

The Sunset of 'Significance Testing'

Using 'Effect Size' Statistics for Better-informed Business Decisions

Susan Schwartz McDonald, Ph.D.
and Michael Poster, Ph.D.

November 2021

NAXION
RESEARCH › CONSULTING

Awash in p-values – and adrift in the data

A key focus of market research is interpretation of *differences*—differences between time periods, between products, between customer segments. Marketing teams who want to know how worried they ought to be, or what actions to take, routinely ask, “Is that difference significant?” or “What’s the p-value?” And while that question is just a matter of straightforward calculations, the *real relevance* of the answer is more elusive. These days, no one is short on data. What we’re often missing are the guidelines to assess what truly matters.

Prominent statisticians and behavioral scientists have engaged in colorful p-thrashing for many years without being able to dislodge it. But change has come at last.



The statistical community has finally decided to cut itself loose from null hypothesis testing ... The insights community has yet to join them

Discomfort with the way p-values are misused or abused – and widely misinterpreted – has led organizations no less authoritative than the American Statistical Association and the American Psychological Association to abandon null hypothesis statistical testing (NHST) in favor of a different estimation framework that shifts the emphasis toward the **magnitude of difference** between any two numbers, **not the probability of** observing that difference by chance. One academic

journal has gone so far as to say it will not publish p-values at all.

While consensus about the need to retire p is overwhelming, this tsunami has yet to reach the shores of the insights community—a curious island off the coast of academic research, with its own dialect and its own priorities. Habits die especially hard when they are put to the service of tracking brand metrics over time, an exercise that favors consistency of approach. The fact that market research data are largely sheltered from public view also means that methods for analyzing them are shielded from public debate. As a result, the boat has not been rocked so much on our shores.



The eternal quest for significance – and what it actually yields

Significance testing is meant to tell us whether the observed difference between two estimates should be treated as *probably* real or the product of chance based on sampling error. If the observed difference fails to reach our designated threshold of significance (e.g., 95%), the difference is deemed “probably not real.” If the difference does reach significance, we can assume it is “real”—but not necessarily *important* or *meaningful*.

It is helpful to distinguish the highly probable from the improbable – in data, as in life generally – but *real* and

meaningful are two different notions.

There are some practical limitations to using NHST for deciding what matters in our data, but its ritualistic use often distracts us from the more important considerations.

The p-value is ripe for replacement because it doesn't mean what we think it means and it doesn't necessarily help us make good decisions.

“Statistically significant” does not equate to “consequential.” Statistical significance suggests only that a difference is probably not the result of sampling error. It doesn't tell us how much we should care about it or what action to take.

Conversely, differences that fail to meet the test of significance (simply because the sample is insufficiently powered) can still be real and potentially quite consequential.

Statistical significance is heavily influenced by sample size. If the sample size is large enough, almost any observed difference will qualify as statistically significant. On the other hand, if the sample size is small – for instance, when the customer universe or the available pool of willing respondents is limited – significant findings are hard to come by, even though there may be real and important differences to consider.

Despite the presumption that significant results have not occurred merely by chance, they frequently fail to replicate. One key reason for wholesale defection from p is a “crisis of replication” that has plagued scientific inquiry for decades, undermining confidence in the conclusions we draw, even from studies that produce “highly significant” results. The reasons for this are complex – too long for this article – but they point to a basic problem. We are lured into a sense of false confidence about our data when we look to p for its credibility.



Statistical significance is a binary idea in a world shaded by gray.

While people may be tempted to blur the line when calls are close, significance testing is a binary idea. A p-value is either significant or it is not. You can, of course, grade on a curve by setting the threshold you want to

