2

STATISTICAL POWER AND UNDERPOWERED STATISTICS

You've seen how it's possible to miss real effects by not collecting enough data. You might miss a viable medicine or fail to notice an important side effect. So how do you know how much data to collect?

The concept of *statistical power* provides the answer. The power of a study is the probability 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.

The Power Curve

Suppose I'm convinced that my archnemesis has an unfair coin. Rather than getting heads half the time and tails half the time, it's biased to give one outcome 60% of the time, allowing

him to cheat at incredibly boring coin-flipping betting games. I suspect he's cheating—but how to *prove* it?

I can't just take the coin, flip it 100 times, and count the heads. Even a perfectly fair coin won't always get 50 heads, as the solid line in Figure 2-1 shows.

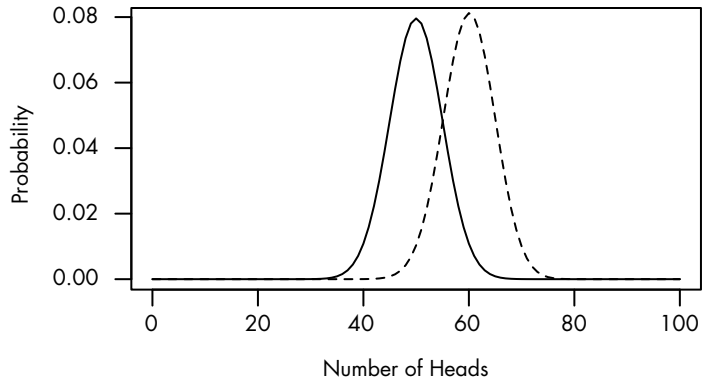

Figure 2-1: The probability of getting different numbers of heads if you flip a fair coin (solid line) or biased coin (dashed line) 100 times. The biased coin gives heads 60% of the time.

Even though 50 heads is the most likely outcome, it still happens less than 10% of the time. I'm also reasonably likely to get 51 or 52 heads. In fact, when flipping a fair coin 100 times, I'll get between 40 and 60 heads 95% of the time. On the other hand, results far outside this range are unlikely: with a fair coin, there's only a 1% chance of obtaining more than 63 or fewer than 37 heads. Getting 90 or 100 heads is almost impossible.

Compare this to the dashed line in Figure 2-1, showing the probability of outcomes for a coin biased to give heads 60% of the time. The curves do overlap, but you can see that an unfair coin is much more likely to produce 70 heads than a fair coin is.

Let's work out the math. Say I run 100 trials and count the number of heads. If the result isn't exactly 50 heads, I'll calculate the probability that a *fair* coin would have turned up a deviation of that size or larger. That probability is my p value. I'll consider a p value of 0.05 or less to be statistically significant and hence call the coin unfair if p is smaller than 0.05.

How likely am I to find out a coin is biased using this procedure? A *power curve*, as shown in Figure 2-2, can tell me. Along the horizontal axis is the coin's true probability of getting heads—that is, how biased it is. On the vertical axis is the probability that I will conclude the coin is rigged.

The *power* for any hypothesis test is the probability that it will yield a statistically significant outcome (defined in this example as $p < 0.05$). A fair coin will show between 40 and 60 heads in 95% of trials, so for an *unfair* coin, the power is the probability of a result *outside* this range of 40–60 heads. The power is affected by three factors:

- **The size of the bias you're looking for.** A huge bias is much easier to detect than a tiny one.
- **The sample size.** By collecting more data (more coin flips), you can more easily detect small biases.
- **Measurement error.** It's easy to count coin flips, but many experiments deal with values that are harder to measure, such as medical studies investigating symptoms of fatigue or depression.

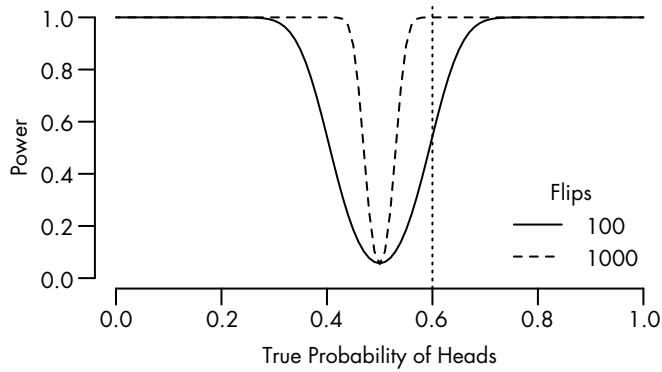

Figure 2-2: The power curves for 100 and 1,000 coin flips, showing the probability of detecting biases of different magnitudes. The vertical line indicates a 60% probability of heads.

