Detecting p-hacking[*]

Graham Elliott[†]   Nikolay Kudrin[‡]   Kaspar Wüthrich[§]

May 25, 2021

Abstract

We theoretically analyze the problem of testing for p-hacking based on distributions of p-values across multiple studies. We provide general results for when such distributions have testable restrictions (are non-increasing) under the null of no p-hacking. We find novel additional testable restrictions for p-values based on t-tests. Specifically, the shape of the power functions results in both complete monotonicity as well as bounds on the distribution of p-values. These testable restrictions result in more powerful tests for the null hypothesis of no p-hacking. When there is also publication bias, our tests are joint tests for p-hacking and publication bias. A reanalysis of two prominent datasets shows the usefulness of our new tests.

Keywords: p-values, p-curve, complete monotonicity, publication bias

[†]Department of Economics, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093. Email: grelliott@ucsd.edu

[‡]Department of Economics, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093. Email: nkudrin@ucsd.edu

[§]Department of Economics, University of California, San Diego, 9500 Gilman Dr. La Jolla, CA 92093. Email: kwuthrich@ucsd.edu

1 Introduction

A researcher's ability to explore various ways of analyzing and manipulating data and then selectively report the ones that yield better-looking results, commonly referred to as *p-hacking*, compromises the reliability of research and undermines the scientific credibility of reported results. Absent systematic replication studies or meta analyses, a popular approach for assessing the extent of p-hacking is to examine distributions of p-values across studies, referred to as *p-curves* (Simonsohn et al., 2014); see Section 2 in Christensen and Miguel (2018) for a review.[1]

We consider the problem of testing the *null hypothesis of no p-hacking* against the *alternative hypothesis of p-hacking* and provide theoretical foundations for developing tests for p-hacking. We characterize analytically under general assumptions the null set of distributions of p-values implied in the absence of p-hacking and provide general sufficient conditions under which, for any distribution of the true effects, the p-curve is non-increasing and continuous in the absence of p-hacking. These conditions are shown to hold for many, but not all popular approaches to testing for effects.

For the leading case where p-curves are based on t-tests, we derive additional previously unknown testable restrictions. Specifically, the p-curves based on t-tests are completely monotone in the absence of p-hacking, and their magnitude and the magnitude of their derivatives are restricted by upper bounds. These restrictions are particularly useful when p-hacking fails to induce an increasing p-curve—for example when researchers engage in specification search across independent tests. In such cases tests based on non-increasingness have no power.

Our theoretical results allow us to develop more powerful statistical tests for p-hacking, which we apply to two large datasets of p-values. We find evidence for p-hacking in settings where the existing tests do not reject the null of no p-hacking.

When there is publication bias, our results characterize the p-curve under the null hypothesis of *no p-hacking and no publication bias*. Our tests become joint tests for p-hacking and publication bias, complementing available methods for identifying publication bias (see, e.g., Andrews and Kasy, 2019, and the references therein).

---

[1]Examples include: Masicampo and Lalande (2012), Leggett et al. (2013), Simonsohn et al. (2014, 2015), Head et al. (2015), de Winter and Dodou (2015), and Snyder and Zhuo (2018). Another strand of the literature uses the distribution of t-statistics to test for p-hacking (e.g., Gerber and Malhotra, 2008; Brodeur et al., 2016b, 2020; Bruns et al., 2019; Vivalt, 2019).

2 The p-curve based on general tests

Here we provide general sufficient conditions under which the p-curve is non-increasing under the null hypothesis of no p-hacking. These results are useful because tests for p-hacking often assume non-increasingness of the p-curve (e.g., Simonsohn et al., 2014, 2015; Head et al., 2015). This assumption has been justified through analytical and numerical examples, which rely on specific choices of tests and distributions of true effects being tested (e.g., Hung et al., 1997; Simonsohn et al., 2014; Ulrich and Miller, 2018). However, such analyses are not sufficient for guaranteeing size control of statistical tests for p-hacking since the true effect distribution is never known. Instead, what is required for size control in a wide range of applications is a characterization of the shape of the p-curve for general tests and effect distributions.

2.1 Setup

Consider a test statistic T that is distributed according to a distribution with cumulative distribution function (CDF) F_h, where h indexes parameters of either the exact or asymptotic distribution of the test. We assume that the parameters h only contain the parameters of interest. This is suitable for settings with large enough samples and asymptotically pivotal test statistics, which are prevalent in applied research.

Suppose researchers are testing the hypothesis

$$H_0 : h \in \mathcal{H}_0 \qquad \text{against} \qquad H_1 : h \in \mathcal{H}_1, \tag{1}$$

where $\mathcal{H}_0 \cap \mathcal{H}_1 = \emptyset$. Let $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$. Denote as F the CDF of the chosen null distribution from which critical values are determined. We assume that the test rejects for large values of the test statistic and denote the critical value for a level p test as $cv(p)$. We will focus on settings with a continuous and strictly increasing F (see Assumption 1 below) and set $cv(p) = F^{-1}(1 - p)$. For any h, we denote by $\beta(p, h) = \Pr(T > cv(p) \mid h) = 1 - F_h(cv(p))$ the rejection rate of a level p test with parameters h. For $h \in \mathcal{H}_1$, this is the power of the test, and we refer to $\beta(p, h)$ as the *power function*.

For the remainder of the paper, we focus on settings where the tests generating the p-values satisfy Assumption 1. This allows us to work with a well-defined density function and provide general results.

Assumption 1 (Regularity). *F and F_h are twice continuously differentiable with uniformly bounded first and second derivatives f, f', f_h and f'_h. $f(x) > 0$ for all $x \in \{cv(p) : p \in (0, 1)\}$. For $h \in \mathcal{H}$, $supp(f) = supp(f_h)$.*[2]

Assumption 1 holds for many tests with parametric F and F_h, including t-tests and Wald-tests. A necessary condition for Assumption 1 is the absolute continuity of F and F_h. This is not too restrictive since, in many cases, F and F_h are the asymptotic distributions of test statistics, which typically satisfy this condition. Further, in cases where the test statistics have a discrete distribution, size does not typically equal level, which could lead to p-curves that violate non-increasingness.

Consider the distribution of the p-values across studies, where we compute p-values from a distribution of T given values of h, which themselves are drawn from a probability distribution Π. We refer to Π as the *distribution of true effects*. The CDF of the p-values is

$$G(p) = \int_{\mathcal{H}} \Pr\left(T > cv(p) \mid h\right) d\Pi(h) = \int_{\mathcal{H}} \beta\left(p, h\right) d\Pi(h). \tag{2}$$

Under Assumption 1, define the p-curve as follows.

Definition 1 (P-curve). *The density of the p-values, the p-curve, is defined as*

$$g(p) := \int_{\mathcal{H}} \frac{\partial \beta\left(p, h\right)}{\partial p} d\Pi(h).$$

In Section 2.2, we analyze the shape of g for general tests and distributions Π.

2.2 Properties of p-curves based on general tests

Here we derive conditions under which the p-curve is non-increasing in the absence of p-hacking for any distribution of true effects. We show that this property holds for most but not all popular statistical tests.

Under Assumption 1, the curvature of the p-curve follows from

$$g'(p) := \frac{dg(p)}{dp} = \int_{\mathcal{H}} \frac{\partial^2 \beta\left(p, h\right)}{\partial p^2} d\Pi(h).$$

The sign of $g'(p)$ is determined by the second derivative of the rejection probability, $\partial^2 \beta\left(p, h\right) / \partial p^2$. As we will show in the proof of Theorem 1 below, the following condition implies that $\partial^2 \beta\left(p, h\right) / \partial p^2$ is non-positive for all $h \in \mathcal{H}$.

---

[2] For a function φ, we define $supp(\varphi)$ to be the closure of $\{x : \varphi(x) \neq 0\}$.

Assumption 2 (Sufficient condition). *For all $(x, h) \in \{cv(p) : p \in (0, 1)\} \times \mathcal{H}$,*

$$f_h'(x)f(x) \geq f'(x)f_h(x).$$

Assumption 2 is a restriction on how the power function changes when the critical value changes, which is governed by the shape of the density. When $\mathcal{H}_0 = \{0\}$ and $F = F_0$ (as, for example, for one-sided t-tests), Assumption 2 is of the form of a monotone likelihood ratio property, which relates the shape of the density of T under the null to the shape of the density of T under alternative h. The next lemma shows that this condition holds for many popular tests. Let Φ denote the CDF of the standard normal distribution.

