The Art and Science of p-value Hacking

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk
@xmjandrews
https://github.com/lawsofthought/smlp2017

August 31, 2017

P-hacking: A definition

- *p-hacking*, broadly defined, is the manipulation, whether intentional or not, of frequentist[1] statistical testing procedures in order to obtain a desired outcome.
- Technically, p-hacking is always a type of undisclosed multiple simultaneous statistical testing, whereby the de facto false positive rate greatly exceeds the nominal false-positive rate, i.e. α.
- Related terms include *data-dredging*, *data-snooping*, *data-fishing*, *cherry-picking*, *significance-chasing*, and so on.
- The original contemporary exposé of this general phenomenon is due to Simmons, Nelson, & Simonsohn (2011), followed by Simonsohn, Nelson, & Simmons (2014).

[1] The Bayesian counterpart to p-hacking is b-hacking.

Frequentist statistical testing

Neyman-Pearson testing

- In advance of data collection, our scientific hypothesis, operationalized as $\mathcal{H}_1$ (e.g. $\mathcal{H}_1 : \theta > 0$), is specified.
- This leads to a corresponding null hypothesis, e.g. $\mathcal{H}_0 : \theta = 0$, a test statistic $\mathsf{T}(\mathcal{D})$, and a critical threshold for the statistic T_{crit}, such that

$$\mathrm{P}(|\mathsf{T}(\mathcal{D})| > \mathsf{T}_{\text{crit}} | \mathcal{H}_0 = \text{True}) = \alpha,$$

  where α is conventionally 0.05.
- We then determine a minimum sample size for $\mathcal{D}$ so that

$$\mathrm{P}(|\mathsf{T}(\mathcal{D})| > \mathsf{T}_{\text{crit}} | \mathcal{H}_1 = \text{True}) \gtrapprox 1 - \beta,$$

  where β is conventionally 0.2 or 0.1.
- We then collect $\mathcal{D}$, calculate $\mathsf{T}(\mathcal{D})$. If $|\mathsf{T}(\mathcal{D})| > \mathsf{T}_{\text{crit}}$ then we reject the null. Otherwise, we do not reject it.
- Following this procedure, in the long run, our false positive rate will be α, and our false negative rate will be β.

P-hacking

- In advance of data collection, we begin with a (perhaps vaguely stated) scientific hypothesis.
- We collect data $\mathcal{D}$.
- We then operationalize our scientific hypothesis as $\mathcal{H}_1^k$, which leads to $\mathcal{H}_0^k$, $T^k(\mathcal{D})$, T^k_{crit}, starting with $k = 1$.
- We calculate $T^k(\mathcal{D})$. If $|T^k(\mathcal{D})| > T^k_{crit}$ then we reject the null and stop.
- Otherwise, if $|T^k(\mathcal{D})| \leqslant T^k_{crit}$, we re-operationalize our scientific hypothesis as $\mathcal{H}_1^{k=2}$ and test again.
- We continue as such indefinitely and stop when we obtain a significant result, and then report *only* that result.
- Following this procedure, in the long run, our false positive rate will be $\gg \alpha$.

P-hacking example 1: Subsetting

Online demo: `https://lawsofthought.shinyapps.io/p_hacking/`

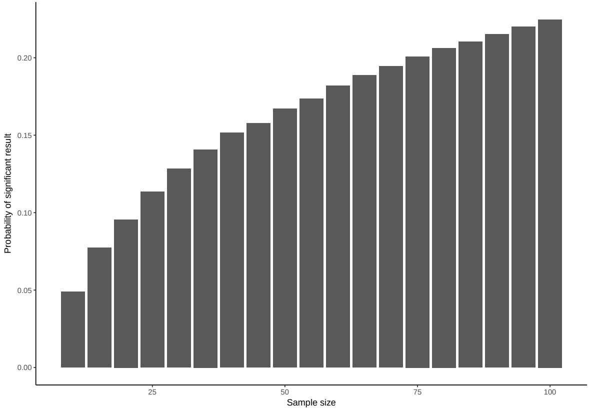
- Let's assume we want to test if two groups of people differ in the mean value of some variable.
- In both groups, there are men and women.
- We can test just the men, just the women, or both.
- In a simulation with $n = 20$ people in each group, with $\alpha = 0.05$, subsetting results in the false positive rate being $\approx 11.8\%$.

