Understanding Statistical Testing

Peter J. Veazie

Abstract
Statistical hypothesis testing is common in research, but a conventional understanding sometimes leads to mistaken application and misinterpretation. The logic of hypothesis testing presented in this article provides for a clearer understanding, application, and interpretation. Key conclusions are that (a) the magnitude of an estimate on its raw scale (i.e., not calibrated by the standard error) is irrelevant to statistical testing; (b) which statistical hypotheses are tested cannot generally be known a priori; (c) if an estimate falls in a hypothesized set of values, that hypothesis does not require testing; (d) if an estimate does not fall in a hypothesized set, that hypothesis requires testing; (e) the point in a hypothesized set that produces the largest p value is used for testing; and (f) statistically significant results constitute evidence, but insignificant results do not and must not be interpreted as evidence for or against the hypothesis being tested.

Keywords
research methods, data processing and interpretation, hypothesis testing, estimation, inference

Introduction
Current concepts of statistical testing can lead to mistaken ideas among researchers such as (a) the raw-scale magnitude of an estimate is relevant, (b) the classic Neyman–Pearson approach constitutes formal testing, which in its misapplication can lead to mistaking statistical insignificance for evidence of no effect, (c) one-tailed tests are tied to point null hypotheses, (d) one- and two-tailed tests can be arbitrarily selected, (e) two-tailed tests are informative, and (f) power-defined intervals or data-specific intervals constitute formal test hypotheses. In this article, I challenge convention regarding hypothesis testing that leads to such mistaken ideas, and I provide a coherent conceptualization and logic for testing that avoids such mistakes.

A recent book and related works by Ziliak and McCloskey (McCloskey & Ziliak, 1996, 2008; Ziliak & McCloskey, 2004a, 2004b, 2008a, 2008b) declare statistical significance is invalid for scientific inquiry. Critics responded (Engsted, 2009; Hoover & Siegler, 2008a, 2008b; Spanos, 2008). Ziliak and McCloskey, and their critic Spanos, imply the raw-scale magnitudes of parameters are relevant to hypothesis testing. This confuses the goals of hypothesis testing with that of parameter estimation; I provide a careful distinction between the two and argue the raw-scale magnitude of an estimate is irrelevant to testing.

The Neyman and Pearson (1933) approach to addressing hypotheses is often offered as a formal logic of testing (Neutens & Rubinson, 2002b; Portney & Watkins, 2000; Rothman & Greenland, 1998; Spanos, 1999b). Such a presentation does not constitute a general framework; I present a general characterization that allows each competing hypothesis to be a set of parameter values.

The two-tailed test is perhaps the most common test. I argue that, as a formal process, the two-tailed test (indeed, a test of any point hypothesis) is almost never informative. Intervals associated with power, confidence intervals (CIs), and data-specific measures are sometimes offered as defining hypotheses (Mayo, 2010; Mayo & Spanos, 2010; Neyman, 1957). I argue they do not constitute formal test hypotheses.

The goal of this article is to provide a coherent understanding and approach to formal statistical hypothesis testing for the researcher who seeks to use this inference tool without confusion. This article does not discuss alternative methods for using empirical evidence in inference such as the
Hypothesis Testing Versus Parameter Estimation

By the term hypothesis, I mean a formal proposition about which its truth or falsity is unknown. An empirical hypothesis is one for which empirical evidence can, in principle, bear on judgments of its truth or falsity. A statistical hypothesis is an empirical hypothesis about distribution parameters of random variables defined by a data generating process.

To properly understand Frequentist statistical hypothesis testing, it is important to understand that the relevant random variables represent the distribution of possible values that a data generating process could obtain, and not actual data. In this sense, data and corresponding estimates are realizations of the underlying random variables but are not themselves random variables. Hence, the sample mean statistic has a distribution of possible values, whereas the mean of a given sample is a number.

A statistical hypothesis should be stated in terms of distribution parameters of random variables, and not in data-specific terms. If a statement includes reference to data, then it will either not be a hypothesis or it will be uninformative. As an example, consider the claim that “there will be a significant result.” In the first case, in terms of the data generating process, there is a particular probability that what is claimed will occur, and the claim is therefore neither true nor false and thereby not a hypothesis. In the second case, in terms of the resulting data, the claim will be either true or false, but it is a proposition by virtue of a necessary numeric characteristic only (of course the result will be either statistically significant or not). The same hypothesis applies to any data generating process, and knowing whether it is true or false is uninformative regarding the data generating process under investigation.

