The P-Value Crisis

A White Paper by Fredrik deBoer

FREDRIK DEBOER

February 19, 2017
fredrikdeboer.com

The P-Value Crisis

A White Paper by Fredrik deBoer

In Brief

P-value, an essential statistical measure used to investigate whether a given quantitative result is the product of a real-world effect, suffers from several drawbacks and limitations. These limitations, when combined with professional incentives for career academics, threaten to undermine the validity of research findings in a variety of fields. Problems can be mitigated with techniques such as pre-registering research questions, reporting effect sizes, and attempting to replicate prior research in order to ascertain the consistency of a perceived effect.

What is P-value?

P-value is one of the most ubiquitous statistics found in conventional quantitative research. We live in a world of **variability**. General statistical trends have outliers and exceptions which can hide their effects, and outliers and exceptions can create the appearance of an overall trend where there in fact is none. Suppose you want to determine if student performance on a standardized test of reading is impacted by eye color. Your **null hypothesis** (a theoretical construct developed by statisticians, designed to help us interpret our findings) is that there is no difference between students of different eye colors on standardized tests of reading. Your **alternative hypothesis** is that there is a difference. You take a random sample and divide the students into groups: blue eyes, brown eyes, green eyes, and grey eyes. You administer your test and find that eye-color groups have different average performance on those tests. Does this mean that students with different eye colors are necessarily different in their academic talents? There are all sorts of reasons that we might see minor differences in test scores, but most of them would hopefully wash out via your random sampling procedure. One challenge that remains, however, is **sampling error**, or the impact on your variable of interest based on the underlying variability of your data. Perhaps the blue eyed score is higher because your random sample happened to select unusually high performing blue-eyed students. We can never entirely eliminate sampling error (unless we're taking a census, which presents its own problems), but we can use p-value to effectively determine how worried about it we should be.

Three quantitative factors contribute to a p-value: the size of your sample (*n*), the size of the effect – that is, the difference between the means of your groups – and the variability of your data, or how spread out the data in your sample is on the variable of interest. Think of variability like this. Remember that we're comparing averages here. Consider that the average of 0 and 10 is 5 and that the average of 5 and 5 is 5. So: for which example is 5 a more representative figure of the represented data? Clearly, the latter. If you have a scale of 0 to 10 and all of your data points lie equally spread out at those extremes, you end up with an average that accurately describes none of your subjects. Measures of variability (or spread) such as **standard deviation** quantitatively define such dynamics. Now consider our test. If we saw that all blue-eyed subjects are consistently performing at the same (somewhat higher) end of the scale, we would trust that the average is a more meaningful description of that population than if some blue-eyed students were performing fantastically well, a smaller number were performing fairly poorly, and the resulting average lied above the average of other groups but still below the high-performing blue-eyed students. This is the influence of spread. The calculation of p-value mathematically combines and controls for these three factors.

The p-value itself is a number between 0 and 1. Practically speaking, in real-world data 0 and 1 are not possible outcomes. P-value is notoriously thankless to define, so I will borrow a conventional definition: p-value is the probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true. That is, how likely are we to have gotten the same result, or a more extreme result in the same "direction" (that is, greater difference in means, whether positive or negative), even when there is no difference between the groups?

In more casual terms you can think of p-value as the likelihood that a given observed quantitative outcome is the result of the underlying variability of our data. Sometimes, this is discussed as whether the effect is "real." A p-value of .10 would indicate that given your underlying explanation of the data – given your model, your tool, your instrument, your definition of the null hypothesis – you would expect to find an difference in means of the same or greater amount as in your observed averages simply as a product of the underlying variability of your data 1 out of 10 times you ran the experiment. A p-value of .05 would be 1 out of 20 times, .01 would be 1 out of 100 times, etc.