P-hacking example 2: Adding a covariate

- In an identical problem to before, instead of subsetting by gender, we simply add it as a covariate.
- We then test if the main effect exists in the presence and absence of the gender covariate, or if there is an interaction between gender and the main effect.
- In another simulation, $n = 20$ people in each group, with $\alpha = 0.05$, this leads to the false positive rate being $\approx 12.1\%$.

P-hacking example 3: Optional stopping

Online demo:
`https://lawsofthought.shinyapps.io/optional_stopping/`



- Collecting data, testing, and then collecting more data if results are not significant, leads to a steady rise in false positive rates.

P-hacking example 4: Removing outliers

- Using an identical problem to before, we remove outliers, or not, before testing.
- Outlier may be defined as any of the following:
  1. Data above/below 2 SDs from mean.
  2. Data above/below 1.5 SDs from mean.
  3. Data in the upper/lower 5% quantiles.
  4. Data in the upper/lower 2.5% quantiles.
  5. The 5 highest/lowest values.
  6. The 2 highest/lowest values
- In a simulation, with $n = 20$ in each group, and $\alpha = 0.05$, this leads to a false positive rate of 13.7%.

P hacking broadside

Combine your p-hack tools for maximum effect

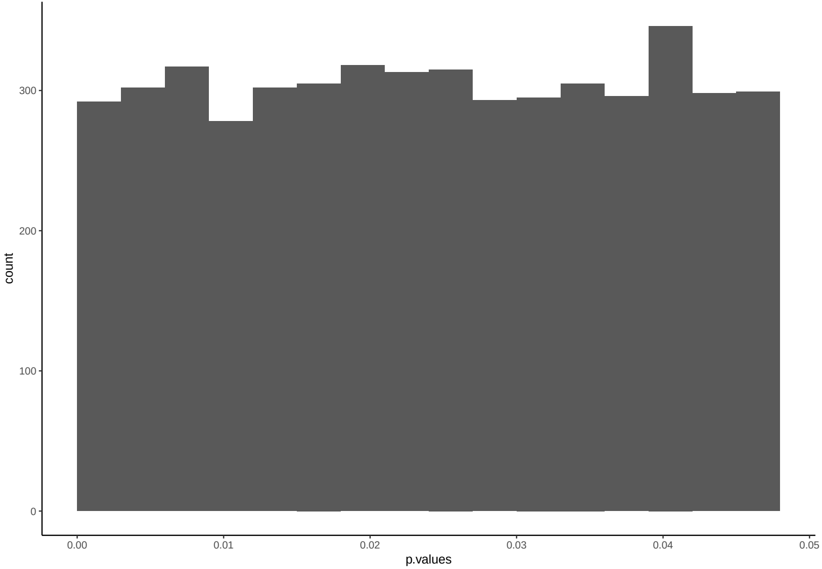
- Using an identical problem to before, we start with two samples of size $n = 20$ and $\alpha = 0.05$.
- Combining our removal of outliers method *and* our covariate method leads to a false positive rate of 33.2%.
- Combining our removal of outliers method *and* our covariate method *and* collecting 10 new data points in each group until significance or 100 in each group leads to a false positive rate of 64.4%.

