The *p*-Value Is Best to Detect Effects

Donald B. Macnaughton
MatStat, 30 Greenfield Ave. Suite 1503, Toronto, ON M2N 6N3, Canada
E-Mail: donmac@matstat.com

ABSTRACT

Researchers study relationships between variables in scientific research as a means to accurate prediction and control. We need a measure of the weight of evidence that a relationship between variables (or other effect) observed in the data from a sample is a *real* (i.e., reproducible) effect in the entities in the underlying population. This paper compares nine measures of the weight of evidence that an effect is real (*p*-value, *t*-statistic, confidence interval, likelihood ratio, Bayes factor, posterior probability that the null hypothesis is true, second-generation *p*-value, *D*-value, and information criteria). The comparisons suggest that the *p*-value is slightly better than the other measures for detecting relationships between variables (or other effects) in populations in scientific research.

KEYWORDS: Hypothesis testing; Significance testing; Statistical inference; Role of statistics in scientific research

1. Introduction

Scientific researchers often study relationships between variables in entities in populations of entities. We study a relationship by first selecting a reasonable *sample* of entities from the population. Then we determine whether a chosen response variable, y, measured in each entity in the sample, "depends" on one or more chosen predictor variables, x, also measured in each entity. If we do this properly, it enables us to infer from the sample whether the relationship of interest exists between the variables in the population behind the sample.

We study relationships between variables because if we can find a useful new relationship, then we can use our knowledge of the relationship to predict or control the value of the response variable for new entities from the population. The ability to predict or control variables (properties of entities) is often useful in a social, theoretical, or commercial sense.

For example, medical researchers study relationships between variables in populations of medical patients. If a medical researcher can find good evidence of a useful (causal) relationship between a medical treatment variable, x, and a measure of patients' health, y, then (after the relationship has been confirmed in independent research) medical doctors can use the knowledge of the relationship to improve the health of new patients from the population.

More generally, a relationship between variables is one type of "effect" that we can study in data in scientific research. Most effects can be viewed as (aspects of) relationships between variables. However, we keep the term "effect" because it is shorter and more general.

A key question in studying any relationship between variables (or other effect) is whether the relationship actually *exists* in the population—whether it is *real* (i.e., reproducible). This is important because we often find in scientific research that a potentially useful relationship between variables apparently *doesn't* exist in the population. (Or at least it doesn't exist strongly enough to be reliably detected by our current measurement methods.) If we have no good evidence that a certain relationship between variables exists, then it is a waste of resources to think or act as if the relationship *does* exist.

This paper discusses some basic principles and nine reliable measures to help researchers to detect relationships between variables in populations of entities. There is substantial controversy about the usefulness of these ideas, with the attention mostly focused on the *p*-value, which is the most popular of the nine measures. The controversy arises because the role of the ideas in scientific research is often misunderstood.

The controversy has led the American Statistical Association (ASA) to publish a careful position paper on *p*-values titled "ASA Statement on Statistical Significance and *P*-Values". The Introduction section of the statement says that:

> While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted (Wasserstein 2016, p. 131).

The main part of the ASA statement gives six widely agreed upon principles that underlie the proper use and interpretation of the *p*-value in scientific research. These important principles are each discussed later below.

The fact that the *p*-value (and the eight other measures) is commonly misused and misinterpreted together with perceived logical problems with the measures has led some statisticians to recommend that we reduce our reliance on the measures by abandoning the use of critical values (thresholds) for the measures (McShane, Gal, Gelman, Robert, and Tackett, 2018). However, these statisticians have failed to present a viable alternative approach for what is arguably a necessary function in scientific research. The function is to provide a reliable way to distinguish the signal we seek (typically a signal that a relationship exists between variables) from the inevitable noise in the data. This function is important because we don't want to draw scientific conclusions from mere noise in data. If we omit using any of the nine measures with a critical value, then how can we reliably and efficiently distinguish likely real effects from likely noise in data?

Some less experienced people think that the nine measures somehow *decide* whether relationships exist between variables, which is a fundamental error. The correct interpretation is that the measures *help* the relevant *research community* (e.g., experimental psychologists) to decide.

This paper focuses on the role or function of the nine measures in scientific research. This leads to discussion of some basic principles of scientific research. To establish a solid foundation, the paper also discusses some basic principles of statistics. To reinforce the ideas for less experienced readers, the paper gives various examples.

The paper illustrates how it is possible to discuss the role or function of the measures of the weight of evidence in scientific research with only passing reference to the highly important underlying mathematics. This is possible because the function of the measures (i.e., to help us to detect relationships between variables) isn't directly mathematical. The mathematics serves to *support* the scientific function. (The mathematics does this extremely well due to its unlimited flexibility and generality.)

A focus on function together with omission of the mathematical details and inclusion of basic ideas and examples makes the ideas easier for less experienced people to understand. Given the broad misuse and misinterpretation of the ideas, focusing on the function and the basic ideas is a sensible approach.

The paper proceeds as follows: Section 2 describes the roles of the "research" and "null" hypotheses in detecting relationships between variables in scientific research. Section 3 explains how researchers use a measure of the "weight of evidence" to attempt to "reject" the null hypothesis and conclude that a relationship exists between the studied variables. Section 4 discusses two serious but controllable errors that the measures of the weight of evidence sometimes make. Section 5 explains the functional operation of each of nine popular measures of the weight of evidence that a relationship exists between variables. Section 6 compares the nine measures on relevant attributes. Section 7 gives conclusions.

2. Statistical Hypothesis Tests to Detect Relationships

A sensible way to detect a relationship between variables is to perform a statistical hypothesis test. We begin the test by stating (at least implicitly) two mutually exclusive and exhaustive hypotheses—the "research" hypothesis and the "null" hypothesis. The research hypothesis says that a relationship exists between a specified predictor variable (x) and the response variable (y) in the entities in the population. In contrast, the null hypothesis says that *no* relationship exists between x and y in the entities.

For example, if we are studying the relationship between physical exercise and heart health in people, then the research hypothesis says that there is a relationship between physical exercise and heart health in the studied population of people. In contrast, the null hypothesis says that there is *no* relationship between the two variables in the population.

More generally, a research hypothesis may say that a relationship exists between a specified set of *multiple* predictor variables, x (now a *vector* of variables), and the response variable, y, in the entities in the population.

Some statisticians refer to a research hypothesis in a scientific research project as the "alternative" or "alternate" hypothesis. However, those terms are inappropriate because they incorrectly suggest that the research hypothesis is *subordinate* to the null hypothesis. Clearly, it is the null hypothesis that is subordinate because it is merely an empty starting point that we hope to escape from. The *research* hypothesis is the essential idea because it is what we hope to prove is true.

The scientific principle of parsimony tells us to keep our ideas as simple as possible while remaining consistent with the known facts (Baker, 2016). The null hypothesis is simpler than the research hypothesis because the null hypothesis has fewer details. Therefore, the standard approach is to begin the study of a new relationship between variables with the assumption that the null hypothesis is true. That is, we begin with the *formal* assumption that there is no relationship whatever between the variables of interest.

Of course, *informally* we usually strongly believe (hope) the opposite—we believe that the research hypothesis is true. We believe that the research hypothesis is true because that is why we are doing the research—we want (among other things) to demonstrate that our carefully chosen research hypothesis is true in the population because that will usefully advance human knowledge.

After formally assuming that the null hypothesis is true, we perform a scientific research project to see if we can find good evidence to enable us to "reject" the null hypothesis. We (a) select a sample of entities from the population, (b) measure the values of the same relevant variables in each entity in the sample, (c) collect the measured values in a data table, and (d) detect a relationship between the variables by examining the data to see if there is good evidence that the relationship exists. If we can find good evidence of the relationship in the sample data, and if we have done everything properly, this enables us to reject the null hypothesis and to (tentatively) decide that the relationship exists in the population.

Some statisticians believe that the null hypothesis is never precisely true in scientific research, as discussed in appendix B.12. However, regardless of whether these statisticians are correct in their belief, statistical hypothesis testing is still sensible, as discussed in appendix C.

3. Hypothesis Testing Methods

Statisticians have invented nine sensible methods to help us to examine appropriate research data to determine whether we have enough evidence to reject the null hypothesis and (tentatively) conclude that a relationship exists between the studied variables. The present section and the next section discuss some general principles behind the methods to prepare for separate discussion of each method in section 5.

All the methods work by (in effect) computing a *measure* of the weight of evidence that the relationship of interest exists. The lower (or, for some measures, the higher) the computed value of the measure for a given data table, the more evidence we have from the data that the studied relationship

exists in the population—the more evidence we have that the relationship is real.

For example, the *p*-value is designed so that if certain reasonable assumptions are adequately satisfied, then the lower the *p*-value that is computed from a data table, the greater the weight of evidence in the data that the studied relationship between the variables is real. Thus we can compute a *p*-value for a relationship between variables from a data table. And if the *p*-value is *low enough*, and if there is no reasonable alternative explanation for the low *p*-value, then we can conclude that we have good evidence that the studied relationship exists between the variables in entities in the population—good evidence that the relationship is real.

Researchers have sensibly defined so-called "critical values" for many of the measures. If the value of a measure of the weight of evidence falls beyond (or is equal to) the critical value, then this is called a positive result. For example, if the *p*-value obtained in a statistical test in a research project is less than (or equal to) a critical value (often 0.05), this is a positive result. A positive result implies (in the absence of a reasonable alternative explanation) that we have found good evidence of a relationship between the studied variables (or good evidence of some other studied effect). Scientific journals are eager to publish interesting properly-obtained positive results about new relationships between variables.

In contrast, if the value of a measure of the weight of evidence *doesn't* fall beyond the critical value, then this is called a negative result. A negative result is almost always disappointing for a researcher because it implies that the research has *failed* to find good evidence of the sought-after relationship between the variables. Scientific journals are rarely interested in publishing negative results because these results usually don't tell us anything beyond what we have already assumed to be true.

Some people refer to the fact that scientific journals generally only publish positive results as "publication bias". This reflects the fact that these people believe that all research results (positive and negative) should be published. However, negative results generally aren't published because readers generally aren't interested in reading about effects that may reflect mere noise in the data. Readers generally don't have time for that. Positive results are much more interesting because they tell us about relationships between variables. Negative results tell us little beyond the fact that the research project failed to find what it was looking for. Appendices K and L in the supplementary material discuss some special cases when negative results are interesting.

Appendix G in the supplementary material discusses a case of a study of a relationship between variables when we don't need a measure of the weight of evidence that the relationship is real.

A given research community (e.g., medical researchers) chooses the critical value for a measure of the weight of evidence based on a sense that the chosen critical value maximizes the long-term payoff of scientific research in the community. These ideas are further discussed in the next section.

The critical values that are chosen by a research community are reflected in the editorial policies of the community's journals. That is, the editor of a journal might specify that the *p*-value for the main research finding in a paper submitted to the journal must be less than or equal to the critical value of 0.05 before the journal will view the results as being convincing enough to *consider* the paper for publication. (Higher-prestige journals often use the stricter critical *p*-value of 0.01.) Similarly, an editor of a Bayesian journal might specify that the Bayes factor for the main research finding in a paper must be greater than 10 before the journal will consider the paper for publication. This practice enables editors to control the rate of publication of false-positive errors in their journals, as discussed below.

A reader suggested that even if the value of a measure of the weight of evidence *doesn't* fall beyond the critical value, but is close to the critical value (e.g., 0.06 in the case of the *p*-value), then the result might still be useful. This is fully correct. However, in the interest of saving time, many scientific research communities opt to use a critical value. This enables us to avoid quibbling about whether results are convincing enough to deserve comprehensive study. This approach is necessary because there are many more research results for potential study than we have time to study. You must be taller than 4 feet to be allowed on this ride.

Of course, if a measure of the weight of evidence for a particular effect fails to fall beyond the critical value, but the researcher continues to believe that the effect is real, then he or she should consider repeating the research project with a more powerful research design to attempt to obtain convincing evidence that the effect exists in the population. If the researcher can obtain such evidence, a journal will be pleased to consider a report of the result for publication.

All the measures of the weight of evidence are based on certain underlying assumptions about the data, with the nature of the assumptions depending on the situation at hand, with the assumptions generally being either identical or similar from measure to measure. These assumptions pertain to how the sample was selected from the population, how the data were analyzed, and whether the data exhibit certain necessary technical features.

The underlying assumptions of a statistical procedure are often adequately satisfied in carefully performed scientific research. But we must always *confirm* that the assumptions are adequately satisfied before we trust a procedure. Macnaughton (2016) discusses how a failure to check assumptions led to an invalid estimate of the Boltzmann constant in physics. Fortunately, modern software for studying relationships between variables automatically provides information in the output to help us to confirm that the computer-checkable assumptions about the data in a data table are adequately satisfied.

As noted in Principle 5 in the ASA Statement on *p*-values, a measure of the weight of evidence that an effect is real doesn't directly tell us anything about either the *strength* (size) or the *importance* of the effect (Wasserstein 2016). Instead, the measure only tells us whether (in the absence of a reasonable alternative explanation) we have good evidence that the effect is *real*, that it *exists*. Failing to distinguish between the existence, the strength, and the importance of relationships between variables is a frequent source of confusion among laypeople. Of course, we can't sensibly discuss the

strength or the importance of an effect until we have first established that it (likely) exists in the population.

4. False-Positive and False-Negative Errors

All the measures of the weight of evidence sometimes make two types of serious errors. First, the values of the measures will occasionally fall beyond their critical value even though there is no relationship (or *effectively* no relationship) between the variables. When this happens, it is called a false-positive error. False-positive errors are highly undesirable in scientific research because they lead us to believe that a relationship exists between variables when *no* relationship (or possibly only an undetectable weak relationship) exists. This leads to a waste of resources for anyone who tries to study or use the nonexistent relationship.

False-positive errors can occur through two independent mechanisms. First, a false-positive error can be caused by random noise in the data. It is easy to show that the noise guarantees that a measure of the weight of evidence will occasionally exceed the critical value even when there is no effect present in the population. If this happens, we will think that we have found a relationship between the variables. But we will be wrong.

Due to the inevitability of random variation, we can't eliminate false-positive errors that are due to random variation in scientific research. However, we can easily control the rate at which these errors occur. We do this by the choice of the critical value for the measure of the weight of evidence. The stricter we set the critical value, the lower the false-positive error rate. For example, the lower we set the critical *p*-value (or the higher we set the critical Bayes factor), the lower the false-positive error rate.

Clearly, using stricter critical values is sensible to reduce false-positive errors. However, using stricter critical values increases research costs (if statistical power is held constant). Therefore, we must compromise to contain costs. The need to compromise has led many statisticians and researchers to agree that critical *p*-values of 0.05 or 0.01 are sensible choices for controlling false-positive errors while achieving reasonable power without driving costs too high. These ideas are expanded in appendix E.

The second way that false-positive errors occur is through errors made by the researcher in conducting the research or in analyzing and interpreting the data. These errors include carelessness and failure to take account of relevant extenuating factors.

In scientific research, we handle all the researcher errors with a single simple but comprehensive rule: There must be *no reasonable alternative explanation* for why the value of a measure of the weight of evidence has fallen beyond its critical value before we can trust it and believe that the studied relationship between variables exists. Scientists are strict about this rule because we wish to be definitive. So if someone finds a reasonable alternative explanation (of any kind, including researcher errors) for a research finding, then this weakens the finding, usually to the point of making it inconclusive. The relevant scientific community decides what is "reasonable" through informal consensus, sometimes after much debate.

Some statisticians pay little or no attention to the importance of eliminating false-positive errors in scientific research. This may be because they are more interested in obtaining positive results than in worrying about errors. Thus they view all results as "positive", even when they may be studying mere noise in the data.

In contrast, researchers worry about making false-positive errors because if a published false-positive research result is at least moderately important, then the error will invariably be exposed (sooner or later) due to the investigative nature of science. Such an error is bad for a researcher's reputation because other researchers will be displeased by the waste of resources it caused. And some people will think that the error may have been caused by carelessness though, as noted, it also may have been caused by bad luck (random noise).

The presence of false-positive errors in scientific research implies that a certain proportion of attempts to replicate research findings will fail. Appendix B.10 discusses the rate of occurrence of false-positive errors in scientific research and appendix B.15 discusses the "replication crisis" or "reproducibility crisis" in modern scientific research.

In contrast to false-positive errors, the measures of the weight of evidence also sometimes make false-*negative* errors. That is, the value of a measure of the weight of evidence may *fail* to fall beyond its critical value even though the studied relationship between variables is present in enough strength in the population. Like false-positive errors, false-negative errors may be caused by random noise in the data or by researcher errors.

Like false-positive errors, false-negative errors are highly undesirable in scientific research, but for a different reason. A false-negative error reflects a failure to discover an extant and perhaps useful relationship between variables, which reflects a loss of potentially useful knowledge. Also, a false-negative error for an important result leads to a loss of satisfaction, prestige, and monetary reward for the researcher.

In theory, we can control the rate of false-negative errors in scientific research by the choice of the critical value of the measure of the weight of evidence. That is, the *less* strict we set the critical value, the lower the false-negative error rate. However, we can't do that in practice because we are already using the critical value to control the socially more important false-*positive* error rate.

False-positive errors are (immediately) socially more important in scientific research than false-negative errors because false-positive errors lead to wasted resources for other researchers who attempt to replicate or use the false result. If the other researchers can't replicate the result, this displeases them and it displeases the rest of the scientific community because they feel that they may have been misled. In contrast, false-negative errors are directly and immediately costly mainly to the original researcher in the sense that a false-negative error leads to a loss of reward. Therefore, we can count on knowledgeable researchers to use appropriate methods to minimize the possibility of false-negative errors in their research.

Researchers reduce the rate of false-negative errors in scientific research by increasing the "power" of the statistical tests that they use to detect relationships. The power tells us what would happen if we were to repeat the same research project over and over, each time drawing a fresh sample of entities from the population. The power of a statistical test is the fraction of the time that the test will successfully detect the studied relationship between variables under the assumption that the relationship has a certain form (as specified by the researcher) in the population.

Researchers who are planning a research project can use statistical power software to compute ahead of time the power of the statistical tests in the research project to ensure that the power is adequate. Ideally, the power of the main statistical tests in a scientific research project should be at least 0.9 for the expected form to ensure that we have a good chance to detect the relationship, if it exists. Sadly, the power of the statistical tests for the expected form in some research projects is substantially less than 0.9, leading to false-negative errors in cases when the relationship under study actually does exist in the population.

A researcher can increase the power of statistical tests by using (a) larger samples, (b) more precise measures of the values of the variables, (c) more (relevant) predictor variables, and (d) more efficient research designs, as discussed in statistics textbooks. However, the methods for increasing statistical power generally increase research costs so, as noted above, we must compromise to contain costs. Careful research design is the relatively inexpensive key to maximize power while controlling costs.

The always-present possibility of false-positive and false negative errors in scientific research is why a measure of the weight of evidence can't *decide* whether a relationship between variables (or other effect) is real. All that the measure can do is tell us whether we have "good evidence". This implies that we usually must "replicate" interesting positive research findings in independent new research before we can reasonably believe that the studied relationship between variables exists in the population. In the absence of a reasonable alternative explanation, a successful replication greatly reduces the chance that a positive result reflects a false-positive error.

The possibility of false-positive and false-negative errors justifies Principle 3 in the ASA Statement on *p*-values, which says that scientific conclusions and business or policy decisions shouldn't only be based only on whether a *p*-value is less than the critical value (Wasserstein 2016). A *p*-value, properly used, can *help* us to make decisions, but it can't make decisions on its own.

False-positive and false-negative errors are traditionally called Type 1 and Type 2 errors respectively. However, those names are inefficient because they have no descriptive content.

5. Details About the Nine Measures

As noted, there are nine common measures of the weight of evidence that an effect is real. (Other sensible measures might also be proposed.) This paper argues that the *p*-value is slightly better than the eight other measures. To support this, let us compare the measures. We first consider some key technical principles that underlie all nine measures.

In any situation in which we wish to test a research hypothesis about a relationship between variables, all the measures of the weight of evidence are similar in the sense that (when applicable) they are all derived (directly or indirectly) from the estimated *distribution* of the same parameter (or test statistic). The parameter is the relevant parameter of an appropriate model equation of the studied relationship between the variables.

For example, suppose we have carefully performed a research project to study the relationship between a set of continuous predictor variables $x_1, x_2, \ldots, x_q$ and a continuous response variable y for the entities in some population. And suppose we have collected the values of the x's and y for our sample in a data table. And suppose it is appropriate to study the relationship between the x's and y in the data with simple linear regression analysis. Then the model equation for the relationship can be written as

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_qx_q + \varepsilon \qquad (1)$$

where:

y is the response variable

$x_1, x_2, \ldots, x_q$ are the q predictor variables

$b_0, b_1, \ldots, b_q$ are the $q + 1$ parameters of the equation (called regression coefficients in this case), and

ε is the error term of the equation.

Suppose we wish to determine whether we have good evidence that the predictor variable associated with the ith term in the equation is related to the response variable. That is, we wish to determine whether we have good evidence of a relationship between the predictor variable x_i and y. We can make this determination by studying the estimated value of parameter b_i for the term, as estimated from our data table.

If there is no relationship between x_i and y, then the b_ix_i term doesn't belong in the equation. In that case, the true value of b_i in the population will be exactly zero because that makes the term vanish from the equation. Here, the value zero is called the "null" value of the parameter—the value that b_i will have in the population if the null hypothesis is or were true.

Thus, in symbols, the null hypothesis in this case is the hypothesis that $b_i = 0$. And the research hypothesis is the hypothesis that $b_i \neq 0$.

So we can decide whether we have good evidence that a relationship exists between x_i and y by determining whether the value of b_i estimated from the sample data is roughly equal to the null value, zero, or whether it is substantially different from zero. (Even if the correct value is exactly zero in the population, the value estimated from the sample will almost never be *exactly* zero due to random noise in the data.) If the estimated value of b_i is substantially different from zero, then we have good evidence that a relationship exists between x_i and y.

Thus we begin by "fitting" the model equation to the data, which gives us the best (in a reasonable mathematical sense) estimated values for each of the b's in the equation. This gives

us the estimated value of b_i, which we represent in the following discussion as $\hat{b}_i$. The "hat" indicates that this is the value estimated from the data for our sample, as opposed to the (unknown) true population value, b_i.

We also compute an estimate of the population standard error, s_i, of the sampling distribution of b_i. This estimate, which (counterintuitively) is readily computable in many research situations, tells us the estimated "average" *spread* of the estimates of the parameter we would get if we were to repeat the research project over and over, each time drawing a new sample from the population. We represent the estimated standard error as $\hat{s}_i$.

Researchers typically use the least-squares procedure to estimate parameter values and standard errors in linear regression (Chatterjee and Hadi, 2012), though other procedures are also sometimes sensibly used.

After we have obtained the estimated values, $\hat{b}_i$ and $\hat{s}_i$, we can have a computer draw a graph of the estimated sampling distribution of $\hat{b}_i$ under the assumption that the null hypothesis is true—under the assumption that $b_i = 0$. This graph illustrates how we compute several of the measures of the weight of evidence that the effect under study is real. Figure 1 is a computer-drawn graph illustrating how this estimated distribution might appear for the estimate of b_i in our research, as computed from our data table.

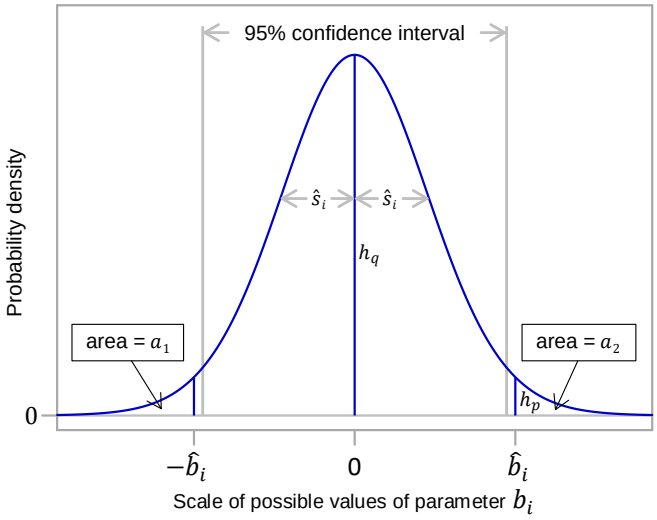

Figure 1. A graph showing the estimated sampling distribution (estimated probability density function) of the estimated value of parameter b_i of a linear regression model equation if the relevant null hypothesis is true.

No numbers except zero are shown on either axis of the figure. This is because the numbers are situation-dependent, but figure 1 is illustrating the general situation. Of course, in any specific situation (including our research) there will be numbers on the two axes. For example, the value $\hat{b}_i$ on the horizontal axis might fall at the value 24.32, which would be the numeric value of the parameter that was estimated from our data.

The horizontal axis of figure 1 shows a range of different possible values of b_i. This is the section of the range in which the unknown population value of b_i almost certainly has its true value.

As noted, the curving blue line shows the estimated sampling distribution function of multiple independent estimates of b_i under the assumption that the null hypothesis is true. This assumption implies that the distribution is *centered* on the null value, which is zero, as shown on the graph.

The value $\hat{s}_i$ is only an estimate of the true s_i. This implies that the shape of the distribution curve on the graph is a statistical *t*-distribution, as mathematically derived from the standard assumptions underlying linear regression analysis (Chatterjee and Hadi, 2012).

The spread (standard error) of the curve in figure 1 is the spread $\hat{s}_i$ that was estimated from the analysis of our research data. This value is shown by the two horizontal gray lines partway up the curve, each indicating the estimated value, $\hat{s}_i$. These two lines illustrate how the estimated standard error defines the estimated "width" of the distribution.

For a *t*-distribution, the value of $\hat{s}_i$ is roughly equal to the horizontal distance from the center of the distribution to either of the two inflection points on the curve (where the curve changes from bulging down to bulging up or vice versa). The value $\hat{s}_i$ estimated from the data plays a pivotal role in the following discussion because it specifies the width of the curve, on which all the conclusions directly or indirectly depend.

The *t*-distribution shown in figure 1 has 30 degrees of freedom, which is a relevant mathematical attribute of the distribution that depends mainly on the sample size—the degrees of freedom is always slightly less than the sample size. However, the graph would look quite similar if the degrees of freedom were different. And the underlying principles would be the same. If (as is typical) a *t*-distribution has 25 or more degrees of freedom, then the shape of the distribution is quite similar to the shape of the normal distribution, with the shape becoming closer and closer without bound to the normal distribution as the number of degrees of freedom increases.

The curving line on the graph shows the estimated *relative rate of occurrence* of different estimated values of b_i we would obtain if the null hypothesis is or were true in the population and if we were to perform the research project over and over, each time drawing a fresh random sample of entities from the population (and if the relevant underlying assumptions are satisfied). Of course, in a real research project, if the relevant null hypothesis is true, we experience only a single instance of the infinitely many instances of the research project that are depicted on the graph.

The curving line descends from its maximum point evenly on both sides of the null value. This tells us that if the null hypothesis is true, then the estimated values of b_i will fall with equal likelihood on either side of the null value, and values that are close to the null value are more likely to be estimated as the value of b_i than values that are farther away.

Of course, if the null hypothesis is *false*, then the distribution won't be centered on the null value, but will be centered on the true (non-zero, unknown) value of b_i in the population. And, of course, we hope that we will find good evidence that

the null hypothesis is false because that can help us to infer that the studied effect exists in the population.

The theory of sampling distributions implies that in figure 1 the *area under the curve* between any two points on the horizontal axis is exactly equal to the theoretical probability that the value of b_i estimated from research data will lie between the two points if the null hypothesis is or were true (and if the assumptions are satisfied). This implies that the total area under the curve is 1.00.

