there is often difficulty deciding what is an appropriate CI because it depends on the testing and stopping intention.

Finally, if an HDI is used as part of a decision rule to assess null values, the decision rule should include a ROPE around the null value. A null value should not be rejected merely if it falls outside a 95 % HDI (unlike 95 % CI's in NHST). Moreover, the decision rule based on HDI and ROPE intervals is not called Bayesian "hypothesis testing," which is a term reserved for a different framework that we describe next.

Bayesian hypothesis test

In a Bayesian point-null hypothesis test, the null hypothesis is expressed as a prior distribution that puts all credibility in an infinitely dense spike at the null value of the parameter, with zero credibility on all other values of that parameter. In other words, the prior distribution for the null hypothesis says that only the null value is available. (The Bayesian framework allows other types of null hypotheses, but here we focus on point nulls for comparability to NHST.) Crucially, the null hypothesis is compared against an alternative prior distribution that spreads credibility over other values of the parameter. Unlike NHST, a Bayesian hypothesis test demands the specification of an alternative hypothesis, in the form of an alternative prior distribution on the parameter. Each hypothesis is indicated by a model-index parameter: $M = 1$ for the null hypothesis and $M = 2$ for the alternative hypothesis. Bayesian inference reallocates credibility across the two hypotheses by reallocating credibility across the values of the model-index parameter.

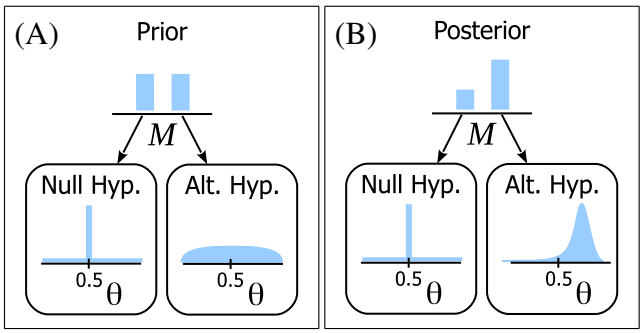

Fig. 6 Bayesian point-null hypothesis testing. The null value of parameter θ is denoted here generically by the tic mark at $\theta = 0.5$. In **A**, the prior distribution shows that the null hypothesis ($M=1$) assumes a spike-shaped prior distribution on θ such that only the null value has non-zero probability, whereas the alternative hypothesis ($M=2$) assumes a broad prior distribution on θ. In **B**, the posterior distribution shows that credibility has been re-allocated across the possible parameter values. For these data, the model-index parameter M shows that the alternative hypothesis ($M=2$) has higher posterior probability, and within the alternative hypothesis the distribution over θ shows that the most credible values of θ are away from the null value

To illustrate this idea, consider again the case of estimating the underlying probability of agreement to a policy statement. Denote the underlying probability of agreement as θ, and suppose that the "null" value of θ is 0.5, indicating exact ambivalence in the population. The null hypothesis has a spike-shaped prior distribution on θ, denoted $p(\theta|null)$, such that the probability density is $p(\theta = 0.5|null) = \infty$ and $p(\theta \neq 0.5|null) = 0$. The alternative hypothesis has a prior distribution over θ denoted $p(\theta|alt)$ that is spread over θ in some meaningful way. The alternative-hypothesis prior on θ could be generically vague and set by default, or the alternative-hypothesis prior on θ could be meaningfully informed by previous data or theory. Although we will illustrate Bayesian null hypothesis testing by using a default alternative prior, in applied practice it is important to use a meaningfully informed alternative prior (Dienes, 2014; Kruschke, 2011a; Vanpaemel and Lee, 2012).

Figure 6 illustrates the parameter space for Bayesian null hypothesis testing. Panel A of Fig. 6 shows the null hypothesis as a distribution over θ that has a spike at $\theta = 0.5$. The alternative hypothesis has a broad distribution over θ. The null and alternative priors are indexed at the top of Panel A by the model index parameter M. The two-bar distribution at the top of panel A indicates the prior probabilities that $M=1$ or $M=2$, which in this illustration are set equal to each other.

Bayesian inference reallocates credibility across the model index M and the parameter θ simultaneously. When new data are taken into account, Bayes' rule prescribes how to shift probabilities across the parameter values. The resulting posterior distribution is illustrated in panel B of Fig. 6. In this case, the data consisted of a higher proportion of agreement than disagreement, and therefore the explicit estimate of θ, illustrated inside the alternative hypothesis in panel B, is peaked somewhat larger than the null value. Simultaneously, the distribution on the model index has shifted so that $M=2$ is more probable than $M=1$. Bayesian null hypothesis testing focuses on the model index, not on the estimate of parameters within the models.

The degree to which the model index shifts, from prior to posterior, is called the *Bayes factor*. With respect to Fig. 6, the Bayes factor can be visualized by seeing how much the two-bar distribution on M shifts from panel A (the prior) to panel B (the posterior). In panel A of Fig. 6, the height of the bar on $M=1$ is the prior probability of the null hypothesis, denoted $p(M=null)$. In panel B of Fig. 6, the height of the bar on $M=1$ is the posterior probability of the null hypothesis, denoted $p(M=null|D)$ where D denotes the observed data. The Bayes factor of null versus alternative is denoted BF_{null} and can be defined as the ratio of the posterior odds to the prior odds:

$$BF_{null} \equiv \frac{p(M=null|D)}{p(M=alt|D)} \Big/ \frac{p(M=null)}{p(M=alt)} \tag{3}$$

Importantly, the Bayes factor does not indicate the posterior probabilities or posterior odds; instead, the Bayes factor indicates the degree of change from the prior odds to the posterior odds of the null-model index. In other words, the posterior odds are the prior odds multiplied by the Bayes factor:

$$\frac{p(M{=}null|D)}{p(M{=}alt|D)} = BF_{null} \times \frac{p(M{=}null)}{p(M{=}alt)} \quad (4)$$

A numerical example of a Bayes factor is shown in the upper-right panel of Fig. 3. In that case, a spike-shaped null hypothesis (at $\theta_{null} = 0.5$) is compared against an alternative hypothesis that has uniform probability across the range of θ. A depiction of an essentially uniform prior was provided in Fig. 4. The uniform distribution for the alternative prior distribution is denoted in Fig. 3 by the notation "alt = dbeta(1,1)" because it is a formal reference to a beta distribution that is equivalent to a uniform distribution. The example in the upper-right panel of Fig. 3 shows that the Bayes factor is approximately 0.22, meaning that the prior odds of the null hypothesis is reduced by a factor of 0.22. For example, if the prior probability of the null hypothesis were 0.8, then the Bayes factor of 0.22 implies that the posterior probability of the null hypothesis would be 0.47. If the prior probability of the null hypothesis were 0.5, then the posterior probability of the null hypothesis would be 0.18. If the prior probability of the null hypothesis were 0.2, then the posterior probability of the null hypothesis would be 0.05. In all cases, these posterior probability of the null hypothesis is with respect to the particular alternative hypothesis being tested.

A common decision rule for Bayesian null-hypothesis testing is based on the Bayes factor (not on the posterior probabilities). According to this decision procedure, the Bayes factor is compared against a decision threshold, such as 10. When $BF_{null} > 10$, the null hypothesis is accepted relative to the particular alternative hypothesis under consideration, and when $BF_{null} < 1/10$, the null hypothesis is rejected relative to the particular alternative hypothesis under consideration. The choice of decision threshold is set by practical considerations. A Bayes factor between 3 and 10 is supposed to indicate "moderate" or "substantial" evidence for the winning model, while a Bayes factor between 10 and 30 indicates "strong" evidence, and a Bayes factor greater than 30 indicates "very strong" evidence (Jeffreys, 1961; Kass & Raftery, 1995; Wetzels et al., 2011). Dienes (2016) suggested a Bayes factor of 3 for substantial evidence, while Schönbrodt et al. (2016) recommended the decision threshold for a Bayes factor be set at 6 for incipient stages of research but set at a higher threshold of 10 for mature confirmatory research (in the specific context of a null hypothesis test for the means of two groups, implying that the decision threshold might be different for different sorts of analyses). Somewhat analogous to considerations for a ROPE, the decision threshold for a Bayes factor depends on the practical aspects of the application.

Basing a decision on the Bayes factor alone can be useful when the prior odds of the models are 50/50 because then the Bayes factor numerically equals the posterior odds, as can be seen immediately from Eq. 4. Otherwise it is important to take into account the prior probabilities of the hypotheses. For example, consider a study of extrasensory perception (ESP), in which people are asked to predict which of two random stimuli will appear in the future. The null hypothesis of chance performance has an extremely high prior probability. Even if the Bayes factor indicates a shift away from the null hypothesis by a factor of 30 or more, the posterior probability of the null hypothesis would remain very high (e.g., Rouder & Morey, 2011; Rouder, Morey, & Province, 2013). As another example in which the Bayes factor alone is not appropriate for making decisions, consider the diagnosis of rare diseases. In this context, the datum is the outcome of a diagnostic test, which could be "positive" to suggest the presence of disease or "negative" to suggest the absence of disease. The actual underlying condition of the patient is one of two states, or models: $M{=}1$ indicates the patient really has the disease, and $M{=}2$ indicates the patient really does not have the disease. Because the disease is rare, the prior probability that $M = 1$ is very small. When the test result is positive, the Bayes factor is the hit rate of the diagnostic test divided by its false alarm rate. Even if this Bayes factor for having the disease is large, the posterior probability of having the disease remains small. Further discussion of the Bayes factor in disease diagnosis is provided at http://tinyurl.com/ModelProbUncertainty.[1]

Making decisions about null values Each cell of the 2×2 table in Figs. 1 or 3 allows the analyst to make a decision about a null value, if the analyst is so inclined, but the decision is based on different information in different cells. In the upper-right cell, the BF indicates the degree of shift from prior odds to posterior odds of a null-hypothesis prior and a particular alternative hypothesis prior. The BF is a continuous value that can be compared against a criterial threshold for making a decision (although we recommend considering the posterior probabilities instead of the BF). In the lower-right cell of Figs. 1 or 3, the posterior distribution shows which values of the parameter are more or less credible. The HDI captures the most credible values, and can be compared against a criterial ROPE around the null value for making decisions. The two Bayesian methods base decisions on the posterior probability of parameter values, with the BF focusing on the between-model index parameter

[1]The full url is http://doingbayesiandataanalysis.blogspot.com/2015/12/lessons-from-bayesian-disease-diagnosis_27.html.

and the HDI+ROPE focusing on the within-model parameter estimate. There is no necessary relation between making a decision via the BF and making a decision via the HDI+ROPE, though often the decisions will agree.

In the frequentist column of Figs. 1 or 3, in the upper-left cell the p value indicates the probability that a null hypothesis would generate imaginary data with a summary statistic as extreme as or more extreme than the actual data's summary statistic, for imaginary data sampled and tested with the same intentions as the actual data. The p value is a continuous value that can be compared against a criterial threshold for making a decision. The threshold represents the rate of false alarms we are willing to tolerate. In the lower-left cell of Figs. 1 or 3, the confidence interval indicates the range of hypothetical parameter values that have p values that do not fall below the decision threshold. Therefore the null value is rejected if and only if it falls outside the CI. In this way the upper-left and lower-left cells are redundant when using them to reject null values. (Frequentist equivalence testing accepts the null value if the 90 % confidence interval falls inside a region of equivalence to the null.)

Another example of frequentist and Bayesian approaches to hypothesis testing and estimation

In this section we present another example of the information in the four cells of Fig. 1, this time applied to metric data from a single group. We will model the data with a normal distribution, which has two parameters: the mean μ and the standard deviation σ. There is a third parameter derived from the mean and standard deviation, called the effect size, that we describe in more detail below. We are interested in testing and estimating all three parameters.

We are presenting this additional example because it illustrates several interesting contrasts between approaches that were not evident in the simpler case presented earlier. In particular, (i) hypothesis testing of the three parameters involves three distinct tests, (ii) Bayesian hypothesis testing can come to different conclusions for mean and effect size unlike traditional frequentist hypothesis testing, and (iii) frequentist parameter estimation involves three sampling distributions whereas Bayesian parameter estimation seamlessly yields complete information about all three parameters in a unified joint distribution.

