Critical Values Robust to P-hacking

Adam McCloskey, Pascal Michaillat

June 2022

P-hacking occurs when researchers engage in various behaviors that increase their chances of reporting statistically significant results. P-hacking is problematic because it reduces the informativeness of hypothesis tests—by making significant results much more common than they are supposed to be in the absence of true significance. Despite its prevalence, p-hacking is not taken into account in hypothesis testing theory: the critical values used to determine significance assume no p-hacking. To address this problem, we build a model of p-hacking and use it to construct critical values such that, if these values are used to determine significance, and if researchers adjust their behavior to these new significance standards, then significant results occur with the desired frequency. Because such robust critical values allow for p-hacking, they are larger than classical critical values. As an illustration, we calibrate the model with evidence from the social and medical sciences. We find that the robust critical value for any test is the classical critical value for the same test with one fifth of the significance level—a form of Bonferroni correction. For instance, for a z-test with a significance level of 5%, the robust critical value is 2.31 instead of 1.65 if the test is one-sided and 2.57 instead of 1.96 if the test is two-sided.

McCloskey: University of Colorado–Boulder. Michaillat: Brown University. We thank Isaiah Andrews, Brian Cadena, Kenneth Chay, Andrew Chen, Garret Christensen, Pedro Dal Bo, Stefano DellaVigna, Peter Hull, Larry Katz, Miles Kimball, Megan Lang, Jonathan Libgober, Edward Miguel, Carlos Martins-Filho, Andriy Norets, Emily Oster, Bobak Pakzad-Hurson, Wenfeng Qiu, Jonathan Roth, Jesse Shapiro, and Yanos Zylberberg for helpful discussions and comments. This work was supported by the Institute for Advanced Study. This paper previously circulated under the title "Incentive-Compatible Critical Values."

1. Introduction

Definition of p-hacking. P-hacking occurs when researchers engage in various behaviors that increase their chances of reporting statistically significant results (Simonsohn, Nelson, and Simmons 2014; Lindsay 2015; Wasserstein and Lazar 2016; Christensen, Freese, and Miguel 2019). Typical p-hacking practices include suppressing inconvenient experiments, halting data collection at a convenient time, dropping inconvenient observations or treatments or outcomes, applying convenient transformations to the data, choosing convenient covariates in regressions, or choosing convenient statistical specifications.

Prevalence of p-hacking. P-hacking is prevalent across scientific fields. Researchers readily admit to engaging in it (John, Loewenstein, and Prelec 2012). It is visible in the lifecycle of scientific studies: significant results are almost certain to be reported, whereas insignificant results are likely to remain unreported (Dwan et al. 2008; Franco, Malhotra, and Simonovits 2014). And its effects appear in meta-analyses: the distributions of test statistics in entire literatures show that researchers tinker with their analysis to obtain significant results (Hutton and Williamson 2000; Head et al. 2015; Brodeur et al. 2016; Vivalt 2019; Brodeur, Cook, and Heyes 2020; Elliott, Kudrin, and Wuthrich 2021).

Problems caused by p-hacking. Despite its prevalence, p-hacking is not taken into account in hypothesis testing theory: the critical values used to determine significance assume no p-hacking. Therefore, the critical values set a standard for significance that is too lax: significance is reached much more often than purported by the test's nominal significance level in the absence of true significance. This is problematic because hypothesis tests are informative only insofar as a true null hypothesis is not rejected more often than the significance level. For instance, hypothesis tests are used to evaluate scientific theories and paradigms (Kuhn 1957; Akerlof and Michaillat 2018). They allow scientists to identify instances when theory does not accord well with empirical observations. Unbridled p-hacking threatens scientific progress. It leads to excessive rejection of established paradigms and to the unwarranted adoption of new paradigms. As such, it threatens the credibility of science. One manifestation of uncontrolled p-hacking is the replication crisis affecting many scientific fields (Prinz, Schlange, and Asadullah 2011; Begley and Ellis 2012; Ioannidis et al. 2014; Open Science Collaboration 2015; Camerer

et al. 2016; Christensen and Miguel 2018; Benjamin et al. 2018).

Existing corrections for p-hacking. A few corrections for p-hacking have been discussed (Anscombe 1954; Lovell 1983; Glaeser 2008). But these corrections take the researcher's p-hacking behavior as fixed, whereas in reality the researcher would change her p-hacking behavior as soon as the correction is implemented. Consider for instance an hypothesis tests with a 5% significance level and corresponding critical value. Classical critical values are constructed such that if the researcher examined one data sample, a true null hypothesis would be rejected no more than 5% of the time. But if a researcher collected more than one data sample, performed hypothesis tests on all the samples, and reported the best result, a true null hypothesis would be rejected more often than 5% of the time. Existing corrections take the number of samples collected as given and compute a more stringent critical value based on this number. This first step is not enough to resolve the problem. Just as researchers may collect more than one sample under the classical critical value, they may collect more samples than anticipated under the new critical value, overwhelming the proposed correction.[1]

This paper's correction for p-hacking. To address this problem, we construct a model of p-hacking and use it to construct critical values of test statistics such that, if these values are used to determine significance, and if researchers optimally p-hack in response to these new significance standards, then significant results will occur with the desired frequency.

Reasons for p-hacking. In our model just as in real life, scientists p-hack because they face strong incentives to do so (Glaeser 2008; Nosek, Spies, and Motyl 2012; Bakker, van Dijk, and Wicherts 2012). First, significant results are more rewarded than insignificant ones. This is because scientific journals prefer publishing significant results (Sterling 1959; Bozarth and Roberts 1972; Begg and Berlin 1988; Csada, James, and Espie 1996; Jennions and Moeller 2002; Song et al. 2000; Ioannidis and Trikalinos 2007; Head et al. 2015; Fanelli, Costas, and Ioannidis 2017; Christensen, Freese, and Miguel 2019; Andrews and Kasy 2019). Publications, in turn, determine a scientist's career path, including

[1] Multiple-testing corrections are not useful in the context of p-hacking because readers cannot observe the number of tests conducted by the researcher. Multiple-testing corrections can only be used when the number of tests is fixed and known. Anytime-valid p-values and confidence intervals cannot be used either because they require that the researcher's data collection process is known to the reader (Jennison and Turnbull 1984; Johari et al. 2021; Howard et al. 2021).

promotions, salary, and honorific rewards (Hagstrom 1965; Skeels and Fairbanks 1968; Merton 1973; Katz 1973; Siegfried and White 1973; Tuckman and Leahey 1975; Hansen, Weisbrod, and Strauss 1978; Sauer 1988; Swidler and Goldreyer 1998; Gibson, Anderson, and Tressler 2014; Biagioli and Lippman 2020). Second, researchers enjoy a lot of flexibility in data collection and analysis. Hence, even when the null hypothesis is true, they have ample opportunity to obtain significant results without violating scientific norms (Cole 1957; Armitage 1967; Leamer 1983; Lovell 1983; Simmons, Nelson, and Simonsohn 2011; Humphreys, de la Sierra, and van der Windt 2013; Huntington-Klein et al. 2021).

P-hacking process. We start our analysis by considering a researcher who sequentially collects independent and identically distributed (iid) data samples, performs a hypothesis test on each sample, and reports the best result. Such data sampling is a useful starting point for several reasons. First, unlike other forms of p-hacking that could be detected through specification curves (Simonsohn, Simmons, and Nelson 2020; Young and Holsteen 2017) or multiverse analysis (Steegen et al. 2016), data sampling is undetectable. It is therefore particularly important to correct this form of p-hacking. Second, the test statistics obtained at each p-hacking step are iid, which makes it simple to compute critical values. Third, the data-sampling process can be calibrated from evidence on the lifecycle of studies in the social and medical sciences, which allows us to quantify the correction required by p-hacking. Fourth, the critical values obtained under data sampling also control the probabilities of type 1 error under many other common forms of p-hacking that induce positive dependence across test statistics: pooling data, removing outliers, searching across regression specifications, or searching across instruments.

P-hacking strategy. We first find researchers' p-hacking strategy using results from optimal stopping theory (Ferguson 2007). The researcher's optimal strategy is to continue collecting new data samples until finding a significant result. Not all studies report significant results, however, because the resources (time, funding, stamina) that a researcher can devote to any project are finite (Chen 2021). If the researcher runs out of resources before obtaining a significant result, she reports an insignificant result.

Probability of type 1 error. We then determine the (random) number of data samples collected by a researcher when the null hypothesis is true and the corresponding probability of type 1 error, as a function of the prevailing critical value. The critical value influences the rate of type 1 error in two ways. First, it determines the probability

that a true null hypothesis is rejected in each data sample. Second, it influences the number of data samples that the researcher collects.

Computation of robust critical value. From these results we compute the critical value such that, if it is used to determine significance and researchers adjust their p-hacking behavior to the new significance standard, then type 1 errors occur at the desired rate—given by the nominal significance level. This critical value is robust to p-hacking, and it is given by a type of Bonferroni correction. For any test and any significance level, the robust critical value is the classical critical value for the same test with the significance level divided by the expected number of p-hacking steps when the robust critical value is in place. Accordingly, the robust critical value is larger than the classical critical value for the same test and significance level. An advantage of the model is that the expected number of p-hacking steps when the robust critical value is in place, and thus the robust critical value itself, are solely determined by two parameters: the significance level and the probability that a p-hacking step is completed before running out of resources.

Numerical illustration. As an illustration, we compute the completion probability using evidence from the lifecycle of studies in the social and medical sciences (Dwan et al. 2008; Franco, Malhotra, and Simonovits 2014). An accurate rule of thumb is that the robust critical value for any test is the classical critical value for the same test with one fifth of the significance level. For instance, for a z-test with a significance level of 5%, the robust critical value is 2.31 instead of 1.65 if the test is one-sided and 2.57 instead of 1.96 if the test is two-sided.

Other forms of p-hacking. In the baseline model, researchers construct a sequence of test statistics by collecting a sequence of independent datasets; therefore, the test statistics are independent. In reality, researchers often construct test statistics that are positively dependent—when they pool data, when they remove outliers, when they examine various regression specifications, or when they examine various instruments. The robust critical values obtained under the independence assumption remain useful in these scenarios because they maintain a type 1 error rate below the significance level.

2. Prevalence of p-hacking, and reasons for it

This section shows that p-hacking is prevalent across scientific fields. It also discusses the reasons behind p-hacking. The first is that p-hacking is rewarded because statistically significant results have greater payoffs than insignificant ones. The second is that p-hacking is not very costly because researchers have a lot of flexibility in their empirical work.

2.1. Prevalence of p-hacking

P-hacking is prevalent in many sciences.

Survey of scientists. A survey of 5964 psychologists at major US universities shows that p-hacking is common: 63% of respondents admit to failing to report all outcomes, 56% admit to deciding whether to collect more data after examining whether the results were significant, 46% admit to selectively reporting studies that "worked", 38% admit to deciding whether to exclude data after looking at the impact of doing so on the results, 28% admit to failing to report all treatments in a study, and 16% admit to stopping data collection earlier than planned after obtaining the desired results (John, Loewenstein, and Prelec 2012, table 1).

Lifecycle of studies. P-hacking is also directly visible in the lifecycle of studies. Franco, Malhotra, and Simonovits (2014, table 3) track 221 experimental studies in the social sciences, from experimental design to publication. They find that 64.6% of the studies reporting insignificant results were never written up, whereas only 4.4% of the studies reporting strongly significant results were not written up. Clearly, scientists report results selectively: significant results are almost certain to be reported, whereas insignificant results are likely to remain unreported. In a large-scale analysis of the lifecycle of clinical trials, Dwan et al. (2008) also find that significant outcomes are more likely to be reported than insignificant outcomes.

Meta-analyses of published studies. Finally the effects of p-hacking appear in meta-analyses of published studies. The distributions of test statistics or p-values across studies in a literature show that researchers tinker with their econometric specifications in order to obtain significant results (Hutton and Williamson 2000; Head et al. 2015;

Brodeur et al. 2016; Vivalt 2019; Brodeur, Cook, and Heyes 2020; Elliott, Kudrin, and Wuthrich 2021).

