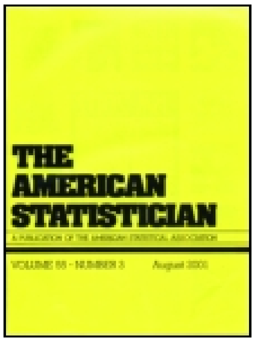


The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

P-Values are Random Variables

Duncan J Murdoch, Yu-Ling Tsai & James Adcock

To cite this article: Duncan J Murdoch, Yu-Ling Tsai & James Adcock (2008) P-Values are Random Variables, The American Statistician, 62:3, 242-245, DOI: 10.1198/000313008X332421

To link to this article: http://dx.doi.org/10.1198/000313008X332421

Published online: 01 Jan 2012.

P-Values are Random Variables

Duncan J. Murdoch, Yu-Ling Tsai, and James Adcock

P-values are taught in introductory statistics classes in a way that confuses many of the students, leading to common misconceptions about their meaning. In this article, we argue that *p*-values should be taught through simulation, emphasizing that *p*-values are random variables. By means of elementary examples we illustrate how to teach students valid interpretations of *p*-values and give them a deeper understanding of hypothesis testing.

KEY WORDS: Empirical cumulative distribution function (ECDF); Histograms; Hypothesis testing; Teaching statistics.

1. INTRODUCTION

Nowadays many authors show students simulations when teaching confidence intervals in order to emphasize their randomness: a true parameter is held fixed, and approximately 95% of the simulated intervals cover it (e.g., Figure 1; similar figures appear in many texts).

In this article, we argue that a simulation-based approach is also a good way to teach students *p*-values. Students will learn the logic behind rejecting the null hypothesis (H_0) when p is small, instead of simply memorizing recipes; they will not learn incorrect interpretations such as "the *p*-value is the probability that H_0 is true." In our approach, it is emphasized that *p*-values are transformations of a test statistic into a standard form, that is, *p*-values are random variables.

A standard definition of the *p*-value is that it is "the probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed." Since test statistics are constructed in order to quantify departures from H_0, we reject H_0 when p is small because "the smaller the *p*-value, the stronger the evidence against H_0 provided by the data." (Both quotes are from Moore 2007, p. 368.) We feel that this standard presentation obscures the randomness of p, and its status as a test statistic in its own right. Defining p as a probability leads students to treat it as a probability, either the probability of H_0 being true or the probability of a type I error: and neither interpretation is valid.

Criticism of invalid interpretations of *p*-values is not new. Schervish (1996) pointed out that logically they do not measure support for H_0 on an absolute scale; Hubbard and Bayarri (2003) pointed out the inconsistency between *p*-values and fixed Neyman–Pearson α levels. Sackrowitz and Samuel-Cahn (1999) emphasized the stochastic nature of *p*-values, and recommended calculation of their expected value under particular alternative hypotheses.

In this article our goal is not to discuss the foundations of inference as those articles do. Instead, we want to present an alternative method of teaching *p*-values, suitable for both introductory and later courses. In early courses students will learn the logic behind the standard interpretation, and later they will be able to follow foundational arguments and judge the properties of asymptotic approximations.

The remainder of this article is organized as follows. Section 2 introduces *p*-values and our recommended way to teach them via simulation and plotting of histograms. Storey and Tibshirani (2003) used histograms of *p*-values from a collection of tests in genome-wide experiments in order to illustrate false discovery rate calculations; in contrast, our histograms are all based on simulated data. Section 3 presents a pair of examples. The first is a simple two-sample t-test, where interpretation of the distribution of p under the null and alternative hypotheses is introduced. Our second example shows a highly discrete test, where plotting histograms breaks down, and more sophistication is needed from the students. Finally, we list a number of other examples where *p*-values can be explored by simulation. The plots for this article were produced by simple R (R Development Core Team 2007) scripts which are available on request.

2. TEACHING *P*-VALUES BY MONTE CARLO SIMULATIONS

Students in introductory classes may have only recently been introduced to the concept of random variables. For these students, we have found that Monte Carlo simulations are an effective way to illustrate randomness. They are likely to be familiar with histograms, so we recommend presenting summary results in that form. A histogram of simulated values sends the message to students that *p*-values are random and they have a distribution that can be studied. Empirical cumulative distribution functions (ECDFs) are another possibility for more sophisticated students, and are preferable with discrete data, as we show later.