The top panel of Fig. 7 shows a histogram of the data, which were generated as a random sample from a normal distribution. The values are supposed to be in the vicinity of typical intelligence quotient (IQ) scores, which are normed for the general population to have a mean of 100 and a standard deviation of 15. We suppose that the data came from a group of subjects who were given a "smart drug" and we would like to know how different this group is from the general population. Typically for this sort of data there are (at least) three parameters of interest, namely, the mean μ, the standard deviation σ, and the *effect size*, which we define here as Cohen's $d = (\mu - 100)/\sigma$, which is the distance between the mean and a reference value, relative to the "noise" in the data (Cohen, 1988). We are interested in the mean because we would like to know how different the central tendency of the smart-drug group is from the general population. We are interested in the standard deviation because we would like to know how different the variability of the smart-drug group is from the general population on the scale of the data. Stressors such as performance drugs can increase the variance of a group because not everyone responds the same way to the stressor (e.g., Lazarus & Eriksen, 1952). Finally, we are interested in the effect size because it indicates the magnitude of the change in central tendency standardized relative to the variability in the group.

Three frequentist hypothesis tests are shown in Fig. 7. A traditional t-test checks the hypothesis that $\mu = 100.0$. The summary statistic for its sampling distribution is $t = (\bar{y} - 100.0)/(s/\sqrt{N})$ where N is the sample size, $\bar{y}$ is the sample mean, and s is the sample standard deviation (using $N-1$). A chi-square test checks the hypothesis that $\sigma = 15.0$. The summary statistic for its sampling distribution is $\chi^2 = (N-1)\, s^2/15.0^2$. Finally, the summary statistic for the effect size is $d = (\bar{y} - 100.0)/s$, which follows a non-central t distribution (e.g., Cumming & Finch, 2001). Notice in this case that the null hypotheses that $\mu = 100.0$ and $\delta = 0.0$ cannot be rejected. (Failure to reject the null hypothesis does not imply accepting the null hypothesis, which could be addressed by equivalence testing.) The null hypothesis that $\sigma = 15.0$ can be rejected. For all three tests, we assume that N is fixed by the stopping intention. If there were some other stopping intention, the p values provided by standard software would not be accurate. Moreover, the p values of the three tests are not corrected for the fact that there are multiple tests on the data. Corrections (enlargements) of the p values would be necessary if the analyst wanted to keep the overall false alarm rate limited to some maximum such as 5 %. Our main point in presenting these tests is to emphasize that they involve distinct summary statistics with distinct sampling distributions.

Frequentist confidence intervals are shown in the lower-left of Fig. 7. The CI's are linked to the p values in the cells above them, because the CI's are defined in terms of the p values. The three CI's involve distinct summary statistics with distinct sampling distributions, as indicated by the cell annotations. Specifically, the CI for the mean comes from the sampling distribution of the mean (equivalently, a central t distribution), the CI for the standard deviation comes from a chi-square sampling distribution, and the CI for the effect size comes from a sampling distribution for

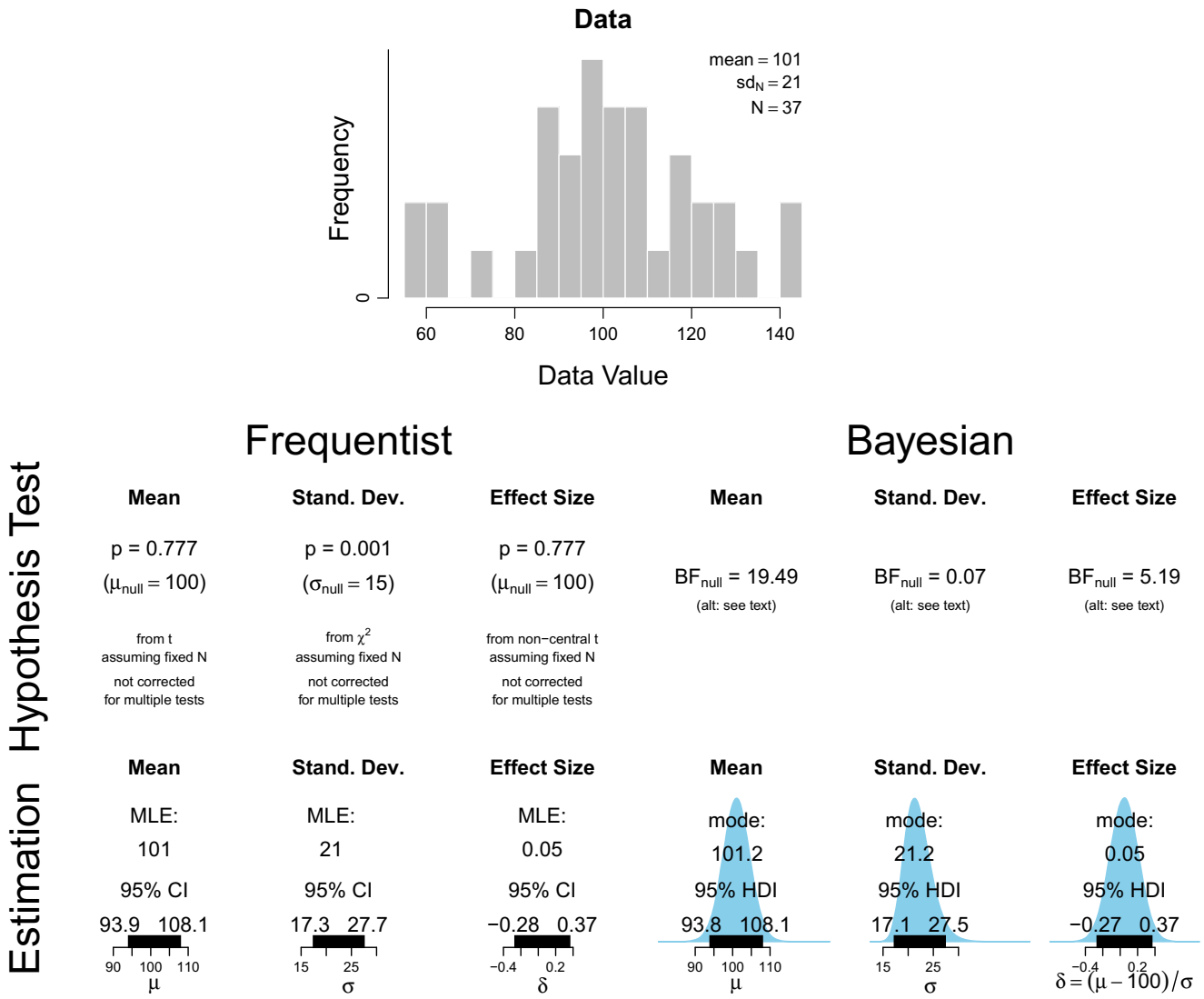

Fig. 7 Data are shown in the *top panel* (the annotated sd_N uses N in its denominator to describe the sample). *Lower panels* show various analyses as described in the main text (*BF* Bayes factor, *MLE* maximum likelihood estimate, *CI* confidence interval, *HDI* highest density interval)

non-central t. The three CI's would change if the stopping intention were changed or if the p values were corrected for multiple testing.

Results from Bayesian hypothesis tests are shown on the right side of Fig. 7. Because we are dealing with two parameters (i.e., μ and σ), the null and alternative hypotheses involve different prior distributions on the two-dimensional joint parameter space. For a hypothesis test regarding the mean, the null prior is an infinitesimally narrow ridge at $\mu = 100.0$ but broad over σ. That is, the null hypothesis is a "spike" prior on the mean parameter, crossed with a vague prior on the standard-deviation parameter. For a hypothesis test regarding the standard deviation, the null prior is an infinitesimally narrow ridge over $\sigma = 15.0$ but broad over μ. For a hypothesis test regarding the effect size, the null hypothesis is an infinitesimally narrow ridge at $\delta = (\mu - 100.0)/\sigma = 0.0$ but broad over σ. In all three cases the alternative hypothesis is broad over both μ and σ, but note that the prior probability density at $\langle \mu, \sigma \rangle$ does not necessarily equal the prior probability density at $\langle (\mu - 100.0)/\sigma, \sigma \rangle$. Bayes factors were computed using the Savage-Dickey method (Lee and Wagenmakers, 2014; Wagenmakers et al., 2010; Wetzels et al., 2009). The alternative prior was moderately diffuse such that it yielded a BF on the effect size quantitatively similar to the BF produced by the Bayes factor package of (Morey et al., 2015) with rscale at its default value of 0.707. This choice of default prior is merely a convenience for the purpose of illustration.

Real research would instead have to use an informed prior that expresses a theoretically meaningful alternative. The three BFs were computed simultaneously using the same alternative prior; future researchers might recommend using different alternative priors for different tests.

Figure 7 shows that the BF for the test of the mean indicates a large shift between prior odds and posterior odds in favor of the null hypothesis relative to the alternative hypothesis. The BF for the effect size also indicates a shift between prior odds and posterior odds in favor of the null hypothesis. If an analyst based a decision on a BF threshold of 10.0 (e.g., Schönbrodt et al., 2016), the null hypothesis for the mean would be accepted relative to the alternative hypothesis but the analyst would remain undecided about the effect size. The BF for the standard deviation, on the other hand, indicates a large shift between prior odds and posterior odds in favor of the alternative hypothesis relative to the null hypothesis. Our main point in presenting these tests is to emphasize that the three Bayesian hypothesis tests are considering distinct prior distributions for distinct null hypotheses (so that the BF for the mean can differ from the BF for the effect size), and that the Bayesian hypothesis test can decide to accept the null relative to a particular alternative not just reject the null.

Finally, the lower right of Fig. 7 shows the result of Bayesian estimation of the parameters, starting with a diffuse prior distribution on the joint parameter space (the same that was used as the alternative hypothesis in the Bayesian null hypothesis tests). As was emphasized earlier, any reasonably diffuse prior yields virtually the same posterior distribution on continuous parameter values for moderate amounts of data. In the posterior distribution, each $\langle \mu, \sigma \rangle$ combination corresponds to an effect size, $\delta = (\mu - 100.0)/\sigma$, so the posterior distribution simultaneously provides a posterior distribution on effect size. The lower right of Fig. 7 shows the marginal posterior distribution on μ, σ, and δ. Notice in particular that the posterior distribution on the mean has a 95 % HDI with a width greater than 14 IQ points, meaning that there is a fairly wide range of credible means despite the fact that the BF favors the null hypothesis. We emphasize that the Bayesian parameter estimates come from a single prior distribution, not from separate hypotheses.

Corrections for multiple tests

In frequentist analysis, a primary goal for the decision rule is keeping the overall false alarm rate (a.k.a., Type I error rate) limited to 5 % (say) across the set of tests. When more tests are included in the analysis, there is increased opportunity for false alarms, as was discussed earlier in the context of Fig. 2. Therefore the decision thresholds for every test must be made more conservative, or, equivalently, the p values of every test must be made larger. The exact amount of increase in p depends on the structural relations of the tests, and hence there are a variety of different corrections available for different typical families of related tests (e.g., Tukey, Scheffé, Dunnett, Bonferonni, etc.; see Maxwell & Delaney, 2004). If error control is the goal of the decision rule, then, by definition, p values must be computed and their dependency on the stopping intention and testing intention must be taken into account.

Bayesian analysis does not base decisions on error control. Indeed, Bayesian analysis does not use sampling distributions. (The only exception to this statement is Bayesian power analysis, which, by definition, considers sampling distributions but not p values. Bayesian power analysis is described later in this article.) Instead of using error rates, Bayesian decisions are based on properties of the posterior distribution on parameter values.

Of course, ignoring errors rates does not make them go away. Any decision rule on noisy data will eventually commit an error on some set of data, because ultimately the errors come from noise that spawns random coincidences of rogue data. Bayesian analysis can mitigate false alarms through model structure instead of through p values. In particular, hierarchical models are seamlessly analyzed in Bayesian software, and hierarchical models allow the data to inform *shrinkage* which pulls in parameter estimates of rogue groups. An example of a hierarchical model with shrinkage is presented later in this article, in an application to meta-analysis.

