Catching up to a 21st century view of statistics in the doing and reporting of research in the Earth sciences

David Jon Furbish[1] and Mark W. Schmeeckle[2]

[1] *Department of Earth and Environmental Sciences, Vanderbilt University, Nashville, Tennessee, USA*

[2] *School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, Arizona, USA*

Rules are not typically a substitute for thinking.

Nancy Cartwright and Jeremy Hardie, *Evidence-Based Policy: A Practical Guide to Doing it Better* (2012)

Abstract. There is a compelling need to re-examine our views and use of statistics in the Earth sciences, and press toward a more informed, measured use of statistical methods in data analysis. This involves moving beyond the false premise that hypothesis testing can be reduced to the dichotomous choice of "significant" or "not significant" decided by arbitrary statistical thresholds, paying increasing attention to the "don'ts" of statistics, and crafting well-reasoned descriptive statistics and analyses with full explanation. Our statistics courses must cover the probabilistic foundation of statistics, not just its applications, giving students the needed insight and thus confidence to critically evaluate their own work as well as what is presented in the literature. Moreover, courses should incorporate philosophical aspects of the theory and practice of statistics. Journals have an important role: to formulate a position and policies on the use and reporting of statistical analyses, thereby providing valuable guidelines for authors and reviewers as well as setting the tone for the expected quality of analyses.

1 Introduction

The world of applied statistics currently is experiencing a rather noisy revolution — what Deborah Mayo, philosopher of science, has dubbed "the statistics wars" (Mayo, 2018) — with far-reaching implications for the doing and reporting of research in the Earth sciences. Two issues are front and center in this revolution. First, frustration is boiling over in reaction to a literature awash in papers that disregard a set of well, and repeatedly, articulated "don'ts" in the use of statistics, with potentially lasting negative impacts in many fields of science. Indeed, "The *ASA* [American Statistical Association] *Statement on p-Values and Statistical Significance* (Wasserstein and Lazar 2016) was developed primarily because after decades, warnings about the don'ts had gone mostly unheeded" (Wasserstein et al., 2019, p. 1). Second, statisticians are carefully reexamining the use of statistics in hypothesis testing, and there is now an open "call for the entire concept of statistical significance to be abandoned" (Amrhein et al., 2019, p. 306; Wasserstein et al., 2019; see also Colquhoun, 2014, 2017; Kennedy-Shaffer, 2019). This is in view of an over-reliance on statistical tests and the false premise that hypothesis testing can be reduced to dichotomous choices concerning significance based on arbitrary statistical thresholds — what Amrhein et al. (2019) call "dichotomania." Whereas this ire in the statistics community currently is focused on the social, behavioral and biological sciences, including medicine, the lessons involved also bear on all natural sciences, including science education.

Many of our colleagues in the Earth sciences who use statistics in their work seem to be unaware of this revolution, despite its noisiness; and we suspect this lack of awareness actually extends far beyond the people with whom we in-

teract. The purpose of this essay therefore is to bring attention to this revolution, and reinforce its basic message with reference to the Earth sciences. There is a clear need to reexamine our formulaic use of statistics and press toward a more informed, measured use of statistical methods in the doing and reporting of analyses.[1] This includes the need for Earth science journals to examine their roles in handling manuscripts that involve statistical analyses. If authors of such manuscripts do not adopt a more measured use of statistical methods, there is the real possibility that some if not many journals will eventually lock step in forcing the point.[2] Herein we summarize the sources of discord, comment on the teaching of statistics in Earth sciences as an essential element of 21st century science, and then briefly consider the role of journals. We finish with five recommendations.

2 Elements of discord

2.1 The perils of p-values

Foremost among the practices that frustrate statisticians are the misinterpretation and misuse of p-values in relation to the idea of significance in hypothesis testing — what Wasserstein et al. (2019) call the "perils of p-values." This begins with the misinterpretation of a p-value calculated from a sample in which, when compared with a so-called critical p-value, the condition $p < 0.05$ is used "to indicate that a certain empirical result has a statistical seal of approval" (Gelman, 2019). For example, it is incorrect to interpret the value 0.05 in terms of the likelihood that the alternative hypothesis is true following rejection of the null hypothesis; in fact this is just the probability of observing a result that is more extreme than that associated with $p = 0.05$ if the null hypothesis is correct. As Gelman notes, "To say that $p = 0.05$ should lead to acceptance of the alternative hypothesis is tempting — a few million scientists do it every year. But it is wrong, and has led to replication crises in many areas..."

More generally, Wasserstein et al. (2019) suggest that the idea of "statistical significance" has become meaningless, and that this phrase should be dropped entirely, "nor should variants such as "statistically different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way." They note that the original intent of this idea as proposed by Edgeworth (1885) was to indicate a result worthy of further scrutiny — that "statistical significance was never meant to imply scientific importance... [yet] the confusion of the two was decried soon after its widespread use (Boring 1919)... [and] a full century later the confusion persists." Although presented as a don't (Wasserstein et al., 2019), this is a call to carefully reexamine our use of statistics in hypothesis testing, and aim at approaches that are more measured than current procedures centered on testing for "significance" using arbitrarily prescribed thresholds such as $p < 0.05$. Frustration with, and criticisms of, the de facto standardization of this approach have been around for decades (Ziliak and McCloskey, 2008; Colquhoun, 2014, 2017; Kennedy-Shaffer, 2019), but only recently has this frustration been vented in the collective voice of the statistics community as summarized in the first paragraph of this essay.

