Magnitude of Excess Success

11

Contents

What You Will Learn in This Chapter

Chapter 10 introduced the Test for Excess Success (TES), which detects some forms of bias in statistical analyses across multiple experiments. Although we found study sets where the TES indicated problems, it could be that the vast majority of scientific investigations are fine. This chapter shows, unfortunately, that this is not the case.

11.1 You Probably Have Trouble Detecting Bias

The basic ideas of the TES are fairly simple. An important point of statistics is that failures are necessary. Even if an effect is real, sometimes a researcher should select a random sample that does not show the effect. Only reporting successful outcomes is problematic because it inflates the reported effect size and can indicate the existence of an effect when

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_11

Table 11.1 Summary statistics for three sets of five simulated experiments

Set A				Set B				Set C			
$n_1 = n_2$	t	p	g	$n_1 = n_2$	t	p	g	$n_1 = n_2$	t	p	g
10	2.48	0.03	1.06	21	2.67	0.01	0.81	16	2.10	0.04	0.72
28	2.10	0.04	0.55	27	4.72	< 0.01	1.26	19	2.19	0.04	0.70
10	3.12	0.01	1.34	22	3.66	< 0.01	1.08	25	2.22	0.03	0.62
15	2.25	0.04	0.80	26	2.74	0.01	0.75	14	2.24	0.04	0.82
12	2.34	0.03	0.92	24	2.06	0.05	0.58	23	2.49	0.02	0.72

One set was created with optional stopping. One set was created with publication bias. One set is valid. Which is the valid set?

it does not exist. Too much replication success is a marker that something has gone wrong in the reporting, analysis, theorizing, or data collection process.

To demonstrate the impact of this kind of interpretation, consider the three sets of simulated results in Table 11.1. Each simulated data set was analyzed with a two-sample t-test. Each set of five studies is based on a different type of simulated experiment. For one set of experiments (valid testing) the population effect size is 0.8. The sample sizes, $n_1 = n_2$, were chosen randomly to be between 10 and 30. All five experiments produced significant results that were fully reported.

Another set of studies in Table 11.1 is based on simulated experiments where the population effect size is 0 (no effect). The sample was generated by an optional stopping approach that started with $n_1 = n_2 = 10$ and increased in steps of one up to a maximum size of 30. A total of twenty experiments were simulated and five of them happened to produce a significant result. Those five significant experiments were reported and the 15 non-significant experiments were not reported.

Another set of studies in Table 11.1 is based on simulated experiments where the true effect size is 0.1. The sample size was randomly chosen between 10 and 30. A total of 100 experiments were simulated and five of them happened to produce significant results. Those five significant experiments were reported and the 95 non-significant experiments were not reported.

The task for the reader is to determine which set of experiments in Table 11.1 corresponds to which simulation condition. Just to be clear, one set of experiments has a very large population effect size that was investigated with five proper experiments that were all fully reported. This is a valid set of experiments. A second set of experiments has no effect at all, but used optional stopping and publication bias to only report five significant results. This is an invalid set of experiments. A third set of experiments has a tiny effect size and used many experiments and publication bias to report only five significant results. This is also an invalid experiment set. Which is the valid experiment set in Table 11.1? We suggest the reader look at the statistics and make a judgment before reading the following text.

Did you spot the valid set? If you find it difficult or you are unsure, you may take some comfort in knowing that you are not alone. Even scientists with substantial experience evaluating statistical data often struggle to identify the valid experiment set in Table 11.1. It is worth recognizing the implication of this observation. With publication bias and optional stopping, scientists often do not know how to distinguish between a set of results where there is no effect at all and a set of results where there is a very large effect size.

Doing a meta-analysis that pools the reported effect sizes gives: for Set A $g^* = 0.82$, for Set B $g^* = 0.89$, and for Set C $g^* = 0.70$. Again, for many scientists, this meta-analytic information hardly helps identify the valid experiment set.

The Test for Excess Success is useful here. The calculations are left as an exercise for the reader, but computing power by using the meta-analytic effect size for each set, and multiplying the power values suggests that the probability of all five experiments producing significant results is: for Set A $p = 0.042$, for Set B $p = 0.45$, and for Set C $p = 0.052$. Indeed, it is Set B that is the valid experiment set.

