frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 03 March 2015
doi: 10.3389/fpsyg.2015.00223

Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing

Jose D. Perezgonzalez*

Business School, Massey University, Palmerston North, New Zealand

Edited by:
Lynne D. Roberts, Curtin University, Australia

Reviewed by:
Brody Heritage, Curtin University, Australia
Patrizio E. Tressoldi, Università di Padova, Italy

****Correspondence:***
Jose D. Perezgonzalez, Business School, Massey University, SST8.18, PO Box 11-222, Palmerston North 4442, New Zealand
e-mail: j.d.perezgonzalez@massey.ac.nz

Despite frequent calls for the overhaul of null hypothesis significance testing (NHST), this controversial procedure remains ubiquitous in behavioral, social and biomedical teaching and research. Little change seems possible once the procedure becomes well ingrained in the minds and current practice of researchers; thus, the optimal opportunity for such change is at the time the procedure is taught, be this at undergraduate or at postgraduate levels. This paper presents a tutorial for the teaching of data testing procedures, often referred to as hypothesis testing theories. The first procedure introduced is Fisher's approach to data testing—tests of significance; the second is Neyman-Pearson's approach—tests of acceptance; the final procedure is the incongruent combination of the previous two theories into the current approach—NSHT. For those researchers sticking with the latter, two compromise solutions on how to improve NHST conclude the tutorial.

Keywords: test of significance, test of statistical hypotheses, null hypothesis significance testing, statistical education, teaching statistics, NHST, Fisher, Neyman-Pearson

INTRODUCTION

This paper introduces the classic approaches for testing research data: tests of significance, which Fisher helped develop and promote starting in 1925; tests of statistical hypotheses, developed by Neyman and Pearson (1928); and null hypothesis significance testing (NHST), first concocted by Lindquist (1940). This chronological arrangement is fortuitous insofar it introduces the simpler testing approach by Fisher first, then moves onto the more complex one by Neyman and Pearson, before tackling the incongruent hybrid approach represented by NHST (Gigerenzer, 2004; Hubbard, 2004). Other theories, such as Bayes's hypotheses testing (Lindley, 1965) and Wald's (1950) decision theory, are not object of this tutorial.

The main aim of the tutorial is to illustrate the bases of discord in the debate against NHST (Macdonald, 2002; Gigerenzer, 2004), which remains a problem not only yet unresolved but very much ubiquitous in current data testing (e.g., Franco et al., 2014) and teaching (e.g., Dancey and Reidy, 2014), especially in the biological sciences (Lovell, 2013; Ludbrook, 2013), social sciences (Frick, 1996), psychology (Nickerson, 2000; Gigerenzer, 2004) and education (Carver, 1978, 1993).

This tutorial is appropriate for the teaching of data testing at undergraduate and postgraduate levels, and is best introduced when students are knowledgeable on important background information regarding research methods (such as random sampling) and inferential statistics (such as frequency distributions of means).

In order to improve understanding, statistical constructs that may bring about confusion between theories are labeled differently, attending to their function in preference to their historical use (Perezgonzalez, 2014). Descriptive notes (notes) and caution notes (caution) are provided to clarify matters whenever appropriate.

FISHER'S APPROACH TO DATA TESTING

Ronald Aylmer Fisher was the main force behind tests of significance (Neyman, 1967) and can be considered the most influential figure in the current approach to testing research data (Hubbard, 2004). Although some steps in Fisher's approach may be worked out a priori (e.g., the setting of hypotheses and levels of significance), the approach is eminently inferential and all steps can be set up a posteriori, once the research data are ready to be analyzed (Fisher, 1955; Macdonald, 1997). Some of these steps can even be omitted in practice, as it is relatively easy for a reader to recreate them. Fisher's approach to data testing can be summarized in the five steps described below.

Step 1–Select an appropriate test. This step calls for selecting a test appropriate to, primarily, the research goal of interest (Fisher, 1932), although you may also need to consider other issues, such as the way your variables have been measured. For example, if your research goal is to assess differences in the number of people in two independent groups, you would choose a chi-square test (it requires variables measured at nominal levels); on the other hand, if your interest is to assess differences in the scores that the people in those two groups have reported on a questionnaire, you would choose a *t*-test (it requires variables measured at interval or ratio levels and a close-to-normal distribution of the groups' differences).

Step 2–Set up the null hypothesis (H_0). The null hypothesis derives naturally from the test selected in the form of an exact statistical hypothesis (e.g., H_0: M1–M2 = 0; Neyman and Pearson, 1933; Carver, 1978; Frick, 1996). Some parameters of this hypothesis, such as variance and degrees of freedom, are estimated from the sample, while other parameters, such as the distribution of frequencies under a particular distribution, are deduced theoretically. The statistical distribution so established thus represents the random variability that is theoretically expected for a statistical

nil hypothesis (i.e., $H_0 = 0$) given a particular research sample (Fisher, 1954, 1955; Bakan, 1966; Macdonald, 2002; Hubbard, 2004). It is called the null hypothesis because it stands to be nullified with research data (Gigerenzer, 2004).

Among things to consider when setting the null hypothesis is its directionality.

Directional and non-directional hypotheses. With some research projects, the direction of the results is expected (e.g., one group will perform better than the other). In these cases, a directional null hypothesis covering all remaining possible results can be set (e.g., H_0: M1–M2 = 0). With other projects, however, the direction of the results is not predictable or of no research interest. In these cases, a non-directional hypothesis is most suitable (e.g., H_0: M1–M2 = 0).

Notes: *H_0 does not need to be a nil hypothesis, that is, one that always equals zero (Fisher, 1955; Gigerenzer, 2004). For example, H_0 could be that the group difference is not larger than certain value (Newman et al., 2001). More often than not, however, H_0 tends to be zero.*

Setting up H_0 is one of the steps usually omitted if following the typical nil expectation (e.g., no correlation between variables, no differences in variance among groups, etc.). Even directional nil hypotheses are often omitted, instead specifying that one-tailed tests (see below) have been used in the analysis.

Step 3-Calculate the theoretical probability of the results under H_0 (*p*). Once the corresponding theoretical distribution is established, the probability (*p*-value) of any datum under the null hypothesis is also established, which is what statistics calculate (Fisher, 1955, 1960; Bakan, 1966; Johnstone, 1987; Cortina and Dunlap, 1997; Hagen, 1997). Data closer to the mean of the distribution (**Figure 1**) have a greater probability of occurrence under the null distribution; that is, they appear more frequently and show a larger *p*-value (e.g., $p = 0.46$, or 46 times in a 100 trials). On the other hand, data located further away from the mean have a lower probability of occurrence under the null distribution; that is, they appear less often and, thus, show a smaller *p*-value (e.g., $p = 0.003$). Of interest to us is the probability of our research results under such null distribution (e.g., the probability of the difference in means between two research groups).

The *p*-value comprises the probability of the observed results and also of any other more extreme results (e.g., the probability of the actual difference between groups and any other difference more extreme than that). Thus, the *p*-value is a cumulative probability rather than an exact point probability: It covers the probability area extending from the observed results toward the tail of the distribution (Fisher, 1960; Carver, 1978; Frick, 1996; Hubbard, 2004).

