A cautionary tale about $p < 0.05$

Pranadharthiharan Narayanan
Nova School of Business and Economics
Sumit Malik
University of Liverpool Management School (ULMS)
Michael Haenlein
ESCP Business School
Lucas Franieck
Nova SBE

Cite as:
Narayanan Pranadharthiharan, Malik Sumit, Haenlein Michael, Franieck Lucas (2025), A cautionary tale about $p < 0.05$. *Proceedings of the European Marketing Academy*, 54th, (125831)

Paper from the 54th Annual EMAC Conference, Madrid, Spain, May 25-30, 2025

A cautionary tale about p < 0.05

Abstract: We demonstrate that p-values are highly sensitive to the selective exclusion of small samples of data, even when outliers are absent. To highlight this issue, we introduce a simple, context-free robustness metric called the "Significance Fadeaway Score" (SFS). This measure quantifies the extent to which statistical significance (i.e., $p < 0.05$) persists as the most extreme data points are removed iteratively (one-at-a time and without replacement). In a pre-registered study, we evaluate SFS of 52 experimental studies published in leading marketing journals over the past five years. Our results reveal that nearly half of these studies lose statistical significance when less than 5% of the sample is selectively removed. These findings challenge the widespread reliance on dichotomous notions of statistical significance and promote the use of complementary metrics like SFS to enhance the transparency and rigor of experimental method in consumer behavior.

Keywords: *robustness | p-values | statistical significance*

Track*: Consumer Behavior*

1. Introduction

The perceived success of randomized experiments using null-hypothesis significance testing (NHST) has historically relied on p-values, which determine "statistical significance" when falling below a predefined threshold (α, typically 0.05). However, this nearly hundred-year-old dichotomous framework for establishing significance has faced growing criticism in recent years (see (McShane et al., 2024) for a comprehensive review). Concerns include widespread misinterpretation of p-values (Greenland et al., 2016), conflation of statistical significance with practical relevance (Cohen, 1992) , the inherent variability of p-values within the same study (McShane et al., 2024), and the susceptibility of p-values to manipulation through “researcher degrees of freedom” or “p-hacking” to achieve statistical significance (Simmons et al., 2011).

Our paper contributes to this ongoing critique by demonstrating a previously under-appreciated vulnerability of p-values: their pronounced sensitivity to the selective removal of data points. This property of p-values has important implications, particularly for researchers in consumer behavior. If the “statistical significance” of experimental results disappears with the removal of only a small fraction of the data, it becomes difficult to argue that the findings are robust, even in the absence of deliberate p-hacking or errors in analyzing data (Broderick et al., 2020). More critically, if this phenomenon is consistently observed across a wide spectrum of published studies, it raises broader questions about the validity of the binary framework of “statistical significance.”

Policymakers and practitioners often base decisions on insights derived from experiments conducted by marketing scholars—decisions that affect populations far beyond the academic community of authors, reviewers, and editors involved in publishing the research. Therefore, it is essential for researchers to rigorously evaluate the sensitivity of p-values both within their own studies and in relation to the broader literature. By doing so, they can better assess the robustness of their findings and ensure that their contributions are credible for policy and practice. At its core, this issue is not necessarily one of p-hacking or questionable research practices (QRPs) (John et al., 2012). It stems from the inferential nature of statistics itself: conclusions drawn from sample data are extrapolated to larger populations using metrics such as p-values. It is therefore essential to ask: how sensitive are these inferences to small exclusions in the sample data? More precisely, *what is the smallest proportion of the sample that can be removed to invalidate the conventional notion of statistical significance ($p<0.05$)?* We address this question by introducing and empirically validating the "Significance Fadeaway

Score" (SFS) — a simple, context-free diagnostic tool for researchers and reviewers to assess the robustness of the reported p-values.

