Inferential Statistics are Descriptive Statistics

Valentin Amrhein[1,2], David Trafimow[3] and Sander Greenland[4]

19 July 2018

[1]Zoological Institute, University of Basel, Basel, Switzerland. [2]Swiss Ornithological Institute, Sempach, Switzerland. [3]Department of Psychology, New Mexico State University, USA. [4]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, USA.
E-mail: v.amrhein@unibas.ch; dtrafimo@nmsu.edu; lesdomes@g.ucla.edu

Abstract. There has been much discussion of a "replication crisis" related to statistical inference, which has largely been attributed to overemphasis on and abuse of hypothesis testing. Much of the abuse stems from failure to recognize that statistical tests not only test hypotheses, but countless assumptions and the entire environment in which research takes place. Honestly reported results *must* vary from replication to replication because of varying assumption violations and random variation; excessive agreement itself would suggest deeper problems, such as failure to publish results in conflict with group expectations or desires. Considerable non-replication is thus to be expected even with the best reporting practices, and generalizations from single studies are rarely if ever warranted. Because of all the uncertain and unknown assumptions that underpin statistical inferences, we should treat inferential statistics as highly unstable local descriptions of relations between assumptions and data, rather than as generalizable inferences about hypotheses or models. And that means we should treat statistical results as being much more incomplete and uncertain than is currently the norm. Rather than focusing our study reports on uncertain conclusions, we should thus focus on describing accurately how the study was conducted, what data resulted, what analysis methods were used and why, and what problems occurred.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

The "crisis of unreplicable research" is not only about alleged replication failures. It is also about perceived non-replication of scientific results being interpreted as a sign of bad science (Baker 2016). Yes, there is an epidemic of misinterpretation of statistics and what amounts to scientific misconduct, even though it is common practice (such as selectively reporting studies that "worked" or that were "significant"; Martinson, Anderson, and de Vries 2005; John, Loewenstein, and Prelec 2012). But all results are uncertain and highly variable, even those from the most rigorous studies.

Indeed, Fisher (1937) wrote that "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon." Neyman and Pearson (1933a) wrote: "As far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis." And Boring (1919) said a century ago, "scientific generalization is a broader question than mathematical description."

Yet today we still indoctrinate students with methods that claim to produce scientific generalizations from mathematical descriptions of isolated studies. Naturally, such generalizations will often fail to agree with those from other studies – and thus statistical inference will fail to replicate. Because our current academic reward system is built on publishing inferences from single studies, it should come as no surprise that many conflicting generalizations are published, and hence that a high proportion of generalizations must be wrong.

A core problem is that both scientists and the public confound statistics with reality. But statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, these models are extremely simplified relative to the complexities of actual study conduct. Statistical results must eventually mislead us when they are used and communicated as if they present this complex reality, rather than a model for it. This is not a problem of our statistical methods. It is a problem of interpretation and communication of results.

In the following, we argue that the crisis of unreplicable research is mainly a crisis of overconfidence in statistical results. We recommend that we should use, communicate, and teach our statistical methods as being descriptive of logical relations between assumptions and data, rather than as allowing specific generalized inferences about universal populations.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

Inferences are not about hypotheses

A statistical model is a set of assumptions, and thus a compound hypothesis, about how the data could have been generated. The model matches reality to the degree that assumptions are met, starting from the assumption that we measured what we think we measured, for example that response time of subjects measures language understanding, and that measurement errors were either absent or adequately accounted for. Such model assumptions are part of what Meehl (1990) calls "auxiliary theories," or what McShane et al. (2018) call "neglected factors."

Thus, statistical models imply countless assumptions about the underlying reality. A null hypothesis such as "the means of these two populations do not differ" is an explicit assumption. Further assumptions that are often explicitly addressed in research reports are that sampling was random or that residuals are independent and identically distributed. Other assumptions may not even be recognized or mentioned in research reports, such as that there was no selection of particular results for presentation. Whether it is assumptions that are reviewed by inspecting residuals, or further assumptions that link statistics to reality: the validity of inferences that we draw from statistical results depends on the entire set of assumptions.

For example, we should think of a p-value as referring not only to the hypothesis it claims to test, such as a null hypothesis. A p-value refers to the entire model including other usually explicit assumptions like randomization of treatment and linearity of effects, and including usually implicit procedural assumptions such as that the equipment for taking measurements was in perfect working order (Greenland et al. 2016; Greenland 2017; Amrhein 2018). Whether recognized or not, these assumptions underpin the usual inferences from a test (Greenland 2018). A small p-value is either a result of random variation, or it indicates that at least one model assumption is violated. But it does not indicate which assumption is violated.

