BY CHRISTOPHER A. SEAMAN, JULIA E. SEAMAN AND I. ELAINE ALLEN

The Significance of Power

Avoid mistakenly rejecting the null hypothesis in statistical trials

THE CONCEPT OF "power" has long been overshadowed in statistical circles by its big brother, "significance." Both parameters, chosen before a test, dictate the sample size and likelihood of making an erroneous conclusion when comparing two groups (see Table 1). The p-value or significance threshold is the first taught and most commonly used statistic.

Significance sets the threshold for mistakenly rejecting the null hypothesis that both groups are similar when the null hypothesis is indeed true. This mistaken result is called a false positive (type I errors or α), and it means you've found a difference between two groups when really they are not different.

Power, on the other hand, focuses on controlling the complementary testing problem to reduce false negatives (type II errors or β). Power determines how likely a test is to reject the null hypothesis when the null hypothesis is false.

When a statistical trial is conducted without enough power, it can lead to problems. A higher-power value indicates a less likely chance for a false negative. The power of a study is directly related to its sample size and effect size variability. In general, the greater the sample size and the lower the variability, the higher a study's power.

Both of these study variables, however, will increase time and cost. To completely eliminate any chance for false negatives and for the best study, you must test every possible option to the absolute physical limit of measurement. In reality, therefore, there is a balancing act between increasing power through more samples and more precise and accurate measurements with budget, time and project logistical constraints.

As power is a required part of testing, the study setup must reduce the risks of being underpowered. If a study is underpowered, the most direct issue that can arise is to refuse to reject the null hypothesis when, in fact, it is false. This means that there is a real difference in the two groups being compared, but the test is unable to detect it.

A more nuanced issue with power comes in the form of reproducibility of results. One study may find significant differences between two groups, but when others attempt to achieve the same results, they do not find a difference. This could be a symptom of underpowered follow-up study.

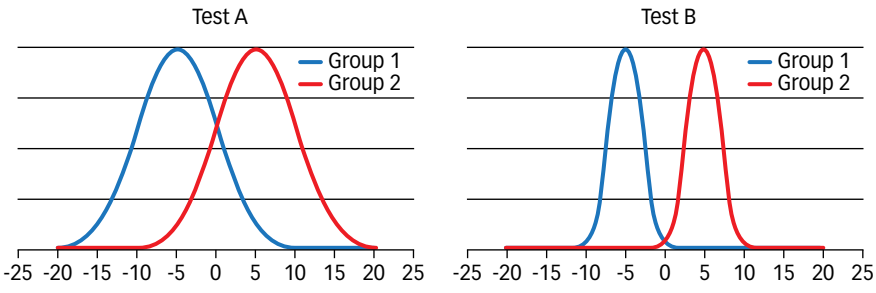
As Figure 1 (p. 52) shows the greater the discrimination between your type I and type II errors, the higher the power and lower the significance level of the study, and the easier it is to discriminate between the null and alternative hypotheses.

In the first test (A), there is a high chance for type I and type II errors, as evidenced by the high overlap between the curves. In the second test (B), the study was designed with higher power (increased sample size and lower measurement variability), and there is much less overlap between the two groups.

Power and statistical significance / TABLE 1

	Null hypothesis should not be rejected	Null hypothesis should be rejected
Reject null hypothesis	False positive = α type I error	True negative = power
Do not reject null hypothesis	True positive = confidence interval	False negative = β type II error

Comparing the difference between 2 distinct groups with low and high power / FIGURE 1



Specifying α and β

Ideally, the sample size of an experiment (or statistical trial) is calculated a priori using investigator-chosen prespecified levels of significance and power. These are often a significance level α = 0.05 (5% error is our threshold or 95% confidence), and 1-β (or power) = 80% or 90%.

In terms of the errors listed in Table 1, this means you may wish to be 90% confident that you are rejecting the null hypothesis when it should be rejected, and only allowing a false negative rate of 5% to reject the null hypothesis when it should not be rejected.

For a sample size calculation, a priori one needs a value of α, a value of 1–β and an effect-size measure. The effect size is often the difference between means of two groups or the percentage change between groups. The sample size calculated using these parameters gives an investigator the minimal detectable effect in the study.

Because the effect size may not be fixed and the variability may not be known, it may be important to calculate a table of sample sizes for a given α and β, varying the effect size and variability of measurement to optimize the study's design. There are sample-size calculators in most statistical packages and online (http://tinyurl.com/sample-size-calc) that can be tailored for the exact type of test being performed.

The power example that follows shows how you might develop a sample-size strategy based on a fixed value of α and β.

Sample size per treatment required for movement difference and variability within appliance / TABLE 2

	Variability						
Movement difference/month	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.5 mm	10	12	12	14	14	16	20
0.75 mm	8	10	10	12	12	14	18
1.0 mm	6	6	6	8	8	10	12
1.25 mm	4	4	4	6	6	8	8

As the effect size (movement difference) increases and variability decreases, the required sample size per group decreases from a maximum of 20 per group (with largest variability and smallest movement difference) to a minimum or four per group at the opposite end of the spectrum.