Hypothesis testing is a process by which we can inform judgments of the truth or falsity of a hypothesis. Formal statistical hypothesis testing is a method that compares data-specific value of a statistic to the statistic’s sampling distribution as implied by the hypothesized values of a statistical hypothesis. There are two largely substitutable methods, in their common usage. One is to define a set of values in the statistic’s range that correspond to sufficiently rare events under the hypothesis-specific distribution (often termed the rejection region); if the data-specific value of the statistic is found to be in this set, the data are considered evidence against the underlying hypothesis. This is a strictly categorical method of testing. The second is to calculate the probability of obtaining data at least as extreme as that actually obtained from the data generating process under the assumption that the statistical hypothesis is true (commonly termed the $p$ value); if the $p$ value is sufficiently small compared with an a priori set level (commonly called the significance level), the data are considered evidence against the hypothesis being tested. This is also a categorical method of testing; however, the $p$ value can also provide a continuous measure of evidence for the hypothesis being tested. The most common use of formal testing is to adopt the categorical approach with its designations of results being “statistically significant” or “statistically insignificant”; I will discuss testing in these terms.

By this definition, the classic Neyman–Pearson test (NPT), in which we set our acceptable Type I and Type II error rates and proceed as if the null were true or false according to our test, is not hypothesis testing: Notwithstanding Neyman and Pearson’s reference to testing, it is a decision rule, as Neyman and Pearson themselves state (Neyman & Pearson, 1933). Its goal is to decide whether or not to act as if a hypothesis is true, not to judge whether the hypothesis is true: The former is suited to model specification; the latter is suited to generating scientific understanding.

However, Fisher’s (1956) approach to statistical inference, in which we use data as evidence for or against the truth of a claim, provides a basis for hypothesis testing by the definition used here. What I provide is in one sense a generalization and in another sense a restriction of Fisher’s approach. It is a generalization because Fisher’s approach has focused primarily on point hypotheses, whereas the logic I present applies to set hypotheses in general. It is a restriction because Fisher’s approach does not explicitly state an alternative, whereas the logic I present addresses sets of hypotheses that partition a parameter space—an idea that Neyman and Pearson (1928a, 1928b) initiated with the introduction of the formal alternative hypothesis.

It is important to distinguish the goals of testing and estimation. The goal of hypothesis testing is to make a judgment regarding the truth or falsity of a hypothesis, whereas the goal of estimation is to make a judgment regarding the value of a parameter.

If I know whether a hypothesis is true or false, I have achieved the goal of hypothesis testing. Suppose I am interested in the hypothesis that “the average annual health care expenditure among men is greater than that among women”: It is either true or false, it cannot be nearly true or mostly false. If an honest omniscient being were to tell me “your hypothesis is true,” then my goal has been achieved. Knowing the magnitudes of the averages or their difference adds nothing more to achieving this goal.

Suppose you are testing the hypothesis that the effect of a policy intervention is zero, and the result is a substantively trivial but statistically significant difference. A criticism is that you have identified a statistically significant yet substantively insignificant effect. Such a critique is not an indictment against your hypothesis test. Your statement that the data provide evidence the hypothesis is false is not
compromised by the raw-scale size of the effect. The two statements, “\(10^{-100} \neq 0\)” and “\(10^{100} \neq 0\),” are both true: There is no degree of truth that varies with the magnitude. What then is the objection? The critic is objecting to your goal of testing the hypothesis and is instead presumably seeking evidence regarding the magnitude of the estimate; apparently to this critic the values in the CI, whether it crosses zero or not, are too small to be useful.

Suppose you are interested in estimating an odds ratio, and your data produce a 95% CI of \([0.9, 2]\). A policy maker may consider the costs and benefits of the program at different values across the interval, or she may take a more rigorous approach and apply statistical decision theory (Berger, 1985). A critic, however, points out that the CI crosses 1 and states that you cannot rule out there is no effect. Such a statement is not an indictment against your estimation. What then is the objection? The critic is objecting to your goal of estimation and is instead presumably seeking to judge whether there is a difference; not being able to rule out an odds ratio of 1 disallows such a judgment.

These examples presume the researcher is interested in either hypothesis testing or estimation, but the researcher may be interested in both. Suppose we are considering the incremental cost–benefit of a program modification. But, regardless of the magnitudes of the cost and benefit differences, if there is a decrease in benefit, then the modification will not be adopted, and if there is an increase in benefit with a corresponding decrease in cost, then the modification will be adopted. In this case, it is only if there is increasing benefit and increasing costs that we need to know the values of these changes to determine whether to adopt the modification. Consequently, it is only when there is sufficient evidence supporting the last hypothesis that we need to pursue the goal of estimation.