A hypothetical example shows that you already have an intuitive sense of p-value. Suppose I came to you and told you that I had found a quarter that I believed to be weighted towards heads. You ask me how I know, and I tell you that I flipped the quarter 10 times, and that 7 out of the 10 times, it came up heads. Since a coin flip is a 50-50 proposition, I conclude that the coin is weighted. Would you find my claim convincing? You would not. Though the most likely outcome may be 5 heads and 5 tails, it's entirely common for a coin to fall to one side or the other more often, over a low number of observations. Indeed, a little math tells us that the exact chances of getting 7 heads in 10 tries is better than 11%. Without doing any math in your head, though, you know that 10 observation simply isn't sufficient to overcome the natural variability in coin flipping. But now suppose I told you that I had flipped the coin 10,000 times and gotten heads 7,000 times. Clearly, this is far more powerful evidence that the coin is weighted, even without having to do the math. This is what I mean by an intuitive sense of p-value.

What constitutes a "good enough" p-value? This is a matter of interpretation, context, and field. The threshold for whether a given p-value represents a statistically significant outcome is referred to as **alpha**. An alpha of .05, for example, means that a result will be considered statistically significant only if the observed result could be expected from the underlying variability less than 5% of the time. Alphas of .01 or .001 are also common. These thresholds are chosen based on a variety of factors: the number of uncontrolled variables, with more uncontrolled variables frequently compelling higher alpha thresholds; the sample size, with lower n studies typically preventing lower alpha levels; and the importance or stakes of the outcomes, with fields like medicine and engineering typically requiring lower alpha levels for reasons of safety and efficacy. If you went to the doctor for an essential medication, and you found that their data showed there was a 5% chance that the medication's perceived advantage over a sugar pill is simply statistical noise, would you feel confident enough to buy the pills? Your answer would probably depend on the severity of the illness and the cost of the medication.

Researchers frequently talk about significance in terms of **Type I error** and **Type II error**, sometimes called "false positives" and "false negatives," respectively. These terms describe two potential problematic outcomes from the use of p-value in evaluating research findings. Type I error describes the possibility of rejecting the null hypothesis – that is, arguing that a perceived effect is the product of more than just sampling error – when in fact there is no real-world difference between your groups. Type II error describes the possibility of accepting the null hypothesis when there is in fact a real underlying difference in the studied populations. (In statistics the ability of a given research plan to reject the null hypothesis when that null hypothesis in fact should be rejected is sometimes referred to as **power**; Type II error is thus an error of inadequate power in a given study.) While the p-value crisis is a product of the prevalence of Type I error, there's no sense in which Type I error is intrinsically worse than Type II. Both give us inaccurate understanding of reality.

P-Value Problems

P-value has proven very useful as an instrument to help researchers address the underlying validity of their conclusions, but it has serious problems, which when multiplied across the corpus of research studies become major challenges to our ability to trust their findings.

Consider the alpha of .05 discussed above. As noted, a more liberal threshold for statistical significance is often necessary in fields where there are a large number of variables at play, or when a given effect is relatively weak, or when the number of observations is relatively small, as are often true in human-subject research. But consider what a .05 p-value means: we would expect to achieve a p-value of that level or lower 5% of the time, or 1 out of 20, simply due to underlying variability – that is, when there is no "real" effect at all. Thus we can expect 5% of statistically significant research results at the .05 alpha to be misleading simply from our basic math. Education is a good example of a field where we typically need higher alpha thresholds thanks to the relatively weak power of effects compared to effects in natural sciences. But if we use educational data with a .05 alpha to drive public policy, we are potentially making high-stakes decisions on effects that have an unacceptably high chance of being statistical noise.

There are other arguments that type I error is even more prevalent than p-values alone indicate, due to various additional factors in the aggregation and reporting of data. I won't pretend to understand the math involved in these claims, but they appear to be convincing. A good rundown of this general issue can be found in a famous article by John P. A. Ioannidis titled "Why Most Published Research Findings are False" (2005).

