Equivalent statistics and data interpretation

Gregory Francis^{1,2}

Published online: 14 October 2016
© Psychonomic Society, Inc. 2016

Abstract Recent reform efforts in psychological science have led to a plethora of choices for scientists to analyze their data. A scientist making an inference about their data must now decide whether to report a p value, summarize the data with a standardized effect size and its confidence interval, report a Bayes Factor, or use other model comparison methods. To make good choices among these options, it is necessary for researchers to understand the characteristics of the various statistics used by the different analysis frameworks. Toward that end, this paper makes two contributions. First, it shows that for the case of a two-sample t test with known sample sizes, many different summary statistics are mathematically equivalent in the sense that they are based on the very same information in the data set. When the sample sizes are known, the p value provides as much information about a data set as the confidence interval of Cohen's d or a JZS Bayes factor. Second, this equivalence means that different analysis methods differ only in their interpretation of the empirical data. At first glance, it might seem that mathematical equivalence of the statistics suggests that it does not matter much which statistic is reported, but the opposite is true because the appropriateness of a reported statistic is relative to the inference it promotes. Accordingly, scientists should choose an analysis method appropriate for their scientific investigation. A

direct comparison of the different inferential frameworks provides some guidance for scientists to make good choices and improve scientific practice.

Keywords Bayes factor · Hypothesis testing · Model building · Parameter estimation · Statistics

Introduction

It is an exciting and confusing time in psychological research. Several studies have revealed that there is much greater flexibility in statistical analyses than previously recognized (e.g., Simmons et al. 2011) and that such flexibility is commonly used (John et al., 2012; LeBel et al., 2013; O'Boyle Jr. et al., 2014) and potentially undermines theoretical conclusions reported in scientific articles (e.g., Francis 2012). The implication for many researchers is that the field needs to re-evaluate how scientists draw conclusions from scientific data. Much of the focus for reform has been directed at the perceived problems with null hypothesis significance testing, and a common suggestion is that the field should move away from a focus on p values and instead report more meaningful or reliable statistics.

These reforms are not small changes in reporting details. The editors of *Basic and Applied Social Psychology* discouraged and then banned the use of traditional hypothesis testing procedures (Trafimow & Marks, 2015) and instead require authors to discuss descriptive statistics. The journal *Psychological Science* encourages authors to eschew hypothesis testing and instead focus on estimation and standardized effect sizes to promote meta-analysis (Cumming, 2014; Eich, 2014). Many journals encourage researchers to design experiments with high power and thereby promote successful replications for real effects (e.g., the “statistical

✉ Gregory Francis
gfrancis@purdue.edu

¹ Department of Psychological Sciences, Purdue University,
703 Third Street, West Lafayette, IN 47907-2004, India

² École Polytechnique Fédéral de Lausanne, Brain Mind
Institute, Lausanne, Switzerland

guidelines” for the publications of the Psychonomic Society). As described in more detail below, Davis-Stober and Dana (2013) recommend designing experiments that ensure a fitted model can outperform a random model. Other scientists suggest that data analysis should promote model comparison with methods such as the difference in the Akaike Information Criterion (ΔAIC) or the Bayesian Information Criterion (ΔBIC), which consider whether a model’s complexity is justified by its improved fit to the observed data (Glover & Dixon 2004; Masson 2011; Wagenmakers 2007). Still other researchers encourage scientists to switch to Bayesian analysis methods, and they have provided computer code to promote such analyses (Dienes, 2014; Kruschke, 2010; Lee & Wagenmakers, 2013; Rouder et al., 2009).

With so many different ideas for improving statistical analyses, some scientists might be confused as to what they should report to convey the information in their dataset that is related to their theoretical ideas. Does a p value, Cohen’s d , confidence interval of Cohen’s d , ΔAIC , ΔBIC , or Bayes factor give the most accurate description of the data set’s information? To the casual reader, it might seem that proponents of a particular method claim their preferred summary statistic is the only one that makes sense and that all other statistics are inferior, worthless, or misrepresent properties of the data. The reality is rather different. Despite their legitimate differences, many of the proposed statistical methods are closely related and derive their properties from the very same information in a set of data. For the conditions that correspond to an independent two-sample t test with equal population variances (and the equivalent two-sample ANOVA, where $F = t^2$), many of the various statistics are mathematically equivalent representations of the information from the data set, and their corresponding analysis methods differ *only* in how they interpret that information. The 16 statistics in Table 1 all contain equivalent information about a data set; and for known sample sizes, it is possible to mathematically transform any given statistic to all of the others. A Web app that performs these transformations is available at <http://psych.purdue.edu/gfrancis/EquivalentStatistics/> and (much faster) R code (R Development Core Team, 2013) is at the Open Science Framework https://osf.io/vrenb/?view_only=48801b00b4e54cd595947f9215092c8d.

Readers intimately familiar with the statistics in Table 1 may not be surprised by their mathematical equivalence because many statistical analyses correspond to distinguishing signal from noise. For less-experienced readers, it is important to caution that mathematical equivalence does *not* imply that the various statistics are “all the same.” Rather, the equivalence of these statistics with regard to the information in the data set highlights the importance of the *interpretation* of that information. The inferences

derived from the various statistics are sometimes radically different.

To frame much of the rest of the paper, it is valuable to notice that some of the statistics in Table 1 are generally considered to be *inferential* rather than *descriptive* statistics, which reflects their common usage. Nevertheless, these inferential statistics can also be considered as descriptive statistics about the information in the data set. Although the various statistics are used to draw quite different inferences about that information, they all offer an equivalent representation of the data set’s information.

The next section explains the relationships between the statistics in Table 1 in the context of a two-sample t test. The subsequent section then describes relationships between the inferences that are drawn from some of these statistics. These inferences are often strikingly different even though they depend on the very same information from the data set. The manuscript then more fully describes the inferential differences to emphasize how they match different research goals. A final section considers situations where none of the statistics in Table 1 are appropriate because they do not contain enough information about the data set to meet a scientific goal of data analysis.

Popular summary statistics are equivalent

Statisticians use the term “sufficient statistic” to refer to a summary of a data set that contains all the information about a parameter of interest (many statistics textbooks provide a more detailed discussion about the definition, use, and properties of sufficient statistics; a classic choice is Mood et al. (1974)). To take a common example, for an independent two-sample t test with known sample sizes, a sufficient statistic for the population standardized effect size

$$\delta = \frac{\mu_2 - \mu_1}{\sigma}, \quad (1)$$

is the estimated standardized effect size, Cohen’s d :

$$d = \frac{\bar{X}_2 - \bar{X}_1}{s}, \quad (2)$$

which uses the estimated mean and pooled standard deviation from the samples. The statistic is “sufficient” because you would learn no more about the population standardized effect size if you had the full data set rather than just the value d (and the sample sizes). It is worth noting that the sign of d is somewhat arbitrary and that interpreting the effect size requires knowing whether the mean from group 1 was subtracted from the mean of group 2 (as above), or the other way around. Interpreting a sufficient statistic requires knowing how it was calculated. In particular, inferential statistics cannot be interpreted without knowledge beyond the information in the data set. The necessary additional

knowledge depends on the particular inference being made, which can dramatically alter the conclusions regarding the information in the data set.

