STATISTICS DONE WRONG

Alex Reinhart
Carnegie Mellon University

LibreTexts™

Statistics Done Wrong

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

TABLE OF CONTENTS

Acknowledgements

Thanks to Dr. James Scott, whose statistics course gave me the background necessary to write this; to Matthew Watson and CharonY, who gave invaluable feedback and suggestions as I wrote my drafts; to my parents, who gave suggestions and feedback; to Dr. Brent Iverson, whose seminar first motivated me to learn about statistical abuse; and to all the scientists and statisticians who have broken the rules and given me a reason to write.

Any errors in explanations are my own.

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

Copyright Note

SECTION OVERVIEW

Introduction

In the final chapter of his famous book *How to Lie with Statistics*, Darrell Huff tells us that "anything smacking of the medical profession" or published by scientific laboratories and universities is worthy of our trust – not unconditional trust, but certainly more trust than we'd afford the media or shifty politicians. After all, Huff filled an entire book with the misleading statistical trickery used in politics and the media, but few people complain about statistics done by trained professional scientists. Scientists seek understanding, not ammunition to use against political opponents.

Statistical data analysis is fundamental to science. Open a random page in your favorite medical journal and you'll be deluged with statistics: t tests, p values, proportional hazards models, risk ratios, logistic regressions, least-squares fits, and confidence intervals. Statisticians have provided scientists with tools of enormous power to find order and meaning in the most complex of datasets, and scientists have embraced them with glee.

They have not, however, embraced statistics *education*, and many undergraduate programs in the sciences require no statistical training whatsoever.

Since the 1980s, researchers have described numerous statistical fallacies and misconceptions in the popular peer-reviewed scientific literature, and have found that many scientific papers – perhaps more than half – fall prey to these errors. Inadequate statistical power renders many studies incapable of finding what they're looking for; multiple comparisons and misinterpreted p values cause numerous false positives; flexible data analysis makes it easy to find a correlation where none exists. The problem isn't fraud but poor statistical education – poor enough that some scientists conclude that most published research findings are probably false.[31]

What follows is a list of the more egregious statistical fallacies regularly committed in the name of science. It assumes no knowledge of statistical methods, since many scientists receive no formal statistical training. And be warned: once you learn the fallacies, you will see them *everywhere*. Don't be alarmed. This isn't an excuse to reject all modern science and return to bloodletting and leeches – it's a call to improve the science we rely on.

CHAPTER OVERVIEW

1: An Introduction to Data Analysis

1.1: Data Analysis

Much of experimental science comes down to measuring changes. Does one medicine work better than another? Do cells with one version of a gene synthesize more of an enzyme than cells with another version? Does one kind of signal processing algorithm detect pulsars better than another? Is one catalyst more effective at speeding a chemical reaction than another?

Much of statistics, then, comes down to making judgments about these kinds of differences. We talk about “statistically significant differences” because statisticians have devised ways of telling if the difference between two measurements is really big enough to ascribe to anything but chance.

Suppose you’re testing cold medicines. Your new medicine promises to cut the duration of cold symptoms by a day. To prove this, you find twenty patients with colds and give half of them your new medicine and half a placebo. Then you track the length of their colds and find out what the average cold length was with and without the medicine.

But all colds aren’t identical. Perhaps the average cold lasts a week, but some last only a few days, and others drag on for two weeks or more, straining the household Kleenex supply. It’s possible that the group of ten patients receiving genuine medicine will be the unlucky types to get two-week colds, and so you’ll falsely conclude that the medicine makes things worse. How can you tell if you’ve proven your medicine works, rather than just proving that some patients are unlucky?

1.2: The Power of p Values

Statistics provides the answer. If we know the *distribution* of typical cold cases – roughly how many patients tend to have short colds, or long colds, or average colds – we can tell how likely it is for a random sample of cold patients to have cold lengths all shorter than average, or longer than average, or exactly average. By performing a statistical test, we can answer the question "If my medication were completely ineffective, what are the chances I'd see data like what I saw?"

That's a bit tricky, so read it again.

Intuitively, we can see how this might work. If I only test the medication on one person, it's unsurprising if he has a shorter cold than average – about half of patients have colds shorter than average. If I test the medication on ten million patients, it's pretty damn unlikely that *all* of them will have shorter colds than average, *unless my medication works.*

The common statistical tests used by scientists produce a number called the p value that quantifies this. Here's how it's defined:

> *The P value is defined as the probability, under the assumption of no effect or no difference (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.*[24]

So if I give my medication to 100 patients and find that their colds are a day shorter on average, the p value of this result is the chance that, if my medication didn't do anything at all, my 100 patients would randomly have, on average, day-or-more-shorter colds. Obviously, the p value depends on the size of the effect – colds shorter by four days are less likely than colds shorter by one day – and the number of patients I test the medication on.

That's a tricky concept to wrap your head around. A p value is not a measure of how right you are, or how significant the difference is; it's a measure of *how surprised you should be* if there is no actual difference between the groups, but you got data suggesting there is. A bigger difference, or one backed up by more data, suggests more surprise and a smaller p value.

It's not easy to translate that into an answer to the question "is there really a difference?" Most scientists use a simple rule of thumb: if p is less than 0.05, there's only a 5% chance of obtaining this data unless the medication really works, so we will call the difference between medication and placebo "significant." If p is larger, we'll call the difference insignificant.

But there are limitations. The p value is a measure of surprise, not a measure of the size of the effect. I can get a tiny p value by either measuring a huge effect – "this medicine makes people live four times longer" – or by measuring a tiny effect with great certainty. Statistical significance does not mean your result has any *practical* significance.

Similarly, statistical *in*significance is hard to interpret. I could have a perfectly good medicine, but if I test it on ten people, I'd be hard-pressed to tell the difference between a real improvement in the patients and plain good luck. Alternately, I might test it on thousands of people, but the medication only shortens colds by three minutes, and so I'm simply incapable of detecting the difference. A statistically insignificant difference does not mean there is no difference at all.

There's no mathematical tool to tell you if your hypothesis is true; you can only see whether it is consistent with the data, and if the data is sparse or unclear, your conclusions are uncertain.

But we can't let that stop us.

CHAPTER OVERVIEW

2: Statistical Power and Underpowered Statistics

2.1: Statistical Power

We've seen that it's possible to miss a real effect simply by not taking enough data. In most cases, this is a problem: we might miss a viable medicine or fail to notice an important side-effect. How do we know how much data to collect?

Statisticians provide the answer in the form of "statistical power." The power of a study is the likelihood that it will distinguish an effect of a certain size from pure luck. A study might easily detect a huge benefit from a medication, but detecting a subtle difference is much less likely. Let's try a simple example.

Suppose a gambler is convinced that an opponent has an unfair coin: rather than getting heads half the time and tails half the time, the proportion is different, and the opponent is using this to cheat at incredibly boring coin-flipping games. How to prove it?

You can't just flip the coin a hundred times and count the heads. Even with a perfectly fair coin, you don't always get fifty heads:

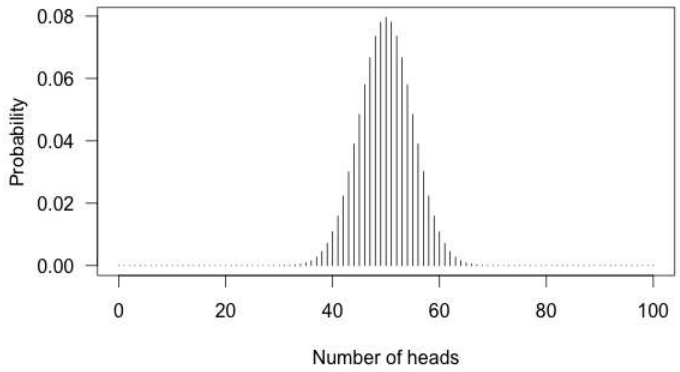

Figure 2.1.1: This shows the likelihood of getting different numbers of heads, if you flip a coin a hundred times.

You can see that 50 heads is the most likely option, but it's also reasonably likely to get 45 or 57. So if you get 57 heads, the coin might be rigged, but you might just be lucky.

Let's work out the math. Let's say we look for a p value of 0.05 or less, as scientists typically do. That is, if I count up the number of heads after 10 or 100 trials and find a deviation from what I'd expect – half heads, half tails – I call the coin unfair if there's only a 5% chance of getting a deviation that size or larger with a fair coin. Otherwise, I can conclude nothing: the coin may be fair, or it may be only a little unfair. I can't tell.

So, what happens if I flip a coin ten times and apply these criteria?

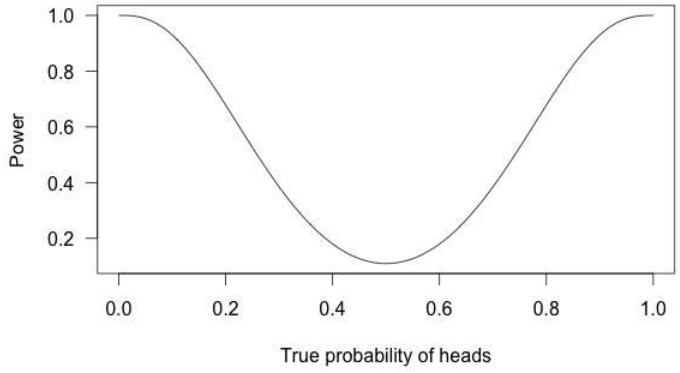

Figure 2.1.2

This is called a *power curve.* Along the horizontal axis, we have the different possibilities for the coin's true probability of getting heads, corresponding to different levels of unfairness. On the vertical axis is the probability that I will conclude the coin is rigged

after ten tosses, based on the p value of the result.

You can see that if the coin is rigged to give heads 60% of the time, and I flip the coin 10 times, I only have a 20% chance of concluding that it's rigged. There's just too little data to separate rigging from random variation. The coin would have to be incredibly biased for me to always notice.

