Explorations in statistics: hypothesis tests and *P* values

Douglas Curran-Everett

Division of Biostatistics and Bioinformatics, National Jewish Health, and Department of Biostatistics and Informatics and Department of Physiology and Biophysics, University of Colorado Denver, Denver, Colorado

Submitted 30 December 2008; accepted in final form 8 February 2009

Curran-Everett D. Explorations in statistics: hypothesis tests and *P* values. *Adv Physiol Educ* 33: 81–86, 2009; doi:10.1152/advan.90218.2008.—Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This second installment of *Explorations in Statistics* delves into test statistics and *P* values, two concepts fundamental to the test of a scientific null hypothesis. The essence of a test statistic is that it compares what we observe in the experiment to what we expect to see if the null hypothesis is true. The *P* value associated with the magnitude of that test statistic answers this question: if the null hypothesis is true, what proportion of possible values of the test statistic are at least as extreme as the one I got? Although statisticians continue to stress the limitations of hypothesis tests, there are two realities we must acknowledge: hypothesis tests are ingrained within science, and the simple test of a null hypothesis can be useful. As a result, it behooves us to explore the notions of hypothesis tests, test statistics, and *P* values.

power; R; significance test; software; test statistic

THIS SECOND ARTICLE in *Explorations in Statistics* (see Ref. 8) provides an opportunity to explore test statistics and *P* values, two concepts integral to the test of a scientific hypothesis. What is a scientific hypothesis? An idea that can be tested. By tradition, this hypothesis is called the null hypothesis. The adjective null can be misleading: this hypothesis need not be one of no difference. In order to test a null hypothesis, we must first define the hypothesis. Using data from the subsequent experiment, we then compute the value of some test statistic and compare that observed value T to some critical value T^* chosen from the distribution of the test statistic that is based on the null hypothesis. If T is more extreme than T^*, that is unusual if the null hypothesis is true, and we are entitled–on statistical grounds–to question the null hypothesis.

Hypothesis tests pervade science, but statisticians have long stressed their limitations (1, 2, 5–7, 15–17, 19–24, 26–29, 31, 32, 42, 45, 49, 50). Despite these limitations, there are two practical considerations that have persisted: the simple test of a null hypothesis can be useful (10, 44), and hypothesis tests are ingrained within science. Before we explore the notions of hypothesis tests, test statistics, and *P* values, it behooves us to understand why hypothesis tests pervade science.

A Brief History of Hypothesis Tests

The earliest known hypothesis test was the Trial of the Pyx, a periodic ritual[1] of the Royal Mint (London) that had become established by 1279 (47). Each time the Mint made coins, a small number of them went into the Pyx, a wooden box. When a Trial was convened, an independent jury of goldsmiths compared these select coins to standards in order to assess whether the coins were within prescribed tolerances for weight and composition. In each Trial of the Pyx, the implicit null and alternative hypotheses, H_0 and H_1, were

H_0: The coins are within the prescribed tolerances.

H_1: The coins are outside the prescribed tolerances.

By the 1700s, astronomers routinely used hypothesis tests to decide if discrepant celestial measurements should be discarded (18, 46). At issue: precise information about the position of the moon for purposes of navigation (46). To do one of these hypothesis tests, an astronomer compared the discrepant value to the *law of errors*, the distribution of errors about the true lunar position (18).[2] For each questionable measurement, the implicit null and alternative hypotheses were

H_0: The measurement is within the limits of error.

H_1: The measurement is outside the limits of error.

As in the Trial of the Pyx, each astronomical hypothesis test had a binary outcome: the measurement either was or was not within some allowable deviation.

From the 1800s through the early 1900s, when a mathematician or physicist wrote about whether some event could be attributed to chance alone, he wrote about the odds of that event (18, 36, 46). The greater the odds, the more likely the event was due to something other than chance–random variation–alone (Fig. 1). Table 1 lists odds that were sufficient to pique scientific interest between 1837 and 1908.

