The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://amstat.tandfonline.com/loi/utas20

The Limited Role of Formal Statistical Inference in Scientific Inference

Raymond Hubbard, Brian D. Haig & Rahul A. Parsa

To cite this article: Raymond Hubbard, Brian D. Haig & Rahul A. Parsa (2019) The Limited Role of Formal Statistical Inference in Scientific Inference, The American Statistician, 73:sup1, 91-98, DOI: 10.1080/00031305.2018.1464947

To link to this article: https://doi.org/10.1080/00031305.2018.1464947

Published online: 20 Mar 2019.

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 91–98: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1464947

The Limited Role of Formal Statistical Inference in Scientific Inference

Raymond Hubbard[a], Brian D. Haig[b], and Rahul A. Parsa[c]

[a]College of Business and Public Administration, Drake University, Des Moines, IA; [b]Department of Psychology, University of Canterbury, Christchurch, New Zealand; [c]Debbie and Jerry Ivy School of Business, Iowa State University, Ames, IA

ABSTRACT

Such is the grip of formal methods of statistical inference—that is, frequentist methods for generalizing from sample to population in enumerative studies—in the drawing of scientific inferences that the two are routinely deemed equivalent in the social, management, and biomedical sciences. This, despite the fact that legitimate employment of said methods is difficult to implement on practical grounds alone. But supposing the adoption of these procedures were simple does not get us far; crucially, methods of formal statistical inference are ill-suited to the analysis of much scientific data. Even findings from the claimed gold standard for examination by the latter, randomized controlled trials, can be problematic.

Scientific inference is a far broader concept than statistical inference. Its authority derives from the accumulation, over an extensive period of time, of both theoretical and empirical knowledge that has won the (provisional) acceptance of the scholarly community. A major focus of scientific inference can be viewed as the pursuit of *significant sameness*, meaning replicable and empirically generalizable results among phenomena. Regrettably, the obsession with users of statistical inference to report *significant differences* in data sets actively thwarts cumulative knowledge development.

The manifold problems surrounding the implementation and usefulness of formal methods of statistical inference in advancing science do not speak well of much teaching in methods/statistics classes. Serious reflection on statistics' role in producing viable knowledge is needed. Commendably, the American Statistical Association is committed to addressing this challenge, as further witnessed in this special online, open access issue of *The American Statistician*.

ARTICLE HISTORY
Received December 2017
Revised February 2018

KEYWORDS
Analytic studies;
Enumerative studies;
Observational studies;
Randomized controlled trials;
Significant difference;
Significant sameness;
Scientific inference;
Statistical inference

"Much of causal inference is beyond statistics" (Shadish and Cook 1999, p. 298).

"Statistical inference …is fundamentally incompatible with 'most' science" (Gunter and Tong 2016–2017, p. 1).

1. Introduction

Employment of formal methods of statistical inference, that is, frequentist methods for generalizing from sample to population in enumerative studies, is ubiquitous in empirical work in the social, management, and biomedical sciences. Indeed, statistical inference and scientific inference generally are taken to be one and the same in these areas. They are not. The present paper argues that "objective" methods of statistical inference play a limited role in scientific inference.

Their prevalence notwithstanding, strictly pragmatic considerations make employment of such statistical methods challenging. That is, the seemingly prosaic steps listed routinely in textbooks for the legitimate application of these tools in fact pose numerous and compounding obstacles hampering their use. More critically, formal statistical inference is geared to the conduct of enumerative studies (Deming 1975) whose object is the estimation of fixed population parameters from random samples. Yet such investigations are the exceptions, not the rule. Most scientific research involves analytic or predictive studies which do not lend themselves to analysis by formal inferential approaches (Deming 1975; Hahn and Meeker 1993). Even the procedure touted as ideal for showcasing the merits of formal statistical inference, the randomized controlled trial, has issues meeting the assumptions justifying its adoption, and can yield troublesome results (Cartwright 2007).

The much broader concept of scientific inference (versus statistical inference) is then introduced. In our telling, this revolves around the search for *significant sameness* (Ehrenberg 1975; Hubbard 2016; Hubbard and Lindsay 2013a, b; Nelder 1986), that is, the discovery of replicable and empirically generalizable findings. These discoveries, and their causes, are fueled in the main by observational (nonrandomized) studies.

Concluding thoughts reiterate the fact that, while occasionally important, overall the part played by formal statistical inference in scientific inference is relatively minor. This, in turn, raises questions concerning the teaching of this topic, and the texts informing it, which unwittingly exaggerate its contribution.

2. How Did We Get to this State of Affairs?

In many fields of inquiry, like the social sciences, management sciences, and areas of biomedicine, the notions of statistical inference and scientific inference are viewed as all but synonymous. This has led to distorted, naïve impressions about how

CONTACT Raymond Hubbard drabbuhyar@aol.com Thomas F. Sheehan Distinguished Professor of Marketing, Emeritus, College of Business and Public Administration, Drake University, Des Moines, IA 50311.

science makes headway which have endured to the present day. Foremost among these is the widespread belief that scientific progress arises from the application of formal methods of statistical inference to random samples of data, something Guttman (1985, p. 3) characterizes as "illogical," and Gigerenzer (2004, p. 587) as "mindless." This, then, begs the question: How did matters get to be this way?