A statistic does not have to be a good estimate of a parameter to be sufficient. Hedges (1981) noted that Cohen's d is biased to overestimate the population effect size when sample sizes are small. He identified an alternative effect size estimate that uses a correction factor to unbiased Cohen's d :

$$g = \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right) d. \quad (3)$$

More generally, any transformation of a sufficient statistic that is invertible (e.g., one can transform one statistic to the other and then back) is a sufficient statistic. This means that the familiar t value is also a sufficient statistic for the population standardized effect size because, for given sample sizes n_1 and n_2

$$t = d \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad (4)$$

and vice-versa, for a known t -value

$$d = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}}. \quad (5)$$

Thus, if a researcher reports a t value (and the sample sizes), then it only takes algebra to compute d or g , so a reader has as much information about δ as can be provided by the data set. It should be obvious that there are an infinite number of sufficient statistics (examples include $d^{1/3}$, $-(t + 3)^5$, and $2g + 7.3$), and that most of them are uninteresting. A sufficient statistic becomes interesting when it describes the information in a data set in a meaningful way or enables a statistical inference that extrapolates from the given data set to a larger situation. Table 1 lists 16 popular statistics that are commonly used for describing information in a data set or for drawing an inference for the conditions of a two-sample t test. As shown below, if the sample sizes are known, then these statistics contain equivalent information about the data set. The requirement of known sample sizes is constant throughout this discussion.

p value

Although sometimes derided as fickle, misleading, misunderstood, and meaningless (Berger and Berry, 1988a; Colquhoun, 2014; Cumming, 2014; Gelman, 2013; Goodman, 2008; Greenland & Poole, 2013; Halsey et al., 2015; Wagenmakers, 2007), the p value itself, as traditionally computed from the tail areas under the sampling distribution for fixed sample sizes, contains as much information about a data set as Cohen's d . This equivalence is because, for given n_1 and n_2 , p values have a one-to-one relation with t values (except for the arbitrary sign that, as noted previously,

can only be interpreted with additional knowledge about the t value's calculation), which have a one-to-one relationship with Cohen's d .

Confidence interval of a standardized effect size

To explicitly represent measurement precision, many journals now emphasize publishing confidence intervals of the standardized effect size along with (or instead of) p values. There is some debate whether confidence intervals are actually a good representation of measurement precision (Morey et al., 2016), but whatever information they contain is already present in the associated p value and in the point estimate of d . This redundancy is because the confidence interval of Cohen's d is a function of the d value itself (and the sample sizes). The computation is a bit complicated, as it involves using numerical methods to identify the upper and lower limits of a range such that the given d value produces the desired proportion in the lower or upper tail of a non-central t distribution (Kelley, 2007). Although tedious (but easily done with a computer), the computation is one-to-one for both the lower and upper limit of the computed confidence interval. Thus, one gains no new information about the data set or measurement precision by reporting a confidence interval of Cohen's d if one has already reported t , d , or p . In fact, either limit of the confidence interval of Cohen's d is also equivalent to the other statistics. The same properties also hold for a confidence interval of Hedges' g .

Post hoc power

Post hoc power estimates the probability that a replication of an experiment, with the same design and sample sizes, would reject the null hypothesis. The estimate is based on the effect size observed in the experiment, so if the experiment rejected the null hypothesis, post hoc power will be bigger than one half. There are reasons to be skeptical about the quality of the post hoc power estimate (Yuan & Maxwell, 2005), and a common criticism is that it adds no knowledge about the data set if the p value is already known (Hoenig & Heisey, 2001). Although true, the same criticism applies to all of the statistics in Table 1. Once one value is known (e.g., the lower limit of the d confidence interval), then all the other values can be computed for the given sample sizes.

Post hoc power is computed relative to an estimated effect size. For small sample sizes, one computes different values depending on whether the estimated effect size is Cohen's d or Hedges' g . However, there is no loss of information in these terms, and knowing either post hoc power value allows for computation of all the other statistics in Table 1.

Table 1 For known sample sizes of an independent two-sample t test, each of these terms is a sufficient statistic for the standardized effect size of the population.

Statistic	Description
Cohen's d or Hedges' g	Estimated standardized effect size
t	Test statistic
p	p value for a two-tailed t test
$d_{95}(\text{lower})$ or $g_{95}(\text{lower})$	Lower limit of a 95 % confidence interval for d or g
$d_{95}(\text{upper})$ or $g_{95}(\text{upper})$	Upper limit of a 95 % confidence interval for d or g
Post hoc power from d or g	Estimated power for experiments with the same sample size
Post hoc v	Proportion of times OLS is more accurate than RLS
Λ	Log likelihood ratio for null and alternative models
ΔAIC , ΔAIC_c	Difference in AIC for null and alternative models
ΔBIC	Difference in BIC for null and alternative models
JZS BF	Bayes factor based on a specified Jeffreys–Zellner–Siow prior

For given sample sizes, it is possible to transform any value to all the others. *OLS* ordinary least squares; *RLS* randomized least squares; *AIC* Akaike Information Criterion; *BIC* Bayesian Information Criterion

Post hoc v

Davis-Stober and Dana (2013) proposed a statistic called v that compares model estimates based on ordinary least squares (OLS, which includes a standard t test) where model parameter estimates are optimized for the given data, against randomized least squares (RLS) where some model parameter estimates are randomly chosen. For a given experimental design (number of model parameters, effect size, and sample sizes), the v statistic gives the proportion of possible parameter estimates where OLS will outperform RLS on future data sets. It may seem counterintuitive that RLS should ever do better than OLS, but when there are many parameters and little data an OLS model may fit noise in the data so that random choices of parameters can, on average, produce superior estimates that will perform better for future data (also see Dawes (1979)).

The calculation of v is complex, but it is a function of the number of regression parameters (two, in the case of a two-sample t test), the sample size ($n = n_1 + n_2$), and the population effect size R^2 , which is the proportion of variance explained by the model. For a two-sample t test the estimate of R^2 can be derived from Cohen's d :

$$R^2 = \frac{d^2}{(n_1 + n_2)^2 / (n_1 n_2) + d^2} \quad (6)$$

Similar to power calculations, v is most commonly recommended as a means of optimizing experimental design. In particular, it does not make sense to run an experiment where $v < 0.5$ because it suggests that, regardless of the experimental outcome, a scientist would most likely produce better estimates by guessing. Nevertheless, v can also be used in a post hoc fashion (Lakens and Evers, 2014)

where the design is analyzed based on the observed standardized effect size. Such an analysis is based on exactly the same information in the data set as all the other statistics in Table 1.

