# 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

The basic ideas behind the statistical *p*-value are reviewed. It is proposed that the function of the *p*-value in scientific research is to provide a measure of the weight of evidence that an effect observed in the data from a sample is a real effect in members of the underlying population. The need in scientific research for a measure of the weight of evidence that an effect is real is assessed. Serious problems with the *p*-value are identified. The *p*-value is compared with seven other measures that perform the same function (*t*-statistic, confidence interval, likelihood ratio, Bayes factor, posterior probability that the null hypothesis is true, *D*-value, and information-criteria). The comparisons imply that the *p*-value is superior to the other measures for detecting effects in scientific research.

KEY WORDS: Hypothesis testing; Significance testing; Basic statistical ideas; Role of statistics in scientific research

## 1. Introduction

Many researchers agree that the *p*-value helps them to detect meaningful "effects" in scientific research data. However, various authors criticize the *p*-value, identifying serious problems with it. The present paper (a) reviews the scientific and statistical ideas behind the *p*-value, (b) summarizes the problems associated with the *p*-value, and (c) compares the *p*-value with seven alternative approaches to detect effects. The goal is to determine which approach is best to detect effects in scientific research data.

To ensure a common base, some of the discussion at the beginning of this paper is basic. The advanced reader's indulgence is requested. Also, since a key role of statistics is to support scientific research, this paper often links statistical ideas to scientific research. Also, the paper focuses on making statistical ideas easy for beginners to understand because many beginners presently misunderstand key ideas.

The rest of the paper proceeds as follows:

- Section 2 describes how a large proportion of scientific research can be reasonably viewed as studying relationships between variables in populations of entities as a means to accurate prediction or control. We study a relationship between variables in a sample of entities selected from the population.
- Section 3 explains how the *p*-value provides a reasonable measure of the *weight of evidence* that an effect (usually a relationship between variables) observed in the research data for a sample is a real effect in members of the population of entities behind the sample.
- Section 4 discusses how we need a measure of the weight of evidence that an effect is real to reduce the occurrence of incorrect scientific conclusions.
- Section 5 reviews the important distinctions in scientific research between (a) the *existence* of an effect (e.g., the existence of a relationship between variables), (b) the *strength* or size of an effect, and (c) the practical or theoretical *importance* of an effect.
- Section 6 lists serious problems with the *p*-value that raise questions about its usefulness.
- Section 7 summarizes comparisons of the *p*-value with seven sensible alternative measures of the weight of evidence that an effect is real. (Details of the comparisons are in an appendix.) The comparisons suggest that the *p*-value is superior to each of the other measures.
- Section 8 draws conclusions.

## 2. Relationships Between Variables as a Means to Accurate Prediction and Control

We can view a large proportion of scientific research as studying relationships between variables as a means to accurate prediction and control. Of course, the *variables* that we study in scientific research reflect the measured values of properties of the entities that we study in the research. A variable can represent any particular property of entities that we might wish to study in any area of life.

A relationship exists in a population of entities between a "predictor" variable $x$ and a "response" variable $y$ if when $x$ changes in value in the entities, $y$ changes in value somewhat "in step" with $x$ in the same entities. For example, if we use an umbrella properly $(x)$ when it is raining, then our dryness $(y)$ will generally be greater. In other words, there is a relationship between using an umbrella properly and staying dry in the population of human beings.

In everyday life we discover many relationships between variables through informal observation. (Many young children are fascinated when they first recognize the practicality of the umbrella relationship.) Throughout our lives we learn

about hundreds of thousands of similar relationships between variables. We learn about many such relationships on an intuitive unspoken level, such as the various relationships between variables that we must learn to ride a bicycle. We use our knowledge of relationships between variables to help us to predict and control our surroundings.

We can use a scientific research project as a powerful formal way to study the relationship between any variables $x$ and $y$ that we might wish to study. A standard research project (or logical portion of a research project) measures the values of selected properties of the entities (i.e., measures the values of variables) in a "sample" of entities that has been selected from the population of interest. The research project collects these measured values in a data table whose data are, on a *technical* level, the main object of study in the research. Of course, on a *conceptual* level, the main objects of study in the research are the entities in the population behind the sample data.

Most data tables have the same standard structure. Each *row* in the table is associated with data for a different entity in the sample. And each *column* is associated with values of a different variable. Each *cell* (intersection of a row and a column) in the table contains the value of the variable associated with the column for the entity in the sample associated with the row.

We analyze (study) the data in a relevant data table to study relationship(s) between selected variables in the entities in the sample. If we do the data collection and analysis properly, this enables us to draw reliable conclusions about the relationship(s) between the variables in the entities in the *entire population* of entities that lie behind the entities in the sample. This generalization is important because we are usually mainly interested in drawing conclusions about relationships between variables in the underlying population, not merely about relationships in the sample.

Usually we can view a scientific research project (or a logical portion of a research project) as studying a particular single response variable $y$ that is in a column of the data table. And usually we can view a research project (or portion) as studying one or more predictor variables $x$ that are also in columns of the table. We wish to determine whether and how the values of $y$ in the entities in the population "depend" on the values of $x$ in the entities.

If we can find a relationship between a set of one or more predictor variables and the response variable in the entities in a population, then we can use the information about the relationship to help us to predict or control the values of the response variable in new entities from the population. For example, medical researchers may study whether a beneficial relationship exists between the amount $(x)$ of a particular drug administered to medical patients with a certain disease, D, and the amount $(y)$ of disease D in the patients. If the researchers can find good evidence that there is a beneficial relationship between the amount of the drug and the amount of the disease, then doctors can use the knowledge of the relationship to reduce the disease in new similar patients (by prescribing the drug for them). Similarly, if psychologists can find a relationship between the style of upbringing $(x)$ of a child and the child's later sense of happiness $(y)$, then parents can use the knowledge of this relationship to help them to raise happy children.

In studying a relationships between variables, it is important to distinguish between "observational" research projects and "experimental" research projects. In an *observational* research project we collect the data by *observing* the values of the predictor variable(s) in the entities in the sample. We also observe the values of the response variable in the entities. We collect these observed values in a data table. We analyze the data in the table and, if possible, we derive a "model equation" for the relationship between the variables, as explained in statistics textbooks. If we properly derive such an equation, then we or others can use it (or a graph or text summarizing it) to *predict* the values of the response variable in new entities from the population. We make the predictions by substituting the values of the predictor variable(s) for a new entity into the equation and then evaluating the expression to yield the predicted value of the response variable for the entity.

In contrast, in an *experimental* research project we systematically *manipulate* the values of one or more predictor variables in the entities in the sample and we subsequently observe the values of the response variable in the entities, again collecting the relevant data in a data table. (We may also measure some predictor variables that *aren't* manipulated in the entities, but are merely observed.) As with observational data, we analyze the data in the table and, if possible, we derive a model equation for the relationship between the variables on the basis of the analysis. If we properly derive such an equation, then we or others can (if socially appropriate) use it (or a summary) to help us to *control* the values of the response variable in new entities from the population. We achieve the control for a new entity by appropriately manipulating the values of the relevant predictor variables in the entity according to the information in the model equation.

The ability to predict or control that is achieved through scientific research is sometimes weak and sometimes strong, but rarely perfect. Therefore, we must take account of "error" in prediction or control. Statisticians have developed an extensive set of techniques to take account of error as discussed in statistics textbooks.

If we find good evidence of a new (real) relationship between variables in a population, and if we publish a report of the relationship, then the knowledge of the relationship (as summarized by the model equation or by a derived table or graph) becomes a new fact in the body of human knowledge. The contribution of new knowledge about a relationship between variables is especially rewarding for a researcher if the relationship has important practical or theoretical implications.

The preceding paragraphs imply that scientific research enables us to develop a model equation of a relationship between variables, which enables us to predict or control the values of response variables in new entities from the relevant population. Prediction and control are clearly useful scientific goals.

A second important goal of scientific research is to achieve an "understanding" of the entities in the population.

(This second goal is obviously assisted by knowledge of relationships between variables in the entities.) It is important to acknowledge the second goal, but this paper focuses on the first goal—prediction or control.

Some researchers don't view scientific research projects that collect data as studying relationships between variables. But, arguably, we can readily view almost any empirical research project that studies the data in a data table as studying one or more relationships between the variables in the entities in the population behind the sample. This unifying point of view substantially increases understanding of scientific research because it enables us to view most research projects through the same sensible logical lens. Appendix A discusses some apparent exceptions to these ideas.

## 3. Hypothesis Testing with *p*-Values to Detect Relationships Between Variables

### *3.1. 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) 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. Statistics textbooks explain how to examine and (without bias) "clean" scientific research data.

### *3.2. Detecting Whether a Relationship Exists*

If we have a data table with properly collected and properly cleaned data, then the first step to study a relationship of interest between the variables is to determine whether we have good evidence that the relationship actually *exists*. This step is important because we sometimes find that a relationship between variables under study *doesn't* exist (or at least doesn't exist in enough strength to be detected). It is sensible to perform this step first because if we can't find good evidence that a particular relationship between variables *exists*, then it is inefficient to study the relationship further in these specific research data (because such study may amount to studying mere noise in the data).

If we have an appropriate table of research data, we can perform a "statistical hypothesis test" to assess whether a relationship exists between selected variables in the table. If we do everything properly, this enables us (by extension) to assess whether the relationship exists in the population.

We begin a hypothesis 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 chosen 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.

More generally, as suggested in section 2, a research hypothesis may say that a relationship of a particular type exists between a specified set of *multiple* predictor variables, $x$, and the response variable, $y$, in the entities in the population.

(Some researchers refer to a research hypothesis in a scientific research project as the "alternative" or "alternate" hypothesis. However, those terms are less appropriate, as discussed in appendix B.1.)

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, we 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 cherished research hypothesis is true in the population (because that will advance human knowledge).

The idea of *formally* beginning with the assumption that the null hypothesis is true refers to our formal thought process. The idea doesn't imply some formal behavior, such as signing a formal document.

After assuming that the null hypothesis is true, we can analyze the data in a relevant data table to determine whether the data show evidence that a relationship exists between selected variables. Technically, we do this analysis by examining a particular "parameter" of a model equation for the relationship between the variables of interest. A parameter of an equation is a particular *number* that we estimate through an analysis of the data table. The estimated numeric value of each parameter of a model equation is an essential part of the detailed specification of the equation, as further explained in appendix B.

If there is *no* relationship in a population between a predictor variable and a response variable, then it is easy to see that the relevant parameter of an appropriate model equation of the relationship between the variables will have a particular "null" value in the population. The null value is typically the value zero.

For example, in a standard linear regression model equation, if there is no relationship between a given predictor variable in the equation and the response variable, then the parameter (regression coefficient) for the term for this predictor variable will have a null value of zero in the population. In contrast, if there *is* a relationship between the variables, then the parameter will (usually) have a value that is different from zero.

Thus we can determine whether a relationship exists between the variables in a given data table by appropriately estimating (through an analysis of the data) the value of the relevant parameter. Then we can check to determine whether this value is "meaningfully" different from the relevant null value. If we find that the estimated value of a parameter is meaningfully different from the null value, then this is good

evidence that a relationship exists in the entities in the population between (a) the predictor variable(s) $x$ associated with the parameter and (b) the response variable $y$.

Statisticians have invented three sensible methods to estimate (from appropriate scientific research data) the values of the parameters of a model equation. These are the least-squares method, the maximum-likelihood method, and the Bayesian method. Each method is optimal (in a different sense) and each method has many details, as explained in statistics textbooks. It is reassuring that the methods (when applicable) generally closely or exactly agree with each other in their estimates of the values of the parameters of a given model equation. The methods agree because they all address the same basic goal, which is to find the best estimates of the values of the parameters of the equation to model the particular relationship between the variables under study.

As noted, we can determine if a relationship exists between variables by checking the estimated value of the relevant parameter to see if it is meaningfully different from the null value. More generally, instead of checking the value of a parameter, we can detect a relationship between variables by checking a "test statistic" derived from the data, such as an *F*-statistic derived in analysis of variance. In this case, we use the same general ideas. That is, we specify the null value for the statistic—the value that would be (on average) expected to occur if the relevant null hypothesis is or were true in the population. Then we check the value of the statistic computed from the data to see if it is meaningfully different from the null value.

### 3.3. The p-Value

This paper considers several effective ways to perform a statistical hypothesis test to help us to check if the value of a parameter or test statistic is meaningfully different from the relevant null value. We first consider how the *p*-value enables us to perform such a test.

The *p*-value is a probability (i.e., a fraction of the time). Here is a definition:

> **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* that the value, as estimated from the research data, will be as discrepant or more discrepant from the relevant null value as it is with 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 of interest, and
> - certain often-satisfied technical assumptions that are required to correctly compute the *p*-value are or were adequately satisfied.

Statisticians have discovered how to correctly compute *p*-values that satisfy this definition.

The definition of the *p*-value implies that the *lower* the *p*-value that is computed (from research data) for a relationship between variables, the less likely it is that we would have obtained this state of affairs if the null hypothesis is or were true. Therefore, the lower the *p*-value, the *more evidence* we have (in the absence of a reasonable alternative explanation) that a relationship exists between the associated variables in the entities in the population. Therefore, if the *p*-value is *low enough*, and if there is no reasonable alternative explanation for the low *p*-value, then we can (tentatively) "reject" the null hypothesis and (tentatively) conclude that the research hypothesis is true. That is, we can (tentatively) conclude that the associated relationship exists between the variables in the population.

Many different statistical hypothesis tests with *p*-values are available to detect relationships between variables. This is because there are several types of variables and there are several types of research projects, and (for technical reasons) different types of variables and research projects generally require different tests.

Fortunately, all of the standard procedures to compute *p*-values for hypothesis tests have been programmed in widely available user-friendly statistical software. Thus a researcher needn't know the 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.

By convention, we tentatively conclude that a relationship exists between variables if the *p*-value for the relationship is less than (or equal to) the chosen "critical" value (and if there is no reasonable alternative explanation for the low *p*-value). And, by convention, the critical *p*-value is often chosen to be 0.05 or 0.01 although other critical values such as 0.005 or lower are sometimes recommended and used, as discussed in appendix C.

If we compute a *p*-value for a relationship between variables, and if the *p*-value is less than (or equal to) the critical *p*-value, then this is called a "positive" result. In computing a *p*-value, a researcher almost always wishes to obtain a positive result because (in the absence of a reasonable alternative explanation) this is good evidence that the research project has found what it was looking for.

In contrast, if we compute a *p*-value for a relationship between variables, and if the *p*-value is *greater* than the critical value (e.g., greater than 0.05), then this is called a "negative result". A negative results is almost always disappointing for a researcher because it means that the research project has *failed* to find good evidence of what it was looking for.

Appendix B.12 shows how the operation of the *p*-value implies that if we do everything properly, and if we use a critical *p*-value of 0.05, then in those cases when a detectable relationship *doesn't* exist between the variables, we will obtain a *p*-value less than (or equal to) 0.05 in 5% of the cases. That is, the results will mistakenly suggest that a relationship *does* exist in 5% of the cases even though the relationship actually *doesn't exist*. This deceptive result is called a false-positive error. In many standard situations the 0.05 false-positive error

rate (or the 0.01 rate) is judged to be acceptable. This is because requiring a lower false-positive error rate (e.g., using a critical *p*-value of 0.005) would drive research costs too high, as discussed in appendix C. Researchers rarely use critical *p*-values that are *higher* than 0.05 because such values are seen as too lenient—allowing too many false-positive errors to occur.

False-positive errors are obviously highly undesirable because they falsely lead us to believe that a non-existent relationship exists between variables, which may lead to a substantial waste of resources. Fortunately, these errors are easy to identify, as discussed below in section 3.6.

Appendix B.12 also shows how the operation of the *p*-value implies that if we do everything properly, and if we use a critical *p*-value of 0.05, then in those cases when a detectable relationship *does* exist between the variables we will obtain a *p*-value greater than 0.05 in a certain proportion of the cases. (The proportion depends on the strength of the relationship and on the design of the research project.) That is, the results will mistakenly imply that we have no evidence of a relationship, even though the relationship exists. This deceptive result is called a false-*negative* error. In this case we have missed discovering a relationship between variables that is actually present in the population.

Like false-positive errors, false-negative errors are also obviously highly undesirable because they amount to a failure to find what we are looking for, even though what we are looking for is there. Fortunately, these errors are easy to reduce by increasing the "power" of statistical tests, as discussed in appendix B.

(False-positive and false-negative errors are sometimes called "Type I" and "Type II" errors respectively. However, this terminology is weak and is confusing for beginners because it has no descriptive content.)

If (through a low *p*-value or through some other reasonable approach) we find good evidence of a relationship between variables in appropriate research data, then we can study the data further to determine the *specifications* of the putative relationship, especially the specifications of the complete form of the model equation for the relationship. Then we can publish a report about the specifications of the apparent relationship we have discovered. Then, assuming our conclusions are correct, we or others can use the knowledge of the relationship to accurately predict or control the values of the response variable in new entities from the population.

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 views of statistical hypothesis testing and statistical inference are discussed by Casella and Berger (2002), Lehmann and Romano (2005), and Cox (2006).

### 3.4. Generalizations

A relationship between variables is one type of "effect" that we can study in scientific research data. More generally, hypothesis tests can test the question of whether some general parameter or test statistic (not necessarily indicative of a relationship between variables) that is computed from the data is significantly different from the relevant null value, which is a more general type of effect. For generality and brevity, this paper sometimes uses the term "effect". However, typically in a scientific research project we can view an effect as reflecting a particular relationship between variables.

The preceding ideas suggest the function of the *p*-value in scientific research: The *p*-value is (in the absence of a reasonable alternative explanation) a reasonable objective measure of the *weight of evidence* that we have successfully observed a real relationship between the variables (or a real effect) in the entities in the population—a measure of the weight or strength of evidence that we have in favor of rejecting the associated null hypothesis (Fisher, 1973, p. 80). In other words, the *p*-value is a measure of the weight of evidence that the research hypothesis under study is true. In the absence of a reasonable alternative explanation, the lower the *p*-value below the critical *p*-value, the greater the weight of evidence that the associated research hypothesis is true and therefore the greater the weight of evidence that the associated effect is real.

Some thoughtful readers who are familiar with the statistical two-sample *t*-test may sensibly wonder whether this test is a test for evidence of the existence of a relationship between variables. This question arises because the *t*-test is often characterized as testing for a *difference between two groups* in the group means of some variable, $y$. However, it is always possible and instructive to view the two-sample *t*-test as testing whether a continuous response variable (the variable $y$) is related to a binary predictor variable, $x$, which is the variable that reflects the conceptual difference between the two groups.

Similarly, the extension of the *t*-test into analysis of variance, multiple regression analysis, and the general linear model can be easily viewed as testing whether the continuous response variable is related to one or more discrete or continuous predictor variables. Similarly, the chi-square test (and related tests) of a two-way contingency table can be sensibly viewed as a test of whether there is a relationship between the discrete row variable and the discrete column variable of the table. In general, one of the two variables can be sensibly viewed as the response variable and the other can be viewed as the predictor variable. These ideas can be easily extended to more complicated contingency tables.

It is noteworthy that the *p*-value is somewhat "crude" in the sense that it makes two types of serious errors—false-positive errors and false-negative errors. Thus one might wonder whether there might be a better way to detect effects—a way that would make fewer errors. Unfortunately, all of the other measures of the weight of evidence that an effect is real make the same or similar errors, as discussed later in this paper. And, apparently, no better approaches are possible.

The fact that the measures of the weight of evidence make errors isn't a bad thing. Rather, the presence of the errors is a reflection of the near-optimal compromise that occurs in modern scientific research between positive results, false-positive

errors, negative results, false-negative errors, and research costs.

### 3.5. The Idea of a Reasonable Alternative Explanation

The preceding discussion refers to the idea of a reasonable alternative explanation of a scientific research finding. This idea is central in the study of relationships between variables in scientific research. This is because finding reasonable alternative explanations helps us to find errors or weaknesses in the research, which might have led us to an incorrect conclusion. If we find such errors or weaknesses, we can then address them in subsequent research. Also, reasonable alternative explanations of research findings sometimes lead us to unexpected insights, which lead to new knowledge.

In view of the importance of reasonable alternative explanations, diligent researchers do their best to design their research projects to eliminate the possibility that such explanations will arise. For example, medical researchers routinely use placebo control groups and they use double-blinding in clinical research to eliminate two reasonable alternative explanations that generally arise if the two procedures aren't used. And diligent researchers spend significant amounts of time trying to think of reasonable alternative explanations of both their own research findings and the research findings of others before they will trust a conclusion derived from scientific research data.

The notion of a reasonable alternative explanation is a completely general idea. That is, *any* explanation whatever can be used as an alternative explanation of a research finding as long as the explanation is "reasonable". This includes innovative unusual explanations, provided only that they are reasonable. The relevant scientific community decides through consensus what is reasonable and what isn't, sometimes after much debate.

Unfortunately, many people who aren't experienced with scientific research are unaware of the possibility of reasonable alternative explanations and they think that *p*-values (or other measures of the weight of evidence) make definitive *decisions* about whether an effect is real. People may think this way because it would indeed be convenient if *p*-values could somehow make correct decisions for us.

Of course, *p*-values can't possibly make decisions because they take no account of the possibility of a reasonable alternative explanations of a research finding. Such an explanation might explain why a *p*-value is low, but without the need to reject the null hypothesis. Also, it is always possible that a low *p*-value merely reflects a fluke in the data (though the lower the *p*-value, the less the chance of a fluke). The mistaken belief that *p*-values make decisions reflects a fundamental misunderstanding of the operation of scientific research.

