In p we trust (too much). Simulations of typical p-value and effect size behavior in advertising research

Tim Smits
Institute for Media Studies, KU Leuven
Ben Van Calster
KU Leuven

Cite as:
Smits Tim, Van Calster Ben (2019), In p we trust (too much). Simulations of typical p-value and effect size behavior in advertising research. *Proceedings of the European Marketing Academy*, 48th, (9521)

Paper presented at the 48th Annual EMAC Conference, Hamburg, May 24-27, 2019.

In *p* we trust (too much). Simulations of typical *p*-value and effect size behavior in advertising research

Abstract:

With this simple illustration we aim to demonstrate the most important issues concerning the continued use of *p*-values and null hypothesis significance testing. We do this based on a set of simulations that pertain to the setting of an advertising researcher investigating a basic effect in a control-versus-experimental between-subjects design. These simulations were performed for different hypothetical true effect sizes and sample sizes and, in a second set of simulations, for some reference values coming from advertising meta-analyses.

Important issues demonstrated in these simulations are the high variability of p-values, the variability of effect sizes, and the over-confidence when significant findings occur. These effects are worse for smaller sample sizes. We also focus on the effect of shifting the significance criterion from .05 to .005 as has been suggested recently. We also reflect on how these issues partially contribute to the replicability crisis.

Keywords: null-hypothesis significance testing, replicability, advertising effects

Track: Advertising & Marketing Communications

1. The history of p-values and significance testing

There are many critical papers about the interpretation of p-values and significance testing in the scientific literature. Many researchers, though, still struggle with using proper alternatives to the flawed “null hypothesis significance testing” (NHST) approach. This paper aims to contribute to a solution by demonstrating (rather than lecturing) what is inherently wrong with the common NHST approach. This is needed given the catch-22 of NHST, as described by George Cobb during a discussion forum of the American Statistical Association (ASA; see Wasserstein & Lazar, 2016, p. 129):

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
>
> A: Because that’s still what the scientific community and journal editors use.
>
> Q: Why do so many people still use $p = 0.05$?
>
> A: Because that’s what they were taught in college or grad school

As has been written elsewhere, but apparently forgotten in many statistics courses, the current use of a strict p-value criterion to denote significance mostly is an amalgamation of the two basic approaches to statistical inference. On the one hand, this pertains to Fisher’s approach, which was rather exploratory. On the other hand, it pertains to the methodology developed by Neyman and Pearson, which was much more confirmatory (see, e.g., Gigerenzer, 2004; Perezgonzalez, 2015). Without dwelling on all the details of these two methodologies, one can summarize them as follows. Fisher conceived his approach with the exploratory design in mind. For this purpose, he considered p-values to be informative, but suggested to use them in an absolute sense, without referring to a cut-off point. Hence, a researcher was supposed to evaluate that p-value given the prior knowledge, given other data etc. One could believe the findings to be significant, but that did not refer to some type of proof. Rather, it should be interpreted as “interesting”. Importantly, p-values were indeed considered as the probability (under a specified statistical model) to observe the finding (as summarized in a statistic) or an even more extreme one, given a null hypothesis. This definition is still true for p-values, but the binary notion of significance is something we came to attach to it in the NHST approach.

Neyman and Pearson, however, had a very strict confirmatory design in mind. Here, researchers should define their hypotheses a priori and in detail. Based on an estimated effect size, one should plan the study and denote regions of rejection for those hypotheses (e.g., one believes the effect size $d = .3$ and based on sample size decides to reject that hypothesis if the observed d is outside a specified interval around .3). Note that this is remarkably similar to the strict pre-registration protocols for replications we have seen emerging as an answer to the replication crisis in our field and others. The researchers then collect data and analyse them as planned. Comparing the results with the initial regions of acceptance and rejection than results in a much firmer and binary conclusion accepting or rejecting the hypotheses.

One can easily see that in current research practice, also specifically within advertising research, we have come to confuse some of these concepts. We do use a p-value, but are inclined to interpret it binary as an acceptance or rejection of a very specific hypothesis (which we often only articulate in a detailed way after doing the analyses). Guidelines about *p*-values such as the ASA's (Wasserstein & Lazar, 2016) are useful, but they might miss the persuasiveness needed to radically or sufficiently change the methodological practices.