Irony abounds in answering this question, which draws heavily from Hubbard (2016, pp. 258–261). This is because the emergent social sciences, and the field of statistics, were concerned primarily with the discovery of significant sameness in their datasets, the very concept we are now trying to rehabilitate!

Early statisticians and social scientists alike subscribed to the French positivism of August Compte, which philosophy posited that all phenomena, both natural and social, are governed by universal laws. Compte maintained that the researcher's job was to detect the laws influencing societal behaviors.

During the 19th century social statisticians uncovered many laws, manifested as empirical regularities, in social data. Chief among them was the Belgian astronomer, Adolphe Quetelet, generally hailed as the father of both sociology (along with Compte) and statistics itself. Quetelet applied probability theory in detecting these empirical regularities. More specifically, he made use of the astronomer's law of error, better known today as the Gaussian or normal curve. Quetelet fit this distribution to biological data, including human birth and death rates, height, weight, and shoe size, and compared them by sex, age, place of residence, and profession. Others found evidence of this approximate normal curve regularity in social data, such as the incidence of crime, drunkenness, illness, insanity, marriage, suicide, and even in musical and poetic talent.

Quetelet demonstrated that although individual behaviors may be unpredictable, they are lawlike in the aggregate, conforming to a normal distribution. It was through this that he viewed the statistical regularity as the cardinal means of understanding social science. In addition, Quetelet regarded Poisson's "law of large numbers" to be the essential axiom of his "social physics." Succinctly put, statistical regularities were the means of discerning the universal laws of society.

Quetelet's belief in the universality of the error law had enormous repercussions on the development of 19th century science and statistics. As an example, the renowned physicist James Maxwell was intrigued by Quetelet's work. In particular, during the 1860s he wanted to know whether Quetelet's normal law might be adapted from its social setting involving large numbers of people to studying the behavior of gas molecules. Maxwell's success in this effort was a landmark in exposing the probabilistic nature of the physical world, previously thought to be deterministic. Likewise, brilliant scientists like Francis Edgeworth, Francis Galton, Gustav Fechner, and Karl Pearson were descendants of Quetelet, who all understood the crucial role played by what we call the pursuit of significant sameness.

Yet the notion of significant sameness lost currency, from which it has never recovered, when Ronald Fisher's ideas began dominating the statistics and social science disciplines in the 20th century. Whereas his predecessors favored the idea of large-scale descriptive studies, Fisher ushered in the age of the significant difference paradigm, with its focus on small sample theory, inferential methods, and statistical significance tests. Another question that arises, then, is this: How was the significant difference paradigm able to displace the emphasis on significant sameness?

In addressing this question, it must be acknowledged that the distinguished individuals mentioned above all had strong mathematical backgrounds. However, the rank and file of members of the Royal Statistical Society of London (of which Quetelet was a founding member in 1834) during the 19th, and well into the 20th, centuries, were not. Most were economists whose charge was the accumulation and sorting of facts generated by population censuses and other government reports. And, of course, the same applied to other fledgling social scientists.

But Fisher's methods seemed to offer something of a "cookbook" approach to statistical/scientific inference, rote and accessible to all. Under Fisher's authoritative sway, researchers came to believe that $p \leq 0.05$ results from null hypothesis significance tests (NHSTs) are the ultimate objective of a study. And it is difficult to argue with them given the force of his recommendations. For example, "It is usual and convenient for experimenters to take 5 per cent as a standard level of significance" (Fisher 1966, p. 13). Or consider his endorsement of the *p*-value's epistemic role in certifying knowledge claims: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this [$p < 0.05$] level of significance" (Fisher 1926, p. 504). And his proclamation that: "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (Fisher 1966, p. 16). [It is true that Fisher later changed his mind regarding the insistence on a fixed 0.05 level, yet the sacredness of this value remains cemented in the research community.] Lastly, he assured that: "Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences" (Fisher 1970, p. 2). In short, Fisher promoted the use of formal statistical inference, particularly significance testing, as an integral component of scientific method. Small wonder that researchers in embryonic disciplines looking for scientific respectability followed suit.

It is easy to appreciate the reasoning behind Fisher's initial impact on data analysis. What is harder to comprehend is that despite withering criticism of its usefulness in the scientific enterprise, NHST (mostly in the form of the incompatible Fisher–Neyman-Pearson hybrid, e.g., Gigerenzer 1993; Goodman 1993; Hubbard and Bayarri 2003) continues to feature prominently in methods texts. Admissible employment of NHSTs requires that the data are a random sample from a clearly defined population, the latter belonging to the Neyman-Pearson component of the hybrid. [Fisher's views on the nature of populations, probability, and random sampling were confusing. See Hubbard (2016, pp. 182–183) for additional discussion.] In a nuts-and-bolts fashion, section 3 assesses the feasibility of meeting these two demands.

3. Practical Difficulties in Implementing Formal Statistical Inference

When presented in statistics and research methods texts, obtaining a random sample from a target population for purposes of statistical inference seems straightforward enough. The following steps from Iacobucci and Churchill's (2010, p. 283) highly

regarded textbook, now in its 11th edition, are typical of the genre: 1. define the target population, 2. identify the sampling frame, 3. select a random sampling procedure, 4. determine the sample size, 5. select the sample elements, and 6. collect the data from the designated elements. In fact, each of these steps can be fraught with difficulties which impede the use of formal methods of statistical inference in scientific work. We comment briefly on these six steps.