In his landmark *The Probable Error of a Mean* (48), William Sealy Gosset, a chemist who wrote under the pseudonym Student because he worked for the Guinness brewery (40), outlined a procedure that would evolve into the one-sample *t*-test. To illustrate this procedure, Gosset used data from a paper published by Cushny and Peebles (11) in *The Journal of Physiology:*

> First let us see what is the probability that [drug *A*] will on the average give increase of sleep. [Looking up the ratio of the sample mean to the sample standard deviation] in the table for ten experiments we find by interpolating. . .the odds are ·887 to ·113 that the mean is positive.
>
> That is about 8 to 1 and would correspond to the normal curve to about 1·8 times the probable error. It is then very likely that [drug *A*] gives an increase of sleep, but would occasion no surprise if the results were reversed by further experiments.

Address for reprint requests and other correspondence: D. Curran-Everett, Div. of Biostatistics and Bioinformatics, M222, National Jewish Health, 1400 Jackson St., Denver, CO 80206 (e-mail: EverettD@NJHealth.org).

[1] In the past, a Trial was announced every 3–4 years. Today, the Trial is convened every year.

[2] The *law of errors* is just a normal distribution, a probability distribution developed independently by De Moivre, Laplace, Legendre, and Gauss (18).

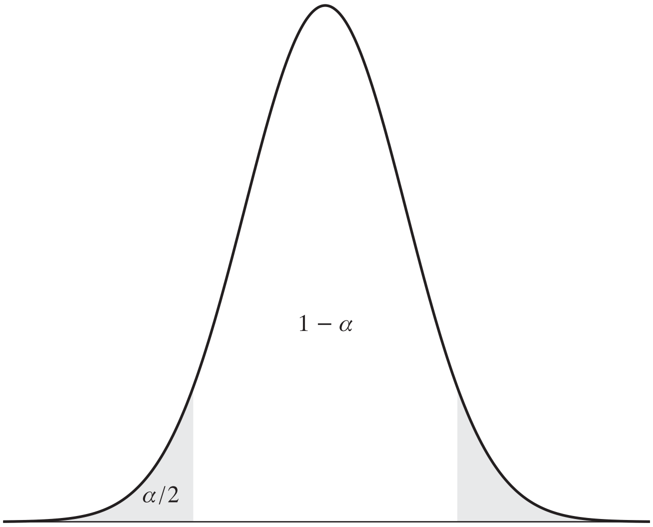

Fig. 1. Normal distribution. The critical significance level α (gray area) is the probability we reject a null hypothesis when it is true. The ratio of $1 - \alpha$ to α is the odds against rejecting a true null hypothesis. Table 1 lists some critical significance levels and odds reported since 1837.

In current jargon, the phrase *1·8 times the probable error* means that $t = 1.8$.

Today, when a scientist writes about whether some event can be attributed to chance, she writes not about the odds but about the probability of that event if chance alone is at work. This convention can be traced to *Statistical Methods for Research Workers* (13), published in 1925 by Sir Ronald Aylmer Fisher (18, 52).[3] In 1919, after 4 years of teaching mathematics and physics in public schools, Fisher went to work as a statistician at the Rothamsted Experimental Station, a facility that conducted research in agronomy and biology (4, 40, 52). It was during his early tenure at Rothamsted that Fisher, inspired in part by Gosset's *The Probable Error of the Mean*, recognized that scientists needed a practical guide to statistical methods. *Statistical Methods for Research Workers* gave them what they needed (35).

Just as others defined the magnitude of a deviation that they regarded as beyond chance, so too did Fisher in *Statistical Methods for Research Workers.* Fisher opted for 1 in 20: only 5% of possible values of the event are more extreme than this benchmark. In *Statistical Methods for Research Workers,* Fisher cites his 0.05 benchmark three times. The latter two, on p. 79 and 101–102, allude to this first (p. 46–47):

> In practical applications we do not so often want to know the frequency at any distance from the centre as the total frequency beyond that distance; this is represented by the area of the tail of the curve cut off at any point. . . A deviation exceeding the standard deviation occurs about once in three trials. Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation only once in 370 trials. . . The value for which $P = \cdot05$, or 1 in 20, is 1·96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.

In our guidelines for reporting statistics (Ref. 9, *guideline* 2), Dale Benos and I wrote that most researchers adhere to tradition and define the critical significance level α–the benchmark for the limits of random variation–to be 0.05. It is this passage in *Statistical Methods for Research Workers* that was the genesis of the tradition.

