Editorial

Statistical Power Analysis and the contemporary “crisis” in social sciences

Journal of Marketing Analytics (2016) 4, 61–65. doi:10.1057/s41270-016-0001-3;
published online 28 November 2016

INTRODUCTION

Statistical Power Analysis has been a largely neglected area. It's a topic that rarely gets much attention, even in graduate education, an unfortunate omission. In the era of “Big Data”, some people might think that statistics and significance testing aren't needed as much. But the reality is that even when you have lots and lots of data available (which makes the tiniest differences “statistically significant”), there are always problems that require deep insight into Statistical Power Analysis.

We don't analyze data to find “nice-to-know” tidbits. When you surface remarkable business facts, the aim is to spur new insights and innovate existing business practices. When companies then attempt to *act* on these analytical findings, they will (almost) always do so in piecemeal, experimental fashion. Implementation is where the rubber meets the road.

An online retailer, for example, might want to test a redesigned landing page: how does this affect click-through rates? A pharmaceutical company wants to test a new compound: are treatment outcomes significantly better? A direct marketing company wants to test several offer letters in parallel (albeit offered to different customers): which one generates the most response?

What all of these examples have in common is that they aim to test, statistically, whether some test group outperforms one or more control groups. But not only is this

potential *difference* of importance, in every case you *also* want to be confident of the converse that you don't abandon a potentially promising new alternative prematurely, even if your initial statistical test is inconclusive.

Note that there are two ways to apply statistics: either to describe what happened in a particular study, or to draw conclusions about the implications of those findings in a broader context. The former is referred to as descriptive statistics; the latter is called inferential statistics. The risks of incorrect inferential statistics were covered extensively in Cohen's classic title of 1977 on Power Analysis. When I use the term statistics, in general I loosely refer to it as inferential statistics, since this is what most people associate it with.

When you test a hypothesis, there are fundamentally two types of errors you can make: either you reject the Null hypothesis, when in fact it is actually true (Type I error), or you fail to reject the Null hypothesis when it is false (Type II error). An example of a Type I error would be when a test result leads to the conclusion that the online retailer's new landing page generates higher click-through rates, when in reality this is not the case. In the second example of a Type I error, you decide that the new compound is superior, but it is not, or you conclude that one of the offer letters generates more response when this increase was actually caused by random variation.

The odds of making a Type I error are generally well known and understood. These odds are reflected in the α or p value. There is an extensive body of research and literature, and various analytical methods are available, if you would like to understand your results better.

Type II errors are less well understood, and historically much less attention has been devoted to this area. This holds in particular for the academic world. For practitioners, the failure to reject a Null hypothesis (often concluding that there is *no* difference between two treatments, for instance) carries a risk and can have substantial negative business implications. Statistical Power Analysis has gotten little attention in the academic world, but it certainly has considerable practical consequences for applied researchers working in commercial fields.

The ‘classic’ scenario where Statistical Power Analysis plays up in commercial settings is when some new treatment or sales tactic is compared alongside the existing approach (“business as usual”). Sometimes, this type of experimental design is referred to as a “champion–challenger” strategy.

From a decision theory perspective, *both* types of errors (Type I and Type II) need to be controlled. You don’t want to risk concluding that there is a difference, when in reality you arrived at this erroneous conclusion due to random variation. But you also don’t want to conclude that there is *no* difference, just because your statistical test doesn’t allow you to discern a (statistically significant) difference that in reality actually *is* there.

Statistical Power is defined as the value for $(1 - \beta)$. There is no commonly agreed threshold we consider acceptable for the value of β . In social sciences, the unspoken rule is that α values (the odds of making a Type I error) of 5 per cent or less are the norm for inclusion in academic papers. Of course, once a significant result has been established, there is little if any point in reporting on β values: the finding has already been established.

A “BIAS” AGAINST STATISTICAL POWER ANALYSIS

Why is it that Statistical Power Analysis has received so little attention? The main reason for this seems to stem from our *biased* exposure to results: most academic journals only publish results when studies find significant results. Otherwise, editors and reviewers for the most part are hardly interested in publication of the findings.

Many research papers test multiple hypotheses in the same study. As might be expected, some of these hypotheses will be significant and some not. Even in those (published) studies, rarely is there any mention of the β values, or their converse $(1 - \beta)$, for the nonsignificant results.

In many contexts, statistical tests are run when there is at least some effect (the population value is greater than zero). If that is the case, then Statistical Power equals the probability that your test will lead to a correct conclusion about the Null hypothesis.

