Educational and Psychological
Measurement
1–18
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164416667981
epm.sagepub.com
SAGE

Tests of Statistical Significance Made Sound

Brian D. Haig[1]

Abstract

This article considers the nature and place of tests of statistical significance (ToSS) in science, with particular reference to psychology. Despite the enormous amount of attention given to this topic, psychology's understanding of ToSS remains deficient. The major problem stems from a widespread and uncritical acceptance of null hypothesis significance testing (NHST), which is an indefensible amalgam of ideas adapted from Fisher's thinking on the subject and from Neyman and Pearson's alternative account. To correct for the deficiencies of the hybrid, it is suggested that psychology avail itself of two important and more recent viewpoints on ToSS, namely the neo-Fisherian and the error-statistical perspectives. The neo-Fisherian perspective endeavors to improve on Fisher's original account and rejects key elements of Neyman and Pearson's alternative. In contrast, the error-statistical perspective builds on the strengths of both statistical traditions. It is suggested that these more recent outlooks on ToSS are a definite improvement on NHST, especially the error-statistical position. It is suggested that ToSS can play a useful, if limited, role in psychological research. At the end, some lessons learnt from the extensive debates about ToSS are presented.

Keywords

tests of statistical significance, null hypothesis significance testing, the neo-Fisherian perspective, the error-statistical perspective, statistical pragmatism, philosophy of statistics

It is well-known that tests of statistical significance (ToSS) are the most widely used means for evaluating hypotheses and theories in psychology. ToSS have been highly

[1] University of Canterbury, Christchurch, New Zealand

Corresponding Author:
Brian D. Haig, Department of Psychology, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand.
Email: brian.haig@canterbury.ac.nz

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

popular in psychology for more than 50 years and in the field of statistics for nearly 80 years. Since the 1960s, a massive critical literature has developed in psychology, and the behavioral sciences more generally, regarding the worth of ToSS (e.g., Harlow, Mulaik, & Steiger, 1997; Hubbard, 2016; Morrison & Henkel, 1970; Nickerson, 2000). Despite the plethora of critiques of ToSS, most psychologists understand them poorly, frequently use them inappropriately, and pay little attention to the controversy they have generated.

The significance testing controversy shows no signs of abating. Calls for replacing ToSS with alternative statistical methods have been prominent in recent debates. For example, an increasing number of methodologists have expressed a strong preference for the use of Bayesian statistics in place of the most popular form of ToSS, commonly known as *null hypothesis significance testing* (NHST; e.g., Dienes, 2011; Kruscke, 2015; Wagenmakers, 2007). Also, the so-called ''new statistics'' of effect sizes, confidence intervals, and meta-analysis, has been assiduously promoted as a worthy package to replace NHST (Cumming, 2014). Some journal editors too have played their part by endorsing alternatives to NHST. For instance, the recent editor of *Psychological Science* endorsed the use of the new statistics wherever appropriate (Eich, 2014), and the current editors of *Basic and Applied Social Psychology* have banned the use of NHST in articles published in their journal (Trafimow & Marks, 2015).

A noteworthy and surprising feature of these calls to do away with NHST is their failure to consider the sensible option of replacing it with defensible accounts of ToSS. The opponents of NHST seem to believe that arguments criticizing the worth of ToSS in its most indefensible form suffice to cast doubt on ToSS in its entirety. However, this is a clear case of faulty reasoning, known as ''the fallacy of the false dichotomy'': Reject NHST in favor of an alternative that does not involve ToSS, even though there are viable accounts of ToSS available for use.

A major objective of this article to bring two credible perspectives on ToSS to the attention of psychologists. I suggest that these alternative renditions of ToSS can play a legitimate, if limited, role in the prosecution of psychological research. In what follows, I provide a brief overview of NHST and point out its primary defects. I then provide an outline of the neo-Fisherian account of ToSS, which breaks from Neyman and Pearson's formulation and presents an update on Fisher's original position. The second option for a better understanding of ToSS is contained in the contemporary philosophy of statistics known as the *error-statistical philosophy*. The article ends with a list of important lessons learnt from the ongoing debates about ToSS that I believe we should carry forward in our thinking on the topic.

NHST: Psychology's Textbook Hybrid

Psychologists tend to assume that there is a single unified theory of ToSS. This assumption is primarily based on treatments of the topic furnished by the writers of statistics textbooks in psychology, who pay little, if any, attention to the work of the founding fathers on the topic. In contrast, it is well-known in professional statistical circles that there are two major historical theories of ToSS: Fisherian and Neyman–Pearsonian (e.g., Fisher, 1925; Neyman & Pearson, 1933). The relation between the two is a matter of some dispute. It is often said that Neyman and Pearson initially sought to build and improve on Fisher's theory, but that they subsequently developed their own theory as an alternative to that of Fisher. However, historians and theorists in statistics differ on how this relationship should be understood.

