When do you use the t distribution?
When do you use the normal distribution?
Why?

Ron Michener
August 2002

Let me begin by admitting that what follows will be too much information for many of you. If you are struggling, and you want a rule, not an explanation, use the flow charts in *Anderson, Sweeny, and Williams* – Figure 8.8 on page 301 and Figure 9.17 on page 370. I am writing this for students who are frustrated by the fact that not all statistics books and not all statistics courses seem to use the same rules, and want a more complete explanation.

To begin, here are the facts for four important cases:

I) Sigma, the population standard deviation, is *known.*

a) *The population is normally distributed.* In this case, the sample mean $\bar{X}$ is exactly normally distributed, regardless of sample size, and the random variable $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ always has an exact standard normal distribution (since it is just a linear transformation of $\bar{X}$).

b) *The population is not normally distributed.* In this case, we must rely on the central limit theorem to ensure that $\bar{X}$ is normally distributed. The central limit theorem says (subject to some mild assumptions) that $\bar{X}$ is *approximately* normally distributed in *large* samples. It guarantees nothing about small samples. If the population itself is distributed in a way that is *approximately* normal (roughly symmetric, with few outliers) the distribution of $\bar{X}$ approaches the normal distribution quite quickly. See figure 7.5 on page 259 for an illustration. The central limit theorem also tells us that as the sample size $n \to \infty$, the distribution of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ approaches the standard normal distribution.

II) Sigma is unknown, and must be estimated.

a) *The population is normally distributed.* In this case, regardless of sample size, the sample mean $\bar{X}$ is exactly normally distributed, and the random variable $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ always has a t distribution with n-1 degrees of freedom.

The reason $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ doesn't have a *normal* distribution is that s in the denominator is a random variable. While the numerator is a normally distributed random variable, the ratio of these two random variables has a t distribution.

b) *The population is not normally distributed.* In this case, the distribution of $\bar{X}$ approaches a normal distribution as sample size $n \to \infty$. The random variable $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ will have some unknown distribution for small sample sizes – and it never has a t distribution. As $n \to \infty$, the distribution of the random variable $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ will approach a standard normal distribution.

There are two other useful facts: one historical and one statistical.

The historical tidbit is that when the first t tables were prepared in the early 20th century, there were no computers and no pocket calculators, and the mathematicians preparing the table had to do so by doing laborious numerical integrations by hand. Since this was time consuming, and the t distribution approaches the normal distribution as $n \to \infty$, they decided to quit once they'd carried the table up to 30 degrees of freedom. For many years, this is where t tables ended, and a generation of statisticians grew up learning that for larger sample sizes, one should simply consult the normal tables. This, more than any deep considerations arising from statistical theory, helped make 30 the cutoff between big and small samples.

Second, statisticians have discovered the t distribution is *robust*. Although the distribution of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ will not be exactly a t distribution unless the population is normally distributed, the t distribution turns out to be pretty close to correct if the population is somewhere near normally distributed and the sample size is not extremely small. What I mean by "pretty close" is that if you construct a large number of 95% confidence intervals based on the t distribution, and the population is only approximately normally distributed, the actual proportion of confidence intervals bracketing the population mean won't be 95%, but rather something like 94.1% or 95.3% - some number not shockingly different from 95%. Many procedures in statistics are not robust, and it is fortunate that the t distribution is.

How does this explain the conflicting advice?

Well, when σ is known, the advice isn't usually conflicting. Generally, one uses the z-table or nothing at all. If the population is normally distributed, then the z-table is exactly correct for all sample sizes. If the population isn't normally distributed, you can still use the z-table for large sample sizes, because you can invoke the central limit theorem. What is a large sample size? The usual rule of thumb is 30.

When σ is not known, the common case, the situation is most ambiguous, and here is where most of the conflicting advice arises.

For small sample sizes, you can't do *anything* without assuming the population is normally distributed. Therefore, when sample sizes are small, it is common for statisticians to make the convenient assumption that the population *is* normally distributed in order to proceed. Of course, if there is clear evidence that the population is wildly non-normal, you must resign yourself to gathering more data, or abandoning the analysis. But statisticians take comfort in the knowledge that the t distribution is robust, and that they are likely to be approximately right even if their assumption of normality is not precisely correct.

For larger sample sizes, you should always use the t distribution if you know the population is normally distributed. But in common practice, you seldom have any assurance that the population is normally distributed. In the small sample case, you were *forced* to assume normality to proceed, but in the large sample case, you have another theorem to fall back on – namely the fact that as $n \to \infty$, the distribution of the random variable $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ will approach that of a standard normal distribution, even when the population is not normally distributed. The A, S, & W approach is to switch over to using the z, based on the limiting distribution of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$, rather than continue to rely on the unproven assumption of normality.

Not everyone thinks this is a good idea. After all, you are using an approximation in either case, whether you stick with the t or switch to the z. If you stick with the t distribution for large sample sizes, you are assuming the population is well approximated by the normal distribution. If you switch to the z distribution, you are assuming that your sample size is large enough to guarantee that $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is well approximated by the limiting standard normal distribution. T values are always a bit bigger than z values, so that t values always produce more conservative estimates (wider confidence intervals, and bigger p values). Some people argue you should stick with the t distribution even in large samples because it gives more conservative answers.

There is another reason for sticking with the t distribution in large samples. When non-normality of a population is a problem, it is usually the case that the population has fatter tails than the normal distribution. The fat tails will make the sample estimate s more variable than it would have been if the distribution were normal, which implies that $\dfrac{\bar{X}-\mu}{s/\sqrt{n}}$ will fluctuate over a wider range than in the case of a normal population. If this is true, confidence intervals based on the t distribution won't be wide enough, and hypothesis tests based on the t distribution will underestimate the true p values. However, in this case, using the z table will give a *worse* approximation. The theoretical justification for using the t table in this case is non-existent, but using the t is nonetheless likely to come closer to the truth than you'd get using the z table.

In short, while I think there is room for reasonable people to differ, I personally think it makes more sense to assume normality and use the t table for large sample sizes. I do not try to impose this view in class, however, because the actual difference between a t value and a z value is small when sample sizes are large, and many students would find it very confusing to have the instructor arguing with the presentation in the textbook.