Yes, a small p-value may arise because the null hypothesis is false. But it can also mean that some mathematical aspect of the model was not correctly specified, that sampling was not a hundred percent random, that we accidentally switched the names of some factor levels, that we unintentionally, or intentionally, selected analyses that led to a small p-value (downward "p-hacking"), that we did not measure what we think we measured, or that a cable in our measuring device was loose (Amrhein 2018). And symmetrically, a large p-value may arise from mistakes and procedural errors, such as selecting analyses that led to a large p-value (upward p-hacking),

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

or using a measurement so noisy that the relation of the measured construct to anything else is hopelessly obscured.

Even the best single studies will be imperfect. Because of varying assumption violations, whether recognized or hidden, their results will vary from replication to replication. Consider the replication project by the Open Science Collaboration (2015): of 97 psychological studies with "significant" results ($p \leq 0.05$) that were repeated, only 35 had $p < 0.05$ in the repetition. This is much less than would have been expected if all original effects were true and only random variation would have caused the different p-values – under these circumstances, with an average power of 92% in the repetitions, 89 repetitions were expected to have $p \leq 0.05$. One explanation by the authors is that the assumption of "no selection of particular results for presentation" could have been violated in the original studies, resulting in inflated effect sizes that could not be replicated.

In addition to such assumption violations, results will vary due to random variation. For example, with a statistical power of 80%, two studies will be traditionally judged as "conflicting," meaning that one is "significant" and the other is not, in one third of the cases if there is a true effect (Greenland et al. 2016; Amrhein, Korner-Nievergelt, and Roth 2017); this means that with anything but nearly 100% statistical power, "significance" and "nonsignificance" cannot be used to judge success or failure of replication (Goodman 1992; Senn 2001, 2002).

The same applies for p-values, for confidence intervals or other interval estimates, and for averages or other point estimates: Unless the sample represents a substantial part of a clearly defined population, statistics will vary from sample to sample due to random variation. Some degree of variation, and hence non-replication, is the norm across honestly reported studies, even when all assumptions are met. Therefore, and because usually some assumptions will not be met, generalizations about hypotheses from single studies should be avoided. Having confidence in generalizations from single studies means having overconfidence in most cases. Inference that could be called trustworthy would require merging information from multiple studies and lines of evidence.

Overconfidence triggers selection bias

Unfortunately, even a combination of studies does not guarantee that inferences will be valid. Published results tend to be biased, for example because they may be selected from unpublished

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

results based on some statistical criterion. Such bad yet common scientific practice introduces bias by accumulating statistical inferences that go into a certain direction, typically emphasizing results that cross some p-value threshold (Amrhein, Korner-Nievergelt, and Roth 2017; Locascio 2017).

We suspect that a major driver of result-selection bias is overconfidence in statistical inference. For decades, scientists were taught to judge which results are trustworthy and which are not, and which results are thus worth being published or not, based on statistics obtained from single studies. Statistics was misused as an automated scientific decision machine, both for statements about hypotheses and for selection of studies for publication. And this made interpretation, publication, and advertising much easier, because everybody assumed that statistical inferences based on p-value thresholds or other rigid criteria would be "reliable," in the sense that a replication would probably meet the same thresholds or criteria again. So if researchers expect that a small p-value or a short confidence interval indicate "reliable" results, while all other results are "unreliable," it is clear that researchers are "prepared to ignore all results which fail to reach this standard" (Fisher 1937, p. 15; one of many published pleas by various authors encouraging selective reporting).

But any selection criterion will introduce bias. If there is a tendency to publish results because the estimates are yellow, because confidence intervals are short, and because p-values are small, then the published literature will become biased towards yellow results with underestimated variances and overestimated effect sizes. The latter effect is the "winner's curse" that is reflected in the findings of the Open Science Collaboration (2015): the average effect size in the original studies was about twice as large as in the replication studies that apparently reported all results and thus did not suffer from selection bias.

Even if authors report all study outcomes, but then select what to discuss and to highlight based on p-value thresholds or other aids to judgement, their conclusions and what is reported in subsequent news and reviews will be biased (Amrhein, Korner-Nievergelt, and Roth 2017). Selective attention based on study outcomes will therefore not only distort the literature but will slant published descriptions of study results – biasing the summary descriptions reported to practicing professionals and the general public.