Let's start with the size of the bias. The solid line in Figure 2-2 shows that if the coin is rigged to give heads 60% of the time, I have a 50% chance of concluding that it's rigged after 100 flips. (That is, when the true probability of heads is 0.6, the power is 0.5.) The other half of the time, I'll get fewer than 60 heads and fail to detect the bias. With only 100 flips, there's just too little data to *always* separate bias from random variation. The coin would have to be incredibly biased—yielding heads more than 80% of the time, for example—for me to notice nearly 100% of the time.

Another problem is that even if the coin is perfectly fair, I will falsely accuse it of bias 5% of the time. I've designed my test

to interpret outcomes with $p < 0.05$ as a sign of bias, but those outcomes *do* happen even with a fair coin.

Fortunately, an increased sample size improves the sensitivity. The dashed line shows that with 1,000 flips, I can easily tell whether the coin is rigged. This makes sense: it's overwhelmingly unlikely that I could flip a fair coin 1,000 times and get more than 600 heads. I'll get between 469 and 531 95% of the time. Unfortunately, I don't really have the time to flip my nemesis's coin 1,000 times to test its fairness. Often, performing a sufficiently powerful test is out of the question for purely practical reasons.

Now counting heads and tails is easy, but what if I were instead administering IQ tests? An IQ score does not measure an underlying "truth" but instead can vary from day to day depending on the questions on the test and the mood of the subject, introducing random noise to the measurements. If you were to compare the IQs of two groups of people, you'd see not only the normal variation in intelligence from one person to the next but also the random variation in *individual* scores. A test with high variability, such as an IQ test requiring subjective grading, will have relatively less statistical power.

More data helps distinguish the signal from the noise. But this is easier said than done: many scientists don't have the resources to conduct studies with adequate statistical power to detect what they're looking for. They are doomed to fail before they even start.

The Perils of Being Underpowered

Consider a trial testing two different medicines, Fixitol and Solvix, for the same condition. You want to know which is safer, but side effects are rare, so even if you test both medicines on 100 patients, only a few in each group will suffer serious side effects. Just as it is difficult to tell the difference between two coins that turn up 50% heads and 51% heads, the difference between a 3% and 4% side effect rate is difficult to discern. If four people taking Fixitol have serious side effects and only three people taking Solvix have them, you can't say for sure whether the difference is due to Fixitol.

If a trial isn't powerful enough to detect the effect it's looking for, we say it is *underpowered*.

You might think calculations of statistical power are essential for medical trials; a scientist might want to know how many patients are needed to test a new medication, 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 detecting a real effect of the expected size. (If the true effect is actually larger, the study will have greater power.)

However, few scientists ever perform this calculation, and few journal articles even mention statistical power. In the prestigious journals *Science* and *Nature*, fewer than 3% of articles calculate statistical power before starting their study.[1] Indeed, many trials conclude that "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.[2] If one of these trials was comparing side effects in two drugs, a doctor might erroneously think the medications are equally safe, when one could very well be much more dangerous than the other.

Maybe this is a problem only for rare side effects or only when a medication has a weak effect? Nope. In one sample of studies published in prestigious medical journals between 1975 and 1990, more than four-fifths of randomized controlled trials that reported negative results didn't collect enough data to detect a *25% difference* in primary outcome between treatment groups. That is, even if one medication reduced symptoms by 25% more than another, there was insufficient data to make that conclusion. And nearly *two-thirds* of the negative trials didn't have the power to detect a 50% difference.[3]

A more recent study of trials in cancer research found similar results: only about half of published studies with negative results had enough statistical power to detect even a large difference in their primary outcome variable.[4] Less than 10% of these studies explained why their sample sizes were so poor. Similar problems have been consistently seen in other fields of medicine.[5,6]

In neuroscience, the problem is even worse. Each individual neuroscience study collects such little data that the median study has only a 20% chance of being able to detect the effect it's looking for. You could compensate for this by aggregating data collected across several papers all investigating the same effect. But since many neuroscience studies use animal subjects, this raises a significant ethical concern. If each study is underpowered, the true effect will likely be discovered only after many studies using many animals have been completed and analyzed—using far more animal subjects than if the study had been done properly in the first place.[7] An ethical review board should not approve a trial if it knows the trial is unable to detect the effect it is looking for.

