Sampling and Statistical Power

Statistics is about trying to estimate values and draw inferences when there is uncertainty about what the true values are. One of the things that affects our uncertainty is the number of samples (i.e. replicates) that we have collected. This is why the sample size (or degrees of freedom) is part of just about any statistical test! So we need to take some time to think about what are true replicates and what are not and how the number of replicates that we have affects our ability to estimate values and draw inferences.

Uncertainty in Statistics

The most fundamental underlying principle of statistics is that we don't need to measure all individuals/populations/communities etc.. Instead we measure some of them and then use the theoretical properties of distributions to estimate values (e.g. means) and draw inferences (e.g. the treatment mean was different from the control mean). This is always done with some uncertainty, which is where P-values come in. If we were to measure all individuals in a population (e.g. the height of all students in our class) then we would know what the average height of the class is without any uncertainty. Unlike our class, it is not reasonable to measure all individuals. Instead we sample individuals from the population and use that sample to draw our inferences (e.g. estimate the mean height). We are trying to estimate the mean height in the population, but now we are using a sample of all possible individuals so we have some uncertainty about what that true value is.

Statistical versus Biological Populations. I will often refer to individuals being sampled from populations as a short-hand. In this case the 'population' is the statistical population from which you want to sample. This doesn't necessarily correspond to a biological population. If you were interested in studying communities you would sample communities from a statistical population of possible communities across which you would like to draw your inferences.

Which individuals do we sample?

Before talking at length about the importance of the number of replicates and what true replicates are it is worth briefly mentioning some issues with respect to how we determine which individuals to sample.

The first step in a sampling design is to determine what your population of interest is. This is the statistical population for which you want to estimate values or draw inferences. Do you want to make conclusions about all birds? Just a particular species? Just one population of this particular species? This is an important decision because it affects the scale at which we sample. Although we often don't want to admit it **the scale at which you sample is the scale at which you are able to draw your conclusions.** We would all prefer to sample in one particular locality because this is easier but if we want to draw global conclusions we need to sample globally.

Steps in Sampling:

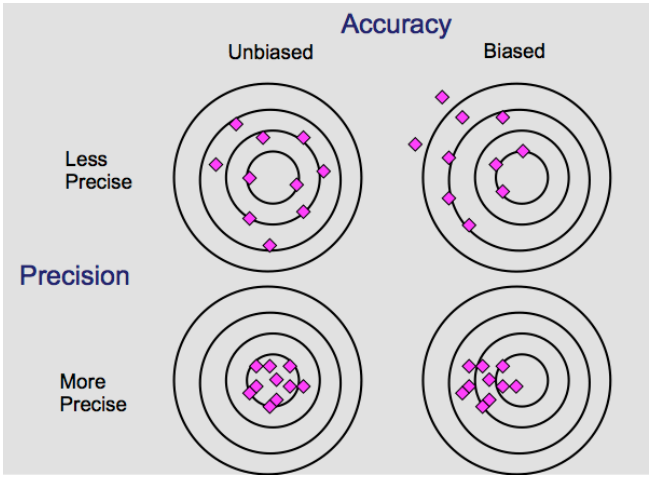
1. Define population of interest. This about what level you want to draw your inferences at.
2. Specify sampling frame (e.g. only live trees, females, etc.).
3. Specification of sampling method.
4. Collect data.
5. Review your sampling procedure.

Next we need to be explicit about the sampling frame. Are we going to restrict our observations to just some individuals within that population (e.g. only females, only trees > 5 cm in dbh).

Next we specify a sampling methods. Are we going to sample randomly, systematically, haphazardly (more on this later). After collecting our data it is important to revisit the sampling protocol. Did it work? Were some individual skipped for some reasons? It is important to consider this because what might have seemed like random sampling might not in fact have been. For example, the animal personality literature has recognized that bold animals are more likely to be caught in traps and hence are more likely to be sampled in most designs.

Why does it matter how many samples I collect?

The *Law of Large Numbers* states that the greater the number of samples the closer the sample estimate (e.g. of the mean) will come to the true value. So obviously if you flip a coin once or twice you are likely to get several instances that suggest that there are no heads on the coin. If you flip a coin 5000 times you are very likely to get 50% heads and 50% tails. This is the *Law of Large Numbers*. Why toast lands butter-side down more often than not is a different story.