One way to reduce this selection bias and related misreporting is to maintain that all results are uncertain. If we obtain a small p-value, a large effect estimate, or a narrow confidence interval – or even all three – we should not be confident about textbook inferences from these results. In

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

one of the next replications, p will be large, the effect estimate will be small, or the confidence interval wide, and thus the textbook inferences will shift dramatically. This caution is even more in force if there is any selective reporting of, or attention to, results.

We should trust in uncertainty and focus on describing accurately how the study was conducted, what problems occurred (e.g., nonresponse of some subjects, missing data), and what analysis methods were used, with detailed data tabulation and graphs, and complete reporting of results. The advent of online supplements eliminates the common excuse of space limitations preventing such detail.

Don't blame the p-value

A clear sign that overconfidence ruled the era of hypothesis testing is that many people still are surprised by the "dance of the p-values" (Cumming 2014), that is, by the way a valid p-value bounces around its range even in the largest of samples. This variability means that $p<0.05$ is no guarantee for $p<0.05$ in a replication (Goodman 1992; Senn 2001, 2002; Gelman and Stern 2006); after all, if the (null) hypothesis tested is correct and experimental conditions are ideal, the p-value will vary uniformly between 0 and 1. And even if our alternative hypothesis is correct, the p-value in the next sample will typically differ widely from our current sample: "The fickle P value generates irreproducible results" (Halsey et al. 2015), at least if reproducibility is defined by whether P is above or below a threshold and the power is not very high.

But the p-value itself is not supposed to be "reliable" in the sense of staying put. Its fickleness indicates variation in the data from sample to sample. If sample averages vary among samples, then p-values will vary as well, because they are calculated from sample averages. And we don't usually take a single sample average and announce it to be the truth. So if instead of simply reporting the p-value, we engage in "dichotomania" (Greenland 2017) and use it to decide which hypothesis is wrong and which is right, such scientifically destructive behavior is our fault, even if socially encouraged; it is not the fault of the p-value.

Further, if we overlook the sensitivity of p-values to possible violations of background assumptions, by assuming that p-values are only about deciding whether to reject "null hypotheses," we are privileging what may be a scientifically irrelevant hypothesis and are engaging in "nullism," a compulsion to test only one hypothesis among many of importance. But again, such bad behavior is our fault, even if socially encouraged. And if we interpret a small p-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

value as providing support for some alternative hypothesis (which currently seems to be a standard interpretation), this too is our fault, not the fault of the p-value.

Ban statistical tests?

It may help to ban some practices, at least temporary and in specific contexts. We ban alcohol drinking[1] before or during driving, despite its general acceptance in relaxed settings. There are even studies claiming positive effects of small doses of alcohol. But use easily becomes abuse, and it is essential to abstain from alcohol in many settings. Reaching for statistical tests to force out "inferences" (whether traditional "$p \leq \alpha$" testing or substitutes like Bayes-factor criteria) is, like drinking alcohol, a culturally ingrained habit. Statistical testing (like alcohol) often gives the wrong impression that complex decisions can be oversimplified without negative consequences, for example by relying on significance tests. And many people are addicted psychologically. These addictions are worth breaking.

At the very least, partial or temporary bans are one way to force researchers to learn how to analyze data in alternative ways (Trafimow and Marks 2015). Hopefully, thinking about advantages and disadvantages of alternatives will lead to more sober interpretation of statistics. One concern, however, is that complete prohibition could lead to misuse and abuse of other methods, such as Bayesian techniques – which have an additional source of nonreplicability insofar as what are acceptable priors can vary dramatically across research groups.

Long live no king

Fixed-cutoff (α-level) hypothesis testing has been king for over 80 years. We propose not to banish the king – a fixed-cutoff decision procedure may be useful, for example, in industrial or laboratory quality control, or "sampling tests laid down in commercial specifications" (Neyman and Pearson 1933b), in which automated decisions to stop a production line or to recalibrate equipment may be necessary. For scientific inference, however, we hope that dichotomania from which such procedures suffer can be cured by abandoning them in favor of data description and direct presentation of p-values – including p-values for alternative hypotheses (Cohen 1994;

[1] We prefer this analogy above comparing p-values or significance tests with guns, as we have heard or read sometimes.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

Ziliak and McCloskey 2008; Hurlbert and Lombardi 2009; Amrhein, Korner-Nievergelt, and Roth 2017; Greenland 2017; Amrhein and Greenland 2018; McShane et al. 2018; Trafimow et al. 2018). Parallel criticisms and remedies apply to tests based on Bayes factors, posterior odds, or any other statistical criterion.