But what if I flip the coin 100 times?

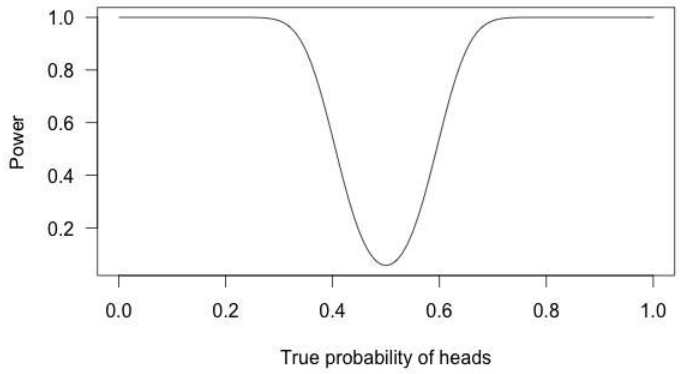

Figure 2.1.3

Or $1,000$ times?

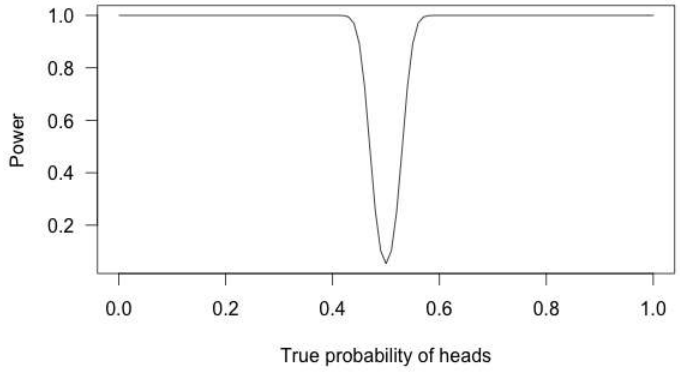

Figure 2.1.4

With one thousand flips, I can easily tell if the coin is rigged to give heads 60% of the time. It's just overwhelmingly unlikely that I could flip a fair coin $1,000$ times and get more than 600 heads.

2.2: The Power of Being Underpowered

After hearing all this, you might think calculations of statistical power are essential to medical trials. A scientist might want to know how many patients are needed to test if a new medication improves survival by more than 10%, and a quick calculation of statistical power would provide the answer. Scientists are usually satisfied when the statistical power is 0.8 or higher, corresponding to an 80% chance of concluding there's a real effect.

However, few scientists ever perform this calculation, and few journal articles ever mention the statistical power of their tests.

Consider a trial testing two different treatments for the same condition. You might want to know which medicine is safer, but unfortunately, side effects are rare. You can test each medicine on a hundred patients, but only a few in each group suffer serious side effects.

Obviously, you won't have terribly much data to compare side effect rates. If four people have serious side effects in one group, and three in the other, you can't tell if that's the medication's fault.

Unfortunately, many trials conclude with "There was no statistically significant difference in adverse effects between groups" without noting that there was insufficient data to detect any but the largest differences.[57] And so doctors erroneously think the medications are equally safe, when one could well be much more dangerous than the other.

You might think this is only a problem when the medication only has a weak effect. But no: in one sample of studies published between 1975 and 1990 in prestigious medical journals, 27% of randomized controlled trials gave negative results, but 64% of these didn't collect enough data to detect a 50% difference in *primary outcome* between treatment groups. Fifty percent! Even if one medication decreases symptoms by 50% more than the other medication, there's insufficient data to conclude it's more effective. And 84% of the negative trials didn't have the power to detect a 25% difference.[17, 4, 11, 16]

In neuroscience the problem is even worse. Suppose we aggregate the data collected by numerous neuroscience papers investigating one particular effect and arrive at a strong estimate of the effect's size. The median study has only a 20% chance of being able to detect that effect. Only after many studies were aggregated could the effect be discerned. Similar problems arise in neuroscience studies using animal models – which raises a significant ethical concern. If each individual study is underpowered, the true effect will only likely be discovered after many studies using many animals have been completed and analyzed, using far more animal subjects than if the study had been done properly the first time.[12]

That's not to say scientists are lying when they state they detected no significant difference between groups. You're just misleading yourself when you assume this means there is no *real* difference. There may be a difference, but the study was too small to notice it.

Let's consider an example we see every day.

2.3: The Wrong Turn on Red

In the 1970s, many parts of the United States began to allow drivers to turn right at a red light. For many years prior, road designers and civil engineers argued that allowing right turns on a red light would be a safety hazard, causing many additional crashes and pedestrian deaths. But the 1973 oil crisis and its fallout spurred politicians to consider allowing right turn on red to save fuel wasted by commuters waiting at red lights.

Several studies were conducted to consider the safety impact of the change. For example, a consultant for the Virginia Department of Highways and Transportation conducted a before-and-after study of twenty intersections which began to allow right turns on red. Before the change there were 308 accidents at the intersections; after, there were 337 in a similar length of time. However, this difference was not statistically significant, and so the consultant concluded there was no safety impact.

Several subsequent studies had similar findings: small increases in the number of crashes, but not enough data to conclude these increases were significant. As one report concluded,

> *There is no reason to suspect that pedestrian accidents involving RT operations (right turns) have increased after the adoption of [right turn on red]…*

Based on this data, more cities and states began to allow right turns at red lights. The problem, of course, is that these studies were underpowered. More pedestrians were being run over and more cars were involved in collisions, but nobody collected enough data to show this conclusively until several years later, when studies arrived clearly showing the results: significant increases in collisions and pedestrian accidents (sometimes up to 100% increases).[27, 48] The misinterpretation of underpowered studies cost lives.

3: Pseudoreplication- Choose Your Data Wisely

Many studies strive to collect more data through replication: by repeating their measurements with additional patients or samples, they can be more certain of their numbers and discover subtle relationships that aren't obvious at first glance. We've seen the value of additional data for improving statistical power and detecting small differences. But what exactly counts as a replication?

Let's return to a medical example. I have two groups of 100 patients taking different medications, and I seek to establish which medication lowers blood pressure more. I have each group take the medication for a month to allow it to take effect, and then I follow each group for ten days, each day testing their blood pressure. I now have ten data points per patient and $1,000$ data points per group.

Brilliant! $1,000$ data points is quite a lot, and I can fairly easily establish whether one group has lower blood pressure than the other. When I do calculations for statistical significance I find significant results very easily.

But wait: we expect that taking a patient's blood pressure ten times will yield ten very similar results. If one patient is genetically predisposed to low blood pressure, I have counted his genetics ten times. Had I collected data from $1,000$ independent patients instead of repeatedly testing 100, I would be more confident that differences between groups came from the medicines and not from genetics and luck. I claimed a large sample size, giving me statistically significant results and high statistical power, but my claim is unjustified.

This problem is known as pseudoreplication, and it is quite common.[38] After testing cells from a culture, a biologist might "replicate" his results by testing more cells from the same culture. Neuroscientists will test multiple neurons from the same animal, incorrectly claiming they have a large sample size because they tested hundreds of neurons from just two rats.

In statistical terms, pseudoreplication occurs when individual observations are heavily dependent on each other. Your measurement of a patient's blood pressure will be highly related to his blood pressure yesterday, and your measurement of soil composition here will be highly correlated with your measurement five feet away. There are several ways to account for this dependence while performing your statistical analysis:

1. Average the dependent data points. For example, average all the blood pressure measurements taken from a single person. This isn't perfect, though; if you measured some patients more frequently than others, this won't be reflected in the averaged number. You want a method that somehow counts measurements as more reliable as more are taken.
2. Analyze each dependent data point separately. You could perform an analysis of every patient's blood pressure on day 5, giving you only one data point per person. But be careful, because if you do this for every day, you'll have problems with multiple comparisons, which we will discuss in the next chapter.
3. Use a statistical model which accounts for the dependence, like a hierarchical model or random effects model.

It's important to consider each approach before analyzing your data, as each method is suited to different situations. Pseudoreplication makes it easy to achieve significance, even though it gives you little additional information on the test subjects. Researchers must be careful not to artificially inflate their sample sizes when they retest samples.

CHAPTER OVERVIEW

4: The p Value and the Base Rate Fallacy

4.1: Prelude to p Values

You've already seen that p values are hard to interpret. Getting a statistically insignificant result doesn't mean there's no difference. What about getting a significant result?

Let's try an example. Suppose I am testing a hundred potential cancer medications. Only ten of these drugs actually work, but I don't know which; I must perform experiments to find them. In these experiments, I'll look for $p < 0.05$ gains over a placebo, demonstrating that the drug has a significant benefit.

To illustrate, each square in this grid represents one drug. The blue squares are the drugs that work:

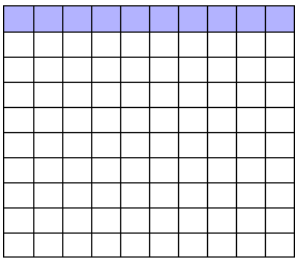

Figure 4.1.1

As we saw, most trials can't perfectly detect every good medication. We'll assume my tests have a statistical power of 0.8. Of the ten good drugs, I will correctly detect around eight of them, shown in purple:

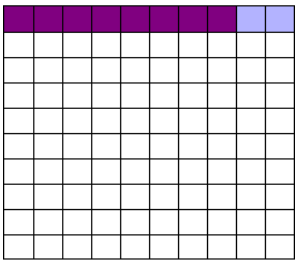

Figure 4.1.2

Of the ninety ineffectual drugs, I will conclude that about 5 have significant effects. Why? Remember that p values are calculated under the assumption of no effect, so $p = 0.05$ means a 5% chance of falsely concluding that an ineffectual drug works.