Difference between *Accuracy* and *Precision*. Precision refers to how similar replicate samples are to one another. Accuracy refers to how similar replicates are to some true value. You might have a very expensive balance that weighs to the nearest 0.0001g but which is always 10 g too heavy. In this case the expensive balance is precise but not very accurate.

How many replicates do we need?

So it is clear that more replicates will give us an estimate that is closer to the true value, but how many do we actually need to collect? How many is enough?

Sample Size Rules of Thumb:

Collect 10 samples for each level of a factor or covariate that you want to include in your model. So for example, if you plan to perform a t-test then have at least 20 samples (10 in each class). If you want to perform an ANOVA with 5 levels then try and have 50 samples. Note that in a two-way ANOVA with two factors each with two levels and you want to look for an interaction (more on this later) you would want to have 40 replicates (10 for each possible combination). If you want to perform a multiple regression with 8 predictor variables in the regression then try and have 80 replicates. **Covariates are easier to measure than replicates are to collect, so pay attention to which covariates you think will be important!**

I can only count so many plants. Should I have many small quadrats or a few larger ones?

This is a common problem. You might be assessing the abundance of some plant in transects, or quadrats or the abundance of a benthic invertebrate in an Ekman grabs. I won't go into a lot of technical detail here, but, in general, biological distributions are patchy (a particular species is found in some places but not others). In this case the variance will often exceed the mean. When this is true you are better off collecting more samples that are each smaller. Just remember that the scale at which you sample cannot be smaller than the biological process you want to study (i.e. don't use a 5cm Ekman grab to try and count 10cm organisms!).

Four Possible Outcomes of a Statistical Test

When performing a statistical test (e.g. t-test comparing two means) there are four possible outcomes. You might reject the null hypothesis when it is in fact false or fail to reject the null hypothesis when it is true. These are both good things because we are correct in both cases. However, we can also

	Reject null	Do Not Reject null
Null correct	Type I error (α)	Correct ($1-\alpha$)
Null wrong	Correct ($1-\beta$)	Type II error (β)

be wrong in two ways. We might reject the null hypothesis when it is in fact true (Type I error) or fail to reject the null hypothesis when it is in fact false (Type II error). These are both bad outcomes because in both cases we have made a mistake. The probability of falsely rejecting the null hypothesis (i.e. saying there is an effect when there really isn't) is referred to as α, which is usually set to be 0.05. That is we are willing to falsely reject the null hypothesis 5% of the time. $1-\alpha$ is usually referred to as *confidence*. When α is 0.05 then we will correctly accept the null hypothesis 95% of the time.

Since we are talking about Power and sample size we are more worried about the opposite problem. That is, we have failed to reject the null hypothesis (and claimed no effect) when there in fact was an effect. This will occur with frequency β. So if the null is in fact wrong we will mistakenly accept it with frequency β and will correctly reject it with frequency $1-\beta$. Here, $1-\beta$ is referred to as *Power*. We want to have a powerful statistical test so that we correctly reject the null hypothesis when it is in fact false (i.e. claim an effect when there is in fact an effect).

Power: A better approach than the rule of thumb.

Power analyses can be extremely useful in project planning and some would argue are essential. They are also often mistakenly used (or asked for by reviewers) when a non-significant result is found. These are called *a priori* and *retrospective* power analyses, respectively.

A prior power analyses can tell you how many samples you need to collect in order to detect a given effect size (i.e. when should I stop collecting data?). Alternatively, some people find themselves with a maximum number of possible samples. This could be because they only have so many incubators, their population size is small or they have limited time and will collect as many samples as possible in that limited time. This is often used as an excuse for not completing an *a priori* power analysis - "I will collect as many as I can!". However, even in this case power analyses are useful because they can tell you what effect size you would be able to detect for a given sample size. In this case the answer might indicate that the experiment is doomed (i.e. could only detect the most massive of effect sizes) before it even begins.

In *retrospective* power analyses we want to try and explain why we did not reject our null hypothesis. Was it because our sample size was too small and the variance in our samples was too high or was it because the effect was non-existent or trivially small.

Power - the basics

In any statistical test there are 5 basic components:

1. Effect size (e.g. how different are two means?)
2. Sample size
3. Variance (or uncertainty)
4. α (Type I error rate)
5. β (Type II error rate)

Power increases with increasing sample size, effect size and α. Power decreases with increasing variance.