A popular view in statistical circles is that there are a number of fundamental points of difference between the two theories, which can be glossed as follows: Both theories adopt fundamentally different outlooks on the nature of scientific method and statistical inference. Fisher argued that an experiment is performed solely to give the data an opportunity to disprove the null hypothesis; no alternative hypothesis is specified, and the null hypothesis is the hypothesis to be nullified. Because one cannot accept the null hypothesis, no provision is made for a statistical concept of power. Fisher subscribed to an inductive conception of scientific method and maintained that significance tests were vehicles of inductive reasoning. For their part, Neyman and Pearson added the requirement of the specification of an alternative hypothesis and replaced Fisher's evidential *p* value with the Type I error rate. Type II error was admitted, and explicit provision was made for a formal statistical concept of power. Most fundamentally, Neyman and Pearson maintained that significance tests are rules of inductive behavior, not vehicles for inductive reasoning. This gloss on the two schools of thought should serve as a background to the following discussion of their hybridization.

In the behavioral sciences, the best-known account of the hybridized form of ToSS, NHST, is that of Gigerenzer (1993). Elaborating on a metaphor first suggested by Acree (1978), Gigerenzer employs Freudian language to identify the psychological tensions of those who use NHST. As he sees it, features of the Neyman-Pearsonian approach to hypothesis testing combine to form the superego of the hybrid logic and prescribe what should be done. The ego of the hybrid logic, which enables ToSS to be carried out, is that of Fisher. For Gigerenzer, there is a third component of the hybrid, which comes from neither Fisher nor Neyman and Pearson, but from the Bayesian desire to assign probabilities to hypotheses on the basis of the data. Gigerenzer likens this to the Freudian id because it is censored by the Neyman–Pearson superego and the Fisherian ego.

The nature of the psychologists' amalgam and its tensions can, on this received view, be redescribed thus: To the bare bones of Fisherian logic, the hybrid adds the notion of Type II error (opposed by Fisher) and the associated notion of statistical power (Fisher preferred the related notion of experimental sensitivity), but only at the level of rhetoric (thereby ignoring Neyman and Pearson), while giving a behavioral interpretation of both Type I and Type II errors (vigorously opposed by Fisher)!

There is, however, a further difference attributed to Fisher and Neyman and Pearson, the conflation of which serves to further characterize the amalgam. The inconsistency involves the equation of Fisher's *p* values with Neyman and Pearson's

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

Type I error rate, in the ubiquitous expression ''$p = \alpha$.'' However, these are said to be fundamentally different things (e.g., Hubbard, 2004). *P* values are measures of evidence, closely tied to the data they summarize, whereas alpha values are rates of error that apply to the tests being used. Fisher, it is said, thought that error rates had no place in his account of significance testing. For their part, Neyman and Pearson are portrayed as thinking that *p* values had no place in their conception of hypothesis testing. However, the claim that the amalgam brings together two ideas that their originators thought were irreconcilable is challenged by the error-statistical perspective, as I shall note later.

As just seen, Gigerenzer employs the psychodynamic metaphor as a device for organizing some of the sources of confusion that he thinks comprise the hybrid in the minds of many psychological researchers, journal editors, and textbook writers. However, like all metaphors, it has its limitations. For one thing, it provides a psychological construal of methodological ideas and their relations that might be more illuminatingly cast in more direct methodological terms. For another, it provides a set of hypotheses about the mind-set (the ''psychic structure'') of researchers who employ NHST that lacks proper empirical confirmation. Evidence from protocol analyses of verbal reports of researchers would be required for such confirmation. In addition, this psychological characterization of psychologists' understanding of the hybrid does not take account of the fact that the confusions contained in the amalgam are exacerbated by a tendency of psychologist to misrepresent further the key features of ToSS in a number of ways. For example, levels of statistical significance are taken as measures of confidence in research hypotheses, information about likelihoods is taken as a gauge of the credibility of the hypotheses under test, and reported levels of significance are taken as measures of the replicability of the findings (e.g., Hubbard, 2016). Additional misunderstandings such as these make a psychological characterization of the hybrid beyond the resources of the Freudian metaphor to provide.

It should be said further that there is not a single agreed-upon characterization of the hybrid NHST, as seems to be supposed in treatments of the topic. Halpin and Stam (2006) examined the formulation of the hybrid in six statistics textbooks in psychology published in the period 1940-1960 and found that it received different characterizations. For example, the textbooks differed in the extent to which they made use of ideas from Neyman and Pearson. Relatedly, the authors discovered that the textbooks took ideas from both Fisher and Neyman and Pearson, but that the journal literature that they reviewed made virtually no use of Neyman and Pearson's ideas.

As just intimated, the view that NHST is an inchoate amalgam of Fisher's and Neyman and Pearson's schools of thought is based on the commonly held belief that the two schools are fundamentally different, and irreconcilable. However, this belief is not held universally among professional statisticians. For example, Lehmann (1993), a former student of Neyman, maintains that although there are some important philosophical differences between the two schools, the strongly voiced differences of opinion between their founders give the misleading impression that the

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

schools are incompatible. Lehmann contends that at a practical level, the two approaches are complementary and that ''*p* values, fixed-level significance statements, conditioning, and power considerations can be combined into a unified approach'' (1993, p. 1248). Spanos too, adopts the view that the two approaches are complementary. In his well-known textbook (Spanos, 1999), he concludes that the Neyman–Pearsonian approach is suited for testing within the boundaries of a postulated model, whereas the Fisherian approach is suited for testing outside the boundaries of the model. As will be seen, the error-statistical philosophy demonstrates that a number of elements of both schools of thought can be incorporated in a wide-ranging, coherent position. However, before presenting and discussing the main features of that philosophy, I consider the more circumscribed neo-Fisherian outlook on ToSS.

