Chapter 6

Introduction to Power Analysis

Overview

A statistical test's *power* is the probability that it will result in statistical significance. Since statistical significance is the desired outcome of a study, planning to achieve high power is of prime importance to the researcher. Because of its complexity, however, an analysis of power is often omitted.

PASS calculates statistical power and determines sample sizes. It does so for a broad range of statistical techniques, including the study of means, variances, proportions, survival curves, correlations, bioequivalence, analysis of variance, log rank tests, multiple regression, and contingency tables.

PASS was developed to meet several goals, including ease of learning, ease of use, accuracy, completeness, interpretability, and appropriateness. It lets you study the influence of sample size, effect size, variability, significance level, and power on your statistical analysis.

Brief Introduction to Power Analysis

Statistical power analysis must be discussed in the context of *statistical hypothesis testing*. Hence, this discussion starts with a brief introduction to statistical hypothesis testing, paying particular attention to topics that relate to power analysis and sample size determination. Although the theory behind hypothesis testing is general, its concepts can be reviewed by discussing simple case: testing whether a proportion is greater than a known standard.

Following the usual terminology of statistical hypothesis testing, define two complementary hypotheses

$$H_0: P \leq P0 \quad \text{vs.} \quad H_0: P > P0$$

where P is the response proportion in the population of interest and $P0$ is the known standard value.

H_0 is called the *null hypothesis* because it specifies that the difference between the two proportions is zero (null).

H_1 is called the *alternative hypothesis*. This is the hypothesis of interest to us. Our motivation for conducting the study is to provide evidence that the alternative (or research) hypothesis is true. We do this by showing that the null hypothesis is unlikely—thus establishing that the alternative hypothesis (the only possibility left) is likely.

Outcomes from a statistical test may be categorized as follows:

1. Reject H_0 when H_0 is true. That is, conclude that H_0 is unlikely when it is true. This constitutes a decision error known as the *Type-I error*. The probability of this error is alpha (α) and is often referred to as the *significance level* of the hypothesis test.
2. Do not reject H_0 when H_0 is false. That is, conclude that H_0 is likely when it is false. This constitutes a decision error known as the *Type-II error*. The probability of this error is beta (β). *Power* is 1 – β. It is the probability of rejecting H_0 when it is false.
3. Reject H_0 when H_0 is false. This is a correct decision.
4. Do not reject H_0 when H_0 is true. This is also a correct decision.

The basic steps in conducting a study that is analyzed with a hypothesis test are:

1. Specify the statistical hypotheses, H_0 and H_1.
2. Run the experiment on a given number of subjects.
3. Calculate the value of a test statistic, such as the sample proportion.
4. Determine whether the sample values favor H_0 or H_1.

Binomial Probability Table

In the current example, suppose that a random sample of ten individuals is selected, i.e., *N* = 10. The number of individuals, *R*, with the characteristic of interest is counted. Hence, *R* is the test statistic. A table of binomial probabilities gives the probability that *R* takes on each of its eleven possible values for various values for *P*.

	P								
R	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	0.349	0.107	0.028	0.006	0.001	0.000	0.000	0.000	0.000
1	0.387	0.376	0.121	0.040	0.010	0.002	0.000	0.000	0.000
2	0.194	0.302	0.233	0.121	0.044	0.011	0.001	0.000	0.000
3	0.057	0.201	0.267	0.215	0.117	0.042	0.009	0.001	0.000
4	0.011	0.088	0.200	0.251	0.205	0.111	0.037	0.006	0.000
5	0.001	0.026	0.103	0.201	0.246	0.201	0.103	0.026	0.001
6	0.000	0.006	0.037	0.111	0.205	0.251	0.200	0.088	0.011
7	0.000	0.001	0.009	0.042	0.117	0.215	0.267	0.201	0.057
8	0.000	0.000	0.001	0.011	0.044	0.121	0.233	0.302	0.194
9	0.000	0.000	0.000	0.002	0.010	0.040	0.121	0.376	0.387
10	0.000	0.000	0.000	0.000	0.001	0.006	0.028	0.107	0.349

Let us discuss in detail the interpretation of the values in this table for the simple case in which a coin is flipped ten times and the number of heads is recorded. The column parameter *P* is the probability of obtaining a head on any one toss of the coin. When dealing with coin tossing, one would usually set *P* = 0.5, but this does not have to be the case. The row parameter *R* is the number of heads obtained in ten tosses of a coin.