SFS quantifies approximately the smallest proportion of data that must be removed for statistical significance ($p < 0.05$) to no longer hold.[1] SFS is calculated after conducting the study, meaning it does not change the experimental design or protocol from a researcher´s perspective. Data points identified as the *most extreme* are iteratively removed one at a time, without replacement, and the p-value of interest is recalculated after each removal (e.g., using t-tests, ANOVA or OLS regression). In a simple between-subjects design with two groups (say, treatment and control), the most extreme data points would refer to the highest outcome in treatment group or lowest outcome in control group if the treatment effect is positive. For a simple illustration of these extreme points, see Figure 1 below. The data removal process continues until the p-value exceeds a pre-defined threshold (e.g., 0.05) *or* if the estimate of interest (e.g., regression coefficient) falls outside the 95% confidence interval of the original estimate. The latter condition is imposed to ensure that the estimate is not biased. The resulting SFS value is straightforward to interpret: an SFS of x indicates that removing x% of the most extreme data points would cause statistical significance to "fade away," resulting in $p > 0.05$. We illustrate the SFS calculation process through simulation example in Study 1. It is worth noting that a low SFS value is possible even in the absence of obvious outliers or small initial p-values.

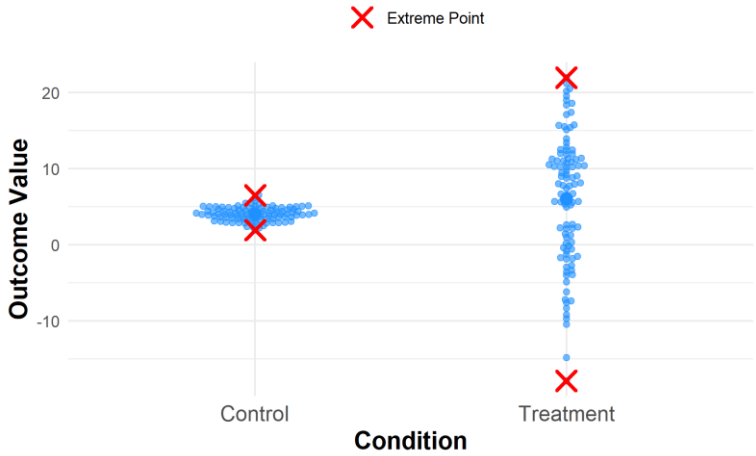

Figure 1. Extreme Points Highlighted for SFS Calculation

[1] It is important to note, however, that identifying the exact smallest subset of data to invalidate significance ($p < 0.05$) is computationally infeasible and classified as a class NP problem in Computer Science. This is because solving the problem would require testing all possible combinations of data points: starting with the removal of one data point (or row) at a time, followed by all possible combinations of two rows, then three, and so on. The number of combinations grows exponentially, making this process computationally prohibitive for larger datasets.

Existing metrics for assessing robustness in the context of our objective (i.e., make $p > 0.05$) are few and much less intuitive than SFS. For example, Broderick et al. (2020) propose a mathematically sophisticated method for identifying the most influential set of data points within a sample. However, once these data points have been identified (e.g., say data in rows 2, 7, and 12), it becomes nearly impossible for a researcher to understand why removing those specific rows would have the most impact on the p-value. Our critique is not a reflection of the mathematical validity of their approach but rather underscores a critical gap in the literature. For researchers in consumer behavior, the need extends beyond computational efficiency; the selective data removal process and the resulting robustness metric must also be intuitively understandable and easy to apply. We believe that SFS achieves this balance between mathematical rigor and interpretability, which could be crucial in making it accessible in applied research contexts.