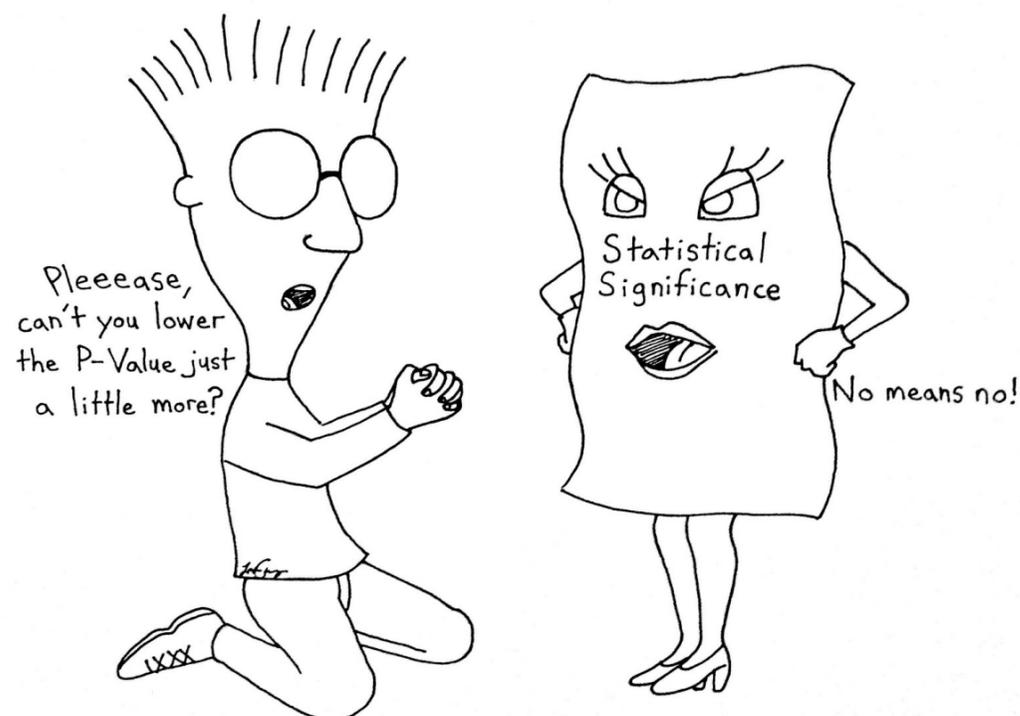
achieve, but because the test is designed to reflect sample size as much as magnitude of difference, there is often a painful arbitrariness to the outcome. A statistical difference deemed significant with $n=100$ might fail to qualify as significant with just $n=99$. To state that a number “tends

toward” significance is a statistical “wink” that violates the basic premise of the test—though it’s in line with the cloudier nature of reality, which routinely plays out on a continuum.

Significance testing does not allow for comparisons across effects or studies, limiting its use in setting priorities. A difference that yields a p-value of .0001, is not more important or more meaningful than a difference whose p-value is .05, even when they appear in the same study. Nor can a statistically significant change in Net Promoter Score (NPS) be compared to a statistically significant difference on a 6 or 7-point rating scale in the same survey. Each statistical test must be considered on its own terms.

Looking to ‘effect size’ for the real measure of meaning

Standardized effect size calculations allow the observer to consider the magnitude and potential meaningfulness of the difference between two values, unconfused by sample size. The most commonly-used effect size statistic is **Cohen’s d**, a calculation introduced over three decades ago by Jacob Cohen, a renowned “quantitative psychologist” whose statistical writing is so lively that even a non-statistician will be charmed. Cohen’s d divides the mean difference between two variables by the pooled standard deviation of those two variables. Because it is a standardized



Source: <https://analyticsindiamag.com/beware-of-statistical-significance-what-the-p-value-really-tells-us/>



measure and thus not directly driven by sample size, Cohen's d makes it possible to:

- Compare means from small sample sizes that are unlikely to yield statistically significant differences
- Compare effect sizes within and across studies – and even across dependent variables – in a way that significance outcomes cannot.
- Assess how meaningful any observed difference might potentially be, based on its actual size.

Cohen suggested thresholds of 0.2, 0.5, and 0.8 for small, medium, and large effects based on data from various scientific disciplines. Others have suggested that effect sizes in the social sciences must exceed 0.4 to be

of practical significance. NAXION has been working to assess whether any adjustments might be needed for market research databases.

Meanwhile, we are using Cohen's suggested thresholds to flag the level of importance or meaningfulness.

Hearing Signals Above Noise

One of our survey databases illustrates how a reliance on p-values as data signals can leave you adrift – or even lead you astray – and shows how standardized effect sizes can provide a sturdier but also more sensitive framework for interpretation. In the example shown, 350 respondents rated two competing products on a series of

12 attributes—the sort of dataset that makes an appearance in so many types of studies, from trackers to market landscape surveys.