Duncan Murdoch is Associate Professor, University of Western Ontario, London, Ontario, Canada N6A 5B7 (E-mail: *murdoch@stats.uwo.ca*). Yu-Ling Tsai is Assistant Professor, University of Windsor, Windsor, Ontario, N9B 3P4 (E-mail: *ytsai@uwindsor.ca*). James Adcock is Lecturer, University of Western Ontario, London, Ontario, Canada N6A 5B7 (E-mail: *jadcock@stats.uwo.ca*). This work was supported in part by an NSERC Research Grant to the first author. The authors thank the editor, associate editor, and referees for helpful comments.

DOI: 10.1198/000313008X332421

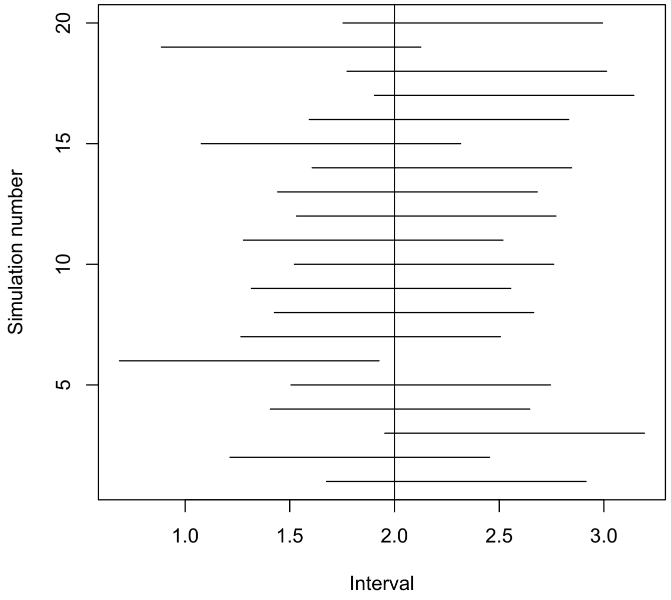


Figure 1. Twenty simulated confidence intervals around a true mean of 2.

3. EXAMPLES

In this section we start with continuous data, where the p-value has a Uniform(0,1) distribution under a simple null hypothesis. Our second example will deal with discrete data.

3.1 Example 1: A Simple t-test

We first consider a one-sample t-test with $N(\mu, \sigma^2)$ data with

$$\begin{aligned} H_0 &: \mu \leq 0 \\ H_a &: \mu > 0. \end{aligned} \qquad (1)$$

The test statistic is $T = \bar{X}/(s/\sqrt{n})$, where $\bar{X}$ is the sample mean, s is the sample standard deviation, and n is the sample size. Under the boundary case $\mu = 0$ in H_0, we have $T \sim t_{(n-1)}$.

We simulated 10,000 experiments with groups of $n = 4$ observations each. The true distribution was $N(\mu, 1)$, with $\mu = -0.5, 0, 0.5$, or 1.

When presenting this in class, we would start by discussing the point null $H_0 : \mu = 0$.

- Under the point null hypothesis (top right of Figure 2), the histogram of p-values looks flat and uniformly distributed over the interval [0, 1]. This result is exactly true: the p-value is the probability integral transform of the test statistic (Rice 2007, p. 63).
- Under the alternative hypothesis (bottom row of Figure 2), the distribution of p-values is not uniform. It will be obvious to students that the chance of observing a p-value less than $\alpha = 0.05$ under the alternative hypothesis is higher than under the null hypothesis, and this effect is more pronounced as μ increases. The concept of power can be introduced at this point. Donahue (1999) gave further discussion about the interpretation of the p-value under H_a.

Once the students have grasped this basic behavior, we introduce the possibility of $\mu < 0$ in H_0.

- If $\mu < 0$, the distribution of the p-values will be concentrated near 1 (top left of Figure 2).
- The behavior under the previously considered cases is identical, illustrating that our hypotheses do not determine the distribution, the parameters do.

3.2 Example 2: Discrete Data

In the previous example, the test statistic was drawn from a continuous distribution. Things become more complicated when the test statistic has discrete support. For example, to test for independence between two categorical variables that label the rows and columns of a two-way table, we may use a chi-square test or Fisher's exact test.