3.1. Define the Target Population

In an important but under-appreciated (only 82 Google Scholar citations as of September 5, 2017) article titled "Assumptions for Statistical Inference," Hahn and Meeker (1993, p. 4) acknowledge that this step sometimes is omitted. Indeed they say that textbooks commonly evade the issue by stating "Assume a random sample …from the population of interest" (Hahn and Meeker 1993, p. 1). Or as Gelman (2016) more recently conceded, statistics courses and textbooks (including his own) mostly treat the data as given.

Shaver and Norton's (1980a, b) examination of empirical papers in notable education journals led them to conclude that few gave an explicit definition or description of the target population. Likewise, Bottai (2014, p. 236) admitted that ill-defined populations are common in medical studies. Finally, Gigerenzer and Marewski (2015, p. 422) wrote that social scientists typically "do not draw random samples from a population or define a population in the first place."

3.2. Identify the Sampling Frame

The sampling frame is a list of the population members from which the sample will be drawn. But such lists may not exist. For example, in his book *Practical Sampling*, Henry (1990, p. 85) noted that getting lists of the general U.S. population is difficult, if not impossible. This helps to explain the relative popularity of *cluster* samples, where sampling frames are required only for the randomly selected clusters (subgroups) employed in the study.

But assessing the external validity (generalizability) of an investigation's results demands the sampling of settings, treatments, and observations as well as people. And while locating comprehensive lists for the latter is challenging enough, it is incredibly bothersome for the other three categories (Shadish, Cook, and Campbell 2002, p. 344).

Even if sampling frames are available they will rarely (ever?) exhibit a one-to-one correspondence with the target populations. Some are obsolete, dated, otherwise varying in quality, or, like the phone book, only rough approximations of a population. Since discrepancies between frames and populations may affect a study's conclusions, Hahn and Meeker (1993, p. 5) urge researchers to examine them. This is seldom done in practice.

3.3. Select a Random Sampling Procedure

In the social and other sciences, the generalizability of findings is conveyed almost exclusively in the language of formal methods of statistical inference. That is, matters of external validity center on the notion of *statistical* generalization (Hubbard and Lindsay 2013a) or the so-called *representative* model of generalization (Cook and Campbell 1979).

The centrality of representativeness was highlighted by survey researchers who wished to generalize from sample *to* population with known probability of error (Mook 1983). And before long, "the inference from a sample to population grew to be considered the most crucial part of research" (Gigerenzer and Marewski 2015, p. 425). It is reflected in methods texts which invariably state that samples must be representative of the relevant population. This has led to an insistence on random (representative) sampling, computing sample averages, and conducting tests of statistical significance.

We have already seen that selecting random samples from target populations is difficult to achieve. The importance of this limitation is reinforced in Section 4.

3.4. Determine the Sample Size

In applying formulae to calculate the sample size necessary to meet desired levels of precision and confidence regarding population parameters, textbook presentations on this topic (e.g., Iacobucci and Churchill 2010, pp. 313–317) sometimes invoke questionable assumptions. For example, it is posited that we know in advance the population variance, σ^2. Or if we don't know σ^2 (usually the case), we can estimate it using the sample variance, s^2, by doing either a pilot study, or possibly dividing the range of the data (if we know it) by 6, to get an estimate of the sample standard deviation, s.

Such practices are uncommon. Researchers rarely calculate required sample sizes (see, e.g., Bezeau and Graves 2001), employing instead whatever data are, or can be made, available. This is further borne out, indirectly, by the fact that much of the empirical social science literature is noticeably underpowered (Cohen 1988; and as summarized by Hubbard 2016, pp. 53–56), and that this state of affairs has been prevalent for some 60 years (Hubbard 2016, pp. 229–230).

3.5. Choose the Sample Elements

In order to draw a *simple* random sample we must be able to number the population elements from 1 to N (population size). But it is commonplace that this supposedly innocuous task is beyond our reach. The same stricture holds if we want to adopt *systematic* sampling.

Matters of selection get more complicated still if we wish to employ, say, a *stratified, cluster*, or *area* random sampling design. Here, in addition to being able to list the population members from 1 to N, we need additional information about each member in terms of the variable(s) used to assign them to a particular stratum, cluster, or area. Yet these kinds of real world details typically are ignored, or downplayed, in the textbooks.

3.6. Collect Data from Designated Elements

Again, an ostensibly simple-sounding task can be forbidding. The impact of unlisted numbers, not-at-homes, and refusals to cooperate, for example, contribute to the major problem of *nonresponse bias*. People are inundated with requests for

information and are increasingly unlikely to comply (Blair and Zinkhan 2006).

Practical considerations alone make application of formal statistical inference problematic. Yet even if this were not the case, the more salient issue is that this methodology offers little assistance in fostering scientific progress.