The Neo-Fisherian Perspective

As its name implies, the neo-Fisherian perspective on ToSS is a reformulation of Fisher's original position. Advocates of this perspective include Cox (2006), Hurlbert and Lombardi (2009), Pace and Salvan (1997), and to some extent in his later years, Fisher himself. In an extensive recent critical review, Hurlbert and Lombardi (2009) comprehensively surveyed the literature on ToSS and recommend a shift in focus from the original ''paleo-Fisherian'' and Neyman–Pearsonian classical frameworks to what they maintain is a more defensible neo-Fisherian alternative. For ease of exposition, and convenient reference for the reader, I largely follow the authors' characterization of the neo-Fisherian position. I briefly identify its major elements and indicate how the authors depart from, and see themselves rejecting, the psychologists' hybrid, while improving on problematic elements of Fisher's original position, and rejecting the Neyman–Pearsonian outlook. That said, Hurlbert and Lombardi in fact retain some elements of the latter position, namely alternative hypotheses, power, and confidence intervals.

1. *Type I error rate is not specified.* In a clear departure from standard practice, critical alphas, or probabilities of Type I error, are not specified. Instead, exact *p* values are reported. The publication of Fisher's statistical tables with fixed *p* values was a matter of pragmatic convenience and should not be taken to imply that ToSS requires fixed *p* values to be chosen. Moreover, the refusal to accept the null hypothesis when an obtained *p* value barely exceeds the adopted value is both rigid and unsound. An alpha value of .051 has the same evidential import as one of .049.
2. *P values are not misleadingly described as ''significant'' or ''nonsignificant.''* There is no requirement that the dichotomous ''significant''/''nonsignificant'' language and thinking be used. Indeed, it is recommended that talk of ''statistically significant'' and ''statistically nonsignificant'' results be

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

dropped. Undoubtedly, Fisher's publication of critical values of test statistics played a major role in the widespread adoption of this misleading language.

3. *Judgment is suspended about accepting the null hypothesis on the basis of high p values.* It is not uncommon for textbook authors, and researchers especially, to think that when a p value is greater than a specified level of significance, one should accept the null hypothesis as true. However, the neo-Fisherian perspective regards it as neither necessary nor sufficient to accept the null hypothesis on the basis of high p values. Factors, such as the strength of experimental conditions, the magnitude of an effect, and power considerations, will have a bearing on whether or not this belief is sound.
4. *The ''three-valued logic'' that gives information about the direction of the effect being tested is adopted.* The logical structure of standard ToSS is a ''two-valued logic'' by which one chooses between two mutually exclusive hypotheses about the direction of an effect. However, Kaiser (1960), Harris, (1997), and others reason that the researcher who adopts the traditional two-tailed test cannot reach a conclusion about the direction of the effect being tested, and one who employs a one-tailed test cannot conclude that the predicted sign of the effect is wrong. Their proposed solution is to adopt a more nuanced ''three-valued logic,'' where a test for just two hypotheses is replaced by a test of three hypotheses that allows for conclusions about effects with either sign, or an expression of doubt and reserved judgment.
5. *Adjunct information about effect sizes and confidence intervals is provided, if appropriate.* It is a common criticism of traditional ToSS to decry the over-emphasis on p values by researchers and their associated neglect of effect sizes and confidence intervals. As noted earlier, some methodologists recommend the abandonment of p value statistics in favor of statistics such as these. However, the neo-Fisherian position retains the emphasis on p values in significance assessments and regards effect sizes and confidence intervals as complements to such tests, rather than as alternatives to them. It is important to remember that effect sizes and confidence intervals are faced with their own challenges. For example, the common practice of reporting effects sizes as ''small,'' ''medium,'' and ''large,'' without interpreting them substantively, is of limited value. Also, confidence intervals are vulnerable to some of the same charges that are levelled against p values, such as the large n problem. This problem arises from the fact that discrepancies from any (simple) null hypothesis, however small, can be detected by a (frequentist) ToSS with a large enough sample size (Spanos, 2014).
6. *A clear distinction is made between statistical and substantive significance.* A source of much confusion in the use and interpretation of ToSS is the conflation of statistical and substantive hypotheses (e.g., Bolles, 1962; Cox, 1958). In the domain of statistical concepts that draws selectively from Fisher and Neyman and Pearson, both the null and the alternative hypotheses are statistical hypotheses. Researchers and textbook writers correctly assume that

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

> rejection of the null implies acceptance of the alternative hypothesis, but they too often err in treating the alternative hypothesis as a research, or scientific, hypothesis rather than as a statistical hypothesis. Substantive knowledge of the domain in question is required to formulate a scientific hypothesis that corresponds to the alternative hypothesis. The neo-Fisherian perspective is directly concerned with testing statistical hypotheses as distinct from scientific hypotheses, and it forbids concluding that statistical significance implies substantive significance. At the same time, it urges researchers to explicitly specify the link between the two, warning that sometimes the former may have a small role in establishing the latter.