Consider a comparison of two drugs for leukemia (Table 1). We want to test if the success and failure probabilities are the same for Prednisone and Prednisone+VCR. The null hypothesis is

$$H_0 : \text{Both treatment groups have equal success probability.}$$

Here there is a choice of reference null distribution, because the response rate under H_0 is a nuisance parameter. We avoid this issue by conditioning on the margins, and use a hypergeometric simulation with parameters expressed in R notation as $m = 21$, $n = 42$, and $k = 52$ (R Development Core Team 2007). For example, a typical simulated table had 17 successes and 4 failures for Prednisone, with 35 successes and 7 failures for Prednisone + VCR.

Both chi-square and Fisher's tests were performed for each simulated table. Both tests were two-sided, with tables of lower probability than the observed one taken as more extreme in Fisher's test. Both tests would consider the simulated table mentioned earlier to be less extreme than the observed one: the chi-square test because the observed counts are closer to the expected counts, and Fisher's test because the probability of the simulated outcome is larger than that of the observed one.

The results of 1,000 simulations are shown in Figure 3. It is no longer true that the distribution of p-values is uniform: they have a discrete distribution, and the histograms are not useful. Westfall and Wolfinger (1997) discussed the effects of discreteness on p-value distributions. Here we study this (as they did,

Table 1. Observed and expected frequencies for leukemia (Tamhane and Dunlop 2000, p. 326).

	Success	Failure	Row Total
Prednisone	14 17.33	7 3.67	21
Prednisone + VCR	38 34.67	4 7.33	42
Column Total	52	11	63