4. Statistical Inference is Ill-Suited for Promoting Scientific Inference

Substantiating the above claim necessitates a distinction between two different kinds of studies, enumerative and analytic. The representative, or statistical inference, model plays a decisive role in what Deming (1975, p. 147) labels *enumerative* studies. Here the goal is to estimate some fixed population parameter based on choosing a random sample from its constituents. As an example: "What is the Grade Point Average of high school juniors in Baltimore, MD?" Canvassing this class ranking supplies the "right" answer.

But Deming's (1975, p. 147) *analytic* studies are concerned with outcomes in the future, so the aim is to offer predictions about some process or cause system. As an example: "Will increasing the length of the school day by one hour raise SAT scores among U.S. high school seniors?" In such cases, representativeness often is impractical and possibly harmful to knowledge development.

To appreciate this, consider Hahn and Meeker's (1993) expansion of Deming's (1975) views. They underline that use of formal statistical inference mandates adherence to two basic requirements. First, the population must be well defined, finite, and unchanging. Second, a random sampling procedure is imperative in the selection of elements. In methods texts this sounds as uncomplicated as following the steps in a basic recipe. In practice, Hahn and Meeker (1993, p. 4) relate that for analytic (and even enumerative) studies, things are not nearly so simple because:

> We define an analytic study to be a study in which one is *not* dealing with a finite, identifiable, unchanging collection of units, and, thus is concerned with a process, rather than a population. … [I]n our experience, the great majority of applications encountered in practice, especially in industrial and in medical and other scientific applications, involve analytic, rather than enumerative studies. Moreover, it is inherently more complex to draw conclusions from analytic than from enumerative studies; analytic studies require the critical (and often unverifiable) added assumption that the process about which one wishes to make inferences is statistically identical to that from which the sample was selected.

In analytic studies random sampling seldom is an attainable option since "one is no longer dealing with 'an aggregate of identifiable units,' [hence] there is no relevant frame from which one can take a random sample" (Hahn and Meeker 1993, p. 5).

Other statisticians, such as Berk and Freedman (2003, p. 2), Chatfield (2002, p. 4), Gunter and Tong (2016–2017, p. 2), and Kass (2011, p. 8) share Hahn and Meeker's (1993, p. 10) position that authentic random samples are the exception rather than the rule. So do methodologists like Gigerenzer (2004, p. 599) and Grice et al. (2017, p. 236), who remark that experimental subjects are hardly ever sampled randomly from adequately defined populations. Or consider Shadish, Cook, and Campbell (2002, p. 91), who claimed that "we have not put much emphasis on random sampling for external validity, primarily because it is so rarely feasible in experiments." Despite this restriction, virtually all data in practice are analyzed as if random samples drawn from clearly specified populations are the norm.

In addition, there is the largely unrecognized admission that the formal statistical inference model is *misleading* in the context of analytic/predictive studies. This is because appraisals of external validity must involve characteristics included in the study along with others which are not (Hubbard and Lindsay 2013a; Shadish, Cook, and Campbell 2002). As Leviton (2001) noted, to be confident in making broad generalizations necessitates sampling from a "super-population" composed of every circumstance imaginable which may impact the result. It is obvious that no single study has this capability and so offers scant evidence of the external validity of an outcome (Leviton 2001, p. 5197).

And paradoxically, seeking representativeness can actually reduce the precision of results in regards to their application or generalizability to particular situations. That is, when making a prediction we must anticipate the different factors (subpopulations) that may affect a result and deal with them directly. This idea is developed further in the next section which addresses randomized controlled trials.

5. Randomized Controlled Trials (RCTs)

5.1. Advantages of RCTs

A major area where the superiority of formal methods of statistical inference are taken mostly for granted is the medical field's adoption of randomized controlled trials (RCTs). Popularized by Fisher (1935), use of randomization is often viewed as a prerequisite for drawing *causal* inferences in scientific experiments. As explained in most methods texts, the random assignment of subjects to the treatment (receiving, say, the new drug) and control (placebo) groups helps ensure their *equivalence* before the experiment commences. After some appropriate passage of time, differences in the mean incidence of illness in the two groups is subjected to a test of statistical significance. If a $p < 0.05$ result is found in favor of the treatment group, the new drug is regarded as "successful." And because the two groups were deemed equivalent before the trial began, improved results for the treatment group must have been caused by the new drug. Otherwise expressed, RCTs are thought to exhibit high levels of *internal validity*.

Indeed, RCTs are touted as the gold standard of experimental research (Jadad and Enkin 2008; Meldrum 2000). A principal reason for this acclaim is because they employ a *deductive* methodology—a causal inference may be implied if the assumptions of the method are met (Cartwright 2007, p. 11). Echoing Cartwright, the gold standard of evidence in the pursuit of evidence-based medicine is believed to come from the execution of large (phase III) RCTs (Williams 2010, p. 107).

But do RCTs constitute the gold standard for causal inference? Recent work disputes this moniker, including the articles by Cartwright (2007) and Williams (2010). In the first place, as discussed throughout this article, there is much information to indicate that meeting the assumptions required to employ formal methods of statistical inference is difficult. As one example

alone, RCTs rarely involve a random sample of patients targeted for the treatment (Williams 2010, p. 109). Second, even if these assumptions could be met, it does not eliminate the potentially serious drawbacks associated with the conduct of RCTs, as explored below.

