Inferential Stats 1: Point Estimation

The purpose of an inferential statistic is to make a statement about the population from which a sample was taken. Assume, for example, that you are interested in the mean level of anxiety suffered by graduate students. You could, I suppose, measure the anxiety level of every grad from every department at every university and thereby learn the true value, but this is impractical. So, instead, you take a sample of grads, measure only their anxiety levels, and then make an *inference* (i.e., a best guess) about the entire population.

Clearly the quality of this inferential process depends on (at least) two things: how well the sample represents the target group, and how well one can calculate an estimate based on a sample. Here we will focus on the second part. The question of external validity -- which is the "how well the sample represents the target group" part -- is not really a statistical issue.

In general, all inferences start with a "best estimate" of the population characteristic of interest. In most cases, the formula for the best guess of something is the same as the formula for the same characteristic of a sample. For example, the best guess for the mean of the population, μ (mu), uses the same formula as that for the mean of a sample, $\overline{X}$ -- i.e., $\Sigma X/N$. In other words, the best guess for the population mean is the sample mean. But this is not always the case; there are attributes of populations that are best estimated by something other than the value for the sample.

Note: if all that you want to do is make a best guess about μ, then, once you have found $\overline{X}$ (i.e., the mean of your sample), you are done. The best guess for μ is $\overline{X}$, end of story. However, because we usually want to make a statement about what μ *really* is -- or, at least, have an idea of how wrong we might be -- we have to do more. Even stronger: in my opinion, just calculating a best guess and not doing any additional work is highly dangerous. Paired with every best guess should at least be an estimate of how wrong you might be. And it's in the estimating of how wrong you might be that the fun comes in.

Point Estimation

The simplest form of inferential statistics is when we want to get an idea about a single parameter of a population, such as its mean, μ. (Note: it is standard to use Greek letters to stand for population values.) This is referred to as *point estimation*, because we are trying to estimate a single value or point that lies somewhere on some continuum. In this case, there is no "null hypothesis" or default estimate or anything else, which might make point estimation seem quite different from all other forms of inferential statistics. In reality, however, the process is extremely similar and is based on the same procedures and assumptions, so it's a good place to start. (Plus: as we'll see later, one form of *t*-test is exactly the same as point estimation.)

As suggested above, point estimation requires that we calculate two things: a best guess and an associated estimate of how wrong we might be. The combination of these two things is often written as: $\overline{X} \pm \varepsilon$, where ε (epsilon) is a particular measure of how wrong we might be, called the "standard error." (Technically, the usual notation is wrong or, at least, very misleading, because it implies that μ must be one of two specific values -- either $\overline{X} + \varepsilon$ or $\overline{X} - \varepsilon$ -- which is

almost never true, but as long as we all know what's really being said, it's OK.)

As mentioned above, calculating the best guess, such as $\overline{X}$, is usually pretty easy and often involves the same formula as the characteristic of the sample. The hard part is calculating ε. But before launching into this, we need to ask why there is any error in the first place. The answer is simple: because we only took a sample … because we didn't measure every value in the population. In other words, because the population from which the sample was taken contains a huge (and maybe infinite) number of values, and the sample is only a random subset of these values, not all samples (even if very large) will be the same. Therefore, any particular value -- such as $\overline{X}$ -- that can be calculated from a sample will vary across samples. The relative frequencies of all possible values of $\overline{X}$ (across all possible samples of a given size) is the **sampling distribution of $\overline{X}$**. The sampling distribution of $\overline{X}$ is an unknown and must be estimated. (In order to truly know the sampling distribution, you'd have to take an infinite number of samples, and that isn't going to happen; that's why it's an unknown.) But, in order to know *exactly* how wrong we might be, we have to "know" the sampling distribution.

☞ Stop for a moment and think about that last part. When we are trying to get an estimate of a population value, such as its mean, we calculate a best guess. We also calculate an estimate of how wrong we might be. But that estimate of how wrong we might be is also a guess, because we can only estimate the sampling distribution. This is the beginning of what could end up as an infinite regress. We might next calculate an estimate of how wrong our estimate of how wrong our first guess was. This would then be followed by an estimate of how wrong our estimate of how wrong our estimate of how wrong our first guess was, etc. Well, at some point, this needs to stop. The place where we stop (fortunately) is very near the beginning. We start with a best guess. Then we also get an estimate of how wrong our best guess might be. And there we stop. All of the rest of the infinite regress is dumped into a single thing and is not explored any further. That single thing, in case you can't wait, is the degrees of freedom.

A single sample (which is all we'll ever have, in practice) usually gives us a rather incomplete picture of the sampling distribution of $\overline{X}$. In particular, while the sample can give us pretty good estimates of the center and the spread of the sampling distribution of $\overline{X}$, it does not provide anything close to exact values (unless the sample is huge) and it does not provide us with any details about the shape of the sampling distribution of $\overline{X}$. This is where the distinction between parametric and nonparametric statistics comes in.

Parametric statistics assume a specific shape for the sampling distribution of $\overline{X}$. In particular, nearly all of these procedures assume that the sampling distribution is normal (aka *Gaussian*; i.e., bell-shaped). A normal distribution is completely specified by just two variables -- the mean and the standard deviation -- because all other parameters are fixed. For example, the skew and kurtosis of a normal are always both zero. When the assumption of normality is combined with good estimates of the center and spread, the entire sampling distribution becomes known, allowing us to calculate a good estimate of how wrong we might be.