The neo-Fisherian paradigm contains a package of pragmatic reforms that overcomes some of the problems of NHST, and it improves on aspects of Fisher's original perspective in some respects. Importantly, it represents a reasoned case for retaining *p*-valued significance testing without the focus on hybrid NHST. Although the neo-Fisherian position shares with the error-statistical approach a distrust of the Bayesian outlook on statistics, it differs from the error-statistical approach in rejecting the Neyman–Pearsonian perspective. However, Hurlbert and Lombardi's (2009) claim that the neo-Fisherian position signals the ''final collapse'' of the Neyman–Pearsonian framework is questionable, for two reasons: First, as noted earlier, some elements of the Neyman and Pearson's outlook are retained by the authors. Second, the founder of the error-statistical approach, Deborah Mayo, maintains that the neo-Fisherian approach does not go far enough (reported in Hurlbert & Lombardi, 2009, p. 326), presumably because of its inability to draw key insights from Neyman and Pearson's outlook, such as the notion of error probabilities. In any case, it will become clear that the error-statistical approach provides a more comprehensive outlook on statistical inference than the neo-Fisherian position does.

The Error-Statistical Perspective

An important part of scientific research involves processes of detecting, correcting, and controlling for error, and mathematical statistics is one branch of methodology that helps scientists do this. In recognition of this fact, the philosopher of statistics and science, Deborah Mayo (e.g., Mayo, 1996), in collaboration with the econometrician, Aris Spanos (e.g., Mayo & Spanos, 2010, 2011), has systematically developed, and argued in favor of, an *error-statistical* philosophy for understanding experimental reasoning in science. Importantly, this philosophy permits, indeed encourages, the local use of ToSS, among other methods, to manage error.

In the error-statistical philosophy, the idea of an experiment is understood broadly to include controlled experiments, observational studies, and even thought experiments. What matters in all these types of inquiry is that a planned study permits one to mount reliable arguments from error. By using statistics, the researcher is able to model ''what it would be like to control, manipulate, and change in situations where

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

we cannot literally'' do so (Mayo, 1996, p. 459). Furthermore, although the error-statistical approach has broad application within science, it is concerned neither with all of science nor with error generally. Instead, it focuses on scientific *experimentation* and error *probabilities*, which ground knowledge obtained from the use of statistical methods.

Development of the Error-Statistical Philosophy

In her initial formulation of the error-statistical philosophy, Mayo (1999) modified, and built upon, the classical Neyman–Pearsonian approach to ToSS. However, in later publications with Spanos (e.g., Mayo & Spanos, 2011), and in writings with David Cox (Cox & Mayo, 2010; Mayo & Cox, 2010), her error-statistical approach has come to represent a coherent blend of many elements, including both Neyman–Pearsonian and Fisherian thinking. For Fisher, reasoning about p values is based on *postdata*, or after-trial, consideration of probabilities, whereas Neyman and Pearson's Type I and Type II errors are based on *predata*, or before-trial, error probabilities. The error-statistical approach assigns each a proper role that serves as an important complement to the other (Mayo & Spanos, 2011; Spanos, 2010). Thus, the error-statistical approach partially resurrects and combines, in a coherent way, elements of two perspectives that have been widely considered to be incompatible. In the post-data element of this union, reasoning takes the form of severe testing, a notion to which I now turn.

The Severity Principle

Central to the error-statistical approach is the notion of a severe test, which is a means of gaining knowledge of experimental effects. An adequate test of an experimental claim must be a severe test in the sense that relevant data must be good evidence for a hypothesis. Thus, according to the error-statistical perspective, a sufficiently severe test should conform to the *severity principle*, which has two variants: A *weak severity principle* and a *full severity principle*. The weak severity principle acknowledges situations where we should deny that data are evidence for a hypothesis. Adhering to this principle discharges the investigator's responsibility to identify and eliminate situations where an agreement between data and hypothesis occurs when the hypothesis is false. Mayo and Spanos (2011) state the principle as follows:

> Data $\mathbf{x}_0$ (produced by process G) do not provide good evidence for hypothesis H if $\mathbf{x}_0$ results from a test procedure with a very low probability or capacity of having uncovered the falsity of H, even if H is incorrect. (p. 162)

However, this negative conception of evidence, although important, is not sufficient; it needs to be conjoined with the positive conception of evidence to be found in the full severity principle. Mayo and Spanos (2011) formulate the principle thus,

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

> Data $\mathbf{x}_0$ (produced by process G) provide good evidence for hypothesis H (just) to the extent that test T has severely passed H with $\mathbf{x}_0$. (p. 162)

With a severely tested hypothesis, the probability is low that test procedure would pass muster if the hypothesis was false. Furthermore, the probability that the data agree with the alternative hypothesis must be very low. The full severity principle is the key to the error-statistical account of evidence and provides the core of the rationale for the use of error-statistical methods. The error probabilities afforded by these methods provide a measure of how frequently the methods can discriminate between alternative hypotheses, and how reliably they can detect errors.

Error-Statistical Methods

The error-statistical approach constitutes an inductive approach to scientific inquiry. However, unlike favored inductive methods that emphasize the broad logical nature of inductive reasoning (notably, the standard hypothetico-deductive method and the Bayesian approach to scientific inference), the error-statistical approach furnishes context-dependent, local accounts of statistical reasoning. It seeks to rectify the troubled foundations of Fisher's account of inductive inference, makes selective use of Neyman and Pearson's behaviorist conception of inductive behavior, and endorses Charles Peirce's (1931-1958) view that inductive inference is justified pragmatically in terms of self-correcting inductive methods.

