Statistics Done Wrong

Alex Reinhart

August 1st, 2012

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 most 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.[1]

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

[1] Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLoS Medicine, 2(8), e124. doi:10.1371/journal.pmed.0020124

excuse to reject all modern science and return to bloodletting and leeches – it's a call to improve the science we rely on.

Updated January 2013 with a relevant example of the base-rate fallacy: survey estimates of gun usage.

Contents

An introduction to 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?

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.[2]

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 colds shorter by a day or more. 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 (or more extreme 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.

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.

[2] Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. Annals of internal medicine, 130(12), 995–1004.

Statistical power and underpowered statistics

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:

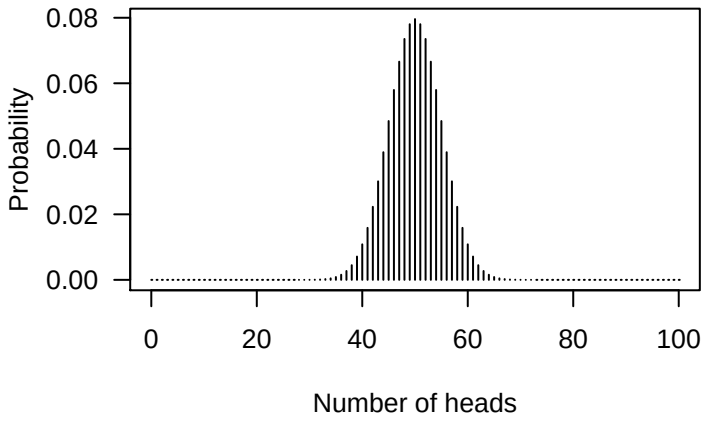


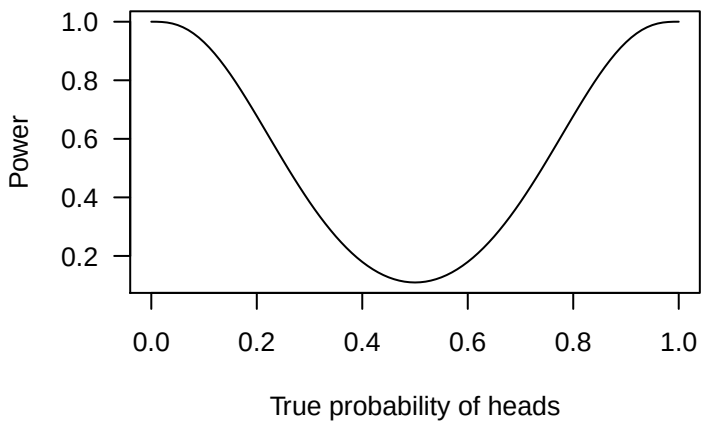
Figure 1: This shows the likelihood of getting different numbers of heads, if you flip a coin a hundred times.

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?



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?

Or 1,000 times? The plots on the next page show the result.

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.

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

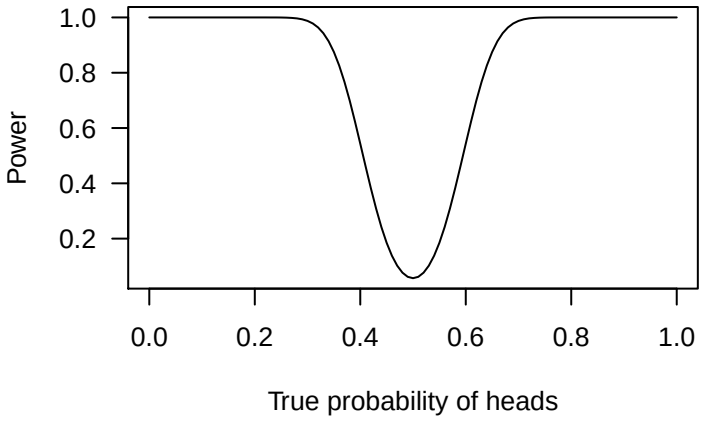


Figure 2: Power after 100 flips.