So I perform my experiments and conclude there are 13 working drugs: 8 good drugs and 5 I've included erroneously, shown in red:

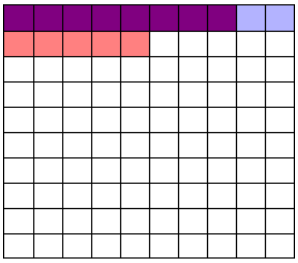

Figure 4.1.3

The chance of any given "working" drug being truly effectual is only 62%. If I were to randomly select a drug out of the lot of 100, run it through my tests, and discover a $p < 0.05$ statistically significant benefit, there is only a 62% chance that the drug is actually effective. In statistical terms, my false discovery rate – the fraction of statistically significant results which are really false positives – is 38%.

Because the *base rate* of effective cancer drugs is so low – only 10% of our hundred trial drugs actually work – most of the tested drugs do not work, and we have many opportunities for false positives. If I had the bad fortune of possessing a truckload of completely ineffective medicines, giving a base rate of 0%, there is a 0% chance that any statistically significant result is true. Nevertheless, I will get a $p < 0.05$ result for 5% of the drugs in the truck.

You often hear people quoting p values as a sign that error is unlikely. "There's only a 1 in $10,000$ chance this result arose as a statistical fluke," they say, because they got $p = 0.0001$. No! This ignores the base rate, and is called the *base rate fallacy*. Remember how p values are defined:

> *The P value is defined as the probability, under the assumption of no effect or no difference (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.*

A p value is calculated under the assumption that the medication *does not work* and tells us the probability of obtaining the data we did, or data more extreme than it. It does *not* tell us the chance the medication is effective.

When someone uses their p values to say they're probably right, remember this. Their study's probability of error is almost certainly much higher. In fields where most tested hypotheses are false, like early drug trials (most early drugs don't make it through trials), it's likely that *most* "statistically significant" results with $p < 0.05$ are actually flukes.

One good example is medical diagnostic tests.

4.2: The Base Rate Fallacy in Medical Testing

There has been some controversy over the use of mammograms in screening breast cancer. Some argue that the dangers of false positive results, such as unnecessary biopsies, surgery and chemotherapy, outweigh the benefits of early cancer detection. This is a statistical question. Let's evaluate it.

Suppose 0.8% of women who get mammograms have breast cancer. In 90% of women with breast cancer, the mammogram will correctly detect it. (That's the statistical power of the test. This is an estimate, since it's hard to tell how many cancers are missed if we don't know they're there.) However, among women with no breast cancer at all, about 7% will get a positive reading on the mammogram, leading to further tests and biopsies and so on. If you get a positive mammogram result, what are the chances you have breast cancer?

Ignoring the chance that you, the reader, are male,[1] the answer is 9%.[35]

Despite the test only giving false positives for 7% of cancer-free women, analogous to testing for $p < 0.07$, 91% of positive tests are false positives.

How did I calculate this? It's the same method as the cancer drug example. Imagine $1,000$ randomly selected women who choose to get mammograms. Eight of them (0.8%) have breast cancer. The mammogram correctly detects 90% of breast cancer cases, so about seven of the eight women will have their cancer discovered. However, there are 992 women without breast cancer, and 7% will get a false positive reading on their mammograms, giving us 70 women incorrectly told they have cancer.

In total, we have 77 women with positive mammograms, 7 of whom actually have breast cancer. Only 9% of women with positive mammograms have breast cancer.

If you administer questions like this one to statistics students and scientific methodology instructors, more than a third fail.[35] If you ask doctors, two thirds fail.[10] They erroneously conclude that a $p < 0.05$ result implies a 95% chance that the result is true – but as you can see in these examples, the likelihood of a positive result being true depends on *what proportion of hypotheses tested are true*. And we are very fortunate that only a small proportion of women have breast cancer at any given time.

Examine introductory statistical textbooks and you will often find the same error. P values are counterintuitive, and the base rate fallacy is everywhere.

Footnotes

[1] Interestingly, being male doesn't exclude you from getting breast cancer; it just makes it exceedingly unlikely.

4.3: Taking up Arms Against the Base Rate Fallacy

You don't have to be performing advanced cancer research or early cancer screenings to run into the base rate fallacy. What if you're doing social research? You'd like to survey Americans to find out how often they use guns in self-defense. Gun control arguments, after all, center on the right to self-defense, so it's important to determine whether guns are commonly used for defense and whether that use outweighs the downsides, such as homicides.

One way to gather this data would be through a survey. You could ask a representative sample of Americans whether they own guns and, if so, whether they've used the guns to defend their homes in burglaries or defend themselves from being mugged. You could compare these numbers to law enforcement statistics of gun use in homicides and make an informed decision about whether the benefits outweigh the downsides.

Such surveys have been done, with interesting results. One 1992 telephone survey estimated that American civilians use guns in self-defense up to 2.5 million times every year – that is, about 1% of American adults have defended themselves with firearms. Now, 34% of these cases were in burglaries, giving us $845,000$ burglaries stymied by gun owners. But in 1992, there were only 1.3 million burglaries committed while someone was at home. Two thirds of these occurred while the homeowners were asleep and were discovered only after the burglar had left. That leaves $430,000$ burglaries involving homeowners who were at home and awake to confront the burglar – $845,000$ of which, we are led to believe, were stymied by gun-toting residents.[28]

Whoops.

What happened? Why did the survey overestimate the use of guns in self-defense? Well, for the same reason that mammograms overestimate the incidence of breast cancer: there are far more opportunities for false positives than false negatives. If 99.9% of people have never used a gun in self-defense, but 1% of those people will answer "yes" to any question for fun, and 1% want to look manlier, and 1% misunderstand the question, then you'll end up *vastly* overestimating the use of guns in self-defense.

What about false negatives? Could this effect be balanced by people who say "no" even though they gunned down a mugger last week? No. If very few people genuinely use a gun in self-defense, then there are very few opportunities for false negatives. They're overwhelmed by the false positives.

This is exactly analogous to the cancer drug example earlier. Here, p is the probability that someone will falsely claim they've used a gun in self-defense. Even if p is small, your final answer will be wildly wrong.

To lower p, criminologists make use of more detailed surveys. The National Crime Victimization surveys, for instance, use detailed sit-down interviews with researchers where respondents are asked for details about crimes and their use of guns in self-defense. With far greater detail in the survey, researchers can better judge whether the incident meets their criteria for self-defense. The results are far smaller – something like $65,000$ incidents per year, not millions. There's a chance that survey respondents underreport such incidents, but a much smaller chance of massive overestimation.

4.4: If at First You Don't Succeed, Try, Try Again

The base rate fallacy shows us that false positives are much more likely than you'd expect from a $p < 0.05$ criterion for significance. Most modern research doesn't make one significance test, however; modern studies compare the effects of a variety of factors, seeking to find those with the most significant effects.

For example, imagine testing whether jelly beans cause acne by testing the effect of every single jelly bean color on acne:

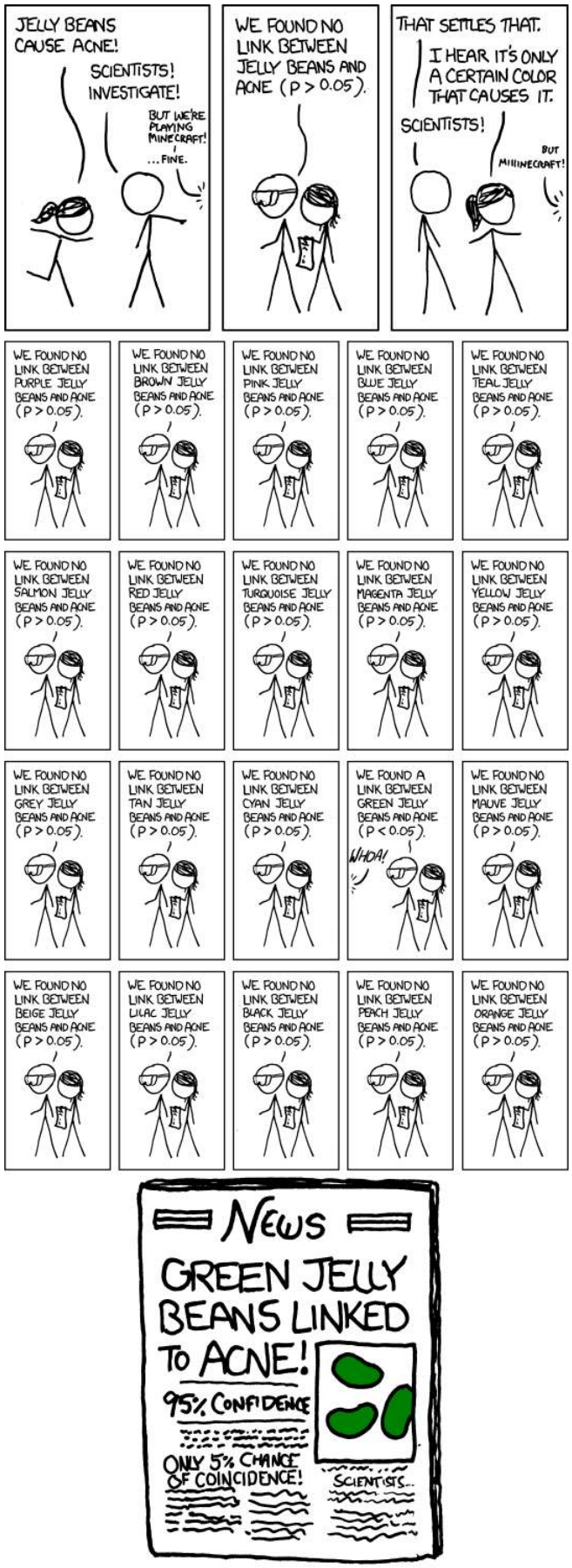

Figure 4.4.1: Cartoon from xkcd, by Randall Munroe. http://xkcd.com/882/

As you can see, making multiple comparisons means multiple chances for a false positive. For example, if I test 20 jelly bean flavors which do not cause acne at all, and look for a correlation at $p < 0.05$ significance, I have a 64% chance of a false positive result.[54] If I test 45 materials, the chance of false positive is as high as 90%.