Statistical Methods for Research Workers filled such a void within the scientific community that two things happened. First, researchers started to use statistical procedures from the book, but without fully understanding the underlying concepts, they sometimes misused the procedures (18, 53). And second, the notion of a significance level of 0.05 approached the level of doctrine despite subsequent but less obvious elaborations by Fisher:

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard of significance at the 5 per cent. point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance.
>
> [Ref. 14 (1926)]

> The attempts that have been made to explain the cogency of tests of significance in scientific research. . . seem to miss the essential nature of such tests. A [person] who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. . . However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance.
>
> [Ref. 16 (1956)]

In contrast to Fisher who defined a null hypothesis but no alternative hypothesis (18), Jerzy Neyman and Egon Pearson advocated a paradigm of hypothesis testing that involved making a decision about competing null and alternative hypotheses (18, 33, 38, 39). In this paradigm, power, the probability that we reject some null hypothesis given that it is false, plays a prominent role. Neyman and Pearson argued that the

Table 1. *Significance levels between 1837 and 1908*

Year	Person	$1 - \alpha$	α	Odds	Reference
1837	Poisson	0.9953	0.0047	212	Matthews (36)
1874	Hirschberg	0.916	0.084	11	
1877	Liebermeister	0.8333	0.1667	5	
1885	Edgeworth	0.93	0.07	13	Edgeworth (12)
	Edgeworth	0.997	0.003	332	
1908	Gosset	0.9801	0.0199	49	Student (48)
	Gosset	0.887	0.113	8	

The odds against rejecting a true null hypothesis is $(1 - \alpha)/\alpha$ (see Fig. 1).

[3] The impact of Fisher on statistics and science is legendary (3, 18, 25, 30, 34, 35, 43, 51–53). His collected papers are posted at http://www.adelaide.edu.au/library/special/digital/fisherj/.

selection of the critical significance level α, the probability that we reject a true null hypothesis, and β, the probability that we fail to reject a false null hypothesis, depended on the costs associated with committing each kind of error (18, 33).

Although the philosophical differences between the Fisher and Neyman-Pearson strategies led to some fierce exchanges between the protagonists, from a practical perspective, the strategies are complementary (18, 33). In fact, in science, these strategies have been blended (18). Despite their philosophical difference, it is fitting that Fisher, Neyman, and Pearson all believed that statistics provided a means by which to learn (33).

With this brief history, we are ready to begin our explorations of contemporary hypothesis tests, test statistics, and *P* values.

The Null Hypothesis: Controlling Mistakes

When we make an inference about a null hypothesis, we can make a mistake. We can reject a true null hypothesis, an error of the first kind, or we can fail to reject a false null hypothesis, an error of the second kind (37, 38).[4] Our challenge is to balance two conflicting objectives: reduce the risk that we find an experimental effect when it does not exist but maintain the likelihood that we detect an experimental effect when it does exist.

The chance that we make an error of the first kind is just the probability that we reject the null hypothesis H_0 given that H_0 is true:

$$\Pr\{\text{reject } H_0 \mid H_0 \text{ is true}\} \quad .$$

We control the chance that we make this kind of error when we define the critical significance level α because

$$\alpha = \Pr\{\text{reject } H_0 \mid H_0 \text{ is true}\} \quad .$$

When we define α, we declare that we are willing to reject a true null hypothesis 100α% of the time. *Guideline* 2 (9) discusses the choice of α so it is appropriate to the goals of your study.

The chance that we make an error of the second kind is the probability that we fail to reject H_0 given that H_0 is false:

$$\Pr\{\text{fail to reject } H_0 \mid H_0 \text{ is false}\} \quad .$$

We control the chance that we make this kind of error when we define the error rate β:

$$\beta = \Pr\{\text{fail to reject } H_0 \mid H_0 \text{ is false}\} \quad .$$

Rather than define β per se, we usually define power, the probability that we reject H_0 given that H_0 is false:

$$\text{power} = 1 - \beta = \Pr\{\text{reject } H_0 \mid H_0 \text{ is false}\} \quad .$$

In general, four things affect power: the critical significance level α, the standard deviation σ of the underlying population, the sample size *n,* and the magnitude of the difference that we want to be able to detect.[5]

Before we begin our actual exploration of test statistics and *P* values, we need to review the software we will use to help us learn about these concepts.