For example, reflecting a common view, Bayarri, Benjamin, Berger, and Sellke (2016, p. 92) write:

> Rejecting the null hypothesis at the 0.05 significance threshold is typically taken to be sufficient evidence to accept the alternative [i.e., research] hypothesis.

They go on to say that this reasoning is erroneous for at least two reasons, and they discuss the two (Bayesian) reasons they have in mind. But they *don't* say that the reasoning is erroneous because it omits the key point that there must be no reasonable alternative explanation for a low *p*-value before it can be viewed as "sufficient evidence" to accept the research hypothesis. Arguably, it isn't permissible to omit this point because omitting it implicitly supports the fundamentally incorrect idea that *p*-values (and parallel Bayes factors) make decisions.

Only humans (or computers with more sophisticated algorithms than mere *p*-values) can make sensible decisions (after carefully evaluating the possibility of alternative explanations). But *p*-values (and other reasonable approaches) can help.

### 3.6. Using Replication to Eliminate False-Positive Errors

As noted in section 3.3, false-positive errors can lead to a substantial waste of resources. Therefore, detection of false-positive errors is important in scientific research. And the methods for detecting and eliminating false-positive errors play an important role in the chain of logic of scientific research. Therefore, let us consider how researchers detect and eliminate false-positive errors.

Reasonable alternative explanations of research findings can be subtle. This means that if we find good evidence of a relationship between variables, and if we can't find a reasonable alternative explanation for the finding, then this *doesn't* imply that no such explanation exists. That is, it doesn't imply that we aren't making a false-positive error.

Thus for any positive research finding there may be a correct alternative explanation for the finding, but we don't (yet) know the explanation. We may not know the explanation because we haven't yet discovered the explanation or because the finding arose through mere chance. In either case, if we accept the finding, we will be making a false-positive error. False-positive errors occur surprisingly often in some areas of scientific research, as illustrated in appendix B.11.

The possibility of false-positive errors leads careful researchers to never claim that their research results "prove" something. Similarly, careful researchers never claim that they have "discovered" something. Instead, we say that the results "suggest" that some conclusion might be drawn. This approach reminds us about the possibility of a false-positive error.

False-positive errors are easy to identify and eliminate. We identify false-positive errors in scientific research through subsequent "replicating" research (ideally performed by another independent researcher) in which we attempt to find independent evidence of the effect under study using a new sample of entities from the population.

Researchers invariably perform appropriate replicating research if a newly discovered relationship between variables is important. If proper replicating research successfully replicates a research finding, this greatly reduces the chance that the finding is a false-positive error. In contrast, if a research project *fails* to replicate an earlier research finding, then this

doesn't *prove* that the earlier finding is incorrect, but (depending on the rigor of the "failing" research) it casts doubt.

The number of successful replications required to satisfy a given research community about the truth of a given research hypothesis depends (in an informal way) on the original believability of the hypothesis in the community and depends on the quality of the work supporting the hypothesis—less believable hypotheses need more successful replications before they will be accepted. And carefully performed and carefully described work needs fewer replications before it will be accepted.

Consider an instructive extreme example: Darrell Bem, an emeritus professor of psychology at Cornell University, found fairly good evidence of extrasensory perception (ESP) in eight different experiments, which he reported in a respected psychological journal (2011). However, because the existence of ESP seems highly unlikely, many readers of his report suspect that the report is reporting false-positive errors. But Bem's research is quite rigorous, so it is difficult to think of a possible reasonable alternative explanation to explain why the research might reflect false-positive errors. Also, it seems unlikely that the results of all eight experiments are flukes because that is statistically unlikely. Also, it seems unlikely that Bem's results reflect scientific fraud because Bem is highly experienced, so he knows the substantial consequences of fraud.

Bem's results are certainly thought-provoking. However, they *don't* imply that we have to believe that ESP exists. Instead, as with all new scientific research, experienced researchers won't believe Bem's results until (if ever) the results are successfully replicated. Bem himself stresses the importance of replication of his results in a subsection of his paper (2011) titled "Issues of Replication". And he offers "replication packages" with detailed information to make it easier for other researchers to duplicate his research to replicate his results.

To date, it appears that nobody has unequivocally replicated Bem's results although, in view of the substantial ramifications of discovering that ESP is real, several researchers have tried. (An up-to-date list of reports of research that has attempted to replicate Bem's research can be found by searching the Science Citation Index or Scopus for journal articles that cite Bem's article. It is also sensible to search the web for reports of replication attempts because reports of failures to replicate a positive result are less likely to be accepted for publication in a scientific journal, but will be published on the web if deemed important, as discussed in appendix J.)

The fact that apparently nobody has unequivocally replicated Bem's results suggests that his research is an instructive example of scientific knowledge accruing at its normal slow pace, with false-positive errors being weeded out when no successful unequivocal replications are reported. And we may never know why Bem's research made what appear to be eight false-positive errors. But, although the cause of the (apparent) false-positive errors is interesting, it is less important, and the important result is that there is presently no replicable evidence that ESP exists.

Of course, for completeness, it is possible that some day somebody will find a way to reliably replicate Bem's results. For example, some researcher may discover that there are certain natural "fields" (e.g., magnetic fields) in Bem's laboratory that must be present to enable ESP. And he or she may discover that Bem's results can be replicated in any laboratory provided that the appropriate fields are present. This discovery will open a new world of knowledge through ESP. This discovery is entirely possible, although most of us think it is unlikely. And we suspect that Bem's results are false-positive errors with a logical explanation, but that explanation is presently unknown (and may be unknown for eternity). Time may tell.

### 3.7. The Asymmetry of Statistical Hypothesis Testing

In scientific research we can never conclude that a particular null hypothesis is *definitely exactly* true. For example, we can never conclude that ESP is definitely exactly impossible. So experienced researchers never "accept" a null hypothesis.

However, we *assume* that a given null hypothesis is true until (if ever) someone proves otherwise because that is a sensible parsimonious way to begin. But even if we have absolutely no evidence that a given null hypothesis is false, we can't therefore conclude that this null hypothesis is *definitely* true. This is because it is always possible that the effect of interest exists in the population, but the effect is weak, so we haven't yet successfully detected it. But researchers will reliably detect the effect with an improved research approach sometime in the future.

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

Rao and Lovric (2016) attempt to prove *analytically* that every null hypothesis is false. However, 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 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 a small amount of (real) extra good fortune if he or she regularly carries a lucky coin. That is, it *conceivable* that some "superstitious" people have actually *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* (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 someone convincingly demonstrates otherwise. Thus we can easily *entertain* the idea of a lucky coin, but we assume that such coins don't work until (if ever) someone provides unequivocal evidence to the contrary.

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

The fact that we can never know whether a null hypothesis is *exactly* true is arguably never a practical problem in scientific research. This is because researchers generally aren't interested in whether a particular null hypothesis is exactly true. And we are generally only interested in conclusively demonstrating that the particular null hypothesis of interest is *false*. This is because the null hypothesis is merely an empty starting point that we hope to escape from.

*If* we can show that a null hypothesis about a relationship between variables is clearly false, then we can use the knowledge of the relationship to predict or control the values of the response variable in new entities from the population, which is often useful. But if we *can't* show that a null hypothesis is clearly false, then it is an error to act as if the relationship or effect is present because we might be deceiving ourselves. Hypothesis tests enable us to determine (using a reasonable convention, and in the absence of a reasonable alternative explanation) if the data provide enough evidence to justify believing that the null hypothesis is false, and therefore the relevant research hypothesis (generally about a relationship between variables) is true.

### 3.8. How Researchers View the p-Value

It is helpful to consider how scientific researchers (as opposed to statisticians) view the *p*-value. Many researchers will agree that you can interpret the *p*-value with the following simple rules:

> If the *p*-value for a sought-after effect is less than 0.05, then you have found reasonable evidence of the effect in the data. And if your finding is interesting enough, clearly described, and has no obvious errors, then a report of the finding will be accepted for publication in a relevant scientific journal (which is what a researcher wants to advance human knowledge and to advance his or her career).
>
> In contrast, if the *p*-value is less than 0.01, then you have found *stronger* evidence of the sought-after effect. In that case, if your finding is highly interesting, clearly described, and has no obvious errors, then the report of the finding will be accepted for publication in a relevant higher-prestige journal.

The preceding rules hide the following important ideas in the concept of "no obvious errors": (a) the idea of a reasonable alternative explanation, (b) the idea of a false-positive error, and (c) the idea that we must confirm that the technical assumptions underlying the *p*-value are adequately satisfied before the *p*-value can be trusted. Furthermore, the rules don't explain what the *p*-value measures although (arguably) that is less important. And although the rules hide some important ideas, they are essentially correct and are easy for a researcher to understand.

Of course, some journals may use different critical *p*-values from 0.05 and 0.01, and some journals may use different criteria altogether. But the intent is invariably the same—to only publish research reports that have sufficient weight of evidence that the discovered effect is real.

Arguably, it isn't necessary for researchers (as opposed to statisticians) to know exactly what the *p*-value measures (which is complicated) provided that they understand the above rules. This is because it is the *function* of the *p*-value that is important—to determine whether we have (in the absence of a reasonable alternative explanation) sufficient evidence of the existence of a relationship between variables (or other effect) in the population of entities under study.

The *p*-value enables us to determine whether we have sufficient evidence that the effect under study is real. The *p*-value is based on a set of rigorous ideas in mathematical statistics. But from the point of view of a researcher, the mathematical details underlying of these ideas are less important, and the complexity of these details tends to obscure the function.

### 3.9. Further Discussion

The preceding subsections discuss how the *p*-value helps us to determine whether or a parameter or test statistic is meaningfully different from the relevant null value, which enables us to determine whether a relationship exists between variables. Section 7 discusses seven alternative approaches to perform the same function. Appendix B expands the ideas in the present section, including (a) further explanation about what the *p*-value measures, (b) further discussion of false-positive and false-negative errors, and (c) discussion of the current controversy about the publication of false-positive errors in the scientific research literature (leading to the "replication crisis" in some fields of scientific research). Appendix C discusses whether there is an *optimal* critical value for a test statistic. Appendix D discusses an approach to teaching *p*-value concepts to beginners.

## 4. Do We Need a Measure of the Weight of Evidence that an Effect Is Real?

The fact that we can sensibly view the *p*-value as a measure of the weight of evidence leads immediately to an important question: Do we *need* a measure of the weight of evidence that an effect (e.g., a relationship between variables)

observed in scientific research is real in the entities in a population? Could scientific research somehow get by without such a measure?

Arguably, we need an objective measure of the weight of evidence that an effect is real. This is because if we don't use *some* measure of the weight of evidence, then we can't be confident that any relationship between variables (or other effect) that we have observed in research data is a *real* (i.e., reproducible) effect in the population and doesn't merely reflect random noise in the sample data. Researchers don't want to believe that a relationship exists between variables if the evidence for the relationship could merely be a reflection of noise. So we need a way to distinguish the signal of a relationship between variables from the inescapable noise in the data.

For example, medical researchers studying the effect of a new drug on patients need a reliable objective measure of the weight of evidence that the observed effect of the drug is a real effect in the relevant population of patients, and isn't merely an effect of random variation in the sample—i.e., an effect that won't be observed in new patients from the population. This is because medical researchers don't want to recommend an ineffective drug because that would put patients' health at increased risk and would waste patients' resources.

We generally can't judge whether an effect is real on the basis of simple intuition because science tries to be as objective as possible because intuition is unreliable. So, in general, we need an objective measure of the weight of evidence that an effect observed in scientific research is a real effect in the entities in the population under study.

An exception to the preceding point is that in some specific cases, especially in the physical sciences, we don't need a measure of the weight of the evidence that an observed effect is real because the evidence of the existence of the effect is so obvious from a graph of the data that there is no doubt that the effect is real. (Of course, in such cases, if we compute the relevant *p*-value, it will be extremely low.) However, modern scientific research generally operates at the leading edge of knowledge where new relationships between variables and other effects are often weak or are complicated and are thus hard to detect. Therefore, we generally need an objective measure of the weight of evidence to assist us to reliably detect real effects.

The measure of the weight of evidence that we use in scientific research might be the *p*-value. But it might also be some other sensible measure of the weight of evidence, as discussed below.

Appendix E discusses another important exception to the idea that we need a measure of the weight of evidence that an effect observed in scientific research is real when we study a relationship between variables.

## 5. The Existence of an Effect Versus the Size of an Effect Versus the Importance of an Effect

It is important to distinguish between (a) determining whether an effect *exists* in a population, (b) determining the *size* (or *strength*) of an effect in the population, and (c) determining the *importance* of an effect. The *p*-value helps us to determine whether an effect *exists*, but it doesn't directly speak to the size (strength) or importance of the effect.

For example, in a large properly designed medical experiment to study a new drug we might obtain a very low *p*-value and therefore obtain very good evidence that the drug has a positive effect on the patients in the relevant population. But we might also find that the *size* of the effect of the drug on the patients is quite small—so small that the effect is of no *practical* significance to the patients and thus isn't worth the cost and side effects of using the drug. It is crucial in scientific research to distinguish statistical significance (indicating whether we have sufficient evidence that the effect under study is real in the first place) from practical significance (indicating whether the effect, if real, is strong enough or important enough to be useful).

If we have confirmed (e.g., through a low *p*-value) that an effect is real in the population, then various measures of the size (or strength) of an effect are available, depending on the types of variables under consideration and depending on other features of the research. For example, the correlation coefficient is a measure of the strength of the straight-line relationship between a pair of continuous variables.

In experimental research a straightforward and easy-to-understand measure of the size of an effect is the simple change in the expected value of the response variable that occurs if a predictor variable is manipulated to have two relevant different values. For example, a medical researcher might report that a daily dose of 125 mg of a certain new blood pressure drug (versus a zero dose) lowers the systolic blood pressure of a certain type of patient by 25 mm on average.

Similarly, if we have confirmed that an effect is real, then the *importance* of an effect is obviously of interest. For we may find good evidence of a relationship between variables *and* we may find that the relationship is a strong relationship. But we may also recognize that the relationship is unimportant in the sense that it has no practical or theoretical ramifications. Then the knowledge of the relationship between the variables is obviously less useful, and the research to study the relationship may even have been a waste of time. Of course, we can avoid this disappointing outcome through careful research planning in which we establish that the expected effect will be important or at least useful if we successfully discover it.

We determine the importance of a relationship (or other effect) by determining its social, theoretical, or commercial implications. For example, the discovery by Jonas Salk of the relationship between his polio vaccine and childhood polio was highly important because it had immediate far-ranging implications for eliminating childhood polio.

Of course, if we *haven't* confirmed that an effect is real in the population, then it doesn't make sense to consider the strength of the effect or to consider the importance of the effect. We must first confirm that an effect is real before it is sensible to consider its other ramifications.

## 6. Problems with the *p*-Value

Although the *p*-value is a reasonable measure of the weight of evidence that an effect it is real, it is subject to several serious problems. These problems lead some researchers to question the usefulness of *p*-values. This section summarizes the main problems.

Section 3.3 gives a standard definition of the *p*-value. This definition is a *conditional* definition, with three complicated conditions. This is highly confusing to beginners, so *p*-values are often misunderstood.

Furthermore, as reflected in the definition, the *p*-value reflects a probability that pertains to the situation in which the null hypothesis is *true*. But researchers are almost always interested in *rejecting* the null hypothesis—in demonstrating that the relevant null hypothesis is *false*. Thus the logic of the *p*-value is roundabout, which is also highly confusing to beginners.

If we use *p*-values to detect relationships between variables, then the *p*-values sometimes make false-positive and false-negative errors. These inescapable (but understandable and controllable) errors add to the complexity of the *p*-value.

Some beginners mistakenly think that the *p*-value is the *probability* that the associated null hypothesis is true. This idea is intuitively sensible, but is incorrect. This error isn't directly a serious problem because, for the same critical *p*-value, the error leads to the same conclusions as we obtain under the correct interpretation of the *p*-value. But indirectly this error is a potential source of confusion because it reflects a fundamental misunderstanding of the logic of hypothesis testing.

Some beginners mistakenly think that the critical *p*-value is the *overall* fraction of the time that a research project will make a false-positive error if this critical value is consistently used. The correct statement is that the critical *p*-value is the fraction of the time that a research project will make a false-positive error *in cases when the null hypothesis is true* (and if the assumptions underlying the *p*-value are adequately satisfied).

Researchers are generally eager to obtain a low *p*-value for their main research hypotheses because a low *p*-value is a widely accepted necessary condition for publication of a new scientific discovery in most statistically oriented scientific journals, as discussed in appendix B.8. Researchers' eagerness to obtain a low *p*-value together with the complexity of the *p*-value makes it prone to usage errors (such as the interrelated errors of cherry picking, data dredging, data hacking, and *p*-hacking). The possibility of these usage errors (which researchers sometimes commit unconsciously) adds to the complexity of the *p*-value.

As discussed in section 3.5, some people are unaware of the possibility of reasonable alternative explanations of a low *p*-value and therefore they mistakenly think that *p*-values make decisions. This thinking leads to further confusion or to outright errors.

As discussed in section 5, beginners sometimes confuse (a) the complicated concept of *statistical* significance associated with the *p*-value and (b) the equally important but independent concepts of *practical* or *theoretical* significance, which depend on the *size* and *importance* of the effect.

The preceding problems with the *p*-value (and other problems) have led to much confusion and controversy about the *p*-value, as noted by Wasserstein and Lazar (2016).

## 7. Alternatives to the *p*-Value

Section 4 concludes that we need a measure of the weight of evidence that an effect observed in scientific research is real. But section 6 discusses how the *p*-value is subject to various serious problems. These problems have led to a sense among some researchers and statisticians that the *p*-value is passé (Morrison and Henkel, 1970; Kline, 2004; Ziliak and McCloskey, 2008; McGrayne, 2011; Nuzzo, 2014; Trafimow and Marks, 2015; Hubbard, 2016).

In view of the problems with the *p*-value, statisticians have invented seven sensible alternative measures of the weight of evidence that an effect observed in scientific research is real. Like the *p*-value, these measures all use variations of the approach of checking whether the estimated value of the relevant parameter is "meaningfully" different from the null value. These measures are the *t*-statistic, the confidence interval, the likelihood ratio, the Bayes factor, the posterior probability that the null hypothesis is true, the *D*-value, and various information criteria. Appendix F compares the *p*-value with the seven measures. These comparisons are a key part of the main argument of this paper. However, the discussion is too long for the body of the paper, so a summary is given here.

Appendix F observes that the various measures, including the *p*-value, are (in situations when they are relevant) all monotonically related to each other in value. (This is because, with other relevant factors held constant, the measures are all monotonically related in value to the effect size.) These 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 will cause the various measures to exhibit exactly the same behavior in determining whether we have enough evidence (in the absence of a reasonable alternative explanation) to tentatively reject the relevant null hypothesis.

The fact that the various measures of the weight of evidence can be calibrated with each other to show the same behavior implies that the seven alternative measures are (when relevant) usually *functionally equivalent* to the *p*-value in operation and in high-level output if equivalent critical values are used. Functionally, the only difference among the eight measures is that they have different scales.

The appendix notes that the seven alternative measures are subject to most of the same problems as the *p*-value. In particular, if we use any of the other measures to measure the weight of evidence (with actual or virtual critical values), then that measure is (with rare but important differences) prone to make the same false-positive errors, false-negative errors, and usage errors (cherry picking, etc.) that occur with the *p*-value. In addition, each measure is subject to various other problems.

For each alternative measure, the appendix presents an argument why the *p*-value is superior to the measure in terms two or three of five criteria, as summarized in table 1.

**Table 1**. Comparisons between the *p*-value and the seven alternative 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 | ✓ | ✓ | = | ✓ | =* |
| *D*-value | ✓ | X | = | = | ✓ |
| information-criterion methods | ✓ | ✓ | ✓ | = | = |

* The Bayes factor and the posterior probability that the null hypothesis is true are occasionally more powerful than the *p*-value, as discussed in sections 4 and 5 of appendix F.

A check mark in a cell of the table indicates that the *p*-value is 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 "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 roughly equivalent on the attribute associated with the column. The information in the table reflects generalizations, and exceptions occur in a few less frequent special cases.

The "more informative" column of the table indicates that the *p*-value is (arguably) more informative than each of the other measures. The *p*-value is more informative because the critical *p*-value gives us a direct estimate of the rate of occurrence of false-positive errors in research in cases when the null hypothesis is true (assuming that the underlying assumptions of the *p*-value are adequately satisfied). The rate of occurrence of false-positive errors is important because these errors are guaranteed to occur some of the time in scientific research, are costly, and their frequency of occurrence is controllable (through the choice of the critical *p*-value).

Some statisticians will disagree that the *p*-value is more informative than the other approaches and they will believe that another measure of the weight of evidence is more informative than the *p*-value. Of course, this depends on how much weight a person puts on the importance of controlling false-positive errors. This paper takes the view that controlling false-positive errors is highly important because these errors lead to a waste of resources. Therefore, both the *p*-value and the critical *p*-value are important information that shouldn't be hidden behind one of the other measures of the weight of evidence.

Appendix F also considers some theoretical arguments why one of the measures might be preferred to the others, but observes that most of the theoretical arguments have flaws or weaknesses.