It is noteworthy that the curve shown in figure 1 is the estimated sampling distribution (under the null hypothesis) of both $\hat{b}_i$ and $\hat{b}_i/\hat{s}_i$ because $\hat{s}_i$ is here a constant. The *shape* of the distribution curve is identical for both distributions. But the units for the scale on the horizontal axis are different—respectively raw units and standardized units. (After the units on the horizontal axis are chosen, the units on the vertical axis are defined by the fact that the area under the curve must be exactly 1.00.)

The curve in figure 1 is sometimes referred to as the "likelihood function" under the null hypothesis. This is because the curve shows, as a reflection of the data at hand, the estimated relative likelihood of different estimated values of b_i being obtained if the null hypothesis is or were true (and if the underlying assumptions are satisfied).

As noted, in this example we assume that we have performed the research project a single time. Let us assume that the value of b_i that we have estimated from the data is the value $\hat{b}_i$ that is shown near the right end of the horizontal axis of the figure. Similarly, the negative of $\hat{b}_i$ is shown near the left end of the horizontal axis. In other research projects the estimated value $\hat{b}_i$ will lie at other places on the horizontal axis relative to the curve (and relative to the keystone $\hat{s}_i$), nearer to or farther away from the null value. But, if the underlying assumptions are adequately satisfied, the principles in the following discussion always apply.

As noted above, we wish to determine whether $\hat{b}_i$ is far enough from zero for us to believe that the null hypothesis is false. We shall see shortly how the ideas behind figure 1 give us several ways to make this determination.

It is important to understand that we can view the value of $\hat{b}_i$ as a measure of the effect size of the relationship (if any) between variables x_i and y. We will observe what happens to each of the measures of the weight of evidence if the absolute value of the estimated effect size, $\hat{b}_i$, increases or decreases (while $\hat{s}_i$ stays fixed). And we will see that all the measures of the weight of evidence can be interpreted as measuring how far the estimated value $\hat{b}_i$ is from zero.

The following discussion draws various conclusions. In real scientific research we can rightly draw these conclusions only if there is no reasonable alternative explanation for the findings. Therefore, to stay practical, we must keep the possibility of alternative explanations in mind. Let us now consider the nine measures of the weight of evidence that an effect is real in a population.

5.1. P-Value

The *p*-value for the effect shown in figure 1 is the sum of the two "tail" areas of the distribution curve, with the tails being defined by the locations of $\hat{b}_i$ and $-\hat{b}_i$. Thus the *p*-value in this example equals $a_1 + a_2$. In the present case, as in most cases (even when the distribution is asymmetric), the tail areas are defined so that $a_1 = a_2$.

For figure 1, the *p*-value (the value of $a_1 + a_2$) was computed by a computer to be 0.039, which is slightly less than the standard critical *p*-value of 0.05. So, in this example, we have good evidence that a relationship exists between the predictor variable x_i, and the response variable, y. That is, $\hat{b}_i$ is far enough away from zero for us to (tentatively) believe that a relationship exists.

Figure 1 enables us to see that the *p*-value is (by definition) the estimated fraction of the time (i.e., the probability) that we will obtain an estimate of b_i at least as extreme as the estimate we have obtained (i.e., $\hat{b}_i$) *if* the null hypothesis is or were true in the population and if we were to repeat the research project over and over, each time drawing a fresh sample of entities from the population (and if the underlying assumptions of the *p*-value are adequately satisfied).

The *p*-value has a key implication: If we use a critical *p*-value of 0.05 (or 0.01, etc.) then, in the long run, if we do everything properly, we will make a false-positive error in 5% (or 1%, etc.) of the statistical tests when the null hypothesis is true in the population. Research communities often use 0.05 or 0.01 as the critical *p*-value, judging that these error rates are acceptable.

It is noteworthy that the rate of publication of false-positive errors in a scientific research community will generally be somewhat or substantially higher than the "average" critical *p*-value used in the community. This counterintuitive fact is illustrated graphically in appendix B.10.

The logic of the *p*-value discussed in the preceding three paragraphs reflects the characterization of the *p*-value in sections 2 and 3 (Principle 1) in the ASA Statement on *p*-values (Wasserstein 2016). Many statisticians and experienced researchers agree that this logic is highly sensible. But, unfortunately, many non-statisticians agree that the logic is hard to understand. Therefore, the *p*-value is a pons asinorum (bridge of fools) in scientific research and statistics. Until a person understands the difficult logic, it is hard to transparently understand how we can use a sufficiently low *p*-value to enable us to *reject* the null hypothesis (in the absence of a reasonable alternative explanation).

Beginners sometimes incorrectly think that the *p*-value measures the probability that the null hypothesis is true or the probability that the effect observed in the data occurred through chance, as noted in Principle 2 in the ASA Statement (Wasserstein 2016). These interpretation errors aren't serious from a scientific viewpoint because they generally lead to the same conclusions as the correct interpretation of the *p*-value. But they are still errors.

Since the probability interpretation of the *p*-value is complicated, it is gratifying that it is possible and conceptually efficient to understand the operation of the *p*-value *without* understanding the probability logic. The *p*-value is simply a

measure of the weight of evidence that an effect is real. The lower the *p*-value below the critical value (and in the absence of a reasonable alternative explanation), the greater the weight of evidence that the effect is real. When interpreting *p*-values in scientific research, many researchers and statisticians use this sensible and easy-to-understand point of view.

The *p*-value is an *informative* measure of the weight of evidence because if we consistently use it with the same critical value, and if we do everything properly, then the critical value tells us the fraction of the time we will make false-positive errors in cases when the null hypothesis is true. That is useful to know and control because false-positive errors are costly because they send other researchers on a bound-to-fail wild-goose chase after a nonexistent effect.

It is easy to see that (because $\hat{s}_i$ is assumed to be constant) if the effect size, $\hat{b}_i$, increases in absolute value, then the tail areas in figure 1 will become smaller, and thus the *p*-value will become lower. Thus the *p*-value is in a monotonic decreasing relationship with the absolute effect size. We will use this important fact later below.

5.2. t-Statistic

The *t*-statistic for the effect shown in figure 1 is the distance of the estimated value of $\hat{b}_i$ from the null value in units of the standard error. Thus the *t*-statistic is $\hat{b}_i/\hat{s}_i$. If we measure (by eye or with a ruler) the distance of $\hat{b}_i$ from zero on the figure and if we also measure the length of $\hat{s}_i$, we see that the value of the *t*-statistic in this case is 2.16. That is, the estimate of parameter b_i is 2.16 standard errors away from zero.

The critical value for the *t*-statistic is often specified as 2.0. Thus, in the example, the *t*-statistic value of 2.16 falls beyond the critical value and thus we have reasonable evidence in this case that a relationship exists. (It is easy to show that a critical *t*-value of 2.0 is roughly equivalent to a critical *p*-value of 0.05 in rejecting or not rejecting the null hypothesis in relevant research situations.)

If the value of the effect size, $\hat{b}_i$, increases in positive value (or becomes more negative in negative value), then (because $\hat{s}_i$ is here assumed to be constant) the value of the *t*-statistic will also increase in absolute value. Therefore, the absolute value of the *t*-statistic is in a monotonic increasing relationship with the absolute effect size.

In terms of how it is constructed, the *t*-statistic is easier to understand than the *p*-value. This is because the *t*-statistic is a simple ratio of two numbers, which turns it into the distance of $\hat{b}_i$ from the null value in standard units. In contrast, the *p*-value is an area, the probability of an event under the null hypothesis. Areas and probabilities are two-dimensional and (due to the greater complexity) are harder to understand than one-dimensional (scalar) distances.

If (under reasonable assumptions) the form of the parameter sampling distribution under the null hypothesis is known, we can use the *p*-value as a measure of the weight of evidence in any situation in which we can use the *t*-statistic. (If the form of the parameter sampling distribution *isn't* known, then we can generally still *compute* the *t*-statistic, but we can't properly interpret it.) However, we can *also* sensibly use the *p*-value as a measure of the weight of evidence in situations when we *can't* use the *t*-statistic, such as in the frequent situations when the test statistic has an *F*-distribution or a chi-square distribution. Thus the *p*-value is more general than the *t*-statistic.

5.3. Confidence Interval

The range of the 95% confidence interval in figure 1 is shown by the two vertical gray lines in the figure. This range is the range that contains 95% of the area under the curve with, conventionally, equal-area tails at each end.

Confidence intervals operate slightly differently from the other measures in the sense that confidence intervals simply give us a Yes or No answer whether the value of the parameter estimate falls beyond the critical value. If the estimated value of the parameter is *outside* the range of the confidence interval, then we take this as good evidence that the relationship between variables under study exists in the population. Thus on figure 1 the parameter estimate, $\hat{b}_i$, is slightly outside the 95% confidence interval and thus we have good evidence that the effect is real.

If the value of the effect size, $\hat{b}_i$, increases in positive value (or decreases in negative value), then $\hat{b}_i$ will be farther outside or closer to being outside the confidence interval. Therefore, in the logically relevant sense, the confidence interval is in a monotonic relationship with the effect size.

Some researchers sensibly center the confidence interval on $\hat{b}_i$ and then check whether the interval overlaps the *null* value. This is effectively the same procedure as described in the preceding paragraphs, but done from a different perspective.

It is easy to show that in standard situations if we use a 95% confidence interval to distinguish between positive and negative results, then this is exactly equivalent to using a critical *p*-value of 0.05 to make the same distinction. Similarly, using a 99% confidence interval is exactly equivalent to using a critical *p*-value of 0.01, and so on.

Confidence intervals are harder to understand than *p*-values because the researcher must consider the scale of the associated parameter. The parameter and its scale (though highly relevant) are somewhat distant from the scientific goal of determining whether a relationship exists between variables. The *p*-value enables us to hide these details and to focus on the *value* of the *p*-value, which is always on the same scale. If the *p*-value is at or below the critical value (and if there is no reasonable alternative explanation), then we have good evidence that the effect is real, which is the scientific question of interest.

The *p*-value is more general than the confidence interval because confidence intervals can't be readily used in more complicated situations, such as with model equations that contain terms for statistical interactions among predictor variables, where the *p*-value operates efficiently.

5.4. Likelihood Ratio

In figure 1 the likelihood ratio is in effect the ratio of the heights of the likelihood function at the value $\hat{b}_i$ on the horizontal axis to the height at the null value, i.e., h_p/h_q. The value h_p is the height of the likelihood function at the estimated value $\hat{b}_i$ if the null hypothesis is true. In contrast, if the specific hypothesis that $b_i = \hat{b}_i$ is true, then it would be correct to superimpose the peak of the distribution function at $\hat{b}_i$ on the horizontal axis. Then the height of the likelihood function at $\hat{b}_i$ on the axis would be h_q.

(Technically, if the research hypothesis is true and the correct value of b_i in the population is $\hat{b}_i$, and the relevant assumptions are satisfied, then the sampling distribution of $\hat{b}_i$ will no longer be a central *t*-distribution, but will be *non*central *t*-distribution. In this case the height of the distribution at the value $\hat{b}_i$ will generally be a slightly different height from h_q, which adds another layer of complexity, which we acknowledge for completeness and then mostly ignore in the present high-level discussion.)

If we measure h_p and h_q on the graph and then compute the ratio of the two heights, we see that the likelihood ratio for the data behind the graph is 0.106 (but note the preceding paragraph).

If the value of the effect size, $\hat{b}_i$, increases in positive value (or becomes more negative in negative value), then the value of the likelihood ratio will decrease because h_p will decrease while h_q remains constant (but note the second preceding paragraph). Therefore, the likelihood ratio is in a monotonic decreasing relationship with the effect size.

If the null hypothesis is true in a given research situation, then the likelihood ratio will generally be close to 1.0. In contrast, if the research hypothesis is true, then the likelihood ratio will generally be lower. Therefore, in theory, we can specify a critical value for the likelihood ratio. And we can decide that we have good evidence that a relationship exists between the relevant variables if the value of the likelihood ratio is less than the specified critical value.

However, critical values for likelihood ratios are rarely used. Instead, we compute the *fraction of the time* that the value of the likelihood ratio will be as low as it is or lower if the null hypothesis is or were true (and if other relevant assumptions are adequately satisfied). But computing this fraction amounts to computing a *p*-value, and thus we can use standard critical *p*-values as the critical values. So using a likelihood ratio can be viewed as merely another sensible path to computing an appropriate *p*-value.

If we study scientific practice, we find that researchers rarely use likelihood ratios either to compute *p*-values or for other approaches for testing for the existence of a relationship between variables. This may be partly because the likelihood-ratio approach often gives essentially the same *p*-values as conventional approaches (Wackerly, Mendenhall, and Scheaffer, 2008, p. 553), but the likelihood ratio concepts are arguably somewhat harder to understand than the *p*-value concepts.

Likelihood-ratios are harder to understand than *p*-values because the ratio of two heights of the likelihood functions of a parameter under the two hypotheses is harder to understand than the probability (fraction of the time) that the parameter estimate will be as discrepant or more discrepant from the null value if the null hypothesis is or were true. This difficulty of understanding the likelihood ratio arises from the difficulty people who aren't statisticians have understanding the concept of the likelihood function for a parameter under a hypothesis.

Of course, the theory of the probability behind the *p*-value is based directly on the likelihood function under the null hypothesis. But we can hide these ideas and work with less experienced people using the idea of the *p*-value as a sensible probability, without referring to the likelihood function itself. This helps to reduce the perceived complexity. But we can't readily hide the likelihood function if we are discussing the ratio of two heights of the function.

The likelihood ratio approach is also harder to understand because it uses *two* distributions—the theoretical probability density of $\hat{b}_i$ under the null hypothesis and the theoretical density of $\hat{b}_i$ under the hypothesis that the population value of b_i is equal to the value estimated from the sample. In contrast, the *p*-value uses only a *single* distribution—the theoretical density of $\hat{b}_i$ under the null hypothesis.

(It is also possible to view the likelihood ratio as being based on the ratio of the heights at two points on the *single* distribution shown in figure 1, though that requires a different analysis.)

The likelihood ratio approach may also be used less often because the mathematical distribution of the likelihood ratio under the null hypothesis is somewhat difficult to compute, and formulas for the distribution are only available in the "asymptotic" sense, which implies that the formulas (and hence the *p*-values derived from the formulas) are only fully correct if the sample size is infinite, which doesn't happen. Fortunately, these asymptotic approaches give "fairly good" answers for typical sample sizes. However, this leads researchers to ask whether "fairly good" is good enough for the situation at hand, and there is presently no easy answer to that question.

Some statisticians (e.g., Cox, 2006, p. 91) use the reciprocal of the likelihood ratio discussed above because the reciprocal is also reasonable and has advantages.

5.5. Bayes Factor

For the Bayes factor, the distribution shown in figure 1 should be viewed as the estimated marginal *posterior* distribution of b_i under the null hypothesis, as derived from Bayesian principles. This distribution may be a *t*-distribution, but it may also be some other type of distribution. But regardless of the type of distribution, in standard situations with a continuous response variable the distribution will typically be (at least roughly) bell-shaped and symmetrical about the null value if the null hypothesis is or were true.

The Bayes factor is similar to the likelihood ratio, but is more complicated. This is because the placement of the center of the distribution under the research hypothesis generally *isn't* on the estimated value of b_i. These ideas are illustrated in figure 2.

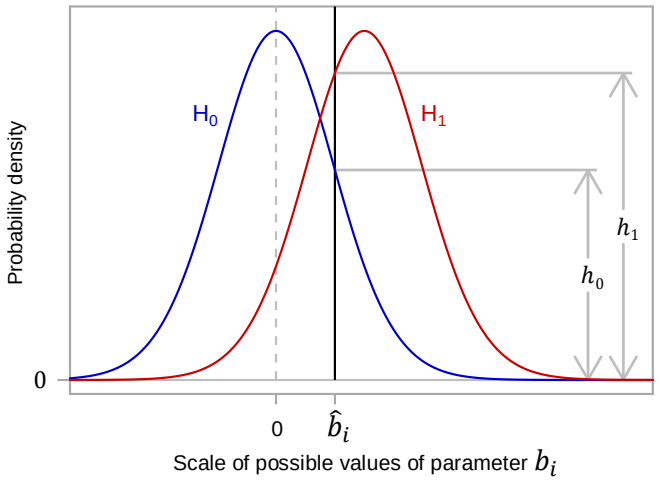

Figure 2. A redrawn version of figure 2 from an article by Bayarri, Benjamin, Berger, and Sellke (2016, hereafter BBBS). This graph shows their interpretation of the Bayes factor. The BBBS notation has been changed to reflect the notation of the present paper. The H_0 stands for the null hypothesis and the H_1 stands for the research hypothesis. As noted in the caption of the BBBS figure, the value of the Bayes factor is h_1/h_0. (The BBBS figure is copied with CC BY 4.0 permission.)

The figure illustrates the complexity of the Bayesian approach because BBBS have drawn the figure with the maximum value of the H_1 (red) distribution offset from the estimated value $\hat{b}_i$ of the parameter. This is because the H_1 distribution is a *specific* distribution that is specified by the researcher, as noted by BBBS in their discussion about the "point alternative" hypothesis they are using (p. 93). This distribution (reflecting a very specific research hypothesis) is at the researcher's complete discretion. The fact that the location and width of the H_1 distribution are at the researcher's discretion adds a degree of arbitrariness to the procedure.

The Bayes factor is also somewhat arbitrary in the sense that it depends on the choice of the prior distribution, and the choice of the prior distribution is itself often arbitrary, even when the choice is "vague".

BBBS discuss (sec. 3.2) a Bayes factor bound, which simplifies things and removes some arbitrariness by specifying (in effect) that the H_1 (red) distribution in figure 2 should be centered on the vertical line at $\hat{b}_i$ on the horizontal axis, as it is in the case of the likelihood ratio. Then the BBBS Bayes factor is equivalent to h_q/h_p on figure 1 (but based on the estimated Bayesian posterior parameter distributions of $\hat{b}_i$ under the two hypotheses, and not on the estimated maximum-likelihood parameter distributions).

BBBS recommend that the Bayes factor should be greater than a critical value of 16 before we can reject the null hypothesis (2016, p. 96). In standard research this approach is substantially stricter than the standard critical *p*-value of 0.05. Thus if we use a critical value of 16 for the Bayes factor then, in the long run, we will make substantially fewer false-*positive* errors with the Bayes factor, but we will also make substantially more false-*negative* errors.

Figure 2 suggests (somewhat obscurely) the relationship between the effect size and the Bayes factor. That is, if we fix the two curves on the graph, and if we then let $\hat{b}_i$ increase or decrease, then we will see that h_1/h_0 (i.e., the Bayes factor) will (at least in certain typical cases) consistently increase or decrease in step. Thus there is a monotonic increasing relationship between the absolute effect size and the Bayes factor.

As with the likelihood ratio, some statisticians sensibly use the reciprocal version of the Bayes factor.

Bayes factors are substantially harder to understand than *p*-values because the Bayes factor uses the additional concepts of the prior distributions of the parameters of the model equation and uses the concept of the point alternative hypothesis illustrated by the H_1 curve in figure 2.

The Bayes factor and the *p*-value are arguably equally general because they can apparently each be computed in any situation in which the other can be computed. In particular, *p*-values can always be computed whenever computer-based resampling tests are possible. And we can simulate and resample any state of affairs that we can model. The computation of the Bayes factor (which pertains to a parameter of a model equation) requires that a model equation be implicitly or explicitly stated. Thus the presence of models behind Bayes factors implies that simulation-based *p*-values are possible in all situations when Bayes factors can be computed. Therefore, arguably, the *p*-value is at least as general as the Bayes factor in scientific applicability. Similarly, a Bayes factor can be computed whenever a *p*-value can be computed because the likelihood function required to compute the Bayes factor will be available.

Figure 2 implies that the Bayes factor depends directly on the widths (e.g., the standard deviations) of the Bayesian posterior distributions of the parameter under the research and null hypotheses. Therefore, it is important to ask whether the widths of these two distributions are scientifically meaningful. The answer to this question is unclear. What is the *scientific* meaning of the width of a Bayesian posterior distribution? (The width *isn't* an estimate of the width of the parameter sampling distribution.) Or does the width have no scientific meaning beyond being the "average" of the widths of the prior distribution and the likelihood function, serving (quite sensibly) only to estimate the value of the associated parameter (through a measure of the central tendency of the posterior distribution under the research hypothesis)?

It is noteworthy that we can compute so-called "posterior intervals" from a posterior distribution, but then the same question arises for them. Since these intervals don't refer to sampling distributions, what do they mean scientifically? Do they have any scientific meaning?

5.6. Posterior Probability that the Null Hypothesis is True

Bayesian statisticians discuss how to compute the "probability" that the null hypothesis is true (Berger and Sellke 1987; Sellke, Bayarri, and Berger 2001; Wagenmakers 2007, pp. 792–794; Held and Ott 2016). This idea is intriguing because this probability is somewhat intuitive. Bayesians refer to this

probability as the "posterior" probability that the null hypothesis is true because it is computed by mathematically combining a "prior" probability that the null hypothesis is true with the data obtained in the research project.

It is easy to show how the posterior probability that the null hypothesis is true can be derived from the Bayes factor (Berger and Sellke 1987, p. 115, left column, third line from bottom). And if we differentiate the Berger and Sellke formula for the relationship between the Bayes factor and the posterior probability that the null hypothesis is true, we find that the derivative is always positive. (Berger and Sellke use the reciprocal Bayes factor.) Thus, as we might expect, the posterior probability that the null hypothesis is true is in a monotonic relationship with the Bayes factor. Thus (due to the transitivity of monotonicity) the posterior probability that the null hypothesis is true is in a monotonic decreasing relationship with the effect size.

Unfortunately, the posterior probability that the null hypothesis is true leads to an apparent contradiction, which is discussed in appendix A. This paper takes the view that the posterior probability that the null hypothesis is true can't be sensibly used in scientific research unless the contradiction is resolved.

5.7. Second-Generation p-Value

Blume, D'Agostino McGowan, Dupont, and Greevy (2018) propose a "second-generation" *p*-value and they note in their abstract that this *p*-value helps to control false-positive (i.e., Type 1) errors. Therefore, it is sensible to view the second-generation *p*-value as a measure of the weight of evidence that an effect observed in scientific research data is real in the underlying population.

The second-generation *p*-value is an appropriate measure of the weight of evidence that an effect is real for researchers who envision an "indifference zone" around the null value of the parameter under consideration. If the estimated value of the parameter is significantly different from the null value according to the traditional *p*-value, but if the confidence interval for the estimated value (when *centered* on the estimated value) overlaps the indifference zone, then Blume et al. say (indirectly) that they are less interested in the associated effect.

It is easy to envision the confidence interval (or perhaps some other reasonable interval) and the indifference zone drawn on figure 1. The confidence interval is centered on the parameter estimate, $\hat{b}_i$, and the indifference zone is centered on the null value, zero. The second-generation *p*-value is a measure of the extent to which these two intervals overlap each other.

If the confidence interval and the indifference zone overlap, then the second-generation *p*-value is essentially the proportion of overlap. (The proportion is measured in terms of the ratio of (a) the width of the overlap to (b) the total width—both halves—of the confidence interval, not to the total width of the indifference zone.) The less the two intervals overlap, the lower the proportion of overlap until there is no overlap and the proportion becomes zero.

A problem with the second-generation *p*-value is that many researchers don't envision an indifference zone around the null value. Instead, researchers who understand the concept generally view such a zone as an unnecessarily-limiting additional concept. That is, we are usually interested in *any* real difference of a parameter from the null value, regardless of how small the difference is. This is because once we know (through a replicated low *p*-value or through some other reasonable measure of the weight of the evidence) that a *real* effect probably exists in the population, and if we think the effect might be important, then we can take further steps to *strengthen* the effect, perhaps by taking account of other relevant variables.

Of course, subject-matter expertise is necessary here to decide whether a weak but (likely) real effect is potentially important and thus deserves further study. But researchers can only apply their subject-matter expertise if they know about the existence of the effect. Thus researchers almost always wish to know about the existence of weak but *real* effects. So they almost never have an indifference zone. Researchers never wish to hide facts.

Blume et al. indirectly suggest that a sensible critical value for the second-generation *p*-value is zero—the borderline between when the two regions do and don't overlap.

The second-generation *p*-value is monotonically related to the effect size because (with other factors held constant) the farther the parameter estimate is from the null value, the lower the proportion of overlap of the two intervals, and therefore the lower the second-generation *p*-value (until it reaches zero).

The need to specify the width of the indifference zone makes the second-generation *p*-value somewhat arbitrary because different researchers may choose to use conceptually different widths, which implies that second-generation *p*-values from different research projects generally aren't fully comparable. However, proponents of the second-generation *p*-value can eliminate this arbitrariness if they can convince researchers to use conceptually equivalent widths of the indifference zone in all research.

As explained above, the interpretation of the proportion of overlap underlying the second-generation *p*-value is relatively simple. However, determining the proportion is only the first step. And to compute the second-generation *p*-value, the proportion must be multiplied by a "small-sample correction factor", as discussed in sections 2.1 and 2.5 in the Blume et al. article. This ingenious but hard-to-understand factor is designed to give the second-generation *p*-value appropriate properties relative to the sample size. This makes the second-generation *p*-value more complicated for beginners, arguably making it harder to understand than the standard *p*-value.

The second-generation *p*-value is less general than the standard *p*-value because the second-generation version (presently) doesn't operate in cases with the *F*-distribution and chi-square distribution, where the standard *p*-value operates efficiently.

The second-generation *p*-value is conceptually equivalent to a standard confidence interval (placed on the estimated parameter value), but with an extra control knob—a knob to control the width of the indifference zone. Researchers will

be inclined to set the knob at zero because that reduces the false-*negative* error rate, which researchers are always eager to reduce. This setting (in effect) takes us back to a standard confidence interval. Of course, in the long run, using a standard confidence will also lead to slightly more false-*positive* errors. But these errors are inevitable (at a controllable rate) regardless of which approach we use. We weed out false-positive errors in scientific research through appropriate replication.

It is noteworthy that the second-generation *p*-value conflates three aspects of an effect—the *existence* of the effect, the *strength* of the effect, and the *importance* of the effect. The indifference zone defines the zone in which overlap with the confidence interval implies that the effect (if it exists) isn't strong enough to be important. As noted in section 3, it is scientifically sensible to keep these concepts separate. And it is sensible to first establish the existence of an effect before we consider its strength or its importance.

5.8. D-Value

Demidenko (2016) proposes the *D*-value as a measure of the weight of evidence that an effect observed in scientific research data is real in the underlying population. Mathematically, in the simplest case, the *D*-value is a transformed version of the associated *p*-value. That is, as Demidenko illustrates in his formulas (2) and (5), the *D*-value uses the same computing formula as the associated (one-sided) *p*-value, except that the *D*-value formula replaces sample size, n, that is used in the *p*-value formula with the numeral 1. In other words, Demidenko has removed the sample size from the formula because he recognized that this removal yields a sensible measure.

Demidenko proposes in both the abstract and the conclusion of his article that a potential role or purpose of the *D*-value in scientific research is "to weigh up the likelihood of events under different scenarios". He also suggests in the last paragraph of the article that we should replace the *p*-value in scientific research with the *D*-value. These points suggest that Demidenko is proposing that we use the *D*-value for the same purpose as we use the *p*-value—as a measure of the weight of evidence that an effect (relationship between variables) observed in scientific research is a real effect in members of the population of entities under study.