R: Basic Operations

In the inaugural article (8) of this series, I summarized R (41) and outlined its installation. The APPENDIX here reviews this process. For this exploration, there is just one additional step: download the script Advances_Statistics_Code_P.R[6] to your Advances folder.

If you use a Mac, highlight the commands in Advances_Statistics_Code_P.R you want to submit and then press ⌘↩. If you use a PC, highlight the commands you want to submit, right-click, and then click Run line or selection. Or, highlight the commands you want to submit and then press Ctrl+R.

The Simulation: Observations and Sample Statistics

When we explored the distinction between standard deviation and standard error (8) we drew a total of 1000 random samples–each with 9 observations–from our population, a standard normal distribution with mean μ = 0 and standard deviation σ = 1 (Fig. 2). These were the observations–the data–for *samples 1*, *2*, and *1000*:

```
> # Sample Observations

    [1]   0.422    1.103    1.006    1.034    0.285   −0.647   1.235   0.912   1.825
    [2]   0.154   −0.654   −0.147    1.715    0.720    0.804   0.256   1.155   0.646
     :
 [1000]   0.560   −1.138    0.485   −0.864   −0.277    2.198   0.050   0.500   0.587
```

Each time we drew a random sample, we calculated the sample statistics listed in Table 2. These were the statistics for *samples 1*, *2*, and *1000*:

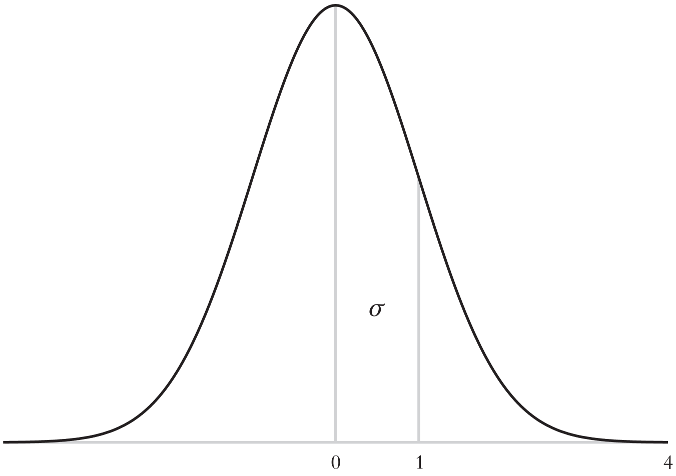

Fig. 2. The population. We can generate a standard normal distribution by transforming some random variable Y to z by the relationship $z = (Y - \mu)/\sigma$, where z represents the number of standard deviations Y is from the mean μ. [Reprinted from Ref. 8.]

[4] Errors of the first kind and second kind are known also as type I and type II errors.

[5] In light of its importance to hypothesis testing and grant applications, we will explore power in a future installment of *Explorations*.

[6] This file is available through the Supplemental Material link for this article at the *Advances in Physiology Education* website.

Table 2. *Sample statistics calculated for each random sample*

Column	Heading	Sample Statistic
1	Sample	Sample number
2	Mean	Mean $\bar{y}$
3	SD	Standard deviation *s*
4	SE	Standard error of the mean SE $\{\bar{y}\} = s/\sqrt{n}$
5	t	Observed value of $t = \bar{y}/\text{SE}\{\bar{y}\}$
6	LCI	Lower confidence interval bound
7	UCI	Upper confidence interval bound

[Reprinted from Ref. 8.]

```
> # Sample   Mean     SD      SE       t       LCI     UCI
       [,1]   [,2]    [,3]    [,4]    [,5]     [,6]    [,7]
          1  0.797   0.702   0.234   3.407    0.362   1.232
          2  0.517   0.707   0.236   2.193    0.079   0.955
          :
       1000  0.233   0.975   0.325   0.718   −0.371   0.838
```

The commands in *lines 35–63* of Advances_Statistics_Code_P.R generate the observations and compute the sample statistics. These commands are identical to those in the first script (8).