Appendix F concludes that the *p*-value is superior to each of the other measures of the weight of evidence that an effect observed in scientific research is real. Curious readers are encouraged to read the appendix.

## 8. Conclusions

Many scientific research projects study relationships between variables as a means to accurate prediction or control of the values of the response variable in new entities from the studied population. In such research we need an efficient measure of the weight of evidence that an effect (such as a relationship between variables) observed in research data for a sample is a real effect in the population of entities behind the sample. We need such a measure to avoid deceiving ourselves and others about an effect that may be either (a) non-existent or (b) so weak that we can't (presently) reliably observe it, and therefore it is *in effect* non-existent. This is important because we wish to avoid wasting resources on effects observed in scientific research that aren't real.

A low *p*-value (or another sensible indicator of sufficient weight of evidence) is good evidence that an effect is real *only if* there is no reasonable alternative explanation for this evidence. Some people are unaware of this point and they therefore erroneously think that a measure of the weight of evidence (such as a *p*-value) *decides* for us whether an effect is real, which is a fundamental misunderstanding.

All of the measures of the weight of evidence are prone to various problems and all are somewhat complicated. But (as

discussed in appendix F) in comparison to the *p*-value, the other available measures of the weight of evidence are less informative. And the other available measures are one or more of (a) harder to understand, (b) less general, (c) more arbitrary, or (d) less powerful. Thus, arguably, the *p*-value is the best available measure of the weight of evidence that an effect (typically a relationship between variables) observed in scientific research is real in the entities in the population of entities under study.

If we find good evidence of the existence of a relationship between variables, then we can derive an appropriate model equation for the relationship. If we do everything properly, and if the analysis hasn't inadvertently made a false-positive error, then we (or others) can use the equation to reliably predict or possibly control the values of the response variable in new entities from the population. This ability is useful in many areas of life.

**Supplementary material**.

The supplementary material contains appendices A through M.

**References**


Arbuthnot, J. (1710), "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, 186–90, reprinted in Kendall, M. G. and Plackett, R. L. (eds) *Studies in the History of Statistics and Probability Volume II*, High Wycombe UK: Griffin, 30–34.

Bakan, D. (1966), "The Test of Significance in Psychological Research," *Psychological Bulletin*, 66, 423– 437.

Baker, A. (2016), "Simplicity," *The Stanford Encyclopedia of Philosophy* (Winter 2016, ed. by E. N. Zalta). 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.

Bem, D. J. (2011), "Feeling of the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect," *Journal of Personality and Social Psychology*, 100, 407–425.

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.

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.

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

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

Cohen, J. (1994), "The earth Is Round (p < .05)," *American Psychologist*, 49, 997–1003.

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

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).

—— (1973), *Statistical Methods for Research Workers* (14th ed.—revised and enlarged). In Fisher (1990).

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

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

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Los Angeles: Sage.

Kline, R. (2004), *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioural Research*, Washington DC: American Psychological Association.

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

McGrayne, S. B. (2011) *The Theory that Would Not Die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*, New Haven CT: Yale University Press.

Morrison, D. E., and Henkel, R. E. (eds) (1970), *The Significance Test Controversy*, New Brunswick NJ: Transaction Publishers.

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.

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

—— (1933b), "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.

Nickerson, R. S. (2000), "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy," *Psychological Methods*, 5, 241–301.

Nuzzo, R. (2014), "Statistical Errors: *P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume," *Nature*, 506, 150–152.

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.

—— (1904), "Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relationship 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 from the Cornell University Library.)

Rao, C. R., and Loris, M. M. (2016), "Testing a Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective," *Journal of Modern Applied Statistical Methods*, 15 (2), 2–21.

Student (1908), "The Probable Error of the Mean," *Biometrika*, 6, 1–25.

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

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

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

—— (1991), "The Philosophy of Multiple Comparisons," *Statistical Science*, 6, 100–116.

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 130–133.

Ziliak, S. T., and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press.

# Supplementary Material for "The *p*-value is best to detect effects"

**Donald B. Macnaughton**

## Contents

## Appendix A: Exceptions to the View that Research Projects Study Relationships Between Variables

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, with each entity in the sample associated with one row of the table and with each variable associated with one column. This appendix discusses research projects that don't directly fit into this pattern.

Sometimes in a real data table each entity is associated with *multiple* rows, and sometimes the same variable appears in *multiple* columns. But (without losing any data) we can easily reorganize the table (by appropriately adding or removing rows and columns) so that each row reflects a unique entity in a sample and each column reflects a unique variable. Alternatively, we can easily adjust our identification of the entities and the variables under study so that each row of the table reflects a unique entity and each column reflects a unique variable.

Also, some cells in a data table in real scientific research may not contain the *correct* variable values, but instead will contain incorrect values due to a measurement or transcription errors (which, of course, we strive to eliminate). Also, some cells in a table may not contain the actual variable values, but instead will contain "missing" values, implying that these values are (for some reason) unavailable.

Also, sometimes the data for a research project aren't in a single table, but are in multiple tables. However, in this case, the multiple tables are invariably joined together (at least in effect) into a standard single table before the data analysis begins.

In a degenerate case, 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 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. This case often arises in the historical sciences such as 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 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 groups (super-variables, so to speak), as in exploratory factor analysis and principal components analysis. For example, this approach is often used to organize the set of questions (technically "items") on a psychological test into new super-variables that are called "scales", with the value of the scale for an entity being a weighted sum of the scores for the entity on two or more highly correlated items.

Similarly, in another special case, we have no response variable, and we use the predictor variables in the table to enable 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. For example, this approach may be used in targeted advertising, enabling a company to divide its customers into groups of similar customers on the basis of earlier purchases and webpage visits and then tailor the advertising presented to each customer according to his or her group membership, such as automobile accessories for one group and handbags for another.

The preceding set of exceptions isn't exhaustive, and 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 (e.g., research projects that use multivariate analysis, canonical correlation analysis, path analysis, and meta-analysis). However, (empirical) scientific research projects can invariably be viewed as studying the values of variables, and the exceptions can often be reasonably viewed as special cases of the study of variables and relationships between variables.

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

Section 3 in the body of this paper presents a high-level discussion of the operation of statistical hypothesis testing with *p*-values. The present appendix expands the ideas for two audiences: (a) for less-experienced readers by providing more information about the details and (b) for more advanced readers by providing an integrated view of important basic issues, some of which are contentious. The discussion is a mixture of simple statistical ideas and basic ideas of scientific research.

This appendix focuses on the first step (after cleaning the data) in the study of a relationship between variables, which is to determine whether we have good evidence that a relationship *exists* between the variables of interest in the entities in the population. Of course, we are interested in determining whether a relationship exists because if we can find good evidence that a particular relationship between variables of interest exists then, in the second step, we can study *details* about the relationship. That is, we can study how we might use knowledge of the relationship for accurate prediction or control.

### *B.1. The Research and Null Hypotheses*

We can use hypothesis testing to determine whether we have good evidence that a relationship exists between variables. We perform a hypothesis test by using an appropriate procedure to examine the data in a data table to look for convincing evidence that a particular relationship of interest exists.

As noted in the body of this paper, the standard approach to hypothesis testing begins by partitioning 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 a general version of the phenomenon that we believe exists in the

population, but nobody has yet properly 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 in scientific research the research hypothesis 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 (although that it is beyond the present scope). In either case, the problem is the same—we need to determine whether we have good evidence that the postulated thing in question exists.

For example, 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 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 contrast, the null hypothesis says that the studied drug has *no* effect on the response variable in the patients—there is *no* relationship between the amount of the drug and the response variable in the population of patients.

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

However, the important pragmatic hypothesis that a drug has a positive effect is outside the general machinery of hypothesis testing. And standard formal hypothesis testing ignores the researcher's preference because sometimes 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 opposite effect.

As noted in the body, some statisticians and researchers refer to the research hypothesis in a scientific research project as the "alternative" hypothesis, which is the name that Neyman and Pearson used for the hypothesis in their original discussion of it (1928). The word "alternate" is also sometimes used. However, these terms are vague and carry the strong but incorrect connotation that the research hypothesis is somehow *subordinate* to the null hypothesis—suggesting that the research hypothesis is somehow unimportant. But researchers are almost always interested in demonstrating that their research hypothesis is *true*, and therefore demonstrating that the associated null hypothesis is *false*. This is because if we find good evidence that a new research hypothesis is true, then the statement of the hypothesis becomes (after confirmation by other researchers) a new scientific fact about the population. Thus a given research hypothesis is much more important than the associated null hypothesis—the null hypothesis is merely a sensible but empty starting point that we hope to escape from. Thus referring to the important research hypothesis as the "alternative hypothesis" or "alternate hypothesis" is inappropriate and misleading.

Similarly, some authors refer to hypothesis testing as "null hypothesis significance testing", often using the acronym NHST. This term is arguably inappropriate because it emphasizes the relatively unimportant null hypothesis—the hypothesis that we are invariably trying to escape from. Therefore, it is arguably 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*. Of course, the two approaches in the preceding sentence are logically equivalent, each reflecting the same idea, but with opposite terminology. But the research hypothesis—our cherished theory—is much more important than the empty-of-content null hypothesis. Thus, arguably, the research hypothesis deserves the emphasis. 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 as scientific research project we have *multiple* predictor variables. In this case, the ideas are the same—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 specific research hypothesis and a corresponding null hypothesis. We examine the research data to determine which of the multiple hypotheses (if any) is or are (likely) true in the entities in the population.

### *B.2. 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? Thus we begin the study of a new relationship between variables with the formal assumption that the associated null hypothesis is true.

Beginning with the assumption that the null hypothesis is true helps us to avoid deceiving ourselves about the existence of a relationship between variables that doesn't exist. Humans are prone to believe in relationships between variables that don't exist. This may be because we need a strong belief in something to motivate us to *study* the thing—we need a strong belief in the existence of a relationship between variables before we will study it. (There is no point in studying a relationship if we think it doesn't exist.)

Thus most researchers who study relationships between variables strongly believe at the beginning of a research project that the relationship between variables they are studying exists. However, a certain percentage of the time (possibly higher than 50%, depending on the discipline) we are wrong and (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 unequivocally *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 as to be undetectable in the present research. It isn't possible to distinguish between the case when a null hypothesis is *precisely* true and 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 undetectable). The inability to distinguish between these cases generally isn't a serious problem because if a relationship between variables is so weak that it is undetectable, then it will usually also be so weak that it isn't useful in any reasonable sense.

(Of course, despite the point in the preceding sentence, we would, if possible, always like to know about any weak relationship between variables. This is because if we know that a weak relationship exists, and if this relationship would be important if it were stronger, then we might be able to perform further research to find a way to strengthen it.)

Summarizing, 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 (because if we can show it is true, this will advance our knowledge).

### *B.3. 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 of this paper, 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

$$y = f(x) + \varepsilon \qquad (1)$$

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.

(Also, $y$ may be a vector, and "multivariate" statistical procedures are available to handle the case when the response variable is a vector. But this case almost never occurs in real scientific research due to the substantial increased complexity and due to a general lack of any demonstrable scientific advantage—it is usually more sensible to study each possible response variable on its own.)

The $f(x)$ in equation (1) 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. Of course, a researcher will try to choose a form for $f(x)$ 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.

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 (1) above. 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 reliably predict that the value of $y$ for this entity will be $f(x')$.

The $\varepsilon$ in equation (1) is the "error" term. It reflects the fact that a model equation can almost never predict the value of $y$ perfectly. The $\varepsilon$ represents the difference between the correct (measured) value of the response variable for an entity and the value of the response variable predicted by $f(x)$. (In more complicated cases $\varepsilon$ can be a sum of two or more error terms representing different errors that occur at different levels of the analysis.) 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.

If we properly derive a model equation for a particular relationship between variables, 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 relationships between the predictor variable(s) and the response variable. 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.

The general multiple linear regression model equation is a good basic example of a model equation of a relationship between multiple predictor variables and a response variable in scientific research. It has the following general form:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_qx_q + \varepsilon \qquad (2)$$

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, which are called the "regression coefficients" in this case, and

$\varepsilon$ is the error term of the equation.

Many other forms of model equation are also available. For example, we may use a nonlinear equation, a logistic equation, a generalized equation, a cell-means equation, or some other form of equation, choosing from the many available types of mathematical functions, and choosing whichever

form seems most appropriate to model the relationship between the variables at hand.

### *B.4. Parameters of Model Equations*

As noted in the preceding subsection, the $b_0, b_1, \ldots, b_q$ in equation (2) are the $q + 1$ parameters (regression coefficients) of the equation. Almost all model equations have parameters, and the parameters are assumed to be fixed *numbers.* As noted in the body of this paper, we can estimate the values of the parameters of a model equation through the analysis of appropriate research data. The numeric values of the parameter estimates are important because they help us to detect relationships between the variables and because they help to specify the exact form of the model equation.

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 theoretical 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 it is estimable with any specified precision if we are prepared to spend sufficient 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 methods for *estimating* parameter values so that (if we do everything properly) the estimated values will be as close as possible to the true values.

As noted in the body, statisticians have invented three somewhat-related general methods to provide good estimates (from appropriate research data) of the true values of the parameters of a model equation of a relationship between variables. These methods are the least-squares method, the maximum-likelihood method, and the Bayesian method. The methods work by analyzing relevant data in a data table and using sensible algorithms to compute optimal estimates of the parameter values.

As also noted in the body, if we apply the three methods to a given applicable data table using a sensible model equation, then the methods almost always all 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 "true" model 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. (This because behind the scenes the software uses the estimated values of the parameters to assist with the computation of the *p*-values, as we shall see.) Thus we can (if we follow the relevant 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 [although it is difficult, perhaps impossible, to separate this component from the variation due to (a) and (b)].

Since we have been discussing the "true" values of parameters, it is also useful to consider 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 H 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 I discusses an instructive exception in the physical sciences in which we know the *exact* true values of certain parameters of model equations.

### *B.5. Detecting Relationships Between Variables by Examining Estimated Parameters*

As noted in the body, 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 *inconsistent* with the null hypothesis. If we can demonstrate that the estimated value of a parameter is inconsistent with the null hypothesis, 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) that is (are) associated with the parameter and (b) the response variable.

In determining whether a relationship exists between variables, the relevant parameter is often a multiplicative parameter (typically a regression coefficient) of a term in the model equation. If the relevant null hypothesis is or were true in the population, then the value of this parameter will be a particular "null" value. In the case of a multiplicative parameter of a term in a model equation, the null value of the parameter is usually the value zero.

It is easy to see that if the correct value of a multiplicative parameter of a term in a model equation is or were zero, then if this value is used in the equation, it causes the associated term in the equation to *vanish*. For example, if the correct value of $b_2$ in equation (2) in appendix B.3 is zero, then if this value is used in the equation, it causes the $b_2x_2$ term to vanish from the equation. In contrast, if the correct value of a multiplicative parameter is or were noticeably different from zero, then this implies that the term (or perhaps a similar term) *belongs* in the equation.

Furthermore, in a large but sensible conceptual leap, if the correct value of a multiplicative parameter of a term in a model equation is noticeably different from zero, then this is good evidence (in the absence of a reasonable alternative explanation) that a relationship exists in the population between (a) the predictor variable(s) associated with the term and (b) the response variable. That is, if the correct value of a multiplicative parameter is noticeably different from zero, then this is good evidence that the associated research hypothesis is true, and the associated null hypothesis is false.

For example, suppose we wish to use equation (2) to study a particular relationship between variables. And suppose we collect some relevant data for the variables in the equation from a representative sample of entities from the population. And suppose we analyze the data to estimate the values of the parameters, and we discover that the estimated value of $b_2$ in the equation is substantially different from zero. Then this suggests (in the absence of a reasonable alternative explanation) that the $b_2x_2$ term (or a similar term) belongs in the equation, which suggests that there is a relationship between the predictor variable $x_2$ and the response variable $y$. Thus we can (in theory) determine whether we have good evidence that a relationship exists between variables by checking whether the correctly estimated value of the relevant multiplicative parameter of the relevant term in the relevant model equation is different from zero.

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 significantly different from the relevant null value. For example, in experimental research (as opposed to observational research) researchers often use a procedure called "analysis of variance" to study the relationships between the predictor variables and the response variable. In analysis of variance we can check whether we have good evidence of a relationship between one or more predictor variables and the response variable by checking whether the relevant "*F*-statistic" is significantly different from its null value. The null value for the *F*-statistic is roughly 1.0.

### *B.6. How to Determine Whether a Relationship Exists Between Variables*

Appendix B.4 names three sensible general methods that we can use to estimate the values of parameters of a model equation from scientific research data. And appendix B.5 implies that we can (in theory) determine if there is a relationship between two variables by determining whether the relevant parameter for the relevant term in the model equation of the relationship is different from the null value. 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, although 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 in the preceding subsection) 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.

Statisticians have invented *p*-values to help us to determine whether the estimated value of a relevant parameter of a relevant model equation of a relationship between variables is significantly different from the null value. The *p*-values enable us to "test" whether we have good evidence that a relationship exists between the studied variables in the entities in the studied population. The following subsections expand the discussion of the *p*-value that is given in the body of this paper.

### *B.7. The p-Value*

As noted in section 3 in the body, the *p*-value is a popular measure of the weight of evidence that the value of a parameter in a model equation is different from the relevant null value. If the *p*-value implies that we have good evidence that the parameter is different from the null value, then (in the absence of a reasonable alternative explanation) this is good evidence that a relationship exists between the predictor variable(s) associated with parameter and the response variable in the entities in the population.

Let us revisit the standard definition of the *p*-value that is given in the body:

> **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 the research data, will be as discrepant or more discrepant from the relevant null value as it is with the present data *if* the following three conditions are or were *all* satisfied:
> - the associated null hypothesis is or were true in the population, and
> - we were to perform the research project *over and over*, each time using a fresh random sample of entities from the population of interest, and
> - certain often-satisfied technical assumptions that are required to correctly compute the *p*-value are or were properly satisfied.

It is important to observe that the *p*-value is a 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 pertaining to the undesirable case 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, although various approaches have been proposed, as discussed in appendix F.

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. Therefore (in another sensible conceptual leap), 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.

Thus in the case of a multiplicative parameter for a term in a model equation, the lower the *p*-value for the parameter (and assuming there is no reasonable alternative explanation for the low *p*-value), the more evidence we have that the associated term (or a similar term) belongs in the equation. That is, the lower the *p*-value, the more evidence we have that *a relationship exists* in the population between the predictor variable(s) associated with the term and the response variable.

Statistics textbooks explain methods to correctly compute *p*-values (from an appropriate data table) as evidence of the existence of an effect (usually an effect that is a relationships between variables). The various methods enable researchers to study the many different types of relationships that can exist. The textbooks also explain the underlying technical assumptions for each method. (The assumptions pertain to how the entities must be sampled from the population and pertain to the presumed distribution of the values of the error term in the model equation.)

As noted in the body, we can use statistical software to automatically compute *p*-values. This means that we don't need to understand the mathematical details of the methods and we need only understand the underlying assumptions. Many different software packages can compute the same *p*-values. It is reassuring that if the various mainstream packages all perform a well-established hypothesis test with the same data, then they all report *exactly* the same *p*-value. They also all agree *exactly* about the parameter estimates and about the value for each of the other well-established statistics pertaining to the analysis.

[The conclusions of the preceding paragraph are excepting rare software bugs and are excepting very small differences introduced by numeric rounding, which arise due to (a) mathematically identical but numerically different algorithms used in different statistical software packages, and (b) slight differences in number representation in different computer hardware. These rounding differences typically only occur in the one, two, or three least-significant digits in typical 15-decimal-digit computations, and thus the differences are virtually always ignorable.]

In some unusual cases no appropriate software is available to compute correct *p*-values for the parameters of a studied model equation of a studied relationship between variables. However, in these cases, it is generally 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 sensibly might 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 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, and therefore the approach isn't used. Instead, we compute the probability for the range of all values that are as far as or farther than the parameter estimate is from the null value if the null hypothesis is or were true, which seems most sensible.

Although the probability that a parameter has a *particular* value is zero, we can compute the probability *density* for a parameter at a particular value, which is it different but related concept (and is generally nonzero). Probability densities are used to assist in developing two important measures of the weight of evidence that an effect is real called the "likelihood ratio" and the "Bayes factor", as discussed in appendix F.

### *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, although some use and recommend lower critical *p*-values, as discussed in appendix C. 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.

As suggested in the body, if we use a lower critical *p*-value, then we decrease the false-positive error rate, which is obviously a good thing. But, unfortunately, if we use a lower critical *p*-value, then (with other factors held constant) we *increase* the false-*negative* error rate—i.e., the rate at which we will fail to discover evidence of an effect even though the effect is present in the population. So researchers generally prefer to use *high* critical *p*-values, such as 0.05 or even 0.1, to avoid false-negative errors (and to reduce research costs).

In contrast, journal editors generally specify that the main *p*-value in a paper be less than a somewhat low critical *p*-value (often 0.01 for higher-prestige journals) before a research paper will be considered for publication. This requirement helps editors to reduce the rate of publication of false-positive errors in their journals.

The lower the *p*-value we obtain in a data analysis below the critical *p*-value, the more evidence we have (in the absence of a reasonable alternative explanation) that the associated research hypothesis is true and therefore the associated null hypothesis is false.

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 inferences from the data about effects in the underlying population.

### *B.9. Reasonable Alternative Explanations*

As noted in the body, reasonable alternative explanations play a key role in scientific research. And most modern scientists won't accept a 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 even researcher or research assistant fraud. Also, certain reasonable alternative explanation are generally specific to each field of study.

