Bayesian explorations in mathematical psychology

Matzke, D.

Chapter 11

Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests

This chapter has been published as:
Ruud Wetzels, Dora Matzke, Michael D. Lee, Jeffrey N. Rouder, Geoffrey J. Iverson, and Eric-Jan Wagenmakers (2011).
Statistical evidence in experimental psychology: An empirical comparison using 855 t tests.
Perspectives on Psychological Science, 6, 291-298.[1]

Abstract

Statistical inference in psychology has traditionally relied heavily on p value significance testing. This approach to drawing conclusions from data, however, has been widely criticized, and two types of remedies have been advocated. The first proposal is to supplement p values with complementary measures of evidence such as effect sizes. The second is to replace inference with Bayesian measures of evidence such as the Bayes factor. The authors provide a practical comparison of p values, effect sizes, and default Bayes factors as measures of statistical evidence, using 855 recently published t tests in psychology. The comparison yields two main results. First, although p values and default Bayes factors almost always agree about what hypothesis is better supported by the data, the measures often disagree about the strength of this support; for 70% of the data sets for which the p value falls between 0.01 and 0.05, the default Bayes factor indicates that the evidence is only anecdotal. Second, effect sizes can provide additional evidence to p values and default Bayes factors. The authors conclude that the Bayesian approach is comparatively prudent, preventing researchers from overestimating the evidence in favor of an effect.

11.1 Introduction

Experimental psychologists use statistical procedures to convince themselves and their peers that the effect of interest is real, reliable, replicable, and hence worthy of academic attention. A representative example comes from Mussweiler (2006), who studied whether particular actions can

[1]The final publication is available at `http://pps.sagepub.com/content/6/3/291.short`.

activate a corresponding stereotype. To test this hypothesis empirically, Mussweiler unobtrusively induced half the participants, the experimental group, to move in a portly manner that is stereotypic for the overweight. The other half, the control group, made no such movements. Next, all participants were given an ambiguous description of a target person and then used a 9-point scale (1 = *not at all*, 9 = *very*) to rate this person on dimensions that correspond to the overweight stereotype (e.g., "unhealthy", "sluggish", "insecure"). To assess whether performing the stereotypic motion affected the rating of the ambiguous target person, Mussweiler computed a t statistic ($t(18) = 2.1$), and found that this value corresponded to a low p value ($p < 0.05$).[2] Following conventional protocol, Mussweiler concluded that the low p value should be taken to provide "initial support for the hypothesis that engaging in stereotypic movements activates the corresponding stereotype" (Mussweiler, 2006, p. 18).

The use of t tests and corresponding p values in this way constitutes a common and widely accepted practice in the psychological literature. It is, however, not the only possible or reasonable approach to measuring evidence and making statistical and scientific inferences. Indeed, the use of t tests and p values has been widely criticized (e.g., J. Cohen, 1994; Cumming, 2008; Dixon, 2003; Howard, Maxwell, & Fleming, 2000; M. D. Lee & Wagenmakers, 2005; G. R. Loftus, 1996; Nickerson, 2000; Wagenmakers, 2007). There are at least two different criticisms, coming from different perspectives and resulting in different remedies. First, many have argued that null hypothesis tests should be supplemented with other statistical measures, such as confidence intervals and effect sizes. Within psychology, this approach to remediation has sometimes been institutionalized, being required by journal editors or recommended by the American Psychological Association (e.g., American Psychological Association, 2010; J. Cohen, 1988; Erdfelder, 2010; Wilkinson & the Task Force on Statistical Inference, 1999).

A second, more fundamental criticism that comes from Bayesian statistics is that there are basic conceptual and practical problems with p values. Although Bayesian criticism of psychological statistical practice dates back at least to Edwards et al. (1963), it has become especially prominent and increasingly influential in the last decade (e.g., Dienes, 2008; Gallistel, 2009; Kruschke, 2010c, 2010a; M. D. Lee, 2008; Myung et al., 2000; Rouder et al., 2009). One standard Bayesian measure for quantifying the amount of evidence from the data in support of an experimental effect is the Bayes factor (Gönen, Johnson, Lu, & Westfall, 2005; Rouder et al., 2009; Wetzels et al., 2009). The measure takes the form of an odds ratio: It is the probability of the data under one hypothesis relative to that under another (Dienes, 2011; Kass & Raftery, 1995; M. D. Lee & Wagenmakers, 2005).