Lemma 1. *Assumption 2 holds when*

(i) *$F(x) = \Phi(x)$, $F_h(x) = \Phi(x - h)$, $\mathcal{H}_0 = \{0\}$, $\mathcal{H}_1 \subseteq (0, \infty)$ (e.g., similar one-sided t-test)*

(ii) *F is the CDF of a half-normal distribution with scale parameter 1, F_h is the CDF of a folded normal distribution with location parameter h and scale parameter 1, $\mathcal{H}_0 = \{0\}$, $\mathcal{H}_1 \subseteq \mathbb{R}\backslash\{0\}$ (e.g., two-sided t-test)*

(iii) *F is the CDF of a χ^2 distribution with degrees of freedom $d > 0$, F_h is the CDF of a noncentral χ^2 distribution with degrees of freedom $d > 0$ and noncentrality parameter h, $\mathcal{H}_0 = \{0\}$, $\mathcal{H}_1 \subseteq (0, \infty)$ (e.g., Wald test[3])*

The following theorem shows that the p-curve is non-increasing and continuously differentiable under the maintained assumptions for any distribution of true effects.

Theorem 1 (Testable restrictions for general tests). *Under Assumptions 1–2, g is continuously differentiable and $g'(p) \leq 0$ for $p \in (0, 1)$.*

The result in Theorem 1 holds for many commonly-used statistical tests such that, in many empirically relevant settings, the p-curve will be non-increasing in the absence of p-hacking. To our knowledge, Theorem 1 provides the first general formal

---

[3]For instance, let $\sqrt{N}(\hat{\theta} - \theta) \overset{a}{\sim} \mathcal{N}(0, V)$, where $\hat{\theta}$ is an estimator of θ based on N observations and $V \in \mathbb{R}^{\dim(\theta) \times \dim(\theta)}$ is known (or can be consistently estimated). Consider the problem of testing $H_0 : R\theta = r$ against $H_1 : R\theta \neq r$, where $R \in \mathbb{R}^{q \times \dim(\theta)}$, $r \in \mathbb{R}^q$, and $rank(R) = q$. Set $T = N(R\hat{\theta} - r)'(RVR')^{-1}(R\hat{\theta} - r)$. This fits our framework with $d = q$ and $h := \lambda'(RVR')^{-1}\lambda$, where $\lambda := \sqrt{N}(R\theta - r)$.

justification for the existing tests for p-hacking that exploit non-increasingness of the p-curve. Theorem 1 further motivates the use of density discontinuity tests as an alternative to tests based on non-increasingness of the p-curve.

The results can be extended to settings with nuisance parameters. In such settings, h contains both the parameters of interest, h_1, as well as additional nuisance parameters, h_2, such that $h = (h_1, h_2)$. Let $\mathcal{H}^1$ and $\mathcal{H}^2$ denote the supports of h_1 and h_2. Allow the null distribution to depend on h_2 with CDF F_{h_2}. The CDF of p-values becomes

$$G(p) = \int_{\mathcal{H}^1 \times \mathcal{H}^2} \beta(p, h_1, h_2) \, d\Pi(h_1, h_2),$$

where $\beta(p, h_1, h_2) = 1 - F_h(cv_{h_2}(p))$ and $cv_{h_2}(p) = F_{h_2}^{-1}(1 - p)$. The results of Theorem 1 extend to the p-curve generated from this distribution after changing the notation to include the dependence on h_2. For $h_2 \in \mathcal{H}^2$, $F_{h_2}, f_{h_2}, f'_{h_2}$ have the same properties as F, f, f' in Assumption 1, and the assumptions on F_h, f_h, f'_h hold for $h = (h_1, h_2)$. Assumption 2 becomes $f'_h(cv_{h_2}(p))f_{h_2}(cv_{h_2}(p)) \geq f'_{h_2}(cv_{h_2}(p))f_h(cv_{h_2}(p))$ for $(h_1, h_2) \in \mathcal{H}^1 \times \mathcal{H}^2$. The proof then follows directly from that of Theorem 1.

In applications, often only a part of the p-curve is examined. The p-curve over subintervals $\mathcal{I} \subset (0, 1)$ is given by $g_{\mathcal{I}}(p) = g(p) / \int_{\mathcal{I}} g(p) dp$ for $p \in \mathcal{I}$. Therefore, the results extend directly to this situation. Moreover, the p-curve constructed from a finite aggregation of different tests satisfying the assumptions of Theorem 1 is continuously differentiable and non-increasing.

The assumptions of Theorem 1 directly suggest p-curves for which the results of Theorem 1 fail. For example, when the tests are non-similar, the p-curve can be non-monotonic in the absence of p-hacking, which arises through a violation of Assumption 2. To illustrate, consider testing $H_0 : h \leq 0$ against $H_1 : h > 0$ using a (non-similar) one-sided t-test, where f is the density of the $\mathcal{N}(0, 1)$ distribution and f_h is the density of the $\mathcal{N}(h, 1)$ distribution. It follows that $f'(x)/f(x) = -x$ and $f'_h(x)/f_h(x) = -(x - h)$, such that Assumption 2 holds when $h \geq 0$ but is violated when $h < 0$. Thus, when the weight in Π on $h < 0$ is large enough, the p-curve can be non-monotonic or increasing. For example, suppose that Π is a normal distribution with mean μ and variance 1, which places some mass on $h < 0$, mixing increasing and decreasing p-curves. Figure 1 shows that the resulting p-curve is non-increasing when $\mu = 0$ and non-monotonic when $\mu = -2.5$.
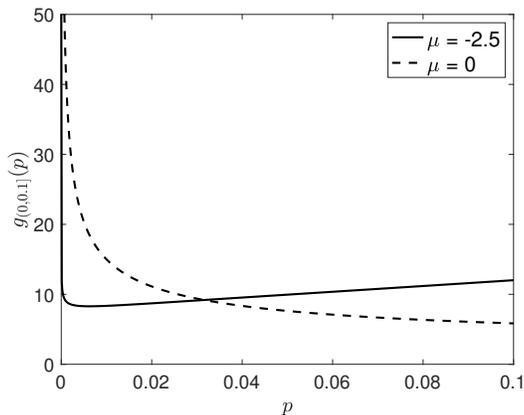
Figure 1. P-curves based on non-similar one-sided t-tests on $(0, 0.1]$. The distribution of true effects Π is a normal distribution with mean μ and variance 1.

3   The p-curve based on t-tests

We now show that for the leading case where p-curves are generated from t-tests with exact or asymptotic normal distributions, there are additional previously unknown testable restrictions. These restrictions allow us to develop more powerful statistical tests for p-hacking (see Section 4.3). In particular, these tests have power in situations where p-hacking does not lead to a violation of non-increasingness.

Consider first the problem of testing a one-sided hypothesis

$$H_0 : h = 0 \qquad \text{against} \qquad H_1 : h > 0, \tag{3}$$

where h is a scalar, $\mathcal{H}_0 = \{0\}$, and $\mathcal{H}_1 = (0, \infty)$. We assume that $T \sim \mathcal{N}(h, 1)$. This holds when using one-sided t-tests to test a hypothesis concerning a scalar parameter θ: $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Let $\sqrt{N}\left(\hat{\theta} - \theta\right) \sim \mathcal{N}\left(0, \sigma^2\right)$, where $\hat{\theta}$ is an estimator of θ based on N observations and σ^2 is assumed to be known. Denote the usual t-statistic as $\hat{t}$ and set $T = \hat{t}$. Defining $h := \sqrt{N}\left((\theta - \theta_0)/\sigma\right)$ this fits (3). More generally, testing problems with limiting normal experiments employed to test hypotheses of the form (3) are common in empirical work (e.g., a one-sided test of a regression parameter using normal critical values).