Log likelihood ratio

Sometimes a statistical analysis focuses on identifying which of two (or more) models best matches the observed data. In the case of a two-sample t test, an observed score for subject i in condition $s \in \{1, 2\}$ is X_{is} , which could be due to a full (alternative) model that has an effect of condition:

$$X_{is} = \mu_s + \epsilon_i \quad (7)$$

where μ_s may differ for the two conditions and ϵ_i is random noise from a Gaussian distribution having a zero mean and unknown variance σ_F^2 . Alternatively, the data might be due to a reduced (null) model with no effect of condition:

$$X_{is} = \mu + \epsilon_i \quad (8)$$

where the mean, μ , is hypothesized to be the same for both conditions. The reduced model is a special case of the full model and has only two parameters (μ and σ_R^2) while the full model has three parameters (μ_1, μ_2, σ_F^2).

The likelihood of observed data is the product of the probability density values for the observations in the data set. For the full model in the conditions of an independent two-sample t test, the likelihood is the product of the

density function values (using the equations for the normal distributions for each value in the two conditions):

$$L_F = \prod_{i=1}^{n_1} \left[\frac{1}{\hat{\sigma}_F^2 \sqrt{2\pi}} \exp \left(-\frac{(X_{i1} - \hat{\mu}_1)^2}{2\hat{\sigma}_F^2} \right) \right] \times \prod_{i=1}^{n_2} \left[\frac{1}{\hat{\sigma}_F^2 \sqrt{2\pi}} \exp \left(-\frac{(X_{i2} - \hat{\mu}_2)^2}{2\hat{\sigma}_F^2} \right) \right] \quad (9)$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are estimates of the means and $\hat{\sigma}_F^2$ is an estimated pooled variance. For the reduced model, $\hat{\mu}_1$ and $\hat{\mu}_2$ are replaced by $\hat{\mu}$ and $\hat{\sigma}_F^2$ is replaced by $\hat{\sigma}_R^2$.

$$L_R = \prod_{i=1}^{n_1} \left[\frac{1}{\hat{\sigma}_R^2 \sqrt{2\pi}} \exp \left(-\frac{(X_{i1} - \hat{\mu})^2}{2\hat{\sigma}_R^2} \right) \right] \times \prod_{i=1}^{n_2} \left[\frac{1}{\hat{\sigma}_R^2 \sqrt{2\pi}} \exp \left(-\frac{(X_{i2} - \hat{\mu})^2}{2\hat{\sigma}_R^2} \right) \right] \quad (10)$$

The means are estimated using the typical sample mean ($\hat{\mu} = \bar{X}$, $\hat{\mu}_1 = \bar{X}_1$, and $\hat{\mu}_2 = \bar{X}_2$), while the variance terms are estimated to maximize the likelihood values (most readers are familiar with the maximum likelihood variance calculations as being the “population” formula for residuals, which uses two means for $\hat{\sigma}_F^2$ and one mean for $\hat{\sigma}_R^2$).

A model with larger likelihood better predicts the observed data. How much better is often measured with the log likelihood ratio:

$$\Lambda = \ln \left(\frac{L_F}{L_R} \right) = \ln(L_F) - \ln(L_R) \quad (11)$$

Because the reduced model is a special case of the full model, the full model with its best parameters will always produce a likelihood value at least as large as the value for the reduced model with its best parameters, so it will always be that $L_F \geq L_R$ and thus that $\Lambda \geq 0$. Similar to the criteria used for judging statistical significance from a p value, one can impose (somewhat arbitrarily) criterion thresholds on Λ for interpreting support for the full model compared to the reduced model.

It might appear that the log likelihood ratio (Λ) calculation is quite different from a p value, but Murtaugh (2014) showed that when comparing nested linear models (the t test being one such case) knowing the total sample size $n = n_1 + n_2$ and either p or Λ allows calculation of the other statistic (essentially the same relationship was noted by Glover and Dixon (2004)). For an independent two-sample t test, it is similarly possible to compute the t statistic from Λ :

$$t = \sqrt{(n-2) \left[\exp \left(\frac{2\Lambda}{n} \right) - 1 \right]}. \quad (12)$$

The corresponding conversion from t to Λ is:

$$\Lambda = \frac{n}{2} \ln \left(\frac{t^2}{n-2} + 1 \right). \quad (13)$$

So the log likelihood ratio contains exactly the same information about a data set as the t value and the other statistics in Table 1.

Model selection using an information criterion

The log likelihood ratio identifies which of the considered models best predicts the observed data. However, such an approach risks over-fitting the data by using model complexity to “predict” variation in the data that is actually due to random noise. Fitting random noise will cause the model to poorly predict the properties of future data sets (which will include different random noise). A systematic approach to dealing with the inherent complexity of the full model compared to the reduced model is to compensate for the number of independent parameters in each model. The Akaike Information Criterion (AIC) (Akaike, 1974) includes the number of independent parameters in a model, m , as part of the judgment of which model is best. For a given model, the AIC value is:

$$AIC = 2m - 2 \ln(L) \quad (14)$$

where m is the number of parameters in the model and L refers to the likelihood of the observed data being produced by the given model with its best parameter values. The multiple 2 is a scaling factor that relates AIC to properties of information theory.

For the AIC calculation, the likelihood is subtracted from the number of parameters, so *smaller* AIC values correspond to better model performance. For the situation corresponding to an independent two-sample t test, $m = 3$ for the full (alternative) model and $m = 2$ for the reduced (null) model, so the difference in AIC for the two models is:

$$\Delta AIC = AIC(\text{reduced}) - AIC(\text{full}) = -2 - 2 \ln(L_R) + 2 \ln(L_F) = -2 + 2\Lambda \quad (15)$$

When deciding whether the full or reduced model is expected to better correspond to the observed data set, one simply judges whether $\Delta AIC > 0$ (choose the full model) or $\Delta AIC < 0$ (choose the reduced, null, model). Unlike the $p < 0.05$ criterion, the zero criterion for ΔAIC is not arbitrary and indicates which model is expected to better predict future data (e.g., Dixon (2013) and Yang (2005)) in the sense of minimizing mean squared error. Notably, the decision of which model is expected to do a better job predicting future data does not indicate that the better model does a *good* job of predicting future data. It could be that neither model does well; and even if a model does predict well, a mixture of models might do even better (e.g., Burnham and Anderson (2004)).

In some practical applications with small sample sizes, model selection based on ΔAIC tends to prefer more complex models, and Hurvich and Tsai (1989) proposed a corrected formula:

$$AICc = AIC + \frac{2m(m+1)}{n-m-1} \quad (16)$$

that provides an additional penalty for more parameters with small sample sizes. As $n = n_1 + n_2$ increases, the additional penalty becomes smaller and $AICc$ converges to AIC . For model selection, one treats $AICc$ in the same way as AIC by computing the difference in $AICc$ for a reduced and full model. For the situation corresponding to an independent two-sample t test, the formula is:

$$\Delta AICc = \Delta AIC + \frac{12}{n-3} - \frac{24}{n-4}. \quad (17)$$

When $\Delta AICc < 0$, the data favor the reduced (null) model and when $\Delta AICc > 0$ the data favor the full (alternative) model. Burnham and Anderson (2002) provide an accessible introduction to the use of AIC methods.

