Random Thoughts on p-values

A.H. Welsh

The Australian National University
Research School of Finance, Actuarial Studies and Statistics

Alan.Welsh@anu.edu.au

Abstract

Much of the recent commentary on p-values has been negative with some journals and societies taking strong stands against their use. They are even blamed for the reproducibility crisis in some fields of applied statistics. Some of this may be due to misunderstanding of what p-values are. We consider a very simple example for which we can do various calculations exactly and use the results of these calculations to gain insight into the properties of p-values. These insights then help clarify the nature and interpretation of p-values.

Introduction

Suppose that $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ are independent and identically distributed nonnegative random variables that we model as exponential with rate parameter θ and, for some $\theta_0 > 0$, we want to test $\mathrm{H}_0 : \theta = \theta_0$ against $\mathrm{H}_A : \theta < \theta_0$.

Let G_n be the distribution function of $2\theta n\bar{y} = 2\theta \sum_{i=1}^n y_i \sim \chi^2_{2n}$. A level α Neyman-Pearson test of H_0 has rejection region

$$\{\boldsymbol{y} : 2\theta_0 n\bar{y} \geq G_n^{-1}(1-\alpha)\} = \{\boldsymbol{y} : 1 - G_n(2\theta_0 n\bar{y}) \leq \alpha\}$$

and the p-value is

$$p(\bar{y}, \theta_0, n) = 1 - G_n(2\theta_0 n\bar{y}).$$

We can interpret the p-value as

- the probability under H_0 of observing a value of the test statistic ($\bar{y}$) at least as extreme as that actually observed
- a test statistic (a convenient transformation of $\bar{y}$) defined on $[0, 1]$ such that small values are evidence against H_0.

As noted by Kuffner and Walker (2019), the second interpretation (a test statistic) makes it more more clear than the first (a probability) that p-values are random.

Distribution of p-values

Let $H_n(\cdot, \theta)$ be the actual distribution function of $2\theta n\bar{y}$ and $h_n(y, \theta) = H_n'(y, \theta)$ be the density of H_n. Write $\theta = \rho\theta_0$ with $0 < \rho \leq 1$. Then, for $0 \leq u \leq 1$, the distribution function of the p-value is

$$\begin{aligned} K(u; \rho, n) &= \Pr\{\boldsymbol{y} : 1 - G_n(2\theta_0 n\bar{y}) \leq u | \theta\} \\ &= \Pr\{\boldsymbol{y} : 2\theta n\bar{y} \geq \rho G_n^{-1}(1-u) | \theta\} \\ &= 1 - H_n\{\rho G_n^{-1}(1-u), \theta\}, \end{aligned}$$

the quantile function is

$$K^{-1}(u; \rho, n) = 1 - G_n\{\rho^{-1} H_n^{-1}(1-u, \theta)\}$$

and the density function is

$$k(u; \rho, n) = \frac{\rho h_n\{\rho G_n^{-1}(1-u), \theta\}}{g_n\{G_n^{-1}(1-u)\}}.$$

When the model is correct, $H_n(y, \theta) = G_n(y)$. Under H_0, $\rho = 1$ and, as is well-known, the p-value has a uniform distribution for all n. Under H_A, $\rho < 1$, the p-value distribution depends on n and, as shown in Figure 1, places more probability in the lower tail.