Power (1-β) ranges from zero to one where higher values represent greater power (better). Values higher than 0.8 are generally considered to be high

Performing a Power Analysis

There are several canned programs and web tools that will allow you to input 4 of the 5 components listed above and get the 5th. Here is an example: Iowa.

Performing a Power Analysis by Hand

I don't expect that many (any?) of you will ever do a power analysis by hand, but I think it is a helpful exercise to work through so that you can see how all the various parts contribute to the calculation.

Let's start with a simple example where we are interested in testing whether a standard (assumed value) has been exceeded by our sample mean. The test statistic for this test is a t test-statistic. The way you would normally see this is:

You might not be familiar with the actual formula but the test statistic for whether a particular value has been exceeded is simply the difference between the sample mean and the assumed standard (Δ_x) divided

$$t_\alpha = \frac{\Delta_x}{s/\sqrt{n}}$$

by the standard error, which is the standard deviation (s) divided by the square root of the sample size (n). What you probably don't know is that this calculation defaults to a power of 0.5 and a $t_{0.5} = 0$. We can expand the basic t-test formula to allow us to let power vary as:

$$t_\alpha + t_{1-\beta} = \frac{\Delta_x}{s/\sqrt{n}}$$

So this is only slightly more complicated and it allows us to do a power analysis for such a test. Notice that the 5 parts of the

equation correspond to the 5 items listed above (effect size, sample size, variance, alpha and beta).

We now have everything that we need to do a power analysis. All we need to do is rearrange this equation to answer either of the following questions:

1. I want a certain amount of power to detect a given effect size. How large does my experiment need to be? That is. given β, Δ_x and α, how large does n need to be?

2. I have a certain budget for a fixed experiment size. What will my power be to detect a given effect size? Given n, Δ_x and α, what is my β ?

Power - A worked example.

Let's say we are planning an experiment to test the effects of Nitrogen on plant biomass and we want to know how many samples we ought to collect. We know from the literature or from a pilot study that average biomass is ~ 103 kg/ha and the sd of this biomass is 16 kg/ha. Note that the latter value is often hard to estimate so sometimes you just need to guess. You want high power so lets say β = 0.10 (power = 0.90). You don't believe in all this P = 0.05 junk so you are going to set your α = 0.10. You are predicting an *increase* in biomass so this is a one-tailed α. Finally, you use your biological intuition to determine that you think a 20% increase in biomass is a biologically meaningful effect.

One of the great benefits of an *a priori* power analysis is that it forces you to think about biologically meaningful effect sizes BEFORE you perform your study. So a 20% increase from 103 kg/ha would be an increase of 20.6 kg/ha.

So we have all the ingredients for our power analysis. We can now take the formula for a t-test to compare the means of two groups (slightly different than above)...

$$t_{\alpha} + t_{1-\beta} = \frac{\Delta_x}{s/\sqrt{n}}$$

...and rearrange it to solve for the n that would be needed.

$$\hat{n} = 2(t_{\alpha} + t_{1-\beta})^2 \frac{s^2}{\Delta_x{}^2}$$

From above, $s^2 = 256$, $\Delta_x^2 = 424.36$, and n is the sample size in each group. Now one complication is that t_{α} and $t_{1-\beta}$ depend on the df, which depends on n (df = 2n-2 in this case). We solve this problem by starting with a large df ($t_{0.1,\ df = large} = 1.28$).

n* = 2*(1.28 + 1.28)2 * 0.60

= 7.86

Remember that n* is the sample size in each group. So our new df = 14 and $t_{0.1,\ df = large} = 1.35$

n** = 2*(1.35 + 1.35)2 * 0.60

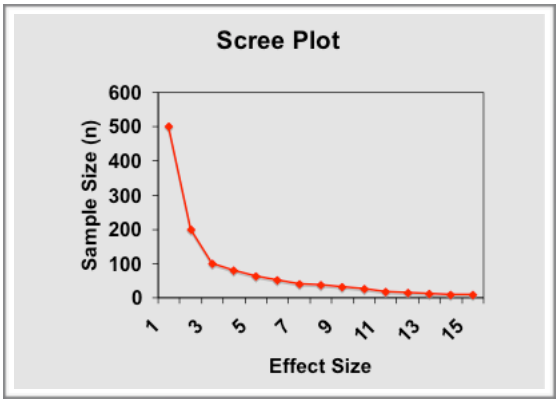
= 8.75

If we repeat this process a couple more times we converge on a value of 8.7. So if we wanted to detect a 20% increase in biomass in our N treated plants with 90% power and an α of 0.1 then we would need to have 9 control plants and 9 experimental plants.