Ultimately, errors can be rectified only by pre-registered replication attempts. Some Bayesian methods in replication analysis are surveyed in a short video available at the following link, http://tinyurl.com/BayesianReplicationAnalysis[2]

Interim summary

We have now provided two full examples of the information provided by hypothesis testing and parameter estimation, using frequentist and Bayesian approaches. The results of the examples are summarized in Figs. 3 and 7. The examples illustrate the very different information provided by the different approaches. Within hypothesis testing, frequentist approaches provide a p value for imaginary data from a null hypothesis, whereas Bayesian approaches provide a comparison between the relative abilities of a null hypothesis and an alternative hypothesis to account for the actual data. Within estimation, frequentist approaches provide the range

[2]The full URL is http://doingbayesiandataanalysis.blogspot.com/2016/05/some-bayesian-approaches-to-replication.html.

of parameter values that would not be rejected, whereas Bayesian approaches provide a full posterior distribution over the joint parameter space. At the least, these examples illustrate that Bayesian approaches provide different and richer information than frequentist approaches.

All of the preceding formed the foundation for the major topics in the remainder of the article, where we will review arguments for shifting emphasis from frequentist to Bayesian methods, and for emphasizing estimation and not only hypothesis testing. We will then explain meta-analysis, randomized controlled trials, and power analysis from a Bayesian perspective. Now would be a good time to stretch, refill your coffee mug, and then return here to read the exciting conclusion!

Arguments for shift from frequentist to Bayesian

The literature across many sciences has articles and books that note and promote a shift away from frequentist to Bayesian methods (e.g., Allenby, Bakken, & Rossi, 2004; Beaumont & Rannala, 2004; Brooks, 2003; Gregory, 2001; Howson & Urbach, 2006; Lindley, 1975; McGrayne, 2011, etc.). Reasons for the shift are numerous; here we focus on only a few foundational issues. First, any set of data has p values and confidence intervals that depend on stopping and testing intentions, but this dependency may not be desirable. Second, frequentist approaches answer a question that most analysts are not asking, whereas Bayesian approaches actually address the question of usual interest.

Sampling and testing intentions influence p values and confidence intervals

Recall the notion of a p value illustrated back in Fig. 2 and defined formally back in Eq. 1. A crucial detail of the process in Fig. 2 is that the simulated samples of data are generated the same way as the actual sample of data. In particular, if the *actual* sample was generated by collecting data until a specific sample size was reached, then the *simulated* samples must be generated by sampling data until that same sample size is reached. This assumption of stopping at fixed N is the usual and tacit assumption. By contrast, actual data are often sampled with other intentions for stopping of data collection. For example, a mail survey will have a random number of respondents. In this case, to generate the cloud of possible outcomes in Fig. 2, we must not use a fixed N for the simulated samples, but we must generate each sample with a random size that mimics the random sample sizes possible in the actual survey. As another example, experimenters in a university setting might post times for available sessions during which people can sign up to participate. Although there is a fixed number of sessions, the number of people who volunteer within that window is a random number. To generate the cloud of possible outcomes in Fig. 2, the analysts must not use a fixed sample size, but must generate each sample with a random size that mimics the random sample sizes possible in the actual experiment procedure. The procedure for collecting data determines when to stop collecting data, and is called the *stopping intention*.

The stopping intention is crucial for computing the p value because the stopping intention determines the cloud of imaginary possibilities relative to which the actual outcome is judged. In other words, different stopping intentions yield different p values for the same actual outcome. In particular, two researchers might happen to collect the same data, but if they had different stopping intentions, then their identical data would have different p values. Detailed numerical examples are provided in the literature, including Kruschke (2013, pp. 588-590) and Kruschke (2015, Ch. 11) and many references cited therein. Despite this dependency of a p value on stopping intention, all standard software packages assume a fixed-N stopping intention.

The p value for a set of data is also affected by the other tests that will contribute to the cloud of possible data summaries. If the actual outcome is going to be compared with several other groups, then there are additional opportunities for the simulated summaries from all the groups to exceed the actual outcome. In other words, the cloud of possibilities expands and the p value increases, even though the actual outcome is unchanged. This inflation of p values with additional tests is often discussed in textbooks and articles, which describe an elaborate set of "corrections" to p values (e.g., Maxwell & Delaney, 2004). Figure 2 annotated the dependence of the cloud of possibilities on the testing intention as well as the stopping intention.

In particular, two researchers might happen to collect the same data, but if they had different testing intentions, then their identical data would have different p values. Notice that the p value is affected only by the intention to do other tests, not by the actual data for the other tests or even whether the other data have yet been collected. This inflation of the p value with intention to do more tests causes many researchers, when reporting an analysis, to pretend disinterest in comparisons that obviously deserve testing.

A confidence interval is defined in terms of p values, and therefore when the stopping and testing intentions change, the CI changes too. The formal definition of a CI was provided in Eq. 2, where the dependence on stopping and testing intentions was explicitly noted. When the p value is inflated by additional tests, the confidence interval becomes correspondingly wider. Because the limits of the CI merely indicate the parameter values at which $p = 0.05$, there is

no distributional information provided by a CI. The contrast between CI and Bayesian posterior distribution was graphically illustrated in Figs. 3 and 7.

The dependence of the p value and CI on stopping and testing intentions does not make much sense when the goal is inferring the credibility of descriptive parameter values. If two researchers collect the same data, but one researcher stopped at fixed N (with a random duration of data collection) while the other researcher stopped Friday at 5:00pm (with a random N), it seems absurd that the researchers should have different estimation intervals from their identical data. But different p values and CI's are required by a correct application of the frequentist approach, and the differences can often be very substantial. Bayesian inference does not depend on stopping and testing intentions.

Frequentist NHST asks the wrong question

When we collect data about some phenomenon of interest and we describe the data with a mathematical model, usually our first question is about what parameter values are credible given the actual data. For example, given a set of IQ scores, we would like to know what values of the mean parameter are credible. Formally, we denote this desired information as

$$p\left(\mu \,\middle|\, D_{actual}\right) \tag{5}$$

where μ indicates a parameter value and D_{actual} indicates the actual data. Equation 5 describes the probability of all possible parameter values given the actual data. This desired information is exactly what Bayesian inference provides in the posterior distribution. But this is not what a frequentist analysis provides. Frequentist analysis provides the probability that summaries of simulated data from a hypothetical value of the parameter would be more extreme than the summary of the actual data. Formally, frequentist analysis provides the p value that was defined in Eq. 1 and is repeated here for convenience:

$$p\left(T(D_{simulated}) \succeq T(D_{actual}) \,\middle|\, \mu, I\right) \tag{6}$$

where $T(D_{simulated})$ is a summary description (such as a t statistic) of simulated data that are sampled according to the same stopping and testing intentions I as the actual data.

Notice that the conditional probability provided by frequentist analysis in Eq. 6 is the *reverse* of the desired conditional probability provided by Bayesian analysis in Eq. 5. For example, the (Bayesian) information we desire may be the probability that it will rain at noon today, but the (frequentist) information provided is the probability that it is noon given that it is raining. Needless to say, p(rain|noon) does not equal p(noon|rain), and analogously $p(\mu \mid D_{actual})$ does not equal $p(T(D_{simulated}) \succeq T(D_{actual}) \mid \mu, I)$. Notice also that the conditional probability provided by frequentist analysis in Eq. 6 involves the sampling intentions I but those intentions are not involved in the Bayesian information in Eq. 5. For example, the Bayesian probability in Eq. 5 does not change if the stopping intention changes from fixed sample size to fixed duration of sampling.

Because researchers want the Bayesian information in Eq. 5, they often misinterpret the results of frequentist analysis. When people find that a frequentist p value is, say, 0.02, they often mistakenly interpret the p value as meaning that the probability that μ equals the null value is 2 % (e.g., J. Cohen, 1994, and references cited therein). In other words, researchers often treat a p value as if it refers to Eq. 5. But the frequentist p value of Eq. 6 has little to say about the probability of parameter values. "[NHST] does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (J. Cohen, 1994, p. 997). Analogously, we so much want to know a distribution of probability across parameter values, as in Eq. 5, that, out of desperation, we believe there is such a distribution on a frequentist confidence interval even though there is none.

The frequentist question can be the right question to ask if the analyst is specifically interested in a decision rule that attempts to control error rates (e.g., Mayo & Spanos, 2011; Mayo, 2016). As Bayesian analysis ignores counterfactual error rates, it cannot control them. "The reason is that [Bayesian methods] condition on the actual data; whereas error probabilities take into account other outcomes that could have occurred but did not." (Mayo, 2016) That is, if the goal is specifically to control the rate of false alarms when the decision rule is applied repeatedly to imaginary data from the null hypothesis, then, by definition, the analyst must compute a p value and corresponding CI's. To answer this sort of question, the analyst must take into account the exact stopping and testing intentions. Consequently, there is a proliferation of methods for computing correct p values and confidence intervals for many different stopping and testing situations (e.g., Sagarin, Ambler, & Lee, 2014). It is important to understand that any such procedure does not yield the credibility of parameter values (as in Eq. 5) but instead yields the probability of imaginary data (as in Eq. 6).

Benefits of Bayesian

Frequentist hypothesis testing can only reject or fail to reject a particular hypothesis. It can never show evidence in favor of a hypothesis. Bayesian hypothesis testing, on the other hand, inherently compares a (null) hypothesis against

one or more alternative hypotheses. In the Bayesian framework, the null hypothesis can be accepted relative to the particular alternative (e.g., Edwards, Lindman, & Savage, 1963; Lee & Wagenmakers, 2014, Ch. 7). One benefit of Bayesian hypothesis testing is that a goal for research could be to reject or to accept a null hypothesis, and publications might be less selectively biased toward rejected null values (Dienes, 2016). As is discussed later, Bayesian methods also allow precision of estimation to be a goal for research more coherently than with frequentist methods.

The examples in Figs. 3 and 7 might make it appear that frequentist and Bayesian estimation tend to look remarkably similar, and all that Bayesian analysis adds is a pretty distribution over the parameters along with legalistic legitimacy to interpret the result as probabilities of parameters, without actually changing any conclusions. Such an impression is a false generalization from the simple examples presented in those figures. There are applications in which frequentist CIs are only roughly approximated or are difficult to compute at all, but Bayesian posterior distributions and their HDIs are seamlessly computed. In many realistic applications with complex models, frequentist approaches are limited by (1) hill-climbing algorithms for finding MLE parameters that sometimes fail to converge, (2) large-N approximations to sampling distributions that provide overly optimistic p values and CI's, and (3) software that constrains the types of model structures and data distributions. On the other hand, modern Bayesian algorithms and software are robust across a wide range of complex models that can be very flexibly specified by the analyst, and the results are exact for any size N no matter how small. Examples of more complex applications of Bayesian methods are presented later in the article (for randomized controlled trials and for meta-analysis of binomial data). Moreover, Bayesian methods allow prior knowledge to be incorporated into estimation and decision making, as is well recognized to be crucial in disease diagnosis and drug testing.

Arguments for shift of emphasis toward estimation with uncertainty—the "New" in the New Statistics

The literature is replete with articles and books that lament hypothesis testing and encourage a shift to estimation with uncertainty. Among the most recent and prominent cautions against NHST is a statement from the American Statistical Association (ASA; Wasserstein & Lazar, 2016), which featured six principles for properly interpreting p values. Principle 3 said "Scientific conclusions and business or policy decisions should not be based only on whether a p value passes a specific threshold. ... The widespread use of 'statistical significance' (generally interpreted as '$p < 0.05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process." Principle 5 said "A p value, or statistical significance, does not measure the size of an effect or the importance of a result. Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p values do not necessarily imply the presence of larger or more important effects, and larger p values do not imply a lack of importance or even lack of effect."

