Suggested Improvements and Challenges

12

Contents

What You Will Learn in This Chapter

The Test for Excess Success highlighted problems with current scientific practice for fields that use hypothesis testing. In addition, errors in statistical reporting (e.g., the reported p-value does not match the reported t-value) are common, results presented at conferences are changed for journal publications, and researchers admit to publication bias and other poor practices. These problems have motivated many researchers to propose counteractions designed to improve scientific practice. In this chapter, we critically evaluate some of these proposals. While there are some positive aspects to many of these proposals, they often have negative characteristics as well, and none of the proposals seem to tackle the more fundamental issues. Along those lines, we do not have a specific proposal that will address all of these problems, but we identify what we think should be the critical long-term goals of science and suggest that scientific practice should reflect those goals.

© The Author(s) 2019

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_12

12.1 Should Every Experiment Be Published?

Some people suggest that scientists have an obligation to publish every experimental result, regardless of statistical significance. Indeed, non-significant experiments contain information about effects (both null and otherwise) that can only be utilized (through meta-analytic methods) if the data got published. More generally, publishing all the data allows readers to draw proper conclusions about effects and avoid overestimating effect sizes that are caused by publication bias.

However, there are several difficulties with publishing all experimental results. First, can experimenters reliably distinguish an experiment that failed methodologically (e.g., a piece of equipment failed) from those experiments that failed due to random sampling? If these different types of failures cannot be distinguished, then the literature becomes cluttered with experimental results that are flawed for a variety of reasons (of course, this might already be the case, but publishing everything may exacerbate the problem).

It is also a bit hard to see how scientists who publish everything should interpret their finding. Can there even be a conclusions section to a paper that simply adds an incremental amount of data to an existing topic? Moreover, when doing a meta-analysis, how does one decide which findings to include? These are all problems that exist right now, but publishing all the data does not remove them, and may make them worse.

Finally, when does the field decide that enough data has been collected? If a meta-analysis gives $p = 0.08$, should more experiments be run until $p < 0.05$? That approach would be problematic because it is optional stopping, just at the level of experiments added to a meta-analysis rather than at the level of individual subjects added to one experiment. Is there ever a moment when the field reports that an effect exists (see Chap. 3, Implication 1a)? What happens when more data changes the decision?

12.2 Preregistration

Several journals now encourage and promote replication studies, often with a requirement for researchers to preregister their experiment, analysis methods, and data collection plans. Some scientists consider preregistration to be the only viable path to move psychology (and other fields that use statistics) out of what they see as a crisis.

The idea of preregistration is that before actually running an experiment a scientist describes the total experimental plan in a place where the scientist cannot alter the original plan (e.g., the Open Science Framework, or AsPredicted.org). This plan describes the stimuli, tasks, experimental methods, number of samples and how they are sampled, the questions to be investigated, and the data analysis plan. After writing down these details, the experiment is run and any deviation from the preregistered plan is noted (perhaps with justification). Proponents of preregistration note that it prevents researchers from generating theoretical ideas or methods of data analysis after looking at the data

(HARKing). With preregistration, it would be obvious that a researcher stopped data collection early or added observations (perhaps due to optional stopping) or that various measures were combined in a way that is different from what was originally planned. If preregistered documents are in a public place, preregistration might also reduce the occurrence of publication bias because there is a public record about the researcher's intention to run the experiment; along similar lines, journals might agree to publish preregistered experiments prior to data collection.

These attributes all seem like good pragmatic reasons for scientists to practice preregistration. However, deeper consideration raises questions about what should be inferred when a researcher sticks to the preregistered plan. Does success for a pre-registered strategy lend some extra confidence in the results or in the theoretical conclusion? Does it increase belief in the process that produced the preregistered experimental design? A consideration of two extremes suggests that it does not.

Extreme Case 1 Suppose a researcher generates a hypothesis by flipping a coin. For example, a drug may increase or decrease working memory. The coin comes up "heads", so the researcher preregisters the hypothesis that the drug will increase working memory. The experiment is subsequently run and finds the predicted effect. Whether the populations truly differ or not, surely such an experimental outcome does not actually validate the process by which the hypothesis was generated (a coin flip). For the experiment to validate the prediction (not just the hypothesis), there needs to be some justification for the theory/process that generated the prediction. Preregistration does not, and cannot, provide such justification; so preregistration seems rather silly for unjustified experimental designs.