The body of the table gives the probability of obtaining a particular value of *R*. One way to interpret this probability value is as follows: conduct a simulation in which this experiment is repeated a million times for each value of *P*. Using the results of this simulation, calculate the proportion of experiments that result in each value of *R*. This proportion is recorded in this table. For example, when the probability of obtaining a head on a single toss of a coin is 0.5, ten flips of a coin would result in five heads 24.6% of the time. That is, as the procedure is repeated (flipping a coin ten times) over and over, 24.6% of the outcomes would be five heads.

Calculating the Significance Level, Alpha

We will now explain how the above table is used to set the significance level (the probability of a type-I error) to a pre-specified value. Recall that a type-I error occurs when an experiment results in the rejection of the null hypothesis when, in fact, the null hypothesis is true. By studying the table, the impact of using different *rejection regions* can be determined. A rejection region is a simple rule that states which values of the test statistic will result in the null hypothesis being rejected.

For example, suppose we want to test *P*0 = 0.5. That is, the null hypothesis is that *P* = 0.5 and the alternative hypothesis is that *P* > 0.5. Suppose the rejection region is *R* equal to 8, 9, or 10. That is, H_0 is rejected if *R* = 8, 9, or 10. From the above table, the probability of obtaining 8, 9, or 10 heads in 10 tosses when *P* = 0.5 is calculated as follows:

$$\begin{aligned}\Pr(R = 8,9,10|P = 0.5) &= \Pr(R = 8|P = 0.5) + \Pr(R = 9|P = 0.5) + \Pr(R = 10|P = 0.5) \\ &= 0.044 + 0.010 + 0.001 \\ &= 0.055\end{aligned}$$

That is, 5.5% of these coin tossing experiments using this decision rule result in a type-I error. By setting the rejection criterion to *R* = 8, 9, or 10, alpha has been set to 0.055.

It is extremely important to understand what alpha means, so we will go over its interpretation again. If the probability of obtaining a head on a single toss of a coin is 0.5, then 5.5% of the experiments that use the rejection criterion of *R* = 8, 9, or 10 will result in the false conclusion that *P* > 0.5.

The key features of this definition that are often overlooked by researchers are:

1. **The value of alpha is based on a particular value of *P*.** Note that we used the assumption "if the probability of obtaining a head is 0.5" in our calculation of alpha. Hence, if the actual value of *P* is 0.4, our calculations based on the assumption that *P* is 0.5 are wrong. Mathematicians call this a conditional probability since it is based on the condition that *P* is 0.5. Alpha is 0.055 if *P* is 0.5.

   Often, researchers think that setting alpha to 0.05 means that the probability of rejecting the null hypothesis is 0.05. Can you see what is wrong with this statement? They have forgotten to mention the key fact that this statement is based on the assumption that *P* is 0.5!

2. **Alpha is a statement about a proportion in multiple experiments.** Alpha tells us what percentage of a large number of experiments will result in 8, 9, or 10 heads. Alpha is a statement about what to expect from future experiments. It is not a statement about *P*. Occasionally, researchers conclude that the alpha level is the probability that *P* = 0.5. This is not what is meant. Alpha is not a statement about *P*. It is a statement about future experiments, given a particular value of *P*.

Interpreting P Values

The term *alpha value* is often used interchangeably with the term *p value*. Although these two terms are closely related, there is an important distinction between them. A *p* value is the largest value of alpha that would result in the rejection of the null hypothesis for a particular set of data. Hence, while the value of alpha is set during the planning of an experiment, the *p* value is calculated from the data after experiment has been run.

Calculating Power and Beta

We will now explain how to calculate the power. Recall that power is the probability of rejecting a false null hypothesis. A false H_0 means that *P* is some value other than *P*0. In order to compute power, we must know the actual value of *P*.

Returning to our coin tossing example, suppose the actual value of *P* is 0.7. What is the power and beta value of this testing procedure? The decision rule is to reject the null hypothesis when *R* is 8, 9, or 10. From the above probability table, the probability of obtaining 8, 9, or 10 heads in 10 tosses of a coin when probability of a head is actually 0.7 is

$$\begin{aligned}\Pr(R = 8,9,10|P = 0.7) &= \Pr(R = 8|P = 0.7) + \Pr(R = 9|P = 0.7) + \Pr(R = 10|P = 0.7) \\ &= 0.233 + 0.121 + 0.028 \\ &= 0.382\end{aligned}$$

This is the power. The value of a type-II error is 1.000 - 0.382, which is 0.618. That is, if *P* is 0.7, then 38.2% of these coin tossing experiments will reject H_0, while 61.8% of them will result in a type-II error.