Statistical models include not only our hypotheses but countless explicit and implicit assumptions (Greenland 2017); as a consequence, traditional hypothesis tests hardly test the hypotheses that we think they test. Yet hypothesis tests still precipitate decisions between significant/nonsignificant or reject/accept, even when such decisions are unnecessary to the research goals or unwarranted based on the available evidence. Hypothesis tests have been destructive agents by allowing overconfident claims from isolated studies to become the prevailing currency in the academic system.

Yes, sometimes we need to make decisions in science, for example whether to further pursue a study or not. For such a decision, we will usually weigh scientific and personal costs and benefits of our decision, applying informed personal judgment (Gigerenzer 1993). But when it comes to weigh evidence against, or in favor of, a scientific hypothesis, statistical tests cannot suffice, and may even be destructive if degraded into a binary form as in reporting tests as significant/non-significant, or in basing conclusions on whether the null value was included in or excluded from an interval. This is especially true when (as almost always) these results are sensitive to doubtful assumptions, such as absence of measurement-error dependencies. And even in the unlikely case that all model assumptions are met, we would still need to consider costs and benefits, as well as the published and unpublished literature, to judge a scientific hypothesis as being largely correct (subject to further evidence to the contrary). We hope that classical hypothesis testing will be retired quickly from research reporting, so that regicide is not necessary.

Empire of diversity

But what comes next? There are countless possibilities. If we want to report a continuous measure of refutational evidence against a model, we could use the traditional p-value in a continuous fashion, or better still the Shannon information or S-value (surprisal) of the test, $-\log_2(p)$, which is unbounded above and thus difficult to misinterpret as a hypothesis probability (Greenland 2017, 2018). If we want to compare the relative support for different models we could use likelihood ratios or Bayesian methods. But we should not lapse back into dichotomous

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

thinking by using some p-value threshold, or by making binary inferences based on confidence intervals or Bayes factors. And we should not pledge uncritical loyalty to posterior probabilities, especially since they rely on the same fundamental assumptions about the data-generating model that hypothesis tests and confidence intervals use.

A 95% confidence interval encompasses a range of hypotheses (effect sizes) that have a p-value exceeding 0.05. Instead of talking about hypothetical coverage of the true value by such intervals, which will fail under various assumption violations, we can think of the confidence interval as a "compatibility interval" (Greenland 2018), showing the effect sizes compatible with the 0.05 predictive criterion for the data, under the model used to compute the interval. Again, whether the interval includes or excludes zero should play no role in its interpretation, because even with only random variation present, the intervals from different data sets will be very different (Cumming 2014).

With additional (and inevitable) nonrandom variation, the true effect size will frequently be outside the interval. In reality, it will not happen that every assumption is met, nor will we be aware of every assumption. Stating that our data "support" any value in the compatibility interval (e.g., a zero effect), or that, upon unlimited replication, the intervals would cover the true value at some rate, or that the interval "measures uncertainty" by indicating the range of possible effect sizes, makes the compatibility interval into an *overconfidence* interval.

The empire of "statistical significance" has its roots in the 19th century writings of Edgeworth (1885) and reached full dominance with the spread of cutoffs for testing, formalized by Jerzy Neyman and Egon Pearson as Type-I error rates. Like the political empires of their period, such significance testing for scientific (as opposed to mechanical) inference is a relic of a bygone era, whose destructive effects reverberates to this day. We hope this era is over. As for what comes next, there is no substitute for accepting methodologic diversity (Good 1957; Cox 1978; Box 1980; Barnard 1996; Little 2006; Senn 2011; Efron and Hastie 2016; Crane 2017; Trafimow and Earp 2017), with careful assessment of uncertainty as the core motivation for statistical practice (Gelman 2016).