Importantly, we evaluate the SFS values from 52 published studies in leading marketing journals such as the Journal of Consumer Psychology (JCP) and Journal of Consumer Research (JCR) within the past five years in a study pre-registered on AsPredicted (https://aspredicted.org/3j8w-yqdt.pdf). We find that nearly the published studies have an SFS value of less than 5%. However, we also found that nearly a quarter of studies in the sample have robust p-values., We are in the process of developing an R package [sfs] to enable researchers to calculate the SFS for their studies and benchmark their results against the dataset of published studies we have analyzed.[2] We want to reiterate that SFS is not meant to be diagnostic of Questionable Research Practices (QRPs). Low SFS values can arise for legitimate reasons such as a low signal-to-noise ratio (i.e., high relative dispersion). However, in the spirit of promoting greater scientific transparency and rigor, we propose that the SFS value be reported alongside p-values and effect sizes in future experimental studies.

[2] We understand that the sample of 52 studies is quite small to be a reasonable benchmark. We are in the process of expanding the list of studies and additionally identify an internal benchmark for each study to assess the relative magnitude of SFS.

2. Study 1

Consider a simulated experimental study with $N = 200$ participants, who are randomly assigned to one of two conditions: a control group (C) and a treatment group (T). Assume a positive treatment effect with the experimental outcome (O) being normally distributed with the following population parameters: $O_C \sim N(4,3)$; $O_T \sim N(4.1,10)$. Importantly, most of the assumptions made here—such as the presence of two conditions, a positive treatment effect, and the outcome being continuous—do not restrict the scope of this exposition. The average outcome and its variability are displayed in Figure 2, Panels A and B, respectively (For similar examples, see Zhang et al., 2023)).

An independent two-sample t-test reveals a p-value of 0.03, indicating a "statistically significant" difference in the population mean between the treatment and control conditions.[3] However, as evident from Figure 2, Panel B, the outcome variability—and potential outliers—within the treatment group raises concerns about the robustness of this result. The critical question is: *How big of a concern is it?* Addressing this question more rigorously than through mere visual inspection serves as a main motivation for developing the Significance Fadeaway Score (SFS) metric.

Note: Error bars signify ±1 standard error

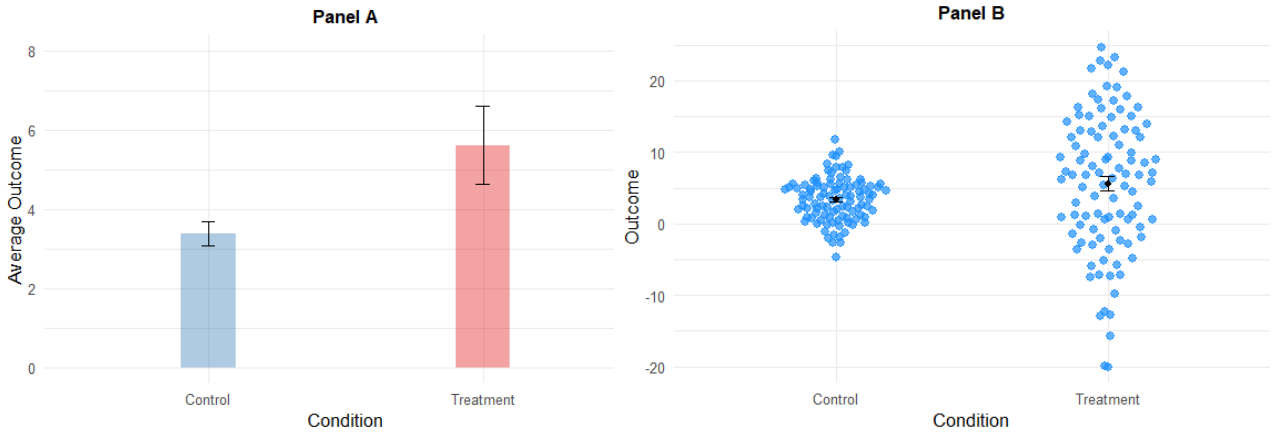

Figure 2. Simulation study results

2.1 Significance Fadeaway Score (SFS) Computation Algorithm

Initially, the direction of the treatment effect is inferred to be either positive $(+)$ or negative $(-)$ by comparing the sample means of the treatment (x_t) and control (x_c) groups. This is shown in Eq. (1) below.

$$Direction = \begin{cases} +, & (x_t - x_c) > 0 \\ -, & (x_t - x_c) < 0 \end{cases} \quad (1)$$

[3] In the simulation example, this result of p = 0.03 is a false-positive.