5.2. Limitations of RCTs

The highest standard of evidence sought in evidence-based medicine is judged to come from large (phase III) RCTs. This is because, in the face of *heterogeneous* patient populations, large samples are necessary to obtain $p < 0.05$ results in comparisons of treatment and control group averages. Pharmaceutical companies view such findings as obligatory in order to win Food and Drug Administration (FDA) approval for the new medicine (Williams 2010, p. 109). The problem is that physicians do not treat the *average* patient, they treat real individuals (Demidenko 2016, p. 34; Kent and Hayward 2007, p. 68; Ruberg, Chen, and Wang 2010, p. 574). As such, averages may have only weak predictive validity for the latter (Williams 2010, p. 106).

Successful ($p < 0.05$) RCT outcomes suggest that patients react in like manner to the effects of the new drug. But this is not the case. It is just as, or more than, likely that reactions depend on the attributes of the patients themselves, implying that only some subset(s) of them may benefit (Williams 2010, p. 108). Here, Williams is reiterating Cartwright's (2007, p. 15) point that, when employing Suppes's (1970) probabilistic theory of causality, the only logical deduction we are entitled to make from a successful RCT is that the treatment causes the outcome in at least *some* members of that population. Disturbingly, Bottai (2014, p. 235), Cartwright (2007, p. 16), and Kent and Hayward (2007, p. 62) warned, this same treatment could have no beneficial outcomes for some subpopulations, and might even be harmful to still others. So much for the value of $p < 0.05$ results.

On the other hand, RCTs with $p > 0.05$ findings, while failing to pass muster with the FDA, can nevertheless be of great help to many patients (Kent and Hayward 2007, p. 62). For instance, the FDA initially denied approval for the drug temozolomide as a treatment for glioblastoma brain tumors because of overall $p > 0.05$ results. This meant delayed access to a new medicine which subsequently became the treatment of choice for patients with brain cancer (Williams 2010, p. 113).

As shown above, because of the focus on the average patient from a heterogeneous population, RCTs likely include patients who may benefit, not benefit, or possibly be harmed by the new drug. This suggests that trial investigators examine their data to see whether these subgroups exist (Kent and Hayward 2007, p. 62; Ruberg, Chen, and Wang 2010, p. 574; Williams 2010, p. 108). Yet this rarely happens. For example, in a review of 108 trials, Kent and Hayward (2007, p. 67) found that only one followed this approach.

The discussion in this section supports the belief that the high levels of internal validity said to be characteristic of large (phase III) RCTs is exaggerated. Moreover, claims regarding the external validity of such RCT results are difficult to justify (Cartwright 2007, p. 19), even though they are routinely overgeneralized (Williams 2010, p. 118). In short, RCT findings can be deficient in terms of *both* their internal and external validities. Or as Cartwright (2007, p. 18) summed it up: "The RCT, with its vaunted rigor, takes us only a very small part of the way we need to go for practical knowledge."

6. Scientific Inference

Unlike its statistical inference counterpart, the concept of scientific inference defies reduction to a series of allegedly neat and tidy methodological steps whose dutiful observance renders the output "science." Believing otherwise is wishful thinking. Scientific inference transcends *by far* the purview of statistical inference. This is why it is inherently more difficult to describe, let alone codify.

Scientific inferences in any given area are not made in a vacuum. Rather, they are offered in light of the totality of empirical and theoretical subject-matter knowledge that has accumulated over a long period of time. In progressive sciences this background knowledge, whose cogency rests on conditional acceptance by members of the field, plays a defining role in evaluating the plausibility of new findings. So scientific inferences are made within a dynamic context of what we believe we know, and hope to know, about our world and beyond. Importantly, these scientific conjectures and judgments do not often involve applications of formal statistical inference. They arise instead from researchers employing rules-of-thumb in their daily, and mostly exploratory (Gunter and Tong 2016–2017), work.

While no single approach to knowledge acquisition is likely to be optimal, a particularly fruitful candidate would be one centered on the search for *significant sameness* among phenomena. This idea is outlined in what follows.

6.1. Looking for Significant Sameness (Replicability)

The academy's historical, and ongoing, fixation with revealing "significant differences" in studies actively impedes cumulative science. In direct opposition, pursuit of "significant sameness" nurtures the latter.

Advocates of significant sameness tend to follow a *realist* philosophy of science (Haig 2014; Haig and Evers 2016; Hubbard 2016). Realism is based on how researchers actually behave, and is concerned with the lengthy process of theory *development*.

Under the auspices of realism, the significant sameness model postulates a facts-before-theory approach. In a nutshell, the argument is as follows: Data from individual studies rarely speak for themselves—they are ephemeral, affected by idiosyncratic boundary conditions, and usually irreproducible. As such, we are unlikely to build sound theories around individual data sets, especially those replete with $p < 0.05$ results. Better to construct theories *after* the discovery of what are variously described as repeatable facts, empirical regularities, or phenomena: their relative stubbornness deserves an explanation. Indeed, it should be emphasized that more Nobel prizes have been awarded for such discoveries than for advancing theories (Haig 2005, p. 384).

Because of widespread unfamiliarity with the issue, it is important to clarify further the distinction between data and phenomena/stubborn facts (Woodward 1989). Theories are mostly concerned with the explanation of phenomena, not data (Haig 2014, p. 34). But phenomena, typically, are not observable, and they are more durable than data. It is the role of data to provide evidence with respect to phenomena (Haig 2014, p. 35).