It's easy to make multiple comparisons, and it doesn't have to be as obvious as testing twenty potential medicines. Track the symptoms of a dozen patients for a dozen weeks and test for significant benefits during any of those weeks: bam, that's twelve comparisons. Check for the occurrence of twenty-three potential dangerous side effects: alas, you have sinned. Send out a ten-page survey asking about nuclear power plant proximity, milk consumption, age, number of male cousins, favorite pizza topping, current sock color, and a few dozen other factors for good measure, and you'll find that *something* causes cancer. Ask enough questions and it's inevitable.

A survey of medical trials in the 1980s found that the average trial made 30 therapeutic comparisons. In more than half of the trials, the researchers had made so many comparisons that a false positive was highly likely, and the statistically significant results they did report were cast into doubt: they may have found a statistically significant effect, but it could just have easily been a false positive.[54]

There exist techniques to correct for multiple comparisons. For example, the Bonferroni correction method says that if you make n comparisons in the trial, your criterion for significance should be $p < 0.05/n$. This lowers the chances of a false positive to what you'd see from making only one comparison at $p < 0.05$. However, as you can imagine, this reduces statistical power, since you're demanding much stronger correlations before you conclude they're statistically significant. It's a difficult tradeoff, and tragically few papers even consider it.

4.5: Red Herrings in Brain Imaging

Neuroscientists do massive numbers of comparisons regularly. They often perform fMRI studies, where a three-dimensional image of the brain is taken before and after the subject performs some task. The images show blood flow in the brain, revealing which parts of the brain are most active when a person performs different tasks.

But how do you decide which regions of the brain are active during the task? A simple method is to divide the brain image into small cubes called voxels. A voxel in the "before" image is compared to the voxel in the "after" image, and if the difference in blood flow is significant, you conclude that part of the brain was involved in the task. Trouble is, there are thousands of voxels to compare and many opportunities for false positives.

One study, for instance, tested the effects of an "open-ended mentalizing task" on participants. Subjects were shown "a series of photographs depicting human individuals in social situations with a specified emotional valence," and asked to "determine what emotion the individual in the photo must have been experiencing." You can imagine how various emotional and logical centers of the brain would light up during this test.

The data was analyzed, and certain brain regions found to change activity during the task. Comparison of images made before and after the mentalizing task showed a $p = 0.001$ difference in a 81mm^3 cluster in the brain.

The study participants? Not college undergraduates paid $\$10$ for their time, as is usual. No, the test subject was one 3.8-pound Atlantic salmon, which "was not alive at the time of scanning."[8]

Of course, most neuroscience studies are more sophisticated than this; there are methods of looking for clusters of voxels which all change together, along with techniques for controlling the rate of false positives even when thousands of statistical tests are made. These methods are now widespread in the neuroscience literature, and few papers make such simple errors as I described. Unfortunately, almost every paper tackles the problem differently; a review of 241 fMRI studies found that they performed 223 unique analysis strategies, which, as we will discuss later, gives the researchers great flexibility to achieve statistically significant results.[13]

4.6: Controlling the False Discovery Rate

I mentioned earlier that techniques exist to correct for multiple comparisons. The Bonferroni procedure, for instance, says that you can get the right false positive rate by looking for $p < 0.05/n$, where n is the number of statistical tests you're performing. If you perform a study which makes twenty comparisons, you can use a threshold of $p < 0.0025$ to be assured that there is only a 5% chance you will falsely decide a nonexistent effect is statistically significant.

This has drawbacks. By lowering the p threshold required to declare a result statistically significant, you decrease your statistical power greatly, and fail to detect true effects as well as false ones. There are more sophisticated procedures than the Bonferroni correction which take advantage of certain statistical properties of the problem to improve the statistical power, but they are not magic solutions.

Worse, they don't spare you from the base rate fallacy. You can still be misled by your p threshold and falsely claim there's "only a 5% chance I'm wrong" – you just eliminate some of the false positives. A scientist is more interested in the false discovery rate: what fraction of my statistically significant results are false positives? Is there a statistical test that will let me control this fraction?

For many years the answer was simply "no." As you saw in the section on the base rate fallacy, we can compute the false discovery rate if we make an assumption about how many of our tested hypotheses are true – but we'd rather find that out from the data, rather than guessing.

In 1995, Benjamini and Hochberg provided a better answer. They devised an exceptionally simple procedure which tells you which p values to consider statistically significant. I've been saving you from mathematical details so far, but to illustrate just how simple the procedure is, here it is:

1. Perform your statistical tests and get the p value for each. Make a list and sort it in ascending order.
2. Choose a false-discovery rate and call it q. Call the number of statistical tests m.
3. Find the largest p value such that $p \leq iq/m$, where i is the p value's place in the sorted list.
4. Call that p value and all smaller than it statistically significant.

You're done! The procedure guarantees that out of all statistically significant results, no more than q percent will be false positives.[7]

The Benjamini-Hochberg procedure is fast and effective, and it has been widely adopted by statisticians and scientists in certain fields. It usually provides better statistical power than the Bonferroni correction and friends while giving more intuitive results. It can be applied in many different situations, and variations on the procedure provide better statistical power when testing certain kinds of data.

Of course, it's not perfect. In certain strange situations, the Benjamini-Hochberg procedure gives silly results, and it has been mathematically shown that it is always possible to beat it in controlling the false discovery rate. But it's a start, and it's much better than nothing.

CHAPTER OVERVIEW

5: When Differences in Significance Aren't Significant Differences

---

5.1: Differences in Significance

"We compared treatments A and B with a placebo. Treatment A showed a significant benefit over placebo, while treatment B had no statistically significant benefit. Therefore, treatment A is better than treatment B."

We hear this all the time. It's an easy way of comparing medications, surgical interventions, therapies, and experimental results. It's straightforward. It seems to make sense.

However, a difference in significance does not always make a significant difference.[22]

One reason is the arbitrary nature of the $p < 0.05$ cutoff. We could get two very similar results, with $p = 0.04$ and $p = 0.06$, and mistakenly say they're clearly different from each other simply because they fall on opposite sides of the cutoff. The second reason is that p values are not measures of effect size, so similar p values do not always mean similar effects. Two results with identical statistical significance can nonetheless contradict each other.

Instead, think about statistical power. If we compare our new experimental drugs Fixitol and Solvix to a placebo but we don't have enough test subjects to give us good statistical power, then we may fail to notice their benefits. If they have identical effects but we have only 50% power, then there's a good chance we'll say Fixitol has significant benefits and Solvix does not. Run the trial again, and it's just as likely that Solvix will appear beneficial and Fixitol will not.

Instead of independently comparing each drug to the placebo, we should compare them against each other. We can test the hypothesis that they are equally effective, or we can construct a confidence interval for the extra benefit of Fixitol over Solvix. If the interval includes zero, then they could be equally effective; if it doesn't, then one medication is a clear winner. This doesn't improve our statistical power, but it does prevent the false conclusion that the drugs are different. Our tendency to look for a difference in significance should be replaced by a check for the significance of the difference.

Examples of this error in common literature and news stories abound. A huge proportion of papers in neuroscience, for instance, commit the error.[44] You might also remember a study a few years ago suggesting that men with more biological older brothers are more likely to be homosexual.[9] How did they reach this conclusion? And why older brothers and not older sisters?

The authors explain their conclusion by noting that they ran an analysis of various factors and their effect on homosexuality. Only the number of older brothers had a statistically significant effect; number of older sisters, or number of nonbiological older brothers, had no statistically significant effect.

But as we've seen, that doesn't guarantee that there's a significant difference between the effects of older brothers and older sisters. In fact, taking a closer look at the data, it appears there's no statistically significant difference between the effect of older brothers and older sisters. Unfortunately, not enough data was published in the paper to allow a direct calculation.[22]

5.2: When Significant Differences are Missed

The problem can run the other way. Scientists routinely judge whether a significant difference exists simply by eye, making use of plots like this one:

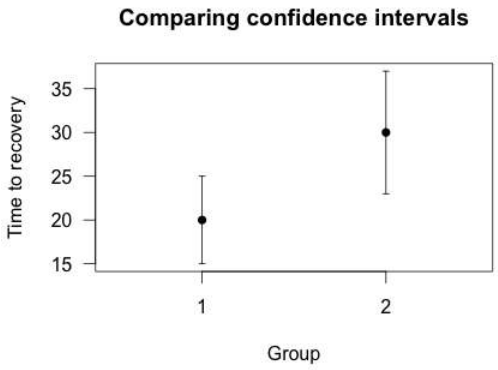

Figure 5.2.1

Imagine the two plotted points indicate the estimated time until recovery from some disease in two different groups of patients, each containing ten patients. There are three different things those error bars could represent:

1. The standard deviation of the measurements. Calculate how far each observation is from the average, square each difference, and then average the results and take the square root. This is the standard deviation, and it measures how spread out the measurements are from their mean.
2. The standard error of some estimator. For example, perhaps the error bars are the standard error of the mean. If I were to measure many different samples of patients, each containing exactly n subjects, I can estimate that 68% of the mean times to recover I measure will be within one standard error of "real" average time to recover. (In the case of estimating means, the standard error is the standard deviation of the measurements divided by the square root of the number of measurements, so the estimate gets better as you get more data – but not too fast.) Many statistical techniques, like least-squares regression, provide standard error estimates for their results.
3. The confidence interval of some estimator. A 95% confidence interval is mathematically constructed to include the true value for 95 random samples out of 100, so it spans roughly two standard errors in each direction. (In more complicated statistical models this may not be exactly true.)