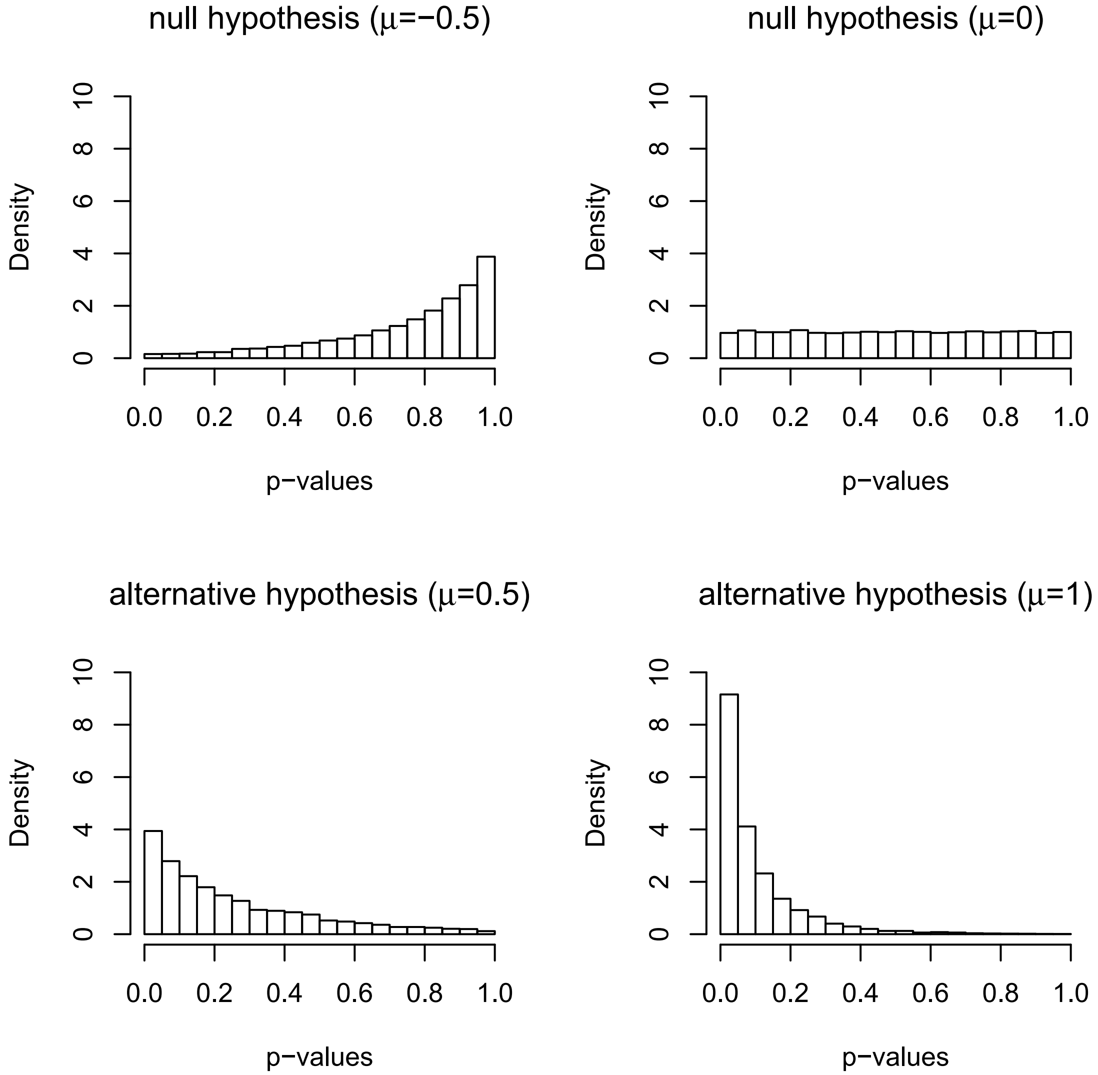


Figure 2. Histograms of 10,000 simulated p-values under the null hypothesis with $\mu = -0.5$ (top left) and $\mu = 0$ (top right), or the alternative $\mu = 0.5$ (bottom left), $\mu = 1$ (bottom right).

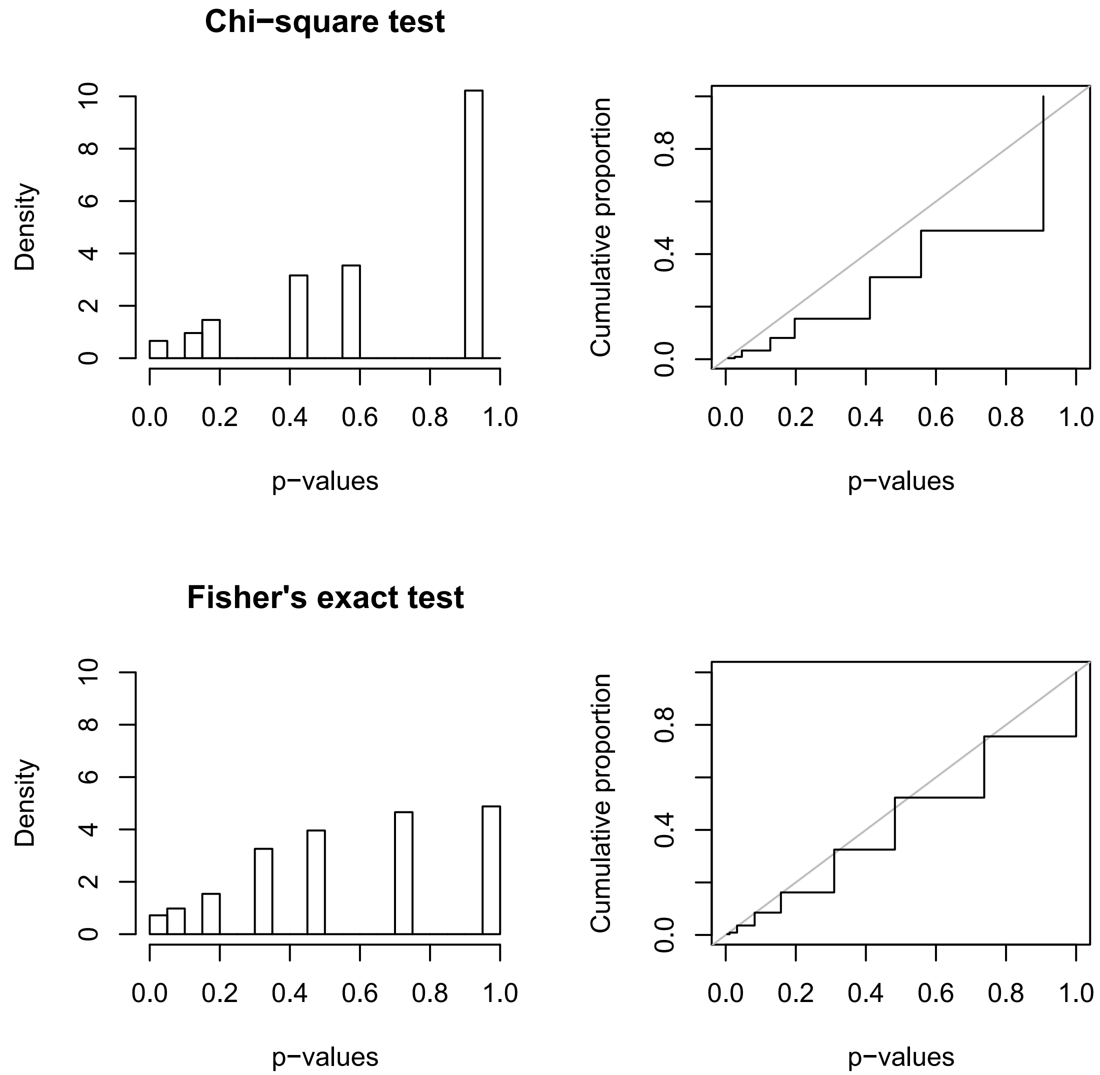
in part) by looking at ECDF plots. The ECDF should be close to the diagonal if the rejection rate matches the critical value. We see that both tests give distributions of p-values that are very discrete, but Fisher's test appears to be slightly better calibrated. Indeed, since the hypergeometric reference distribution is the distribution for which it is "exact," the true CDF should just touch the diagonal at each stairstep: students can see the Monte Carlo error by the fact that it sometimes oversteps.