Unfortunately, there is sometimes 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. And 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, as illustrated graphically in the left-hand panel in the figure in appendix B.12.

If in this situation the researcher *reports* only a single instance of the research project in which a 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, or variations of it such as "data dredging", is sometimes committed by less experienced or less vigilant researchers in their (admirable but poorly reasoned) efforts to obtain a positive result.

It is important to note that it is fully permissible for a researcher to perform a research project 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 *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 usage errors such as cherry picking and data dredging in scientific research through proper training. The training should point out that false-positive errors aren't good for a researcher's career because they are invariably exposed if a research result is important. Also, because false-positive errors can lead to a significant waste of resources, they can lead to strong criticism or censure of the researcher by the research community.

(Pons and Fleischmann were more or less drummed out of chemistry for their expensive apparent false-positive cold-fusion error, as discussed by Huizenga, 1993. This error may have been caused partly by cherry picking and partly by measurement problems.)

As noted in the body, the relevant scientific community decides whether a research finding is believable through evolving informal consensus in the community through formal and informal discussion about the finding. 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 due 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 technique 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.

### *B.10. Positive Results and Negative Results*

As noted in the body, if we perform 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 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.

Negative results occur often in scientific research, but receive much less publicity than positive results. For example, 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 a formal experiment (published in 1982) to look for evidence of a relationship between laetrile and cancer. But the experiment 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 obtained negative results. Therefore, all mainstream medical researchers now believe that there is no beneficial relationship between laetrile and cancer in cancer patients (National Cancer Institute, 2017).

Although negative results occur often, we don't hear much about them except in a few high-profile examples (such as the laetrile example). 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 general 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 when positive effects of laetrile from informal clinical experience were reported. Positive results (when they are correct) often lead to useful applications (e.g., a cure for cancer), but negative results don't.

In view of the lack of interest in negative results, and in view of the many positive results that are vying for the limited space in scientific journals, most 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 J discusses journals and registries that do provide information about negative results. Appendix K discusses some instructive exceptions to the general rule that negative results won't be accepted for publication in a mainstream scientific journal.

### *B.11. False-Positive and False-Negative Errors*

Regardless of which approach we use to decide whether a relationship exists between variables, we must take account of the possibility of two types of errors—false-positive errors and false-negative errors. As noted in the body of this paper, a false-positive error occurs if (through chance or through some other reason) we obtain a positive result and therefore conclude that a particular relationship exists between variables, but actually 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 usually 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 they have an understandable and inescapable bias toward the research hypothesis.

False-positive errors are costly in the sense that if they occur (and if the result is important), then they lead other researchers to try to replicate the research finding in order to confirm and extend our knowledge about the relationship between the variables. But if a false-positive error has occurred and therefore the *null* hypothesis is (actually or in effect) true then, unfortunately, this replication research merely amounts to a wild-goose chase that necessarily must fail—an undesirable (but unfortunately unavoidable) waste of resources.

It is noteworthy that if a false-positive error occurs, but the result is *unimportant*, then nobody may try to replicate the result and therefore the false finding will remain uncorrected in the research literature. This is undesirable, but doesn't do much harm (because the result is unimportant). And if an uncorrected false-positive result someday *becomes* important, then other researchers will try to replicate it, and will fail, and the result will therefore be discredited.

It is (at least in theory) 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

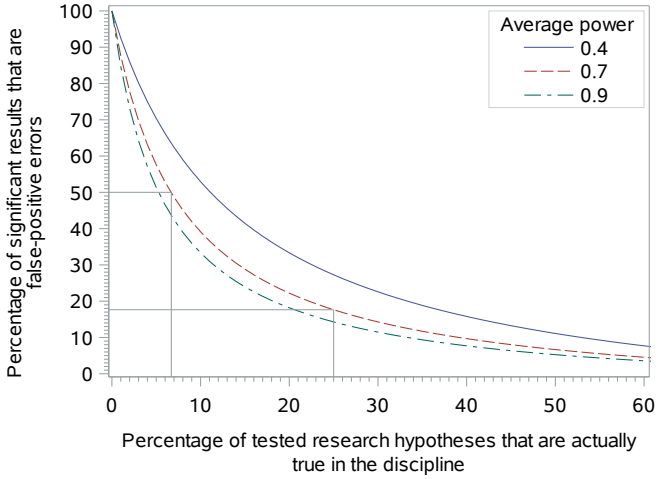


Figure B.1. A graph showing the percentage of statistically significant results that reflect false-positive errors in research projects in a discipline (e.g., in medical research) as a function of the percentage of tested research hypotheses that are actually true in the discipline. The computer code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

The three curving lines on the graph show the percentage of false-positive errors that will occur in a scientific discipline as a function of the percentage of tested research hypotheses that are actually true across the discipline. The lines show this function for three different hypothetical averaged statistical powers of the statistical tests performed in the discipline. For example, if the average power of statistical tests in a discipline is 0.7, then the dashed red line shows the relationship between the variables.

For simplicity, the graph in figure B.1 is based on the assumption that the "average" critical *p*-value used in all research in the discipline is 0.05. However, the graph can be readily redrawn for other assumed average critical *p*-values, 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.)

The graph is based on the assumption that we know the average power of statistical tests that are used in a given discipline. However, in actual scientific research in a given discipline we never know the average power of statistical tests in the discipline. Similarly, we don't know the percentage of tested research hypotheses that are actually true in a given discipline. However, it seems likely that in any given research discipline the percentage of tested research hypotheses that are actually true lies somewhere between 1% and 60%, and it seems likely that the average power of statistical tests in the discipline lies somewhere between 0.4 and 0.9, which are the regions covered by the graph.

The graph is also based on the assumption that research is done without other types of error beyond false-positive and false-negative errors due to chance. But various other errors (e.g., cherry picking, carelessness, fraud) can occur in scientific research, which have a net effect of causing the lines to be somewhat different in reality from the lines on the graph. But the lines on the graph *resemble* reality, even if they don't reflect it perfectly, and thus the graph helps us to understand false-positive errors.

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 research projects in the discipline is 0.7, and if we use a critical *p*-value of 0.05 in the research projects in the discipline, and if there are no extenuating factors, then roughly 18% of the positive research results in the discipline will reflect false-positive errors.

Ioannidis (2005) suggested that more than half of the research findings published in medical research articles reflect false-positive errors. The graph shows that this will be the case if all positive results are published and if the percentage of tested research hypotheses in medical research that are actually true is less than around 6.8% and if the average power of medical research projects is around 0.7, and if medical research projects use a critical *p*-value of 0.05. (Medical research projects generally use a lower critical *p*-value, typically 0.01, or even 0.001. Usage errors also cause some false-positive errors in medical research. Thus the percentage of research hypotheses that are actually true in medical research could be greater than 6.8%, but still yield a 50 percent false-positive error rate.)

As suggested in the body of this paper, the existence of false-positive errors in the research literature isn't a serious problem *if* we keep the possibility of these errors in mind. This is because if a positive scientific research result is potentially important, then we can easily eliminate (well, *almost* eliminate) the possibility that this result is a false-positive error through a successful carefully performed independent replication of the result. Most important scientific research results must be replicated before experienced researchers will believe them (because experienced researchers are highly aware of the possibility of false-positive errors). Thus replication is an important part of the efficient operation of the scientific method.

The preceding paragraphs discuss the idea of a false-*positive* error in a scientific research project. In a similar serious problem, *p*-values sometimes lead us to make false-*negative* errors in research projects. 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 particular 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".

False-negative errors are costly in the sense that if a false-negative error occurs (and if the result is important), then we lose the benefits that we would have obtained if we had discovered the relationship between the variables. In particular, society loses the social or commercial benefits that would have arisen through discovery of the relationship. And the researcher loses the benefits of emotional gratification, honor, and financial reward for discovering a useful new relationship between variables.

Thus both false-positive and false-negative errors are clearly undesirable in scientific research. Unfortunately, both

types of error are always possible in a scientific research project. We deal with the possibility of these errors by using various sensible methods to minimize their occurrence.

In particular, if we use the *p*-value to detect relationships between variables, then we can reduce the rate of occurrence of false-positive errors by using a lower critical *p*-value. That is, the lower we set the critical *p*-value, the lower the rate (across multiple research projects) of occurrence of false-positive errors in the research (but the higher the rate of false-negative errors).

Similarly, for a given critical *p*-value, we can control the rate of false-negative errors by controlling the "power" of the hypothesis tests—the higher we set the power of the tests, the lower the rate of false-negative errors. We can increase the power of hypothesis tests by (a) increasing the sample size, (b) using more precise measurement methods, (c) studying more relevant predictor variables, and (d) using more efficient research designs, as discussed in statistics textbooks about research design.

If we perform a large number of hypothesis tests in a research project, as often occurs in modern data analysis with "big data", then false-positive errors become more likely, 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 a large number of 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).

Of course, if we wish to perform scientific research using a low critical *p*-value and if we wish to use hypothesis tests with high power, then this increases the research cost, so we must strike a reasonable compromise. Experienced researchers use statistical principles to design their research to sensibly control the rate of false-positive errors, while maximizing the power of the statistical tests to detect the sought-after relationships (i.e., minimizing false-negative errors), while minimizing the cost.

### *B.12. 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 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.

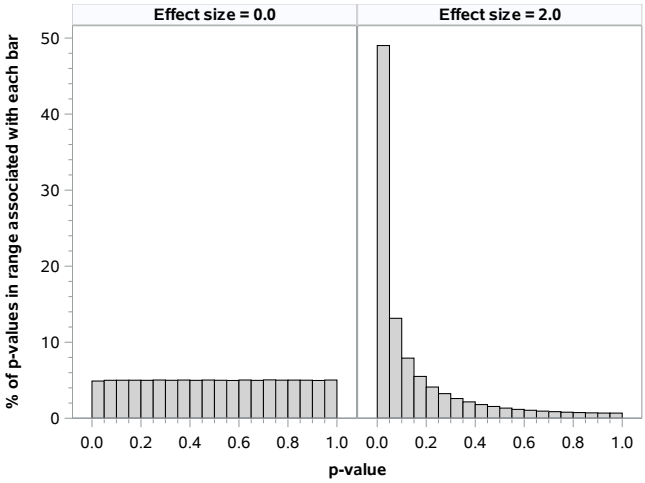


Figure B.2. Two histograms, each showing the distribution of the *p*-values if we repeat a particular research project over and over, each time collecting fresh data. These histograms were computed through a computer simulation. The computer code to generate the histograms (with an explanation of the logic) is in this paper's supplementary material.

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 each histogram is a scale of percentages ranging between zero and 50 percent. If you 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 percent.

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.

(Technically, 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 impercep-

tible deviations from perfect uniformity in the left-hand histogram are artifacts of the necessarily discrete 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 percent 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 percent 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 percent 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 all 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 percent of the time if we repeat the research project over and over.

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 zero point and 05 *even though* the fact that the effect size is 2.0 implies that (behind the scenes) the null hypothesis is false. Thus in this research project we 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 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 in 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 *research design* 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 percent of the time.

Designing research projects to decrease false-negative errors is conceptually easy. That is, assuming that the effect under study is actually real in the population, we can reduce the false-negative error rate in scientific research by (a) using more precise measuring instruments to measure the response and predictor variables, (b) using more relevant predictor variables, and (c) using a more efficient research design (as discussed in statistics textbooks). Of course, generally these approaches increase research costs, so we must compromise to contain costs. In view of these points, diligent researchers spend substantial time designing their research projects to maximize the chance of non-equivocal positive results before they begin any operational work.

### B.13. Do p-Values Make Publication Decisions?

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

Recall that mainstream scientific journals almost always only publish articles that report *positive* results—they almost never publish articles that report *negative* results (because negative results are generally uninteresting). Many journals that report scientific research results include regular discussions of analyses of research data. Given that journals generally only publish positive results, an editor may sensibly use a critical *p*-value as a *screening rule* to determine whether a result is sufficiently "positive" to be considered for publication in their journal (Estes 1997, Cox, 2014, Jager and Leek, 2014). That is, as suggested in appendix B.8, 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 (often 0.01 for higher-prestige journals). Journal editors use a critical *p*-value as a screening rule because this helps them to control the rate of publication of false-positive errors in their journals.

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. And although *p*-values may *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.

### B.14. The Publication of False-Positive Errors

It is an unfortunate fact that (regardless of which method we use to detect relationships between variables) some false-positive errors do slip through a journal's screening process and are published, unrecognized as false-positive errors by both the researcher and by the journal's editors and reviewers. Currently, there is substantial scientific 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).

There are two main causes of the publication of false-positive errors. First, some false-positive errors occur due to random noise in the data that has (by chance) led to a statistically significant result. Not much can be done about these false-positive errors because chance is uncontrollable. Second, some false-positive errors are obtained due to researchers' negligence, such as through cherry picking. We can reduce these false-positive errors through proper training.

False-positive errors are less frequent in the physical sciences than in the biological and social sciences. This may be because there is often a higher signal-to-noise ratio in the data in the physical sciences. That is, real effects observed in data in the physical sciences tend to be far above the noise, and thus are usually clearly real effects.

Appendix B.11 discusses the rate of *occurrence* of false-positive errors in scientific research. The rate of *publication* of false-positive errors in a given field depends on the rate of occurrence of the errors in the field and depends on the critical *p*-value used by journal editors in the field (or depends on some other sensible criterion of the weight of evidence required for consideration for publication, as discussed in section 7 and appendix F). That is, the lower the critical *p*-value used by an editor, the lower the rate of publication of false-positive errors published in the journal.

Unfortunately, it is difficult (arguably impossible) to directly determine the rate of occurrence of false-positive errors 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.

As noted in section 3.6 in the body, we identify false-positive errors by trying to replicate the associated result, but failing. A single failure to replicate a positive result generally isn't definitive in determining that the associated research hypothesis is false because there are usually several possible reasons why a replication attempt failed. For example, certain types of carelessness in research make it likely that a research project will obtain a negative result. Also, slightly different research conditions between the original research and the replicating research may lead to a negative result. Also, a failure to replicate a positive result amounts to a *negative* result and, as noted in appendix B.10, journals generally don't publish negative results because they are less interesting. This explains why journals are generally unwilling to publish a report of a single failure to replicate a research finding.

Of course, most journals will publish a report summarizing *multiple* independent failures of careful research to replicate a particular phenomenon. Then the weight of opinion in the field about the phenomenon will swing back toward the null hypothesis.

The inevitability of false-positive errors in some published research results leads experienced researchers to consider new positive research results with healthy skepticism until the results are properly confirmed (replicated) in independent research. Of course, *important* relationships between variables are always quickly confirmed (or disconfirmed) by other researchers as they try to enhance knowledge about the phenomenon.

### B.15. Comparing Hypothesis Testing with Karl Popper's Idea of Falsification

Let us compare hypothesis testing with Karl Popper's ideas about falsification in scientific research (1980, 1989, 1992). Popper suggested that a theory isn't a valid scientific theory unless it can be falsified. 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 falsification of 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 in theory 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.

However, the 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 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 the opposite relationship—e.g., a decreasing relationship instead of an increasing relationship). (The rareness of discovery of opposite relationships may occur because researchers generally think carefully

about the relationships they study, which makes it less likely that they will find the opposite.)

Thus in actual scientific research we rarely appeal to the idea of falsifying a *research* hypothesis. In contrast, we regularly appeal to the idea of falsifying the *null* hypothesis. That is, in scientific research we provide support for a theory by obtaining good evidence that the relevant *null* hypothesis is false. If (and, arguably, only if) we can convincingly falsify or reject the null hypothesis, then we can accept that a relationship between variables exists in the population and that 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. Actually though, Popper's theory doesn't appear to recognize the null hypothesis, although the concept is easily added to the theory.

### *B.16. Can We Make Any p-Value Arbitrarily Low?*

Demidenko (2016) correctly notes 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 in section 3.7 of the body of this paper.) This point leads Demidenko to suggest that *p*-values aren't useful.

However, Demidenko's point, although possibly theoretically correct, doesn't reflect a practical problem. 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 real 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 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 an observed effect is real in the underlying population. Without that (or 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.

## Appendix C: Is There an Optimal Critical Value for a Test Statistic?

Suppose that we have performed a scientific research project, and suppose that we have found some evidence that the sought-after effect is present in the population. And suppose we have been unable to find a reasonable alternative explanation for this evidence. How can we decide whether the evidence of the effect is convincing enough for us to reject the null hypothesis? Or, in more practical terms, how can we choose the best critical values for a given test statistic? For example, how can we choose the best critical value for the *p*-value, for the Bayes factor, or for the posterior probability that the null hypothesis is true?

*In theory,* there is an optimal critical value for a given test statistic (e.g., an optimal critical value for the *p*-value) used in any field of science. This is the critical value that, if used consistently, maximizes the total benefit-cost ratio of all scientific research performed in the field.

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 optimal value. These attributes include (a) the rate of study (in good but incorrect faith) of false research hypotheses in the field, (b) the payoff of positive results in the field when such results are obtained, and (c) various other attributes, such as research costs in the field.

Unfortunately, it appears that we can't reasonably measure the attributes (a), (b), and (c) in any field of science in a practical sense. Therefore, it is apparently impossible to know the optimal critical value for a test statistic in a field of science according to the preceding ideas. However, it is sensible to be aware of these ideas because arguably they represent the ideal.

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 on the basis of 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, a researcher who thinks that the critical *p*-value of 0.01 is too high can opt to use a critical *p*-value value of, say, 0.005 in his or her research or in his or her interpretation of other researchers' research.

As noted in the body of this paper, in the case of the *p*-value, in standard situations many researchers agree that it is sensible to use a critical *p*-value of 0.05 or 0.01. Similar considerations about conventional values apply for the Bayes factor (Kass and Raftery 1995; Spiegelhalter, Abrams, and Myles, 2004, p. 55) and the other measures of the weight of evidence, although statisticians haven't agreed on conventional critical values for some of the measures.

In situations where comparisons are possible, it is interesting to compare the different measures of weight of evidence (when used with their conventional critical values) in terms of their false-positive and false-negative error rates. If we do this, we see that the conventional critical values for the Bayes factor are stricter than the conventional critical *p*-values. This implies that, for a given research design, the Bayes factor with a conventional critical value will make fewer false-positive errors than the *p*-value with a conventional crit-

ical value. But, of course, this also implies that the Bayes factor will make more false-*negative* errors than the *p*-value if both use their conventional critical values.

As noted in appendix B.8, researchers prefer low false-*negative* error rates (i.e., they prefer critical values that *aren't* strict) because this makes it easier and less expensive 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 helps to reduce the 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 report of a research project is to be accepted for *consideration* for publication in a journal. This is the first step toward being accepted for *publication* in the journal. Generally, each journal that is statistically oriented will indicate that a paper will only be considered for publication in the journal if the relevant test statistic is equal to or better than the journal's critical value. For example, nowadays higher-impact journals that are statistically oriented generally require that the main *p*-value in a research project be less than or equal to 0.01 before the report of the research project will be considered for publication.

Some statisticians recommend lower conventional critical *p*-values (Johnson, 2013; Bayarri, Benjamin, Berger, and Sellke 2016; Johnson, Payne, Wang, Asher, and Mandal 2017). This 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 sufficient weight of evidence to consider the paper for publication.

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 lower critical values in statistical hypothesis tests. This will lead to fewer false-positive results in the literature (although 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 the present appendix, it is difficult or impossible to evaluate "too many" objectively.

Arguably, it is sensible to be somewhat lenient in setting critical values for publication in scientific hypothesis testing so that interesting results can be published and therefore brought to light. This allows other researchers to know about relationships between variables that *may* exist. This suggests that setting the critical *p*-value at 0.005 may be too strict.

Of course, even if we use a very strict critical value for a test statistic, we must keep firmly in mind that a certain (generally unknown) percentage of published positive research results that satisfy this critical value are actually false-positive results, as shown in the figure in appendix B.11. However, the presence of these false-positive results in the research literature isn't a serious problem because, as noted, if a research result is potentially important, then we can easily eliminate (well, *almost* eliminate) the possibility that the result is a false-positive result if somebody performs a successful careful independent replication of it.

We must replicate a research result even if the result is highly statistically significant. We must replicate to reduce the possibility that the result is a fluke, and to reduce (through a re-examination of the issues) the possibility that there is a reasonable alternative explanation for the result. Arguably, regardless of how low the relevant *p*-value is, we must *always* properly replicate a research result before we can trust it because mistakes happen.

Thus we see again that the *p*-value (and each other measure of the weight of evidence) doesn't make decisions in any important sense. Instead, the *p*-value is merely a sensible measure to help researchers to decide what to believe. The decision about whether an effect exists is made by the relevant scientific community. The community makes the decision by evaluating a broad range of information, often including (a) relevant *p*-values (or other sensible measures of the weight of evidence), (b) knowledge of successful or unsuccessful replications, and (c) careful consideration of possible alternative explanations.

## Appendix D: Teaching *p*-Value Concepts to Beginners

The *p*-value is arguably the most commonly used method for detecting relationships between variables in scientific research. Therefore, if a person wishes to understand the study of relationships between variables, then he or she must understand the *p*-value.

The concept of the *p*-value is somewhat complicated. Therefore, proper understanding of the *p*-value is a pons asinorum (bridge of fools) on the road to a proper understanding of statistics. Unfortunately, some beginners fail to cross the bridge.