The replacement for significance testing

We should not "look for a magic alternative to NHST [null hypothesis significance testing], some other objective mechanical ritual to replace it. It doesn't exist" (Cohen 1994). What needs to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

change is not necessarily the statistical methods we use, but how we select our results for interpretation and publication, and what conclusions we draw. Fisher (1937) said we should not make decisions about scientific hypotheses based on single studies. Neyman and Pearson (1933a) said we cannot decide about the truth or falsehood of a scientific hypothesis based on a statistical test alone. So why would we want a mechanical decision procedure for single studies, if not for selecting results for interpretation or publication? As we described above, every selection criterion would introduce bias. We therefore join others who have advised that we should, to the extent feasible:

(a) Target results for publication and interpretation *before* data are collected and the analysis begins, using a defined protocol,

(b) Before analyzing data (and preferably before collecting them), make an analysis plan (i.e., a pre-analysis protocol), setting out how data will be analyzed; and, in the publication, show what results the protocol produced before displaying the results of any analyses deviating from the predefined protocol,

(c) Emphasize and interpret our estimates rather than tests, explicitly discussing both the lower and upper limits of our interval estimates,

(d) Report statistics – including p-values – precisely rather than merely as inequalities,[2]

(e) Not use the word "significant" to describe results, as it has produced far too much confusion between statistical, scientific, and policy meanings,

(f) acknowledge that our statistical results describe relations between assumptions and the data *in our study*, and that scientific generalization from single studies is unwarranted.

As an example, consider a study by Brown et al. 2017, who say that "exposure (...) was not associated with autism (...) (HR=1.61, 95% CI=0.997–2.59)." As is often the case, the authors misused the confidence interval as a significance test, and they claim to have demonstrated a zero association because the lower limit of the interval was slightly below a null effect (which would mean a hazard-rate ratio of HR=1), ignoring that the upper limit exceeded 2.50. A more correct summary of the results would have been: "Our estimate of the hazard-rate ratio was 1.61, and thus exposure could be associated with autism; however, possible hazard-rate ratios that are compatible with our data, given our model, ranged from 0.997 (essentially no association) to 2.59 (a relatively strong association)." If applicable, this could then be followed by a discussion of

[2] An exception: If a p-value is below the limit of numerical accuracy of the data, an inequality would be called for, but the precision would be context dependent: e.g., $p<10^{-8}$ is typical in genomics, $p<0.001$ is common in ecology.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

why the authors seem to think such potential hazard-rate ratios might be negligible, and how strong they judge their evidence not only based on the width of the confidence interval, but also in view of possible shortcomings of their study, of their prior knowledge about the study system, and of possible costs of their interpretation for the health of the patients.

Had the authors found an interval of 1.09–2.59 rather than 0.997–2.59, the reporting should have been largely similar. Even with an interval of 0.900–2.59, the description of the results should not be very different – the point estimate would still be a HR well above 1, indicating a possible positive association. What would need to change with the latter interval is the description that not only relatively large positive, but also small negative associations would fall within the interval.

Anything goes

So what do we conclude from a study like Brown et al. (2017)? If we would interpret an interval of 1.09–2.59 and of 0.997–2.59 in the same way, does that mean that the floodgates of "anything goes" are wide open? Yes, the floodgates should be open – for reporting our results. Everything should be published if whatever we measured made sense theoretically because it was connected in a potentially useful way to some *a priori* research question. If after doing the study, the researchers feel their measure did not make sense or their methods were faulty and publish that information, at least other researchers can learn that lesson without repeating the error.

However, the floodgates should be closed for drawing conclusions. Because their interval estimate touched the null value, Brown et al. (2017) tried to construct a conflict with previously observed associations that were nearly the same (hazard ratios around 1.6) but had interval estimates that did not include the null value. We think the most daring conclusion that could be drawn is that the new study was largely consistent with previous studies, but that a null value could also be compatible with the data, given the model. Although the study by Brown et al. (2017) had an overall sample size of near 36,000 children, a fair conclusion would have been to refrain from drawing conclusions, independent of the width of the confidence interval – even given that all assumptions are correct, it is largely a matter of luck whether an interval estimate includes the null value, and as the authors themselves note there are potential biases that the intervals do not account for.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

Given all the unmodeled uncertainties, it would be good to plan and publish such single studies as "prospective meta-analyses" (Ioannidis 2010), each producing results that are as unbiased as possible. In the words of Trafimow et al. (2018): "It is desirable to obtain precise estimates in those studies, but a more important goal is to eliminate publication bias by including wide confidence intervals and small effects in the literature, without which the cumulative evidence will be distorted."

If we are researchers ...