Demidenko notes that in his standard two-group medical example the *D*-value is the proportion of patients in the sample who got worse after the treatment. This proportion is much easier to understand than the corresponding *p*-value for the hypothesis that the drug has a real effect on patients in the population. This ease of understanding of the *D*-value is a good reason the *D*-value might be an effective replacement for the complicated *p*-value as a measure of the weight of evidence that an effect is real.

However, from a theoretical point of view, it doesn't make sense to use the *D*-value as a measure of the weight of evidence that an effect is real because the *D*-value doesn't take account of the sample size. And, as suggested by Demidenko, it seems more sensible to view the *D*-value as a measure of the *strength* or *effect size* of a relationship between variables. Demidenko refers to "effect size on the probability scale" (2016, sec. 3.1) and "the effect size expressed in terms of the probability of group separation" (2016, sec. 6).

Measures of the strength or size of an effect are *not* good measures of the weight of evidence that an effect is real because the value of a proper measure of strength must be independent of the size of the sample that is used to estimate the value. Measures of strength must be independent of the sample size because the strength is a property of the underlying effect in the population, not a property of the sample. (The sample size *is* relevant for estimating the *precision* of an estimate of strength, but not in computing the estimate of strength itself.)

In contrast, the sample size is directly relevant in determining the weight of evidence (provided by a research result) that an effect is real. That is, for a given observed effect size, a larger sample gives us a greater weight of evidence that the effect is real than a smaller sample. This is due to the idea that (assuming proper sampling) the larger the sample, the more representative the sample test statistic (e.g., Student's *t*-statistic) is of the correct value of the statistic in the entire population (due to the law of large numbers). And the more representative a test statistic is of the correct value, the more confidence we can have in conclusions drawn from the value of the statistic.

In the linear regression example, the *D*-value is in a monotonic decreasing relationship with the effect size, as implied by the definition of the *D*-value in section 5 of Demidenko's article (2016).

Even though the *D*-value doesn't take account of the sample size, we could still define critical values for the *D*-value to enable us to use it as a sensible measure of the weight of evidence that an effect is real. However, for the *D*-value to work like the other measures, the appropriate critical values would need to be a function of the sample size. This would make using the *D*-value as a measure of the weight of evidence that an effect is real substantially more complicated than using a measure that can sensibly use a single fixed critical value.

Therefore, since the *D*-value doesn't take account of the sample size, it isn't an efficient measure of the weight of evidence that an effect is real. Therefore, it isn't sensible to evaluate the *D*-value as a measure of the weight of evidence that an effect is real in a population. And, contrary to Demidenko's recommendation, it isn't sensible to consider replacing the *p*-value with the *D*-value because the two measures perform different functions. The *p*-value is a measure of the weight of evidence that an effect is *real*, but the *D*-value is sensibly viewed as a measure of the effect *size* (under the assumption that the effect is real).

5.9. Information Criteria

Researchers sometimes use an information-criterion method to (in effect) provide a measure of the weight of evidence that an effect observed in scientific research data is real in the underlying population, as discussed by Konishi and Kitagawa (2008). We can use a stepwise regression procedure (e.g., forward stepwise regression or the stepwise lasso [Efron, Hastie,

Johnstone, and Tibshirani, 2004]) to select the terms for candidate regression equations and we can use an information criterion (e.g., the Schwarz Bayesian Criterion) to help us to decide which of the equations is best (in the sense of yielding the lowest value of the information criterion). Or we can use the information criterion itself to select terms, on each step selecting the term that causes the value of the information criterion to be reduced the most. These methods yield model equations for relationships between variables that are similar to or identical to the model equations yielded by the other methods.

The information-criterion methods work somewhat differently from the other methods for determining whether we have enough evidence that an effect is real. An information-criterion method typically works *automatically* in a computer program to derive a model equation for a relationship between variables so that the equation contains all the predictor variables that the information-criterion method has decided are related to the response variable.

The information-criterion methods don't use explicit critical values because they have (in effect) *implicit* critical values that are built in. These implicit methods decide which terms to include in the model equation based on a sensible mathematical formula for computing the value of the information criterion for an equation, which amounts to a mathematical score. SAS gives the formulas of some of the main information criteria (2018). The scores in the information-criterion methods reflect how well the equation fits the data at hand—the better the fit, generally the lower (better) the score. The scores also take account of the complexity of the model equation—the less complex the equation, generally the lower the score.

The information-criterion methods operate by simply selecting the model equation with the lowest score from the set of possible equations. This is complicated slightly by the fact that the different information criteria (e.g., the Schwarz Bayesian Information Criterion, the Akaike Information Criterion, and Mallows C_p) sometimes disagree with each other about which equation is best. Then the researcher wonders which information criterion to use. However, this is more a theoretical problem than a real problem because the measures often agree with each other. The Schwarz Bayesian criterion is popular, perhaps because it is conservative in the sense that it tends to require slightly stronger evidence than the other criteria that a term belongs in the equation before the term will be selected for inclusion.

Arguably, the information-criterion methods are *harder* to understand than the *p*-value because the multiple formulas for the criteria are somewhat daunting. On the other hand, at a higher level the information-criterion methods can be viewed as being *easier* to understand than the *p*-value because they operate as an automatic black box (implemented by programming the formulas into a computer program) for determining sensible model equations.

When used appropriately, the information-criterion methods are in monotonic relationships with the effect size. That is, if we were somehow able to change the size of an effect in the population (with other factors, including the total sum of squares, held constant), then the formulas imply that the expected value of the relevant information criterion will change accordingly.

Researchers use the information-criterion methods for detecting relationships less frequently than other methods, perhaps because the information-criterion methods have only been developed for certain standard situations with continuous response variables. Also, unlike the other methods, the information-criterion methods don't allow easy control of false-positive and false-negative error rates.

5.10. The Monotonic Relationships Among the Nine Measures

Each of the preceding nine subsections concludes that its measure of the weight of evidence is monotonically related to the absolute effect size. This implies that the measures are all monotonically related to each other (due to the transitivity of monotonicity). Thus if one of the measures of the weight of evidence is somehow made to go up or down in value in a particular research project (with other relevant factors, typically the total sum of squares, held constant), then all the other measures will go up or down in value (or down or up) in step. This is an instance of mathematical clockwork.

The monotonic relationships between the values of the measures imply that in almost any given research situation, all the applicable measures can be *calibrated* with each other to have equivalent critical values. This calibration (through the choice of critical values) will cause the various measures to exhibit identical behavior in indicating whether we have enough evidence (in the absence of a reasonable alternative explanation) to tentatively reject the relevant null hypothesis.

The measures can't be calibrated with each other in *all* situations because there are exceptions. But, to the extent that the exceptions are known, they are unimportant. For example, the second-generation *p*-value isn't *strictly* monotonic with the other measures because if the effect is large enough, this *p*-value goes to exactly zero. Similarly, we can't calibrate the information criteria to behave equivalently with the behavior of another method because the information criteria don't have adjustable critical values. But we can generally calibrate the other measures to behave equivalently with the information criteria.

In discussing the calibration of the measures, there is no suggestion that we *should* calibrate the measures with each other. (That would be complicated.) Instead, the idea of calibrating the measures with each other is a conceptual exercise to illustrate the parallel operation of the measures.

The fact that we can generally calibrate the various measures to behave equivalently implies that the nine measures are (when so calibrated) equally powerful for detecting relationships between variables. Thus the nine measures are (when usable) generally *functionally equivalent* to one another in high-level output if equivalent critical values are used. Operationally, the nine measures have somewhat different mathematical bases and they have different (but highly correlated) scales. But they are functionally equivalent.

Bayesian methods can take account of an *informative* prior distribution to increase their statistical power. However,

frequentist methods can also efficiently take account of such a distribution through the methods of meta-analysis (Hedges and Olkin 1985; Borenstein, Hedges, Higgins, and Rothstein 2009).

The monotonic relationships between the measures of the weight of evidence generally aren't *linear* relationships. But the relationships are almost all smooth, as suggested by figures 1 and 2.

The form of the monotonic relationships between the measures sometimes depends on the research situation. For example, in one research situation there will be one monotonic relationship between a given *p*-value and the associated Bayes factor as the effect size changes. But in another research situation there will be *another* monotonic relationship between the *p*-value and the Bayes factor as the effect size changes. The differing relationships between the measures are due to other factors that sometimes play a role in the relationships, such as the prior distribution and the sample size.

This paper hasn't proven that all the measures of the weight of evidence are always in monotonic relationships with the effect size. This point is relevant because it is conceivable that there is a non-monotonic relationship in certain cases with the likelihood ratio or the Bayes factor. However, it seems likely that the relationships are monotonic in all *practical* scientific research situations, which is what is important for the present discussion.

The fact that the nine measures are functionally equivalent implies that they are equally good at detecting relationships between variables. Therefore, it is efficient to decide which of the measures is most reasonable based on one or more secondary attributes, and then to work solely with that measure. This allows us to minimize the number of concepts that we must juggle. Section 6 compares the nine measures to help us to decide which is most reasonable.

5.11. Generalization

For simplicity, the preceding discussion about the nine measures is in terms of a linear regression model equation. This presumes that the value of the response variable in the studied relationship between variables is a continuous variable and it presumes that the values of the response variable can be sensibly modeled with a linear combination of the values (or functions of the values) of the predictor variables. However, the discussion readily generalizes to most (all?) other types of model equations of relationships between variables, such as other types of relationships with continuous response variables (e.g., nonlinear model equations, cell-means model equations, or generalized linear or nonlinear model equations). The discussion also readily generalizes to model equations with discrete response variables (e.g., logistic model equations, loglinear model equations).

The generalization of the discussion is possible because all model equations have parameters (which are occasionally hidden). And most parameters have null values. And if the correct value of a parameter in nature is equal to the null value, this implies that the associated effect is absent.

For many parameters of model equations, we have sensible ways of estimating the correct population values of the parameters from appropriate data and sensible ways of determining the likely sampling distribution of the values under the null hypothesis, which will be similar to the distribution shown above in figure 1. And for many parameters we are interested in determining whether the population value of the parameter is different from the null value because if we can find good evidence of a real difference, this implies that we have good evidence that the associated relationship between variables (or other effect) exists—good evidence that the relationship (effect) is real in the population.

The *p*-value, likelihood ratio, and Bayes factor have been generalized for various types of model equations. Some of the other measures have been generalized in some cases.

6. Comparing the Measures of the Weight of Evidence that an Effect Is Real

Let us compare the *p*-value with the eight other measures. Table 1 summarizes some comparisons using criteria discussed in the preceding section and using one other important criterion.

Table 1. Comparisons of the *p*-value with the eight other measures of the weight of evidence that an effect observed in scientific research is real.

Measure of weight of evidence ↓	The *p*-value is				
	more informative	easier to understand	more general	less arbitrary	more powerful
t-statistic	✓	X	✓	=	=
confidence interval	✓	✓	✓	=	=
likelihood ratio	✓	✓	=	=	=
Bayes factor	✓	✓	=	✓	=
posterior probability null hypothesis is true	✓	✓	=	✓	=
second-generation *p*-value	✓	✓	✓	✓	=
D-value	✓	X	=	=	=
information-criterion methods	✓	?	✓	=	=

Table 1 presents the comparisons from the winner's (i.e., the *p*-value's) perspective because that is pedagogically efficient. If the reader thinks that a different measure of the weight of evidence is best, then it is a good exercise to regenerate the table with the preferred measure swapped with the *p*-value in their locations in the table.

A check mark in a cell in the body of the table indicates that the *p*-value is (arguably) superior to the measure associated with the row in terms of the attribute associated with the column. For example, the check mark in the "more informative" column for the *t*-statistic implies that the *p*-value is more informative than the *t*-statistic. In contrast, an X in a cell indicates that the *p*-value is inferior to the measure associated with the row in terms of the attribute. For example, the X in the "easier to understand" column for the *t*-statistic indicates that the *p*-value is harder to understand than the *t*-statistic. An equals sign in a cell indicates that the *p*-value and the measure associated with the row are (exactly or roughly) equivalent on the attribute associated with the column.

The "more informative" column of the table indicates that the *p*-value (arguably) has a more informative *scale* than each of the other measures. The *p*-value is more informative because the critical *p*-value (when consistently properly used) gives us a direct estimate of the rate of occurrence of false-positive errors in research in cases when the null hypothesis is true. The rate of occurrence of false-positive errors is important in scientific research because these errors are guaranteed to occur some of the time, are socially costly (because they send researchers attempting to replicate the effect on a wild-goose chase), and their rate of occurrence is partially controllable (through the choice of the critical value).

Some readers may disagree that the *p*-value is more informative than the other measures and they may believe that another measure of the weight of evidence is more informative than the *p*-value. The judgement here depends mainly on how much weight a person puts on the importance of controlling the rate of occurrence of false-positive errors in scientific research. In choosing among measures of the weight of evidence that an effect is real, what (if anything) is socially more important in scientific research than controlling the rate of occurrence of socially costly false-positive errors?

Each checkmark and X in the "easier-to-understand", "more general", and "less arbitrary" columns in the body of the table is supported by discussion above in sections 5.1–5.9. The "more powerful" column of the table only contains equals signs, which implies that the *p*-value and the eight other measures are generally equally powerful for detecting effects. This is justified by the idea that the nine measures can generally be calibrated to have equivalent critical values, as discussed above in section 5.10.

No attempt is made to justify the cells with equals signs in the "more general" and "less arbitrary" columns of the table. Instead, these *equalities* are (like null hypotheses) merely assumed to be correct. Thus each equality (and each inequality) in the table is completely open to focused refutation. And, for any cell in the body of the table, it is a useful exercise to try to refute the cell's claim.

We might sensibly compare the nine measures of the weight of evidence on criteria that are different from the five criteria in table 1. However, it is hard to devise other sensible criteria. For example, we might compare the measures on the criterion of mathematical beauty, though that seems less important from a practical perspective.

In view of the central role of empiricism in scientific research, it is highly sensible to ask whether we could perform an empirical or quantitative comparison of the nine measures. For example, we might compare the nine measures in terms of the frequency with which they each make false-positive and false-negative errors. However, we can't do that because, as discussed above in section 5.10, the nine measures can be (with minor exceptions) calibrated to perform equivalently, so they are effectively *equivalent* in performance (in terms of indicating a positive or negative result and thus in terms of statistical power).

It would be very sensible to empirically compare the measures in cognitive terms, particularly in terms of ease of student understanding. But such empirical educational comparisons are difficult (perhaps impossible) due to the difficulty in removing the confounding effects of the teaching approach that we use to teach each method.

Furthermore, even if we could somehow remove the confounding, the paucity of practical real effects discovered in education research despite numerous careful attempts implies that negative results are sadly the norm in education research. Any small real effects are presumably swamped by the noise due to the high student-to-student variation.

[More powerful comparative approaches in education research have recently become available using pre-recorded online courses that (a) have tightly controllable and easily variable content, (b) enable easy random assignment of treatments, (c) can have consistent student evaluation measures, and (d) can have large enrollments, yielding higher statistical power. Perhaps education researchers can compare various approaches to teaching each of the nine measures to find the approach and the measure that is best in terms of (a) measures of students' attitudes toward the various measures and (b) attitudes toward the use of the measures in scientific research.]

We could easily empirically compare the measures in terms of *popularity* by surveying the scientific research literature to determine the frequency of usage of each measure. In this case experience suggests that the *p*-value would win handily. However, we would like an empirical comparison that goes beyond mere popularity.

So far, the author has been unable to think of sensible useful empirical comparisons between the measures that could *realistically* be done. Can the reader think of realistic ways that we could empirically compare the nine measures?

In view of the apparent lack of *empirical* criteria on which we can compare the nine measures, the *logical* arguments that are collected in this paper must apparently suffice. (Appendix H discusses some further theoretical arguments about the preferred measure.) The paper holds that the logical arguments are enough to conclude that (a) we need a formal efficient way to detect effects in scientific research, and (b) the *p*-value is the best way.

It is noteworthy that the "logical" arguments given in sections 5 and 6 and summarized in table 1 are somewhat subjective because the arguments are based on certain (basic) judgments. Thus, if a researcher believes that the logical arguments favoring the *p*-value aren't convincing, and if he or she believes that another measure of the weight of evidence that an effect of is real is better than the *p*-value, then they should use the other measure. A researcher is completely free to use whichever measure they find is most practical for them (while bearing in mind that if they wish to publish their research, some measures will likely be more acceptable to journals in their field than others).

And, of course, if a researcher believes that none of the measures of the weight of evidence is useful, then the researcher should omit using the measures. But if we omit using a measure of the weight of evidence, then how can we reliably and efficiently distinguish real effects from noise in data?

7. Conclusions

Many scientific research projects study relationships between variables in populations of entities. Such study, if successful, gives us the ability to accurately predict or control the values of the response variable in new entities from the population. This ability is often socially, theoretically, or commercially useful.

In studying an effect (e.g., a relationship between variables), we need an efficient measure of the weight of evidence that the effect observed in the research data for a sample is a real (i.e., reproducible) effect in the population behind the sample. We need such a measure to avoid deceiving ourselves and others about an effect that may be either (a) nonexistent or (b) so weak that we can't (presently) reliably observe it, and therefore it is effectively non-existent. This helps us to avoid wasting resources on effects observed in scientific research that aren't real.

This paper has discussed nine measures of the weight of evidence that an effect is real. With minor exceptions, the measures can all be calibrated with each other to make the same pronouncements about whether there is enough evidence in the data to reject the relevant null hypothesis and thereby (tentatively) conclude that the studied effect is real. Therefore, the nine measures are all functionally equivalent. And the only difference between them is operational—they use different mathematical approaches and different scales.

All the measures of the weight of evidence that an effect is real sometimes make false-positive and false-negative errors. Unfortunately, we can't completely eliminate these errors due to our always-limited resources. However, we can use statistical methods to help us to control the rates of occurrence of the errors.

A low-enough *p*-value (or another sensible indicator of enough weight of evidence) is good evidence that an effect is real *only if* there is no reasonable alternative explanation for this evidence.

As summarized in table 1, the *p*-value has certain advantages over the other measures pertaining to information content, ease of understanding, generality, and lack of arbitrariness. These advantages arguably outweigh the disadvantages of the *p*-value. No other reasonable attributes for comparing the measures appear to be available. Therefore, the *p*-value is the best available measure of the weight of evidence that an effect (usually a relationship between variables) observed in scientific research is real in the entities in the studied population.

Supplementary Material

The supplementary material expands some of the preceding ideas. Appendix A describes an apparent contradiction that arises from the concept of the posterior probability that a null hypothesis is true. Appendix B gives more details about using *p*-values to detect relationships. Appendix C discusses some well-known criticisms of the *p*-value. Appendix D compares hypothesis testing with Karl Popper's idea of falsification. Appendix E discusses the optimal critical value for a test statistic. Appendix F proposes a way to teach *p*-value ideas to

beginners. Appendices G through M discuss miscellaneous topics about detecting and studying relationships between variables. Appendix N discusses some exceptions to the idea that scientific research projects study relationships between variables. Appendix O discusses whether the ideas in this paper are "real".

Acknowledgements

The author thanks Ross Macnaughton and Milo Schield for helpful suggestions that led to substantial improvements to the paper. The author also thanks the American Statistical Association and Executive Director Ron Wasserstein for the October 2017 Symposium on Statistical Inference, which greatly stimulated the development of the paper.

References

Baker, A. (2016), "Simplicity," *The Stanford Encyclopedia of Philosophy* (Winter 2016, ed. E. N. Zalta). Available at https://plato.stanford.edu/archives/win2016/entries/simplicity/

Bayarri, M. J., Benjamin, D. J., Berger, J. O., and 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. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112–139.

Blume, J. D., D'Agostino McGowan, L., Dupont, W. D., and Greevy, R. A. (2018), "Second-generation *p*-values: improved rigor, reproducibility, & transparency in statistical analyses," *PLoS ONE* 13(3): e0188299.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009), *Introduction to Meta-Analysis*, Chichester, UK: Wiley.

Chatterjee, S., and Hadi, A. S. (2012), *Regression Analysis by Example*, Hoboken, NJ: Wiley.

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

Demidenko, E. (2016), "The *p*-Value You Can't Buy," *The American Statistician*, 70, 33–38.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression" (with discussion), *Annals of Statistics*, 32, 407–499.

Hedges, L. V., and Olkin, I. (1985), *Statistical Methods for Meta-Analysis*, San Diego, CA: Academic Press.

Held, L., and Ott, M. (2016), "How the Maximal Evidence of *P*-Values Against Point Null Hypotheses Depends on Sample Size," *The American Statistician*, 70, 335–341.

Konishi, S., and Kitagawa, G. (2008), *Information Criteria and Statistical Modelling*, New York: Springer.

Macnaughton, D. B. (2016), "Comment on 'A Low-Uncertainty Measurement of the Boltzmann Constant'," *Metrologia*, 53, 108–115.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2018), "Abandon Statistical Significance (version 3)". https://arxiv.org/abs/1709.07588v3

SAS (2018), "GLMSELECT Procedure > Details > Criteria Used in Model Selection Methods," in *SAS/STAT 14.3 User's Guide*. Available at https://support.sas.com/documentation/onlinedoc/stat/index.html

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of *p* Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71.

Wackerly, D. D., Mendenhall, W. III, and Scheaffer, R. L. (2008), *Mathematical Statistics with Applications*, Belmont CA: Brooks/Cole.

Wagenmakers, E. -J. (2007), "A Practical Solution to the Pervasive Problems of *p* Values," *Psychonomic Bulletin & Review*, 14, 779–804.

Wasserstein, R. L. (ed.) (2016), "ASA Statement on Statistical Significance and *P*-values," *The American Statistician*, 70, 131–133.

The *p*-Value Is Best to Detect Effects

Donald B. Macnaughton

Supplementary Material

Table of Contents

The appendices are in a different order from the order in which they are referenced in the body. The order is sensible for readers who wish to read the appendices from beginning to end. Other readers can locate appendices of interest through studying the preceding table of contents.

Appendix A: The Jeffreys-Lindley Paradox

This appendix will be of more interest to statisticians than to general readers.

The posterior probability that the null hypothesis is true (as discussed in section 5.6 in the body of this paper) leads to a puzzling paradox: Consider the task of assessing from research data whether a regression coefficient in a model equation is different from the null value of zero—the problem that was discussed in section 5 in the body.

In this situation, the posterior probability that the null hypothesis is true is (like the other measures of the weight of evidence) a function of the parameter estimate. This function is derived under reasonable assumptions by Berger and Sellke (1987, equation 1.1). In their example 1, they base their derivation on the assumption that the variance, σ^2, of the sampling distribution of the parameter is *known* and they refer to the effect-size function of the parameter of interest as *t*. Their *t*-statistic is closely related to the conventional *t*-statistic (discussed in section 5.2 of this paper) for which the variance of the sampling distribution of the parameter is *unknown* (and thus is estimated from the data).

In either case, the *t*-statistic is the standardized distance of the parameter estimate from the null value of the parameter. It is "standardized" in the sense that it is the distance in raw units divided by the (known or estimated) standard error of the estimate. This division by the standard error has the useful effect of making the *t*-statistic dimensionless and comparable from one research situation to the next, as discussed in section 5.2.

(Technical Aside: For the following discussion it is noteworthy that under the assumption that the variance of the sampling distribution of the parameter is *known*, the Berger and Sellke *t*-statistic takes complete account of the sample size in the sense that the sample size is appropriately used in the computation of the estimate of the standard error used in the denominator of the *t*-statistic. Under the assumption that the variance is *unknown*, the *t*-statistic takes *almost* complete account of the sample size, with the shape of the distribution curve varying slightly depending on the degrees of freedom, which depend mainly on the sample size. However, although it is useful to review these points for clarity, they arguably aren't relevant for the present discussion. That is, the phenomenon described in the following paragraphs occurs regardless of whether the variance of the sampling distribution is known.)

It is of interest to study the relationship between the Berger and Sellke *t*-statistic and the posterior probability that the null hypothesis is true. Figure A.1 shows (for three different sample sizes) the relationship according to Berger and Sellke's function.

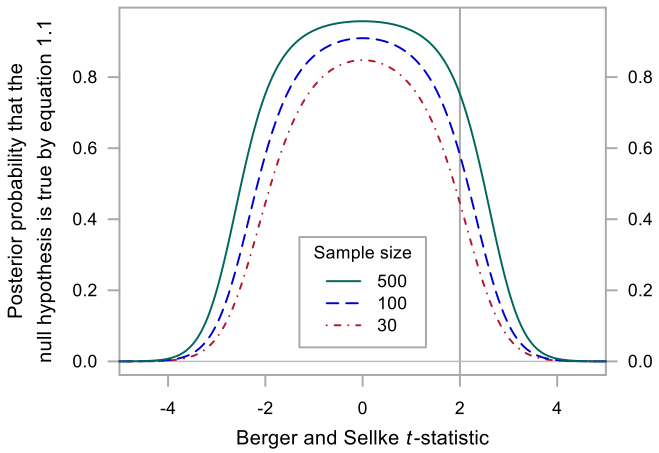

Figure A.1. The relationship between the Berger and Sellke *t*-statistic and the posterior probability that the null hypothesis is true for three different sample sizes assuming that the prior probabilities that the research and null hypotheses are true are both 0.5. The figure was generated using Berger and Sellke's equation 1.1. The R program to generate this figure is in the supplementary material for this paper.

The figure shows that Berger and Sellke's function behaves appropriately in the sense that the *higher* the value of the *t*-statistic is above zero (or the lower the value of the *t*-statistic is below zero), the *lower* the "probability" that the null hypothesis is true, as we would expect. However, the function appears to behave inappropriately in the sense that for a given value of the *t*-statistic, the *greater* the sample size, the *higher* the "probability" that the null hypothesis is true.

For example, the vertical line at 2 on the horizontal axis of the figure tells us that if the value of the Berger and Sellke *t*-statistic is 2.0 and if the sample size is 30, then the "probability" that the null hypothesis is true is roughly 0.45. But if the value of the value of the *t*-statistic is 2.0 and the sample size is 100, then the "probability" that the null hypothesis is true is roughly 0.58. And if the value of the *t*-statistic is 2.0 and sample size is 500, then the "probability" that the null hypothesis is true is roughly 0.75.

These results are counterintuitive because we would think that for a given value of the *t*-statistic (i.e., a given standardized distance of a parameter estimate from the null value), the larger the sample size, the more evidence we have that the null hypothesis is *false*. But the figure is showing that for a given value of the *t*-statistic, the larger the sample size, the more evidence we have that the null hypothesis is *true*.

Furthermore, in the case when the *t*-statistic is 2.0, if the sample size is 500, then the conventional critical *t*-value of 2.0 implies that we can (in the absence of a reasonable alternative explanation) just barely reject the null hypothesis. But the Berger and Sellke formula is telling us that the probability that the null hypothesis is true is 0.75, implying that it is more likely than not that the null hypothesis is *true*. This is in *direct opposition* to the conventional measures of the weight of evidence.

The idea that for a given value of the *t*-statistic, a larger sample should give us *more* evidence that the null hypothesis

is *false* is derived from the law of large numbers (also discussed in section 5.8). This law implies that the larger the sample, the closer we can expect (on average) the standardized parameter value estimated from the sample data (i.e., the Berger and Sellke *t*-statistic in the present case) to be to the correct value of the parameter for the entire population. This, in turn, implies that, for a larger sample, the distance of the parameter estimate from the null value is a more reliable estimate of the true value of this distance in the population. But if for a larger sample we have a more reliable estimate of the value of the parameter, and if this estimate is different from the null value, then this should cause the "probability" that the null hypothesis is true to be somewhat *lower*, not higher, than for a smaller sample.