In our simulation example, the treatment effect is positive. The computation of the SFS metric begins with iterative removal of extreme data points and recalculating p-values.

1. **First Iteration:**
    - The treated subject with the highest outcome (i.e., the most extreme treatment point) is removed from the treatment group, and a new p-value ($p1$) is computed using a *t*-test comparing the (modified) treatment group to the (original) control group.
    - Simultaneously, the control subject with the lowest outcome (i.e., the most extreme control point) is removed from the control group, and a new p-value ($p2$) is computed from a *t*-test comparing the (modified) control group to the (original) treatment group.
    - The algorithm saves the higher of the two p-values ($\max(p1,p2)$).

2. **Subsequent Iterations:**
    - Based on the comparison of $p1$ and $p2$:
        - If $p1 \geq p2$, the treated subject with the highest outcome is removed from the sample to update the (original) treatment group.
        - If $p1 < p2$, the control subject with the lowest outcome is removed from the sample to update the (original) control group.
    - The same procedure is repeated from the first iteration: removing extreme points, recomputing p-values, and saving the higher of the two-resulting p-values.

3. **Ensuring Unbiased Inference:**
    - After each removal and subsequent t-test, the estimated treatment effect (e.g., mean difference) must remain within the original 95% confidence interval (CI) of the estimate. This ensures that the inference process remains unbiased.
    - If at any iteration, the estimate falls outside the original 95% CI (i.e., for both extreme data point removals), the SFS computation stops immediately.

4. **Stopping Conditions:**
    - The SFS computation halts if either:
        - The p-value from any iteration exceeds a predefined threshold (e.g., 0.05, 0.1, or 0.5), or
        - The estimate of the treatment effect lies outside the original 95% CI.

With a stopping threshold of p = 0.5 specified, the p-values obtained from 10% sample exclusions in the simulation study are shown below in Figure 3, in what we call the *SFS curve*. The

results show that even if 2% of most extreme subjects are excluded from the simulated experimental study, statistical significance ($p = 0.05$) no longer persists.

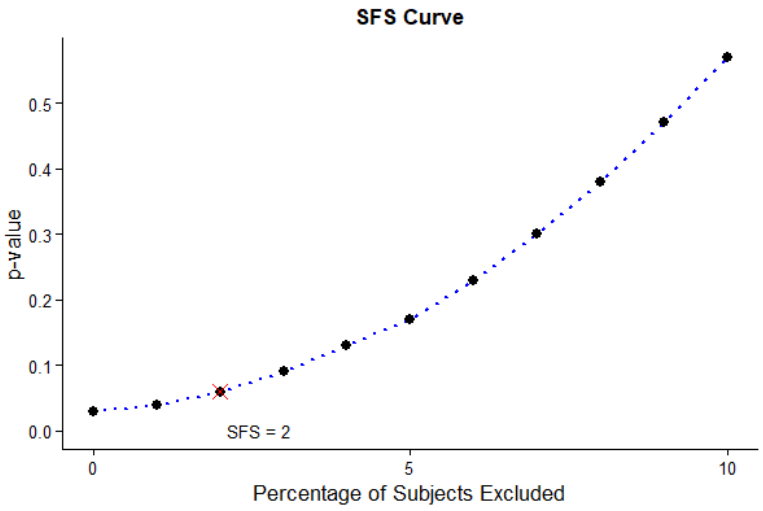

Figure 3. SFS Curve

While the simulation study illustrates the SFS calculation process with a concrete example, the more pressing question is: *How robust are the p-values reported in published experimental studies in marketing?* We empirically explore this issue in Study 2.