... and we obtained a large p-value, or a compatibility interval that includes a null effect, our interval will show that the null hypothesis is only one of the many practically different hypotheses that are compatible with the data (Rothman, Greenland, and Lash 2008; Greenland et al. 2016). Unlike Brown et al. (2017) suggest with their hazard-rate ratios, we cannot claim statistics demonstrate there is no effect whatsoever, because even if the data remain consistent with a zero effect, they remain consistent with many other effects as well. A "proof of the null hypothesis" such as "the earth is flat ($p>0.05$)" is therefore not possible (Greenland 2011; Amrhein, Korner-Nievergelt, and Roth 2017). And we should remember there are lots of additional hypotheses outside the compatibility interval that will also be compatible with our data, due to methodologic limitations that we have not modeled.

Thus, we should not be overconfident about our "weak evidence." Almost never will we have found absolutely no effect. Let us free our negative results by allowing them to be potentially positive (Amrhein, Korner-Nievergelt, and Roth 2017). And symmetrically, let us free our positive results by allowing them to be potentially negative. If we believe we have "strong evidence" because our p-value is small, our point estimate is large, or our interval estimate is not near the null, we are placing too much faith in our inferential statistics. The limits of our interval estimate will usually show that any practical conclusion is uncertain anyway. Keep in mind, however, that interval estimates do not "measure" uncertainty – they give a rough estimate of uncertainty, given that all assumptions about our model are true. And remember the "dance of the confidence intervals" (Cumming 2014) shows a valid interval will bounce around from sample to sample even if all assumption are true, due to random variation.

Because of all the known and unknown assumptions about our model in such studies, we should treat every inferential statistic, including point estimates, interval estimates, p-values, and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

posterior probabilities, as being descriptive of the relation of statistical models to the data. It is hard enough to describe the known assumptions about our model. We should not draw inference and generalize based on assumptions we cannot be certain about or we do not even think about.

For example, a p-value is merely the probability of one particular test statistic being as or more extreme than observed in our particular study, given that the current model is true. There is no inferential meaning that must be attached to that. For the next set of data, the p-value will be different. A small p-value is just a warning signal that the current model could have a problem, so we should check our model assumptions (including an assumption such as "the means of these two populations do not differ", i.e., our tested hypothesis). And this assumption checking does not only mean inspecting residuals, but also checking the extent of deviations of our study from a perfect randomized experiment or random survey, whether from failures of protocol, measurement, equipment, or any of the innumerable details that real research must confront.

Science includes learning about assumption violations, then addressing those violations and improving the performance of our models about reality. Statistics can help by formalizing parts of the models, and by assisting in careful assessment of uncertainty. We should thus communicate our limited conclusions about our study system, not our generalized inferences about some ill-defined universal population. And presentation decisions should not be based on p-values, nor on any other statistic. Presentations that start with analysis plans that were formulated before the analysis (pre-analysis protocols) can help strengthen both the validity and credibility of our inferences. If we think we did a good study, we should be modest about our conclusions, but be proud about our painfully honest and thorough description and discussion of our methods and of our data.

If we are science writers and journalists ...

... we should continue writing about isolated experiments and replications. Single studies are the life blood of science. If we think we found a good study, or a bad study, we may report it. But let us try not to be impressed by what researchers say is surprising about their study – surprising results are often products of data dredging or random error, and are thus less reproducible (Open Science Collaboration 2015). So surprising results will usually not point to general scientific discoveries, although they may still be valuable because they lead to new insights about study problems and violations of assumptions.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

Then too, we should not overemphasize what researchers say was unsurprising, since that may largely reflect their conformity to group expectations rather than what the data would actually show under close scrutiny. Indeed, we might consider not asking researchers about surprising or unsurprising results, but instead ask which results appeared most boring because they were shown several times before and thus seem to be trustworthy. More generally, we should not fall for overconfident claims by researchers, science writers, or journalists. Rather, we should try to uncover overconfident claims and the bad incentives leading to overconfident claims.

Clear signs of overconfidence are formulations like "we proved" or "we disproved" or "we rejected" a hypothesis, or "we have shown" a relation exists or is explained in some manner. So are "there was *no* effect / *no* association / *no* difference" (which almost always would be an impossible proof of the null hypothesis), and "our study confirms / validates / invalidates / refutes previous results" (because a single study can only add a further data point to the larger picture). If we find any of those or related phrases, we should question the interpretations being offered in the paper and search for arguments provided by the authors. If the main argument for a conclusion is that the results were "significant" or "not significant", this does not automatically mean that the study is bad. But it does flag the paper as likely providing an unreliable interpretation of the reported results.

