The American Statistician

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

Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science

Christopher Tong

To cite this article: Christopher Tong (2019) Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science, The American Statistician, 73:sup1, 246-261, DOI: 10.1080/00031305.2018.1518264

To link to this article: https://doi.org/10.1080/00031305.2018.1518264

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 246–261: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1518264

Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science

Christopher Tong

United States Department of Agriculture, Center for Veterinary Biologics, Ames, IA

ABSTRACT

Scientific research of all kinds should be guided by statistical thinking: in the design and conduct of the study, in the disciplined exploration and enlightened display of the data, and to avoid statistical pitfalls in the interpretation of the results. However, formal, probability-based statistical inference should play no role in most scientific research, which is inherently exploratory, requiring flexible methods of analysis that inherently risk overfitting. The nature of exploratory work is that data are used to help guide model choice, and under these circumstances, uncertainty cannot be precisely quantified, because of the inevitable model selection bias that results. To be valid, statistical inference should be restricted to situations where the study design and analysis plan are specified prior to data collection. Exploratory data analysis provides the flexibility needed for most other situations, including statistical methods that are regularized, robust, or nonparametric. Of course, no individual statistical analysis should be considered sufficient to establish scientific validity: research requires many sets of data along many lines of evidence, with a watchfulness for systematic error. Replicating and predicting findings in new data and new settings is a stronger way of validating claims than blessing results from an isolated study with statistical inferences.

ARTICLE HISTORY
Received February 2018
Revised June 2018

KEYWORDS
Many sets of data; Model uncertainty; Optimism principle; Researcher degrees of freedom; Shoe leather; Triangulation

1. Introduction

Statistical inferences are claims made using probability models of data generating processes, intended to characterize unknown features of the population(s) or process(es) from which data are thought to be sampled. Examples include estimates of parameters such as the population mean (often attended by confidence intervals), hypothesis test results (such as p-values), and posterior probabilities. Such methods are often intended to quantify (and thus to tame) uncertainty, and to evaluate the plausibility of chance as an explanation of a data pattern. The widespread use of statistical inference methods in scientific research has recently been scrutinized and questioned, for reasons outlined in the "ASA Statement on Statistical Significance and P-values" (Wasserstein and Lazar 2016). The *ASA Symposium on Statistical Inference* (Bethesda, Maryland, USA; October 11–13, 2017) was expected to "lead to a major rethinking of statistical inference, aiming to initiate a process that ultimately moves statistical science—and science itself—into a new age," according to the call for articles for this special issue of *The American Statistician*. This article is intended to offer one such rethinking. The core of our perspective is indeed the essential link between statistical science and science itself.

Much of the recent discussion of statistical inference focuses on null hypothesis testing, p-values, and even the very notion of statistical significance. Unease with these commonly used procedures and concepts has a lengthy history, eloquently discussed by others. We do not add to this dimension of the conversation; it suffices to cite the ASA Statement itself (Wasserstein and Lazar 2016) and two recent commentaries (Argamon 2017; McShane et al. 2017). Among these criticisms, McShane and Gelman (2017) succinctly stated that null hypothesis testing "was supposed to protect researchers from over-interpreting noisy data. Now it has the opposite effect." The *ASA Symposium on Statistical Inference* was intended to move the conversation to what should be done instead.

However, as a prelude to a discussion of remedies, we will argue here, as Andrew Gelman (2016) did, that nearly *all* forms of statistical inference share serious deficiencies, except in cases when the study protocol and statistical model are fully prespecified. We shall argue that these issues stem largely from the *Optimism Principle* (Picard and Cook 1984) that is an inevitable byproduct of the necessarily flexible data analysis and modeling work that attends most scientific research. Moreover, we contend that the well-established distinction between exploratory and confirmatory objectives provides a framework for understanding the proper roles of flexible versus prespecified statistical analyses. Unfortunately, we think that in much of the current use of inferential methods in science, except in specialized fields such as human clinical trials, this distinction is absent. This absence has enabled the widespread dissemination of biased statistical inferences and encouraged a *Cult of the Isolated Study* (Nelder 1986) that short-circuits the iterative nature of research. Statistical inference should not be used to avoid wearing "shoe leather" (Freedman 1991), a metaphor for the hard work of gathering more and better data, discovering and dealing with systematic sources of error, and building a scientific argument along many lines of evidence. The statistical contribution to science must focus on data production, data description and

CONTACT Christopher Tong ctong@mailaps.org United States Department of Agriculture, Center for Veterinary Biologics, P. O. Box 844, Ames, IA 50010.

exploration, and statistical thinking, rather than statistical inference. None of the concepts we discuss are new, but collectively this perspective argues for abandoning much of conventional statistical practice and teaching. The perspective outlined here has little in common with many of the proposals discussed at the *ASA Symposium*, although a few have specific points of commonality.

2. Statistical Inference and the Optimism Principle

A stereotypical framework for statistical inference in a simple univariate setting is as follows (Fisher 1922). We are presented with a dataset, and we have reason to believe (or are prepared to assume) that the data were generated as an independent, identically distributed (IID) sample from an underlying population. This population may be characterized by a probability model, and our goal is to make inferences about parameters characterizing this unknown model. Such inferences will be attended by uncertainty, since we do not have access to the population, but only IID samples drawn from it. Examples of statistical inferences include estimates of parameters or other properties of the population, confidence intervals (confidence sets, credible sets) for such estimates, and hypothesis tests about such parameters or properties. More generally, the probability model may include both fixed and random components, and the latter may have variance-covariance structures more complex than the IID assumption. There are several contending schools of thought about how such inferences should be made, of which the frequentist and Bayesian are the most widely known (see *Appendix* A.1). What all these ideologies have in common is that *they express uncertainty through a probability claim*, such as a confidence level, Type 1 error, *p*-value, posterior probability, likelihood function, and so on. Our focus on statistical inference in this article will mainly be on these probabilistic expressions of uncertainty. This is distinct from both the common language and general scientific uses of the term "inference."

On the first page of the first chapter of their recent book, *Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science*, Efron and Hastie (2016, p. 3) presented the sample mean and its standard error. They then stated that

> It is a surprising, and crucial, aspect of statistical theory that the same data that supplies an estimate can also assess its accuracy.

This principle is indeed typical of statistical teaching and practice. However, when it comes to *prediction*, Efron and Hastie (2016, p. 227) instead advocated a procedure common in machine learning, namely, *data-splitting*. The available data are randomly split into a training set and a test set. The test set is hidden in a vault, and the training set is used to train/fit the predictive algorithm or model. Once that algorithm/model is developed, the test set is unveiled and run through the algorithm to obtain an "honest" assessment of its performance on data it has never seen. (We return to data splitting and more sophisticated approaches in Section 5.) All model building activities, such as variable selection, must be confined to the training set (e.g., Ambroise and McLachlan 2002; Reunanen 2003). The use of data-splitting helps to diagnose *overfitting*, the tendency of a predictive algorithm to optimize its performance on the training data, at the expense of its ability to perform well on data *outside* the training set. Generalizability, to data from the future or to subjects not sampled, for instance, is usually of greater ultimate interest to the user than training data performance.

Statistical inferences also can suffer from this overfitting problem when "the same data that supplies an estimate" are used "to assess its accuracy." In fact, Efron and Hastie (2016) alluded to this in chap. 20 of their book, titled "Inference after model selection." There they wrote that in the past, such inferences were "typically done ignoring the model selection process" but that "Electronic computation has opened the door to a more honest analysis of estimation accuracy, one that takes account of the variability induced by data-based model selection" (ibid, p. 394). They conceded that while "Currently, there is no overarching theory for inference after model-selection" they instead provide "promising analyses of individual situations" (ibid, pp. 395).

Let us unpack these ideas further. Harrell (2015, p. ix) observed that

> Using the data to guide the data analysis is almost as dangerous as not doing so.

This seems like an oxymoron, but there is wisdom here. Statistical model building is usually a multi-step, interactive process because the model is not completely prespecified prior to data collection. Consequently, the model building/model criticism/model selection process may include the following steps that depend on the data at hand, are often partly subjective, and are ideally informed by knowledge of earlier studies and/or existing scientific theory.

- Screening the data for unusual, extreme, incomplete, or inconsistent data records, for possible adjustment or removal;
- Consideration of rescaling or transforming some variables (e.g., logarithm? standardize?) or even a change of coordinates (e.g., principle components analysis);
- Variable selection;
- Decisions to keep or remove interaction terms, or higher-order polynomials;
- Graphs and tables of data to identify patterns, expected or otherwise, for possible inferential reporting;
- Graphical examination of residuals from fitted models, to assess goodness of fit, and applying remedial measures if needed; and
- Methodology selection for fitting the model (e.g., should a robust regression be used instead of least squares?).

Each of these activities offers a chance to improve the fit of the model by repeated comparison with the data—and therefore a chance to *overfit* the data. Simmons, Nelson, and Simonsohn (2011) called these opportunities *researcher degrees of freedom*, and when abused to fish for publishable *p*-values, *p-hacking*. Wicherts et al. (2016) cataloged 34 kinds of researcher degrees of freedom spanning design, conduct, analysis, and reporting of psychological studies; they state that their list is "in no way exhaustive."

The resulting inferences from the final model tend to be biased, with uncertainties underestimated, and statistical sig-

nificance overestimated, a phenomenon dubbed the *Optimism Principle* by Picard and Cook (1984). They quoted Mosteller and Tukey (1977, p. 37):

> Testing the procedure on the data that gave it birth is almost certain to overestimate performance, for the optimizing process that chose it from among many possible procedures will have made the greatest use possible of any and all idiosyncrasies of those particular data . …As a result, the procedure will likely work better for these data than for almost any other data that will arise in practice.

Nearly seven decades ago Koopmans (1949), who was later a 1975 Nobel laureate in economics, alluded to this problem, and it has been studied extensively by statisticians and econometricians since the early 1980s. A classic result was given by Freedman (1983), who showed that a commonly taught and practiced procedure of screening available predictor variables for inclusion in a regression model, using a statistical test, can result in the inclusion of bogus variables with high statistical significance. He showed this in the extreme case where all the candidate predictor variables consist of Gaussian noise, unrelated to the response variable (which is also Gaussian noise). In other words, it is possible to obtain a seemingly informative linear model, with decent R^2 and several statistically significant predictor variables, from data that is utter nonsense. This finding was later dubbed "Freedman's paradox" (Raftery, Madigan, and Hoeting 1993).