Modern technological advances have in fact increased the problems with p-values, as they allow for **data snooping**. Once upon a time, the kind of math utilized in statistical analysis was laborious. It took a great deal of time and effort to look at certain relationships to find statistically significant results. Nowadays, however, any computer can calculate thousands of relationships in a matter of seconds. This increases the temptation that we will simply look at any possible number of combinations for significant results. With modern spreadsheet software, it's trivially easy to compared all number of variables together to find results that are significant. But this maximizes the problems inherent to p-value. If we do that by looking at relationships between 20 different variables, we are more likely than not to find at least one relationship that is statistically significant to a .05 level simply do to the underlying variability of our data. This, clearly, is a dangerous situation.

The p-value crisis cannot be understood without reference to the professional incentives of academics and researchers. In professional research, particularly in a university setting, the key to success is **publication**, particularly in peer-reviewed journals (that is, academic publications that submit research to external review by subject-matter experts). Without a strong publishing record, academics are unlikely to earn tenure, promotion to higher levels, or disciplinary prestige. Academic publishing demonstrates a strong and consistent bias towards those whose findings are significant. Under those conditions, it is to be expected that professional and personal pressures would cause researchers to pursue significant results even in ways that increases the risk of unsupportable findings. Sometimes this may take the form of out-and-out research fraud. Much more commonly, researchers will bend rules and influence their results in largely offhand or subconscious ways, rationalizing their behaviors away under the extreme pressure that a tenure push can create. Sometimes, efforts to massage data to reach a given p-value (such as by adding just enough to your *n* to achieve the p-value you want) is referred to as **p-value hacking.**

Not Just Psychology, Not Just the Social Sciences

The p-value crisis is often associated with the social sciences in general and the fields of psychology and education specifically. This is largely due to the inherent complexities of human-subject research, which typically involves many variables that researchers cannot control; the inability to perform true control-grouped experimental studies due to practical or ethical limitations; and the relatively high alpha thresholds associated with these fields, typically .05, which are necessary because effects studied in the social sciences are often weak compared to those in the natural or applied sciences.

However, it is important to be clear that the p-value problem exists in all manner of fields, including in some that are among the "hardest" of scientific disciplines. In a 2016 story for Slate, Daniel Engber writes of much cancer research, "much of cancer research in the lab—maybe even most of it—simply can't be trusted. The data are corrupt. The findings are unstable. The science doesn't work," because of p-value and associated problems. In a 2016 article for the *Proceedings for the National Academy of Sciences of the United States*, Eklund, Nichols, and Knutsson found that inferences drawn from fMRI brain imaging are frequently invalid, sharing concerns voiced in a 2016 *eNeuro* article by Katherine S. Button about replication problems across the biomedical sciences. A 2016 paper by Erik Turkhemier, an expert in genetic heritability of behavioral traits, discussed the ways that even replicable weak associations between genes and behavior prevent researchers from drawing meaningful conclusions about the relationship between genes and behavior. In a 2014 article for *Science*, Erik Stokstad expressed concerns that ecology literature was more and more likely to list p-values, but that the actual explained effects were becoming weaker and weaker, and that p-values were not adequately contextualized through reference to other statistics.

Clearly, we can't reassure ourselves that p-value problems are found only in the "soft" sciences. There is a far broader problem with basic approaches to statistical inference that affect a large number of fields.

Methods to Mitigate P-Value Problems

The first step in mitigating p-value's problems lies in the development and planning of a research study. Because of the inherent and inescapable presence of uncertainty in interpreting a study based on inferential statistics, we have to trust to **theory** when crafting our empirical studies. By that I mean that we should have a clear, logical explanation for an effect we believe might exist *before* we investigate if that effect is perceived in the numbers. Suppose I'm crafting an educational study and I have some theoretical reason for thinking that college students will perform better on a math test if they consume coffee beforehand than those who don't. That idea would derive from some sort of observation or understanding about how coffee affects the human brain, productivity, learning, etc. In other words, I would have a theoretical basis for a claim that I could then investigate empirically. The purpose of my data collection would be to confirm that theoretical idea. This doesn't eliminate problems with statistical inference, but it does at least give us more confidence that an observed effect stems from a real-world phenomenon. In contrast, if I merely use computerized tools to hunt for any significant effect, I am much more likely to find a spurious result and then come up with a rationalization for why it's real.