The Bayesian Information Criterion (BIC) is another model selection approach where the penalty for the number of parameters includes the sample size (Schwarz, 1978). For a model with m parameters, the formula is:

$$BIC = m \ln(n) - 2 \ln(L). \quad (18)$$

Model selection is, again, based on the difference in BIC for two models. For the conditions corresponding to a two-sample t test, the difference between the BIC for the null and full models is:

$$\Delta BIC = -\ln(n) + 2\Lambda. \quad (19)$$

When $\Delta BIC < 0$, the data favor the reduced (null) model and when $\Delta BIC > 0$, the data favor the full (alternative) model. Sometimes the strength of support for a model is compared to (somewhat arbitrary) thresholds for classification of the evidence such as “barely worth a mention” ($0 < \Delta BIC \leq 2$) or “strong,” ($\Delta BIC \geq 6$) among others (Kass & Raftery, 1995).

Since Λ can be computed from t and the sample sizes for an independent two-sample t test, it is trivial to compute ΔAIC , $\Delta AICc$, and ΔBIC from the same information. Thus, for the case of an independent two-sample t test, all of these model selection statistics are based on exactly the same information in the data set that is used for a standard hypothesis test and an estimated standardized effect size.

JZS Bayes factor

A Bayes factor is similar to a likelihood ratio, but whereas the likelihood values for Λ are calculated relative to the model parameter values that maximize likelihood, a Bayes

factor computes mean likelihood across possible parameter values as weighted by a prior probability distribution for those parameter values.

For the conditions of a two-sample t test, Rouder et al. (2009) proposed a Bayesian analysis that uses a Jeffreys–Zellner–Siow (JZS) prior for the alternative hypothesis. They demonstrated that such a prior has nice scaling properties and that it leads to a Bayes factor value that can be computed relatively easily (unlike some other priors that require complex simulations, the JZS Bayes factor can be computed with numerical integration), and they have provided a Web site and an R package for the computations. Rouder et al. suggested that the JZS prior could be a starting point for many scientific investigations and they have extended it to include a variety of experimental designs (Rouder et al., 2012).

Rouder et al. (2009) showed that for a two-sample t test, the JZS Bayes factor for a given prior can be computed from the sample sizes and t value. The calculation is invertible so there is a one-to-one relationship between the p value and the JZS Bayes factor value. Thus, for known sample sizes, a scientist using a specific JZS prior for a two-sample t test has as much information about the data set if the p value is known as if the JZS Bayes factor (and its prior) is known. Calculating a Bayes factor requires knowing the properties of the prior, which means knowing the value of a scaling parameter, r , which defines one component of the JZS prior distribution. Since interpreting a Bayes factor requires knowing the prior distribution(s) that contributed to its calculation, this information should be readily available for computation from a t value to a BF and vice versa. The one-to-one relationship between the p value and the JZS Bayes factor value holds for every value of the JZS prior scaling parameter, so there are actually an infinite number of JZS Bayes factor values that are equivalent to the other statistics in Table 1 in that they all describe the same information about the data set. These different JZS Bayes factor values differ from each other in how they interpret that information; in the same way, a given JZS Bayes factor value provides a different interpretation of the same information in a data set than a p value or a ΔAIC value.

There are additional relationships between p values and Bayes factors in special situations. Marsman and Wagenmakers (2016) show that one-sided p values have a direct relationship to posterior probabilities for a Bayesian test of direction. Bayarri et al. (2016) argue that the p value provides a useful bound on the Bayes factor, regardless of the prior distribution.

Discussion

If the sample sizes for a two-sample t test are known, then the t value, p value, Cohen’s d , Hedges’ g , either limit of

a confidence interval of Cohen's d or Hedges' g , post hoc power derived from Cohen's d or Hedges' g , v , the log likelihood ratio, the ΔAIC , the $\Delta AICc$, the ΔBIC , and the JZS Bayes factor value all provide equivalent information about the data set. This equivalence means that debates about which statistic to report must focus on how these different statistics are interpreted. Likewise, discussions about post hoc power and v are simply different perspectives of the information that is already inherent in other statistics.

The equivalence of the p value, as typically calculated, with other summary statistics indicates that it is not appropriate to suggest that p values are meaningless without a careful discussion of the interpretation of p values. Some critiques of hypothesis testing do include such careful discussions, but casual readers might mistakenly believe that the problems of hypothesis testing are due to limitations in the information extracted from a data set to compute a p value rather than its interpretation. The mathematical equivalence between p values and other statistics might explain why the use of p values and hypothesis tests persist despite their many identified problems. Namely, regardless of interpretation problems (e.g., of relating a computed p value to a type I error rate even for situations that violate the requirements of hypothesis testing), the calculated p value contains all of the information about the data set that is used by a wide variety of inferential frameworks.

If a researcher simply wants to *describe* the information in the data set, then the choice of which statistic to report should reflect how well readers can interpret the format of that information (Cumming and Fidler, 2009; Gigerenzer et al., 2007). For example, readers may more easily understand the precision of measured effects when the information is formatted as the two end points of a 95 % confidence interval of Cohen's d rather than as a log likelihood ratio. Likewise, the information in a data set related to the relative evidence for null and alternative models may be better expressed as a JZS Bayes factor than as a Cohen's d value. Mathematical equivalence of these values does not imply equivalent communication of information; you must know both your audience and your message.

Scientists often want to do more than describe the information in a data set, they also want to *interpret* that information relative to a theory or application. In some cases, they want to convince other scientists that an effect is "real." In other cases, they want to argue that a measurement should be included in future studies or models because it provides valuable information that helps scientists understand other phenomena or predict performance. Despite their mathematical equivalence regarding the information in the data set, the different statistics in Table 1 are used to address dramatically different questions about interpreting that information. To help readers understand these different interpretations and their properties, the following two

sections compare the different statistics and then describe contexts where the various statistics are beneficial. The comparison grounds the statistics to cases that most psychological scientists already know: p values and standardized effect sizes. The consideration of different contexts should enable scientists to identify whether the interpretation of a reported statistic is appropriate for their scientific study.

Relationships between inferential frameworks

For the conditions corresponding to a two-sample t test, the mathematical equivalence of the statistics in Table 1 suggests that one can interpret the decision criteria for an inferential statistic in one framework in terms of values for a statistic in another framework. Given that traditional hypothesis testing is currently the approach taken by most research psychologists and that many calls for reform focus on reporting standardized effect sizes, it may be fruitful to explore how other inferential decision criteria map on to p values and d values.

Figure 1 plots the p values that correspond to criterion values that are commonly used in other inference frameworks as a function of sample size ($n_1 = n_2 = 5$ to 400). Figure 1a plots the p -values that correspond to JZS BF values (using a recommended default prior with a scale of $r = 1/\sqrt{2}$) often taken to provide minimal "positive" support for the alternative hypothesis ($BF = 3$, lower curve) or for the null hypothesis ($BF = 1/3$, upper curve) (e.g., Rouder et al. (2009) and Wagenmakers (2007)). The middle curve indicates the p values corresponding to equal support for either conclusion ($BF = 1$). It is clear that with regard to providing minimal support for the alternative hypothesis, the JZS BF criterion is more conservative than the traditional $p < 0.05$ criterion, as the p -values corresponding to $BF = 3$ are all less than 0.025, and that even smaller p values are required to maintain this level of support as the sample size increases.