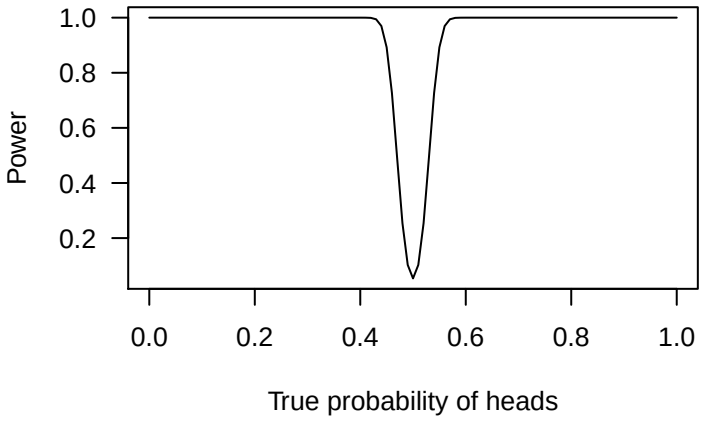


Figure 3: Power after 1,000 flips.

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.[3] 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, 64% of randomized controlled medical trials didn't collect enough data to detect a 50% difference 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 trials didn't have the power to detect a 25% difference.[4]

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.

[3] Tsang, R., Colley, L., & Lynd, L. D. (2009). Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. Journal of clinical epidemiology, 62(6), 609–616. doi:10.1016/j.jclinepi.2008.08.005

[4] D Moher, C S Dulberg, and G A Wells. (1994). Statistical Power, Sample Size, and Their Reporting in Randomized Controlled Trials. Journal of the American Medical Association, 272(2), 122–124. doi:10.1001/jama.1994.03520020048013

Bedard, P. L., Krzyzanowska, M. K., Pintilie, M., & Tannock, I. F. (2007). Statistical Power of Negative Randomized Controlled Trials Presented at American Society for Clinical Oncology Annual Meetings. Journal of Clinical Oncology, 25(23), 3482–3487. doi:10.1200/JCO.2007.11.3670

Brown, C. G., Kelen, G. D., Ashton, J. J., & Werman, H. A. (1987). The beta error and sample size determination in clinical trials in emergency medicine. Annals of emergency medicine, 16(2), 183–187.

Chung, K. C., Kalliainen, L. K., & Hayward, R. A. (1998). Type II (beta) errors in the hand literature: the importance of power. The Journal of hand surgery, 23(1), 20–25. doi:10.1016/S0363-5023(98)80083-X

Pseudoreplication: choose your data wisely

Many studies strive to collect more data through replication: by repeating their measurements they can be more certain of their numbers, and can 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.[5] 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.

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.

The p value and the base rate fallacy

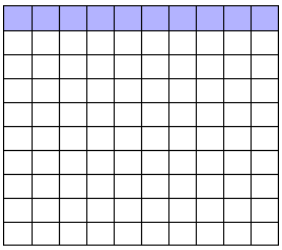
You've already seen that p values are hard to interpret. Getting a statistically insignificant result doesn't mean there's no difference. What about getting a significant result?

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

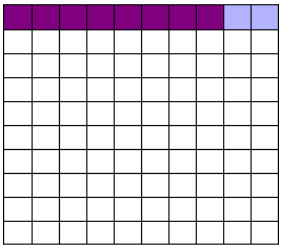
[5] Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neuroscience, 11, 5. doi:10.1186/1471-2202-11-5

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:

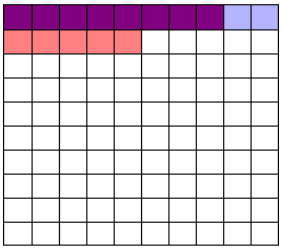


As we saw, most trials can't perfectly detect every good medication. We'll assume my tests have a statistical power of 0.8. Of the ten good drugs, I will correctly detect around eight of them, shown in purple:



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:



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.

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.

