Explorations in statistics: power

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 5 January 2010; accepted in final form 2 February 2010

Curran-Everett D. Explorations in statistics: power. Adv Physiol Educ 34: 41–43, 2010; doi:10.1152/advan.00001.2010.—Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This fifth installment of Explorations in Statistics revisits power, a concept fundamental to the test of a null hypothesis. Power is the probability that we reject the null hypothesis when it is false. Four things affect power: the probability with which we are willing to reject—by mistake—a true null hypothesis, the magnitude of the difference we want to be able to detect, the variability of the underlying population, and the number of observations in our sample. In an application to an Institutional Animal Care and Use Committee or to the National Institutes of Health, we define power to justify the sample size we propose.

This fifth article in Explorations in Statistics (see Refs. 3–6) revisits power, a concept fundamental to the test of a null hypothesis and to the calculation of a confidence interval. Power is also an element essential to most applications to Institutional Animal Care and Use Committees and to grant agencies. What is power? Power is the probability that we reject the null hypothesis when it is false: that we find an effect when it is there. Before we explore the notion of power, we want to review briefly the null hypothesis.

The Null Hypothesis

As we discovered in our second exploration (5), 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 (10, 11). Our challenge is to 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 the probability that we reject the null hypothesis $H_0$ given that $H_0$ is true: $Pr[\text{reject } H_0 \mid H_0 \text{ is true}]$. We control this kind of error when we define the error rate $\alpha$:

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

This means we will reject a true null hypothesis 100$\alpha$% of the time. Guideline 2 (Ref. 7) discusses the choice of $\alpha$.

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: $Pr[\text{fail to reject } H_0 \mid H_0 \text{ is false}]$. We control this kind of error when we define the error rate $\beta$:

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

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

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

A Basic Illustration of Power

Suppose we define the null and alternative hypotheses, $H_0$ and $H_1$, as

$$H_0: \mu = \mu_0$$

$$H_1: \mu = \mu_1 ,$$

where $\mu$ is the population mean and $\mu_1 \leq \mu_0$ (Fig. 1). That is, if we reject the null hypothesis, we conclude our sample observations came from a population whose mean is less than some reference value $\mu_0$.

Power is the probability that we reject $H_0$ given that $H_0$ is false. If $H_0$ is false, then the population mean $\mu = \mu_1$. And if we reject $H_0$, then the observed value of the test statistic $z$ is more extreme than the critical value $z_\alpha^*$: $z > z_\alpha^*$. This means Eq. 1 can be written

$$\text{power} = Pr[z < z_\alpha^* \mid \mu = \mu_1] , \tag{2}$$

where

$$z = \frac{\bar{y} - \mu_0}{SE(\bar{y})} .$$

If we substitute for $z$ and rearrange Eq. 2, we can express power as

$$\text{power} = Pr[\bar{y} < \mu_0 + z_\alpha^* SE(\bar{y}) \mid \mu = \mu_1] . \tag{3}$$

Figure 1 depicts Eq. 3.

If $H_0$ is false, then we can also calculate

$$z = \frac{\bar{y} - \mu_1}{SE(\bar{y})} \quad \text{so that} \quad \bar{y} = \mu_1 + z SE(\bar{y}) .$$

If we replace $\bar{y}$ with this expression and rearrange Eq. 3,

$$\text{power} = Pr \left[ z < \frac{\mu_0 + z_\alpha^* SE(\bar{y}) - \mu_1}{SE(\bar{y})} \right] . \tag{4}$$

Last, if we replace $SE(\bar{y})$ with $\sigma/\sqrt{n}$, we can simplify Eq. 4 to

$^{2}$ If the one-tailed error rate $\alpha = 0.10$, then $z_{0.10}^* = -1.282$. If the one-tailed error rate $\alpha = 0.01$, then $z_{0.01}^* = -2.326$. 

$^{1}$ In our second exploration (5) we referred to $\alpha$ as the critical significance level. Error rate and critical significance level are synonymous (2).

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).
error rates: of observations in our sample. Figure 2 depicts power for two the variability of the underlying population; and outlined its installation. For this exploration, there is just one R: Basic Operations

estimate the power we had to detect the difference we observed. In this situation, we can think about power after an experiment in which we failed to some meaningful difference (for example, see Ref. 1). We also things–so we can estimate the sample size sufficient to detect before we do an experiment. We define power–among other [H9251/H11005] = \mu_0 + z^* \text{SE}(\bar{y})].

power = \Pr \left( z < \frac{\mu_1 - \mu_0}{\sigma} \sqrt{\frac{1}{n}} \right). \quad (5)

In this example, four things affect power: \alpha, the probability with which we are willing to reject a true null hypothesis; |\mu_1 - \mu_0|, the magnitude of the difference we want to be able to detect; \sigma, the variability of the underlying population; and n, the number of observations in our sample. Figure 2 depicts power for two error rates: \alpha = 0.10 and \alpha = 0.01.

When do we want to think about power? In an ideal world, before we do an experiment. We define power–among other things–so we can estimate the sample size sufficient to detect some meaningful difference (for example, see Ref. 1). We also can think about power after an experiment in which we failed to find a statistically significant difference. In this situation, we estimate the power we had to detect the difference we observed.

R: Basic Operations

In the first article (3) of this series, I summarized R (12) and outlined its installation. For this exploration, there is just one extra step: download Advances_Statistics_Code_Power.R³ to your Advances folder (see Ref. 3).