**A Logic of Statistical Testing**

The logic presented here requires the a priori specification of hypotheses in terms of mutually exclusive and mutually exhaustive sets of possible values for distribution parameters of random variables reflecting a data generating process. A priori specification is an epistemological requirement for results to have evidential value; however, as discussed in this section, this requirement does not apply to the act of testing. See Figure 1 for examples regarding a parameter that can possibly take values on the extended real line (i.e., any number from negative infinity to positive infinity): Panel A depicts the set of hypotheses that underlie typical one-tailed tests; Panel B depicts the hypotheses that underlie two-tailed tests; Panel C represents how three hypotheses might be expressed. Each set of values represents a hypothesis regarding the parameter. For example, partitioning the real line into the set of negative values and the set of nonnegative values can represent a comprehensive set of hypotheses (e.g., \(H_1: \mu < 0\) and \(H_2: \mu \geq 0\)). The set of hypotheses can include substantively derived hypotheses as well as a catchall negation of these hypotheses. We wish to determine in which set of values is the true parameter (i.e., which hypothesis regarding the value of the parameter is true).

If an estimate \(\hat{\theta}\) (i.e., the value the estimator yields when applied to specific data) is in a given hypothesis-specified set of values, then it conforms to the corresponding hypothesis. Because \(H\) and \(\neg H\) (in which “\(\neg\)” denotes logical negation and \(\neg H\) can be interpreted as “not \(H\)” represent two sets of possible values that are mutually exclusive but together make up the full set of possible values, if the estimate \(\hat{\theta}\) conforms to one of the hypotheses, it will conform to only that hypothesis (e.g., any point on the real lines depicted in Figure 1 can only be in one of the hypothesized sets). If the set of possible parameter values is a proper subset of the estimator’s range (e.g., whole numbers are a proper subset of all real numbers), then it is possible for \(\hat{\theta}\) not to conform to any hypothesis (e.g., the estimate might be a fraction when the hypotheses are sets of whole numbers). If an estimate conforms to a hypothesis, then it is a plausible result if the hypothesis were true; indeed, for \(\theta = \hat{\theta}\), then for an unbiased estimator with a symmetric unimodal distribution, \(\hat{\theta}\) is the most likely result, given the data.

I define an estimate \(\hat{\theta}\), and by extension the underlying data, as being consistent with a hypothesis if the estimate conforms to the hypothesis or if there exists at least one element in the corresponding set of values that would define a data generating process that could plausibly have produced data at least as extreme as the obtained estimate. For hypothesis \(H\), to which the data do not conform, \(\hat{\theta}\) is consistent with \(H\) if there exists an element \(\theta\) in the set of values
corresponding to H for which the $p$-value($\hat{\theta}; \theta$) is large. The term $p$-value($\hat{\theta}; \theta$) denotes the probability of the specified data generating process, with an actual parameter value of $\theta$, producing data having at least as extreme an estimated value as $\hat{\theta}$ (this is the usual definition of a $p$ value). An estimate, and therefore the data, is inconsistent with hypothesis H if the estimate is not consistent with H: which is to say, if $\hat{\theta}$ does not conform to H and for all elements $\theta$ in the set of values corresponding to H the $p$-value($\hat{\theta}; \theta$) is small. Judgments regarding what constitutes a large or small $p$ value are typically made in comparison with a threshold value termed a significance level. Notice that to be consistent with a hypothesis, the estimate need only correspond to a large $p$ value for one value in the hypothesized set of values; to be inconsistent with a hypothesis, the estimate needs to correspond with a small $p$ value for all values in the hypothesized set of values.