However, a proper understanding of the *p*-value is *guaranteed* if a student studies enough real-life examples of scientific research projects that study relationships between variables when the null hypothesis is and isn't rejected. Of course, 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. 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.

It is also important for students to study examples of reasonable alternative explanations and examples of false-positive and false-negative errors. Again, the examples should be practical so that students can see the payoff of proper scientific research.

It is arguably less appropriate to teach beginning students the underlying mathematics of *p*-values or the mathematics of model equations. This is because the mathematical ideas are somewhat complicated, and the computer can readily do all the math. Instead, beginning students must understand the *function* of the *p*-value and the *function* of the model equation in scientific research. Therefore, it is sensible to focus on (a) the usefulness of relationships between variables for accurate prediction or control, (b) the use of the *p*-value (or other valid methods) to detect relationships between variables, (c) the use of a properly derived model equation of a relationship between variables for accurate prediction or control, and (d) many practical examples of relationships between variables with obvious payoffs so that students' thinking about relationships becomes intuitive.

In discussing a practical example of a relationship between variables, it is helpful to present beginning students with well formatted computer output. The first part of the output should show five or so rows of the data that were used in the analysis. The columns of the data should be clearly labelled (with carefully chosen names, typically multiple words) to help to ensure that students will understand the response variable and the predictor variable(s) that are under study. Students must understand the nature of the data being analyzed or the analyses will be disconnected from the real world and thus harder to understand. Usually showing only five rows together with the count of the actual number of rows in the full table is enough because the aim is to give students the *gist* of the data, not to overwhelm them with a sea of numbers.

Next the output can show the results of the analysis of the data, showing descriptive statistics, test statistics, *p*-values, possibly other measures of the weight of evidence that an effect is real, and carefully drawn 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 some of the statistics are included for thoroughness, but are much less important than others.

For students who like to think, it is recommended that they be given exercises to design research projects in areas of interest to them. They should first choose a response variable they would like to learn to predict or control. Then they can choose one or more predictor variables that the response variable might be related to. (Students need guidance here because some variables they choose are impractical.) Then they can decide whether they must merely *observe* the predictor variable(s) in an observational research project or whether they can manipulate 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, specifying how the entities in the sample will be selected from the population, specifying how the response and predictor variables will be measured, specifying the detailed steps to perform the research project, discussing expected outcomes, and discussing possible alternative explanations for any results they might obtain. Students' proposed research projects can be presented to the class and constructively criticized by the teacher and by the other students, showing students how scientific logic works.

If there is enough time, and if the students' research projects are performable, then the students can actually perform the research projects and obtain relevant data. Then they can, with the teacher's help, analyze the data to look for evidence of the sought-after relationships between variables.

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. But from a practical point of view, programming is surprisingly complicated. This is because there are many minor but necessary details in 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 the programming details tend to become the center of attention as students strive to master them. But understanding the 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 E: A Case When We Don't Need a Measure of Weight of Evidence

The discussion in section 4 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 appendix discusses an instructive exception.

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 have a continuous response variable. If a variable *isn't* a continuous variable, then it is a "discrete" variable, having a (usually finite) discrete set of possible values, perhaps as few as only two 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, which amounts to a decrease in the precision of the predictions made by the model equations. (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 actually 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, although 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 researchers generally strive to eliminate unnecessary predictor variables from a model equation if the response variable in the equation is a continuous 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.

Consider now the case when the response variable is a discrete variable. If a discrete response variable has a small number of possible (discrete) 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 a large number of possible values (e.g., more than 100 or so values), then things change.

In particular, 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 a particular 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 deriving the model equation. For example, in handwriting recognition, the predictor variables are a set of variables that describe an image of handwriting, and 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 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 a large number of possible values). Arguably, this unifying view of pattern-recognition problems 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 hand-held electronic devices and personal computers. Users have found that using speech-recognition software to enter text and commands is significantly more convenient than typing the text 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 by the software inside the computer and is typically actually 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).

The relationships between variables in pattern-recognition software may mathematically emulate the complicated electrochemical relationships between variables 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 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 point 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 a large number of possible values. And

pattern-recognition systems generally take account of *all* of the chosen 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. And *all* the chosen predictor variables are included in the network of equations because each predictor variable will play a relevant role some of the time.

In view of the preceding points, this paper holds that methods for detecting good evidence of the existence of relationships between (a) one or more predictor variables and (b) a response variable are useful. But it acknowledges that such methods are unnecessary in some cases, such as the case with a discrete response variable when this variable has a large number of possible values, as in pattern-recognition problems.

## Appendix F: Details About Alternatives to the *p*-Value

This appendix compares the *p*-value with seven other sensible measures of the weight of evidence that an effect observed in scientific research is real in the population of entities under study.

First, we consider some similarities among the measures. All of the measures of the weight of evidence (including the *p*-value) are similar in the sense that (with certain limitations) they can all be easily viewed as performing the same function—as performing a test of the research hypothesis that a certain effect (typically a relationship between variables) exists in the population of entities behind the sample.

In any given situation in which we wish to test a research hypothesis, all the measures of the weight of evidence are also similar in the sense that (when applicable) they are all derived from the estimated *sampling distribution* of the same parameter (or test statistic), which is generally a relevant parameter of the relevant model equation, as discussed in the case of the *p*-value in appendices B.4–B.8. In view of the mathematical details of how these measures are derived (from research data) from the same estimated parameter sampling distribution, it is easy to show that these measures (including the standard *p*-value) are all monotonically related to each other.

That is, the expected values of these measures (and the actual estimated values) will all change in monotonic synchrony with one another if (with other things being constant) the *size of the effect* under study were somehow made larger or smaller in the population. (The measures are, in effect, all connected to each other by a complicated set of mathematical gears.) Therefore, although the *scales* are different, the measures are all roughly *functionally equivalent* in providing a measure of the weight of evidence that an effect observed in scientific research data is a real effect in the underlying population.

The monotonic relationships between the various measures generally aren't *linear* relationships. But the relationships are all smooth, as illustrated graphically in appendix G.

It is noteworthy that 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 relationship between in 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 given *p*-value and the Bayes factor as the effect size changes. The differing relationships between the measures of the weight of evidence are due to other factors that play a role in the relationships. For example, in the case of the Bayesian approaches, the prior distribution and the sample size both play roles in the relationships. (The sample size is involved through the Jeffreys-Lindley paradox, which reflects the fact that certain Bayesian measures are dependent on the sample size, as discussed in appendix L.)

Furthermore, with certain important exceptions (discussed below), the measures of the weight of evidence usually all have roughly the same statistical *power* for finding good evidence that an effect is real in the population (assuming that we use an "equivalent" critical value with each measure).

It is important to ask whether any of the alternatives to the *p*-value can escape from the problems with the *p*-value that are discussed in section 6 in the body of this paper. Unfortunately, due to performing the same function, and due to the monotonic relationships between the other measures and the *p*-value, it is easy to see that many of the same general problems arise with the alternatives to the *p*-value (though sometimes the problems assume different forms). For example, all of the approaches sometimes make false-positive and false-negative errors and all are prone to usage or interpretation errors, just like the *p*-value.

Of course, the problem of the definitional complexity of the *p*-value isn't directly present with the other measures because they are differently defined. But a parallel problem of definitional complexity arises across the board—each measure is somewhat hard to understand.

All of the measures of the weight of evidence can use a "critical value" to help us to decide whether we have enough evidence that an effect is real. (The concept of "critical value" takes a somewhat different form with confidence intervals, but is logically equivalent.) If the value of the measure is on one side of (or equal to) the chosen critical value, then (by widely accepted convention) this is good evidence (in the absence of a reasonable alternative explanation) that the effect under study is real in the population—good evidence that we can reject the null hypothesis. In contrast, if the value of the measure is on the *other* side of the critical value, then we have *no* good evidence that the observed effect is real.

All of the measures of the weight of evidence are based on certain underlying assumptions. In particular, a key assumption underlying all the measures is that the entities in the sample are sufficiently representative of the entities in the population of interest for us to draw meaningful conclusions from the sample data about the population. Ideally, the sample should be a "random" sample from the population because then if everything is done properly, correct generalizability is mathematically all but guaranteed. Statisticians have derived

efficient approaches to draw random samples from populations to assist in the study of relationships between variables, as discussed in statistics textbooks.

Although random sampling is preferred, it is somewhat expensive, so researchers sometimes use a "convenience" sample, especially if the population is reasonably homogeneous. In this case, generalizability is more tenuous. However, we can safely conclude that the results of an analysis can be generalized to other entities that are "sufficiently similar" to the entities in the sample, although the idea of "sufficiently similar" is somewhat vague.

Also, if the response variable is continuous, often a key assumption underlying the computation of measures of the weight of evidence that an effect is real is that the "errors" in predictions made by the model equation are uncorrelated and have a normal distribution with constant variance. (Procedures also exist for cases when this assumption is violated.)

Fortunately, the relevant assumptions underlying a properly applied measure of the weight of evidence are often adequately satisfied in proper scientific research. However, regardless of which method(s) we use to study a relationship between variables (or to study some other effect), we must confirm that the relevant assumptions underlying the method(s) are adequately satisfied before we can trust the analysis. Unfortunately, some beginners are unaware of the underlying assumptions of statistical methods and therefore use the methods without confirming that the assumptions are adequately satisfied. This can lead to serious errors, as illustrated by Macnaughton (2016).

Fortunately, confirming that the underlying assumptions of a statistical method are adequately satisfied is becoming much easier. This is because some modern statistical software performs the necessary computations for checking the assumptions as an automatic default part of the analysis, presenting the results of the computations in an easy-to-understand carefully labelled format in the computer output. This makes it much easier for users (and encourages them) to check whether the underlying assumptions are adequately satisfied. For example, the main linear regression program in the SAS system now (by default) automatically computes and displays a sensible set of regression diagnostic plots, and can automatically compute many other diagnostic plots and statistics on request. Some other software systems feature similar automatically generated or easy-to-generate diagnostic information, as one can see by checking software manuals.

Let us compare the *p*-value with the other measures of the weight of evidence to try to decide which one is best. The following seven subsections consider comparisons between (a) the *p*-value as a measure of the weight of evidence and (b) each of the seven alternatives to the *p*-value as a measure of the weight of evidence that an effect observed in scientific research is real.

### *F.1. Student's t-Statistic*

Consider Student's *t*-statistic as a measure of the weight of evidence that an effect observed in scientific research data is real in the underlying population. The *t*-statistic is the ratio of (a) the distance of the estimated value of the relevant parameter from the relevant null value to (b) the estimated standard error of the parameter estimate.

Since the null value of a parameter is typically zero, the distance of the parameter estimate from the null value is typically simply the estimated value of the parameter itself. Therefore, the *t*-statistic is typically simply the ratio of the parameter estimate to its estimated standard error. Of course, Student's *t*-statistic is the mathematical signal-to-noise ratio for the effect under study.

(The *t*-statistic was invented by Gosset [1908] who wrote under the pen name of "Student" to help to hide his invention from his employer's competitors.)

Researchers in the physical sciences sometimes use Student's *t*-statistic (perhaps not naming it) as a measure of the weight of evidence that an observed effect is real in the underlying population, sometimes using a critical value of 2. If the absolute value of the *t*-statistic for a parameter estimate is greater than the critical value of 2 (i.e., the estimated value of the parameter is more than two standard errors away from zero—i.e., the signal-to-noise ratio for the parameter is greater than 2), then (by convention, and in the absence of a reasonable alternative explanation) we have good evidence that the effect under study is real.

Some researchers in the physical sciences use critical values for the *t*-statistic that are much higher than 2, sometimes as high as 5, represented as $5\sigma$. Using a critical value for the *t*-statistic of 5 is (assuming 30 degrees of freedom) equivalent to using a critical *p*-value of roughly 0.000023. Using such a high critical value for the *t*-statistic (or using such a low critical value for the *p*-value) generally requires that the researcher use a very large sample in order to obtain acceptable evidence of the existence of an effect, which obviously increases research costs, as illustrated by Della Negra, Jenni, and Virdee, 2012. However, using such a strict critical value makes it highly unlikely that any discovered positive effect reflects a false-positive error.

We can see the mathematical linkage between the *t*-statistic and the *p*-value by noting that most introductory statistics courses explain how (in relevant situations) the *t*-statistic is the mathematical basis for computing the relevant *p*-value (in Student's *t*-test). It is easy to show that the larger the absolute value of the *t*-statistic, the smaller the derived *p*-value. And it is easy to show that using a critical *t*-value of 2 is roughly equivalent (in the sense of providing a positive or negative result) to using a critical *p*-value of 0.05.

Mathematically, the *t*-statistic is substantially easier to understand than the associated *p*-value because the *t*-statistic comes first in the mathematical development of the ideas. And the *p*-value (in the *t*-statistic case) is a mathematical extension that is built atop the *t*-statistic.

However, the *p*-value is slightly purer than the *t*-statistic as a measure of the weight of evidence that an effect is real because the *p*-value takes full account of the sample size (through the relevant underlying degrees of freedom), which the *t*-statistic doesn't do. Thus in mathematical and logical senses, the *p*-value is slightly more sensible than the *t*-statistic as a measure of the weight of evidence that an effect observed in scientific research is real in the underlying population.

Also, the *p*-value is substantially more general than the *t*-statistic because the *p*-value readily operates in areas where the *t*-statistic doesn't apply. In particular, in addition to summarizing the tail-area probability of the distribution of the *t*-statistic, the *p*-value can also summarize the tail-area probabilities of the distributions of other important test statistics that are used in scientific research, such as the *F*-statistic and the chi-square statistic. The *p*-value enables these summaries to all be on the same relatively-easy-to-understand probability scale that ranges between zero and one.

In summary, in cases when it is applicable, the *t*-statistic is a sensible measure of the weight of evidence that an effect observed in scientific research is real in the underlying population, with an easy-to-understand conceptual derivation in terms of the signal-to-noise ratio. The *t*-statistic is an important *waypoint* to understanding the *p*-value. But (a) the fact that the *t*-statistic doesn't take full account of the sample size, and (b) the lack of generality of the *t*-statistic imply that the *p*-value is preferred to the *t*-statistic as a *universal* measure of the weight of evidence that an effect observed in scientific research is real in the members of the population of entities under study.

### *F.2. Confidence Interval*

Consider the confidence interval as a measure of the weight of evidence that an effect observed in scientific research data is real in the underlying population. We can sensibly view the operation of a confidence interval graphically by marking the estimated value of the relevant parameter on a number line and also marking the null value of the parameter on the line. Then we can superimpose the relevant confidence interval (as computed from appropriate research data) on the line, centering the interval on the null value. Then we can examine the line to see if the confidence interval overlaps the estimated parameter value. This is illustrated in appendix G.

(Some researchers use a sensible equivalent way of doing this graphical operation, centering the same confidence interval on the estimated *parameter* value and then checking to see if the interval overlaps the *null* value.)

If we use a confidence interval as a measure of the weight of evidence that an effect is real, then the analogue of the idea of 'critical value' is whether the confidence interval for the parameter estimate (often a 95% confidence interval) overlaps the estimated value of the parameter. If the confidence interval overlaps the estimated parameter value, then this implies that we have *no* good evidence that the parameter estimate is different from the null value in the population. In contrast, if the appropriate confidence interval *doesn't* overlap the estimated parameter value, then we have (in the absence of a reasonable alternative explanation) good evidence that the value of the parameter is different from the null value in the population.

We can see the mathematical linkage between the confidence interval and the *p*-value by noting that using a 95% confidence interval is *exactly* equivalent (in the sense of providing a positive or negative result in the standard two-tail case) to using a critical *p*-value of 0.05. Of course, the fact that 95% and 5% (0.05) add to 100% isn't a coincidence.

The graphical interpretation of confidence intervals substantially assists understanding and suggests that confidence intervals may be easier to understand than *p*-values. However, the *conceptual* derivation of confidence intervals is obscure for beginners because they wonder:

(a) Where does the confidence interval for the parameter actually come from? (Answer: It comes from the inferred sampling distribution and estimated standard error of the parameter estimate, although that isn't easy for beginners to understand.)

(b) How does lack of overlap of the confidence interval of the estimated parameter value pertain logically to rejecting the null hypothesis?

These underlying questions imply that confidence intervals are conceptually more complicated than their graphical representation might suggest.

Furthermore, in the case of simple comparison of means (which in the simplest case is equivalent to the two-sample *t*-test), we may use the idea that confidence intervals overlap *one another*, or the idea that the confidence interval for the *difference* between the means overlaps zero. These different approaches to confidence intervals add to the complexity of confidence intervals because beginners have difficulty shifting among the approaches. In contrast, if we use the *p*-value in these cases, then the individual approaches are all hidden. Hiding the various approaches enables beginners to focus on the important scientific question (of whether a relationship exists in the population between the variables), as opposed to focusing on somewhat complicated and distracting issues of statistical methodology. Beginners should master the scientific ideas first because otherwise the statistical procedures intended to support the scientific ideas seem like complicated unjustified esoteric rituals.

Furthermore, unlike the somewhat-well-known probability scale of the *p*-value, the scale of a confidence interval is the scale of the parameter under study, which in the standard regression case is a rate—a mathematical first derivative. This scale is difficult for beginners to interpret. And, unlike the *p*-value, the scale of the confidence interval generally changes to a different scale for each different parameter, which requires the user to reorient to a new scale each time, which is difficult for beginners.

Of course, the scale of the confidence interval can be converted to the scale of the *t*-statistic, which turns it into an easier-to-interpret standardized scale. But then it is sensible to use the *p*-value associated with *t*-statistic for the reasons given in the preceding section.

Furthermore, the computation of the *p*-value is mathematically straightforward in more complicated research situations such as in the case when we wish to determine if there is a statistical *interaction* between two or more predictor variables with respect to their joint relationship to the response variable. But it is generally difficult or impossible to use confidence intervals to study interactions.

Of course, all the above problems with confidence intervals generally also exist in convoluted form with *p*-values because confidence intervals and *p*-values are tightly linked mathematically. But if we use the *p*-value, and if we interpret

the *p*-value using the definition above in appendix B.7, then the problems with confidence intervals are sensibly hidden from beginners. The problems are hidden under the covering concept of "as discrepant from the null value as …". It is sensible to hide these technical matters from beginners until they have first mastered the relevant scientific ideas because otherwise the technical matters seem to have no purpose. And, arguably, it is sensible, to the extent possible, to hide these technical matters from *everybody*, so that we can focus on the scientific problem, as opposed to the technical details of statistics.

In summary, a confidence interval is a sensible measure of the weight of evidence that an effect observed in scientific research is real in the underlying population, with a somewhat-easy-to-understand graphical interpretation. However, confidence intervals (a) are harder to understand from the point of view of scientific function than *p*-values, (b) lack a center-of-focus common scale like the *p*-value, and (c) are difficult or impossible to use in more complicated situations. Therefore, the *p*-value is preferred to the confidence interval as a universal measure of the weight of evidence that an effect observed in scientific research is real in the members of the population of entities under study.

### *F.3. Likelihood Ratio*

Consider the likelihood ratio as a measure of the weight of evidence that an effect observed in scientific research is real in the underlying population. The next three paragraphs give a technical description of the likelihood ratio, which some readers may wish to skip.

The likelihood ratio for an effect is the ratio of the heights of two estimated maximum-likelihood marginal probability density functions for the parameter at the value of the parameter. These concepts are illustrated graphically in appendix G.

The numerator of the likelihood ratio is the height of the estimated (probability) density function for the parameter for the effect that we obtain if the *null* hypothesis for the parameter is or were true in the population and if we measure the height *at* the estimated value of the parameter. In computing this height, the values of any other parameters in the model equation are set at their maximum-likelihood estimated values.

The denominator of the likelihood ratio is the height of the estimated density function for the parameter that we obtain if the *research* hypothesis for the parameter is or were true *and* if the true value of the parameter in the population is equal to the actual *estimated* value of the parameter *and* if we measure the height of the implied density function at the estimated value of the parameter. As before, the values of the other parameters are set at their maximum-likelihood estimated values.

Under the preceding definition, the numerator of the likelihood ratio is always less than the denominator. This implies that the likelihood ratio always lies between 0.0 and 1.0 (Wackerly, Mendenhall, and Scheaffer, 2008, p. 550).

Some researchers (e.g., Cox, 2006, p. 91) use the inverse of the likelihood ratio discussed in the preceding paragraphs because the inverse is also reasonable. The fact that the likelihood ratio has two possible definitions is unfortunate because it leads to confusion. Therefore, in any discussion of the likelihood ratio it is important to say at the beginning which approach is being used. The present paper uses the approach to the likelihood ratio discussed in the first four paragraphs of this subsection because this approach appears to be somewhat more popular. But the inverse approach is arguably somewhat more intuitive because it is in an increasing relationship with the effect size.

If the null hypothesis is true in a given research situation, then the likelihood ratio will be close to 1.0. In contrast, if the *research* hypothesis is true, then the likelihood ratio will 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, in actual practice, this approach isn't often 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). Of course, computing this fraction amounts to computing a *p*-value. So using a likelihood ratio can be viewed as merely another sensible path to computing an appropriate *p*-value.

However, if we study actual practice in scientific research, 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 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 conventional approaches.

The likelihood-ratio concepts are harder to understand because the ratio of two heights of the estimated marginal probability density functions of a parameter under the two hypotheses is harder to understand than the probability [fraction of the time] that a particular estimated parameter value 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 probability density function (likelihood function) for a parameter.