Wherefore Poor Power?

Curiously, the problem of underpowered studies has been known for decades, yet it is as prevalent now as it was when first pointed out. In 1960 Jacob Cohen investigated the statistical power of studies published in the *Journal of Abnormal and Social Psychology*[8] and discovered that the average study had only a power of 0.48 for detecting medium-sized effects.* His research was cited hundreds of times, and many similar reviews followed, all exhorting the need for power calculations and larger sample sizes. Then, in 1989, a review showed that in the decades since Cohen's research, the average study's power had actually *decreased*.[9] This decrease was because of researchers becoming aware of another problem, the issue of multiple comparisons, and compensating for it in a way that reduced their studies' power. (I will discuss multiple comparisons in Chapter 4, where you will see that there is an unfortunate trade-off between a study's power and multiple comparison correction.)

So why are power calculations often forgotten? One reason is the discrepancy between our intuitive feeling about sample sizes and the results of power calculations. It's easy to think, "Surely these are enough test subjects," even when the study has abysmal power. For example, suppose you're testing a new heart attack treatment protocol and hope to cut the risk of death in half, from 20% to 10%. You might be inclined to think, "If I don't see a difference when I try this procedure on 50 patients, clearly the benefit is too small to be useful." But to have 80% power to detect the effect, you'd actually need *400* patients—200 in each control and treatment group.[10] Perhaps clinicians just don't realize that their adequate-seeming sample sizes are in fact far too small.

Math is another possible explanation for why power calculations are so uncommon: analytically calculating power can be difficult or downright impossible. Techniques for calculating power are not frequently taught in intro statistics courses. And some commercially available statistical software does not come with power calculation functions. It is possible to avoid hairy mathematics by simply simulating thousands of artificial datasets with the effect size you expect and running your statistical tests on the simulated data. The power is simply the fraction of datasets for which you obtain a statistically significant result. But this approach requires programming experience, and simulating realistic data can be tricky.

*Cohen defined "medium-sized" as a 0.5-standard-deviation difference between groups.

Even so, you'd think scientists would notice their power problems and try to correct them; after five or six studies with insignificant results, a scientist might start wondering what she's doing wrong. But the average study performs not one hypothesis test but many and so has a good shot at finding *something* significant.[11] As long as this significant result is interesting enough to feature in a paper, the scientist will not feel that her studies are underpowered.

The perils of insufficient power do not mean that scientists are lying when they state they detected no significant difference between groups. But it's misleading to assume these results mean there is no *real* difference. There may be a difference, even an important one, but the study was so small it'd be lucky to notice it. Let's consider an example we see every day.

Wrong Turns on Red

In the 1970s, many parts of the United States began allowing 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 traffic agencies to consider allowing right turns on red to save fuel wasted by commuters waiting at red lights, and eventually Congress required states to allow right turns on red, treating it as an energy conservation measure just like building insulation standards and more efficient lighting.

Several studies were conducted to consider the safety impact of the change. In one, a consultant for the Virginia Department of Highways and Transportation conducted a before-and-after study of 20 intersections that had begun 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. But this difference was not statistically significant, which the consultant indicated in his report. When the report was forwarded to the governor, the commissioner of the Department of Highways and Transportation wrote that "we can discern no significant hazard to motorists or pedestrians from implementation" of right turns on red.[12] In other words, he turned *statistical* insignificance into *practical* insignificance.

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].