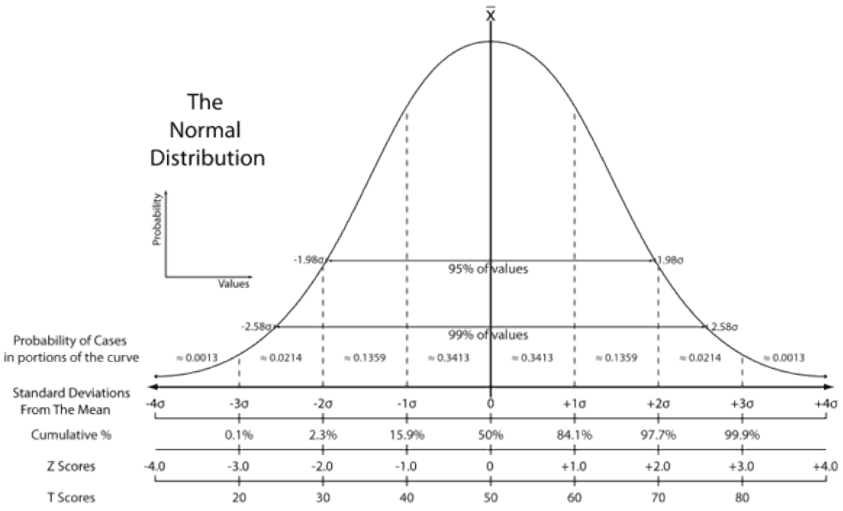
Probably the most effective and important tool for mitigating p-value problems is **replication**. In replication, researchers attempt to verify the findings of prior researchers by running the same investigations as those prior researchers as faithfully as possible. Often, this will also entail increasing the sample size of the previous study, helping to further decrease the influence of sampling error on our results. If other researchers are able to achieve similar results as prior researchers by applying the same techniques to a different sample of the same tested

population, the chance that this outcome is a statistically anomaly is decreased. (But never eliminated.) Once enough replications for a given research project have been conducted, a **meta-analysis** can be produced. A meta-analysis is a type of study that looks at previous studies and uses original statistical investigation into their outcomes to help determine just how much confidence we can have in a given reported effect. In order for either replication or meta-analysis to be conducted, researchers must be expansive and rigorous in detailing various aspects of their procedures. Unfortunately, there is a widely-perceived bias against replications in publishing standards of many top journals, as well as a tendency to see replication as less useful than original research when it comes to tenure and promotion.

Another useful technique for mitigating problems with statistical inference lies in **pre-registration**. In pre-registration, researchers report their study plans to some agency or publication *before* they gather and analyze their data. They are then obligated to report their findings in the manner initially intended and whether or not they found a significant result. This lessens the chance that they will use p-value hacking, as that behavior will be more apparent with the information provided in pre-registration. Additionally, researchers will feel some obligation to report their findings even in the event that no significant effect is discovered. This enables us to have a better sense of how common such studies are. It also gives us valuable information about relationships and effects that we have reason to believe but which do not show up in empirical data. This is important as there is no a priori reason why a lack of an effect is less interesting or useful than a positive effect.