At the heart of the matter is the false dichotomous choice. From Wasserstein et al. (2019, p. 2):

> "For example, no p-value can reveal the plausibility, presence, truth, or importance of an association or effect. There-

[1] It is important to acknowledge at the outset that the use of statistics and statistical methods varies widely across disciplines in the Earth sciences, reflecting differences in the types of problems and data sets. Elements of this essay thus have differing relevance across these disciplines.

[2] Indeed, this top-down process has started. At least one major social science journal, *Basic and Applied Social Psychology*, no longer publishes papers involving the "null hypothesis significance testing procedure" and associated reporting of "p-values, t-values, F-values, statements about "significant" differences or lack thereof, and so on" (Trafimow and Marks, 2015, p. 1). The journal *Estuaries and Coasts* now instructs authors to avoid the idea of "statistical significance" and points to the essay by Smith (2020).

> fore, a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important. Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant. Yet the dichotomization into "significant" and "not significant" is taken as an imprimatur of authority on these characteristics... [and as] Gelman and Stern (2006) famously observed, the difference between "significant" and "not significant" is not itself statistically significant."

The problems and issues associated with statistical procedures centered on this dichotomization in hypothesis testing are now being spelled out in a quickly growing literature. For those not familiar with the brouhaha, we recommend starting with the Comment by Amrhein et al. (2019) that recently appeared in *Nature* (567, 305–307) entitled, "Retire statistical significance," then back step to the Editorial by Wasserstein et al. (2019) that appeared in *The American Statistician* (73, 1–19) entitled, "Moving to a world beyond "$p < 0.05$"," then back step again to the article by Wasserstein and Lazar (2016) that appeared in *The American Statistician* (70, 129–133) entitled, "The ASA's statement on p-values: context, process, and purpose." Add to this the assessment of the state of affairs provided by the 2019 National Academies report focused on the social and behavioral sciences, "Reproducibility and Replicability in Science," which does not mince words. These items, together with the literature cited within them, provide an informed overview of the problems and issues, and of what happens next. Kennedy-Shaffer (2019) provides a thoughtful historical perspective on the current state of affairs. For our part, the next steps involve reading key papers in the statistics literature, engaging with statisticians, and deep thinking on insightful non-formulaic uses of statistics in data analysis. This cannot be about letting the statisticians fight it out and then tell us what to do when the dust settles while we meanwhile do business-as-usual statistics.

2.2 Basic examples

There may be merit in a separate, systematic listing of don'ts pertaining to analyses in the Earth sciences, paralleling those that Wasserton and Lazar (2016) and Wasserton et al. (2019) refer to, as some don'ts in the Earth sciences require specific context to go with explanation.[1] In this vein, students should be offered specific examples from the literature regarding what not to do with statistics as a standard part of our courses and research group interactions.[3] However, for the purpose of this essay — and while recognizing that don'ts don't solve the problem — here we offer a few generic examples, ranging from silly to serious, to provide a sense of the attention needed.

Examples among the silly include: referring to random variables (or variables more generally) as parameters; reporting estimated values of parameters and coefficients to more significant digits than those of the measured quantities; referring to natural variability in a random variable as "error" (e.g., as in an "error bar"); plotting (discrete) probability mass functions as continuous functions; misinterpreting a (continuous) probability density as a probability without regard to the dimensions of the density; and referring to smoothed, non-normalized histograms as probability density functions.

Front and center among the serious is the frequently offered misinterpretation of a so-called confidence interval to describe the uncertainty associated with a sample estimate of a population parameter, for example, the population mean. A 90% confidence interval does *not* imply that there is a 90% chance that the population parameter falls within the calculated interval. Instead, this interval describes the uncertainty associated with a sampling method. A 90% confidence interval represents the proportion of a great number of imagined independent samples

[3] This pedagogical element is front and center in the 2020 book by C. Berstrom and J. West, *Calling Bullshit*, and in their college course with the same title: https://www.callingbullshit.org/syllabus.html

whose calculated intervals would include the parameter, assuming homogeneity in the process. A confidence interval provides no information regarding the proximity of the sample estimate to the population value. Moreover, if by chance a calculated interval does not include the true value, then the true value could be just outside the interval or it could be far outside the interval. One simply cannot know this. We further suggest that thoughtful assessment and estimation of sampling error should be a standard part of analyses, paying attention to such things as sample size, possible sources of bias, the possibility of mixtures of populations, etc.

We must avoid reporting p-values based on an assumed sampling distribution (e.g., the Student's t-distribution) when the data are insufficient to justify the distribution or when the assumptions involved are demonstrably not met. For example, many geophysical quantities are for physical reasons *not* normally distributed, a condition of some parametric tests.[4] In many situations our sample data do not satisfy the basic, cornerstone assumption of being independent and identically distributed, and we suspect that efforts aimed at assessing these qualities and then, for example, removing effects of serial correlation with resampling, are uncommon. That a data set does not satisfy the assumptions of standard inferential statistics should not necessarily be viewed as a weakness. Rather, having a clear sense of one's data, including that they do not satisfy oft used assumptions, is a position of strength for electing to conduct specific analyses and report certain statistics instead of defaulting to formulaic procedures out of a sense of unease, and "because that's what other people did."

