The Assumption(s) of Normality

This is very complicated, so I'll provide two versions. At a minimum, you should know the short one. It would be great if you knew them both.

Short version: in order to do something as magical as provide a specific probability for observing a particular mean or a particular difference between two means, our statistical procedures must make some assumptions. One of these assumptions is that the sampling distribution of the mean is normal. That is, if you took a sample, calculated its mean, and wrote this down; then took another (independent) sample (from the same population) and got its mean and wrote it down; and did this an infinite number of times; then the distribution of the values that you wrote down would always be a perfect bell curve. While maybe surprising, this assumption turns out to be relatively uncontroversial, at least when each of the samples is large, such as $N \geq 30$. But in order to use the same statistical procedures for all sample sizes and in order for the underlying procedures to be as straight- forward as they are, we must expand this assumption to saying that all populations from which we take samples are normal. In other words, we have to assume that the data inside each of the samples are normal, not just that the means of the samples are normal. This is a very strong assumption and it probably isn't always true, but we have to assume this to use our procedures. Luckily, there are simple ways to protect ourselves from the problems that would arise if these assumptions are not true.

Now, the long version….

Nearly all of the inferential statistics that psychologists use (e.g., *t*-tests, ANOVA, simple regression, and MRC) rely upon something that is called the "Assumption of Normality." In other words, these statistical procedures are based on the assumption that the value of interest (which is calculated from the sample) will exhibit a bell-curve distribution function if oodles of random samples are taken and the distribution of the calculated value (across samples) is plotted. This is why these statistical procedures are called *parametric*. By definition, parametric stats are those that make assumptions about the shape of the sampling distribution of the value of interest (i.e., they make assumptions about the skew and kurtosis parameters, among other things; hence the name). The shape that is assumed by all of the parametric stats that we will discuss is *normal* (i.e., skew and kurtosis are both zero). The only statistic of interest that we will discuss here is the mean.

What is assumed to be normal?

When you take the parametric approach to inferential statistics, the values that are assumed to be normally distributed are the means across samples. To be clear: the Assumption of Normality (note the upper case) that underlies parametric stats does not assert that the observations within a given sample are normally distributed, nor does it assert that the values within the population (from which the sample was taken) are normal. (At least, not yet.) The core element of the Assumption of Normality asserts that the distribution of sample means (across independent samples) is normal. In technical terms, the Assumption of Normality claims that *the sampling distribution of the mean is normal* or that *the distribution of means across samples is normal.*

Example: Imagine (again) that you are interested in the average level of anxiety suffered by graduate students. Therefore, you take a group of grads (i.e., a random sample) and measure their levels of anxiety. Then you calculate the mean level of anxiety across all of the subjects. This final value is the sample mean. The Assumption of Normality says that if you repeat the above sequence many many many times and plot the sample means, the distribution would be normal. Note that I never said anything about the distribution of anxiety levels within given samples, nor did I say anything about the distribution of anxiety levels in the population that was sampled. I only said that the distribution of sample means would be normal. And again, there are two ways to express this: "the distribution of sample means is normal" and/or "the sampling distribution of the mean is normal." Both are correct as they imply the same thing.

Why do we make this assumption?

As mentioned in the previous chapter, in order to know how wrong a best guess might be and/or to set up a confidence interval for some target value, we must estimate the sampling distribution of the characteristic of interest. In the analyses that we perform, the characteristic of interest is almost always the mean. Therefore, we must estimate the sampling distribution of the mean.

The sample, itself, does not provide enough information for us to do this. It gives us a start, but we still have to fill in certain blanks in order to derive the center, spread, and shape of the sampling distribution of the mean. In parametric statistics, we fill in the blanks concerning shape by assuming that the sampling distribution of the mean is normal.

Why do we assume that the sampling distribution of the mean is normal, as opposed to some other shape?

The short and flippant answer to this question is that we had to assume something, and normality seemed as good as any other. This works in undergrad courses; it won't work here.

The long and formal answer to this question relies on **Central Limit Theorem** which says that: *given random and independent samples of N observations each, the distribution of sample means approaches normality as the size of N increases, regardless of the shape of the population distribution*. Note that the last part of this statement removes any conditions on the shape of population distribution from which the samples are taken. No matter what distribution you start with (i.e., no matter what the shape of the population), the distribution of sample means becomes normal as the size of the samples increases. (I've also seen this called "the Normal Law.")

✂ The long-winded, technical version of Central Limit Theorem is this: if a population has finite variance σ^2 and a finite mean μ, then the distribution of sample means (from an infinite set of independent samples of N independent observations each) approaches a normal distribution (with variance σ^2/N and mean μ) as the sample size increases, regardless of the shape of population distribution.