Accompanying commentaries on the ASA statement also highlighted a shift of emphasis from (frequentist) null hypothesis testing to estimation with uncertainty: "... statistical tests should never constitute the sole input to inferences or decisions about associations or effects. Among the many reasons are that, in most scientific settings, the arbitrary classification of results into 'significant' and 'non-significant' is unnecessary for and often damaging to valid interpretation of data; and that estimation of the size of effects and the uncertainty surrounding our estimates will be far more important for scientific inference and sound judgment than any such classification." (Greenland et al., 2016)" To avoid fallacious black-and-white thinking from null hypothesis testing, "... we can and should advise today's students of statistics that they should avoid statistical significance testing, and embrace estimation instead." (Rothman, 2016)"

The literature offers numerous reasons in support of a shift away from (frequentist) null hypothesis testing to estimation with uncertainty. In this section we focus on only three reasons: Null hypotheses are often false *a priori* (so it's pointless to test them), null-hypothesis tests ignore magnitude and uncertainty (which can lead to misinterpretation of results), and null hypothesis testing impedes meta-analysis and cumulative science.

When we use the term "hypothesis testing," we mean *point-value* hypothesis testing, for which the hypothesis being tested is a specific value of a parameter such as a null value, as in the examples presented earlier in the article. There are other types of non-point hypotheses that can be tested, but the typical routine hypothesis test (either frequentist or Bayesian) involves a point-null hypothesis.

Null hypotheses are often false *a priori*

One argument against using null hypotheses is that in many applications they are straw men that can be rejected with enough data. The premise of the argument is that most factors of interest have some non-zero relation or effect on other variables of interest, even if the effect is very small. With enough data to cancel out noise variance, any non-zero effect can be detected. Thus, if a theory merely posits *any* non-null effect, then the theory becomes easier to confirm

as more data are collected. But scientific theories should work the other way around: Because theoretical predictions almost surely deviate from reality by some amount however small, it should be easier to *dis*confirm a theory as data accumulate. This contradiction between the way science should work and the way traditional null-hypothesis testing does work is sometimes known as Meehl's paradox (Meehl, 1967; 1978; 1997).

A premise of the argument from Meehl's paradox is that most factors of interest have a non-zero effect, that is, some non-zero departure from the predicted value (e.g., null value) even if the discrepancy is small. To what extent is this premise true?

Some factors might plausibly have exactly zero effect, and theories about the factor might be committed to an exactly zero effect size. For example, we might assert that extra-sensory perception, in the sense of foretelling the future via temporally-backward causality, is theoretically impossible and therefore has exactly zero effect size. In this case, any deviation from zero, no matter how small, could be a major upheaval in current theories of physics. Various authors have discussed situations in which a point null is theoretically important to consider (e.g., Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Examples in cognitive psychology were provided by Lee and Wagenmakers (2014, Ch. 7). In these situations, when the null hypothesis is genuinely plausible, there are a variety of reasons to prefer Bayesian hypothesis testing over frequentist hypothesis testing, as was discussed previously.

But theories that are plausibly exactly correct (including an exactly correct null hypothesis) may be relatively rare. For example, Anderson, Burnham, and Thompson (2000, p. 913) stated that "The most curious problem with null hypothesis testing, as the primary basis for data analysis and inference, is that nearly all null hypotheses are false on *a priori* grounds (Johnson, 1995)." Anderson et al. (2000) surveyed 500 articles randomly sampled from several years of two respected journals. They found on average about two dozen reported p values per article. Despite the many thousands of hypothesis tests, "In the 347 sampled articles in *Ecology* containing null hypothesis tests, we found few examples of null hypotheses that seemed biologically plausible. Perhaps 5 of 95 articles in *Journal of Wildlife Management* contained $\geq$ 1 null hypothesis that could be considered ... plausible" (Anderson et al., 2000, p. 915). Some *a priori* false null hypotheses that have been tested in the ecology literature include "the occurrence of sheep remains in coyote (*Canis latrans*) scats differed among seasons (p = 0.03, n = 467), (2) duckling body mass differed among years ($p < 0.0001$), and (3) the density of large trees was greater in unlogged forest stands than in logged stands (p = 0.02)" (Johnson, 1999). We are not aware of analogous surveys of the literature in psychological sciences, but we strongly suspect there would be analogous results. The general point about straw-man null hypotheses has been made for decades by many authors (e.g., Savage, 1957).

In a Bayesian framework, the implausibility of a null hypothesis is expressed as a low prior probability of the null hypothesis relative to a high prior probability of the alternative hypothesis. The prior probabilities of the hypotheses were illustrated graphically in panel A of Fig. 6 as the heights of the bars on the model index M. To say that a null hypothesis is false *a priori* is to say that the height of the bar on the null model index $M = 1$ is infinitesimally small, and the height of the bar on the alternative model index $M = 2$ is essentially 1.0. Therefore the posterior probabilities of the models must favor the alternative (non-null) model regardless of the Bayes factor. This conclusion can be understood directly from Eq. 4: If the prior odds are zero then the posterior odds are zero, regardless of the Bayes factor. This type of argument also applies, in reverse, when the null hypothesis has an extremely high prior probability. For example, as was mentioned previously for the case of ESP, the null hypothesis of no effect has an extremely high prior probability and therefore the posterior probability of the alternative hypothesis is very small even if the Bayes factor strongly suggests a shift away from the null (Rouder and Morey, 2011; Rouder et al., 2013).

In a Bayesian context, putting a high prior probability on a model is not a claim that the model is true or correct in an ontological sense. Instead, Bayesian model probabilities are indicators of relative descriptive abilities within the limited space of models being considered. Thus, to say that the prior probability of the null model is virtually zero and the prior probability of the alternative model is virtually one is not to say that the alternative model is correct; rather, it says that within the limited space of models under consideration the null model is not a plausible description of the data relative to the other models.

Null hypothesis tests ignore magnitude and uncertainty

A more important general problem of point-value hypothesis testing is that the result of a hypothesis test reveals nothing about the magnitude of the effect or the uncertainty of its estimate, which we should usually want to know. Null hypothesis testing, in frequentist or Bayesian forms, has three undesirable consequences: (i) a null hypothesis can be rejected by a trivially small effect; (ii) a null hypothesis can be rejected even though there is high uncertainty in its magnitude; and (iii) a null hypothesis can be accepted by a Bayes factor even though the interval estimate of the magnitude includes a fairly wide range of non-null values.

The severity of problems (ii) and (iii) declines with more conservative decision criteria that demand larger N, but the problems can be dramatic when using the typical weak criterion of $p < .05$ in a frequentist test. A corresponding Bayes-factor criterion is roughly 3 (i.e., $BF_{null} > 3$ accepts the null hypothesis and $BF_{null} < 1/3$ accepts the alternative hypothesis), but there is no direct mathematical equivalence to $p < .05$ (Wetzels et al., 2011). While Dienes (2016) suggests that a BF of 3 is "substantial," other proponents of Bayesian hypothesis tests (e.g., Schönbrodt et al., 2016) recommend higher decision thresholds such as 6 or 10 depending on the nature of the research.

As examples of the three problems mentioned above, consider a single group of metric values sampled from a normal distribution. (i) Suppose the sample has $N = 1,200$ with sample mean of 101.5 and standard deviation of 15.0. Relative to a null hypothesis mean of 100.0, it turns out that $p = 0.001$ and the Bayes factor on the effect size (using the BayesFactor R package, Morey et al., 2015, with default alternative hypothesis rscale=0.707) is 0.08 with respect to the null hypothesis, that is, 12.5 in favor of the alternative hypothesis. Thus, the BF rejects the null hypothesis even with a fairly strict decision threshold of 10. But despite *rejecting* the null, the estimated effect size indicates only a small effect, with the 95 % HDI extending from 0.04 to 0.15 (which is less than a conventionally "small" effect size of 0.2; J. Cohen, 1988). Therefore, rejecting a null hypothesis by itself tells us nothing about the magnitude of the effect and whether the effect is eye-rollingly trivial or eye-poppingly whopping. (ii) Suppose the sample has $N = 20$, with sample mean of 111.0 and standard deviation of 15.0. Relative to a null hypothesis mean of 100.0, it is the case that $p = 0.005$ and the Bayes factor on the effect size is 0.10 (that is, 10.0 in favor of the alternative), rejecting the null hypothesis. Despite rejecting the null hypothesis, the magnitude of the effect is very uncertain, with the 95 % HDI extending from 0.20 to nearly 1.2, that is from "small" to very large. (iii) Suppose the sample has $N = 125$ and the sample mean exactly matches the null mean. Then the Bayes factor exceeds 10 in favor of the null hypothesis and the 95 % HDI on the effect size spans ± 0.18. While the width of the HDI is not enormous because the high criterion on the BF demanded a relatively large sample size, the HDI does suggest that a non-negligible range of non-zero effect sizes remains plausible. When lower decision criteria are used for the BF (e.g., 3.0 as used by Dienes, 2016), the HDI is noticeably wider. For these three generic illustrations we computed BFs using the BayesFactor R package (Morey et al., 2015) with default alternative hypothesis rscale=0.707. In applied research, it would be important to use meaningfully informed alternative priors, which would result in different values for the BF's (Dienes, 2014; Kruschke, 2011a; Vanpaemel and Lee, 2012).

One of the key problems with null-hypothesis testing is that it easily leads to fallacious "black and white thinking" that ignores the magnitude and uncertainty of the effect. When a null hypothesis is *not* rejected (by a p value) or even when a null hypothesis is accepted (by a Bayes factor), people can mistakenly believe there is zero effect even though the results may actually allow for a reasonable range of non-zero effect magnitudes. "... a non-significant result is often in practice taken as evidence for a null hypothesis. For example, to take one of the most prestigious journals in psychology, in the 2014 April issue of the Journal of Experimental Psychology: General, in 32 out of the 34 articles, a non-significant result was taken as support for a null hypothesis (as shown by the authors claiming no effect), with no further grounds given for accepting the null other than that the p value was greater than 0.05." (Dienes, 2016, p. 2) When a null hypothesis *is* rejected, people can mistakenly believe that the effect is solidly large, even though the results may actually allow for a large range of effect magnitudes, including very small effect sizes (e.g., Kline, 2004). On the other hand, if the estimated effect size and its interval of uncertainty are explicitly presented, we can judge not only whether the null value is near the interval of uncertainty, but we can also judge the importance of the effect size and how securely its magnitude has been ascertained.

Null hypothesis testing hinders cumulative science and meta-analysis

One of the main casualties inflicted by the black-and-white thinking of hypothesis testing is cumulative science. (Schmidt, 1996) clearly explained the problem and provided a compelling example, which we now briefly summarize. Suppose there is a real effect being investigated that has a moderate (non-null) effect size. We repeatedly generate experimental samples of data from that fixed non-null effect. Using typical low-power sample sizes, experiments will show that about one third produce significant results (with $p < .05$) and the other two thirds produce insignificant results (with $p > .05$). There are two traditional interpretations of the set of results. One interpretation tallies the black-and-white decisions, notes that the majority of studies found no effect, and concludes therefore that there is no effect. This conclusion is wrong, of course, because all the data were generated by the same non-null effect size. The second traditional interpretation notes that while the majority of studies found no effect, some of the studies did find an effect, and therefore follow-up research must discover moderator variables that cause the effect to appear sometimes and not other times. This conclusion is also wrong, because the data for all studies were generated the same way. Schmidt (1996, p. 126) concluded that "Researchers must

be disabused of the false belief that if a finding is not significant, it is zero. This belief has probably done more than any of the other false beliefs about significance testing to retard the growth of cumulative knowledge in psychology."

An emphasis on null-hypothesis testing impedes cumulative science another way, by biasing which data get published. In the example of the previous paragraph, all of the (simulated) studies were considered, including all the studies that produced non-significant results. But scientific journals (i.e., reviewers and editors) are reluctant to publish non-significant results (e.g., Rosenthal, 1979). Therefore what actually gets published is a biased and unrepresentative sample. This bias has been pointed out many times, and recently by Cumming (2014, p. 22), who said, "Indeed, NHST has caused some of its worst damage by distorting the results of meta-analysis, which can give valid results only if an unbiased set of studies is included; ... selective publication biases meta-analysis." Dienes (2016) recommended basing publication on Bayesian null-hypothesis testing when a decision is reached in either direction, so that both rejected and accepted null hypotheses would be published. To discourage black and white thinking, we further recommend an emphasis on effect size and uncertainty. As is discussed later in the article, if publication were based on achieving a reasonable degree of precision (and assuming that precise estimates are accurate estimates, which is usually the case), regardless of whether a null hypothesis is rejected or accepted or undecided, then there would not be bias in published data.