With this background, it seems that psychological statistical practice currently stands at a three-way fork in the road. Staying on the current path means continuing to rely on p values. A modest change is to place greater focus on the additional inferential information provided by effect sizes and confidence intervals. A radical change is struck by moving to Bayesian approaches, such as Bayes factors. The path that psychological science chooses seems likely to matter. It is not just that there are philosophical differences between the three choices. It is also clear that the three measures of evidence can be mutually inconsistent (e.g., Berger & Sellke, 1987; Rouder et al., 2009; Wagenmakers, 2007; Wagenmakers & Grünwald, 2006; Wagenmakers et al., 2010).

In this article, we assess the practical consequences of choosing among inference by p values, by effect sizes, and by Bayes factors. By practical consequences, we mean the extent to which conclusions of extant studies change according to the inference measure that is used. To assess these practical consequences, we reanalyzed 855 t tests reported in articles from the 2007 issues

[2] The findings suggest that Mussweiler (2006) conducted a one-sided t test. In the remainder of this article, we conduct two-sided t tests.

of *Psychonomic Bulletin & Review* (PBR) and *Journal of Experimental Psychology: Learning, Memory and Cognition* (JEP:LMC). For each t test, we compute the p value, the effect size, and the Bayes factor and study the extent to which they provide information that is redundant, complementary, or inconsistent. On the basis of these analyses, we suggest the best direction for measuring statistical evidence from psychological experiments.

11.2 Three Measures of Evidence

In this section, we describe how to calculate and interpret the p value, the effect size, and the Bayes factor. For concreteness, we use Mussweiler's (2006) study on the effect of action on stereotypes. The mean score of the control group, M_c, was 5.8 on a weight-stereotype scale ($s_c = 0.69, n_c = 10$), and the mean score of the experimental group, M_e, was 6.4 ($s_e = 0.66, n_e = 10$).

The p Value

The interpretation of p values is not straightforward, and their use in hypothesis testing is heavily debated (J. Cohen, 1994; Cortina & Dunlap, 1997; Cumming, 2008; Dixon, 2003; Frick, 1996; Gigerenzer, 1993, 1998; Hagen, 1997; Killeen, 2005, 2006; Kruschke, 2010c, 2010a; M. D. Lee & Wagenmakers, 2005; G. R. Loftus, 1996; Nickerson, 2000; Schmidt, 1996; Wagenmakers & Grünwald, 2006; Wainer, 1999). The p value is the probability of obtaining a test statistic (in this case, the t statistic) at least as extreme as the one that was observed in the experiment, given that the null hypothesis is true and the sample is generated according to a specific intended procedure, such as fixed sample size. Fisher (1935) interpreted these p values as evidence against the null hypothesis. The smaller the p value, the more evidence there was against the null hypothesis. Fisher viewed these values as self-explanatory measures of evidence that did not need further guidance. In practice, however, most researchers (and reviewers) adopt a 0.05 cutoff: p values less than 0.05 constitute evidence for an effect, and those greater than 0.05 do not. More fine-grained categories are possible, and Wasserman (2004, p. 157) proposes the gradations shown in Table 11.1. Note that Table 11.1 lists various categories of evidence against the null hypothesis. A basic limitation of null hypothesis significance testing is that it does not allow a researcher to gather evidence in favor of the null (Dennis, Lee, & Kinnell, 2008; Gallistel, 2009; Rouder et al., 2009; Wetzels et al., 2009).