A last issue to address for now is that the use of NHST is also partly deemed responsible for the replicability crisis, recently inspiring some researchers to lower the significance criterion value to .005 (Benjamin et al., 2018). This of course decreases the number of false positive findings. However, others claimed that researchers should justify their alpha criterion rather than mechanically applying the one or the other threshold (Lakens et al., 2018).

With the current demonstration, we aim to tackle the issues we think are the most compelling reasons to rethink the NHST in advertising research. We use simulated data from a number of different standard situations that might resemble the advertising researcher's reality. Of course, as a researcher in a particular study, we are only confronted with the data and do not know the true effect size. We therefore demonstrate p-value and effect size behavior pertaining to populations differing in effect size, ranging from no effect (Cohen's *d* = 0) to a large effect (*d* = .8), a size that is not that common according to advertising research norms. We also, of course, simulate different sample sizes because, contrary to the true effect size, this is a factor that researchers can control themselves. Researchers typically understand that small sample sizes have adverse effects in a research setting, but we want to demonstrate how exactly they affect the behavior of *p*-values and effect sizes. In a second set of simulations, we run specific simulations based on some of the well-documented effects in advertising research, to further illustrate the relevance to the advertising research field.

2. Simulation in general

2.1. Methodology of Simulation 1

Data were simulated using R3.2.2. All simulations pertain to a very simple setup where we have one binary independent variable possibly affecting one dependent variable. This is the most basic experimental design, typically analyzed with a t-test. This is an adaptation of similar simulations recently published to illustrate these issues in a clinical setting with binary outcomes (Van Calster, Steyerberg, Collins & Smits, 2018). For each of a set of standard situations we took a population effect size (an average difference between the control group and the experimental group) for granted and sampled "studies" from this population. We included the following effect sizes: Cohen's *d* = [0, 0.2, 0.4, 0.6, 0.8]. We varied the sample size from a small N situation (15 participants per condition), over more realistic sample sizes that are often suggested in methodology courses (N = 30 or 50 per condition) to a large N situation (100 participants per condition). Hence, each of these situations is defined by a population effect size and a decided-on sample size. For each situation we then simulated 1000 experiments and calculated the relevant statistics: the *t*-test's *p*-value and the sample's observed effect size *d*, and the observed effect sizes for the subset of simulated experiments that turn out to be significant. The latter is done separately with the significance criterion set to .05 and to .005, as such enabling an illustration of the consequences of adopting this criterion as suggested by Benjamin et al (2018).

2.2. Results of Simulation 1

Given the demonstrative goal of this paper, we summarize the findings mostly visually (see Figure 1). We will present these more extensively, but for the purpose of this extended abstract, we will focus on some crucial issues that should be apparent from this visual presentation.

First, *p-values are highly variable*. In general, the smaller the sample size and the smaller the true effect size, the more variable the *p*-value was within the 1000 simulations for those parameter settings. This also results in a high variability of the proportion of significant findings within each subset of simulations. For the small effect size (*d* = .2) with small sample size (N = 15) simulation only 9% of simulations was significant. In contrast, for the larger effect sizes (*d* = .6 or higher), a sample size of 50 or more per condition did result in the typically aimed for power of at least 80%.

Second, it is interesting to evaluate the *variability of observed effect sizes*. The observed effect size in these simulation is an unbiased estimate of the true effect size, but the observed variability again is impressive (when compared to how researchers typically imagine them). Again, this variability is much more pronounced in the small sample size cases.

Third, and as a corollary of the above, we consider the *observed effect size for significant (p<.05) results*. A researcher might believe that the observed effect size is an unbiased predictor of the true effect size. As Figure 1 clearly demonstrates, this simply is not true when we consider the subset of significant findings only. Within this subset, there is a large overestimation in these observed effect sizes relative to the true effect size and this again is much more pronounced for smaller sample sizes and smaller true effect sizes.

Finally, we can also evaluate what the *proposed .005 significance cut-off* would imply for the overestimation of effect sizes. Figure 1 clearly demonstrates that applying this criterion value results in even stronger over-estimation of effect sizes for significant findings. It follows that if we would only report these more extreme findings in literature, future replication attempts could be even more biased in their power analysis.