With these 1000 sets of sample statistics, we are ready to explore hypothesis tests, test statistics, and *P* values.

Hypothesis Tests: Test Statistics and P Values

When we began our statistical explorations, we wanted in part to estimate μ, the mean of our population (see Ref. 8). Only because we defined our population do we know that $\mu = 0$. Had this been a real experiment, we might have wanted to learn if some intervention affected the physiological thing[7] we cared about. In this case, we would have constructed the null and alternative hypotheses, H_0 and H_1, as

H_0: The intervention has no effect.

H_1: The intervention has an effect.

More formally, we would have written these hypotheses as

$$H_0\colon\ \mu = 0$$

$$H_1\colon\ \mu \neq 0 \quad ,$$

which translate to these statements: the sample observations are consistent with having come from a population that has a mean μ of 0, and the sample observations are consistent with having come from a population that has a mean μ other than 0. This is silly, right? We know that $\mu = 0$. We can use our knowledge that the null hypothesis is true, however, to explore the concepts behind test statistics and *P* values.

The basis for some test statistic is a comparison between what we observe in the experiment and what we expect if the null hypothesis is true. In our first theoretical experiment, what we observed was the sample mean $\bar{y} = 0.797$. Now the question is, what do we expect if the null hypothesis is true? We already know the answer. When we took 1000 samples from a population with $\mu = 0$, the sample means varied (see Ref. 8, Fig. 5), but a typical sample mean differed from the population mean by a distance of 1 SD $\{\bar{y}\}$, the standard deviation of the sample means (8). This is identical to the standard error of the mean SE $\{\bar{y}\}$. Therefore, if the null hypothesis is true, we expect the typical variation in the sample mean to be SE$\{\bar{y}\}$.

One test statistic with which we can assess whether our sample observations are consistent with having come from a population with $\mu = 0$ is the familiar *t* statistic:

$$t = \frac{\bar{y}}{\text{SE}\{\bar{y}\}} \quad , \quad \text{where SE}\,\{\bar{y}\} = s/\sqrt{n} \quad ,$$

s is the sample standard deviation, and *n* is the number of observations in the sample.[8] In a manner similar to the statistic *z* (8), *t* represents the number of standard deviations the sample mean $\bar{y}$ is from the population mean μ. If the sample mean $\bar{y}$ is far enough away from the population mean μ, then that is unusual if the null hypothesis is true.

In our first sample,

$$t = \frac{\bar{y}}{\text{SE}\{\bar{y}\}} = \frac{0.797}{0.234} = 3.407 \quad .$$

That's fabulous, but exactly how do we interpret a *t* value of 3.407? We interpret it within the context of a true null hypothesis: if the null hypothesis is true, how usual is this value of *t*? If the null hypothesis is true, we expect to observe a value of

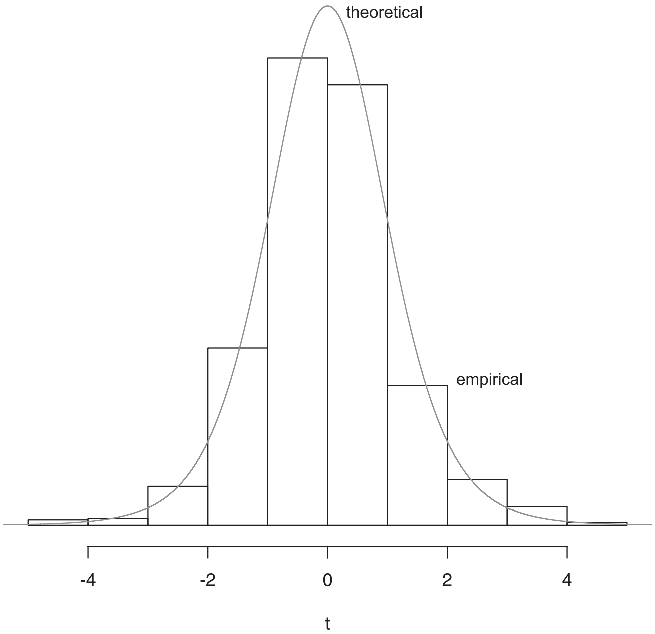

Fig. 3. Empirical (black) and theoretical (gray) distributions of *t* for 9 observations. The empirical distribution is composed of 1000 values of *t*. The commands in *lines 89–90* of Advances_Statistics_Code_P.R calculate the theoretical distribution of *t*, and the commands in *lines 92–101* create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 3: first line to Figure 3: last line.