In other words, as long as each sample contains a very large number of observations, the sampling distribution of the mean ***must*** be normal. So if we're going to assume one thing for all situations, it has to be a normal, because the normal is always correct for large samples.

The one issue left unresolved is this: how big does N have to be in order for the sampling distribution of the mean to always be normal? The answer to this question depends on the shape of the population from which the samples are being taken. To understand why, we must say a few more things about the normal distribution. As a preview: if the population is normal, than any size sample will work, but if the population is outrageously non-normal, you'll need a decent-sized sample.

The **First Known Property** of the Normal Distribution says that: *given random and independent samples of N observations each (taken from a normal distribution), the distribution of sample means is normal and unbiased (i.e., centered on the mean of the population), regardless of the size of N.*

✂ The long-winded, technical version of this property is: if a population has finite variance σ^2 and a finite mean μ and is normally distributed, then the distribution of sample means (from an infinite set of independent samples of N independent observations each) must be normally distributed (with variance σ^2/N and mean μ), regardless of the size of N.

Therefore, if the population distribution is normal, then even an N of 1 will produce a sampling distribution of the mean that is normal (by the First Known Property). As the population is made less and less normal (e.g., by adding in a lot of skew and/or messing with the kurtosis), a larger and larger N will be required. In general, it is said that Central Limit Theorem "kicks in" at an N of about 30. In other words, as long as the sample is based on 30 or more observations, the sampling distribution of the mean can be safely assumed to be normal.

✂ If you're wondering where the number 30 comes from (and whether it needs to be wiped off and/or disinfected before being used), the answer is this: Take the worst-case scenario (i.e., a population distribution that is the farthest from normal); this is the exponential. Now ask: if the population has an exponential distribution, how big does N have to be in order for the sampling distribution of the mean to be close enough to normal for practical purposes? Answer: around 30. (Note: this is a case where extensive computer simulation has proved to be quite useful. No-one ever "proved" that 30 is sufficient; this rule-of-thumb was developed by having a computer do what are called "Monte Carlo simulations" for a month or two.) (Note, also: observed data in psychology and neuroscience are rarely as "bad" as a true exponential and, so, Ns of 10 or more are almost always enough to correct for any problems, but we still talk about 30 to cover every possibility.)

At this point let's stop for a moment and review. 1. Parametric statistics work by making an assumption about the shape of the sampling distribution of the characteristic of interest; the particular assumption that all of our parametric stats make is that the sampling distribution of the mean is normal. (To be clear: we assume that if we took a whole bunch of samples, calculated the mean for each, and then made a plot of these values, the distribution of these means would be normal.) 2. As long as the sample size, N, is at least 30 and we're making an inference about the mean, then this assumption must be true (by Central Limit Theory plus some simulations), so all's well if you always use large samples to make inferences about the mean.

The remaining problem is this: we want to make the same assumption(s) for all of our inferential procedures and we sometimes use samples that are smaller than 30. Therefore, as of now, we are not guaranteed to be safe. Without doing more or assuming some more, our procedures might not be warranted when samples are small.

This is where the second version of the Assumption of Normality (caps again) comes in. By the First Known Property of the Normal, if the population is normal to start with, then the means from samples of any size will be normally distributed. In fact, when the population is normal, even an N of 1 will produce a normal distribution (since you're just reproducing the original distribution). So, if we assume that our populations are normal, then we're always safe when making the parametric assumptions about the sampling distribution, regardless of sample size.

To prevent us from having to use one set of statistical procedures for large (30+) samples and another set of procedures for smaller samples, the above is exactly what we do: we assume that the population is normal. (This removes any reliance on the Monte Carlo simulations [which is good, because simulations annoy people who always want proofs].) The one thing about this that (rightfully) bothers some people is that we know -- from experience -- that many characteristics of interest to psychologists are not normal. This leaves us with three options: 1. Carry on regardless, banking on the idea that minor violations of the Assumption of Normality (at the sample-means level) will not cause too much grief -- the fancy way of saying this is "we capitalize of the robustness of the underlying statistical model," but it really boils down to looking away and whistling. 2. Remember that we only need a sample size as big as 30 to guarantee normality if we started with the worst-case population distribution -- viz., an exponential -- and psychological variables are rare this bad, so a sample size of only 10 or so will probably be enough to "fix" the non-normalness of any psych data; in other words, with a little background knowledge concerning the shape of your raw data, you can make a good guess as to how big your samples need to be to be safe (and it never seems to be bigger than 10 and is usually as small as 2, 3, or 4, so we're probably always safe since nobody I know collects samples this small). 3. Always test to see if you are notably violating the Assumption of Normality (at the level of raw data) and do something to make the data normal (if they aren't) before running any inferential stats. The third approach is the one that I'll show you (after one brief digression).