Chatfield (1995) used the term *model selection bias* to describe the distorted inferences that result when using the same data that determines the form of the final model to also produce inferences from that model. Parameter estimates themselves may be biased (e.g., Hjorth 1989; Berk, Brown, and Zhao 2010); their uncertainties (standard errors and confidence intervals) underestimated; and prediction intervals will be too narrow. Thus, the highly precise claims of statistical inference tend to be misleading if taken at face value; indeed, Gelman (2016) used the term "uncertainty laundering" to describe this behavior.

Closely related is the term *model uncertainty*, which reflects the fact that a statistical model can rarely be prespecified (as known) without alteration in light of the data, with the important exception of statistical models used in Phase III clinical trials (discussed further in Section 3). "It is indeed strange that we often admit uncertainty by searching for a best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true" Chatfield (1995) says. Moreover, this source of uncertainty, which is invisible to conventional statistical inference, can be the largest component of uncertainty. We refer readers who wish to pursue the sizable literature on model uncertainty and the Optimism Principle to two influential review/discussion articles, Draper (1995) and Chatfield (1995), and the more recent discussions by Berk, Brown, and Zhao (2010), Gelman and Loken (2014), and Holmes (2018).

While these problems are often attributed to misuses of statistical methodology, Gelman and Loken (2014) noted that even *without* a conscious effort to abuse statistical inference (e.g., testing multiple hypotheses but only reporting the significant ones), it is still possible for a null effect to appear highly significant:

> Given a particular data set, it can seem entirely appropriate to look at the data and construct reasonable rules for data exclusion, coding, and analysis that can lead to statistical significance. In such a case, researchers need to perform only one test, but that test is conditional on the data …with the same effect as if they had deliberately fished for those results.

The choice of what test to carry out, even if only one such test is made, is data-dependent. Had a different dataset obtained, a different choice could have been made. "This error carries particular risks in the context of small effect sizes, small sample sizes, large measurement error, and high variation," they wrote. However, such issues also exist in "big data" problems, such as functional imaging in neuroscience, where damaging reuse of the data is known as "double dipping" (Kriegeskorte et al. 2009). Gelman and Loken (2014) used the term "the garden of forking paths" to describe the many potential data-dependent choices a researcher can make during data analysis, resulting in "a sort of invisible multiplicity: data-dependent analysis choices that did not appear to be degrees of freedom because the researchers analyze only one data set at a time."

Gelman and Loken (2014) did *not* decry the researcher's ability to refine hypotheses in light of the data, for on the contrary, such activity is actually "good scientific practice," they said. The only trouble is that the use of statistical inference to guard against being fooled by randomness just does not work when a necessarily flexible data analysis procedure is pursued. Gelman and Loken (2014) "did not want demands of statistical purity to strait-jacket our science," and we completely agree. Making the distinction between exploratory and confirmatory objectives helps us understand where either flexibility or statistical purity are called for in study design and analysis.

3. Exploratory and Confirmatory Objectives in Scientific Research

The obvious way to avoid the difficulties of overfitting and produce valid statistical inferences is to completely prespecify the study design and statistical analysis plan prior to the start of data collection. This can only be done once a great deal is known about the scientific problem at hand. Hence, most scientific research occurs long before it is possible to entertain such rigid prespecification.

Tukey (e.g., 1969, 1977) made a distinction between exploratory and confirmatory analyses. We contend that most scientific research is *exploratory* in nature: the design, conduct, and analysis of a study are necessarily flexible, and must be open to the discovery of unexpected patterns that prompt new questions and hypotheses. In this context, statistical modeling can be exceedingly useful for elucidating patterns in the data, and researcher degrees of freedom can be helpful and even essential, though they still carry the risk of overfitting. The price of allowing this flexibility is that the validity of any resulting statistical inferences is undermined. In particular, inferences from such models provide no ability to evaluate "statistical significance"—the implausibility of chance as an explanation of the data. Formal statistical inferences are only valid and appropriate for *confirmatory* analyses, where rigid prespecification of design and analysis methods can be made,

and which constitute only the latter stages of the iterative learning process that characterizes most scientific research (Box 1976, 1999). As medical researchers Mogil and Macleod (2017) put it, "most preclinical research articles describe a long chain of experiments, all incrementally building support for the same hypothesis."

In George Box's discussion of Draper (1995), he made an observation that provides the motivating philosophy of this article:

> Statistics has no reason for existence except as a catalyst for scientific enquiry in which only the last stage, when all the creative work has already been done, is concerned with a final fixed model and a rigorous test of conclusions. The main part of such an investigation involves an inductive-deductive iteration with input coming from the subject-matter specialist at every stage. This requires a continuously developing model in which the identity of the measured responses, the factors considered, the structure of the mathematical model, the number and nature of its parameters and even the objective of the study change. With its present access to enormous computer power and provocative and thought-provoking graphical display, modern statistics could make enormous contributions to this—the main body of scientific endeavour. But most of the time it does not.

There is, of course, one arena of science where the exploratory/confirmatory distinction is clearly made, and attitudes toward statistical inferences are sound: the phased experimentation of medical clinical trials. The epistemological framework of exploratory and confirmatory objectives articulated in the ICH guidelines, E8 and E9 (International Conference on Harmonisation 1997, 1998), seems to be unique in scientific research. This framework helps to separate therapeutic exploratory (typically Phase II) with therapeutic confirmatory (typically Phase III) objectives (see *Appendix* A.2). The latter are intended to inform licensing decisions, and imply comprehensive prespecification of (and adherence to) the study protocol, standard operating procedures, and statistical analysis plan, including writing the analysis software code, prior to collecting any data. Exploratory objectives may be pursued in earlier phase trials, where data-driven choices may influence trial conduct and analysis in a flexible way. However ICH E9 states, "Such trials cannot be the basis of the formal proof of efficacy, though they may contribute to the total body of relevant evidence." Although statistical inferences are reported in these earlier phase trials, their cognitive status is very different: they are considered much less definitive than inferences from later phase trials, and licensing decisions are *not* typically made solely on their basis. Moreover, even for Phase III trials, usually at least *two* are required for a New Drug Application. As Piantadosi (2017) noted, "Medicine is a conservative science and behavior usually does not change on the basis of one study."

Nonetheless, the ICH guidance acknowledges that a clinical trial can have both exploratory and confirmatory objectives. In a confirmatory trial, once the prespecified analysis of the primary endpoint is complete, a large, rich, and expensive dataset typically remains at hand. It seems responsible to use that dataset for further exploratory work—to generate hypotheses for further testing in later experiments. A *subgroup analysis*, for example, may seek to identify a subpopulation (perhaps defined by the presence of a particular biomarker at elevated levels) for which efficacy is particularly pronounced (or suppressed). Statistical inferences are often reported, but like Phase II inferences, they only serve as fodder for proposing and designing future confirmatory trials, rather than for defining the current label indication. A succinct perspective on such inferences is given by Sir Richard Peto, often quoted (e.g., Freedman 1998) as saying "you should always do subgroup analysis and never believe the results."

This shift in attitude about the cognitive status of statistical inferences in exploratory versus confirmatory analyses is wholly absent from much of the rest of scientific endeavor. Typically, flexible data analysis methods are used to generate biased statistical inferences that are taken at face value, and used to justify publication decisions. However, in an interview (Shell 2016), the then-Editor of *Science*, and now President of the U.S. National Academy of Science, Marcia McNutt, stated:

> At *Science*, the paradigm is changing. We're talking about asking authors, 'Is this hypothesis testing or exploratory?' An exploratory study explores new questions rather than tests an existing hypothesis. But scientists have felt that they had to disguise an exploratory study as hypothesis testing and that is totally dishonest. I have no problem with true exploratory science. That is what I did most of my career. But it is important that scientists call it as such and not try to pass it off as something else. If the result is important and exciting, we want to publish exploratory studies, but at the same time make clear that they are generally statistically underpowered, and need to be reproduced.

This call for reproduction is reminiscent of the phased clinical trials framework, where multiple studies are carried out, in series and sometimes in parallel, to build an evidence base for a licensing decision of a medical product. Piantadosi (2017) reminded us that "Readers of clinical trials tend to protect themselves by reserving final judgment until findings have been verified independently or assimilated with other knowledge." A single set of data can rarely give definitive results.

4. From the Cult of the Isolated Study to Triangulation

The treatment of statistical inferences from exploratory research as if they were confirmatory enables what Nelder (1986) called *The Cult of the Isolated Study*, so that

> The effects claimed may never be checked by painstaking reproduction of the study elsewhere, and when this absence of checking is combined with the possibility that the original results would not have been reported unless the effects could be presented as significant, the result is a procedure which hardly deserves the attribute 'scientific.'

The use of iterative experimentation and analysis, described by Box above, is obviously more labor intensive, time-consuming, and expensive than worshipping at the Cult of the Isolated Study, but it is more likely to converge to reproducible results. However, we must depart from Tukey's (1977) view that "exploratory and confirmatory can—and should—proceed side by side," and

Gelman's (2003) comment that they "can both be applied at various stages of the analysis." Instead, we agree with Andrew Ehrenberg's (1990) hope for the future of statistics: *Many Sets of Data* (MSOD), which "seems the only way in which we can produce results that are generalizable, lawlike, and predictable—which in fact hold for many different sets of data." Chatfield (1995) added that "The (over?) emphasis on analyzing single sets of data permeates the statistical literature and is a serious disease of statistical teaching." As Freedman (1991) also noted:

> Generally, replication and prediction of new results provide a harsher and more useful validating regime than statistical testing of many models on one data set. Fewer assumptions are needed, there is less chance of artifact, more kinds of variation can be explored, and alternative explanations can be ruled out.

That said, simple replication is usually not sufficient. Many scientific theories have implications that can be tested in multiple ways, as illustrated by John Snow's work on the 1854 cholera outbreak in London, a major step in developing the germ theory of disease. Freedman (1991) wrote of this episode that "The force of the argument results from the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use to get the data." Moreover "He made steady progress from shrewd observation through case studies to analysis of ecological data. In the end, he found and analyzed a natural experiment." There were many sets of data of different kinds, but no statistical inferences were involved. A second example is the discovery of the link between smoking and lung cancer, built on many epidemiological studies of varying designs, again discussed by Freedman (1999). He wrote that because of the doubtfulness of the usual statistical model assumption that patients in a case–control study constitute a random sample,

> Scientifically, the strength of the case against smoking rests not so much on the P-values, but more on the size of the effect, on its coherence and on extensive replication both with the original research design and with many other designs. Replication guards against chance capitalization and, at least to some extent, against confounding—if there is some variation in study design.