From Bayesian null hypothesis testing to Bayesian estimation

There are some clear advantages of Bayesian hypothesis testing over NHST. Most prominently, because an explicit alternative hypothesis is posited, the Bayesian approach can produce the relative probabilities of hypotheses, unlike NHST which only yields the probability of simulated data from the null hypothesis (as was expressed in Eq. 1). In particular, Bayesian hypothesis testing can indicate that the null hypothesis is more credible than the alternative hypothesis, which NHST can never do. This is highly desirable for theoretical domains in which "proving" the null is the goal (e.g., Dienes, 2014, 2016; Gallistel, 2009; Lee & Wagenmakers, 2014; Rouder et al., 2009; Wagenmakers, 2007).

But because the Bayes factor does not reveal magnitude and uncertainty, it is easy for the meaning to slip away and for only the dichotomous accept/reject decision to remain. Gigerenzer and Marewski (2015, p. 423, 437) warned against default Bayes factors becoming the same "mindless null ritual" as NHST: "The automatic calculation of significance levels could be revived by similar routines for Bayes factors. That would turn the [Bayesian] revolution into a re-volution — back to square one. ... The real challenge in our view is to prevent ... replacing routine significance tests with routine interpretations of Bayes factors." Cumming (2014, p. 15) made a similar remark: "Bayesian approaches to estimation based on credible intervals, to model assessment and selection, and to meta-analysis are highly valuable (Kruschke, 2011b). I would be wary, however, of Bayesian hypothesis testing, if it does not escape the limitations of dichotomous thinking."

Beyond the general problems, Bayesian hypothesis testing has some problems unique to its formulation. First, the BF is extremely sensitive to the choice of prior distribution for the alternative hypothesis. Therefore in realistic application it is important to use a theoretically meaningful and informed distribution for both the prior and alternative hypotheses, not merely generic defaults, and it is important to check that the BF does not change much if the prior distributions are changed in reasonable ways (e.g., Dienes, 2014, 2016; Kruschke, 2011a; Lee & Wagenmakers, 2014; Vanpaemel & Lee, 2012). Second, the BF does not indicate the posterior odds, and users must remember to take into account the prior odds of the hypotheses. If the null hypothesis has a minuscule prior probability, then BF_{null} must be enormous to compensate and produce a posterior probability that favors the null. Of course that reasoning goes the other way, too, so that if the null hypothesis has an enormous prior probability, then BF_{null} must be exceedingly tiny to compensate and produce a posterior probability that favors the alternative.

In summary, in those situations that you really want to test a null hypothesis, it is better to do it by Bayesian model comparison than by frequentist NHST because the Bayesian approach can provide information about the relative probabilities of hypotheses. For Bayesian hypothesis testing, it is important to use prior distributions that are theoretically meaningful and that previous data could have generated; be wary of using a generic default prior. (Bayesian estimation, on the other hand, is typically far less sensitive to default priors.) It is important to incorporate the prior probabilities of the hypotheses and not rely only on the Bayes factor to make decisions. Perhaps most importantly, do not fall into black-and-white thinking; also estimate the magnitude and uncertainty of the parameters with a goal of precision and a meta-analytic thinking.

Meta-analysis, randomized controlled trials, and power analysis: Better done Bayesian

The previous sections have shown that the usual goals of hypothesis testing and estimation *in general* are better achieved with a Bayesian approach than with a frequentist

approach. The following sections visit three other specific emphases of the "New Statistics" (Cumming, 2014): Meta-analysis, randomized controlled trials, and power analysis.

Meta-analysis: Better done Bayesian

To discuss meta-analysis, we first discuss hierarchical models, because models for meta-analysis are a case of hierarchical models. Hierarchical models are natural descriptions for many kinds of data. For example, in a variety of studies, there might be many individuals in each treatment, and each individual provides many data values. In these cases, the model has descriptive parameters for each individual and higher-level parameters that describe the distribution of individual-level parameter values. As another example, consider a study that includes many different predictors in a large multiple regression. The distribution of regression-coefficient values across predictors can be described by a higher-level distribution.

Hierarchical models are especially useful because the low-level and high-level parameters are estimated simultaneously and are mutually constraining. When data from many low-level units inform the high-level distribution, the high-level distribution constrains the low-level parameters to be mutually consistent. This causes the more extreme low-level cases to be "shrunken" toward the mode(s) of the group. Shrinkage helps prevent false alarms caused by random conspiracies of rogue outlying data. Essentially, the data from the other individuals are acting as simultaneous prior information to rein in estimates of outlying individuals. Bayesian hierarchical models are explained by Kruschke and Vanpaemel (2015) and Ch. 9 of Kruschke (2015), among other resources.

HDI is seamless for complex hierarchical models, unlike the CI Bayesian methods are especially convenient for complex hierarchical models for two reasons. First, the interpretation of the parameters is seamless because we simply read off whatever we want to know from the posterior distribution, including modal parameter values and their HDIs. There is no need for auxiliary assumptions for generating p values and confidence intervals from sampling distributions. (In frequentist analyses of hierarchical models, p values and confidence intervals are typically only roughly approximated.) Second, modern Bayesian software allows very flexible specification of complex hierarchical models. For example, attentional parameters in a cognitive model of classification can be useful to characterize what individual respondents attend to in a classification task. The cognitive model can be easily implemented in Bayesian software and its parameters estimated. Kruschke and Vanpaemel (2015) reported estimates of attentional allocation by individual human subjects in a classification task, grouped by clinical symptoms. The estimates of individuals showed shrinkage toward two modes of a bimodal group distribution, in which some subjects paid most attention to one stimulus dimension while other subjects paid most attention to another stimulus dimension. Another example of the flexibility of Bayesian hierarchical modeling is a hierarchical conditional-logistic regression model for describing behavior of players in a public goods game who chose different types of punishments to apply to free loaders (Liddell & Kruschke, 2014). Every subject's multinomial choice data were modeled with conditional-logistic regression. The distribution of individuals' regression coefficients was modeled by a heavy-tailed t-distribution, which accommodated some widely outlying individuals without distorting the estimates of typical respondents. Parameter estimates and HDIs are produced seamlessly for these complex hierarchical models. It would be difficult to generate accurate CIs for these models in a freqentist framework.

Various software packages have the abilities we have been touting. The most popular, and free of charge, are JAGS (Plummer, 2003, 2012), BUGS (2013), and Stan (Stan Development Team, 2012). A thorough introduction is provided by Kruschke (2015).

Meta-analysis as Bayesian hierarchical modeling A core premise of meta-analytic thinking is that "Any one study is most likely contributing rather than determining; it needs to be considered alongside any comparable past studies and with the assumption that future studies will build on its contribution" (Cumming, 2014, p. 23). That premise is based on the fact that any one study is merely a finite random sample of data, and different random samples will show different trends due to random variation. Therefore, we should combine the results of all comparable studies to derive a more stable estimate of the true underlying effect.

Describing variation of data *across* studies is a modeling problem, just like describing variation of data *within* a study is a modeling problem. The structure of data across multiple studies is naturally described by a hierarchical model: Each study has individual parameters, and a higher-level distribution describes the variation of those parameters across studies. The top-level distribution describes the central tendency of the trend across studies, and the uncertainty of that trend. This type of hierarchical structure is often referred to as a *random-effects model*, with the idea being that each individual study has its own effect that is randomly selected from the overarching population. "... what we should routinely prefer ... is the random-effects model, which assumes that the population means estimated by the different studies are randomly chosen from a superpopulation" (Cumming, 2014, p. 22)

As an example of Bayesian meta-analysis, consider data from 22 clinical trials of beta-blockers for reducing mortality after myocardial infarction, from Yusuf et al. (1985) and reported in Gelman et al. (2013, Sec. 5.6). In each of the studies, heart-attack patients were randomly assigned to a control group or a group that received a heart-muscle relaxant called a beta blocker. For each group, the number of patients and the number of deaths was recorded. Across studies, group size ranged from about 40 to 1,900. The typical proportion of deaths in the control group, $\text{deaths}_{\text{control}}/\text{patients}_{\text{control}}$, was just under 9 %, but varied across studies.

The underlying probability of death in the control group of study s is denoted $\theta_{C[s]}$, and the underlying probability of death in the treatment group of study s is denoted $\theta_{T[s]}$. The difference in probability of death between the two groups is indicated by the *log odds ratio*,

$$\rho_{[s]} = \text{logit}\left(\theta_{T[s]}\right) - \text{logit}\left(\theta_{C[s]}\right) \tag{7}$$

where the logit function is the inverse logistic function, and logit is also called the *log odds*. (The parameter ρ is called the log odds ratio because is can be re-written as $\rho = \log\left([\theta_T/(1-\theta_T)]/[\theta_C/(1-\theta_C)]\right)$, which is the logarithm of the ratio of the odds in the two groups.) The log odds ratio is a natural way to express the difference between the groups because ρ is symmetric with respect to which outcome (e.g., heart attack or no heart attack) is the event being counted and with respect to which group is the target group, by merely changing the sign of ρ. A key feature of Eq. 7 is that it can be re-arranged to express the dependency of the death rate in the treatment group on the death rate in the control group:

$$\theta_{T[s]} = \text{logistic}\left(\rho_{[s]} + \text{logit}\left(\theta_{C[s]}\right)\right) \tag{8}$$

The idea expressed by Eq. 8 is that the treatment effect $\rho_{[s]}$ shifts the rate of occurrence relative to the control rate.

The goal in meta-analysis is to combine the information across studies to derive a more precise and stable estimate of the relative risk of heart attack when using beta blockers. We assume that each study is representative of (i) an underlying probability of heart attack across the control groups and (ii) an underlying effect of treatment. In particular, we model the distribution of treatment effects across studies as a normal distribution with mean μ_ρ and standard deviation σ_ρ. Both of these parameters are estimated from the data. Analogously, we model the distribution of control rates as a beta distribution with mode ω_C and concentration κ_C. The computer programs are available at https://osf.io/j6364/, and more details can be found at http://tinyurl.com/BayesMetaTwoProp[3] Our primary interest is the magnitude and uncertainty of the overall treatment effect as expressed by the parameter μ_ρ, but we are also interested in the estimated effect for each study as expressed by $\rho_{[s]}$.

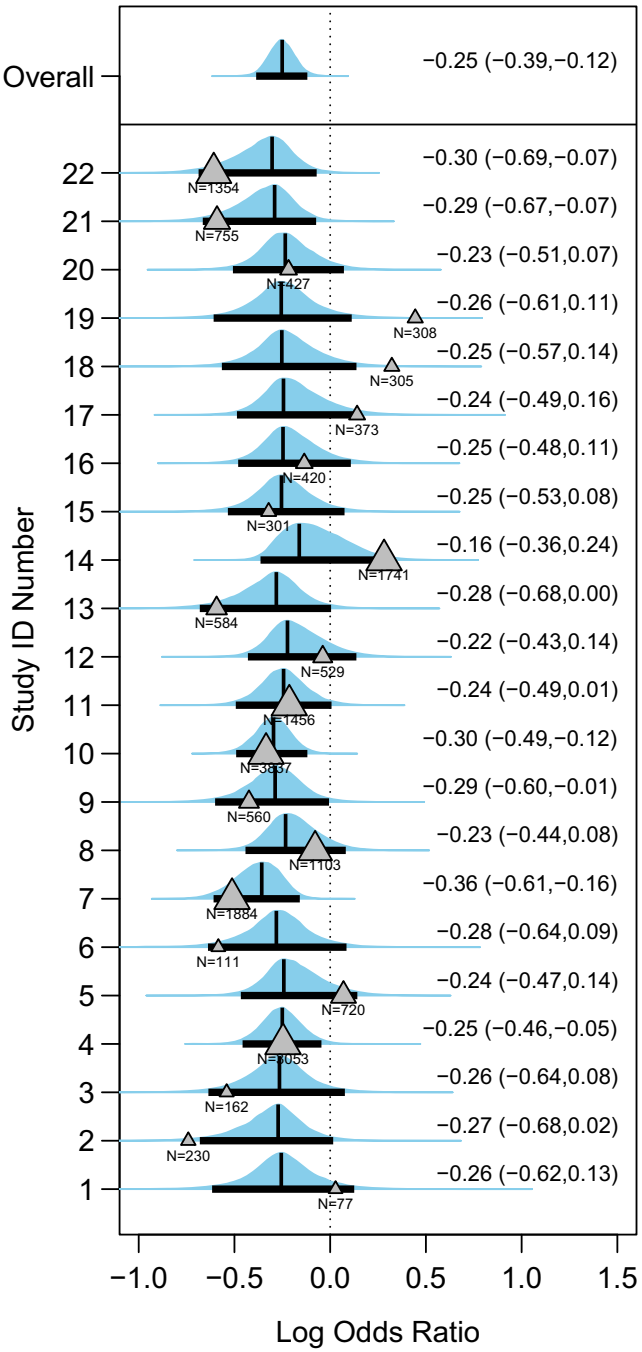

Fig. 8 Meta-analysis of effect of beta-blockers on heart-attack fatalities. The *horizontal axis* shows the effect of beta-blockers expressed by the value of the parameter ρ described in Eqs. 7 and 8. The *vertical axis* shows the study ID number and the overall estimate at top. For each study, the *triangle* shows the study-specific sample value of ρ along with total sample size (N) in the study. The distributions show the posterior distribution of each parameter, with a *horizontal bar* marking the 95 % HDI and a *vertical line* marking the mode of the distribution. Numerical values of the mode and 95 % HDI are displayed on the *right side*

[3]The full URL is http://doingbayesiandataanalysis.blogspot.com/2016/11/bayesian-meta-analysis-of-two.html.