Figure 2 presents the possibilities for the single-parameter two-hypotheses case (H and ¬H) defined on the real line. Panel A depicts the estimate being consistent with both hypotheses: $\hat{\theta}$ conforms to H, and there also exists at least one parameter value in the set of values corresponding to ¬H that could plausibly produce data with an estimate at least as extreme as that obtained (if we consider the area under the curve to the right of $\hat{\theta}$ as being large). In this case, the data cannot adjudicate between them, and we cannot rule out either hypothesis. Surely, $\hat{\theta}$, providing a statistically insignificant test result, still provides some evidence in favor of ¬H? No. Even though $\hat{\theta}$ is consistent with some values in ¬H, it is in fact in H and thereby it is even “more consistent” with values in H. Consider, for example, the true value being exactly that of the estimate $\hat{\theta}$, a value that is in H. If anything, the fact that $\hat{\theta}$ is anywhere in H provides more evidence in favor of H than ¬H regardless of being consistent with values in ¬H (i.e., regardless of the insignificant test of ¬H). Making a judgment based on this fact alone, however, would not be an exercise in formal statistical testing as it would not be properly accounting for the fact that the data generating process could have produced estimates at least as extreme as $\hat{\theta}$ if the true value were actually in ¬H.

Panel B depicts the estimate being consistent with H but inconsistent with ¬H: Hypothesis H could well have produced the data, but ¬H is not likely to have (if we consider the area under the curve to the right of $\hat{\theta}$ as being small). This constitutes evidence for H and evidence against ¬H. Panel C depicts the estimate being consistent with ¬H but inconsistent with H: Hypothesis H is not likely to have produced such an estimate (if we consider the area under the curve to the left of $\hat{\theta}$ as being small), but ¬H could have. This situation constitutes evidence against H and evidence for ¬H. Panel D is not possible because in this example the estimate must conform to, and thereby be consistent with, at least one of the hypotheses.

Because data are consistent with a hypothesis to which the estimate conforms, it is not necessary to statistically test such a hypothesis; however, a statistical test is required of those hypotheses to which the estimate does not conform. In this case, because the estimate will conform to only one hypothesis, if there are $N$ hypotheses, then $N − 1$ hypotheses must be statistically tested. In the case where the estimate does not conform to any hypothesis, all hypotheses must be tested.

**Steps in the Application of the Testing Logic**

The application of this logic proceeds in four steps as stated in Table 1. The interpretation of results depends on which of two cases apply.

**Table 1. Steps in the Application of the Logic of Statistical Testing.**

| Step | Description |
|------|-------------|
| 1    | Determine the hypothesis-specific partition of the parameter space associated with the data generating process. How this is achieved depends on the substance and logic of the research being pursued and is not merely a question of statistics. |
| 2    | Obtain the parameter estimate using data from the data generating process. |
| 3    | Identify the hypothesis to which the estimate conforms. |
| 4    | Test whether the estimate is consistent with the remaining hypotheses. |

Case 1: Data are consistent with all hypotheses. If the data are consistent with all hypotheses, then there exists at least one plausible parameter value in each of the hypothesized set of values. In this case, the data do not provide evidence for or against any of the hypotheses.

Case 2: Data are inconsistent with at least one hypothesis. If the data are consistent with one hypothesis but inconsistent with all others, then the data provide evidence for the
hypothesis and evidence against the others. When there are more than two hypotheses, the data can provide evidence for or against sets of hypothesis. In such a case, the data cannot adjudicate between the hypotheses in the set of hypotheses with which the data are consistent but can rule out the hypotheses with which the data are inconsistent.

The One-Tailed Test

Hypotheses regarding a single parameter often take the form of directional hypotheses such as $H_0: \theta \leq 0$ versus $H_1: \theta > 0$. Which hypothesis should we statistically test? Suppose at Step 3, we estimate $\hat{\theta} = -0.6$. Being negative, $\hat{\theta}$ conforms to $H_0$ and is consequently consistent with this hypothesis. We do not, therefore, need to statistically test $H_1$. But the question remains whether it is likely that if $\theta$ were in fact positive, the data generating process would produce data with an estimate at least as extreme as $-0.6$. We answer this question by statistically testing $H_1$. Notice that we do not know a priori which hypotheses will be subject to statistical testing; we must first know to which hypothesis, if any, the estimate conforms.

Because the hypotheses specify sets of possible values, they are not necessarily expressed in terms of the specific parameter value used to calculate the $p$ value. If the hypotheses are $H_0: \theta \leq 0$ versus $H_1: \theta > 0$, and if in Step 3 we obtain $\hat{\theta} = 0.4$, which conforms to $H_1$, then we need to statistically test $H_0$. But which of the infinite number of values in $H_0$ do we use to calculate a $p$ value? To determine whether there exists a value in $H_0$ consistent with $\hat{\theta}$, we only need a $p$ value for a value in $H_0$ with which the data must be consistent if there exists any such point. This will be the value nearest to the estimate, on the metric of statistical distance. For a discrete parameter space, this value will be in the set of values corresponding to the hypothesis being tested; for a continuous parameter space, this point will be on the boundary between the hypotheses’ sets of values. For example, consider an estimate of $-0.6$ and a test of $H_1: \theta > 0$ for a continuous parameter space. We would use $\theta = 0$ to calculate the $p$ value even though this value is not in $H_1$. The reason is that for any value in $H_1$ near 0, there are an infinite number of values between it and 0. However, using $\theta = 0$ to calculate a $p$ value will produce the same quantity as a point in $H_1$ infinitely close to 0.