It is extremely important to understand what beta means, so we will go over its interpretation again. If the probability of obtaining a head on the toss of a coin is 0.7, then 61.8% of the experiments that use the rejection criterion of *R* = 8, 9, or 10 will result in the false conclusion that *P* = 0.5.

The key features of this definition that are often overlooked by researchers are:

1. **The value of beta is based on a particular value of *P*.** Note that we used the assumption "if the probability of obtaining a head is 0.7" in our calculation of beta. Hence, if the actual value of *P* is 0.6, our calculation based on the assumption that *P* was 0.7 is wrong.
2. **Beta is a statement about the proportion of experiments.** Beta tells us what percentage of a large number of experiments will result in 8, 9, or 10 heads. Beta is a statement about what we can expect from future experiments. It is not a statement about *P*.
3. **Beta depends on the value of alpha.** Since the rejection region (8, 9, or 10 heads) depends on the value of alpha, beta depends on alpha.
4. **You cannot make both errors at the same time.** A type-II error can only occur when a type-I error did not occur, and vice versa.

Specifying Alternative Values of the Parameters

We have noted a great deal of confusion about specifying the values of the parameters under the alternative hypothesis. The alternative hypothesis is usually that the value of one parameter is different from another. The hypothesis does not usually specify how different. It simply gives the direction of the difference. The power is calculated at specified alternative values. These values should be considered as values at which the power is calculated, not as the true value.

Effect Size

The *effect size* is the size of the change in the parameter of interest that can be detected by an experiment. For example, in the coin tossing example, the parameter of interest is *P*, the probability of a head. In calculating the sample size, we would need to state what the baseline probability is (probably 0.5) and how large of a deviation from *P* that we want to detect with our experiment. We would expect that it would take a much larger sample size to detect a deviation of 0.01 than it would to detect a deviation of 0.40.

Selecting an appropriate effect size is difficult because it is subjective. The question that must be answered is: what size change in the parameter would be of interest? Note that, in power analysis, the effect size is not the actual difference. Instead, *the effect size is the change in the parameter that is of interest* or is to be detected. This is a fundamental concept that is often forgotten after the experiment is run.

After an experiment is run that leads to non-significance, researchers often ask, "What was the experiment's power?" and "How large of a sample size would have been needed to detect significance?" To compute the power or sample size, they set the effect size equal to the amount that was seen in their experiment. This is incorrect. *When performing a power analysis after an experiment has completed, the effect size is still the change in the parameter that would be of interest to other scientists*. It is not the change that was actually observed!

Often, the effect size is stated as a percentage change rather than an absolute change. If this is the case, you must convert the percentage change to an absolute change. For example, suppose that you are designing an experiment to determine if tossing a particular coin has exactly a 50% chance of yielding a head. That is, *P*0 is 0.50. Suppose your gambling friends are interested in whether a certain coin has a 10% greater chance. That is, they are concerned with the case where *P* is 0.55 or greater. The effect size is |0.50 - 0.55| or 0.05.

Types of Power Analyses

There are several types of power analyses. Often, power analysis is performed during the design phase of a study to determine the sample size. This type of study would determine the value of *N* for set values of alpha and beta. Another type of power analysis is a post hoc analysis, which is done after the study is concluded. A post hoc analysis studies such questions as:

1. What sample size would have been needed to detect a specific effect size?
2. What is the smallest effect size that could be detected with this sample size?
3. What was the power of the test procedure?

These and similar questions may be answered using power analysis. By considering these kinds of questions after a study is concluded, you can gain important insights into how to make your research more efficient and effective.

Nuisance Parameters

Statistical hypotheses usually make statements about one or more parameters from a set of one or more probability distributions. Often, the hypotheses leave other parameters of the probability distribution unspecified. These unspecified parameters are called 'nuisance' parameters.