The error-statistical approach employs a wide variety of error-statistical methods to link experimental data to theoretical hypotheses. These include the panoply of standard frequentist statistics that use error probabilities assigned on the basis of the relative frequencies of errors in repeated sampling, such as ToSS and confidence interval estimation, which are used to collect, model, and interpret data. They also include computer-intensive resampling methods, such as the bootstrap, Monte Carlo simulations, nonparametric methods, and ''noninferential'' methods for exploratory data analysis. In all this, ToSS have a minor, though useful, role.

A Hierarchy of Models

In the early 1960s, Patrick Suppes (1962) suggested that science employs a hierarchy of models that ranges from experimental experience to theory. He claimed that theoretical models, which are high on the hierarchy, are not compared directly with empirical data, which are low on the hierarchy. Rather, they are compared with models of the data, which are higher than data on the hierarchy. The error-statistical approach similarly adopts a framework in which three different types of models are interconnected and serve to structure error-statistical inquiry: primary models, experimental models, and data models. Primary models break down a research question into a set of local hypotheses that can be investigated using reliable methods. Experimental models structure the particular models at hand and serve to link

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

primary models to data models. And, data models generate and model raw data, as well as checking whether the data satisfy the assumptions of the experimental models. The error-statistical approach (Mayo & Spanos, 2010) has also been extended to primary models and theories of a more global nature. The hierarchy of models employed in the error-statistical perspective exhibits a structure similar to the important threefold distinction between data, phenomena, and theory (Woodward, 1989; see also Haig, 2014). These similar threefold distinctions accord better with scientific practice than the ubiquitous coarse-grained data-theory/model distinction.

Error-Statistical Philosophy and Falsificationism

The error-statistical approach shares a number of features with Karl Popper's (1959) falsificationist theory of science. Both stress the importance of identifying and correcting errors for the growth of scientific knowledge, both focus on the importance of hypothesis testing in science, and both emphasize the importance of strong tests of hypotheses. However, the error-statistical approach differs from Popper's theory in a number of respects: It focuses on *statistical* error and its role in *experimentation*, neither of which were considered by Popper. It employs a range of statistical methods to test for error. And, in contrast with Popper, who deemed deductive inference to be the only legitimate form of inference, it stresses the importance of inductive reasoning in its conception of science. This error-statistical stance regarding Popper can be construed as a constructive interpretation of Fisher's oft-cited remark that the null hypothesis is never proved, only possibly disproved.

Error-Statistical Philosophy and Bayesianism

The error-statistical philosophy is arguably the major alternative to the reigning Bayesian philosophy of statistical inference. Indeed, in her first major presentation of the error-statistical outlook, Mayo often used Bayesian ideas as a foil in its explication (Mayo, 1996). For one thing, the error-statistical approach rejects the Bayesian insistence on characterizing the evidential relation between hypothesis and evidence in a universal and logical manner in terms of Bayes's theorem via conditional probabilities. It chooses instead to formulate the relation in terms of the substantive and specific nature of the hypothesis and the evidence with regard to their origin, modeling, and analysis. This is a consequence of a commitment to a contextual approach to testing using the most appropriate methods available. Furthermore, the error-statistical philosophy rejects the classical Bayesian commitment to the subjective nature of fathoming prior probabilities in favor of the more objective process of establishing error probabilities understood in frequentist terms. It also finds the turn to ''objective'' Bayesianism unsatisfactory, but it is not my purpose in this article to rehearse those arguments against that form of Bayesianism. Finally, the error-statistical outlook employs probabilities to measure how effectively *methods* facilitate the detection of error, and how those methods enable us to choose between

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

alternative hypotheses. Bayesians are not concerned with error probabilities at all. Instead, they use probabilities to measure *belief* in hypotheses or degrees of confirmation. This is a major point of difference between the two philosophies.

Virtues of the Error-Statistical Approach

The error-statistical approach has a number of strengths, which I enumerate at this point without justification (1) it boasts a philosophy of statistical inference, which provides guidance for thinking about, and constructively using, common statistical methods, including ToSS, for the conduct of scientific experimentation. Statistical methods are often employed with a shallow understanding that comes from ignoring their accompanying theory and philosophy; (2) it has the conceptual and methodological resources to enable one to avoid the common misunderstandings of ToSS, which afflict so much empirical research in the behavioral sciences; (3) it provides a challenging critique of, and alternative to, the Bayesian way of thinking in both statistics and current philosophy of science; moreover, it is arguably the major modern alternative to the Bayesian philosophy of statistics; (4) finally, the error-statistical approach is not just a philosophy of statistics concerned with the growth of experimental knowledge. It is also regarded by Mayo and Spanos as a general philosophy of science. As such, its authors employ error-statistical thinking to cast light on vexed philosophical problems to do with scientific inference, modeling, theory testing, explanation, and the like. A critical evaluation by prominent philosophers of science of the early extension of the error-statistical philosophy to the philosophy of science more generally can be found in Mayo and Spanos (2010).