The puzzling result illustrated by the figure is an example of the "Jeffreys-Lindley paradox", which Berger and Sellke (1987) discuss in the context of their equation 1.1 that was used to generate the figure. However, despite the many published "explanations" of the Jeffreys-Lindley paradox, the fact that the posterior probability that the null hypothesis is true is a counterintuitive *increasing* function of the sample size for a given value of the *t*-statistic suggests that this phenomenon isn't merely a "paradox", but is a contradiction.

We can't readily attribute the paradox or contradiction to the fact that the prior probabilities assigned to the research and null hypothesis are both 0.5. For if we change these probabilities in multiple small increments, then the locations of the lines on the graph will change in step, but the fact that larger sample sizes are associated with higher lines on the graph for a given value of the *t*-statistic won't change. This will be correct at least within some limited but likely wide range of the prior probabilities.

This apparent contradiction tells us that something is wrong here because the probabilities are misbehaving. This raises the question whether the posterior probability that the null hypothesis is true is scientifically meaningful. The apparent contradiction also raises the question of whether the Bayes factor is scientifically meaningful because the posterior probability that the null hypothesis is true is derived directly from the Bayes factor.

Appendix B: Details About Hypothesis Testing with *p*-Values to Detect Relationships

Sections 2 through 4 in the body of this paper present a high-level discussion of statistical hypothesis testing. The present appendix expands the ideas for less-experienced readers, focusing on the *p*-value. The discussion is a mixture of simple statistical ideas and basic ideas of scientific research.

B.1. First, Clean the Data

If we wish to study a relationship between one or more predictor variables and a response variable using the data in a data table, then we must first perform a crucial housekeeping step. In this step we carefully identify and correct errors in the values in the table. This (mundane) "cleaning" step is important because data errors occur surprisingly often in scientific research, and the errors will obviously distort any analyses we do of the data. Omit data cleaning at your peril. Some statistics textbooks explain how to examine and (without bias) clean scientific research data.

B.2. The Research and Null Hypotheses

As noted in the body, we can use hypothesis testing to determine whether we have good evidence that a relationship exists between variables. We first partition the possibilities about the phenomenon under study into two mutually exclusive and exhaustive hypotheses—the *research* hypothesis and the *null* hypothesis. The research hypothesis describes the general version of the phenomenon that we believe exists in the population, but we haven't yet reliably observed. In contrast, the null hypothesis describes the "null" situation—the situation in which the phenomenon that is under study *doesn't* exist in the population. We perform a hypothesis test of appropriate research data to help us decide which of the two hypotheses is (likely) true.

Typically, the (main) research hypothesis in a scientific research project states that a relationship between certain variables *exists* in the entities in the population of entities we are studying. But, more generally, a research hypothesis can assert the existence of something that isn't a relationship between variables, such as the existence of a particular type of entity, such as a fundamental physical particle. In either case, the problem is the same—we need to determine whether we have good evidence that the postulated thing (relationship or other entity) exists.

Thus in medical research to test a new drug, the research hypothesis says that the drug has an effect on the patients—a detectable relationship *exists* in the population of patients between the variables "drug dose" and "patient response", where "patient response" is a relevant measure of the wellness or illness of a patient. Note how the research hypothesis simply says that the drug has an effect on the response variable in the patients, but with no details about the effect.

In drug research there is invariably a further presumed hypothesis, which is that the drug under study has a *beneficial* effect on the patients, as opposed to a detrimental effect. This hypothesis is present because the goal of drug research isn't merely to find an effect, but is to find a useful beneficial effect—an effect that makes patients better, not worse. This point generalizes to many areas of scientific research—researchers often have a strong preference for discovering one type of effect, a "beneficial" effect, as opposed to the opposite "detrimental" effect because a beneficial effect will have a positive payoff, as opposed to a negative payoff.

However, the important pragmatic hypothesis that a given treatment has a beneficial effect is outside the general machinery of hypothesis testing. And standard formal hypothesis testing ignores the researcher's preference because occasionally when we analyze the relevant research data we find good evidence of the *opposite* effect to what we expected. Thus the standard conservative form of hypothesis testing is impartial and allows equally for the possibility of an effect that is opposite to what we expect.

The preceding point has the important technical implication that a "two-sided" or "two-tail" statistical test should generally be used instead of a one-sided test. That is, we take account of both the upper and the lower tails of the distribution shown in figure 1 in the body. A one-sided test is permissible only in situations in which we are completely confident that the estimated value in nature of the parameter under consideration can only occur on one side of the null value. That is, if parameter estimates above (or perhaps below) the null value are impossible, then a one-sided test is permissible. Here, the impossibility is determined by the situation being measured, and not merely by limitations of the measuring instrument. This situation occurs, but it is rare, so it is sensible to almost always use two-sided statistical tests.

Some researchers are attracted to a one-sided *p*-value because it is generally only half as big as the corresponding two-sided *p*-value. So switching from a two-sided *p*-value to a one-sided *p*-value may cause the *p*-value to jump from being *above* the critical value to being *below* it. That is, choosing to use a one-sided *p*-value may enable the researcher to satisfy the conventional criterion for publication. However, arguably, that is inappropriately bending the rules.

It is noteworthy that some authors refer to hypothesis testing as "null hypothesis significance testing", sometimes using the acronym NHST. This term is arguably inappropriate because it emphasizes the relatively unimportant null hypothesis—the hypothesis that we are trying to escape from. Therefore, it is more sensible to emphasize that we are attempting to show good evidence that the relevant *research* hypothesis is noticeably *true*, rather than attempting to show that the opposing less important *null* hypothesis is noticeably *false*. Thus it is sensible to call the procedure "research hypothesis testing" or simply "(statistical) hypothesis testing".

The preceding discussion refers to the relationship between a *single* predictor variable and a response variable. However, as noted in the body, often in a scientific research project we have *multiple* predictor variables. In this case the ideas are similar but slightly more general—we are interested in determining whether there is a relationship between (a) one or more of the predictor variables and (b) the response variable in the entities in the underlying population. And, for each possible relationship between the variables we will have a research hypothesis and a corresponding null hypothesis. We examine the research data to determine which of the multiple research hypotheses (if any) is or are (likely) true in the entities in the population.

B.3. The Beginning Assumption that the Null Hypothesis Is True

As noted in the body, the widely accepted scientific principle of parsimony (also called Occam's or Ockham's razor) tells us to keep things as simple as possible while remaining consistent with all the known facts (Baker, 2016). A sensible justification of this principle is the rhetorical question: Why make things more complicated than need be—why make things up that we don't know are true?

As also noted in the body, most researchers who study relationships between variables strongly believe at the beginning of a research project that the relationship between variables (or other effect) they are studying exists. However, a certain percentage of the time (possibly higher than 50%, depending on the scientific discipline) we are wrong. That is, unfortunately, the relationship between variables we are studying *doesn't* exist, and the null hypothesis is (actually or in effect) true. Thus, by convention, to reduce errors in scientific research, we aren't allowed to *formally* believe that a relationship between variables exists until someone has properly *demonstrated* that it exists.

The preceding paragraph refers to the idea that a null hypothesis may be "in effect" true. This important idea enables us to take account of the possibility that a null hypothesis may be false, but the associated relationship between variables is extremely weak—so weak that it is undetectable in the present research. It isn't possible to distinguish between (a) the case when a null hypothesis is *precisely* true and (b) the case when the null hypothesis is false, but it is *in effect* true (i.e., a relationship between the variables exists, but it is too weak to be detectable). Our inability to distinguish between these two cases generally isn't a serious problem because if a relationship between variables is so weak that it is undetectable, then this implies that it is so weak that it isn't useful in any reasonable (i.e., noticeable) sense.

We begin the study of a new relationship between variables with the formal assumption that the null hypothesis about the relationship is true. But informally we usually strongly believe and hope that our research hypothesis is true. If we have chosen the research hypothesis sensibly, and if we can show through our research that the research hypothesis is (likely) true, then this will increase human knowledge.

B.4. Model Equations

Hypothesis testing uses a sensible mathematical procedure to help us to decide whether we can reject a given null hypothesis and (tentatively) conclude that a relationship exists between (a) selected predictor variable(s) and (b) the response variable in the entities in the population. As noted in the body, the procedure is based on a study of a "model equation" of the relationship between the variables. The model equation states the mathematical form of the relationship between variables that we believe (hope) exists.

We can write the general form of a model equation as

$$y = r(x) + \varepsilon \qquad (2)$$

where y is the response variable and x is the predictor variable(s). The x may symbolize either a single predictor variable or a vector of two or more predictor variables.

The $r(x)$ in equation (2) is a mathematical function of x. This function may be any (single-valued) mathematical function—the choice of the function is at the researcher's discretion. A researcher will try to choose a form for $r(x)$ so that it best mimics the true form of the relationship between the predictor variable(s), x, and the response variable, y, in the population. Statistics textbooks discuss approaches to selecting the best function for a model equation of a relationship between variables.

(This paper follows the convention used by Efron and Hastie [2016] to use the notation $r(x)$ instead of $f(x)$ for the function because $f(x)$ is by convention used in statistics to represent a density function.)

If we have derived a model equation properly, then we can use it to make predictions. For example, suppose that we have derived a specific form of model equation (2). And suppose we measure the numeric values x of the properties of a new entity from the population, and suppose that the specific numeric values can be represented symbolically as x'. Then we can predict that the value of y for this entity will be $r(x')$, which will translate into a real predicted numeric value of y.

The ε in equation (2) is the "error" term. It reflects the fact that a model equation can almost never predict the value of y perfectly. The ε represents the difference between the correct (measured) value of the response variable for an entity and the value of the response variable predicted by $r(x)$:

$$\varepsilon = y - r(x).$$

The error term is viewed as varying at random from entity to entity in the population, with the distribution of the values of the term typically being a random normal distribution, as explained in statistics textbooks.

If we find good evidence of a relationship between variables and if we then properly derive a model equation, $r(x)$, for the relationship, then the predictions made by the equation for new entities from the population will be good predictions in the sense of being more accurate and more precise than other predictions that don't take account of the relationship (and don't take account of other relevant relationships). The increase in accuracy and precision of predictions may be substantial or it may be minimal, depending on the strength of the relationship between the variables, and depending on the design of the research project we use to derive the equation.

Equation (2) is completely general, but real model equations are more specific. For example, recall the model equation discussed in section 5 of the body of this paper:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_qx_q + \varepsilon \qquad (1)$$

where:

y is the response variable

$x_1, x_2, \ldots, x_q$ are the q predictor variables

$b_0, b_1, \ldots, b_q$ are the $q + 1$ parameters of the equation (regression coefficients), and

ε is the error term of the equation.

As noted in section 5.11 in the body, many other forms of model equation are also available. We use statistical procedures to help us to choose the form that works best to model the relationship between the variables we are studying.

B.5. Parameters of Model Equations

The $b_0, b_1, \ldots, b_q$ in equation (1) are the $q + 1$ parameters of the equation. Almost all model equations have parameters, and the parameters are assumed to be fixed (i.e., constant) *numbers*. We can estimate the values of the parameters of a model equation through the analysis of appropriate research data. The estimated numeric values of the parameters are important because they help us to specify the exact form of the equation and because they help to detect relationships between the variables.

It is sensible to conceive of "true" values of the parameters of a model equation in the underlying population. The "true" value of a parameter of a model equation in the population is the numeric value of the parameter that we would estimate if our measuring instruments could measure with perfect precision and if we were able to perform the research project under study on a sample that includes *every entity in the population*.

The preceding paragraph implies that the true value of a parameter is meaningful because we can estimate it with any specified precision if we are prepared to spend enough resources. But the paragraph also implies that the exact true value of a parameter is generally unknowable because (a) we almost never have perfect measuring instruments and (b) we almost never have enough resources to study every entity in the relevant population. Fortunately, statisticians have discovered efficient methods for *estimating* parameter values from a data table so that (if we do everything properly) the estimated values will be as close as possible to the true values.

Statisticians have invented three sensible somewhat-related general methods to provide good estimates (from an appropriate scientific research data table) of the true values of the parameters of an appropriate model equation. These methods are the least-squares method, the maximum-likelihood method, and the Bayesian methods. Each method is optimal (according to its own sensible definition of "optimal"). And each method has many details, as explained in statistics textbooks.

It is reassuring that if we apply the three methods to a given applicable data table using a sensible model equation, then the methods usually give identical or highly similar estimates of the values of the parameters of the equation. This is because, at root, each method is trying to satisfy the same basic goal, which is to correctly estimate the true values of the parameters of the chosen equation for the studied relationship between variables in the entities in the population of entities under study.

Statisticians and programmers have programmed the parameter-estimation methods into easy-to-use software—generally the same software that we use to compute *p*-values. Thus we can (if we follow the rules) easily correctly perform these methods by supplying the data table and a few simple instructions to the software and then running the software. The software analyzes the data and provides "best" estimates of the numeric values of the parameters of the model equation of interest in easy-to-understand computer output.

The fact that the obtained parameter estimates are only *estimates* of the true values implies that if we perform the same research project to estimate the same parameter values two or more times, each time collecting fresh data, then the obtained numeric values of the estimate for a given parameter will *vary* (by "small" amounts) from one instance of the research project to the next. This variation has three sources: (a) possible variation in relevant *unmeasured* variables that vary from one instance of the research project to the next due to possible minor differences in the research conditions, (b) random measurement error in the measurement of the values of the response and predictor variables in the entities in each instance

of the research project, and (c) possibly a true *random* component of the variation [though it is difficult, perhaps impossible, to separate this component from the variation due to (a) and (b)].

The preceding discussion hints at the idea of the "true" model equation for a relationship between variables. Here is a sensible empirical definition based on the principle of parsimony:

> **Definition:** The **true** model equation with the **true** values of the parameters of a relationship between variables is the simplest equation and parameter values that makes the very best predictions of the values of the response variable from the values of the available predictor variables for new entities from the population.

This definition doesn't enable us to directly *identify* the true equation for a given relationship between variables. But the definition tells us how to *zero in* on the true equation (through trying different forms of the equation with relevant data and selecting the simplest form that reliably works best).

Appendix I below discusses two instances when the true values of the parameters of a model equation aren't viewed as *fixed* values, but are viewed as varying. As noted, we usually view the values of parameters that we work with in scientific research as *estimated* values. Appendix J discusses an instructive exception in the physical sciences in which we know the *exact* true values of certain parameters of model equations.

B.6. Detecting Relationships Between Variables by Examining Estimated Parameters

We determine whether a relationship exists between variables by determining whether the research data imply that the *estimated* value of a relevant parameter of the relevant model equation is meaningfully different from the *null* value of the parameter. As noted in the body, if we can demonstrate that the estimated value of a parameter is meaningfully different from (i.e., inconsistent with) the null value, then this implies that it is unlikely that the null hypothesis is true in the population (Cox, 2006, pp. 42, 197–198). This, in turn, implies that it is likely that a relationship exists between (a) the predictor variable(s) x associated with the parameter and (b) the response variable y. The following discussion expands these ideas.

Appendix B.5 names three sensible general methods that we can use to estimate the values of parameters of a model equation from scientific research data. Therefore, in theory, we can determine whether there is a relationship between two variables by collecting appropriate data (i.e., by collecting values of the two variables from members of a representative sample of entities from the population). Then we can use one or more of the parameter-estimation methods to estimate (from the data) the value of the relevant parameter of the appropriate term in an appropriate model equation for the relationship between the two variables. (We can choose an "appropriate" model equation through careful examination of scatterplots or other graphs of the data.) Then we can check whether the estimated value of the parameter is different from the null value. If we find that the estimated value is different from the null value, this suggests that we can reject the null hypothesis and conclude that a relationship exists between the two variables.

However, though the preceding ideas are theoretically correct, there is a further complication: If we estimate the value of a parameter of a model equation from appropriate scientific research data, then (as discussed above in appendix B.5) the estimated value will vary from one research project to the next. This implies that the estimated value of a parameter will virtually never be *exactly* equal to the null value, even when there is no relationship whatever between the variables in the population. This phenomenon occurs even in realistic *artificial* data in which (by construction) there is absolutely no relationship between the variables. The phenomenon is due to inescapable random noise in data.

Therefore, if we wish to determine whether a relationship exists between certain variables, we can't simply check whether the estimated value of the relevant parameter of a relevant model equation is *different* from the null value (because the estimated value will almost always be different from the null value). Instead, we must check whether the estimated value is *significantly* different from the null value—far enough away from the null value to be well above the noise.

More generally, we can determine whether we have good evidence that a relationship exists between variables by checking whether an appropriate test statistic (which may or may not be a parameter of a model equation) is meaningfully different from the relevant null value. For example, we may check whether an *F*-statistic in analysis of variance is meaningfully different from its null value of approximately 1.0.

Statisticians have invented the *p*-value (and the eight other measures) to help us to determine whether the estimated value of a relevant parameter (or test statistic) is significantly different from the null value. This enables us to "test" whether we have good evidence that a relationship exists between the studied variables in the entities in the studied population.

B.7. The p-Value

Consider a standard definition of the *p*-value:

> **Definition:** The ***p*-value** for the estimated value of a parameter of a model equation (or the *p*-value for the value of a relevant test statistic) is the *fraction of the time* (i.e., the probability) that the value, as estimated from relevant research data, will be as discrepant or more discrepant from the relevant null value as the value estimated from the present data *if* the following three conditions are or were all satisfied:
>
> - the associated null hypothesis is or were true in the population
> - we were to perform the research project *over and over*, each time using a fresh random sample of entities from the population, and
> - certain often-satisfied technical assumptions that are required to correctly compute the *p*-value are or were satisfied.

The definition implies that the *p*-value is the probability of the relevant event occurring *if the null hypothesis is or were true*. This initially may seem odd because we are highly interested

in proving that the null hypothesis is *false*. So why are we computing probabilities that pertain to the undesirable situation when the null hypothesis is true?

The answer is that this approach is (arguably) logically the most sensible approach, even though it is roundabout. The approach is most sensible because many researchers agree that nobody has proposed a *better* approach, though various alternative approaches have been proposed, as discussed in sections 5 and 6 in the body.

Consider the logic of the *p*-value. The definition implies that the lower the *p*-value, the *less likely* it is that a parameter estimate as far from the null value (or farther) as was obtained would be obtained if the null hypothesis is or were true in the population (and if the underlying assumptions are satisfied). Therefore, in a sensible conceptual leap, if the assumptions are adequately satisfied, the lower the *p*-value, the *more evidence* we have that the value of the associated parameter is different from the null value in the population.

But if the value of a parameter of a model equation in the population is different from the null value, then this implies that *a relationship exists* in the population between the predictor variable(s) associated with the parameter and the response variable. Thus, the lower the *p*-value below the critical value (and in the absence of a reasonable alternative explanation), the more evidence we have that a relationship (or other studied effect) exists—the more evidence we have that the relationship is real.

Statistics textbooks explain methods to correctly compute *p*-values from the data in an appropriate data table. The various methods enable researchers to study the many different types of relationships between variables that can exist. The textbooks also explain the underlying technical assumptions for each method.

Fortunately, all the standard procedures to compute *p*-values for hypothesis tests have been programmed in user-friendly statistical software. Thus a researcher needn't understand the mathematical details of how to compute *p*-values. Instead, if a researcher knows the name of the appropriate test, then he or she can easily compute a correct *p*-value for a relationship between the variables by supplying the relevant data and a few simple instructions to the appropriate software, and then "running" the software. The software will analyze the data, apply the requested test, and compute the correct *p*-value from the data and will also compute various other important statistics. Thus we need only understand the underlying assumptions—the software will look after all the math.

If you would like to perform a statistical test for a relationship between variables in a data table, but are uncertain about the best statistical test, and if you are at a college or university, then the statistical consulting group at your institution may be able to help. If such help is available, you can greatly decrease the chance of problems by consulting with them *before* you finalize your research design (i.e., well before you begin collecting data) because they may be able to help you to substantially improve the design.

Many different statistical software systems can compute the same *p*-values. It is reassuring that if the various mainstream systems all perform a well-established hypothesis test with the same data table, then they all report *exactly* the same *p*-value. They also all agree *exactly* about the parameter estimates and about the values for each of the other well-established statistics pertaining to the analysis.

The conclusions of the preceding paragraph are excepting rare programming errors. The conclusions are also excepting small differences caused by computer rounding errors. These extremely small errors are generally in the last one, two, or three significant decimal digits in numbers that generally have fifteen-decimal-digit precision.

In some unusual cases no appropriate software is available to compute correct *p*-values for the parameters of a model equation of a studied relationship between variables. However, if the research was done properly, in these cases it is usually easy for a statistician to write a simple custom program to compute appropriate *p*-values through randomization tests or through a Monte Carlo simulation of the research situation under study.

A thoughtful reader might sensibly ask why we don't compute the probability that a parameter will have the *exact* value that it has. The answer is that the probability that a parameter has a particular *exact* value in a continuous possible range of values can be shown to be always zero. Therefore, we can only compute the probability that the parameter estimate lies in some *range* of values.

We could (under various assumptions) compute the probability that the parameter lies in a small range around its estimated value, but then we must specify the width of the range, which is theoretically possible, but seems arbitrary. We could also compute the probability *density* under the null hypothesis at the estimated value, although that isn't done, perhaps because it is less intuitive for many people. Instead of the preceding two approaches, we compute the probability (under the assumption that the null hypothesis is true) for the range of all values that are as far as or farther than the parameter estimate is from the null value. This approach is sensible because it helps us to control the false-positive error rate, as discussed in section 6 of the body of this paper.

The modern use of the *p*-value to detect relationships between variables is an amalgamation and an evolution of the work of John Arbuthnot (1710), Daniel Bernoulli (1734), Karl Pearson (1900, 1904), William Sealy Gosset (1908), Ronald Aylmer Fisher (1925, 1935), and coauthors Jerzy Neyman and Egon Sharpe Pearson (1928, 1933a, 1933b). Modern technical views of statistical hypothesis testing and statistical inference are discussed by Casella and Berger (2002), Lehmann and Romano (2005), and Cox (2006).

B.8. The Critical p-Value

As noted in the body, researchers often specify a "critical" *p*-value. This is the value that the *p*-value obtained in a research project must be less than or equal to before we will conclude that we have (in the absence of a reasonable alternative explanation) reasonable evidence that the relationship between variables we are studying exists in the population—enough evidence to allow us to reject the null hypothesis. By convention, researchers often use a critical *p*-value of 0.05 or 0.01, though some use and recommend lower critical *p*-values, as discussed below in appendix E. Of course, regardless of which critical

p-value a *researcher* appeals to in reporting and interpreting his or her research, each individual *reader* of the research report is free to use their own critical *p*-value in interpreting the research.

The critical *p*-value is sometimes referred to as the "significance level" of the statistical test and is sometimes symbolized by the Greek letter alpha, α. The critical *p*-value is also sometimes referred to as the "alpha level". However, these terms are arguably inferior to the term "critical *p*-value" because they carry less explanatory content.

It is important to remember that if a particular *p*-value is less than the critical value and if therefore a report of a putative relationship between variables is published in a research journal, this doesn't imply that the studied relationship between the variables definitely exists, for it may reflect a false-positive error. All that it implies is that in the opinion of the editors and referees who reviewed the paper, the evidence is strong enough that it is sensible to publish the results of the research project so that other researchers can know about the results. If other researchers think that the results are important, then they will attempt to replicate and extend the results. If the replications are successful, this strengthens the evidence that the relationship exists in the population.

As noted in the body, the procedure of computing a *p*-value and then determining whether it is less than or equal to the critical *p*-value is a statistical hypothesis test of the research hypothesis. This is also often referred to as "statistical inference" because we are making tentative inductive inferences from the data about effects in the underlying population.

As a convention, it is sensible to work with *p*-values that are rounded to two significant digits. This is sensible because more significant digits generally aren't meaningful, as can be seen by studying the typical variation that occurs in *p*-values if a research project is performed over and over, as illustrated below in appendix B.13. Using only *one* significant digit arguably isn't precise enough to instill confidence.

B.9. Positive Results and Negative Results

As noted in the body, if we compute a *p*-value in a proper test of a research hypothesis using appropriate scientific research data, then there are only two possible outcomes of the test, either a "positive result" (when the *p*-value is *less than or equal to* the critical *p*-value) or a "negative result" (when the *p*-value is *greater than* the critical *p*-value.

A *positive* result implies (in the absence of a reasonable alternative explanation) that we have good evidence of the existence of the effect or phenomenon we are looking for—good evidence that the research hypothesis is true—typically good evidence that the relevant relationship between variables exists in the population.

In contrast, a *negative* result implies that we have found no good evidence of the existence of the effect or phenomenon we are looking for.

We can also obtain a negative research result if we initially obtain a positive result, but then we discover a reasonable alternative explanation for the result. The reasonable alternative explanation turns the positive result into a negative result because the alternative explanation implies that the result is equivocal, and scientific research strives to be decisive.

Consider an example of negative results: In the 1950's some medical practitioners strongly believed (based on informal clinical experience) that laetrile (derived from apricot pits) could cure cancer. This led to initial public excitement about laetrile. These informal positive results led to a formal experiment (published in 1982) to look for evidence of a relationship between laetrile and cancer. But the experiment obtained a negative result—it found no good evidence of a relationship between the amount of laetrile administered to cancer patients and the amount of cancer in the patients. And virtually all other careful research to study the effects of laetrile on cancer has also obtained a negative result. Therefore, all mainstream medical researchers now believe that there is no beneficial relationship between laetrile and cancer in cancer patients (National Cancer Institute, 2018).

Though negative results occur often, we rarely hear about them. This is because negative results are generally uninteresting, only telling us that the research project failed to find what it was looking for. Scientists and the public are much more interested in *positive* results in scientific research—results in which a new effect or phenomenon is discovered. For example, the general interest in positive results was reflected in the initial excitement about laetrile. Positive results (when they are correct) often lead to useful applications (e.g., a cure for cancer). But negative results usually lead nowhere.

In view of the lack of usefulness of negative results, and in view of the many positive results that are vying for the limited space in scientific journals, most scientific journals will almost never publish the report of a research project whose main finding is a negative result. This is sometimes a source of frustration for researchers who believe that their negative results are important. Appendix K discusses journals and registries that do provide information about negative results. Appendix L discusses some instructive exceptions to the general rule that negative results won't be accepted for publication in a scientific journal.

B.10. False-Positive and False-Negative Errors

Regardless of which approach we use to help us to decide whether a relationship exists between variables, we must take account of the possibility of false-positive and false-negative errors. As noted, a false-positive error occurs if (through chance or through some other reason) we obtain a positive result and therefore conclude that a relationship exists between variables, but behind the scenes *no* detectable relationship exists between the variables in the population. Using the terminology of signal detection theory, a false-positive error is sometimes called a "false alarm".

If a research project makes a false-positive error, then the researcher often doesn't recognize this at the time. Instead, the researcher generally believes that the positive result implies that the research hypothesis is true. The researcher believes this because that is what he or she is trying to prove. So the researcher generally has an understandable and inescapable bias in favor of the research hypothesis.

False-positive errors are costly in the sense that if a false-positive error is published, and if the result is important enough, then this will lead other researchers to try to replicate the research finding to confirm and extend our knowledge about the effect. But if the original result is a false-positive error, then this replication research merely amounts to a wild-goose chase that necessarily must fail—an undesirable (but once the error is published, unavoidable) waste of resources.

It is theoretically possible to compute the rate of occurrence of false-positive errors in a scientific discipline. This rate depends on (a) the rate of study of true research hypotheses in the discipline, (b) the average critical *p*-value used in the discipline, and (c) the average power of the statistical tests used in the discipline (Jager and Leek 2014, fig. 1). This dependence is illustrated in figure B.1.