Note: *P-values provide information about the theoretical probability of the observed and more extreme results under a null hypothesis assumed to be true (Fisher, 1960; Bakan, 1966), or, said otherwise, the probability of the data given a true hypothesis—P(D|H); (Carver, 1978; Hubbard, 2004). As H_0 is always true (i.e., it shows the theoretical random distribution of frequencies under certain parameters), it cannot, at the same time, be false nor falsifiable a posteriori. Basically, if at any point you say that H_0 is false, then you are also invalidating the whole test and its results. Furthermore, because H_0 is always true, it cannot be proved, either.*

Step 4–Assess the statistical significance of the results. Fisher proposed tests of significance as a tool for identifying research results of interest, defined as those with a low probability of occurring as mere random variation of a null hypothesis. A research result with a low *p*-value may, thus, be taken as evidence against the null (i.e., as evidence that it may not explain those results satisfactorily; Fisher, 1960; Bakan, 1966; Johnstone, 1987; Macdonald, 2002). How small a result ought to be in order to be considered statistically significant is largely dependent on the researcher in question, and may vary from research to research (Fisher, 1960; Gigerenzer, 2004). The decision can also be left to the reader, so reporting exact *p*-values is very informative (Fisher, 1973; Macdonald, 1997; Gigerenzer, 2004).

Overall, however, the assessment of research results is largely made bound to a given level of significance, by comparing whether the research *p*-value is smaller than such level of significance or not (Fisher, 1954, 1960; Johnstone, 1987):

- If the *p*-value is approximately equal to or smaller than the level of significance, the result is considered statistically significant.
- If the *p*-value is larger than the level of significance, the result is considered statistically non-significant.

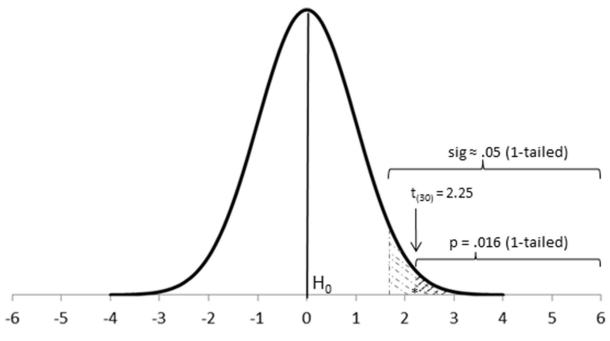

FIGURE 1 | Location of a *t*-value and its corresponding *p*-value on a theoretical *t* distribution with 30 degrees of freedom. The actual *p*-value conveys stronger evidence against H_0 than sig ≈0.05 and can be considered highly significant.

Among things to consider when assessing the statistical significance of research results are the level of significance, and how it is affected by the directionality of the test and other corrections.

Level of significance (sig). The level of significance is a theoretical *p*-value used as a point of reference to help identify statistically significant results (**Figure 1**). There is no need to set up a level of significance a priori nor for a particular level of significance to be used in all occasions, although levels of significance such as 5% (sig ≈0.05) or 1% (sig ≈0.01) may be used for convenience, especially with novel research projects (Fisher, 1960; Carver, 1978; Gigerenzer, 2004). This highlights an important

property of Fisher's levels of significance: They do not need to be rigid (e.g., *p*-values such as 0.049 and 0.051 have about the same statistical significance around a convenient level of significance of 5%; Johnstone, 1987).

Another property of tests of significance is that the observed *p*-value is taken as evidence against the null hypothesis, so that the smaller the *p*-value the stronger the evidence it provides (Fisher, 1960; Spielman, 1978). This means that it is plausible to gradate the strength of such evidence with smaller levels of significance. For example, if using 5% (sig ≈0.05) as a convenient level for identifying results which are just significant, then 1% (sig ≈0.01) may be used as a convenient level for identifying highly significant results and 1‰ (sig ≈0.001) for identifying extremely significant results.

Notes: *Setting up a level of significance is another step usually omitted. In such cases, you may assume the researcher is using conventional levels of significance.*

If both $H_0$and sig are made explicit, they could be joined in a single postulate, such as H_0: M_1–$M_2 = 0$, sig ≈0.05.

Notice that the p-value informs about the probability associated with a given test value (e.g., a t value). You could use this test value to decide about the significance of your results in a fashion similar to Neyman-Pearson's approach (see below). However, you get more information about the strength of the research evidence with p-values.

Although the p-value is the most informative statistic of a test of significance, in psychology (e.g., American Psychological Association, 2010) you also report the research value of the test—e.g., $t_{(30)} = 2.25$, $p = 0.016$, 1-tailed. Albeit cumbersome and largely ignored by the reader, the research value of the test offers potentially useful information (e.g., about the valid sample size used with a test).

Caution: *Be careful not to interpret Fisher's p-values as Neyman-Pearson's Type I errors (α, see below). Probability values in single research projects are not the same than probability values in the long run (Johnstone, 1987), something illustrated by Berger (2003)—who reported that $p = 0.05$ often corresponds to $\alpha = 0.5$ (or anywhere between $\alpha = 0.22$ and $\alpha > 0.5$)—and Cumming (2014)—who simulates the "dance" of p-values in the long run, commented further in Perezgonzalez (2015).*

One-tailed and two-tailed tests. With some tests (e.g., *F*-tests) research data can only be tested against one side of the null distribution (one-tailed tests), while other tests (e.g., *t*-tests) can test research data against both sides of the null distribution at the same time. With one-tailed tests you set the level of significance on the appropriate tail of the distribution. With two-tailed tests you cover both eventualities by dividing the level of significance between both tails (Fisher, 1960; Macdonald, 1997), which is commonly done by halving the total level of significance in two equal areas (thus covering, for example, the 2.5% most extreme positive differences and the 2.5% most extreme negative differences).

Note: *The tail of a test depends on the test in question, not on whether the null hypothesis is directional or non-directional. However, you can use two-tailed tests as one-tailed ones when testing data against directional hypotheses.*

Correction of the level of significance for multiple tests. As we introduced earlier, a *p*-value can be interpreted in terms of its expected frequency of occurrence under the specific null distribution for a particular test (e.g., $p = 0.02$ describes a result that is expected to appear 2 times out of 100 under H_0). The same goes for theoretical *p*-values used as levels of significance. Thus, if more than one test is performed, this has the consequence of also increasing the probability of finding statistical significant results which are due to mere chance variation. In order to keep such probability at acceptable levels overall, the level of significance may be corrected downwards (Hagen, 1997). A popular correction is Bonferroni's, which reduces the level of significance proportionally to the number of tests carried out. For example, if your selected level of significance is 5% (sig ≈0.05) and you carry out two tests, then such level of significance is maintained overall by correcting the level of significance for each test down to 2.5% (sig ≈0.05/2 tests ≈0.025, or 2.5% per test).

Note: *Bonferroni's correction is popular but controversial, mainly because it is too conservative, more so as the number of multiple tests increases. There are other methods for controlling the probability of false results when doing multiple comparisons, including familywise error rate methods (e.g., Holland and Copenhaver, 1987), false discovery rate methods (e.g., Benjamini and Hochberg, 1995), resampling methods (jackknifing, bootstrapping— e.g., Efron, 1981), and permutation tests (i.e., exact tests—e.g., Gill, 2007).*

Step 5–Interpret the statistical significance of the results. A significant result is literally interpreted as a dual statement: Either a rare result that occurs only with probability *p* (or lower) just happened, or the null hypothesis does not explain the research results satisfactorily (Fisher, 1955; Carver, 1978; Johnstone, 1987; Macdonald, 1997). Such literal interpretation is rarely encountered, however, and most common interpretations are in the line of "The null hypothesis did not seem to explain the research results well, thus we inferred that other processes—which we believe to be our experimental manipulation—exist that account for the results," or "The research results were statistically significant, thus we inferred that the treatment used accounted for such difference."