In the Bayes factor inference framework, $BF \leq 1/3$ is often taken to indicate at least minimal support for the null hypothesis. Figure 1a shows that p values corresponding to this criterion start rather large (around 0.85) for sample sizes of $n_1 = n_2 = 17$ and decrease for larger sample sizes (for smaller sample sizes it is not possible to produce a JZS BF value less than $1/3$ with the default JZS scale). This relationship between p values and JZS BF values highlights an important point about the equivalence of the statistics in Table 1. Traditional hypothesis testing allows a researcher to reject the null, but does not allow a researcher to accept the null. This restriction is due to the inference framework of hypothesis testing rather than a limitation of the information about the data that is represented by a p value. As Fig. 1a demonstrates, it is possible to translate a decision

Fig. 1 p values that correspond to various criterion values for other inference frameworks. **a** p values that correspond to JZS $BF=3$ (bottom curve), JZS $BF = 1$ (middle curve), and JZS $BF=1/3$ (top curve). **b** p values that correspond to $\Delta AIC = 0$ (top curve) and $\Delta AICc = 0$

(bottom curve). **c** p values that correspond to $\Delta BIC = -2$ (top curve), $\Delta BIC = 0$ (middle curve), and $\Delta BIC = 2$ (bottom curve). **d** Log-log plots of p values that correspond to the different criterion values

criterion for accepting the null from the Bayes factor inference framework to p values. The justification for such a decision rests not on the traditional interpretation of the p value as related to type I error, but on the mathematical equivalence that p values share with JZS BF values for the comparison of two means.

Figure 1b plots the p values that correspond to $\Delta AIC = 0$ (top curve) and $\Delta AICc = 0$ (bottom curve), which divides support for the full (alternative) or reduced (null) model. Perhaps the most striking aspect of Fig. 1b is that these information criterion frameworks are more lenient than the typical $p < 0.05$ criterion for supporting the alternative hypothesis. For samples larger than $n_1 = n_2 = 100$, the decision criterion corresponds to $p \approx 0.16$, so any data set with $p < 0.16$ will be interpreted as providing better support for the alternative than the null. For still smaller sample sizes, a decision based on ΔAIC is even more lenient, while a decision based on $\Delta AICc$ is more stringent.

Figure 1c plots the p values that correspond to $\Delta BIC = -2$ (top curve) and $\Delta BIC = 2$ (bottom curve), which are commonly used criteria for providing minimal evidence for the reduced (null) or full (alternative) models (Kass and Raftery, 1995), respectively. Relative to the traditional $p < 0.05$ criterion to support the alternative hypothesis, inference from ΔBIC is more lenient at small sample sizes and more stringent at large sample sizes. The middle curve in Fig. 1c indicates the p values corresponding to $\Delta BIC = 0$, which gives equal support to both models.

To further compare the different inference criteria, Fig. 1d plots the p values that correspond to the criteria for the different inference frameworks on a log-log plot. It becomes clear that the ΔAIC and $\Delta AICc$ criteria are generally the most lenient, while the JZS BF criterion for minimal evidence for the alternative is more stringent. The p values corresponding to the ΔBIC criteria for the alternative are liberal at small sample sizes and become the most

stringent for large sample sizes. For the sample sizes used in Fig. 1, the typical $p = 0.05$ criterion falls somewhere in between the other criteria.

Figure 2 shows similar plots that identify the Cohen's d values that correspond to the different criteria, and the story is much the same as for the critical p values. To demonstrate minimal support for the alternative, the JZS BF requires large d values, while ΔAIC and $\Delta AICc$ require relatively small d values. Curiously, the $BF = 3$, $\Delta AIC = 0$, $\Delta BIC = 2$, and $p = 0.05$ criteria, which are used to indicate minimal support for the alternative, correspond to d values that are almost linearly related to sample size in log-log coordinates. These different inferential framework criteria differ in the slope and intercept of the d value line, but are otherwise using the same basic approach for establishing minimal support for the alternative hypothesis as a function of sample size. Additional simulations indicate that the

linear relationship holds up to at least $n_1 = n_2 = 40,000$. R code to generate the critical values for Figs. 1 and 2 is available at the Open Science Framework.

It should be stressed that although a criterion value is an integral part of traditional hypothesis testing (it need not be $p < 0.05$, but there needs to be some criterion to control the type I error rate), some of the “minimal” criterion values used in Figs. 1 and 2 for the other inference frameworks are rules of thumb and are not a fundamental part of the inference process. While some frameworks do have principled criteria (0 for ΔAIC , $\Delta AICc$, and ΔBIC and 1 for BF) that indicate equivalent support for a null or alternative model, being on either side of these criteria need not indicate definitive conclusions. For example, it would be perfectly reasonable to report that a JZS $BF = 1.6$ provides some support for the alternative hypothesis. Many researchers find such a low level of support insufficient

Fig. 2 Cohen's d values that correspond to various criterion values for various inference frameworks. **a** d values that correspond to JZS $BF = 3$ (top curve), JZS $BF = 1$ (middle curve), and JZS $BF = 1/3$ (bottom curve). **b** d values that correspond to $\Delta AIC = 0$ (bottom curve) and

$\Delta AICc = 0$ (top curve). **c** d values that correspond to $\Delta BIC = -2$ (bottom curve), $\Delta BIC = 0$ (middle curve), and $\Delta BIC = 2$ (top curve). **d** Log-log plots of d -values that correspond to the different criterion values

for the kinds of conclusions scientists commonly want to make (indeed, some scientists find the minimal criterion of $BF = 3$ to be insufficient), but the interpretation very much depends on the context and impact of the conclusion. In a similar way, a calculation of $\Delta AIC = 0.1$ indicates that the alternative (full) model should do (a bit) better at predicting new data than the null (reduced) model, but this does not mean that a scientist is obligated to reject the null model. Such a small difference could be interpreted as indicating that both models are viable and that more data (and perhaps other models) are needed to better understand the phenomenon of interest.

What statistic should be reported?

From a mathematical perspective, in the context of a two-sample t test, reporting the sample sizes and any statistic from Table 1 provides equivalent information for all of the other statistics and enables a variety of decisions in the different inference frameworks. However, that equivalence does not mean that scientists should feel free to report whatever statistic they like. Rather, it means that scientists need to think carefully about whether a given statistic accommodates the purpose of their scientific investigations. The following sections consider the different statistics and the benefits and limitations associated with their interpretations.

p values

Within the framework of classic hypothesis testing, the p value is interpreted to be the probability of observing an outcome at least as extreme as what was observed if the null hypothesis were true. Under this interpretation, repeated hypothesis tests can control the probability of making a type I error (rejecting a true null hypothesis) by requiring p to be less than 0.05 before rejecting the null hypothesis. As many critics have noted, even under the best of circumstances, the p value is often misunderstood and a decision to reject the null hypothesis does not necessarily justify accepting the alternative hypothesis (e.g., Rouder et al. (2009) and Wagenmakers (2007)). Moreover, the absence of a small p value does not necessarily support the null hypothesis (Hauer, 2004).