The chosen null distribution is the standard normal distribution, $F = \Phi$. A level p test rejects the null hypothesis when T is larger than $cv_1(p) := \Phi^{-1}(1 - p)$. Note that $cv_1(p) \geq 0$ for $p \in (0, 1/2]$. Then $\beta(p, h) = 1 - \Phi(cv_1(p) - h)$ and the CDF of

p-values is

$$G_1(p) = 1 - \int_{[0,\infty)} \Phi\left(cv_1(p) - h\right) d\Pi(h). \tag{4}$$

We also consider the two-sided version of this test. Here the hypothesis is

$$H_0 : h = 0 \qquad \text{against} \qquad H_1 : h \neq 0 \tag{5}$$

with $\mathcal{H}_0 = \{0\}$ and $\mathcal{H}_1 = \mathbb{R}\backslash\{0\}$. The two-sided test statistic T is assumed to have a folded normal distribution. This holds when using a two-sided t-test with $T = |\hat{t}|$ for testing a two-sided hypothesis about $\theta : H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. More generally, testing problems with limiting normal experiments employed to test hypotheses of the form (5) are also common in empirical work.

The chosen null distribution is the half normal distribution with scale parameter 1. A level p test rejects the null hypothesis when T is larger than $cv_2(p) := \Phi^{-1}\left(1 - \frac{p}{2}\right)$. The CDF of the p-values is

$$G_2(p) = 2 - \int_{\mathbb{R}} \left[\Phi\left(cv_2(p) - h\right) + \Phi\left(cv_2(p) + h\right)\right] d\Pi(h). \tag{6}$$

In addition to the results of Section 2.2, previously unknown testable restrictions for p-curves based on t-tests follow from the shape of the power functions for these tests. These additional restrictions enable us to better pin down the space of potential p-curves when there is no p-hacking, allowing us to construct more powerful statistical tests for p-hacking. They also enable distinguishing non-increasing p-curves, which can arise from certain types of p-hacking, from curves where there is no p-hacking.

The p-curve based on one-sided t-tests testing hypothesis (4) is

$$g_1(p) = \int_{[0,\infty)} \exp\left(hcv_1(p) - \frac{h^2}{2}\right) d\Pi(h). \tag{7}$$

For two-sided t-tests testing hypothesis (6), the p-curve is

$$g_2(p) = \int_{\mathbb{R}} \frac{1}{2}\left[\exp\left(hcv_2(p) - \frac{h^2}{2}\right) + \exp\left(-hcv_2(p) - \frac{h^2}{2}\right)\right] d\Pi(h). \tag{8}$$

Our next theorem shows that the p-curves (7) and (8) are completely monotone. A function ξ is completely monotone on an interval $\mathcal{I}$ if $0 \leq (-1)^k \xi^{(k)}(x)$ for every $x \in \mathcal{I}$ and all $k = 0, 1, 2, \ldots$, where $\xi^{(k)}$ is the kth derivative of ξ.

Theorem 2 (Complete monotonicity). *(i) The p-curve g_1 is completely monotone on $(0, 1/2]$. (ii) The p-curve g_2 is completely monotone on $(0, 1)$.*

Complete monotonicity yields additional restrictions that can be exploited to improve the power of statistical tests for p-hacking. Whilst available for one- and two-sided t-tests, not all tests yield completely monotonic p-curves. For example, a direct calculation shows that complete monotonicity may fail for tests based on χ^2 distributions with more than two degrees of freedom (e.g., Wald tests).

The next theorem presents additional testable restrictions in the form of upper bounds on the p-curves and their derivatives.

Theorem 3 (Upper bounds)**.**

(i) *The p-curves g_1 and g_2 are bounded from above:*

$$g_1(p) \leq 1_{\{p \leq 1/2\}} \exp\left(\frac{cv_1(p)^2}{2}\right) + 1_{\{p > 1/2\}} =: \mathcal{B}_1^{(0)}(p), \tag{9}$$

$$g_2(p) \leq 1_{\{p < 2(1-\Phi(1))\}} \tilde{\mathcal{B}}_2^{(0)} + 1_{\{p \geq 2(1-\Phi(1))\}} =: \mathcal{B}_2^{(0)}(p), \tag{10}$$

where

$$
\begin{aligned}
\tilde{\mathcal{B}}_2^{(0)}(p) &:= \frac{1}{2}\left[\exp\left(h^*(p)cv_2(p) - \frac{h^*(p)^2}{2}\right) + \exp\left(-h^*(p)cv_2(p) - \frac{h^*(p)^2}{2}\right)\right] \\
&\leq \exp\left(\frac{cv_2(p)^2}{2}\right),
\end{aligned}
$$

and $h^*(p)$ is the non-zero solution to

$$\varphi(cv_2(p), h) := (cv_2(p) - h)\exp(cv_2(p)h) - (cv_2(p) + h)\exp(-cv_2(p)h) = 0.$$

(ii) *The derivatives of g_1 and g_2 are bounded from above. For $s = 1, 2$ and $k = 1, 2, 3, \ldots$, then $(-1)^k g_s^{(k)}(p) \leq \mathcal{B}_s^{(k)}(p)$, where $\mathcal{B}_s^{(k)}$ is defined in Appendix B.3.*

As with the results in Theorem 2, the results in Theorem 3 yield additional restrictions, allowing more powerful tests for p-hacking.[4] The bounds in Theorem 3 do not only rule out large humps around significance cutoffs such as 0.01, 0.05, and 0.1 but also restrict the magnitude of the p-curves near zero. For the two-sided test, tests for p-hacking can be either constructed using the sharper (but not explicit) bound $\tilde{\mathcal{B}}_2^{(0)}(p)$ or the simpler explicit bound $\exp\left(\frac{cv_2(p)^2}{2}\right)$.

---

[4]One can use similar arguments as in Theorem 3 to derive bounds for p-curves based on other specific tests such as Wald tests.

The bounds of Theorem 3 are particularly useful when p-hacking fails to induce an increasing p-curve, a situation where tests based on non-increasingness of the p-curve have no power. Intuitively we might suspect this happens when all researchers p-hack but this simply shifts mass of the p-curve to the left, rather than inducing humps. A concrete example is when researchers run a finite number of $M > 1$ independent analyses and report the smallest p-value, for example, when engaging in specification search across independent subsamples or data sets. The resulting p-curve under p-hacking is $g^p(p; M) = M(1 - G^{np}(p))^{M-1}g^{np}(p)$, where G^{np} and g^{np} are the CDF and density of p-values in the absence of p-hacking.[5] Note that g^p is non-increasing (completely monotone) whenever g^{np} is non-increasing (completely monotone).[6] Thus, g^p will not violate the testable implications of Theorems 1–2, so tests based on these restrictions do not have power. However, g^p can violate the bounds in Theorem 3 whenever $M(1 - G^{np}(p))^{M-1} > 1$. For example, consider the one-sided case and let Π be a half-normal distribution with scale parameter 1. Figure 2 shows that g^p violates the upper bound in Theorem 3 to an extent that depends on M.
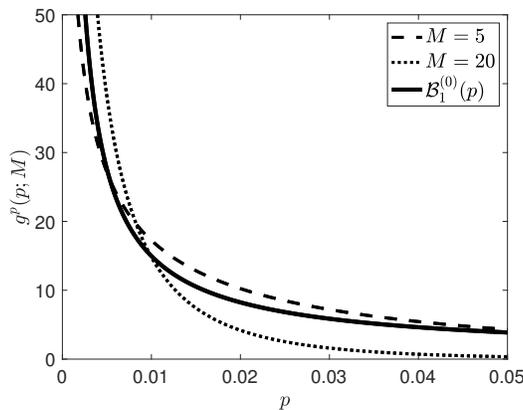
Figure 2. Comparison of the p-curve from specification search based on one-sided t-tests and the upper bound in Equation (9).

Upper bounds also help with testing for p-hacking with non-similar tests. In Section 2.2, we show that non-increasingness may fail for non-similar one-sided t-tests,

---

[5]This generalizes the example in Ulrich and Miller (2015), who studied the special case where all null hypotheses are true such that $G(p) = p$.

[6]Since the products of completely monotone functions are completely monotone, complete monotonicity of $g^p(p; M)$ follows from complete monotonicity of $1 - G^{np}(p)$ and $g^{np}(p)$.

in which case tests of p-hacking based on non-increasingness may well reject because of non-similarity rather than p-hacking. Since upper bounds can also be derived for non-similar tests, we can still use bounds on the p-curve and its derivatives to test for p-hacking.[7]

Finally, the characterizations in Theorems 2–3 imply related characterizations of p-curves over subintervals $\mathcal{I} \subset (0,1)$, $g_{s,\mathcal{I}}(p) = g_s(p)/\int_{\mathcal{I}} g_s(p)dp$. In particular, complete monotonicity of g_s implies the complete monotonicity of $g_{s,\mathcal{I}}$, because the sign of $g_{s,\mathcal{I}}^{(k)}$ equals the sign of $g_s^{(k)}$ for $k = 0, 1, 2 \dots$. Moreover, (conservative) upper bounds on $g_{s,\mathcal{I}}(p)$ for $\mathcal{I} = (0, \alpha]$ are given by the upper bounds in Theorem 3, re-scaled by α since $G_s(\alpha) \geq \alpha$ for $s = 1, 2$.