Another Reason to Assume that the Population is Normal

Although this issue is seldom mentioned, there is another reason to expand the Assumption of Normality such that it applies down at the level of the individual values in the population (as opposed to only up at the level of the sample means). As hinted at in the previous chapter, the mean and the standard deviation of the sample are used in very different ways. In point estimation, the sample mean is used as a "best guess" for the population mean, while the sample standard deviation (together with a few other things) is used to estimate how wrong you might be. Only in the final step (when one calculates a confidence interval or a probability value), do these two things come back into contact. Until this last step, the two are kept apart.

In order to see why this gives us another reason to assume that populations are normal, note the following two points. First, it is assumed that any error in estimating the population mean is independent of any error in estimating how wrong we might be. (If this assumption is not made, then the math becomes a nightmare ... or so I've been told.) Second, the **Second Known Property** of the Normal Distribution says that: *given random and independent observations (from a normal distribution), the sample mean and sample variance are independent*. In other words, when you take a sample and use it to estimate both the mean and the variance of the population, the amount by which you might be wrong about the mean is a completely separate (statistically independent) issue from how wrong you might be about the variance. As it turns out, the normal distribution is the only distribution for which this is true. In every other case, the two errors are in some way related, such as over-estimates of the mean go hand-in-hand with either over- or under-estimates of the variance.

Therefore, if we are going to assume that our estimates of the population mean and variance are independent (in order to simplify the mathematics involved, as we do), and we are going to use the sample mean and the sample variance to make these estimates (as we do), then we need the sample mean and sample variance to be independent. The only distribution for which this is true is the normal. Therefore, we assume that populations are normal.

Testing the Assumption of Normality

If you take the idea of "assuming" seriously, then you don't have to test the shape of your data. But if you happen to know that your assumptions are sometimes violated -- which, starting now, you do, because I'm telling you that sometimes our data aren't normal -- then you should probably do something before carrying on.

There are at least two approaches to this. The more formal approach is to conduct a statistical test of the Assumption of Normality (as it applies to the shape of the sample). This is most-often done using either the Kolmogorov-Smirnov or the Shapiro-Wilk Test, which are both non-parametric tests that allow you to check the shape of a sample against a variety of known, popular shapes, including the normal. If the resulting *p*-value is under .05, then we have significant evidence that the sample is not normal, so you're "hoping" for a *p*-value of .05 or above.

✂ Some careful folks say that you should reject the Assumption of Normality if the *p*-value is anything under .10, instead of under .05, because they know that the K-S and S-W tests are not very good at detecting deviations from the target shape (i.e., these tests are not very powerful). I, personally, use the .10 rule, but you're not obligated to join me. Just testing for normality at all puts you in the 99th percentile of all behavioral researchers.

So which test should you use … K-S or S-W? This is a place where different sub-fields of psychology and neuroscience have different preferences and I'll discuss this in class. (In brief, those who always work with large samples, such as those who use surveys, use K-S, while those who often use small samples, such as those studying information processing, use S-W.) For now, I'll explain how you can get both using SPSS.

The easiest way to conduct tests of normality (and a good time to do this) is at the same time that you get the descriptive statistics. Assuming that you use **Analyze… Descriptive Statistics… Explore…** to do this, all you have to do is go into the Plots sub-menu and (by clicking Plots on the upper right side of the Explore window) and then put a check-mark next to Normality plots with tests. Now the output will include a section labeled **Tests of Normality**, with both the K-S and S-W findings.

If you would like to try the K-S test now, please use the data in *Demo11A.sav* from the first practicum. Don't bother splitting up the data by Experience; for now, just rerun Explore with Normality plots with tests turned on. The *p*-values for mACC_DS1 are .125 for K-S and .151 for S-W. The *p*-values for mACC_DS5 are .200 for K-S and .444 for S-W. All of this implies that these data are normal (enough) for our standard procedures, no matter which test or criterion you use.

Other people use informal rules-of-thumb to decide whether their data is normal enough, such as only worrying when either skew or kurtosis is outside the range of ±2.00. I'm not a fan of this approach and won't say much more about it.

As to what you're supposed to do when your data aren't normal, that's in the next chapter.