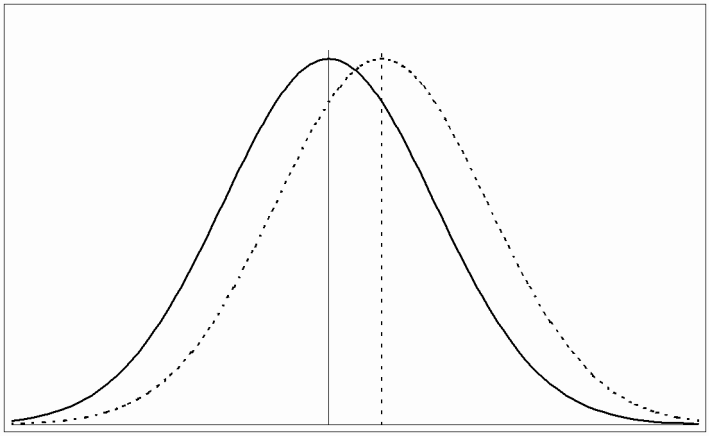
Effect Size

Another important step, for some fields, involves reporting **effect size**. Effect size refers to techniques to represent the unit-independent size of the difference between observed averages for different groups (like test and control). That is, effect sizes allow us to compare research findings from studies that use different scales and types of variables. A variable on a scale from 1-10 is not very useful to compare to, in unadjusted numbers, to a variable on a scale from 1-24. Instead, effect sizes are calculated to represented differences in unit-independent terms – specifically, in reference to the (pooled) standard deviation of the data. Reference to standard deviation is useful because the **normal distribution** (sometimes referred to as the Bell Curve) tells us how much of the data in any given distribution can be found relative to the average. Take a look at the image below.



This graphic shows where we can expect normally-distributed data points to fall relative to the mean. The mean is the line running through the center, marked $\bar{x}$. As you'd expect, most of the data points fall close to that figure – that's why it's the average! In fact, in normally distributed data, about 68% of the data points can be expected to fall within one standard deviation from the mean on either side of the average – that is, one standard deviation bigger or smaller. Around 95% of the data will fall within two standard deviations, and around 99% of the data will fall within three. If data is normally distributed, in other words, we can use standard deviation to express distances along the scale. Someone who scores on the exact average of a sample of test scores, for example, has scored one standard deviation below someone who outscores 84.1% of the other test takers.

Now consider that we are comparing different averages when we do most inferential statistics. We'll therefore get two different averages and two different distributions. Effect size helps us tell how different they are. Look at the graphic below.



Here we have two different distributions with two different averages represented by solid and dotted lines. As we can see, the dotted line sample has a higher average and a comparatively higher distribution overall. There is, however, a lot of overlap – that is, many data points in the solid-line sample are higher than many data points in the dotted-line sample, despite the averages. This, again, is a consequence of living in a world of variability. If those two samples used different units and scales, how could we compare the averages and distributions? Through reference to the percentiles and standard deviations in the first graphic. The numbers might reveal to us that the dotted-line average is .5 of a standard deviation higher than the solid-line average. Whether this is statistically significant to a given alpha will depend on the sample size and variability; how practically significant this is depends on the field and research context. In educational data, a .5 SD difference might be huge, whereas in pharmaceutical testing, it might be judged to be inadequate.

Let's go back to the example of the coin. Again remember that the basic question is whether there is a real difference between the coin that I believe to be a trick coin and the regular coin. Let's suppose again that we've conducted the test previously described and found that, after 10,000 observations, the coin came up heads 7,000 times and tails 3,000 times. This is definitely evidence that the coin is in fact significantly different than the regular coin. But let me ask: if I gave you the chance of flipping the coin where a heads earned you a million dollars but tails resulted in your death, would you take that bet? Probably not! After all, there's still a 30% chance that you'll die. Even though we have strong confidence that there is a real difference between the coin and a regular coin – even though we've rejected the null hypothesis with a great deal of confidence – the effect simply isn't powerful enough to make a life-or-death decision based on that difference. If the coin had come up heads 9,999 times and tails 1 time, you might very well take the bet. This shows that you also have an intuitive understanding of effect size.