Non-significant results may be ignored (Fisher, 1960; Nunnally, 1960), although they can still provide useful information, such as whether results were in the expected direction and about their magnitude (Fisher, 1955). In fact, although always denying that the null hypothesis could ever be supported or established, Fisher conceded that non-significant results might be used for confirming or strengthening it (Fisher, 1955; Johnstone, 1987).

Note: *Statistically speaking, Fisher's approach only ascertains the probability of the research data under a null hypothesis. Doubting or denying such hypothesis given a low p-value does not necessarily "support" or "prove" that the opposite is true (e.g., that there is a difference or a correlation in the population). More importantly, it does not "support" or "prove" that whatever else has been done in the research (e.g., the treatment*

used) explains the results, either (Macdonald, 1997). For Fisher, a good control of the research design (Fisher, 1955; Johnstone, 1987; Cortina and Dunlap, 1997), especially random allocation, is paramount to make sensible inferences based on the results of tests of significance (Fisher, 1954; Neyman, 1967). He was also adamant that, given a significant result, further research was needed to establish that there has indeed been an effect due to the treatment used (Fisher, 1954; Johnstone, 1987; Macdonald, 2002). Finally, he considered significant results as mere data points and encouraged the use of meta-analysis for progressing further, combining significant and non-significant results from related research projects (Fisher, 1960; Neyman, 1967).

HIGHLIGHTS OF FISHER'S APPROACH

Flexibility. Because most of the work is done a posteriori, Fisher's approach is quite flexible, allowing for any number of tests to be carried out and, therefore, any number of null hypotheses to be tested (a correction of the level of significance may be appropriate, though—Macdonald, 1997).

Better suited for ad-hoc research projects. Given above flexibility, Fisher's approach is well suited for single, *ad-hoc*, research projects (Neyman, 1956; Johnstone, 1987), as well as for exploratory research (Frick, 1996; Macdonald, 1997; Gigerenzer, 2004).

Inferential. Fisher's procedure is largely inferential, from the sample to the population of reference, albeit of limited reach, mainly restricted to populations that share parameters similar to those estimated from the sample (Fisher, 1954, 1955; Macdonald, 2002; Hubbard, 2004).

No power analysis. Neyman (1967) and Kruskal and Savage (Kruskal, 1980) were surprised that Fisher did not explicitly attend to the power of a test. Fisher talked about sensitiveness, a similar concept, and how it could be increased by increasing sample size (Fisher, 1960). However, he never created a mathematical procedure for controlling sensitiveness in a predictable manner (Macdonald, 1997; Hubbard, 2004).

No alternative hypothesis. One of the main critiques to Fisher's approach is the lack of an explicit alternative hypothesis (Macdonald, 2002; Gigerenzer, 2004; Hubbard, 2004), because there is no point in rejecting a null hypothesis without an alternative explanation being available (Pearson, 1990). However, Fisher considered alternative hypotheses implicitly—these being the negation of the null hypotheses—so much so that for him the main task of the researcher—and a definition of a research project well done—was to systematically reject with enough evidence the corresponding null hypothesis (Fisher, 1960).

NEYMAN-PEARSON'S APPROACH TO DATA TESTING

Jerzy Neyman and Egon Sharpe Pearson tried to improve Fisher's procedure (Fisher, 1955; Pearson, 1955; Jones and Tukey, 2000; Macdonald, 2002) and ended up developing an alternative approach to data testing. Neyman-Pearson's approach is more mathematical than Fisher's and does much of its work a priori, at the planning stage of the research project (Fisher, 1955; Macdonald, 1997; Gigerenzer, 2004; Hubbard, 2004). It also introduces a number of constructs, some of which are similar to those of Fisher. Overall, Neyman-Pearson's approach to data testing can be considered tests of acceptance (Fisher, 1955; Pearson, 1955; Spielman, 1978; Perezgonzalez, 2014), summarized in the following eight main steps.

A PRIORI STEPS

Step 1–Set up the expected effect size in the population. The main conceptual innovation of Neyman-Pearson's approach was the consideration of explicit alternative hypotheses when testing research data (Neyman and Pearson, 1928, 1933; Neyman, 1956; Macdonald, 2002; Gigerenzer, 2004; Hubbard, 2004). In their simplest postulate, the alternative hypothesis represents a second population that sits alongside the population of the main hypothesis on the same continuum of values. These two groups differ by some degree: the effect size (Cohen, 1988; Macdonald, 1997).

Although the effect size was a new concept introduced by Neyman and Pearson, in psychology it was popularized by Cohen (1988). For example, Cohen's conventions for capturing differences between groups—d (**Figure 2**)—were based on the degree of visibility of such differences in the population: the smaller the effect size, the more difficult to appreciate such differences; the larger the effect size, the easier to appreciate such differences. Thus, effect sizes also double as a measure of importance in the real world (Nunnally, 1960; Cohen, 1988; Frick, 1996).

When testing data about samples, however, statistics do not work with unknown population distributions but with distributions of samples, which have narrower standard errors. In these cases, the effect size can still be defined as above because the means of the populations remain unaffected, but the sampling distributions would appear separated rather than overlapping (**Figure 3**). Because we rarely know the parameters of populations, it is their equivalent effect size measures in the context of sampling distributions which are of interest.

As we shall see below, the alternative hypothesis is the one that provides information about the effect size to be expected. However, because this hypothesis is not tested, Neyman-Pearson's procedure largely ignores its distribution except for a small percentage of it, which is called "beta" (β; Gigerenzer, 2004). Therefore, it is easier to understand Neyman-Pearson's procedure if we peg the effect size to beta and call it the expected minimum effect size (MES; **Figure 3**). This helps us conceptualize better how Neyman-Pearson's procedure works (Schmidt, 1996): The minimum effect size effectively represents that part of the main hypothesis that is not going to be rejected by the test (i.e., MES

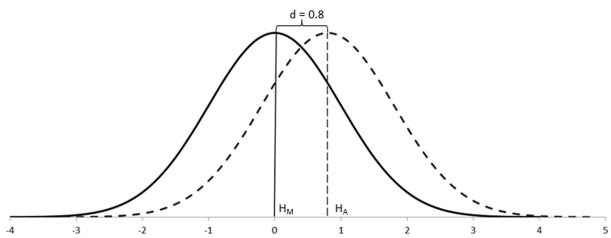

FIGURE 2 | A conventional large difference—Cohen's *d* = *0.8*—between two normally distributed populations, as a fraction of one standard deviation.

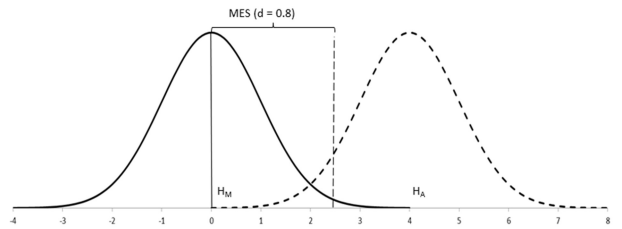

FIGURE 3 | Sampling distributions ($N = 50$ each) of the populations in Figure 2. MES ($d = 0.8$), assuming $\beta = 0.20$, is $d = 0.32$ (i.e., the expected difference in the population ranges between $d = 0.32$ and infinity).

captures values of no research interest which you want to leave under H_M; Cortina and Dunlap, 1997; Hagen, 1997; Macdonald, 2002). (Worry not, as there is no need to perform any further calculations: The population effect size is the one to use, for example, for estimating research power.)