The TES is a formal analysis, but there are rules of thumb that can be quickly used to gauge the validity of an experiment set. One thing to look at is the relationship between the sample size and the effect size. In a valid experiment set these numbers are unrelated (larger sample sizes lead to more precise estimates of effect sizes, but do not affect the magnitude of the effect size). For Set B in Table 11.1, a correlation between the sample size and the effect size gives a modest $r = 0.25$, which reflects random variation in the effect sizes across the experiments. In contrast for Sets A and C, $r = -0.86$ and $r = -0.83$, respectively. This relationship is easily understood when optional stopping is involved: a sample can be large only if the effect size happens to be small (if the estimated effect size were large a small sample would have been significant). One sees similar relationships in other sets of experiments, for example, the studies purporting to find evidence of precognition in Table 10.1 of Chap. 10 have a correlation between sample size and effect size of $r = -0.89$.

Another marker of a problematic data set is having many p-values close to, but always below, the criterion for statistical significance. Experiments run with optional stopping very often produce statistics with a p-value just below the criterion. In contrast, valid experiments with a real effect and appropriate sample sizes generally produce very small p-values; and values close to the criterion should be rare. One can see that Set B in Table 11.1 has almost all very tiny p-values, while the other experiment sets have many p-values between 0.02 and 0.05. Such a distribution of p-values should be a flag that something is odd about the set of experiments.

11.2 How Extensive Are These Problems?

So far we have established that some experiment sets have results that seem too good to be true, and that such findings should undermine our confidence in the validity of the original conclusions. The existence of some problematic experiment sets does not,

Table 11.2 Results of the TES analysis for articles in *Science*

Year	Short title	Success probability
2006	Deliberation-Without-Attention Effect	0.051
2006	Psychological Consequences of Money	0.002
2006	Washing Away Your Sins	0.095
2007	Perception of Goal-Directed Action in Primates	0.031
2008	Lacking Control Increases Illusory Pattern Perception	0.008
2009	Effect of Color on Cognitive Performance	0.002
2009	Monkeys Display Affiliation Toward Imitators	0.037
2009	Race Bias via Televised Nonverbal Behavior	0.027
2010	Incidental Haptic Sensations Influence Decisions	0.017
2010	Optimally Interacting Minds	0.332
2010	Susceptibility to Others' Beliefs in Infants and Adults	0.021
2010	Imagined Consumption Reduces Actual Consumption	0.012
2011	Promoting the Middle East Peace Process	0.210
2011	Writing About Worries Boosts Exam Performance	0.059
2011	Disordered Contexts Promote Stereotyping	0.075
2012	Analytic Thinking Promotes Religious Disbelief	0.051
2012	Stop Signals Provide Inhibition in Honeybee Swarms	0.957
2012	Some Consequences of Having Too Little	0.091

however, indicate that these kinds of problems are pervasive; it could be that such problems are rare. While we might have concerns about the specific studies that seem too good to be true, we would not necessarily worry about the entire field.

A way of examining the extent of these kinds of problems is to systematically analyze a specified set of studies. *Science* is one of the top academic journals; it has over 100,000 subscribers and is a major stepping-stone for any young scientist hoping to land a tenure-track position or be approved for tenure. One might hope that such a journal publishes the best work in any given field, especially given its very low acceptance rate of around 7%. The journal's on-line search tool reported 133 research articles that were classified as psychology or education and were published between 2005 and 2012. We applied the TES analysis to each of the 18 articles that had four or more experiments and provided sufficient information to estimate success probabilities.

Table 11.2 reports the estimated success probabilities for these 18 studies. Surprisingly, 15 out of 18 (83%) of the Science articles reported results that seem too good to be true (i.e., success probability is less than 0.1). The reader will probably recognize several of the short titles in Table 11.2 because many of these findings were described in the popular press and some have been the basis for policy decisions regarding education, charity, and dieting.

One study in Table 11.2 ("Disordered Contexts Promote Stereotyping") deserves special discussion. The lead author on this study was Diederik Stapel, a Dutch social psychologist who was found guilty of publishing fraudulent data. Indeed, the data in his *Science* paper was not gathered in a real experiment but was generated with a spreadsheet by the lead author (the other author was unaware of the fraud). You might think that a fraudster would insure that the data looked believable, but the reported (fake!) findings actually seem too good to be true. Very likely Stapel generated fake data that looked like real data from published experiments; unfortunately, the (presumably real) published data also often seems to be too good to be true.