The hard truth is that we, as journalists, cannot decide whether a result from a single study can be generalized – and the same is usually true for the authors of the study. Instead, an important role for statistics in research is the summary and accumulation of information. If replications do not find the same results, this is not necessarily a crisis, but is part of a natural process by which science evolves. The goal of scientific methodology should be to direct this evolution toward ever more accurate descriptions of the world and how it works, not toward ever more publication of inferences, decisions, or conclusions.

References

Amrhein, V. (2018), "Inferential statistics is not inferential," *sci five, University of Basel*, http://bit.ly/2oLY7t9.

Amrhein, V., and Greenland, S. (2018), "Remove, rather than redefine, statistical significance," *Nature Human Behaviour*, 2, 4.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017), "The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research," *PeerJ*, 5, e3544. doi: 10.7717/peerj.3544.

Baker, M. (2016), "Is there a reproducibility crisis?" *Nature*, 533, 452–454.

Barnard, G.A. (1996), "Fragments of a statistical autobiography," *Student*, 1, 257–268.

Boring, E.G. (1919), "Mathematical vs. scientific significance," *Psychological Bulletin*, 16, 335–338. doi: 10.1037/h0074554.

Box, G.E.P. (1980), "Sampling and Bayes' inference in scientific modeling and robustness," *Journal of the Royal Statistical Society, Series A*, 143, 383–430. doi: 10.2307/2982063.

Brown, H.K., Ray, J.G., Wilton, A.S., Lunsky, Y., Gomes, T., and Vigod, S.N. (2017), "Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children," *Jama-Journal of the American Medical Association*, 317, 1544–1552. doi: 10.1001/jama.2017.3415.

Cohen, J. (1994), "The earth is round (p < .05)," *American Psychologist*, 49, 997–1003. doi: 10.1037/0003-066x.50.12.1103.

Cox, D.R. (1978), "Foundations of statistical inference: the case for eclecticism," *Australian Journal of Statistics*, 20, 43–59.

Crane, H. (2017), "Why 'Redefining statistical significance' will not improve reproducibility and could make the replication crisis worse," arXiv:1711.07801.

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

Edgeworth, F.Y. (1885), "Methods of statistics," *Journal of the Statistical Society of London*, Jubilee Volume, 181–217.

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

Fisher, R.A. (1937), *The Design of Experiments*, second ed, Edinburgh: Oliver and Boyd.

Gelman, A. (2016), "The problems with p-values are not just with p-values," *The American Statistician, supplemental material to the ASA statement on p-values and statistical significance*. doi: 10.1080/00031305.2016.1154108.

Gelman, A., and Stern, H. (2006), "The difference between 'significant' and 'not significant' is not itself statistically significant," *The American Statistician*, 60, 328–331. doi: 10.1198/000313006x152649.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

Gigerenzer, G. (1993), "The superego, the ego, and the id in statistical reasoning," in *A Handbook for Data Analysis in the Behavioral Sciences*, edited by G. Keren and C. Lewis, 311–339, Hillsdale: Lawrence Erlbaum Associates.

Good, I.J. (1957), "Some logic and history of hypothesis testing," in *Philosophical Foundations of Economics*, edited by J.C. Pitt, 149–174, Dordrecht, Holland: D. Reidel. Reprinted as Ch. 14 in Good, I.J. (1983), *Good Thinking*, 129–148, Minneapolis: U. Minnesota Press.

Goodman, S.N. (1992), "A comment on replication, p-values and evidence," *Statistics in Medicine*, 11, 875–879. doi: 10.1002/sim.4780110705.

Greenland, S. (2011), "Null misinterpretation in statistical testing and its impact on health risk assessment," *Preventive Medicine*, 53, 225–228. doi: 10.1016/j.ypmed.2011.08.010.

Greenland, S. (2017), "Invited commentary: The need for cognitive science in methodology," *American Journal of Epidemiology*, 186, 639–645. doi: 10.1093/aje/kwx259.

Greenland, S. (2018), "The unconditional information in *P*-values, and its refutational interpretation via *S*-values," *under submission*.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.C., Poole, C., Goodman, S.N., and Altman, D.G. (2016), "Statistical tests, confidence intervals, and power: A guide to misinterpretations," *European Journal of Epidemiology*, 31, 337–350. doi: 10.1007/s10654-016-0149-3.

Halsey, L.G., Curran-Everett, D., Vowler, S.L., and Drummond, G.B. (2015), "The fickle *P* value generates irreproducible results," *Nature Methods*, 12, 179–185. doi: 10.1038/nmeth.3288.