The likelihood ratio approach is also harder to understand because it uses *two* distributions—the estimated probability density function of the parameter under the null hypothesis and the estimated probability density function under the hypothesis that the population value of the parameter is equal to the value estimated from the sample. In contrast, the *p*-value uses only a *single* distribution—the estimated sampling distribution of the values of the parameter (if the research is repeated over and over) if the null hypothesis is or were true. And the *p*-value is able to keep this distribution in the background, which helps to reduce the perceived complexity.

The likelihood ratio approach may also be used less often because the mathematical distribution of the likelihood ratio is sometimes difficult to compute, and formulas for the distri-

bution 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 of course never happens. Fortunately, these asymptotic approaches give "fairly good" answers for typical sample sizes. However, this leads researchers ask whether "fairly good" is good enough for the particular situation at hand, and there is presently no easy answer to that question.

In summary, the likelihood ratio is a sensible measure of the weight of evidence that an effect observed in scientific research is real in the underlying population. But likelihood ratios are rarely used in practice. The rare direct use of the likelihood ratio together with the greater complexity of the likelihood ratio suggests that the *p*-value is preferred to the likelihood ratio as a measure of the weight of evidence that an effect observed in scientific research is real in the underlying population.

### *F.4. Bayes Factor*

Consider the Bayes factor as a measure of the weight of evidence that an effect observed in scientific research data is real in the underlying population. The next paragraph gives a technical description, which some readers may wish to skip.

The Bayes factor for a parameter of a model equation is a ratio of two numbers that is similar to the likelihood ratio. In the case of the likelihood ratio we determine the two numbers in the ratio by (for each) mathematically *maximizing* the likelihood across all of the parameters of the model equation under the relevant hypothesis. For one of the numbers, the *null* hypothesis is the relevant hypothesis and for the other number the *research* hypothesis (in the specific form directly implied by the data) is the relevant hypothesis. In contrast, in the case of the Bayes factor we determine conceptually nearly the same two numbers by (for each) mathematically *integrating* the likelihood across all of the parameters of the equation under the relevant hypothesis, taking direct account in the integration of the "prior distribution" of each parameter.

The Bayes factor is based on the same set of concepts as the *p*-value, *t*-statistic, confidence interval, and likelihood ratio. But the Bayes factor takes account of an *extra* concept called the "prior distribution(s)" of the parameter(s) of the model equation. The prior distribution of a parameter is the estimated distribution of the parameter that we have somehow obtained *prior* to (or at least independently of) the research project. We include this distribution in the analysis because if the knowledge of the distribution is valid and reliable, then it increases the accuracy and precision of the results of the analysis.

If there are multiple parameters in a model equation, then in a Bayesian analysis we generally specify a prior distribution for each of them.

The idea of the prior distributions of the parameters of a model equation adds another puzzling layer of complexity above the complexity of the previously considered approaches. This extra complexity is also present in the mathematics of the Bayesian approach. This extra complexity makes the Bayes factor and the Bayesian approach substantially harder for researchers to understand.

Furthermore, if we wish to compute a Bayes factor for a parameter of a model equation, then we must *supply* the prior distribution of that parameter and the prior distributions of all of the other parameters in the equation. Many researchers find it difficult to supply prior distributions for the parameters of a model equation because their research is at the leading edge of knowledge in their field, and therefore usually no prior information is available. And many researchers are uncomfortable trying to *guess* what the prior distribution might be, believing that guessing is somewhat arbitrary.

Of course, we do allow guesswork in the initial *framing* of a scientific research hypothesis. This is because (educated) guesswork in framing the hypothesis (or hypotheses) is necessary to define the research. But once we have guessed or postulated the research hypothesis, no more guessing is allowed, And the goal of the research is to determine if our guessed research hypothesis is true.

There is a noteworthy exception to the point in the preceding paragraph: We also allow guesswork in scientific research in "power" computations in which we guess what we think the effect size for a relationship between variables will be. Then we compute the power of a given statistical test for detecting an effect of that size. (The power is the fraction of the time that the test would successfully detect an effect of the specified size if the proposed research project were performed repeatedly.) But standard power computations are used only in research *design*. Thus (unlike the Bayes factor) the guesswork in power computations plays no direct role in the analysis of research data.

Also, the standard approach to computing the Bayes factor requires that we specify a *specific* research hypothesis in which the values of the parameters of the model equation have *specific* numeric values. For example, Bayarri, Benjamin, Berger, and Sellke specify a specific "point alternative" [i.e., research] hypothesis in computing Bayes factors (2016, p. 93). Specifying such a specific hypothesis seems somewhat arbitrary. (Alternatively, we can compute Bayes factors by specifying the relevant values of the parameters under the research hypothesis using the values of the parameters estimated from the data, akin to the approach used in computing the likelihood ratio.)

The Bayes factor is on a scale of odds. Thus if we obtain a Bayes factor of, say, 6.3 in favor of the research hypothesis, and if we have done everything properly, then the Bayes factor is telling us that the odds that the research hypothesis is true are estimated to be 6.3 to 1. Of course, these estimated odds will vary if we repeat the research project over and over, just like the *p*-value will vary, as illustrated in the figure in appendix B.12. In fact, the Bayes factor will vary in close synchrony with the *p*-value (as the effect size varies) because the two concepts are mathematically tightly linked.

Many people find that the concept of odds is somewhat harder to understand than the concept of probability, which is the basis of the concept of the *p*-value. People find odds harder to understand because the concept of "odds" is *built atop* the concept of "probability" in the sense that an odds is the *ratio* of two probabilities.

Of course, the likelihood ratio is also an odds. But the Bayes factor is more versatile because it allows us to bring the

prior distribution into the analysis, which is sometimes an important advantage—an advantage that outweighs the added complexity.

Like the *p*-value, the Bayes factor uses a critical value in the sense that statisticians have published recommended cutoff (critical) values to help researchers to determine if a Bayes factor provides good evidence of an effect. For example, Kass and Raftery (1995, p. 777) suggest that Bayes factors greater than 3 are "positive" evidence that an effect is real and values greater than 20 are "strong" evidence that an effect is real. Similarly, Bayarri, Benjamin, Berger, and Sellke (2016, p. 96) recommend that a Bayes factor must be greater than 16 for us to decide that we have good evidence that an effect is real.

As with the likelihood ratio, the reciprocal of any Bayes factor is *also* a Bayes factor. Of course, if a Bayes factor increases as the effect size increases, then in the reciprocal of the Bayes factor decreases. Good (1958) and Kass and Raftery (1995) use what we refer to below as the "traditional" version of the Bayes factor. This version is in an increasing relationship with the effect size (and is therefore in a decreasing relationship with the *p*-value). However, Jeffreys (1961), Spiegelhalter, Abrams and Myles (2004, pp. 55, 132) and Held and Ott (2016) use the inverse form. Thus, in view of the possibility of ambiguity, it is important for any discussion of the Bayes factor to indicate which version is being used. The present paper uses the traditional version.

It is noteworthy that automatically computed values of Bayes factors are unavailable in some modern commercial statistical software. (If automatically computed Bayes factors are unavailable in a software system, we can often still compute them *manually* through custom programming, but that is somewhat complicated.) However, a few commercial software products (e.g., Stata) can automatically compute Bayes factors, and some specialized software routines can also compute Bayes factors, such as some specialized packages available for the freeware data-analysis language R (Morey, Rouder, and Jamil 2015; Park 2017).

The inability of some commercial data-analysis software to directly compute Bayes factors is noteworthy because software vendors know that they must stay current to remain competitive. And it is relatively easy to compute Bayes factors analytically in standard situations, as noted by Kass and Raftery (1995, sec. 4). The inability of some commercial data-analysis software to compute Bayes factors may be partly due to the fact that it would be somewhat hard to implement the necessary specification of the prior distributions of parameters to the software by the user, and perhaps partly due to low user demand for the ability to compute Bayes factors, perhaps due to the complexity of the concepts, or due to the difficulty in many analyses in choosing reasonable prior distributions that are beyond mere guessing.

It is important to note that the Bayes factor has a significant but rarely relevant advantage over the *p*-value: For a given false-positive error rate, the Bayes factor will always have higher statistical power for detecting relationships between variables than the standard *p*-value if we have a sufficiently precise *informative* prior distribution of the relevant parameter (or test statistic). This is because the Bayes factor takes proper account of the meaningful information in an informative prior distribution, which the standard *p*-value doesn't do.

Thus it is sensible to use the Bayes factor as a measure of the weight of evidence instead of the standard *p*-value if we have an informative prior distribution. (In some such cases, if we have an informative prior distribution, we also can use the alternative data-analysis procedure of meta-analysis.) But, unfortunately, in scientific research we rarely have informative prior distributions for the relevant parameters (or test statistics) because, as noted, in typical scientific research we are working at a leading edge of human knowledge where objective informative prior knowledge is almost always unavailable.

One important case when we do have somewhat reliable prior information is the case of replication research. That is, if we are attempting to replicate an earlier research result, we could use the posterior distribution obtained in the earlier research as the prior distribution in our current research. However, it seems inappropriate to let the research that we are trying to replicate have any influence over the replicating research because (for the sake of impartiality) we would like the replicating research to be completely independent of the original research. So many researchers will agree that the Bayesian approach is ruled out in this case.

In order to remove some of the arbitrariness from the Bayesian approach, some Bayesian statisticians recommend that we use a "noninformative" prior distribution for each parameter, which is a "default" distribution, so we don't have to guess what the prior distribution is. In this case, since there is no situation-specific information in the prior distributions, it is sensible to ask whether we could completely dispense with the prior distributions and instead use the simpler traditional "frequentist" approach.

However, there are certain cases when the Bayesian approach can be shown to be unquestionably superior, such as when we have an informative prior distribution or when (for technical reasons) the relevant *p*-value or parameter estimates can't be computed. Thus it is sometimes quite sensible to use the Bayesian approach instead of the frequentist approach.

However, in view of the complexity of the Bayesian approach, any research project that uses it should *justify* the use. This is because if the researcher *can't* acceptably justify the use of the Bayesian approach in simple language, then (if one accepts the principal of parsimony, and if one doesn't prefer complexity for its own sake) it would be more sensible to use the simpler frequentist approach.

In summary, the Bayes factor is a reasonable measure of the weight of evidence that an effect observed in scientific research is real in the underlying population. But the *p*-value is easier to understand, is not based on an often-speculative prior distribution, and is theoretically as powerful as the Bayes factor in almost all situations. Therefore (except possibly in rare cases with a sufficiently informative prior distribution, or in any other case when the Bayesian approach is demonstrably superior), the *p*-value is preferred to the Bayes factor as a measure of the weight of evidence that an effect observed in scientific research is real in the members of the population of entities under study.

### *F.5. Posterior "Probability" that the Null Hypothesis Is True*

The Bayesian approach enables us to compute the posterior "probability" (or bounds on that "probability") that a particular null hypothesis is true, as discussed by Berger and Sellke (1987), Sellke, Bayarri, and Berger (2001), Wagenmakers (2007, pp. 792-4), and Held and Ott (2016). This idea is intriguing because the *probability* that the null hypothesis is true seems (at least on the surface) substantially easier to understand than the *p*-value.

Some statisticians refer to the probability that the null hypothesis is true as "the probability of the null hypothesis". However, that idea is somewhat ambiguous. So, arguably, it is clearer to refer to the idea as the probability that the null hypothesis is true.

Perhaps the first thing to note about the "probability" that the null hypothesis is true is that many researchers believe that the null hypothesis in any scientific research project is *never* exactly true for an effect in a population, as discussed in section 3.7 in the body of this paper. If the null hypothesis is *never* (or almost never) exactly true, then the probability that the null hypothesis is exactly true is always (or almost always) zero. Thus almost any non-zero estimate of this probability is presumably incorrect.

However, we can bypass this issue by using the idea that the relevant null hypothesis may be "in effect" true. Here, by "in effect" we mean that a very weak relationship may exist between the variables. But if such a relationship exists, then it is too weak for us to detect (at least with our present research), so there is no evidence that the relationship or effect is real, so the null hypothesis is *in effect* true, even though behind the scenes it may not be *exactly* true.

Of course, also behind the scenes, the actual correct "probability" that the null hypothesis is true (or in effect true) is always either 0.0 or 1.0 because in any given research situation we believe that the null hypothesis is either true (or, equivalently, in effect true) or it is false. We are fully confident that either the null hypothesis or the research hypothesis is true and thus the other is false (because one is the logical negation of the other, and due to the logical law of "excluded middle").

Next, it is noteworthy that researchers don't normally think in terms of the *probability* that the null hypothesis is true. This is because, as discussed above, we believe that in reality the null hypothesis is either true (or in effect true) or it is false. So we normally aren't interested in a somewhat vague assessment of the "probability" that the null hypothesis is true. Instead, we want a reasonable approach that will enable us to *decide* (in the absence of a reasonable alternative explanation) whether to believe that the null hypothesis is true (or is perhaps in effect true) or to believe that the null hypothesis is false. (Of course, we generally hope that we can obtain good evidence that the null hypothesis is *false* and, equivalently, our research hypothesis is definitely true because this gives us new knowledge.)

The concept of probability both in scientific research and in statistical theory is usually based on the notion of frequency. Here "frequency" means the fraction of the time that some event will occur if we repeat the relevant conditions to observe the possible occurrence over and over, each time collecting fresh data. Thus many probabilities in scientific research and in the field of statistics (including *p*-values) can be interpreted in terms of the relative frequency that some event occurs (or would occur), based on actual or theoretical counting or enumeration of events. But the posterior "probability" that the null hypothesis is true (as determined by Bayes' theorem) generally isn't determined by frequency considerations or by counting or enumeration of events. So it isn't a standard type of probability.

Suppose that the "probability" that the null hypothesis is true *were* determined by frequency considerations, and suppose that we were to obtain a result indicating that the probability that the particular null hypothesis under study is true is, say, 0.52. Then the frequency considerations would enable us to say that the null hypothesis would be true roughly 52% of the time if we were to perform multiple independent instances of the relevant research project(s). But in the present case we *can't* use the frequency interpretation, so we are faced with the question of what a "probability" of 0.52 means. In view of this issue, the word "probability" is generally in quotation marks in this subsection to remind us that we are considering a different interpretation of probability from the standard frequency interpretation.

Some Bayesian statisticians might argue that the posterior "probability" that the null hypothesis is true *is* a frequency-based probability in a very general sense. But if we adopt this point of view, we still have the problem that this probability doesn't reflect reality in the particular field of scientific research in which we are working. This is because the prior probability that a null hypothesis is true in a given field of science depends on the percentage of tested research hypotheses in the field that are actually true (as discussed in appendix B.11), and that rate almost certainly varies from one field of science to the next. And the rate of study of true research hypotheses will substantially influence the correct value of the probability that the null hypothesis is true in a given research situation.

Unfortunately, the rate of study of true research hypothesis in a given field of science is arguably difficult or impossible to determine. This is because there is (due to sensible cost considerations) insufficient tracking of negative results. But knowledge of the rate of negative results in a given field of science would be necessary to help us to determine the rate of study of true research hypotheses in the field.

Since the probability that the null hypothesis is true in a given field of science depends on the rate of study of true research hypotheses in the field, but since that rate is difficult or impossible to determine, therefore we can't have a reliable indicator of the probability that the null hypothesis is true if we view probability in a frequency sense. So in a given research project in a given field of science we can't expect an estimate of the "probability" that in a particular null hypothesis is true to be a reliable estimate of a meaningful relative frequency.

Given that we can't use the frequency interpretation of probability, how can we interpret the posterior "probability" that a given null hypothesis is true? A sensible approach is to

interpret the "probability" as simply being an abstract *scale* that ranges between zero and one. This scale gives us another measure of the weight of evidence that the research (or null) hypothesis is true. That is, the lower the posterior "probability" that the null hypothesis is true, the greater the weight of evidence we have that the research hypothesis is true.

It is noteworthy that the scale of the posterior "probability" that a given null hypothesis is true has the same range (between 0.0 and 1.0) as the scale of the *p*-value. And it is easy to see that the *p*-value and the associated posterior "probability" that the null hypothesis is true will rise and fall in synchrony with each other across scientific research projects—the lower the *p*-value, the lower the posterior "probability" that the null hypothesis is true (because both measures depend monotonically on the effect size). However, in a given research situation, if both values are computed, the posterior "probability" that the null hypothesis is true will usually be higher than the *p*-value. This fact doesn't reflect a problem because the two scales are addressing the believability of the null hypothesis from different conceptual perspectives, so there is no contradiction in the fact that the two measures of the weight of evidence generally have different (but highly correlated) values.

In a given research situation, if we wish to compute the posterior "probability" that a given null hypothesis is true, then we enter the prior "probability" that the hypothesis is true into the formula provided by Bayes' theorem and we also enter the data from the relevant research project into the formula. The formula takes the two inputs and converts them (in a mathematically sensible way) into the posterior "probability" that the null hypothesis is true. That is, the formula uses the data to improve on the prior "probability" that the null hypothesis is true, thereby giving us the posterior "probability" that the null hypothesis is true. (Typically, we use a computer program to evaluate the formula because the evaluation is somewhat complicated.) If we do everything properly, the posterior "probability" is a better estimate of the actual correct "probability" that the null hypothesis is true.

(Held and Ott [2016] eliminate the need to know the prior probability that the null hypothesis is true by proposing that we compute the *minimum* posterior "probability" that the null hypothesis is true, which is a statistic that can be computed and which is independent of the prior "probability". However, a problem with this approach is that researchers generally don't want to know the *minimum* possible "probability" that the null hypothesis is true. Instead, if a researcher is interested at all in this "probability", then he or she wants a *correct* estimate of the "probability" that the null hypothesis is true.)

In entering the prior "probability" that the null hypothesis is true into the formula, researchers typically have no knowledge of the actual correct value of this "probability", so any value that we enter is typically somewhat speculative. However, if we ask a researcher at the beginning of a research project what they think is the prior "probability" that the relevant null hypothesis is true, they will generally say that they think this "probability" is quite low. This is because if they thought the "probability" was *high*, then they wouldn't be doing the research because there would be no point. That is, if the "probability" that the null hypothesis is true is high, then the research would likely lead to a negative result, so (with rare exceptions, as in equivalence testing) the research would have no potential payoff.

As noted, researchers generally can't reliably know the prior probability that the null hypothesis is true. Therefore, conscientious researchers using this approach generally use a mathematically *vague* prior "probability", such as 0.5. This reduces the chance of biasing the analysis, but still enables Bayes' theorem to operate—enables the formula to compute the posterior "probability" that the null hypothesis is true.

In a given research situation if there is a real relationship in the population between the variables, then the presence of the relationship will generally cause the posterior "probability" that the null hypothesis is true to move away from the prior "probability" to a point that is closer to the actual correct value of zero. However, although the "probability" will often *tend* to move in the proper direction, there will invariably be noise in the data which will make the "probability" also move at random.

We can use the posterior "probability" that the null hypothesis is true as a measure of the weight of evidence that the effect under study exists in the underlying population. Thus we could define a critical value for this "probability" and say that if the "probability" that the null hypothesis is true is less than the critical value, then we can (tentatively) conclude (in the absence of a reasonable alternative explanation) that the null hypothesis is *false* and therefore the research hypothesis is true.

However, presently there appear to be no scientifically defensible critical values for the posterior "probability" that the null hypothesis is true. Therefore, if we wish to use this measure to detect relationships between variables, then we must use our intuition to decide whether the "probability" is low enough to give us enough evidence to conclude that an effect is real. Unfortunately, using intuition here is difficult because, as noted, the scale of this measure is an abstract scale, and isn't a probability scale that we can interpret in the standard frequency sense.

It is conceivable that statisticians will study the mathematical aspects of the posterior "probability" that the null hypothesis is true and will derive sensible critical values for this statistic. (Or statisticians may derive sensible critical values for the movement of the statistic from the prior "probability" or sensible critical values for the *minimum* posterior probability that the null hypothesis is true.)

The critical value for the posterior "probability" that the null hypothesis is true might be based on appropriately controlling false-positive and false-negative errors, or it might be based on other sensible principles. However, in view of the Jeffreys-Lindley "paradox" discussed in appendix L, the critical value (of the posterior probability that the null hypothesis is true) that enables us to control the rate of false-positive errors will vary (slightly) from one research situation to the next, depending on various aspects of the situation, such as the sample size. This variation in the critical value doesn't invalidate the approach. But the variation makes the approach more complicated than the *p*-value approach, which is able to consistently and sensibly use a *single* critical value that (if properly used) has the same logical interpretation (in terms of

false-positive errors) in all hypothesis tests for all sample sizes.

In summary, the Bayesian posterior "probability" that a null hypothesis is true is a sensible measure of the weight of evidence that an observed effect is real in the underlying population. However, the concept of "probability" used with this concept is puzzling because it is different from the standard frequency interpretation of probability. And until someone defines sensible critical values for this "probability", we must guess the appropriate critical values to help us to decide whether we can conclude that an effect is real in the underlying population. Therefore, the *p*-value (with its well-established and easy-to-interpret conventional critical values of 0.05 or 0.01) is preferred to the posterior "probability" that the null hypothesis is true as an easy-to-understand measure of the weight of evidence that an effect (e.g., a relationship between variables) observed in scientific research is real.

### *F.6. D-Value*

Consider the *D*-value (Demidenko 2016) 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 (2016) illustrates in his formulas (2) and (5)] the *D*-value uses exactly 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 including the sample size is inappropriate for the purpose he envisions for the *D*-value.

