Stat 321 - Day 39
Confidence Intervals Based on the *t*-Distribution

You have learned that a confidence interval for a population mean μ has the form: $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$. One huge problem with this expression is that if the population mean μ is unknown and needs to be estimated, it is not likely that the population standard deviation σ would be known. A natural solution is to simply replace σ with its sample counterpart *s*: $\bar{x} \pm z_{\alpha/2}s/\sqrt{n}$. But does this expression produce an interval that will succeed in capturing the population mean μ 100(1-α)% of the time?

Example: Chest Measurements of Militiamen (cont.)
The file `militia.mtw` contains chest measurements (in inches) for 5738 Scottish militiamen in the early 19th century. We will treat these observations as our population, take samples of size *n*=5 from it, and form 95% confidence intervals for the population mean.

(a) Examine a histogram of the population values, and describe the distribution. Then use the `describe` command to compute the population mean and standard deviation. Record them below with appropriate symbols.

(b) Create the following macro and save it onto the desktop or onto your disk:

```
sample 5 c1 c2
let c3(k1)=mean(c2)
let c4(k1)=std(c2)
let k1=k1+1
```

In the session window, initialize `k1` (`let k1=1`) and execute the macro 1000 times. For each sample, the sample mean is stored in C3 and the sample standard deviation is stored in C4. You should name C3 'xbar' and C4 's'.

(c) Now create the bounds for the $\bar{x} \pm z_{\alpha/2}\, s/\sqrt{n}$ expression from each of your 1000 samples, and count how many of these (alleged) 95% confidence intervals succeed in capturing the population mean μ:

```
MTB> let c5=c3-1.96*c4/sqrt(5)
MTB> let c6=c3+1.96*c4/sqrt(5)
MTB> let c7=(c5<39.849 & c6>39.849)
MTB> tally c7
```

What percentage of your 1000 intervals succeed in capturing the population mean? Is it close to 95%?

You should have found that this procedure does not produce intervals that succeed in capturing μ 95% of the time, so we need to adjust the procedure. The key is to study and compare the probability distributions of $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ and $\frac{\bar{x}-\mu}{s/\sqrt{n}}$.

(d) When the population standard deviation σ is known, we can standardize by dividing by σ: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. Do this for each sample:

```
MTB> let c8=(c3-39.832)/(2.05/sqrt(5))
```

Produce a histogram or dotplot to display this distribution. Construct a normal probability plot of the distribution (Graph > Probability Plot). What you should observe is that this distribution is approximately a standard normal distribution. Name C8 'zs'.

(e) When the population standard deviation σ is unknown, we standardize by dividing by the sample standard deviation *s*: $\frac{\bar{x}-\mu}{s/\sqrt{n}}$. Do this for each sample:

```
MTB> let c9=(c3-39.832)/(c4/sqrt(5))
```

Compare the dotplots and numerical summaries of the two distributions.

```
MTB> dotplot c8 c9;                  MTB> describe c8 c9
SUBC> same.
```

Also construct a normal probability plot of the distribution in C9. (Note: Enter both C8 and C9.) Comment on how the distributions differ and on how the distribution in C9 deviates from the standard normal distribution.

In fact, $(\bar{x}-\mu)/(s/\sqrt{n})$ has a *Student's t* distribution with *n-1 degrees of freedom* **if** the population of observations follows a normal distribution. We get a different *t* distribution for each value of degrees of freedom. As the degrees of freedom increase, the *t* distribution approaches the standard normal distribution (see Figure 7.6 in your text). Critical values for the *t* distribution are recorded in Table A.5 (see the inside back cover of your book).

A 100(1-α)% confidence interval for μ is therefore given by: $\bar{x} \pm t_{\alpha/2,n-1}\, s/\sqrt{n}$. This procedure is appropriate when one has a random sample and when *either* the population has a normal distribution or the sample size is large ($n \geq 30$ as a rule-of-thumb).

(f) With a sample size of 5, what are the degrees of freedom? What is the critical value for a 95% confidence interval? How does this *t* critical value compare to the *z* critical value?

(g) Mimic what you did in columns 5-7 to examine the "coverage rate" of these (alleged) 95% confidence intervals by changing the critical value. Roughly what percentage of these *t*-intervals succeed in capturing μ? Is this closer to the claimed 95%?

Example: Sleeping Times
You will be asked to report how long you slept last night (in hours).

(h) Examine visual displays of this distribution, and comment on its key features.

(i) Use Minitab to construct a 90% confidence interval for the mean sleeping time last night among all Cal Poly students. [*Hints*: Use `Stat > Basic Statistics > One Sample t`. Double click on the column containing the data to select it as the variable, and use the Options button to set the confidence level (click OK). Click Graphs and select a dotplot or histogram of the data. Click OK in both windows.] Report the interval, and write a sentence or two interpreting it.

(j) Does this interval suggest that the mean sleeping time last night among all Cal Poly students was less than eight hours? Less than seven hours? Explain.

(k) Comment on whether this *t*-procedure is valid here. [*Hints*: First ask whether the sample is a random one from the population. Then ask whether the sample size is large or whether the data appear to be normally distributed.]

(l) Count how many and what proportion of the sample sleeping times fall within the interval. [*Hints*: Create an indicator variable using the `let` command to determine whether each value falls above the lower bound and below the upper bound (see above). Then use `tally`.] Is this close to 90%? Should it be? Explain.

This last question should remind you that the confidence interval estimates the value of the population *mean*. If we repeatedly take random samples and form a confidence interval like this, then in the long run 90% of those intervals would contain the actual value of the population mean. The interval does not aim to estimate the value of an *individual* observation, however.

Example: Robustness of *t*-intervals

The *t*-interval procedures that you have learned depend on the population having a normal distribution. Is the *t*-procedure still valid when this condition is not satisfied?

(m) Open the Java applet called "t-intervals for different population shapes." Start with the distribution listed as "normal" with μ=5, σ=10, and a sample size of n=5. Leave the confidence level at 95%, and change the number of intervals to 100. Click on "Sample," and keep clicking on "Sample" until you have generated a total of 1000 intervals. How many and what percentage of these intervals succeed in capturing the population mean (look at the "running total")? Is this close to 95%?

(n) Click on "Reset," and change the distribution shape to "Uniform." Leave a=5 and b=10. Also leave the sample size at n=5, confidence level at 95%, and number of intervals at 100. Click on "Sample" until you have generated 1000 intervals. How many and what percentage of these intervals succeed in capturing the population mean? Is this close to 95%? Would you say that the *t*-interval does fairly well even when the population shape is uniform and the sample size is n=5?

(o) Change the sample size to n=25, and repeat (b).

(p) Change the population shape to "exponential," and change the sample size back to n=5. Repeat (b) and comment on how well (or poorly) the *t*-procedure does in this situation.

(q) Change the sample size to n=30 and repeat (d).

A statistical procedure (such as the *t*-interval) is said to be *robust* if it performs well even when its technical conditions are not satisfied.

(r) Summarize what your analysis reveals about how robust the *t*-interval procedure is. Explain your answer.