Of course, these studies were underpowered. But more cities and states began to allow right turns on red, and the practice became widespread across the entire United States. Apparently, no one attempted to aggregate these many small studies to produce a more useful dataset. Meanwhile, more pedestrians were being run over, and more cars were involved in collisions. Nobody collected enough data to show this conclusively until several years later, when studies finally showed that among incidents involving right turns, collisions were occurring roughly 20% more frequently, 60% more pedestrians were being run over, and twice as many bicyclists were being struck.[13,14,*]

Alas, the world of traffic safety has learned little from this example. A 2002 study, for example, considered the impact of paved shoulders on the accident rates of traffic on rural roads. Unsurprisingly, a paved shoulder reduced the risk of accident—but there was insufficient data to declare this reduction statistically significant, so the authors stated that the cost of paved shoulders was not justified. They performed no cost-benefit analysis because they treated the insignificant difference as meaning there was no difference at all, despite the fact that they had collected data suggesting that paved shoulders improved safety! The evidence was not strong enough to meet their desired p value threshold.[12] A better analysis would have admitted that while it is plausible that shoulders have no benefit at all, the data is also consistent with them having substantial benefits. That means looking at *confidence intervals*.

Confidence Intervals and Empowerment

More useful than a statement that an experiment's results were statistically insignificant is a confidence interval giving plausible sizes for the effect. Even if the confidence interval includes zero, its width tells you a lot: a narrow interval covering zero tells you that the effect is most likely small (which may be all you need to know, if a small effect is not practically useful),

*It is important to note that accidents involving right turns are rare: these changes amount to fewer than 100 deaths per year in the United States.[15] A 60% increase in a small number is still small—but nonetheless, a statistical error kills dozens of people each year!

while a wide interval clearly shows that the measurement was not precise enough to draw conclusions.

Physicists commonly use confidence intervals to place bounds on quantities that are not significantly different from zero. In the search for a new fundamental particle, for example, it's not helpful to say, "The signal was not statistically significant." Instead, physicists can use a confidence interval to place an upper bound on the rate at which the particle is produced in the particle collisions under study and then compare this result to the competing theories that predict its behavior (and force future experimenters to build yet bigger instruments to find it).

Thinking about results in terms of confidence intervals provides a new way to approach experimental design. Instead of focusing on the power of significance tests, ask, "How much data must I collect to measure the effect to my desired precision?" Even a powerful experiment can nonetheless produce significant results with extremely wide confidence intervals, making its results difficult to interpret.

Of course, the sizes of our confidence intervals vary from one experiment to the next because our data varies from experiment to experiment. Instead of choosing a sample size to achieve a certain level of power, we choose a sample size so the confidence interval will be suitably narrow 99% of the time (or 95%; there's not yet a standard convention for this number, called the *assurance*, which determines how often the confidence interval must beat our target width).[16]

Sample size selection methods based on assurance have been developed for many common statistical tests, though not for all; it is a new field, and statisticians have yet to fully explore it.[17] (These methods go by the name *accuracy in parameter estimation*, or *AIPE*.) Statistical power is used far more often than assurance, which has not yet been widely adopted by scientists in any field. Nonetheless, these methods are enormously useful. Statistical significance is often a crutch, a catchier-sounding but less informative substitute for a good confidence interval.

Truth Inflation

Suppose Fixitol reduces symptoms by 20% over a placebo, but the trial you're using to test it is too small to have adequate statistical power to detect this difference reliably. We know that small trials tend to have varying results; it's easy to get 10 lucky patients who have shorter colds than usual but much harder to get 10,000 who all do.

Now imagine running many copies of this trial. Sometimes you get unlucky patients, 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%, so you stop the trial and say, "Look! It works!" You can plot these outcomes in Figure 2-3, which shows the probability that each trial will yield a certain effect size estimate.

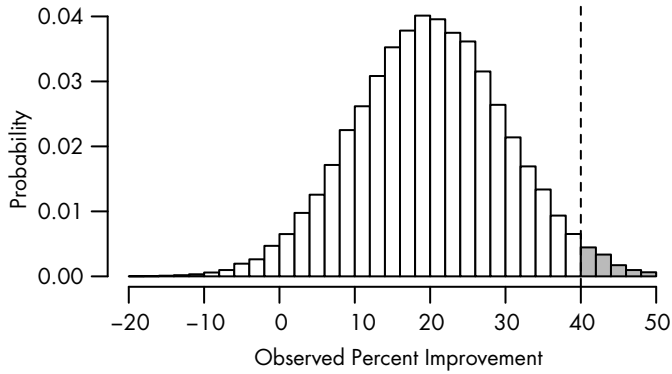

Figure 2-3: If you run your trial thousands of times, you will see a broad distribution of effect sizes in terms of percent reduction in symptoms. The vertical dotted line indicates the effect size which is large enough to be statistically significant. The true improvement is 20%, but you see effects from 10% losses to 50% gains. Only the lucky trials are statistically significant, exaggerating the effect size.