Note: *A particularity of Neyman-Pearson's approach is that the two hypotheses are assumed to represent defined populations, the research sample being an instance of either of them (i.e., they are populations of samples generated by repetition of a common random process—Neyman and Pearson, 1928; Pearson, 1955; Hagen, 1997; Hubbard, 2004). This is unlike Fisher's population, which can be considered more theoretical, generated ad-hoc so as for providing the appropriate random distribution for the research sample at hand (i.e., a population of samples similar to the research sample—Fisher, 1955; Johnstone, 1987).*

Step 2–Select an optimal test. As we shall see below, another of Neyman-Pearson's contributions was the construct of the power of a test. A spin-off of this contribution is that it has been possible to establish which tests are most powerful (for example, parametric tests are more powerful than non-parametric tests, and one-tailed tests are more powerful than two-tailed tests), and under which conditions (for example, increasing sample size increases power). For Neyman and Pearson, thus, you are better off choosing the most powerful test for your research project (Neyman, 1942, 1956).

Step 3–Set up the main hypothesis (H_M). Neyman-Pearson's approach considers, at least, two competing hypotheses, although it only tests data under one of them. The hypothesis which is the most important for the research (i.e., the one you do not want to reject too often) is the one tested (Neyman and Pearson, 1928; Neyman, 1942; Spielman, 1973). This hypothesis is better off written so as for incorporating the minimum expected effect size within its postulate (e.g., H_M: $M_1–M_2 = 0 \pm$ MES), so that it is clear that values within such minimum threshold are considered reasonably probable under the main hypothesis, while values outside such minimum threshold are considered as more probable under the alternative hypothesis (**Figure 4**).

Caution: *Neyman-Pearson's H_M is very similar to Fisher's H_0. Indeed, Neyman and Pearson also called it the null hypothesis and often postulated it in a similar manner (e.g., as H_M: $M_1–M_2 = 0$). However, this similarity is merely superficial on three accounts: H_M needs to be considered at the design stage (H_0 is rarely made explicit); it is implicitly designed to incorporate any value below the MES—i.e., the a priori power analysis of a test aims to capture such minimum difference (effect sizes are not part of Fisher's approach); and it is but one of two competing explanations for the research results (H_0 is the only hypothesis, to be nullified with evidence).*

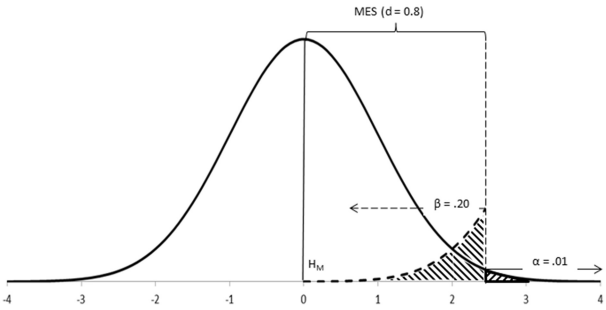

FIGURE 4 | Neyman-Pearson's approach tests data under H_M using the rejection region delimited by α. H_A contributes MES and β. Differences of research interest will be equal or larger than MES and will fall within this rejection region.

The main aspect to consider when setting the main hypothesis is the Type I error you want to control for during the research.

Type I error. A Type I error (or error of the first class) is made every time the main hypothesis is wrongly rejected (thus, every time the alternative hypothesis is wrongly accepted). Because the hypothesis under test is your main hypothesis, this is an error that you want to minimize as much as possible in your lifetime research (Neyman and Pearson, 1928, 1933; Neyman, 1942; Macdonald, 1997).

Caution: *A Type I error is possible under Fisher's approach, as it is similar to the error made when rejecting H_0 (Carver, 1978). However, this similarity is merely superficial on two accounts: Neyman and Pearson considered it an error whose relevance only manifests itself in the long run because it is not possible to know whether such an error has been made in any particular trial (Fisher's approach is eminently ad-hoc, so the risk of a long-run Type I error is of little relevance); therefore, it is an error that needs to be considered and minimized at the design stage of the research project in order to ensure good power—you cannot minimize this error a posteriori (with Fisher's approach, the potential impact of errors on individual projects is better controlled by correcting the level of significance as appropriate, for example, with a Bonferroni correction).*

Alpha (α). Alpha is the probability of committing a Type I error in the long run (Gigerenzer, 2004). Neyman and Pearson often worked with convenient alpha levels such as 5% ($\alpha = 0.05$) and 1% ($\alpha = 0.01$), although different levels can also be set. The main hypothesis can, thus, be written so as for incorporating the alpha level in its postulate (e.g., H_M: $M_1–M_2 = 0\pm$ MES, $\alpha = 0.05$), to be read as the probability level at which

the main hypothesis will be rejected in favor of the alternative hypothesis.

Caution: *Neyman-Pearson's α looks very similar to Fisher's sig. Indeed, Neyman and Pearson also called it the significance level of the test and used the same conventional cut-off points (5, 1%). However, this similarity is merely superficial on three accounts: α needs to be set a priori (not necessarily so under Fisher's approach); Neyman-Pearson's approach is not a test of significance (they are not interested in the strength of the evidence against H_M) but a test of acceptance (deciding whether to accept H_A instead of H_M); and α does not admit gradation—i.e., you may choose, for example, either $\alpha = 0.05$ or $\alpha = 0.01$, but not both, for the same test (while with Fisher's approach you can have different levels of more extreme significance).*

The critical region (CR_{test}) and critical value (CV_{test}, $Test_{crit}$) of a test. The alpha level helps draw a critical region, or rejection region (**Figure 4**), on the probability distribution of the main hypothesis (Neyman and Pearson, 1928). Any research value that falls outside this critical region will be taken as reasonably probable under the main hypothesis, and any research result that falls within the critical region will be taken as most probable under the alternative hypothesis. The alpha level, thus, also helps identify the location of the critical value of such test, the boundary for deciding between hypotheses. Thus, once the critical value is known—see below—, the main hypothesis can also be written so as for incorporating such critical value, if so desired (e.g., H_M: $M_1–M_2 = 0 \pm$ MES, $\alpha = 0.05$, $CV_t = 2.38$).

Caution: *Neyman-Pearson's critical region is very similar to the equivalent critical region you would obtain by using Fisher's sig as a cut-off point on a null distribution. However, this similarity is rather unimportant on three accounts: it is based on a critical value which delimits the region to reject H_M in favor of H_A, irrespective of the actual observed value of the test (Fisher, on the contrary, is more interested in the actual p-value of the research result); it is fixed a priori and, thus, rigid and immobile (Fisher's level of significance can be flexible—Macdonald, 2002); and it is non-gradable (with Fisher's approach, you may delimit several more extreme critical regions as areas of stronger evidence).*

Step 4–Set up the alternative hypothesis (H_A). One of the main innovations of Neyman-Pearson's approach was the consideration of alternative hypotheses (Neyman and Pearson, 1928, 1933; Neyman, 1956). Unfortunately, the alternative hypothesis is often postulated in an unspecified manner (e.g., as H_A: $M_1–M_2 \neq 0$), even by Neyman and Pearson themselves (Macdonald, 1997; Jones and Tukey, 2000). In practice, a fully specified alternative hypothesis (e.g., its mean and variance) is not necessary because this hypothesis only provides partial information to the testing of the main hypothesis (a.k.a., the effect size and β). Therefore, the alternative hypothesis is better written so as for incorporating the minimum effect size within its postulate (e.g., H_A: $M_1–M_2 \neq 0 \pm$ MES). This way it is clear that values beyond such minimum effect size are the ones considered of research importance.