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,[6] the answer is 9%.[7]

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

[7] Krämer, W., & Gigerenzer, G. (2005). How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities. Statistical Science, 20(3), 223–230. doi:10.1214/088342305000000296

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 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.[8] If you ask doctors, two thirds fail.[9] 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.

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

[8] Krämer, W., & Gigerenzer, G. (2005). How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities. Statistical Science, 20(3), 223–230. doi:10.1214/088342305000000296

[9] Bramwell, R., & West, H. (2006). Health professionals' and service users' interpretation of screening test results: experimental study. British Medical Journal. doi:10.1136/bmj.38884.663102.AE

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 been left. That leaves 430,000 burglaries involving homeowners who could confront the burglar – 845,000 of which, we are led to believe, were stymied by gun-toting residents.[10]

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.

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:

[10]Hemenway, D. (1997). Survey Research and Self-Defense Gun Use: An Explanation of Extreme Overestimates. The Journal of Criminal Law and Criminology, 87(4), 1430-1445.

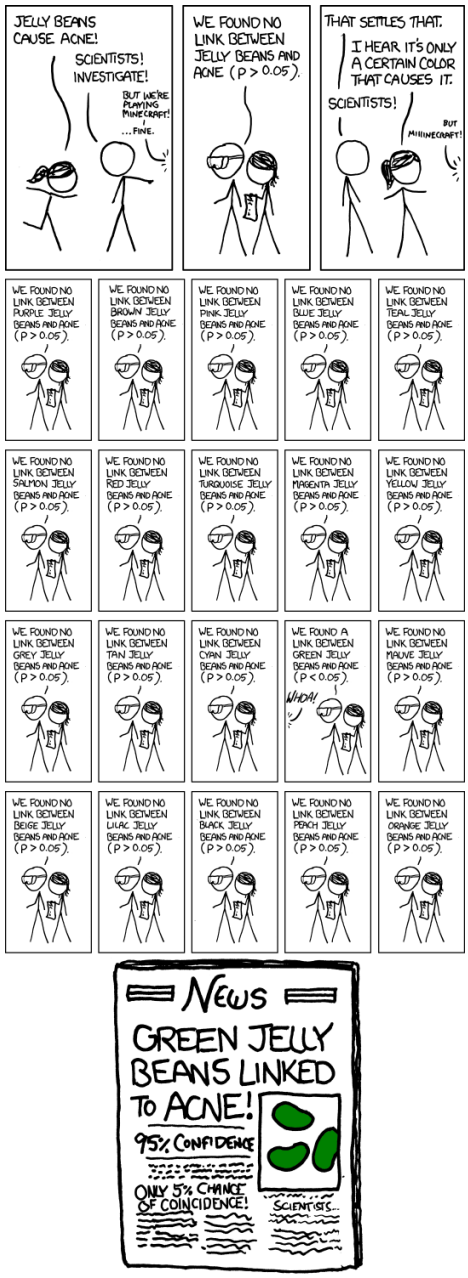


Figure 4: Cartoon from xkcd, by Randall Munroe. http://xkcd.com/882/ (copyright info)

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.[11] 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.

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 $p < 0.05$.[12] 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.

Red herrings in brain imaging

Neuroscientists do massive numbers of comparisons regularly. They often perform fMRI studies, where an 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. (Of course, most studies are more sophisticated

[11] Smith, D. G., Clemens, J., Crede, W., Harvey, M., & Gracely, E. J. (1987). Impact of multiple comparisons in randomized clinical trials. The American Journal of Medicine, 83(3), 545–550.

[12] Weisstein, Eric W. "Bonferroni Correction." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/BonferroniCorrection.html

than this; there are methods of looking for clusters of voxels which all change together, among other techniques.)

Trouble is, there are thousands of voxels to compare and many opportunities for false positives. Neuroscientists set their threshold for significance low, at $p < 0.001$, but that may not be enough.

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."[13]

When differences in significance aren't significant differences

"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.[14]

Imagine a study comparing walrus diets. One group of walruses is fed their ordinary diet, while two other groups are fed new, more nutritious diets. The researchers weigh the walruses after a month and find that nutritious diet A caused the walruses to gain about 25 kilograms more than the ordinary diet, while nutritious diet B caused the walruses to only gain about 10 kg more.

