DigitalCommons
@WayneState

Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 7

5-2016

Does One Size Fit All? A Case for Context-Driven Null Hypothesis Statistical Testing

Grayson L. Baird
Rhode Island Hospital, grayson_baird@brown.edu

Lisa L. Harlow
University of Rhode Island

Does One Size Fit All? A Case for Context-Driven Null Hypothesis Statistical Testing

Cover Page Footnote

This research was supported in part by G20RR030883 from the National Institutes of Health.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 100-122.

ISSN 1538 − 9472

Does One Size Fit All? A Case for Context-Driven Null Hypothesis Statistical Testing

Grayson L. Baird
Rhode Island Hospital
Providence, RI

Lisa L. Harlow
University of Rhode Island
Kingston, RI

Rodgers (2010a) asserted that the practice of null hypothesis statistical testing (NHST) follows a mechanistic and rule-based epistemology. This concern is addressed using historical and modern sources as evidence for NHST as a dynamic, context-driven framework for empowering researchers in scientific inquiry.

Keywords: Null hypothesis statistical testing, NHST, context-driven

Introduction

Rodgers (2010a; 2010b) brought to light many important issues pertaining to what he called a "quiet revolution" (p. 2) concerning statistics in practice. Rodgers noted the practice of null hypothesis statistical testing (NHST) follows a "mechanistic" (p. 10) and "rule-based" (p. 1) epistemology. The intent of this article is to elaborate on this idea and to consider how the current NHST framework is applied in a somewhat rigid and prescriptive fashion. Specifically, the automatic and what Cohen (1994) called "ritual" (p. 997) practices of NHST is examined relative to what was suggested in original sources by foundational theorists (e.g., Fisher, 1926; 1928; 1935; 1973; Neyman & Pearson, 1933a; 1933b; Yule & Kendall, 1950). In addition, original and contemporary sources are provided as evidence for NHST as a dynamic, context-driven framework for empowering researchers in scientific inquiry. Although ritualism may be pervasive throughout many aspects of NHST, the scope of this paper is limited to considering only the selection of the critical value and the value of the null hypothesis.

Grayson L. Baird is a research statistician with the Lifespan Biostatistics Core at Rhode Island Hospital. Email him at: grayson_baird@brown.edu.

The Critical Value (*α*) and Level of Significance

Rodgers (2010a) described the practice of NHST as a set of procedures, applied mechanistically. Researchers today often collect data and test their hypotheses by deriving test statistics and corresponding *p*-values, from which statistical significance of results is ascribed if the derived *p*-value is less than a fixed threshold, conventionally 0.05 (for a concise history of 0.05 level of significance, see Cowles & Davis, 1982). This fixed threshold is used as a conventional cutoff value for determining if a result is statistically significant, above random variation, assuming the null. In practice, 0.05 (or alternatively, sometimes 0.01 and 0.001, see Skipper, Guenther, & Nass, 1967) is almost the universal definition of significance regardless of the subject area, the nature and size of the sample, the quality of the measurement, the quality and nature of the design, the hypothesized and actual effect size, or the research question itself.

Although the practice of using 0.05 is pervasive, a great deal of criticism towards NHST results from the use of an arbitrary and traditional cutoff value to determine significance (see Mudge, Baker, Edge, & Houlahan, 2012). For instance, early on, Selvin (1958) noted “reciting the magic phrase ‘significant at the 0.01 level’ is often a substitute for hard thinking about the quality of one's data” (p. 86). Ironically, this ritualistic practice of determining significance does not appear to be in accordance with testing espoused by either Neyman and Pearson or Fisher. Specifically, when discussing errors of the first and second kind (i.e., Type I error (PI), rejecting a null hypothesis that should be retained, and Type II error (PII), holding onto a null hypothesis that should be rejected, respectively), Neyman and Pearson (1933a) noted:

> These two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second. We are reminded of the old problem considered by Laplace of the number of votes in a court of judges that should be needed to convict a prisoner. Is it more serious to convict an innocent man or to acquit a guilty?... From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator. (p. 296)

Neyman and Pearson (1933b) also noted: "we attempt to adjust the balance between the risks PI and PII, to meet the type of problem before us" (p. 497). Here, Neyman and Pearson described a system whereby the researcher plays an active role in evaluating significance, in the context of minimizing and thus balancing errors of the first and second kind, which are inversely related to each other, relative to the conditions of the study at hand. Therefore, significance level is not an arbitrary and universal value, but rather a value that achieves a meaningful and appropriate balance of Type I versus Type II errors, determined by the researcher with the specific conditions of the study in mind. Neyman and Pearson (1933b) stressed the influence of context for deciding if a small or large critical value is warranted.