These three options are all different. The standard deviation is a simple measurement of my data. The standard error tells me how a statistic, like a mean or the slope of a best-fit line, would likely vary if I take many samples of patients. A confidence interval is similar, with an additional guarantee that 95% of 95% confidence intervals should include the "true" value.

In the example plot, we have two 95% confidence intervals which overlap. Many scientists would view this and conclude there is no statistically significant difference between the groups. After all, groups 1 and 2 *might not* be different – the average time to recover could be 25 in both groups, for example, and the differences only appeared because group 1 was lucky this time. But does this mean the difference is not statistically significant? What would the p value be?

In this case, $p < 0.05$. There is a statistically significant difference between the groups, even though the confidence intervals overlap.[1]

Unfortunately, many scientists skip hypothesis tests and simply glance at plots to see if confidence intervals overlap. This is actually a much more conservative test – requiring confidence intervals to not overlap is akin to requiring $p < 0.01$ in some cases.[50] It is easy to claim two measurements are not significantly different even when they are.

Conversely, comparing measurements with standard errors or standard deviations will also be misleading, as standard error bars are shorter than confidence interval bars. Two observations might have standard errors which do not overlap, and yet the difference between the two is not statistically significant.

A survey of psychologists, neuroscientists and medical researchers found that the majority made this simple error, with many scientists confusing standard errors, standard deviations, and confidence intervals.[6] Another survey of climate science papers found

that a majority of papers which compared two groups with error bars made the error.[37] Even introductory textbooks for experimental scientists, such as *An Introduction to Error Analysis*, teach students to judge by eye, hardly mentioning formal hypothesis tests at all.

There are, of course, formal statistical procedures which generate confidence intervals which *can* be compared by eye, and even correct for multiple comparisons automatically. For example, Gabriel comparison intervals are easily interpreted by eye.[19]

Overlapping confidence intervals do not mean two values are not significantly different. Similarly, separated standard error bars do not mean two values *are* significantly different. It's always best to use the appropriate hypothesis test instead. Your eyeball is not a well-defined statistical procedure.

Footnotes

[1] This was calculated with an unpaired t test, based on a standard error of 2.5 in group 1 and 3.5 in group 2.

CHAPTER OVERVIEW

6: Stopping Rules and Regression to the Mean

6.1: Rules of the Game

Medical trials are expensive. Supplying dozens of patients with experimental medications and tracking their symptoms over the course of months takes significant resources, and so many pharmaceutical companies develop “stopping rules,” which allow investigators to end a study early if it’s clear the experimental drug has a substantial effect. For example, if the trial is only half complete but there’s already a statistically significant difference in symptoms with the new medication, the researchers may terminate the study, rather than gathering more data to reinforce the conclusion.

When poorly done, however, this can lead to numerous false positives.

For example, suppose we’re comparing two groups of patients, one with a medication and one with a placebo. We measure the level of some protein in their bloodstreams as a way of seeing if the medication is working. In this case, though, the medication causes no difference whatsoever: patients in both groups have the same average protein levels, although of course individuals have levels which vary slightly.

We start with ten patients in each group, and gradually collect more data from more patients. As we go along, we do a t test to compare the two groups and see if there is a statistically significant difference between average protein levels. We might see a result like this simulation:

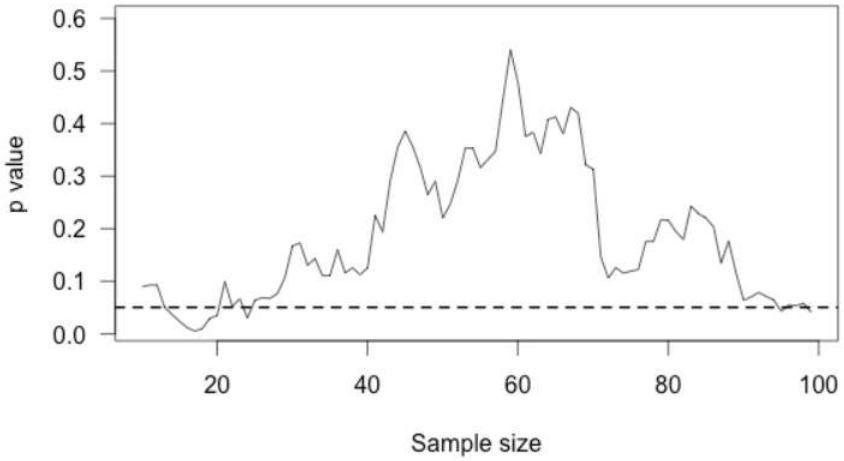

Figure 6.1.1

This plot shows the p value of the difference between groups as we collect more data, with the horizontal line indicating the $p = 0.05$ level of significance. At first, there appears to be no significant difference. Then we collect more data and conclude there is. If we were to stop, we’d be misled: we’d believe there is a significant difference between groups when there is none. As we collect yet more data, we realize we were mistaken – but then a bit of luck leads us back to a false positive.

You’d expect that the p value dip shouldn’t happen, since there’s no real difference between groups. After all, taking more data shouldn’t make our conclusions worse, right? And it’s true that if we run the trial again we might find that the groups start out with no significant difference and stay that way as we collect more data, or start with a huge difference and quickly regress to having none. But if we wait long enough and test after every data point, we will eventually cross *any* arbitrary line of statistical significance, even if there’s no real difference at all. We can’t usually collect infinite samples, so in practice this doesn’t always happen, but poorly implemented stopping rules still increase false positive rates significantly.[53]

Modern clinical trials are often required to register their statistical protocols in advance, and generally pre-select only a few evaluation points at which they test their evidence, rather than testing after every observation. This causes only a small increase in the false positive rate, which can be adjusted for by carefully choosing the required significance levels and using more advanced statistical techniques.[56] But in fields where protocols are not registered and researchers have the freedom to use whatever methods they feel appropriate, there may be false positive demons lurking.

6.2: Truth Inflation

Medical trials also tend to have inadequate statistical power to detect moderate differences between medications. So they want to stop as soon as they detect an effect, but they don't have the power to detect effects.

Suppose a medication reduces symptoms by 20% over a placebo, but the trial you're using to test it does not have adequate statistical power to detect this difference. We know that small trials tend to have varying results: it's easy to get ten lucky patients who have shorter colds than usual, but much harder to get ten thousand who all do.

Now imagine running many copies of this trial. Sometimes you get unlucky patients, and so you don't notice any statistically significant improvement from your drug. Sometimes your patients are exactly average, and the treatment group has their symptoms reduced by 20% – but you don't have enough data to call this a statistically significant increase, so you ignore it. Sometimes the patients are lucky and have their symptoms reduced by much more than 20%, and so you stop the trial and say "Look! It works!"

You've correctly concluded that your medication is effective, but you've inflated the size of its effect. You falsely believe it is much more effective than it really is.

This effect occurs in pharmacological trials, epidemiological studies, gene association studies ("gene A causes condition B"), psychological studies, and in some of the most-cited papers in the medical literature.[30, 32] In fields where trials can be conducted quickly by many independent researchers (such as gene association studies), the earliest published results are often wildly contradictory, because small trials and a demand for statistical significance cause only the most extreme results to be published.[33]

As a bonus, truth inflation can combine forces with early stopping rules. If most drugs in clinical trials are not quite so effective to warrant stopping the trial early, then many trials stopped early will be the result of lucky patients, not brilliant drugs – and by stopping the trial we have deprived ourselves of the extra data needed to tell the difference. Reviews have compared trials stopped early with other studies addressing the same question which did not stop early; in most cases, the trials stopped early exaggerated the effects of their tested treatments by an average of 29%.[3]

Of course, we do not know The Truth about any drug being studied, so we cannot tell if a particular study stopped early due to luck or a particularly good drug. Many studies do not even publish the original intended sample size or the stopping rule which was used to justify terminating the study.[43] A trial's early stoppage is not automatic evidence that its results are biased, but it *is* a suggestive detail.

6.3: Little Extremes

Suppose you're in charge of public school reform. As part of your research into the best teaching methods, you look at the effect of school size on standardized test scores. Do smaller schools perform better than larger schools? Should you try to build many small schools or a few large schools?

To answer this question, you compile a list of the highest-performing schools you have. The average school has about $1,000$ students, but the top-scoring five or ten schools are almost all smaller than that. It seems that small schools do the best, perhaps because of their personal atmosphere where teachers can get to know students and help them individually.

Then you take a look at the worst-performing schools, expecting them to be large urban schools with thousands of students and overworked teachers. Surprise! They're all small schools too.

What's going on? Well, take a look at a plot of test scores vs. school size:

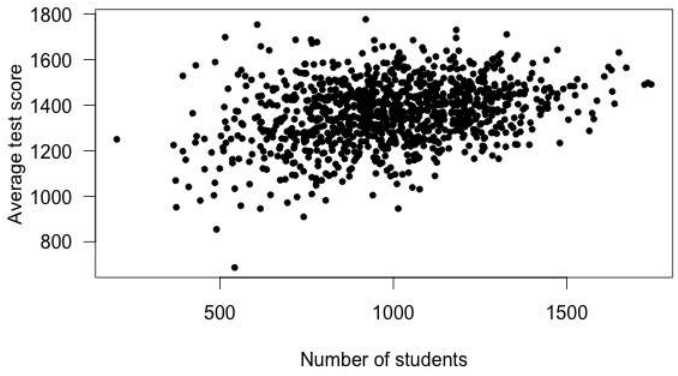

Figure 6.3.1

Smaller schools have more widely varying average test scores, entirely because they have fewer students. With fewer students, there are fewer data points to establish the "true" performance of the teachers, and so the average scores vary widely. As schools get larger, test scores vary less, and in fact *increase* on average.