3. Study 2

3.1 Methods

We calculated the SFS values of 52 research studies published in the last five years (i.e., from 2019 to March 2024). Study 2 was pre-registered on AsPredicted (link). To identify the research studies, we searched for published articles that included the keywords "OSF" and "experiments" in leading marketing journals i.e., Journal of Consumer Research (28 studies), Journal of Consumer Psychology (12 studies), Journal of Marketing (6 studies), and Journal of Marketing Research (6 studies)[4]. The datasets for these studies were collected from the Open Source Framework (OSF) platform. These studies employed a between-subjects design, mostly having two conditions (80.77%), were conducted online (73.08%), and with sample size ranging from 104 to 1703 (average = 413). 25 of these 52 studies were pre-registered. Our analysis was based on the sample size reported in the published manuscripts and we did not exclude any additional data points. Prior to calculating the SFS metric, we confirmed the means of each study to ensure consistency with reporting in the respective journals. Our

[4] The selected studies are recent and have published the data on OSF. Therefore, we believe that estimates of SFS from this sample would be conservative, meaning the SFS of reported studies in broader literature is likely to be lower (i.e., less robust p-values).

analysis primarily aimed to quantify the sensitivity of statistical significance (as reported in the studies using *p-value* < .05) to selective removal of extreme data points. Accordingly, our variable of interest was the *p*-value of the main outcome variable, which was either continuous or categorical.

The analysis of SFS for each published study involved several steps. First, for each published study, we identified the most extreme data point i.e., the highest or lowest outcome value in either treatment (coded: 1) or control/ alternate condition (coded: 0). Second, the identified data point was removed, one-at-a-time and without replacement. Following the removal, we recalculated the *p*-value and used a *t*-test to compare the modified treatment and control groups. Third, we continued removing the extreme data points until one of the two conditions was met: (a) the *p*-value exceeded 0.05, or (b) the estimate of mean difference between the two groups remained outside the 95% confidence interval (CI) of the original estimate. Fourth, we recorded the percentage of sample required to surpass the threshold *p*-value of 0.05. For example, if removing 2% of the experiment's sample caused the *p*-value to exceed 0.05, the observed SFS was recorded as 2%.

3.2 Results and Discussion

Figure 4 presents the SFS curves for the 25 out of 52 randomly selected studies (due to space constraints). Our analysis reveals that nearly half of the 52 studies lose statistical significance with the selective removal of less than 5% of the sample. However, approximately a quarter of the studies demonstrate high robustness, with their SFS curves appearing as flat lines with a slope of zero. These studies remain unaffected by small sample exclusions, underscoring their resilience to such perturbations.

Overall, these results validate our proposition that p-values are highly sensitive to small sample exclusions, challenging the reliance of the binary framework traditionally used to determine statistical significance. These findings reiterate the need to move-away from the dichotomous reporting of p-values ((McShane et al., 2019, 2024)) and instead include complimentary robustness metrics such as SFS along with confidence intervals and effect sizes.