In addition to those concerns, there are practical issues that make the p value and its associated inference difficult to work with. For situations common to scientific investigations, the intended control of the type I error rate is easily lost because the relation of the p value to the type I error rate includes a number of restrictions on the design of the experiment and sampling method. Perhaps the most stringent restriction is that the sample size must be fixed prior to data collection. When this restriction is not satisfied, the

decision-making process will reject true null hypotheses at a rate different than what was nominally intended (e.g., 0.05). In particular, if data collection stops early when $p < 0.05$ (or equivalently additional data are added to try to get a $p = 0.07$ to become $p < 0.05$), then the type I error rate can be much larger than what was intended (Berger and Berry, 1988b; Strube, 2006). This loss of type I error control is especially troubling because it makes it difficult for scientists to accumulate information across replication experiments. Ideally, data from a new study would be combined with data from a previous study to produce a better inference, but such combining is not allowed with standard hypothesis testing because the sample size was not fixed in advance (Scheibehenne et al., 2016; Ueno et al., 2016). You can calculate the p value for the combined data set and it is a summary statistic of the information in the combined data set that is equivalent to a standardized effect size; but the p value magnitude from the combined data set cannot be directly related to a type I error rate.

When experiments are set up appropriately, the p value can be interpreted in relation to the intended type I error rate, but it is difficult to satisfy these requirements in common scientific settings such as exploring existing data to find interesting patterns and gathering new data to sort out empirical uncertainties from previous studies.

Standardized effect sizes and their confidence intervals

A confidence interval of Cohen's d or Hedges' g is often interpreted as providing some information about the uncertainty in the measurement of the population effect size (but see Morey et al. (2016) for a contrary view), and an explicit representation of this uncertainty can be useful in some situations (even though Table 1 indicates that the uncertainty is already implicitly present when reporting the effect size point estimate and the sample sizes).

There are surely situations where the goal of research is to estimate an effect (standard or otherwise), but descriptive statistics by themselves do not promote model building or hypothesis testing. As Morey et al. (2014) argue, scientists often want to test theories or draw conclusions, and descriptive statistics do not by themselves support those kinds of investigations, which require some kind of inference from the descriptive statistics.

AIC and BIC

Model selection based on AIC (for this discussion including AIC_c) or BIC seeks to identify which model will better predict future data. That is, the goal is not necessarily to select the “correct” model, but to select the model that maximizes predictive accuracy. The AIC and BIC approaches can be applied to some more complex situations

than traditional hypothesis testing, and thereby can promote more complicated model development than is currently common. For example, a scientist could compare *AIC* or *BIC* values for the full and reduced models described above against a model that varies the estimates for both the means and variances (e.g., $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$). In contrast to traditional hypothesis testing, the models being compared with *AIC* or *BIC* need not be nested. For example, one could compare a model with different means but a common variance against a model with a common mean but different variances.

One difficulty with *AIC* and *BIC* analyses is that it is not always easy to identify or compute the likelihood for complicated models. A second difficulty is that the complexity of some models is not fully described by counting the number of parameters. These difficulties can be addressed with a minimum description length analysis (Pitt et al., 2002) or by using cross-validation methods where model parameters are estimated from a subset of the data and the resulting model is tested by seeing how well it predicts the unused data. *AIC* analyses approximate average prediction error as measured by leave-one-out cross validation, where parameter estimation is based on all but one data point, whose value is then predicted. Cross validation methods are very general but computationally intensive. They also require substantial care so that the researcher does not inadvertently sneak in knowledge about the data set when designing models. For details of cross-validation and other related methods see Myung et al. (2013). Likewise, there is a close relationship between ΔBIC and a class of Bayes Factors (with certain priors), and ΔBIC is often treated as an approximation to a Bayes Factor (Nathoo and Masson, 2016; Kass & Raftery, 1995; Masson, 2011; Wagenmakers, 2007). The more general Bayes factor approach allows consideration of complex models where counting parameters is not feasible.

Since both *AIC* and *BIC* attempt to identify the model that better predicts future data, which one to use largely depends on other aspects of the model selection. Aho et al. (2014) argue that the *AIC* approach to model selection is appropriate when scientists are engaged in an investigation that involves ever more refinement with additional data and when there is no expectation of discovering the “true” model. In such an investigative approach, the selected model is anticipated to be temporary, in the sense that new data or new models will (hopefully) appear that replace the current best model. Correspondingly, the appropriate conclusion when $\Delta AIC > 0$ is that the full (alternative) model tentatively does better than the reduced (null) model, but with recognition that new data may change the story and that a superior model likely exists that has not yet been considered. Aho et al. (2014) further argue that *BIC* is appropriate when scientists hope to identify which candidate model corresponds to the “true” model. Indeed, *BIC*

has the nice property that when the true model is among the candidate models a *BIC* analysis will almost surely identify it as the best model for large enough samples. This consistency property is not a general characteristic of *AIC* approaches, which sometimes pick a too complex model even for large samples (e.g., Burnham and Anderson 2004; Wagenmakers and Farrell 2004). In one sense, a scientist’s decision between using *AIC* or *BIC* should be based on whether he/she is more concerned about under fitting the data with a too simple model (choose *AIC*) or overfitting the data with a too complex model (choose *BIC*).

Although *AIC* and *BIC* analyses are often described as methods for model selection, they can be used in other ways. Even when the considered models are not necessarily true, *AIC* and *BIC* can indicate the relative evidential value between those models (albeit in different ways). Contrary to how $p < 0.05$ is sometimes interpreted, the decision of an *AIC* or a *BIC* analysis need not be treated as “definitive” but as a starting point for gathering more data (either as a replication or otherwise) and for developing more complicated models. In many cases, if the goal is strictly to predict future data, researchers are best served by weighting the predictions of various models rather than simply selecting the prediction generated by the “best” model (Burnham & Anderson, 2002).

Bayes factors

AIC and *BIC* use observed data to identify model parameters and model structures (e.g., whether to use one common mean or two distinct means) that are expected to best predict future data. The prediction is for future data because the model parameters are estimated from the observed data and therefore cannot be used to predict the observed data on which they are derived. Rather than predict observed data, a Bayes factor identifies which model best predicts the observed data (Wagenmakers et al., 2016). The prior information used to compute the Bayes factor defines models before (or at least independent of) observing the data and characterizes the uncertainty about the model parameter values. Unlike *AIC* and *BIC*, there is no need for a Bayes Factor to explicitly adjust for the number of model parameters because a model with great flexibility (e.g., many parameters) automatically makes a less precise prediction than a simpler model.

The trick for using Bayes factors is identifying good models and appropriate priors. The primary benefit for a Bayesian analysis is when informative priors strongly influence the interpretation of the data (e.g., Li et al. (2011) and Vanpaemel (2010). Poor models and poor priors can lead to poor analyses; and these difficulties are not alleviated by simply broadening the prior to express ever more uncertainty (Gelman, 2013; Rouder et al., 2016).

