A Short Note on P-Value Hacking
Nassim Nicholas Taleb
Tandon School of Engineering

Abstract—We present the expected values from p-value hacking as a choice of the minimum p-value among \( m \) independents tests, which can be considerably lower than the "true" p-value, even with a single trial, owing to the extreme skewness of the meta-distribution. We first present an exact probability distribution (meta-distribution) for p-values across ensembles of statistically identical phenomena. We derive the distribution for small samples \( 2 < n \leq n' \approx 30 \) as well as the limiting one as the sample size \( n \) becomes large. We also look at the properties of the "power" of a test through the distribution of its inverse for a given p-value and parametrization.

The formulas allow the investigation of the stability of the reproduction of results and "p-hacking" and other aspects of meta-analysis.

P-values are shown to be extremely skewed and volatile, regardless of the sample size \( n \), and vary greatly across repetitions of exactly same protocols under identical stochastic copies of the phenomenon; such volatility makes the minimum \( p \) value diverge significantly from the "true" one. Setting the power is shown to offer little remedy unless sample size is increased markedly or the p-value is lowered by at least one order of magnitude.

P-value hacking, just like an option or other members in the class of convex payoffs, is a function that benefits from the underlying variance and higher moment variability. The researcher or group of researchers have an implicit "option" to pick the most favorable result in \( m \) trials, without disclosing the number of attempts, so we tend to get a rosier picture of the end result than reality. The distribution of the minimum \( p \)-value and the "optionality" can be made explicit, expressed in a parsimonious formula allowing for the understanding of biases in scientific studies, particularly under environments with high publication pressure.

Assume that we know the "true" \( p \)-value, \( p_s \), what would its realizations look like across various attempts on statistically identical copies of the phenomena? By true value \( p_s \), we mean its expected value by the law of large numbers across an \( m \) ensemble of possible samples for the phenomenon under scrutiny, that is \( \frac{1}{m} \sum_{i=1}^{m} p_i \rightarrow p_s \) (where \( \rightarrow \) denotes convergence in probability). A similar convergence argument can be also made for the corresponding "true median" \( p_M \). The distribution of \( n \) small samples can be made explicit (albeit with special inverse functions), as well as its parsimonious limiting one for \( n \) large, with no other parameter than the median value \( p_M \). We were unable to get an explicit form for \( p_s \) but we go around it with the use of the median.

It turns out, as we can see in Fig. 3 the distribution is extremely asymmetric (right-skewed), to the point where 75% of the realizations of a "true" \( p \)-value of .05 will be <.05 (a borderline situation is \( 3 \times \) as likely to pass than fail a given protocol), and, what is worse, 60% of the true \( p \)-value of .02 and .12 will be below .05. This implies serious gaming and "p-hacking" by researchers, even under a moderate amount of repetition of experiments.

Although with compact support, the distribution exhibits the attributes of extreme fat-tailedness. For an observed \( p \)-value of, say, .02, the "true" \( p \)-value is likely to be >.1 (and very possibly close to .2), with a standard deviation >.2 (sic) and a mean deviation of around .35 (sic, sic). Because of the excessive skewness, measures of dispersion in \( L^1 \) and \( L^2 \) (and higher norms) vary hardly with \( p_s \), so the standard deviation is not proportional, meaning an in-sample .01 \( p \)-value has a significant probability of having a true value > .3.

Second version, January 2018, First version was March 2015.
So clearly we don’t know what we are talking about when we talk about p-values.

Earlier attempts for an explicit meta-distribution in the literature were found in [1] and [2], though for situations of Gaussian subordination and less parsimonious parametrization. The severity of the problem of significance of the so-called “statistically significant” has been discussed in [3] and offered a remedy via Bayesian methods in [4], which in fact recommends the same tightening of standards to p-values ≈ .01. But the gravity of the extreme skewness of the distribution of p-values is only apparent when one looks at the meta-distribution.

For notation, we use \( n \) for the sample size of a given study and \( m \) the number of trials leading to a p-value.