However, some of Fisher's writings may be viewed as promoting a fixed level of significance. For instance, Fisher (1928) noted:

> [Regarding] the value for which *P* = .05, or 1 in 20… it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. (p. 45)

In this statement, Fisher appears to have advocated a significance level of 0.05. However, also around this time, Fisher (1926) wrote:

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance. (p. 504)

This statement reveals that a significance level of 0.05 is viewed by Fisher as a "low" standard and other levels of significance may be used. Of the two aforementioned statements, attention should be drawn to Fisher's use of the words "convenient" and "prefers" when he described choosing a level of significance. Here, Cochran (1976) suggested that Fisher appeared to be promoting a significance level of 0.05 based on preference but not advocating 0.05 as an exclusive level of significance. Also apparent in Fisher's aforementioned statement is that his confidence in experimental results rested with the quality of the design. Further evidence of Fisher's reluctance to assign an official level of significance but instead consider significance in light of the conditions of the research can be seen in the following statement some years later:

> …the attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to supposed frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A [scientist] who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. . . . However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, [they] reject hypotheses; [the scientist] rather gives [their] mind to each particular case in the light of … evidence and [one's] ideas. (1973, p. 45)

Fisher (1973) suggested a level of significance that may reliably indicate statistical significance over the long run, although he was quick to condition this statement by noting that a universal or fixed level of significance used in all situations would not make sense. He noted, "In choosing the grounds upon which a general hypothesis should be rejected, personal judgment may and should properly be exercised" (p. 50).

This is evidence neither Neyman and Pearson nor Fisher advocated any universal or canonical level of significance, but rather entrusted that researchers would define a level of significance that was relevant to their field of research and appropriate for the conditions of the study. Specifically, it is possible to see that Neyman and Pearson advocated testing as a dynamic procedure where the researcher actively engages in evaluating "what is significant" by balancing the costs of committing Type I verses Type II errors relative to the context of the research. It is also possible to see that Fisher was advocating testing also as a dynamic procedure, where the researcher actively engages in evaluating what is significant by considering the conditions of the particular study and the nature of the research question.

It is difficult to imagine either Neyman and Pearson or Fisher as supporters of mechanistic thinking in general. Pearson (1955) noted that "from the start we shared Professor Fisher's view that in scientific inquiry, a statistical test is 'a means of learning'" (p. 206). Neyman asserted "for a satisfactory performance of a statistician's duty… it is necessary that [they] fully understand the circumstances of experiments, whatever their nature, to which statistical methods are applied"

(Reid, 1982, p. 183). Fisher specifically noted that "tests of significance are used as an aid to judgment, and should not be confused with automatic acceptance tests" (Fisher, 1928). None of the aforementioned statements presented suggest that any of these theorists intended for or engendered a ritualistic application of statistics.

Fisher and Neyman and Pearson advocated two separate frameworks. For example, Neyman-Pearson theory advocated setting the critical value (alpha) a priori of analysis, whereas Fisher advocated reporting significance level after analysis. Fisher's framework tested the null only, whereas the Neyman-Pearson framework tested two or more hypotheses, thus allowing errors of the first and second kind to be controlled for and power to be estimated. A more in-depth review of these and other differences can be found in Gigerenzer (2004). Although a greater elaboration of these differences is beyond the scope of this review, it is important to establish that neither framework appears to be advocating an arbitrary and universal threshold of significance, such as 0.05.

Fisher's and Neyman and Pearson's treatments of significance as a contextual judgment appears to be in agreement with other original theorists. For instance, Yule and Kendall (1950) noted:

> In the examples we have given…our judgment whether *P* was small enough to justify us in suspecting a significant difference…has been more or less intuitive. Most people would agree…that a probability of only 0.0001 is so small that the evidence is very much in favour of the supposition that the dice were biased…Suppose we had obtained *P* = 0.1…Where, if anywhere, can we draw the line? The odds against the observed event which influence a decision one way or the other depend to some extent on the caution of the investigator. Some people (not necessarily statisticians) would regard odds of ten to one as sufficient. Others would be more conservative and reserve judgment until the odds were much greater. It is a matter of personal taste. (p. 471)

This discussion of significance by Yule and Kendall appeared to have advocated a way of determining significance based on a researcher's intuition, caution, scientific background, and "personal taste." It should be noted that they went on to mention that there are two values of *P*, 0.05 and 0.01, which are used widely to provide a "rough line of demarcation" for level of significance (p. 472). Yule and Kendall do not appear to have promoted a strict level of significance; rather, they appear to have advocated a significance level based on contextual considerations and later mentioning 0.05 and 0.01 as rough thresholds commonly used. Thus, significance level based on contextual considerations appears to be an established tradition early on in the field of statistics.

The convention of statistical significance being achieved if and only if *p* is less than an arbitrary and recognized cutoff is perhaps the most illustrative instance of rather rote thinking in the current practice of applied statistics in psychology. The current NHST framework, in practice, allows arbitrary, traditional, and preordained cut off values to determine the significance of the results rather than allowing the significance of the results to be determined by researchers with the conditions of the study and the research question in mind. Thus, pervasively defining significance at 0.05 has led the process of inference away from a scientific basis, as noted by Morrison and Henkel (1969):

> If, indeed, .05 (or any other level) is 'sacred'…then what do we do in sociology surely is much more akin to religion than science and we might as well forget empirical work and get on with the development of more rituals. (p. 137)

Fortunately, there is support for NHST, as a framework, which empowers researchers to evaluate the significance of their results relative to the context of their research. Aguinis et al. (2010) asserted that conventional cutoffs ignore the relative seriousness of committing a Type I versus II error for a given study. For instance, researchers studying the possible effects of a new drug could be committing a Type I error if the drug was found to appear effective although it was later found to have very serious side effects that reduced the benefit of the treatment and potentially endangered the patients. Alternatively, researchers may commit a Type II error when testing a drug with little or no side effects if the power in their study is small (due, for example, to a small sample size and/or a small effect), which could have serious consequences by potentially removing a viable treatment from consideration by patients needing new options. Thus, in some research contexts, failing to reject the null when the null is false may be more serious than rejecting the null when the null is true (or vice versa). As a consequence, the widely used convention of maintaining arbitrary cutoffs disallows the researcher to appreciate and control for the relative seriousness of committing either Type I or II errors in a given research situation.