Accordingly, the progression is data → phenomena → theory, and it underpins the significant sameness model.

The quest for significant sameness endorses the pivotal role of replication research in cultivating healthy science. This is because successful replications are the primary means available for identifying phenomena. So we must ask: How do we gauge whether a replication is successful? We can do so by employing the criterion of overlapping confidence intervals (or Bayesian equivalents) around point values of interest, but always in a heuristic fashion and not as some cut-and-dried rule. In this endeavor, statistical analysis plays a vital role.

Additionally, replication research is crucial in placing bounds on the application of stubborn facts, that is, exploring their empirical generalizability. Moreover, empirical generalization focuses on projecting quantitative results *across* multiple, relatively homogeneous, (sub)populations. Of note, failure to generalize results can further enrich theory building by explaining why this limitation occurred. Counter to popular misconceptions, replication research in no way legislates "brute empiricism." Rather, its practice is responsible for eliciting facts, determining their empirical range, and assisting in the initial and subsequent refinement of causal explanations for them. This requires the conduct of many studies over a long period of time. Science in action is a process, not a product.

Scientific inference is better viewed as being grounded in *abductive* (explanatory) reasoning. Abduction—sometimes termed inference to the best explanation (Harman 1965)—takes as its locus the studying of facts and proposing a theory about the causal mechanisms generating them. Thus, abduction is a method of scientific inference geared toward the development of feasible and best explanations for the stubborn facts we possess. Like detective work, this approach mirrors the behavior of practicing scientists. And it is not beholden to methods of formal statistical inference.

Of course, the significant sameness model is not without its own limitations. For example, it might be objected that this model could be potentially expensive to apply in practice. There is merit to this argument as things stand today. If, however, the editorial-reviewer biases favoring the publication of "novel," $p < 0.05$, results, and against the publication of replication research—which biases institutionalize bad science—could be successfully confronted, then substantial resources would be freed to promote significant sameness.

As well, there are shortcomings associated with replication studies which could cause some editors, reviewers, and authors to question their usefulness (Hubbard 2016, pp. 156–157). For instance, let's say that Smith attempts to replicate Jones's initial work and gets different results. This situation could degenerate into the pair blaming one another for the conflicting outcomes. Jones could accuse Smith of failing to conduct an accurate replication—a charge understandable given that limited journal space often precludes a full disclosure of the methods used. Smith could respond that Jones made mistakes. And on and on. Yet while all of this is possible, there is no escaping the fact that "Replicability is almost universally accepted as the most important criterion of genuine scientific knowledge" (Rosenthal and Rosnow 1984, p. 9).

6.2. The Indispensable Role of Observational Studies

It is necessary at this juncture to revisit sampling issues. Since the 1960s and 1970s, when various user-friendly software packages became widely available, use of formal methods of statistical inference have monopolized empirical work in large areas of academe. At this same time, the impression is easily gained that nonrandomized (observational) studies came to be seen as lacking the exactitude of their randomized counterparts, and therefore constituted inferior research. This, despite the fact that virtually all scientific advances over the centuries have accrued—and continue to do so—without the "benefit" of randomization.

Scientific knowledge is gleaned chiefly from nonrandom samples. For instance, the process of discovering the scope and limits of empirical generalizations, mentioned earlier, relies on the use of purposive or judgment samples (Hubbard and Lindsay 2013a, p. 1384), where researcher choices are central (Rosenbaum 2001, p. 223). Yet there is no reason for concern. Nonrandom samples, allied with a replication strategy, can yield robust findings. Which is to say that, over time, via replication research, the point estimates and confidence intervals of additional (new) results about the phenomena at hand may be compared with the increasingly sturdy benchmarks established by their numerous predecessors to check for consistency. Consistency gained in this manner leads to deserved support for the veracity of the findings.

Beyond this, it should not be forgotten that entire disciplines such as archaeology, dendrochronology, and paleontology have been forged on the basis of *convenience* or *accidental*, never mind purposive, samples (Hubbard 2016, p. 180). And striking breakthroughs in other fields have arisen from the employment of such samples. For example, Freedman (1999) tells of how the use of convenience samples was key to establishing the causal links between contaminated water and the spread of cholera, and cigarette smoking and lung cancer. Contrary to the entrenched beliefs of many in the scientific community, "most of what we know about causation …is derived from observational studies" (Freedman 1999, p. 255).

In the context of RCTs, Williams (2010) concurs. A major criticism of the generalizability of results from a phase II trial (which does not feature randomization or a control group) is that, owing to possible selection bias, they may be due to characteristics in the patient sample, rather than the treatment. He adds that because phase III samples may not be representative of the universe of patients targeted for the treatment, coupled with the vagaries in individual patient attributes, the claimed advantages of phase III over phase II trials mostly evaporate. Along with Bottai (2014, p. 236), Williams (2010, p. 109) concludes that neither phase II nor phase III trials eliminate the worry about results not generalizing beyond the patients involved in that specific trial. Consequently, he favors the use of small trials, because when it comes to predicting the efficacy of a treatment on a particular patient, phase III RCTs are often "only marginally better than simply guessing" (Williams 2010, p. 111).