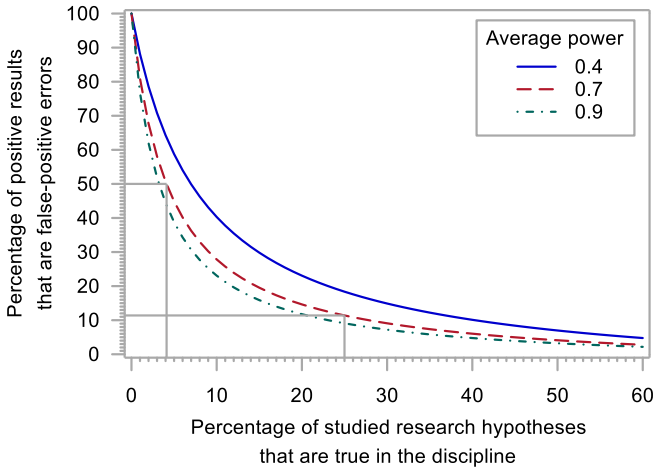

Figure B.1. A graph showing the percentage of statistically significant results that reflect false-positive errors in research projects in a discipline (e.g., in medical research) as a function of the percentage of studied research hypotheses that are true in the discipline. The R program to generate this graph with an explanation of the logic is in the supplementary material for this paper.

The three curving lines on the graph show how the percentage of positive results that are false-positive errors in a scientific discipline depends on the percentage of studied research hypotheses that are true in the discipline. The lines show this function for three different hypothetical averaged statistical powers of the statistical tests performed in the discipline. (The power is assumed to be computed across all the research projects in the discipline when the research hypothesis is at least minimally true.) For example, if the average power of statistical tests in a discipline is 0.7, and if the average critical *p*-value used in the discipline is 0.03, then the dashed red line shows the relationship between the percentage of true research hypotheses and the percentage of false-positive errors.

The vertical line at 25 on the horizontal axis of the graph implies that if 25% of the research hypotheses that are studied in a research discipline are actually true, and if the average power of statistical tests in research projects in the discipline is 0.7, and if the average of the critical *p*-values used in the research projects in the discipline is 0.03, and if there are no extenuating factors, then roughly 11% of the positive research results in the discipline will (of mathematical necessity) be false-positive errors.

The graph is based on the assumption that the "average" critical *p*-value used in all research in the discipline is 0.03. However, the graph can be readily redrawn for another assumed average critical *p*-value and will be similar. (The lower the average critical *p*-value used in a discipline, the closer the three lines move to the lower left corner of the graph.)

As noted, the graph has three predictor variables associated with it: the percentage of research hypotheses studied in the discipline that are true, the average critical *p*-value in the discipline, and the average power in the discipline. Unfortunately, we generally can't use the graph to determine the percentage of positive results that are false-positive errors for a given scientific discipline because we generally don't reliably know the value of *any* of the three predictor variables for the graph for the discipline. However, the graph is still useful because it helps us to see how things work.

The graph is based on the assumption that research is done without other types of error beyond false-positive and false-negative errors due to chance. However, various researcher errors (e.g., cherry picking [see below], incorrect analyses, carelessness, fraud) can occur in scientific research. These errors have a net effect of causing the point for a given scientific discipline to be somewhat different from its theoretical value on the graph.

Ioannidis (2005) suggests that more than half of the research findings published in medical research articles reflect false-positive errors. The horizontal line at 50 on the vertical axis shows that this will be the case if all positive results are published and if the percentage of studied research hypotheses in medical research that are true is less than around 4.1% and if the average power of statistical tests in medical research projects is around 0.7 and if, on average, medical research projects use a critical *p*-value of 0.03, and if there are no researcher errors. However, the preceding assumptions about the power and the critical *p*-value may be unrealistic. Also, researcher errors also cause some false-positive errors in medical research. Thus the percentage of studied research hypotheses that are true in medical research could be greater than 4.1%, but still yield a 50% false-positive error rate.

Fortunately, it isn't necessary to know the rate of occurrence of false-positive errors in a scientific research discipline. This is because regardless of the rate of occurrence of the errors in the discipline, we can identify and eliminate the errors through appropriate replication, as discussed in section 4 in the body of this paper. All potentially important research results must be replicated before we can trust them because errors happen.

If we perform many hypothesis tests in a research project (as often occurs in modern data analysis with "big data") and if the null hypothesis is sometimes true (which is often the case, at least in effect), then false-positive errors become more problematical merely because we are performing so many hypothesis tests. For example, if we use the standard approach with a critical *p*-value of 0.05, then this implies that (even if we do everything properly) a false-positive error will occur roughly 5% of the time when there is (at least in effect) no

relationship between the relevant variables. Therefore, experienced researchers who perform many hypothesis tests in a research project use special procedures to control the rate of false-positive errors, as discussed by Benjamini and Hochberg (1995) and Efron and Hastie (2016, ch. 15).

The preceding paragraphs discuss the idea of a false-*positive* error in a scientific research project. In a similar serious problem, research projects sometimes make false-*negative* errors. As noted in the body, a false-negative error occurs if we obtain a negative result and therefore conclude that we have no evidence of the existence of a relationship between a pair (or larger set) of variables when, in fact, a relationship of the hypothesized form actually *does* exist in the population. A false-negative error is sometimes called a "failed alarm".

The probability of a false-negative error for a given statistical test for a given effect size is related to the power of the test. The greater the power, the lower the probability of a false-negative error. In fact, if the underlying assumptions are properly satisfied, then for a given effect size,

False-negative error probability = 1 − Power.

As suggested in the body, we can reduce the rate of false-positive errors by using a lower critical *p*-value, but this generally increases the cost of research. Similarly, we can reduce the rate of false-negative errors by using statistical test with higher power, but this also increases the cost. Therefore, we must compromise to contain costs. Diligent researchers plan their research projects to yield an efficient compromise between positive results, false-positive errors, negative results, false-negative errors, and research costs. In practice, this amounts to (a) using a critical *p*-value equal to the critical *p*-value that is standard in the field of study, (b) maximizing the power of the statistical tests under the available resources through careful research design, and (c) and ensuring that there will be no reasonable alternative explanations of the results that will invalidate the work, also through careful research design. Careful research design is the key to optimizing scientific research.

Appendix M below contrasts false-positive and false-negative errors with parameter sign errors ("Type S") and parameter magnitude errors ("Type M") in scientific research.

Colquhoun (2017) discusses a conceptually sensible statistic that he calls the "false-positive risk". (He omits the hyphen, but it is useful because it eliminates a minor ambiguity.) The false-positive risk in a scientific research project is the estimated probability that a positive research result obtained in the project is a false-positive error.

The false-positive risk is closely related to the variable shown on the vertical axis of the graph in figure B.1 above. However, Colquhoun's approach is based on first determining a prior probability that the null hypothesis is true. This is akin to choosing a particular percentage on the horizontal axis of the graph. Then other assumptions are made and then a mathematical procedure similar to the procedure behind the graph is used to calculate the estimated value of the variable on the vertical axis of the graph—the estimated percentage of positive results that are false-positive errors, which is sensibly called the false-positive risk.

Colquhoun's approach uses the likelihood ratio in its computations. It is noteworthy that Colquhoun defines the likelihood ratio on page 7 of his paper as $y_1/2y_0$, which can be expressed in terms of the relevant probability density functions as $f_1(x)/2f_0(x)$. This is different from the standard definition of the likelihood ratio, which is $f_1(x)/f_0(x)$, as discussed by Cox (2006, p. 91). Both Colquhoun and Cox use a reciprocal version of the likelihood ratio discussed in section 5.4 of the present paper, which is noteworthy to reduce confusion, though it has no substantive effect on the relevant ideas.

As noted, Colquhoun's approach is (in effect) based on choosing a point on the horizontal axis of figure B.1. However, as Colquhoun notes (2017, p. 7), we rarely (if ever) have a valid value for the prior probability that the null hypothesis is true or false. This is because failures to obtain a positive result in a scientific research project generally aren't tracked in a scientific discipline because many members of the discipline have sensibly judged that the difficult task of tracking these failures would be unreliable and isn't worth the effort. (This might change if mandatory research registries become the norm, as discussed in appendix K.) Without the tracking information, we can't determine the correct value on the horizontal axis for a scientific discipline. And, in general, we can only guess the value of this percentage.

Thus, arguably, if we wish to consider these concepts, it is sensible to use the graphical approach shown in figure B.1 and not try to guess (or somehow otherwise determine) the prior probability that the null hypothesis is true in a particular research situation to enable us to compute the false-positive risk. If we use the graphical approach, this allows us to see the entire range of possibilities instead of focusing on a specific and possibly unrealistic case.

Colquhoun's figure 4 shows graphs of the relationship between his computed false-positive risk and the obtained *p*-value. The lines for different sample sizes all cross each other on each graph. While not a disproof of Colquhoun's approach, the crossing lines are counterintuitive because isobar-like lines generally don't cross one another on a graph.

Colquhoun proposes in section 7 of his paper a sensible way around the problem that we don't know the percentage of studied research hypotheses that are true in a discipline. He suggests that we define a sensible critical value for the false-positive risk, and he proposes 0.05 for this value. Then, using the data at hand, we compute the prior probability that the null hypothesis is false that must obtain to enable us to obtain this false-positive risk. If this computed prior probability is low enough (e.g., less than a critical value of 0.5), then we can conclude according to this criterion that we have enough evidence that the effect is real. This is thus a way of detecting relationships between variables, performing the same function as the nine measures discussed in the body of this paper.

It is easy to show that Colquhoun's approach is in a monotonic relationship with the effect size when other factors are held constant. Therefore, we can, through the choice of critical values, calibrate Colquhoun's approach to behave equivalently (in decreeing a positive or negative result) with the nine measures discussed in the body. Colquhoun's approach is conceptually more complicated than the other measures in the

sense that it has effectively two critical values—a critical value for the false-positive risk and a critical value for the prior probability that the null hypothesis is false.

For comparison, if we use Colquhoun's approach with a critical value of 0.05 for the false-positive risk and with a critical value of 0.5 for the prior probability that the null hypothesis is false, then this leads to a much stricter statistical test than a test based on a critical *p*-value of 0.05. This implies that the statistical tests based on Colquhoun's critical values will make substantially fewer false-positive errors, but substantially more false-negative errors than the statistical tests based on a *p*-value with critical value of 0.05.

Of course, we can calibrate Colquhoun's approach to behave equivalently to the other approaches in deciding whether we have enough evidence to reject the null hypothesis. We do this by adjusting Colquhoun's critical values.

Colquhoun suggests that a researcher makes "a fool" of him- or herself if their research commits a false-positive error (2017, pp. 3). However, research that commits a false-positive error doesn't somehow turn the researcher into a fool. Instead, a false-positive error only shows that the researcher was either an unfortunate victim of chance or was somewhat careless—it may be difficult to tell which.

B.11. Reasonable Alternative Explanations

As noted, reasonable alternative explanations play a key role in scientific research. Most scientists won't accept a positive conclusion suggested by a scientific research result if there is a reasonable alternative explanation for the result. Instead, they will ask for or perform further research to determine which of the possible explanations is the correct explanation.

There is a wide range of possible standard types of reasonable alternative explanations of a research finding, including hidden variables, confounding, data-collection errors, data-analysis errors, equipment failure, and scientific fraud. Also, each field of study typically has certain unique reasonable alternative explanations for research findings that must be considered.

It is noteworthy that outright fraud in scientific research is rare because most researchers know that all consequential scientific fraud is exposed sooner or later due to the investigative nature of science. And researchers know that fraud leads to severe consequences, typically loss of employment, loss of respect, and possible litigation.

Although fraud is rare, other types of reasonable alternative explanations often arise in scientific research. Researchers take careful pains to try to ensure that no reasonable alternative explanations can be proposed to explain away their research findings.

Sometimes there is a correct reasonable alternative explanation for a research finding, but the explanation is undetectable because the report of the research project omits the relevant information. For example, suppose that a researcher (in good faith) performs a research project over and over, each time adjusting the research conditions somewhat, hoping to find a set of conditions in which the effect under study will be observed. But suppose that behind the scenes the research hypothesis is *false* and thus the null hypothesis is true. If the researcher performs the research project enough times, then the definition of the *p*-value implies that some of the instances will obtain statistically significant results. This phenomenon is illustrated graphically in the left-hand panel in the figure in appendix B.13 below.

If in this situation the researcher *reports* only a single instance of the research project in which the significant result was obtained and doesn't report the fact that the research project was performed over and over, then (if other aspects of the research report are satisfactory) readers of the report will interpret the positive result as good evidence that the research hypothesis is true even though there is a reasonable alternative explanation for the result and the result is actually a false-positive error.

Selecting and reporting positive results from a large set of research results without reporting the negative results is called "cherry picking" the results. This *p*-value usage error, which is also called "data dredging", is sometimes committed by less experienced or less vigilant researchers in their (dedicated but poorly reasoned) efforts to obtain a positive result.

More generally, the practice of "*p*-hacking" is any procedure that takes steps to obtain lower *p*-values, but violates the assumptions underlying the *p*-value. We reduce these errors in scientific research by ensuring that all of the underlying assumptions of our statistical procedures are adequately satisfied and by trying to transparently document everything important that we do pertaining to the research. A commitment to transparency makes a researcher's work more credible and wins his or her peers' respect. The importance of full reporting and transparency in scientific research reflects Principle 4 in the ASA Statement on *p*-values (Wasserstein 2016).

If a researcher has a strong hunch about a relationship between variables, then it is fully permissible for the researcher to perform a research project to look for the relationship over and over, adjusting the conditions each time in the hope of finding conditions that yield a positive result. But if the researcher finds some conditions that appear to yield a positive result, then he or she should independently *replicate* this result with these conditions one or more times to confirm that the positive result hasn't occurred through mere chance. Some researchers don't do that, and instead publish a report of their positive result, to their later regret when their false-positive finding can't be replicated.

We can reduce false-positive errors caused by *p*-value usage errors through proper training of researchers. The training should point out that false-positive errors about important effects lead to a waste of resources. Therefore, an egregious false-positive error can lead to strong criticism or censure of the researcher by the research community.

For example, Stanley Pons and Martin Fleischmann were figuratively drummed out of chemistry for their apparent false-positive cold-fusion error, which sent many interested researchers on a costly wild-goose chase, as discussed by Huizenga (1993). Their now-believed-incorrect observation of excess energy from their cold-fusion cells may have been caused partly by cherry picking and partly by measurement problems.

The Pons and Fleischman error was egregious because their report describes several experiments that were highly successful in demonstrating cold fusion. But nobody has been able to reliably replicate any of these experiments, though many researchers have tried, at significant expense.

As noted in the body, the relevant research community decides whether a research finding is believable through evolving informal consensus in the community through formal and informal discussion about the finding and about attempts to replicate it. If nobody in the research community can think of a reasonable alternative explanation for the finding, and (as is usually required) if the finding has been successfully replicated, then the community will, in time, accept the finding.

What is the relationship between the *p*-value and the idea of a reasonable alternative explanation? The *p*-value (and the various other measures of weight of evidence that an effect is real) is merely a sensible method to tentatively eliminate *chance* as a reasonable alternative explanation of evidence that an effect observed in scientific research is real in the entities in the population of entities under study.

The importance of considering reasonable alternative explanations in drawing conclusions about scientific research reflects Principle 6 in the ASA Statement (Wasserstein 2016).

B.12. The Asymmetry of Statistical Hypothesis Testing

In scientific research we can never conclude that a given null hypothesis is *exactly* true. This is because even if the null hypothesis *is* exactly true, we can't empirically *demonstrate* that it is true. We can't demonstrate that there is no effect at all because it is always possible that an effect is present in the population, but it is too small for us to detect with our current research approach and measuring instruments, but we will detect it later.

Thus, for example, we can't somehow empirically demonstrate that extrasensory perception is exactly impossible. So experienced researchers never "accept" a null hypothesis. However, we *assume* that a given null hypothesis is true until (if ever) someone reliably proves otherwise. Thus we *assume* that extrasensory perception is exactly impossible because this is a sensible efficient (parsimonious) way to proceed.

In a related but opposing idea, some researchers and statisticians believe that the null hypothesis is *never* precisely true in a population (Berkson 1938; Bakan 1966; Colquhoun 1971, p. 95; Tukey 1989, p. 176, 1991, p. 100; Cohen, 1994, p. 1000; Jones and Tukey 2000; Nickerson 2000, p. 263). However, this belief is speculative because it can't be empirically confirmed. This is because it is impossible to study *every* null hypothesis in the universe and to somehow *confirm* that they are all false.

With strong faith in analysis, Rao and Lovric (2016) attempt to prove *analytically* that every null hypothesis in the universe is false. Unfortunately, their proof is tenuous due to the tenuous links between the set of analytical premises they use and the real world.

Furthermore, despite some researchers' belief to the contrary, some null hypotheses are probably *exactly* true in nature. For example, many readers will agree that there is almost certainly no *direct* relationship in people between carrying a "lucky" coin and having good luck. (There may be an *indirect* relationship for some people, such as in the sense that believing in a coin causes them to positively pursue more opportunities, which leads them to better "luck".) So in this example the null hypothesis (that there is no direct relationship in people between carrying a certain coin and good luck for them) is probably absolutely true.

But again, we can't know with certainty that the "lucky coin" null hypothesis is true. And it is *conceivable* (though most of us think it highly unlikely) that a person might obtain an amount of (real) extra good fortune if he or she carries a lucky coin. That is, it is *conceivable* that some "superstitious" people have correctly observed this (real) relationship between variables—a relationship that the rest of us think doesn't exist. And these "superstitious" people wisely use their knowledge of the relationship to increase their good luck (by carrying lucky coins).

As scientists, we *assume* that *anything* that is *logically* possible (even a truly lucky coin) is possible because we can learn more if our minds are open to any logically possible hypothesis. But we *also* always assume that the relevant null hypothesis is true until (if ever) someone demonstrates otherwise.

The lucky coin example illustrates the asymmetry of statistical hypothesis testing—we can never use empirical research to prove that a null hypothesis is *true*. But we *can* use empirical research to prove (beyond a reasonable doubt) that a given null hypothesis is *false* (assuming, of course, that the hypothesis actually *is* false).

B.13. The Distribution of the p-Value Under the Null and Research Hypotheses

The logic behind the *p*-value implies that if we were to perform the same scientific research project over and over, each time using a fresh sample of entities from the population, and if we were to compute the *p*-value for the same hypothesis test each time, then the value of the *p*-value would generally be different each time. It is instructive to study the distribution of the *p*-values that (under standard conditions) we will get if we repeat the same research project over and over. In other words, we study the relative frequency with which different *p*-values will occur. Of course, the distribution of *p*-values we get will depend on whether the research hypothesis or the null hypothesis is true.

It is easy to show that if we repeat a research project over and over, and if the null hypothesis is true (i.e., the relevant "effect size" is zero in the population), and if the assumptions underlying the *p*-value are satisfied, then the *p*-values will occur in a "uniform" distribution. This uniform distribution is illustrated in the left-hand histogram in figure B.2.

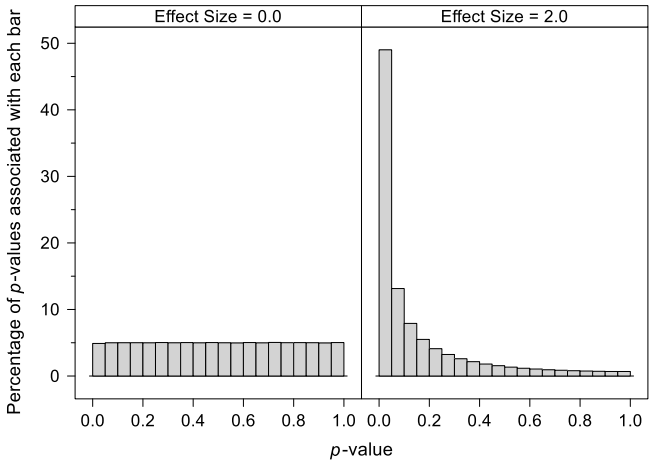

Figure B.2. Two histograms, each showing the distribution of the *p*-values that we will obtain if we repeat a particular research project over and over, each time collecting fresh data. Each histogram was obtained through a computer simulation. The SAS program to generate the underlying data and the R program to draw the figure are in the supplementary material for this paper.

The scale of the horizontal axis of each histogram in the figure ranges between 0.0 and 1.0 because that is the possible range of values of a *p*-value. The scale on the vertical axis of the histograms is a scale of percentages ranging between zero and 50%. If we add together the heights of the 20 bars on each histogram, the sum of the heights of the bars on each histogram is exactly 100%.

The histogram on the left summarizes the *p*-values in a data table that was generated to contain roughly 2.5 million simulated *p*-values under the assumption that the *null* hypothesis is true (i.e., the effect size is 0.0). Similarly, the histogram on the right summarizes the *p*-values in a data table that was generated to contain roughly 2.6 million simulated *p*-values under the assumption that the *research* hypothesis is true, and the effect size is 2.0.

(Technical details: The effect size in this example is the value of the non-centrality parameter of the noncentral *t*-distribution that was used to generate the data behind each histogram. The two histograms in the figure are based on a statistical test of a coefficient in a standard linear regression analysis, assuming the *t*-statistic for the coefficient has 30 degrees of freedom. However, similar histograms can be generated through a computer simulation for the *p*-values for *any* mathematically describable statistical hypothesis test.)

The histogram on the left shows that if the null hypothesis is true in the underlying population (and if the underlying assumptions of the computation of the *p*-value are adequately satisfied), then we can expect that the *p*-values we obtain if we repeat the research project over and over will be spread perfectly evenly between 0.0 and 1.0. (The almost imperceptible deviations from perfect uniformity in the left-hand histogram are artifacts of the discrete and thus slightly imperfect computer procedure that was used to generate it.)

The height of each bar on the left-hand histogram is theoretically exactly 5%. Thus this histogram tells us that in research projects in which the null hypothesis is true, if we divide the *p*-value range into 20 adjacent segments of equal width, then it will be equally likely for these research projects that the *p*-value will lie in any one of the segments—the *p*-value will lie in each segment 5% of the time. This implies that if the null hypothesis is true, and if the underlying assumptions are satisfied, we will obtain *p*-values that are less than 0.05 exactly 5% of the time. That is, if we do everything properly, in research projects in which the null hypothesis is (unfortunately) true (or in effect true), our statistical test will (at random) make a false-positive error roughly 5% of the times that we perform the test.

The histogram on the right is computed under the assumption that the population effect size is 2.0, which implies that the null hypothesis is *false*, and thus the *research* hypothesis is true. We see that in this case if we repeat the research project over and over (and if the underlying assumptions of the computation of the *p*-value are adequately satisfied), then the *p*-values tend to fall closer to the lower end of the range—i.e., closer to 0.0 than to 1.0. Clearly, this is exactly what we want because if the null hypothesis is false, then we want the *p*-value to be low because this *tells* us that the null hypothesis is false.

Consider the height of the leftmost bar on the right-hand histogram, which is the bar for the case when the *p*-value is less than 0.05. We see that the bar contains roughly 49% of the 2.6 million *p*-values behind the histogram. This bar tells us that, under the assumed conditions, we can expect the *p*-value to be less than 0.05 roughly 49% of the time if we repeat the research project over and over. Thus, using the definition of power from section 4 in the body, the power of the statistical test in this situation is 0.49. (This power is much lower than the 0.90 recommended in the body.)

The histogram on the right implies that in the research project under discussion the *p*-value will be *greater* than 0.05 in roughly $100 - 49 = 51\%$ of the time. The *p*-value will be greater than 0.05 *even though* the fact that the effect size is 2.0 implies that (behind the scenes) the null hypothesis is false. Thus this research project would make a false-*negative* error roughly 51% of the time if we were to perform the research project and over and over, each time (unfortunately) obtaining a *p*-value in the range between 0.05 and 1.00, telling us that we don't have enough evidence to reject the null hypothesis.

So, in the long run, in the situation illustrated in the histogram on the right, the *p*-value makes a false-negative error slightly more than half of the time. Thus a thoughtful reader might reasonably wonder if there might be a better way to analyze the data to detect a relationship between the variables that would (without detriment) lead to the correct positive result more often. Unfortunately, nobody has found a better way, and apparently there *is* no better way. That is, there is no obvious way (for a given effect size and a given research design) to decrease the false-negative error rate without also unacceptably *increasing* the false-positive error rate. This is because the two rates are tightly bound together, both being a

function of (a) the design of the research project under consideration, (b) the effect size, (c) the statistical procedure (e.g., the *p*-value) we have chosen to use to decide if we have sufficient evidence to (tentatively) reject the null hypothesis, and (d) the particular critical value that we have chosen to use with the procedure (e.g., 0.05 for the *p*-value).

Of course, the *design of the research project* is the key here. And in the research behind the right-hand histogram in the figure we could redesign the research project so that it has a more powerful statistical test, and *then* we would be more likely to find good evidence of the relationship—we would make false-negative errors less than 51% of the time. We can increase the power of statistical tests using the methods discussed in section 4 of the body.

B.14. Do p-Values Make Publication Decisions?

Demidenko suggests that a paper reporting a scientific research project will be accepted for publication in a scientific journal merely if the paper has a low-enough *p*-value for its main result (2016, sec. 2). This suggestion is based on a misunderstanding.

As noted in section 3 in the body of the present paper and in appendix B.9 above, scientific journals almost always only publish papers that report *positive* results—they almost never publish papers that report negative results (because negative results are generally uninteresting). Therefore, editors often sensibly use a critical *p*-value as a *screening rule* to determine whether a result is "positive" enough to be considered for publication in their journal (Estes 1997, Cox, 2014, Jager and Leek, 2014). That is, a paper won't be considered for publication unless the *p*-value for the main research finding in the paper is less than or equal to the journal's critical *p*-value (typically 0.05 or 0.01).

Thus, a low *p*-value in a hypothesis test for the main research finding of a research project is a *necessary* condition that must be satisfied before many journals will consider a paper reporting research results for publication. As illustrated by Demidenko, this leads some people to confuse things and to think that a low *p*-value is a *sufficient* condition for publication. However, a low *p*-value in a hypothesis test is never a sufficient condition for publication of a paper in a reputable journal.

Thus though *p*-values often *participate* in publication decisions, they don't *make* publication decisions. Instead, the editor of a journal will decide to accept a paper for publication only if (with rare exceptions) it has a sufficiently low *p*-value for its main finding *and* if the paper satisfies the journal's many other mandatory criteria for acceptance. These include the criterion that the main research finding must be "interesting" and the criterion that (except in unusual circumstances) there must be no reasonable alternative explanation for the low *p*-value for the main research finding.

Thus the *p*-value is merely one of many criteria that researchers and journals use to evaluate scientific research results. And if we are evaluating the results of a new research project, then we can consider the *p*-value criterion at any point during the evaluation. However, it is sensible to consider the *p*-value criterion *first* because this step can be done quickly (often in less than five minutes). And if the *p*-value obtained by the main result in a research project is *greater* than the critical *p*-value, then this implies that the result is inconclusive. Then it is generally sensible for a journal to abandon further consideration of the result because the high *p*-value implies that the result may merely reflect noise. That enables the journal to escape from performing the long process of considering other criteria in evaluating the result, which saves the journal's time.

The fact that journals generally apply the *p*-value criterion first in evaluating a research project leads some statisticians and researchers to think that the *p*-value has an undeserved special role, causing it to take priority over the "currently subordinate factors" (McShane, Gal, Gelman, Robert, and Tackett, 2018, p. 2). But the *p*-value doesn't have an undeserved role. And it is no more important than the other factors or criteria that researchers and journals use to evaluate scientific research results. We apply the *p*-value criterion first merely because that saves our time.

