Sampling Theory for Sample Means

Assumptions:

Population
Sample obtained from the population via random sampling

Random Sampling can be “with replacement” or “without replacement”.
If population is much larger than sample, “with” or “without” does not matter
Otherwise:
Need correction factor for finite population when sampling without replacement.

$$\boxed{SD_{sample\,.mean} = \frac{SD_{population}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}}$$

Important Feature of a Sample Mean – its variability
Why impt? Because it tells you how close the sample mean is to the population mean (we are usually more interested in the population mean than the sample means.)
Note the logical connection between variability of a sample mean and the distance of the sample mean from the population mean.

We measure the variability of the sample mean by its SD.
(Note that this usually requires some indirect method because in a typical sampling situation we only have one sample mean.)

We **estimate** the variability of the sample mean by first estimating the variability of the sample values by its SD, and then dividing it by the square root of the sample size. Ignoring the correction factor we have, exactly,

$$SD_{samplemean} = SD_{population} / \sqrt{n}$$

But usually in a sampling situation we do not know $SD_{population.}$ So in this case we estimate it with the SD of the sample: SD_{sample}.

$$estimatedSD_{samplemean} = SD_{sample} / \sqrt{n}$$

For example, if I have 25 sample values as a result of a random sample of size 25 from a population, and I calculate its mean to be 3.0 and its SD to be 1.5, the precision of this sample mean is estimated by its SD which is 1.5/√25 = 0.3, so we could say that the population mean is estimated as 3.0 but only with an SD of 0.3. Sometimes this is written 3.0±0.3 (which is a little ambiguous as it stands since the 0.3 in this case is 1 SD but sometimes we like 2 SDs – more on this below.)

Approximate Normality of Sample Means:

An important result in the theory of sample means is that the distribution of sample means taken from a given population is approximately Normal, with the approximation getting better and better as the sample size increases. For our purposes, we will always assume this approximation is good enough. This is a useful result since we know that normal distributions have 95% of their values within 2 SDs of the population mean. So if we write mean± 2SDs, we have an interval that includes the population mean 95% of the time.

In the previous paragraph, we talked about the "distribution" of sample means. But usually we only have one mean since we only have one sample (of size 25 say). So where is the "distribution"? We need to imagine the sampling process having been performed many times so that the one we have data for is just one of those times. The reason for this mental exercise is that, if we are using the sample mean to estimate the population mean, we need to be aware of the fact that every sample will have a different mean, and there will very likely be error in our estimate. Thinking of the distribution of sample means allows us to think about how far from the population mean the sample mean might be. In other words, we need to think of this imaginary sampling to describe the accuracy of our estimate – the SD of the sample mean.

The interval mean ± 2SDs is called a **95% Confidence Interval for the Population Mean.** Note that it is an interval estimate of a single numerical value. It will be "correct" 19 times out of 20.