Table 11.1 Evidence Categories for p Values (adapted from Wasserman, p. 157, 2004).

p value			Interpretation
	$<$	0.001	Decisive evidence against H_0
0.001	–	0.01	Substantive evidence against H_0
0.01	–	0.05	Positive evidence against H_0
	$>$	0.05	No evidence against H_0

For the data from Mussweiler (2006), we compute a p value based on the t test. The t test is designed to test whether a difference between two means is significant. First, we calculate the t statistic:

$$t = \frac{M_e - M_c}{\sqrt{s^2_{pooled}\left(\frac{1}{n_e} + \frac{1}{n_c}\right)}} = \frac{6.42 - 5.79}{\sqrt{0.46\left(\frac{1}{10} + \frac{1}{10}\right)}} = 2.09,$$

where M_c and M_e are the means of both groups, n_c and n_e are the sample sizes, and s^2_{pooled} estimates the common population variance:

$$s^2_{pooled} = \frac{(n_e - 1)s_e^2 + (n_c - 1)s_c^2}{n_e + n_c - 2}.$$

Next, the t statistic with $n_e + n_c - 2 = 18$ degrees of freedom results in a p value slightly larger than 0.05 (≈ 0.051). For our concrete example, Table 11.1 leads to the conclusion that the p value is on the cusp between "no evidence against H_0" and "positive evidence against H_0".

The Effect Size

Effect sizes quantify the magnitude of an effect and serves as a measure of how much the results deviate from the null hypothesis (J. Cohen, 1988; Thompson, 2002; Richard, Bond, & Stokes-Zoota, 2003; Rosenthal, 1990; Rosenthal & Rubin, 1982). For the data from Mussweiler (2006), the effect size d is calculated as follows:

$$d = \frac{M_e - M_c}{s_{pooled}} = \frac{6.42 - 5.79}{0.68} = 0.93.$$

Note that in contrast to the p value, the effect size is independent of sample size; increasing the sample size does not increase effect size but instead allows it to be estimated more accurately.

Effect sizes are often interpreted in terms of the categories introduced by J. Cohen (1988), as listed in Table 11.2, ranging from "small" to "very large". For our concrete example, $d = 0.93$, and we conclude that this effect is large to very large. Interestingly, the p value was on the cusp between the categories "no evidence against H_0" and "positive evidence against H_0" whereas the effect size indicates the effect to be strong.

Table 11.2 Evidence Categories for Effect Sizes as Proposed by J. Cohen (1988).

Effect Size			Interpretation
	$<$	0.2	Small effect size
0.2	–	0.5	Small to medium effect size
0.5	–	0.8	Medium to large effect size
	$>$	0.8	Large to very large effect size

The Bayes Factor

In Bayesian statistics, uncertainty (or degree of belief) is quantified by probability distributions over parameters. This makes the Bayesian approach fundamentally different from the classical "frequentist" approach, which relies on sampling distributions of data (Berger & Delampady, 1987; Berger & Wolpert, 1988; Jaynes, 2003; Lindley, 1972).

Within the Bayesian framework, one may quantify the evidence for one hypothesis relative to another. The Bayes factor is the most commonly used (although certainly not the only possible) Bayesian measure for doing so (Jeffreys, 1961; Kass & Raftery, 1995). The Bayes factor is the probability of the data under one hypothesis relative to the other. When a hypothesis is a simple point, such as the null, then the probability of the data under this hypothesis is simply the likelihood evaluated at that point. When a hypothesis consists of a range of points, such as all positive effect

sizes, then the probability of the data under this hypothesis is the weighted average of the likelihood across that range. This averaging automatically controls for the complexity of different models, as has been emphasized in Bayesian literature in psychology (e.g., Pitt et al., 2002; Rouder et al., 2009).

We take as the null that a parameter α is restricted to 0 (i.e., $H_0 : \alpha = 0$), and take as the alternative that α is not zero (i.e., $H_A : \alpha \neq 0$). In this case, the Bayes factor given data D is simply the ratio

$$BF_{A0} = \frac{p\left(D \mid H_A\right)}{p\left(D \mid H_0\right)} = \frac{\int p\left(D \mid H_A, \alpha\right) p\left(\alpha \mid H_A\right) \, \mathrm{d}\alpha}{p\left(D \mid H_0\right)},$$

where the integral in the denominator takes the average evidence over all values of α, weighted by the prior probability of those values $p\left(\alpha \mid H_A\right)$ under the alternative hypothesis.