Aguinis et al. (2010) therefore proposed a "customer-centric" (p. 517) approach to science, where the customer (i.e., the researcher) controls the probability of committing a Type I and Type II error based on the relative seriousness of committing these errors and given the nature of the research and research question. Thus alpha is chosen by the researcher based on the context of the research and predicated on the researcher's preference and rationale behind the relative seriousness of committing a Type I versus II error.

A similar proposal by Baker and Mudge (2012; also see Mudge et al., 2012) called for researchers to explicitly consider the relative costs of Type I and II errors when determining a value of alpha. Baker and Mudge pointed out that when sample variability is high and/or the sample size is low, the habitual use of designating alpha at 0.05 leads researchers to unrealistically test for real effects and, as a consequence, increase the rate of Type II errors (false negatives). They advocated using an optimal alpha value that takes into account the relative costs of committing Type I and II errors using power analysis (e.g., Cohen, 1988); however, instead of determining a needed value for power, effect size, or sample size, alpha is being determined. Thus, researchers can calculate an optimal alpha value by specifying a meaningful (a priori) critical effect size, sample size, and different values of power. In practice, once the observed sample size and a meaningful effect size have been specified, if a Type II error is more serious than a Type I, then simply increase the value of power which will decrease the probability of a Type II error, thereby increasing the probability of a Type I error. Conversely, if a Type I error is more serious than a Type II, then simply decrease the value of power which will in turn decrease the probability of a Type I error but will increase the probability of a Type II error.

However, Baker and Mudge (2012) held that, for most studies, alpha should be a value that minimizes the overall probability or cost of making a mistake; thus the selected alpha value should minimize the combined probabilities of a Type I and II error. Specifically, they noted that "If we consider minimising the chances of errors to be the goal for good decision-making, we can choose an optimal decision-making threshold (optimal α level) that minimises the average of α and β (Type I and Type II errors) at the smallest potentially meaningful effect size" (p. 30). They also asserted that researchers should report the sample size, observed variability of the data, exact *p*-values, specified power value and effect size used in determining each optimal alpha value so that other researchers can re-evaluate results using different optimal alpha values based on their own notions of relative cost of Type I and II error and critical effect sizes. Thus, instead of convention, the context of the study (e.g., relative seriousness of Type I and II error, or the goal of reducing both errors optimally, the variability of the data, the observed sample size, the a priori desired or hypothesized effect size, and power) must be considered when setting an alpha value.

Cascio and Zedeck (1983) proposed directly assessing the "apparent relative seriousness" (ARS) of Type I and II errors with the following equation:

$$\mathrm{ARS} = \left[\frac{\mathrm{P}\left(H_1\right) + \beta}{1 - \mathrm{P}\left(H_1\right) + \alpha} \right] \tag{1}$$

where P(H_1) is the probability that the null is false, 1 – P(H_1) is the probability that the null is true, β is the probability of a Type II error, and α is the probability of a Type I error. In essence, if alpha is held constant at 0.05 and power 0.80 then, as the probability of the null being false increases, the relative importance of alpha increases dramatically; thus when P(H_1) = 0.1, 0.2, 0.4, 0.5, or 0.7, the alpha value is 0.44, 1, 2.7, 4, or 9.3 times more serious than β. The ARS equation allows researchers to directly assess relative seriousness of error with aspects of context within their study.

Using this framework, Murphy and Myors (2004) proposed operationalizing an appropriate alpha value based on the same aspects of research context. Therefore, instead of assessing ARS based on an arbitrary alpha value, they proposed determining a specific alpha value using a researcher's desired relative seriousness (DRS) value of committing a Type I versus Type II error (desired ARS value), given:

$$\alpha_{\text{desired}} = \left[\frac{\mathrm{P}\left(H_1\right)\beta}{1 - \mathrm{P}\left(H_1\right)} \right] \bullet \left[\frac{1}{\mathrm{DRS}} \right] . \tag{2}$$

Alpha is based on the context of the research, where the balance of Type I over Type II error is specified by the researcher. Moreover, the researcher's confidence in the alternative being true along with the researcher's notion that a rejection may be false is realized mathematically. Confidence in the alternative being true and the probability of Type II error may be due to the quality of the sample, the quality and control of the design, researcher's experience, previous research, etc. The benefit here is this approach produces an alpha level that fits the needs of the researcher relative to the conditions of the study (to the degree the conditions of the study can be translated into those parameters) and can be justified a priori.

Both the historical and contemporary authors mentioned revealed a need for determining an alpha value appropriate for the context of the research instead of using conventional and universal cutoff values. However, alpha (or level of significance) is not the only value thoughtlessly selected in the application of NHST.

The Value of Null

Another common criticism of NHST is that the hypothesis being tested is often limited to the hypothesis of no effect, often called the nil hypothesis (see Cohen, 1994). There are several research situations where testing the nil is appropriate given the research question. Specifically, if one is interested in examining any effect or difference above zero only, then the nil is a logical hypothesis of comparison (e.g., effect of experimental manipulation between two randomly assigned groups). Although this type of research question may be seen as overly simplistic to many researchers, testing the nil can nevertheless legitimately address the question of interest. While testing the nil hypothesis may be statistically sound, however, the habitual practice of testing only the nil in all research contexts, as a default value rather than a null value of interest, is another illustration of ritualistic practice in the application of NHST.