4   Statistical tests for p-hacking

Here we consider tests for p-hacking based on a sample of n p-values. We consider three types of tests that differ with respect to the specification of the null hypothesis (the null space of p-curves). As a result, the different tests will differ with respect to the violations of the null of no p-hacking that they are able to detect.

In the absence of publication bias, our tests are tests for p-hacking; when there is also publication bias, they are joint tests for p-hacking and publication bias in general.

4.1   Tests for non-increasingness of the p-curve

Theorem 1 shows that, under general conditions, the p-curve is non-increasing. Consider the following testing problem

$$H_0 : g \text{ is non-increasing} \quad \text{against} \quad H_1 : g \text{ is not non-increasing.} \quad (11)$$

Popular tests based on hypothesis testing problem (11) include the Binomial test (e.g., Simonsohn et al., 2014; Head et al., 2015) and Fisher's test (Simonsohn et al., 2014). Here we describe two alternative and more powerful tests.

Histogram-based tests. Let $0 = x_0 < x_1 < \cdots < x_J = 1$ be an equidistant partition of the unit interval. Define the population proportions as $\pi_j := \int_{x_{j-1}}^{x_j} g(p)dp$, $j = 1, \dots, J$. When g is non-increasing, $\Delta_j := \pi_{j+1} - \pi_j$ is non-positive

---

[7]For instance, for $p \leq 1/2$, the upper bound on the p-curve for non-similar one-sided t-tests coincides with that in Part (i) of Theorem 3.

for all $j = 1, \ldots, J - 1$. Thus, the null hypothesis in testing problem (11) can be reformulated as $H_0 : \Delta_j \leq 0$ for all $j = 1, \ldots, J-1$. To test this hypothesis, we apply the conditional chi-squared test of Cox and Shi (2020). We describe the implementation of this test in Section 4.3 and Appendix A, where we propose more general tests that nest the histogram-based test for non-increasingness.

LCM test based on concavity of the CDF of p-values. Under the null hypothesis (11), the CDF of p-values is concave. This observation allows us to apply tests based on the least concave majorant (LCM) (e.g., Carolan and Tebbs, 2005; Beare and Moon, 2015; Fang, 2019). LCM-based tests assess concavity of the CDF based on the distance between the empirical CDF of p-values, $\hat{G}$, and its LCM, $\mathcal{M}\hat{G}$, where $\mathcal{M}$ is the LCM operator.[8] We consider the test statistic $T = \sqrt{n}\|\mathcal{M}\hat{G} - \hat{G}\|_\infty$. The uniform distribution is least favorable for LCM tests (e.g., Kulikov and Lopuhaä, 2008; Beare, 2021), in which case T converges weakly to $\|\mathcal{M}B - B\|_\infty$, where B is a standard Brownian Bridge on $[0, 1]$.

4.2 Tests for continuity

Theorem 1 shows that the p-curve is continuous in the absence of p-hacking. Tests for continuity of the p-curve at significance thresholds α such as $\alpha = 0.05$, thus, provide an alternative to the tests based on non-increasingness of the p-curve. Consider the following testing problem:

$$H_0 : \lim_{p\uparrow\alpha} g(p) = \lim_{p\downarrow\alpha} g(p) \qquad \text{against} \qquad H_1 : \lim_{p\uparrow\alpha} g(p) \neq \lim_{p\downarrow\alpha} g(p) \qquad (12)$$

Testing (12) requires estimating two densities at the boundary point α. Traditional kernel density estimators are not suitable for this task because they suffer from boundary bias (e.g., Karunamuni and Alberts, 2005). A popular approach to overcome this problem is to use local linear density estimators that rely on prebinning the data (e.g., McCrary, 2008). We apply the density discontinuity test of Cattaneo et al. (2020) with data-driven bandwidth selection (Cattaneo et al., 2021), which is based on boundary adaptive local polynomial density estimators and avoids prebinning.

---

[8]For a function f, the LCM operator is defined as $\mathcal{M}f = \inf\{g : g \text{ is concave and } f \leq g\}$ (e.g., Beare and Moon, 2015, Definition 2.1).

4.3 Tests for K-monotonicity and upper bounds

Theorem 2 shows that p-curves based on t-tests are completely monotone, and Theorem 3 establishes upper bounds on the p-curves and their derivatives. Here we develop tests based on these testable restrictions.

We say a function ξ is K-monotone on some interval $\mathcal{I}$ if $0 \leq (-1)^k \xi^{(k)}(x)$ for every $x \in \mathcal{I}$ and all $k = 0, 1, \ldots, K$, where $\xi^{(k)}$ is the kth derivative of ξ. By definition, a completely monotone function is K-monotone. Consider the null hypothesis

$$H_0 : g_s \text{ is } K\text{-monotone and } (-1)^k g_s^{(k)} \leq \mathcal{B}_s^{(k)}, \text{ for } k = 0, 1, \ldots, K, \qquad (13)$$

where $s = 1$ for one-sided t-tests, $s = 2$ for two-sided t-tests, and $\mathcal{B}_s^{(k)}$ is defined in Theorem 3. Hypothesis (13) implies restrictions on the population proportions $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_J)'$, which can be expressed as $H_0 : A\boldsymbol{\pi}_{-J} \leq b$, where $\boldsymbol{\pi}_{-J} := (\pi_1, \ldots, \pi_{J-1})'$.[9] The matrix A and vector b are defined in Appendix A.2.[10]

We estimate $\boldsymbol{\pi}_{-J}$ using the sample proportions $\hat{\boldsymbol{\pi}}_{-J}$.[11] This estimator is $\sqrt{n}$-consistent and asymptotically normal with mean $\boldsymbol{\pi}_{-J}$ and non-singular (if all proportions are positive) covariance matrix $\Omega = \mathrm{diag}\{\pi_1, \ldots, \pi_{J-1}\} - \boldsymbol{\pi}_{-J}\boldsymbol{\pi}'_{-J}$. Following Cox and Shi (2020), we test the null by comparing $T = \inf_{q: Aq \leq b} n(\hat{\boldsymbol{\pi}}_{-J} - q)'\hat{\Omega}^{-1}(\hat{\boldsymbol{\pi}}_{-J} - q)$ to the critical value from a χ^2 distribution with $rank(\hat{A})$ degrees of freedom, where $\hat{A}$ is the matrix formed by the rows of A corresponding to active inequalities.

5 Empirical applications

The analyses were done using R (R Core Team, 2020) and Stata (StataCorp., 2019).

5.1 P-hacking in economics journals

Here we reanalyze the data collected by Brodeur et al. (2016b), which contain information about 50,078 t-tests from 641 papers published in the AER, QJE, and JPE

---

[9]The upper bounds on $\boldsymbol{\pi}$ implied by hypothesis (13) are not sharp in general. Sharp bounds can be obtained by directly extremizing the proportions and their differences; see Appendix A.1.

[10]We use $\boldsymbol{\pi}_{-J}$ because the variance matrix of the estimator of $\boldsymbol{\pi}$ is singular by construction and we want to express the left-hand side of our moment inequalities as a combination of "core" moments.

[11]Given a sample of n p-values, $\{P_i\}_{i=1}^n$, the sample proportions are defined as $\hat{\pi}_i = \frac{1}{n} \sum_{i=1}^n 1\{x_{i-1} < P_i \leq x_i\}, i = 1, \ldots, J$.

2005–2011 (Brodeur et al., 2016a). We convert t-statistics into p-values associated with two-sided t-tests based on the standard normal distribution.[12] After excluding observations with missing information, there are 49,838 tests from 640 papers.