[7] For example, L-ascorbic acid transport, differential gene expression, TNF-α, or venous capacitance in trout (see Ref. 8).

[8] We can use a *t* statistic to also assess whether two sets of sample observations are consistent with having come from the same or different populations. In this situation, we calculate *t* in the same manner but replace the single sample mean $\bar{y}$ with the difference between sample means, $\bar{y}_2 - \bar{y}_1$ (see Ref. 44).

$|t|$ at least as big as 3.407 just 1 time in 200 ($P = 0.005$).[9] By virtue of our simulation, we have 1000 values of *t*; the magnitudes of 14 values are at least as extreme as 3.407:

```
> # Sample     Mean     SD      SE        t       LCI       UCI
      [,1]     [,2]   [,3]    [,4]     [,5]      [,6]      [,7]
         1    0.797  0.702   0.234    3.407     0.362     1.232
        28    0.790  0.507   0.169    4.681     0.476     1.104
        34   −0.668  0.451   0.150   −4.442    −0.948    −0.389
       215    1.100  0.880   0.293    3.751     0.555     1.646
       238    0.923  0.736   0.245    3.763     0.467     1.379
       283   −0.576  0.387   0.129   −4.465    −0.816    −0.336
       536    0.826  0.682   0.227    3.635     0.404     1.249
       616   −0.633  0.487   0.162   −3.895    −0.935    −0.331
       645    0.671  0.558   0.186    3.607     0.325     1.017
       800   −0.981  0.726   0.242   −4.053    −1.431    −0.531
       804    1.037  0.707   0.236    4.405     0.599     1.475
       880    0.804  0.674   0.225    3.581     0.387     1.222
       925   −0.801  0.574   0.191   −4.182    −1.157    −0.445
       981   −0.744  0.610   0.203   −3.663    −1.122    −0.366
```

The commands in *lines 81–83* of Advances_Statistics_Code_P.R print the statistics for the samples in your simulation. Your number of samples will differ.

If we treat these 1000 values of *t* as observations, their empirical distribution is centered at −0.04 (median), just less than the theoretical value of 0 (Fig. 3). Five percent of the 1000 values of *t* are less than −1.773 and five percent are greater than 1.979, close to the theoretical percentiles of −1.860 and 1.860. The commands in *lines 109–110* of Advances_Statistics_Code_P.R return these values. Your values will differ slightly.

In most experiments, we use a single sample and calculate a single test statistic in order to make inferences about some null hypothesis. Suppose we established beforehand a critical signficance level–a benchmark for uncommonness–of $\alpha = 0.10$ (9). If the null hypothesis is true, the test statistic $t = 3.407$ ($P = 0.005$) from our first sample is more unusual–less likely to occur–than our benchmark. As a result, we would reject the null hypothesis and conclude that the sample observations were consistent with having come from a population that had a mean μ other than 0. And we would be wrong. We know the null hypothesis is true: we drew our observations from a population that had a mean of 0 (see Fig. 2).

So how can we make sense of this? By realizing that when we draw a single random sample from some population–when we do a single experiment–we can have enough unusual observations so that it just appears the observations came from a different population.

Summary

As this exploration has demonstrated, a test statistic compares what we observe in an experiment to what we expect to see if the null hypothesis is true. The *P* value associated with the magnitude of that test statistic answers the question, if the null hypothesis is true, what proportion of possible values of the test statistic are at least as extreme as the one I got? Although the statistical test of a null hypothesis is useful–it helps guard against an unwarranted conclusion, or it helps argue for a real experimental effect (7, 44)–the only question it can answer is a trivial one: is there anything other than random variation going on here? The answer to this question is a simple yes or no. Science is less yes-or-no and more how-much.