The Null as a Value

Although nil hypothesis testing is often used, it is not the only hypothesis available to researchers within the NHST framework. Originally, Fisher proposed the null hypothesis as the hypothesis of interest in which we were trying to disprove; it is the hypothesis we are trying to nullify (see Bakan, 1966; Cohen, 1994; Gigerenzer, 2004). Thus, the hypothesis to be nullified can refer to any null value, including but not limited to the nil. Specifically, Fisher (1935) asserted "we may, however, choose any null hypothesis we please, provided it is exact" (p. 20). This is perhaps best illustrated in his infamous Lady tasting tea problem: the Lady asserted that she could discriminate between cups of tea where the milk was infused either before or after the tea was poured. Her claim is tested with 8 cups of tea, 4 containing tea with milk infused prior to pouring and 4 after. The Lady is presented with the cups in random order and is blinded to their preparation. Fisher noted that the null could be either that the Lady has no sensory discrimination in detecting how tea was prepared regarding milk or that she has perfect sensory discrimination (Fisher, 1935, p. 13). Thus, the null could be 0.5 or 1.0, revealing that the null need not be the nil (i.e., 0.5).

The Neyman-Pearson (see Neyman, 1950; 1957) hypothesis testing framework specifically required researchers to designate the values of the null and an alternative (H_1 and H_2), where the null is preferably called the "hypothesis tested" and "it is immaterial which of the two alternatives H_1 and H_2 is labeled the hypothesis tested" (Neyman, 1950, p. 259). To illustrate this point, Neyman considered two hypotheses in regards to Fisher's Lady tasting tea problem. He noted of (a) $p \neq 1/2$ and (b) $p = 1/2$ that one of these hypotheses will be “the hypothesis tested” and the other “the alternative hypothesis” (Neyman, 1950, p. 273). Neyman went on to say that which claim will be regarded as the hypothesis tested and which the alternative depends on the situation and the balance of errors of the first and second kind: if we were the Lady, we would want the hypothesis tested to be (a), as the more important error to avoid is having her claim refused (avoid rejecting (a) if (a) were true); if we were the jury, we would want (b), given that the more important error to avoid is the granting of an unjustified claim (avoid rejecting (b) if (b) were true). Here, context plays into which hypothesis is the “null” in concert with balancing errors of the first and second kind. In another example, Neyman (1942) provided general guidance for selecting the hypothesis to be tested; he noted that the null hypothesis should be the hypothesis whereby the errors of the first kind are of greater importance relative to errors of the second kind. In this example, he specifically chose a non-nil hypothesis (i.e., “the actual toxicity of the drug does exceed the prescribed safety limit”) given the relative importance of a Type I error (p. 304).

The two examples above concerning the Lady tasting tea experiment reveal that although Fisher and Neyman and Pearson explicitly promoted two different frameworks, neither advocated that the null always be defined as the nil. Indeed, as illustrated by both Fisher and Neyman, in theory and application, the null can be defined as any value; instead of the nil, or a value of zero, being the standard, it is just one possible hypothesis to test within the greater NHST framework (see Murphy & Myors, 1999).

Apart from reducing the involvement of researchers in the decision process, the default use of the nil as the null hypothesis can also limit application and theory. For instance, Serlin (1987) asserted that use of the nil hypothesis provides weak evidence for many theories given that it is often believed a priori that populations do in fact differ at least somewhat. In application, always testing the nil can be problematic because, most often, samples differ from each other to some degree, regardless if they come from the same population or different populations, due to sampling error alone. Meehl (1990) went so far as to call the use of nil hypothesis testing a “weak use” (p. 116) of a significance test, because the nil is (literally) always false. In addition, nil hypothesis testing is limited in detecting only if a difference or relationship exists, above zero, without regard to magnitude (Murphy & Myors, 2004). Finally, by only testing a single null value, statistical significance can be achieved by simply increasing the sample to a sufficient size (Serlin & Lapsley, 1985).

Fortunately, these issues associated with exclusively defining the null as the nil are largely unnecessary. As mentioned, nil hypothesis testing can provide weak evidence for theories in which differences between populations or relationships between variables are anticipated or known to exist. Hodges and Lehmann (1954) noted that "when we formulate the hypothesis that the sex ratio is the same in two populations, we do not really believe that it could be exactly the same, and would only wish to reject equality if they are sufficiently different" (p. 261). One way to test for these "sufficient differences" lies in testing some value for the null other than zero. Murphy and Myors (1999) advocated an alternative to nil hypothesis testing which they termed "minimum-effect" testing. This framework is predicated on testing against a "negligibly small or trivial" effect, rather than testing for zero. Thus, depending on the context of the study, minimum effects testing can test more realistic hypotheses, rather than the "straw man" nil (Serlin & Lapsley, 1985, p. 74), which may be untenable in many research situations.

Another benefit of minimum effects testing is that it allows researchers to test both the presence of an effect and the magnitude of said effect by creating an upper and lower bound; thus, a range of null values can be tested instead of a specific value only. A minimum effect null is no longer a point hypothesis but rather a range between the minimum effect specified and the nil. Thus, if we set a null to 3% of variance accounted for and we reject this null, then we are more confident that a real effect exists because we are no longer testing a null of 0% variance accounted for. Moreover, by testing a non-nil null, when we do reject the null, we now have some information about the magnitude of said effect (e.g., the effect is above 3% variance accounted for). The benefits of using a minimum effect are apparent; however, the drawback of using a minimum effect is it increases the risk of committing a Type II error.

