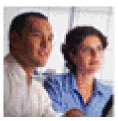
On Power

by Floyd Bullard
North Carolina School of Science and Mathematics
Durham, North Carolina



Teaching the Concept of Power

Annually, AP Statistics teachers struggle to help their students understand the concept of power in tests of significance. It seems very daunting when you read a text that describes how to calculate the power of a test against an alternate hypothesis or that shows how to graph power curves. In my experience, these two approaches to teaching power are sufficiently difficult for students that only the brightest really can see the concepts through the calculations. The rest may or may not learn to do the calculations correctly, but even those who do sometimes learn merely a mechanical process and do not understand what power really is.

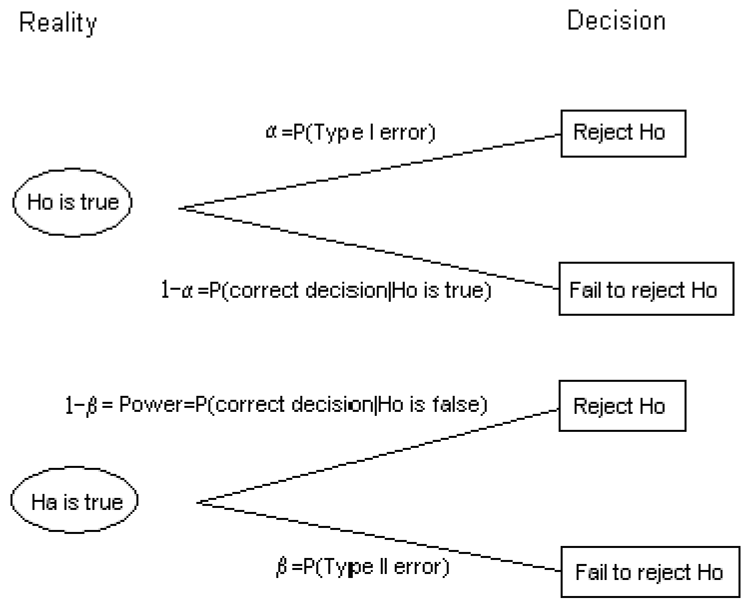
Happily, power is not all that difficult a concept, and the AP Statistics curriculum requires students to understand only the concept of power and what affects it. Students are not expected to compute the power of a test of significance against a particular alternate hypothesis.

What Is Meant by Power?

Power may be expressed in several different ways, and it might be worthwhile sharing more than one of them with your students, as one definition may "click" with a student where another does not. Here are a few different ways to describe what power is:

- Power is the probability of rejecting the null hypothesis when in fact it is false.
- Power is the probability of making a correct decision (to reject the null hypothesis) when the null hypothesis is false.
- Power is the probability that a test of significance will pick up on an effect that is present.
- Power is the probability that a test of significance will detect a deviation from the null hypothesis, should such a deviation exist.
- Power is the probability of avoiding a Type II error.

Note: Technically, power is simply the probability of rejecting the null hypothesis, but it is easier for students to think about it in the case where the null is false, and many introductory texts define power as such. Given this definition, it is important for students to understand that power, like the probability of Type I and Type II errors, is implicitly a conditional probability. In the case of power, it is conditional upon the falsehood of the null hypothesis, or it is conditional upon what alternate hypothesis is in fact true. The following tree diagram may help students appreciate the fact that α, β, and power are all conditional probabilities.



Naturally, we don't want our students merely to parrot a phrase about power, as that is little better than having no concept of power at all. I have found it helpful as we discuss power to continually restate what power means throughout discussions, using different language each time. For example, if we do a test of significance at level α = 0.1, I might say, "That's a pretty big alpha level. This test is ready to reject the null at the drop of a hat. Is this a very powerful test?" (Yes, it is. Or at least, it's more powerful than it would be with a smaller alpha value.) If a student answers a question about Type II errors and says that the consequences of a Type II error are very severe, then I may follow up with the question, "So you really want to avoid Type II errors, huh? What does that say about what we require of our test of significance?" (We want a very powerful test.)

What Affects Power?

There are four things that primarily affect the power of a test of significance. They are:

- **The significance level α of the test.** If all other things are held constant, then as α increases, so does the power of the test. This is because a larger α means a larger rejection region for the test and thus a greater probability of rejecting the null hypothesis. That translates to a more powerful test. The price of this increased power is that as α goes up, so does the probability of a Type I error should the null hypothesis in fact be true. In practice, α is not chosen with power in mind. Rather, the significance level is chosen based on the severity of the consequences of a Type I error or the credibility of the alternate hypothesis. But students should still know that α has an effect on power.
- **The sample size *n*.** As *n* increases, so does the power of the significance test. This is because a larger sample size narrows the distribution of the test statistic. The hypothesized distribution of the test statistic and the true distribution of the test statistic (should the null hypothesis in fact be false) become more distinct from one another as they become narrower, so it becomes easier to tell whether the observed statistic comes from one distribution or the other. The price paid for this increase in power is of course the higher cost in time and resources required for collecting more data. There is usually a sort of "point of diminishing returns" up to which it is worth the cost of the data to gain more power, but beyond which the extra power is not worth the price.
- **The inherent variability in the population(s).** As the variability increases, the power of the test of significance decreases. One way to think of this is that a test of significance is like trying to detect the presence of a "signal," such as the effect of a treatment, and the inherent variability in the population is "noise" that will drown out the signal if it is too great. Researchers can't always control the variability in the population, but they can sometimes reduce it through especially careful data collection. The design of a study may also reduce variability, and one primary reason for choosing such a design is that it allows for increased power without necessarily having exorbitantly costly sample sizes. For example, a matched pairs design usually reduces variability by "subtracting out" some of the variability that individual subjects bring to a study. Often, researchers will do a preliminary study before conducting a full-blown study intended for publication. There are several reasons for this, but one of the most important is so they can assess the inherent variability within the populations they are studying. An estimate of that variability allows them to determine the sample size they will require for a future test having a desired power. A test lacking statistical power could likely result in a costly study that produces no significant findings.
- **The difference between the hypothesized value of a parameter and its true value.** This is sometimes called the "magnitude of the effect" in the case when the parameter of interest is the difference between parameter values (say, means) for two treatment