In the next installment of this series, we will explore confidence intervals. A confidence interval provides the same statistical information as the *P* value from a hypothesis test, but it circumvents the drawbacks inherent to a hypothesis test. In essence, a confidence interval helps answer the question, is the experimental effect big enough to be relevant?

APPENDIX

Regardless of whether you use a Mac or a PC, there are two preliminary steps to perform: first, on your Desktop, create a folder called Advances, and second, download and install R.

If you use a Mac, download R from

`http://cran.us.r-project.org/bin/macosx/` .

After you have installed R, double-click on Advances_Statistics_Code_P.R to open it.

If you use a PC, download R from

`http://cran.us.r-project.org/bin/windows/base/` .

After you have installed R, a shortcut for R will exist on your Desktop. To simplify the process of starting R from within your Advances folder, move this shortcut into your Advances folder, right-click on the shortcut, and then click Properties. Paste the full address (path) of your Advances folder–this path will vary depending on the Windows operating system you use–into the Start in: location (see Ref. 8, Fig. 1) and then click OK. Now double-click on the R shortcut to open R. To open Advances_Statistics_Code_P.R, click File | Open script. . . or click the Open script icon 📂, select the script filename, and then click Open. Advances_Statistics_Code_P.R will open in the R Editor.

ACKNOWLEDGMENTS

I thank Prof. John Ludbrook (Department of Surgery, The University of Melbourne, Victoria, Australia) for helpful comments and suggestions.

REFERENCES

1. **Berger JO, Sellke T.** Testing a point null hypothesis: the irreconcilability of *P* values and evidence. *J Am Stat Assoc* 82: 112–122, 1987.
2. **Berkson J.** Tests of significance considered as evidence. *J Am Stat Assoc* 37: 325–335, 1942.
3. **Box JF.** *RA Fisher: the Life of a Scientist*. New York: Wiley, 1978.
4. **Box JF.** Gosset, Fisher, and the *t* distribution. *Am Stat* 35: 61–66, 1981.
5. **Cohen J.** The Earth is round ($p < .05$). *Am Psychol* 49: 997–1003, 1994.
6. **Cox DR.** Some problems connected with statistical inference. *Ann Math Statistics 29:* 357–372, 1958.
7. **Cox DR.** Statistical significance tests. *Br J Clin Pharmacol* 14: 325–331, 1982.
8. **Curran-Everett D.** Explorations in statistics: standard deviations and standard errors. *Adv Physiol Educ* 32: 203–208, 2008.
9. **Curran-Everett D, Benos DJ.** Guidelines for reporting statistics in journals published by the American Physiological Society. *J Appl Physiol* 97: 457–459, 2004.
10. **Curran-Everett D, Taylor S, Kafadar K.** Fundamental concepts in statistics: elucidation and illustration. *J Appl Physiol* 85: 775–786, 1998.
11. **Cushny AR, Peebles AR.** The action of optical isomers. II. Hyoscines. *J Physiol* 32: 501–510, 1905.
12. **Edgeworth FY.** The calculus of probabilities applied to psychical research. *Proc Soc Psychical Res* 3: 190–199, 1885.
13. **Fisher RA.** *Statistical Methods for Research Workers.* London: Oliver and Boyd, 1925.
14. **Fisher RA.** The arrangement of field experiments. *J Minist Agric GB* 33: 503–513, 1926.
15. **Fisher RA.** Note on Dr. Berkson's criticism of tests of significance. *J Am Stat Assoc* 38: 103–104, 1943.
16. **Fisher RA.** *Statistical Methods and Scientific Inference.* London: Oliver and Boyd/Longman Group, 1956.

[9] The commands in *lines 75–76* of Advances_Statistics_Code_P.R return this value.