Although Murphy and Myors (1999) admitted that establishing a suitable minimum-effect value may be difficult initially, the benefits of such testing could greatly increase the meaningfulness of results. Thus they advocated a system whereby the hypothesis being tested is not determined for the researcher by convention, but rather the researcher determines a hypothesis relevant for the given research question and relative to the conditions of the study (e.g., a priori desired or hypothesized effect size, confirmation vs. exploratory study, theory concerning the population(s) being tested, etc.).

Use of the non-nil null also should not be applied in a rote manner. As Knapp and Sawilowsky (2001) warned, some effects are inherently small; thus, by using an arbitrary non-nil null, the chances of these (albeit) small effects being missed are increased, if not certain, depending on the non-nil value. Therefore, the value of any null, nil or not, must be guided by context. As a consequence, this framework empowers researchers to operationalize their research questions by evaluating and designating a value of sufficient difference or relationship (minimum effect) germane to and appropriate for the area of focus.

The Null as an Interval

Originally, Hodges and Lehmann (1954) proposed testing "sufficient differences or relationships" by using a range of possible values for the null hypothesis rather than testing a single null value. Later, Meehl (1990) proposed what he called a "strong use of hypothesis test" whereby the null is a specific value a researcher asserts as their theory, and therefore as the null they are testing against their assertion (p. 79). Serlin and Lapsley's (1985) framework advocated testing one's own theory as the null, along with using what they call a "good-enough belt" around a "complex null hypothesis" (p. 79). Instead of testing a nil hypothesis exclusively, they recommended testing a null value that represents one's theory (which could include the nil) and has a beltor width (denoted as Δ) around the value of the chosen null value. For example, instead of testing a null value against one's hypothesized value, researchers instead designated their hypothesized values as the null, and use good-enough belts to test a range of possible null values (e.g., 2.5 ± 0.5); thus one can think of good-enough belts as a type of confidence interval for the null value (see Serlin, 1987). Serlin and Lapsley (1985) noted that, by using good-enough belts, the imprecision of estimating the population is reduced because a range is being tested instead of a single all-or-nothing value. Moreover, they noted that instead of simply testing a direction, researchers are testing the magnitude of the change in direction.

A major criticism of the NHST is that the null can almost always be rejected when the sample size is sufficiently large. This problem, sometimes referred to as "infinite precision" (Serlin & Lapsley, 1985, p. 74), is a function of infinite (or very large) sample size whereby natural differences between populations can be detected even if they are not meaningful (Serlin & Lapsley, 1985). Conversely, by testing a range of possible null values, the almost inevitable rejection of the null due to increasing sample size is reduced. Serlin and Lapsley (1985) noted that the value of Δ must be chosen by the researcher a priori and "reflects the state of the art or the error in the best 'known experimental technique' in the field" (p. 79). The framework proposed by Serlin and Lapsley empowers researchers to determine a range of meaningful null values instead of mechanistically testing a single all-or-nothing value that is more easily rejected with a large enough sample. Thus they advocated a framework where the researcher, not ritual, decides the hypotheses of interest and where large samples do not automatically guarantee significance.

In summary, the inherent limitations associated with testing the nil hypothesis without ample consideration of a desired effect are largely unnecessary given the context-based alterative frameworks mentioned (although some may consider these to be alternatives to NHST itself; see Denis, 2003). Specifically, Murphy and Myors (1999) advocated a framework that empowers the researcher to evaluate the significance of hypotheses by determining a (minimum effect) null value that is meaningful to the researcher and appropriate for the context of the research. What is more, Serlin and Lapsley (1985) advocated a framework that empowers researchers to both specify a hypothesis of interest (including but not limited to the nil) while also determining a range or interval of possible values (a good-enough belt) where the null may still hold. Neither framework allows the researcher to blindly test a nil hypothesis by default (the dangers of which are clearly illustrated by Sawilowsky, 2003). These frameworks therefore empower researchers to specify their hypotheses in concert with the context of their research areas and questions.

Discussion

Many have observed that the current application of NHST is ritualistic (see Cohen, 1994) and mechanistic (Rodgers, 2010a; 2010b). Gigerenzer (2004) even labeled this phenomenon as “the null ritual” (p. 33). Indeed, a ritualistic approach to NHST, where the null hypothesis value and critical value are predetermined by convention, may actually impede researchers from testing the hypotheses appropriate for their particular research questions. In addition, rote selection of the nil and critical values may induce researchers to inadvertently ignore many important conditions of their study, such as the hypothesized effect size and the relative seriousness of Type I versus Type II error. As a consequence, the null ritual, not the researcher, ends up determining the significance of hypotheses and even the hypotheses themselves without regard to the context of the research. If used in this fashion, the application of NHST is indeed in danger of becoming a rite or ceremonial practice, much akin to those of the cargo cults where the deliverance of a *p*-value smaller than 0.05 is tantamount to a cargo box (see Feynman, 1985).

