Confidence Interval Illustration

Suppose we have a random process following a Normal PDF, with a mean of 100 units, and standard deviation ("sigma") of 3 units. We take samples from this distribution of size 10, independently, 200 times. For each sample we estimate the mean, the sigma, and calculate a "95% confidence interval" for the mean. We usually would only do this calculation process once, in which case we would say something like "There is a 95% probability that the true mean of the random process lies within this confidence interval." Actually, the true mean either does, or does not, lie within that single interval. There is no probability about it- the mean is either in the specific interval we found, or it isn't.

The correct interpretation of a confidence interval is that, *repeated many times*, the exact **procedure** we are using will create intervals that will include the true mean value with the specified probability. The procedure is: (a) taking samples of a given size; (b) estimating the mean and sigma of that sample; (c) constructing a "confidence interval" on the mean using a t-statistic at a specified probability level, usually 95%, two-sided (two-sided since our estimated mean may be greater or less than the true mean).

Below is a figure that shows 200 replications of the procedure; The red intervals do not include the true mean of 100, and there are 10 of these in this run. So 190/200 intervals do include the mean, for a probability in this run of exactly 95%. This will vary a bit, of course, as we repeat this experiment, but remain close to 95%. This shows that the estimation process with a sample size of 10, and the t-statistic for (0.975, 9df), 2.262 (larger than the z-statistic of 1.96), does indeed create intervals that include the true mean with 95% probability. Note in this context that t*he t-statistic is **always** used when the variability (sigma) is estimated directly from a given dataset; the z-statistic is only used if the variability is somehow considered independently "known" and thus is not estimated from the dataset.* ***Sample size has nothing to do with the choice of z vs. t.***

It happens, as another of my applets ("z vs. t") shows, that as the sample size increases the t-distribution looks very much like the z (standard Normal). As a matter of convenience in olden times before computers and calculators, statistical tables in books would only take the t-distribution up to df of about 30 or so, because, *numerically*, the tabulated values became indistinguishable from z, so we were instructed to go from the t-table to the z-table, above df of 30. *This does not mean that the mathematical form of the t-distribution magically transforms into the mathematical form of the standard Normal at df 30*, as elementary statistics textbooks, and, unfortunately, some teachers, imply. See the z-vs-t applet discussion for more on this.