As just noted, the error-statistical perspective addresses a wide-range of misunderstandings of ToSS and criticisms of error-statistical methods more generally. Mayo and Spanos (2011) address a baker's dozen of these challenges and show how their error-statistical outlook on statistics corrects the misunderstandings, and counters the criticisms, of ToSS. These include the allegation that error-statistical methods preclude the use of background knowledge, the contention that the fallacies of rejection and acceptance are perpetuated by ToSS, the claim that confidence-interval estimation should replace ToSS, and the charge that testing model assumptions amounts to unwarranted data-mining. Mayo and Spanos's (2011) reply to these challenges constitutes an important part of the justification of the error-statistical perspective. Because of space limitations, I briefly consider the claims about the fallacies of acceptance and rejection only.

Fallacies of rejection involve the misinterpretation of statically significant differences. The best known example of such a fallacy is the conflation of statistical and substantive significance, which was discussed earlier. This conflation is frequently made by psychological researchers when they employ ToSS. The misinterpretation involves accepting the correctness of a substantive hypothesis solely on the basis of confirming a statistical hypothesis. This is more likely to happen with a Fisherian use of statistical tests because it carries with it no rival statistical hypothesis to

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

compare with the null hypothesis. Of course, the provision of a statistical alternative to the null, in the manner of Neyman and Pearson, might help put a brake on those who would otherwise commit the fallacy. The error-statistical perspective incorporates this feature of Neyman and Pearson's approach, explicitly stresses the importance of the distinction between statistical and substantive hypotheses, and urges that it be respected when reasoning back and forth between the data, experimental, and primary models described earlier.

Fallacies of acceptance involve taking statistically insignificant differences as grounds for believing that the null hypothesis is true. The basic mistake here is to think that an absence of evidence against the null hypothesis can be taken as evidence for the null hypothesis, as for example when the test used has insufficient power to detect the existing discrepancies. Crucially, the error-statistical approach appeals to the strategy of severe testing to guard against the fallacies of acceptance and rejection. It does this by using postdata assessments of evidence based on the reasoning involved in severe testing. The severity involved formalizes the intuition that p values have different evidential import, depending on the size of the sample, or, more generally, the power of the test under consideration (see Mayo & Spanos, 2006, 2011 for details).

What Should We Think About Tests of Significance?

Before concluding this article, I enumerate some of the important lessons that I believe can be taken from the extensive debates about the nature and merits of ToSS. Some of these draw from the statistics literature, others from scientific methodology, more generally. These are necessarily presented in brief form. Not all the material relevant to these lessons has been canvassed in the body of the article, but I summon up the chutzpah to present them, nonetheless.

1. *NHST should not be employed in research.* NHST, understood as the variable, inchoate amalgam of elements of Fisherian and Neyman-Pearsonian thinking, should be abandoned because of its incoherence. Its presence in textbooks and research publications has done, and continues to do, untold damage to psychology. The reasoning in research articles that appeals to the illogic of NHST is either impossible to fathom, or the conclusions it gives rise to are unjustified. Psychology's defective statistics education has provided a shallow understanding of ToSS that has resulted in its researchers mechanically employing the hybrid NHST without sufficient awareness of its origins and problems. Moreover, psychology has remained blind to the possibilities of combining elements of different schools of statistical thought in defensible hybrid packages.
2. *Defensible forms of ToSS should be employed, where appropriate*. It is a mistake to believe that we should give up, or ban, ToSS because of the unsatisfactory nature of its most popular form, NHST. Psychologists are

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

almost entirely unaware that there are credible forms of ToSS, primary among which are the neo-Fisherian and the error-statistical perspectives. Unfortunately, psychology has yet to show an awareness of the fact that these are viable replacements for NHST that can do useful work in data analysis and scientific inference. Methodologists in psychology have a duty to inform themselves about these alternatives to NHST and make considered recommendations about them for researchers in the field. Relatedly, advocates of alternatives to NHST, including some Bayesians (e.g., Wagenmakers, 2007) and the new statisticians (e.g., Cumming, 2014), have had an easy time of it by pointing out the flaws in NHST and showing how their preferred approach does better. However, I think it is incumbent on them to consider plausible versions of ToSS, such as the neo-Fisherian and error-statistical approaches, when arguing for the superiority of their own positions.

3. *There are a number of legitimate research goals for ToSS.* More specifically, ToSS can do useful local work in different research contexts that involves separating signal from noise. These include pattern detection in exploratory contexts (recommended by Fisher), assistance in judgments about the presence of experimental effects (again, recommended by Fisher [though frequently misused by scientists]), and strong probes designed to detect error in hypotheses under test (a key feature of the error-statistical perspective). Seldom, will it be appropriate to rely on p values exclusively (Senn, 2001). Rather, it will mostly be appropriate to employ effect sizes and confidence intervals as complements to ToSS, but that too will depend on context. Generally speaking, I maintain that these supplements should not be used as replacements for ToSS. Finally, the claim made by some opponents of ToSS that such tests are seldom used in the physical sciences (e.g., McCloskey & Ziliak, 1996) is false (Hoover & Siegler, 2008). ToSS have been, and continue to be, used to good purpose by many researchers in the physical sciences. An instructive example of their informed and rigorous use in physics is the recent discovery of a Higgs boson (van Dyk, 2014).
4. *Maintaining the distinction between statistical and substantive hypotheses is of paramount importance.* As noted earlier, both the neo-Fisherian and error-statistical perspectives stress the importance of distinguishing between statistical and substantive hypotheses. Despite the fact that ToSS assess statistical hypotheses only, psychologists frequently take them to have direct implications for substantive hypotheses. Moreover, statistical hypotheses play a subservient role to substantive hypotheses and theories, which are the major focus of scientific attention. This is one of a number of reasons why ToSS should have a lesser role to play in the assessment of scientific hypotheses and theories than psychology has generally accorded them.
5. *An attitude of strong methodological pluralism should be adopted.* The totalizing tendency to be found among some Bayesian statisticians (e.g., Lindley, 2000) and advocates of the Bayesian way in psychology, who