Returning to our example with $\hat{\theta} = 0.4$ and testing the hypothesis $\theta \leq 0$, notice from Figure 3 that if $\theta$ is plausible given $\hat{\theta} = 0$ (i.e., has a large $p$ value under the distribution depicted by the solid line), then we have at least one point in $H_0$ that makes $\hat{\theta}$ consistent with $H_0$—the very point we tested. However, if $\hat{\theta}$ is implausible given $\theta = 0$ (i.e., has a small $p$ value), then it must be implausible for any other point further away from $\hat{\theta}$ in $H_0$ (e.g., the distribution depicted by the dashed line) and therefore there is no point in $H_0$ that would make $\hat{\theta}$ consistent with $H_0$. Consequently, the test of whether $\hat{\theta}$ is consistent with the hypothesis that $\theta \leq 0$ is achieved using the $p$ value associated with the value $\theta = 0$, but the hypotheses being considered remain as originally stated: $H_0: \theta \leq 0$ and $H_1: \theta > 0$.

Directional hypotheses are sometimes introduced as a point hypothesis and a directional hypothesis such as $H_0: \theta = 0$ versus $H_1: \theta > 0$ (Neutens & Rubinson, 2002a; Spanos, 1999a). As indicated by the preceding discussion, this specification is not a general description of directional hypotheses and would only be appropriate when the parameter space is legitimately restricted to the indicated range and the point hypothesis is part of an a priori specification of the partition.

The Two-Tailed Test

Perhaps the most common statistical test is the two-tailed test of a scalar parameter. In this case, the hypotheses include a single point and its negation. Step 1 of the preceding logic is to express the hypotheses that constitute the relevant partition of the parameter space: for example, $H_0: \theta = 0$ versus $H_1: \theta \neq 0$. However, we know a priori that the estimate is almost certainly going to conform to $H_1$; The chance of an estimate equaling 0 (to the precision of the computer, and certainly to the infinite decimal place) is almost certainly 0. This has two consequences: First, we know a priori that we will almost certainly be statistically testing the null hypothesis $H_0$. Second, it is not possible to accrue evidence for the point hypothesis $H_0$ in a continuous parameter space. To have evidence for $H_0$, one would need an estimated value in $H_0$ that would be unlikely under all points in $H_1$. Even if we obtained an estimate exactly equal to 0, for any finite sample there will be some positive number $c$ such that if $\theta = 10^{-c}$ (which,
The rejection region of the statistic can be between \( \theta_0 \) and \( \delta \), that \( \theta \) is no more than \( \theta_0 \pm \delta \) (Neyman, 1957). Do these results significantly different from \( \theta_0 \), we might infer that \( \theta \) is at least \( \theta_0 \pm \delta \); if results are not significant, then we might infer that \( \theta \) is no more than \( \theta_0 \pm \delta \) (Neyman, 1957). Do these results, however, constitute a formal test of the implied set of hypothesis \( \theta \in [\theta_0 - \delta, \theta_0 + \delta] \) and \( \theta \notin [\theta_0 - \delta, \theta_0 + \delta] \)? No. The rejection region of the statistic can be between \( \theta_0 \) and \( \delta \), or greater than \( \delta \) but not statistically significantly different from \( \delta \) (see Figure 4, in which \( \theta_0 = 0 \)). If we wish to test these new hypotheses, we would follow the steps in the testing logic, which would lead to collecting data and a test centering the sampling distribution on either \( \theta_0 - \delta \) or \( \theta_0 + \delta \) (whichever is closest to the new estimate). But, given the less than perfect power in the new data generating process, we will end up with another discernible deviation \( \delta^* \) around this point. If we again follow the a posteriori interpretation above, we would be lead to the implied hypotheses \( \theta \in [\theta_0 - \delta - \delta^*, \theta_0 + \delta + \delta^*] \) and \( \theta \notin [\theta_0 - \delta - \delta^*, \theta_0 + \delta + \delta^*] \). As we keep pursuing these new implied hypotheses, taking into account the discernible deviation due to power, we ultimately (through infinite iterations, assuming the sequence of standard errors has a nonzero lower bound) get to the implied statement that \( \theta \in (-\infty, \infty) \), which we presume to be true a priori—We have arrived at a trivial truth rather than an informative hypothesis.