Although NHST may often be applied in practice without regard to context, there is little evidence that hypothesis testing was ever intended to be used in this fashion by original theorists. Neyman and Pearson, Fisher, Yates, and Kendall all wrote about determining significance relative to the judgment of the researcher in concert with the context of the research itself; none appear to have advocated for the definition of the null as the nil hypothesis exclusively. In addition, contemporary authors reviewed here offer innovative ways of conceptualizing the application of NHST to better suit the context of research while breaking away from habitually testing the point nil hypothesis. By implementing the concepts from these sources, both traditional and contemporary, researchers are engaged in what could be described as "context-driven NHST" or CD-NHST. Instead of being driven by convention, which may or may not have much relevance, CD-NHST places the researcher in the driver's seat of inference. In so doing, CD-NHST is in part responding to the changes in quantitative thinking and training called for by Rodgers (2010a; 2010b) and others (e.g., Cumming, 2012; Harlow, Mulaik & Steiger, 1997; Kline, 2011). Rodgers (2010a) noted:

> The treatment of the null and alternative hypotheses, of Type I and Type II errors, and of power needs to change to accommodate the focus on the researcher's model, rather than the null (nil) hypothesis. (p. 10)

CD-NHST not only addresses the issues brought up by Rodgers (2010a), but a happy by-product of CD-NHST is that, as a general framework, it inherently promotes replication and meta-analysis. Because CD-NHST requires more thought and detail, studies using CD-NHST could therefore yield an abundance of data for replication and meta-analytic studies. Specifically, with thoughtful and specific critical values, null values, and null ranges based on justified contextual reasons and all being reported, researchers can have access to a wealth of data to perform well-informed replications and meta-analyses. More importantly, CD-NHST as a framework relies on designating values from previous studies, thereby relying, to some degree, on replication itself.

Although the bulk of this discussion emphasizes empowering researchers by placing the selection of critical values and null hypothesis value(s) into the hands of researchers rather than being determined by common practice, this viewpoint is not without controversy. As noted by Cortina and Landis (2010), by having alpha set by externally determined criteria, corroboration between the hypothesis and data is compelling because the evidence is determined independently of the researcher; one may assume this thinking also extends to the selection of the null hypothesis value(s) as well. Conversely, Hubbard and Ryan (2000) asserted that conventional cutoffs only provide an illusion of objectivity that "makes life tidier" rather than requiring researcher's to use subjective judgment. Indeed, although setting alpha to a default value may be objective, one should always remember inference remains subject to the conditions of the study. Both points presented by Cortina and Landis (2010) and Hubbard and Ryan (2000) are important; thus a delicate balance must be struck between researchers evaluating their hypothesis and remaining objective in their evaluation. By empowering researchers to make context-driven decisions regarding the application of NHST, we at the same time risk inviting a certain level of subjectivity into the analysis.

One possible solution in balancing active evaluation and biased subjectivity would be to encourage researchers to establish critical values, null values, and null ranges a priori of data analysis or even data collection. This would allow researchers to participate in determining their hypotheses and the significance of said hypotheses, without the data and results influencing these decisions. A second way to encourage researchers to engage in context-driven NHST without biasing their results could be achieved by having researchers justify specifically why they are using a particular critical value or null hypothesis value (based on previous research, theory, etc.).

A third step would be to encourage researchers to report as much detail as possible in their articles. Specifically, by researchers reporting specific *p*-values (McGrath, 2011), confidence intervals (Cumming, 2012), effect sizes (Grissom & Kim, 2012), and power analyses (e.g., Cohen, 1988; and see Denis, 2003), readers can form their own conclusions from a given study. Beale (1972) asserted that “The *p* level is for the reader's use, and [the reader] alone should be the one who decides whether the *p* level reported is significant” (p. 1080). Reporting specific *p*-values has also been proposed by contemporary authors (see Aguinis et al. 2010; Baker & Mudge, 2012). Careful consideration must be used here in distinguishing the utility of alpha and *p*-values for CD-NHST. It is essential that authors establish an appropriate (and hopefully context-driven) alpha value which allows the authors to evaluate and conclude if a given result occurs above random variation, assuming the null; reporting *p*-values allows readers to evaluate the results for themselves, though this practice does not remove the real need for authors to establish a justifiable alpha value (Knapp & Sawilowsky, 2001). In addition, confidence intervals can play a special role in reporting as they contain information concerning estimation with inference, which exceeds the utility of a *p*-value alone. In summary, these three suggestions can help to promote researchers in engaging in context-driven NHST while also attempting to minimize the bias inherent in researchers’ decisions.

Given the aforementioned arguments, it is not difficult to at least question the wisdom of a “one-size-fits-all” approach when using NHST. However, what exactly is CD-NHST in application? Current researchers and statisticians (e.g., Aguinis et al, 2010; Cohen, 1994; Cumming, 2012; Mudge et al., 2012; Murphy & Myors, 2004) have aptly decried the perpetuation of the exclusive and, admittedly, somewhat mindless use of the 0.05 critical or a nil difference of zero when making statistical inferences and original theorists (e.g., Fisher, 1926; 1935; 1973; Neyman & Pearson, 1933a; 1933b; Neyman, 1950; Yule & Kendall, 1950) never seemed to have promoted it in the first place. In contrast, context-driven NHST requires researchers to specify the values they use within the NHST framework and to be able to justify these values based on the context of their research. Within CD-NHST applications, it is important to clarify what context means in specific and various research settings. Context can include (but is certainly not limited to) the nature of the research area (both major field and subfields), the research question, the sampling methodology, the study design, the sample size, the measurement of the data, the ethical implications regarding the research and sample, and the quality of the data, along with researcher judgment and experience.