P-value distribution (p-curves) under null

- The distribution of p-values (p-curves) in any given body of work will be a function of the true effect size, which may be zero, and the extent of p-hacking.
- Whether we can use p-curves in meta-analysis to assess the extent and consequences of p-hacking, as recommended by Simonsohn et al. (2014), is a matter of debate, see
  - Gelman & O'Rourke (2013)
  - Head, Holman, Lanfear, Kahn, & Jennions (2015)
  - Bishop & Thompson (2016)
  - Bruns & Ioannidis (2016)
  - Hartgerink (2017)

P-curve under null

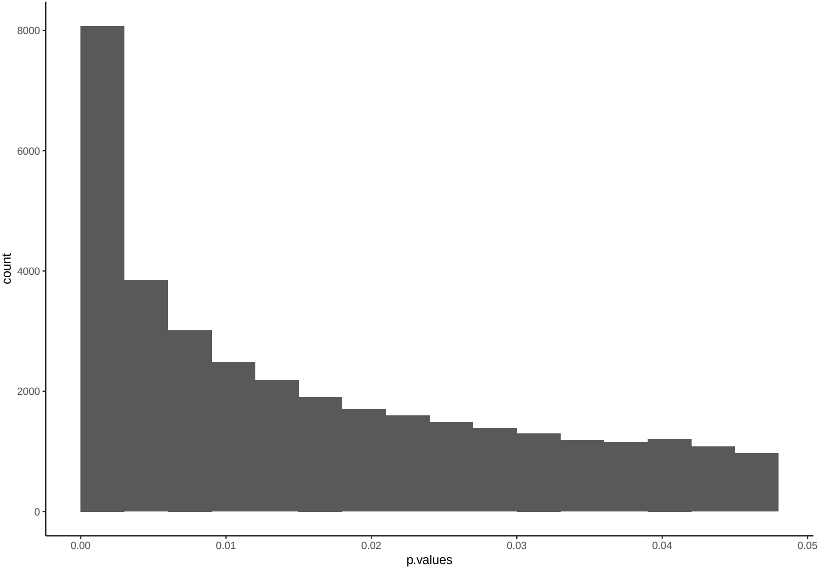
When null is true, the distribution of p-values is uniform.



P-value distribution (p-curve) under non-null

Medium effect

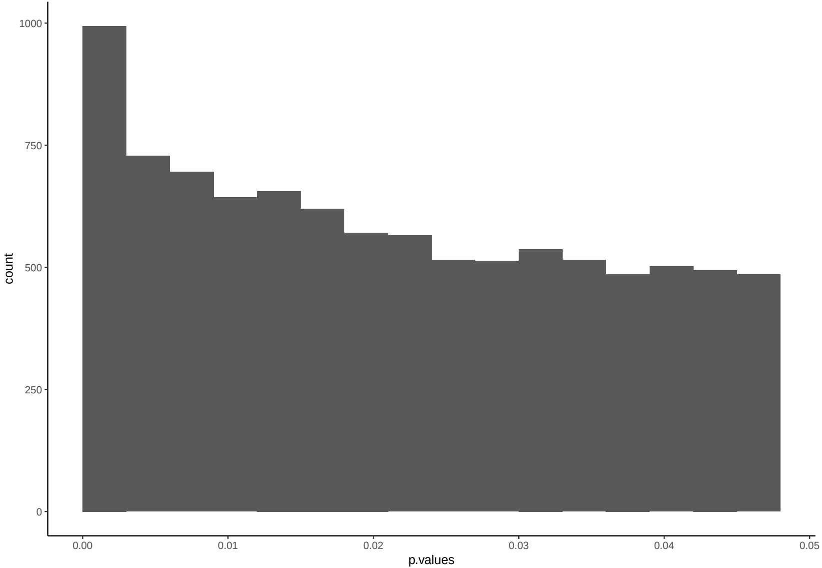
When null is false, the distribution of p-values is right skewed.