For example, a common clinical hypothesis is that the response proportions of two drugs are equal. The null hypothesis is that the difference between these two drugs is zero. The alternative is that the difference is non-zero. Note that the actual values of the two proportions are not stated in the hypothesis—just their difference. The actual values of the proportions will be needed to compute the power. That is, different powers will result for the case when $P1 = 0.05$ and $P2 = 0.25$ and for the case $P1 = 0.50$ and $P2 = 0.70$. In this example, the proportion difference ($D = P1 - P2$) is the parameter of interest. The baseline proportion, $P1$, is a nuisance parameter.

Another example of a nuisance parameter occurs when using the t-test to test whether the mean is equal to a particular value. When computing the power or sample size for this test, the hypothesis specifies the value of the mean. However, the value of the standard deviation is also required. In this case, the standard deviation is a nuisance parameter.

When performing a power analysis, you should state all your assumptions, including the values of any nuisance parameters that were used. When you do not have any idea as to reasonable values for nuisance parameters, you should use a range of possible values so that you can analyze how sensitive the results are to the values of the nuisance parameters. Also, do not be tempted to use the nuisance parameter's value from a previous (or pilot) study. Instead, a reasonable strategy is to compute a confidence interval and use the confidence limit that results in the largest sample size.

Choice of Test Statistics

Many hypothesis tests can be tested with a variety of test statistics. For example, statisticians often have to decide between the t-test and the Wilcoxon test when testing means. Similarly, when testing whether two proportions are equal, they have to decide whether to use a z-test or an exact test. If they choose a z-test, they have to decide whether to apply a continuity correction.

In most cases, each test statistic will have a different power. Thus, it should be obvious that *you must compute the power of the test statistic that will be used in the analysis*. A sample size based on the t-test will not be accurate for a nonparametric test.

The next question is usually "Which test statistic should I use?" You might say "They one that requires the smallest sample size." However, other issues besides power must be considered. For example, consideration must be given to whether the assumptions of the test statistic will be met be the data. If your data is binary, it is probably unreasonable to assume that they are continuous.

These are simple principles, but they are often overlooked.

Types of Hypotheses

Hypothesis tests work this way. If the null hypothesis if rejected, the alternative hypothesis is concluded to be true. However, if null hypothesis is not rejected, no conclusion is reached--the null hypothesis *is not* concluded to be true. The only way that a conclusion is reach is if the null hypothesis is rejected.

Because of this, it is very important that the null and alternative hypotheses be constructed so that the conclusion of interest is associated with the alternative hypothesis. That way, if the null hypothesis is rejected, the study reaches the desired conclusion.

There are several types of hypotheses. These include inequality, equivalence, non-inferiority, and superiority hypotheses. In the statistical literature, these terms are used with completely different meanings, so it is important to define what is meant by each. We have tried to adopted names that are associated with the alternative hypothesis, since this is the hypothesis of interest.

It is important to note that even though two sets of hypotheses may be similar, they often will have different power and sample size requirements. For example, an equivalence test (see below) appears to be the simple reverse of a two-sided test of inequality, yet the equivalence test requires a much larger sample size to achieve the same power as the inequality test. Hence, you cannot select the sample size for an inequality test and then later decide to run an equivalence test.

Each of the sections to follow will give a brief definition along with an example based on the difference between two proportions.

Inequality Hypothesis

The term “inequality” represents the classical one-sided and two-sided hypotheses in which the alternative hypothesis is simply that the two values are unequal. These hypotheses are called tests of superiority by Julious (2004), emphasizing the one-sided versions.

Two-Sided

When the null hypothesis is rejected, the conclusion is simply that the two parameters are unequal. No statement is made about how different. For example, 0.501 and 0.500 are unequal, as are 0.500 and 0.800. Obviously, even though the former are different, the difference is not large enough to be of practical importance in most situations.

$$H_0: p_1 - p_2 = 0 \quad \text{vs.} \quad H_1: p_1 - p_2 \neq 0 \quad \text{or} \quad H_1: (p_1 - p_2 < 0) \quad \text{or} \quad (p_1 - p_2 > 0)$$

One-Sided

These tests offer a little more information than the two-sided tests since the direction of the difference is given. Again, no indication is made about how much higher (or lower) the superior value is to the inferior.

$$H_0: p_1 - p_2 \leq 0 \quad \text{vs.} \quad H_1: p_1 - p_2 > 0 \quad \text{or} \quad H_0: p_1 - p_2 \geq 0 \quad \text{vs.} \quad H_1: p_1 - p_2 < 0$$

Non-Inferiority Hypothesis

These tests are a special case of the one-sided inequality tests. The term 'non-inferiority' is used to indicate that one treatment is not worse than another treatment. That is, one proportion is not less than another proportion by more than a trivial amount called the 'margin of equivalence'.