Scree Plots

Sometimes we don't have a known effect size that we are aiming for. Instead we want to see how various sample sizes affect the detectable effect size. We can do this by performing a power analysis for many different effect sizes. If we plot n against the effect size this is called a *Scree Plot*. It is called this because it looks like the rocks (scree) that pile up at the base of a mountain. The main point of a scree plot is that it is non-linear, so there is an area of high-returns and an area of low-returns. For example, if

your sample size is below 100 you will see a pretty large decrease the detectable effect size for a given increase in sample size. However, once the sample size gets above 100 you will get only a marginal decrease in the detectable effect size for even very large increases in sample size (e.g. 200 to 500).



Retrospective Power Analyses

Reviewers often mistakenly ask for and authors often mistakenly provide an estimate of power for a non-significant analysis that has already been performed. This is called "Observed Power". **This approach is NOT useful.** It will always yield low power. This is because there are only 5 pieces to the puzzle. If you analysis has already yielded a P-value greater than 0.05 the **by definition** you did not have adequate power to detect the observed effect size with the observed sample size at an α of 0.05. So 'Observed Power' is just another way of stating your P-value. More specifically, recall our equation above...

$$t_{\alpha} + t_{1-\beta} = \frac{\Delta_x}{s/\sqrt{n}}$$

When the effect size, sd and n are fixed then t_{α} (and hence P) are related negatively to power. There is a graph in the Hoenig & Heisey (2001) paper that shows this nicely.

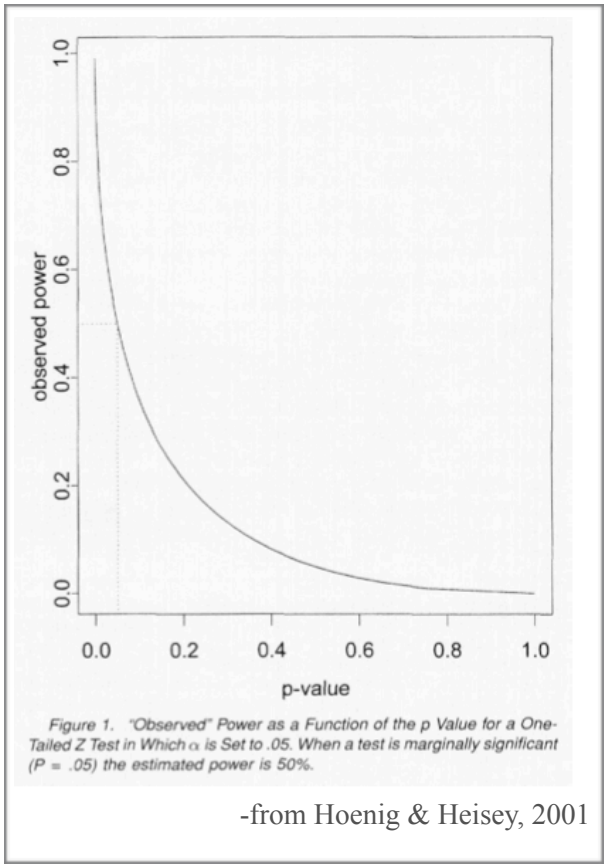


Figure 1. "Observed" Power as a Function of the p Value for a One-Tailed Z Test in Which α is Set to .05. When a test is marginally significant (P = .05) the estimated power is 50%.

-from Hoenig & Heisey, 2001

So when are a posteriori power analyses useful? They can be useful to determine what effect size could have been detected in your study given sd, n. In this way you can determine whether your experiment could have rejected the null hypothesis in interesting cases.

Could my experiment reject the null hypothesis in interesting cases?

In my view most interesting retrospective questions can be answered with effect sizes and consideration of biological significance.

Another way in which retrospective power analyses are useful is if you turn them around and use them as prospective power analyses for future studies.

Power Summary

- Power is the probability of rejecting the null when it is false
- Power is increased by sample size, effect size, α
- Power is decreased by variance
- *A Priori* power analyses are essential!
- *A Posteriori* (retrospective) power analyses are often asked for but are of limited use