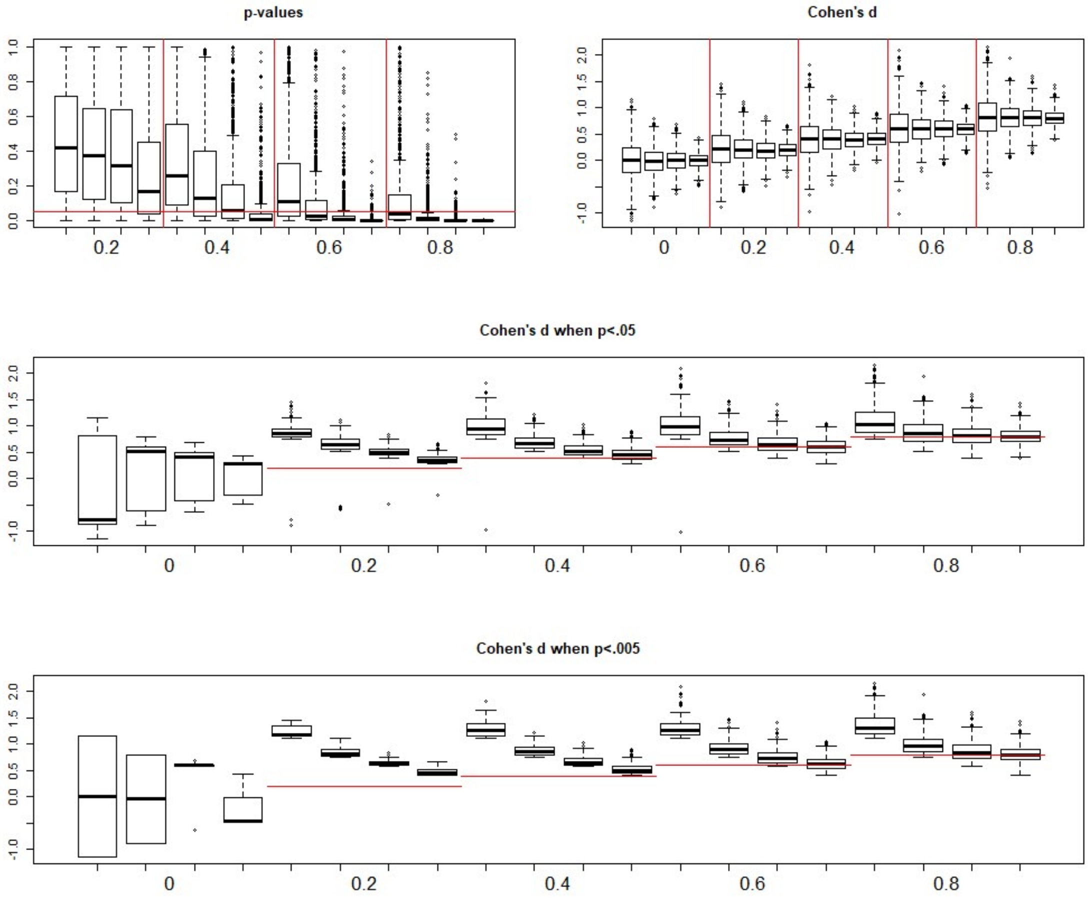

Figure 1. *p*-values, observed Cohen's *d* effect sizes, and effect sizes for significant findings (when using the .05 or the .005 criterion; with a reference line for the true effect size) for different assumed effect sizes (*d* = 0, 0.2, 0.4, 0.6 or 0.8) and different simulated sample sizes (15, 30, 50, or 100 participants per condition).

3. Simulation with advertising effects benchmarks

3.1 Reference values from the advertising literature

For the sake of brevity in this extended abstract, we only look at three different benchmarks and we apply these benchmark values to further substantiate the claim that the illustrations do apply to an advertising research setting. In selecting benchmark data, we look for meta-analyses that suggest established findings within the domain and that pertain to experimental studies. Before discussing the cases, it must be noted that the meta-meta-analysis by Eisend and Tarrahi (2016) suggested an overall effect size of $r = .2$, which roughly corresponds with a $d = .41$. The authors did highlight the fact that experimental studies typically result in somewhat higher observed effect sizes (maybe due to the processes we highlighted above), but this result does already suggest that many cases within the advertising research tradition may suffer from the issues highlighted above.