In addition he wrote, "Great care was taken to exclude alternative explanations for the findings. Even so, the argument depends on a complex interplay among many lines of evidence." A third example is the invention of powered flight by the Wright Brothers, discussed by Box (1999). The Wright Brothers conducted a lengthy series of experiments, using wind tunnels, kites, gliders, and finally powered aircraft. No formal statistical modeling or inference was used. Such examples support the assertions made by William Feller (1969):

> The aim of basic research is not to produce statistically valid results but to study new phenomena. An evaluation of experimental findings depends on many factors, such as compatibility with other results, predictions to which it leads and so on—such evidence can rarely be evaluated statistically.

Munafo and Davey Smith (2018) define *triangulation* as "the strategic use of multiple approaches to address one question. Each approach has its own unrelated assumptions, strengths and weaknesses. Results that agree across different methodologies are less likely to be artifacts."

A particular weakness of the Isolated Study is that systematic errors may contaminate an entire study but remain hidden if no further research is done. An example is the notorious report by the OPERA collaboration of a detection of faster-than-light neutrinos, with a claimed statistical significance of six sigmas. The team even reproduced this finding using their own apparatus. However, the collaboration later discovered an error due to an improperly connected fiber optic cable. Correcting for this error eliminated the superluminal claim (the OPERA collaboration 2012). Several other examples from physics and astronomy, results with unambiguous statistical significance that "crumbled to dust" due to systematic errors, are given by Seife (2000).

Lithgow, Driscoll, and Phillips (2017) began an instructive tale with initially nonreproducible results on worm lifetimes among different labs, an example of systematic lab-to-lab differences swamping any claims of statistical significance. Three different labs, to achieve reproducible results among themselves, examined and standardized a host of experimental sources of variation, such as

- Lighting and temperature in the labs.
- Amount of heat emitted by microscopes in different labs.
- Stirring versus rocking in the cell isolation procedure.
- Procedure for picking up worms to place them in a new agar dish. (Gentler technicians added a day to worm lifetime!)
- Positioning of flasks in autoclave runs.
- Defining the worm lifetime: starting from the laying of an egg, or from the hatching of an egg?

Lithgow, Driscoll, and Phillips (2017) concluded that such factors had to be addressed more vigorously at the study design stage to improve reproducibility. More commonly the influence of such technical artifacts is discovered post hoc, for instance, through comparison of datasets from different labs, if they are discovered at all. (See also the related discussion of batch effects in Section 7.4.)

Regrettably, statisticians have continued attempts to salvage the validity of statistical inference from the Isolated Study and the "final model." This makes such systematic errors an inherent risk of the Isolated Study paradigm (see Youden 1972; Bailey 2018). Before we discuss positive ideas on how to proceed, we first examine (and dispense with) a few examples of such salvaging attempts.

5. Technical Solutions and Their Deficiencies

Many have tried to address the concerns with the Optimism Principle using a set of methods that we call "technical," which here means attempts to use mathematics and computation to salvage valid statistical inference in the Isolated Study. A general framework for all of them is that the data represent, at least approximately, a representative sample of the population of interest, an assumption that itself calls for validation in many scientific contexts—validation that can only be sought with MSOD.

The most widely known class of such methods is based on adjusting for multiple inferences. These range from the simple Bonferroni inequality to the modern methods of false discovery rate and false coverage rate (e.g., Dickhaus 2014). However, in a case study of human microbiome data discussed by Holmes (2018), "if one counts the number of possible analyses on the same data—allowing for the choice of up to nine outliers, different transformations of the data, choice from 40 different possible distances, and five different ordination methods—the result is more than 200 million possibilities. No multiple hypothesis correction can protect the user." Moreover, these methods do not account for other researcher degrees of freedom, data-dependent choices that influence model selection, which Gelman and Loken (2014) called "invisible multiplicity."

A second class of methods incorporates resistance to overfitting into the statistical modeling process, often through an optimization procedure that penalizes model complexity, an approach sometimes called *regularization*. A signature example is the *lasso* (Tibshirani 1996). While such methods can certainly mitigate overfitting, it remains unclear how badly mis-calibrated the resulting statistical inferences remain. Similar arguments could be made for *robust* statistical methods (designed to be less sensitive to some model assumptions) and *nonparametric* methods (designed to minimize model assumptions). Such methods must still fail to guarantee reliable statistical inference, because they cannot eliminate model uncertainty and systematic error, though they are safer to use than conventional statistical methodology in exploratory data analysis, which we discuss in Section 7.4.

Among other proposed remedies are *data splitting*, to which we alluded earlier in the context of prediction, and various versions of *cross-validation* (see also Stone 1974; Picard and Cook 1984; Faraway 2016). These methods are also widely used in machine learning (e.g., Hastie, Tibshirani, and Friedman 2009). Data splitting can also be considered for more traditional statistical inferences like hypothesis testing (e.g., Cox 1975; Dahl et al. 2008). (Cross-validation is discussed in *Appendix* A.3. Another approach, using bootstrap methodology, is discussed briefly in *Appendix* A.4.) Unfortunately, such procedures (or their variants) are still vulnerable to the Optimism Principle, because random splitting implies that "left-out" samples are similar to the "left-in" samples (Gunter and Tong 2017). Put another way, if "a dataset from model A happens to have features which suggest model B, then the resampled data are also likely to indicate model B rather than the true model A" (Chatfield 1995). Obtaining "more than one set of data, whenever possible, is a potentially more convincing way of overcoming model uncertainty and is needed anyway to determine the range of conditions under which a model is valid" (Chatfield 1995).

Another widely advocated category of technical solutions is *model averaging*, which comes in both Bayesian (Hoeting et al. 1999; Fragoso, Bertoli, and Louzada 2018) and frequentist (Hjort and Claeskens 2003) flavors. This approach acknowledges the futility of assuming a single "true" model; instead many models are fit to one (isolated) dataset, and their outputs are averaged in some fashion. All entertained models must be known and specified; the Bayesian version also requires their priors. This approach doubles down on the notion of reusing a single set of data, and inherits all the hazards of the Cult of the Isolated Study.

As "post model selection inference" is currently an active area of statistical research, there are still other different approaches, some of which are reviewed by Holmes (2018). A common theme of such methods is that an initial full model needs to be prespecified. Once this is done, inferences from a selected final model can be made that account for variable selection, if that model is a submodel of the initially specified one. The work of Berk et al. (2013), Lee et al. (2016), and the Bayesian methods discussed by Efron and Hastie (2016), seem to share this common feature. Such proposals do not capture the full range of "invisible multiplicity" and model uncertainty that we discussed earlier, which cannot be boiled down to just variable selection. Taylor and Tibshirani (2015) conceded as much: "The challenge of correcting for the effects of selection is a complex one, because the selective decisions can occur at many different stages in the analysis process." Like the others cited here, they focus on "more limited problems." Another theme is conditional inference, conditioning on either the selection itself (Lee et al. 2016) or on the training data (Leeb 2009). Such conditioning severely limits the interpretation and generalizability of statistical inferences in the setting of exploratory research and many sets of data.

Only through the iterative learning process, using multiple lines of evidence and many sets of data, can systematic error be discovered, and model refinement be continually guided by new data. Moreover, a retreat into mathematical and computational "remedies" can distract us from interactively confronting the natural phenomena under investigation by acquiring more and better data, under an increasingly wider range of conditions, and with constantly improving experimental methodology. For example, Galileo's late 16th century "Leaning Tower of Pisa" experiment, showing that bodies of different mass and composition fall at the same rate, has been tested under various scenarios for over four centuries, including by Apollo 15 astronaut David Scott (Allen 1972) and, most recently, on board the MICROSCOPE satellite (Touboul et al. 2017; see further references therein for other historical experiments). Incidentally, this example also suggests that our use of the term "confirmatory" is a verbal shorthand, not meant to imply definitive confirmation. All scientific findings are tentative. A "confirmatory" study and analysis are simply those designed so that previously established models may be fit and prespecified effects may be estimated with greater rigor than in a flexible setting. (Experience with late-stage clinical trials shows that even the knowledge they generate must still be considered tentative, e.g., Gauch 2009.)

6. More Thoughtful Solutions

A second set of solutions considers the larger framework for scientific research, not just statistical inference methodology.

One strategy requires *preregistering* both the research hypotheses to be tested and the statistical analysis plan prior to data collection, much as in a late-stage clinical trial (e.g., Nosek et al. 2018). If the temptation to adjust the analysis in light of the data can be resisted, this approach inherently avoids the flexible, data-dependent analysis choices that can

lead to overfitting, regardless of the style of inference being practiced (e.g., frequentist or Bayesian). Indeed, Nosek et al. (2018) argued that such preregistration helps clarify the distinction between hypothesis-generating and hypothesis-testing activities. Hypothesis-generating activities are not ruled out, but are intended for future confirmatory hypothesis-testing, as in the clinical trials framework.

While such an approach is commendable, the simple fact is that much—indeed most—scientific research cannot fit this paradigm (e.g., Scott 2013; Goldin-Meadow 2016). As we have emphasized, most scientific research is (and should be) highly exploratory: this means that not enough can be known a priori to prespecify a fully formed statistical analysis plan. Moreover, as Gelman and Loken (2014) said, we "do not want demands of statistical purity to strait-jacket our science" by limiting the ability to explore modeling alternatives. If science is to be evidence-driven, research must be open to what the data tell us, rather than rigidly committed to a prespecified analysis plan based on prior expectations. As Feller (1969) wrote, "No statistics should stand in the way of an experimenter keeping his eyes open, his mind flexible, and on the lookout for surprises."

A variation on this theme is *preregistered replication*, where a *replication* study, rather than the original study, is subject to strict preregistration (e.g., Gelman 2015). A broader vision of this idea (Mogil and Macleod 2017) is to carry out a whole series of exploratory experiments *without* any formal statistical inference, and summarize the results by descriptive statistics (including graphics) or even just disclosure of the raw data. When results from this series of experiments converges to a single working hypothesis, it can *then* be subjected to a preregistered, randomized, and blinded, appropriately powered confirmatory experiment, carried out by another laboratory, in which valid statistical inference may be made. (Unlike with preregistered replication, the confirmatory study here need not be a literal replication of one of the earlier exploratory studies.) The key is that publication of the exploratory experiments would require that they be accompanied by the confirmatory study in the same manuscript. Mogil and Macleod (2017) explained how this publication model may change the incentives for research for the better. The proposal is appealing, but there are still many challenges and situations (such as observational data studies) that might not fit it well (Gunter and Tong 2017). For example, when a confirmatory study fails to go according to plan, troubleshooting should be prioritized over statistical validity. When this happens, the study simply reverts to becoming another in the series of exploratory studies, rather than confirmatory, as originally intended.