Anything we want to know about the parameter estimates can be directly read off the full posterior distribution, and without any need for additional assumptions to create sampling distributions for p values and confidence intervals. Figure 8 shows a forest plot of the results. The top row shows the posterior distribution of the overall effect μ_ρ, and the lower rows show the posterior distributions for the study-specific effects $\rho_{[s]}$. Each distribution is marked with a horizontal bar that indicates the 95 % HDI, and with a vertical line that marks its mode. The numerical values of the mode and 95 % HDI limits are displayed on the right side of the graph. The graph also plots the study-specific sample values of ρ as triangles, with the size of the triangle indicating the sample size of the study. The top of Fig. 8 indicates that overall, across studies, the log odds ratio has a modal estimate of −0.25, with a 95 % HDI from −0.38 to −0.11, and the distribution is clearly well below zero.

The meta-analysis simultaneously estimates the effects in each of the 22 studies. Because of the hierarchical structure of the random-effects model, the estimate of each individual study is informed by the data from the other 21 studies. The 22 studies inform the overarching estimate, which in turn shrinks the individual study estimates toward what is typical across studies. The posterior distributions of trial-specific $\rho_{[s]}$ in Fig. 8 reveal the shrinkage. For example, the posterior distribution for $\rho_{[22]}$ is noticeably skewed to the left, and the posterior distribution for $\rho_{[14]}$ is noticeably skewed to the right. Shrinkage is strong, as indicated dramatically by the posterior 95 % HDI of $\rho_{[s]}$ for many studies not even including the triangle that indicates the sample value of $\rho_{[s]}$ (e.g., studies 2, 18, etc.).

Brophy et al. (2001) pointed out that Bayesian meta-analysis also seamlessly produces a posterior distribution on the difference of probabilities of heart attack in the two groups, as shown in Fig. 9. In the model, the over-arching probability of death in the control group is denoted ω_C, and the over-arching probability of death in the treatment group is denoted ω_T. The left panel of Fig. 9 shows the joint posterior distribution of ω_C and ω_T. The right panel of Fig. 9 shows the marginal difference of the probabilities, $\omega_T - \omega_C$,

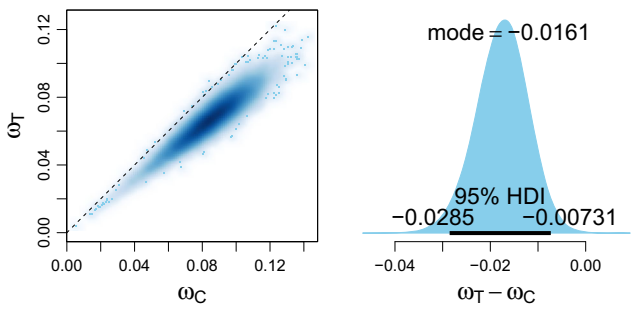

Fig. 9 Posterior distribution on overall probability of death in treatment and control groups

which has a modal values of about −0.016 with 95 % HDI from about −0.028 to −0.007, that is, about 1.6 lives saved per 100 heart attacks. This result provides clinicians with important information to decide if the number of lives saved is worth the cost and possible side effects of the treatment.

In general, Bayesian methods are well suited for meta-analytic modeling because meta-analytic models are a kind of hierarchical model and Bayesian methods are exceptionally useful for hierarchical models. The Bayesian approach makes it especially easy and direct to interpret the results, without need for auxiliary assumptions and approximations for constructing confidence intervals or p values. Moreover, software for Bayesian estimation makes it straightforward to set up complex hierarchical models. Pitchforth and Mengersen (2013, p. 118) said that "Bayesian meta-analysis has advantages over frequentist approaches in that it provides a more flexible modeling framework, allows more appropriate quantification of the uncertainty around effect estimates, and facilitates a clearer understanding of sources of variation within and between studies (Sutton and Abrams, 2001)."

For some textbook treatments of Bayesian meta-analysis, see Berry, Carlin, Lee, and Müller (2011, Sec. 2.4), Gelman et al. (2013, Sec. 5.6), and Woodworth (2004, Ch. 11). A brief example of Bayesian meta-analysis applied to smoking and lung-cancer was presented by Ntzoufras (2009, Sec. 9.2.4). Pitchforth and Mengersen (2013) gave examples using the software WinBUGS, based on the precedent provided by Sutton and Abrams (2001) who gave an example of a three-level hierarchical meta-analysis. A mathematical discussion of Bayesian meta-analysis was given by Hartung et al. (2008).

Randomized controlled trials: Better done Bayesian

Randomized controlled trials (RCTs) are an important experimental design that can present challenges for generating and interpreting confidence intervals, as was emphasized by Cumming (2014, p. 19) in his article about the New Statistics. In this section we show that analysis of RCTs is seamless and straightforward in a Bayesian framework.

In an RCT, subjects are randomly assigned to a control condition or a treatment condition. There could be several different types of control conditions (e.g., do nothing or administer placebo) and several different types of treatment condition. These conditions constitute a factor with conditions that vary between subject. Within each condition, every subject might be measured on a series of occasions or in a series of orthogonal within-subject conditions, such as pre-treatment, post-treatment, follow-up 1, and follow-up 2. This particular type of design, which crosses a within-subject factor with a between-subject factor, is called a *split-plot* design. Split-plot designs are discussed in a frequentist

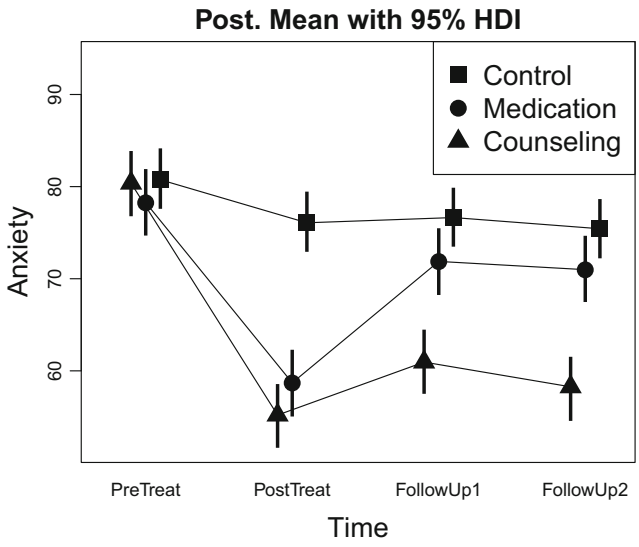

Fig. 10 Fictitious data for a randomized controlled trial (RCT) in a split-plot design. The within-subject factor (Time: Pre-Treatment, Post-Treatment, Follow-Up 1, and Follow-Up 2) is plotted on the *horizontal axis*. Each level of the between-subject factor (Treatment: Control, Medication, and Counseling) is marked by a separate *curve*. Symbols show posterior estimated cell means with 95 % HDIs

framework by Maxwell and Delaney (2004, Ch. 12) and Howell (2013, Sec. 14.7), and in a Bayesian framework by Kruschke (2015, Ch. 20).

As an example of a RCT with a split-plot structure, consider the fictitious data summarized in Fig. 10. The structure of the scenario is analogous to the one presented by Cumming (2014, p. 19). We consider two treatments for anxiety, viz. medication and counseling, and a control treatment (e.g., no intervention). Every subject had their anxiety measured at four times: pre-treatment, post-treatment, follow-up 1, and follow-up 2. There were 15 subjects in the control condition, 12 in medication, and 13 in counseling. Figure 10 shows the overall trends, where it can be seen that all three groups had similar pre-treatment anxiety levels. After treatment, anxiety levels in the medication and counseling treatments appear to be lower than in the control condition, but anxiety levels seem to rise during follow-up after medication. Please note that these are completely fictitious data, merely for illustration of analysis methods.

We did a Bayesian analysis of the data, described by a model that is directly analogous to the usual frequentist analysis-of-variance (ANOVA) model. Our model also had a hierarchical structure that imposed modest shrinkage on the effects; details are explained in Ch. 20 of Kruschke (2015). The computer programs are available at https://osf.io/j6364/. We assumed that the residual variance was normally distributed and homogeneous across cells (which was true for the program that generated the simulated data). It is easy to relax these assumptions in Bayesian software to accommodate outliers and heterogeneous variances, unlike in frequentist approaches.

The Bayesian analysis produces a posterior distribution over a joint space of 60 parameters, including a baseline (one parameter), a main effect of treatment (three parameters), a main effect of time (four parameters), an interaction of treatment by time ($3 \times 4 = 12$ parameters), and separate additive effects for each individual subject (40 parameters), that simultaneously respect the sum-to-zero constraints of the ANOVA-like model (see Section 20.5.2 of Kruschke, 2015). The 60-dimensional posterior distribution provides the relative credibility of all possible parameter-value combinations. In particular, the Bayesian analysis yields a posterior distribution on the estimated cell means, as summarized in Fig. 10. Each point shows the central tendency of the posterior distribution on the cell mean, $\mu_{\text{treat,time}}$, and the vertical bar through the point shows the 95 % HDI of the cell mean.

The structure of the split-plot RCT in Fig. 10 suggests many interesting comparisons. We might be interested in comparisons across levels of the between-subject factor (Treatment), or comparisons across levels of the within-subject factor (Time), or interaction contrasts that assess how much the effect of one factor changes across levels of the other factor, or "simple" contrasts within single levels of either the within-subject or between-subject factor. All of these comparisons can be simply "read off" the posterior distribution. If we are interested in the difference between $\mu_{\text{treat1,time1}}$ and $\mu_{\text{treat2,time2}}$, we just look at the posterior distribution of the difference, $\mu_{\text{treat1,time1}} - \mu_{\text{treat2,time2}}$, which is directly computable from the *joint* distribution across the parameter space.

Figure 11 shows many such comparisons, including marginal contrasts on the between-subject factor, marginal contrasts on the within-subject factor, interaction contrasts for particular combinations of levels of each factor, and "simple" comparisons of cells within levels of one factor or the other. All of these comparison are computed merely by looking at the corresponding difference of cell means in the joint posterior distribution.