This example used simulated data, but it's based on real (and surprising) observations of Pennsylvania public schools.[59]

Another example: In the United States, counties with the lowest rates of kidney cancer tend to be Midwestern, Southern and Western rural counties. How could this be? You can think of many explanations: rural people get more exercise, inhale less polluted air, and perhaps lead less stressful lives. Perhaps these factors lower their cancer rates.

On the other hand, counties with the highest rates of kidney cancer tend to be Midwestern, Southern and Western rural counties.

The problem, of course, is that rural counties have the smallest populations. A single kidney cancer patient in a county with ten residents gives that county the highest kidney cancer rate in the nation. Small counties hence have vastly more variable kidney cancer rates, simply because they have so few residents.[21]

7: Researcher Freedom- Good Vibrations?

There's a common misconception that statistics is boring and monotonous. Collect lots of data, plug the numbers into Excel or SPSS or R, and beat the software with a stick until it produces some colorful charts and graphs. Done! All the statistician must do is read off the results.

But one must choose *which* commands to use. Two researchers attempting to answer the same question may perform different statistical analyses entirely. There are many decisions to make:

1. Which variables do I adjust for? In a medical trial, for instance, you might control for patient age, gender, weight, BMI, previous medical history, smoking, drug use, or for the results of medical tests done before the start of the study. Which of these factors are important, and which can be ignored?
2. Which cases do I exclude? If I'm testing diet plans, maybe I want to exclude test subjects who came down with uncontrollable diarrhea during the trial, since their results will be abnormal.
3. What do I do with outliers? There will always be some results which are out of the ordinary, for reasons known or unknown, and I may want to exclude them or analyze them specially. Which cases count as outliers, and what do I do with them?
4. How do I define groups? For example, I may want to split patients into "overweight", "normal", and "underweight" groups. Where do I draw the lines? What do I do with a muscular bodybuilder whose BMI is in the "overweight" range?
5. What about missing data? Perhaps I'm testing cancer remission rates with a new drug. I run the trial for five years, but some patients will have tumors reappear after six years, or eight years. My data does not include their recurrence. How do I account for this when measuring the effectiveness of the drug?
6. How much data should I collect? Should I stop when I have a definitive result, or continue as planned until I've collected all the data?
7. How do I measure my outcomes? A medication could be evaluated with subjective patient surveys, medical test results, prevalence of a certain symptom, or measures such as duration of illness.

Producing results can take hours of exploration and analysis to see which procedures are most appropriate. Papers usually explain the statistical analysis performed, but don't always explain why the researchers chose one method over another, or explain what the results would be had the researchers chosen a different method. Researchers are free to choose whatever methods they feel appropriate – and while they may make the right choices, what would happen if they analyzed the data differently?

In simulations, it's possible to get effect sizes different by a factor of two simply by adjusting for different variables, excluding different sets of cases, and handling outliers differently.[30] The effect size is that all-important number which tells you how much of a difference your medication makes. So apparently, being free to analyze how you want gives you enormous control over your results!

The most concerning consequence of this statistical freedom is that researchers may choose the statistical analysis most favorable to them, arbitrarily producing statistically significant results by playing with the data until something emerges. Simulation suggests that false positive rates can jump to over 50% for a given dataset just by letting researchers try different statistical analyses until one works.[53]

Medical researchers have devised ways of preventing this. Researchers are often required to draft a clinical trial protocol, explaining how the data will be collected and analyzed. Since the protocol is drafted before the researchers see any data, they can't possibly craft their analysis to be most favorable to them. Unfortunately, many studies depart from their protocols and perform different analysis, allowing for researcher bias to creep in.[15, 14] Many other scientific fields have no protocol publication requirement at all.

The proliferation of statistical techniques has given us many useful tools, but it seems they have been put to use as blunt objects. One must simply beat the data until it confesses.

8: Everybody Makes Mistakes

Until now, I have presumed that scientists are capable of making statistical computations with perfect accuracy, and only err in their choice of appropriate numbers to compute. Scientists may misuse the results of statistical tests or fail to make relevant computations, but they can at least calculate a p value, right?

Perhaps not.

Surveys of statistically significant results reported in medical and psychological trials suggest that many p values are wrong, and some statistically insignificant results are actually significant when computed correctly.[25, 2] Other reviews find examples of misclassified data, erroneous duplication of data, inclusion of the wrong dataset entirely, and other mixups, all concealed by papers which did not describe their analysis in enough detail for the errors to be easily noticed.[1, 26]

Sunshine is the best disinfectant, and many scientists have called for experimental data to be made available through the Internet. In some fields, this is now commonplace: there exist gene sequencing databases, protein structure databanks, astronomical observation databases, and earth observation collections containing the contributions of thousands of scientists. Many other fields, however, can't share their data due to impracticality (particle physics data can include many terabytes of information), privacy issues (in medical trials), a lack of funding or technological support, or just a desire to keep proprietary control of the data and all the discoveries which result from it. And even if the data were all available, would anyone analyze it all to spot errors?

Similarly, scientists in some fields have pushed towards making their statistical analyses available through clever technological tools. A tool called Sweave, for instance, makes it easy to embed statistical analyses performed using the popular R programming language inside papers written in LaTeX, the standard for scientific and mathematical publications. The result looks just like any scientific paper, but another scientist reading the paper and curious about its methods can download the source code, which shows exactly how all the numbers were calculated. But would scientists avail themselves of the opportunity? Nobody gets scientific glory by checking code for typos.

Another solution might be replication. If scientists carefully recreate the experiments of other scientists and validate their results, it is much easier to rule out the possibility of a typo causing an errant result. Replication also weeds out fluke false positives. Many scientists claim that experimental replication is the heart of science: no new idea is accepted until it has been independently tested and retested around the world and found to hold water.

That's not entirely true; scientists often take previous studies for granted, though occasionally scientists decide to systematically re-test earlier works. One new project, for example, aims to reproduce papers in major psychology journals to determine just how many papers hold up over time – and what attributes of a paper predict how likely it is to stand up to retesting.[1] In another example, cancer researchers at Amgen retested 53 landmark preclinical studies in cancer research. (By "preclinical" I mean the studies did not involve human patients, as they were testing new and unproven ideas.) Despite working in collaboration with the authors of the original papers, the Amgen researchers could only reproduce six of the studies.[5] Bayer researchers have reported similar difficulties when testing potential new drugs found in published papers.[49]

This is worrisome. Does the trend hold true for less speculative kinds of medical research? Apparently so: of the top-cited research articles in medicine, a quarter have gone untested after their publication, and a third have been found to be exaggerated or wrong by later research.[32] That's not as extreme as the Amgen result, but it makes you wonder what important errors still lurk unnoticed in important research. Replication is not as prevalent as we would like it to be, and the results are not always favorable.

Footnotes

[1] The Reproducibility Project, at http://openscienceframework.org/reproducibility/

CHAPTER OVERVIEW

9: Hiding the Data

9.1: Handling Data

> *"Given enough eyeballs, all bugs are shallow."*
>
> *—Eric S. Raymond*

We've talked about the common mistakes made by scientists, and how the best way to spot them is a bit of outside scrutiny. Peer review provides some of this scrutiny, but a peer reviewer doesn't have the time to extensively re-analyze data and read code for typos – reviewers can only check that the methodology makes good sense. Sometimes they spot obvious errors, but subtle problems are usually missed.[52]

This is why many journals and professional societies require researchers to make their data available to other scientists on request. Full datasets are usually too large to print in the pages of a journal, so authors report their results and send the complete data to other scientists if they ask for a copy. Perhaps they will find an error or a pattern the original scientists missed.

Or so it goes in theory. In 2005, Jelte Wicherts and colleagues at the University of Amsterdam decided to analyze every recent article in several prominent journals of the American Psychological Association to learn about their statistical methods. They chose the APA partly because it requires authors to agree to share their data with other psychologists seeking to verify their claims.

Of the 249 studies they sought data for, they had only received data for 64 six months later. Almost three quarters of study authors never sent their data.[61]

Of course, scientists are busy people, and perhaps they simply didn't have the time to compile their datasets, produce documents describing what each variable means and how it was measured, and so on.

Wicherts and his colleagues decided they'd test this. They trawled through all the studies looking for common errors which could be spotted by reading the paper, such as inconsistent statistical results, misuse of various statistical tests, and ordinary typos. At least half of the papers had an error, usually minor, but 15% reported at least one statistically significant result which was only significant because of an error.

Next, they looked for a correlation between these errors and an unwillingness to share data. There was a clear relationship. Authors who refused to share their data were more likely to have committed an error in their paper, and their statistical evidence tended to be weaker.[60] Because most authors refused to share their data, Wicherts could not dig for deeper statistical errors, and many more may be lurking.

This is certainly not proof that authors hid their data out of fear their errors may be uncovered, or even that the authors knew about the errors at all. Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing "look over there."[1]

Footnotes

[1] Joke shamelessly stolen from the alternate text of http://xkcd.com/552/.

9.2: Just Leave out the Details

Nitpicking statisticians getting you down by pointing out flaws in your paper? There's one clear solution: don't publish as much detail! They can't find the errors if you don't say how you evaluated your data.

I don't mean to seriously suggest that evil scientists do this intentionally, although perhaps some do. More frequently, details are left out because authors simply forgot to include them, or because journal space limits force their omission.

It's possible to evaluate studies to see what they left out. Scientists leading medical trials are required to provide detailed study plans to ethical review boards before starting a trial, so one group of researchers obtained a collection of these plans from a review board. The plans specify which outcomes the study will measure: for instance, a study might monitor various symptoms to see if any are influenced by the treatment. The researchers then found the published results of these studies and looked for how well these outcomes were reported.