Because the p-values may be correlated within papers, we use cluster-robust estimators of the variance of the sample proportions for the Cox and Shi (2020) tests. In addition, we apply all tests to random subsamples with one p-value per paper, allowing us to use exact tests in the presence of within-paper correlation. To test for p-hacking, we focus on p-values smaller than 0.15. We consider a Binomial test on $[0.04, 0.05]$, Fisher's test, a histogram-based test for non-increasingness (CS1), a histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives (CS2B), the LCM test, and a density discontinuity test at 0.05.[13]

Figure 3 shows the results before and after de-rounding and based on the full sample and random subsamples. There is a large number of very small p-values, which is sometimes interpreted as indicative of evidential value (e.g., Simonsohn et al. (2014); in our notation, this is a large mass of Π away from zero). The data exhibit a noticeable mass point at $\hat{t} = 2$ (there are 427 such observations), which translates into a mass point in the p-curve at $p = 0.046$.[14] To analyze the impact of rounding, we also apply the tests to the de-rounded data provided by Brodeur et al. (2016b).[15]

In what follows, we say that a test rejects the null of no p-hacking if its p-value is smaller than 0.1. Based on the original raw (rounded) data on all p-values, all tests reject the null except Fisher's test and the density discontinuity test. There are no rejections based on the random subsample, suggesting that the tests may be underpowered in small samples.

We find different results based on the de-rounded data.[16] There are no rejections based on the full sample of p-values. This finding suggests that the rejections based

---

[12]The original data contain p-values for less than 10% of observations. Where available, we work with the reported p-values.

[13]For the Binomial test, we split $[0.04, 0.05]$ into two subintervals $[0.04, 0.045]$ and $(0.045, 0.05]$. Under the null of no p-hacking, the fraction of p-values in $(0.045, 0.05]$ should be smaller than or equal to 0.5, which we assess using an exact Binomial test. For CS1 and CS2B, we use 30 bins when testing based on all p-values and 15 bins when testing based on random subsamples of p-values.

[14]This mass point could be due to low precision reporting (Brodeur et al., 2016b), but also due to p-hacking, publication bias, or a combination thereof.

[15]The de-rounded data were constructed by randomly redrawing estimates and standard errors; see Section II in Brodeur et al. (2016b) for a detailed description.

[16]Note that the (sub)sample sizes for the rounded and de-rounded data differ due to de-rounding.
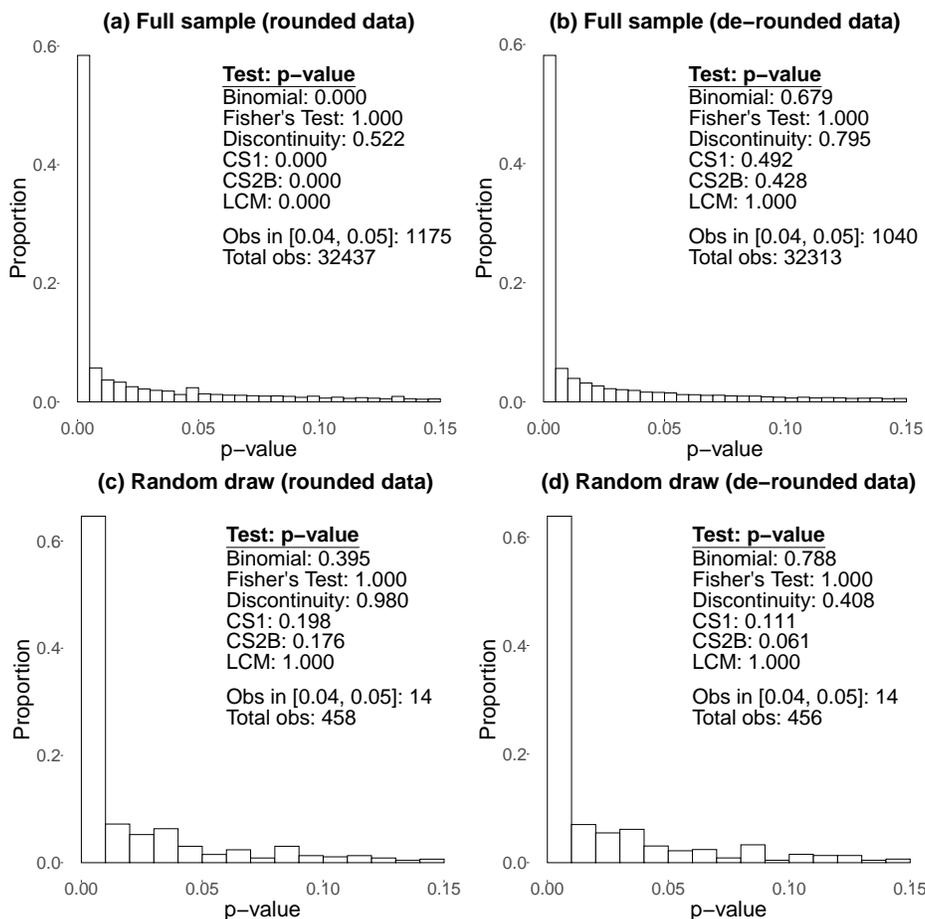
Figure 3. *P*-curves and *p*-values from testing for *p*-hacking. The tests for *p*-hacking are described in Section 4. Data: Brodeur et al. (2016a).

on the raw data are mainly due to the mass point just below 0.05 and shows that de-rounding may substantially affect empirical conclusions.

Based on the random subsample of de-rounded *p*-values, only the CS2B test rejects the null of no *p*-hacking. The CS1 test comes close to rejecting ($p = 0.11$). These two tests yield the smallest *p*-values across all four samples.

5.2 P-hacking across different disciplines

Here we reanalyze the data collected by Head et al. (2015), which contain *p*-values obtained from text-mining open access papers in the PubMed database (Head et al., 2016). There are *p*-values from 21 different disciplines. We focus on biology, chemistry, education, engineering, medical and health sciences, and psychology and cognitive

science. The data contain p-values from the abstracts and the results sections in the main text. We use p-values from the results sections, allowing us to work with larger samples and present results for p-values smaller than 0.15.

Since the data do not only contain t-tests, we consider tests based on non-increasingness and continuity of the p-curve (Theorem 1): a Binomial test on $[0.04, 0.05]$, Fisher's test, a histogram-based test for non-increasingness (CS1), the LCM test, and a density discontinuity test at 0.05.[17] To account for within-paper dependence of p-values, we use a cluster-robust variance estimator for the CS1 test, and also present results based on random subsamples with one p-value per paper.

(a) Full sample (rounded data)

Test: p–value
Binomial: 1.000
Fisher's Test: 1.000
Discontinuity: 0.000
CS1: 0.000
LCM: 0.000

Obs in [0.04, 0.05]: 38462
Total obs: 352817

(b) Full sample (de–rounded data)

Test: p–value
Binomial: 1.000
Fisher's Test: 1.000
Discontinuity: 0.162
CS1: 0.000
LCM: 0.065

Obs in [0.04, 0.05]: 28318
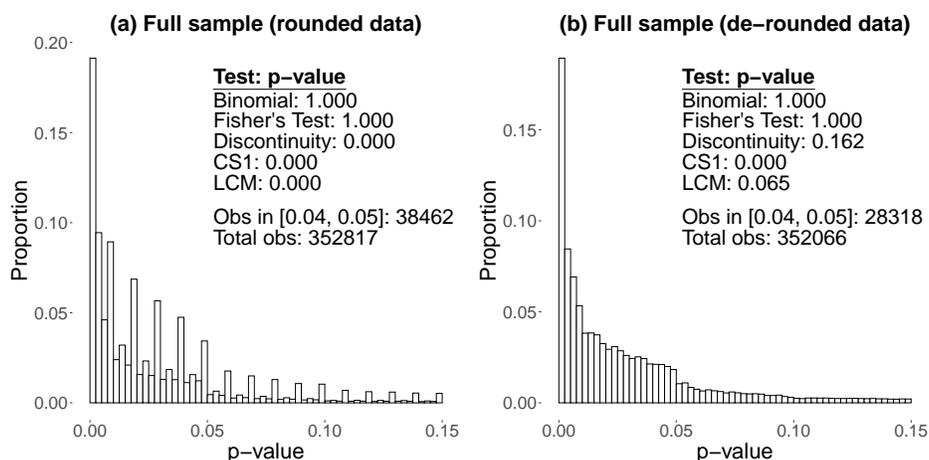Total obs: 352066

Figure 4. P-curves and p-values from testing for p-hacking for medical and health sciences. The tests for p-hacking are described in Section 4. Data: Head et al. (2016).