You've correctly concluded Fixitol is effective, but you've inflated the size of its effect because your study was underpowered.

This effect, known as *truth inflation*, *type M error* (*M* for magnitude), or the *winner's curse*, occurs in fields where many researchers conduct similar experiments and compete to publish the most "exciting" results: pharmacological trials, epidemiological studies, gene association studies ("gene A causes condition B"), and psychological studies often show symptoms, along with some of the most-cited papers in the medical literature.[18,19] In fast-moving fields such as genetics, the earliest published results are often the most extreme because journals are most interested in publishing new and exciting results. Follow-up studies tend to show much smaller effects.[20]

Consider also that top-ranked journals, such as *Nature* and *Science*, prefer to publish studies with groundbreaking results—meaning large effect sizes in novel fields with little prior research. This is a perfect combination for chronic truth inflation. Some evidence suggests a correlation between a journal's impact factor (a rough measure of its prominence and importance) and the factor by which its studies overestimate effect sizes. Studies that produce less "exciting" results are closer to the truth but less interesting to a major journal editor.[21,22]

When a study claims to have detected a large effect with a relatively small sample, your first reaction should not be "Wow, they've found something big!" but "Wow, this study is underpowered!"[23] Here's an example. Starting in 2005, Satoshi Kanazawa published a series of papers on the theme of gender ratios, culminating with "Beautiful Parents Have More Daughters." He followed up with a book discussing this and other "politically incorrect truths" he'd discovered. The studies were popular in the press at the time, particularly because of the large effect size they reported: Kanazawa claimed the most beautiful parents have daughters 52% of the time, but the least attractive parents have daughters only 44% of the time.

To biologists, a small effect—perhaps one or two percentage points—would be plausible. The *Trivers–Willard Hypothesis* suggests that if parents have a trait that benefits girls more than boys, then they will have more girls than boys (or vice versa). If you assume girls benefit more from beauty than boys, then the hypothesis would predict beautiful parents would have, on average, slightly more daughters.

But the effect size claimed by Kanazawa was extraordinary. And as it turned out, he committed several errors in his statistical analysis. A corrected regression analysis found that his data showed attractive parents were indeed 4.7% more likely to have girls—but the confidence interval stretched from 13.3% more likely to 3.9% *less* likely.[23] Though Kanazawa's study used data from nearly 3,000 parents, the results were not statistically significant.

Enormous amounts of data would be needed to reliably detect a small difference. Imagine a more realistic effect size—say, 0.3%. Even with 3,000 parents, an observed difference of 0.3% is far too small to distinguish from luck. You'd be lucky to obtain a statistically significant result just 5% of the time. These

results will be those that exaggerate the true effect by at least a factor of 20, and 40% of them will produce a wild overestimate in favor of boys instead of girls.[23]

So even if Kanazawa had performed a perfect statistical analysis, he still would have occasionally gotten lucky with a paper like "Engineers Have More Sons, Nurses Have More Daughters"* and given a wild overestimate of a true, tiny effect. Studies of the size he conducted are simply *incapable* of detecting effects of the size you'd expect in advance. A prior power analysis would have told him this.

Little Extremes

Truth inflation arises because small, underpowered studies have widely varying results. Occasionally you're bound to get lucky and have a statistically significant but wildly overestimated result. But this wide variation can cause trouble even when you're not performing significance tests. 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 10 schools are almost all smaller than that. It seems that small schools do the best, perhaps because 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 the plot of test scores versus school size in Figure 2-4. Smaller schools have wider variation in test scores because they have fewer students. With fewer students, there are fewer data points to establish the "true" performance of the teachers; a few anomalous scores can sway the school's average significantly. As schools get larger, test scores vary less and in fact *increase* on average.[24]

*A real paper, which he published in 2005 in the *Journal of Theoretical Biology*.