[13] Bennett, C., Baird, A., Miller, M., & Wolford, G. (2010). 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), 1–5.

[14] Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. The American Statistician, 60(4), 328–331. doi:10.1198/000313006X152649

We want to establish how much weight gain we can expect on average from each diet. If we fed these diets to all the walruses in the universe, what would the average weight gain be? Now, we don't have many walruses, so it's hard to answer that – individual walruses vary quite a bit, and can gain weight for reasons other than a new diet. (Perhaps the male walruses are bulking up for swimsuit season.) Accounting for this variation, we calculate that diet B's effect is statistically insignificant: there's too much variation between walruses to conclude that the 10 kg weight gain was caused by the diet. Diet A, however, causes a statistically significant weight gain, and was probably effective.

A researcher might conclude "diet A caused a statistically significant weight gain, while diet B did not; clearly diet A is more fattening than diet B." Other walrus keepers might read the paper and decide to feed diet A to their underweight and sick walruses, since it's more effective.

But is it? Not necessarily.

Because we have limited data, there's some inherent error in our numbers. We can calculate what results would also be consistent with the data; for example, the "true" effect of diet A might be 35 kg or 17 kg of weight gain, and it's plausible that with our small sample of walruses we'd still see the results we did. Collecting more data would help us pin down the true effects more precisely.

Statistics supplies tools for quantifying this error. If we calculate the uncertainties of each of our measurements, we might find it plausible that both diets have exactly the same effect. Diet B has a statistically insignificant effect because it's entirely plausible that it causes a weight gain of 0 kilograms – but it's also plausible that it causes a gain of 20 kg and we got some unusually skinny walruses in our sample. Similarly, it's entirely plausible that diet A *also* causes a gain of 20 kg and we got some unusually gluttonous walruses in our study. Without more data we cannot be sure.

Our data is insufficient to conclude there is a statistically significant difference between diets A and B. While one diet produces statistically significant results and the other doesn't, there's not a statistically significant difference between the two. They might both be equally effective. Be careful comparing the significance of two results. If you want to compare two treatments or effects, compare them directly.

Examples of this error in common literature and news stories abound. A huge proportion of papers in neuroscience, for instance, commit the error.[15] You might also remember a study a few years ago suggesting that men with more biological older brothers are more likely to be homosexual.[16] How did they reach this conclusion? And why older brothers and not older sisters?

[15] Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience, 14(9), 1105–1109. doi:10.1038/nn.2886

[16] Bogaert, A. F. (2006), "Biological Versus Nonbiological Older Brothers and Men's Sexual Orientation," in Proceedings of the National Academy of Sciences, 103, pp. 10771–10774. doi:10.1073/pnas.0511152103

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.[17]

Stopping rules and regression to the mean

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.

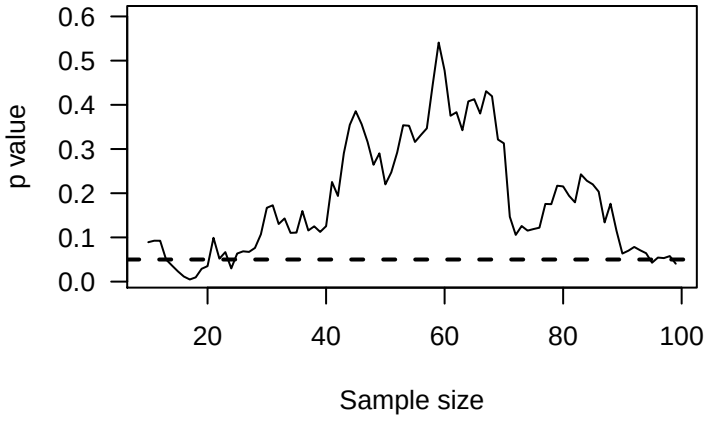
When poorly done, however, this can lead to numerous false positives.