The left panel of Figure 4 shows a histogram of the raw data on all p-values for the medical and health sciences (the largest subsample). A substantial fraction of p-values is rounded to two decimal places, which results in sizable mass points at $0.01, 0.02, \ldots, 0.15$. Rounding makes the p-curve non-monotonic and discontinuous even in the absence of p-hacking and, thus, invalidates the testable restrictions in Theorem 1. Therefore, we also show results based on de-rounded data.[18] In an earlier version of this paper (Elliott et al., 2020), we show that de-rounding restores

---

[17]For CS1, we use 60 bins (all data) and 30 bins (random subsamples) for biological and medical and health sciences given the large sample sizes, and 30 and 15 bins for the other disciplines.

[18]We de-round the data as follows. To each observed p-value rounded up to the kth decimal point we add a random number generated from the uniform distribution supported on the interval $[\underline{u}, 0.5] \cdot 10^{-k}$, where $\underline{u} = 0$ for zero p-values and $\underline{u} = -0.5$ for non-zero p-values.

16

the non-increasingness but not the continuity of the p-curve. The right panel of Figure 4 shows the impact of de-rounding on the shape of the p-curve. We note that density discontinuity tests are poorly suited here because rounding induces substantial discontinuities, which remain even after de-rounding. This means that rejections of the null can be either due to rounding or due to p-hacking.

In what follows, define a rejection of the null of no p-hacking for p-values smaller than 0.1. Table I presents the results for the full sample of p-values. For the original (rounded) data, the CS1 and the LCM test reject the null for all disciplines. De-rounding leads to fewer rejections. The CS1 test only rejects for biological sciences, engineering, and medical and health sciences; the LCM test rejects for medical and health sciences. This shows that rounding and de-rounding can substantially affect empirical results. The Binomial and Fisher's test do not reject the null for any discipline, which demonstrates the importance of using our more powerful tests.

TABLE I. Testing results based on full sample of p-values

Test	Discipline					
	Biological sciences	Chemical sciences	Education	Engineering	Medical and health sciences	Psychology and cognitive sciences
	Rounded					
Binomial	1.000	0.342	0.975	0.999	1.000	1.000
Fisher's Test	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	0.000	0.000	0.159	0.000	0.000	0.172
CS1	0.000	0.000	0.000	0.000	0.000	0.000
LCM	0.000	0.000	0.000	0.000	0.000	0.000
Obs in [0.04, 0.05]	7692	296	220	396	38462	1621
Total obs	74746	2631	1993	3262	352817	15189
	De-rounded					
Binomial	0.993	0.133	0.467	0.975	1.000	0.811
Fisher's Test	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	0.005	0.117	0.245	0.849	0.162	0.406
CS1	0.028	0.530	0.884	0.084	0.000	0.836
LCM	0.936	1.000	1.000	1.000	0.065	0.653
Obs in [0.04, 0.05]	5720	234	144	250	28318	1161
Total obs	74550	2628	1988	3258	352066	15130

Notes: Table reports p-values from applying different tests for p-hacking based on the full sample of p-values for rounded and de-rounded data. The tests for p-hacking are described in Section 4. Data: Head et al. (2016).

Table II shows the results based on random samples with one p-value per pa-

per. We find that the CS1 test (biological sciences, engineering, medical and health sciences) and the LCM test (all disciplines except chemical sciences) reject the null based on the rounded data. None of the tests based on non-increasingness rejects the null based on the de-rounded data. A comparison to the results based on all p-values shows that the sample sizes required for detecting p-hacking may be quite large.

TABLE II. Testing results based on random subsamples of one p-value per paper

Test	Discipline					
	Biological sciences	Chemical sciences	Education	Engineering	Medical and health sciences	Psychology and cognitive sciences
			Rounded			
Binomial	0.510	0.157	0.439	0.904	1.000	0.670
Fisher's Test	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	0.113	0.083	0.103	0.000	0.000	0.157
CS1	0.000	0.637	0.232	0.078	0.000	0.734
LCM	0.000	0.265	0.035	0.002	0.000	0.000
Obs in [0.04, 0.05]	1482	63	42	85	6270	185
Total obs	13829	482	366	619	56892	1730
			De-rounded			
Binomial	0.178	0.116	0.286	0.712	0.976	0.465
Fisher's Test	1.000	1.000	1.000	1.000	1.000	1.000
Discontinuity	0.571	0.085	0.997	0.287	0.557	0.637
CS1	0.992	0.688	0.481	0.731	0.872	0.747
LCM	1.000	1.000	1.000	0.999	0.846	1.000
Obs in [0.04, 0.05]	1053	45	28	51	4536	128
Total obs	13788	482	365	619	56753	1716

Notes: Table reports p-values from applying different tests for p-hacking based on random subsamples of p-values for rounded and de-rounded data. The tests for p-hacking are described in Section 4. Data: Head et al. (2016).

Finally, the density discontinuity test rejects for at least three disciplines based on the full sample and the random subsamples. After de-rounding, it only rejects for biological sciences (full sample) and chemical sciences (random subsample). These rejections are expected because of the prevalence of rounding-induced discontinuities.

6   Conclusion

We provide theoretical foundations for testing for p-hacking based on the distribution of p-values across scientific studies. We establish general results on the p-curve,

providing conditions under which a null set of p-curves can be shown to be non-increasing. For p-values based on t-tests, we derive previously unknown additional restrictions on the p-curve when there is no p-hacking. These restrictions lead to the suggestion of more powerful tests that can be used to test the absence of p-hacking. A reanalysis of two datasets from the literature shows that the new tests based on additional restrictions are useful in testing for p-hacking.

References

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94.

Beare, B. K. (2021). Least favorability of the uniform distribution for tests of the concavity of a distribution function. *Stat*, page e376. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.376.

Beare, B. K. and Moon, J.-M. (2015). Nonparametric tests of density ratio ordering. *Econometric Theory*, 31(3):471–492.

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–60.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016a). Replication data for: Star wars: The empirics strike back. Nashville, TN: American Economic Association [publisher], 2016. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12. https://www.openicpsr.org/openicpsr/project/113633/version/V1/view (last accessed 09/23/2020).

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016b). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Bruns, S. B., Asanov, I., Bode, R., Dunger, M., Funk, C., Hassan, S. M., Hauschildt, J., Heinisch, D., Kempa, K., König, J., Lips, J., Verbeck, M., Wolfschütz, E., and Buenstorf, G. (2019). Reporting errors and biases in published empirical findings: Evidence from innovation research. *Research Policy*, 48(9):103796.

Carolan, C. A. and Tebbs, J. M. (2005). Nonparametric tests for and against likelihood ratio ordering in the two-sample problem. *Biometrika*, 92(1):159–171.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Cattaneo, M. D., Jansson, M., and Ma, X. (2021). *rddensity: Manipulation Testing Based on Density Discontinuity*. R package version 2.2.

Christensen, G. and Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80.

Cox, G. and Shi, X. (2020). Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models. *arXiv:1907.06317v2*.

de Winter, J. C. and Dodou, D. (2015). A surge of *p*-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, 3:e733.

Elliott, G., Kudrin, N., and Wüthrich, K. (2020). Detecting p-hacking. *arXiv:1906.06711v3*.

Fang, Z. (2019). Refinements of the kiefer-wolfowitz theorem and a test of concavity. *Electron. J. Statist.*, 13(2):4596–4645.

Gerber, A. and Malhotra, N. (2008). Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quarterly Journal of Political Science*, 3(3):313–326.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2016). Data from: The extent and consequences of p-hacking in science. Dryad, Dataset. https://datadryad.org/resource/doi:10.5061/dryad.79d43 (last accessed 09/29/2020).

Hung, H. M. J., O'Neill, R. T., Bauer, P., and Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 53(1):11–22.

Karunamuni, R. and Alberts, T. (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191 – 212.

Kulikov, V. N. and Lopuhaä, H. P. (2008). Distribution of global measures of deviation between the empirical distribution function and its concave majorant. *Journal of Theoretical Probability*, 21(2):356–377.

Leggett, N. C., Thomas, N. A., Loetscher, T., and Nicholls, M. E. R. (2013). The life of p: "just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology*, 66(12):2303–2309.

Masicampo, E. J. and Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, 65(11):2271–2279.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, 144(6):1146–1152.