Caution: *Neyman-Pearson's H_A is often postulated as the negation of a nil hypothesis (H_A: $M_1–M_2 \neq 0$), which is coherent with a simple postulate of H_M (H_M: $M_1–M_2 = 0$). These simplified postulates are not accurate and are easily confused with Fisher's approach to data testing—H_M resembles Fisher's H_0, and H_A resembles a mere negation of H_0. However, merely negating H_0 does not make its negation a valid alternative hypothesis—otherwise Fisher would have put forward such alternative hypothesis, something which he was vehemently against (Hubbard, 2004). As discussed earlier, Neyman-Pearson's approach introduces the construct of effect size into their testing approach; thus, incorporating such construct in the specification of both H_M and H_A makes them more accurate, and less confusing, than their simplified versions.*

Among things to consider when setting the alternative hypothesis are the expected effect size in the population (see above) and the Type II error you are prepared to commit.

Type II error. A Type II error (or error of the second class) is made every time the main hypothesis is wrongly retained (thus, every time H_A is wrongly rejected). Making a Type II error is less critical than making a Type I error, yet you still want to minimize the probability of making this error once you have decided which alpha level to use (Neyman and Pearson, 1933; Neyman, 1942; Macdonald, 2002).

Beta (β). Beta is the probability of committing a Type II error in the long run and is, therefore, a parameter of the alternative hypothesis (**Figure 4**, Neyman, 1956). You want to make beta as small as possible, although not smaller than alpha (if β needed to be smaller than α, then H_A should be your main hypothesis, instead!). Neyman and Pearson proposed 20% ($\beta = 0.20$) as an upper ceiling for beta, and the value of alpha ($\beta = \alpha$) as its lower floor (Neyman, 1953). For symmetry with the main hypothesis, the alternative hypothesis can, thus, be written so as for incorporating the beta level in its postulate (e.g., H_A: $M_1–M_2 \neq 0 \pm$ MES, $\beta = 0.20$).

Step 5–Calculate the sample size (N) required for good power (1–β). Neyman-Pearson's approach is eminently a priori in order to ensure that the research to be done has good power (Neyman, 1942, 1956; Pearson, 1955; Macdonald, 2002). Power is the probability of correctly rejecting the main hypothesis in favor of the alternative hypothesis (i.e., of correctly accepting H_A). It is the mathematical opposite of the Type II error (thus, 1–β; Macdonald, 1997; Hubbard, 2004). Power depends on the type of test selected (e.g., parametric tests and one-tailed tests increase power), as well as on the expected effect size (larger ES's increase power), alpha (larger α's increase power) and beta (smaller β's increase power). A priori power is ensured by calculating the correct sample size given those parameters (Spielman, 1973). Because power is the opposite of beta, the lower floor for good power is, thus, 80% ($1–\beta = 0.80$), and its upper ceiling is 1–alpha ($1–\beta = 1–\alpha$).

Note: *H_A does not need to be tested under Neyman-Pearson's approach, only H_M (Neyman and Pearson, 1928, 1933; Neyman, 1942; Pearson, 1955; Spielman, 1973). Therefore, the procedure looks similar to Fisher's and, under similar circumstances (e.g., when using the same test and sample size), it will lead to the same results. The main difference between procedures is that Neyman-Pearson's H_A provides explicit information to the test; that is, information about ES and β. If this information is not*

taken into account for designing a research project with adequate power, then, by default, you are carrying out a test under Fisher's approach.

Caution: *For Neyman and Pearson, there is little justification in carrying out research projects with low power. When a research project has low power, Type II errors are too big, so it is less probable to reject H_M in favor of H_A, while, at the same time, it makes unreasonable to accept H_M as the best explanation for the research results. If you face a research project with low a priori power, try the best compromise between its parameters (such as increasing α, relaxing β, settling for a larger ES, or using one-tailed tests; Neyman and Pearson, 1933). If all fails, consider Fisher's approach, instead.*

Step 6–Calculate the critical value of the test (CV_{test}, or $Test_{crit}$). Some of above parameters (test, α and N) can be used for calculating the critical value of the test; that is, the value to be used as the cut-off point for deciding between hypotheses (**Figure 5**, Neyman and Pearson, 1933).

A POSTERIORI STEPS

Step 7–Calculate the test value for the research (RV_{test}). In order to carry out the test, some unknown parameters of the populations are estimated from the sample (e.g., variance), while other parameters are deduced theoretically (e.g., the distribution of frequencies under a particular statistical distribution). The statistical distribution so established thus represents the random variability that is theoretically expected for a statistical main hypothesis given a particular research sample, and provides information about the values expected at different locations under such distribution.

By applying the corresponding formula, the research value of the test (RV_{test}) is obtained. This value is closer to zero the closer the research data is to the mean of the main hypothesis; it gets larger the further away the research data is from the mean of the main hypothesis.

Note: *P-values can also be used for testing data when using Neyman-Pearson's approach, as testing data under H_M is similar to testing data under Fisher's H_0 (Fisher, 1955). It implies calculating the theoretical probability of the research data under the distribution of H_M—$P(D|H_M)$. Just be mindful that p-values go in the opposite way than RVs, with larger p-values being closer to H_M and smaller p-values being further away from it.*

Caution: *Because of above equivalence, you may use p-values instead of CV_{test} with Neyman-Pearson's approach. However, p-values need to be considered mere proxies under this approach and, thus, have no evidential properties whatsoever (Frick, 1996; Gigerenzer, 2004). For example, if working with a priori $\alpha = 0.05$, $p = 0.01$ would lead you to reject H_M at $\alpha = 0.05$; however, it would be incorrect to reject it at $\alpha = 0.01$ (i.e., α cannot be adjusted a posteriori), and it would be incorrect to conclude that you reject H_M strongly (i.e., α cannot be gradated). If confused, you are better off sticking to CV_{test}, and using p-values only with Fisher's approach.*

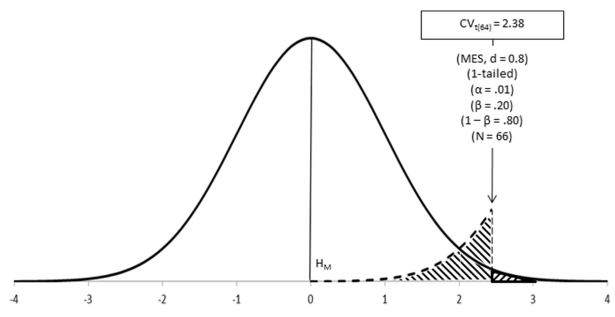

FIGURE 5 | Neyman-Pearson's test in action: CV_{test} is the point for deciding between hypotheses; it coincides with the cut-off points underlying α, β, and MES.