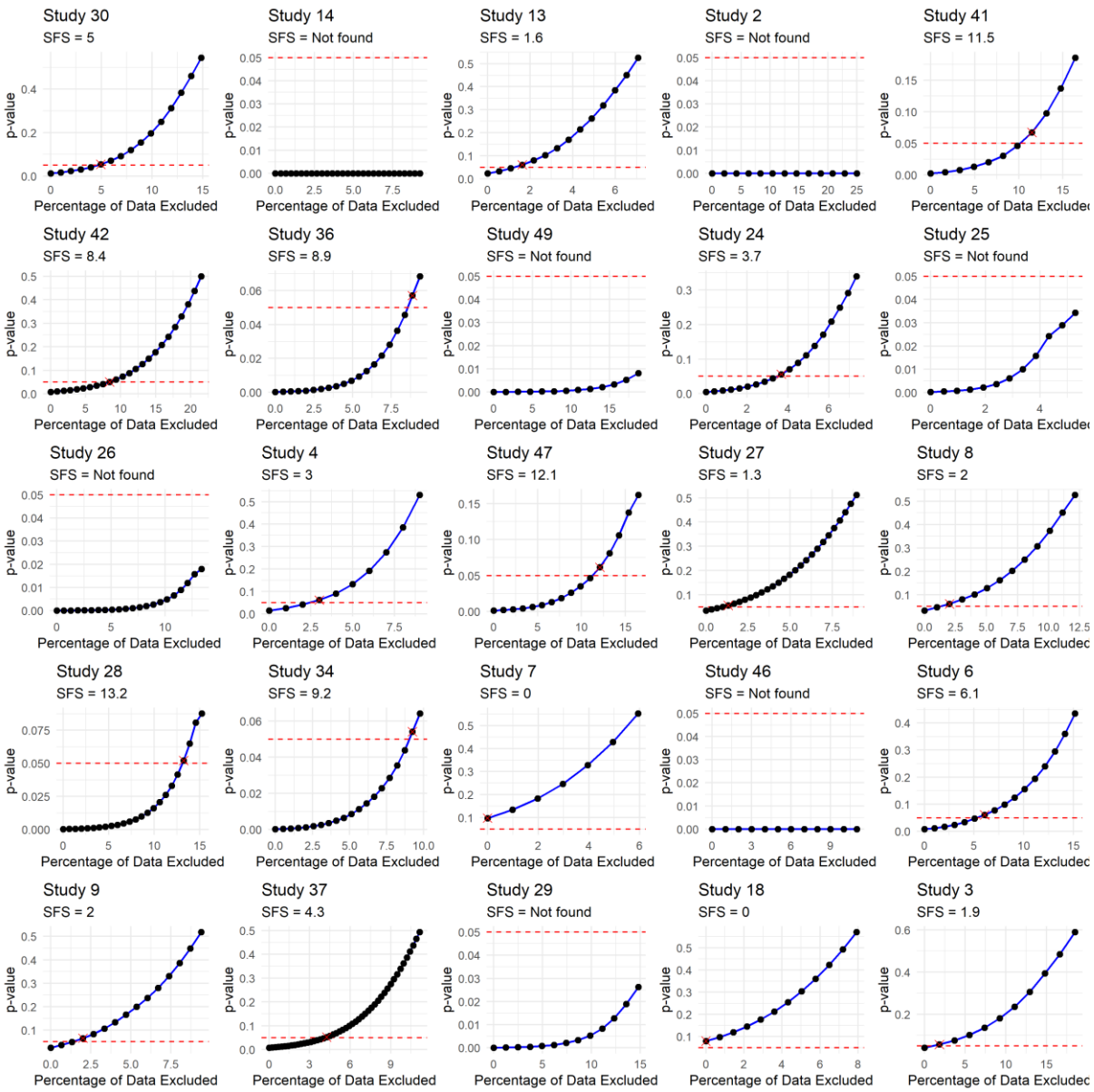

Figure 4. SFS curves for randomly selected 25 published studies in marketing

Note. SFS = Not Found refers to cases where the SFS computation stopped without p-value crossing the threshold of 0.05.

4. References

Broderick, T., Giordano, R., & Meager, R. (2020). An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? *https://arxiv.org/abs/2011.14999v5*

Cohen, J. (2004). Things I have learned (so far). *Methodological Issues & Strategies in Clinical Research.*, 315–333.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532.

McShane, B. B., Bradlow, E. T., Lynch, J. G., & Meyer, R. J. (2024). "Statistical Significance" and Statistical Reporting: Moving Beyond Binary. *Journal of Marketing*, *88*(3), 1–19.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, *73*(sup1), 235–245.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366.

Zhang, S., Heck, P. R., Meyer, M. N., Chabris, C. F., Goldstein, D. G., & Hofman, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(33), e2302491120.