1. For a random variable (r.v.) *X* with mean *μ* and standard deviation *σ*:
   1. is the sample mean statistic, which is a point estimator of the population mean *μ*.
   2. is the value of the sample mean, which is a point estimate of the population mean *μ*.
   3. is the sample standard deviation statistic, which is a point estimator of the population s.d. *σ*.
   4. is the value of the sample standard deviation, which is a point estimate of the population s.d. *σ*.

That is, when we don’t know the value of *μ*, we can use as an estimate, and when don’t know the value of *σ*, we can use the value of as an estimate.

1. Because and are sample statistics, they are random variables, and differ from sample to sample.
2. has a normal distribution with mean and standard deviation (where *n* is the sample size) if one of the following holds:
3. *X* is normal (regardless of sample size *n*)
4. *X* has any distribution but sample size *n* > 30 (according to the Central Limit Theorem)
5. 95% CI on is defined as: .
6. Recall that 1.96 comes from the standard normal table. For any normal random variable, 95% of all the values lie within 1.96 standard deviations of the mean. In the formula , we’re simply saying that 95% of all values of the normal r.v. lie within 1.96 standard deviations (calculated as ) of the mean .
7. The calculations of the CI are done using obtained from the sample data, as . Assume that is calculated to be some value *a*, and is calculated to be some value b. Then we can say that is between *a* and *b* with 95% confidence.
8. A note on interpretation: We CANNOT say that the probability that is between *a* and *b* is 95%. But what does it mean that the value of is between *a* and *b* with 95% confidence? It means that if you have a very large number of samples drawn from r.v. *X*, and a sample mean and a confidence interval calculated from each of these samples, then 95% of these samples will contain *μ*, and the remaining 5% will not. There’s nothing really special about the *a* or *b* we get from our sample, and there’s no way of really knowing whether the true value of *μ* is between these values. However, because 95% of all possible 95% confidence intervals (from all samples of size *n* drawn from *X*) contain *μ*, if our sample is not one of the 95%, and *μ* is not between *a* and *b*, we are simply unlucky.
9. Again, in practice, is rarely known when is not known. So we have to use the value of the sample s.d., *s* as a point estimate of *σ* in the formulas for the confidence interval. So:
10. When *n* > 30, *s* is generally close to *σ*. So, regardless of the distribution of the r.v. *X*, the 95% confidence interval formula becomes:

95% CI =

1. When *X* is normal and n < 30, *s* is no longer close to *σ* (i.e., it’s the best available estimate of *σ*, but still not that great), so the quantity no longer has a standard normal (*z*) distribution, but rather, a *t* distribution with *ν = n – 1* degrees of freedom. The formula for the 95% CI becomes: , and can be obtained from Excel using the formula *=tinv(0.05, n – 1*). Note that we put in 0.05 and NOT 0.025 in the Excel formula. In R, we can write *abs(qt(0.025, n-1))*.
2. Because the *t*-distribution works well when *n* is small, and is essentially non-distinguishable from the *z*-distribution when n is large, most statisticians use the formula for the CI in (10) above regardless of the sample size.
3. However if X is *not* normal and n < 30, the formulas for calculating the confidence interval can be complex and are beyond the scope of this course.

***Example***

Imagine once again that we have a sample of babies whose weights X are known to be normally distributed. The mean and the standard deviation are not known. However, we do have a sample of *n* = 100 babies; in that sample, the sample mean is = 8 lbs, and the sample standard deviation *s* = 2 lbs.

Q: Find the 95% CI on the population mean .

A: Since *n* = 100 > 30, we can say that *s* is a good point estimate of . That is, the quantity has a standard normal distribution (i.e., *z*-distribution). So the 95% CI can be calculated as . That is, we can be 95% *confident* that the true value of the population mean is between 7.608 and 8.392. However, we CANNOT say that the probability that is between 7.608 and 8.392 is 95%.

We could also use the *t*-distribution to calculate the formula for . That way, the 95% CI can be calculated as .

In Excel, we compute as *= tinv(0.05,99)* = 1.984. The same value can be computed in R by writing *abs(qt(0.025,99))*. We then plug this value into the formula: , which is only a tiny bit wider than the 95% CI computed using the *z*-distribution.

Now let’s imagine that instead of having a sample of *n* = 100 babies, we have a sample of *n* = 25 babies. Since 25 < 30, we can no longer use the formula for calculating the CI; instead, we must compute the CI by relying on the t-distribution formula, . In Excel, can be calculated as *= tinv(0.05,24)* = 2.064. In R, we can calculate this as *abs(qt(0.025,24))*. Plugging in this value, Note that this confidence interval is *wider* than the one where *n* = 100, and in general, as *n* increases, the confidence interval on becomes narrower.