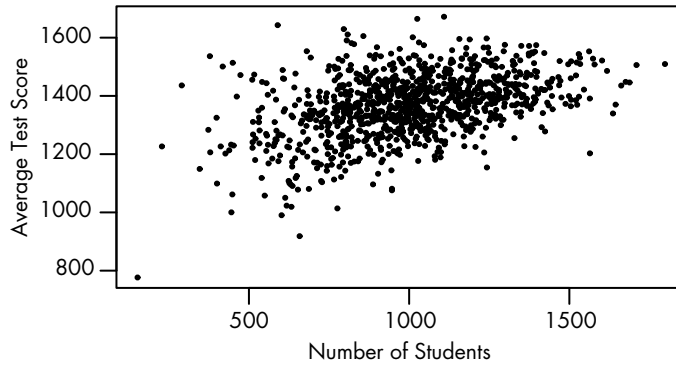

Figure 2-4: Schools with more students have less random variation in their test scores. This data is simulated but based on real observations of Pennsylvania public schools.

Another example: in the United States, counties with the lowest rates of kidney cancer tend to be Midwestern, Southern, and Western rural counties. Why might this be? Maybe rural people get more exercise or inhale less-polluted air. Or perhaps they just lead less stressful lives.

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 10 residents gives that county the highest kidney cancer rate in the nation. Small counties hence have much more variation in kidney cancer rates simply because they have so few residents.[25] The confidence intervals for their cancer rates are correspondingly larger.

A popular strategy to fight this problem is called *shrinkage*. For counties with few residents, you can "shrink" the cancer rate estimates toward the national average by taking a weighted average of the county cancer rate with the national average rate. When the county has few residents, you weight the national average strongly; when the county is large, you weight the county strongly. Shrinkage is now common practice in constructing cancer rate maps, among other applications.* Unfortunately, it biases results in the opposite direction: small

*However, shrinkage is usually implemented using more sophisticated methods than a simple weighted average.

counties with truly abnormal cancer rates are estimated to have rates much closer to the national average than they are.

There's no single fix to this problem. The best alternative is to sidestep it altogether: rather than estimating rates by county, you could use congressional districts, which in the United States are designed to have roughly equal populations. Congressional districts are much larger than counties, though, and frequently they have strange shapes because of gerrymandering. Maps based on districts may not be statistically misleading but are still difficult to interpret.

Of course, enforcing equal sample sizes isn't always an option. Online shopping sites, for instance, need to sort products based on customer ratings, but they can't force equal numbers of customers to rate every product. Another example is a discussion website like reddit, which can sort comments by user ratings; comments can receive vastly different numbers of votes depending on when or where or by whom they were posted. Shrinkage is helpful in dealing with these situations. An online store can use a weighted average of a product's ratings and some global average. Products with few ratings will be treated as generically average, while products with thousands of votes are sorted by their true individual ratings.

For sites like reddit that have simple up-and-down votes rather than star ratings, one alternative is to generate a confidence interval for the fraction of positive votes. The interval starts wide when a comment has only a few votes and narrows to a definite value ("70% of voters like this comment") as comments accumulate; sort the comments by the bottom bound of their confidence intervals. New comments start near the bottom, but the best among them accumulate votes and creep up the page as the confidence interval narrows. And because comments are sorted by the proportion of positive votes rather than the total number, new comments can compete with those that have already accumulated thousands of votes.[26,27]

TIPS

- Calculate the statistical power when designing your study to determine the appropriate sample size. Don't skimp. Consult a book like Cohen's classic *Statistical Power Analysis for the Behavioral Sciences* or talk to a statistical consultant. If the sample size is impractical, be aware of the limitations of your study.
- When you need to measure an effect with precision, rather than simply testing for significance, use assurance instead of power: design your experiment to measure the hypothesized effect to your desired level of precision.

- Remember that "statistically insignificant" does not mean "zero." Even if your result is insignificant, it represents the best available estimate given the data you have collected. "Not significant" does not mean "nonexistent."
- Look skeptically on the results of clearly underpowered studies. They may be exaggerated due to truth inflation.
- Use confidence intervals to determine the range of answers consistent with your data, regardless of statistical significance.
- When comparing groups of different sizes, compute confidence intervals. These will reflect the additional certainty you have in larger groups.