In offering formulations of how quantities are functionally related, we must pay attention to the idea that the average of a product is not necessarily equal to the product of the averages, and that exponentiation of a random variable gives a new random variable with altogether different moments. For example, if we propose that the random variable z is a function of the random variables x and y according to, say, $z \sim x^\alpha$, $z \sim xy$ or $z \sim x/y$, the temptation is strong to assume that such relations hold with expected values, say, $\overline{z} \sim \overline{x}^\alpha$, $\overline{z} \sim \overline{x}\,\overline{y}$ or $\overline{z} \sim \overline{x}/\overline{y}$, which is generally incorrect.[5] The algebra of expected values (and higher moments) of random variables differs from the algebraic manipulation of deterministic quantities.

We should avoid the fitting of sample data to probability distributions using automated goodness-of-fit algorithms that search and compare various distribution forms[6] with little regard to the principle of parsimony or to the physical basis for the choice of distribution, and absent careful examination of the tail behavior (e.g., using quantile-quantile and exceedance probability plots). Similarly, we must avoid being quick to claim the existence of a heavy-tail distribution based on small data sets or, worse, tail censured data. We must avoid overinterpreting histograms of relatively small data sets in distinguishing "distinct" modes that may or may not represent mixed distributions, but instead reflect happenstance or non-random sampling that gives the appearance of more than one mode.

We must avoid the algebraic inversion of an ordinary least-squares regression equation, linear or nonlinear, relating y to x in order to predict x from y. We must stop the indefensible practice of arbitrarily designating (and excluding) data points as so-called "outliers" without clear physical reasons and without subjecting all retained data to the same criteria used to identify outliers, or worse, using canned algorithms

[4] Commentary on parametric versus nonparametric statistics and methods is far beyond the scope of this essay. The point here pertains only to p-values and the conditions assumed by some tests.

[5] For example, if $z = Ax^\alpha$ with coefficient A, then the expected value $\overline{z} = A\overline{x^\alpha}$. Similarly, if $z = Axy$, then $\overline{z} = a\overline{xy} = A\overline{x}\,\overline{y} + A\mathrm{cov}(x, y)$. If $z = Ax/y$, then $\overline{z} = A\overline{xw} = A\overline{x}\,\overline{w} + A\mathrm{cov}(x, w)$ with $w = 1/y$. Random variables formed as reciprocals of random variables must be treated carefully, as their behavior can be pathological. But see Furbish et al. (2021a) for an example of the treatment of such a reciprocal.

[6] Such algorithms typically involve different options for fitting, for example, maximum likelihood estimation or regression of the empirical cumulative distribution. But some methods may not be suited to specific distribution forms, and censured data can be problematic.

to "choose" outliers based on non-physical statistical criteria.[7] We must not be quick to dismiss the possible information contained in a relation between variables with a moderately strong correlation r simply because the calculated p-value is "too large," or conversely, claim that such a relation with a minute correlation merits special attention because the p-value is tiny — when in fact its smallness is merely attributable to large N. Significance and so-called statistical significance are *not* the same. We know to avoid using regression to estimate quantities (e.g., intercepts) that are physically impossible; we know to avoid reporting the strength of a least-squares fit (e.g., using r^2 or p-values) when the data represent a spurious or induced correlation (e.g., an empirical cumulative distribution function);[8] we know to interpret residual variance in semi-log or log-log fits differently than that associated with the original arithmetic space; and we know to avoid being quick to claim discovery of a "scaling" relation from data that span a limited domain and range (Korup et al., 2012), notably in the absence of an underlying theory (Stumpf and Porter, 2012).

To this we add an observation. We sometimes seem to be quick to collect data. This can involve established, standardized procedures with ample justification and prior attention to appropriate analyses of the data — which is a good thing. But in the absence of established procedures, our efforts could well benefit from thoughtful upfront consideration of sampling strategy and design — with anticipated analyses in mind — while recognizing that adjustments in sampling might be desirable as data are collected (Wilson et al., 2021). This is particularly important when, for example, time or budgetary constraints limit the sampling effort (Furbish et al., 2018). Adaptive sampling may be particularly valuable in field-based and experimental work where real-time data evaluation is possible, for example, assessing convergence in the variability of measured quantities.

We stop here with this sketch of the issues involved, noting that these examples represent just the basics. Similar concerns may carry over into more advanced analyses, including time series and geospatial analysis, and problems where resampling methods or Bayesian analyses are the appropriate choices. Meanwhile, there is merit in pausing, backing up, and adopting a more critical view of what we are doing with statistics and why.

2.3 Misplaced significance

Reconsider Gelman's comment centered on "...a statistical seal of approval." This phrase in fact implicates a troublesome mindset with which we all are familiar — that statistical tests in support of an analysis, including statements regarding "significance," are viewed as lending a certain credibility or implied rigor that otherwise would not exist in the absence of the tests. This mindset is manifest in numerous ways. Here are two examples, starting with the manuscript review process.