An alternative —but formally equivalent— conceptualization of the Bayes factor is as a measure of the change from prior model odds to posterior model odds, brought about by the observed data. This change is often interpreted as the *weight of evidence* (Good, 1983; Good, 1985). Before seeing the data D, the two hypotheses H_0 and H_A are assigned prior probabilities $p(H_0)$ and $p(H_A)$. The ratio of the two prior probabilities defines the *prior odds*. When the data D are observed, the prior odds are updated to *posterior odds*, which is defined as the ratio of the posterior probabilities $p(H_0 \mid D)$ and $p(H_A \mid D)$:

$$\frac{p(H_A \mid D)}{p(H_0 \mid D)} = \frac{p(D \mid H_A)}{p(D \mid H_0)} \times \frac{p(H_A)}{p(H_0)}. \tag{11.1}$$

Equation 11.1 shows that the change from prior odds to posterior odds is quantified by $p(D \mid H_A)/p(D \mid H_0)$, the Bayes factor BF_{A0}.

Under either conceptualization, the Bayes factor has an appealing and direct interpretation as an odds ratio. For example, $BF_{A0} = 2$ implies that the data are twice as likely to have occurred under H_A than under H_0. Jeffreys (1961), proposed a set of verbal labels to categorize the Bayes factor according to its evidential impact. This set of labels, presented in Table 11.3, facilitates scientific communication but should only be considered an approximate descriptive articulation of different standards of evidence (Kass & Raftery, 1995).

Table 11.3 Evidence Categories for the Bayes Factor BF_{A0} (Jeffreys, 1961).

Bayes factor			Interpretation
	$>$	100	Decisive evidence for H_A
30	–	100	Very strong evidence for H_A
10	–	30	Strong evidence for H_A
3	–	10	Substantial evidence for H_A
1	–	3	Anecdotal evidence for H_A
	1		No evidence
1/3	–	1	Anecdotal evidence for H_0
1/10	–	1/3	Substantial evidence for H_0
1/30	–	1/10	Strong evidence for H_0
1/100	–	1/30	Very strong evidence for H_0
	$<$	1/100	Decisive evidence for H_0

Note. We replaced the label "worth no more than a bare mention" with "anecdotal". Note that, in contrast to p values, the Bayes factor can quantify evidence in favor of the null hypothesis.

In general, calculating Bayes factors is more difficult than calculating p values and effect sizes. However, psychologists can now turn to easy-to-use Web pages to calculate the Bayes factor for many common experimental situations or use software such as WinBUGS (Lunn et al., 2000; Wetzels et al., 2009; Wetzels, Lee, & Wagenmakers, 2010).[3] In this article, we use the Bayes factor calculation described in Rouder et al. (2009). Rouder et al.'s development is suitable for one-sample and two-sample designs, and the only necessary input is the t value and sample size.

The Bayes factor that we report in this article is the result of a default Bayesian t test (for details see Rouder et al., 2009). The test is default because it applies regardless of the phenomenon under study: For every experiment, one uses the same prior on effect size for the alternative hypothesis, the Cauchy(0,1) distribution. This prior has statistical advantages that make it an appropriate default choice (for example, it has excellent theoretical properties in the limit, when $N \to \infty$ and $t \to \infty$; for details see Liang et al., 2008).

The default test is easy to use and avoids informed specification of prior distributions that other researchers may contest. Conversely, one may argue that the informed specification of priors is the appropriate way to take problem-specific prior knowledge into account. Bayesian statisticians are divided over the relative merits of default versus informed specifications of prior distributions (Press, Chib, Clyde, Woodworth, & Zaslavsky, 2003). In our opinion, the default test provides an excellent starting point of analysis, one that may later be supplemented with a detailed problem-specific analysis (see Dienes, 2011, 2008; Kruschke, 2010a, 2010b, 2011, for additional discussion of informed priors).

In our concrete example, the resulting Bayes factor for $t = 2.09$ and a sample size of 20 observations is $BF_{A0} = 1.56$. Accordingly, the data are 1.56 times more likely to have occurred under the alternative hypothesis than under the null hypothesis. This Bayes factor falls into the category "anecdotal". In other words, this Bayes factor indicates that although the alternative hypothesis is slightly favored, we do not have sufficiently strong evidence from the data to reject or accept either hypothesis.