The hypothesized effect size, due to theory or past research, is also fundamental in driving CD-NHST, as it can influence what alpha value is selected, the sample size needed, and the value of the null. Likewise, the desired level of power for a study is essential in both contributing context and requiring context. Specifically, desired or hypothesized effect size and desired level of power are fundamental in determining an appropriate alpha value (balance of errors) and null value. In general, effect sizes (e.g., to determine magnitude of effects) and power considerations (e.g., study design of detecting real effects) along with confidence intervals (e.g., to illustrate uncertainty around estimates) have long been championed as essential components and/or supplements to NHST (e.g., Denis, 2003; Harlow, 2010; Robinson & Levin, 2010). These and other broad contextual considerations are suggested in a matrix in Table 1. This matrix is presented only to stimulate additional and deeper research context considerations and how they relate to the alpha value, null value and range, and should not be viewed as an exhaustive list, or even worse as a replacement ritual.

In general, contextual aspects of research help guide researchers in deciding which statistical tools to use (e.g., CD-NHST, modeling, Bayesian, etc.) and how to implement these tools to evaluate research questions (see Gigerenzer, 2004). Indeed, as Abelson (1997) asserted:

Table 1. Research context matrix

	Example Considerations	Critical Value	Value of Null	Range of Null
		α	H0	Δ+/-
Research Question	i. Specific Hypothesis			
	ii. No Hypothesis			
Study Type	i. Pilot			
	ii. Exploratory			
	iii. Confirmatory			
Measurement	i. Precise data (small variability/Reliable)			
	ii. Noisy data (Large variability/ less reliable)			
Field of Research	i. Biological Psychology (e.g., precise biomarkers)			
	ii. Clinical Psychology (e.g., self-report)			
Design	i. Experimental			
	ii. Observational			
	iii. Correlational			
Sampling	i. Probability-sampling			
	ii. Non-probability sampling			
	iii. Clinical sample			
Sample Size & Power	i. Small sample (related: underpowered)			
	ii. Large sample (related: overpowered)			
	iii. Level of power desired			
Cost	i. Type I error more costly relative to Type II error			
	ii. Type II error more costly relative to Type I error			
	iii. Cost of sample			
Seriousness	i. Type I error more serious relative to Type II error			
	ii. Type II error more serious relative to Type I error			
Replication	i. Study is a replication of another study			
	ii. Study is first of its kind			
	iii. Study will probably not be replicated			
Previous Data	i. Previous results indicate for this study…			
	ii. No previous results for this study…			
Effect size	i. Hypothesized magnitude of the effect, based on theory or past research.			

Note that each category is not mutually exclusive. For example, measurement variability is often closely related to field of research

> Good methodologists should be open to the possibility that a method does not apply in a particular case, or that more information is required. Statistical methods are better conceived as options than as commandments. Each method has areas of application in which it is typically useful, and areas in which it is weak or open to criticism. (p. 14)

Earlier, Neyman remarked:

> It may be useful to point out that although we are frequently witnessing controversies in which authors try to defend one or another system of the theory of probability as the only legitimate [one], I am of the opinion that several such theories may be and actually are legitimate, in spite of their occasionally contradicting one another. Each of these theories is based on some system of postulates, and so line as the postulates forming one particular system do not contradict each other and are sufficient to construct a theory, this is as legitimate as any other (Reid, 1982, p. 136).

Once the appropriate type of analysis is selected, researchers can use the context of the research to then guide and inform which values to use in the selected analysis. Although this holds for modeling and especially Bayesian analysis, which takes into account prior information, only the conventional NHST situation has been considered for the purposes of this paper. However, the field would benefit greatly from future work examining the issues regarding research context and other quantitative approaches such as statistical modeling (e.g., Harlow, 2010; McGrath, 2011; Rodgers, 2010a; 2010b).

Some may hold that NHST should be abandoned as an evaluative framework in science because it is often employed in a formulaic way. However, the argument presented here reveals that, regardless of how NHST may be commonly applied, it need not be used in a mechanistic way. Indeed, judging from original sources, it is questionable if null hypothesis testing or significance testing were ever designed to be used in the way they are applied today. Given the ability to designate alpha values and null values with context in mind, it is difficult to see why NHST is credited with being a mechanistic epistemological framework in the first place (see Rodgers, 2010a).

In closing, there are a number of errors that researchers must keep in mind when engaged in research. For instance, errors of the first kind are achieved when we incorrectly reject the null hypothesis whereas errors of the second kind are achieved when we incorrectly accept the null hypothesis (Neyman & Pearson, 1933b). These are the familiar errors that must be considered when selecting alpha. Mosteller (1948, p. 61) proposed an error of a third kind, whereby we correctly reject the null, but for the wrong reason. Later Marascuilo and Levin (1970, p. 398) proposed that errors of the fourth kind are achieved when we correctly reject the null hypothesis but give the wrong interpretation. It is proposed here that errors of the infinite kind are achieved when we correctly or incorrectly reject or accept the null hypothesis, but do so without context. That is, a limitless supply of error is available when we conclude without context.

Acknowledgements

This research was supported in part by G20RR030883 from the National Institutes of Health.