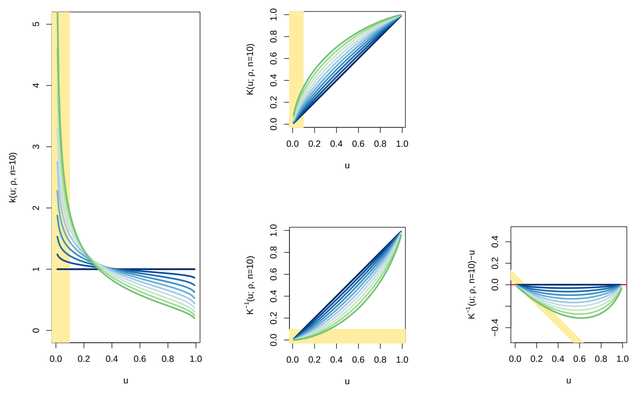

Figure 1: Density, distribution, quantile and detrended quantile functions of p-values when ρ = 1 (blue), 0.975, 0.95, 0.925, 0.90, 0.875, 0.85, 0.825, 0.80, 0.775 (green) and n = 10. The buff strip corresponds to the rejection region for a level $\alpha = 0.1$ test. Detrending rotates the quantile function through 45°.

The power of the test (probability of rejecting H_0) is

$$\text{power}(\alpha, n, \rho) = K(\alpha; \rho, n) = 1 - H_n\{\rho G_{2n}^{-1}(1-\alpha), \theta\}$$

For $\rho = 1$, $\text{power}(\alpha, n, 1) = \alpha$ for all n and α. Decreasing ρ and increasing α and/or n all increase the power; see Figure 2.

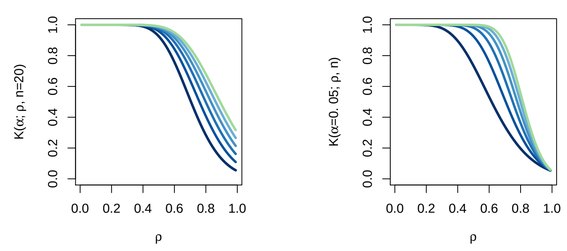

Figure 2: Power of the test when (i) α = 0.05 (blue), 0.1, 0.15, 0.2, 0.25, 0.3 (green) and n = 10 and (ii) α = 0.05 and n = 10 (blue), 20, 30, 40, 50, 60 (green). Increasing α and/or n increases the power.

Inflating the level of the test

Inflating the level of a test increases the probability of finding effects both when there are none (type I error) and when they are present. The level can be inflated when H_0 is true if $H_n \neq G_n$ and the distribution of the p-value puts more weight in the lower tail. This can occur if we use a poor approximation to compute the p-value, assume the wrong model or represent the analysis process incorrectly (e.g. ignore selection).

Incorrect model 1: If the true distribution is gamma(κ, λ), then $2\lambda n\bar{y} \sim$ gamma$(2n\kappa/2, 1/2) \sim \chi^2_{2n\kappa}$ and hence

$$2\theta n\bar{y} = \frac{\theta}{\lambda} 2\lambda n\bar{y} \sim \frac{\theta}{\lambda}\chi^2_{2n\kappa}.$$

The meaning of the null hypothesis H_0 depends on how we identify θ. We can make different choices:

1. If we identify the rate as the reciprocal of the mean (because the p-value is derived from $1/\bar{y}$), we have $\theta = \lambda/\kappa$ so
$$2\theta n\bar{y} \sim \frac{1}{\kappa}\chi^2_{2n\kappa}$$
and $H_n(y; \theta) = G_{n\kappa}(\kappa y)$. Under H_0, we have $\lambda = \kappa\theta_0$.
2. If we identify the rate as the reciprocal of the scale (because the rate is defined this way in the exponential distribution) and measure the scale by the standard deviation, we have $\theta = \lambda/\kappa^{1/2}$ so
$$2\theta n\bar{y} \sim \frac{1}{\kappa^{1/2}}\chi^2_{2n\kappa}$$
and $H_n(y; \theta) = G_{n\kappa}(\kappa^{1/2} y)$. Under H_0, we have $\lambda = \kappa^{1/2}\theta_0$.
3. If we identify the rate as the rate parameter of the gamma distribution, we have $\theta = \lambda$ so $H_n(y; \theta) = G_{n\kappa}(y)$. Under H_0, we have $\lambda = \theta_0$.

The three cases are shown in Figure 3.