I. DERIVATION OF THE METADISTRIBUTION OF P-VALUES

Proposition 1. Let \( P \) be a random variable \( \in [0,1] \) corresponding to the sample-derived one-tailed p-value from the paired T-test statistic (unknown variance) with median value \( \mathbb{M}(P) = p_M \in [0,1] \) derived from a sample of \( n \) size. The distribution across the ensemble of statistically identical copies of the sample has for PDF

\[
\varphi(p; p_M) = \begin{cases} 
\varphi(p; p_M)_L & \text{for } p < \frac{1}{2} \\
\varphi(p; p_M)_H & \text{for } p > \frac{1}{2}
\end{cases}
\]

\[
\begin{align*}
\varphi(p; p_M)_L &= \lambda_p^{\frac{1}{2}}(n-1) \left( -\lambda_p (\lambda_{p_M} - 1) \right) \\
&\times \sqrt{1 - \lambda_p (\lambda_{p_M} - 1) - 2 \sqrt{1 - \lambda_p (1 - \lambda_p) \sqrt{1 - \lambda_{p_M} (1 - \lambda_{p_M})}}} \\
&\times \frac{1}{\lambda_p} \left( \frac{1}{\lambda_p} - 2 \sqrt{1 - \lambda_p (1 - \lambda_p) \sqrt{1 - \lambda_{p_M} (1 - \lambda_{p_M})}} + 1 \right)^{n/2}
\end{align*}
\]

\[
\begin{align*}
\varphi(p; p_M)_H &= (1 - \lambda_p)^{\frac{1}{2}}(n-1) \left( -\lambda_p (1 - \lambda_{p_M} - 1) \right) \\
&\times \sqrt{1 - \lambda_p (1 - \lambda_p) \sqrt{1 - \lambda_{p_M} (1 - \lambda_{p_M})}} \\
&\times \frac{(\lambda_p' - \lambda_{p_M})}{\lambda_p'} \sqrt{1 - \lambda_{p_M} (1 - \lambda_{p_M})} + 1 \right)^{n/2}
\end{align*}
\]

where \( \lambda_p = I_{2p}^{-1}(\frac{1}{2}, \frac{1}{2}) \), \( \lambda_{p_M} = I_2^{-1}(\frac{1}{2}, \frac{1}{2}), \lambda_p' = I_{2p-1}^{-1}(\frac{1}{2}, \frac{1}{2}) \), and \( I_{\cdot}^{-1}(\cdot, \cdot) \) is the inverse beta regularized function.

Remark 1. For \( p = \frac{1}{2} \) the distribution doesn’t exist in theory, but does in practice and we can work around it with the sequence \( p_{MS} = \frac{1}{2} \pm \frac{1}{2} \), as in the graph showing a convergence to the Uniform distribution on \([0,1]\) in Figure 2. Also note that what is called the “null” hypothesis is effectively a set of measure 0.

Proof. Let \( Z \) be a random normalized variable with realizations \( \zeta \), from a vector \( \vec{v} \) of \( n \) realizations, with sample mean \( m_v \), and sample standard deviation \( s_v \), \( \zeta = \frac{m_v - m_h}{s_v} \) (where \( m_h \) is the level it is tested against), hence assumed to \( \sim \) Student T with \( n \) degrees of freedom, and, crucially, supposed to deliver a mean of \( \zeta \),

\[
f(\zeta; \hat{c}) = \frac{n}{(n-\zeta)^{\frac{n+1}{2}}} \frac{n^{n/2}}{\sqrt{\Gamma(n/2) B(\frac{n}{2}, \frac{1}{2})}}
\]

where \( B(\cdot, \cdot) \) is the standard beta function. Let \( g(.) \) be the one-tailed survival function of the Student T distribution with zero mean and \( n \) degrees of freedom:

\[
g(\zeta) = P(Z > \zeta) = \begin{cases} 
\frac{1}{2} I_{\zeta n + \frac{1}{2}}(\frac{n}{2}, \frac{1}{2}) & \zeta \geq 0 \\
\frac{1}{2} \left( I_{\zeta^2} \left( \frac{n}{2}, \frac{1}{2} \right) + 1 \right) & \zeta < 0
\end{cases}
\]

where \( I_{\cdot} \) is the incomplete Beta function.