Students presented with this example will learn that not all p-values are achievable with discrete data: it is not unusual to see large gaps in their distributions. Advanced students will learn a way to study the quality of asymptotic approximations. An instructor could follow up our conditional simulation with an unconditional one, and illustrate reasons why one might prefer the chi-square p-value over Fisher's.

3.3 For Further Study

Besides the above examples, the same approach can be used in many other situations, such as the following:

- Display histograms based on varying sample sizes to show that the null distribution remains uniform, but power increases with sample size.
- Compare p-values under alternative test procedures, and in situations where the assumptions underlying the test are violated. Both distortions to the null distribution and changes to the power of the test could be explored. How robust are one or two sample t-tests? What is the effect of Welch's correction for unequal variances on the p-values arising in a two-sample t-test, in situations where the correction is applied even though the variances are equal, and where we do not use the correction when it should have been used? How do nonparametric tests compare to parametric ones?
- Monte Carlo p-values can be illustrated in cases where the null distribution is obtained by bootstrapping.
- As with the chi-square test in Example 2, we can explore the accuracy of asymptotic approximations in other tests by studying the distributions of nominal p-values.



Chi−square test

Fisher's exact test

Figure 3. Histogram and ECDF plots of 1,000 null p-values for chi-square (top) and Fisher's exact tests (bottom). The entries of the two-way table are hypergeometrically distributed.

- For the multiple testing problem, we could illustrate the distribution of the smallest of n p-values, and the distribution of multiplicity-adjusted p-values, under various assumptions about the joint distributions of the individual tests.

4. CONCLUDING REMARKS

To emphasize the point that a p-value is *not* the probability that H_0 is true, an instructor need only point to the top right plot in Figure 2: here H_0 is certainly true, but the p-value is uniformly distributed between 0 and 1.

We believe students should look at the distributions of simulated p-values. When they understand that a p-value is a random variable, they will better understand the reasoning behind hypothesis testing, the proper interpretation of the results, and the effects of violated assumptions.

[Received August 2007. Revised April 2008.]

REFERENCES

Donahue, R. M. J. (1999), "A Note on Information Seldom Reported via the *P* Value," *The American Statistician*, 53, 303–306.

Hubbard, R., and Bayarri, M.J. (2003), "Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing," *The American Statistician*, 57, 171–182.

Moore, D. S. (2007), *The Basic Practice of Statistics* (4th ed.), New York: Freeman.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing.

Rice, J. A. (2007), *Mathematical Statistics and Data Analysis* (3rd ed.), Belmont, CA: Duxbury.

Sackrowitz, H., and Samuel-Cahn, E. (1999), "P values as Random Variables—Expected P-Values," *The American Statistician*, 53, 326–331.

Schervish, M. J. (1996), "P Values: What They are and What They are Not," *The American Statistician*, 50, 203–206.

Storey, J., and Tibshirani, R. (2003), "Statistical Significance for Genome-Wide Studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Tamhane, A. C., and Dunlop, D. D. (2000), *Statistics and Data Analysis: From Elementary to Intermediate,* Englewood Cliffs, NJ: Prentice-Hall.

Westfall, P. H., and Wolfinger, R. D. (1997), "Multiple Tests with Discrete Distributions," *The American Statistician*, 51, 3–8.