In summary, then, we have argued that the paradigms and assumptions of statistical inference do not fit the inherently exploratory nature of science, and therefore should rarely be applied. We next examine the many positive contributions of statistical methodology.

7. Enabling Good Science

7.1. A Taxonomy of Statistical Activity

We adapt a taxonomy of statistical activity that has previously been used, in different forms, by Cox (1957) and Moore (1992):

- Data production. The planning and execution of a study (either observational or experimental).
- Descriptive and exploratory analysis. Study the data at hand.
- Generalization. Make claims about the world beyond the data at hand.

Data production includes the experimental design, sampling plan, measurement and data collection procedures, and the operational conduct of the study. Descriptive and exploratory analysis includes summary statistics, statistical graphics and tables, and disciplined data exploration. Generalization includes both prediction and statistical inferences. Much of current statistical teaching disproportionately focuses on statistical inference, just one piece of this taxonomy, one that we have argued is the least appropriate for most scientific research. As Freedman (1995) lamented:

> I wish we could learn to look at the data more directly, without the fictional models and priors. On the same wish list: We should stop pretending to fix bad designs and inadequate measurements by modeling.

However, statistical modeling can still contribute to all three sectors of the taxonomy. For example, statistical models help us understand the advantage of factorial designs over one-factor-at-a-time designs; succinctly characterize potential patterns observed in the dataset, such as linear or curved relationships among variables; and produce statistical inferences and predictions for confirmatory analyses. Also contributing to all three sectors is the amorphous notion of *statistical thinking*, which we try to partially characterize in Section 7.5.

The first step of statistical thinking is to understand the *objective* of the study, its *context*, and its *constraints*, so that planning for study design and analysis can be fit for purpose. In this respect, the use of statistical inference as a universal mechanism for scientific validity must be replaced by mainly noninferential statistical methods that are discipline- and problem-specific (Gigerenzer and Marewski 2015). Despite this, the thoughts below may yet be of broad general interest to data analysts in many disciplines. Only the first two items in the taxonomy are discussed here, since as we argue above, generalization best emerges from the iterative learning process described by Box (Section 3). (Statistically literate readers may skip Section 7.2, which is intended for nonspecialist readers.)

7.2. Data Production

Feller (1969) pronounced that "The purpose of statistics in laboratories should be to save labor, time, and expense by efficient experimental designs" rather than null hypothesis significance testing. Some useful basic principles of study design include *concurrent control*, *replication*, *randomization*, *blinding*, and *blocking*. Each of these is not always mandatory (or possible), but they should be carefully considered. In a comparative study, for instance, a group of subjects exposed to an experimental intervention is compared to a control group. Attributing causality of outcomes to the intervention can then be made only if the two groups were treated the same in other respects. Use of a *concurrent control* group helps compensate for certain sources of

systematic errors that may occur, for instance, if the two groups are widely separated in time and/or space. *Random allocation* of subjects to treatment groups helps "even out" other systematic differences between groups that may not be known or measured. *Blinding* of subjects, and others involved in the study, to their assigned treatment reduces the likelihood of unconscious bias. Together, these mechanisms also account for the placebo effect, when present. In many observational studies, *random sampling* (a guarantee that each subject in the target population has an equal chance of being selected) helps to reduce sampling bias in the results. Finally, *randomization of treatment and measurement order* helps to average out the corresponding systematic effects, such as learning effects in a study involving radiologists' interpretation of medical images. Together, these concepts are aimed at bias suppression, including bias due to unknown or unobserved sources, that contribute to misleading conclusions. The impact of such simple procedures can be substantial: Couzin-Frankel (2013) observed that only about 1/3 of mouse studies of stroke therapeutics report randomization or blinding, but studies reporting neither one "gave substantially and significantly higher estimates of how good these drugs were," according to interviewee Malcolm Macleod. In one case the same drug had twice the effectiveness in a study *without* randomization as in one that *did* randomize.

Adequate *replication* allows us to get a handle on experimental variation, such as variation between and within subjects, and variation due to the measurement process. The notion of an *experimental unit* helps us understand the replication structure of the experiment. The experimental unit is the smallest subset of experimental materials or subjects that can be randomly allocated to separate treatments. For instance, if experiments are done with mice in a cage, and treatments are administered through a common food tray in each cage, then the experimental unit is the cage, not the individual mice (which are usually the units of *observation* or *measurement*). Treatments can only be administered at the cage level, not the mouse level. Such a design is often necessary since mice are social, and ideally should live with their litter-mates, not in single-animal cages. (Statistical models can accommodate such nuances during the analysis.) In many such designs, several units of replication may be identified, and the identification of *one* of these as the experimental unit for analysis purposes is an important and sometimes controversial issue impacting scientific interpretability. As another example, a cluster-randomized trial of a new pedagogical method could have schools as units of allocation, classes (within schools) as units of intervention, and students (within classes) as units of measurement (Murray 1998). In biology, we often make the distinction between *biological* replicates (samples from different people or animals) and *technical* replicates (replicate measurements on the same sample); often both terms must be enriched depending on the context.

The concept of *blocking* recognizes that experimental units or runs are not all equally alike. There are subgroups of units or runs whose data are more similar to each other than with those of other subgroups of units or runs. Examples include plots of land in close proximity, in an agronomy study involving multiple fields; litter-mates in animal studies involving multiple litters of animals; experimental runs made on the same day versus those from other days; and clinical trial subjects enrolled close to each other in time and at the same clinical site. These known sources of variation can be accounted for in both the study design and analysis, to improve the quality of information obtained from the experiment. Randomization should still be applied *within* these subgroups ("blocks") when they have been identified. As the classic text, Box et al. (2005), advised: "Block what you can and randomize what you cannot." (In observational studies, *stratified sampling* plays an analogous role to blocking.)

Confounding occurs when the treatment groups differ systematically in some way besides the intervention. Randomization, blinding, and blocking help to eliminate some sources of confounding. When prognostic covariates are known for enrolled subjects, methods for allocating subjects to treatments that seek approximate balance on such covariates can further reduce the risk of confounding (e.g., Lock Morgan and Rubin 2012; Kallus 2018).

More advanced notions of experimental design include matched-pairs designs, factorial designs, split-plotting, crossover trials, and many other concepts that we have not the space to discuss here, but may be appropriate in different situations. Beyond these standard topics of study design, other aspects of data production must be considered. Measurement processes must be well-defined, standardized, and validated, with quality assurance procedures in place. Data acquisition and storage systems should have appropriate resolution and reliability. (We once worked with an instrument that allowed the user to retrieve stored time series data with a choice of time-resolution. Upon investigation, we found that the system was artificially interpolating data, and reporting values not actually measured, if the user chose a high resolution.) Again, many issues will be discipline- and problem-specific. In biological research for instance, authentication of cell lines and validation of antibodies can address major sources of systematic error (see Harris 2017).

Earlier we described "researcher degrees of freedom" that are characteristic of flexible data analysis. Other researcher degrees of freedom can affect study design and execution. An instructive example for the latter is the decision to terminate data collection. Except in clinical trials, where this decision is tightly regulated and accounted for in the subsequent analysis (e.g., Chow and Chang 2012), many researchers have no formal termination rule, stopping when funding is exhausted, lab priorities shift, apparent statistical significance is achieved (or becomes clearly hopeless), or for some other arbitrary reason, often involving unblinded interim looks at the data. *Any* formal or informal rule for terminating data collection, whether an explicit calculation of statistical inference is used or not, has inherent risks of chasing a false positive for too long, or burying a false negative too soon. In the exploratory setting, probability claims about such risks (e.g., "power analysis") are just as invalid as the statistical inferences that obtain during and after data collection.

Principles of study design are a major positive contribution of statistical methodology dating from the time of Fisher (1926), and we feel that much of the reproducibility crisis in science (e.g., Harris 2017) can be addressed using this sector of statistical activity, quite apart from the inference issues that have dominated statisticians' conversations about reproducible research. Many general topics are covered in standard texts on experimental design (e.g., Winer, Brown, and Michels 1991; Kuehl 2000; Box et al. 2005; Montgomery 2017) and sampling

(e.g., Cochran 1977; Thompson 2012) though again each discipline has its own specific issues of data production to deal with. Coleman and Gunter (2014) provided a brief introduction to some of the main ideas of experimental design, and emphasized how varying multiple variables simultaneously can be used profitably, contrary to the conventional wisdom of varying one variable at a time, while keeping all others fixed. Despite the evident utility of all these ideas—and their relevance to many of the problems of nonreproducibility—many scientists seem largely unaware of or confused by them. We therefore think this is fertile ground for improving statistical education for scientists and should be an important component of such efforts.

7.3. Data Description

Data description and exploration are related activities that focus on the data at hand, without attempting to make inferences to the world beyond. Consider first data description. Many data analytical problems do not lend themselves to statistical inference at all: when there is no sense in which the data are even approximately a sample from a population; or when a probability model (conveying some notion of randomness) is irrelevant (Mallows and Walley 1980) or even misleading, as in Taleb's (2007) "ludic fallacy." However almost all data analysis requires at least some data reduction or description, which may include statistical summaries, tables, and visualizations. Moses (1992) warned us that

> Good statistical description is demanding and challenging work: it requires sound conceptualization, and demands insightfully organizing the data, and effectively communicating the results; not one of those tasks is easy. To mistakenly treat description as 'routine' is almost surely to botch the job.

One of the best developed branches of data description is statistical graphics and data visualization, the subject of many books (e.g., Cleveland 1993, 1994; Robbins 2013) and software packages. Scientists could use help in this arena. Perhaps the most egregious yet ubiquitous graphical method found in life science articles is a set of barplots featuring one-sided error bars (sometimes called "skyscraper" or "dynamite" plots), used to represent univariate data from different comparison groups. Such plots are poor representations of the data (Pikounis 2001; Koyama 2011). Krzywinski and Altman (2014) advocated using *boxplots* (Tukey 1977) as a less distorting alternative. Superimposing a representation of the actual data points on a boxplot mitigates some of its deficiencies, such as its inability to indicate bimodality.