For example, suppose that a new drug is being developed that is less expensive and has fewer side effects than the standard drug. Producers must show that its effectiveness is no worse than the drug it is to replace.

When testing two proportions in which a higher proportion is better, the non-inferiority of treatment 1 as compared to treatment 2 is expressed as

$$H_0: p_1 - p_2 \leq -\delta \quad \text{vs.} \quad H_1: p_1 - p_2 > -\delta \qquad \text{or} \qquad H_0: p_1 \leq p_2 - \delta \quad \text{vs.} \quad H_1: p_1 > p_2 - \delta$$

where $\delta > 0$ is called the margin of equivalence. Note that when H_0 is rejected, the conclusion is that the first proportion is not less than the second proportion by more than δ.

Perhaps an example will help introduce this type of test. Suppose that the current treatment for a disease works 70% of the time. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. A promising new treatment has been developed to the point where it can be tested. One of the first questions that must be answered is whether the new treatment is as good as the current treatment. In other words, do at least 70% of subjects respond to the new treatment?

Because of the many benefits of the new treatment, clinicians are willing to adopt the new treatment even if it is slightly less effective than the current treatment. They must determine, however, how much less effective the new treatment can be and still be adopted. Should it be adopted if 69% respond? 68%? 65%? 60%? There is a percentage below 70% at which the difference between the two treatments is no longer considered ignorable. After thoughtful discussion with several clinicians, it is decided that if a response of at least 63% is achieved, the new treatment will be adopted. The difference between these two percentages is called the *margin of equivalence*. The margin of equivalence in this example is 7% (which is ten percent of the original 70%).

The developers must design an experiment to test the hypothesis that the response rate of the new treatment is at least 0.63. The statistical hypothesis to be tested is

$$H_0: p_1 - p_2 \leq -0.07 \quad \text{vs.} \quad H_1: p_1 - p_2 > -0.07$$

Notice that when the null hypothesis is rejected, the conclusion is that the response rate is at least 0.63. Note that even though the response rate of the current treatment is 0.70, the hypothesis test is about a response rate of 0.63. Also, notice that a rejection of the null hypothesis results in the conclusion of interest.

Non-Zero Null Hypothesis

These tests are a special case of the one-sided inequality tests. The term 'non-zero null' is used to indicate that one treatment is better than another by more than a trivial amount. For example, suppose that a new drug is being developed that is thought to have superior performance to the existing drug. Producers must show that its effectiveness is better than the drug it is to replace.

When testing two proportions in which a higher proportion is better, the superiority of treatment 1 over treatment 2 is expressed as

$$H_0: p_1 - p_2 \leq \delta \quad \text{vs.} \quad H_1: p_1 - p_2 > \delta \qquad \text{or} \qquad H_0: p_1 \leq p_2 + \delta \quad \text{vs.} \quad H_1: p_1 > p_2 + \delta$$

where $\delta > 0$ is called the difference margin. Note that when H_0 is rejected, the conclusion is that the first proportion is higher than the second proportion by more than δ.

Equivalence Hypothesis

The term 'equivalence' is used here to represent tests designed to show that response rates of two treatments do not differ by more than a trivial amount called the 'margin of equivalence'. These tests are the reverse of the two-sided inequality test.

The typical set of hypotheses are

$$H_0: p_1 - p_2 \leq \delta_L \quad \text{or} \quad p_1 - p_2 \geq \delta_U \quad \text{vs.} \quad H_1: \delta_L < p_1 - p_2 < \delta_U$$

where $\delta_L < 0$ and $\delta_U > 0$ are called the *equivalence limits*.

Suppose 70% of subjects with a certain disease respond to a certain drug. The company that produces the drug has decided to open a new facility in another city. They must show that the drug produced in the new facility is equivalent (all most the same) as that produced in existing facilities. After thoughtful discussion with several clinicians and regulatory agencies, it is decided that if the response rate of the drug produced at the new facility is between 65% and 75%, the new facility will go into production. In this case, the *margin of equivalence* is 5%.

The statistical hypothesis to be tested is

$$H_0: |p_1 - p_2| \geq 0.05 \quad \text{vs.} \quad H_1: |p_1 - p_2| < 0.05$$