These two factors combined, lack of exposure during graduate education and (very) limited mention in academic publications together make that students and even practitioners are often unfamiliar with Statistical Power Analysis. As is obvious from this decision matrix, there are two ways to err, either Type I or Type II, but more importantly, the underlying process that is represented in this decision matrix is often poorly understood.

STATISTICAL POWER ANALYSIS IN THE AGE OF BIG DATA

We are entering the age of Big Data. Volumes of data are growing at unprecedented rates. Prices of storage and computing power continue to drop, and broad adoption of cloud computing causes further spiraling down of costs. As our lives turn increasingly digital, and new

		What Is True in the Population?	
		Treatments Have No Effect	Treatments Have An Effect
Conclusion Reached in a Study	No Effect	Correct Conclusion ($p = 1 - \alpha$)	Type II error ($p = \beta$)
	Treatment Effect	Type I error ($p = \alpha$)	Correct Conclusion ($p = 1 - \beta$)

↑
Power

Figure 1: From Murphy *et al* (2009)

technologies lead to even more data creation, we find new ways to apply all of these data every single day. Research agency McKinsey refers to this era of Big Data as the fourth industrial revolution.

It hasn't been long that more data are being generated about us than we produce ourselves during a lifetime. Gartner estimates that every day 2 ExaBytes of data are being generated. Note that the entire Library of Congress holds "only" about 15 TeraBytes of text¹!

As the volumes of data we analyze continue to grow and grow, increasingly smaller effect sizes can be surfaced (i.e., demonstrated with significant difference). In social sciences, we were used to running experiments on hundreds and sometimes thousands of subjects. In the era of Big Data, we can collect data on millions of events, yet continue to rely on the same statistics, and also largely the same criteria for labeling findings as "significant."

Indeed, the same criteria apply in our decision matrix. There is an important difference, though. As the volume of data grows, the N we use for making statistical comparisons, our statistical tests are still valid,

but we need to reconsider if traditionally used criteria for referring to a finding as significant might need to be revisited. The implicit norm of 5 per cent as the hallmark of a statistically significant finding, of course, also implies that you are willing to make an error of falsely assuming some effect, when in effect there isn't one. And our willingness to take that chance is reflected in the p value = α .

There are fundamentally three factors that contribute to the statistical power of a particular test: the magnitude of the effect as it exists in the population, our choice for the statistical test we are employing (which is mostly related to our choice for a methodological design), and the size of our experimental groups, N . These three factors together form an iron triangle that drives the distribution of cells in our decision matrix shown in Figure 1.

For many common statistical tests, statistical power goes up roughly as a function of the square root of N . As the number of records in very large databases goes up, this dramatically impacts the likelihood of finding an effect if you keep your Type I threshold constant at 5 per cent.

MINIMAL EFFECT SIZE

In many settings, and this certainly pertains to the business world, our most pressing concern isn't really whether there *is* in fact a (any) difference at all between two treatments. In business settings, we are typically not so much interested to know *whether* there is zero difference, as we are in discovering if the effect of some treatment exceeds a certain ("Minimal") threshold. That is where the use of the word "Minimal" comes in: we want to know by means of statistical testing if the difference between two (or more) treatments exceeds some purposely chosen threshold (Minimal Effect Size).

Bear in mind that when we formulate the Null hypothesis to state that there is no ("zero") difference between the experimental and control groups, the choice of the number "0" as the critical value is completely arbitrary. In most settings, certainly in a business context, when a Null hypothesis gets tested, we actually assume beforehand that there is some effect. Under this assumption, power ($1 - \beta$) often translates into the probability that your statistical test leads to a correct conclusion about the Null hypothesis.

Sometimes, it makes sense to distinguish between significance and relevance of findings. For business purposes, a difference *may* be statistically significant, yet not quite large enough to be relevant, which is relevant in the sense of worth acting on. Notably in a context where you have access to very large volumes of data, then statistical tests can become very powerful. Powerful here equates being able to flag relatively small differences between groups, yet determine that those differences are statistically significant. It is possible to find results in this way that business stakeholders deem insufficiently large to act upon.

Of course the converse can hold, too. As an example, when we test the effect of a new medicine that can save lives, the results in an early trial may prove nonsignificant and thus indecisive. This scenario occurs quite

frequently, especially with the so-called early clinical trials, testing new compounds on small groups of experimental subjects.