The pattern of results in *Science* does not seem to be unique. A TES analysis for articles in the journal *Psychological Science* found a similar rate of excess success (36 out of 44, 82%, seem too good to be true). The problems do not seem to be restricted to psychology, as some papers on epigenetics and neuroscience show similar problems.

The overall implication of the analyses in Table 11.2 and other similar studies is that top scientists, editors, and reviewers do not understand what good scientific data looks like when an investigation involves multiple experiments and tests. At best, much of what is considered top experimental work in psychology, and other fields that depend on statistics, will probably prove unreplicable with similar kinds of experiments and analyses.

11.3 What Is Going On?

At this point it might be prudent to step back and consider how science got into its current situation. Certainly there is much pressure for scientists to report successful experimental outcomes (how else to get a faculty position or a grant?), but most (at least many) scientists seem to genuinely care about their field of research and they believe they are reporting valid and important findings that can have a positive impact on society. The implication seems to be that many scientists do not understand how statistical analyses contribute to interpretations of their empirical data. The following discussion is necessarily speculative, but it seems worthwhile to discuss some common confusions.

11.3.1 Misunderstanding Replication

As mentioned in Chap. 10, successful replication is often seen as the "gold standard" for scientific work. What many scientists seem to fail to appreciate, though, is that proper experiment sets show successful replication at a rate that matches experimental power (success probabilities, more generally). An emphasis on replication success blinded scientists from noticing that published experiments with moderate or low power were nevertheless nearly always working. Just due to random sampling, experiments with low power should not always work. Here, are some reasons why low powered studies so frequently deliver significant results even though they should not.

11.3.2 Publication Bias

Scientists may report experimental outcomes that support a certain theoretical perspective and not report experimental outcomes that go against that perspective. If every experiment provided a clear answer to the scientific question, this kind of behavior would be a type of fraud. However, it is often difficult to know whether an experiment has "worked." Given the complexity of many experiments (e.g., a cell culture must grow properly before you can claim to show some inhibitory effect of a suspected chemical), there are many reasons an experiment can fail. Scientists may treat an experiment that fails to show a desired outcome as one that suffers from a methodological flaw rather than one that provides a negative answer to the research question. Some scientists may have many studies that were improperly labeled as "pilot studies" but should actually have been treated as negative answers.

11.3.3 Optional Stopping

Empirically focused sciences constantly look for more data. Such an approach is valuable, but it often conflicts with the characteristics of hypothesis testing. For example, we noted in Chap. 10 how optional stopping inflated the Type I error rate of hypothesis tests. This problem is very difficult to solve within the hypothesis testing framework. For example suppose a scientist notes a marginal ($p = 0.07$) result in Experiment 1 and decides to run a new Experiment 2 to check on the effect. It may sound like the scientist is doing careful work, however, this is not necessarily true. Suppose Experiment 1 produced a significant effect ($p = 0.03$), would the scientist still have run Experiment 2 as a second check? If not, then the scientist is essentially performing optional stopping across experiments, and the Type I error rate for any given experiment (or across experiments) is unknown.

Indeed, the problem with optional stopping is not the actual behavior preformed by the scientist (e.g., the study with a planned sample size gives $p = 0.02$) but with what he would have done if the result turned out differently (e.g., if the study with a planned sample size gives $p = 0.1$, he would have added 20 more subjects). More precisely, if you do not know what you would have done under all possible scenarios, then you cannot know the Type I error rate for your analysis.

11.3.4 Hypothesizing After the Results Are Known (HARKing)

What may be happening for some scientific investigations is that scientists gather data from many experiments and then try to put together a coherent story that binds together the different results. That may sound like good scientific practice because it stays close to the data, but this approach tends to produce theories that are *too* close to the data and end up tracking noise along with any signal. A post hoc story can almost always be created to

justify why an effect appears or disappears. Likewise, findings that do not fit into a story can be labeled as irrelevant and properly (in the mind of the scientist) discarded from the experiment set.

This kind of post hoc reasoning applies for measures within an experiment as well as across experiments. A scientist may run multiple measures, identify one that seems to work across multiple experiments and conclude that this measure is the best. Again, this may seem like good scientific practice (and it can be, if done properly), but it often leads to selection of one measure on the basis of random sampling variation. The other measures may have been just as good (or better), but happened to not show the effect (or maybe they properly showed that the effect was not there).