P-curve under non-null

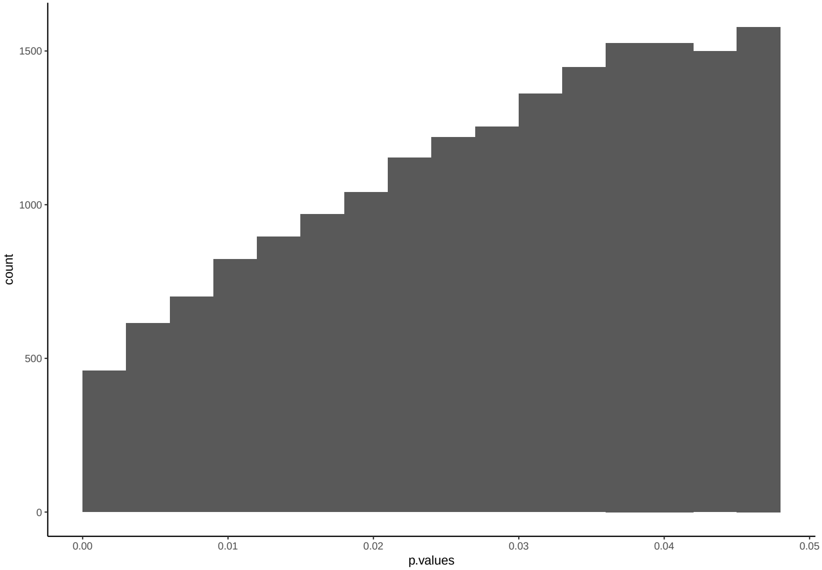
Low effect

When null is false, the distribution of p-values is right skewed.



P-curve under null with p-hacking

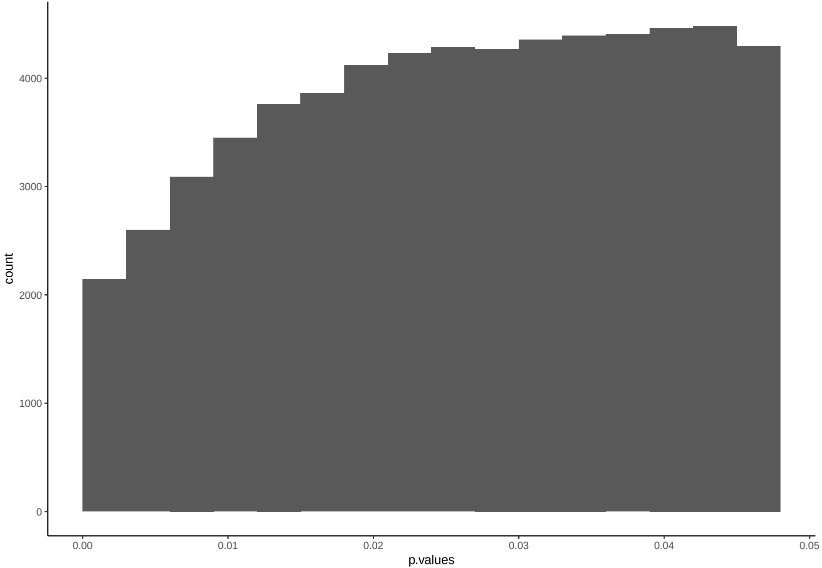
When null is true, the distribution of *p-hacked* p-values is left-skewed.