It can be challenging to conduct these comparisons in NHST because each type of test requires the choice of an appropriate error term for the test's F ratio. "You can test the effects of the [within subjects] factor separately for each group or test the effects of the [between-subjects] factor separately at each level of the [within-subjects] factor or conduct tests in both ways... . Unfortunately the interaction of the two factors in a mixed design complicates the choice of an error term for each of the simple effects... . (B. H. Cohen, 2008, p. 549)" The choice of error term influences the p value and confidence interval for the comparison. The frequentist CI for comparisons tends to be a bit wider than the Bayesian HDI. For example, the 95 % CI for panel A of Fig. 11 extends from 7.6 to 13.4, as opposed to the 95 %

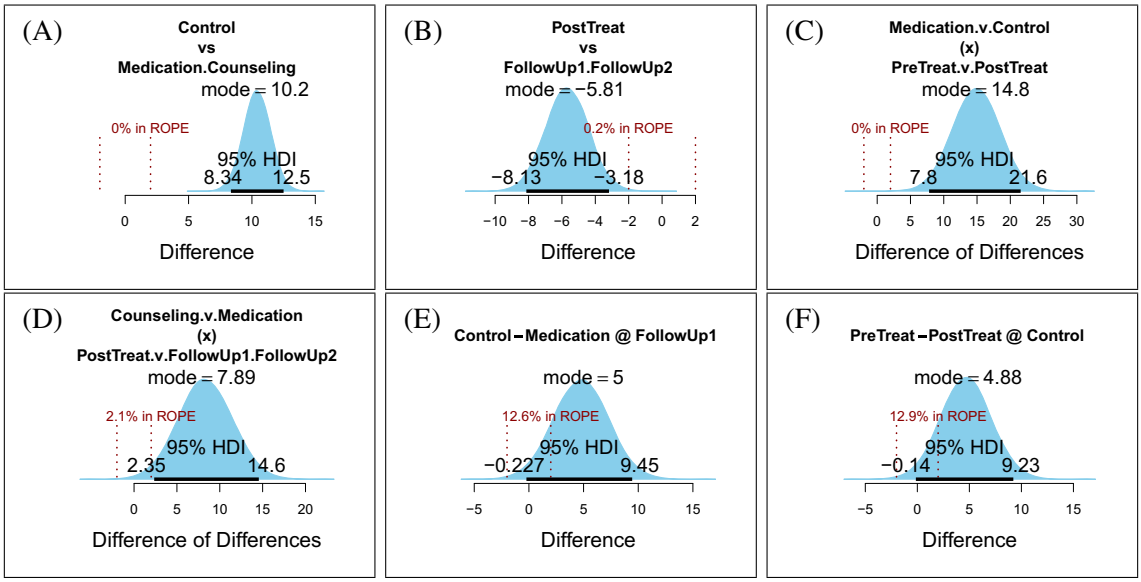

Fig. 11 Posterior distributions of selected comparisons for data in the split-plot RCT shown in Fig. 10. The *horizontal axes* indicate differences in anxiety-scale units. A region of practical equivalence (ROPE) is marked around zero difference, at ±2 units on the anxiety scale. The ROPE here was chosen arbitrarily for purposes of illustration; a ROPE should be based on the clinical significance of differences on the anxiety scale. **A** Marginal contrast on the between-subject factor. **B** Marginal contrast on the within-subject factor. **C**, **D** Interaction contrasts. **E** Simple between-subject contrast at a fixed level of the within-subject factor. **F** Simple within-subject contrast at a fixed level of the between-subject factor. *HDI* highest density interval

HDI which extends from about 8.3 to 12.5. Furthermore, the magnitudes of the Bayesian estimates have been shrunken a bit by the hierarchical model. For example, the 95 % CI for the interaction contrast in panel C of Fig. 11 extends from 9.1 to 23.1 around a central estimate of 16.1, as opposed to the 95 % HDI which extends from about 7.8 to 21.6 around a central estimate of about 14.8. For the interaction contrast in panel D, the 95 % CI extends from 2.87 to 15.37 around a central estimate of 9.12, as opposed to the 95 % HDI which extends from about 2.3 to 14.6 around a central estimate of 7.9.

Panels E and F of Fig. 11 show two "simple" contrasts of individual cells within a level of the between-subjects factor or within a level of the within-subjects factor. The posterior distribution of the differences was computed the same way as all the other Bayesian contrasts, merely by computing the difference of the cell means at the cells of interest. In NHST, however, these contrasts require consideration of different error terms. "However, because the sphericity assumption is generally considered quite risky, especially for pairwise comparisons, it is strongly recommended that you base your error term only on the two levels being tested. (This is equivalent to performing a simple matched t test between a pair of [within-subject] levels for one of the groups.)" (B. H. Cohen, 2008, p. 550) Again the Bayesian HDI's are tighter than the frequentist CI's: For panel E, the 95 % CI extends from −1.66 to 10.89 while the 95 % HDI extends from about −0.2 to 9.5, and for panel f the 95 % CI extends from −0.91 to 9.05 while the 95 % HDI extends from about −0.1 to 9.2.

A frequentist would also want to correct the p values and confidence intervals because of doing multiple comparisons. As described in the context of Fig. 2, when more tests are conducted, the cloud of simulated test statistics expands, consequently enlarging the p value for every test. The exact correction would depend on the particular set of tests being conducted. Bayesian analysis, on the other hand, does not set decision thresholds on the basis of false-alarm rates. Instead, Bayesian analysis considers only the posterior distribution based on the actual data. False alarms can still occur, of course, because false alarms are caused by rogue data. But Bayesian analysis can attenuate false alarms in ways other than corrections on p values. In particular, the model used here has hierarchical structure that imposes data-driven shrinkage across levels of a factor, which reins in outlying cells.

In summary, Bayesian analysis yields a posterior distribution from which the answer to any comparison can be directly viewed, even in complex split-plot RCT designs. Frequentist analyses, on the other hand, require careful selection of appropriate error terms for F ratios and corrections for multiple comparisons.

Planning for precision and other goals: Better done Bayesian

When the focus of data analysis is on null-hypothesis testing, then the goal of research is to reject or to accept a null

hypothesis. But when the focus of data analysis is on estimation, uncertainty, and meta-analysis, then a natural goal for research is to achieve precision of estimation. The difference in the goal is important because the goal determines how we plan the research. Traditional planning involves computing the probability of achieving the goal of rejecting the null hypothesis, which is called statistical *power*. But we can instead compute the probability of achieving the goal of a precise estimate. This is called planning for precision or for accuracy in parameter estimation (AIPE; Kelley, 2013; Maxwell, Kelley, & Rausch, 2008). For many model parameters, precision achieves accuracy, so we will use the term precision instead of the term accuracy.

The procedure for traditional power analysis is diagrammed in the upper part of Fig. 12. The analyst hypothesizes a specific point value for a parameter such as effect size. This point value is supposed to be the analyst's best guess for the true, non-null effect size. Then random samples of simulated data are generated from the hypothesis. Every sample of simulated data is created according to the intended stopping rule that will be used for the real data. For example, the stopping rule could be a fixed sample size, N. Because of random variability in the sample, only some of the simulated samples will have data extreme enough to reject the null hypothesis. The probability of rejecting the null hypothesis is called the traditional "power" for the choice of sample size and posited effect size. If the power is not big enough, then a larger N is considered.

In frequentist approaches to planning for precision, a similar scheme is used, but the goal is to achieve a width of CI that is less than some maximum. For a particular sample size N, we compute the probability that the width of the CI will not exceed the desired maximum. If the probability is not big enough, then a larger N is considered.

A key problem with the frequentist approach to power, or the probability of achieving any goal, is that the hypothesized value is punctate, that is, only one specific point value. This assumption conflicts with the fact that we do not know the exact value of the parameter that best describes the world. If we did know the exact value, then we would not be doing the research. Our hypothesis would be better represented as a distribution across possible parameter values. A Bayesian approach naturally takes this uncertainty into account.

The Bayesian procedure for computing the probability of achieving any goal is diagrammed in the lower part of Fig. 12. The procedure begins by hypothesizing a probability distribution over parameter values. Typically, the hypothesized distribution is simply the posterior distribution from a Bayesian analysis of previous data. The previous data could be real data or idealized data. This approach is particularly appealing because it is often easier for theorists to hypothesize idealized data than to hypothesize probability distributions on complex parameter spaces. That is, a theorist can have an informed idea of what the data in a proposed study should look like without knowing what those data imply for a multi-dimensional parameter distribution in a complex model. The amount of data expresses the degree of certainty, because the posterior distribution becomes narrower (more certain) as the amount of data increases. The theorist has real or idealized data, and then a Bayesian analysis of the data produces a corresponding distribution over parameter values that expresses the corresponding hypothesis for power analysis.

With the hypothetical distribution over parameters in place, the next step is to randomly sample a set of representative parameter values, as indicated in the lower part of Fig. 12. This sampling of parameter values is already

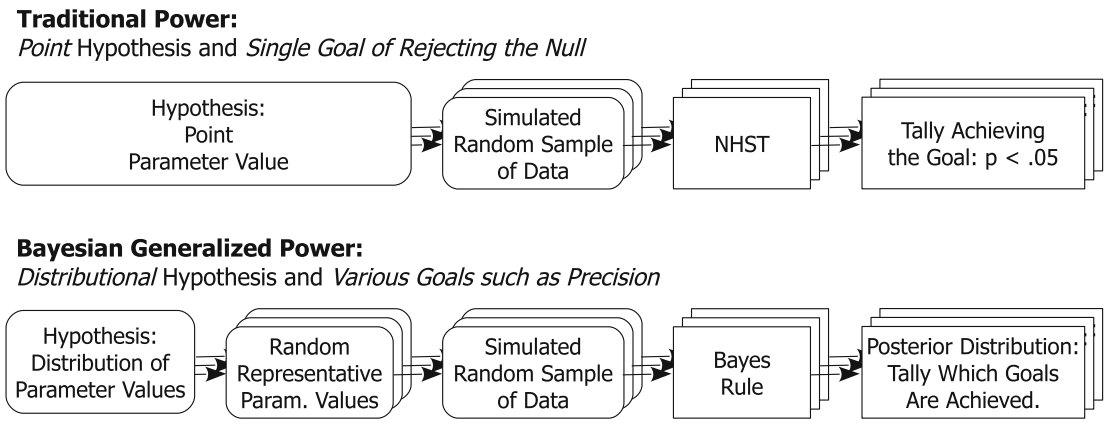

Fig. 12 Flowcharts for procedures in traditional power analysis (*upper chart*) and Bayesian generalized power (*lower chart*). In the Bayesian approach, the hypothesized distribution over parameter values is usually the posterior distribution from previous data, either real or idealized (*NHST* null hypothesis significance test)

provided directly by the Bayesian computational method called Markov chain Monte Carlo (MCMC), which is routinely used in modern software. The parameter values are then used to generate simulated data. Finally, the simulated data are given a Bayesian analysis just as the real data would be, and the posterior distribution is checked for whether the goal is achieved. The process repeats, and the proportion of times that the goal is achieved is the estimate of its probability, that is, the power of the study's procedure for that goal. This estimated probability takes into account the uncertainty in the hypothesis. The process is also completely general and can be used for any goal, and multiple goals can be checked simultaneously. The Bayesian approach works seamlessly for all models. The process shown in Fig. 12 is simulated directly in Bayesian software. Repeated random data sets are generated until the estimate of power is as precise as desired.

The goals of research are closely linked to criteria for publication. When the focus is on rejecting a null hypothesis, then rejection of the null often becomes the criterion for publication (and vice-versa: when rejection of the null is the publication criterion, then it becomes the focus of research). Unfortunately, that particular criterion creates biases in the data that get published because of the file-drawer problem. We might instead consider focussing on Bayesian hypothesis testing and judge that publication is merited whenever the null hypothesis is rejected *or accepted* (Dienes, 2016). That criterion for publication could usefully redress some of the biases caused by the file-drawer problem of NHST. Unfortunately a focus on hypothesis testing can engender fallacious black-and-white thinking and impede meta-analytic thinking. For example, many of the individual beta-blocker studies in the meta-analysis of Fig. 8 have default Bayes factors that neither reject nor accept the null hypothesis. (Bayes factors for the individual beta-blocker studies can be computed exactly using extensions of formulas described on pages 160 and 305 of Kruschke, 2011b.) What should be the publication status of indecisive studies? Should they remain in the file drawer, unpublished? Moreover, Bayes factors should use meaningfully informed priors instead of defaults, and the magnitude (and even direction) of a Bayes factor can vary substantially with the prior, and different studies might be analyzed with different alternative hypotheses. Instead of filtering publication according to whether or not a null hypothesis was rejected or accepted, publication might be based on whether or not a reasonable precision was achieved relative to the practical constraints of the study. The posterior precision is relatively invariant for vague priors. Basing publication on adequate precision will help solve the file-drawer problem because all suitably representative studies will get published, not just those studies that happen to sample enough data to reject or accept a null hypothesis relative to some particular alternative hypothesis.