In contrast, **nonparametric statistics** do not assume a certain shape for the sampling distribution of $\overline{X}$. Therefore, these methods must use some other procedure (which we won't cover in any detail, but I'll give you an idea of how they work in lecture).

! Regardless of how one derives an estimate of the sampling distribution for $\overline{X}$, the same logic is employed under both parametric and nonparametric statistics. In other words, these procedures only differ in terms of how the shape of the sampling distribution for $\overline{X}$ is estimated; everything else is the same. Thus, parametric and nonparametric stats are not as different as you may have heard.

Now that we have completely specified the sampling distribution for $\overline{X}$ (which implies that we could make a plot of it, if we wished), we can now express how wrong we might be in a single number. It will then be a relatively simple last step to create a confidence interval for the true value of μ, if we wish. The only remaining issue is this: *how much confidence do we want to have?* And before you give the obvious answer of 100%, keep the following point in mind: because most sampling distributions are assumed to be normal, and because a normal distribution has infinitely long tails in both directions, the only way to be 100% confident is to say that μ must be somewhere between $-\infty$ and $+\infty$, which isn't very useful. (You already knew that the mean was between $-\infty$ and $+\infty$ before you took your sample, eh?) So we usually choose some high value that isn't 100% -- in psychology and neuroscience (and most other fields), we use 95%. Why? Because 95% confidence is the same as (or the complement to) the 5% rule that we use for statistical significance. Yeah, the 5% rule might be arbitrary, but at least we're being consistent.

Degrees of Freedom

In the process described above, we used the sample to do two things: (1) to get a value of $\overline{X}$ (our best guess for μ), and (2) to get an estimate of how wrong we might be, which, when combined with parametric assumptions or non-parametric procedures, will allow us to create a range or interval with a 95% chance of containing μ. These are separate and independent acts. To correct for errors in estimating μ using $\overline{X}$, we set up a confidence interval for μ, instead of just claiming that μ must be equal to $\overline{X}$ -- i.e., we admit that $\overline{X}$ is only a guess. To correct for errors in estimating how wrong we might be, we express the "quality" of this second part as a number, which is the degrees of freedom. In the case of point estimation, the degrees of freedom (df) is equal to $N-1$ (i.e., one less than the sample size).

OK, so how do we use the value of df to adjust our statements to match the quality of our guess as to how wrong we might be? In other words: How do we use this one number to avoid the infinite-regress issue mentioned above? In a nutshell, the way that the value of df is used is to adjust the shape of the sampling distribution. When df is ∞, we leave it as normal, since there can't be any error if we measured everything. When df is less than ∞, as it always will be in practice, we spread the distribution out a bit more, to cover our ~~*bleep!*~~ … er, to cover all possibilities. This spread-out version of the normal is the *t* distribution.

☞ Note: yes, the *t* distribution is slightly different in shape from the normal distribution (although it becomes the same as a normal as the degrees of freedom increase). This causes some people to believe that our parametric statistics don't actually assume that the sampling distribution of $\overline{X}$ is normal after all, at least when the sample size is small. This is not correct. The parametric statistics that we use always assume that the sampling distribution of $\overline{X}$ is normal; the use of the *t* distribution (instead of the *z* or the normal distribution) is to correct for errors in

estimating the spread. Again, we are absorbing all of the possible errors that would arise in the infinite regress mentioned above into this one thing. Not because we've stopped believing that the sampling distribution is normal, but because we need to be ready for the possibility that the sampling distribution that we created isn't wide enough to include the true value because we under-estimated the amount of error. That's why all *t* distributions are wider than the *z* distribution.

Side-note: Many textbooks include examples where the population variance is known, but the mean is not, under which the *z* distribution can be used. I don't do this, because this is pretty much ridiculous. I can think of no situation where you would know the variance of a population and not know its mean. I kind of see what they are trying to do -- gently move you towards the whole picture in small steps -- but I find the idea of talking about a best guess in a case where you know exactly how wrong your best guess might be to be a tad misleading. I think we should face the infinite-error regress problem from the start.

Creating a Confidence Interval

As mentioned above, one way to combine your best guess with an estimate of how wrong you might be is by giving the mean and the standard error in the form of $\overline{X} \pm \varepsilon$. (You might also want to attach the df to this, too, for completeness' sake.) An alternative (and less-and-less popular) way to do roughly the same thing is to set up an interval that has a 95% chance of including μ. This is where the *t* distribution comes in directly. The 95% CI is centered on $\overline{X}$ (since $\overline{X}$ is the best guess) and extends both downwards and upwards by an amount that is equal to the standard error multiplied by the value of ***t*** critical for the degrees of freedom. As the degrees of freedom become very, very high, ***t*** critical becomes 1.96, which is a magical number that ought to be memorized. When the value of df is lower (and, thus, more realistic), ***t*** critical is higher. With 100 df, for example, it is 1.98. With 10 df, it is 2.23. With only 1 df, it is 12.71. Thus, smaller samples have wider confidence intervals, not only because they have larger standard errors, but also because the odds of under-estimating the standard error is higher.

tl;dr

A point estimate has three components: a best guess, an estimate of how wrong this guess might be, and a measure of how confident we should be in our estimate of how wrong we might be. It's a very bad habit to give a best guess without, at least, also giving the estimate of how wrong you might be. Also, when parametric statistics are used, which is a huge majority of the time in our fields, the estimate of how wrong we might be is based on the Assumption of Normality … the next topic to cover.