We frequently hear anecdotes from colleagues, and have stories of our own, that involve being asked by reviewers or editors to unnecessarily conduct analyses aimed at providing "significance levels" or related quantities in order to "strengthen" our conclusions. Bacchetti (2002) also raises the concern that reviewers often complain about sound statistical methods because they are operating from a mindset of finding fault with manuscripts and because the reviewers' limited understanding of statistics leads them to blindly apply rules and reject analyses that may have perfectly good reasons for not following those rote rules. Notably egregious examples include being asked to: report r^2 values or p-values, where in fact calculating such values would be precisely the incorrect thing to do; perform parametric tests on data that do not satisfy the required assumptions; perform perfunctory regression on data that already are plotted with theoretical curves; fit data to a specific distribution without a physical justification for doing so; and justify in statistical terms the colloquial use

[7] That such algorithms are available in software packages, for example R, does *not* justify their use.

[8] The text by Davis (2002) contains examples of geophysical data that involve spurious or induced correlations. Numerous online resources also are available.

of the word "significant."[9]

This mindset spills over into the writing of research proposals. For example, we now see null hypotheses offered (evidently) to give the impression of statistical credibility in the research design. Yet a so-called null that is merely a negative existential statement of what is expected from the measurements and analysis serves no meaningful purpose.[10] Such presentations unfortunately reinforce the impression that dichotomous hypothesis testing is viewed as an "imprimatur of authority" — that scientific discovery should involve answering a series of yes-no questions adjudicated by statistical tests.[11] Clearly stated, reasoned hypotheses certainly are desirable,[12] together with statements regarding any caveats and anticipated problems. Equally desirable are descriptions of anticipated approaches to data analysis wherein statistical analyses, if appropriate, are considered part of the research (e.g., experimental) design at the outset — with the recognition that elements of the project, including analyses, are likely to change as the research progresses.

We must move away from the misconception that dichotomous statistical tests, associated measures of "significance" or "confidence," or hollow hypothesis statements, lend credibility or implied rigor that supersedes thorough data analysis, with or without statistics. Let us instead use statistics for what they are intended — as one tool to explore data in the process of discovery and understanding of natural phenomena. Statistics certainly have importance in data analysis, but statistics in themselves do not imply gravitas in reasoning.

3 Elements of education

3.1 The Probability Foundation

Amidst the noise, one can discern nascent melodic chords devoted to education. Indeed, our experience suggests that an important part of the problem, and its eventual solution, resides in how we teach statistics to our students — whether we offer courses in our own Earth science programs or ask our students to take courses in other programs, for example, biostatistics courses (which generally are decidedly not geared toward many Earth science problems, and understandably so). Specifically, and judging from numerous conversations with student and faculty colleagues, we typically make the mistake of glossing over the probabilistic underpinnings of statistics, and instead offer a formulaic treatment of statistical methods, perhaps punctuated with clever "tricks" that we have learned along the way concerning specific aspects of statistics.[13] As a consequence our students lack the deeper understanding of statistics, accessed through an understanding of probability, that is necessary to critically evaluate their own work as well as what is presented in the literature. We end up propagating the same mistakes and misinterpretations offered in the literature, perhaps assuaging any tickles of discomfort with what we read by subscribing to the imaginary rule that if it appeared in the literature following peer review, it must be right — or at least is not fatally flawed. Here is an example, returning to the idea of p-values, to illustrate the point.

Some students who take our (Vanderbilt) graduate course on probability and statistics in Earth sciences previously have taken one or more

[9] Perhaps an eventual, nice outcome of the revolution will be that we rescue the word "significant" for its intended ordinary use.

[10] An hypothesis structure offered with the null H_0 normally invokes a formal statistical test. Moreover, the juxtaposition of such a null with a statement of expected results, nominally the alternative hypothesis H_1, signals the intent to claim the alternative as being correct if the null is rejected — a statistical no-no.

[11] We should not imagine that problems in the Earth sciences parallel the degree of precision demanded in particle physics that arises from the tight coupling between theory and experiments leading to "five sigma" testing; see Section 3.1.

[12] It is worth noting that a formulation of how a system works, expressed mathematically, is a formal hypothesis. Depending on its structure, constructing a meaningful null may not be straightforward.

[13] A colleague suggests that the demise of probability in our courses started with the birth of software packages like SPSS, and that statistics courses centered around learning such packages, for example R, risk reinforcing the illusion that their use can replace critical thinking.

courses on statistics. Yet these students, and indeed, many seasoned researchers who make use of the idea of p-values in appealing to so-called significance testing, often cannot articulate what a p-value actually is[14] — that it is an exceedance probability given by the sampling distribution of a random variable formed as a sample statistic. To confound matters, folks often do not know what a sampling distribution is, nor the assumptions that go into its creation. Unfortunately, online resources that immediately place p-values within the context of hypothesis testing and the idea of critical p values do not help matters. This approach is both misleading and counterproductive, as it does not show the probabilistic basis of how a p-value is obtained nor the decided merits of p-values outside of the context of standard procedures of hypothesis testing. To be clear, the statistics community is not asking us to abandon p-values. Exceedance probabilities (i.e., p-values) in fact can be an important element of solid statistical analysis. Rather, we are being asked to toss the idea of so-called critical p-values in relation to the misguided view that hypothesis testing must somehow involve a dichotomous outcome as expressed in words such as "significant" versus "not significant," and "reject" versus "cannot reject."[15]