Can we consider a similar construction of implied hypotheses using data-specific concepts such as CIs or severity measures (Mayo, 2010; Mayo & Spanos, 2010)? For example, based on the CI we might consider our results as formally testing the implied hypotheses \( \theta \in [ \text{Lower Limit}, \text{Upper Limit} ] \) and \( \theta \notin [ \text{Lower Limit}, \text{Upper Limit} ] \)? No. Because, as argued above, such statements refer to the data and thereby simply do not constitute informative hypotheses about data generating processes. Although the a posteriori consideration of power, CIs, or severity measures can inform inference, they do not deliver their epistemic value through formal hypothesis testing.

We could, however, use these data-specific values as part of a formal test. For example, regarding the CI, we could use the underlying interval estimator \((L, U)\) as a test statistic and would thereby require its sampling distribution to calculate the probability of the statistic taking on values at least as extreme as the calculated data-specific CI. Here it is important to remember that the CI as calculated from data is not a statistic (i.e., not a random variable) upon which a formal test can be based: It is a single realization from the distribution of an underlying interval statistic \((L, U)\). A formal test would be based on the distribution of \((L, U)\). If such a test were constructed, the logic of its use would follow that as described in this article. Nonetheless, these types of data-specific values are not commonly used in formal statistical testing as defined here.

**Why Test Hypotheses?**

Statistical hypothesis testing is common when a researcher wishes to determine a substantive claim. If the truth or falsity of the substantive claim can be identified with the truth or
falsity of a statistical hypothesis, then hypothesis testing can be used to inform judgments about the substantive claim. This is the basis for hypothesis-driven science. For example, Kan (2007) derived and tested statistical hypotheses from the claim that time inconsistent preferences with hyperbolic discounting explains lack of self-control among smokers. Cook, Orav, Liang, Guadagnoli, and Hicks (2011) tested the hypothesis that disparities in the placement of implantable cardioverter-defibrillators (ICD) can be explained by the underutilization of ICD implantation among clinically appropriate racial/ethnic minorities and women and the overutilization of the procedure among clinically inappropriate Whites and men. Veazie (2006) derived and tested statistical hypotheses from the claim that variation in individuals’ perceptions of those with chronic medical conditions is explained by Ames’s (2004a, 2004b) theory of social inferences. It is the fact that the estimate is consistent with a hypothesized set of parameter values and inconsistent with others that constitutes evidence for the hypothesis, not the raw-scale distance (i.e., magnitude) from the boundary between such sets.

Statistical hypothesis testing is also used when the goal of estimation is of interest only if the parameter is in a particular range of values. The cost–benefit example presented above is one such case. Another is when researchers determine whether a variable predicts an outcome. Stating that something “will either go up or not go up” clearly does not constitute an informative prediction, and stating that something “will either go up or down” (e.g., an inference from a significant two-tailed test) is not much better. Consequently, identifying predictors typically requires isolating a direction. In this case, it is reasonable for the researcher to first address the three hypotheses that (1) the parameter is greater than zero, (2) the parameter is less than zero, and (3) the parameter is equal to zero. Because, in a continuous parameter space, the third hypothesis is a point on the boundary between the first two, testing this set of hypotheses reduces in practice to essentially testing the disjunction of one of the first two with the third. If the estimate conforms to (1), then it is statistically tested against the disjunction of (2) and (3). If the estimate conforms to (2), then it is statistically tested against the disjunction of (1) and (3). If an adequate judgment regarding the truth or falsity of these hypotheses can be made, then the researcher continues with the estimation goal and interprets point or interval estimates accordingly.

Discussion

The objective of this article was to present a coherent Frequentist logic of testing. To do so, I distinguished the goal of hypothesis testing from that of estimation, and presented a logic for the former that does not confuse it with the latter. The key points include (a) hypotheses are expressed as a partition of the parameter space specifying the distribution of random variables associated with a data generating process, (b) which of the a priori specified hypotheses are statistically tested cannot generally be known before the parameter is estimated (the exception being when a point hypothesis is involved), (c) the parameter estimate is consistent with a hypothesis to which the parameter estimate conforms and thereby this hypothesis does not require statistical testing, (d) all hypotheses to which the estimate does not conform are subject to statistical testing to rule them out as alternative explanations, (e) the element in the hypothesis’ set of values that produces the largest p value is used to test the hypothesis, and (f) except in the case of a point hypothesis, an estimate can provide either evidence for or against hypotheses (or sets of hypotheses), or remain ambiguous regarding them.