Step 8–Decide in favor of either the main or the alternative hypothesis. Neyman-Pearson's approach is rather mechanical once the a priori steps have been satisfied (Neyman and Pearson, 1933; Neyman, 1942, 1956; Spielman, 1978; Macdonald, 2002). Thus, the analysis is carried out as per the optimal test selected and the interpretation of results is informed by the mathematics of the test, following on the a priori pattern set up for deciding between hypotheses:

- If the observed result falls within the critical region, reject the main hypothesis and accept the alternative hypothesis.
- If the observed result falls outside the critical region and the test has good power, accept the main hypothesis.
- If the observed result falls outside the critical region and the test has low power, conclude nothing. (Ideally, you would not carry out research with low power—Neyman, 1955).

Notes: *Neyman-Pearson's approach leads to a decision between hypotheses (Neyman and Pearson, 1933; Spielman, 1978). In principle, this decision should be between rejecting H_M or retaining H_M (assuming good power), as the test is carried out on H_M only (Neyman, 1942). In practice, it does not really make much difference whether you accept H_M or H_A, as appropriate (Macdonald, 1997). In fact, accepting either H_M or H_A is beneficial as it prevents confusion with Fisher's approach, which can only reject H_0 (Perezgonzalez, 2014).*

Reporting the observed research test value is relevant under Neyman-Pearson's approach, as it serves to compare the observed value against the a priori critical value—e.g., $t_{(64)} = 3.31$, 1-tailed $> CV_t = 2.38$, thus accept H_A. When using a p-value as a proxy for CV_{test}, simply strip any evidential value off p—e.g., $t_{(64)} = 3.31$, $p < \alpha$, 1-tailed.

Neyman-Pearson's hypotheses are also assumed to be true. H_M represents the probability distribution of the data given a true hypothesis—$P(D|H_M)$, while H_A represents the distribution of the data under an alternative true hypothesis—$P(D|H_A)$, even when it is never tested. This means that H_M and H_A cannot be, at the same time false, nor proved or falsified a posteriori. The only way forward is to act as if the conclusion reached by the test was true—subject to a probability α or β of making a Type I or Type II error, respectively (Neyman and Pearson, 1933; Cortina and Dunlap, 1997).

HIGHLIGHTS OF NEYMAN-PEARSON'S APPROACH

More powerful. Neyman-Pearson's approach is more powerful than Fisher's for testing data in the long run (Williams et al., 2006). However, repeated sampling is rare in research (Fisher, 1955).

Better suited for repeated sampling projects. Because of above, Neyman-Pearson's approach is well-suited for repeated sampling research using the same population and tests, such as industrial quality control or large scale diagnostic testing (Fisher, 1955; Spielman, 1973).

Deductive. The approach is deductive and rather mechanical once the a priori steps have been set up (Neyman and Pearson, 1933; Neyman, 1942; Fisher, 1955).

Less flexible than Fisher's approach. Because most of the work is done a priori, this approach is less flexible for accommodating tests not thought of beforehand and for doing exploratory research (Macdonald, 2002).

Defaults easily to Fisher's approach. As this approach looks superficially similar to Fisher's, it is easy to confuse both and forget what makes Neyman-Pearson's approach unique (Lehman, 1993). If the information provided by the alternative hypothesis—ES and β—is not taken into account for designing research with good power, data analysis defaults to Fisher's test of significance.

NULL HYPOTHESIS SIGNIFICANCE TESTING

NHST is the most common procedure used for testing data nowadays, albeit under the false assumption of testing substantive hypotheses (Carver, 1978; Nickerson, 2000; Hubbard, 2004; Hager, 2013). NHST is, in reality, an amalgamation of Fisher's and Neyman-Pearson's theories, offered as a seamless approach to testing (Macdonald, 2002; Gigerenzer, 2004). It is not a clearly defined amalgamation either and, depending on the author describing it or on the researcher using it, it may veer more toward Fisher's approach (e.g., American Psychological Association, 2010; Nunnally, 1960; Wilkinson and the Task Force on Statistical Inference, 1999; Krueger, 2001) or toward Neyman-Pearson's approach (e.g., Cohen, 1988; Rosnow and Rosenthal, 1989; Frick, 1996; Schmidt, 1996; Cortina and Dunlap, 1997; Wainer, 1999; Nickerson, 2000; Kline, 2004).

Unfortunately, if we compare Fisher's and Neyman-Pearson's approaches vis-à-vis, we find that they are incompatible in most accounts (**Table 1**). Overall, however, most amalgamations follow Neyman-Pearson procedurally but Fisher philosophically (Spielman, 1978; Johnstone, 1986; Cortina and Dunlap, 1997; Hubbard, 2004).

NHST is not only ubiquitous but very well ingrained in the minds and current practice of most researchers, journal editors and publishers (Spielman, 1978; Gigerenzer, 2004; Hubbard, 2004), especially in the biological sciences (Lovell, 2013; Ludbrook, 2013), social sciences (Frick, 1996), psychology (Nickerson, 2000; Gigerenzer, 2004) and education (Carver, 1978, 1993). Indeed, most statistics textbooks for those disciplines still teach NHST rather than the two approaches of Fisher and of Neyman and Pearson as separate and rather incompatible theories (e.g., Dancey and Reidy, 2014). NHST has also the (false) allure of being presented as a procedure for testing substantive hypotheses (Macdonald, 2002; Gigerenzer, 2004).

In the situations in which they are most often used by researchers, and assuming the corresponding parameters are also the same, both Fisher's and Neyman-Pearson's theories work with the same statistical tools and produce the same statistical results; therefore, by extension, NHST also works with the same statistical tools and produces the same results—in practice, however, both approaches start from different starting points and lead to different outcomes (Fisher, 1955; Spielman, 1978; Berger, 2003). In a nutshell, the differences between Fisher's and Neyman-Pearson's theories are mostly about research philosophy and about how to interpret results (Fisher, 1955).

The most coherent plan of action is, of course, to follow the theory which is most appropriate for purpose, be this Fisher's or Neyman-Pearson's. It is also possible to use both for achieving different goals within the same research project (e.g., Neyman-Pearson's for tests thought of a priori, and Fisher's for exploring the data further, a posteriori), pending that those goals are not mixed up.

However, the apparent parsimony of NHST and its power to withstand threats to its predominance are also understandable. Thus, I propose two practical solutions to improve NHST: the first a compromise to improve Fisher-leaning NHST, the second a compromise to improve Neyman-Pearson-leaning NHST. A computer program such as G*Power can be used for implementing the recommendations made for both.

IMPROVING FISHER-LEANING NHST

Fisher's is the closest approach to NHST; it is also the philosophy underlying common statistics packages, such as SPSS. Furthermore, because using Neyman-Pearson's concepts within NHST may be irrelevant or inelegant but hardly damaging, it requires little re-engineering. A clear improvement to NHST comes from incorporating Neyman-Pearson's constructs of effect size and of a priori sample estimation for adequate power. Estimating effect sizes (both a priori and a posteriori) ensures that researchers consider importance over mere statistical significance. A priori estimation of sample size for good power also ensures that the research has enough sensitiveness for capturing the expected effect size (Huberty, 1987; Macdonald, 2002).

IMPROVING NEYMAN-PEARSON-LEANING NHST

NHST is particularly damaging for Neyman-Pearson's approach, simply because the latter defaults to Fisher's if important constructs are not used correctly. An importantly damaging issue is the assimilation of *p*-values as evidence of Type I errors and the subsequent correction of alphas to match such *p*-values (roving α's, Goodman, 1993; Hubbard, 2004). The best compromise for improving NHST under these circumstances is to compensate a posteriori roving alphas with a posteriori roving betas (or, if so preferred, with a posteriori roving power). Basically, if you are adjusting alpha a posteriori (roving α) to reflect both the strength of evidence (sig) and the long-run Type I error (α), you should also adjust the long-run probability of making a Type II error (roving β). Report both roving alphas and roving betas for each test, and take them into account when interpreting your research results.