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 the simulation shown.

The 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.

[17] Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. The American Statistician, 60(4), 328–331. doi:10.1198/000313006X152649



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.[18]

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.[19] 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.

[18]Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science, 22(11), 1359–1366. doi:10.1177/0956797611417632

[19]Todd, S., Whitehead, A., Stallard, N., & Whitehead, J. (2001). Interim analyses and sequential designs in phase III studies. British journal of clinical pharmacology, 51(5), 394–399. doi:10.1046/j.1365-2125.2001.01382.x

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.[20] 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.[21]

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

[20] Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. Epidemiology, 19(5), 640–648. doi:10.1097/EDE.0b013e31818131e7

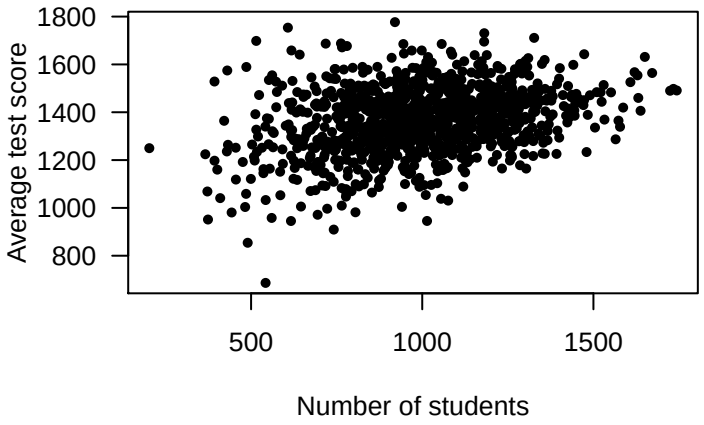
Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. JAMA, 294(2), 218–228. doi:10.1001/jama.294.2.218

[21] Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. Journal of clinical epidemiology, 58(6), 543–549. doi:10.1016/j.jclinepi.2004.10.019

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:



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.[22]

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

[22] Wainer, H. (2007). The Most Dangerous Equation. American Scientist, 95(2), 49. doi:10.1511/2007.65.1026

highest kidney cancer rate in the nation. Small counties hence have vastly more variable kidney cancer rates, simply because they have so few residents.[23]

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?

[23] Gelman, A., & Price, P. N. (1999). All maps of parameter estimates are misleading. Statistics in medicine, 18(23), 3221–3234.

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.[24] 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.[25]

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.[26] 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.

[24] Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. Epidemiology, 19(5), 640–648. doi:10.1097/EDE.0b013e31818131e7

[25] Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science, 22(11), 1359–1366. doi:10.1177/0956797611417632

[26] Chan, A.-W., Hróbjartsson, A., Jørgensen, K. J., Gøtzsche, P. C., & Altman, D. G. (2008). Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. British Medical Journal, 337, a2299. doi:10.1136/bmj.a2299

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles. JAMA, 291(20), 2457–2465. doi:10.1001/jama.291.20.2457

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 trials suggest that many p values are wrong, and some statistically insignificant results are actually significant when computed correctly.[27] 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.[28]

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

[27] Gøtzsche, P. C. (2006). Believability of relative risks and odds ratios in abstracts: cross sectional study. British Medical Journal, 333(7561), 231–234. doi:10.1136/bmj.38895.410451.79

[28] Baggerly, K. A., & Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. The Annals of Applied Statistics, 3(4), 1309–1334. doi:10.1214/09-AOAS291

Gøtzsche, P. C. (1989). 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.

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.[29] 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.[30]

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.[31] 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.

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

[29] The Reproducibility Project, at http://opensciencframework.org/reproducibility/

[30] Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. Nature, 483(7), 531–533. doi:10.1038/483531a

[31] Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. JAMA, 294(2), 218–228. doi:10.1001/jama.294.2.218

> 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.

Contact

I've tried my best, but inevitably this guide will contain errors and omissions. If you spot an error, have a question, or know a common fallacy I've missed, email me at stats@refsmmat.com.

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.