P-curve under non-null with p-hacking

Low effect

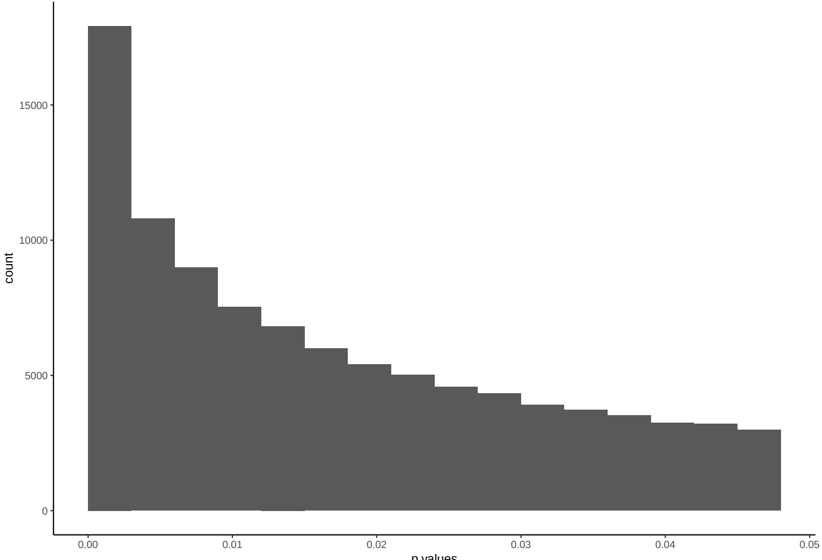
When effect is low, the distribution of *p-hacked* p-values is left-skewed.



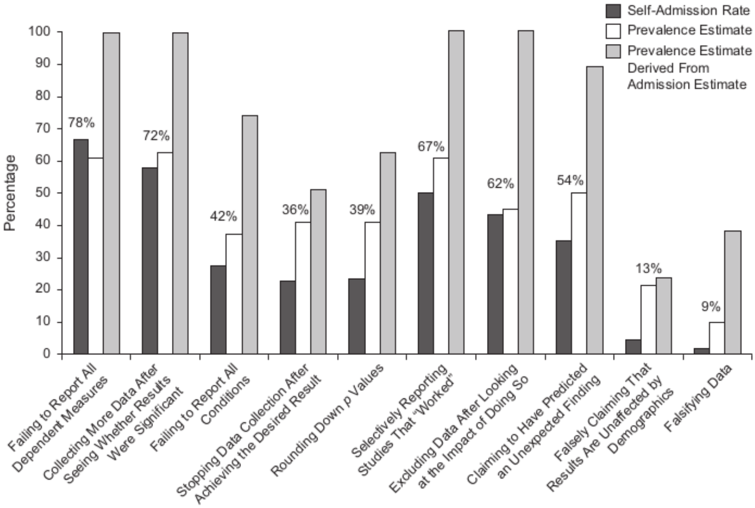
P-curve under non-null with p-hacking

Medium effect

When effect is medium, the distribution of *p-hacked* p-values is right-skewed.



Prevalence of P-hacking



From John, Loewenstein, & Prelec (2012).

Why P-hacking is so toxic for science

- P-hacking is easy to do (you just need some ethical laxity and some stamina)
- It is hard to detect
- It can dramatically increase the false positive rate
- False positives are hard to detect and hard to eliminate
- False positives add noise to the literature, and result in wasted resources when used as the basis for future research
- P-hacking may be self-perpetuating: Results are p-hacked because some effects are assumed to be real (on the basis of p-hacked literature)

How to eliminate p-hacking?

- P-hacking is an ethical problem, rather than a statistical issue.
- P-hacking can be eliminated by changing ethical standards:
  - *Honesty in reporting*: The explicit recommendations in Simmons et al. (2011) are largely recommendations for a cultural shift away from selective reporting.
  - *Pre-registration*: It immediately eliminates *harking* and greatly reduces researcher degrees of freedom
  - *Open (raw) data and analysis code*: Disclosing all the original data (especially as recommended by Rouder (2016)) and the processing/analysis pipeline can make tricks easier to identify, and allows alternative analyses to be performed

References I

Bishop, D. V., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ*, *4*, e1715.

Bruns, S. B., & Ioannidis, J. P. (2016). P-curve and p-hacking in observational research. *PLoS One*, *11*(2), e0149144.

Gelman, A., & O'Rourke, K. (2013). Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics*, kxt034.

Hartgerink, C. H. (2017). Reanalyzing head et al.(2015): Investigating the robustness of widespread p-hacking. *PeerJ*, *5*, e3068.

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

References II

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior Research Methods*, *48*(3), 1062–1069.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

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