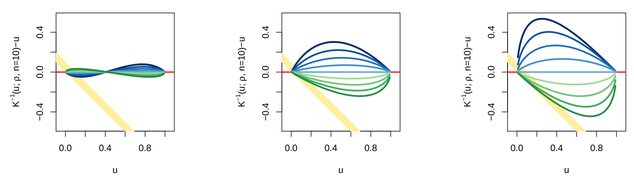

Figure 3: Detrended quantile functions of the p-value distribution under H_0 when the true distribution is gamma(κ, λ) with κ = 0.6 (blue), 0.7, 0.8, 0.9, 1.0, 1.2, 1.3, 1.4 (green) for $n = 10$. The rate θ is taken to be 1/mean, 1/sdev, and λ. For (i), large κ is conservative but small kappa increases α slightly; for (ii) and (iii), small κ is conservative but large κ increases α considerably.

Incorrect model 2: For a second example consider the ϵ mixture distribution

$$(1-\epsilon)\,\text{exponential}(\eta) + \epsilon\,\text{contamination}(\mu_c, \sigma_c^2)$$

which has moments

$$\begin{aligned} \mu &= \frac{1}{\eta} + \epsilon\Big(\mu_c - \frac{1}{\eta}\Big) \\ \sigma^2 &= \frac{1}{\eta^2} + \epsilon\Big(\sigma_c^2 - \frac{1}{\eta^2}\Big) + \epsilon(1-\epsilon)\Big(\mu_c - \frac{1}{\eta}\Big)^2. \end{aligned}$$

The exact distribution of the sample mean is now not known so we use the lognormal approximation $\log(\bar{y}) \sim N\big(\log(\mu), \sigma^2/n\mu^2\big)$ under which the distribution function of $T = 2\theta n\bar{y}$ is

$$H_n(y; \theta) = \Pr(2\theta n\bar{y} \leq y) = \Phi\Big\{\frac{\mu n^{1/2}}{\sigma}\log\Big(\frac{y}{2\theta\mu n}\Big)\Big\}.$$

To make the approximate distribution under H_0 exactly uniform, we compute the p-value using the log-normal approximation $1 - G_n(2\theta_0 n\bar{y}) \approx 1 - \Phi\{n^{1/2}\log(\theta_0\bar{y})\}$. The results depend on whether we identify θ with $1/\mu$, $1/\sigma$ or η respectively.

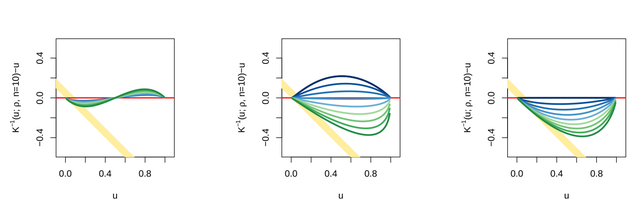

Figure 4: Detrended quantile functions of the p-value distribution under H_0 when the true distribution is an ϵ-mixture of an exponential(η) distribution and a distribution with mean μ_c and variance σ_c^2 with ϵ = 0 (blue), 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4 (green) for $n = 10$. The rate θ is taken to be 1/mean, 1/sdev, and η. For (i) and (iii), large ϵ increases α; for (ii), small ϵ is conservative but large ϵ increases α considerably.

Ignoring selection: Suppose we do m independent studies and then use the smallest p-value. The distribution function of the p-value is

$$\begin{aligned} K_m(u; \boldsymbol{\rho}, \boldsymbol{n}) &= \Pr\{\boldsymbol{y} : \min_{j \leq m} p_j \leq u | \theta\} \\ &= 1 - \prod_{j=1}^m H_{n_j}\{\rho G_{n_j}^{-1}(1-u), \theta\}. \end{aligned}$$

If H_0 holds and the model is correct, $H_{n_j} = G_{n_j}$, we have

$$K_m(u; 1, \boldsymbol{n}) = 1 - (1-u)^m,$$

the distribution function of the beta$(m, 1)$ distribution.