Mallows (1983) provided an interesting perspective on a Theory of Description. He noted that "A good descriptive technique should be appropriate for its purpose; effective as a mode of communication, accurate, complete, and resistant." By *effective as a mode of communication*, he included the notions of "familiarity, relative to the target audience, simplicity, and honesty." By *accurate*, he meant "a measure of the closeness to which the description approximates the data," and proposed a discrepancy measure. By *complete*, he meant "the degree to which there is an absence of structure in the residual or undescribed variation in the data." To elaborate, statistical models may be deployed to examine patterns in the data. Deviations between the observed data and the model's fitted values are called *residuals*, and can themselves be examined for patterns suggesting systematic effects not fully captured in the fitted model. By *resistant*, Mallows (1983) meant that methods should minimize "sensitivity …to small changes in the data—either small perturbations in all the data, or arbitrarily large perturbations in a small part of the data." Resistance to distortion by extreme values, potential outliers, is a common example. In this respect the *median* is often used in place of the mean, and the *median absolute deviation* in place of the standard deviation. For curve fitting, *least absolute error* is more resistant than *least squared error*. Mallows acknowledged that his criteria are sometimes in competition with each other. Median absolute deviations are not widely familiar outside of statistics, so using them prioritizes resistance over effectiveness in communication. Nonetheless, his criteria provide a useful guide when designing a data description for a given audience.

Though we might not quantify *uncertainty* using probability statements, we can attempt to convey the observed *variability* of the data at hand, while acknowledging that it does not fully capture uncertainty. For univariate data, quantities such as Tukey's (1977) five-number summary (the minimum, first quartile, median, third quartile, and maximum) or other empirical quantiles partially describe the distribution of the data, as does a graph of the empirical distribution function. In addition, resistant measures of dispersion, such as the median absolute deviation, Gini's mean difference (David 1968), and others (e.g., Rousseeuw and Croux 1993) are available to characterize observed variability of the data at hand. However, the use of such data summaries is not free of assumptions (e.g., unimodality, in some cases symmetry), so they are descriptive only in relation to these assumptions, not in an absolute sense. They cannot serve as measures of uncertainty due to the selection bias inherent in making such assumptions.

The relationship among many variables is less easily characterized by such simple methods; in such cases, "a statistical model is often the best descriptive tool, even when it's not used for inference" (Harrell 2018). For example, bivariate data can be examined with scatterplots and simple curve fits, such as the *loess smoother*, a type of nonparametric regression that incorporates resistance to atypical values (Cleveland 1979). The loess fit is determined by a user-selected parameter called the *span*, and varying this parameter controls the tradeoff between variability captured by the curve fit and variability left over to the residuals. These residuals may be graphed, and when devoid of clear patterns, univariate measures of dispersion such as those just mentioned could be applied to quantify their dispersion from the fitted curve. Of course, the user's choice of the span parameter illustrates that model selection bias remains a hazard. Any quantification of observed variability only describes the relationship between the model and the data at hand. (See Appendix A.5 for a discussion of standard errors as descriptive statistics.)

The use of variability as a benchmark for measuring effect size is often emphasized by statisticians, but is rarely the only benchmark of interest when assessing practical/clinical significance. There are even cases where "variability is so small (or the

data are so abundant) that variability is not the central issue" (Mallows 1998). Two examples (depending on the context, as always) include the power spectrum in time series (as noted by Tukey 1962) and the mean kinetic temperature in applied thermodynamics (see Tong and Lock 2015); examples from materials science are given by Wenmackers and Vanpouke (2012). The selection of appropriate benchmark(s) and the subsequent design of meaningful descriptive statistics are highly problem-dependent and require both a grasp of the subject matter and creative thinking.

7.4. Disciplined Data Exploration

Effective data *exploration* goes beyond data *description*. According to Tukey (1973), exploratory analysis of the data is not "just descriptive statistics," but rather an "actively incisive rather than passively descriptive" activity, "with a real emphasis on the discovery of the unexpected." Later, Tukey (1977) added, "Exploratory data analysis is detective work—numerical detective work—or counting detective work—or graphical detective work." The founding text is Tukey (1977), which was followed by a series known as the *Statistician's Guide to Exploratory Data Analysis* (Hoaglin, Mosteller, and Tukey 1983, 1985, 1991). Pearson (2011) provided a contemporary perspective on exploratory data analysis.

An example of how exploratory analysis may be essential for scientific inquiry is in the detection of and adjustment for *batch effects*. Leek et al. (2010) defined batch effects as "sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study." Their examples include data obtained under varying experimental processing conditions, perhaps from multiple laboratories using different protocols. They found that

> For example, multiple laboratory comparisons of microarray experiments have shown strong laboratory-specific effects. In addition, in nearly every gene expression study, large variations are associated with the processing date, and in microarray studies focusing on copy number variation, large effects are associated with DNA preparation groups. The processing group and date are therefore commonly used to account for batch effects. However, in a typical experiment these are probably only surrogates for other sources of variation, such as ozone levels, laboratory temperatures and reagent quality. Unfortunately, many possible sources of batch effects are not recorded, and data analysts are left with just processing group and date as surrogates.

Leek et al. (2010) also observed that "In gene expression studies, the greatest source of differential expression is nearly always across batches rather than across biological groups, which can lead to confusing or incorrect biological conclusions owing to the influence of technical artefacts." While known batch effects can be accounted for at the study design stage, unknown batch effects may well remain. Statistical methods for the detection of batch effects rely on (noninferential) exploratory analysis using multivariate statistical methods, such as principal components analysis and hierarchical clustering. Adjusting the data for such effects often makes (noninferential) use of linear statistical models. Of course, the comparison of many sets of data, from different laboratories and varying experimental protocols, further helps elucidate batch effects.

Diaconis (1985) warned that undisciplined exploratory analysis of a dataset lends itself to finding spurious patterns as well as real ones, as we discussed earlier. If such patterns are accepted as gospel without considering that they *may* have arisen by chance, he considers it *magical thinking*, which he defines as "our inclination to seek and interpret connections and events around us, together with our disinclination to revise belief after further observation." (See also the related discussion of Grolemund and Wickham 2014.) Statistical methods mentioned earlier that resist overfitting (regularized methods, such as the lasso), are less sensitive to some model assumptions (robust estimators, e.g., Maronna, Martin, and Yohai 2006), or make few such assumptions (nonparametric estimators, e.g., Wasserman 2006) help guard against being fooled by randomness, though their associated probabilistic claims about uncertainty remain unreliable. (A new class of statistical methods with the property of *differential privacy* may also be worthy of further consideration; see Dwork et al. 2016.)

7.5. Statistical Thinking

Mallows (1998) reviewed several definitions of statistical thinking, and provided his own:

> Statistical thinking concerns the relation of quantitative data to a real-world problem, often in the presence of variability and uncertainty. It attempts to make precise and explicit what the data has to say about the problem of interest.

Statistical thinking begins with a relentless focus on *fitness for purpose* (paraphrasing Tukey 1962: seeking approximate answers to the right questions, not exact answers to the wrong ones), sound attitudes about data production and its pitfalls, and good habits of data display and disciplined data exploration. An incomplete list of other attributes of statistical thinking includes the following.

- Evidence, often in the form of data, matters, "for without data, everyone is an expert" (Snee 1986). As Frederick Mosteller said, "It is easy to lie with statistics, but a whole lot easier without them" (quoted in Holmes 2018).
- All observations are variable; all conclusions are uncertain. (However, as discussed above, variability is not always the aspect of the data that should most be emphasized.)
- A holistic view of the data. For instance, we study measurement *processes*, not just measurement *instruments*, which comprise only one component of the whole process of measurement, which also involves personnel, reagents, environmental conditions, etc. Another example is the *Intention to Treat* (ITT) principle in clinical trials, whereby a patient's data are analyzed according to the random assignment of the patient to treatment group, rather than whether the patient actually complied with the protocol (e.g., Piantadosi 2017). This reflects interest in treatment *policy*, including the occurrence of realistic phenomena such as errors and noncompliance.

- Statistical significance is *not* a measure of practical, clinical, or scientific significance.

Statisticians have a useful language and a mathematical apparatus for decomposing sources of variation, both systematic and random. These include the basic bias-variance decomposition, the discussion of replication structure in Section 7.2, and so on, up to complex hierarchical models with both nested and crossed factors and nontrivial correlation structures among variables. Unfortunately, in many cases such statistical models may be difficult to fit, due to lack of data or unrealistic model assumptions. Any inferences from such models are subject to model selection biases and the Optimism Principle. Nonetheless, the paradigm represented by such models is a guide to precise quantitative thinking characteristic of the statistical approach to data analysis. Statisticians are also taught to be sensitive to sources of bias and variability, as well as confounding, issues nicely illustrated in the discussion of batch effects given by Leek et al. (2010), an article in which statistical inference plays almost no role.

Statistical thinking also involves a keen awareness of the pitfalls of data analysis and its interpretation, including:

- The correlation versus causation fallacy.
- The distinction between interpolation and extrapolation.
- The distinction between experimental and observational data.
- Regression to the mean.
- Simpson's paradox, and the ecological fallacy.
- The curse of dimensionality.

We have only outlined some of the elements of statistical thinking, and many readers may have more to add.

8. Discussion

To summarize the argument of this article:

- Most scientific research is exploratory, not confirmatory. The research process is iterative, requiring many sets of data and many lines of evidence (Freedman's "shoe leather" and triangulation). Avoid the Cult of the Isolated Study.
- Attention to statistical issues in the design and execution of the study should be the primary concern. Remain on guard for systematic error.
- Methods with alleged generality, such as the p-value or Bayes factor, should be avoided in favor of discipline- and problem-specific solutions that can be designed to be fit for purpose.
- Formal statistical inference may only be used in a confirmatory setting where the study design and statistical analysis plan are specified prior to data collection, and adhered to during and after it. That is the only setting where we may rely on the Efron and Hastie (2016) principle of using the same data that produces an estimate to assess its precision; in any other setting, statistical inferences are undermined by the Optimism Principle.
- Statistical analysis of exploratory research data should rely only on descriptive methods (summary statistics, tables, and graphics) and disciplined data exploration, which often involves statistical modeling. The latter is often enabled by statistical methods that are regularized, robust, and/or nonparametric, which are safer to use than conventional methodology, but still do not fully eliminate the Optimism Principle. Bear in mind the Harrell (2015) maxim, "Using the data to guide the data analysis is almost as dangerous as not doing so" and Diaconis' (1985) warning about magical thinking.
- The framework outlined by Mogil and Macleod (2017) is an example of an approach to research and publication that seems consistent with the exploratory/confirmatory distinction outlined here.
- Exploratory analysis of data from a confirmatory study (subsequent to the completion of the preplanned analysis) should also be entertained, for the purpose of hypothesis *generation* rather than hypothesis *testing*. Subgroup analysis for late-stage clinical trials is a signature example.
- Statistical thinking provides a framework for critical thinking that can benefit every stage of the research program.