When testing hypotheses, researchers should report whether there is either evidence for or against hypotheses. Moreover, ambiguous findings (i.e., insignificant findings) should not be reported as evidence from a formal test for a hypothesis. For example, the common practice of treating insignificant results of a formal two-tailed test as evidence that there is no effect should be avoided. Instead, it should be acknowledged that the data cannot distinguish hypotheses or cannot rule out certain alternatives.

In this article, I focused on hypotheses about a single parameter. The presented logic naturally extends to hypotheses regarding multiple parameters as well (e.g., hypotheses regarding two parameters θ and γ such as H1: [θ > 0 and γ > 0] and H2: [θ ≤ 0 or γ ≤ 0]). See the online appendix for a description of hypothesis testing with multiple parameters.

For clarity of presentation, I adopted the standard concept of a threshold (i.e., significance level) to categorically determine whether data are consistent with a hypothesis; an approach that leads to the common use of categorical statements such as having a “significant result” or an “insignificant result.” This does not preclude the determination of a set of thresholds to define multiple categories of evidence such as weak evidence (e.g., perhaps .1 ≤ p > .05), moderate evidence (e.g., perhaps .05 ≤ p > .01), and strong evidence (e.g., perhaps p ≤ .01). Note, however, that such thresholds are arbitrary, relative to a scientist’s judgment, or conventional, relative to expectations of a community of scholars (e.g., as broad as a discipline or field of study, and as narrow as a specific journal). It should also be clear that it is not necessary to adopt formal thresholds at all in the application of the presented logic: A scientist may directly interpret the evidential value of the p value: For example, notwithstanding the conventional .05 significance level, a scientist may consider p values of .052 and .048 as essentially equivalent in their evidential bearing, perhaps judging both indicate the data are inconsistent with the hypothesis being tested. Moreover, the logic described here can be applied by considering the p value as a continuous measure of consistency with a value contained in the hypotheses with which the data do not conform. Nonetheless, unlike Bayesian methods, the logic of formal Frequentist hypothesis testing does not imply statements of mathematical probability reflecting subjective
beliefs. Consequently, the $p$ value (i.e., the probability that a data generating process would produce a statistic value as extreme as that observed given hypothesized distributional characteristics) requires interpretation by the scientist in light of the context and scientific goals.

A final point of clarification may be helpful. I have mentioned the need for a priori specification of hypothesis but also the fact that one cannot determine which hypothesis (or hypotheses) will be statistically tested before observing the estimate (except when including a point hypothesis). These do not conflict. The first is an epistemic requirement for using test results as evidence. The second is a logical consequence of the first bearing on the process of establishing evidence. The first means you should not use the data to determine whether you are addressing, for example, the hypotheses $H_1: \theta \leq 0$ and $H_2: \theta > 0$, or you are addressing the hypotheses $H_1: \theta = 0$ and $H_1: \theta \neq 0$. This specification should be determined a priori. Notice this precludes arbitrarily doubling your power given your results. Suppose, however, I were to decide ahead of time that I will statistically test $H_1: \theta > 0$ and I subsequently obtain an estimate $\hat{\theta} = 5$. It is unclear how I would formally test $H_1: \theta > 0$ given the estimate. What value in $H_1$ do I base the test on and how would it be structured? There is no useful answer. How can the fact that $\hat{\theta}$ is in the hypothesized set of values provide evidence against the hypothesized values? It remains, however, to rule out $H_1$, but this is not my a priori specified statistical test. The steps in Table 1 avoid this issue because the hypothesis that is statistically tested, but not the hypothesis specification, depends on the obtained result and is not determined a priori.

By following the logic of formal statistical testing presented here, a researcher does not confuse the goal of testing with that of estimation and can thereby avoid the conflict inherent in interpreting results of testing and interpreting raw-scale magnitudes of estimation. However, a limitation of hypothesis testing is that it provides evidence solely for the truth or falsity of the specified hypotheses. It is the responsibility of the researcher to justify knowing this fact. For the case in which knowing the truth or falsity of a hypothesis is not important, formal hypothesis testing is not an appropriate goal—Estimation, or other informative means, without the pretense of formal testing, may be the better objective.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research and/or authorship of this article.