We now look for the distribution of \( g \circ f(\zeta) \). Given that \( g(.) \) is a legt Borel function, and naming \( p \) the probability as a random variable, we have by a standard result for the transformation:

\[
\varphi(p, \tilde{\zeta}) = \frac{f(g^{-1}(p))}{g'(g^{-1}(p))}
\]

We can convert \( \tilde{\zeta} \) into the corresponding median survival probability because of symmetry of \( Z \). Since one half the observations fall on either side of \( \tilde{\zeta} \), we can ascertain that the transformation is median preserving: \( g(\tilde{\zeta}) = \frac{1}{2} \), hence \( \varphi(p_M, \cdot) = \frac{1}{2} \). Hence we end up having \( \{ \tilde{\zeta} : \frac{1}{2} I_{\zeta^2} \left( \frac{n}{2}, \frac{1}{2} \right) = p_M \} \) (positive case) and \( \{ \tilde{\zeta} : \frac{1}{2} \left( I_{\zeta^2} \left( \frac{n}{2}, \frac{1}{2} \right) + 1 \right) = p_M \} \) (negative case). Replacing we get Eq[1] and Proposition 1 is done.

\[ \square \]

We note that \( n \) does not increase significance, since p-values are computed from normalized variables (hence the universality of the meta-distribution); a high \( n \) corresponds to an increased convergence to the Gaussian. For large \( n \), we can prove the following proposition:

Proposition 2. Under the same assumptions as above, the limiting distribution for \( \varphi(.) \):

\[
\lim_{n \to \infty} \varphi(p; p_M) = e^{-\text{erfc}^{-1}(2p_M)(\text{erfc}^{-1}(2p_M) - 2	ext{erfc}^{-1}(2p))}
\]

where \( \text{erfc}(.) \) is the complementary error function and \( \text{erfc}^{-1}(.) \) its inverse.

The limiting CDF \( \Phi(.) \)

\[
\Phi(k; p_M) = \frac{1}{2} \text{erfc} \left( \text{erfc}^{-1}(1 - 2k) - \text{erfc}^{-1}(1 - 2p_M) \right)
\]

Proof. For large \( n \), the distribution of \( Z = \frac{m_v}{s_v} \) becomes that of a Gaussian, and the one-tailed survival function \( g(.) = \frac{1}{2} \text{erfc} \left( -\sqrt{2} \zeta \right), \zeta(p) \to \sqrt{2} \text{erfc}^{-1}(p) \).

This limiting distribution applies for paired tests with known or assumed sample variance since the test becomes a Gaussian variable, equivalent to the convergence of the T-test (Student T) to the Gaussian when \( n \) is large.

N. N. Taleb
Remark 2. For values of \( p \) close to 0, \( \varphi \) in Eq. 2 can be usefully calculated as:

\[
\varphi(p; p_M) = \sqrt{2\pi p_M} \left( \log \left( \frac{1}{2\pi p_M^2} \right) \right) e^{-\log(2\pi \log(1/2\pi p_M)) - 2\log(p) - \log(2\pi \log(1/2\pi p_M)) - 2\log(p_M)} + O(p^2). \tag{4}
\]

The approximation works more precisely for the band of relevant values \( 0 < p < \frac{1}{2\pi} \).

From this we can get numerical results for convolutions of \( \varphi \) using the Fourier Transform or similar methods.

II. P-VALUE HACKING

We can and get the distribution of the minimum p-value per \( m \) trials across statistically identical situations thus get an idea of "p-hacking", defined as attempts by researchers to get the lowest p-values of many experiments, or try until one of the tests produces statistical significance.

**Proposition 3.** The distribution of the minimum of \( m \) observations of statistically identical p-values becomes (under the limiting distribution of proposition 2):

\[
\varphi_m(p; p_M) = m e^{erf^{-1}(2pM)}(2erf^{-1}(2p) - erf^{-1}(2pM))^{m-1} \left( 1 - \frac{1}{2}erf(\frac{erf^{-1}(2p) - erf^{-1}(2pM))}{2} \right) \tag{5}
\]

**Proof.** \( P(p_1 > p, p_2 > p, \ldots, p_m > p) = \prod_{i=1}^m \Phi(p_i) = \Phi(p)^m \). Taking the first derivative we get the result. □