Consistent with the description of scientific inference sketched above, Shadish, Cook, and Campbell (2002, pp. 353–354) list five principles of generalized causal inferences that scientists employ every day in their work with mostly purposive samples. First is the idea of *surface similarity*, as when scientists

generalize by evaluating the likenesses among phenomena. Second is the *ruling out of irrelevancies*, meaning scientists generalize by deciding on those aspects of persons, settings, treatments, and results which are immaterial because they do not alter a generalization. Third is *making discriminations*, that is, finding limits to a generalization. Fourth is *interpolation* and *extrapolation*, where scientists generalize within the range of sampled persons, settings, treatments, and outcomes, together with those beyond them. Fifth is *causal explanation*, in which scientists generalize by developing and testing explanatory theories. It must be underscored that none of these five principles espouse the virtues of formal statistical inference. But this is not problematic. As Shadish, Cook, and Campbell (2002, p. 356) note, even though "these purposive sampling methods are not backed by a statistical logic that justifies formal generalizations …they are more practical than formal probability sampling …and are by far the most widely used sampling methods for generalized causal inference problems."

7. Conclusions

It is quite extraordinary that a model of inference that resists prescribed implementation, and anyway plays a quite limited part in furthering scientific progress, nevertheless has dominated the pages of journals in the social, management, and biomedical sciences for decades. Clearly, some serious re-thinking about statistics' role in scientific advance is needed. Fortunately, help looks to be on the way.

Recently the American Statistical Association (ASA) drew attention to problems associated with statistical significance and *p*-values (Wasserstein 2016; Wasserstein and Lazar 2016) in the context of science's "reproducibility crisis" (Peng 2015). This is a welcome and bold initiative offering as it does an unprecedented opportunity to publicly debate the value of statistical methods in the creation of robust scientific knowledge. It was followed by the ASA's Symposium on Statistical Inference in October 2017, which in turn led to this TAS Special Issue, "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$," an online, permanently open access issue of *The American Statistician*.

It seems certain that these initiatives will require major educational reforms in statistics classes and textbooks (Gelman 2016). Berry (2016, p. 1, p. 4) maintains that since the collective credibility of statisticians in the science community presently is jeopardized, such reforms are vital. They must reflect the circumscribed role of formal statistical inference in scientific inference.

References

Berk, R. A., and Freedman, D. (2003), "Statistical Assumptions as Empirical Commitments," in *Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger*, (2nd ed.), eds. T. G. Blomberg and S. Cohen, Hawthorne, NY: Aldine de Gruyter, pp. 235–254. [94]

Berry, D. A. (2016), "*P*-values Are Not What They're Cracked Up to Be," online supplemental comment to Wasserstein (2016). [97]

Bezeau, S., and Graves, R. (2001), "Statistical Power and Effect Sizes of Clinical Neuropsychology Research," *Journal of Clinical and Experimental Neuropsychology*, 23, 399–406. [93]

Blair, E., and Zinkhan, G. M. (2006), "Nonresponse and Generalizability in Academic Research," *Journal of the Academy of Marketing Science*, 34, 4–7. [94]

Bottai, M. (2014), "Inferences and Conjectures About Average and Conditional Treatment Effects in Randomized Trials and Observational Studies," *Journal of Internal Medicine*, 276, 229–237. [93,95,96]

Cartwright, N. (2007), "Are RCTs the Gold Standard?" *BioSocieties*, 2, 11–20. [91,94,95]

Chatfield, C. (2002), "Confessions of a Pragmatic Statistician," *The Statistician*, 51, 1–20. [95]

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum. [93]

Cook, T. D., and Campbell, D. T. (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston, MA: Houghton Mifflin. [93]

Demidenko, E. (2016), "The *p*-Value You Can't Buy," *The American Statistician*, 70, 33–38. [95]

Deming, W. E. (1975), "On Probability As a Basis For Action," *The American Statistician*, 29, 146–152. [91,94]

Ehrenberg, A. S. C. (1975), *Data Reduction* (Rev. reprint), London, UK: Charles Griffin. [91]

Fisher, R. A. (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513. [92]

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd. [94]

——— (1966), *The Design of Experiments* (8th ed.), Edinburgh: Oliver and Boyd. [92]

——— (1970), *Statistical Methods for Research Workers* (14th ed.), New York: Hafner. [92]

Freedman, D. (1999), "From Association to Causation: Some Remarks on the History of Statistics," *Statistical Science*, 14, 243–258. [96]

Gelman, A. (2016), "The Problems with *P*-Values are Not Just With *P*-Values," online supplemental comment to Wasserstein (2016). [93,97]

Gigerenzer, G. (1993), "The Superego, the Ego, and the Id in Statistical Reasoning," in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, eds. G. Keren and C. A. Lewis, Hillsdale, NJ: Lawrence Erlbaum, pp. 311–339. [92]

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socio-Economics*, 33, 587–606. [92,94]

Gigerenzer, G., and Marewski, J. N. (2015), "Surrogate Science: The Idol of a Universal Method for Scientific Inference," *Journal of Management*, 41, 421–440. [93]

Goodman, S. N. (1993), "*p* Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate," *American Journal of Epidemiology*, 137, 485–496. [92]