A counterargument to our position is that inferential statistics (p-values, confidence intervals, Bayes factors, and so on) could still be used, but considered as just elaborate descriptive statistics, without inferential implications (e.g., Berry 2016; Lew 2016). We do not find this a compelling way to salvage the machinery of statistical inference. Divorced from the probability claims attached to such quantities (confidence levels, nominal Type I errors, and so on), there is no longer any reason to privilege such quantities over descriptive statistics that more directly characterize the data at hand. The danger is that both cultural inertia and the seductive appearance of quantified uncertainty may continue to incentivize the inappropriate reporting of statistical inferences.

A second counterargument is that, as George Box (1999) reminded us, "All models are wrong, but some are useful." Statistical inferences may be biased per the Optimism Principle, but they are reasonably approximate (it might be claimed), and paraphrasing John Tukey (1962), we are concerned with *approximate* answers to the right questions, not *exact* answers to the wrong ones. This line of thinking also fails to be compelling, because we cannot safely estimate how large such approximation errors can be. In a single study, hypothesis test selection can lead to underestimation of error by an order of magnitude (e.g., Huber 1985). Recall also Freedman's Paradox, a "final model" with bogus significant variables. Of course unknown systematic error is completely unaccounted for. Many sets of data and triangulation are more reliable ways to explore the approximation error in our (tentative) conclusions.

One reviewer of this article characterized our view as "the proposed solution for imperfect variance estimation is no variance estimation," and then asked "Is no quantification of uncertainty truly better than imperfect quantification?" We think many readers will share this question. In Section 7.3, we made a distinction between characterizing observed variability and quantifying uncertainty. In the exploratory/learning phases of research, it is way too early to pretend to be able to quantify uncertainty. However, we *can* describe how variable the data at hand are, relative to one or more models, using well-chosen descriptive statistics and graphs, though this variability should

not be conflated with uncertainty. Uncertainty also includes model uncertainty, due to model selection bias (the Optimism Principle) and the potential for systematic error, which both require many sets of data to fully evaluate. There is no scientifically sound way to quantify uncertainty from a single set of data, in isolation from other sets of data comprising an exploratory/learning process.

Tukey (1962) once suggested that "it might be well if statisticians looked to see how data was actually analyzed by many sorts of people." In fact, some statisticians have shown just such an interest, for example, Box (1999) and Freedman (1991, 1999), as discussed earlier. For the most part, however, statistics teaching and practice ignores the fact that most of the great historical discoveries in science, engineering, and medicine were made without the crutch of statistical inference (Gigerenzer and Marewski 2015). Kepler's laws of planetary motion, the periodic table of the elements, the germ theory of disease, plate tectonics, the molecular structure of DNA, and the quantization of energy are other examples where data and modeling, without statistical inferences, were crucial for discovery. Financial accounting, a data profession much older and larger than statistics, often depends on estimation in an uncertain setting, but does not report any precise uncertainty quantification in the balance sheet. This brings to mind an observation made about certain research in materials science: "Even if the studies had reported an error value, the trustworthiness of the result would not depend on that value alone" (Wenmackers and Vanpouke 2012).

Nevertheless, we do think that the discipline of statistics has much to offer science. By focusing on the methods that are of broadest usefulness to science, engineering, and medicine—rather than on an obsession with statistical inference and the Cult of the Isolated Study—statistical thinking has much to contribute to scientific work. By emphasizing principles of data production, data description, enlightened data display, disciplined data exploration, and exposing statistical pitfalls in interpretation, there is much that statisticians can do to ensure that statistics is "a catalyst to iterative scientific learning" (Box 1999). *Statistical inference enables bad science; statistical thinking enables good science.*

Appendix: Further Elaborations

A.1. Frequentist and Bayesian Schools of Inference

There are several schools of thought on how statistical inference should be carried out (e.g., Barnett 1999). For instance, the survey by Geisser (2006) focuses on the four most prominent: frequentist, Bayesian, likelihoodist, and fiducial. The first two of these are by far the most often used in practice, and we describe them briefly here. Bear in mind that within each school of thought, there is further variation of both perspective and methodology.

The frequentist approach is based on the interpretation of probability as a fixed long-run frequency, imagined as resulting from replication of the data generating process ad infinitum. This requires defining a fixed probability space a priori, which in turn requires knowing the intentions of the investigator (Berger and Berry 1988), a setting wholly inconsistent with flexible data analysis (Gunter and Tong 2017). Frequentist inference is only conceivable for prespecified confirmatory analyses.

The Bayesian approach views probability as a subjective degree of belief. A Bayesian analysis begins by specifying prior probability distributions for the parameters in a stochastic model, then updating these probabilities in light of the data (using Bayes' theorem, and a model-dependent quantity known as the *likelihood function*). It represents a computationally intensive set of procedures attempting to capture the intuition that prior knowledge should be combined with current data to make current knowledge. This sounds deceptively consistent with the iterative learning process advocated in this article. However, it is not. The placement of prior distributions on model parameters is a highly artificial way of embedding prior knowledge; few scientists formulate their beliefs in terms of real-valued (i.e., infinitely precise) probability functions. The more natural (and more important) way that prior knowledge is incorporated (for both frequentists and Bayesians) is in the choice of study design (including what variables to measure, and how to measure them) and the class of models the analyst is prepared to entertain (Harrell 2018). Harrell (2018) noted that for the Bayesian "the model choice does not 'wear off' nearly as much as the prior does as the sample size gets large," so this choice is more impactful than the choice of priors. However, this model choice remains subject to researcher degrees of freedom and the Optimism Principle.

In the iterative learning process described by Box (Section 3), there is no fixed model whose parameters can be repeatedly updated, because the model itself changes as the research program evolves. In the most extreme cases, such as alleged scientific revolutions (Kuhn 1970), the incommensurate paradigms cannot transfer information to each other through prior probabilities. Bayesian inference cannot guide an entire research program, but like frequentist inference, may be of value within the confines of a confirmatory analysis.

Harrell (2018) observed that even in confirmatory analyses, model uncertainty remains, because we never know if the model is misspecified. He argued that Bayesian inference can accommodate some degree of model uncertainty in this setting: "If the model contains a parameter for everything we know we don't know (e.g., a parameter for the ratio of variances in a two-sample t-test), the resulting posterior distribution for the parameter of interest will be flatter, credible intervals wider, and confidence intervals wider. This makes them more likely to lead to the correct interpretation, and makes the result more likely to be reproducible." This approach may be considered by those whose response to model uncertainty is to double down on probability modeling. Others might instead back off, and rely instead on our suggestions for exploratory analysis: regularization, robustness, and nonparametricness. The differences in interpretation among the various styles of statistical inference are another factor to consider.

A.2. Phased Clinical Trials

The framework of phased clinical trials is an exemplar of scientific epistemology. The ICH E8 states:

> The cardinal logic behind serially conducted studies of a medicinal product is that the results of prior studies should influence the plan of later studies. Emerging data will frequently prompt a modification of the development strategy.
> […]
> Drug development is ideally a logical, step-wise procedure in which information from small early studies is used to support and plan larger, more definitive studies. To develop new drugs efficiently, it is essential to identify characteristics of the investigational medicine in the early stages of development and to plan an appropriate development based on this profile.

> Initial trials provide an early evaluation of short-term safety and tolerability and can provide pharmacodynamic and pharmacokinetic information needed to choose a suitable dosage range and administration schedule for initial exploratory therapeutic trials. Later confirmatory studies are generally larger and longer and include a more diverse patient population. …Throughout development, new data may suggest the need for additional studies that are typically part of an earlier phase.

To oversimplify, the typical stages of clinical trials are as follows:

- Phase I. This represents the first administration of the drug in humans, usually to a small number of healthy volunteers. The goal is to evaluate tolerability of the dose range, observe potential side effects, and evaluate clinical pharmacokinetics, such as drug absorption, distribution, metabolism, and excretion (ADME).
- Phase II. These studies are typically therapeutic exploratory studies. Patients are enrolled under "relatively narrow criteria" to obtain data on drug efficacy, safety, and dose response. Flexibility in design and analysis may be entertained.
- Phase III. These studies are typically therapeutic confirmatory studies. A large number of patients, from a wider population than Phase II, are enrolled to confirm drug efficacy and safety. Very tight specification of study protocols, standard operating procedures, and statistical analysis plans should be made prior to data collection, and adhered to during and after it. There is rightful indignation when the prespecified analysis is altered for publication purposes (see *http://compare-trials.org/*).

For a more detailed exposition, consult the guidance documents ICH E8 and E9 (International Conference on Harmonisation 1997, 1998) and textbooks such as Piantadosi (2017). (See also Scheiner 1997, for another perspective.)

The intellectual heritage of phased clinical trials is not statistical, but statisticians have helped sharpen the reasoning this framework embodies. Carpenter (2010, chap. 4) traced the origins of phased clinical trials to pharmacologists at the U.S. Food and Drug Administration (FDA), such as A.J. Lehman and O. Garth Fitzhugh, and clinical oncologists at the National Cancer Institute (NCI), who made the distinction between Phases I and II in the 1950s. The notion of Phase III was formulated by FDA physicians, attorneys, and other officials, around the time that Congress passed the 1962 Kefauver-Harris amendment to the 1938 Federal Food, Drug, and Cosmetic Act. Phases I, II, and III first appear in formal regulatory guidance in the FDA's draft Investigational New Drug rules of January 1963. Carpenter writes of the latter, "The authorship of phased experimentation appears to have been collaborative, with [Julius] Hauser and [Frances] Kelsey assuming crucial leadership roles." Of particular interest to statisticians, the drafting of the ICH E9 guidance in the 1990s is described briefly by Lewis (1999).

A.3. Cross-Validation

Hastie, Tibshirani, and Friedman (2009, p. 241) wrote that "Probably the simplest and most widely used method for estimating prediction error is cross-validation." As an example, we consider five-fold cross-validation (5CV), which proceeds as follows.

1. Randomly partition the dataset into five roughly equal sized subsets.
2. For each of the five subsets separately, perform model fitting/criticism/selection on an aggregation of the other four subsets, and then use the left-out fifth to evaluate performance of the model, as in data splitting. The model fitting procedure should be coded to run automatically (without human intervention) so that it can be called in each iteration. (Of course, such an automated procedure could not reflect the human judgment usually applied in actual model building.)
3. After this has been done five times, each data record has one predicted value, from a model built on 4/5 of the dataset, that can be compared with the actual value. Prediction performance can be then be evaluated by comparing this whole set of predicted values with the actual values. The resulting performance estimate does not correspond to any particular one of the five models that was trained, but to the automated modeling *procedure* that was coded.