Like *AIC* and *BIC*, Bayes factor approaches can be applied to situations where *p* values hardly make sense. They also support a gradual accumulation of data in a way that *p* values do not; for example they are insensitive to optional stopping (Rouder, 2014). Thus, when scientists can identify good models and appropriate priors for those models, then Bayes factor analyses are an excellent choice.

Discussion

There is no “correct” statistical analysis that applies to all purposes of statistical inference. Although the statistics in Table 1 provide equivalent information about a data set, they motivate different interpretations of the data set that are appropriate in different contexts. Classic hypothesis testing focuses on control of type I error, but such control requires specific experimental attributes that are difficult to satisfy in common situations. Standardized effect sizes and their confidence intervals characterize important information in a data set, but they do not directly promote model building, prediction, or hypothesis testing. *AIC* approaches are most useful for exploratory investigations where tentative conclusions are drawn from the currently available data and models. *BIC* and Bayes factor approaches tend to be most useful for testing well-characterized models.

Are we reporting appropriate information about a data set?

The equivalence of the statistics in Table 1 motivates reconsideration of whether the information in the data set represented by these statistics is relevant for good scientific practice. The answer is surely “yes” in some situations, but this section argues that there are situations where other statistics would do better.

Let us first consider the situations where the statistics in Table 1 should be part of what is reported. The standardized effect size represents the difference of means relative to the standard deviation (for known sample sizes, the other statistics provide equivalent information, but this representation is most explicit in the estimated standardized effect size). This signal-to-noise ratio is important when one wants to predict future data because predictability depends on both the estimated difference of means and the variability that will be produced by sampling. Thus, if the goal is to understand how well you can *predict* future data values, then some of the statistics in Table 1 should be reported, because they all describe (in various ways) the signal-to-noise ratio in the data set.

It is worthwhile to consider a specific example of this point. If a new method of teaching is estimated to improve

reading scores by $\bar{X}_2 - \bar{X}_1 = 5$ points (the signal), the estimated standardized effect size will be large if the standard deviation (noise) is small (e.g., $s = 5$ gives $d = 1$) but the estimated standardized effect size will be small if the standard deviation is large (e.g., $s = 50$ gives $d = 0.1$). If the estimates are from an experiment with $n_1 = n_2 = 150$, then for the $d = 1$ case, $\Delta AIC = 65.4$, which means that the full model (with separate means for the old and new teaching methods) is expected to better predict future data than the reduced model (with one mean used to predict both teaching methods). On the other hand, if $d = 0.1$, then $\Delta AIC = -1.2$, which indicates that future data is better predicted by the reduced model.

Likewise, the standardized effect size is very important for understanding replications of empirical studies. Large values of δ for the population indicate that support for an effect can be found with relatively small sample sizes. It also means that other scientists are likely to replicate (in the sense of meeting the same statistical criterion) a reported result because the experimental power will be large for commonly used sample sizes. It is not necessarily true, but a large standardized effect size often also means that an effect is robust to small variations in experimental context because the signal is large relative to the sources of noise in the environment.

Prediction and replication are fundamental scientific activities, so it is understandable that scientists are interested in the standardized effect size. But what is important for scientific investigations may not necessarily be important for other situations. If we return to the different effect sizes described above, it is certainly easier for a scientist to demonstrate an improvement in reading scores when $\delta = 1$ than when $\delta = 0.1$, but an easily found effect is not necessarily important and a difficult to reproduce effect might be important in some contexts.

In many practical situations, the effect that matters is not the standardized effect size, but an unstandardized effect size (Baguley, 2009). If a teacher wants to increase test scores by at least 7 points, it does not help to only know that the estimated standardized effect size is $d = 1$. Rather, the teacher needs a direct estimate of the increase in test scores, $\bar{X}_2 - \bar{X}_1$, for a new method 2 compared to an old method 1. The teacher may also want some estimate of the variability of the test scores, s , which would give some sense of how likely the desired change was to occur for a sample of students. Since the statistics in Table 1 are all derived from (or equivalent to) the estimated standardized effect size, they do not provide sufficient information for the teacher to make an informed judgment in this case. This ambiguity suggest that, at the very least, these statistics need to be supplemented with additional information from the data set.

For example, if the teacher considers the sample means and standard deviations to be representative of the population means and standard deviations; then for a class of 30 students, the teacher can predict the probability of an average increase in test scores of at least 7 points. Namely, when $\bar{X}_2 - \bar{X}_1 = 5$, as above, one would predict the standard deviation of the sampling distribution ($s/\sqrt{30}$) and find the area under a normal distribution for scores larger than 7. If $s = 5$, this probability is 0.014; while if $s = 50$, this probability is 0.41. Note that these probabilities require knowing both the mean and standard deviation and cannot be calculated from just the standardized effect size. Moreover, this is a situation where a smaller standardized effect is actually in the teacher's favor because the larger standard deviation means that there is a better chance (albeit still less than 50 %) of getting a sufficiently large difference in the sample means. Oftentimes a better approach is to use Bayesian methods that characterize the uncertainty in the sample estimates for the means and standard deviations by using probability density functions (e.g., Kruschke (2010) and Lindley (1985)).

Depending on the teacher's interests, one might also want to consider the costs associated with the possibility of reducing test scores and then weigh the potential gains against the potential costs. More generally, a consideration of the utility of various outcomes might be useful to determine whether it is worth the hassle of changing teaching methods given the estimated potential gains. In a decision of this type, where costs (e.g., ordering new textbooks, learning a new teaching method) are being weighed against gains (e.g., anticipated increases in mean reading scores), the concept of statistical significance does not play a role. Instead, one simply contrasts expected utilities for different choices based on the available information in the data set and, for Bayesian approaches, from other sources that influence the priors. One cannot substitute statistical significance or arbitrary thresholds for a careful consideration of the expected utilities of different choices. Gelman (1998) provides examples of this kind of decision making process that are suitable for classroom demonstrations.

These properties are true even when the conclusions stay within the research lab. If a scientist computes $\Delta AICc = 3$ ($p = 0.025$) for a study with $n_1 = n_2 = 150$ should he continue to pursue this line of work with additional experiments (e.g., to look for sex effects, attention effects, or income inequality effects)? The data so far suggest that a model treating the two groups as having different means will lead to better predictions of future data than a model treating the two groups as having a common mean, so that sounds fairly promising. But the scientist has to consider many other factors, including the difficulty of running the studies, the importance of the topic, and whether the scientist has anything better to work on. Depending on the

utilities associated with these factors, further investigations might be warranted, or not. A Bayesian analysis would also allow the researcher to consider other knowledge sources (framed as priors) that might improve the estimated probabilities of different outcomes and thereby lead to better decisions.

Ultimately, the statistics in Table 1 summarize just one aspect of the information in a data set: the signal to noise ratio. It is valuable information for some purposes, but it is not sufficient for other purposes. Both creators (scientists) and consumers (policy makers) of statistics need to think carefully about the pros and cons associated with different decisions and weight them accordingly with statistical information. In many cases, this analysis will involve information that is not available from the statistics in Table 1. Such uses should further motivate scientists to publicly archive their raw data so that the data can be used for analyses not restricted to the statistics in Table 1.