A first case pertains to the widely popular and often studied technique of *endorsement* advertising. Knoll and Matthes (2017) recently published a meta-analysis and reported an average effect size of $d = .24$ of the endorsement versus no endorsement manipulation on attitudes towards the advertised products. This effect size is supposedly based on studies with, at most, 81 participants per condition (data from personal communication with Matthes; some studies had more than only 2 conditions, thus 81 is an overestimation erring on the safe side for the sake of our illustration).

A second case pertains to the more recent, but equally popular technique to use *advergames* to market to children, a technique predominantly used in the food domain. Folkvord and Van ‘t Riet (2018) published a meta-analysis on this topic and found an overall effect size of Hedges’ $g = .30$. Based on their article, this is based on experiments with, at most, 137 participants per condition, but again many of these studies had more than just two conditions (thus making the 137 per condition an exaggeration; again erring on the safe side when using this in the illustrations). Due to this rather large sample size, we assume that Cohen’s d will approximate Hedges’ g for this case.

A final case pertains to the effect of exposure to *food advertising* on children’s subsequent consumption. A meta-analysis (Boyland et al., 2016) suggested a Cohen’s $d = .56$. Based on their detailed report, this is based on studies with an average of 53 participants per condition.

We now follow a similar simulation approach as above to simulate data, but use the reference values of these three cases to demonstrate our main points. To do so, we will assume that the reported observed effect sizes from these meta-analyses correspond to the real effect sizes. This is, of course, a mere assumption and the first set of simulations already suggests that this could be an overestimation as well. Still, to err on the safe side, we follow this assumption. Together, this imply that we will probably underestimate the real effects of *p*-value variability should a researcher do a conceptual replication in one of these domains, following the sample size tradition within that line of research.

3.2 Results of Simulation 2

Based on the reference values, the simulations suggest rather low replicability when the replications would follow the typical sample sizes. The endorsement effect only replicated in 33% of the simulations, the advergame effect in 69%, and the food advertising effect replicated in 83% of simulations (due to its stronger effect size and despite its lower typical sample size). Lowering alpha to .005 reduced these successful replications to 11%, 37%, and 50%, respectively. Given that publication is biased towards significant findings, this implies that at least half of the "typical" studies in these domains would not even be publishable under the stricter alpha criterion. But even according to current standards of an alpha .05 criterion, 17% to 67% of studies following a typical design of these rather established effects would not result in significant findings.

Figure 2 further illustrates the previous issues. First, there is a large variability in *p*-values, thus affecting the replicability of the original meta-analysis findings. Second, even for these established effects, we see a large variability in observed effect sizes. Third, within the subset of significant simulations, there is a clear overestimation of the true effect size. It follows from the above that this overestimation further increases when applying the .005 alpha criterion.