11.3.5 Flexibility in Analyses

Modern software programs allow scientists to try a wide variety of analyses in search of statistical significance. The data do not show a significant result? Try transforming the data with a logarithm, or take the inverse of the data values and try again. Still no significance? Try removing outliers that are greater than three standard deviations from the mean, or 2.5 standard deviations from the mean. Remove floor effects or ceiling effects (Chap. 2), or data from participants who do not meet some other criterion. If you have multiple measures in your experiment you can combine them in a wide variety of ways (average them, take the max, multiply them, do a principle components analysis). While exploring your data is a perfectly good scientific exercise for an exploration study, it increases your Type I error rate in proper experiments. For this reason, if you tried various analysis for your data and found for an analysis a significant result, you need to replicate the experiment with this analysis and an independent sample.

Standard choices in analysis seem to encourage this kind of flexibility. Recall from Sect. 6.7 that a 2×2 ANOVA will have a 14% chance of producing at least one significant result (a main effect or an interaction) for a truly null data set. You can counteract this property by having a good understanding of precisely which tests are appropriate for your investigation. You would then (properly) ignore the outcomes of other tests reported by the ANOVA.

11.3.6 Misunderstanding Prediction

Scientific arguments seem very convincing when a theory predicts a novel outcome that is then verified by experimental data. Indeed, many of the *Science* articles analyzed in Table 11.2 include a phrase similar to "as predicted by the theory there was a significant difference." Such statements are very strange on two levels. First, even if an effect is real a hypothesis test is not going to produce a significant result every time. Sampling variability means that there will be some data sets that do not show the effect. At best a theory can

only predict the probability of a significant result given a certain sample size. Second, for a theory to predict the probability of success (typically, this is power), the theory must indicate an effect size for a given experimental design and sample size(s). None of the articles analyzed in Table 11.2 included any discussion of theoretically predicted effect sizes or power.

What this means is that the phrase "as predicted by the theory there was a significant difference" is empty of content for those papers. There may be a theory, but it is not the kind of theory that is able to predict the probability of an experiment producing a significant result. (Presumably, if it were that kind of theory, the scientists working with it would have discussed those details.) So, there actually is no prediction at all. The field seems to be in the bizarre situation where theoretical predictions that are not actually predictions seem to work every time. It indicates success at what should be a fundamentally impossible task.

11.3.7 Sloppiness and Selective Double Checking

Mistakes are inevitable in any kind of activity as complicated as science. Data entry errors, calculation errors, copy-and-paste errors can all lead to wrong interpretations. Although scientists are trained to check and re-check everything, these types of errors seem to be very common in published work. A computer program called STATCHECK can analyze published papers to check whether the reported statistics make sense. For example, if a paper reports $t(29) = 2.2$, $p = 0.01$ then there is definitely an error because $t = 2.2$ and $df = 29$ corresponds to $p = 0.036$. In a study of thousands of psychology articles, STATCHECK found that roughly half of published articles have at least one error of this type. Nearly 10% of articles had at least one reporting error that changed whether a reported result was significant.

These kinds of reporting errors may be indicative of a more general degree of sloppiness in research practices. Such a lack of care can mean that even well-intentioned researchers report untrustworthy findings. Worse still, the degree of care in data handling may correspond to whether the reported results match researcher's hopes of expectations. For example, if due to some error in data entry, a t-test finds $t(45) = 1.8$, $p = 0.08$, the researcher may re-check the data entry procedure, find the mistake and re-run the analysis to get $t(45) = 2.3$, $p = 0.03$. On the other hand, if a data entry error leads to a significant outcome such as $t(52) = 2.4$, $p = 0.02$, then the researcher may not re-check the data entry procedure even though doing so might have revealed the error and produced a result like $t(52) = 1.7$, $p = 0.1$.

Because there are so many places in scientific investigations where errors can occur, it is easy for scientists to unintentionally bias their results by selectively re-checking undesirable outcomes and selectively trusting desirable outcomes.

Take Home Messages

1. There is too much replication in many fields including medicine, biology, psychology, and likely many more.
2. It seems that many scientists use techniques they should better avoid: optional stopping, publication bias, HARKing, flexibility in the analysis, and many more.