References

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*(1), 12-15. doi: 10.1111/j.1467-9280.1997.tb00536.x

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H. & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13*(3), 515-539. doi: 10.1177/1094428109333339

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423-437. doi: 10.1037/h0020412

Baker, L. F. & Mudge, J. F. (2012), Making statistical significance more significant. *Significance, 9*(3), 29-30. doi: 10.1111/j.1740-9713.2012.00574.x

Beale, D. K. (1972). What's so significant about .05? *American Psychologist, 27*(11), 1079-1080. doi: 10.1037/h0038057

Cascio, W. F., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology, 36*(3), 517-526. doi: 10.1111/j.1744-6570.1983.tb02233.x

Cochran, W. G. (1976). Early development of techniques in comparative experimentation. In D. B. Owen (Ed.), *On the history of statistics and probability* (pp. 1-26) New York, NY: Dekker.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994). The Earth is round (p < .05). *American Psychologist, 49*(12), 997-1003. doi: 10.1037/0003-066X.49.12.997

Cortina, J., & Landis, R. (2010). The Earth is *not* round (p = .00). *Organizational Research Methods*, *14*(2), 332-349. doi: 10.1177/1094428110391542

Cowles, M. & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*(5), 553-558. doi: 10.1037/0003-066X.37.5.553

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Denis, D. (2003). Alternatives to null hypothesis significance testing. *Theory & Science*, *4*(1). Retrieved from http://theoryandscience.icaap.org/content/vol4.1/02_denis.html

Feynman, R. (1985). *"Surely you're joking, Mr. Feynman!": Adventures of a curious character*. New York, NY: W. W. Norton & Company.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, *33*, 503-513.

Fisher, R. A. (1928). *Statistical methods for research workers* (2nd ed.). Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd ed.) London, UK: Collins Macmillan.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606. doi: 10.1016/j.socec.2004.09.033

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge.

Harlow, L. L., Mulaik, S., & Steiger, J. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harlow, L. L. (2010). On scientific research: The role of statistical modeling and hypothesis testing. *Journal of Modern Applied Statistical Methods*, *9*(2), 348-358. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol9/iss2/4/

Hodges, J. L. & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society. Series B*

(Methodological), 16(2), 261-268. Available from http://www.jstor.org/stable/2984052

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. Educational and Psychological Measurement, 60(5), 661-681. doi: 10.1177/00131640021970808

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Knapp, T. R. & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education, 70*(1), 65-79. doi: 10.1080/00220970109599498

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. *American Educational Research Journal, 7*(3), 397-421. Available from http://www.jstor.org/stable/1161635

McGrath, R. E. (2011). *Quantitative models in psychology*. Washington, DC: American Psychological Association.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108-141. doi: 10.1207/s15327965pli0102_1

Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *American Sociologist, 4*(2), 131-140. Available from http://www.jstor.org/stable/27701482

Mosteller, F. (1948). A *k*-sample slippage test for an extreme population. *The Annals of Mathematical Statistics, 19*(1), 58-65. Available from http://www.jstor.org/stable/2236056

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS ONE, 7*(2). doi: 10.1371/journal.pone.0032734

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum effect tests in the general linear model. *Journal of Applied Psychology, 84*(2), 234-248. doi: 10.1037/0021-9010.84.2.234

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Mahwah, NJ: Erlbaum.

Neyman, J. (1942). Basic ideas and theory of testing statistical hypotheses. *Journal of the Royal Statistical Society, 105*(4), 292-327. doi: 10.2307/2980436

Neyman, J. (1950). *First course in probability and statistics*. New York, NY: Henry Holt.

Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *Revue de l'Institut International de Statistique, 25*(1/3), 7-22. doi: 10.2307/1401671

Neyman, J., & Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, 231A*, 289-337.

Neyman, J., & Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society, 29*(04), 492-510. doi: 10.1017/S030500410001152X

Pearson, E. S. (1955). Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society*. *Series B (Methodological), 17*(2), 204-207. Available from http://www.jstor.org/stable/2983954

Reid, C. (1982). *Neyman – from life*. Berlin, Germany: Springer.

Robinson, D. H., & Levin, J. R. (2010). The not-so-quiet revolution: cautionary comments on the rejection of hypothesis testing in favor of a "causal" modeling alternative. *Journal of Modern Applied Statistical Methods, 9*(2), 332-339. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol9/iss2/2/

Rodgers, J. L. (2010a). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*(1), 1-12. doi: 10.1037/a0018326

Rodgers, J. L. (2010b). Statistical and mathematical modeling versus NHST? There's no competition! *Journal of Modern Applied Statistical Methods, 9*(2), 340-347. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol9/iss2/3/

Sawilowsky, S. S. (2003). Deconstructing arguments from the case against hypothesis testing. *Journal of Modern Applied Statistical Methods, 2*(2), 467-474. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol2/iss2/19/

Selvin, H. C. (1958). Reply to Gold's comment on "A critique of tests of significance". *American Sociological Review, 23*(1), 86. Available from http://www.jstor.org/stable/2089071

Serlin, R. C. (1987). Hypothesis testing, theory building, and the philosophy of science. *Journal of Counseling Psychology, 34*(4), 365-371. doi: 10.1037/0022-0167.34.4.365

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*(1), 73-83. doi: 10.1037/0003-066X.40.1.73

Skipper, J. K., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist, 2*(1), 16-18. Available from http://www.jstor.org/stable/27701229

Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London, UK: Griffin.