Extreme Case 2 Suppose a researcher generates a hypothesis and has an effect size derived from a quantitative theory that has previously been published in the literature. The researcher preregisters this hypothesis and the corresponding experimental design. The subsequent experiment finds the predicted difference. Such an experimental finding may be interpreted as strong validation of the hypothesis and of the quantitative theory, but it does not seem that preregistration has anything to do with such validation. Since the theory has previously been published, other researchers could follow the steps of the original researcher and derive the very same predicted effect size and thereby conclude that the experimental design was appropriate. In a situation such as this it seems unnecessary to preregister the experimental design because its justification is derived from existing ideas.

Most research situations are neither of these extremes, but scientists often design experiments using a mix of vague ideas, intuition, curiosity, past experimental results, and quantitative theories. It is impossible to gauge the quality of the experimental design for the vague parts; and preregistration does not change that situation. For those parts of the predicted hypotheses (and methods and measures) that are quantitatively derived from existing theory or knowledge, it is possible to gauge the quality of the experiment from

readily available information; and preregistration does not add anything to the quality of the design.

Preregistration does force researchers to commit to making a real prediction and then creating an experiment that properly tests that prediction. This is a laudable goal. But such a goal does not make sense if researchers do not have any hope of achieving it. When researchers design their experiments based on vague ideas, they are doing exploratory work, and it is rather silly to ask such researchers (or even to invite them) to make predictions. If forced to do so, such researchers may generate some predictions, but those predictions will not be meaningful with regard to the process by which they were generated. At best, such studies would provide information about a scientist's intuition, but researchers are generally not interested in whether scientists can generate good guesses. They run confirmatory studies to test aspects of theoretical claims.

At a practical level, many researchers who are motivated to preregister their hypotheses may quickly realize that they cannot do it because their theories are not sufficiently precise. That might be a good discovery for those researchers, and it may lead to better science in the long term. Likewise, preregistration does deal with some types of researcher degrees of freedom, such as optional stopping, dropping unsuccessful conditions, and hypothesizing after the results are known (HARKing). But these are exactly the issues that are handled by good justification for experimental design.

In summary, writing down the justifications for an experimental design may be a good activity for scientists to self-check the quality of their planned experiment. It may also be good to write down all the details and justifications of an experiment because it is easy to forget the justifications later. Moreover, when attempting to be so precise, it may often be the case that scientists recognize that part of their work is exploratory. Recognizing the exploratory parts of research can help guide how scientists interpret and present their empirical findings. However, justification for an experimental design should be part of a regular scientific report about the experiment; so there seems to be no additional advantage to publishing the justification in advance as a preregistration.

12.3 Alternative Statistical Analyses

There is a long history of criticism about hypothesis testing, and such criticisms often include alternative analyses that are claimed to lead to better statistical inference. While not denying that traditional hypothesis testing has problematic issues, and while personally favoring some alternative statistical methods, it is important to understand just what a method does and then determine whether it is appropriate for a particular scientific investigation.

For example, critiques of hypothesis testing sometimes claim that the p-value used in hypothesis testing is meaningless, noisy, or pointless. Depending on the situation, there may be merit to some of these concerns, but the general point cannot be true because the p-value is often based on exactly the same information in a data set (the estimated

signal-to-noise-ratio) as other statistics. For example, when an analysis is based on a two-sample t-test with known sample sizes n_1 and n_2, it is possible to transform the t value to many other statistics. An on-line app to do the conversion is at http://psych.purdue.edu/~gfrancis/EquivalentStatistics/.

This equivalence of information across the statistics suggests that what matters for using a particular statistic is the inference that is being made. If that inference is what you are interested in, then you should use it. Different choices of which statistic to use can give very different answers because they are addressing different questions. Just for illustration and without full explanation, we show a few examples to highlight the idea (it is not necessary to understand what the following terms exactly mean). If $n_1 = n_2 = 250$ and $d = 0.183$, then (if all other requirements are satisfied) the following are all valid inferences from the data:

- $p = 0.04$, which is less than the typical 0.05 criterion. The result is statistically significant.
- $CI_{95} = (0.007, 0.359)$ is a confidence interval for Cohen's d; it is often interpreted as describing some uncertainty about the true population effect size.
- $\Delta AIC = 2.19$, refers to the difference of the Akaike Information Criterion for null and alternative models. The value suggests that a model with different means better predicts future data than a model with a common mean for the two populations.
- $\Delta BIC = -2.03$, refers to the difference of the Bayesian Information Criterion for null and alternative models. The value provides evidence that the null model is true.
- JZS BF= 0.755, refers to a Bayes Factor based on a specific Jeffreys-Zellner-Siow prior, which provides weak evidence that the null model is true.