Hurlbert, S.H. and Lombardi, C.M. (2009), "Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici*, 46, 311–349. doi: 10.5735/086.046.0501.

Ioannidis, J.P.A. (2010), "Meta-research: the art of getting it wrong," *Research Synthesis Methods* 1, 169–184. doi: 10.1002/jrsm.19.

John, L.K., Loewenstein, G., and Prelec, D. (2012), "Measuring the prevalence of questionable research practices with incentives for truth telling," *Psychological Science*, 23, 524–532. doi: 10.1177/0956797611430953.

Little, R.J. (2006), "Calibrated Bayes: A Bayes/frequentist roadmap," *American Statistician*, 60, 213–223. doi: 10.1198/000313006x117837.

Locascio, J. (2017), "Results blind science publishing," *Basic and Applied Social Psychology*, 39, 239–246. https://doi.org/10.1080/01973533.2017.1336093.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

Martinson, B.C., Anderson, M.S., and de Vries, R. (2005), "Scientists behaving badly," *Nature*, 435, 737–738. doi: 10.1038/435737a.

McShane, B.B., Gal, D., Gelman, A., Robert, C., and Tackett, J.L. (2018), "Abandon statistical significance," *arXiv*:1709.07588.

Meehl, P.E. (1990), "Why summaries of research on psychological theories are often uninterpretable," *Psychological Reports*, 66, 195–244. doi: 10.2466/pr0.66.1.195-244.

Neyman, J., and Pearson, E.S. (1933a), "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289–337. doi: 10.1098/rsta.1933.0009.

Neyman, J., and Pearson, E.S. (1933b), "The testing of statistical hypotheses in relation to probabilities a priori," *Mathematical Proceedings of the Cambridge Philosophical Society*, 29, 492–510. doi: 10.1017/S030500410001152X.

Open Science Collaboration (2015), "Estimating the reproducibility of psychological science," *Science*, 349, aac4716. doi: 10.1126/science.aac4716.

Rothman, K., Greenland, S., and Lash, T.L. (2008), *Modern Epidemiology*, 3rd Edition, Ch. 10, Philadelphia, PA: Lippincott Williams & Wilkins.

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

Senn, S.J. (2002), "Letter to the Editor" re: Goodman 1992, *Statistics in Medicine*, 21, 2437–2444. doi: 10.1002/sim.1072.

Senn, S.J. (2011), "You may believe you are a Bayesian but you are probably wrong," *Rational Markets and Morals*, 2, 48–66.

Trafimow, D., Amrhein, V., Areshenkoff, C.N., Barrera-Causil, C., Beh, E.J., Bilgiç, Y., Bono, R., Bradley, M.T., Briggs, W.M., Cepeda-Freyre, H.A., Chaigneau, S.E., Ciocca, D.R., Carlos Correa, J., Cousineau, D., de Boer, M.R., Dhar, S.S., Dolgov, I., Gómez-Benito, J., Grendar, M., Grice, J., Guerrero-Gimenez, M.E., Gutiérrez, A., Huedo-Medina, T.B., Jaffe, K., Janyan, A., Karimnezhad, A., Korner-Nievergelt, F., Kosugi, K., Lachmair, M., Ledesma, R., Limongi, R., Liuzza, M.T., Lombardo, R., Marks, M., Meinlschmidt, G., Nalborczyk, L., Nguyen, H.T., Ospina, R., Perezgonzalez, J.D., Pfister, R., Rahona, J.J., Rodríguez-Medina, D.A., Romão, X., Ruiz-Fernández, S., Suarez, I., Tegethoff, M., Tejo, M., van de Schoot, R., Vankov, I., Velasco-Forero, S., Wang, T., Yamada, Y., Zoppino,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018

F.C., and Marmolejo-Ramos, F. (2018), "Manipulating the alpha level cannot cure significance testing," *Frontiers in Psychology*, 9, 699. doi: 10.3389/fpsyg.2018.00699.

Trafimow, D., and Earp, B.D. (2017), "Null hypothesis significance testing and Type I error: The domain problem," *New Ideas in Psychology*, 45, 19–27. doi: 10.1016/j.newideapsych.2017.01.002.

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. doi: 10.1080/01973533.2015.1012991.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26857v2 | CC BY 4.0 Open Access | rec: 19 Jul 2018, publ: 19 Jul 2018