11.3 Comparing p Values, Effect Sizes and Bayes Factors

For our concrete example, the three measures of evidence are not in agreement. The p value was on the cusp between the categories "no evidence against H_0" and "positive evidence against H_0", the effect size indicates a large to very large effect size, and the Bayes factor indicates that the data support the null hypothesis almost as much as they support the alternative hypothesis. If this example is not an isolated one, and the measures differ in many psychological applications, then it is important to understand the nature of those differences.

To address this question, we studied all of the empirical results evaluated by a t test in the 2007 volumes of *PBR* and *JEP:LMC*. This sample was comprised of 855 t tests from 252 articles. These articles covered 2,394 journal pages and addressed many topics that are important in modern experimental psychology. Our sample suggests, on average, that an article published in *PBR* and *JEP:LMC* contains about 3.4 t tests, which amounts to one t test for every 2.8 pages. For simplicity we did not include t tests that result from multiple comparisons in ANOVA designs (for a Bayesian perspective on multiple comparisons see Scott & Berger, 2006). Even though our t tests are sampled from the field of experimental/cognitive psychology, we expect our findings to generalize to many other subfields of psychology, as long as the studies in these subfields use the same level

[3] A Web page for computing a Bayes factor online is `http://pcl.missouri.edu/bayesfactor`, and a Web page to download a tutorial and a flexible R/WinBUGS function to calculate the Bayes factor can be found at `www.ruudwetzels.com`.

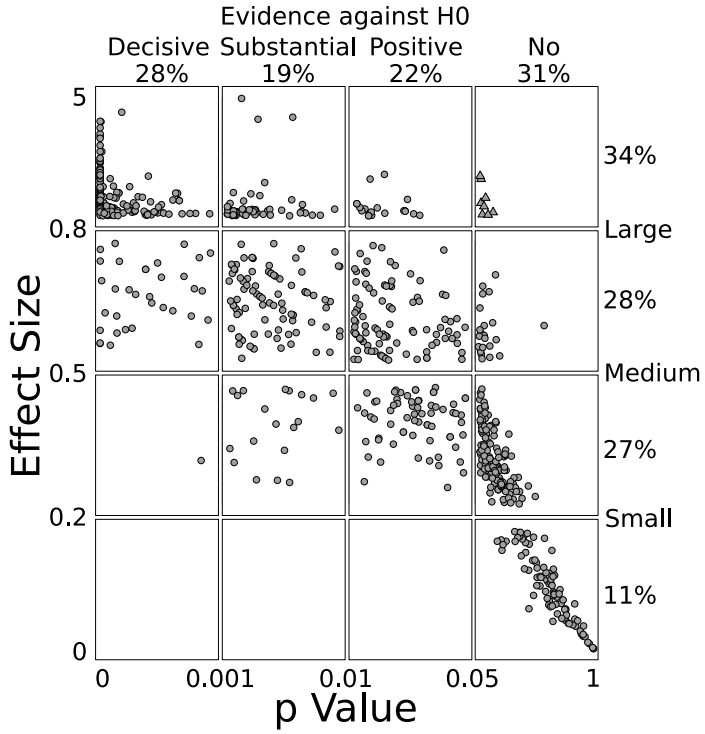

Figure 11.1 The relationship between effect size and p values. Points denote comparisons (855 in total). Points denoted by circles indicate relative consistency between the effect size and p value, whereas those denoted by triangles indicate gross inconsistency. The scale of the axes is based on the decision categories, as given in Table 11.1 and Table 11.2.

of statistical significance, approximately the same number of participants, and approximately the same number of trials per participant (Howard et al., 2000).

In the next sections we describe the empirical relation between the three measures of evidence, starting with the relation between effect sizes and p values.

Comparing Effect Sizes and p Values

The relationship between the obtained p values and effect sizes is shown as a scatter plot in Figure 11.1. Each point corresponds to one of the 855 comparisons. Different panels are introduced to distinguish the different evidence categories, as given in Table 11.1 and Table 11.2.