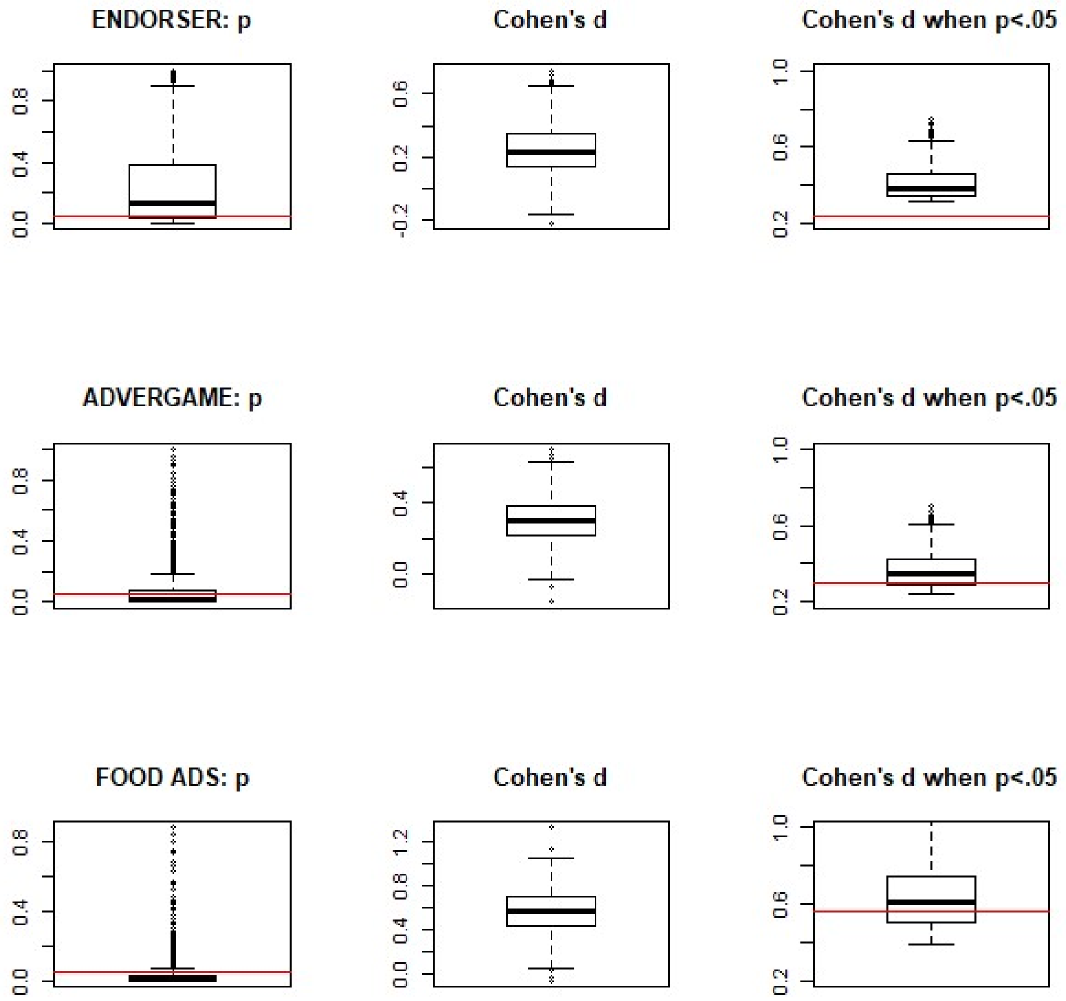

Figure 2. *p*-values (with a p=.05 reference line), observed Cohen's *d* effect sizes, and effect sizes for significant findings (*p* < .05, with a reference line for the meta-analytic effect-size used as a "true" effect size in the simulation) for three different advertising effects: Endorsement advertising (*d* = .24; *N* = 81 per condition), Advergames (*d* = .3; *N* = 137 per condition), and Food ads (*d* = .56, *N* = 51 per condition)

4. GENERAL DISCUSSION

With this simple illustration, we aimed to demonstrate the most important issues concerning the continued use of p-values and null hypothesis significance testing. Important issues are the high variability of p-values, the variability of effect sizes, and the over-confidence when significant findings occur. These effects are worse for smaller sample sizes.

One way to deal with this issue is to systematically increase sample sizes and to refer to confidence intervals around the effect sizes. Furthermore, we (authors and reviewers alike) should realize that significance testing doesn’t have too much contribution in a strict exploratory setting and that we should rather learn how to interpret statistics such as the p-value in these exploratory settings, much like these statistics were once intended to be used.

References.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Boyland, E. J., Nolan, S., Kelly, B., Tudur-Smith, C., Jones, A., Halford, J. C., & Robinson, E. (2016). Advertising as a cue to consume: a systematic review and meta-analysis of the effects of acute exposure to unhealthy food and nonalcoholic beverage advertising on intake in children and adults. *The American Journal of Clinical Nutrition*, *103*(2), 519-533.

Eisend, M. & Tarrahi, F. (2016) The Effectiveness of Advertising: A MetaMeta-Analysis of Advertising Inputs and Outcomes. *Journal of Advertising, 45*:4, 519-531, DOI: 10.1080/00913367.2016.1185981

Folkvord, F., & van 't Riet, J. (2018). The persuasive effect of advergames promoting unhealthy foods among children: A meta-analysis. *Appetite*.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606.

Knoll, J., & Matthes, J. (2017). The effectiveness of celebrity endorsements: A meta-analysis. *Journal of the Academy of Marketing Science*, 45(1), 55-75.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Buchanan, E. M. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168.

Perezgonzalez JD (2015) Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology,* **6**:223. doi: 10.3389/fpsyg.2015.00223

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology*, *66*, 47-54.

Van Calster, B., Steyerberg, E. W., Collins, G. S., & Smits, T. (2018). Consequences of relying on statistical significance: Some illustrations. *European Journal of Clinical Investigation*, *48*(5), e12912.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician*, *70*(2), 129-133.