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 (e.g., a 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 another reason why the *D*-value might be a good 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* of a relationship between variables or equivalently, as Demidenko notes, "effect size on the probability scale" (2016, sec. 3.1). Or, as also suggested by Demidenko, the *D*-value is "the effect size expressed in terms of the probability of group separation" (2016, sec. 6).

If we view the *D*-value as a measure of the strength of a relationship between variables—i.e., as a measure of effect size—then it is similar to and highly correlated with other measures of the strength of an effect, such as eta-squared, omega-squared, and $r^2$, as listed by Sheskin (2007, pp. 129–130, sec. VII). Measures of the strength 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 of the measure. Measures of strength must be independent of the sample size because the strength is a property of the underlying effect in the population, and the strength isn't 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 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.

Despite the fact that 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).

### *F.7. Information-Criterion Methods*

For completeness, another sensible method for detecting relationships between variables is to use an "information-criterion" method, as discussed by Konishi and Kitagawa (2008). These methods yield model equations for relationships between variables that are similar to or identical to the

model equations yielded by the other methods. When used appropriately, these methods are in monotonic relationships with the other methods for measuring weight of evidence that an effect is real in the sense that if we were somehow able to change the size of an effect in the population, then the expected value of the relevant information criterion will change appropriately. However, these methods operate at a different level from the other measures—at a level that is midway between the level of the individual effects and the level of the entire equation. Researchers generally use the information-criterion methods for detecting relationships less frequently than other methods, perhaps because the information-criterion methods don't allow easy computation and control of false-positive and false-negative error rates like some of the other methods.

### *F.8. Graphical Methods*

As suggested in section 4 in the body of this paper, in cases where relationships between variables are strong and relatively simple, we don't need sophisticated statistical techniques to detect relationships between variables. Instead, we can reliably and easily detect strong relationships between variables using appropriate graphs of the relationships (ideally including clearly labelled error indicators—e.g., error bars—on the graphs to show us the nature of the noise in the data). But scientific researchers often study relationships between variables that *aren't* strong or that aren't simple, and such relationships are hard to reliably detect with graphs. In these cases, the statistical methods for detecting relationships discussed above are useful.

### *F.9. Some Theoretical Arguments About the Preferred Measure*

The preceding discussion covers eight sensible measures (including the *p*-value) of the weight of evidence in support of a research hypothesis—in support of the hypothesis that an effect discovered in scientific research is real in the underlying population. The discussion concludes that the *p*-value is slightly superior to the other measures in the sense of being easier to understand, or more general, or less arbitrary. However, despite these advantages of the *p*-value, we can still ask whether one of the measures might in some sense be *theoretically* more correct than the others. That is, does one of the measures gives us the "true" measure of the weight of evidence in favor of the research hypothesis? This section evaluates six arguments why one of the measures might be theoretically superior 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 are seeking objectivity, this first approach is ruled out.

Consider a second argument for 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 opinions about which of the measures of the weight of evidence is most "natural", although the 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 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 interesting argument for why one of the measures of the weight of evidence that an effect is real is superior to the others: Suppose that we choose any given measure of the weight of evidence that an effect is real (e.g., the Bayes factor), and suppose we choose a sensible critical value for the measure. Suppose that we then calibrate all of the other measures to have critical values that correspond to that value in a given research situation. That is, when our chosen measure declares that an effect is statistically significant, then all of the other measures will also declare that the effect is statistically significant. (It is theoretically easy to do this calibration either analytically or through computer simulations due to the monotonic relationships between the various measures.) Then 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 critical value, we will find that the various measures will in some borderline cases disagree with each other about whether there is sufficient evidence to reject the null hypothesis.

This phenomenon is illustrated in the case of the *p*-value and Bayesian approaches by Kass and Raftery (1995, sec. 8.2), Wagenmakers (2007, pp. 792-794), and Held and Ott (2016). The Kass and Raftery example shows that for a given fixed weight of evidence under the Bayesian approach, 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.

Figure 6 in Wagenmakers' (2007) article shows the relationship between the sample size and the posterior probability that the null hypothesis is true. This figure 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, which bypass the problem of correctly specifying the prior distribution because minimum Bayes factors are independent of the choice of the prior distribution. Held and Ott illustrate that the same *p*-value corresponds to a different minimum Bayes factors depending on the sample size.

The preceding three examples correctly show that 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, say, 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, of course, 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 of computing the weight of evidence are inconsistent with the *p*-value, and thus the Bayesian methods are "incorrect".

Thus there are 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 the measures of weight of evidence aren't linear (as illustrated by Spiegelhalter et al, 2004, p. 132) and due to the Jeffreys-Lindley paradox (which is discussed in appendix L). 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.

However, the existence of the inconsistencies doesn't 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). Instead, it only demonstrates that there are smooth inconsistencies in critical values between the measures if we change the sample size or if we change other aspects of the research situation. Arguably, these inconsistencies are trivial. (The differences would *not* be trivial in cases when the differences are demonstrably substantial in a practical sense but, so far, it appears that no such cases have been proposed). Therefore, until demonstrated otherwise, these minor inconsistencies are arguably ignorable. And these inconsistencies certainly don't imply that one of the measures of the weight of evidence is the true measure, and the various other measures are therefore incorrect.

Consider a fourth interesting argument for why one of the measures of the weight of evidence that an effect is real is preferred to the others: It is arguably sensible to say that the preferred measure of the weight of evidence is the measure that has the lowest false-positive and false-negative error rates. (False-positive and false-negative errors are explained in appendices B.11 and B.12). This is sensible because scientists wish to make as few errors as possible in detecting relationships between variables.

Appendix F.4 says that the Bayes factor is more powerful (i.e., they will make fewer false-negative errors for a given false-positive error rate) if we have *informative* prior distributions for the relevant parameters. (The posterior probability that the null hypothesis is true would also be theoretically more powerful in the presence of informative prior distributions if sensible critical values were defined for this probability.) But (because we are often working at the leading edge of knowledge) we are only rarely in research situations in which we have informative prior distributions for the parameters. So in the following discussion we ignore the case of informative prior distributions.

So (ignoring the case of informative prior distributions) does one of the measures of the weight of evidence tend to have lower rates of false-positive or false-negative errors than the others? Interestingly, it turns out that all of the measures can be calibrated to have exactly the same error rates.

That is, in theory (and assuming we lack an informative prior distribution) we can choose critical values for each of the measures so that each measure declares that good evidence of a relationship is present if and only if all of the *other* measures make the same declaration. If the measures are calibrated this way, then in a given set of research situations the various measures of weight of evidence will all make *exactly* the same false-positive and false-negative errors. So none of the measures of the weight of evidence would have a lower error rate than any of the others.

It is noteworthy that if we do such calibration, we will find that the critical values for some (but not all) of the measures will depend on certain other aspects of the research situation. For example, the *p*-value automatically takes sensible account of the sample size, but some of the other measures (e.g., the *t*-statistic and the *D*-value) don't take full account of the sample size. So for the calibration to work perfectly in a given research project, the critical values of these measures will be a function of the sample size in the research project. This makes the mathematics of the calibration somewhat complicated. But the complexity of the mathematics doesn't negate the idea that in a given research situation if the measures are properly calibrated, then they will all make exactly the same declaration about whether we can (in the absence of a reasonable alternative explanation) sensibly reject the null hypothesis.

Of course, the preceding ideas aren't meant to imply that the various measures of the weight of evidence *should* be calibrated with each other. And we can define independent critical values for any of the measures at our sole discretion, provided only that the critical values are scientifically sensible. And if we do define such critical values, then we can readily calculate (sometimes analytically or always with a computer simulation) the rate of false-positive and false-negative errors we will make using these critical values in different research situations.

The preceding ideas imply that (if we ignore the rare case of informative prior distributions) we can calibrate the measures of the weight of evidence to have critical values so that all of the measures will make the same false-positive and false-negative errors. Therefore, under such calibration, none of the measures is superior to the others in terms of its false-positive or false-negative error rates. So (except in rare cases with an informative prior distribution) we can't choose the "best" measure of the weight of evidence that an effect is real on the basis of lower false-positive and false-negative error rates.

Of course, in real scientific research the measures of the weight of evidence generally aren't calibrated with each other so, as noted, they will disagree with each other in some borderline cases. If we are concerned about this, then in any research situation we are free to compute several (or all) appli-

cable sensible measures of the weight of evidence that an effect is real. Then we can compare the computed values against sensible critical values for each measure. This enables us to consider the results from various sensible points of view. This may also help us to decide which of the measures is most reasonable, perhaps on the basis of ease of computation, or on the basis of the ease of understanding, or on some other sensible basis.

Typically, in a scientific research project we will find that the various measures (whether calibrated or not) will *agree* about whether we have enough evidence to reject the null hypothesis. This is because all of the measures are sensible if they are used properly. For example, if we use both the *t*-statistic and the *p*-value to look for evidence of a relationship between variables, and if we use a critical value for the *t*-statistic of 2.0, and if we use a critical value for the *p*-value of 0.05, then then the *t*-statistic and the *p*-value will *almost* always agree with each other about whether there is sufficient evidence to conclude that the studied effect is real. But if we encounter a situation in which some of the measures *disagree*, then this disagreement tells us that we are in a borderline situation, so our conclusions must be tentative until the results are properly replicated.

Thus with the exception of rare cases when we have an informative prior distribution, we see that the various measures of the weight of evidence can be calibrated with each other to make exactly the same false-positive and false-negative errors. This implies that none of the measures of the weight of evidence is superior to the others in the theoretical sense of making fewer errors.

Consider a fifth interesting argument for 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 somewhat strict conventional critical value for Bayes factors make it less likely that the Bayesian approach will make false-positive errors. (But, of course, the stricter critical value make 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.)

The question of whether there is an optimal critical value for a test statistic is discussed further in appendix C.

Consider a sixth interesting argument why one of the measures of the measures of the weight of evidence that an effect is real is superior to the others: Berger and Berry (1988) correctly note that the *p*-value is the fraction of the time that we will get a parameter estimate (or test statistic) 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 is or were true and if we 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 (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, the *p*-value is a report of the estimated false-positive error rate (under the obtained research results) when the null hypothesis is true (and provided that the conditions underlying the *p*-value are adequately satisfied). But if we report the estimated false-positive error rate, then we are reporting the calculated rate of occurrence of parameter estimates as great as *or greater than* the value actually obtained (assuming that the null hypothesis is true). Therefore, if we want to report to the false-positive error rate, we *have* to report the frequency of occurrence of parameter estimates that are greater than (or equal to) the value actually estimated.

Arguably, it is important for researchers to be aware of the false-positive error rate for the statistical test of a given research hypothesis. This is because false-positive errors in research are expensive because if an obtained result is (theoretically, socially, or commercially) important, and if the result is actually a false-positive error, then this leads to wasted resources in trying to replicate or use the imaginary effect. So it is sensible to be attentive to false-positive errors. Therefore, it is sensible to have the measure of the weight of evidence directly *report* the theoretical rate at which false-positive errors will occur under the obtained research results if the null hypothesis is true. Therefore, the *p*-value is sensible.

### *F.10. Which Measure of the Weight of Evidence Is Best?*

The preceding subsection considered some *theoretical* arguments why one of the eight measures of the weight of evidence that an effect observed in scientific research is real might be best. We saw that most of these arguments are inconclusive about whether one of the measures is best. But the *p*-value has the advantage that (through the critical *p*-value) it reports the false-positive error rate, which is useful because false-positive errors can lead to an expensive waste of resources and therefore they must be identified and eliminated.

In addition, the details in subsections F.1 through F.7 suggest that the *p*-value is slightly superior to the seven other approaches in the sense of sometimes being easier to understand (in terms of the count and complexity of the required concepts), or sometimes somewhat more general, or sometimes somewhat less arbitrary. (As noted in subsection F.4, the Bayesian approach is more powerful than the *p*-value approach if we have a reliable *informative* prior distribution, but that occurs only rarely.)

In view of the preceding points, the *p*-value is a sensible and arguably slightly superior universal criterion for determining whether (in the absence of a reasonable alternative explanation) we have good evidence that an effect observed in a scientific research project is real in the population of entities under study.

## Appendix G: The Relationships Between the Measures of the Weight of Evidence that an Effect Is Real

Appendix F discusses eight different measures (including the *p*-value) of the weight of evidence that a real effect (usually a relationship between variables) exists in the entities in the population of entities under study. In general, these measures can all be computed from properly collected scientific research data. Appendix F also says that in a given research situation the various measures of the weight of evidence that a particular effect is real are monotonically related to each other. The present appendix supports this point and shows graphically how some of the measures are computed.

Suppose that we have performed an appropriate research project to study the relationship between a set of predictor variables $x$ and a response variable $y$. And suppose we have collected the values of $x$ and $y$ in a data table. And suppose we wish to study the relationship between $x$ and $y$ with linear regression analysis. Finally, suppose that we wish to determine whether we have good evidence that the predictor variable(s) associated with the $i$th term in the model equation is (are) related to the response variable. We can make this determination by studying the estimated "sampling distribution" of the estimated parameter for the term. Using the notation for equation (2) in appendix B.3, we refer to this parameter as $b_i$.

We begin by computing the estimated value of $b_i$. We also compute the estimated standard error of the sampling distribution of $b_i$. Straightforward methods have been derived to estimate both of these values from properly collected research data in standard research situations. These two values, along with the sample size, the relevant prior information (which is only required for Bayesian approaches), and theoretical considerations enable us to compute the values of the measures of the weight of evidence. (Some measures of the weight of evidence don't apply in some situations, although all of the measures apply in the standard regression situation.)

After we have obtained the estimated value of $b_i$ and its estimated standard error, we can have the computer draw a graph of the sampling distribution of the parameter under the assumption that the null hypothesis is true. This graph illustrates several of the measures of the weight of evidence that the effect under study is real. Figure G.1 illustrates how this distribution might appear for parameter $b_i$ in our research, as computed from our data table.

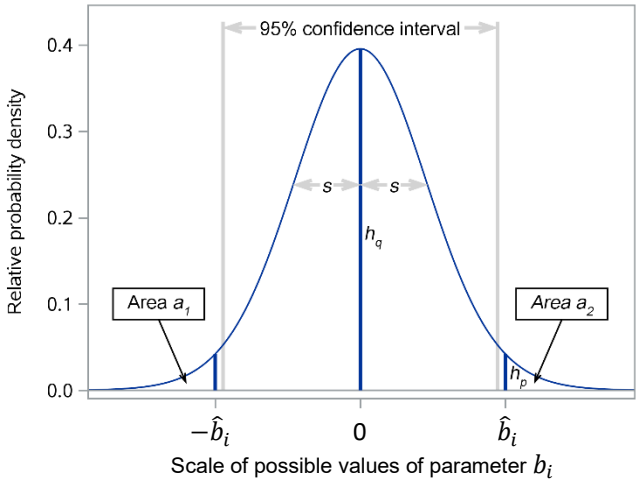


Figure G.1. A graph showing the estimated sampling distribution (estimated probability density function) of a parameter $b_i$ of a linear regression model equation assuming that the relevant null hypothesis is true. The computer code to generate this graph is available in the supplementary material of this paper.

In computing of the estimate of parameter $b_i$ and the estimate of the standard error of the estimate we must decide which *other* terms to include in the model equation because the estimate of $b_i$ and the estimate of its standard error depend on which other terms are included. This is a complicated matter which is dealt with various approaches to "variable selection" in the study of relationships between variables. We ignore the matter here to focus on the single parameter $b_i$. However, we must properly deal with this matter in real research.

The horizontal axis of the graph represents a section of the theoretically possible range of values of $b_i$—the section of the range in which the parameter likely has its true value. The curving blue line shows the estimated sampling distribution function of the parameter under the assumption that the null hypothesis is true. This assumption implies that the estimated sampling distribution is *centered* on the null value, zero, as shown on the graph. The shape of the distribution curve on the graph is mathematically (rigorously) derived from the standard assumptions underlying linear regression analysis, as discussed, for example, by Chatterjee and Hadi (2012).

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

The curving blue line is drawn so that the area under the curve between any two points on the horizontal axis is *exactly* equal to the estimated probability that the value of the parameter estimated from research data will lie between the two

points if the null hypothesis is true. This implies that the total area under the curve (if we go out to "infinity" in both directions) is exactly equal to 1.00 if we use the correct behind-the-scenes height and width units of the graph to measure the area.

The blue line descends from its maximum point evenly on both sides of the null value. This tells us that (assuming that the null hypothesis is true) the estimated values of the parameter might fall 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 the parameter in the research project than values that are farther away from the null value.

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) value of the parameter in the population. And, of course, we would like the null hypothesis to be false because in standard scientific research that is what we are hoping to discover.

The curving blue line was drawn by a computer after it was told (a) the null value (zero in this case), (b) the estimated standard error of the parameter estimate (which defines the "width" of the curve), (c) the type of distribution the (standardized) parameter has (in this case, a central *t*-distribution, as dictated by statistical theory in regression analysis under the standard assumptions) and (d) the "degrees of freedom" of the *t*-distribution.

No numeric values are shown on the horizontal axis of the figure because the numbers are less relevant. It is the *shape* of the curve and the *location of the parameter estimate* relative to the curve that are important. (We consider the location of the parameter estimate in a moment.)

As noted, the curving line on the graph shows the sampling distribution function of a *t*-distribution (under the assumption that the null hypothesis is true). The computer drew this graph under the assumption that the relevant *t*-distribution has 30 degrees of freedom—the degrees of freedom are directly related to the sample size—the larger the sample, the greater the degrees of freedom. (In standard situations, the degrees of freedom is usually between 1 and 10 or so less than the number of values of the response variable in the data, i.e., between 1 and 10 or so less than the sample size.) Of course, the *t*-distribution becomes indistinguishable from a normal distribution if the degrees of freedom is large enough (i.e., greater than 40 or so depending on how fussy we are about "indistinguishable").

The spread (standard error) of the distribution in the figure is the spread that was estimated from the analysis of the research data. The curve on the figure is completely defined by the combination of (a) the identity of the null value (zero), (b) the estimated standard error of the parameter estimate, and (c) the assumption of the *t*-distribution (with appropriate degrees of freedom).

As noted, the computer used the estimated standard error of the parameter estimate to draw the figure. The value of the estimated standard error is shown on the figure as $s$, shown by the two horizontal lines partway up the curve, each indicating the value of $s$, illustrating how the estimated standard error is a measure of the "width" of the curve.

For a *t*-distribution, the value of $s$ is roughly (but not exactly) equal to the horizontal distance from the center of the distribution to either of the two "points of inflection" on the curve. The two points of inflection are the points where the curve changes from curving down to curving up or vice versa.

As noted, in this example we assume that we have performed the research project a single time. Let us assume that the value of the parameter that we half estimated from the data is the value $\hat{b}_i$, which is shown near the right end of the horizontal axis of the figure. The "hat" on the $b_i$ is a standard notation to indicate that the value is the *actual* value that we have estimated from the data, as opposed to the *theoretical* estimated value. 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 blue curve, 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 in appendix F.1, Student's *t*-statistic is the distance of the parameter from the null value in *units of the standard error* of the parameter. So if you use a ruler to measure the distance of $\hat{b}_i$ from the null value of zero on the figure (e.g., measuring the distance in centimeters or inches), and if you measure the length of $s$ (using the same units), then you can compute the *t*-statistic as $\hat{b}_i/s$. And if you actually measure these values on the figure with a ruler, you will see that the value of $\hat{b}_i/s$ in this case is approximately 2.16. In other words, the estimated value of $\hat{b}_i$ is 2.16 standard errors away from zero.

Thus the estimated value of $\hat{b}_i$ together with the estimated standard error $s$ of the sampling distribution causes the *t*-statistic to be a little greater than 2.0, which is a standard critical value for the *t*-statistic, as discussed in appendix F.1. Therefore, (if the figure were based on real data and in the absence of a reasonable alternative explanation) $b_i$ is far enough away from the null value that we have reasonable evidence of the existence of a relationship between the predictor variable(s) associated with parameter $b_i$ and the response variable.

If this were a real research project, we might stop studying the $b_i$ parameter at this point, happy that the *t*-value of 2.16 is greater than 2.0, which implies that we have found reasonable evidence of a relationship between the variable(s) associated with the *i*th term in the equation and the response variable. However, let us study figure G.1 further to see how it illustrates the various measures of the weight of evidence of the relationship between the variables.

Sometimes graphs like figure G.1 are shown with the horizontal axis specified in units of the estimated *standard error* of parameter $b_i$ as opposed to the *raw* units of $b_i$. We use this approach because the standard-error units directly represent the values of the *t*-statistic, which are in effect always the same units and are therefore easier to interpret than the actual units of the parameter, which vary from application to application. It is easy to see that if we were to show standard-error units on the horizontal axis of figure G.1, then the $\hat{b}_i$ near the right end of the axis falls at the value (discussed above) of 2.16 on the axis. (If the horizontal axis represents the values of the *t*-statistic, then the numbers on the *vertical* axis are no longer *relative* values but are the actual values of the height

of the curve—the values that cause the area under the curve to be exactly 1.0.)

Figure G.1 enables us to determine the *p*-value that is associated with $\hat{b}_i$. The *p*-value is simply the sum of the areas of the two "tail" sections of the distribution, where the tail is the area under the curve where the value of the parameter is less than $-\hat{b}_i$ or greater than $\hat{b}_i$. For some readers it may be intuitive, and it is easy to show theoretically, that that sum of the two tail areas (i.e., $a_1 + a_2$) is the theoretical probability (fraction of the time) that the obtained estimated absolute parameter value will be as great as or greater than the value $\hat{b}_i$ obtained in the current research *if* the null hypothesis is or were actually true in the population.