Figure 11.1 suggests that p values and effect sizes capture roughly the same information in the data. Large effect sizes tend to correspond to low p values, and small effect sizes tend to correspond to large p values. The two measures, however, are far from identical. For instance, a p value of 0.01 can correspond to effect sizes ranging from about 0.2 to 1, and an effect size near 0.5 can

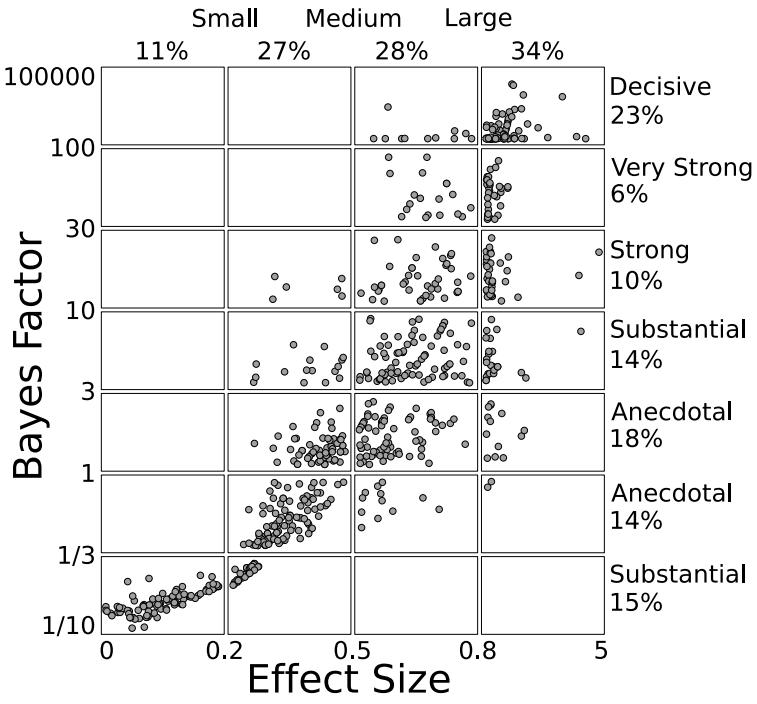

Figure 11.2 The relationship between Bayes factor and effect size. Points denote comparisons (855 in total). The scale of the axes is based on the decision categories, as given in Table 11.2 and Table 11.3.

correspond to p values ranging from about 0.001 to 0.05. The triangular points in the top-right panel of Figure 11.1 highlight gross inconsistencies. These 8 studies have a large effect size, above 0.8, but their p values do not indicate evidence against the null hypothesis. A closer examination revealed that these studies had p values very close to 0.05, and were comprised of small sample sizes.

Comparing Effect Sizes and Bayes Factors

The relationship between the obtained Bayes factors and effect sizes is shown in Figure 11.2. Much as with the comparison of p values with effect sizes, it seems clear that the default Bayes factor and effect size generally agree, though not exactly. No striking inconsistencies are apparent: No study with an effect size greater than 0.8 coincides with a Bayes factor below 1/3, nor does a study with very low effect size below 0.2 coincide with a Bayes factor above 3. The two measures, however, are not identical. They differ in the assessment of strength of evidence. Effect sizes above 0.8 range all the way from anecdotal to decisive evidence in terms of the Bayes factor. Also note that small to medium effect sizes (i.e., those between 0.2 and 0.5) can correspond to Bayes factor evidence in favor of either the alternative or the null hypothesis.

This last observation highlights that Bayes factors may quantify support for the null hypothesis. Figure 11.2 shows that about one-third of all studies produced evidence in favor of the null hypothesis. In about half of these studies favoring the null, the evidence is substantial. Because of the file-drawer problem (i.e., only significant effects tend to get published) this is an underestimate of the true amount of null findings and their Bayes factor support.