Suppose you have invented an easier appliance for orthodontic braces that accelerates the movement of teeth. Your hypothesis is that the new appliance will show between 0.5 and 1.25 mm more movement per month than the old appliance, and the variability will be between 0.20 and 0.50 mm per month. You will test this with a two-sided paired t-test with a significance level of α = 0.05 and 80% power.

To achieve those set levels, you must have four to 20 subjects per treatment. To confidently declare that the new braces are better, you need the fewest patients if you have the greatest movement with the lowest variability. Conversely, you need the most patients if there is little difference in movement and a lot of variability (see Table 2).

Dangers of overpowering, underpowering

Using the braces example, suppose you decide that you expect the variability to be 0.25-0.30 mm and the movement to be 0.75 mm per month, and you enroll 10 patients per group for the study. At the end of the study, however, the movement is only 0.50 mm per month and the variability greater than 0.40 mm.

Using your sample-size calculations, you needed to enroll 14 to 20 patients in the study. This indicates the sample size was too small, and the study is underpowered. The completed 10-patient study will not have a statistically significant result at your set significance and power levels.

Using this example, a more likely reason for an underpowered study is that

you only have funding for 10 patients per group, regardless of the power and sample size estimates. After completing a study, you can calculate the actual power of the study given the actual effect size and variability. This is the post-hoc power and—while seldom reported in published results—can be helpful when designing future studies, especially when the original study is underpowered.

Overpowering a study by increasing the sample size has been called a waste of resources when it involves the use of humans or animals. But can too many observations ever be a bad thing?

Studies are increasing in size as the prevalence of big data is seen in all areas of investigation. If the sample size is very large, everything may be statistically significant—but these results may not be their results are not clinically meaningful or actionable.

In some rare instances, overpowering a study can be useful for examining outcomes more precisely. It may give investigators a result, however, that is statistically significant but not an important difference. Table 3 illustrates these pitfalls.

Pitfalls of underpowering and overpowering a study / TABLE 3

	Not a statistically significant difference	Statistically significant difference
Important difference	Underpowered = too small of a sample	True negative
Unimportant difference	True positive	Overpowered = too large of a sample

Studies are increasing in size as the **prevalence of big data** is seen in all areas of investigation.

important results from the study.

As the sample size increases, the effect size and variability shrink, which gives the results more precision than can be measured with the available tools or have clinical usefulness in a new intervention.

In extreme cases, investigators can actually game the system and claim statistical significance simply by virtue of having an extremely large sample size, but

Well-balanced power

Although overshadowed by the p-value, power is an important aspect of study design, controlling the error of false negatives. As many studies are carefully designed to avoid falsely rejecting the null hypothesis, these studies also must ensure they do not falsely accept the null hypothesis through power.

Power is associated with significance levels, sample sizes and effect size variability. The best study will be well balanced among all four parameters, adjusting for the restrictions from the reality of data availability, budgets and deadlines. QP

ADDITIONAL RESOURCES

Button, Katherine S., John P.A. Ioannidis, et al., "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews: Neuroscience,* May 2013, Vol. 14, pp. 365-376.

Gelman, Andrew and David Weakliem, "Of Beauty, Sex and Power," *American Scientist,* July-August 2009, Vol. 97, No. 4, pp. 310-316.

Hochster, Howard S., "The Power of 'P': On Overpowered Clinical Trials and 'Positive' Results," *Gastrointestinal Cancer Research,* 2008, Vol. 2, No. 2, pp. 108-109.

Seaman, Christopher and I. Elaine Allen, "Different, Equivalent or Both," *Quality Progress,* July 2006, pp. 77-79.

Seaman, Julia and I. Elaine Allen, "Not Significant, But Important," *Quality Progress,* August 2011, pp. 58-59.

THREE'S NOT A CROWD

Read another Statistics Roundtable column from this trio of authors. "So Many Variables, So Few Observations" appeared in the September 2013 edition of QP. Visit http://tinyurl.com/seaman-stats-round to access the article.

CHRISTOPHER A. SEAMAN is a data scientist at Atlassian in San Francisco and a statistical consultant for the Quahog Research Group in Oakland, CA. He has a master's degree in mathematics from the Graduate Center of the City University of New York.

JULIA E. SEAMAN is a doctoral student in pharmacogenomics at the University of California-San Francisco, and a statistical consultant for the Babson Survey Research Group at Babson College in Wellesley, MA. She earned a bachelor's degree in chemistry and mathematics from Pomona College in Claremont, CA. Seaman is a member of ASQ.

I. ELAINE ALLEN is professor of biostatistics at the University of California-San Francisco and emeritus professor of statistics at Babson College. She is also director of the Babson Survey Research Group. She earned a doctorate in statistics from Cornell University in Ithaca, NY. Allen is a member of ASQ.