Using Excel to Construct Confidence Intervals

This handout explains how to construct confidence intervals in Excel for the following cases:

1. Confidence Intervals for the **mean** of a population when the data is numerical and the standard deviation is *known*;
2. Confidence Intervals for the **mean** of a population when the data is numerical and the standard deviation is *unknown*;
3. Confidence Intervals for the **proportion** in a population when the data is categorical.

The terms used in this handout are as follows. A 99% Confidence interval for a parameter of the population (say, the **mean**) means that in 99% of the cases (99% of the samples, if you were to repeat the sampling process many times) the true parameter (in our case, the mean) will fall inside this interval. Remember, it's *"wrong"* to say that "there's a 99% probability that the interval contains the true parameter" ;-).

A confidence interval always has a *confidence level*, for example 99%, 95%, etc. We will also need a number for α, which can be interpreted as the *"unconfidence level",* i.e., α= 100-confidence level. So, for a 95% confidence level, we have α=5% or 0.05. In the remainder, *n* stands for the sample size. In general, a confidence interval *always* takes on the following form:

Point estimate ± margin of error

1. Confidence Interval for the Mean of a Population

In this case the point estimate (i.e., the best *"guess"* for the mean of the population) is the sample mean, written as $\overline{X}$ ("X bar"). The confidence interval is then:

$$\overline{X} \pm Z_{\alpha/2} \sigma / \sqrt{n}$$

So, the margin of error is $Z_{\alpha/2}\sigma/\sqrt{n}$, and the $Z_{\alpha/2}$ stands for the value of the standard normal distribution (the one with mean 0 and standard deviation 1) that has a probability of $\alpha/2$ on the right (this is called a "tail probability" of $\alpha/2$, drawing a little picture of the normal distribution will make this nomenclature clear). Since we assumed that the population standard deviation σ is known, and we should know what the size was of the sample we took (otherwise, we're in big trouble), the confidence interval can readily be computed. The easiest way to accomplish this in Excel is to use the following formula to compute the **margin of error**:

=NORMINV(α/2,0,σ/SQRT(*n*))

which will then give you a negative number [1].

<u>Example</u>

Suppose that our sample of size 25 yields $\overline{X}$=7.3, and we know that σ=2 in the population. Then, a 90% confidence interval has α=0.10, and α/2=0.05. The Excel formula becomes:

=NORMINV(0.05,0,2/SQRT(25))

and Excel returns the value: -0.657941, so our confidence interval becomes:

$$7.3 \pm 0.657941$$

which is the interval [6.642,7.958]. So, this interval is interpreted as the interval such that in 90% of the samples (if we were to take a lot of samples), the true mean (written as μ) lies within in this interval, and remember it's *wrong* to think of this as "there's a 90% probability that the true mean is between 6.642 and 7.958" ;-).

[1] You can safely ignore the sign.

2. Confidence Interval for the Mean of a Population when σ is unknown

Above we assumed that we didn't know the true mean of the population, but we did know its *standard deviation*. This assumption is only valid in *very rare* instances (read "never"). However, we could get an estimate for σ by using the standard deviation computed from the sample, denoted by *s* (use the EXCEL STDEV() formula). Since there's more "uncertainty" in this case (*s* will probably not be exactly equal to the *true* population standard deviation σ), the result will be that our confidence interval will become wider than in the first case. This is done by using values from the *t*-distribution rather than the (standard) normal distribution when computing the margin of error. It should be intuitive that the uncertainty about σ is reduced when we take a larger sample, i.e., a larger sample will probably give us a better estimate for σ than a smaller sample. So, the value for *t* will depend on how big a sample we took. For the same confidence, we will get smaller values for *t* when using a big sample as compared to a smaller sample. This where the *degrees of freedom* (written as d.f.) come in. At this point, we do not have to worry too much about the precise meaning of the degrees of freedom other than that for confidence intervals, the degrees of freedom is always one less than the sample size, or:

$$\text{d.f.} = n - 1$$

The mathematical formula for the confidence interval now becomes:

$$\overline{X} \pm t_{\alpha/2}(n-1) \; {}^{s}\!/\!_{\sqrt{n}}$$

The term $t_{\alpha/2}(n-1)$ is called the "*t*-value with a tail probability of α/2 and (*n*-1) degrees of freedom". Its value can be computed using Excel with the following formula (Note that Excel gives a two-tailed value for *t*, i.e., we do not divide α in half):

=TINV(α,(*n*-1))

To compute the margin of error, we will need to multiply the above quantity for the *t* value by *s* and divide by the square root of *n* (the sample size). Or, the complete Excel formula to compute the margin of error becomes:

=TINV(α,(*n*-1))*s/sqrt(n)

with: α: the "unconfidence level"
n: the sample size
s: the sample standard deviation

<u>Example</u>

Let's continue our example from above, so $\overline{X}$=7.3, *n*=25, and we computed a sample standard deviation *s*=1.98 from our sample [2]. The confidence level was set at 90%, so α=0.10 and α/2=0.05. In the first step, we compute the *t* value:

=TINV(0.1,24)

and Excel returns the value 1.71. Multiplying by *s* (1.98) and dividing by the square root of *n* gives us the margin of error of 0.678. Compare this value with the margin of error of 0.657941 computed under the assumption that σ was known. Also note that the sample standard deviation was *smaller* than the true standard deviation, but we still ended up with a bigger margin of error (and hence wider confidence interval), because the *t* value corrected for this approximation used. Our 90% confidence interval now becomes [6.62, 7.98].

[2] Remember, the Excel function STDEV computes this value for you.

3. Confidence Interval for the Proportion

Since we know that the *sample proportion* (written as p_s) is normally distributed around the true population proportion *p* with a standard deviation of $\sqrt{\frac{p(1-p)}{n}}$ (*n* being the sample size), the mathematical formula for a 100(1-α) confidence interval becomes:

$$p_s \pm Z_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}}$$

So, the margin of error is $Z_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}}$ with $Z_{\alpha/2}$ the value of the standard normal distribution that has a probability of $\alpha/2$ (a "tail probability". See 1.) on the right. The following Excel formula computes this margin of error:

= NORMINV(α/2,0,SQRT(*Ps*(1-Ps)/n*))

where *Ps* stands for the *sample proportion,* i.e., the one computed from sample data.

Example

Suppose that a sample of 200 registered voters was taken. 85 of those responded that they would vote to re-elect the present president. We want to construct a 95% confidence interval for the true proportion of registered voters who will vote for the current president. For a 95% confidence interval, α=0.05, and α/2=0.025. The value for Ps is $85/200$ =0.425, so the margin of error is given by:

= NORMINV(0.025,0,SQRT(0.425*(1-0.425)/200))

which yields -0.06851 or about 6.85%. So, this poll would be reported in the newspapers as "in a recent […] poll, only 42.5% of the voters would re-elect the current president. The poll had a margin of error of ±6.85%." The actual confidence interval then becomes [35.65%, 49.35%], so we can be fairly confident (95% confident) that less than half the voters will vote for the present president. As usual, it would be *wrong* to say that “there’s a 95% probability that the number of voters who will vote for the president is less than half.” ;-)