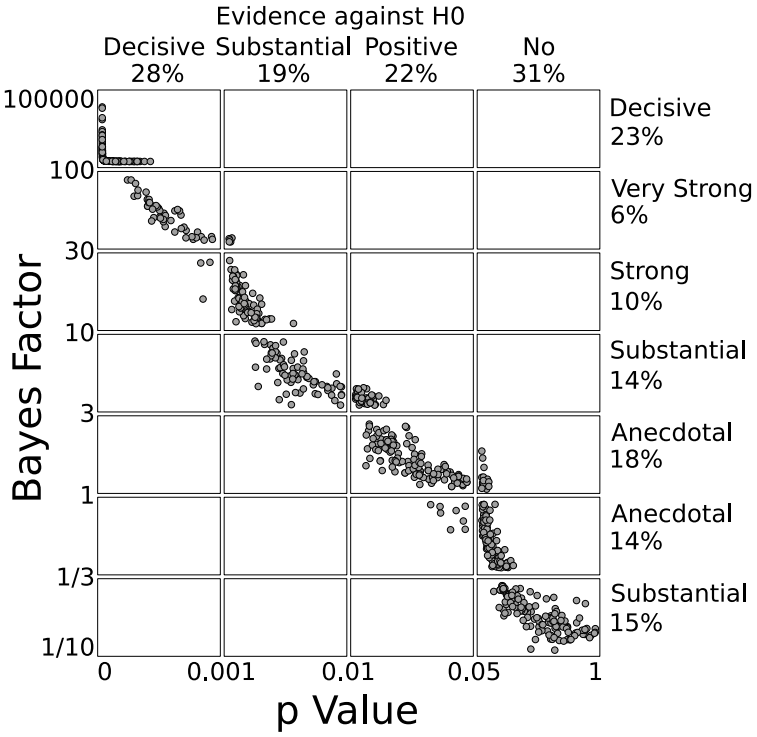

Figure 11.3 The relationship between Bayes factor and p value. Points denote comparisons (855 in total). The scale of the axes is based on the decision categories, as given in Table 11.1 and Table 11.3.

Comparing p Values and Bayes Factors

The relationship between the obtained Bayes factors and p values is shown in Figure 11.3, again using interpretative panels. It is clear that default Bayes factors and p values largely covary with each other. Low Bayes factors correspond to high p values and high Bayes factors correspond to low p values, a relationship that is much more exact than for our previous two comparisons. The main difference between default Bayes factors and p values is one of calibration; p values accord more evidence against the null than do Bayes factors. Consider the p values between 0.01 and 0.05, values that correspond to "positive evidence" and that usually pass the bar for publishing in academia. According to the default Bayes factor, 70% of these experimental effects convey evidence in favor of the alternative hypothesis that is only "anecdotal". This difference in the assessment of the strength of evidence is dramatic and consequential.

11.4 Conclusions

We compared p values, effect sizes and default Bayes factors as measures of statistical evidence in empirical psychological research. Our comparison was based on a total of 855 different t statistics from all published articles in two major empirical journals in 2007. In virtually all studies, the three different measures of evidence are broadly consistent: Small p values correspond to large effect sizes and large Bayes factors in favor of the alternative hypothesis. Despite the fact that the measures of evidence reach the same conclusion about what hypothesis is best supported by the data, however, the measures differ with respect to the strength of that support. In particular, we noted that p values between 0.01 and 0.05 often correspond to what, in Bayesian terms, is only anecdotal evidence favor of the alternative hypothesis. The practical ramifications of this are considerable.

Practical Ramifications

Our results showed that when the p value falls in the interval from 0.01 to 0.05, there is a 70% chance that the default Bayes factor indicates the evidence for the alternative hypothesis to be only anecdotal or "worth no more than a bare mention"; this means that the data are no more than three times more likely under the alternative hypothesis than they are under the null hypothesis. Hence, for the studies under consideration here, it seems that a p value criterion more conservative than 0.05 is appropriate. Alternatively, researchers could avoid computing a p value altogether and instead compute the Bayes factor. Both methods help prevent researchers from overestimating the strength of their findings and help the field from incorporating ambiguous findings as if they were real and reliable (Ioannidis, 2005).

As a practical illustration, consider a series of recent experiments on precognition (Bem, 2011). In nine experiments with over 1,000 participants, Bem intended to show that precognition exists, that is, that people can foresee the future. And indeed, eight out of nine experiments yielded a significant result. However, most p values fell in the ambiguous range of 0.01 to 0.05, and, across all nine experiments, a Bayes factor analysis indicates about as much evidence for the alternative hypothesis as against it (Kruschke, 2011; Wagenmakers et al., 2011). We believe that this situation typifies part of what could be improved in psychological research today. It is simply too easy to obtain a p value below 0.05 and subsequently publish the result.