Different procedures have different trade-offs, and the advantages or disadvantages of different procedures will continue to be discussed and clarified in coming years. What is clear is that the scheme in the lower part of Fig. 12 is more general, more appropriate, and more informative than the traditional frequentist scheme for power analysis. In particular, the general scheme represents the hypothesis as a (Bayesian posterior) distribution over parameter values instead of only as a point value, allows simulating any sampling procedure instead of only fixed N, and encourages considering multiple goals for research such as Bayesian hypothesis testing and Bayesian precision (HDI width) instead of only p values.

Bayesian power and planning for precision is discussed in the video at http://tinyurl.com/PrecisionIsGoal.[4] A working package and full explanation of Bayesian power and planning for precision for comparing two groups is detailed in Kruschke (2013). Details of the general procedure are explained at length in Chapter 13 of Kruschke (2015).

Summary and conclusion

Trafimow and Marks (2015) banned NHST from the journal *Basic and Applied Social Psychology*. In other words, the editors banned the methods in the top-left cell of Fig. 1. But Trafimow and Marks (2015) also expressed doubt about frequentist confidence intervals and Bayesian approaches involving Bayes factors. In other words, they expressed doubt about the lower-left and upper-right cells of Fig. 1. The remaining cell, at the convergence of Bayesian methods applied to estimation, uncertainty, and meta-analysis, was not mentioned by Trafimow and Marks (2015). We believe that this convergence alleviates many of the problems that the editors were trying to avoid.

Bayesian estimation can help us remember the dance

Cumming (2014) encouraged analysts to keep in mind that their particular set of data is merely one random sample, and another independent random sample might show very different trends. He called the random variation from sample to sample a "dance" and emphasized that CIs can change dramatically from one random sample to the next. In Bayesian analysis there is also a dance across different random samples. Indeed, the posterior distribution revealed by a Bayesian analysis is always explicitly conditional on the data, by definition: $p(\mu|D)$ in Eq. 5. If the data change, then the posterior distribution changes. There is a dance of HDIs.

[4]The full url is https://www.youtube.com/playlist?list=PL_mlm7M63Y7j641Y7QJG3TfSxeZMGOsQ4.

We think that the Bayesian framework is amenable to remembering the dance of random samples. This claim might seem implausible insofar as the Bayesian framework de-emphasizes thoughts of random samples by never using a hypothetical sampling distribution for computing p values (as in Fig. 2). On the other hand, a Bayesian framework does emphasize that the results are conditional on the particular (random) data obtained, and that the inference is merely the best we can do given the data we happen to have. Bayesian power analysis (Fig. 12) also explicitly simulates the dance of random samples. Perhaps most relevantly, the Bayesian framework is helpful for remembering the dance because it facilitates meta-analysis. The core premise of meta-analysis is acknowledging the variation of results across different samples.

Summary: Bayesian estimation does everything the New Statistics desires, better

We agree with many of the claims made by Cumming (2014) and others about the advantages of estimation and meta-analysis over null hypothesis tests, but we hope to have shown that when hypothesis testing is theoretically meaningful then it is more coherently done in a Bayesian framework than in a frequentist framework. We have attempted to show that Bayesian estimation, Bayesian meta-analysis, and Bayesian planning achieve the goals of the New Statistics more intuitively, directly, coherently, accurately, and flexibly than frequentist approaches.

Acknowledgments For comments on previous versions of this article, the authors gratefully acknowledge Geoff Cumming, Zoltan Dienes, Gregory Hickock, Michael D. Lee, Joachim Vandekerckhove, and an anonymous reviewer. Correspondence can be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, or via electronic mail to johnkruschke@gmail.com. More information can be found at http://www.indiana.edu/~kruschke/.

References

Allenby, G.M., Bakken, D.G., & Rossi, P.E. (2004). The hierarchical Bayesian revolution: How Bayesian methods have changed the face of marketing research. *Marketing Research*, *16*, 20–25.

Anderson, D.R., Burnham, K.P., & Thompson, W.L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management*, *64*(4), 912–923.

Beaumont, M.A., & Rannala, B. (2004). The Bayesian revolution in genetics. Nature Reviews Genetics.

Berry, S.M., Carlin, B.P., Lee, J.J., & Müller, P. (2011). *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC Press.

Brooks, S.P. (2003). Bayesian computation: A statistical revolution. *Philosophical Transactions of the Royal Society of London. Series A*, *361*(1813), 2681–2697.

Brophy, J.M., Joseph, L., & Rouleau, J.L. (2001). β-blockers in congestive heart failure: A Bayesian meta-analysis. *Annals of Internal Medicine*, *134*, 550–560.

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

Cohen, B.H. (2008). *Explaining psychological statistics*, 3rd edn. Hoboken, New Jersey: Wiley.

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

Cohen, J. (1994). The world is round (p <: 05), American Psychologist, 997–1003.

Cox, D.R. (2006). *Principles of statistical inference*. Cambridge, UK: Cambridge University Press.

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics*, *29*(3), 89–93.

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, *25*(1), 7–29.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie / Journal of Psychology*, *217*(1), 15–26.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 530–572.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. doi:10.1016/j.jmp.2015.10.003.

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

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

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

Gallistel, C.R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian data analysis third edition,3, Boca Raton, Florida, CRC Press.

Gigerenzer, G., & Marewski, J.N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., & Altman, D.G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. The American Statistician. Retrieved from doi:10.1080/00031305.2016.1154108#tabModule.

Gregory, P.C. (2001). A Bayesian revolution in spectral analysis, AIP (American Institute of Physics) Conference Proceedings, 568, 557. Retrieved from doi:10.1063/1.1381917.

Hartung, J., Knapp, G., & Sinha, B.K. (2008). Bayesian meta-analysis, Statistical Meta-analysis with Applications, 155–170, Hoboken, NJ, Wiley.

Hobbs, B.P., & Carlin, B.P. (2008). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, *18*(1), 54–80.

Howell, D.C. (2013). *Statistical methods for psychology 8th edition*, 8th edn. Wadsworth / Cengage Learning: Belmont, CA.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach*, 3rd edn. Open Court: Chicago.

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

Johnson, D.H. (1995). Statistical sirens: The allure of nonparametrics. *Ecology*, *76*, 1998–2000.

Johnson, D.H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management*, *63*, 763–772.

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

Kelley, K. (2013). Effect size and sample size planning. In Little, T.D. (Ed.) *Oxford Handbook of Quantitative Methods (Vols. Volume 1, Foundations, pp. 206–222)*. New York: Oxford University Press.

Kline, R.B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kruschke, J.K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

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

Kruschke, J.K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi:10.1037/a0029146.

Kruschke, J.K. (2015). *Doing Bayesian data analysis, Second Edition: A tutorial with R, JAGS, and Stan*. 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. doi:10.1177/1094428112457829.

Kruschke, J.K., & Liddell, T.M. (2015). Bayesian data analysis for newcomers. (in preparation).

Kruschke, J.K., & Vanpaemel, W. (2015). Bayesian estimation in hierarchical models. In Busemeyer, J.R., Townsend, J.T., Wang, Z.J., & Eidels, A. (Eds.) *Oxford Handbook of Computational and Mathematical Psychology*: Oxford University Press.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710.

Lazarus, R.S., & Eriksen, C.W. (1952). Effects of failure stress upon skilled performance. *Journal of Experimental Psychology*, *43*(2), 100–105. doi:10.1037/h0056614.

Lee, M.D., & Wagenmakers, E.-J. (2014). Bayesian cognitive modeling: A practical course, Cambridge, England, Cambridge University Press.

Lesaffre, E. (2008). Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU Hospital for Joint Diseases*, *66*(2), 150–154.

Liddell, T.M., & Kruschke, J.K. (2014). Ostracism and fines in a public goods game with accidental contributions: The importance of punishment type. *Judgment and Decision Making*, *9*(6), 523–547.

Lindley, D.V. (1975). The future of statistics: A Bayesian 21st century. *Advances in Applied Probability*, *7*, 106–115.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, Florida: CRC Press.

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

Maxwell, S.E., Kelley, K., & Rausch, J.R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

Mayo, D.G. (2016). Don't throw out the error control baby with the bad statistics bathwater: A commentary. The American Statistician. Retrieved from doi:10.1080/00031305.2016.1154108#tabModule.

Mayo, D.G., & Spanos, A. (2011). Error statistics. In Bandyopadhyay, P.S., & Forster, M.R. (Eds.) *Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics*, (pp. 153–198): Elsevier.

McGrayne, S.B. (2011). The theory that would not die, Yale University Press.

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

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of consulting and clinical Psychology*, *46*(4), 806.

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

Morey, R.D., Rouder, J.N., & Jamil, T. (2015). BayesFactor package for R. http://cran.r -project.org/web/packages/BayesFactor/index.html.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGHoboken S.* NJ: Wiley.

Pitchforth, J.O., & Mengersen, K.L. (2013). Bayesian meta-analysis, Case Studies in Bayesian Statistical Modelling and Analysis Alston, C.L., Mengersen, K.L., & Pettitt, A.N. (Eds.), Wiley.

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 (dsc 2003), Vienna, Austria, ISSN 1609-395X.

Plummer, M. (2012). JAGS version 3.3.0 user manual [Computer software manual].

Poole, C. (1987). Beyond the confidence interval. *American Journal of Public Health*, *77*(2), 195–199.

Rogers, J.L., Howard, K.I., & Vessey, J.T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565.

Rosenthal, R. (1979). The "file drawer problem"? and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Rothman, K.J. (2016). Disengaging from statistical significance. The American Statistician. Retrieved from 10.1080/00031305.2016.1154108#tabModule.

Rouder, J.N., & Morey, R.D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin and Review*, *18*, 682–689.

Rouder, J.N., Morey, R.D., & Province, J.M. (2013). A Bayes factor meta-analysis of recent extrasensory perception experiments: Comment on Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin*, *139*(1), 241–247.

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 and Review*, *16*, 225–237.

Sagarin, B.J., Ambler, J.K., & Lee, E.M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 293–304.

Savage, I.R. (1957). Nonparametric statistics. *Journal of the American Statistical Association*, *52*, 331–344.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129.

Schönbrodt, F.D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2016). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences, Psychological Methods. doi:10.1037/met0000061.

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680.

Schweder, T., & Hjort, N.L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics*, *29*, 309–332.

Serlin, R.C., & Lapsley, D.K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83.

Serlin, R.C., & Lapsley, D.K. (1993). Keren, G., & Lewis, C. (Eds.) *Rational appraisal of psychological research and the good enough principle*, (pp. 199–228). Hillsdale, NJ: Erlbaum.

Singh, K., Xie, M., & Strawderman, W.E. (2007). Confidence distribution (CD) distribution estimator of a parameter, Complex Datasets and Inverse Problems, 54, 132–150, Beachwood, OH, Institute of Mathematical Statistics Liu, R., et al. (Eds.)

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

Stan Development Team (2012). Stan: A C++ library for probability and sampling, version 1.1. Retrieved from http://mc-stan.org/citations.html.

Sullivan, K.M., & Foster, D.A. (1990). Use of the confidence interval function. *Epidemiology*, *1*(1), 39–42.

Sutton, A.J., & Abrams, K.R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*(4), 277–303.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2.

Vanpaemel, W., & Lee, M.D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin and Review*, *19*, 1047–1056.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*, 2nd edn. Boca Raton: Chapman and Hall/CRC Press.

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

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*(3), 291–298.

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 and Review*, *16*(4), 752–760.

Woodworth, G. (2004). Biostatistics: A Bayesian introduction, Wiley.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases*, *27*(5), 335–371.