To run R commands. If you use a Mac, highlight the commands you want to submit and then press \$2H6126H11001H9251H20841H11001H9251H11005H20841H11002 (command key+enter). 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 Simulations

If we want to explore the notion of power, we need some data. Suppose you read an article in your favorite physiology journal in which the authors report that some response decreased an average of 0.60 units (SD 1.26) after an intervention. Because of experimental constraints, the authors studied just nine subjects. The authors report that t = −1.44 (P = 0.19) and that the 90% confidence interval was [−1.38, 0.17]. They conclude that the intervention failed to affect the response: that the sample observations were consistent with having come from a population whose mean \mu = 0.

In the DISCUSSION, the authors note that the response could have decreased by as much as 1.38 units, a change that is scientifically meaningful. The commands in lines 23–35 of Advances_Statistics_Code_Power.R read in the nine observations

and generate the sample statistics above.

In fact, we drew these fictitious observations not from a normal distribution with mean \mu_0 = 0 but from a normal distribution with mean \mu_1 = −0.5. This raises a question: how much power did we–or the authors of that mythical physiology article–have to detect that the observations came from a population with a mean of −0.5 units? We can answer this question using power.t.test (Advances_Statistics_Code_Power.R, lines 40–45):

³ The commented-out command in line 21 of Advances_Statistics_Code_Power.R generates nine observations from a population with mean \mu_1 and standard deviation \sigma = 1.
Table 1. Power estimates from R  

<table>
<thead>
<tr>
<th>Error Rate α</th>
<th>Difference Δμ</th>
<th>Variability σ</th>
<th>Sample Size n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.25</td>
<td>0.5</td>
<td>9</td>
</tr>
<tr>
<td>0.05</td>
<td>0.30</td>
<td>0.5</td>
<td>18</td>
</tr>
<tr>
<td>0.10</td>
<td>0.35</td>
<td>0.5</td>
<td>36</td>
</tr>
</tbody>
</table>

All else being equal, power increases when the error rate α increases, the difference Δμ increases, the variability σ decreases, and the sample size n increases. In this example, the difference Δμ is the effect size |μ1 − μ0|/σ.

\[
power.t.test(n = n0bs, \delta = PopMean_Hl - PopMean_HO, sd = PopSD, sig.level = Alpha, type = 'one.sample')
\]

We defined n0bs, PopMean_Hl, PopMean_HO, PopSD, and Alpha in lines 7–15 of our R script. R returns 0.3927 = 0.39 for power. No wonder the authors failed to find that the observations came from a population with a mean of −0.5! Had the authors used an error rate of 0.01, power would have been a measly 0.09 units (lines 50–55). The next question is, if the authors could have studied more subjects, how many would they have needed to detect that the sample observations came from a population with a mean of −0.5 units, assuming a typical power of 0.80? We can answer this question using power.t.test if we replace the argument n with power (lines 60–65):

\[
power.t.test(power = 0.80, \delta = PopMean_Hl - PopMean_HO, sd = PopSD, sig.level = Alpha, type = 'one.sample')
\]

R returns that 26.1375 = 27 subjects would have been required. Had this been a real experiment, we would have used the sample mean \(\bar{y} = 0.60\) to estimate PopMean_Hl and the sample standard deviation s = 1.26 to estimate PopSD.

Last, we can use power.t.test (lines 70–105) to explore the impact on power of those four things we mentioned: α, the probability with which we reject a true null hypothesis; |μ1 − μ0|, the magnitude of the difference we want to detect; σ, the variability of the population; and n, the number of observations in our sample. Table 1 lists the results. Power increases when the error rate α increases, when the magnitude of the difference we want to detect increases, when the variability σ decreases, and when the number of observations increases. We can use Eq. 5 or Fig. 1 to explain these results, but we can also use words.

When the error rate α increases, we are more likely to reject a true null hypothesis, but this means we are also more likely to reject a null hypothesis that is false: we are more likely to find a difference that does exist.

When the magnitude of the difference we want to detect increases, we are more likely to find that difference: it is easier to find a knitting needle than a sewing needle.

When the variability σ decreases or when the number of observations increases, the standard deviation of the theoretical distribution of the sample mean, \(s\sqrt{n}\), decreases (3): there is more separation between the null and alternative distributions of the sample mean.

R has power functions for a one-way analysis of variance (power.anova.test), a two-sample test of proportions (power.prop.test), and a one- or two-sample t test (power.t.test). These suffice for many experimental situations. A commercial software package such as PASS (9), however, provides power and sample size calculations for more sophisticated designs.

Summary

As this exploration has demonstrated, power—the probability that we reject the null hypothesis when it is false—is a vital component of any statistical analysis. If we do an experiment that has insufficient power, we may fail to find a scientific difference that truly exists. We have seen that four things affect power: the probability with which we are willing to reject—by mistake—a true null hypothesis, the magnitude of the difference we want to detect, the variability of the underlying population, and the number of observations in our sample.

In the next installment of this series, we will explore correlation, a technique that estimates the magnitude of the straight-line relationship between two variables. Like any statistical procedure, correlation can clarify, or it can confuse; this happens not because of the procedure per se but because of its interpretation. In the next exploration, we will see why.

DISCLOSURES

No conflicts of interest are declared by the author.

REFERENCES