B.15. The Publication and Correction of False-Positive Errors

As noted in appendices B.8 and B.10, some scientific research results reflect false-positive errors. Currently, there is substantial interest in the problem of the publication of false-positive errors in scientific research, with some authors viewing the publication of these errors as a "replication crisis", a scandal, as discussed by Ioannidis (2005), Palus (2018), and in a *Nature* editorial (2018).

Appendix B.10 discusses the rate of occurrence of false-positive errors in *positive* results (as opposed to in *all* research results, positive and negative) in scientific research. The rate of *publication* of false-positive errors in each field will be similar to the rate of occurrence of false-positive errors in positive results. This is because generally only positive results are published, as discussed in appendices B.8 and B.9, and because, without the help of replication, false positive results are generally indistinguishable from true positive results.

As discussed in appendix B.10, it is difficult (arguably impossible) to directly determine the rate of occurrence of false-positive errors in scientific research in a given field. Therefore, the rate of publication of false-positive errors in a field can (apparently) only be determined (to a limited extent and in hindsight) by the study of failures to replicate published positive results in the field.

It is important to note that a single failure to replicate a positive result generally isn't definitive in determining that the earlier research reflects a false-positive error. This is because there are usually several possible reasons why a replication attempt failed. For example, the inevitable slightly different research conditions between the original research and the replicating research may lead to a negative result in the replicating research. Also, certain types of carelessness in research make it likely that a research project will obtain a negative result. Also, a failure to replicate a positive result amounts to a *negative* result and, as noted in appendix B.9, journals generally don't publish negative results because they

are less interesting. These ideas explain why journals are generally unwilling to publish a report of a single failure to replicate a positive research finding.

But if negative results aren't published, then how can the research community *know* about failures to replicate a positive result to expose it as a false-positive error? The answer is that news about failures to replicate an interesting positive result spreads informally. This is because scientists are always interested in knowing about these failures in their field. (But scientists are generally less interested in the details.)

News about failures to replicate is spread in personal discussions, in social media, in newsletters, at scientific meetings, or in journal articles or letters to the editor summarizing the results of *several* independent failures to replicate a given effect. Also, a few important negative results are published in their own journal articles, as discussed in appendix L. Scientists have recognized that this form of communication is sufficient to identify and correct false-positive errors. This is because scientists are devoted to the truth, so they don't want to believe false hypotheses. So the tide quickly turns against a positive result (false or not) if word gets around that nobody can replicate it.

The inevitability of some false-positive errors in published research results leads experienced researchers to consider new positive research results with polite skepticism until the results are properly replicated in independent research. Of course, other researchers in the field are quick to try to replicate any *important* new relationship between variables as they try to enhance knowledge about the phenomenon.

It is noteworthy that if a false-positive error occurs, and if the result is published, but the result is *unimportant*, then nobody may try to replicate the result. Therefore, the false-positive result will remain uncorrected in the research literature. This is unfortunate, but doesn't do much harm (because the result is unimportant).

Appendix C: Some Criticisms of the *p*-Value

This appendix discusses some noteworthy criticisms of the *p*-value. (The criticisms also sometimes apply to the other measures of the weight of evidence.) Let us first consider some criticisms proposed by McShane, Gal, Gelman, Robert, and Tackett (2018).

As noted in appendix B.12, it has been suggested that the null hypothesis may never be precisely true in nature. If the null hypothesis is never precisely true, then McShane et al. suggest that this implies that the null hypothesis is "implausible" (2018, p. 4). Therefore, perhaps hypothesis testing is illogical.

Expressing the same ideas at a more statistical level, if the null hypothesis is true, then the value of the relevant parameter of the model equation in the population is *exactly* equal to the null value, which implies an effect size of *exactly* zero. But (despite the lucky coin example in appendix B.12) some statisticians think (perhaps due to their sensible intuitions about variability of measured values) it is implausible that an effect size could ever be *exactly* zero in a population.

In considering these ideas, we can first note that the null value of a standard parameter lies in the middle of a highly plausible range of values. So even if it may be implausible that the effect size could ever be *exactly* zero, the null value lies in a plausible range. Thus, conceptually, the null value is equally as *plausible* as any other value in the range.

But, having established that, we can still ask whether the *precise* null hypothesis might ever be true in nature. And (despite the lucky coin example) it appears that the best we can answer is that it is *conceivable* that the null hypothesis might never be precisely true. This leads to the question: What happens to the *p*-value and hypothesis testing if the null hypothesis is never precisely true?

Perhaps surprisingly, it is irrelevant for the *p*-value and hypothesis testing whether the null hypothesis is ever precisely true. This is because we don't care whether it is ever precisely true because the null hypothesis is merely specifying a useful *hypothetical* situation. In many cases, the null hypothesis is false, but the effect is quite weak and therefore the effect is completely undetectable with our current measurement systems. In these cases the null hypothesis is *in effect* true, as discussed above in appendix B.3.

Thus even though the null hypothesis may never be precisely true, experience with negative results in scientific research suggests that the null hypothesis is still *in effect* true in a substantial proportion of cases. However, having established that point for clarity of concepts, it is noteworthy that we don't care about whether a given null hypothesis is either precisely true *or* in effect true—we only care whether the null hypothesis is clearly *false*.

If we can empirically show (e.g., with a low *p*-value and in the absence of a reasonable alternative explanation) that a null hypothesis is almost certainly *false*, then this implies that the effect under study almost certainly exists the population, which is a standard goal of research. If we can find good evidence of a new useful real effect, then this advances human knowledge. In doing that, it doesn't matter whether the null hypothesis is ever precisely true in a population.

Criticizing the *p*-value from a different direction, McShane et al. note that the parameter associated with the null hypothesis is assumed to have zero systematic error (2018, p. 4), which may also seem implausible. However, the null hypothesis is stating the hypothesized parameter value *for the population*, and the (unknown) parameter value in the population (whether it is the null value or not) has zero error—i.e., it doesn't vary. (Or, at least, it doesn't vary within the time frame of most scientific research, but see appendix I below.)

On another tack, McShane et al. (2018, p. 6) echo Rosnow and Rosenthal (1989), who say that dividing the scale of the value of the measure of the weight of evidence into two categories (i.e., "statistically significant" and "not statistically significant") has "no ontological basis". By this they (correctly) note that the values of most of the measures of the weight of evidence lie on a continuum of values. And in nature there is no critical value dividing the continuum for a measure into two categories. Since the two categories of the continuum don't exist in nature (except in the sense that they are invented by humans), it seems unnatural and thus inappropriate to these authors to use a critical value to break the continuum into two categories.

However, as noted in the body, it is often efficient to break a continuum into two categories. For example, you must be taller than four feet to be allowed on this ride. This criterion is quick and practical. Similarly, if the value of the measure of the weight of evidence falls beyond (or equals) the critical value, then (by convention, and in the absence of a reasonable alternative explanation), the evidence is strong enough to be (tentatively) believable. Therefore, if other important conditions are also satisfied, the evidence is strong enough to be published.

Coming from another direction, McShane et al. say that the use of *p*-values and the other measures of the weight of evidence encourages statisticians and researchers to use "dichotomous thinking" (2018, p. 7), which is clearly inappropriate. By this they mean that if a measure of the weight of evidence is on one side of (or equal to) the critical value, then we conclude that the effect under study exists. But if the measure of the weight of evidence is on the other side, then we conclude that the effect doesn't exist.

Applying dichotomous thinking to the existence of effects is inappropriate because in the case of a positive result, the statistical test may have made a false-positive error. Similarly, in the case of a negative result, the statistical test may have made a false-negative error. So our conclusions must be more tentative. So dichotomous thinking is highly inappropriate in deciding whether effects exist.

In contrast, dichotomous thinking is quite sensible with respect to the *publication* of a report of a research result in a scientific journal. If the value of the measure of the weight of evidence falls beyond (or equals) the critical value and if the other requirements for publication are adequately satisfied, then it is quite sensible to decide that a report of the research result is worth publishing, even though the result doesn't necessarily imply that the effect exists. Publishing the report gives other researchers in the field the opportunity to study the result and (if the effect exists) gives them the opportunity to successfully replicate it.

Berger and Berry (1988) criticize the *p*-value on the basis of its definition. They correctly note that the *p*-value tells us the fraction of the time that we will obtain a parameter estimate (or test statistic) of interest that is as discrepant *or more discrepant* from the null value as the result that was actually obtained in the research *if* the null hypothesis associated with the parameter is or were true and if we were to repeat the research project over and over (and if the assumptions underlying the *p*-value are adequately satisfied).

Berger and Berry focus on the idea of "more discrepant" and they suggest that we *aren't interested* in parameter estimates (or test statistics) that are *more* discrepant from the null value than the actual discrepancy of the parameter we have estimated in the research. And they suggest that we are only interested in the *obtained* estimated parameter value, and how discrepant *it* is from the null value. Therefore, they suggest that taking account of cases when the parameter estimate is *more* discrepant from the null value than the obtained estimate is illogical, and therefore the *p*-value is illogical.

However, by telling us the fraction of the time the parameter estimate will be as discrepant or *more* discrepant than the value at hand if the null hypothesis is true, the *p*-value gives us a reasonable measure of how discrepant our result is from the null value. This measure is reasonable because it is meaningfully comparable from one research project to the next.

The *p*-value is also reasonable because the critical *p*-value that the *p*-value is compared against is (if used consistently and if everything is done properly) the theoretical fraction of the time that we will make false-positive errors when the null hypothesis is true. Knowing and controlling this fraction is important because false-positive errors about important effects lead to a waste of resources. So the *p*-value is reasonable in both logical and practical senses.

In another criticism, Berger and Berry (1988) note that the *p*-value depends on the rule that we use for stopping the collection of data in a research project, which is called the "stopping rule". For example, we might specify that the stopping rule is: We will stop collecting data after 50 entities have been recruited for participation in the research project. (That is, the sample size is a fixed value according to the research design.) Or we might specify that the stopping rule is: We will stop collecting data after we have collected data for three weeks, regardless of how many entities are in the sample. (That is, the sample size is a "random variable" according to the research design.)

Berger and Berry discuss two research projects that are identical except that they have different stopping rules. And they consider the instructive case when the two research projects both obtain exactly the same data table. They observe that the *p*-values for the same hypothesis test from these two research projects will generally be somewhat different. They illustrate this point with a carefully (and sensibly) concocted example in which two research projects with identical data tables but with different stopping rules yield *p*-values of 0.049 and 0.085. So the result of one of the research projects is statistically significant at the 0.05 level, but the other result isn't statistically significant. This difference arises between the two *p*-values even though both analyses are based on exactly the same data table, and the only difference is in the two stopping rules.

As Berger and Berry note, we obtain different *p*-values under the two stopping rules due to the definition of the *p*-value. The *p*-value tells us the fraction of the time that a research project will obtain a parameter estimate that is as discrepant or more discrepant from the null value as the actual estimate at hand *if* the associated null hypothesis is or were true in the population and *if* we were to repeat the research project *over and over* (and if the underlying assumptions of the analysis are adequately satisfied). Under this definition, it is easy to show analytically that the correct *p*-value for a given effect in a data table will generally depend on the stopping rule that we use in the research project.

The fact that the two research projects yield different *p*-values for the same data (depending on the stopping rule) isn't a contradiction and isn't a shortcoming of the *p*-value. Instead, the fact is merely a mathematical consequence of the *p*-value definition, a consequence especially of the idea of repeating the research project over and over. This consequence is quite reasonable, as we will see momentarily.

Berger and Berry note that a standard Bayesian analysis of a data table *doesn't* depend on the stopping rule, which they

suggest makes the Bayesian analysis more sensible. They argue that a given data table should offer the *same* evidence in support of rejecting the null hypothesis regardless of the rule that was used to stop the data collection. In other words, they imply that *all* the relevant evidence is in the data table, and there is no relevant information in how the table was obtained.

To help us to consider this issue, we must consider our goal. This paper assumes that the goal in hypothesis testing is to find real (i.e., reproducible) effects in populations. And an important secondary goal is to control the rate of false-positive errors (because these errors are costly).

If it is important to control the rate of false-positive errors, then it is sensible to use the *p*-value to detect effects because the *p*-value is directly relevant to controlling the false-positive error rate. That is, as noted, if we consistently use the same critical *p*-value in a set of research projects, and if we do everything properly, then the critical *p*-value is the fraction of the time that we will make a false-positive error across the research projects in cases when the null hypothesis is true (and if the underlying assumptions of the *p*-value are adequately satisfied in each case).

In contrast, a researcher or statistician might believe that it is more important to satisfy some *other* goal. In that case, it might be sensible to omit the use of *p*-values and instead use some other statistical approach that will help to satisfy the goal, perhaps a Bayesian approach. But, arguably, there aren't any general goals in scientific research that are more important than finding interesting real effects while controlling the false-positive error rate (while trying to simultaneously [a] maximize the power of the tests, [b] obtain the model equation that makes the best predictions, and [c] minimize the research costs).

Berger and Berry object to the *p*-value depending on the stopping rule because they think that the resulting *p*-value depends on the researcher's *intentions* (as the intentions are reflected in the stopping rule). That is, if the researcher intends or intended to stop the research project one way, then he or she will obtain one *p*-value, but if they intend or intended to stop it the other way, then they will obtain the other *p*-value, even though both *p*-values are computed from the same data. Berger and Berry think it is inappropriate that the *p*-value should depend on mere intentions in the researcher's mind.

But the *p*-value isn't about intentions. It merely reports a *fact*—the fraction of the time we will obtain a result as discrepant or more discrepant as the result at hand if the null hypothesis is or were true and if (using the stopping rule we used in the research) we were to repeat the research project over and over (and if the underlying assumptions of the analysis are adequately satisfied). This fact gives us a sensible measure of the weight of evidence that the underlying effect is real. And the probability scale used by the *p*-value helps us to control the rate of false-positive errors in scientific research.

So the fact that the *p*-value depends on the stopping rule is sensible (to enable us to control the false-positive error rate). And from the practical perspective of scientific research, the statistical likelihood principle (that all the relevant information about a parameter or effect is in the likelihood function) is correct, but a given data table can have essentially different likelihood functions for a given parameter, with the correct likelihood function depending on the stopping rule.

Mayo (2014) and her discussants give a penetrating technical discussion of the preceding and related ideas.

Which stopping rule should we use in scientific research? In general, we should use the rule that satisfies the constraints of the research project and that leads to the most powerful key statistical tests, where the power of statistical tests is discussed in section 4 of the body of this paper.

These ideas also apply to post hoc comparisons and multiple testing. For example, Dienes (2011) suggests that the Bayesian approach enables researchers to (a) ignore stopping rules, and (b) perform multiple post hoc comparisons or other forms of multiple testing *without* taking any frequentist consideration of the multiple testing. However, if we wish to control the rate of false-positive errors, then it is easy to show analytically that for sensible rigor we must take account of the stopping rule in the analysis, and we must make certain adjustments if we perform post hoc comparisons or if we perform other forms of multiple testing.

In summary, it is sensible to retain control of the false-positive error rate in scientific research because false-positive errors are costly. If it is used properly, the *p*-value helps researchers to control the false-positive error rate. Therefore, the *p*-value is sensible.

Criticizing the *p*-value from another direction, some statisticians point to the fact that the *p*-values from different research projects that address the same research question sometimes disagree with each other (Greenland, 2017, p. 640). That is, we might perform two research projects to study the same research question and one research project might yield a *p*-value of, say, 0.02, but the other research project might yield a *p*-value of, say, 0.08. So, if we are using a critical *p*-value of 0.05, then which of the two research projects should we believe? This situation suggests to some statisticians that using *p*-values is irrational.

However, another sensible way to view this situation is to conclude that the research results under consideration are equivocal. That is, the contradictory results *may* mean that the effect under study *isn't* real in the population, and the positive result in one of the two cases is merely a false-positive error. Or the results *may* mean that the effect *is* real in the population, but the effect is weak, and the negative result is a false-negative error. So, in such cases, if the effect under study is important enough, then an interested researcher should consider performing more powerful research to see if he or she can obtain better evidence that the effect is real.

So the example doesn't imply that using *p*-values is irrational. However, the example clearly illustrates how the use of the *p*-value (and each of the other measures for detecting relationships) sometimes leads to false-positive and false-negative errors.

Demidenko (2016) criticizes the *p*-value by correctly noting that we can (at least in theory) make any *p*-value in scientific research arbitrarily low by merely increasing the sample size. (This conclusion is based on the widely believed but unprovable premise that the null hypothesis is never *exactly* true in scientific research, as discussed above in appendix B.12.)

This point leads Demidenko to suggest that *p*-values aren't useful.

However, Demidenko's point, though possibly theoretically correct, doesn't reflect a practical issue. This is because researchers generally can't *afford* the enormous sample sizes that would be required in some cases to obtain arbitrarily low *p*-values. So, disappointingly, in scientific research we often obtain *high p*-values—*p*-values that are greater than 0.05.

If a properly computed *p*-value is less than (or equal to) the critical *p*-value, and in the absence of a reasonable alternative explanation, this enables us to tentatively conclude that data based on an affordable sample provide enough evidence that the observed effect is real in the underlying population. Without that (and without an equivalent procedure), in some cases we may deceive ourselves. Thus, contrary to Demidenko's point, *p*-values are useful because they help us to reliably determine (in the absence of a reasonable alternative explanation) if we have enough evidence that an effect is real (i.e., reproducible in the population).

Appendix D: Comparing Hypothesis Testing with Karl Popper's Idea of Falsification

Karl Popper suggested that a theory isn't a valid scientific theory unless it can be falsified (1980, 1989, 1992). He used this sensible principle to support the ideas that Freudian theory, Marxist theory, and astrology aren't scientific theories. That is, none of the three theories can be readily falsified. That is, careful thinkers have been unable to find aspects of these theories that can be readily tested with some form of objective test, with the possibility of falsifying the theory through the test. In contrast, any accepted *scientific* theory (e.g., the theory of relativity) can in theory easily be empirically falsified if certain research findings (pertaining to relationships between variables) are or were obtained.

The ideas about relationships between variables discussed in the present paper are consistent with Popper's falsification approach. That is, all theories (research hypotheses) about relationships between variables could be falsified by showing that the relationship of interest *doesn't* exist or by showing that the relationship exists, but goes in the "opposite direction" to what the theory predicts. That is, the theory might predict that there is an *increasing* relationship between two variables. In this case as one variable increases, the other variable also tends to increase. But empirical research might reveal that the relationship is a *decreasing* relationship—as one variable increases, the other variable tends to *decrease*.

However, the two forms of falsification discussed in the preceding paragraph occur only rarely in the study of relationships between variables. This is because (a) it is generally agreed that it is impossible to prove that a relationship between compatible variables *doesn't* exist, and (b) although effects that are opposite to what we expect occur occasionally, they are rare. And if we fail to find evidence that a research hypothesis is supported, then we almost always find that there is *no* good evidence of a relationship between the variables under study (as opposed to finding good evidence of *no* relationship or the *opposite* relationship). (The rareness of discovery of opposite relationships may arise because researchers generally think carefully about the relationships they study, which makes it less likely that they will find the opposite.)

The fact that clear falsifications of research hypotheses rarely happen might seem to suggest that Popper's falsifiability criterion is merely a theoretical criterion that isn't often used in practical scientific research. However, if we turn things around and focus on falsifying the *null* hypothesis, then this type of falsification happens often and is what many scientific researchers are trying to do. We wish to falsify the null hypothesis to support the idea that our research hypothesis is true.

If (and, arguably, only if) we can convincingly falsify or reject the null hypothesis, then we can tentatively conclude that a relationship between variables (or other effect) exists in the population and therefore the theory associated with the research hypothesis is supported.

Thus Popper's theory of falsification and the notion of statistical hypothesis testing discussed in this paper are consistent if we assume that the falsification is performed of the *null* hypothesis, as opposed to falsification of the *research* hypothesis. This is somewhat different from Popper's approach because he doesn't discuss falsifying a null hypothesis. But the two approaches are consistent. Arguably, science proceeds by carefully falsifying various null hypotheses, thereby providing (in the absence of a reasonable alternative explanation) good evidence that the associated research hypotheses are true.

Appendix E: The Optimal Critical Value for a Test Statistic

McShane, Gal, Gelman, Robert, and Tackett suggest that the critical *p*-value of 0.05 is "entirely arbitrary" (2018, p. 6). Therefore, they conclude that the use of the *p*-value is illogical. However, the *p*-value isn't illogical because there is always a theoretical optimal critical *p*-value. Similarly, there is a theoretical optimal critical value for the confidence interval, for the Bayes factor, and for the other measures of the weight of evidence that an effect is real. (The presumed optimal critical value is "built in" to the information criteria.)

As suggested in section 3 in the body, the optimal critical value for a measure of the weight of evidence that an effect is real in a given field of science is the critical value that, if used consistently across the field, maximizes the total social, theoretical, or commercial long-term *benefits* or *payoff* of all research performed in the field given the available research resources. This criterion for the optimal critical value is sensible because, arguably, no other requirement for scientific research is more important than maximizing the long-term benefits of the research across a field of science.

If we use the preceding view, then the optimal critical value for a test statistic will be different in different fields of science because different fields have differing values of the relevant attributes that determine the maximal benefits. These attributes include (a) the rate of study (in good but incorrect faith) of false research hypotheses in the field, (b) the average payoff of positive results in the field when such results are

obtained, (c) the average costs of false-positive and false-negative errors in the field, and (d) various other attributes, such as research costs in the field.

Unfortunately, although the preceding ideas are sensible, it appears that we can't reasonably measure the attributes (a), (b), (c), or (d) in any field of science in a practical sense. Therefore, it is apparently impossible to use cost-benefit principles to directly determine the optimal critical value for a test statistic in a field of science. However, it is useful to be aware of these ideas because they provide a sensible conceptual definition of the optimal critical value for a test statistic.

The fact that we can't know the optimal critical value for a test statistic in a field of science has led to the choice of *general* critical values based on consensus among experienced researchers. These general critical values seem reasonable in the sense that they give us a reasonable balance of (a) positive results, (b) false-positive errors, (c) negative results, (d) false-negative errors, and (e) research costs. And these critical values give researchers a level playing field—a consistent criterion that we can use in all standard scientific research to determine whether research results are (in the absence of a reasonable alternative explanation) believable.

Also, by giving us a scale, the critical-value approach enables researchers who think that the conventional critical value is unreasonable to choose their own critical value. For example, if a researcher thinks that the critical *p*-value of 0.05 is too lenient (i.e., allows too many false-positive errors to occur), then they can opt to use a critical *p*-value value of, say, 0.01 in their research or in their interpretation of other researchers' research.

Researchers prefer low false-*negative* error rates (i.e., they prefer critical values that *aren't* strict) because this makes it easier and less costly for their research to obtain statistical significance and thereby (if everything else is satisfactory) be published. But journal editors prefer low false-*positive* error rates (i.e., they prefer critical values that *are* strict) because this reduces the rate of publication of misleading false-positive errors in the research literature.

Journal editors are the final arbiters of the critical value for a test statistic in the sense that a key hurdle for any standard scientific research project is to have a report of the results accepted for *consideration* for publication in a journal. This is the first step toward being accepted for *publication* in the journal. As discussed in section 3 in the body, most journals that are statistically oriented will indicate that a paper will only be considered for publication in the journal if the value of the relevant measure of the weight of evidence falls beyond (or is equal to) the journal's critical value. This standard saves time by eliminating quibbling about whether a research result is convincing enough to deserve attention.

Some statisticians recommend lower conventional critical *p*-values (Johnson, 2013; Bayarri, Benjamin, Berger, and Sellke 2016; Johnson, Payne, Wang, Asher, and Mandal 2017). This recommendation is based on the perception that "too many" false-positive results are being published in the scientific research literature.

Benjamin, Berger, …, and Johnson (72 authors, 2017) recommend that a critical *p*-value of 0.005 be used for "claims of new discoveries". However, interestingly, these authors distinguish their recommended critical value from the critical *p*-value that is used as a screening rule for publication. And in their "Concluding remarks" section they "emphasize" that journals can continue to use a critical *p*-value of 0.05 (or lower, at each journal's discretion) as a screening rule to determine whether the results in a paper provide enough weight of evidence to consider the paper for publication. This screening role is, arguably, the key role of the *p*-value in scientific research—to decide whether the evidence of an effect is good enough to warrant consideration for publication in a scientific journal.

It is noteworthy that researchers generally don't use critical values to *decide* whether a "new discovery" has been made. This is because, as noted in section 4 in the body, *p*-values can't make decisions because false-positive errors can always occur. Instead, the research community decides whether a "new discovery" has been made based on many factors, of which a low *p*-value is often an important consideration.

Lakens et al. (88 authors, 2018) rebut the Benjamin, Berger, …, and Johnson (2017) article.

It *may* be true that "too many" false-positive results are being published in the research literature. If so, then it is mandatory to use stricter critical values in statistical hypothesis tests. This will lead to fewer false-positive results in the literature (though it will also increase the cost of scientific research if we wish to maintain equivalent statistical power in hypothesis tests).

Unfortunately, it is difficult or impossible to determine objectively whether "too many" false-positive results are being published in the research literature. This is because, as noted at the beginning of this appendix, it is difficult or impossible to evaluate "too many" objectively.

It seems likely that (due to a form of natural selection by the crowd) the standard critical *p*-values of 0.05 and 0.01 are near-optimal for most scientific research in the sense that they lead to the greatest overall scientific payoff—a good proportion of published true positive results together with an acceptable mix of (published) false-positive errors and (unpublished) false-negative errors. Of course, the important errors will be discovered and corrected in later research.

Appendix F: Teaching *p*-Value Concepts to Beginners

It is (arguably) necessary to understand *p*-values to fully understand the principles of scientific research. But, as noted, *p*-values are somewhat hard to understand. Fortunately, a proper understanding of *p*-values is *guaranteed* if a student studies enough realistic examples of scientific research projects that study relationships between variables when the null hypothesis is and isn't rejected. The concept of "enough examples" depends on (a) the student's initial level of understanding, (b) the student's ability, and (c) the quality of the examples.

It is also important for students to study examples of reasonable alternative explanations and examples of false-positive and false-negative errors. There must be enough examples to ensure that students will never make the common and

tempting error of assuming that *p*-values make decisions. Examples work best if they are *practical* in the sense that there is an easily recognized meaningful social, theoretical, or commercial payoff if the studied relationship between variables is found to be real.

Students must be aware that a certain (unknown, but hopefully low) proportion of the positive results in the scientific literature are false-positive errors.

Students must learn the fundamental distinction in scientific research between an observational research project (in which the predictor variables are merely *observed* in the entities in the sample) and an experiment (in which one or more of the predictor variables are *manipulated* in the entities). In general, we must perform an experiment if we wish to determine whether a *causal* relationship exists between variables, though there are occasional exceptions.

It is important to distinguish between (a) using a *p*-value to detect a relationship between variables and (b) studying a newly detected relationship between variables through developing a model equation and with graphs. Detecting relationships between variables comes first, and we can use *p*-values to help with this detection. However, beginning students needn't understand the mathematical details of how *p*-values are computed because these details aren't necessary to understand the *function* of the *p*-value in helping us to detect relationships between variables.

Similarly, beginning students needn't understand the details of how to derive a model equation from a data table because that is arguably too complicated for most introductory courses. Instead, students should understand that straightforward methods are available to derive equations, as described in statistics textbooks.

But students should know what the various forms of derived model equations look like (with the parameter values expressed as numbers, not as abstruse algebraic symbols). And students should understand that if you substitute the value(s) of the predictor variable(s) for a new entity from the studied population into a properly derived model equation, and if you then evaluate the resulting expression, then you get a sensible predicted value of the response variable for the entity.