When researchers publish ambiguous results as if they were real and reliable, this damages the field as a whole: Time, effort, and money will be invested to replicate the phenomenon, and, when replication fails, the burden of proof is almost always on the part of the researcher who, after all, failed to replicate a phenomenon that was demonstrated to be present (with a p value in between 0.01 and 0.05).

Thus, our empirical comparison shows that the academic criterion of 0.05 is too liberal. Note this problem would not be solved by opting for a stricter significance level, such as 0.01. It is well known that the p value decreases as the sample size n increases. Hence, if psychologists switch to a significance level of 0.01 but inevitably increase their sample sizes to compensate for the stricter statistical threshold, then the phenomenon of anecdotal evidence will start to plague p values even when these p values are lower than 0.01. Therefore, we make a case for Bayesian statistics in the next section.

A Case for Bayesian Statistics

We have compared the conclusions from the different measures of evidence. It is easy to make a case for Bayesian statistical inference in general, based on arguments already well documented in statistics and psychology (e.g., Dienes, 2008; Jaynes, 2003; Kruschke, 2010c, 2010a; M. D. Lee & Wagenmakers, 2005; Lindley, 1972; Wagenmakers, 2007). We briefly mention three arguments here.

First, unlike null hypothesis testing, Bayesian inference does not violate basic principles of rational statistical decision making such as the stopping rule principle or the likelihood principle (Berger & Delampady, 1987; Berger & Wolpert, 1988). This means that the results of Bayesian inference do not depend on the intention with which the data were collected. As stated by Edwards et al. (1963, p. 193), "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience".

Second, Bayesian inference takes model complexity into account in a rational way. Specifically, the Bayes factor has the attraction of not assigning a special status to the null hypothesis and so makes it theoretically possible to measure evidence in favor of the null (e.g., Dennis et al., 2008; Gallistel, 2009; Kass & Raftery, 1995; Rouder et al., 2009).

Third, we believe that Bayesian inference provides the kind of answers that researchers care about. In our experience, researchers are usually not that interested in the probability of encountering data at least as extreme as those that were observed, given that the null hypothesis is true and the sample was generated according to a specific intended procedure. Instead, most researchers want to know what they have learned from the data about the relative plausibility of the hypotheses under consideration. This is exactly what is quantified by the Bayes factor.

These advantages notwithstanding, the Bayes factor is not a measure of the mere size of an effect. Hence, the measure of effect size confers additional information, particularly when small numbers of participants or trials are involved. So, especially for these sorts of studies, there is an argument for reporting both a Bayes factor and an effect size. We note that, from a Bayesian perspective, the effect size can naturally be conceived as a (summary statistic of) the posterior distribution of a parameter representing the effect, under an uninformative prior distribution. In this sense, a standard Bayesian combination of parameter estimation and model selection could encompass all of the useful measures of evidence we observed (for an example of how Bayes factor estimation can be incorporated in a Bayesian estimation framework, see, for instance, Kruschke, 2011).

Our final thought is that reasons for adopting a Bayesian approach now are amplified by the promise of using an extended Bayesian approach in the future. In particular, we think the hierarchical Bayesian approach, which is standard in statistics (e.g., Gelman & Hill, 2007), and is becoming more common in psychology (e.g., Kruschke, 2010c, 2010b; M. D. Lee, 2011; Rouder & Lu, 2005) could fundamentally change how psychologists identify effects. Hierarchical Bayesian analysis can be a valuable tool both for meta-analyses and for the analysis of a single study. In the meta-analytical context, multiple studies can be integrated, so that what is inferred about the existence of effects and their magnitude is informed, in a coherent and quantitative way, by a domain of experiments. In the context of a single experiment, a hierarchical analysis can be used to take variability across participants or items into account.

In sum, our empirical comparison of 855 t tests shows that three often-used measures of evidence —p values, effect sizes, and Bayes factors— almost always agree about what hypothesis is better supported by the data. The measures often disagree about the strength of this support: for those data sets with p values in between 0.01 and 0.05, about 70% are associated with a Bayes factor

that indicates the evidence to be only anecdotal or "worth no more than a bare mention" (Jeffreys, 1961). This analysis suggests that many results that have been published in the literature are not established as strongly as one would like.