How are effect sizes useful for overcoming p-value problems? Remember that p-value is a means to establish our degree of trust that an observed quantitative effect is the product of the underlying variability of our data. It is not intended to represent the size of that effect. Yes, bigger effects will have lower p-values and are thus more likely to be judged significant at a given alpha, but as noted sample size and variability play a large role. For these reasons, it's easy for an effect to be significant to a very conservative alpha even if the practical power of the effect is very small. For example, I do a lot of work in corpus linguistics, where I am often working with computer programs and corpora of written texts that number in the tens of thousands. Since each of these texts can count as observations, the *n* involved in these studies is often quite large. It's often common for me to find relationships that are significant to a .01 alpha or better, but which are quite weak. In other words, we could feel confident that these relationships are "real," but also that practically speaking they do not describe a relationship that is particularly meaningful in real-world terms. This is the advantage of effect size: used in conjunction with p-value it gives us not just a sense of whether the effect is likely to be the product of sampling error but also whether it's actually meaningful in real-world terms. This alone cannot eliminate problems with statistical inference, but it helps, and in particular it might help lessen the career pressure of finding statistically significant results at all costs.

Research Nihilism is Not the Answer

It can be tempting, given all of these issues, to take a nihilistic approach to social science research. Given the inevitability of bias, error, and distortion, it would perhaps be natural to throw up one's hands and decide that inferential statistics are too misleading to be helpful. But this would be a mistake. As a species with both an intrinsic desire to understand the world and various limitations on our ability to see it clearly, we should use many different means to make observations into the nature of reality. Statistical research can't be the only such tool that we use, but it can be a powerful weapon in our arsenal if we use it with skepticism and care. Combined with qualitative research, narrative work, humanistic inquiry, historical investigation, and many other ways of understanding, research based on inferential statistics remains a key method for separating true from false – if we recognize its inherent limitations and build appropriate checks and balances into its use.

Further Reading

- Wikipedia articles on statistics tend to be consistently accurate and strong (which is still not true of all subjects). The article on the replication crisis is particularly worth your time.
- StatsDirect.com has a good, brief breakdown on p-value.
- The physician Scott Alexander's blog, SlateStarCodex, frequently considers issues of research problems that result from the use of inferential statistics, particularly regarding medicine.
- Andrew Gelman of Columbia considers why psychology bears the brunt of criticism about p-value problems, despite being relatively healthy in terms of communal understanding of these problems. In the other direction, *The Atlantic's* Ed Yong has an uncharacteristically careless take on the issue which seems not to recognize just how broad this problem is.
- Gail Sullivan and Richard Feinn published a strong primer on effect size and the necessity of reporting it alongside p-values in the Journal of Graduate Education.
- Here's a research aggregator of life sciences papers that only runs studies that failed to replicate past research. Similarly, here's a journal that publishes nothing but research that has failed to reject the null.

References

Button, K. S. (2016). Statistical Rigor and the Perils of Chance. eNeuro, 3(4), ENEURO-0030.
Engber, D. (2016). Cancer Research Is Broken. Slate.com.
Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences, 201602413.
Stokstad, E. (2014). Is ecology explaining less and less?. Science.
Turkheimer, E. (2016). Weak genetic explanation 20 years later: Reply to Plomin et al.(2016). Perspectives on Psychological Science, 11(1), 24-28.

About the Author

Dr. Fredrik deBoer is the Academic Assessment Manager at Brooklyn College, where he works with faculty to gather, analyze, and report data on student learning. He received his PhD from Purdue University in 2015, where he studied writing assessment, standardized tests of college students, and applied linguistics. He is also a freelance writer whose work has appeared in *the New York Times*, *the Los Angeles Times*, *the Washington Post*, *the Guardian*, *the New York Times Magazine*, *Harper's Magazine*, *Foreign Policy*, *the New Republic*, *Politico*, and many others. He lives in Brooklyn.

The author can be contacted at fredrik.deboer@gmail.com. If you notice an error, please contact me and I will issue a correction. If you find this resource useful and would like to tip me for it, you can donate via PayPal at https://www.paypal.me/FredrikDeBoer.