Roughly half of the outcomes never appeared in the scientific journal papers at all. Many of these were statistically insignificant results which were swept under the rug.[1] Another large chunk of results were not reported in sufficient detail for scientists to use the results for further meta-analysis.[14]

Other reviews have found similar problems. A review of medical trials found that most studies omit important methodological details, such as stopping rules and power calculations, with studies in small specialist journals faring worse than those in large general medicine journals.[29]

Medical journals have begun to combat this problem with standards for reporting of results, such as the CONSORT checklist. Authors are required to follow the checklist's requirements before submitting their studies, and editors check to make sure all relevant details are included. The checklist seems to work; studies published in journals which follow the guidelines tend to report more essential detail, although not all of it.[46] Unfortunately the standards are inconsistently applied and studies often slip through with missing details nonetheless.[42] Journal editors will need to make a greater effort to enforce reporting standards.

We see that published papers aren't faring very well. What about *unpublished* studies?

Footnotes

[1] Why do we always say "swept under the rug"? Whose rug is it? And why don't they use a vacuum cleaner instead of a broom?

9.3: Science in a Filing Cabinet

Earlier we saw the impact of multiple comparisons and truth inflation on study results. These problems arise when studies make numerous comparisons with low statistical power, giving a high rate of false positives and inflated estimates of effect sizes, and they appear everywhere in published research.

But not every study is published. We only ever see a fraction of medical research, for instance, because few scientists bother publishing "We tried this medicine and it didn't seem to work."

Consider an example: studies of the tumor suppressor protein TP53 and its effect on head and neck cancer. A number of studies suggested that measurements of TP53 could be used to predict cancer mortality rates, since it serves to regulate cell growth and development and hence must function correctly to prevent cancer. When all 18 published studies on TP53 and cancer were analyzed together, the result was a highly statistically significant correlation: TP53 could clearly be measured to tell how likely a tumor is to kill you.

But then suppose we dig up *unpublished* results on TP53: data that had been mentioned in other studies but not published or analyzed. Add this data to the mix and the statistically significant effect vanishes.[36] After all, few authors bothered to publish data showing no correlation, so the meta-analysis could only use a biased sample.

A similar study looked at reboxetine, an antidepressant sold by Pfizer. Several published studies have suggested that it is effective compared to placebo, leading several European countries to approve it for prescription to depressed patients. The German Institute for Quality and Efficiency in Health Care, responsible for assessing medical treatments, managed to get unpublished trial data from Pfizer – three times more data than had ever been published – and carefully analyzed it. The result: reboxetine is not effective. Pfizer had only convinced the public that it's effective by neglecting to mention the studies proving it isn't.[18]

This problem is commonly known as publication bias or the file-drawer problem: many studies sit in a file drawer for years, never published, despite the valuable data they could contribute.

The problem isn't simply the bias on published results. Unpublished studies lead to a duplication of effort – if other scientists don't know you've done a study, they may well do it again, wasting money and effort.

Regulators and scientific journals have attempted to halt this problem. The Food and Drug Administration requires certain kinds of clinical trials to be registered through their website ClinicalTrials.gov before the trials begin, and requires the publication of results within a year of the end of the trial. Similarly, the International Committee of Medical Journal Editors announced in 2005 that they would not publish studies which had not been pre-registered.

Unfortunately, a review of 738 registered clinical trials found that only 22% met the legal requirement to publish.[47] The FDA has not fined any drug companies for noncompliance, and journals have not consistently enforced the requirement to register trials. Most studies simply vanish.

10: What Have We Wrought?

I've painted a grim picture. But anyone can pick out small details in published studies and produce a tremendous list of errors. Do these problems matter?

Well, yes. I wouldn't have written this otherwise.

John Ioannidis's famous article "Why Most Published Research Findings are False"[31] was grounded in mathematical concerns rather than an empirical test of research results. If most research articles have poor statistical power – and they do – while researchers have the freedom to choose among multitudes of analyses methods to get favorable results – and they do – when most tested hypotheses are false and most true hypotheses correspond to very small effects, we are mathematically determined to get a multitude of false positives.

But if you want empiricism, you can have it, courtesy of John Ioannidis and Jonathan Schoenfeld. They studied the question "Is everything we eat associated with cancer?"[51][1] After choosing fifty common ingredients out of a cookbook, they set out to find studies linking them to cancer rates – and found 216 studies on forty different ingredients. Of course, most of the studies disagreed with each other. Most ingredients had multiple studies claiming they increased *and* decreased the risk of getting cancer. Most of the statistical evidence was weak, and meta-analyses usually showed much smaller effects on cancer rates than the original studies.

Of course, being contradicted by follow-up studies and meta-analyses doesn't prevent a paper from being cited as though it were true. Even effects which have been contradicted by massive follow-up trials with unequivocal results are frequently cited five or ten years later, with scientists apparently not noticing that the results are false.[55] Of course, new findings get widely publicized in the press, while contradictions and corrections are hardly ever mentioned.[23] You can hardly blame the scientists for not keeping up.

Let's not forget the merely biased results. Poor reporting standards in medical journals mean studies testing new treatments for schizophrenia can neglect to include the scale they used to evaluate symptoms – a handy source of bias, as trials using unpublished scales tend to produce better results than those using previously validated tests.[40] Other medical studies simply omit particular results if they're not favorable or interesting, biasing subsequent meta-analyses to only include positive results. A third of meta-analyses are estimated to suffer from this problem.[34]

Another review compared meta-analyses to subsequent large randomized controlled trials, considered the gold standard in medicine. In over a third of cases, the randomized trial's outcome did not correspond well to the meta-analysis.[39] Other comparisons of meta-analyses to subsequent research found that most results were inflated, with perhaps a fifth representing false positives.[45]

Let's not forget the multitude of physical science papers which misuse confidence intervals.[37] Or the peer-reviewed psychology paper allegedly providing evidence for psychic powers, on the basis of uncontrolled multiple comparisons in exploratory studies.[58] Unsurprisingly, results failed to be replicated – by scientists who appear not to have calculated the statistical power of their tests.[20]

We have a problem. Let's work on fixing it.

Footnotes

[1] An important part of the ongoing Oncological Ontology project to categorize everything into two categories: that which cures cancer and that which causes it.

CHAPTER OVERVIEW

11: What Can be Done?

I've discussed many statistical problems throughout this guide. They appear in many fields of science: medicine, physics, climate science, biology, chemistry, neuroscience, and many others. Any researcher using statistical methods to analyze data is likely to make a mistake, and as we've seen, most of them do. What can we do about it?

11.1: Statistical Education

Most American science students have a minimal statistical education – perhaps one or two required courses, or even none at all for many students. And even when students have taken statistical courses, professors report that they can't apply statistical concepts to scientific questions, having never fully understood – or simply forgotten – the appropriate techniques. This needs to change. Almost every scientific discipline depends on statistical analysis of experimental data, and statistical errors waste grant funding and researcher time.

Some universities have experimented with statistics courses integrated with science classes, with students immediately applying their statistical knowledge to problems in their field. Preliminary results suggests these methods work: students learn and retain more statistics, and they spend less time whining about being forced to take a statistics course.[41] More universities should adopt these techniques, using conceptual tests to see what methods work best.

We also need more freely available educational material. I was introduced to statistics when I needed to analyze data in a laboratory and didn't know how; until strong statistics education is more widespread, many students will find themselves in the same position, and they need resources. Projects like OpenIntro Stats are promising, and I hope to see more in the near future.

11.2: Scientific Publishing

Scientific journals are slowly making progress towards solving many of the problems I have discussed. Reporting guidelines, such as CONSORT for randomized trials, make it clear what information is required for a published paper to be reproducible; unfortunately, as we've seen, these guidelines are infrequently enforced. We must continue to pressure journals to hold authors to more rigorous standards.

Premier journals need to lead the charge. *Nature* has begun to do so, announcing a new checklist which authors are required to complete before articles may be published. The checklist requires reporting of sample sizes, statistical power calculations, clinical trial registration numbers, a completed CONSORT checklist, adjustment for multiple comparisons, and sharing of data and source code. The guidelines cover most issues covered in *Statistics Done Wrong*, except for stopping rules and discussion of any reasons for departing from the trial's registered protocol. *Nature* will also make statisticians available to consult for papers as needed.

If these guidelines are enforced, the result will be much more reliable and reproducible scientific research. More journals should do the same.

11.3: Your Job

Your task can be expressed in four simple steps:

1. Read a statistics textbook or take a good statistics course. Practice.
2. Plan your data analyses carefully and deliberately, avoiding the misconceptions and errors you have learned.
3. When you find common errors in the scientific literature – such as a simple misinterpretation of p values – hit the perpetrator over the head with your statistics textbook. It's therapeutic.
4. Press for change in scientific education and publishing. It's our research. Let's not screw it up.

12: Conclusion

Beware false confidence. You may soon develop a smug sense of satisfaction that *your* work doesn't screw up like everyone else's. But I have not given you a thorough introduction to the mathematics of data analysis. There are many ways to foul up statistics beyond these simple conceptual errors.

Errors will occur often, because somehow, few undergraduate science degrees or medical schools require courses in statistics and experimental design – and some introductory statistics courses skip over issues of statistical power and multiple inference. This is seen as acceptable despite the paramount role of data and statistical analysis in the pursuit of modern science; we wouldn't accept doctors who have no experience with prescription medication, so why do we accept scientists with no training in statistics? Scientists need formal statistical training and advice. To quote:

> *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."*
>
> *—R. A. Fisher, popularizer of the p value*

Journals may choose to reject research with poor-quality statistical analyses, and new guidelines and protocols may eliminate some problems, but until we have scientists adequately trained in the principles of statistics, experimental design and data analysis will not be improved. The all-consuming quest for statistical significance will only continue.

Change will not be easy. Rigorous statistical standards don't come free: if scientists start routinely performing statistical power computations, for example, they'll soon discover they need vastly larger sample sizes to reach solid conclusions. Clinical trials are not free, and more expensive research means fewer published trials. You might object that scientific progress will be slowed needlessly – but isn't it worse to build our progress on a foundation of unsound results?