Note that while Brand A is performing significantly better on all these

dimensions, looking at p-values alone doesn't allow us to prioritize, and eyeballing mean differences can be misleading. *While two of the largest mean differences (Attributes 1 and 5) look comparable, they do differ notably in their effect sizes.*

Mean Brand Differences: Significance Testing vs Effect Size

Attribute	Brand A Rating Minus Brand B	P < .05	Cohen's d
1	0.9	✓	0.73
2	1.0	✓	0.68
3	1.0	✓	0.66
4	0.8	✓	0.62
5	1.0	✓	0.48
6	0.7	✓	0.56
7	0.8	✓	0.56
8	0.6	✓	0.40
9	0.5	✓	0.39
10	0.6	✓	0.45
11	0.2	✓	0.22
12	0.2	✓	0.17

Of course, to complete the picture, we always need to understand attribute importance – as judged by driver modeling or some other measure of customer salience – and we need to know how responsive to marketing efforts any of these attributes might actually be.

In some of our larger data sets, the use of Cohen’s d has enabled us to train a sharper lens on the real magnitude of effect rather than the size of the sample, helping us spot the most meaningful trends and deprioritize “significant” effects that may not be measurable in dollars.

It’s important to note that where very large samples are driving small but

significant results, replicability becomes especially important, lest enterprise business decision-makers overreact to a small, one-time *blip* in an attempt to be highly responsive to the marketplace. The act of p-chasing has consequences.

Aligning Enterprise Decision Science with Other Sciences in the Quest for Meaning

By design, p-values give greater weight to larger samples on the presumption that a sufficient number of observations reduces error—and, indeed, very small samples *are* especially prone to the effects of outliers. On the other hand, a

myopic focus on p-values can divert our attention from other critical considerations in interpreting and prioritizing critical findings.

Of course, no statistic – not NHST and not Cohen’s d – is able to deliver “canned,” ready-to-consume meaning right off-the-shelf. Using data for effective decision support requires an interpretive framework customized to the risks and challenges of each enterprise, and a choice of statistical tools that better match our market research objectives. It’s time for the decision sciences to join the rest of the scientific community and seriously consider replacing “p” with “d”.

Suggested Reading

For those with an appetite to read more, here are some authoritative but easily digestible articles:

Amrhein, V., Greenland, S. and McShane, B. (2019). Retire Statistical Significance. *Nature*, 57, 305-307.

Cohen, J. (1994). The Earth is Round ($p < .05$). *American Psychologist*, 49 (12), 997–1003.

Szucs, D., and Ioannidis, J.P.A. (2017). When Null Hypothesis Significance Testing is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience*, 11: 390.

Cumming, G. (2013). The New Statistics: Why and How. *Psychological Science*, 25 (1): 7-29.



About the Authors



Susan Schwartz McDonald, Ph.D.

CEO

215.496.6850

smcdonald@naxionthinking.com

Susan's career focus has been on the development and protection of robust brands, and the research methodologies needed to support them. She has contributed to the evolution of many standard research techniques, and she writes frequently on industry topics and issues of broader interest. Her commentary on language and culture can be found at www.smartmouth.blog.

Susan holds MA and PhD degrees from UPenn's Annenberg School of Communication.



Michael Polster, Ph.D.

SVP, Life Sciences Practice

215.496.6913

mpolster@naxionthinking.com

Michael is a neuropsychologist and methodologist who provides critical decision support to clients commercializing new therapies and healthcare solutions. He provides intellectual leadership on a diverse set of topics ranging from launch strategy and life cycle management to health economics and regulatory compliance. Michael earned a PhD from Cambridge University.

About NAXION

NAXION is a nimble, broadly resourced boutique that relies on advanced research methods, data integration, and sector-focused experience to guide strategic business decisions and shape the destiny of brands. Our century-long history of innovation has helped to propel the insights discipline, and continues to inspire contributions to the development and effective application of emerging data science techniques. For information on what's new at NAXION and how we might help you with your marketing challenges, please visit www.naxionthinking.com

© NAXION, Inc. All rights reserved.