NAXION
thinking

Beyond Significance Testing: How to Measure the *Meaningfulness* of Statistical Change

A Tool to Supplement p-values

There's seldom a shortage of data these days. What's often missing is conviction about which data trends we should care.

A key focus of market research is the interpretation of statistical differences and whether they actually matter. When marketing teams want to know how pleased or worried to be about a change, or whether a difference between two numbers confers leverage, they routinely ask researchers, "Is it *significant*?" By which they mean: Is this difference *real* or merely the product of chance? Although the question can be answered based on the simple calculation of a *p*-value, the *relevance* of the answer is less straightforward. *P*-values have long been a North Star for guiding decisions based on data, but they can also lead us astray. It's important to understand what *p-values* offer, what their limitations may be, and what else we can do when *p* alone cannot provide adequate decision support.

Significance testing has infiltrated even our popular culture – so much so that an American voter who keeps an eye on polls is familiar with terms like "statistically significant" and "within the margin of error." Ironically, researchers and statisticians have been growing more prone to challenging its use in a variety of situations. Uneasiness about the way *p*-values can be misinterpreted (and abused) has led prominent organizations like the American Statistical Association and the American Psychological Association to largely abandon the use of Null Hypothesis Testing (NHST) in favor of a different estimation framework that shifts the emphasis *toward* the *magnitude* of difference between numbers and *away from* the *probability of observing* that difference by chance. One respected academic journal has gone so far as to say it will not publish *p*-values at all. So after years of colorful p-thrashing by statisticians themselves, there is growing consensus in the sciences that we need a shift in focus from significance to meaningfulness. But despite a much-trumpeted focus on methodological innovation, the "insights industry" island has been largely insulated from the debate about *p* for several good reasons.

Habits die hard everywhere, and the market research community is no exception. In our neighborhood, the *p*-habit has long been reinforced by the need for consistency of approach in the interpretation of tracking metrics. Moreover, since most market research studies are largely proprietary, research conclusions and methods are not necessarily subject to peer review or academic debate. As a result, many highly experienced market researchers have had little or no exposure to the charges leveled against null hypothesis testing. That has meant less opportunity to become familiar with alternative statistical approaches.

So why do we need alternatives to p?

There are some practical limitations to NHST that spring from its intellectual origins and its method of calculation. NHST 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% probability), the difference is deemed "*probably not real*." If the difference *does* reach significance, we can assume it is *"probably real."* In data, as in life generally, it is helpful to distinguish the highly probable from the improbable, but "probably real" and "meaningful" are two very different notions – which is to say, "statistically significant" does not equate to "consequential." Significance does not tell us how much to care or whether to take action. Conversely, differences that fail to meet the test of significance can still be real and potentially quite consequential. The inability to address meaningfulness, a fundamental limitation of NHST, is closely related to its other deficiencies.

Statistical significance is heavily influenced by sample size, which can mislead us about what is consequential in the data.

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 – statistically significant findings are hard to come by, even though there may be real and important differences to consider. The rationale for reflecting sample size so heavily in significance calculation is rooted in the premise that sample size can correct for sampling error but the temptation to trust statistics based on large sample sizes alone needs to be resisted.

Despite the statistical *presumption* that results deemed 'significant' have not occurred merely by chance, they frequently fail to replicate – which is a key reason for the recent wholesale defection from *p* in the scientific community.

This "crisis of replication" has plagued scientific inquiry for decades, undermining confidence in the conclusions drawn even from studies that produce "highly significant" results. The reasons for this apparent anomaly have to do with the nature of *p* as a concept and a calculation, as well as the way in which researchers tend to frame and test hypotheses in pursuing the holy grail of significance. The crisis of replication demonstrates how easy it is to be lured into a sense of false confidence about data when we look only to *p* to establish its credibility.

Statistical significance is a binary idea in a world shaded by gray, even though people may be tempted to blur the line when calls are close.

In NHST, a *p*-value is either statistically significant or it is not. You can, of course, grade on a curve by lowering the bar (e.g., from .05 to .10). But because both sample size and magnitude of difference influence the outcome, decisions made on the target threshold can seem (and can be) arbitrary. 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.

Finally, 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. Unfortunately, researchers don't necessarily appreciate or live by that rule; they are often tempted to draw inferences and set priorities based on comparisons of *p*-values.

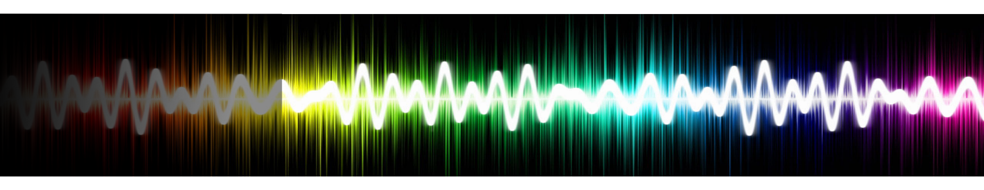
While p remains important, there are other tools available to help us decide what's meaningful in our data.

Standardized "effect size" calculations allow the observer to consider the magnitude and potential meaningfulness of the difference between two values unconflated by sample size—so long as there is confidence in the estimate the sample has provided. (In other words, you trust your instrument and your respondents to deliver valid information.) 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 may 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 measure, and thus not directly driven by sample size, Cohen's d makes it possible to:

- *Compare effect sizes from smalls 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 for small, medium, and large effects based on data from his experiments with psychology students. Others have suggested that effect size in the social sciences must exceed 0.41 to be of practical significance. NAXION is currently using Cohen's suggested thresholds to flag the level of importance or meaningfulness, but we are working to assess whether any adjustments might be needed for market research databases.

Hearing Signals Above Noise



Cohen's d has helped us spot meaningful trends that might have gone overlooked due to sample size, or given us grounds to deprioritize "significant" effects that are unlikely to yield a meaningful return.

By back-stopping significance testing with Cohen's *d,* we've been able to reinterpret what looked like a significant preference for one product over another as, instead, a dead heat and, in other cases, have confirmed that a non-significant trends in market tracking were meaningful enough in d-terms to warrant serious attention. In an upcoming webinar, we will work through some actual data examples.

Wear a Belt with Your Suspenders

Of course, no statistic is able to deliver canned, ready-to-consume meaning right off the shelf. There's always art to the behavioral science. Using data for effective decision support requires an interpretative framework customized to the priorities and challenges of each enterprise. *The case we are making is not to abandon p values. It's to look at "d" (or in the case of percentages and ratios, Cohen's h) for a more complete and more nuanced picture* – especially in verticals like healthcare, B2B, technology, and even in broad consumer categories when subsegments of special interest may be hard to find or afford. Effect sizes can give added decision support when your differences don't meet a target significance threshold, and they help avoid over-reaction to blips driven by large sample size. It's time for decision science to join the other sciences by adding a few letters to our statistical alphabet.

About the Authors

Susan Schwartz McDonald, Ph.D.
CEO, NAXION
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 M.A. and Ph.D. 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 strategic decision support to clients commercializing new therapies and healthcare solutions. His areas of expertise include launch strategy and life cycle management as well health economics and regulatory compliance. Michael earned a Ph.D. from Cambridge University and continues to write and speak regularly on research methodology.

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 that 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

Updated, Fall 2024

NAXION
thinking