

t vs z, why does it matter

When we use the z , the Standard Normal, curve, we are assuming, not only that the data, or at least the sample mean, follows a normal distribution, BUT ALSO the true variance of the data, σ^2 , is known. This never happens in real life, so what can we do? We can estimate the true standard deviation, σ , with the sample standard deviation, s . But there is a problem: 1. we're now estimating 2 things, AND 2. the sample sd, s , can underestimate as often as overestimate. To compensate, we use a t instead of a z .

The distribution of t is quite similar to the z , the Standard Normal. It is centered at *zero*, but instead of defining the spread by the standard deviation, σ , it is defined by the *degrees of freedom* or just *df*. For the one-sample case, the *df* of $t = n - 1$, the sample size minus 1. [We lose one degree of freedom because now if we know \bar{x} , s , and $n - 1$ of the observations, the last (n^{th}) observation is fixed.] As our sample size and hence the *df* increases, the t distribution gets taller and less spread out. This equates to s getting closer to the true value, σ , as we get more data. When $s = \sigma$, the t confidence interval will be wider than the z since we are unsure of the true the variability of the data. There are times when our estimate, s , is so much less than σ that even using the t doesn't quite give us a wide enough interval, but this is rare. What happens when our interval is too narrower? Well, we may not cover the true mean!