Thus, this data produces a significant result ($p < 0.05$) and favors the alternative model ($\Delta AIC > 0$), but it also provides some evidence that the null model is true ($\Delta BIC < 0$ and JZS BF<1). These conclusions might seem contradictory; but the conclusions are different because the questions are different. If you want to base decisions about effects by using a process that controls the Type I error rate, then the p-value provides the answer you are looking for. If you find the range of a confidence interval useful for representing uncertainty in an estimate of the standardized effect size, then the confidence interval provides what you want. If you want to estimate whether a model based on a common mean or a model with different means better predicts future data, then the ΔAIC value provides an answer. If you want to determine whether the data provides evidence for the null (common mean) or alternative (two different means) model, then the ΔBIC or the JZS BF provides the answer. Note that the ΔAIC, ΔBIC, and BF approaches have an option to *accept* the null hypothesis. Standard null hypothesis testing never accepts the null because absence of proof is not proof of absence (Chap. 3, Implication 3a).

12.4 The Role of Replication

Many scientists consider replication to be the final arbiter of empirical issues. If a finding replicates (perhaps by an independent lab), then the result is considered proven. A failure to replicate raises questions that have to be resolved (one group or the other must have made a mistake). Chapters 9–11 suggest that this view of replication is too simplistic when statistics are used. Simply due to random sampling, even well done studies of real effects will not always produce a successful replication.

Intuitions about the role of replication in science are largely derived from its role in the physical sciences. For example, acceleration of a feather in free-fall is the same as for a hammer, but only in a vacuum where air resistance does not impede their movement. The latter part of the previous sentence is especially important because it emphasizes that the outcome is dependent on the conditions and details of the experiment. For example, to successfully replicate free-fall acceleration in a vacuum it is necessary to have accurate measurements of distance and time; and a photogate timer is superior to an experimenter with a stopwatch. In addition, Newtonian physics posits that it does not matter whether the experiment is performed in the morning or afternoon, by male or female experimenters, or uses a dog treat and a battleship instead of a feather and a hammer.

Replication success in physics is nearly always determined relative to a theory. There is much experimental evidence that Newtonian physics is largely correct and that the type of object is irrelevant to free-fall acceleration, provided one has appropriate conditions and measurement precision. Under such situations, replication failures become especially interesting because they indicate a problem in the experimental set up (perhaps the vacuum has failed) or in the theory (photons have a constant velocity even under the effects of gravity, which leads to Einstein's relativity theory). Science is rife with stories where replication successes provided overwhelming support for a theory (replication as confirmation) and also where replication failures drive theory development.

In contrast to psychology, sociology, biology, and many more disciplines, an important characteristic of replication in the physical sciences is that the experimental outcome is (nearly) deterministic. Great care goes into identifying and reducing sources of noise. For example, a naïve experimental physicist might use the left and right hands to release objects, which would introduce some random difference in the release time. A better free-fall experiment would involve a mechanical device that was calibrated to insure simultaneous release of the two items, thereby largely removing one source of noise. For many phenomena in physics, only the motivation and resources to remove uncertainty limits this kind of careful control.

The situation is rather different for experimental psychology, medicine, and related fields. Some sources of noise can be reduced (e.g., by improving measurement scales or training subjects to perform better) but the limits imposed by the topic often exceed the motivation and resources of the experimenter. More importantly, there is often natural variability across the effect being measured (e.g., some people show an effect while other

people do not), that is, variability is part of the phenomenon rather than being added noise (Chap. 3, Implication 4). As a result, these fields often have no choice but to utilize statistical methods, such as null hypothesis testing, and they will sometimes produce inconsistent outcomes simply due to sampling variability.

For these fields to use replication in the way it is used by physics, studies need to work nearly every time (high power) and/or there needs to be a theory that distinguishes between sampling variability and measurement variability (Chap. 3, Implication 4a).

12.5 A Focus on Mechanisms

As we have seen throughout this book, problems of statistics are ubiquitous in many sciences. These problems are not only problems of bad statistical practice as outlined in the last chapters. The problems are often of conceptual nature. As we have seen in Chap. 3 Implications 4, science that is mainly based on statistics can often not disentangle true variability and noise and, thus, it remains an open question whether a significant effect is true in general or just holds true for a subpopulation. In addition as shown in the last subsection, failures of an experiment do not tell too much. We have the feeling that the problems with statistics are keeping scientists so busy that they may have lost focus on the perhaps most fundamental aspects of science: specification and understanding of the mechanisms that produce a particular outcome. Without such understanding it is impossible to predict future outcomes or to be confident that an empirical finding will be replicated in a new setting.

To have high confidence in any empirical result requires a corresponding theory that specifies the necessary and sufficient conditions. Even experimental findings that are generally reproducible cannot have a high level of confidence without a corresponding theoretical explanation because one cannot be sure that a new experimental condition will show the same result.