Here a brief lesson from physics is useful. Much is made of the "five sigma" criterion used in physics as a threshold for declaring the "discovery" of a new phenomenon, for example, the Higgs boson in 2012 and more recently in relation to the Fermilab Muon g-2 experiment (Wolchover, 2021a). This five sigma indeed corresponds with a p-value, equal to about 3×10^{-7}, which physicists have agreed must be reached before a discovery can be claimed. The sampling distribution is based on detailed theory of expected particle behavior, accounting for all known sources of variability in behavior (consider Borsanyi et al., 2021). The tight coupling between theory and experiments in this situation is *not* what we have in the Earth sciences. Our systems generally are far messier, our sampling distributions pertain to geophysical quantities that imprecisely characterize system behavior, and our data sets typically are tiny relative to those generated at the Large Hadron Collider and the Fermilab. Moreover, to be clear, physicists do not claim that the Higgs boson *does* exist. Rather, they claim only that there is a small probability that the Higgs boson does *not* exist. The five sigma value "is the probability that if the particle does not exist, the data that CERN scientists collected in Geneva, Switzerland, would be at least as extreme as what they observed" (Lamb, 2012). The choice of five sigma comes from having experienced three sigma "discoveries" that later were demonstrated to be effects of statistical variability, not new phenomena. The idea of "statistical significance" in this context is irrelevant, and we should not appeal to this as an example in support of retaining dichotomous testing based on p-values.

Concerning the topics above, and more, numerous texts, papers and essays are available as primary or secondary resources that systematically develop the probabilistic foundation of statistics, not just its applications. In the ideal this material should include a calculus based treatment of applicable topics,[16] for example, the properties of probability distributions and

[14] This seems to be widespread. Consider the FiveThirtyEight posting by C. Aschwanden entitled, "Not Even Scientists Can Easily Explain P-values" (https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/). Moreover, a recent viral exchange on Twitter involved a famous scientist asking what folks recommended for essential content in a second statistics course. One serious response in effect suggested the topic: No, that is not what a p-value is.

[15] As a counterpoint, Ioannidis (2019) suggests that "[s]tatistical significance sets a convenient obstacle to unfounded claims... [such that] removing the obstacle could promote bias," and among others advocates more stringent thresholds of significance, where "rules are set before data collection and analysis." This is aimed, however, at issues of replication in the fields of medicine and social sciences, not the Earth sciences, and it does not address the misuse of the concept of "significance."

[16] If this is perceived as not being an option for math anxious students, we recommend examining the classic 1910 text by S. P. Thompson, *Calculus Made Easy*, which is freely accessible (https://calculusmadeeasy.org/). The entry of this book in the References below includes its humorous self-deprecating Prologue.

related functions, emphasizing the physical interpretation of the mathematics for selected distributions, continuous and discrete. Similarly, this material should cover the *probabilistic basis* of selected methods, for example, the generation of sampling distributions, resampling techniques, correlation and regression (and related methods), the ensemble interpretation of time and space series analysis, basic elements of Monte Carlo methods, error propagation, etc.[17] We suggest that such a foundation offers to our students a position of strength, where they can engage "...more fully [in] critical thinking rather than rote applications of formulae... [particularly] as statisticians and scientists consider a world beyond $p < 0.05$" (Kennedy-Shaffer, 2019).

To this we add an observation. Our experience suggests that sometimes we are quick to encourage students to use sophisticated techniques — perhaps in part because these techniques are readily available and quickly implemented with statistics packages — without offering the students the opportunity to gain a full understanding of the foundational elements of the techniques. Similarly, we sometimes are quick to adopt statistical measures presented in the literature without critically assessing what they are actually measuring, or whether they are the right choice for our work. We must instead aim at helping ensure that, when our students are in situations of presenting their work, they can capably describe — with confidence — the basis and justification of the statistical analyses they are conducting.

3.2 The philosophy part

We further suggest that a deeper epistemological issue is involved which deserves explicit attention in the teaching of statistics. This begins with the observation that mathematical knowledge is cumulative, whereas knowledge in all other fields, science and non-science, is evolutionary. Students fully appreciate this point. However, because statistics is quantitative, and because we generally associate quantitativeness with the practice of mathematics, there is the real risk that the "knowledge" of statistics will be conflated with the knowledge of mathematics — that this knowledge is cumulative and does not change its mind, and that the "rules" of doing statistics are like the unchanging rules of mathematics. Yet views on the meaning and use of statistics change, as suggested by the material in the earlier sections of this essay. Recent statistics texts are decidedly different in approach and presentation from texts of the late 20th century. This reflects that the field of statistics is not static nor just about applied mathematics and methods. It is deeply philosophical. For example, consider what the clever 20th century logician and philosopher Bertrand Russel had to say about probability:

> "Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means." (see citation of Bell, 1945 in Hájek, 2019)

Indeed, the topic begs a deep discussion of what probability and statistics are and are not, and this discussion must be part of the fabric of courses in statistics. Here are a few examples.

