The Meta-Distribution of Standard P-Values

Nassim Nicholas Taleb
Tandon School of Engineering, New York University and Real World Risk Institute, LLC.

***Abstract*—We present an explicit and parsimonious probability distribution (meta-distribution) for p-values across ensembles of statistically identical phenomena, having for sole parameter the median "true" p-value, as well as the distribution of the minimum p-value among m independents tests. P-values are 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.**

he convenience of the formulas allows the investigation of scientific results, particularly meta-analyses.

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 $p_M \triangleq \frac{1}{m}\sum_{\leq m} p_i \xrightarrow{P} p_s$ (where $\xrightarrow{P}$ denotes convergence in probability). A similar convergence argument can be also made for the corresponding "true median". The main result of this paper is that distribution of p-values (meta-distribution) can be made explicit, in a parsimonious formula, with no other parameter than the median value p_M. We cannot get an explicit form for p_s but we go around it with the use of the median.

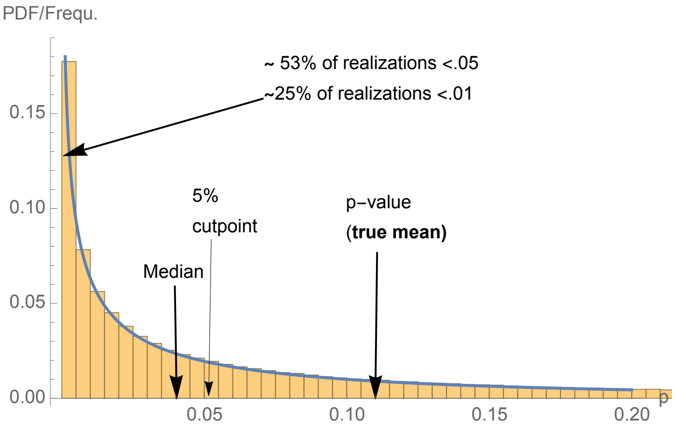

Fig. 1. The probability distribution of a one-tailed p-value with expected value .11 generated by Monte Carlo (histogram) as well as analytically with $\varphi(.)$ (the solid line). We draw all possible subsamples from an ensemble with given properties. The excessive skewness of the distribution makes the average value considerably higher than most observations, hence causing illusions of "statistical significance".

It turned out, as we can see in Fig. 1 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× as likely to pass than fail a given protocol), and, what is worse, 60% of the true p-value of .12 will be below .05. There has been quite a bit of confusion over the past 100 years of binary hypothesis testing, as well as serious misinterpretation.

Earlier attempts for an explicit meta-distribution in the literature were found in [1] and [2], though for situations of Gaussian subordination and less pars imonious 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. Theorems and derivations

Theorem 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*

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

$$\varphi(p;p_M)_L = \lambda_p^{\frac{1}{2}(-n-1)} \sqrt{-\frac{\lambda_p\left(\lambda_{p_M}-1\right)}{\left(\lambda_p-1\right)\lambda_{p_M}-2\sqrt{\left(1-\lambda_p\right)\lambda_p}\sqrt{\left(1-\lambda_{p_M}\right)\lambda_{p_M}}+1}} \left(\frac{1}{\frac{1}{\lambda_p}-\frac{2\sqrt{1-\lambda_p}\sqrt{\lambda_{p_M}}}{\sqrt{\lambda_p}\sqrt{1-\lambda_{p_M}}}+\frac{1}{1-\lambda_{p_M}}-1}\right)^{n/2}$$

$$\varphi(p;p_M)_H = \left(1-\lambda'_p\right)^{\frac{1}{2}(-n-1)} \left(\frac{\left(\lambda'_p-1\right)\left(\lambda_{p_M}-1\right)}{\lambda'_p\left(-\lambda_{p_M}\right)+2\sqrt{\left(1-\lambda'_p\right)\lambda'_p}\sqrt{\left(1-\lambda_{p_M}\right)\lambda_{p_M}}+1}\right)^{\frac{n+1}{2}} \tag{1}$$

where $\lambda_p = I_{2p}^{-1}\left(\frac{n}{2},\frac{1}{2}\right)$, $\lambda_{p_M} = I_{1-2p_M}^{-1}\left(\frac{1}{2},\frac{n}{2}\right)$, $\lambda'_p = I_{2p-1}^{-1}\left(\frac{1}{2},\frac{n}{2}\right)$, and $I_{(.)}^{-1}(.,.)$ 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 $p_{m_k} = \frac{1}{2} \pm \frac{1}{k}$, 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.*

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 theorem:

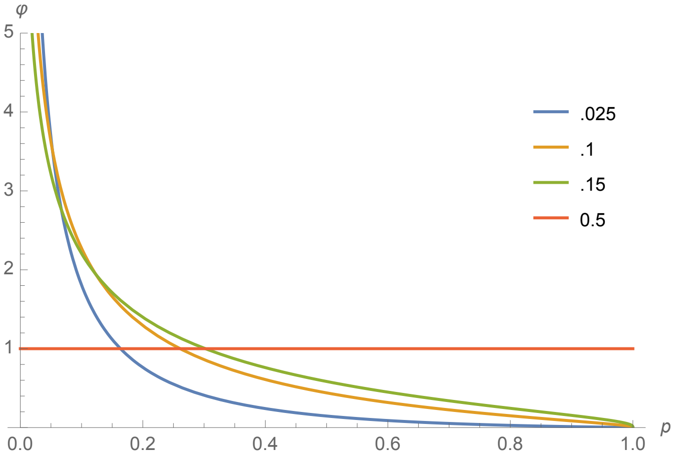

Fig. 2. The probability distribution of p at different values of p_M. We observe how $p_M = \frac{1}{2}$ leads to a uniform distribution.

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

$$\lim_{n\to\infty} \varphi(p;p_M) = e^{\left(-\mathit{erfc}^{-1}(2)\left(\mathit{erfc}^{-1}(2p_M)-2\mathit{erfc}^{-1}(2p)\right)\right)} \quad (2)$$

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

The limiting CDF $\Phi(.)$

$$\Phi(k;p_M) = \frac{1}{2}\mathit{erfc}\left(\mathit{erf}^{-1}(1-2k) - \mathit{erf}^{-1}(1-2p_M)\right) \quad (3)$$

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.

Remark 2. *For values of* p *close to 0,* φ *in Equ. 2 can be usefully calculated as:*

$$\varphi(p;p_M) = \sqrt{2\pi}p_M\sqrt{-\left(\log\left(2\pi p_M^2\right)\right)}$$
$$e^{\sqrt{-\frac{(2\pi\log)\log}{2\pi p^2}-2(p\log)}\sqrt{-2(\log p_M)-\log\left(-2\pi\left(\log(2\pi p_M^2)\right)\right)}}+O(p^2). \quad (4)$$

The approximation works more precisely for values $0 < p < \frac{1}{2\pi}$.

Proof. Let Z be a random normalized variable with realizations ζ, from a vector $\vec{v}$ of n realizations, with sample mean m_v, and sample standard deviation s_v, $\zeta = \frac{m_v - m_h}{\frac{s_v}{\sqrt{n}}}$ (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 $\bar{\zeta}$,

$$f(\zeta;\bar{\zeta}) = \frac{\left(\frac{n}{(\bar{\zeta}-\zeta)^2+n}\right)^{\frac{n+1}{2}}}{\sqrt{n}B\left(\frac{n}{2},\frac{1}{2}\right)}$$

where B(.,.) 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:

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

where $I_{(.,.)}$ is the incomplete Beta function.

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

$$\varphi(p,\bar{\zeta}) = \frac{f\left(g^{(-1)}(p)\right)}{\left|g'\left(g^{(-1)}(p)\right)\right|}$$

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

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

From this we can get numerical results for convolutions of φ using the Fourier Transform or similar methods.

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.

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

$$\varphi_m(p;p_M) = m\, e^{\mathit{erfc}^{-1}(2p_M)\left(2\mathit{erfc}^{-1}(2p)-\mathit{erfc}^{-1}(2p_M)\right)}$$
$$\left(1-\frac{1}{2}\mathit{erfc}\left(\mathit{erfc}^{-1}(2p)-\mathit{erfc}^{-1}(2p_M)\right)\right)^{m-1} \quad (5)$$

Proof. $P\left(p_1 > p, p_2 > p, \ldots, p_m > p\right) = \bigcap_{i=1}^{n} \Phi(p_i) = \bar{\Phi}(p)^m$. Taking the first derivative we get the result. □

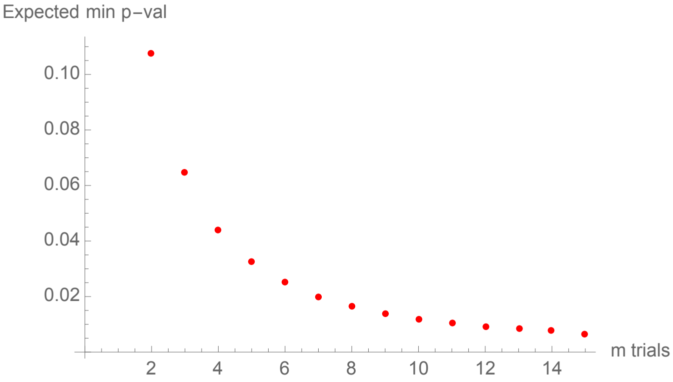

Fig. 3. The "p-hacking" value across m trials for $p_M = .15$ and $p_s = .22$.

II. Application and Conclusion

- One can safely see that under such stochasticity, to get what people mean by 5% confidence (and the inferences they get from it), they need 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.

Acknowledgment

Marco Avellaneda, Yaneer Bar-Yam, friendly people 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. 19 313–19 317, 2013.

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