But if this experimental drug displays lower mortality rates in these early clinical trials, the empirically established "best estimate" of its effect size might well be "relevant" – relevant as in commercially viable for the pharmaceutical company that is trying to go to market with a new medicine. This finding is clearly *relevant* (albeit not statistically significant) from a business perspective. Despite the fact that in the current sample the effect size doesn't allow a conclusive decision (yet), the effect may be deemed large enough to merit further investments. That might be finding additional research subjects to attempt to replicate the early findings, or investing in research to establish whether it is worthwhile to continue developing this medicine or treatment, or not.

From this example, it should be clear that you need to distinguish between *relevance* and *significance* of findings. Sometimes, a finding can be relevant, even if the test is not significant, and this can occur provided that the treatment effect is of the desired direction, but not large enough to produce statistically significant findings given the current sample size.

As I explained, a finding can also be significant, yet not relevant, in a business context that usually equates to an effect that does not merit further action. Even though there is a significant result, that you expect to uphold, that does not mean it can be commercially exploited. In many industries, there are start-up costs, barriers to competition, etc., that preclude corporations from exploiting findings from their research.

CONCLUSION

Researchers, and in particular academic researchers, have been focusing mostly on finding significant results, rather than worry about the odds of *not* finding a significant

result when in fact the experimental treatment does have some effect. The academic tradition that mostly only significant findings get published has led to a bias that I have alluded to.

Somewhat ironically, social sciences are currently going through a bit of a crisis. Many findings that had been the basis for theory in the social sciences have shown not to be replicated. Last year, a group of some 250 researchers attempted to replicate 100 studies that had been previously published in three psychological journals (Open Science Collaboration, 2015).

Since then, there have been papers published both in favor of as well as opposing these results (e.g.: Gilbert *et al.*, 2016). The reason why the original Open Science Collaboration from 2015 paper stirred up so much controversy was that several of these studies they failed to replicate had been the basis for rather elaborate theory formation. Needless to say, this paints a rather disturbing picture of the state of affairs in science.

In my opinion, there are a few lessons to be learned. Despite the fact that the statistics for both the original studies as well as the meta analyses that were later performed have been known and documented for a few decades, a surprising number of discussions fail to reference the pertinent texts. I find this highly alarming, and also indicative for the state of affairs in science. Some would like to think, maybe hope, that this merely holds for social sciences, but I am not so optimistic.

Another point that can be taken from this contemporary crisis is that theories are never proven or disproven by a single study. Yet, that appears to have been the culprit that made the Open Science Collaboration paper so alarming: for many theories, there was an alarming low number of confirmatory studies. Especially when a study is used as a cornerstone for theory formation, it behooves the researcher to replicate some variations of the original study to firmly establish its concepts.

An experimental finding is always stochastic. It represents the outcome of a hypothesis test that has been demonstrated with a specified chance attributed to it. Therefore, it should always be considered as merely one step in the process to theory formation. If I understand and read the facts correctly, the current crisis in social sciences seems to have been caused by all-too-opportunistic conclusions that were drawn from relatively isolated research findings. In other words, the “merely one step” toward formulation of theory part was neglected.

In particular when research findings happen to either pique particular interest, or, when they happen to conveniently fit the existing frame of mind, one needs to be extra cautious not to make broad inferences. Without solid knowledge of Statistical Power Analysis, making broad, sweeping inferences is risky at best. As Jerry Weinberg has said, “a crisis is often merely the end of an illusion”, and it appears that many scientists have been clinging to their illusion in the face of less-than-compelling mastery of well-established statistical concepts like Statistical Power Analysis.

NOTE

1. ExaByte amounts to 1,000,000 TeraBytes.

REFERENCES

- Cohen, J. (1977) *Statistical Power Analysis for the Social Sciences*. New York, NY: Academic Publishers.
- Gilbert, D.T., King, G., Pettigrew, S. and Wilson T.D. (2016) Comment on estimating the reproducibility of psychological science. *Science* 351(6277): 1037.
- Murphy, K.R., Myors, B. and Wolach, A. (2009) *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. 3rd edn. New York, NY: Taylor & Francis Group, LLC.
- Open Science Collaboration. (2015) Estimating the reproducibility of psychological science. *Science*. doi:[10.1126/science.aac4716](https://doi.org/10.1126/science.aac4716)

Tom Breur
Marlborough, MA, USA
E-mail: tombreur2@gmail.com