To any science students: invest in a statistics course or two while you have the chance. To researchers: invest in training, a good book, and statistical advice. And please, the next time you hear someone say "The result was significant with $p < 0.05$, so there's only a 1 in 20 chance it's a fluke!", please beat them over the head with a statistics textbook for me.

Disclaimer: The advice in this guide cannot substitute for the advice of a trained statistical professional. If you think you're suffering from any serious statistical error, please consult a statistician immediately. I shall not have any liability from any injury to your dignity, statistical error or misconception suffered as a result of your use of this website.

Use of this guide to justify rejecting the results of a scientific study without reviewing the evidence in any detail whatsoever is grounds for being slapped upside the head with a very large statistics textbook. This guide should help you find statistical errors, not allow you to selectively ignore science you don't like.

Index

D

dire

Glossary

Sample Word 1 | Sample Definition 1

Bibliography

[1] K. A. Baggerly, K. R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 3:1309–1334, 2009.

[2] M. Bakker, J. M. Wicherts. The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43:666–678, 2011.

[3] D. Bassler, M. Briel, V. M. Montori, M. Lane, P. Glasziou, Q. Zhou, D. Heels-Ansdell, S. D. Walter, G. H. Guyatt. Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects: Systematic Review and Meta-regression Analysis. *JAMA*, 303:1180–1187, 2010.

[4] P. L. Bedard, M. K. Krzyzanowska, M. Pintilie, I. F. Tannock. Statistical Power of Negative Randomized Controlled Trials Presented at American Society for Clinical Oncology Annual Meetings. *Journal of Clinical Oncology*, 25:3482–3487, 2007.

[5] C. G. Begley, L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483:531–533, 2012.

[6] S. Belia, F. Fidler, J. Williams, G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10:389–396, 2005.

[7] Y. Benjamini, Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 289–300, 1995.

[8] C. Bennett, A. Baird, M. Miller, G. Wolford. Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results*, 1:1–5, 2010.

[9] A. F. Bogaert. Biological versus nonbiological older brothers and men's sexual orientation. *PNAS*, 103:10771–10774, 2006.

[10] R. Bramwell, H. West. Health professionals' and service users' interpretation of screening test results: experimental study. *BMJ*, 2006.

[11] C. G. Brown, G. D. Kelen, J. J. Ashton, H. A. Werman. The beta error and sample size determination in clinical trials in emergency medicine. *Annals of Emergency Medicine*, 16:183–187, 1987.

[12] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, M. R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 2013.

[13] J. Carp. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*, 63:289–300, 2012.

[14] A.-W. Chan, A. Hróbjartsson, M. T. Haahr, P. C. Gøtzsche, D. G. Altman. Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles. *JAMA*, 291:2457–2465, 2004.

[15] A.-W. Chan, A. Hróbjartsson, K. J. Jørgensen, P. C. Gøtzsche, D. G. Altman. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ*, 337:a2299, 2008.

[16] K. C. Chung, L. K. Kalliainen, R. A. Hayward. Type II (beta) errors in the hand literature: the importance of power. *The Journal of Hand Surgery*, 23:20–25, 1998.

[17] M. D, D. CS, W. GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*, 272:122-124, 1994.

[18] D. Eyding, M. Lelgemann, U. Grouven, M. Härter, M. Kromp, T. Kaiser, M. F. Kerekes, M. Gerken, B. Wieseler. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ*, 341:2010.

[19] K. R. Gabriel. A simple method of multiple comparisons of means. *Journal of the American Statistical Association*, 73:724–729, 1978.

[20] J. Galak, R. A. LeBoeuf, L. D. Nelson, J. P. Simmons. Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, 103:933–948, 2012.

[21] A. Gelman, P.N. Price. All maps of parameter estimates are misleading. *Statistics in Medicine*, 18:3221–3234, 1999.

[22] A. Gelman, H. Stern. The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. *The American Statistician*, 60:328–331, 2006.

[23] F. Gonon, J.-P. Konsman, D. Cohen, T. Boraud. Why Most Biomedical Findings Echoed by Newspapers Turn Out to be False: The Case of Attention Deficit Hyperactivity Disorder. *PLoS ONE*, 7:e44275, 2012.

[24] S. N. Goodman. Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, 130:995–1004, 1999.

[25] P. C. Gøtzsche. Believability of relative risks and odds ratios in abstracts: cross sectional study. *BMJ*, 333:231–234, 2006.

[26] P. C. Gøtzsche. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials*, 10:31–56, 1989.

[27] E. Hauer. The harm done by tests of significance. *Accident Analysis & Prevention*, 36:495–500, 2004.

[28] D. Hemenway. Survey Research and Self-Defense Gun Use: An Explanation of Extreme Overestimates. *The Journal of Criminal Law and Criminology*, 87:1430–1445, 1997.

[29] K. Huwiler-Müntener, P. Jüni, C. Junker, M. Egger. Quality of Reporting of Randomized Trials as a Measure of Methodologic Quality. *JAMA*, 287:2801–2804, 2002.

[30] J. P. A. Ioannidis. Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19:640–648, 2008.

[31] J. P. A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine*, 2:e124, 2005.

[32] J. P. A. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294:218–228, 2005.

[33] J. P. A. Ioannidis, T. A. Trikalinos. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, 58:543–549, 2005.

[34] J. J. Kirkham, K. M. Dwan, D. G. Altman, C. Gamble, S. Dodd, R. Smyth, P. R. Williamson. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, 340:c365–c365, 2010.

[35] W. Krämer, G. Gigerenzer. How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities. *Statistical Science*, 20:223–230, 2005.

[36] P. A. Kyzas, K. T. Loizou, J. P. A. Ioannidis. Selective Reporting Biases in Cancer Prognostic Factor Studies. *Journal of the National Cancer Institute*, 97:1043–1055, 2005.

[37] J. R. Lanzante. A cautionary note on the use of error bars. *Journal of climate*, 18:3699–3703, 2005.

[38] S. E. Lazic. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis?. *BMC Neuroscience*, 11:5, 2010.

[39] J. LeLorier, G. Gregoire, A. Benhaddad. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, 1997.

[40] M. Marshall, A. Lockwood, C. Bradley, C. Adams, C. Joy, M. Fenton. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. *The British Journal of Psychiatry*, 176:249–252, 2000.

[41] A. M. Metz. Teaching Statistics in Biology: Using Inquiry-based Learning to Strengthen Understanding of Statistical Analysis in Biology Laboratory Courses. *CBE Life Sciences Education*, 7:317–326, 2008.

[42] E. Mills, P. Wu, J. Gagnier, D. Heels-Ansdell, V. M. Montori. An analysis of general medical and specialist journals that endorse CONSORT found that reporting was not enforced consistently. *Journal of Clinical Epidemiology*, 58:662–667, 2005.

[43] V. M. Montori, P. J. Devereaux, N. Adhikari. Randomized trials stopped early for benefit: a systematic review. *JAMA*, 294:2203–2209, 2005.

[44] S. Nieuwenhuis, B. U. Forstmann, E.-J. Wagenmakers. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, 14:1105–1109, 2011.

[45] T. V. Pereira, J. P. A. Ioannidis. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology*, 64:1060–1069, 2011.

[46] A. C. Plint, D. Moher, A. Morrison, K. Schulz, e. al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical journal of Australia*, 185:263–267, 2006.

[47] A. P. Prayle, M. N. Hurley, A. R. Smyth. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. *BMJ*, 344:d7373, 2011.

[48] D. F. Preusser, W. A. Leaf, K. B. DeBartolo, R. D. Blomberg, M. M. Levy. The effect of right-turn-on-red on pedestrian and bicyclist accidents. *Journal of Safety Research*, 13:45–55, 1982.

[49] F. Prinz, T. Schlange, K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets?. *Nature Reviews Drug Discovery*, 10:328–329, 2011.

[50] N. Schenker, J. F. Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55:182–186, 2001.

[51] J. D. Schoenfeld, J. P. A. Ioannidis. Is everything we eat associated with cancer? A systematic cookbook review. *American Journal of Clinical Nutrition*, 97:127–134, 2013.

[52] S. Schroter, N. Black, S. Evans, F. Godlee, L. Osorio, R. Smith. What errors do peer reviewers detect, and does training improve their ability to detect them?. *JRSM*, 101:507–514, 2008.

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

[54] D. G. Smith, J. Clemens, W. Crede, M. Harvey, E. J. Gracely. Impact of multiple comparisons in randomized clinical trials. *The American Journal of Medicine*, 83:545–550, 1987.

[55] A. Tatsioni, N. G. Bonitsis, J. P. A. Ioannidis. Persistence of Contradicted Claims in the Literature. *JAMA*, 298:2517–2526, 2007.

[56] S. Todd, A. Whitehead, N. Stallard, J. Whitehead. Interim analyses and sequential designs in phase III studies. *British Journal of Clinical Pharmacology*, 51:394–399, 2001.

[57] R. Tsang, L. Colley, L. D. Lynd. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *Journal of Clinical Epidemiology*, 62:609–616, 2009.

[58] E. Wagenmakers, R. Wetzels. Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 2011.

[59] H. Wainer. The Most Dangerous Equation. *American Scientist*, 95:249–256, 2007.

[60] J. M. Wicherts, M. Bakker, D. Molenaar. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLoS ONE*, 6:e26828, 2011.

[61] J. M. Wicherts, D. Borsboom, J. Kats, D. Molenaar. The poor availability of psychological research data for reanalysis. *American Psychologist*, 61:726–728, 2006.

Detailed Licensing

Overview

Title: Statistics Done Wrong (Reinhart)

Webpages: 48

All licenses found:

- CC BY 4.0: 93.8% (45 pages)
- Undeclared: 6.3% (3 pages)

By Page