Students need to know that mathematicians, scientists and philosophers have for centuries strived to clarify the very idea of probability, whose interpretations, wrapped up in ideas of countable and uncountable sets (Wolchover, 2021b), continue to be debated (Hájek, 2019; Wallace, 2012). Students should fully grasp classical versus frequentist interpretations of probability, how these interpretations are related, and how they are connected with Bayesian concepts (see Hájek, 2019). Students should know that the principle of indifference — the foundation of the classical interpretation of probability first formalized by Laplace and Bernoulli — has been both naïvely and rigorously contested (Shackel, 2007; Norton, 2008). Students should be exposed to so-called paradoxes in probability and

[17] Our one-semester course at Vanderbilt involves using numerous simple numerical codes specifically designed to illustrate and exercise most of the mathematical concepts and methods covered, including the "don'ts" mentioned above.

statistics (Bertrand, 1889; Jaynes, 1973; Marinoff, 1994; Rowbottom, 2013). They should be engaged in discussions on the principle of parsimony (or Occam's razor) as a philosophical style of thinking and as a technical guide in statistical analyses. Add to this the closely related idea of remaining as faithful to what is *not* known as we are to what *is* known about a problem (Jaynes, 1957), and how this bears on the idea of a *prior* in Bayesian analysis (Jaynes, 1988). Students deserve honest discussions of sampling and confirmation biases, and of what is meant by an *outlier*, including the possibility that this idea is fictitious. Students should be asked to ponder the perhaps counterintuitive idea that one of the best uses of statistics is to collect data for which one does not need to calculate statistics. More broadly, students deserve exposure to ideas of hypothesis testing (or falsification) as viewed through the lens of critical rationalism espoused by Karl Popper and others. They deserve discussions (with demonstrations) of why observations are *theory-laden* and the practical implications of this. Students deserve to engage with, for example, Isaac Asimov's views on the *relativity of wrong* and the meaning of *hard-to-vary explanations* as articulated by David Deutsch in relation to our use of statistics in pursuing scientific discoveries. Indeed, the list of possible topics is a cornucopia. The key point is this: to present probability and statistics as rote methods absent their rich philosophical and historical backdrop is a lost opportunity for our students (Kennedy-Shafer, 2019).

4 Role of journals

Journal editors and editorial boards have an important role in helping move the community toward a more measured view and use of statistics in the doing and reporting of research. The first part of this role is to examine the issues surrounding the use (and misuse) of statistics, check the pulse of community thinking on these issues, and communicate in these terms with editors of journals associated with other professional societies. This effort informs the second part, namely, formulating a journal position on the use and reporting of statistical analyses, with full explanation of the position chosen. We suggest that, at a minimum, stated journal policies on the use and reporting of statistical analyses can provide valuable guidelines for contributing authors and reviewers, as well as set the tone for the expected quality of analyses. Journal policies then can be periodically revisited as the science moves beyond $p < 0.05$.

In moving beyond familiar formulaic statistical analyses, we need to anticipate a lengthening of papers, perhaps by judiciously expanding the use of appendixes and supplementary materials.[18] This may well require that editors and editorial boards critically reexamine the current fashionable trend of shortening papers. Currently the reporting of, say, a significance test is reasonably straightforward and requires relatively little space because of the standardization of such methods. The future will increasingly involve thorough descriptive statistics and analyses with all assumptions and steps clearly stated and described, with appropriate documentation, in support of interpretations and conclusions presented. This effort must include attention to reproducibility, consistent with growing expectations for making data sets and specialized numerical codes readily accessible to the community.

5 Recommendations

Here are five recommendations deriving from the ideas presented above.

1) For those of us who use statistics in our research and find this essay unsettling — that the matters of using basic statistics are unsettled — it is essential to learn what the brouhaha is all about and what the key issues are. The editorials and essays described above, which are aimed at scientists, not just statisticians, are a great start. Know that, while it might appear that the "...old, rotting timbers... holding up the edifice of modern scientific research..." are

[18] For example, see Appendix A in Furbish et al. (2021b).

being torn out, folks are simultaneously "...offering solid construction materials to replace them" (Wasserstein et al., 2019, p. 1). All will be well.

2) We must avoid focusing on the misguided idea of "statistical significance" and its variants. It is essential to recalibrate our thinking and focus on thoughtfully crafted non-formulaic presentations of the statistics of scientific results. As suggested by Amrhein et al. (2019, p. 307), "[d]ecisions to interpret or to publish results will not be based on statistical thresholds. People will spend less time with statistical software, and more time thinking."

3) In contributing work to journals, do not be shy during the review process about pushing back on reviewers and editors if they suggest or insist that you incorporate unnecessary or unjustified statistical analyses centered on business-as-usual hypothesis or significance testing.[19] Bring to their attention the editorials, comments and papers cited here (as well as other relevant literature) in support of your push back, and defend your position thoughtfully. This of course means that any statistics you do present must be justified and thorough. Conversely, it is essential to be open to insightful perspectives and recommendations for data analysis.