A good graph is a key to proper understanding of a relationship between variables because a good graph shows the relationship at a glance. So students must learn how to understand graphs. By convention, the response variable is always plotted on the vertical axis of a graph because this makes it easier for the viewer to orient to the graph. Beginners often inadvertently omit axis labels from graphs, which substantially reduces the ability of a graph to communicate.

For the main examples in a statistics course, it is helpful to begin by explaining the goal of the research to the students. The goal must be explained in terms that enable students to appreciate the usefulness of the study of the relationship. Then the teacher can present students with well-formatted computer output showing an organized complete analysis of a relevant data table. The first part of the output should show the first five or so rows of the properly rounded raw data. The columns of the data should be clearly labelled (with carefully chosen variable labels, typically multiple words, perhaps including the units of measurement) to assist understanding. These illustrative data rows, together with the count of the number of rows in the full table, enable students to understand the nature of the data being analyzed.

Next the output can show the results of the analysis of the data, showing descriptive statistics, test statistics, *p*-values, and possibly other measures of the weight of evidence that an effect is real. The output should include carefully chosen graphs to illustrate any relationships between variables that are (apparently) discovered. The teacher can explain to the students what each item in the output tells us, explaining that the software includes many items for thoroughness, but some items are less important than others.

After students have learned the basic ideas about the study of relationships between variables in scientific research, it is recommended that they be assigned to groups and then each group be asked to design a research project in their area of interest. (The students won't be asked to *perform* the research project because that would be too complicated and too time-consuming for beginning students.) They should first choose a response variable they would like to learn to predict or control. The teacher should encourage the students to use *continuous* response variables if possible because these variables contain more information in their values than *discrete* response variables. And analysis procedures for continuous response variables are generally better known than analysis procedures for discrete response variables.

Next, the students can choose one or more predictor variables that the response variable might be related to. (Predictor variables may be continuous or discrete.) Then the students can decide whether they must merely observe the predictor variable(s) in an observational research project or whether they can manipulate some or all of the predictor variable(s) in an experiment.

Next, the students can design an observational research project or an experiment to study the relationship of interest. The students can specify how the entities in the sample will be selected from the population and how the response and predictor variables will be measured in the entities. (Some predictor variables may not be measured directly in the entities, but are measured in the entities' environment.) Students can specify the detailed steps to perform the research project, discussing the expected outcome, and discussing possible alternative explanations for any results they might obtain. Students can present their research designs to the class. Then the teacher and the other students can constructively criticize the designs, helping students to see how scientific logic works.

Next, the teacher can retire to his or her office and use a computer to generate (cook up) some data for each group. Generating data (with random number generators) is relatively easy with modern statistical software, especially if you clone one of the many example programs for generating data on the web. Search for "generate data with [the name of your statistical software]".

The teacher should generate the data so that it yields a plausible "nice" set of results. Perhaps typically the results should show modestly (not highly) statistically significant evidence that the central effect under study is real. This implies

that the teacher must adjust the data-generation procedure until it gives nice analysis results. This is easy to do by modifying the model equation in the data generation program. You first set the values of the parameters of the data-generation equation to sensible (null or non-null) values. Then you adjust the overall standard error of the equation to give a sensible main *p*-value that is slightly less than 0.05. Of course, the smaller the standard error, the smaller the *p*-value. This may require 20 or 30 or more different generations of the data to get everything right, but that goes quickly after you have done it a few times. Use a fixed "seed" for the random number generator(s) to give reproducible results.

After cooking up appropriate data, the teacher can perform a formal full analysis of the data to look for evidence of the sought-after relationship(s) between the variables. The teacher should use careful variable and value labelling in the output with appropriate references to the students' own terminology, which will help the students to relate to the analysis. After completing the analysis, the teacher can present the computer output from the analysis to the students, as if the students had obtained the data through performing the research. Then the teacher and students can discuss the interpretation of the computer output.

After considering the computer output, each group of students can write a report summarizing the conduct of the research and summarizing the results. Students must understand that the report must explain all the "relevant" details about how the research was done because this makes it easier for other researchers to successfully replicate the results, which is in everyone's interest. Many less experienced researchers omit relevant information from their research reports, thereby leading to difficulties in replicating the results by other researchers, which leads the other researchers to wonder if the results reflect a false-positive error.

A challenge in the preceding approach is that if the students have free range in designing their research projects, then the teacher may find a surprising array of statistical analysis challenges. And not every teacher is academically equipped to handle all these challenges (because there is a broad set of methods for data analysis). However, this problem can be avoided if the teacher ensures that the students design their research projects in ways that he or she can comfortably analyze.

It is recommended that most courses for beginners *not* include any data-analysis computer programming. Of course, the programming is *conceptually* simple—we give the data and some simple instructions to the computer, and the computer analyzes the data and generates the relevant output. What could be easier than that?

However, from a practical point of view, programming for data analysis is surprisingly complicated. This is because there are many minor but necessary details in software installation, syntax, options, and data management, all of which must be correctly handled before a program will work properly. Thus in a course for beginners that includes computer programming, the multitude of programming details tend to become the center of attention as students strive to master them. But understanding the *scientific research* ideas and understanding the computer *output* is much more important than understanding the programming details, which are important, but aren't central, and which can come in later courses for students who wish to learn more about scientific research.

Appendix O below discusses an approach to understanding the reality of the ideas discussed in this paper, which may be of interest to beginners or others who are curious about the philosophy behind the ideas.

Appendix G: A Case When We Don't Need a Measure of the Weight of Evidence

The earlier discussion suggests that we generally need a measure of the weight of evidence that an effect discovered in scientific research is real in the population of entities of interest. This helps us to avoid deceiving ourselves. This appendix discusses an instructive important exception to that point.

For this discussion it is useful to split the study of relationships between variables into two cases—the case in which the response variable is a *continuous* variable and the case in which the response variable is a *discrete* variable. A variable is a "continuous" variable if it can (at least in theory) have any value within some continuous range of values (where the range is usually a *numeric* range, though it needn't be). A large percentage of scientific research projects that study relationships between variables have a continuous response variable.

If a variable isn't a continuous variable, then it is a "discrete" variable, having a (usually finite) set of discrete possible values. Some discrete variables have only two possible values, such as "yes" and "no", or "true" and "false", or "one" and "zero". Many other discrete variables have between three and ten or so possible values. But we will see in a moment that some discrete variables can have a very large number of possible values.

If the response variable in a scientific research project is a continuous variable, and if the model equation is a sum of terms (as is typical with continuous response variables), then an elementary statistical theorem shows that the variance of the sum of the terms in the equation is equal to the sum of the variances of the individual terms (plus double the sum of the unique covariances, if relevant). Therefore, adding terms for unnecessary predictor variables to an additive prediction equation always tends to increase the variance and therefore increases the standard error of the predicted values of the response variable. This increase in variance amounts to a decrease in the precision of the predictions made by the model equation.

(The increase in variance from adding a term will often be small and may be nullified or reversed if a real relationship between the added predictor variable[s] and the response variable exists. Negative covariances are generally too small to cancel out the extra variance from adding a term.)

Also, in the case of a continuous response variable, if we add unnecessary terms to a model equation, then the uncertainty of the parameter estimates for terms already in the equation is almost always increased, as proven in the case of linear regression analysis by Sen and Srivastava (1990, sec. 11.2.3).

This increase in uncertainty of parameter estimates is arguably undesirable, though it is also possible to argue that the uncertainty of the parameter estimates is less important, and it is the uncertainty in the *predictions* that is relevant.

Also, adding unnecessary predictor variables to a model equation violates the principle of parsimony. Also, adding unnecessary predictor variables to a model equation implies that we must *measure* the unnecessary variables whenever we wish to use the equation, and this measurement of unnecessary variables adds an unnecessary extra cost.

Therefore, adding *unnecessary* terms to a model equation with a continuous response variable is undesirable. Thus we wish to avoid including a predictor variable in a model equation with a continuous response variable unless we have good evidence that this variable is related to the response variable. Thus if the response variable is continuous, we need a measure of the weight of evidence that a term belongs in the equation to help us to determine whether we should include or exclude each available term for the equation.

(We can bypass the *direct* need for a measure of weight of evidence with a continuous response variable if we use an automatic procedure for selecting relevant predictor variables. However, in this case we are still using a measure of the weight of evidence of an effect, but a measure is built into the automatic procedure we are using, so the measure is present, but less visible.)

Consider now the case when the response variable is a discrete variable. If a discrete response variable has a small number of possible values, then the arguments above for omitting unnecessary terms generally still apply, and including unnecessary predictor variables in an equation will unnecessarily increase cost and complexity. But if a discrete response variable has many possible values (e.g., more than 100 or so values), then things change.

Consider the problem of computer pattern recognition, which is an important extreme case. This problem is easily viewed as the study of a relationship between variables in which the predictor variables are variables that describe an observed state of nature as recorded in a data table, and the response variable is some form of a proper "description" of the pattern observed in data, which is also recorded in the data table when we are first deriving the model equation. For example, in handwriting recognition, the values of the predictor variables are values that describe an image of handwriting, and the value of the response variable is a character string that is a digital representation of the text in the handwriting. Handwriting recognition software uses an internal representation of the relationship between the variables to predict the digital character string from a handwriting image.

Similarly, in general visual image recognition, the predictor variables are a set of variables describing an image (typically, the color and intensity of each pixel in the image) and the response variable is a plain-language description of the image, such as "a woman throwing a Frisbee in a park" (LeCun, Bengio, and Hinton, 2015). Similarly, in speech recognition the predictor variables are a set of variables describing the time-varying pitch and intensity of the sounds of spoken words that are received by the system's microphone, and the response variable is a character string of the text for the words that the system "heard".

Pattern recognition problems often *aren't* viewed as studying relationships between variables. But these problems can be readily viewed as studying relationships by viewing the inputs to such systems as the (rather complicated) values of predictor variables and by viewing the output as the value of a discrete response variable (with many possible values). Arguably, viewing pattern-recognition as a type of study of relationships between variables helps to increase understanding.

Modern pattern-recognition software systems are surprisingly practical in the sense that some such systems are now more efficient for many users than traditional systems. For example, many users accept and use speech-recognition systems for text entry and for command entry in electronic devices and computers. Users have found that using speech-recognition software to enter text and commands is significantly more convenient than typing the information on a keyboard, even for fast typists with good keyboards.

As noted, in the case of continuous response variables, we usually derive a model equation in which we omit irrelevant predictor variables. But if we examine modern pattern-recognition software (e.g., neural network software) in which a model equation for the relationship between the variables is developed by the software, the software typically makes no direct attempt to identify and omit "irrelevant" predictor variables from the broad set of predictor variables it is allowed to use. This is because a predictor variable that is irrelevant in one situation may be highly relevant in another.

Generally, we never see the internal model equation in pattern-recognition software. This is because the equation is developed inside the computer by the software and is typically a highly complicated and difficult-to-interpret *network* of equations that have been "naturally" selected through "training" of the software with many earlier instances (samples) of the various types of patterns under study. Thus the precise nature of the relationship between the variables is obscure. However, if we study the low-level details of the software, we see that the response variable is mathematically connected to the predictor variables by a large complicated network of mathematical relationships (equations) that the software has derived.

The relationships between variables in pattern-recognition software may mathematically emulate the complicated electrochemical relationships (networks) that occur at a low level between the neurons of a biological living brain.

Thus, at a high level, pattern-recognition software works in the sense that it merely observes certain regularities in the data it was trained with and it uses these regularities to develop in a complicated internal model equation or rule to predict the values of the response variable in new entities from the population of entities (e.g., images or utterances) that it is designed to interpret.

For the present discussion, the main conclusion is that we can view pattern-recognition systems as studying relationships between (a) a set of predictor variables and (b) a discrete response variable with many possible values. And pattern-recognition systems generally take account of *all* the available predictor variables, making no attempt to determine whether

certain of them are irrelevant in predicting the values of the response variable. Thus pattern-recognition systems generally don't use or need a measure of the weight of evidence that a relationship exists between variables.

In view of the preceding points, methods for determining the weight of evidence of the existence of relationships between (a) one or more predictor variables and (b) a response variable are clearly often useful, especially with continuous response variables. But such methods are unnecessary in some cases, such as the case with a discrete response variable when this variable has many possible values, as in pattern-recognition problems.

Appendix H: Some Theoretical Arguments About the Preferred Measure

Section 5 in the body of this paper discusses nine sensible measures of the weight of evidence in support of a research hypothesis—in support of the hypothesis that an effect observed in scientific research is real in the underlying population. The discussion concludes that the *p*-value is slightly superior to the other measures in various senses, as summarized in table 1 in the body.

However, despite the advantages of the *p*-value, we can still ask whether one of the measures might in some sense be *theoretically* more or less correct than the others. That is, does one of the measures gives us the "true" (or a false) measure of the weight of evidence in favor of the research hypothesis? This appendix evaluates four arguments why one of the measures might be theoretically superior or inferior to the others.

First, it could be argued that the measure of weight of evidence that is most nearly *linearly* related to a standard measure of the effect size in the region of the critical value is the true measure. But there are generally various available measures of the effect size in a given research situation, and these measures generally aren't linearly related to each other as the effect size changes. Therefore, we would need to choose one of the measures of effect size and say that it is the "true" measure of effect size before we could use the linearity argument to choose the best measure of the weight of evidence that an effect is real. But choosing one of the measures of effect size as the "true" measure seems somewhat arbitrary. So, if we seek objectivity, this first argument is ruled out.

Consider a second argument why one of the measures of the weight of evidence that an effect is real is superior to the others: It could be argued that one of the measures is more "natural" than the others. In fact, many statisticians have strong opinions about which of the measures of the weight of evidence is most "natural", though opinions vary.

The idea of appealing to the "naturalness" of the measure of the weight of evidence is sensible *if* we have a reliable measure of "naturalness". Unfortunately, we don't appear to have such a measure, so we must fall back on intuitions. But intuitions are unreliable. So, again, if we seek objectivity, it seems difficult to appeal to the concept of "naturalness" in choosing the best measure of the weight of evidence that an effect is real.

Consider a third argument why one of the measures of the weight of evidence that an effect is real is superior to the others: Suppose that we choose a measure of the weight of evidence that an effect is real. For example, let us choose the Bayes factor. And suppose we choose a sensible critical value for the measure, say, 16.

Suppose that we calibrate all the other measures to have critical values that correspond to a Bayes factor of 16 in a specified research situation. That is, when the Bayes factor (by being greater than or equal to 16) declares that an effect is statistically significant in this situation, then we give all the other measures critical values so that they will also declare that the effect is statistically significant.

Next, suppose that we go to *another* research situation (e.g., the same research situation but with a different sample size). Then, if we use the same calibrated critical values, we will find that some of the other measures will in some borderline cases disagree with the Bayes factor about whether there is enough evidence to reject the null hypothesis.

This phenomenon is illustrated in the case of the *p*-value and Bayesian approaches in three journal articles. First, Kass and Raftery (1995, sec. 8.2) show that for a given Bayes factor, we would under the standard *p*-value approach need to use a *different* critical *p*-value to reach the same conclusion, depending on the sample size.

Similarly, Wagenmakers (2007, pp. 792–794) discusses the relationship between the sample size and the posterior probability that the null hypothesis is true. His figure 6 shows that for a research result that just obtains statistical significance (i.e., the *p*-value is exactly equal to 0.05), the posterior "probability" that the null hypothesis is true depends on the sample size.

Held and Ott (2016) illustrate the phenomenon using minimum Bayes factors. They illustrate that the same *p*-value corresponds to a different minimum Bayes factors depending on the sample size.

The preceding three examples show how there are inconsistencies between (a) the *p*-value and (b) either the Bayes factor or the posterior "probability" that the null hypothesis is true. Thus if we assume that the Bayes factor provides a "correct" measure of the weight of evidence, then corresponding critical *p*-values will vary with the sample size. Therefore, *p*-values are inconsistent with the "correct" measure, and thus the *p*-value is an "incorrect" measure of the weight of evidence.

But, we can readily reverse things. And if we assume that the *p*-value is the "correct" measure of weight of evidence, then the Bayesian methods for computing the weight of evidence are inconsistent with the *p*-value, and thus the Bayesian methods are "incorrect".

Thus there are (smooth) inconsistencies between some of the measures of the weight of evidence pertaining to critical values if we move from one research situation to another (such as by changing the sample size). This is because the (monotonic) relationships between some of the measures of weight of evidence aren't linear (as illustrated by Spiegelhalter, Abrams, and Myles, 2004, p. 132) and due to the Jeffreys-Lindley paradox (which is discussed in appendix A). Thus in a new research situation one measure may cross the critical-

value boundary ahead of another as the sample size (or some other relevant attribute of the research situation) changes.

The inconsistencies between the Bayesian methods and the *p*-value are often scientifically inconsequential. But they raise puzzling scientific questions with larger samples due to the Jeffreys-Lindley paradox. Efron and Hastie give an instructive example of the inconsistencies in varying sample sizes between a function of the *p*-value and a form of the Bayes factor (2016, table 13.5). In examples like this, even if an effect is small, if it is *real*, then researchers invariably want to know about it. Of course, a small real effect may not be directly useful. But we still want to know about it because it might lead us to a way to *make* it useful.

The key point in this discussion of the third argument is that the existence of the inconsistencies doesn't somehow imply that one of the methods is the true method and therefore the other methods are inferior (because they are slightly inconsistent with the "true" method).

Consider a fourth argument why one of the measures of the weight of evidence that an effect is real is superior to the others: Some researchers say that the Bayes factor is preferred to the *p*-value because the conventional critical value for the Bayes factor is stricter than the conventional critical values for the *p*-value (Ioannidis 2008; Wetzels et al. 2011; Bayarri, Benjamin, Berger, and Sellke 2016). They recommend using the Bayes factor because the stricter conventional critical value for Bayes factors make it less likely that the Bayesian approach will make false-positive errors. (But, as discussed in section 5.5 in the body of this paper, the stricter critical value makes it *more* likely that the Bayesian approach will make false-*negative* errors.) Therefore, in view of the "replication crisis" in scientific research, these researchers suggest that we should use the Bayes factor because (if we use it with a conventional critical value) it will lead us to make fewer false-positive errors.

However, if a researcher or an editor wishes to reduce the rate of false-positive errors in research, then he or she needn't switch to using Bayes factors. Instead, they can simply use a stricter critical value for the measure of weight of evidence that they are already using. For example, if a researcher or editor is using the *p*-value as a measure of the weight of evidence, and if they are using a critical *p*-value of 0.01, and if they wish to use a stricter test, then they can switch to using a lower critical *p*-value, such as 0.005 or 0.001. (But, unfortunately, this will necessarily increase the rate of false-negative errors or it will necessarily increase research costs, exactly as switching to the Bayes factor with conventional critical values would do.)

Appendix E above discusses the idea of the optimal critical value for a test statistic.

In summary, this appendix has discussed four theoretical arguments why one of the measures of the weight of evidence might be superior or inferior to the others. But in each case, a sensible rebuttal is proposed.

Appendix I: Should We Allow the True Values of Parameters of Model Equations to Vary?

Recall the general regression model equation discussed above in appendix B.4:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_qx_q + \varepsilon \qquad (1)$$

The x's in the equation are the q predictor variables and the b's are the $q + 1$ parameters. Virtually all model equations have zero or (usually) more predictor variables and one or more parameters.

As noted in appendix B.5, we usually view the true values of parameters of model equations as being *fixed* (i.e., constant) numbers in the population. (Of course, the *estimates* of the values in a given equation generally vary slightly from one research project to the next.) The idea that parameters have fixed values in the population is especially evident in the physical sciences, as discussed below in appendix J.

However, it is also possible and sometimes sensible to view the *true* values of parameters or effects of a model equation as themselves varying "slowly" over time. But in that case, we generally view the parameters as being fixed within the time frame of reference under study.

In contrast, some statisticians suggest that we should allow the *true* values of the parameters of a model equation to *vary* instead of assuming that they have constant fixed values. For example, Gelman recommends that we move "beyond the worldview in which effects are constant …" (2015, p. 633). Although Gelman uses the word "effects", it appears that he means what the present paper refers to as "parameters". (The values of the parameters define the effects.) This suggests that a modern approach to data analysis would allow the *true* values of the parameters of a model equation of a relationship between variables to vary from one research project to the next.

Although the approach with varying true parameter values is more complicated, the idea seems sensible in a given situation *if* we can show that the approach is useful. For example, if we can show that if we allow parameter values to vary, then this enables model equations to make better predictions than if we use fixed parameter values, then clearly the approach would be sensible.

However, if we allow the true values of the parameters of a model equation to vary, then we can *model* the variation in the values of a parameter with a "second-level" model equation. That is, any parameter with varying values can be the response variable in a second-level equation. And whatever causes or is related to the variation in this new response variable can be the *predictor* variable(s) in the second-level equation. And this second-level equation will itself almost certainly have parameters with *fixed* true values.

(If the second-level model equation *also* has parameters with varying values, then we can use a third-level model equation with fixed parameters to model the variation in the values of the parameters of the second-level equation, and so on. And, presumably, though not necessarily, we would encounter fixed parameter values at some point in the sequence of model equations.)

However, if we have a second-level model equation that models the variation in the values of a parameter, then we can

substitute the right-hand side of the second-level (or the right-hand side of a yet higher-level) model equation into the original model equation in place of the associated parameter. (In this substitution operation we omit the *error* term associated with the second-level equation, leaving the prediction errors in the first-level model equation to be modelled by the error term in *that* equation.) This will generate a new version of the original equation, except that all the parameters in the new equation associated with the term with the varying parameter will now have *fixed* true values in the population.

Thus though having parameter values that vary in the population is theoretically permissible, we can (at least in theory) often convert varying parameters to fixed parameters by replacing them with a more complicated set of terms (with fixed parameter values). And, importantly, this approach will lead to model equations that make better predictions. Thus arguably we don't need to develop statistical procedures to *directly* handle varying parameter values (unless someone convincingly shows that the approach with varying parameter values is somehow more efficient than trying to *model* the varying parameter values).

The approach described in the preceding paragraphs won't work if the variation in the values of a varying parameter is "truly" random variation that depends on no other variables. This is because if a parameter is truly random, then there will be no model equation that can account for the parameter's varying values. However, in this case, it is arguably sensible to conceptually move the variation out of the *parameter* and into the error term of the original model equation, and to let the parameter value itself be the mean (or some other sensible measure of central tendency) of the varying distribution. This is sensible because it collects all the random variation together in the error term, which makes things simpler when we wish to use the model equation to predict or control the values of the response variable. Of course, if we can *demonstrate* that some of the variation somehow rightfully *belongs* in the parameter itself, as opposed to belonging in the error term, then this variation is arguably best left in the parameter.

Thus it seems sensible to view the true values of the parameters of a model equation as being *fixed* values in the population, with the provision that a parameter may have varying values if a significant advantage of that can be clearly demonstrated.

Appendix J: A Case When We Know the Exact Values of Parameters

As noted, researchers usually view the values of the parameters of a model equation as being *fixed* numeric values in the population that are constant from one instance of a research project to the next. But if we perform scientific research, we are only able to obtain *estimates* of the *true* values, and the estimates will vary somewhat from one instance of a research project to the next.

The view that the true parameter values are fixed in the population (i.e., in nature) is highly evident in the physical sciences where researchers study the fundamental physical constants, such as the gravitational constant, the molar gas constant, and the Planck constant (Mohr, Newell, and Taylor 2016). These constants can all be readily viewed as *parameters* of model equations of relationships between variables. Physical scientists view these constants as being fixed (unvarying) over time and (they presume) throughout the universe, as implied by the name "constants". Physical scientists have performed various careful research projects to estimate the correct values of these parameters.

But in an interesting reversal, at the most basic level of the physical sciences, the true values of certain parameters *aren't* estimated from data, but are instead specified by human fiat. Then various concepts are defined in terms of these specified-by-fiat values (Mohr, Newell, and Taylor, 2016, sec. II).

For example, in Einstein's model equation, $E = mc^2$, the E is the amount of energy in a piece of matter and the m is the mass of the piece of matter. We can use this equation to determine the amount of energy in a piece of matter if we know its mass. And we can likewise use the equation to determine the mass of a piece of matter if we know its energy.

The c^2 in Einstein's equation is the parameter of the equation. Einstein has shown that the value of this parameter is equal to the square of the speed of light in a vacuum.

(Einstein's equation is astonishing because we ask what does the speed of light have to with mass or energy? And how could the square of the speed of light be the exact correct value for the parameter for this model equation? Of course, Einstein has answered these questions to physicists' complete satisfaction.)

The speed of light in a vacuum, c, is a special type of parameter because (since 1983) its value has been specified by human fiat (based on earlier estimates and based on almost universal agreement among physical scientists). The value is specified to be *exactly* 299,792,458 meters per second (BIPM, 2006). Physical scientists specify the speed of light in a vacuum by fiat because this effectively and exactly defines the standard unit of length, the meter. That is, the meter is defined to be exactly 1/299,792,458 of the distance that light will travel in a vacuum in one second. So, instead of defining the unit of length and then determining the speed of light in terms of that unit, physical scientists specify the speed of light, and then they define the unit of length in terms of that speed.

The definition of the meter refers to the measurement of time, specifically the measurement of one second of time. Thus the definition of the meter requires that we have a good definition of the unit of time, the second, which is now also exactly specified (BIPM, 2006).

Physical scientists chose to define the unit of length in terms of the speed of light because they believe it is sensible to view the speed of light in a vacuum as being constant in nature (i.e., constant in all instances in the population of cases when light travels in a vacuum). Thus this constant value is a reasonable foundation for other physical constants—constants that must be estimated from data. Some other parameter values that are now or will likely soon be specified exactly are the Planck constant, the Boltzmann constant, and the Avogadro constant.

We can see the difference between the *estimated* parameter values in the physical sciences and the parameter values specified by fiat by noting that all estimated parameter values have (perhaps behind the scenes) an associated estimate of

their precision or uncertainty. For example, the key article specifying the currently accepted values of the more than 300 fundamental physical constants reflects the fact that almost all the constants have been *estimated* from appropriate research data, and thus each of these constants has an associated uncertainty, which is shown in the "Relative standard uncertainty" column in most of the tables in the article (Mohr, Newell, and Taylor, 2016). But a few of the fundamental constants have *exact* fixed values, such as the speed of light in a vacuum and the molar mass of carbon 12. These constants have no associated estimate of their uncertainty, as illustrated in table I in the Mohr, Newell, and Taylor article.

Physical scientists specify the values of a small number of basic parameters and measurement units by fiat because they have decided that this is the most efficient way to develop variables and measurement of variables in the physical sciences. Physical scientists have chosen the *particular* set of parameters and measurement units to be specified by fiat because these parameters and measurement units are viewed as an easy-to-understand, relatively easy-to-use, and (hopefully) unshakable foundation on which measurements in the physical sciences can rest.

The method of specifying certain parameter values and measurement units in physical science by fiat is closely akin to specifying a small set of axioms in a logical or mathematical system and then deriving a set of propositions from the axioms. The method is also closely akin to specifying a "basis set" of vectors for a subspace of a vector space in linear algebra. Multiple basis sets are possible for a given vector subspace, just as it would be possible to choose different sets of parameters to be the basis of physical science.

Gauss appears to have been the first physical scientist to specify a parameter value by fiat. As discussed by Roche (1998), Gauss was the first scientist to put Newton's second law of motion in modern form as $F = ma$. Here, F stands for the net force exerted on a physical object, m stands for the mass of the object, and a stands for the resulting acceleration of the object due to the force. Since the acceleration is typically the response variable in this relationship between variables, the law is also sensibly specified as $a = F/m$.

Gauss in effect specified that the parameter of this equation has the value one (1.0), thereby implicitly specifying a definition of the units of force. Here, by specifying that the value of the parameter of the model equation for Newton's second law is the numeral one, Gauss wasn't defining the *concept* of force—he was merely defining the *units* of force.