**References**

Ames, D. R. (2004a). Inside the mind reader’s tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology*, 87, 340-353. doi:10.1037/0022-3514.87.3.340

Ames, D. R. (2004b). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology*, 87, 573-585. doi:10.1037/0022-3514.87.5.573

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York, NY: Springer-Verlag.

Cook, N. L., Orav, E. J., Liang, C. L., Guadagnoli, E., & Hicks, L. S. (2011). Racial and gender disparities in implantable cardioverter-defibrillator placement: Are they due to overuse or underuse? *Medical Care Research and Review*, 68, 226-246. doi:10.1177/0022178310379421

Engsted, T. (2009). Statistical vs. economic significance in economics and econometrics: Further comments on McCloskey and Ziliak. *Journal of Economic Methodology*, 16, 393-408. doi:10.1080/13501780903337339

Fisher, R. A. (1956). *Statistical methods and scientific inference*. New York, NY: Hafner.

Hoover, K. D., & Siegler, M. V. (2008a). The rhetoric of “signifying nothing”: A rejoinder to Ziliak and McCloskey. *Journal of Economic Methodology*, 15, 57-68. doi:10.1080/13501780801913546

Hoover, K. D., & Siegler, M. V. (2008b). Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology*, 15, 1-37. doi:10.1080/13501780801913298

Kan, K. (2007). Cigarette smoking and self-control. *Journal of Health Economics*, 26, 61-81.

Mayo, D. G. (2010). Learning from error, severe testing, and the growth of theoretical knowledge. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 28-57). Cambridge, UK: Cambridge University Press.

Mayo, D. G., & Spanos, A. (2010). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of the philosophy of science: Philosophy of statistics* (Vol. 7, pp. 153-198). New York, NY: Elsevier.

McCloskey, D. N., & Ziliak, S. T. (1996). The standard error of regressions. *Journal of Economic Literature*, 34, 97-114.

McCloskey, D. N., & Ziliak, S. T. (2008). Signifying nothing: Reply to Hoover and Siegler. *Journal of Economic Methodology*, 15, 39-55. doi:10.1080/13501780801913413

Neutens, J. J., & Rubinson, L. (2002a). Analyzing and interpreting data: Inferential analysis. In *Research techniques for the health sciences* (3rd ed., pp. 272-273). New York, NY: Benjamin Cummings.

Neutens, J. J., & Rubinson, L. (2002b). *Research techniques for the health sciences* (3rd ed.). San Francisco, CA: Benjamin Cummings.

Neyman, J. (1957). The use of the concept of power in agricultural experimentation. *Journal of the Indian Society of Agricultural Statistics*, IX, 9-17.

Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A, 175-240.

Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, 20A, 263-294.
Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289-337. doi:10.1098/rsta.1933.0009

Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia, PA: Lippincott-Raven.

Spanos, A. (1999a). *Hypothesis testing probability theory and statistical inference: Econometric modeling with observational data*. Cambridge, UK: Cambridge University Press.

Spanos, A. (1999b). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge, UK: Cambridge University Press.

Spanos, A. (2008). *Review of Stephen T. Ziliak and Deirdre N. McCloskey’s the cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: The University of Michigan Press.

Veazie, P. J. (2006). Projection, stereotyping, and the perception of chronic medical conditions. *Chronic Illness*, 2, 303-310.

Ziliak, S. T., & McCloskey, D. N. (2004a). Significance redux. *Journal of Socio-Economics*, 33, 665-675. doi:10.1016/j.socec.2004.09.038

Ziliak, S. T., & McCloskey, D. N. (2004b). Size matters: The standard error of regressions in the American Economic Review. *Journal of Socio-Economics*, 33, 527-546.

Ziliak, S. T., & McCloskey, D. N. (2008a). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Ziliak, S. T., & McCloskey, D. N. (2008b). Science in judgment, not only calculation: A reply to Aris Spanos’s review of The Cult of Statistical Significance. *Erasmus Journal for Philosophy and Economics*, 1, 165-170.

**Author Biography**

**Peter J. Veazie** is Associate Professor in the Department of Public Health Sciences, Chief of the Division of Health Policy and Outcomes Research, and Director of the Health Services Research and Policy Doctoral Program at the University of Rochester. His research focuses on medical and healthcare decision making, health and quality of life outcomes, and research methods.