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

argue for the uptake of Bayesian rationality across the board (e.g., Dienes, 2011), should be resisted. The local use of statistics that are fit for purpose is much to be preferred. Similarly, the suggestion of the new statisticians that data analysts should, wherever possible, seek parameter estimates for effect sizes and confidence intervals, underappreciates the need for a strong methodological pluralism in which a host of quite different research goals are pursued by employing different statistical methods. Psychology stands to benefit from a greater use of additional statistical methods, such as exploratory data analysis, computer intensive resampling methods, and robust statistics, to mention only a few.

6. *Statistical pragmatism is a viable stance.* Arguably, an attitude of statistical pragmatism should be encouraged in our use of statistics. Thus, a blending of insights from seemingly opposed schools of statistical thought, which has been built on different philosophical outlooks, is both possible, and sometimes desirable, at the level of practice. For example, thoughtful Bayesian/frequentist compromises that exploit the insights of both statistical traditions are common in contemporary statistics and some sciences, though they are absent from psychology. Andrew Gelman's heterodox view of Bayesian statistics (e.g., Gelman & Shalizi, 2013) is a good example of the statistical pragmatism I have in mind: It involves the contextual use of Bayesian statistics without buying into the usual inductive Bayesian philosophy of science. Instead, it involves something like a Popperian hypothetico-deductive testing of models, which, moreover, Gelman thinks is consistent with the error-statistical philosophy. This is an example of a ''principled'' form of pragmatism, in the sense that it comprises an explicitly thought-out philosophy of statistics.
7. *Adopting a broad perspective on statistics is important.* A broad perspective on statistics is needed to counter the widespread tendency among both scientists and methodologists to view statistics through a narrow lens. Arguably, the error-statistical and Bayesian outlooks are the two most prominent approaches in this regard. The error-statistical approach adopts a broad perspective on the use of statistics in science, as its overview in this article makes clear. It has a well-developed philosophy, is concerned with much more than data analysis (e.g., the design of experiments and the validation of model assumptions), and encourages the use of a wide range of statistical methods. The Bayesian outlook on statistics can also be viewed in broad compass, especially if it is joined with a Bayesian philosophy of science and its attendant theory of confirmation—something that most Bayesian statisticians are reluctant do. Further work on the comparative evaluation of the error-statistical and Bayesian perspectives is to be encouraged.
8. *There is a need to go beyond standard hypothetico-deductivism in science.* The dominant ''significant difference'' paradigm, with its use of hybridized forms of NHST embedded in an impoverished view of the hypothetico-deductive method, is of questionable value. This paradigm contrasts with

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

the error-statistical perspective and its conception of hypothetico-deductive testing, augmented by a statistical-inductive approach with strong tests. Moreover, hypothesis and theory testing in science is far from all-important. Taken together, the tasks of theory construction, including theory generation, theory development, and multicriterial theory appraisal, are much more important than just testing for predictive success. One viable replacement for NHST is the ‘‘significance sameness’’ paradigm developed by Hubbard and Lindsay (e.g., Hubbard, 2016). This paradigm seeks to establish empirical generalizations using effect sizes, confidence intervals, and replication practices, where appropriate, before seeking to understand them through the abductive construction of explanatory theories. Related outlooks on the construction of explanatory theories are to be found in Grice (2011) and Haig (2014).

9. *There is a need for different sorts of statistics textbooks.* Psychology needs better statistics textbooks, written by specialists who have a good appreciation of modern statistical theory, as well as an understanding of how statistics operate in the prosecution of successful science. To date, statistics textbooks in psychology have been written mainly by nonspecialists, who have made limited use of statistical theory, who have presented NHST as though it was a justified whole, and who have shown a reluctance to replace it with better alternatives. Spanos's *Probability Theory and Statistical Inference* (1999), mentioned earlier, is a good example of a textbook that exhibits the desirable features just mentioned. Moreover, his book provides an instructive account of the historical development of ToSS and shows how the Fisherian and Neyman–Pearsonian outlooks can be regarded as complementary. One might expect that its next edition will embrace the fuller-bodied error-statistical outlook.
10. *Statistical methods should be taught through methodology.* Finally, and importantly, I strongly believe that our understanding of ToSS, and other statistical methods, should be enhanced by a greater familiarity with the full range of interdisciplinary contributions to methodology, in addition to our knowledge of statistical practice. Important among these are statistical theory, the philosophy and history of statistics, and statistical cognition. To take just one of these, the value of the philosophy of statistics as an aid to our understanding of ToSS has been considerably underrated by researchers and methodologists in psychology. The error-statistical perspective presented in this article is in fact a full-blown philosophy of statistics. As such, it brings with it a deep understanding of the role of ToSS and associated methods, which is made possible by an extensive knowledge of the nature of science and its statistical practices, the history and conceptual foundations of statistics, and the philosophy of science more generally (Mayo, 2011, 2012). Philosophy these days is said to be naturalized—that is to say, it is regarded as continuous with science, arguably a *part* of science and is concerned with

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

foundational issues *in* science. So located, the philosophy of statistics is well-positioned to contribute in important ways to our understanding of statistical theory and practice. Because of this, it deserves to be part of any curriculum that aspires to provide a genuine education in statistics.