4) Because of the importance of statistics in the doing of science, a rigorous course on this topic is an essential element of a 21st century curriculum in the Earth sciences. Let us move away from formulaic treatments and methods, and instead invest in laying the probabilistic foundation that allows students to take a deeper intellectual ownership of the material, growing their ability to critically assess their own work and what is presented in the literature, and successfully navigate a world beyond $p < 0.05$. Moreover, explicit coverage of the philosophical aspects of probability and statistics must be part of the fabric of our courses.

5) Editors of Earth science journals should, as an explicitly stated policy following a grace period, reject manuscripts that present material in terms of "statistical significance" or variants of this idea, or present the results of hypothesis testing in terms of dichotomous choices set by arbitrary statistical thresholds — and instead insist on presentation of solid, thorough descriptive statistics and analysis, with all assumptions and steps clearly stated and described. Similarly, program officers of research funding agencies should highlight the merits of incorporating any anticipated statistical analyses into the research design at the outset, while acknowledging the likely evolution of project details, including the types of analyses used.

***

Numerous insightful recommendations, many at a very practical level, are presented in the literature cited above and the references listed therein. And there is of course value in engaging with statisticians interested in Earth science problems. Such engagements can be particularly valuable in relation to experimental and sampling design. Our own journeys have been piecemeal efforts to learn the theory and practice of probability and statistics in relation to our teaching and research.

Acknowledgements. We appreciate thoughtful discussions with colleagues concerning the topics in this essay, notably Rachel Bain, Brandt Gibson and Jonathan Gilligan, who helped with wording in several places — although the views expressed are ours.

[19] Colleagues point out that doing this might be easier for established scientists than for early career scientists, particularly students — that the process of navigating reviewer and editor reactions and expectations can be frustrating if not intimidating. And we know of cases in which authors have acquiesced to unreasonable reviewer and editor "demands," despite the authors' better judgment. This is unfortunate, and reflects larger issues that can include situations where reviewers and editors misunderstand their roles and responsibilities in the review process. Nonetheless, our experience suggests that dedicated editors are open to well-reasoned arguments with the intention of getting things right. This topic merits a separate essay.

References

[1] Amrhein, V., Greenland, S., and McShane, B. 2019. Retire statistical significance, *Nature*, 567, 305–307.

[2] Bacchetti, P. 2002. Peer review of statistics in medical research: the other problem, *British Medical Journal*, 324, 1271–1273.

[3] Bell, E. T. 1945. *The Development of Mathematics*, 2nd edition, New York, McGraw-Hill Book Company.

[4] Bergstrom, C. T. and West, J. D. 2020. *Calling Bullshit: The Art of Skepticism in a Data-Driven World*, Penguin Random House, New York.

[5] Bertrand, J. 1889. Calcul des probabilités, Gauthier-Villars, p. 5–6.

[6] Boring, E. G. 1919. Mathematical vs. scientific significance, *Psychological Bulletin*, 16, 335–338.

[7] Borsanyi, S., Fodor, Z., Guenther, J. N., Helbling, C., Katz, S. D., Lellouch, L., Lippert, T., Miura, K., Parato, L., Szabo, K. K., Stokes, F., Toth, B. C., Torok, Cs., and Varnhorst, L. 2021. Leading hadronic contribution to the muon magnetic moment from lattice QCD, *Nature*, 593, 51–55. https://doi.org/10.1038/s41586-021-03418-1

[8] Cartwright, N. and Hardie, J. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*, Oxford University Press.

[9] Colquhoun, D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values, *Royal Society Open Science*, 1, 140216, doi: doi: 10.1098/rsos.140216.

[10] Colquhoun, D. 2017. The reproducibility of research and the misinterpretation of p-values, *Royal Society Open Science*, 4, 171085, doi: 10.1098/rsos.171085.

[11] Davis, J. C. 2002. *Statistics and Data Analysis in Geology* (3rd Edition), John Wiley & Sons, New York, ISBN: 978-0-471-17275-8.

[12] Edgeworth, F. Y. 1885. Methods of statistics, *Journal of Statistical Society, London*, Jubilee Volume, 181–217.

[13] Furbish, D. J., Roering, J. J., Doane, T. H., Roth, D. L., Williams, S. G. W., and Abbott, A. M. 2021a. Rarefied particle motions on hillslopes – Part 1: Theory, *Earth Surface Dynamics*, 9, 539–576, https://doi.org/10.5194/esurf-9-539-2021.

[14] Furbish, D. J., Schumer, R., and Keen-Zebert, A. 2018. The rarefied (non-continuum) conditions of tracer particle transport in soils, with implications for assessing the intensity and depth dependence of mixing from geochronology, *Earth Surface Dynamics*, 6, 1169–1202, https://doi.org/10.5194/esurf-6-1169-2018.

[15] Furbish, D. J., Williams, S. G. W., Roth, D. L., Doane, T. H., and Roering, J. J. 2021b. Rarefied particle motions on hillslopes – Part 2: Analysis, *Earth Surface Dynamics*, 9, 577–613, https://doi.org/10.5194/esurf-9-577-2021.

[16] Gelman, A. 2019. Of chaos, storms and forking paths: the principles of uncertainty, *Nature*, 569, 628–629.

[17] Gelman, A. and Stern, H. 2006. The difference between 'Significant' and 'not significant' is not itself statistically significant, *The American Statistician*, 60, 328–331.