Grice, J., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., Medellin, E., and Vest, A. (2017), "Four Bad Habits of Modern Psychologists," *Behavioral Sciences*, 7, 1–21. [94]

Gunter, B., and Tong, C. (2016–2017), "A Response to the ASA Statement on Statistical Significance and *P*-values," *NV–ASA Newsletter*, 14, 1–3. [91,94,95]

Guttman, L. (1985), "The Illogic of Statistical Inference for Cumulative Science," *Applied Stochastic Models and Data Analysis*, 1, 3–10. [92]

Hahn, G. J., and Meeker, W. Q. (1993), "Assumptions for Statistical Inference," *The American Statistician*, 47, 1–11. [91,93,94]

Haig, B. D. (2005), "An Abductive Theory of Scientific Method," *Psychological Methods*, 10, 371–388. [95]

——— (2014), *Investigating the Psychological World: Scientific Method in the Behavioral Sciences*, Cambridge, MA: MIT Press. [95]

Haig, B. D., and Evers, C. W. (2016), *Realist Inquiry in Social Science*, Thousand Oaks, CA: Sage. [95]

Harman, G. H. (1965), "Inference to the Best Explanation," *Philosophical Review*, 74, 88–95. [96]

Henry, G. T. (1990), *Practical Sampling*, Newbury Park, CA: Sage. [93]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Thousand Oaks, CA: Sage. [91,92,93,95,96]

Hubbard, R., and Bayarri, M. J. (2003), "Confusion Over Measures of Evidence (*p*'s) Versus Errors (α's) in Classical Statistical Testing," (with comments), *The American Statistician*, 57, 171–182. [92]

Hubbard, R., and Lindsay, R. M. (2013a), "From Significant Difference to Significant Sameness: Proposing a Paradigm Shift in Business Research," *Journal of Business Research*, 66, 1377–1388. [91,93,94,96]

Hubbard, R., and Lindsay, R. M. (2013b), "The Significant Difference Paradigm Promotes Bad Science," *Journal of Business Research*, 66, 1393–1397. [91]

Iacobucci, D., and Churchill, G. A., Jr. (2010), *Marketing Research: Methodological Foundations* (10th ed.), Mason, OH: South-Western. [92,93]

Jadad, A. R., and Enkin, M. W. (2008), *Randomized Controlled Trials: Questions, Answers and Musings* (2nd ed.), Malden, MA: Blackwell. [94]

Kass, R. E. (2011), "Statistical Inference: The Big Picture," *Statistical Science*, 26, 1–9. [94]

Kent, D., and Hayward, R. (2007), "When Averages Hide Individual Differences in Clinical Trials," *American Scientist*, 95, 60–68. [95]

Leviton, L. C. (2001), "External Validity," in *International Encyclopedia of the Social and Behaviour Sciences*, eds. N. J. Smelser and P. B. Baltes, Oxford, UK: Elsevier, pp. 5195–5200. [94]

Meldrum, M. L. (2000), "A Brief History of the Randomized Controlled Trial: From Oranges and Lemons to the Gold Standard," *Hematology/Oncology Clinics of North America*, 14, 745–760. [94]

Mook, D. G. (1983), "In Defense of External Invalidity," *American Psychologist*, 38, 379–387. [93]

Nelder, J. A. (1986), "Statistics, Science and Technology," (with comments), *Journal of the Royal Statistical Society*, Series A, 149, 109–121. [91]

Peng, R. (2015), "The Reproducibility Crisis in Science: A Statistical Counterattack," *Significance*, 12, 30–32. [97]

Rosenbaum, P. R. (2001), "Replicating Effects and Biases," *The American Statistician*, 55, 223–227. [96]

Rosenthal, R., and Rosnow, R. L. (1984), *Essentials of Behavioral Research: Methods and Data Analysis*, New York: McGraw–Hill. [96]

Ruberg, S. J., Chen, L., and Wang, Y. (2010), "The Mean Does Not Mean as Much Anymore: Finding Sub-Groups for Tailored Therapeutics," *Clinical Trials*, 7, 574–583. [95]

Shadish, W. R., and Cook, T. D. (1999), "Comment—Design Rules: More Steps Toward a Complete Theory of Quasi-Experimentation," *Statistical Science*, 14, 294–300. [91]

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston, MA: Houghton Mifflin. [93,94,96]

Shaver, J. P., and Norton, R. S. (1980a), "Populations, Samples, Randomness, and Replication in Two Social Studies Journals," *Theory and Research in Social Education*, 8, 1–10. [93]

Shaver, J. P., and Norton, R. S. (1980b), "Randomness and Replication in Ten Years of the *American Educational Research Journal*," *Educational Researcher*, 9, 9–15. [93]

Suppes, P. (1970), *Probabilistic Theory of Causality*, Atlantic Highlands, NJ: Humanities Press. [95]

Wasserstein, R. L. (2016), "ASA Statement on Statistical Significance and *P*-Values," *The American Statistician*, 70, 131–133. [97]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–131. [97]

Williams, B. A. (2010), "Perils of Evidence-Based Medicine," *Perspectives in Biology and Medicine*, 53, 106–120. [94,95,96]

Woodward, J. (1989), "Data and Phenomena," *Synthese*, 79, 393–472. [95]