Conclusion

Although this article is broad-brush in nature, I hope that it will stimulate both psychological researchers and their institutions to think further and deeper about the nature of ToSS and their proper place in research. In more than 50 years of preoccupation with these tests, psychology has concentrated its gaze on teaching, using, and criticizing NHST in its muddled hybrid form. It is high time for the discipline to bring itself up-to-date with best thinking on the topic, and employ sound versions of ToSS in its research.

Acknowledgments

The author wishes to thank Ray Hubbard and Deborah Mayo for their helpful comments on an earlier version of this article.

Declaration of Conflicting Interests

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

Funding

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

References

Acree, M. C. (1978). *Theories of statistical inference in psychological research: A historico-critical study* (University Microfilms No. H790 H7000). Ann Arbor, MI: University Microfilms International.

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, *11*, 639-645.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, *29*, 357-372.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge, England: Cambridge University Press.

Cox, D. R., & Mayo, D. G. (2010). Objectivity and conditionality in frequentist inference. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 276-304). New York, NY: Cambridge University Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274-290.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3-6.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8-38.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum.

Grice, J. W. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. San Diego, CA: Academic Press.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. Cambridge: MIT Press.

Halpin, P. F., & Stam, H. J. (2006). Inductive inference or inductive behavior: Fisher and Neyman-Pearson approaches to statistical testing in psychological research (1940-1960). *American Journal of Psychology*, *119*, 625-653.

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

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145-174). Mahwah, NJ: Lawrence Erlbaum.

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

Hubbard, R. (2004). Alphabet soup: Blurring the distinction between p's and α's in psychological research. *Theory & Psychology*, *14*, 295-327.

Hubbard, R. (2016). *Corrupt research: The case for reconceptualising empirical management and social science*. Thousand Oaks, CA: Sage.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici*, *46*, 311-349.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, *67*, 160-167.

Kruscke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam, Netherlands: Elsevier.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, *88*, 1242-1249.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*, 293-319.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago, IL: University of Chicago Press.

Mayo, D. G. (2011). Statistical science and philosophy of science: Where do/should they meet in 2011 (and beyond)? *Rationality, Markets and Morals*, *2*, 79-102.

Mayo, D. G. (2012). Statistical science meets philosophy of science, part 2: Shallow versus deep explorations. *Rationality, Markets and Morals*, *3*, 71-107.

Mayo, D. G., & Cox, D. (2010). Frequentist statistics as a theory of inductive inference. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 247-304). New York, NY: Cambridge University Press.

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, *57*, 323-357.

Mayo, D. G., & Spanos, A. (Eds.). (2010). *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*. New York, NY: Cambridge University Press.

Mayo, D. G., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of philosophy of Science: Vol. 7. Philosophy of statistics* (pp. 153-198). Amsterdam, Netherlands: Elsevier.

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

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy: A reader*. Chicago, IL: Aldine.

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

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.

Pace, L., & Salvan, A. (1997). *Advanced series on statistical science and applied probability: Vol. 4. Principles of statistical inference from a neo-Fisherian perspective*. Singapore: World Scientific.

Peirce, C. S. (1931-1958). *The collected papers of Charles Sanders Peirce* (Vols. 1-8; C. Hartshorne & P. Weiss [Eds., Vols. 1-6], & A. W. Burks [Ed., Vols. 7-8]). Cambridge, MA: Harvard University Press.

Popper, K. R. (1959). *The logic of scientific discovery*. London, England: Hutchinson.

Senn, S. (2001). Two cheers for *P*-values? *Journal of Epidemiology and Biostatistics*, *6*, 193-204.

Spanos, A. (1999). *Probablity theory and statistical inference: Economic modeling with observational data*. Cambridge, England: Cambridge University Press.

Spanos, A. (2010). On a new philosophy of frequentist inference: Exchanges with David Cox and Deborah G. Mayo. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 315-330). New York, NY: Cambridge University Press.

Spanos, A. (2014). Recurring controversies about *P* values and confidence intervals revisited. *Ecology*, *95*, 645-651.

Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology, and philosophy of science: Proceedings of the 1960 International Congress* (pp. 252-261). Stanford, CA: Stanford University Press.

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

Van Dyk, D. A. (2014). The role of statistics in the discovery of a Higgs Boson. *Annual Review of Statistics and Its Applications*, *1*, 41-59.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779-804.

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*, 393-472.

Downloaded from epm.sagepub.com at University Libraries | Virginia Tech on November 1, 2016