Interestingly, Gauss' decision to set the parameter of Newton's second law to the numeral one has ever since confused many physics students who (despite conventional explanations) are still puzzled why the law appears to have no parameter. They are puzzled because they know intuitively that in the real world usually things don't come out as perfectly as the model equation suggests, and there is always a parameter (coefficient of proportionality) to make the units conform. Of course, the parameter is present (as a multiplier) in Gauss' expression of Newton's second law, but the value of the parameter is (by Gauss' fiat) 1.0, so the parameter is invisible.

Gauss would likely interpret this matter in different terms because the concept of a parameter of a model equation of a relationship between variables wasn't as clear in his time as it is now.

Appendix K: Approaches to Publishing Negative Results

As noted in appendix B.9, most scientific journals won't accept a report of a research project if the main result is a negative result. However, some researchers sensibly believe that individual negative results should be published because these results tell us what has been tried in research, but has failed. Thus some researchers have established journals or registries that *do* accept reports of negative results. These journals and registries can be found by searching the Internet for "negative results" or "research registry".

Here are some arguments in favor of publishing negative results and in favor of research registries:

- The publication of negative results helps researchers to avoid repeating research projects that have failed, thereby conserving resources.
- The publication of negative results provides useful cautionary information.
- The requirement that all research be registered before it is begun, including a statement of the research hypothesis and the research design makes it more difficult for researchers to publish serendipitous secondary findings that may have arisen through cherry picking or other researcher errors.
- For researchers who would like to study negative results, the requirement that all research be pre-registered before it is begun provides an indirect way to track down negative results. That is, we can identify research projects that have been registered but were subsequently never published. The omission of publication of the results of a registered research project suggests that the research may have obtained a negative result. (If the result had been positive, then a report of the positive result would likely have been published because that would be in everyone's interest.) We can also search the relevant registry to determine whether relevant negative results are recorded in the registry because some (though not many) researchers will make the effort to report their negative results in the registry.

Here are some arguments against publishing negative results and against research registries:

- In general, negative results are less interesting than positive results because negative results don't tell us anything beyond what we have already assumed—that the null hypothesis appears to be true.
- There are various possible reasons to explain why a research project obtained a negative result, including (a) the effect may not exist (which is the obvious reason), (b) the research may have failed to establish the (as yet unknown) conditions that are required for the effect to appear, (c) the researcher have may have been careless, which tends to lead to a negative result, or (d) a false-negative error due to chance may have occurred. Thus a

negative result doesn't necessarily imply that the effect under study *doesn't* exist (though some less experienced people may mistakenly interpret it that way).

- Most researchers know better than to publish serendipitous findings without first confirming them. This is because most researchers know that publishing a false-positive error will harm their reputation when the error is discovered, which is inevitable if the result is of at least moderate importance.
- It is highly unlikely that any researcher would ever *exactly* repeat the conditions of an unknown failed research project. And the difference in conditions between the "repeating" research project and the original research project might lead the second researcher to obtain a positive result. Thus, in general, negative results don't tell us much, but see appendix L.

If a researcher obtains a negative result, and if the researcher can't find a venue for publishing the result, and if the researcher thinks the result is important, then the researcher can avoid the so-called "file-drawer" problem by publishing the details of the research in an appropriate Internet archive, perhaps announcing the publication in relevant email lists or in other relevant social media. This enables other interested researchers in the field to be aware of the result.

It is sensible for any researcher planning a new research project to search journals of negative results, research registries, and the Internet for reports of similar research because the reports may contain useful information.

Venues that report negative results receive less readership due to general lack of interest in negative results because most researchers don't have enough time to read about all the *positive* results in their field, let alone the usually less-well-curated and generally less interesting negative results. Negative results are sometimes viewed as "failures" and may be embarrassing to some researchers. (A researcher shouldn't be embarrassed by a negative result because no researcher can expect that all his or her research hypotheses will be upheld.) And researchers usually get no reward for publishing a report of their negative results. So most researchers sensibly view it as a waste of time to prepare a proper report of a research project that obtained a negative result. Therefore, they won't spend the necessary time unless they are somehow coerced. Time will tell whether repositories of negative results and whether research registries are effective enough to justify their cost.

Appendix L: Examples of the Publication of Important Negative Results

As noted, scientific journals almost never publish reports of research that obtained a negative result because negative results are generally uninteresting. However, there are instructive exceptions when negative results *are* interesting and are therefore published in scientific journals.

For example, the famous Michelson-Morley experiment in physics (1887) studied the relationship between the *direction* of light travel and the *speed* of light. This careful experiment failed to find any good evidence of a relationship between the direction and the speed of light, which is a negative result that was surprising at the time of the research.

The report of the negative result of the Michelson-Morley experiment was published in the prestigious *American Journal of Science* and was widely discussed. The result was important because the expected *size* of the expected effect (i.e., the difference in the speed of light as a function of direction of light travel) was known, which is unusual in scientific research—we usually don't know the expected effect size ahead of time. (It was possible to compute the minimum possible size of the effect from the speed of the Earth in its orbit around the Sun.)

The "failure" of the sufficiently powerful Michelson-Morley experiment to discover the expected relationship of the expected size between the direction and the speed of light helped physicists to rule out the possibility of the existence of a stationary "luminiferous ether" as a necessary medium for the transmission of light. (The ether was thought to be necessary for the transmission of light, just as air, or some other gas, liquid, or solid, is a necessary medium for the transmission of sound—sound won't travel through a vacuum, but light will.) Prior to the Michelson-Morley experiment, many physical scientists believed that the stationary ether probably existed and was only waiting for someone to find good evidence of it (Wikipedia contributors, 2018).

The general point that we can take from the Michelson-Morley experiment is that negative results *are* interesting if (a) a particular effect is expected by many researchers in a field, (b) the expected effect size is at least roughly known and (c) the research project that obtained the negative result is clearly powerful enough and carefully enough performed that it ought to detect an effect of the expected size, if such an effect is present. This case is rare in scientific research, but does occur. In this case if the effect is important, then a report of a negative result in carefully performed research will often be accepted for publication in a relevant scientific journal. Ji (2017) discusses a modern example.

Another instructive example of publication of negative results arises in the cold fusion case, which is discussed above in appendix B.11. Some of the negative results that were obtained in the attempts to replicate the Pons and Fleischmann positive result were published, as noted by Huizenga (1993, app. III). This is because the Pons and Fleischmann result, if correct, was extremely important because it suggested an inexpensive, clean, and safe way to produce large amounts of energy. Therefore, many people wanted to know whether the claimed effect was real, and therefore these people were interested in any reports about attempts to replicate the effect, regardless of whether the reports were positive or negative.

Appendix M: Parameter Sign and Magnitude Errors

Appendix B.10 discusses false-positive and false-negative errors. Gelman and Tuerlinckx (2000) and Gelman and Carlin (2014) discuss two additional general types of errors that can occur in scientific research. These authors use a terminology that is consistent with the cryptic Type 1 and Type 2 (or Type

I and type II) terminology for false-positive and false-negative) errors respectively. They refer to the additional errors as Type S and Type M errors. Here, S stands for "sign" and M stands for "magnitude". However, in the interest of simple self-explanatory terminology, this paper refers to these errors respectively as "sign errors" and "magnitude errors".

As noted, false-positive and false-negative errors can occur when we are drawing conclusions about the *existence* of an effect. That is, an analysis will sometimes *incorrectly* conclude that an effect exists in the population when it doesn't (detectably) exist, which is a false-positive error. Similarly, an analysis will sometimes *incorrectly* conclude that there is no good evidence that an effect exists in the population, when the effect does (detectably) exist, which is a false-negative error.

The sign and magnitude errors introduced by Gelman and his co-authors don't pertain to the existence of effects, but pertain to the *estimation of parameter values*—they reflect errors about the correct *sign* of a parameter estimate, and errors about the correct *magnitude* of an estimate.

If we make a sign error in scientific research, then this implies that we have used a statistical procedure to estimate the value of a parameter, but the procedure has estimated the value with the wrong sign, a positive sign instead of a negative sign or vice versa. A sign error is a serious misleading error because it is telling us the "opposite" to what is true. Fortunately, in cases when we have a statistically significant result, and assuming there is no reasonable alternative explanation, sign errors are rare. But the laws of probability imply that sign errors can occur, so we should be aware of the possibility. Researchers eliminate sign errors through appropriate replication.

If we use a statistical procedure to estimate the value of a parameter, then the estimated value will almost never be exactly equal to the true value of the parameter in the population, which results in a magnitude error. (The unknowable "true" value of a parameter is defined as the value that causes the model equation to make the very best predictions, where "best" can be defined in various sensible theoretical senses.) Parameter magnitude errors are less misleading than sign errors, but we should be aware of them because they are inevitable. Fortunately, the theory behind the distribution of parameter estimates implies that magnitude errors will more often be small than large, as illustrated in figure 1 in the body of this paper. (For technical reasons, parameter estimates will *on average* often tend to be slightly higher in absolute value than the true population value.) Magnitude errors generally aren't a serious problem if we remember that parameter estimates are only estimates, and all estimates are subject to magnitude errors.

If we are uncertain about the magnitude of a parameter estimate, then we can perform a new research project to obtain another independent estimate of the magnitude. Also, statistical procedures are available to assist us to combine the various estimates of the same parameter value into an overall estimate which (if we do everything properly) will be more precise and more accurate overall than any of the individual estimates.

Researchers and research projects can make other errors in addition to false-positive errors, false-negative errors, sign errors, and magnitude errors. For example, researchers may specify a model equation for a relationship between variables that is somehow incompatible with the relationship that exists in the population, which is called a model error. Similarly, there may be confounding errors, data dredging errors, and so on.

Appendix N: Exceptions to the Idea that Research Projects Study Relationships Between Variables

Section 1 in the body of this paper says that we can view most scientific research projects that collect and study data as studying relationships between variables in data tables. This appendix discusses some apparent exceptions to this point of view.

Readers who are familiar with the two-sample *t*-test will know that it is a statistical test of whether a continuous variable has a significantly different average in two different groups of entities. Does this test study the existence of a relationship between variables? Yes. The response variable is the continuous variable mentioned in the first sentence of this paragraph. And the predictor variable is a "binary" or two-valued variable that reflects the difference between the two groups. Thus we can readily view the two-sample *t*-test as a test of the existence of a relationship between two variables. And the extension of the two-sample *t*-test into multiway analysis of variance is readily viewed as the study of the relationship between two or more discrete predictor variables and a continuous response variable.

In a degenerate case of the study of a relationship between variables, we may study a single variable (column) in a data table in isolation. In this case we have a response variable and *zero* predictor variables, which is logically and mathematically the limiting case of a relationship between variables when the number of predictor variables is reduced to zero.

In a second degenerate case, we may (in effect) study a single entity (row) in a data table in isolation because we are unable to obtain multiple rows for the table due to a lack of available data. (In this case we often don't use a table to hold the data.) This case often arises in the historical sciences such as in archaeology, paleontology, and evolutionary biology, which must often work with a sample size of one. This case also arises in some branches of the social sciences, such as in some areas of anthropology (when the main entity of study may be a single society) and in traditional clinical psychiatry (when the main entity of study is usually a single psychiatric patient). In cases with a sample size of one, the research is limited to a description of the value of each variable for the entity and possible comparison of the values with other related entities that are drawn from different populations. And relationships between the variables can't sensibly be studied because we need a sample of at least ten or so entities from a population before we can sensibly study a relationship between variables in the entities. (Properly collected larger samples are better in the sense of being better able to detect relationships and being better able to enable reliable prediction or control.)

In another special case, we have no response variable and we merely study a set of several predictor variables in a data table. Our goal is to find a way to organize the *variables* (columns of the table) into sensible groups of highly correlated variables (super-variables, so to speak), as in exploratory factor analysis and principal components analysis. We can use the appropriate weighted average of highly correlated variables as a more precise measure of the property that the variables all measure. We can then use this property as the response variable or as a predictor variable in other scientific research.

Similarly, in another special case, we have no response variable, and we use the predictor variables in the table to assist us to organize the *entities* (rows of the table) into groups of similar entities according to the values of the variables for the entities, as in cluster analysis. This enables us to assign new entities from the population to relevant groups, which is sometimes useful to facilitate dealing with the entities. This, in effect, *defines* a new variable that identifies the different groups.

Other exceptions to the view that research projects study the relationship between (a) one or more predictor variables and (b) a single response variable also exist. Some examples are research projects that use the less-frequently-used statistical methods of multivariate analysis, path analysis, and canonical correlation analysis. These approaches don't directly study a relationship between one or more predictor variables and a single response variable. But they can all be readily viewed as sensible variations or extensions of the idea of studying relationships between variables.

Thus, though there are exceptions, we often find that a scientific research project (or a portion of a research project) has a response variable of central interest and zero or more predictor variables that the researcher believes can help us to predict or control the values of the response variable. Sometimes these concepts aren't explicit, so we may have to puzzle a little to identify the variables and their roles. Identifying all the relevant variables and their roles in a research project greatly facilitates understanding.

Appendix O: Are the Ideas Discussed in this Paper "Real"?

This paper refers to the idea of a parameter of a model equation "having a value" in a population of entities. What does that mean?

That is, do model equations and parameters of equations exist in the external world? Is there, metaphorically, a great book somewhere in the sky that drives everything, specifying:

(a) all the true types of entities and properties of entities in the universe
(b) the true systems of measurement of the values of properties
(c) all the true model equations of all the relationships between the properties, and
(d) the true values of all the parameters of the equations?

Or have humans merely imposed these ideas on the world, and some or all the ideas have no basis in reality? Or is there some other sensible explanation?

We presently don't decisively know the answers to these questions, though most people believe that at least some of the concepts reflect what is "real". But, unfortunately, we haven't found a way to look behind the curtain (if any) to see the true reality.

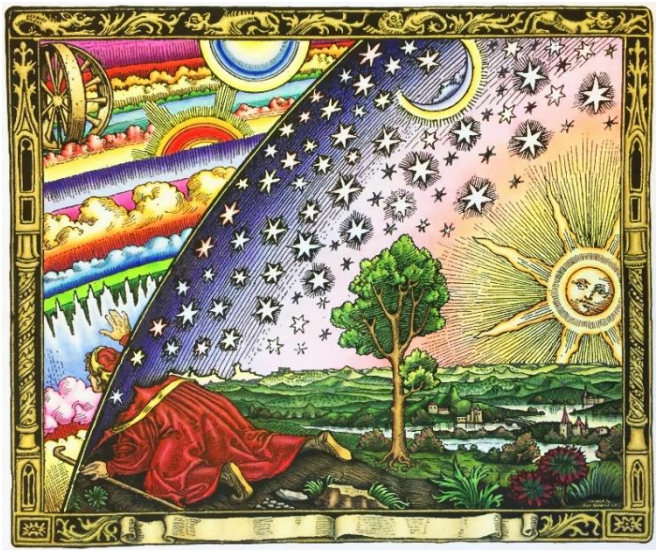

Figure O.1. The Flammarion engraving, as colored by Houston Physicist. This entrancing image is discussed in Wikipedia. (The figure is copied with CC BY-SA 4.0 permission.)

The idea of looking behind the curtain at the true reality is highly attractive. But, unfortunately, even if we could look behind the curtain, there might still be another curtain concealing the true true reality. (That needn't discourage us from looking for the curtains because the search itself is highly informative and highly satisfying.)

Fortunately, we needn't decide whether the ideas are "real" in some absolute sense. This is because we (humans) have found that the ideas are *useful*. The ideas are useful because they assist us with accurate prediction and control, regardless of the true reality of the ideas. The usefulness is due to the basic *stability over time* that we have observed in the ideas that we can see behind the noise. The stability is in the ideas of entities, properties, measurement, variables, relationships between variables, model equations, parameters of model equations, and values of parameters of model equations.

The stability begins with the idea of entities. As infants we somehow see how to efficiently organize the jumble of incoming sensations that we experience. We carry out the organization by using the concepts of entities and properties. Perhaps the first entity we recognize is usually "mother", who has properties associated with actions, looks, sounds, and smells. If mother is nearby, then crying is almost guaranteed to get her attention, which may be the first relationship between variables that we learn to use.

We don't ask whether "mother" is "real". This is because of course she is real, because (for most children) she is stable in our senses across time. The long-term stability of the ideas assures us that the ideas are real, or at least real enough for us

to sensibly *believe* that they are real. We experience this stability in the many entities, properties, and relationships that we use in our daily thinking.

To return to the first question in this appendix, we see that empirical research has shown that the stability of the ideas extends to the values of parameters of properly derived model equations. Research has shown that parameters of model equations typically (though not always) have stable unchanging values across time (except for changes associable with measurement noise). Therefore, the idea of a parameter having a value in a population of entities is a sensible real (i.e., generally stable across time) idea.

References

Arbuthnot, J. (1710), "II An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes," *Philosophical Transactions of the Royal Society of London*, 27, 186–190. Reprinted in *Studies in the History of Statistics and Probability Volume II*, eds. M. G. Kendall and R. L. Plackett, High Wycombe UK: Griffin, 30–34. https://doi.org/10.1098/rstl.1710.0011 and https://www.york.ac.uk/depts/maths/histstat/arbuthnot.pdf

Bakan, D. (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66, 423–437. https://doi.org/10.1037/h0020412

Baker, A. (2016), "Simplicity," *The Stanford Encyclopedia of Philosophy* (Winter 2016, ed. E. N. Zalta). https://plato.stanford.edu/archives/win2016/entries/simplicity/

Bayarri, M. J., Benjamin, D. J., Berger, J. O., and Sellke, T. M. (2016), "Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses," *Journal of Mathematical Psychology*, 72, 90–103. https://doi.org/10.1016/j.jmp.2015.12.007

Benjamin, D. J., Berger, J. O., …, and Johnson, V. E. (2017), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300. https://www.jstor.org/stable/2346101

Berger, J. O., and Berry, D. A. (1988), "Statistical Analysis and the Illusion of Objectivity," *American Scientist*, 76, 159–165. http://www.jstor.org/stable/27855070

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112–139. https://doi.org/10.1080/01621459.1987.10478397

Berkson, J. (1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test," *Journal of the American Statistical Association*, 33, 526–536. https://doi.org/10.1080/01621459.1938.10502329

Bernoulli, D. (1734), "Recherches Physiques et Astronomiques, Sur Le Problème Proposé pour la Seconde Fois par l'Académie Royale des Sciences de Paris," Recueil des pièces qui ont remporté les prix de l'Académie royale des sciences, depuis leur fondation jusqu' à présent. Avec les Pièces qui y ont concouru. Tome Troisième Contenant les Pièces depuis 1734 jusque en 1737. Paris: l'Académie Royale des Sciences de Paris. Note: Todhuner (1865, 222–223) gives an English explanation of Bernoulli's contribution. http://doi.org/10.3931/e-rara-48836

BIPM [Bureau International des Poids et Mesures] (2006), "The International System of Units". https://www.bipm.org/en/publications/si-brochure/

Casella, G., and Berger, R. L. (2002), *Statistical Inference* (2nd ed.), Delhi, India: Cengage Learning India.

Cohen, J. (1994), "The Earth Is Round (p < .05)," *American Psychologist*, 49, 997–1003. https://dx.doi.org/10.1037/0003-066X.49.12.997

Colquhoun, D. (1971), *Lectures on Biostatistics: An Introduction to Statistics with Applications in Biology and Medicine*, Oxford, UK: Clarendon Press.

—— (2017), "The Reproducibility of Research and the Misinterpretation of *p*-Values," *Royal Society Open Science*, 4: 171085. http://doi.org/10.1098/rsos.171085

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

—— (2014), "Comment on a paper by Jager and Leek," *Biostatistics*, 15, 16–18. https://doi.org/10.1093/biostatistics/kxt033

Demidenko, E. (2016), "The *p*-Value You Can't Buy," *The American Statistician*, 70, 33–38. https://doi.org/10.1080/00031305.2015.1069760

Dienes, Z. (2011), "Bayesian Versus Orthodox Statistics: Which Side Are You On?" *Perspectives on Psychological Science*, 6, 274–290. https://doi.org/10.1177/1745691611406920

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference*, New York: Cambridge University Press.

Estes, W. K. (1997), "Significance Testing in Psychological Research: Some Persisting Issues," *Psychological Science*, 8, 18–20. https://doi.org/10.1111/j.1467-9280.1997.tb00538.x

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd. The 14th edition of this seminal work appears in Fisher (1990).

—— (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd. The 8th edition of this seminal work appears in Fisher (1990).

—— (1990), *Statistical Methods, Experimental Design, and Scientific Inference*, ed. J. H. Bennett, Oxford: Oxford University Press.

Gelman, A. (2015), "The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective," *Journal of Management*, 41, 632–643. https://doi.org/10.1177/0149206314525208

Gelman, A., and Carlin, J. (2014), "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., and Tuerlinckx, F. (2000), "Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures," *Computational Statistics*, 15, 373–390. https://doi.org/10.1007/s001800000040

Gosset, W. S. (1908) [see Student (1908)].

Greenland, S. (2017), "Invited Commentary: The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–645. https://doi.org/10.1093/aje/kwx259

Held, L., and Ott, M. (2016), "How the Maximal Evidence of *P*-Values Against Point Null Hypotheses Depends on Sample Size," *The American Statistician*, 70, 335–341. https://doi.org/10.1080/00031305.2016.1209128

Huizenga, J. R. (1993), *Cold Fusion: The Scientific Fiasco of the Century*, New York: Oxford University Press.

Ioannidis, J. P. A. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124

—— (2008), "Effect of Formal Statistical Significance on the Credibility of Observational Associations," *American Journal of Epidemiology*, 168, 374–383. https://doi.org/10.1093/aje/kwn156

Jager, L., and Leek, J. T. (2014), "An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature" (with discussion), *Biostatistics*, 15, 1–45. https://doi.org/10.1093/biostatistics/kxt007

Ji, X. (2017), "Dark matter remains elusive," *Nature* 542, 172. https://doi.org/10.1038/542172a

Johnson, V. E. (2013), "Revised Standards for Statistical Evidence," *PNAS*, 110, 19313–19317. https://doi.org/10.1073/pnas.1313476110

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2017), "On the Reproducibility of Psychological Science," *Journal of the American Statistical Association*, 112, 1–10. https://doi.org/10.1080/01621459.2016.1240079

Jones, L. V., and Tukey, J. W. (2000), "A Sensible Formulation of the Significance Test," *Psychological Methods*, 5, 411–414. https://doi.org/10.1037/1082-989X.5.4.411

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Lakens, D. et al. (2018), "Justify Your Alpha," *Nature Human Behavior*, 2, 168–171. https://doi.org/10.1038/s41562-018-0311-x

LeCun, Y., Bengio, Y., and Hinton, G. (2015), "Deep Learning," *Nature*, 521, 436–444. https://dx.doi.org/10.1038/nature14539

Lehmann, E. L., and Romano, J. P. (2005), *Testing Statistical Hypotheses* (3rd ed.), New York: Springer.

Mayo, D. (2014), "On the Birnbaum Argument for the Strong the Likelihood Principle" (with discussion), *Statistical Science*, 29, 227–266. https://dx.doi.org/10.1214/13-STS457

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2018), "Abandon Statistical Significance (version 3)". https://arxiv.org/abs/1709.07588v3

Michelson, A. A., and Morley, E. W. (1887), "On the Relative Motion of the Earth and the luminiferous Ether," *American Journal of Science* (third series), 34, 333–345. https://doi.org/10.2475%2Fajs.s3-34.203.333

Mohr, P. J., Newell, D. B., and Taylor, B. N. (2016), CODATA recommended Values of the Fundamental Physical Constants: 2014," *Reviews of Modern Physics*, 88, 1–73. https://doi.org/10.1103/RevModPhys.88.035009

National Cancer Institute (2018), "Laetrile/Amygdalin – Health Professional Version". https://www.cancer.gov/about-cancer/treatment/cam/hp/laetrile-pdq

Nature editors (2018), "Checklists work to improve science" [editorial], *Nature*, 556, 273–274. https://doi.org/10.1038/d41586-018-04590-7

Neyman, J., and Pearson, E. S. (1928), "On the use and interpretation of certain test criteria for purposes of statistical inference, Part I," *Biometrika*, 20A, 175–240. https://doi.org/10.1093/biomet/20A.1-2.175

—— (1933a), "The Testing of Statistical Hypotheses in Relation to Probabilities A Priori," *Joint Statistical Papers*. Cambridge UK: Cambridge University Press. pp. 186–202.

—— (1933b), "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337. https://doi.org/10.1098/rsta.1933.0009

Nickerson, R. S. (2000), "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy," *Psychological Methods*, 5, 241–301. http://psycnet.apa.org/doi/10.1037/1082-989X.5.2.241

Palus, S. (2018), "Make Research Reproducible," *Scientific American*, 319 (4) October, 56–59. https://doi.org/10.1038/scientificamerican1018-56

Pearson, K. (1900), "On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling," *Philosophical Magazine Series 5*, 50, 157–175. See https://en.wikipedia.org/wiki/Philosophical_Magazine for links.

—— (1904), "Mathematical contributions to the theory of evolution. XIII. On [the theory of] Contingency and its Relation to Association and Normal Correlation," *Draper's Company Research Memoirs: Biometric series I.* London: Dulau and Co. and Department of Applied Mathematics, University College, University of London. (This book is available in PDF from the Cornell University Library.) https://newcatalog.library.cornell.edu/catalog/5670968

Popper, K. R. (1980), *The Logic of Scientific Discovery*, London: Routledge.

—— (1989), *Conjectures and Refutations: The Growth of Scientific Knowledge*, London: Routledge.

—— (1992), *Realism and the Aim of Science*, London: Routledge.

Rao, C. R., and Lovric, M. M. (2016), "Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective," *Journal of Modern Applied Statistical Methods*, 15 (2), 2–21. https://doi.org/10.22237/jmasm/1478001660

Roche, J. J. (1998), *The Mathematics of Measurement: A Critical History*, London: Athlone Press.

Rosnow, R. L., and Rosenthal, R. (1989), "Statistical procedures and the justification of knowledge in psychological science," *American Psychologist*, 44 (10), 1276–1284. https://doi.org/10.1037/0003-066X.44.10.1276

Sen, A., and Srivastava, M. (1990), *Regression Analysis: Theory, Methods, and Applications*, New York: Springer.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Chichester UK: Wiley.

Student (1908), "The Probable Error of the Mean," *Biometrika*, 6, 1–25. https://doi.org/10.1093/biomet/6.1.1

Todhunter, I. (1865), *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace.* New York: Hardgrass.

Tukey, J. W. (1989), "SPES in the Years Ahead," in *American Statistical Association Proceedings of the Sesquicentennial Invited Book Sessions*, pp. 175–182.

—— (1991), "The Philosophy of Multiple Comparisons," *Statistical Science*, 6, 100–116. https://doi.org/10.1214/ss/1177011945

Wagenmakers, E. -J. (2007), "A Practical Solution to the Pervasive Problems of *p* Values," *Psychonomic Bulletin & Review*, 14, 779–804. https://doi.org/10.3758/BF03194105

Wasserstein, R. L. (ed.) (2016), "ASA Statement on Statistical Significance and *P*-values," *The American Statistician*, 70, 131–133.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E. -J. (2011), "Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 *t* Tests," *Perspectives on Psychological Science*, 6, 291–298. https://doi.org/10.1177/1745691611406923

Wikipedia contributors (2018), "Michelson-Morley experiment," Accessed on April 5, 2018. https://en.wikipedia.org/wiki/Michelson%E2%80%93Morley_experiment