Snyder, C. and Zhuo, R. (2018). Sniff tests in economics: Aggregate distribution of their probability values and implications for publication bias. NBER WP 25058.

StataCorp. (2019). *Stata Statistical Software: Release 16*. College Station, TX.

Ulrich, R. and Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, 144:1137–1145.

Ulrich, R. and Miller, J. (2018). Some properties of p-curves, with an application to gradual publication bias. *Psychological Methods*, 23(3):546–560.

Vivalt, E. (2019). Specification searching and significance inflation across time, methods and disciplines. *Oxford Bulletin of Economics and Statistics*, 81(4):797–816.

A   Additional details Section 4.3

A.1   Bounds on proportions and their differences

The bounds on the proportions and their differences implied by hypothesis (13) are not sharp in general. Here we derive sharp bounds by directly extremizing the proportions and their differences.

For the one-sided t-tests, the population proportion, π_j, can be written as

$$
\begin{aligned}
\pi_j = \int_{x_{j-1}}^{x_j} g_1(p)dp &= \int_{x_{j-1}}^{x_j} \int_{[0,\infty)} e^{-h^2/2} e^{hcv_1(p)} d\Pi(h)dp \\
&= \int_{[0,\infty)} \left(\int_{x_{j-1}}^{x_j} e^{-h^2/2} e^{hcv_1(p)} dp \right) d\Pi(h) \\
&= \int_{[0,\infty)} \left(\int_{cv_1(x_j)}^{cv_1(x_{j-1})} \phi(t-h)dt \right) d\Pi(h) \\
&= \int_{[0,\infty)} \lambda_{1,j}(cv_1,h)d\Pi(h),
\end{aligned}
$$

where $\lambda_{1,j}(cv, h) := \Phi(cv(x_{j-1}) - h) - \Phi(cv(x_j) - h)$. For the two-sided t-tests, $\pi_j = \int_{x_{j-1}}^{x_j} g_2(p)dp = \int_{\mathbb{R}} \lambda_{2,j}(cv_2, h)d\Pi(h)$, where $\lambda_{2,j}(cv, h) := \lambda_{1,j}(cv, h) + \lambda_{1,j}(cv, -h)$.

Since $\lambda_{1,j}(cv_1, h)$, as a function of h, attains its maximum at $h_j^* = \frac{cv_1(x_{j-1}) + cv_1(x_j)}{2}$, for the one-sided t-tests $\pi_j \leq 2\Phi\left(\frac{cv_1(x_{j-1}) - cv_1(x_j)}{2}\right) - 1 := \vartheta_{1,j}^{(0)}$. In case of the two-sided t-tests, the bound, $\vartheta_{2,j}^{(0)} := \max_{h \in \mathbb{R}} \lambda_{2,j}(cv_2, h)$, can be calculated numerically.

For the bounds on the kth differences of π's, note that, for $j = 1, \ldots, J - k$, $\Delta_j^k = \sum_{i=0}^k (-1)^i \binom{k}{i} \pi_{k+j-i}$ and therefore

$$|\Delta_j^k| \leq \vartheta_{s,j}^{(k)} := \max_{h \in \mathcal{H}_{(s)}} \left\{ \sum_{i=0}^k (-1)^{i+k} \binom{k}{i} \lambda_{s,k+j-i}(cv_s, h) \right\}, \quad j = 1, \ldots, J - k,$$

where $\mathcal{H}_{(1)} = [0, \infty)$, $\mathcal{H}_{(2)} = \mathbb{R}$, and $s = 1$ and $s = 2$ for the one- and two-sided t-tests, respectively. These bounds can be computed numerically.

A.2   Null hypothesis

The null hypothesis formulated in terms of the proportions is

$$H_0: \ 0 \leq (-1)^k \Delta^k \leq \vartheta_s^{(k)}, \ \sum_{j=1}^J \pi_j = 1, \ \text{ for all } k = 0, \ldots, K, \tag{14}$$

where Δ^k is a $(J-k) \times 1$ vector of kth differences of π's, $\Delta^0 = \pi$, $\vartheta_s^{(k)} := (\vartheta_{s,1}^{(k)}, \ldots, \vartheta_{s,J-k}^{(k)})'$ is the vector of upper bounds on $|\Delta^k|$ (cf. Appendix A.1), $s = 1$ for one-sided tests, and $s = 2$ for two-sided tests. The inequalities in (14) are interpreted element-wise.

Let D_m be $(m - 1) \times m$ differencing matrix of the following form:

$$D_m := \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix}.$$

In addition, define the $J \times 1$ vector $e_J := (0, \ldots, 1)'$, $(J - 1) \times 1$ vector $i_{J-1} := (1, \ldots, 1)'$, and matrix $F := [-I_{J-1}, i_{J-1}]'$. Using this notation, we can write $(-1)^k \Delta^k = D^k \pi$, $k = 1, \ldots, K$, where $D^k := (-1)^k D_{J-k+1} \times \cdots \times D_J$. Note that the restrictions under the null are equivalent to $\mathcal{D}_K \pi \geq c$ and $\pi = e_J - F\pi_{-J}$, where $\mathcal{D}_K = [-1, 1]' \otimes [I_J, D^{1'}, \ldots, D^{K'}]'$ and $c = [\vartheta_s^{(0)'}, \ldots, \vartheta_s^{(K)'}, 0'_{(K+1)(J-K/2) \times 1}]'$. The symbol $\otimes$ denotes the Kronecker product. We can thus express the null hypothesis (14) as $H_0: \ A\pi_{-J} \leq b$, where $A := \mathcal{D}_K F$ and $b := \mathcal{D}_K e_J - c$.

When testing on a subinterval $(0, \alpha]$, the bounds need to be re-scaled. We use a consistent (under the null) estimator of $G(\alpha)$ to re-scale the bounds. In particular, we use bounds $\tilde{\vartheta}_{s,j}^{(k)} = \vartheta_{s,j}^{(k)}/\hat{G}(\alpha)$, where $\hat{G}(\alpha)$ is the fraction of p-values below α.

B   Proofs

B.1   Proof of Lemma 1

Note that for claim (i) $\{cv(p) : p \in (0, 1)\} = \mathbb{R}$ and for claims (ii) and (iii) $\{cv(p) : p \in (0, 1)\} = (0, \infty)$.

Claim (i): In this case $f(x) = \phi(x)$ and $f_h(x) = \phi(x - h)$. It follows that, for all $h \geq 0$, $f'_h(x)f(x) - f'(x)f_h(x) = h\phi(x)\phi(x - h) \geq 0$.

Claim (ii): In this case $f(x) = 2\phi(x)$ and $f_h(x) = \phi(x - h) + \phi(x + h)$, where $x \geq 0$. After taking derivatives and collecting terms we get

$$f'_h(x)f(x) - f'(x)f_h(x) = 2\phi(x)h(\phi(x - h) - \phi(x + h)) = 2\phi(x)\phi(x + h)h(e^{2xh} - 1) \geq 0,$$

because $h(e^{2xh} - 1) \geq 0$ for any h.

Claim (iii): In this case $f(x) := f(x; d) = \frac{1}{2^{d/2}\Gamma(d/2)}x^{d/2-1}e^{-x/2}$ and $f_h(x) = \sum_{j=0}^{\infty} \frac{e^{-h/2}(h/2)^j}{j!} f(x; d + 2j)$, where $x > 0$. Note that $f'(x; d) = f(x; d)\left((d - 2)x^{-1} - 1\right)/2$. After taking derivatives and collecting terms we get

$$
\begin{aligned}
f'_h(x)f(x) - f'(x)f_h(x) &= \sum_{j=0}^{\infty} \frac{e^{-h/2}(h/2)^j}{2j!} f(x; d + 2j)f(x; d)\left[((d + 2j - 2)x^{-1} - 1) - ((d - 2)x^{-1} - 1)\right] \\
&= \sum_{j=0}^{\infty} \frac{e^{-h/2}(h/2)^j}{j!} f(x; d + 2j)f(x; d)jx^{-1} \geq 0,
\end{aligned}
$$

since every term in the last sum is non-negative. □

B.2   Proof of Theorem 1

Recall that $\beta(p, h) = 1 - F_h(cv(p))$, where $cv(p) = F^{-1}(1 - p)$. Under Assumption 1,

$$
\begin{aligned}
\frac{\partial^2 \beta(p, h)}{\partial p^2} &= \frac{f'_h(cv(p))cv'(p)f(cv(p)) - f'(cv(p))cv'(p)f_h(cv(p))}{f(cv(p))^2} \\
&= \frac{cv'(p)}{f(cv(p))^2}\left[f'_h(cv(p))f(cv(p)) - f'(cv(p))f_h(cv(p))\right].
\end{aligned}
$$

Non-increasingness of g now follows by Assumption 2 and because $cv'(p)/f(cv(p))^2 \leq 0$. Continuous differentiability is implied by Assumption 1. □

B.3   Proofs of Theorems 2 and 3

Note that the p-curves for the one-sided and two-sided t-tests are given by

$$g_1(p) = \int_{[0,\infty)} \Psi(cv_1(p), h) \exp\{-h^2/2\} d\Pi(h), \tag{15}$$

$$g_2(p) = \frac{1}{2} \int_{\mathbb{R}} (\Psi(cv_2(p), h) + \Psi(cv_2(p), -h)) \exp\{-h^2/2\} d\Pi(h) \tag{16}$$

where $\Psi(x, y) := \exp\{xy\}$. We start by proving an auxiliary lemma about $\Psi(x, y)$.