A.4. Bootstrap Methodology

Bootstrap methodology is a signature achievement of statistical inference in the late 20th century, and it illustrates the maxim (disputed here) that the same data used to provide an estimate can be used to assess its precision (Efron 1982). However, our interest in the bootstrap here is in its use to examine model selection bias in various ways (e.g., Efron 1983; Gong 1986; Dijkstra 1988; Efron 2014). A bootstrap resampling procedure for assessing prediction performance could operate as follows. Generate multiple bootstrap samples, which are samples with replacement from the original dataset. On each of these, fit a statistical model, then assess the model's performance on the original dataset. (Like cross-validation, this requires coding an automated model building procedure that can be called iteratively.) An overall estimate of performance is obtained by averaging over the results from all bootstrap samples. An enhanced approach is to use the bootstrap to evaluate the expected bias due to overfitting. Harrell (2015), sec. 5.3, provided a more thorough discussion than we can give here. However, the criticisms we raised earlier about data splitting and cross-validation mostly extend to such bootstrap methods as well. As Harrell (2018) observed, all these methods attend to *internal* not *external* validation, and it is external validation that matters most to the user. However, Harrell (2018) argued that among the *internal* validation methods, the bootstrap is least wasteful of the data.

A.5. Standard Errors as Descriptive Statistics

For univariate data, the standard error is often used as an error bar for the mean. In regression modeling, regression coefficients are sometimes reported with standard errors to characterize their precision.

As Motulsky (2014) observed, the standard error is *not* a measure of variability, since it shrinks with sample size regardless of how variable the data are. Rather, the standard error is a measure of *precision* for a point estimate under certain probability models. From this perspective, the standard error is associated with inferential probability statements derived from a statistical model, and such statements are often invalid due to the Optimism Principle. For this reason, we advise avoiding the standard error even as a descriptive statistic. Regardless, often the variability of the data themselves (think standard deviation, not standard error) can be a practically or clinically meaningful benchmark. (For the cognoscenti, Cohen's D may be more meaningful than a t-statistic, though neither properly characterizes uncertainty.)

A.6. Recommended Reading

Even in a manuscript of this length, we have not been able to rehearse the compelling arguments made by earlier writers with full justice. It would be pointless to do so when the original writers have been so eloquent. Readers interested in specific topics are invited to track down the literature cited in sections of this article that interest them. More generally, if we had to recommend just three articles that capture

the spirit of the overall approach outlined here, they would be (in chronological order) Freedman (1991), Gelman and Loken (2014), and Mogil and Macleod (2017).

Acknowledgments

This article was written on the author's personal time, not as part of his official duties. Many of the ideas in this article were developed jointly with Bert Gunter, to whom the author also owe thanks for considerable improvements to the manuscript. The author is grateful to several readers for valuable discussions and critique, particularly Prof. Frank Harrell and the anonymous reviewers, and for feedback from audiences when he presented some of these ideas in Reno (October 2017) and Indianapolis (May 2018). The author takes the sole responsibility for the final content. The views expressed do not necessarily represent the policies, views, or opinions of the authors's employer.

References

Allen, J. P. (1972), "Summary of Scientific Results," Apollo 15 Preliminary Science Report NASA SP-289, NASA Manned Spacecraft Center, Washington, DC. [251]

Ambroise, C., and McLachlan, G. J. (2002), "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data," *Proceedings of the National Academy of Sciences*, 99, 6562–6566. [247]

Argamon, S. E. (2017), "Don't Strengthen Statistical Significance—Abolish It," *American Scientist*, Macroscope blog [online], available at *https://www.americanscientist.org/blog/macroscope/dont-strengthen-statistical-significance-abolish-it* [246]

Bailey, D. (2018), "Why Outliers Are Good for Science," *Significance*, 15, 14–19. [250]

Barnett, V. (1999), *Comparative Statistical Inference* (3rd ed.), Chichester, UK: Wiley. [257]

Berger, J. O., and Berry, D. A. (1988), "Statistical Analysis and the Illusion of Objectivity," *American Scientist*, 76, 159–165. [257]

Berk, R., Brown, L., and Zhao, L. (2010), "Statistical Inference After Model Selection," *Journal of Quantitative Criminology*, 26, 217–236. [248]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-Selection Inference," *Annals of Statistics*, 41, 802–837. [251]

Berry, D. (2016), "P-values Are Not What They're Cracked Up To Be," Online supplement to Wasserstein & Lazar (2016). [256]

Box, G. E. P. (1976), "Science and Statistics," *Journal of the American Statistical Association*, 71, 791–799. [249]

——— (1999), "Statistics as a Catalyst to Learning by Scientific Method Part II—A Discussion," *Journal of Quality Technology*, 31, 16–29. [249,250,256,257]

Box, G. E. P., Hunter, J. S., and Hunter, W. G. (2005) *Statistics for Experimenters: Design, Innovation, and Discovery* (2nd ed.), Hoboken, NJ: Wiley. [253]

Carpenter, D. (2010), *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*, Princeton, NJ: Princeton University Press. [258]

Chatfield, C. (1995), "Model Uncertainty, Data Mining and Statistical Inference" (with discussion), *Journal of the Royal Statistical Society*, Series A, 158, 419–466. [248,250,251]

Chow, S.-C., and Chang, M. (2012), *Adaptive Design Methods in Clinical Trials* (2nd ed.), Boca Raton, FL: CRC Press. [253]

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836. [254]

——— (1993), *Visualizing Data*, Summit, NJ: Hobart Press. [254]

——— (1994), *The Elements of Graphing Data* (2nd ed.), Summit, NJ: Hobart Press. [254]

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley. [254]

Coleman, D., and Gunter, B. (2014), *A DOE Handbook: A Simple Approach to Basic Statistical Design of Experiments*, Seattle, WA: Amazon CreateSpace. [254]

Couzin-Frankel, J. (2013), "When Mice Mislead," *Science*, 342, 922–925. [253]

Cox, D. R. (1975), "A Note on Data-Splitting for the Evaluation of Significance Levels," *Biometrika*, 62, 441–444. [251]

Cox, G. M. (1957), "Statistical Frontiers," *Journal of the American Statistical Association*, 52, 1–12. [252]

Dahl, F. A., Grotle, M., Benth, J. J. S., and Natvig, B. (2008), "Data Splitting as a Countermeasure Against Hypothesis Fishing: With a Case Study of Predictors for Low Back Pain," *European Journal of Epidemiology*, 23, 237–242. [251]

David, H. A. (1968), "Gini's Mean Difference Rediscovered," *Biometrika*, 55, 573–575. [254]

Diaconis, P. (1985), "Theories of Data Analysis: From Magical Thinking Through Classical Statistics," in *Exploring Data Tables, Trends, and Shapes*, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: Wiley, pp. 1–36. [255,256]

Dickhaus, T. (2014), *Simultaneous Statistical Inference*, Berlin: Springer. [251]

Dijkstra, T. K. (ed.) (1988), *On Model Uncertainty and its Statistical Implications* (*Lecture Notes in Economics and Mathematical Systems*, Vol. 307), Berlin: Springer. [258]

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty" (with discussion), *Journal of the Royal Statistical Society*, Series B, 57, 45–97. [248,249]

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2016), "Preserving Statistical Validity in Adaptive Data Analysis," [online], available at *https://arxiv.org/abs/1411.2664v3* [255]

Efron, B. (1982), *The Jacknife, the Bootstrap and Other Resampling Plans*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [258]

——— (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331. [258]

——— (2014), "Estimation and Accuracy After Model Selection," *Journal of the American Statistical Association*, 109, 991–1007. [258]

Efron, B., and Hastie, T. (2016), *Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science*, New York: Cambridge University Press. [247,251,256]

Ehrenberg, A. S. C. (1990), "A Hope for the Future of Statistics: MSOD," *The American Statistician*, 44, 195–196. [250]

Faraway, J. (2016), "Does Data Splitting Improve Prediction?" *Statistics and Computing*, 26, 49–60. [251]

Feller, W. (1969), "Are Life Scientists Overawed by Statistics? (Too Much Faith in Statistics)," *Scientific Research*, 4, 24–29. [250,252]

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London A*, 222, 309–368. [247]

——— (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513. [253]

Fragoso, T. M., Bertoli, W., and Louzada, F. (2018), "Bayesian Model Averaging: A Systematic Review and Conceptual Classification," *International Statistical Review*, 86, 1–28. [251]

Freedman, D. A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152–155. [248]

——— (1991), "Statistical Models and Shoe Leather," *Sociological Methodology*, 21, 291–313. [246,250,257,259]

——— (1995), "Issues in the Foundations of Statistics: Probability and Statistical Models," *Foundations of Science*, 1, 19–39. [252]

——— (1998), "Oasis or Mirage?," *Chance*, 21, 59–61. [249]

——— (1999), "From Association to Causation: Some Remarks on the History of Statistics," *Statistical Science*, 14, 243–258. [250,257]

Gauch, R. R. (2009), *It's Great! Oops, No It Isn't: Why Clinical Research Can't Guarantee the Right Medical Answers*, Berlin: Springer. [251]

Geisser, S. (2006), *Modes of Parametric Statistical Inference*, Hoboken, NJ: Wiley. [257]

Gelman, A. (2003), "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing," *International Statistical Review*, 71, 369–382. [250]

——— (2015), "Statistics and Research Integrity," *European Science Editing*, 41, 13–14. [252]