Conclusions

The “crisis of confidence” (Pashler and Wagenmakers, 2012) being experienced by psychological science has produced a number of suggestions for improving data analysis, reporting results, and developing theories. An important part of many of these reforms is adoption of new (to mainstream psychology) methods of statistical inference. There is promise to these new approaches, but it is important to recognize connections between old and new methods; and, as Table 1 indicates, there are close mathematical relationships between various statistics. It is equally important to recognize that these close relationships do not mean that the different analysis methods are “all the same.” As described above, different analysis methods address different approaches to scientific investigation, and there are situations where they strikingly disagree with each other. Such disagreements are to be expected because scientific investigations variously involve discovery, theorizing, verification, prediction, and testing. Even for the same data set, different scientists can properly interpret the results in different ways for different purposes. For example, Fig. 1d shows that for $n_1 = n_2 > 200$, data that generate $\Delta BIC = -2$ can be interpreted as evidence for the null model even though the data reject the null hypothesis ($p = 0.046$) and the full model is expected to better predict future data than the null model ($\Delta AIC = 1.99$). As Gigerenzer (2004) noted, “no method is best for every problem.” Instead, researchers must carefully consider what kind of statistical analysis lends itself to their scientific goals. The statistical tools are available to improve scientific practice in psychology, and informed scientists will greatly benefit by putting them to good use.

References

- Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, 95, 631–636.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Baguley, T. (2009). Standardized or simple effect size: What should be reported. *The British Psychological Society*, 100, 603–617.
- Bayarri, M.J., Benjamin, D.J., Berger, J.O., & Sellke, T.M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72, 90–103.
- Berger, J., & Berry, D. (1988a). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159–165.
- Berger, J., & Berry, D. (1988b). The relevance of stopping rules in statistical inference (with discussion). In *Statistical Decision Theory and Related Topics 4* (S. S. Gupta and J. Berger, eds.), 1, 29–72. New York: Springer.
- Burnham, K.P., & Anderson, D.R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*, 2nd edition. New York: Springer.
- Burnham, K.P., & Anderson, D.R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33, 261–304.
- Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *R. Soc. Open Sci.*, 1, 140216. <http://dx.doi.org/10.1098/rsos.140216>.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29.
- Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie*, 217, 15–26.
- Davis-Stober, C.P., & Dana, J. (2013). Comparing the accuracy of experimental estimates to guessing: A new perspective on replication and the “crisis of confidence” in psychology. *Behavior Research Methods*, 46, 1–14.
- Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571–582.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, 781. doi:10.3389/fpsyg.2014.00781
- Dixon, P. (2013). The effective number of parameters in post hoc models. *Behavior Research Methods*, 45, 604–612.
- Eich, E. (2014). Business not as usual. *Psychological Science*, 25, 3–6.
- Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, 19, 975–991.
- Gelman, A. (1998). Some class-participation demonstrations for decision theory and Bayesian statistics. *The American Statistician*, 52, 167–174.
- Gelman, A. (2013). P values and statistical practice. *Epidemiology*, 24, 69–72.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606.
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L.M., & Woloshin, S. (2007). Helping doctors and patients to make sense of health statistics. *Psychological Science in the Public Interest*, 8, 53–96.
- Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, 11, 791–806.
- Goodman, S. (2008). A dirty dozen: Twelve P-value misconceptions. *Seminars in Hematology*, 45(3), 135–140. doi:10.1053/j.seminhematol.2008.04.003.
- Greenland, S., & Poole, C. (2013). Living with P values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, 24, 62–68.
- Halsey, L.G., Curran-Everett, D., Vowler, S.L., & Drummond, G.B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, 12, 179–185.
- Hauer, E. (2004). The harm done by tests of significance. *Accident Analysis & Prevention*, 36, 495–500.
- Hedges, L.V. (1981). Distribution theory for Glass’s estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128.
- Hoening, J.M., & Heisey, D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55, 1–6.
- Hurvich, C.M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297–307.
- John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23, 524–532.
- Kass, R.E., & Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, 20. <http://www.jstatsoft.org/v20/a08/>.
- Kruschke, J.K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press/Elsevier Science.
- Lakens, D., & Evers, E.R.K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, 9, 278–292.
- LeBel, E.P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K.R., Ratliff, K.A., & Tucker Smith, C. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science*, 8, 424–432.
- Lee, M.D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.
- Lindley, D.V. (1985). *Making Decisions*, 2nd edition. London: Wiley.
- Li, Y., Sawada, T., Shi, Y., Kwon, T., & Pizlo, Z. (2011). A Bayesian model of binocular perception of 3D mirror symmetric polyhedra. *Journal of Vision*, 11(4:11), 1–20.
- Marsman, M., & Wagenmakers, E.-J. (2016). Three insights from a Bayesian interpretation of the one-sided P value. *Educational and Psychological Measurement*. in press.
- Masson, M.E.J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, 43, 679–690.
- Mood, A.M., Graybill, F.A., & Boes, D.C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.
- Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23, 103–123. doi:10.3758/s13423-015-0947-8.
- Morey, R.D., Rouder, J.N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential to psychological science: A Comment on Cumming. *Psychological Science*, 24, 1291–1292.
- Murtaugh, P.A. (2014). In defense of P values. *Ecology*, 95, 611–617.
- Myung, I.I., Cavagnaro, D.R., & Pitt, M.A. (2013). Model selection and evaluation. In Batchelder, W.H., Colonius, H., Dzhaferov, E., & Myung, I.I. (Eds.) *New Handbook of Mathematical Psychology, Vol. 1: Measurement and Methodology*. London: Cambridge University Press.
- Nathoo, F.S., & Masson, M.E.J. (2016). Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *Journal of Mathematical Psychology*, 72, 144–157.

- O'Boyle Jr., E.H., Banks, G.C., & Gonzalez-Mulé, E. (2014). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*. doi:10.1177/0149206314527133.
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, 7, 528–530.
- Pitt, M.A., Myung, I.J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109, 472–491.
- R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org/>.
- Rouder, J.N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21, 301–308.
- Rouder, J.N., Morey, R.D., Speckman, P.L., & Province, J.M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56, 356–374.
- Rouder, J.N., Morey, R.D., Verhagen, J., Province, J.M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference?. *Topics in Cognitive Science*, 8, 520–547.
- Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225–237.
- Scheibehenne, B., Jamil, T., & Wagenmakers E.-J. (2016). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*, 27, 1043–1046.
- Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.
- Strube, M.J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods*, 38, 24–27.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37, 1–2.
- Ueno, T., Fastrich, G.M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General*, 5, 643–654.
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apology for the Bayes factor. *Journal of Mathematical Psychology*, 54, 491–498.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14, 779–804.
- Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11, 192–196.
- Wagenmakers, E.-J., Morey, R.D., & Lee, M.D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25, 169–176.
- Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92, 937–950.
- Yuan, K.H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, 30, 141–167.