17. **Gibbons JD, Pratt JW.** *p*-values: interpretation and methodology. *Am Stat* 29: 20–25, 1975.
18. **Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L.** *The Empire of Chance*. London: Cambridge Univ. Press, 1989.
19. **Goodman S.** Commentary: the *P*-value, devalued. *Int J Epidemiol* 32: 699–702, 2003.
20. **Goodman SN.** A comment on replication, *P*-values and evidence. *Stat Med* 11: 875–879, 1992.
21. **Goodman SN.** *p* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 137: 485–496, 1993.
22. **Goodman SN.** Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann Intern Med* 130: 995–1004, 1999.
23. **Greenland S.** Re: "*p* Values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate". *Am J Epi* 139: 116–117, 1994.
24. **Healy MJR.** Significance tests. *Arch Dis Child* 66: 1457–1458, 1991.
25. **Hotelling H.** The impact of RA Fisher on statistics. *J Am Stat Assoc* 46: 35–46, 1951.
26. **Hubbard R, Lindsay RM.** Why *P* values are not a useful measure of evidence in statistical significance testing. *Theory Psychol* 18: 69–88, 2008.
27. **Inman HF.** Karl Pearson and RA Fisher on statistical tests: a 1935 exchange from *Nature. Am Stat* 48: 2–11, 1994.
28. **Jones LV.** Statistical theory and research design. *Annu Rev Psychol* 6: 405–430, 1955.
29. **Jones LV, Tukey JW.** A sensible formulation of the significance test. *Psychol Methods* 5: 411–414, 2000.
30. **Kendall MG.** Ronald Aylmer Fisher, 1890–1962. *Biometrika* 50: 1–15, 1963.
31. **Kruskal WH.** Tests of significance. In: *International Encyclopedia of the Social Sciences,* edited by Sills DL. New York: Macmillan & The Free Press, 1968, vol. 14, p. 238–250.
32. **Lang JM, Rothman KJ, Cann CI.** That confounded *P*-value. *Epidemiology* 9: 7–8, 1998.
33. **Lehmann EL.** The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc* 88: 1242–1249, 1993.
34. **Ludbrook J.** RA Fisher's life and death in Australia, 1959–1962. *Am Stat* 59: 164–165, 2005.
35. **Mather K.** RA Fisher's *Statistical Methods for Research Workers:* an appreciation. *J Am Stat Assoc* 46: 51–54, 1951.
36. **Matthews JR.** *Quantification and the Quest for Medical Certainty*. Princeton, NJ: Princeton Univ. Press, 1995.
37. **Neyman J.** *First Course in Probability and Statistics*. Berkeley, CA: Univ. of California, 1948.
38. **Neyman J, Pearson ES.** On the use and interpretation of certain test criteria for purposes of statistical inference, part 1. *Biometrika* 20A: 175–240, 1928.
39. **Neyman J, Pearson ES.** On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond Ser A* 231: 289–337, 1933.
40. **Pearson ES.** *Student: a Statistical Biography of William Sealy Gosset*. New York: Oxford Univ. Press, 1990.
41. **R Development Core Team.** *R: a Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2008; http://www.R-project.org.
42. **Rozeboom WW.** The fallacy of the null-hypothesis significance test. *Psychol Bull* 57: 416–428, 1960.
43. **Savage LJ.** On rereading RA Fisher. *Ann Stat* 4: 441–500, 1976.
44. **Snedecor GW, Cochran WG.** *Statistical Methods* (6th ed.). Ames, IA: Iowa State Univ. Press, 1967.
45. **Sterne JAC, Smith GD.** Sifting the evidence–what's wrong with significance tests? *Br Med J* 322: 226–231, 2001.
46. **Stigler SM.** *The History of Statistics: the Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard Univ. Press, 1986.
47. **Stigler SM.** *Statistics on the Table: the History of Statistical Concepts and Methods*. Cambridge, MA: Harvard Univ. Press, 1999.
48. **Student.** The probable error of a mean. *Biometrika* 6: 1–25, 1908.
49. **Weinberg CR.** It's time to rehabilitate the *P*-value. *Epidemiology* 12: 288–290, 2001.
50. **Wilkinson L and the Task Force on Statistical Inference.** Statistical methods in psychology journals. *Am Psychol* 54: 594–604, 1999.
51. **Yates F.** The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *J Am Stat Assoc* 46: 19–34, 1951.
52. **Yates F, Mather K.** Ronald Aylmer Fisher, 1890–1962. *Biogr Mem Fellows R Soc Lond* 9: 91–120, 1963.
53. **Youden WJ.** The Fisherian revolution in methods of experimentation. *J Am Stat Assoc* 46: 47–50, 1951.