Table 1 | Equivalence of constructs in Fisher's and Neyman-Pearson's theories, and amalgamation of constructs under NHST.

Concept	Fisher		Neyman-Pearson
Test object	Data—P(D|H_0)	=	Data—P(D|H_M)
NHST	➥	Data as if testing a falsifiable hypothesis—P(H_0|D)	➥
Approach	A posteriori	≠	A priori
NHST	➥	A posteriori, sometimes both	➥ (partly)
Research goal	Statistical significance of research results	≠	Deciding between competing hypotheses
NHST	➥	Statistical significance, also used for deciding between hypotheses	➥
Hs under test	H_0, to be nullified with evidence	≈	H_M, to be favored against H_A
NHST	➥	Both ($H_0 = H_M$)	➥
Alternative hypothesis	Not needed (implicitly, "No H_0")	≠	Needed. Provides ES and β
NHST	➥	H_A posed as 'No H_0' (ES and β sometimes considered)	➥ (partly)
Prob. distr. of test	As appropriate for H_0	=	As appropriate for H_M
NHST	➥	As appropriate for H_0	➥
Cut-off point	Sig identifies noteworthy results; can be gradated; can be corrected a posteriori	≠	Common to CV_{test}, α, β, and MES; cannot be gradated; cannot be corrected a posteriori
NHST	➥	Sig = α, can be gradated, can be corrected a posteriori	➥ (partly)
Sample size calculator	None	≠	Based on test, ES, α, and power (1 − β)
NHST	➥	Either	➥
Statistic of interest	*p*-value, as evidence against H_0	≠	CV_{test} (*p*-value has no inherent meaning but can be used as a proxy instead)
NHST	➥	*p*-value, used both as evidence against H_0 and a proxy to accept H_A	➥
Error prob.	α possible, but irrelevant with single studies	≠	α = Type I error prob. β = Type II error prob.
NHST	(partly) ➥	*p*-value = α = Type I error in single studies (β sometimes considered)	➥ (partly)
Result falls outside critical region	Ignore result as not significant	≠	Accept $H_M$if good power; conclude nothing otherwise
NHST	➥	Either ignore result as not significant; or accept H_0; or conclude nothing	➥
Result falls in critical region	Reject H_0	≠	Accept H_A (= Reject H_M in favor of H_A)
NHST	➥	Either	➥
Interpretation of results in critical region	Either a rare event occurred or H_0 does not explain the research data	≠	H_A explains research data better than H_M does (given α)
NHST		H_A has been proved / is true; or H_0 has been disproved / is false; or both	

(Continued)

Table 1 | Continued

Concept	Fisher		Neyman-Pearson
Next steps	Rejecting H_0 does not automatically justify not H_0. Replication needed, meta-analysis is useful.	≠	Impossible to know whether α error has been made. Repeated sampling of same population needed, Monte Carlo is useful.
NHST		None (results taken as definitive, especially if significant); further studies may be sometimes recommended (especially if results are not significant)	

Caution: NHST is very controversial, even if the controversy is not well known. A sample of helpful readings on this controversy are Christensen (2005); Hubbard (2004); Gigerenzer (2004); Goodman (1999), Louçã (2008, http://www.iseg.utl.pt/departamentos/economia/wp/wp022008deuece.pdf), Halpin and Stam (2006); Huberty (1993); Johnstone (1986), and Orlitzky (2012).

CONCLUSION

Data testing procedures represented a historical advancement for the so-called "softer" sciences, starting in biology but quickly spreading to psychology, the social sciences and education. These disciplines benefited from the principles of experimental design, the rejection of subjective probabilities and the application of statistics to small samples that Sir Ronald Fisher started popularizing in 1922 (Lehmann, 2011), under the umbrella of his tests of significance (e.g., Fisher, 1954). Two mathematical contemporaries, Jerzy Neyman and Egon Sharpe Pearson, attempted to improve Fisher's procedure and ended up developing a new theory, one for deciding between competing hypotheses (Neyman and Pearson, 1928), more suitable to quality control and large scale diagnostic testing (Spielman, 1973). Both theories had enough similarities to be easily confused (Perezgonzalez, 2014), especially by those less epistemologically inclined; a confusion fiercely opposed by the original authors (e.g., Fisher, 1955)—and ever since (e.g., Nickerson, 2000; Lehmann, 2011; Hager, 2013)—but something that irreversibly happened under the label of null hypothesis significance testing. NHST is an incompatible amalgamation of the theories of Fisher and of Neyman and Pearson (Gigerenzer, 2004). Curiously, it is an amalgamation that is technically reassuring despite it being, philosophically, pseudoscience. More interestingly, the numerous critiques raised against it for the past 80 years have not only failed to debunk NHST from the researcher's statistical toolbox, they have also failed to be widely known, to find their way into statistics manuals, to be edited out of journal submission requirements, and to be flagged up by peer-reviewers (e.g., Gigerenzer, 2004). NHST effectively negates the benefits that could be gained from Fisher's and from Neyman-Pearson's theories; it also slows scientific progress (Savage, 1957; Carver, 1978, 1993) and may be fostering pseudoscience. The best option would be to ditch NHST altogether and revert to the theories of Fisher and of Neyman-Pearson as—and when—appropriate. For everything else, there are alternative tools, among them exploratory data analysis (Tukey, 1977), effect sizes (Cohen, 1988), confidence intervals (Neyman, 1935), meta-analysis (Rosenthal, 1984), Bayesian applications (Dienes, 2014) and, chiefly, honest critical thinking (Fisher, 1960).

REFERENCES

American Psychological Association. (2010). *Publication Manual of the American Psychological Association, 6th Edn*. Washington, DC: APA.

Bakan, D. (1966). The test of significance in psychological research. *Psychol. Bull.* 66, 423–437. doi: 10.1037/h0020412

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57, 289–300.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Stat. Sci.* 18, 1–32. doi: 10.1214/ss/1056397485

Carver, R. P. (1978). The case against statistical significance testing. *Harv. Educ. Rev.* 48, 378–399. doi: 10.1080/00220973.1993.10806591

Carver, R. P. (1993). The case against statistical significance testing, revisited. *J. Exp. Educ.* 61, 287–292. doi: 10.1080/00220973.1993.10806591

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *Am. Stat.* 59, 121–126. doi: 10.1198/000313005X20871

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. New York, NY: Psychology Press.

Cortina, J. M., and Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychol. Methods* 2, 161–172. doi: 10.1037/1082-989X.2.2.161

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Dancey, C. P., and Reidy, J. (2014). *Statistics Without Maths for Psychology, 6th Edn*. Essex: Pearson Education.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* 68, 589–599. doi: 10.1093/biomet/68.3.589

Fisher, R. A. (1932). Inverse probability and the use of likelihood. *Proc. Camb. Philos. Soc.* 28, 257–261. doi: 10.1017/S0305004100010094

Fisher, R. A. (1954). *Statistical Methods for Research Workers, 12th Edn.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 17, 69–78.