(Of course, the *p*-value will be valid only if the underlying assumptions of the *p*-value are adequately satisfied, and only if there is no reasonable alternative explanation for the low *p*-value. And, of course, the other measures of the weight of evidence have similar assumptions that must be satisfied for the use of the measures to be valid.)

In the example in the figure, the actual sum of the two tail areas (if we properly assume that the graph goes to "infinity" in both directions) is approximately 0.039, as determined by a computer using the relevant formula. Thus the *p*-value for $\hat{b}_i$ in this example is 0.039. The fact that the *p*-value is less than 0.05 implies that if this were a real research result, then (in the absence of a reasonable alternative explanation) we could (tentatively) reject the null hypothesis and conclude that we have reasonable evidence of the existence of a relationship between the predictor variable(s) associated with parameter $b_i$ and the response variable. Of course, this is (due to the direct linkage between the *t*-statistic and the *p*-value) the same conclusion that we drew four paragraphs above from the *t*-statistic.

Some readers may wonder why we compute the *p*-value using *both* tails of the parameter distribution under the null hypothesis. Why not use just one of the tails—i.e., the tail on the side that the actual measured value of the parameter lies? We use both tails because if the null hypothesis is or were true, then it is generally equally possible that the value of the parameter could fall on either the low side or the high side of the null value. Thus in computing the fraction of the time that the parameter estimate will be greater than or equal to the obtained value (under the null hypothesis), it is sensible to take account of both sides of the null value—sensible to include the area of both tails.

Some researchers are tempted to take account of only one of the tails because this has the effect of cutting the *p*-value for this particular parameter value in half, and researchers are almost always eager to obtain lower *p*-values. However, in general, this approach isn't permissible because, as noted, if the null hypothesis is true, then the estimated value of the parameter might lie on either side of the null value.

In figure G.1 the 95% confidence interval is the interval centered on the null value that is labelled "95% confidence interval". Recall that the area under the curve is 1.0 square unit. If we measure the area under the curve inside the lower and upper limits of the 95% confidence interval using the behind-the-scenes units of the horizontal and vertical axes, then we will find that the area is exactly 0.95 square units, as the name is intended to imply.

Assuming that the actual value of the parameter estimated in the research is the value $\hat{b}_i$ on the horizontal axis, we see that in the present example the 95% confidence interval *doesn't* overlap the estimated parameter value of $\hat{b}_i$. Thus, as discussed above in appendix F.2, we can (in the absence of a reasonable alternative explanation) reject the null hypothesis and conclude that we have reasonable evidence of a relationship between the predictor variable(s) associated with parameter $\hat{b}_i$ and of the response variable. Of course, this is the same conclusion that we obtained immediately above using the *t*-statistic and using the *p*-value.

In figure G.1 the likelihood ratio is the ratio of the heights of the density function at the estimated value of the parameter 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 value of the parameter if the null hypothesis is true and if the estimated value of the parameter is $\hat{b}_i$. 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 the parameter in the population is $\hat{b}_i$, then the distribution of the parameter 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 ignore in the present conceptual discussion. Of course, in mathematical treatments we generally take this complexity into account for the sake of mathematical consistency.)

If you measure $h_p$ and $h_q$ on the graph and then compute the ratio of the two heights, you will see that the likelihood ratio is approximately 0.106 (but note the preceding paragraph). It is also easy to see that if the estimated value of $\hat{b}_i$ increases (i.e., if $\hat{b}_i$ moves to the right along the axis), then the value of the likelihood ratio will decrease because $h_p$ will decrease while $h_q$ remains constant (but note the preceding paragraph).

In the case of the Bayes factor, the sampling distribution shown in figure G.1 should be viewed as the estimated marginal *posterior* distribution of the parameter 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 the distribution will often be (at least roughly) bell-shaped and symmetrical about the null value if the null hypothesis is true.

In the case of the Bayes factor, things are somewhat more complicated than the likelihood ratio case. This is because the placement of the center of the distribution under the research hypothesis generally isn't on the estimated value of the parameter. These ideas are illustrated in figure G.2

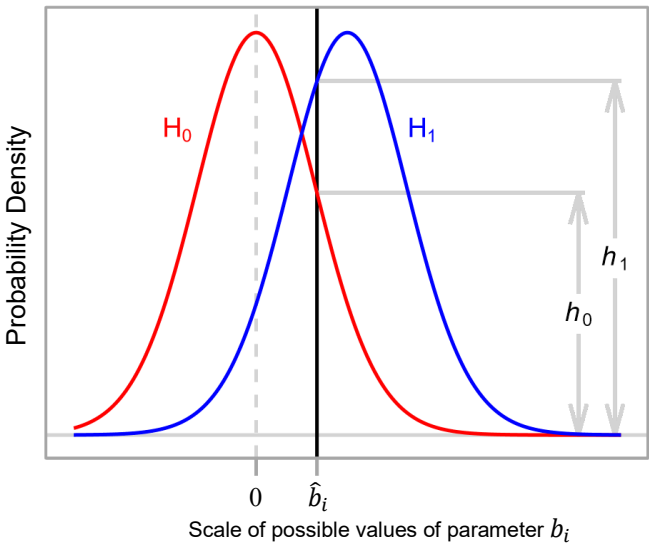


Figure G.2. A redrawn version of figure 2 from an article by Bayarri, Benjamin, Berger, and Sellke (2016, hereafter BBBS) showing their graphical interpretation of the Bayes factor. The BBBS notation has been changed to reflect the notation of the present paper. As noted in the caption of the BBBS figure, the value of the Bayes factor is $h_1 / h_0$. The computer code to generate this graph is available in the supplementary material of the present paper. (Permission to reproduce the BBBS figure is granted in the article.)

The figure helps us to see the complexity of the Bayesian approach because BBBS have drawn the figure with the maximum value of the $H_1$ distribution offset from the observed 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 an air of arbitrariness to the procedure.

We could simplify things and remove some arbitrariness by specifying that the $H_1$ (blue) distribution 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. And we can specify that the width of the distribution should be the width as estimated from the data. Then the Bayes factor would be equivalent to $h_q / h_p$ on the earlier figure G.1 (but based on the estimated Bayesian posterior parameter distributions, and not on the estimated frequentist parameter distributions).

Figure G.2 suggests (somewhat obscurely) the relationship between the Bayes factor and the effect size. That is, if we fix the red and blue curves on the graph, and if we then let $\hat{b}_i$ (which is a measure of the effect size) 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. That is, the relationship between the Bayes factor and the effect size is monotonic. This phenomenon is illustrated in the output from the computer program that generates figure G.2.

It is typically the case that the Bayesian posterior distribution is wider than the associated frequentist parameter distribution, with the relative widths depending on the prior probability distribution of the parameter. If the Bayesian distribution is wider than the frequentist distribution, this causes the estimated value of the parameter (i.e., $\hat{b}_i$ in the example) to be (relative to the distribution curve) closer to the null value, which implies that the standard Bayes factor is generally smaller than the reciprocal of the associated likelihood ratio. The fact that the Bayesian posterior distribution is typically wider than the frequentist distribution is arguably neither good nor bad, but should be noted for proper understanding.

The discussion of figures G.1 and G.2 enables us to see how some of the measures of the weight of evidence are monotonically related to each other. That is, with other things being constant, if the value of $\hat{b}_i$ (a raw measure of the effect size) increases, then the *t*-statistic will increase, the *p*-value will decrease, the parameter estimate will be further outside (or closer to being outside) the 95% confidence interval, the likelihood ratio will decrease, and the Bayes factor will increase. Thus all of the referenced measures are in monotonic relationships with the effect size. This implies, in turn, that all these measures of the weight of evidence are in monotonic relationships with each other (due to the transitivity of monotonicity). Of course, in the Bayesian case we are using a different graph, but the frequentist and Bayesian graphs are linked by the fact that they are both showing the same effect size on the horizontal axis. (The concepts pertaining to the confidence interval are slightly different, but the conclusion is the same.)

In the case of the posterior probability that the null hypothesis is true, we can't picture this probability on one of the figures. However, as noted by Berger and Sellke [1987, equations (2.2 and (2.4)], the posterior probability that the null hypothesis is true is a simple function of the Bayes factor. And if we differentiate the function that expresses the relationship between the posterior probability that the null hypothesis is true and the Bayes factor, we see that the derivative is always negative (or, if we are using the inverse of the standard Bayes factor, always positive) when the prior probability is the permissible range, which implies that there is a monotonic relationship between the Bayes factor and the posterior probability that the null hypothesis is true. This implies, in turn, that the posterior probability that the null hypothesis is true is in a monotonic relationship with all of the other measures of the weight of evidence.

The *D*-value is closely related to the *p*-value derived from figure G.1 because the *D*-value uses the same formula as the *p*-value, except that the sample size in the formula is replaced by the numeral 1, as discussed by Demidenko (2016, sec. 5). Thus, as with the *p*-value, the *D*-value is a function of the *t*-statistic. Thus as $\hat{b}_i$ increases, the tail area of the distribution function for the *D*-value will decrease, and thus the *D*-value itself will decrease. Therefore, the *D*-value is monotonically related to the *t*-statistic, and therefore the *D*-value is (due to transitivity of monotonicity) also monotonically related to the other measures of the weight of evidence. (But, as noted in appendix F.6, the *D*-value is better viewed as a measure of

strength than as a measure of the weight of evidence because it doesn't take proper account of the sample size.)

A similar argument applies to the information criteria—as the discrepancy of the parameter estimate from the null value increases (with other factors held constant), the expected value of the relevant information criterion will decrease in value, which implies that (if the decrease is great enough) the relevant term will be selected for inclusion in the model equation by the algorithm used by the information theory approach to select terms for inclusion in the model equation. Of course, selection of a term for inclusion in the model equation implies that the algorithm has decided that there is enough evidence to (tentatively) conclude that a relationship exists between the predictor variable(s) associated with the parameter and the response variable.

(For completeness, it is noteworthy that this paper hasn't proven that the measures of the weight of evidence are always in monotonic relationships with the effect size. And, in fact, pathological cases exist when non-monotonic relationships occur, such as a Bayes factor based on two central *t*-distributions, one offset from the other. In this case, if the *t*-statistic becomes large enough, then the Bayes factor is no longer in a monotonic relationship with the *t*-statistic, as illustrated in the output from the program used to generate figure G.2. Any measure of the weight of evidence that an effect is real that isn't in a monotonic relationship with the effect size won't be in a monotonic relationship with the other measures of the weight of evidence. It seems less likely but possible that this non-monotonicity could also occur in some non-pathological cases.)

The preceding discussion implies that all of the measures of the weight of evidence are (at least in the standard cases) in monotonic relationships with each other if the effect size increases or decreases under standard assumptions and with other things being equal. Therefore, the various measures of the weight of evidence are in a sense *equivalent* to each other. They are equivalent in the sense that (with some less important exceptions) they could be calibrated with one another to make the same declarations about whether (in the absence of a reasonable alternative explanation) we have enough evidence to conclude that an effect is real. And the main difference between the measures in a practical sense is merely that they operate using different scales.

It is straightforward to generalize the preceding discussion beyond linear regression analysis, although that is beyond the present scope. It seems likely that the monotonic relationships between the measures of the weight of evidence will be present in many or most cases.

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

As noted in appendix B.4, we usually view the true values of parameters of model equations as being fixed values in the population (although the *estimates* of the values generally vary from one research project to the next). The idea that parameters have fixed values is especially evident in the physical sciences, as discussed below in appendix I. 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 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". 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 model 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. Thus although having parameter values that vary 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). Thus arguably we don't need to develop statistical procedures to *directly* handle varying parameter values (unless it can be convincingly shown that the approach with varying parameter values is somehow more efficient than trying to *model* the varying parameter values).

Of course, 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 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 somehow *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 numbers, with the provision that a parameter may have varying values if a significant advantage of that can be clearly demonstrated.

## Appendix I: 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, as implied by the name "constants". Physical scientists have performed various careful research projects to estimate the correct values of these parameters.

However, 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 which, astonishingly, Einstein has shown is equal to the square of the speed of light in a vacuum.

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 (on the basis of earlier estimates and on the basis of 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), and thus this constant value is a reasonable foundation for other physical constants—constants that must be estimated from data. Also, this constant value is in theory relatively easy to reproduce anywhere in the universe, which satisfies our preference for generality. 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 of 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 perceived as an easy-to-understand, relatively easy-to-use, and (hopefully) an unshakable foundation on which measurements in the physical sciences can rest. That is, these specified-by-fiat parameter values and measurement units are the basis for developing the measurement system for physical phenomena

which, in turn, serves as a basis for deriving relationships between variables pertaining to physical phenomena, and for deriving estimates of the values of the *other* parameters (physical constants) of the associated model equations for relationships between variables in entities of the various types of entities that are studied in the physical sciences.

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 developing 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 $F = ma$. Gauss in effect specified that the parameter of this equation is the numeral one (1.0), thereby implicitly specifying a definition of the units of force. It is noteworthy that 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, and he was merely defining the *units* of force.

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

## Appendix J: Approaches to Publishing Negative Results

As noted in appendix B.10, most scientific journals won't accept reports of research results when the main result is a negative result. However, some researchers sensibly believe that negative results should be published because these results tell us what has been tried in research but has failed. Thus because most scientific journals won't publish negative results, some researchers have established journals or registries that enable reporting of negative results. These journals and registries can be found by searching the Internet for "negative results" or "research registry".

The following are arguments in favor of publishing negative results:

- The publication of negative results helps researchers to avoid repeating research that has failed, thereby conserving resources.
- The publication of negative results allows researchers to report innovative methods that might be useful in other research.
- 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 protocol makes it more difficult for researchers to publish serendipitous findings that may have arisen through chance or through data dredging.

The following are arguments against publishing negative results.

- In general, negative results are less interesting than positive results.
- If a new research procedure is truly innovative, then it is generally best not to discuss it in a paper reporting a negative result, but to use it in further research and to publish a description of the innovative research procedure in the report of that research.
- There are many possible reasons to explain why a research project obtained a negative result, including the possibility of carelessness on the researcher's part and the possibility of a simple false-negative error. Thus a negative result doesn't necessarily mean that the relationship between variables under study definitely *doesn't* exist (although some readers may mistakenly interpret it that way).
- It is highly unlikely that any researcher would ever *exactly* repeat an unknown failed research project, and the slight differences between the "repeating" research project and the original research project might lead the second researcher to obtain a positive result.

If a researcher obtains a negative result, and if there is no specific 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 on his or her own website or in a general Internet archive, perhaps announcing the publication in relevant email lists. This enables other interested researchers in the field to learn about the result.

It is sensible for any researcher planning new research to search journals of negative results, research registries, and the Internet for 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, and thus they won't spend the necessary time unless they are somehow coerced. Time will tell whether repositories of negative results are useful enough to justify their cost.

## Appendix K: An Example of the Publication of an Important Negative Result

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 mainstream 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, 2017).

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 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. Ji (2017) discusses a modern example.

## Appendix L: The Jeffreys-Lindley Paradox

The posterior "probability" that the null hypothesis is true (as discussed in appendix F.5) leads to a puzzling paradox. To illustrate, consider the task of assessing from research data whether a regression coefficient in a model equation is different from the null value of zero. Here, the sampling distribution of the parameter of interest is typically sensibly assumed to be a normal distribution with a mean of zero if the null hypothesis is true, and with a mean that is different from zero if the null hypothesis is false. (In this case, for sensible technical reasons, we typically *model* the estimated sampling distribution of the parameter with Student's *t*-distribution.)

In this situation, the posterior "probability" that the null hypothesis is true is (like the associated *p*-value) a function of the relevant *t*-statistic. This function is derived under reasonable assumptions by Berger and Sellke (1987, equation 1.1).

It is of interest to study the relationship between the *t*-statistic and the posterior "probability" that the null hypothesis is true. Figure L.1 shows (for three different sample sizes) the relationship according to Berger and Sellke's equation. The figure is based on the assumption that the prior probability that the *research* hypothesis is true is equal to the prior probability that the *null* hypothesis is true, and (because the two hypotheses are exhaustive) thus both prior probabilities are equal to 0.5.

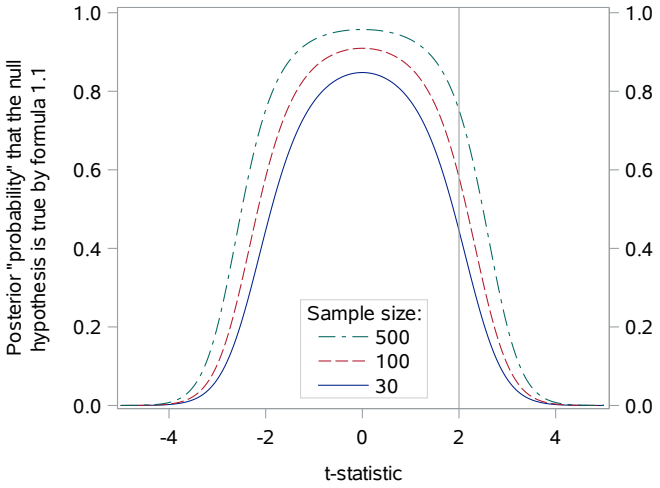


Figure L.1. The relationship between the posterior "probability" that the null hypothesis is true and the *t*-statistic for three different sample sizes assuming that the prior probabilities that the research and null hypothesis are true are both 0.5. The figure was generated using a formula given by Berger and Sellke (1987 equation 1.1). The computer code to generate this figure (with an explanation of the logic) is available in the supplementary material for this paper.

The figure shows that the Berger and Sellke formula 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 formula 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 *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*.

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. 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 *t*-statistic in the present case) to be to the correct value of the parameter in 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*. 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 paradox also raises the parallel 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.

This paradox or contradiction is also illustrated in an article by Held and Ott (2016). Their figure 2 shows that for a given effect size (as approximately reflected in the associated *p*-value) a smaller sample consistently gives a lower Bayes factor, implying that for the same effect size a *smaller* sample gives us *greater* evidence that the null hypothesis is false, which is counterintuitive.

(Held and Ott work with Bayes factors that are the inverse of the standard Bayes factor. The sample size is involved in the computation of the *p*-values in the Held and Ott figure 2, but in the case of the *p*-values a smaller sample gives *less* weight of evidence, not greater, so this point can't somehow negate the points in the preceding paragraph.)

Interestingly, the apparent contradiction behind the Jeffreys-Lindley paradox doesn't rule out the use of the posterior "probability" that the null hypothesis is true as a measure of the weight of evidence that an effect is real. However, the apparent contradiction makes interpretation of the measure more complicated because if we wish to satisfy the sensible goal of controlling the rate of false-positive errors, then (at least in theory—the effect may sometimes be small) we would need to use different critical values for the posterior probability that the null hypothesis is true depending on the sample size.

## Appendix M: Computer Programs

This computer programs that were used to generate the figures in the earlier appendices are available in the Supplementary Information for this paper on the journal's website.

## References


Baker, A. (2016), "Simplicity," *The Stanford Encyclopedia of Philosophy* (Winter 2016, ed. by E. N. Zalta). 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.

Benjamin, D. J., Berger, J. O, …, and Johnson, V. E. (2017), "Redefine Statistical Significance," *Nature Human Behaviour*, http://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.

Berger, J. O., and Berry, D. A. (1988), "Statistical Analysis and the Illusion of Objectivity," *American Scientist*, 76, 159–165.

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.

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

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

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.

Della Negra, M., Jenni, P., and Virdee, T. S. (2012) "Journey in the search for the Higgs boson: The ATLAS and CMS experiments at the Large Hadron Collider," *Science*, 338, 1560–1568.

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

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.

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

Good, I. J. (1958), "Significance Tests in Parallel and in Series," *Journal of the American Statistical Association*, 53, 799-813.

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

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.

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.

—— (2008), "Effect of Formal Statistical Significance on the Credibility of Observational Associations," *American Journal of Epidemiology*, 168, 374–383.

Jager, L., and Leek, J. T. (2014), "An Estimate of the Science -Wise False Discovery Rate and Application to the Top Medical Literature," *Biostatistics*, 15, 1–12.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.) Oxford UK: Clarendon.

Ji, X. (2017) "Dark matter remains elusive," *Nature* 542, 172.

Johnson, V. E. (2013), "Revised Standards for Statistical Evidence," *PNAS*, 110, 19313–19317.

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.

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

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

LeCun, Y., Bengio, Y., and Hinton, G. (2015), Deep Learning," *Nature*, 521, 436–444.

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

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.

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.

Morey, R. D., Rouder, J. N., and Jamil, T. (2015), "BayesFactor package [computer software]" Available for download at https://www.r-project.org/

National Cancer Institute (2017), "Laetrile/Amygdalin – Health Professional Version". Available at https://www.cancer.gov/about-cancer/treatment/cam/hp/laetrile-pdq#section/_1

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.

Park, J. H. (2017), "CRAN Task View: Bayesian Inference". Available at https://cran.r-project.org/view=Bayesian For Bayes factor computing routines search the page for "Bayes factor".

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.

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

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.

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

Sheskin, D. J. (2007), *Handbook of Parametric and Nonparametric Statistical Procedures*, Boca Raton, FL: Chapman & Hall/CRC.

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.

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.

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.

Wikipedia contributors (2017) "Michelson-Morley experiment," Accessed on January 23, 2017.