——— (2016), "The Problems with P-values Are Not Just With P-values," online supplement to Wasserstein & Lazar (2016). [246,248]
Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465. [248,251,252,259]
Gigerenzer, G., and Marewski, J. (2015), "Surrogate Science: The Idol of a Universal Method for Scientific Inference," *Journal of Management*, 41, 421–440. [252,257]
Goldin-Meadow, S. (2016), "Why Preregistration Makes Me Nervous," *Association for Psychological Science Observer*, [online], available at *https://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous* [252]
Gong, G. (1986), "Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression," *Journal of the American Statistical Association*, 81, 108–113. [258]
Grolemund, G., and Wickham, H. (2014), "A Cognitive Interpretation of Data Analysis" (with discussion), *International Statistical Review*, 82, 184–213. [255]
Gunter, B., and Tong, C. (2017), "What Are the Odds!? The 'Airport Fallacy' and Statistical Inference," *Significance*, 14, 38–41. [251,252,257]
Harrell, F. E., Jr. (2015), *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2nd ed.), New York: Springer. [247,256,258]
——— (2018), "Improving Research Through Safer Learning from Data," Statistical Thinking blog [online], available at *http://www.fharrell.com/post/improve-research/* [254,257,258]
Harris, R. (2017), *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*, New York: Basic Books. [253]
Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), New York: Springer. [251,258]
Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [251]
Hjorth, U. (1989), "On Model Selection in the Computer Age," *Journal of Statistical Planning and Inference*, 23, 101–115. [248]
Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (eds.) (1983), *Understanding Robust and Exploratory Data Analysis*, New York: Wiley. [255]
——— (1985), *Exploring Data Tables, Trends, and Shapes*, New York: Wiley. [255]
——— (1991), *Fundamentals of Exploratory Analysis of Variance*, New York: Wiley. [255]
Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [251]
Holmes, S. (2018), "Statistical Proof? The Problem of Irreproducibility," *Bulletin of the American Mathematical Society*, 55, 31–55. [248,251,255]
Huber, P. J. (1985), "Data Analysis: In Search of an Identity," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (Vol. 1), Wadsworth, pp. 65–78. [256]
International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1997), *ICH Harmonised Tripartite Guideline: General Considerations for Clinical Trials*, E8. [249,258]
——— (1998), *ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials*, E9. [249,258]
Kallus, N. (2018), "Optimal *A Priori* Balance in the Design of Controlled Experiments," *Journal of the Royal Statistical Society*, Series B, 80, 85–112. [253]
Koopmans, T. C. (1949), "Identification Problems in Economic Model Construction," *Econometrica*, 17, 125–144. [248]
Koyama, T. (2011), "Dynamite Plots," [online], available at *http://biostat.mc.vanderbilt.edu/wiki/Main/DynamitePlots* [254]
Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I., (2009), "Circular Analysis in Systems Neuroscience: The Dangers of Double Dipping," *Nature Neuroscience*, 12, 535–540. [248]
Krzywinski, M., and Altman, N. (2014), "Visualizing Samples with Box Plots," *Nature Methods*, 11, 119–120. [254]
Kuehl, R. O. (2000), *Design of Experiments: Statistical Principles of Research Design and Analysis* (2nd ed.), Pacific Grove, CA: Duxbury. [253]
Kuhn, T. S. (1970), *The Structure of Scientific Revolutions* (2nd ed.), Chicago, IL: University of Chicago Press. [257]
Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), "Exact Post-Selection Inference, with Application to the Lasso," *Annals of Statistics*, 44, 907–927. [251]
Leeb, H. (2009), "Conditional Predictive Inference Post Model Selection," *Annals of Statistics*, 37, 2838–2876. [251]
Leek, J. T., Scharpf, R. B., Corrada Bravo, H., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010), "Tackling the Widespread and Critical Impact of Batch Effects in High-throughput Data," *Nature Reviews Genetics*, 11, 733–739. [255,256]
Lew, M. (2016), "Three Inferential Questions, Two Types of P-value," Online supplement to Wasserstein & Lazar (2016). [256]
Lewis, J. A. (1999), "Statistical Principles for Clinical Trials (ICH E9) An Introductory Note on an International Guideline," *Statistics in Medicine*, 18, 1903–1904. [258]
Lithgow, G. J., Driscoll, M., and Phillips, P. (2017), "A Long Journey to Reproducible Results," *Nature*, 548, 387–388, available at *https://www.nature.com/news/a-long-journey-to-reproducible-results-1.22478?WT.mc_id=FBK_NatureNews&sf108251523=1* [250]
Lock Morgan, K., and Rubin, D. B. (2012), "Rerandomization to Improve Covariate Balance in Experiments," *Annals of Statistics*, 40, 1263–1282. [253]
McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2017), "Abandon Statistical Significance," [online], available at *https://arxiv.org/abs/1709.07588*. [246]
McShane, B. B., and Gelman, A. (2017), "Abandon Statistical Significance," *Nature*, 551, 558, available at *https://www.nature.com/articles/d41586-017-07522-z* [246]
Mallows, C. L. (1983), "Data Description," in *Scientific Inference, Data Analysis, and Robustness*, eds. G. E. P. Box, T. Leonard, and C.-F. Wu, New York: Academic Press, pp. 135–151. [254]
——— (1998), "The Zeroth Problem," *The American Statistician*, 52, 1–9. [255]
Mallows, C. L., and Walley, P. (1980), "A Theory of Data Analysis?" in *Proceedings of the Business and Economics Statistics Section*, American Statistical Association, pp. 8–14. [254]
Maronna, R., Martin, D., and Yohai, V. (2006), *Robust Statistics: Theory and Methods*, Hoboken, NJ: Wiley. [255]
Mogil, J. S., and Macleod, M. R. (2017), "No Publication Without Confirmation," *Nature*, 542, 409–411, available at *https://www.nature.com/news/no-publication-without-confirmation-1.21509* [249,252,256,259]
Montgomery, D. C. (2017), *Design and Analysis of Experiments* (9th ed.), Hoboken, NJ: Wiley. [253]
Moore, D. S. (1992), "What is Statistics?" in *Perspectives on Contemporary Statistics*, eds. D. C. Hoaglin, and D. S. Moore, Washington, DC: Mathematical Association of America, pp. 1–17. [252]
Moses, L. E. (1992), "The Reasoning of Statistical Inference," in *Perspectives on Contemporary Statistics*, eds. D. C. Hoaglin, and D. S. Moore, Washington, DC: Mathematical Association of America, pp. 107–122. [254]
Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley. [248]
Motulsky, H. J. (2014), "Common Misconceptions About Data Analysis and Statistics," *Journal of Pharmacology and Experimental Therapeutics*, 351, 200–205. [258]
Munafo, M. R., and Davey Smith, G. (2018), "Robust Research Needs Many Lines of Evidence," *Nature*, [online], available at *https://www.nature.com/articles/d41586-018-01023-3* [250]
Murray, D. M. (1998), *Design and Analysis of Group-Randomised Trials* (*Monographs in Epidemiology and Biostatistics*, Vol. 27), New York: Oxford University Press. [253]
Nelder, J. A. (1986), "Statistics, Science and Technology," *Journal of the Royal Statistical Society*, Series A, 149, 109–121. [246,249]
Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018), "The Preregistration Revolution," *Proceedings of the National Academy of Sciences*, 115, 2600–2606. [251,252]
Pearson, R. K. (2011), *Exploring Data in Engineering, the Sciences, and Medicine*, New York: Oxford University Press. [255]
Piantadosi, S. (2017), *Clinical Trials: A Methodologic Perspective* (3rd ed.), Hoboken, NJ: Wiley. [249,255,258]

Picard, R. R., and Cook, R. D. (1984), "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575–583. [246,248,251]

Pikounis, B. (2001), "One-Factor Comparative Studies," in *Applied Statistics in the Pharmaceutical Industry*, eds. S. P. Millard, and A. Krause, New York: Springer, pp. 17–40. [254]

Raftery, A., Madigan, D., and Hoeting, J. (1993), "Model Selection and Accounting for Model Uncertainty in Linear Regression Models," Technical Report No. 262, University of Washington (Seattle), Department of Statistics. [248]

Reunanen, J. (2003), "Overfitting in Making Comparisons Between Variable Selection Methods," *Journal of Machine Learning Research*, 3, 1371–1382. [247]

Robbins, N. B. (2013), *Creating More Effective Graphs*, Wayne, NJ: Chart House. [254]

Rousseeuw, P. J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1273–1283. [254]

Scheiner, L. B. (1997), "Learning Versus Confirming in Clinical Trials," *Clinical Pharmacology and Therapeutics*, 61, 275–291. [258]

Scott, S. (2013), "Pre-Registration Would Put Science in Chains," *Times Higher Education Supplement*, [online], available at *https://www.timeshighereducation.com/comment/opinion/pre-registration-would-put-science-in-chains/2005954.article* [252]

Seife, C. (2000), "CERN's Gamble Shows Perils, Rewards of Playing the Odds," *Science*, 289, 2260–2262. [250]

Shell, E. R. (2016), "Hurdling Obstacles: Meet Marcia McNutt, Scientist, Administrator, Editor, and Now National Academy of Sciences President," *Science*, 353, 116–119. [249]

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366. [247]

Snee, R. D. (1986), "In Pursuit of Total Quality," *Quality Progress*, 20, 25–31. [255]

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Series B, 36, 111–147. [251]

Taleb, N. N. (2007), *The Black Swan: The Impact of the Highly Improbable*, New York: Random House. [254]

Taylor, J. E., and Tibshirani, R. (2015), "Statistical Learning and Selective Inference," *Proceedings of the National Academy of Sciences*, 112, 7629–7634. [251]

The OPERA Collaboration (2012), "Measurement of the Neutrino Velocity with the OPERA Detector in the CNGS Beam," *Journal of High Energy Physics* [online], available at *https://doi.org/10.1007/JHEP10(2012)093* [250]

Thompson, S. K. (2012), *Sampling* (3rd ed.), Hoboken, NJ: Wiley. [254]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [251]

Tong, C., and Lock, A. (2015), "A Computational Procedure for Mean Kinetic Temperature Using Unequally Spaced Data," in *Proceedings of the Biopharmaceutical Section*, American Statistical Association, pp. 2065–2070. [255]

Touboul, P., Metris, G., Rodrigues, M., Andre, Y., Baghi, Q., Berge, J., Boulanger, D., Bremer, S., Carle, P., Chhun, R., et al. (2017), "MICROSCOPE Mission: First Results of a Space Test of the Equivalence Principle," *Physical Review Letters*, 119, 231101. [251]

Tukey, J. W. (1962), "The Future of Data Analysis," *Annals of Statistics*, 33, 1–67. [255,256,257]

——— (1969), "Analyzing Data: Sanctification or Detective Work?," *American Psychologist*, 24, 83–91. [248]

——— (1973), "Exploratory Data Analysis as Part of a Larger Whole," in *Proceedings of the Eighteenth Conference on the Design of Experiments in Army Research Development and Testing* (Part I), Durham, NC: Army Research Office, pp. 1–10. [255]

——— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley. [248,249,254,255]

Wasserman, L. (2006), *All of Nonparametric Statistics*, New York: Springer. [255]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on P-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [246]

Wenmackers, S., and Vanpoucke, D. E. P. (2012), "Models and Simulations in Material Science: Two Cases Without Error Bars," *Statistica Neerlandica*, 66, 339–355. [255,257]

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., and van Assen, M. A. L. M. (2016), "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking," *Frontiers in Psychology*, 7, 1832. [247]

Winer, B. J., Brown, D. R., and Michels, K. M. (1991), *Statistical Principles in Experimental Design* (3rd ed.), Boston, MA: McGraw-Hill. [253]

Youden, W. J. (1972), "Enduring Values," *Technometrics*, 14, 1–11. [250]