Consider the potential differences between two Magnetic Resonance Imaging (MRI) machines shown in Fig. 12.1. The MRI machine at the top right is located in Lausanne, Switzerland (top left), while the MRI machine at the bottom right is located in West Lafayette, Indiana (bottom left). How do engineers know that these two machines work similarly? There are many differences between Lausanne and West Lafayette that could, potentially, alter the behavior of the MRI machines. Lausanne has nearby mountains, a lake, stone buildings, and typical residents eat fondue and speak French. West Lafayette has nearby soybean fields, a river, brick buildings, and typical residents eat hamburgers and speak English. How should we know that these differences do not make the MRI machines behave differently? It is not enough to know that other MRI machines seem to function similar to each other; after all, every new machine is in a new environment and it is not feasible to test every possible environment.

Engineers have confidence in the behavior of the MRI machines because they understand how the machines work. For example, modern MRI machines use the properties of

Fig. 12.1 Two MRI machines in Lausanne, Switzerland (top) and West Lafayette, Indiana (bottom)

superconductivity, which was experimentally discovered in 1911. Even though superconductivity could be reproduced in many settings, it was not until the 1930s that a quantitative theory explained superconductivity. Further work in the 1950s explained superconductivity in terms of a superfluid of Cooper pairs, thereby connecting superconductivity to condensed matter physics and quantum mechanics. This kind of understanding allows scientists to identify the necessary and sufficient conditions to produce superconductivity and to predict its properties in MRI machines and elsewhere. Engineers have confidence that MRI machines will work not because previous studies have shown that they do work but because the field's theoretical understanding of superconductivity (and many other aspects of MRI machines) predicts that they will work despite some environmental differences.

As another example, consider the plague, which killed nearly one-third of people in Europe centuries ago. French scientist Paul-Louis Simond established in 1898 that fleas from rats transmitted the plague. Careful experiments justified this mechanism (he also identified the bacteria *Yersinia pestis*, which was infecting the fleas), as he showed that when fleas jumped from an infected rat to a healthy rat, the plague was transmitted. The rat-plague connection served as a (incomplete) mechanism: rats bring the plague. The implication of such a mechanism is clear, to reduce the occurrence of the plague,

reduce the number of rats: keep cats and dogs in the home to serve as rat predators, keep food stuff in sealed packages, set rat traps, avoid contact with live or dead rats (sadly, in the Great Plague of London in 1665, one suspected mechanism was that dogs and cats were spreading the plague, so they were exterminated in great numbers; which probably increased the rat population). When a case of the plague appeared in San Francisco in 1900, the scientific advice was to kill rats (but political conflict prevented this good advice from being generally applied). Note that the rat-plague mechanism does not have to make a quantitative prediction about exactly how many lives will be saved by various actions; it says that almost any action to reduce contact with rats is going to help control the plague. It also indicates what kinds of actions are unlikely to be helpful (e.g., isolation of an infected household). Of course, theories with more accurate mechanisms are even better. Nowadays the plague is kept in check by antibiotics, which directly attack the underlying bacterial cause.

As a final example, consider the effect of anesthesia. The effect of isofluran, one type of anesthetic, has been known for more than a century, and it enables complicated surgeries that would otherwise be impossible. Although the effects are very reliable, the mechanism by which isofluran induces anesthesia is unknown. Moreover, there are occasional failures where patients wake up in the middle of surgery or subsequently have memories about the surgery. If we understood the mechanisms by which isofluran induces anestheisia, we might be able to anticipate and compensate for these failures. Without a theory we do not know whether a failure is noise or hints at some hidden mechanism. In the meantime, doctors use anesthesia with the knowledge that they need to be ready should there be a failure.

Identifying mechanisms and justifying their role in a scientific phenomenon is very difficult to do, but it should be the long-term goal of every scientist. Some scientists may never actually achieve that goal, as they (valuably) spend their time gathering data and testing out ideas; activities that may help future scientists identify and justify mechanisms. Until justified mechanisms exist, scientists can never be confident that a particular effect will show up in a new setting.

In some sense, the long-term goal of science is to (as much as possible) remove the role of statistics by finding the "right" factors and thus reducing variability (see Chap. 6). An understanding of mechanisms promotes new hypotheses that can be rigorously investigated with good experimental designs. For example, an understanding of how vitamins affect bodily organs would explain why vitamins improve the health of some people but hurt the health of other people. Hence, with deeper insights into mechanisms many of the problems and concerns raised throughout this book largely disappear. Thus, to rejuvenate scientific practice the goal should not be to reform statistics but to not need it.

Take Home Messages

1. Many suggestions, such as preregistration, to improve statistical practice do not address the fundamental problems.
2. Good science involves more than just statistics.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.