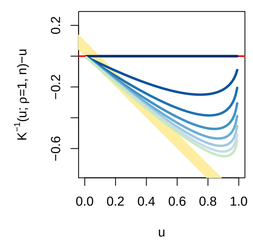

Figure 5: Quantile function of the p-value distribution under H_0 when the p-value is the smallest from m = 1 (blue), 2, 3, 4, 5, 6, 7, 8 (light blue) p-values for all n. Increasing m increases α.

Accumulation of evidence

The test should be recast from testing

- $\mathrm{H}_0 : \theta = \theta_0$ against $\mathrm{H}_0 : \theta < \theta_0$ using n independent observations (which sounds reasonable) to
- testing $\mathrm{H}_0 : k(u) = 1$ against $\mathrm{H}_0 : k(u) = k(u; \rho, n)$, $0 < \rho < 1$, for which using a single observation $p(\bar{y}, \theta_0, n)$ is obviously less reasonable.

We need to replicate studies, i.e. try to get p-values p_j from $j = 1, \ldots, m$ independent replicate studies. A set of simulated independent p-values is shown in Table 1.

0.332 (10)	0.381 (5)	0.412 (20)	0.576 (10)	0.581 (12)
0.028 (8)	0.033 (20)	0.081(15)	0.101 (10)	0.447 (15)
0.012 (20)	0.041 (15)	0.165 (12)	0.347 (15)	0.555 (20)

Table 1: Simulated p-values from $m = 15$ independent studies with $\rho = 0.8$. The sample sizes are in brackets after each p-value and significant results are higlighted in blue.

There are many ways to combine the p-values. Fisher (1925) suggested using

$$-2\sum_{j=1}^m \log(p_j) \sim \chi^2_{2m} \qquad \text{under } \mathrm{H}_0.$$

Fisher's statistic gives p-values of 0.619, 0.031 and 0.003 for the first $m = 5, 10$ and 15 p-values from Table 1 respectively. Graphical methods are usually more informative so we suggest using a detrended uniform QQ-plot of the m p-values. To aid interpretation, we include pointwise upper and lower bounds in the plot. Since the kth uniform order statistic has a beta$(k, m+1-k)$ distribution, we use the detrended 0.975 and 0.025 quantiles of this distribution as pointwise limits under H_0. Plots for the first $m = 5, 10$ and 15 p-values are shown in Figure 6.

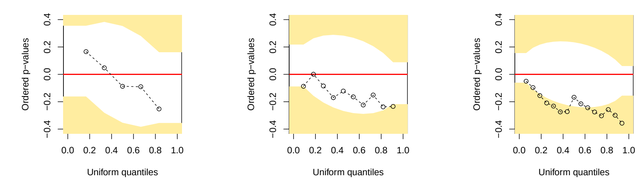

Figure 6: Detrended uniform QQ-plots of the p-values shown in Table 1 from the first $m = 5$, 10 and 15 independent studies.

The departure from uniformity is clear in all three figures (c.f. Fisher's combined p-value). In this example, it is less the lack of small p-values than the fact that there are not enough large ones that means the uniform distribution does not hold!

Conclusions

- A p-value is a statistic with a sampling distribution.
- The information about H_0 is in the sampling distribution of the p-value, not the value itself.
- A Neyman-Pearson test reaches a decision with $m = 1$, but we can make more sound inferences by exploring the sampling distribution of p-values.
- To explore a distribution, we need independent replicate observations (in this case, p-values).
- It is easy for things to go wrong (e.g. assume incorrect distribution, ignore p-hacking) and then make incorrect inferences.

References

[1] R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh, 1925.

[2] T. A. Kuffner and S.G. Walker. Why are p-values controversial. *The American Statistician*, 73:1–3, 2019.

[3] D. J. Murdoch, Y.-L. Tsai, and J. Adcock. p-values are random variables. *The American Statistician*, 62:242–245, 2008.