groups. The larger the effect, the more powerful the test is. This is because when the effect is large, the true distribution of the test statistic is far from its hypothesized distribution, so the two distributions are distinct and it's easy to tell which one an observation came from. The intuitive idea is simply that it's easier to detect a large effect than a small one. This principle has two consequences that students should understand that are essentially two sides of the same coin. On the one hand, it's important to understand that a subtle but important effect (say, a modest increase in the life-saving ability of a hypertension treatment) may be demonstrable but could require a powerful test with a large sample size to produce statistical significance. On the other hand, a small, unimportant effect may be demonstrated with a high degree of statistical significance if the sample size is large enough. Because of this, too much power can almost be a bad thing, at least so long as many people continue to misunderstand the meaning of statistical significance. For your students to appreciate this aspect of power, they must understand that statistical significance is a measure of the *strength of evidence of the presence of an effect.* It is *not* a measure of the magnitude of the effect. For that, a confidence interval should be constructed.

Two Classroom Activities for Teaching About Power

The two activities described below are similar in nature. The first one relates power to the magnitude of the effect, and the second one relates power to sample size. Both are described for classes of 20 students, but you can modify them as needed for smaller or larger classes or for classes in which you have fewer resources available. Both of these activities involve tests of significance on a single population proportion, but the principles are true for nearly all tests of significance.

Activity One: Relating Power to the Magnitude of the Effect

In advance of the class, you should prepare 20 bags of poker chips or some other token that comes in more than one color. Each of the bags should have a different number of blue chips in it, ranging from 0 out of 20 to 20 out of 20. Do not include 10 out of 20. These bags represent populations with different proportions. Distribute one bag to each student but tell the students not to look in the bag. Then instruct them to draw 20 chips from the bag with replacement, shaking well after each draw. (Sampling with replacement from a small bag of chips simulates sampling without replacement from a very large population.) Have them count the number of blue chips out of the 20 that they observed in their sample and then perform a test of significance whose null hypothesis is that the bag contains 50 percent blue chips and whose alternate hypothesis is that it does not. They should use a significance level of $\alpha = 0.10$. While they are sampling, make a column on the blackboard of their population proportions, from 0 to 1.00, with the heading "Population Proportion." Beside it, make a column with the heading "Reject p = 0.50? Yes/No" and have the students complete their own row of the table as they complete their significance tests. (After doing their tests, they are to look in their bags to see what their population proportions are.) You and your students should see in the completed table that as the true proportion gets farther from 0.50, the test becomes more likely to reject the null hypothesis. "A big effect is easier to detect than a small one."

Activity Two: Relating Power to Sample Size

For this activity, prepare 20 paper bags that are all identical, containing 13 blue chips and 7 nonblue chips. Pair the students up. One pair gets sample sizes n = 20 and n = 115. The next pair gets sample sizes n = 25 and n = 110. And so on, so that the last pair gets the sample sizes n = 65 and n = 70. As before, the students are to sample with replacement up to their sample size and then perform a test of significance to see whether they reject the null hypothesis of p = 0.50 at the $\alpha = 0.10$ significance level. As before, they complete a table on the board with Yes or No answers, in which the first column, completed by you, is the sample size. You and your students should see in the completed table that as sample size increases, so does the probability of rejecting the null hypothesis of p = 0.50. Thus, as sample size increases, so does statistical power.

Focus on Concepts

The AP Statistics curriculum is designed primarily to help students understand statistical concepts and become critical consumers of information. Being able to perform statistical computations is of, at most, secondary importance and for some topics, such as power, is not expected of students at all. Students should know what power means and what affects the power of a test of significance. The activities described here can help students understand power better. If you teach a 50-minute class, you should spend one or at most two class days teaching power to your students. Don't get bogged down with calculations. They're important for statisticians, but they're best left for a later course. Getting the concepts down is all that is appropriate for the introductory-level AP Statistics course.

Floyd Bullard received his bachelor's degree in mathematical sciences from the Johns Hopkins University in 1991 and his master's degree in statistics from the University of North Carolina at Chapel Hill in 1999. He has taught high school math as a Peace Corps volunteer in Bénin, West Africa, at the Horace Mann School in New York, and most recently at the North Carolina School of Science and Mathematics (NCSSM) in Durham, North Carolina. He is now on a leave of absence from NCSSM to study in a doctoral program in statistics at Duke University, after which

he plans to return to teaching. Floyd is a Reader for the AP Statistics Exam.

See also...

- AP Statistics Course Home Page
- Teachers' Resources
- Personal Profile