[18] Hájek, A. 2019. Interpretations of Probability, The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (Ed.), http://plato.stanford.edu/archives/fall2019/entries /probability-interpret/.

[19] Ioannidis, J. P. A. 2019. Retiring statistical significance would give bias a free pass, *Nature*, 567, 461.

[20] Jaynes, E. T. 1957. Information theory and statistical mechanics, *Physical Review*, 106, 620–630.

[21] Jaynes, E. T. 1973. The well-posed problem, *Foundations of Physics*, 3, 477–493.

[22] Jaynes, E. T. 1988. The relation of Bayesian and maximum entropy methods, in *Maximum-Entropy and Bayesian Methods in Science and Engineering*, vol. 1, edited by G. J. Erickson and C. R. Smith, pp. 25–29, Kluwer Acad., Dordrecht, Netherlands.

[23] Jaynes, E. T. 1988. The relation of Bayesian and maximum entropy methods, in *Maximum-Entropy and Bayesian Methods in Science and Engineering*, vol. 1, edited by G. J. Erickson and C. R. Smith, pp. 25–29, Kluwer Academic, Dordrecht, Netherlands.

[24] Kennedy-Shaffer, L. 2019. Before $p < 0.05$ to beyond $p < 0.05$: using history to contextualize p-values and significance testing, *The American Statistician*, 73, 82–90.

[25] Korup, O., Görüm, T., and Hayakawa, Y. 2012. Without power? Landslide inventories in the face of climate change, *Earth Surface Processes and Landforms*, 37, 92–99.

[26] Lamb, E. 2012. 5 Sigma What's That, *Scientific American*, https://blogs. scientificamerican.com/observations /five-sigmawhats-that/

[27] Marinoff, L. 1994. A resolution of Bertrand's paradox, *Philosophy of Science*, 61, 1–24.

[28] Mayo, D. G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, Cambridge.

[29] National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*, Washington, DC, The National Academies Press, https://doi.org/10.17226/25303.

[30] Norton, J. D. 2008. Ignorance and indifference, *Philosophy of Science*, 75, 45–68.

[31] Rowbottom, D. P. 2013. Bertrand's paradox revisited: Why Bertrand's 'solutions' are all inapplicable, *Philosophia Mathematica*, 21, 110–114.

[32] Shackel, N. 2007. Bertrand's paradox and the principle of indifference, *Philosophy of Science*, 74, 150–175.

[33] Smith, E. P. 2020. Ending reliance on statistical significance will improve environmental inference and communication, *Estuaries and Coasts*, 43, 1–6, https://doi.org/10.1007/s12237-019-00679-y

[34] Stumpf, M. P. H. and Porter, M. A. 2012. Critical truths about power laws, *Science*, 335, 665–666, https://doi.org/10.1126/science.1216142.

[35] Thompson, S. P. 1910. *Calculus Made Easy: Being a Very-Simplest Introduction to Those Beautiful Methods of Reckoning which are Generally Called by the Terrifying Names of the Differential Calculus and the Integral Calculus*, Macmillan Company, New York.

> **Prologue** [from Thompson's text]
>
> Considering how many fools can calculate, it is surprising that it should be thought either a difficult or a tedious task for any other fool to learn how to master the same tricks.
>
> Some calculus-tricks are quite easy. Some are enormously difficult. The fools who write the textbooks of advanced mathematics — and they are mostly clever fools — seldom take the trouble to show you how easy the easy calculations are. On the contrary, they seem to desire to impress you with their tremendous cleverness by going about it in the most difficult way.
>
> Being myself a remarkably stupid fellow, I have had to unteach myself the difficulties, and now beg to present to my fellow fools the parts that are not hard. Master these thoroughly, and the rest will follow. What one fool can do, another can.

[36] Trafimow, D. and Marks, M. 2015. Editorial, *Basic and Applied Social Psychology*, 37, 1–2.

[37] Wallace, D. 2012. *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*, Oxford University Press, Oxford.

[38] Wasserstein, R. and Lazar, N. 2016. The ASA's statement on p-values: context, process, and purpose, *The American Statistician*, 70, 129–133.

[39] Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. 2019. Moving to a world beyond "$p < 0.05$", *The American Statistician*, 73, 1–19.

[40] Wilson, C. G., Qian, F., Jerolmack, D. J., Roberts, S., Ham, J., Koditschek, D., and Shipley, T. F. 2021. Spatially and temporally distributed data foraging decisions in disciplinary field science, *Cognitive Research: Principles and Implications*, 6, 29, https://doi.org/10.1186/s41235-021-00296-z.

[41] Wolchover, N. 2021a. 'Last hope' experiment finds evidence for unknown particles, *Quanta Magazine*, https://www.quantamagazine.org/last-hope-experiment-finds-evidence-for-unknown-particles-20210407/

[42] Wolchover, N. 2021b. How many numbers exist? Infinity proof moves math closer to an answer, *Quanta Magazine*, https://www.quantamagazine.org/how-many-numbers-exist-infinity-proof-moves-math-closer-to-an-answer-20210715/

[43] Ziliak, S. T. and McCloskey, D. N. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, University of Michigan Press, Ann Arbor.