Lemma 2. *For $k \geq 1$, the kth derivative of $\Psi(cv_s(p), h)$ is*

$$\Psi^{(k)}(cv_s(p), h) = (-1)^k \frac{h \sum_{j=0}^{k-1} A_j^k(cv_s(p))[cv_s(p) + h]^j}{s^k(\phi(cv_s(p)))^k} \Psi(cv_s(p), h),$$

where coefficients $A_j^k(cv_s(p))$ are polynomials in $cv_s(p)$ with non-negative coefficients and $s = 1$ for one-sided and $s = 2$ for two-sided t-tests.

Proof. By direct computation, the first derivative of $\Psi(cv_s(p), h)$ with respect to p is $\Psi^{(1)}(cv_s(p), h) = -\frac{h}{s\phi(cv_s(p))} \Psi(cv_s(p), h)$. We use induction to derive the kth derivative of $\Psi(cv_s(p), h)$. Suppose that for $k > 1$

$$\Psi^{(k)}(cv_s(p), h) = (-1)^k \frac{h \sum_{j=0}^{k-1} A_j^k(cv_s(p))[cv_s(p) + h]^j}{s^k(\phi(cv_s(p)))^k} \Psi(cv_s(p), h),$$

where coefficients $A_j^k(cv_s(p))$ are polynomials in $cv_s(p)$ with non-negative coefficients. Define $B_0^k = (k-1)cv_s(p)A_0^k(cv_s(p))$, $B_j^k = (k-1)cv_s(p)A_j^k(cv_s(p)) + A_{j-1}^k(cv_s(p))$ for $j = 1, \ldots, k-1$, and $B_k^k = A_{k-1}^k(cv_s(p))$; $C_j^k = \partial A_j^k(cv_s(p))/\partial cv_s(p) + (j+1)A_{j+1}^k(cv_s(p))$ for $j = 0, \ldots, k-2$, $C_{k-1}^k = \partial A_{k-1}^k(cv_s(p))/\partial cv_s(p)$, and $C_k^k = 0$. Now differentiate $\Psi^{(k)}(cv_s(p), h)$ with respect to p to get

$$
\begin{aligned}
\Psi^{(k+1)}(cv_s(p), h) &= (-1)^{k+1} \frac{h^2 \sum_{j=0}^{k-1} A_j^k(cv_s(p))[cv_s(p) + h]^j}{s^{k+1}(\phi(cv_s(p)))^{k+1}} \Psi(cv_s(p), h) \\
&+ (-1)^{k+1} \frac{(hcv_s(p)k) \sum_{j=0}^{k-1} A_j^k(cv_s(p))[cv_s(p) + h]^j}{s^{k+1}(\phi(cv_s(p)))^{k+1}} \Psi(cv_s(p), h) \\
&+ (-1)^{k+1} \frac{h \sum_{j=0}^{k-1} (\partial A_j^k(cv_s(p))/\partial cv_s(p))[cv_s(p) + h]^j}{s^{k+1}(\phi(cv_s(p)))^{k+1}} \Psi(cv_s(p), h) \\
&+ (-1)^{k+1} \frac{h \sum_{j=1}^{k-1} j A_j^k(cv_s(p))[cv_s(p) + h]^{j-1}}{s^{k+1}(\phi(cv_s(p)))^{k+1}} \Psi(cv_s(p), h) \\
&= (-1)^{k+1} \frac{\Psi(cv_s(p), h)}{s^{k+1}(\phi(cv_s(p)))^{k+1}} \left\{ h \sum_{j=0}^{k} (B_j^k + C_j^k)[cv_s(p) + h]^j \right\}.
\end{aligned}
$$

24

Since $A_j^k(cv_s(p)), j = 0, \ldots, k-1$ are polynomials with non-negative coefficients, B_j^k and C_j^k are also polynomials with non-negative coefficients for every $j = 0, \ldots, k$. It follows that

$$\Psi^{(k+1)}(cv_s(p), h) = (-1)^{k+1} \frac{h \sum_{j=0}^{k} A_j^{k+1}(cv_s(p))[cv_s(p) + h]^j}{s^{k+1}(\phi(cv_s(p)))^{k+1}} \Psi(cv_s(p), h),$$

where $A_j^{k+1}(cv_s(p)) = B_j^k + C_j^k, j = 0, \ldots, k$. This completes the induction step. $\square$

Using Lemma 2, we now proof Theorem 2 and Theorem 3.

Proof of Theorem 2. Lemma 2 and equations (15)–(16) directly imply that $0 \le (-1)^k g_1^{(k)}(p)$, for $p \in (0, 1/2]$ and $0 \le (-1)^k g_2^{(k)}(p)$, for $p \in (0, 1)$ for $k = 1, 2, \ldots$. The result for the two-sided case follows from the fact that $h\{[cv_2(p) + h]^j \Psi(cv_2(p), h) - [cv_2(p) - h]^j \Psi(cv_2(p), -h)\} \ge 0$ for every $j \in \mathbb{N}$ and every $h \in \mathbb{R}$. $\square$

Proof of Theorem 3. Consider first the one-sided t-test. Lemma 2 implies that

$$(-1)^k g_1^{(k)}(p) \le \mathcal{B}_1^{(k)}(p) := \max_{h \ge 0} \left\{ |\Psi^{(k)}(cv_1(p), h)| \exp\{-h^2/2\} \right\},$$

where the inequality holds for every $p \in (0, 1)$ and the maximum is finite for every $p \in (0, 1)$ since $|\Psi^{(k)}(cv_1(p), h)| \exp\{-h^2/2\}$ is finite for every $h \ge 0$ and converges to zero as h goes to infinity. For the upper bound on $g_1(p)$, note that for $p \in (0, 1/2]$, $\max_{h \ge 0} \{|\Psi(cv_1(p), h)| \exp\{-h^2/2\}\} = \Psi(cv_1(p), cv_1(p)) \exp\{-cv_1^2(p)/2\} = \exp\{cv_1^2(p)/2\}$. For $p > 1/2$ and $h \ge 0$, $hcv_1(p) - cv_1^2(p)/2 < 0$ and hence $g_1(p) \le 1$.

For two-sided tests, by the above arguments and symmetry, we have

$$(-1)^k g_2^{(k)}(p) \le \mathcal{B}_2^{(k)}(p) := \max_{h \in \mathbb{R}} \left\{ |\Psi^{(k)}(cv_2(p), h) + \Psi^{(k)}(cv_2(p), -h)| \exp\{-h^2/2\}/2 \right\},$$

where the upper bound is finite for every $p \in (0, 1)$.

For the upper bound on $g_2(p)$, one can show that for $p \ge 2(1 - \Phi(1))$, the first-order condition for maximizing $|\Psi(cv_2(p), h) + \Psi(cv_2(p), -h)| \exp\{-h^2/2\}/2$ has only one solution, $h_o = 0$. By checking second-order conditions we can verify that 0 is the maximum. For $p < 2(1 - \Phi(1))$, 0 becomes local minimum, and there are two additional non-zero symmetric solutions to the first-order condition that satisfy the second-order condition for a maximum and result in identical values of the objective function.

$\square$