2.2. Rewards from significant results

Researchers hunt for significant results because such results are more rewarded than insignificant results. The reason is twofold. First, a study presenting significant results is more likely to be published than one presenting insignificant results. Second, a published study yields higher rewards than an unpublished study.

Publication bias. Indeed, scientific journals prefer publishing significant results. Such publication bias was first identified in psychology journals (Sterling 1959; Bozarth and Roberts 1972). It has since been observed across the social sciences (Christensen, Freese, and Miguel 2019, chapter 3), medical sciences (Begg and Berlin 1988; Song et al. 2000; Ioannidis and Trikalinos 2007; Dwan et al. 2008), biological sciences (Csada, James, and Espie 1996; Jennions and Moeller 2002), and many other disciplines (Fanelli, Costas, and Ioannidis 2017). Andrews and Kasy (2019, p. 2767) assess the magnitude of the bias in two literatures: experimental economics and psychology. They find that results significant at the 5% level are 30 times more likely to be published than insignificant results.

Rewards from publication. Publications, in turn, determine a scientist's career path, including promotion (Skeels and Fairbanks 1968) and salary (Katz 1973; Siegfried and White 1973; Tuckman and Leahey 1975; Hansen, Weisbrod, and Strauss 1978; Sauer 1988; Swidler and Goldreyer 1998; Gibson, Anderson, and Tressler 2014). In some countries, scientists are also rewarded with cash bonuses as high as $30,000 for publication in top journals (Biagioli and Lippman 2020, p. 6). Publications yield not only material rewards but also honorific rewards (Hagstrom 1965). One such reward is eponymy, "the practice of affixing the name of the scientist to all or part of what he has found" (Merton 1957). Beyond eponymy are prizes, medals, memberships in academies of sciences, and fellowships in learned societies (Merton 1957).

Rewards from significant results. Accordingly, researchers have an incentive to obtain significant results by p-hacking. Formally, let V be the random variable giving the rewards from a completed study.[2] The expected rewards from a study with significant

[2]There are several sources of randomness. The study may not be published at all. Or it may be published in one of many possible journals, from the most prestigious to the most obscure. Even when it

results are

$$v^s = \mathbb{E}(V \mid \text{significant}),$$

and those from a study with insignificant results are

$$v^i = \mathbb{E}(V \mid \text{insignificant}).$$

Using the law of iterated expectations, we find

$$\begin{aligned} v^s = {} & \mathbb{E}(V \mid \text{published \& significant}) \times \mathbb{P}(\text{published} \mid \text{significant}) \\ & + \mathbb{E}(V \mid \text{unpublished \& significant}) \times \mathbb{P}(\text{unpublished} \mid \text{significant}). \end{aligned}$$

We note that $\mathbb{P}(\text{unpublished} \mid \text{significant}) + \mathbb{P}(\text{published} \mid \text{significant}) = 1$, and we assume that conditional on the publication status, the rewards are independent from statistical significance. Then we obtain

$$\begin{aligned} v^s = {} & \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished})\right] \times \mathbb{P}(\text{published} \mid \text{significant}) \\ & + \mathbb{E}(V \mid \text{unpublished}). \end{aligned}$$

Following the same logic, we find

$$\begin{aligned} v^i = {} & \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished})\right] \times \mathbb{P}(\text{published} \mid \text{insignificant}) \\ & + \mathbb{E}(V \mid \text{unpublished}). \end{aligned}$$

Accordingly, the expected gain from obtaining a significant result is

$$\begin{aligned} v^s - v^i = {} & \left[\mathbb{P}(\text{published} \mid \text{significant}) - \mathbb{P}(\text{published} \mid \text{insignificant})\right] \\ & \times \left[\mathbb{E}(V \mid \text{published}) - \mathbb{E}(V \mid \text{unpublished})\right]. \end{aligned} \tag{1}$$

Empirically, significant results are more likely to be published than insignificant ones:

$$\mathbb{P}(\text{published} \mid \text{significant}) > \mathbb{P}(\text{published} \mid \text{insignificant}).$$

Moreover, a published study yields higher rewards than an unpublished one:

$$\mathbb{E}(V \mid \text{published}) > \mathbb{E}(V \mid \text{unpublished}).$$

is published in a journal of a given standing, the study's impact may vary.

These facts together with (1) imply that it is beneficial to obtain a significant result:

$$v^s > v^i.$$

2.3. Opportunities for p-hacking

Researchers have a lot of flexibility in data collection and analysis (Huntington-Klein et al. 2021). This flexibility affords them opportunities to obtain significant results, even when the null hypothesis is true. Indeed, researchers have found that it is easy to obtain significant results when the null hypothesis is true, without violating prevailing scientific norms in biology (Cole 1957), medical science (Armitage 1967, section 4), economics (Leamer 1983; Lovell 1983), psychology (Simmons, Nelson, and Simonsohn 2011), and political science (Humphreys, de la Sierra, and van der Windt 2013).

3. Model of p-hacking

We develop a simple model of p-hacking and incorporate it into statistical hypothesis testing theory. The researcher samples data, with the aim of reaching a significant result. Sampling data and conducting hypothesis tests takes time, stamina, and money, which are all in finite supply. Because researchers must report results before running out of time, stamina, and money, not all researchers can obtain significant results.

3.1. Hypothesis test

The researcher tests a null hypothesis H_0 against an alternative hypothesis H_1. The data are governed by a different probability distribution under each hypothesis. The researcher sets the test's significance level to $\alpha \in (0, 1)$. The significance level gives the desired probability of type 1 error—the error that occurs when a true null hypothesis is rejected. Common significance levels are 10%, 5%, and 1%.

3.2. Test statistic

To conduct the hypothesis test, the researcher collects a dataset. From this dataset she constructs a test statistic T, whose realization is t. Under H_0, the cumulative distribution function of the test statistic under is F, its survival function is $S = 1 - F$, and its inverse

survival function is $Z = S^{-1}$.[3]

3.3. Classical critical value

The null hypothesis is rejected when the test statistic falls into the critical region. We assume that the critical region takes the standard form (z, ∞), where z is the critical value. If the researcher obtains a test statistic $t > z$, the null hypothesis is rejected: the result is significant. But if she obtains a test statistic $t \leq z$, the null hypothesis cannot be rejected: the result is insignificant. Accordingly, the probability of type 1 error is $S(z)$. The classical critical value is set such that the probability of type 1 error in one single test equals the significance level:

$$S(z) = \alpha, \tag{2}$$

or equivalently $z = Z(\alpha)$.

3.4. Rewards from significant results

The first nonclassical element of the model are the rewards accruing to significant results. Based on the evidence in section 2, we assume that the expected rewards v^s from a study with significant results are higher than the expected rewards v^i from a study with insignificant results.

3.5. Opportunities for p-hacking

Researchers have ample opportunity to p-hack. However, their resources—time, money, manpower, stamina—are not infinite. Hence, they cannot systematically obtain significant results (Chen 2021). We assume that it takes a random amount of resources to sample data and conduct hypothesis tests, and the researcher must keep the cumulative resources used below a random limit L. Once the researcher has exhausted more resources than L, she must stop working on the project. The resource limit captures the many resource constraints faced by researchers: limited access to data, limited funding, limited coauthor time, limited time before publication of similar results by competing research teams, limited time and stamina to work on specific projects, or limited time

[3] For simplicity we focus on simple null hypotheses. For composite null hypotheses, we would use the distribution under the null hypothesis's configuration that is the easiest to reject. For example, when testing $H_0 : \mathbb{E}(X) \leq \mu_0$ versus $H_1 : \mathbb{E}(X) > \mu_0$, we would use the distribution of the test statistic at the point $\mathbb{E}(X) = \mu_0$.

before the opportunity to work on more promising projects arises. Following Ferguson (2007, p. 4.12), we assume that the resource limit has an exponential distribution with rate $\lambda > 0$, so $\mathbb{P}(L > l) = \exp(-\lambda l)$ for any $l > 0$.

3.6. P-hacking process

P-hacking steps. The steps of the p-hacking process are denoted by $n = 0, 1, 2, \ldots, \infty$, with step 0 corresponding to not starting the research project. It takes a random amount of resources to sample data and conduct hypothesis tests. The cumulative amounts of resources required to complete the different steps of the p-hacking process are D_1, D_2, ... given by a renewal process independent of the resource limit L. That is, the resources required at each step, D_1, $D_2 - D_1$, $D_3 - D_2$, ..., are iid according to a distribution independent of L.

First step. If resources are exhausted before the first step is completed, $L < D_1$, the researcher is not able to obtain any results. If the resources are not exhausted when the first step is completed, $L > D_1$, the researcher is able to collect a first sample of data and complete a first hypothesis test. The test statistic obtained in the first test is T_1, which is independent of the resource variables. The researcher then decides to submit this result to a scientific journal, or to engage in another step of p-hacking.

Second step. If the researcher moves to step 2, she starts by collecting a second sample. This second sample has the same size as the first, and it is drawn from the same underlying population. The researcher conducts the same hypothesis test on this second sample. Once again, if resources are exhausted before the second step is completed, $L < D_2$, the researcher must stop the project before obtaining the second test statistic. As the researcher stores all the data that she collects and keeps track of all the statistical analyses, she is able to recall past results. While the researcher does not obtain a second test statistic, she may still submit the statistic T_1 to a scientific journal. If resources are not exhausted, $L > D_2$, the researcher obtains a second test statistic, T_2. The statistics T_1 and T_2 are iid. She may submit the best result from the two hypothesis tests, $\max\{T_1, T_2\}$. If she does not want to submit the result, she may engage in another step of p-hacking.

Nth step. More generally, at step n, the researcher collects a nth sample, of the same size and drawn from the same underlying population as all the previous samples. She conducts the same test on that sample. If resources are exhausted before step n is

completed, the researcher can only report the best result obtained up to the previous stage, $\max\{T_1, \ldots, T_{n-1}\}$. If resources are not exhausted before step n is completed, the researcher obtains the nth statistic, T_n, which is iid with $T_1, T_2, \ldots, T_{n-1}$. She may then submit the best of the n test statistics, $\max\{T_1, \ldots, T_n\}$, or she may proceed to the next step of p-hacking.

Infinite p-hacking. Step ∞ corresponds to collecting infinitely many samples, conducting infinitely many tests, and never reporting any result.

3.7. Completion probability

Following Ferguson (2007, p. 4.13), we introduce the index of the first p-hacking step that cannot be completed before resources are exhausted: $K = \min\{n \geq 1 : D_n > L\}$. Let γ be the probability that the first step can be completed:

$$\gamma = \mathbb{P}(D_1 < L) = \mathbb{E}(\exp(-\lambda D_1)).$$

The index K is independent of the test statistics $T_1, T_2, \ldots$, and it has a geometric distribution with success probability $1 - \gamma$, so $\mathbb{P}(K > k) = \gamma^k$ for $k = 0, 1, 2, \ldots$.

3.8. Payoffs

No results. If the researcher does not start the research project, she receives a payoff $y_0 = 0$. If resources are exhausted before the end of the first step, the researcher does not obtain any result, so she receives the same payoff of $y_1 = 0$. If the researcher never concludes the research project and keeps on p-hacking forever, she also receives a payoff $y_\infty = 0$. In all other cases, she receives a positive payoff.

Exhausted resources. The researcher is not able to continue p-hacking once the project resources are exhausted. To capture this constraint, we set to zero all payoffs once resources are exhausted: $y_n = 0$ in any step $n > K$. With these payoffs, the researcher never continues past step K. At step K, the researcher cannot obtain a new test statistic, but she can submit for publication the best test statistic from the previous $K-1$ hypothesis tests, $\max\{T_1, \ldots, T_{K-1}\}$. If that statistic is significant, the payoff is $y_K = v^s$; if that statistic is not significant, the payoff is $y_K = v^i$.

Non-exhausted resources. Any step $n < K$ can be completed before running out of resources, so the researcher can submit the best statistic from the n previous tests, $\max\{T_1, \ldots, T_n\}$. If that statistic is significant, the payoff is $y_n = v^s$; if not, the payoff is $y_n = v^i$.

Possible extensions. Here research is costless to the researcher: her university or lab covers all the costs. But the results remain essentially the same if the researcher incurs a research cost—be it a monetary cost or a psychological cost (appendix A). Moreover, the researcher does not discount the future. A significant result yields the same payoff irrespective of when it is obtained. But the results are not modified if the researcher discounts future payoffs (appendix B).

4. Optimal stopping time

The researcher p-hacks as long as she wishes. At each step, after observing all previous test statistics, she may decide to stop and receive a payoff, or she may decide to continue to the next step. If she is able to complete the next p-hacking step, she will observe another test statistic. The researcher's problem is to choose a time to stop p-hacking to maximize expected payoffs. We now solve this problem.

4.1. Researcher's problem

The stopping rule chosen by the researcher, the critical value z, and the random research events determine the random time $N(z)$ at which the researcher stops p-hacking. The problem of the researcher is to choose a stopping time to maximize expected payoffs.

4.2. Reported statistic

As long as she is able to complete at least one hypothesis test, the researcher reports a random statistic $R(z)$ upon stopping. This is the best test statistic that she has been able to obtain through p-hacking. It may be significant or insignificant, and the researcher may be able to publish it or not.

4.3. Characteristics of the optimal stopping time

An optimal stopping time $N(z)$ exists because two conditions are satisfied (Ferguson 2007, chapter 3). Let Y_n denote the random payoff received by the researcher when she

stops at time n. First, $Y_n \leq v^s$ a.s., so $\sup_n Y_n < \infty$ a.s. Second, because the resources inevitably run out, $Y_n \xrightarrow{as} 0 = y_\infty$ as $n \to \infty$. Furthermore, the optimal stopping time is given by the principle of optimality of dynamic programming: it is optimal to stop as soon as the payoff is at least as high as the best payoff that can be expected by continuing.

4.4. Finding the optimal stopping time

We find the optimal stopping time by considering the various situations faced by the researcher.

Starting the research project. If the researcher does not start the research project, she receives $Y_0 = 0$. In contrast, if she starts she earns a non-negative payoff: 0 if resources are exhausted before the first step is completed; v^i if she obtains an insignificant result; or v^s if she obtains a significant result. Hence it is always optimal to start the research project.

Continuing after insignificant results. How does the researcher behave when she still has resources to allocate to the project? A first possibility is that the result at step n and all the results before that are insignificant. Since the best result found by the researcher is insignificant, the researcher earns $Y_n = v^i$ by stopping at step n. All possible payoffs are more than the payoff received for an insignificant result, v^i, so all expected payoffs are more than v^i. Since the researcher is expected to obtain more than v^i by continuing, it is not optimal to stop without obtaining a significant result.

Stopping after a significant result. If the result of test n is significant, the best result found by the researcher is significant, so the researcher earns $Y_n = v^s$ by stopping at step n. All possible payoffs are less than the payoff received for a significant result, v^s, so all expected payoffs are less than v^s. Hence, the researcher cannot do better by continuing. It is therefore optimal to stop at step n and report $R(z) = \max\{T_1, \ldots, T_n\} > z$. In fact, the principle of optimality indicates that it is optimal to stop the first time that a significant result occurs.

Stopping when resources are depleted. Once resources are depleted, the researcher is not able to continue p-hacking. Hence, the researcher stops at step K if she had not stopped before. There are two possibilities. If $K = 1$, resources are depleted before even the first step, so the researcher has nothing to report. If $K > 1$, the researcher submits

the best test statistic that she has collected. This best result is necessarily insignificant, otherwise she would have stopped before. So she reports $R(z) = \max\{T_1, \ldots, T_{K-1}\} \leq z$.

Summary. The optimality principle gives the following results:

LEMMA 1. *The researcher stops when she obtains a significant result or when she runs out of resources, whichever comes first. In the former case the researcher reports a significant result; in the latter case she reports an insignificant result. So there is p-hacking: the researcher never stops at insignificant results, unless she runs out of resources to support the project.*

5. Critical value robust to p-hacking

Based on the researcher's p-hacking strategy, we compute the robust critical value. This critical value ensures that the probability of type 1 error remains below the significance level even when the researcher's behavior adjusts to the critical value itself.

5.1. Distribution of optimal stopping time

We compute the distribution of the optimal stopping time. Since the distribution is used to calculate the critical value, we compute it under the null hypothesis.

Probability to reach significance at step n. Under the null hypothesis, the probability that the test statistic at step n reaches the critical value z is simply given by the test statistic's survival function: $\mathbb{P}(T_n > z) = S(z)$, where $\mathbb{P}$ denotes the probability measure under H_0.

Probability to continue p-hacking at step n. The researcher continues p-hacking after any step if she has not run out of resources during that step, which happens with probability γ, and the latest result is insignificant, which happens with probability $1 - S(z)$. The two events are independent, so the probability that the researcher continues p-hacking is $\gamma[1 - S(z)]$. Conversely, the probability that the researcher stops p-hacking at any step is

$$1 - \gamma[1 - S(z)]. \tag{3}$$

Distribution of the stopping time. At each step, the probability of stopping p-hacking is constant, given by (3). The optimal stopping time therefore has a geometric distribution

with success probability (3). The probability that the optimal stopping time is $n \geq 1$ is

$$\mathbb{P}(N(z) = n) = \left[\gamma - \gamma S(z)\right]^{n-1} \left[1 - \gamma + \gamma S(z)\right].$$

Expected number of p-hacking steps. Given that the optimal stopping time has a geometric distribution with success probability (3), we obtain the following result:

PROPOSITION 1. *Under the null hypothesis, the expected number of p-hacking steps is*

$$\mathbb{E}(N(z)) = \frac{1}{1 - \gamma[1 - S(z)]}, \tag{4}$$

where $\mathbb{E}$ denotes the expectation operator under H_0. P-hacking is prevalent ($\mathbb{E}(N(z)) > 1$). Moreover, researchers p-hack more when the standards for significance are more stringent (higher z).

Given that classical critical values are defined by (2), we infer the following result:

COROLLARY 1. *Under the null hypothesis and with classical critical values, the expected number of p-hacking steps is*

$$\mathbb{E}(N(z)) = \frac{1}{1 - (1 - \alpha)\gamma}. \tag{5}$$

P-hacking is more prevalent when the significance level is lower (lower α).

P-hacking under the alternative hypothesis. In (5), $1 - \alpha$ represents the probability to obtain an insignificant result at each step when the classical critical value is used to determine significance and the null hypothesis is true. When the alternative hypothesis is true instead, the probability to obtain an insignificant result at each step becomes β, where $1-\beta$ is the power of the hypothesis test. Hence, if the alternative hypothesis is true, the expected number of p-hacking steps is simply $1/(1 - \beta\gamma)$. In many fields, hypothesis tests are acceptable only if their power is above 80% (Duflo, Glennerster, and Kremer 2007; Christensen 2018). Setting power to $1 - \beta = 80\%$, we find that the expected number of p-hacking steps under the alternative is $1/(1-0.2 \times \gamma) < 1/(1-0.2) = 1.25$: there is almost no p-hacking. This is unsurprising. If the alternative hypothesis is true and the study is well powered, the null hypothesis is rejected most of the time, which makes p-hacking unnecessary. Hence, if we see a lot of p-hacking, either the alternative hypothesis is false, or the alternative hypothesis is true but tests have little power (Ioannidis 2005).

5.2. Probability of type 1 error

Next, we compute the probability of type 1 error as a function of the critical value.

PROPOSITION 2. *When the critical value is set to z, the probability of finding a type 1 error in a reported study is*

$$S^*(z) = \frac{S(z)}{1 - \gamma[1 - S(z)]}. \tag{6}$$

The probability of type 1 error is larger when researchers p-hack ($S^*(z) > S(z)$). In fact, the probability of type 1 error grows linearly with the expected number of p-hacking steps ($\mathbb{E}(N(z))$):

$$S^*(z) = S(z) \times \mathbb{E}(N(z)). \tag{7}$$

The proof is not difficult: it relies on appropriate applications of the law of total probability and Bayes' rule. But it is not particularly interesting so we relegate it to appendix C. Given that classical critical values are defined by (2), we infer the following result:

COROLLARY 2. *Under classical critical values, the probability of type 1 error is larger than the significance level α:*

$$S^*(z) = \frac{\alpha}{1 - (1 - \alpha)\gamma} > \alpha. \tag{8}$$

When researchers p-hack, the probability of type 1 error given by classical critical values is larger than the significance level α. Hence, the standards for significance set by classical critical values are too low: significance is reached more often than purported by the test's significance level. This is problematic, because hypothesis tests are only informative insofar as true null hypotheses are not rejected more often than the significance level.

5.3. Robust critical value

Effects of critical value on type 1 error rate. Changing the critical value z has two effects on the probability of type 1 error (equation (7)). First, there is a mechanical effect, whereby a higher critical value makes it less likely that a given test statistic exceeds it ($S(z)$ is decreasing in z). Second, there is a behavioral effect, whereby the optimal stopping time and thus reported test statistic are altered by the critical value. When

the critical value is larger, researchers p-hack more in hope of reaching significance ($\mathbb{E}(N(z))$ is increasing in z). The behavioral effect was not taken into account by previous corrections for p-hacking (Anscombe 1954; Lovell 1983; Glaeser 2008). The novelty of this analysis is to propose a critical value that accounts for it.

Computing the robust critical value. The robust critical value is such that the probability of type 1 error equals the significance level α when researchers p-hack. Since the probability of type 1 error with p-hacking is given by (6), the robust critical value z^* is implicitly defined by

$$\frac{S(z^*)}{1-\gamma+\gamma S(z^*)} = \alpha. \tag{9}$$

From this implicit definition we obtain the following results (details of the proof are relegated to appendix C):

PROPOSITION 3. *For any hypothesis test with significance level α, the robust critical value is given by*

$$z^* = Z\left(\alpha \cdot \frac{1-\gamma}{1-\alpha\gamma}\right), \tag{10}$$

where Z is the inverse survival function of the test statistic. The robust critical value is always larger than the classical critical value $Z(\alpha)$.

P-hacking under the robust critical value. The robust critical value corrects for the distortion introduced by p-hacking without eliminating p-hacking. In fact, because the significance standards imposed by the robust critical value are more stringent than classical standards, researchers p-hack more under the robust critical value. Combining (4) and (9), we obtain the following corollary:

COROLLARY 3. *The average number of p-hacking steps under the robust critical value is*

$$\mathbb{E}(N(z^*)) = \frac{1-\alpha\gamma}{1-\gamma}. \tag{11}$$

5.4. Bonferroni correction

Our correction for p-hacking can be formulated as a type of Bonferroni correction. The classical significance level corresponding to the robust critical value is the desired

significance level divided by the amount of p-hacking under the robust critical value:

COROLLARY 4. *Achieving a significance level α under p-hacking requires to set the critical value at the level that would be appropriate for a significance level*

$$\alpha^* = \frac{\alpha}{\mathbb{E}(N(z^*))} \tag{12}$$

under classical conditions.

This relation is obtained by evaluating (7) at z^*, and using $\alpha^* = S(z^*)$ and $S^*(z^*) = \alpha$. Unlike in a typical Bonferroni correction, however, the number of p-hacking steps used for the correction is not observed, and it is not the number of p-hacking steps prevailing under a standard critical value. Rather, it is the average number of p-hacking steps under the robust critical value when the null hypothesis is true. Thanks to the model, we can link this number of steps to the probability γ, which can be calibrated by observing the lifecycle of scientific studies (section 6).

5.5. Influence of the completion probability

Finally, we discuss how the results are influenced by the completion probability—the main parameter of the model.

Higher completion probability. The following corollary establishes that the robust critical value is higher when the completion probability is higher.

COROLLARY 5. *Consider a situation with a higher completion probability (higher γ). For a given critical value (z), researchers p-hack more (higher $\mathbb{E}(N(z))$), so type 1 errors are more likely (higher $S^*(z)$). As a result, the robust critical value is higher (higher z^*).*

The corollary indicates that critical values should be higher for research teams with more resources—more time, more money, or more manpower. Research teams with more resources are less likely to be forced to interrupt a study before completion, so they can p-hack more. To control their type 1 error rate properly, a higher critical value is required. The corollary also implies that critical values should be raised when technological progress makes p-hacking easier. An example of such progress is the advent of online surveys and online experiments in the social sciences, which have greatly simplified the task of collecting data. Finally, the corollary implies that critical values should be higher in fields in which p-hacking is easier.

Completion probability of 1. The next corollary establishes that the robust critical value continues to exist but reaches infinity when the completion probability reaches 1.

COROLLARY 6. *Assume that the completion probability reaches 1 ($\gamma \to 1$). For a given critical value (z), researchers conduct $1/\alpha$ steps of p-hacking on average ($\mathbb{E}(N(z)) \to 1/\alpha$), and the probability of type 1 error reaches 1 ($S^*(z) \to 1$). The robust critical value continues to exist but it reaches infinity ($z^* \to \infty$). The average number of p-hacking steps under the robust critical value also reaches infinity ($\mathbb{E}(N(z^*)) \to \infty$).*

The corollary indicates that if researchers can complete any number of tests, they will sample data until they reach significance. Since all null hypotheses are eventually rejected, the probability of type 1 error is 1. At this limit, researchers sample data to reach a foregone conclusion (Anscombe 1954). Yet, the robust critical value continues to exist: it becomes arbitrarily large to offset the arbitrarily large amount of p-hacking.

6. Numerical illustration

As an illustration, we compute the correction required if researchers in the medical and social sciences p-hack by sampling data. We calibrate the completion probability γ from the lifecycle of studies in the medical and social sciences.

6.1. Completion probability in the medical and social sciences

We measure the completion probability γ from data on the lifecycle of studies. In the model, with probability $1-\gamma$, the first research step cannot be completed before running out of resources, so the researcher does not obtain any result. The probability $1 - \gamma$ therefore is the share of studies that never reported results, and the probability γ is the share of studies that reported some results.

Social sciences. Franco, Malhotra, and Simonovits (2014, table 2) report that among 249 social-science studies funded by the National Science Foundation, 20 went missing: they did not report any results and were never written up. An additional 3 studies were written up but without any results. So 23/249 = 9.2% of the studies did not obtain any results, which implies a completion probability of $\gamma = 1 - 9.2\% = 90.8\%$. In addition to these 23 studies, 45 studies were never written up. Because some of the 45 studies had isolated significant or strongly significant results, and because a study that is left unwritten contributes neither to the authors' CV nor to the literature, it seems likely

that researchers were forced to abandon their project by external forces. If we tally these studies, we find that (23 + 45)/249 = 27.3% of the studies were not completed, which lowers the completion probability to $\gamma = 1 - 27.3\% = 72.7\%$. Overall, depending on how we count unwritten studies, the completion probability ranges between 72.7% and 90.8% in the social sciences, with a midpoint of 81.8%.

Medical sciences. Dwan et al. (2008) review 16 metastudies that each follow a cohort of medical studies. The studies are followed from protocol approval to publication, so we can determine the fraction of approved studies that were abandoned. The 16 metastudies consider a total of 6903 approved studies. We focus on the 4563 studies whose fate is known—either by surveying the scientists who conducted the studies or by searching the literature. Of these studies, 658 were never started, or 658/4563 = 14.4%. This number implies a completion probability of $\gamma = 1 - 14.4\% = 85.6\%$. In addition, not all the studies that started were completed. Of the 3905 studies that started, 228 were still ongoing when the cohort studies were written and 388 were stopped early. Hence, of the 3905 – 228 = 3677 studies that started and stopped, 388/3677 = 10.6% stopped early. The vast majority of the studies that stopped early did not include any analysis, so it seems likely that researchers were forced to abandon their project by external forces. Adding the studies that stopped early to those that never started, we find that $14.4\% + (85.6\% \times 10.6\%) = 23.5\%$ of the approved studies were not completed, which yields a completion probability of $\gamma = 1 - 23.5\% = 76.5\%$. Overall, depending on how we count studies that stopped early, the completion probability ranges between 76.5% and 85.6% with a midpoint of 81.1%.

Summary. Combining the evidence from the social and medical sciences, we simply calibrate the completion probability to $\gamma = 80\%$. We use the range 72.7%–90.8% to assess the sensitivity of the findings to the value of γ.

6.2. Obtaining the robust critical value by Bonferroni correction

We now compute the robust critical value from the Bonferroni correction (12). We apply the correction using the completion probability of $\gamma = 80\%$ observed in the medical and social sciences.

Formula. Since the significance level α is always less than 10%, and since γ is less than 1, $1 - \alpha\gamma$ is close to 1, and the average number of p-hacking steps under the robust

critical value is close to $1/(1-\gamma)$ (equation (11)). This gives a very simple Bonferroni correction to deal with p-hacking. From (12), we see that the classical significance level α^* required to deal with p-hacking is approximately $1-\gamma$ times the desired significance level α:

$$\alpha^* \approx (1-\gamma)\alpha.$$

Numerical application. Since $\gamma = 80\%$ in the social and medical sciences, the classical significance level required to deal with p-hacking is just one fifth of the desired significance level: $\alpha^* = (1-0.8) \times \alpha = \alpha/5$. For instance, achieving a 5% type 1 error rate under p-hacking requires using the critical value that yields a 1% type 1 error rate under classical conditions. This rule of thumb works for any hypothesis test. To achieve a 5% type 1 error rate for other plausible completion probabilities, the classical significance level is between 1.4% (when $\gamma = 72\%$) and 0.5% (when $\gamma = 90\%$).

Comparison with the Benjamin et al. (2018) proposal. To address the lack of reproducibility of scientific studies, Benjamin et al. (2018) argue that researchers should replace the standard significance level of 5% by a lower significance level of 0.5%. Our analysis provides a theoretical underpinning for this proposal. We have seen that a tenfold reduction in significance level is appropriate for $\gamma = 90\%$, which is at the top end of plausible completion probabilities. Hence, a critical value constructed from a significance level of 0.5% would be robust to p-hacking for all plausible completion probabilities in the medical and social sciences.

6.3. Additional numerical results

We now use the model to provide additional numerical results. We fix the significance level at 5%.

Prevailing p-hacking. The amount of p-hacking under classical critical values is given by (5). For the completion probability of 80% and a significance level of 5%, the expected number of p-hacking steps is 4.2 (figure 1). Moreover, the amount of p-hacking is increasing with the completion probability. When the completion probability increases from 72.7% to 90.8%, the average number of p-hacking steps grows from 3.2 to 7.3.

Prevailing probability of type 1 error. The probability of type 1 error under classical critical values is given by (8). For the completion probability of 80%, although the

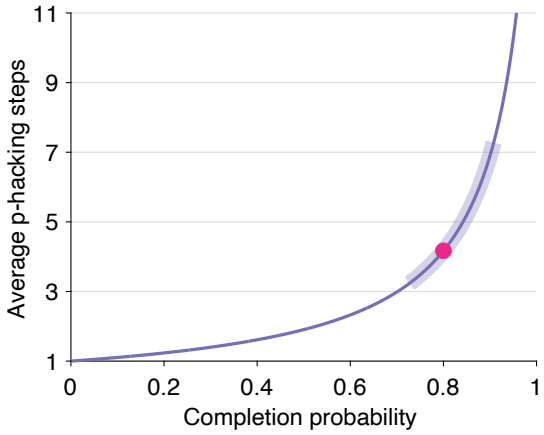


FIGURE 1. Amount of p-hacking at 5% significance level

The thick line indicates plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink point indicates our preferred calibration of the completion probability: $\gamma = 80\%$. The curve is obtained from (5) with $\alpha = 5\%$.

significance level is 5%, the probability of type 1 error is 21% (figure 2). So p-hacking quadruples the probability of type 1 error in this case. Moreover, the distortion caused by p-hacking is more severe when the completion probability is larger—because then there is more p-hacking. When the completion probability increases from 72.7% to 90.8%, the probability of type 1 error increases from 16% to 36%.

Robust critical value in one-sided z-test. We calculate the robust critical value when the significance level is 5% and the underlying test statistic has a standard normal distribution under H_0, as in the common z-test, or in a t-test conducted from a large sample. We begin by calculating the robust critical value for a one-sided z-test. The critical value is given by (10) where $\alpha = 5\%$ and Z is the inverse survival function for the standard normal distribution: $Z(x) = \Phi^{-1}(1 - x)$ where Φ is the standard normal cumulative distribution function. For the completion probability of $\gamma = 80\%$, the robust critical value is 2.31 (figure 3A). It is larger than the corresponding classical critical value of 1.65 but not by a tremendous amount.

Robust critical value in two-sided z-test. Next we calculate the robust critical value for a two-sided z-test. The critical value is now given by (10) where $\alpha = 5\%$ and Z is the inverse survival function for the standard half-normal distribution: $Z(x) = \Phi^{-1}(1 - x/2)$. We use the standard half-normal distribution here because conducting a two-sided test when the test statistic follows a standard normal distribution is equivalent to conducting

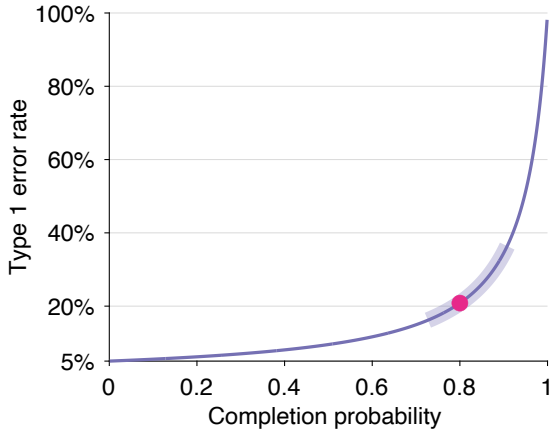


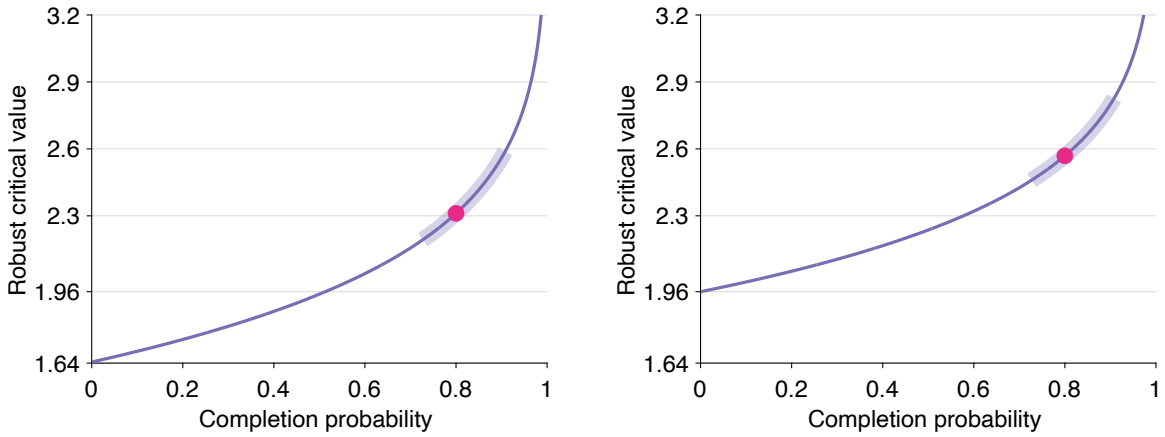
FIGURE 2. Type 1 error rate at 5% significance level

The thick line indicates plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink point indicates our preferred calibration of the completion probability: $\gamma = 80\%$. The curve is obtained from (8) with $\alpha = 5\%$.

a one-sided test with the absolute value of the test statistic—which follows a standard half-normal distribution under H_0. For the completion probability of 80%, the robust critical value is 2.57 (figure 3B). It is larger than the corresponding classical critical value of 1.96 but not by much.

Sensitivity to the completion probability. The robust critical value is increasing with the completion probability, but despite the broad range of p-hacking intensity induced by plausible completion probabilities—from 3 to 7 p-hacking steps on average (figure 1)—the range of plausible robust critical values is fairly narrow. For one-sided tests, the robust critical value remains between 2.19 and 2.59 for all plausible values of the completion probability (figure 3A). For two-sided tests, the robust critical value remains between 2.46 and 2.83 (figure 3B). This is reassuring: robust critical values are not very different in fields with different completion probabilities and p-hacking intensity. It also means that the robust critical value does not depend much on the exact value of the completion probability in a given field.

P-hacking under robust critical value. The average number of p-hacking steps under the robust critical value is given by (11). For the completion probability of 80% and a significance level of 5%, the expected number of p-hacking steps is 4.8 (figure 4A). Moreover, the amount of p-hacking is increasing with the completion probability. When the completion probability increases from 72.7% to 90.8%, the average number of p-



A. One-sided z-test

B. Two-sided z-test

FIGURE 3. Critical value robust to p-hacking for z-test at 5% significance level

The thick lines indicate plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink points indicate our preferred calibration of the completion probability: $\gamma = 80\%$. A: The curve is obtained from (10) where $\alpha = 5\%$ and Z is the inverse survival function for the standard normal distribution. B: The curve is obtained from (10) where $\alpha = 5\%$ and Z is the inverse survival function for the standard half-normal distribution.

hacking steps grows from 3.5 to 10.4. P-hacking is more prevalent under robust critical values than under classical critical values (figure 4B). For instance, under a completion probability of 80% and significance level of 5%, the average number of p-hacking steps increases from 4.2 under the classical critical value to 4.8 under the robust critical value.

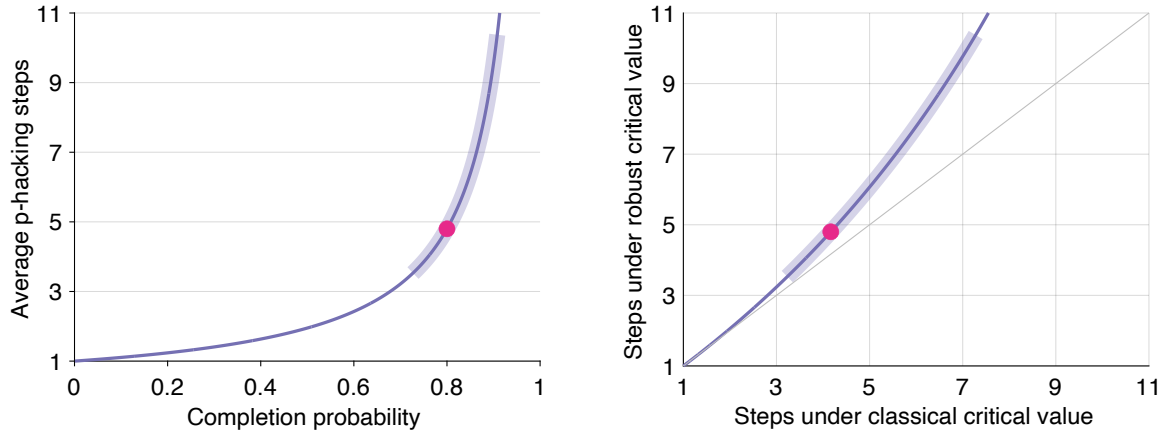
7. Other forms of p-hacking

We have so far assumed that the test statistics sequentially formed by the researcher are independent. However, a p-hacker often forms test statistics that are positively dependent—for instance, when pooling data. We now show that with positive dependence, the robust critical values obtained under the independence assumption continue to maintain control of the type 1 error rate.

7.1. General p-hacking process

PROPOSITION 4. *Rather than assuming that the sequence of test statistics $T_1, \ldots, T_n$ are independent, we assume that*

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) \leq \mathbb{P}(T_n > z) = S(z) \tag{13}$$



A. Average number of p-hacking steps under robust critical value

B. Comparison with the average number of p-hacking steps under classical critical value

FIGURE 4. P-hacking under robust critical value

The thick lines indicate plausible calibrations of the completion probability: $\gamma \in [72.7\%, 90.8\%]$. The pink points indicate our preferred calibration of the completion probability: $\gamma = 80\%$. A: The curve is obtained from (11) with $\alpha = 5\%$. B: The curve is obtained from (5) and (11) with $\alpha = 5\%$ and $\gamma \in (0, 1)$.

for all $z \geq 0$. *Then the probability of type 1 error under the robust critical value* (10) *does not exceed the significance level* α.

In the common case of sequential t-tests, a simple condition on the covariances between successive t-statistics guarantees that proposition 4 applies:

PROPOSITION 5. *Suppose the sequence of test statistics are distributed as follows under* H_0: $(T_1, \ldots, T_n) \sim \mathcal{N}(0, \Omega(n))$, *where all the variances* $\Omega_{1,1}(n)$, …, $\Omega_{n,n}(n)$ *equal 1 and all covariances* $\Omega_{1,n}(n)$,…, $\Omega_{n-1,n}(n)$ *are non-negative. Then condition* (13) *is satisfied so proposition 4 applies.*

The proofs of the propositions are relegated to appendix C, but the intuition is simple. The optimal p-hacking strategy described by lemma 1 remains the same. Indeed, the derivation of the optimal stopping time does not rely on the independence of the test statistics; it remains valid even if the test statistics are dependent. What do change are the stochastic properties of the optimal stopping time and of the reported test statistic. However, under (13), we can guarantee that the robust critical value given by (10) keeps the type 1 error rate below the significance level.

The distributional assumption in proposition 5 is satisfied by the large-sample joint distribution of a sequence of positively correlated t-statistics under the null hypothesis.

Such positive correlation appears under several common forms of p-hacking, as we now discuss.

Suppose that the researcher constructs a general estimator of the form

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m_n} X_{nj} W_{nj}}{\sum_{j=1}^{m_n} X_{nj}^2} \tag{14}$$

at step n, where m_n is equal to the sample size used in step n. In the subsections that follow, we show that several common estimators in applied work take the form of (14). Under standard moment conditions on two sets of m_n approximately iid data points $(X_{n1}, \ldots, X_{nm_n})$ and $(W_{n1}, \ldots, W_{nm_n})$, a bivariate central limit theorem implies the following distributional approximation for large m_n:

$$\frac{1}{\sqrt{m_n}} \begin{pmatrix} \sum_{j=1}^{m_n} [X_{nj} W_{nj} - \mathbb{E}(X_n W_n)] \\ \sum_{j=1}^{m_n} [X_{nj}^2 - \mathbb{E}(X_n^2)] \end{pmatrix} \sim \mathcal{N}(0, \Sigma_n)$$

with

$$\Sigma_n = \begin{pmatrix} \mathbb{E}(X_n^2 W_n^2) - \mathbb{E}(X_n W_n)^2 & \mathbb{E}(X_n^3 W_n) - \mathbb{E}(X_n^2)\,\mathbb{E}(X_n W_n) \\ \mathbb{E}(X_n^3 W_n) - \mathbb{E}(X_n^2)\,\mathbb{E}(X_n W_n) & \mathbb{E}(X_n^4) - \mathbb{E}(X_n^2)^2 \end{pmatrix}.$$

In turn, the delta method implies that for large m_n,

$$\sqrt{m_n}(\hat{\mu}_n - \mu_n) \sim \mathcal{N}(0, \sigma_n^2) \tag{15}$$

with $\mu_n = \mathbb{E}(X_n W_n)/\mathbb{E}(X_n^2)$ and $\sigma_n^2 = [\mathbb{E}(X_n^2 W_n^2)\,\mathbb{E}(X_n^2)^3 - 2\,\mathbb{E}(X_n^3 W_n)\,\mathbb{E}(X_n^2)\,\mathbb{E}(X_n W_n) + \mathbb{E}(X_n^4)\,\mathbb{E}(X_n W_n)^2]/\mathbb{E}(X_n^2)^4$.

By using an estimator of the form (14), (15) shows that the researcher is implicitly testing the null hypothesis $H_{0,n} : \mu_n = \mu_{0,n}$ at step n – the estimand μ_n and its hypothesized value $\mu_{0,n}$ may differ across steps n, depending upon the context (see the examples below). Under standard moment conditions, the researcher can consistently estimate the large-sample variances σ_n^2, by some estimator $\hat{\sigma}_n^2$. This enables the formation of t-statistics with standard normal distributions under $H_{0,n}$ in large samples:

$$T_n = \frac{\sqrt{m_n}(\hat{\mu}_n - \mu_{0,n})}{\hat{\sigma}_n} \sim \mathcal{N}(0, 1).$$

Given $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$ are consistent for σ_i^2 and σ_j^2, $\operatorname{cov}(T_i, T_j) \approx \sqrt{m_i m_j}\operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j)/(\sigma_i \sigma_j) \geq 0$ if and only if $\operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$. Therefore for estimators of the form (14), the conditions

of Proposition 5 hold in large samples when the standard normal approximation for each T_i holds jointly with the others and $\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ for each $i, j = 1, \ldots, n$.

We now provide several common examples of estimators of the form (14) for which these two conditions typically hold.

7.2. Pooling data

The researcher studies a mean parameter $\mu = \mathbb{E}(W)$ for some random variable W. The null hypothesis is $H_0 : \mu = \mu_0$, which does not differ across steps. The alternative hypothesis is $\mu > \mu_0$. At each step the researcher adds data to the previous sample; the additional data are independent and collected from the same underlying population. In step n the researcher constructs an estimate $\hat{\mu}_n$ of the parameter by taking a mean from the pooled sample:

$$\hat{\mu}_n = \frac{1}{m_n} \sum_{j=1}^{m_n} W_j, \tag{16}$$

where m_n is the size of the pooled sample, and $W_1, \ldots, W_{m_n}$ are iid random variables with mean μ. In terms of the notation in (14), we have $X_{nj} = 1$ and $W_{nj} = W_j$ for all n and j.

Since the researcher accumulates data at each step, $m_i > m_j$ for all $i > j$. Hence, using (16) for $i \geq j$, we obtain

$$\text{cov}(\hat{\mu}_i, \hat{\mu}_j) = \frac{1}{m_i m_j} \sum_{r=1}^{m_j} \sum_{k=1}^{m_i} \text{cov}(W_r, W_k) = \frac{\text{var}(W)}{m_i} \geq 0. \tag{17}$$

Here we used the assumption that $W_1, \ldots, W_{m_n}$ are iid, so that $\text{cov}(W_r, W_k) = 0$ for all $r \neq k$ and $\text{cov}(W_r, W_r) = \text{var}(W)$ for all r. Furthermore, any finite set of $\hat{\mu}_i$'s have an approximate joint normal distribution in large samples by a standard multivariate central limit theorem. Therefore, the conditions of proposition 5 are satisfied when the researcher p-hacks by pooling data.

7.3. Removing outliers

Consider the simple case of a researcher successively removing outliers from a given dataset of size m. At step n, the researcher discards all data points further away from some value χ than some given value c_n and discards more data points at each step so

that $c_n < c_q$ for $n > q$. In this scenario, in step n the researcher constructs an estimate $\hat{\mu}_n$ of the parameter by taking a mean from the trimmed sample:

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m} W_j \mathbf{1}(|W_j - \chi| \leq c_n)}{\sum_{j=1}^{m} \mathbf{1}(|W_j - \chi| \leq c_n)}, \tag{18}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function, and $W_1, \ldots, W_m$ are iid random variables. The researcher is implicitly testing a different null hypothesis $H_{0,n} : \mu_n = \mu_{0,n}$ at each step n in this example, where $\mu_n = \mathbb{E}(W\mathbf{1}(|W - \chi| \leq c_n))/\mathbb{P}((|W - \chi| \leq c_n))$. In terms of the notation in (14), we have $X_{nj} = \mathbf{1}(|W_j - \chi| \leq c_n)$, $W_{nj} = W_j$ and $m_n = m$ for all n and j.

Any finite set of $\sum_{j=1}^{m} W_j \mathbf{1}(|W_j - \chi| \leq c_i)$'s and $\sum_{j=1}^{m} \mathbf{1}(|W_j - \chi| \leq c_i)$'s have an approximate joint normal distribution in large samples so that the delta method implies the same for any finite set of $\hat{\mu}_i$'s in this example. In addition, the joint normality of the $\hat{\mu}_i$'s and the delta method further provide the approximate covariance between any two $\hat{\mu}_i$'s in large samples, as the proof of the following proposition shows.

PROPOSITION 6. *For $\hat{\mu}_n$ defined by* (18) *and a sequence W_1, W_2, ... of iid random variables, for any $i \geq j$, $m \operatorname{cov}(\hat{\mu}_i, \hat{\mu}_j)$ converges to*

$$\frac{\operatorname{var}(W||W - \chi| \leq c_i) + \mathbb{E}(W||W - \chi| \leq c_i)\,\mathbb{E}(W||W - \chi| \leq c_j)\,\mathbb{P}(|W - \chi| > c_i)\,\mathbb{P}(|W - \chi| > c_j)}{\mathbb{P}(|W - \chi| \leq c_j)}$$

as $m \to \infty$.

This proposition tells us when the conditions of proposition 5 should hold in large samples. For example, these conditions hold if $\mathbb{E}(W||W - \chi| \leq c_i)$ and $\mathbb{E}(W||W - \chi| \leq c_j)$ have the same sign. It is natural to expect this latter condition to hold in reasonable applications of outlier removal, that is, for reasonable choices of χ and c_n's. For example, suppose $\mathbb{E}(W) = \chi$, so that outliers are considered with respect to deviations from the mean. Then if W is symmetrically distributed, this condition will hold for any choice of c_n since $\mathbb{E}(W||W - \chi| \leq c_n) = \chi$.

7.4. Examining various regression specifications

For this example, we assume that the researcher uses ordinary least squares in the standard linear regression model to estimate an effect of interest. A typical effect of interest would be the population value of a regression coefficient. The researcher uses different regression specifications at each p-hacking step, so the parameter of interest

differs at each step.

Specifically, at step n the researcher uses ordinary least squares to estimate a regression coefficient in a regression of W_n on X_n from two sets of m iid data points $(W_{n1}, \ldots, W_{nm})$ and $(X_{n1}, \ldots, X_{nm})$ so

$$\hat{\mu}_n = \frac{\sum_{j=1}^{m} X_{nj} W_{nj}}{\sum_{j=1}^{m} X_{nj}^2}. \tag{19}$$

Here, X_n represents the regressor of interest after it has been projected off of the space spanned by the covariates included in the nth regression model, following the procedure described in the Frisch-Waugh-Lovell theorem.

Clearly the least squares estimator in (19) takes the structure of (14) with $m_n = m$ for all n and therefore satisfies (15) when, for example, W_n and X_n have finite fourth moments. In this context, the conditions of Proposition 5 therefore hold if $\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ for each $i, j = 1, \ldots, n$. This latter condition is natural in the context of regression if the regressor X_n and regressand W_n measure similar quantities across each step. In other words, if the researcher estimates similar population regression coefficients at each p-hacking step, the coefficient estimates should be expected to be positively correlated in large samples. The condition $\text{cov}(\hat{\mu}_i, \hat{\mu}_j) \geq 0$ is also testable from observed data: the delta method allows one to compute the approximate covariances between any two $\hat{\mu}_i$'s in large samples for any choices of W_i and X_i. Indeed, proposition 6 serves as an example of such an exercise.

7.5. Examining various instruments

By modifying some of the definitions in the previous example, we can also cover the case in which the researcher uses two-stage least squares to estimate the effect of interest. Assuming that the instruments are both strong and valid, we can simply modify the definition of X_n to equal the regressor of interest after all regressors have been projected onto the space spanned by the instruments used at the nth p-hacking step, and then the resulting regressor of interest has been projected off of the space spanned by the covariates included in the nth regression model. If the researcher uses the same dependent variable and second stage covariates at each step and only changes the set of instruments used, and if the regression model is correctly specified, the null hypotheses are identical at each step since each μ_n simply equals the true second stage regression coefficient.

8. Conclusion

To conclude, we summarize our results. We also compare our solution to p-hacking with the registration of pre-analysis plans.

8.1. Summary

Model of p-hacking. This paper models p-hacking in hypothesis tests. It then uses the model to construct critical values that correct the excessive rate of type 1 error caused by p-hacking. Unlike classical critical values, these robust critical values deliver the promised rate of type 1 error. Once the robust critical values are in place, researchers continue to p-hack, but readers can be confident that true null hypotheses are not rejected more often than the advertised significance level.

Numerical illustration. Robust critical values are larger than classical critical values. For instance, when we calibrate the p-hacking process with evidence from the social and medical sciences, we find that the robust critical value for any test and any significance level is the classical critical value for the same test with roughly one fifth of the significance level. For instance, for a z-test with a significance level of 5%, the robust critical value is 2.31 instead of 1.65 if the test is one-sided and 2.57 instead of 1.96 if the test is two-sided.

8.2. Comparison with pre-analysis plans

Using robust critical values. The most popular solution to p-hacking is to ask researchers to register a pre-analysis plan (Miguel et al. 2014; Christensen and Miguel 2018; Nosek et al. 2018; Christensen, Freese, and Miguel 2019). Pre-analysis plans prevent many forms of p-hacking (Adda, Decker, and Ottaviani 2020). Yet they are not without limitations (Olken 2015; Coffman and Niederle 2015; Banerjee et al. 2020; Abrams, Libgober, and List 2021). First, pre-analysis plans prevent scientists from engaging in exploratory analysis, although exploration is an important source of scientific discovery. Second, pre-analysis plans are sometimes impractical, either because it is hard to formulate hypotheses before seeing the data, or because the scientist is already familiar with the data. Using critical values robust to p-hacking therefore seems more appropriate than imposing pre-analysis plans when scientific exploration plays a key role, and with observational data.

Combining pre-analysis plans with robust critical values. Even in settings where pre-analysis plans are appropriate, it might make sense to use them in conjunction with critical values robust to p-hacking. The reason is that pre-analysis plans cannot prevent the data sampling that we consider in our model. Furthermore, they cannot prevent researchers from strategically selecting a subset of the data that they collected in order to achieve significant results (Lang and Qiu 2021). Using robust critical values in conjunction with pre-analysis plans would eliminate the excessive type 1 error rates caused by strategic data selection.

References

Abrams, Eliot, Jonathan Libgober, and John A. List. 2021. "Research Registries and the Credibility Crisis: An Empirical and Theoretical Investigation." https://perma.cc/NJG8-55QV.

Adda, Jerome, Christian Decker, and Marco Ottaviani. 2020. "P-hacking in Clinical Trials and How Incentives Shape the Distribution of Results Across Phases." *Proceedings of the National Academy of Sciences* 117 (24): 13386–13392.

Akerlof, George A., and Pascal Michaillat. 2018. "Persistence of False Paradigms in Low-Power Sciences." *Proceedings of the National Academy of Sciences* 115 (52): 13228–13233.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–2794.

Anscombe, Francis J. 1954. "Fixed-Sample-Size Analysis of Sequential Observations." *Biometrics* 10 (1): 89–100.

Armitage, Peter. 1967. "Some Developments in the Theory and Practice of Sequential Medical Trials." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by Lucien M. Le Cam and Jerzy Neyman, vol. 4: 791–804. Berkeley, CA: University of California Press.

Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–554.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics." NBER Working Paper 26993.

Begg, Colin B., and Jesse A. Berlin. 1988. "Publication Bias: a Problem in Interpreting Medical Data." *Journal of the Royal Statistical Society (Series A)* 151 (3): 419–445.

Begley, C. Glenn, and Lee M. Ellis. 2012. "Raise Standards for Preclinical Cancer Research." *Nature* 483: 531–533.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Bjorn Brembs, Lawrence Brown, Colin Camerer et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6–10.

Biagioli, Mario, and Alexandra Lippman. 2020. "Metrics and the New Ecologies of Academic

Misconduct." In *Gaming the Metrics: Misconduct and Manipulation in Academic Research*, edited by Mario Biagioli and Alexandra Lippman, 1–23. Cambridge, MA: MIT Press.

Bozarth, Jerold D., and Ralph R. Roberts. 1972. "Signifying Significant Significance." *American Psychologist* 27 (8): 774–775.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–3660.

Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: the Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jurgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 433–1436.

Chen, Andrew Y. 2021. "The Limits of P-hacking: Some Thought Experiments." *Journal of Finance* 76 (5): 2447–2480.

Christensen, Garret. 2018. "Manual of Best Practices in Transparent Social Science Research." https://github.com/garretchristensen/BestPracticesManual/blob/65b77b1991e9b6d5360d3fc6aa2bb7528bedf7ff/Manual.pdf.

Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. Oakland, CA: University of California Press.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–980.

Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (3): 81–98.

Cole, LaMont C. 1957. "Biological Clock in the Unicorn." *Science* 125 (3253): 874–876.

Csada, Ryan D., Paul C. James, and Richard H. M. Espie. 1996. "The 'File Drawer Problem' of Non-Significant Results: Does It Apply to Biological Research?" *Oikos* 76 (3): 591–593.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, edited by T. Paul Schultz and John A. Strauss, vol. 4, 3895–3962. Amsterdam: Elsevier.

Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, Carrol Gamble et al. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLoS ONE* 3 (8): e3081.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wuthrich. 2021. "Detecting P-hacking." *Econometrica*. https://www.econometricsociety.org/system/files/18583-3.pdf.

Fanelli, Daniele, Rodrigo Costas, and John P. A. Ioannidis. 2017. "Meta-Assessment of Bias in Science." *Proceedings of the National Academy of Sciences* 114 (14): 3714–3719.

Ferguson, Thomas S. 2007. "Optimal Stopping and Applications." https://web.archive.org/web/20200812154935/https://www.math.ucla.edu/~tom/Stopping/Contents.html.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–1505.

Gibson, John, David L. Anderson, and John Tressler. 2014. "Which Journal Rankings Best Explain Academic Salaries? Evidence from the University of California." *Economic Inquiry* 52 (4): 1322–1340.

Glaeser, Edward L. 2008. "Researcher Incentives and Empirical Methods." In *The Foundations of Positive and Normative Economics: A Hand Book*, edited by Andrew Caplin and Andrew Schotter, chap. 13. New York: Oxford University Press.

Hagstrom, Warren. 1965. *The Scientific Community*. New York: Basic Books.

Hansen, W. Lee, Burton A. Weisbrod, and Robert P. Strauss. 1978. "Modeling the Earnings and Research Productivity of Academic Economists." *Journal of Political Economy* 86 (4): 729–741.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-hacking in Science." *PLoS Biology* 13 (3): e1002106.

Howard, Steven R., Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. 2021. "Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences." *Annals of Statistics* 49 (2): 1055–1080.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21 (1): 1–20.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3): 944–960.

Hutton, J. L., and Paula R. Williamson. 2000. "Bias in Meta-Analysis Due to Outcome Variable Selection Within Studies." *Applied Statistics* 49 (3): 359–370.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124.

Ioannidis, John P. A., Sander Greenland, Mark A. Hlatky, Muin J. Khoury, Malcolm R. Macleod, David Moher, Kenneth F. Schulz, and Robert Tibshirani. 2014. "Increasing Value and Reducing Waste in Research Design, Conduct, and Analysis." *Lancet* 383 (9912): 166–175.

Ioannidis, John P.A., and Thomas A. Trikalinos. 2007. "An Exploratory Test for an Excess of Significant Findings." *Clinical Trials* 4 (3): 245–253.

Jennions, Michael D., and Anders P. Moeller. 2002. "Publication Bias in Ecology and Evolution: An Empirical Assessment Using the 'Trim and Fill' Method." *Biological Reviews* 77 (2): 211–222.

Jennison, Christopher, and Bruce W. Turnbull. 1984. "Repeated Confidence Intervals for Group Sequential Clinical Trials." *Controlled Clinical Trials* 5 (1): 33–45.

Johari, Ramesh, Pete Koomen, Leonid Pekelis, and David J. Walsh. 2021. "Always Valid Inference: Continuous Monitoring of A/B Tests." *Operations Research*. https://doi.org/10.1287/opre.2021.2135.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–532.

Katz, David A. 1973. "Faculty Salaries, Promotions, and Productivity at a Large University." *American Economic Review* 63 (3): 469–477.

Kuhn, Thomas S. 1957. *The Copernican Revolution*. Cambridge, MA: Harvard University Press.

Lang, Megan, and Wenfeng Qiu. 2021. "Cherry Picking." https://doi.org/10.31222/osf.io/as9zd.
Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73 (1): 31–43.
Lindsay, D. Stephen. 2015. "Replication in Psychological Science." *Psychological Science* 26 (12): 1827–1832.
Lovell, Michael C. 1983. "Data Mining." *Review of Economics and Statistics* 65 (1): 1–12.
Merton, Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* 22 (6): 635–659.
Merton, Robert K. 1973. *The Sociology of Science*. Chicago: University of Chicago Press.
Miguel, Edward, Colin Camerer, Katherine Casey, Joshua Cohen, Kevin M. Esterling, Alan Gerber, Rachel Glennerster, Don P. Green, Macartan Humphreys, Guido Imbens et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.
Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606.
Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspectives on Psychological Science* 7 (6): 615–631.
Olken, Benjamin A. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.
Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.
Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" *Nature Reviews Drug Discovery* 10: 712.
Sauer, Raymond D. 1988. "Estimates of the Returns to Quality and Coauthorship in Economic Academia." *Journal of Political Economy* 96 (4): 855–866.
Siegfried, John J., and Kenneth J. White. 1973. "Financial Rewards to Research and Teaching: A Case Study of Academic Economists." *American Economic Review* 63 (2): 309–315.
Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–1366.
Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143 (2): 534.
Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. "Specification Curve Analysis." *Nature Human Behaviour* 4 (11): 1208–1214.
Skeels, Jack W., and Robert P. Fairbanks. 1968. "Publish or Perish: An Analysis of the Mobility of Publishing and Nonpublishing Economists." *Southern Economic Journal* 35 (1): 17–25.
Song, F., A. J. Eastwood, S. Gilbody, L. Duley, and A. J. Sutton. 2000. "Publication and Related Biases: A Review." *Health Technology Assessment* 4 (10): 1–115.
Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11 (5): 702–712.

Sterling, Theodore D. 1959. “Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance–Or Vice Versa.” *Journal of the American Statistical Association* 54 (285): 30–34.

Swidler, Steve, and Elizabeth Goldreyer. 1998. “The Value of a Finance Journal Publication.” *Journal of Finance* 53 (1): 351–363.

Tuckman, Howard P., and Jack Leahey. 1975. “What Is an Article Worth?” *Journal of Political Economy* 83 (5): 951–967.

Vivalt, Eva. 2019. “Specification Searching and Significance Inflation Across Time, Methods and Disciplines.” *Oxford Bulletin of Economics and Statistics* 81 (4): 797–816.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. “The ASA’s Statement on P-values: Context, Process, and Purpose.” *American Statistician* 70 (2): 129–133.

Young, Cristobal, and Katherine Holsteen. 2017. “Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis.” *Sociological Methods & Research* 46 (1): 3–40.

Appendix A. Adding a monetary or psychological cost of research

We extend the p-hacking model of section 3 by introducing a cost of research, incurred at each new p-hacking step. The cost could be monetary or psychological. We find that the robust critical value is not modified by this extension. Hence, our model of p-hacking is robust to the presence of research costs.

A.1. Assumptions

We introduce an expected cost of doing research, c. The cost could be monetary or psychological; it is incurred at each p-hacking step. Because we focus on fields in which research occurs, we assume that c is low enough relative to the rewards from research, v^i and v^s, such that it is optimal for researchers to engage in research.

A.2. Optimal stopping time and robust critical value

Significant result. Since it is optimal to engage in research, the researcher starts a first p-hacking step. With probability γ, the step can be completed, and the researcher obtains a test result. If the result is significant, the researcher obtains v^s, so she stops immediately. Indeed, she cannot obtain a higher payoff by continuing. The same is true at any future step: any time a researcher obtains a significant result, she immediately stops, since it is impossible to obtain a higher payoff in the future.

High research cost. What does the researcher decide if the result is insignificant? It depends on the research cost c. If the cost is high enough, the researcher stops right away. This happens when the possibility of obtaining a significant result in the future does not compensate the research cost. In that case, there is no p-hacking: the researcher conducts one step of the study and stops, irrespective of the result. The robust critical value is then just the classical critical value.

Low research cost. Since p-hacking is prevalent in reality, the most realistic scenario is that the research cost is low enough so that the researcher starts a new step upon obtaining a first insignificant result. In that case, because the researcher faces exactly the same situation after each step, the researcher continues to p-hack until she obtains a significant result.

Summary. If the research cost is low enough that p-hacking occurs, the presence of the research cost does not modify the researcher's behavior. It is optimal for the researcher to p-hack until she reaches a significant result. Accordingly, everything remains the same in the model—including the robust critical value.

A.3. Computing the cost boundaries

We now compute the expected payoffs from doing research, the cost below which it is optimal to p-hack, and the cost below which it is optimal to engage in research. The expectations of the payoffs depend on the distribution of the test statistic, which in turn depends on which hypothesis is true. We assume that the researcher is conservative and computes the payoff expectations under the null hypothesis.

Continuation value of research. We first compute the continuation value of research for a researcher who has already recorded an insignificant result. We denote this value V^i. Because the researcher's situation is invariant in time, the continuation value is the same at each p-hacking step.

When a researcher decides to continue p-hacking, three scenarios are possible. With probability $1-\gamma$, the researcher cannot complete the p-hacking step and must submit an insignificant result. She then collects v^i. With probability γ, she can complete the p-hacking step. Then with probability $S(z^*)$, her result is significant and she collects v^s. With probability $1-S(z^*)$, her result is insignificant once again and the continuation value at this point is V^i. In any case, she must incur a cost c to conduct the research step.

Aggregating these scenarios, we obtain the following continuation value:

$$V^i = (1-\gamma)v^i + \gamma S(z^*)v^s + \gamma[1-S(z^*)]V^i - c.$$

Hence the continuation value is

$$V^i = \frac{(1-\gamma)v^i + \gamma S(z^*)v^s - c}{1-\gamma[1-S(z^*)]}. \tag{A1}$$

Condition for p-hacking. From the continuation value (A1), we compute the cost below which it is optimal to p-hack. When a researcher has obtained one insignificant result, it is optimal to continue p-hacking if $V^i > v^i$. After a few steps of algebra, this condition

becomes

$$c < \gamma S(z^*)(v^s - v^i).$$

Hence, it is optimal to p-hack if the cost of each p-hacking step is below the threshold

$$c^p = \gamma S(z^*)(v^s - v^i).$$

Of course, the cost threshold is higher when significant results are more rewarded relative to insignificant results.

Condition for research. From the continuation value (A1), we also compute the cost below which it is optimal to engage in research. Given that we have normalized the outside option of the researcher to 0, it is optimal to engage in research if the expected value from it is positive.

When a researcher decides to start research, three scenarios are again possible. With probability $1-\gamma$, the researcher cannot complete the first research step and cannot submit any result; she then collects 0. With probability γ, she can complete the first research step. Then with probability $S(z^*)$, her result is significant and she collects v^s. With probability $1 - S(z^*)$, her result is insignificant and the continuation value at this point is V^i. In any case, she must incur a cost c to conduct the research step.

Aggregating these scenarios, we obtain the initial continuation value:

$$V^r = (1-\gamma) \times 0 + \gamma S(z^*)v^s + \gamma[1 - S(z^*)]V^i - c.$$

We rewrite the initial continuation value as

$$V^r = \gamma V^i + \gamma S(z^*)(v^s - V^i) - c.$$

Using the value of V^i given by (A1), we finally obtain

$$V^r = \frac{\gamma S(z^*)}{1-\gamma[1-S(z^*)]}v^s + \frac{(1-\gamma)\gamma[1-S(z^*)]}{1-\gamma[1-S(z^*)]}v^i - \frac{1}{1-\gamma[1-S(z^*)]} \cdot c. \tag{A2}$$

It is optimal to start a research project if $V^r > y_0 = 0$. From (A2), this condition becomes

$$c < \gamma S(z^*)v^s + (1-\gamma)\gamma[1-S(z^*)]v^i.$$

Hence, it is optimal to start research if the cost of each research step is below the

threshold

$$c^r = \gamma S(z^*)v^s + (1-\gamma)\gamma[1-S(z^*)]v^i.$$

The cost threshold is higher when scientific results are more rewarded.

The threshold to engage in research is higher than the threshold to engage in p-hacking:

$$c^r = c^p + \gamma[1-\gamma(1-S(z^*))] > c^p.$$

Hence, for all cost between c^p and c^r, researchers engage in research but do not p-hack.

Appendix B. Adding time discounting

We now introduce time discounting into the p-hacking model of section 3. When the researcher discounts the future, a result submitted early is more valuable than the same result submitted later. Yet, the researcher's behavior and robust critical value are not modified.

B.1. Assumptions

We introduce a discount factor, $\delta \in (0,1)$. The discount factor cost is incurred at each new p-hacking step, so the value of a research result obtained at step n is discounted by δ^n. Because the returns to research are positive without discounting, they also are positive with discounting, so it is optimal for researchers to engage in research.

B.2. Optimal stopping time

As in appendix A, we find that the optimal stopping time is the same as in the basic model.

Significant result. Any time a researcher obtains a significant result, she immediately stops, since it is impossible to obtain a higher payoff in the future.

High discounting. What does the researcher decide if the result is insignificant? It depends on the value of the discount factor δ. If discounting is high enough, the researcher is better off stopping right away. This happens when the possibility of obtaining a significant result in the future does not compensate the time discounting. In that case, there

is no p-hacking: the researcher conducts one step of the study and stops, irrespective of the result. The robust critical value is then just the classical critical value.

Low discounting. Since p-hacking is prevalent in reality, the most realistic scenario is that discounting is low enough so the researcher starts a new step upon obtaining a first insignificant result. In that case, because the researcher faces exactly the same situation after each step, the researcher continues to p-hack until she obtains a significant result.

Summary. If time discounting is low enough that p-hacking occurs, the presence of discounting does not modify the researcher's behavior. It is optimal for the researcher to p-hack until she reaches a significant result. Accordingly, everything remains the same in the model—including the robust critical value.

B.3. Computing discounting boundary

Given that all the properties of the model remain the same with discounting, we can use previous results to compute the discount factor below which it is optimal to p-hack.

The key step is computing the continuation value of research for a researcher who has already recorded an insignificant result. We denote this value V^i. Because the researcher's situation is invariant in time, this continuation value is the same at each new p-hacking step.

When a researcher decides to continue p-hacking, three scenarios are possible. With probability $1-\gamma$, the researcher cannot complete the new p-hacking step and must submit an insignificant result; she then collects δv^i. With probability γ, she can complete the new p-hacking step. Then with probability $S(z^*)$, her result is significant and she collects δv^s; with probability $1-S(z^*)$, her result is insignificant once again and the continuation value at this point is δV^i.

Aggregating these scenarios, we obtain the following continuation value:

$$V^i = (1-\gamma)\delta v^i + \gamma S(z^*)\delta v^s + \gamma[1-S(z^*)]\delta V^i.$$

Hence the continuation value is

$$V^i = \delta \frac{(1-\gamma)v^i + \gamma S(z^*)v^s}{1-\delta\gamma[1-S(z^*)]}.$$

Then, when a researcher has obtained one insignificant result, it is optimal to con-

tinue p-hacking if $V^i > v^i$. After a few steps of algebra, this condition becomes

$$\delta > \frac{v^i}{v^i + \gamma S(z^*)(v^s - v^i)}.$$

Hence, it is optimal to p-hack if the discount factor is above the threshold

$$\delta^p = \frac{v^i}{v^i + \gamma S(z^*)(v^s - v^i)}.$$

Of course, the discounting threshold is lower when significant results are more rewarded relative to insignificant results. If insignificant results are not rewarded at all, then researchers p-hack irrespective of discounting.

Appendix C. Proofs

We provide proofs that are omitted in the main text.

C.1. Proof of proposition 2

We start by computing the probability that the reported test statistic $R(z)$ exceeds a critical value z under the null hypothesis. From the law of total probability:

$$\mathbb{P}(R(z) > z) = \sum_{j \geq 1} \mathbb{P}(R(z) > z \mid N(z) = j)\, \mathbb{P}(N(z) = j), \tag{A3}$$

where, according to Bayes' rule,

$$\mathbb{P}(R(z) > z \mid N(z) = j) = \frac{\mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j - 1)}{\mathbb{P}(N(z) = j \mid N(z) > j - 1)}. \tag{A4}$$

Then we compute the conditional probability given by (A4). The fact that $N(z) > j - 1$ means that the project resources have not been exhausted during the first $j - 1$ steps, and that the $j - 1$ test statistics collected have not been significant. Conditional on $N(z) > j - 1$, three events may happen.

First, with probability $1 - \gamma$, resources are exhausted during step j. If $j > 1$, then $N(z) = j$ and the researcher reports an insignificant result: $R(z) \leq z$. If $j = 1$, the researcher does not report any result.

Second, with probability γ, resources are not exhausted during step j. This creates

two subcases. With probability $\gamma S(z)$, the test statistic T_j obtained during step j is significant. Then $N(z) = j$ and $R(z) = T_j > z$. With probability $\gamma[1 - S(z)]$, the test statistic T_j obtained during step j is insignificant. In that case, $N(z) > j$.

From this case-by-case description, we see that the probability that the researcher stops at step j given that she has already completed $j - 1$ steps is

$$\mathbb{P}(N(z) = j \mid N(z) > j - 1) = 1 - \gamma + \gamma S(z). \tag{A5}$$

And the probability that the researcher reports a significant result at step j given that she has already completed $j - 1$ steps is

$$\mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j - 1) = \mathbb{P}\Big(T_j > z, K > j \mid N(z) > j - 1\Big) = \gamma S(z). \tag{A6}$$

Combining (A4), (A5), and (A6), we find that the probability to report a significant result given that the researcher stops the research project at step j is

$$\mathbb{P}(R(z) > z \mid N(z) = j) = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}. \tag{A7}$$

The probability (A7) is independent of j, which greatly simplifies (A3):

$$\mathbb{P}(R(z) > z) = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)} \left[\sum_{j \geq 1} \mathbb{P}(N(z) = j)\right] = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}.$$

Finally, we compute the probability of reporting a significant result given that any result is reported. This conditional probability is given by

$$\mathbb{P}(R(z) > z \mid L > D_1) = \frac{\mathbb{P}(R(z) > z)}{\mathbb{P}(L > D_1)} = \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)} \cdot \frac{1}{\gamma}.$$

To compute the ratio, we use the fact that with probability γ, resources are not exhausted before the end of the first step, so some results will be reported, either significant or insignificant.

Therefore, when the critical value is set to z, the probability of type 1 error in a reported study is

$$S^*(z) = \mathbb{P}(R(z) > z \mid L > D_1) = \frac{S(z)}{1 - \gamma + \gamma S(z)}.$$

C.2. Proof of proposition 3

First, to compute the robust critical value, we rewrite the definition (9) as

$$S(z^*) = \alpha \cdot \frac{1-\gamma}{1-\alpha\gamma}.$$

The inverse of the survival function S is the function Z. Inverting S here, we obtain the explicit expression for the robust critical value:

$$z^* = Z\left(\alpha \cdot \frac{1-\gamma}{1-\alpha\gamma}\right). \tag{A8}$$

Equation (A8) indicates that the robust critical value always exists. Since $\alpha \in (0,1)$ and $\gamma \in (0,1)$, the ratio $(1-\gamma)/(1-\alpha\gamma)$ is in $(0,1)$. Hence, the argument of the inverse survival function Z in (10) satisfies

$$0 < \alpha \cdot \frac{1-\gamma}{1-\alpha\gamma} < \alpha.$$

Accordingly, the argument is in $(0,1)$. As the domain of the inverse survival function is $(0,1)$, the robust critical value exists.

From (A8), we can compare the robust critical value to a classical critical value. A classical critical value is defined by $z = Z(\alpha)$, while the robust critical value is defined by (A8). Since the inverse survival function is strictly decreasing, and since the argument of the inverse survival function in (10) is strictly less than α, we infer that the robust critical value is strictly larger than the classical critical value: $z^* > z$.

Unsurprisingly, the robust critical value is strictly decreasing in the significance level α. Indeed, the argument of the inverse survival function in (A8) is strictly increasing in the significance level $\alpha \in (0,1)$. Since the inverse survival function itself is strictly decreasing, we infer that the robust critical value is strictly decreasing in the significance level.

C.3. Proof of proposition 4

The proof proceeds as the proof of proposition 2, with some adjustments. In particular, note that (A3) and (A4) continue to hold and the probability that resources are exhausted at any step k continues to be $1-\gamma$. However conditional on $N(z) > j-1$, the probability the test statistic obtained during step k is significant is now bounded above by $\gamma S(z)$

since

$$\text{(A9)} \qquad \mathbb{P}(T_n > z \mid N(z) > j-1) = \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) \leq S(z).$$

Therefore, (A5) no longer holds but can be replaced by

$$\text{(A10)} \qquad \mathbb{P}(N(z) = j \mid N(z) > j-1) = 1 - \gamma + \gamma\, \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z).$$

Similarly, (A6) no longer holds but can be replaced by

$$\text{(A11)} \qquad \mathbb{P}(R(z) > z, N(z) = j \mid N(z) > j-1) = \gamma\, \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z).$$

Since the function $x \mapsto x/(1-\gamma+x)$ is increasing in $x > 0$ for all $\gamma < 1$, (A9), (A10), (A11) and (A4) imply

$$\mathbb{P}(R(z) > z \mid N(z) = j) \leq \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}$$

so that (A3) implies

$$\mathbb{P}(R(z) > z) \leq \frac{\gamma S(z)}{1 - \gamma + \gamma S(z)}.$$

Applying (9), we obtain the statement of the proposition:

$$\mathbb{P}\big(R(z^*) > z \mid L > D_1\big) = \frac{\mathbb{P}(R(z^*) > z^*)}{\mathbb{P}(L > D_1)} \leq \frac{\gamma S(z^*)}{1 - \gamma + \gamma S(z^*)} \cdot \frac{1}{\gamma} = \frac{S(z^*)}{1 - \gamma + \gamma S(z^*)} = \alpha.$$

C.4. Proof of proposition 5

We show (13) holds by showing the conditional probability on the left-hand side is less than the unconditional probability on the right-hand side after further conditioning on any realized value of an additional statistic.

Note that the normally-distributed random vector

$$A(n) = [T_1, \ldots, T_{n-1}] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n$$

is independent of T_n since

$$\begin{aligned}
\operatorname{cov}(A(n), T_n) &= \operatorname{cov}([T_1, \ldots, T_{n-1}] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n, T_n) \\
&= \operatorname{cov}([T_1, \ldots, T_{n-1}], T_n) - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)] \operatorname{var}(T_n, T_n) \\
&= [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)] - [\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)] = 0.
\end{aligned}$$

Using the vector $A(n)$, we describe the conditioning event in (13) as follows:

$$\begin{aligned}\{T_1, \ldots, T_{n-1} \leq z\} &= \left\{[\Omega_{1,n}(n), \ldots, \Omega_{n-1,n}(n)]T_n \leq z - A(n)\right\} \\ &= \left\{T_n \leq \min_{1\leq j\leq n-1:\Omega_{j,n}(n)>0} \frac{z - A_j(n)}{\Omega_{j,n}(n)}, \max_{1\leq j\leq n-1:\Omega_{j,n}(n)=0} A_j(n) \leq z\right\}.\end{aligned}$$

Since $A(n)$ and T_n are independent, the conditional distribution of the nth t-statistic given the conditioning event in (13) and the realized value of $A(n)$ is a standard normal truncated from above:

$$T_n \mid \left\{T_1, \ldots, T_{n-1} \leq z, A(n) = a\right\} \sim \xi \mid \xi \leq \mathcal{U}(a),$$

where $\xi \sim \mathcal{N}(0, 1)$ and

$$\mathcal{U}(a) = \min_{1\leq j\leq n-1:\Omega_{j,n}(n)>0} \frac{z - a_j}{\Omega_{j,n}(n)}.$$

Using the properties of the truncated normal distribution, we characterize the conditional probability of type 1 error for the nth t-statistic given non-rejection by the previous t-statistics in the sequence and the realized value of $A(n)$ as

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) = \begin{cases} 1 - \frac{\Phi(z)}{\Phi(\mathcal{U}(a))} & \text{if } z \leq \mathcal{U}(a), \\ 0 & \text{if } z > \mathcal{U}(a) \end{cases}$$

for all a, where Φ denotes the cumulative distribution function of a standard normal random variable. Therefore for any values of a and z,

$$\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) \leq 1 - \Phi(z).$$

But for $F_A(\cdot)$ equal to the cumulative distribution function of $A(n)$,

$$\begin{aligned}\mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z) &= \int_{\mathbb{R}^{n-1}} \mathbb{P}(T_n > z \mid T_1, \ldots, T_{n-1} \leq z, A(n) = a) dF_A(a) \\ &\leq 1 - \Phi(z) = \mathbb{P}(T_n > z)\end{aligned}$$

and we obtain the statement of the proposition.

C.5. Proof of proposition 6

A multivariate central limit theorem and delta method immediately imply

$$
\begin{aligned}
m\,\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_j) \to\ & \frac{\mathrm{cov}(W\mathbf{1}(|W-\chi| \le c_i), W\mathbf{1}(|W-\chi| \le c_j))}{\mathbb{P}(|W-\chi| \le c_i)\,\mathbb{P}(|W-\chi| \le c_j)} \\
& - \frac{\mathbb{E}(W\mathbf{1}(|W-\chi| \le c_j)\,\mathrm{cov}(W\mathbf{1}(|W-\chi| \le c_i), \mathbf{1}(|W-\chi| \le c_j))}{\mathbb{P}(|W-\chi| \le c_i)\,\mathbb{P}(|W-\chi| \le c_j)^2} \\
& - \frac{\mathbb{E}(W\mathbf{1}(|W-\chi| \le c_i)\,\mathrm{cov}(W\mathbf{1}(|W-\chi| \le c_j), \mathbf{1}(|W-\chi| \le c_i))}{\mathbb{P}(|W-\chi| \le c_j)\,\mathbb{P}(|W-\chi| \le c_i)^2} \\
& + \frac{\mathbb{E}(W\mathbf{1}(|W-\chi| \le c_i)\,\mathbb{E}(W\mathbf{1}(|W-\chi| \le c_j)\,\mathrm{cov}(\mathbf{1}(|W-\chi| \le c_j), \mathbf{1}(|W-\chi| \le c_i))}{\mathbb{P}(|W-\chi| \le c_j)^2\,\mathbb{P}(|W-\chi| \le c_i)^2}
\end{aligned}
$$

as $m \to \infty$. Using the definition of covariance and the facts that for $f(w) = w$ or $f(w) = w^2$,

$$
\begin{gathered}
\mathbb{E}(f(W)||W-\chi| \le c_i) = \mathbb{E}(f(W)\mathbf{1}(|W-\chi| \le c_i))/\,\mathbb{P}(|W-\chi| \le c_i), \\
\mathbf{1}(|W-\chi| \le c_i)\mathbf{1}(|W-\chi| \le c_j) = \mathbf{1}(|W-\chi| \le c_i)
\end{gathered}
$$

since $c_i < c_j$, standard algebra then yields the result of the proposition.