Outside the limiting distribution: we integrate numerically for different values of \( m \) as shown in figure [I] So, more precisely, for \( m \) trials, the expectation is calculated as:

\[
\mathbb{E}(p_{min}) = \int_0^1 -m \varphi(p; p_M) \left( \int_p^1 \varphi(u) \, du \right) \, dp \tag{6}
\]

III. OTHER DERIVATIONS

**Inverse Power of Test**

Let \( \beta \) be the power of a test for a given p-value \( p \), for random draws \( X \) from unobserved parameter \( \theta \) and a sample size of \( n \). To gauge the reliability of \( \beta \) as a true measure of power, we perform an inverse problem:

[Diagram of \( \beta^{-1}(X) \)]

**Proposition 4.** Let \( \beta_c \) be the projection of the power of the test from the realizations assumed to be student T distributed and evaluated under the parameter \( \theta \). We have

\[
\Phi(\beta_c) = \begin{cases} 
\Phi(\beta_c)_L & \text{for } \beta_c < \frac{1}{2} \\
\Phi(\beta_c)_H & \text{for } \beta_c > \frac{1}{2}
\end{cases}
\]

where

\[
\Phi(\beta_c)_L = \sqrt{1 - \gamma_1 \gamma_2^{-1/2}} \\
\left( -\frac{2\sqrt{\beta_c}\gamma_1^{-1} - \sqrt{\gamma_1}}{\sqrt{\gamma_1}} - \frac{2\sqrt{\beta_c} - \sqrt{\gamma_1}}{\sqrt{\gamma_1}} \right) \frac{n+1}{2} \\
\Phi(\beta_c)_H = \sqrt{\gamma_2}(1 - \gamma_2)^{-1/2} B \left( \frac{1}{2}, \frac{n}{2} \right) \\
\left( -\frac{2\gamma_2^{-1} + \sqrt{\gamma_2}}{\gamma_2^{-1/2}} + \frac{2\gamma_2^{-1} - \sqrt{\gamma_2}}{\gamma_2^{-1/2}} - \frac{1}{\sqrt{\gamma_2}} \right)^{-1/2} \\
\sqrt{-(\gamma_2 - 1)\gamma_2 B \left( \frac{1}{2}, \frac{1}{2} \right)} \tag{7}
\]

where \( \gamma_1 = I_{2\beta_c}^{-1} \left( \frac{n}{2}, \frac{1}{2} \right), \gamma_2 = I_{2\beta_c-1}^{-1} \left( \frac{n}{2}, \frac{1}{2} \right), \text{ and } \gamma_3 = I_{(1,2p_c-1)}^{-1} \left( \frac{n}{2}, \frac{1}{2} \right) \).
IV. Application and Conclusion

- One can safely see that under such stochasticity for the realizations of p-values and the distribution of its minimum, to get what a scientist would expect from a 5% confidence level (and the inferences they get from it), one needs a p-value of at least one order of magnitude smaller.

- Attempts at replicating papers, such as the open science project [5], should consider a margin of error in its own procedure and a pronounced bias towards favorable results (Type-I error). There should be no surprise that a previously deemed significant test fails during replication—in fact it is the replication of results deemed significant at a close margin that should be surprising.

- The "power" of a test has the same problem unless one either lowers p-values or sets the test at higher levels, such as .99.

Acknowledgment

Marco Avellaneda, Pasquale Cirillo, Yaneer Bar-Yam, friendly people on twitter, less friendly verbagastic psychologists on twitter, ...

References

[1] H. J. Hung, R. T. O’Neill, P. Bauer, and K. Kohne, “The behavior of the p-value when the alternative hypothesis is true,” *Biometrics*, pp. 11–22, 1997.

[2] H. Sackrowitz and E. Samuel-Cahn, “P values as random variables—expected p values,” *The American Statistician*, vol. 53, no. 4, pp. 326–331, 1999.

[3] A. Gelman and H. Stern, “The difference between “significant” and “not significant” is not itself statistically significant,” *The American Statistician*, vol. 60, no. 4, pp. 328–331, 2006.

[4] V. E. Johnson, “Revised standards for statistical evidence,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 48, pp. 19313–19317, 2013.

[5] O. S. Collaboration et al., “Estimating the reproducibility of psychological science,” *Science*, vol. 349, no. 6251, p. aac4716, 2015.