Fisher, R. A. (1960). *The Design of Experiments, 7th Edn.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference, 3rd Edn.* London: Collins Macmillan.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science* 345, 1502–1505. doi: 10.1126/science.1255484

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychol. Methods* 1, 379–390. doi: 10.1037/1082-989X.1.4.379

Gigerenzer, G. (2004). Mindless statistics. *J. Soc. Econ.* 33, 587–606. doi: 10.1016/j.socec.2004.09.033

Gill, P. M. W. (2007). Efficient calculation of p-values in linear-statistic permutation significance tests. *J. Stat. Comput. Simul.* 77, 55–61. doi: 10.1080/10629360500108053

Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* 137, 485–496.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: the p-value fallacy. *Ann. Intern. Med.* 130, 995–1004. doi: 10.7326/0003-4819-130-12-199906150-00008

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *Am. Psychol.* 52, 15–24. doi: 10.1037/0003-066X.52.1.15

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: a methodological perspective. *Theor. Psychol.* 23, 251–270. doi: 10.1177/0959354312465483

Halpin, P. F., and Stam, H. J. (2006). Inductive inference or inductive behavior: fisher and Neyman–Pearson approaches to statistical testing in psychological research (1940-1960). *Am. J. Psychol.* 119, 625–653. doi: 10.2307/20445367

Holland, B. S., and Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics* 43, 417–423. doi: 10.2307/2531823

Hubbard, R. (2004). Alphabet soup. Blurring the distinctions between p's and α's in psychological research. *Theor. Psychol.* 14, 295–327. doi: 10.1177/0959354304043638

Huberty, C. J. (1987). On statistical testing. *Educ. Res.* 8, 4–9. doi: 10.3102/0013189X016008004

Huberty, C. J. (1993). Historical origins of statistical testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks. *J. Exp. Educ.* 61, 317–333. doi: 10.1080/00220973.1993.10806593

Johnstone, D. J. (1986). Tests of significance in theory and practice. *Statistician* 35, 491–504. doi: 10.2307/2987965

Johnstone, D. J. (1987). Tests of significance following R. A. Fisher. *Br. J. Philos. Sci.* 38, 481–499. doi: 10.1093/bjps/38.4.481

Jones, L. V., and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychol. Methods* 5, 411–414. doi: 10.1037/1082-989X.5.4.411

Kline, R. B. (2004). *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research.* Washington, DC: APA.

Krueger, J. (2001). Null hypothesis significance testing. On the survival of a flawed method. *Am. Psychol.* 56, 16–26. doi: 10.1037/0003-066X.56.1.16

Kruskal, W. (1980). The significance of Fisher: a review of R. A. Fisher: the life of a scientist. *J. Am. Stat. Assoc.* 75, 1019–1030. doi: 10.2307/2287199

Lehman, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypothesis: one theory or two? *J. Am. Stat. Assoc.* 88, 1242–1249. doi: 10.1080/01621459.1993.10476404

Lehmann, E. L. (2011). Fisher, Neyman, and the creation of classical statistics. New York, NY: Springer.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 1: Probability; Part 2: Inference*. Cambridge: Cambridge University Press.

Lindquist, E. F. (1940). *Statistical Analysis in Educational Research*. Boston, MA: Houghton Mifflin.

Louçã, F. (2008). *Should the Widest Cleft in Statistics - How and Why Fisher Opposed Neyman and Pearson.* WP/02/2008/DE/UECE, Technical University of Lisbon. Available online at: http://www.iseg.utl.pt/departamentos/economia/wp/wp022008deuece.pdf

Lovell, D. P. (2013). Biological importance and statistical significance. *J. Agric. Food Chem.* 61, 8340–8348. doi: 10.1021/jf401124y

Ludbrook, J. (2013). Should we use one-sided or two-sided p values in tests of significance? *Clin. Exp. Pharmacol. Physiol.* 40, 357–361. doi: 10.1111/1440-1681.12086

Macdonald, R. R. (1997). On statistical testing in psychology. *Br. J. Psychol.* 88, 333–347. doi: 10.1111/j.2044-8295.1997.tb02638.x

Macdonald, R. R. (2002). The incompleteness of probability models and the resultant implications for theories of statistical inference. *Underst. Stat.* 1, 167–189. doi: 10.1207/S15328031US0103_03

Newman, I., Fraas, J., and Herbert, A. (2001). "Testing non-nil null hypotheses with T tests of group means: a monte carlo study," in *Annual Meeting Mid-Western Educational Research Association* (Chicago, IL).

Neyman, J. (1935). On the problem of confidence intervals. *Ann. Math. Stat.* 6, 111–116. doi: 10.1214/aoms/1177732585

Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *J. R. Stat. Soc.* 105, 292–327. doi: 10.2307/2980436

Neyman, J. (1953). First *Course in Probability and Statistics*. New York, NY: Henry Holt.

Neyman, J. (1955). The problem of inductive inference. *Commun. Pure Appl.Math.* III, 13–46. doi: 10.1002/cpa.3160080103

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 18, 288–294.

Neyman, J. (1967). R. A. Fisher (1890-1962): an appreciation. *Science* 156, 1456–1460. doi: 10.1126/science.156.3781.1456

Neyman, J., and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika* 20A, 175–240. doi: 10.2307/2331945

Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. A* 231, 289–337. doi: 10.1098/rsta.1933.0009

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301. doi: 10.1037/1082-989X.5.2.241

Nunnally, J. (1960). The place of statistics in psychology. *Educ. Psychol. Meas.* 20, 641–650. doi: 10.1177/001316446002000401

Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organ. Res. Methods* 15, 199–228. doi: 10.1177/1094428111428356

Pearson, E. S. (1955). Statistical concepts in the relation to reality. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 17, 204–207.

Pearson, E. S. (1990). *'Student' a Statistical Biography of William Sealy Gosset*. Oxford: Claredon Press.

Perezgonzalez, J. D. (2014). A reconceptualization of significance testing. *Theor. Psychol.* 24, 852–859. doi: 10.1177/0959354314546157

Perezgonzalez, J. D. (2015). Confidence intervals and tests are two sides of the same research question. *Front. Psychol.* 6:34. doi: 10.3389/fpsyg.2015.00034

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research*. Beverly Hills, CA: Sage.

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

Savage, I. R. (1957). Nonparametric statistics. *J. Am. Stat. Assoc.* 52, 331–344. doi: 10.1080/01621459.1957.10501392

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol. Methods* 1, 115–129. doi: 10.1037/1082-989X.1.2.115

Spielman, S. (1973). A refutation of the Neyman-Pearson theory of testing. *Br. J. Philos. Sci.* 24, 201–222. doi: 10.1093/bjps/24.3.201

Spielman, S. (1978). Statistical dogma and the logic of significance testing. *Philos. Sci.* 45, 120–135. doi: 10.1086/288784

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychol. Methods* 4, 212–213. doi: 10.1037/1082-989X.4.2.212

Wald, A. (1950). *Statistical Decision Functions*. Ney York, NY: Wiley.

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. Guidelines and explanations. *Am. Psychol.* 54, 594–604. doi: 10.1037/0003-066X.54.8.594

Williams, R. H., Zimmerman, D. W., Ross, D. C., and Zumbo, B. D. (2006). *Twelve British Statisticians*. Raleigh, NC: Boson Books.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 January 2015; accepted: 13 February 2015; published online: 03 March 2015.

*Citation: Perezgonzalez JD (2015) Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front. Psychol. **6**:223. doi: 10.3389/fpsyg.2015.00223*

This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology.