Confidence Intervals

I. What are confidence intervals?

- simply, an interval for which we have a certain confidence.

  - for example, we are 90% certain that an interval contains the true value of something we're interested in.

  - a more specific example:

    - we are 90% certain that the interval 5'6" to 6'3" contains the true mean height for men (numbers made up).

    - this tells us that we can be reasonably confident that the true average height for men is somewhere between these two numbers.

  - usually we want confidence intervals for means, but we can also construct them for variances, medians, and other quantities. Some of these can be quite useful.

- a neat example from the text:

  - "an invisible man walking a dog"

    - we might not really be interested in the dog (well, some might be), but rather in the man.

    - we can "see" the dog - this is our y-bar, or sample mean.

    - we can't see our man - this is our population mean.

    - But, we know that the dog spends most of his time near the man - in fact, the further away from the man the dog gets, the less time he spends. If we draw a curve for amount of time spent, it'll be a normal curve, peaking where the man is [illustrate].

II. Some more properties of the normal curve.

- We already discussed "reverse lookup", and one exam problem asked about the 90th percentile for litter sizes in lions.

- Since we're interested in numbers like 90%, 95%, 99%, etc., we need to look these up in our normal tables. For instance,

  - $\Pr(Z < z) = .90 \Rightarrow z = 1.28$
  - $\Pr(Z < z) = .95 \Rightarrow z = 1.65$
  - $\Pr(Z < z) = .99 \Rightarrow z = 2.33$

- But we're interested in an "interval", so we really want:

  - $\Pr(-z < Z < z) = .90 \Rightarrow -z = -1.65$ and $z = 1.65$

[note: we look up the value for Pr($Z < z$) = .95, and then take advantage of symmetry. If, for example, we want 90% of the area in the middle of our curve, that means each end of our curve has 5% that we "don't want". So we look up the value for 95%, and then just add a negative sign for the lower part of our curve [illustrate].]

- Pr($-z < Z < z$)= .95 => $-z$ = -1.96 and z = 1.96

- Pr($-z < Z < z$)= .99 => $-z$ = -2.58 and z = 2.58

- The values 1.64, 1.96, and 2.58 will come up a lot! (you might want to memorize them!)

III. Constructing confidence intervals.

- So where are we?

- Well, suppose we took a sample and calculated $\bar{y}$. What we want to figure out is the probability that a certain interval around $\bar{y}$ includes μ. Or, putting it another way, we'd like something like:

- Pr$\{y_1 < \bar{Y} < y_2\}$ = .90, where y_1 and y_2 are such that they would have a 90% chance of including the true value for μ.

- But we know how to do this! (It's just not obvious yet). Let's start with the following:

Pr $\{-1.96 < Z < 1.96\}$ = 0.95

- we know this (just see above). Now we also know that the following has a standard normal distribution:

$$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

- now we'll substitute for Z in the first equation and isolate μ:

$$Pr\left\{-1.96 < \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} < 1.96\right\} = 0.95$$

$$= Pr\left\{\frac{-1.96 \times \sigma}{\sqrt{n}} < \bar{Y} - \mu < \frac{1.96 \times \sigma}{\sqrt{n}}\right\} = 0.95 \qquad \textit{multiplying by } \frac{\sigma}{\sqrt{n}}$$

$$= Pr\left\{-\bar{Y} - \frac{-1.96 \times \sigma}{\sqrt{n}} < -\mu < -\bar{Y} + \frac{1.96 \times \sigma}{\sqrt{n}}\right\} = 0.95 \qquad \textit{subtracting } \bar{Y}$$

$$= Pr\left\{\bar{Y} - \frac{-1.96 \times \sigma}{\sqrt{n}} < \mu < \bar{Y} + \frac{1.96 \times \sigma}{\sqrt{n}}\right\} = 0.95 \qquad \textit{see note in text below}$$

- comment: in the last step we multiplied by -1, therefore reversed the inequalities; then we re-arranged things within the probability symbol (so it looks a little like we never reversed the inequalities).

- This finally yields:

$$\bar{y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

as an interval that will contain μ for 95% of our samples.

- Problem: we DO NOT know σ. Well, substitute s for σ!

- Unfortunately, this means that our Z is no longer normally distributed; in other words, we can't use our normal tables.

- Why? Here's a partial explanation:

Z has a normal distribution and is equal to the following

$$Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

Everything in this equation (except $\bar{Y}$) is a constant, and $\bar{Y}$ has a normal distribution, so Z has a normal distribution.

But if we substitute s for σ, we have:

$$Z = \frac{\bar{Y} - \mu}{s / \sqrt{n}}$$

"s" is NOT a constant, it's a random variable (like $\bar{Y}$), and has it's own distribution. And it's not normally distributed.

This means Z does not have a normal distribution.

Without delving deep into theory, the above quantity has a t-distribution:

$$t = \frac{\bar{Y} - \mu}{s / \sqrt{n}}$$

- the *t*-distribution:

The *t*-distribution looks a lot like the normal, but generally has fatter tails, and an additional parameter, the greek "nu"or ν (NOT μ). Nu represents the "degrees of freedom". [Incidentally, as ν --> infinity, the *t*-distribution --> normal.]

- So what about this *t*-distribution?? see figure 6.7 on p. 191 **[6.7, p. 187]** *{6.3.2, p. 178}*. [illustrate].

- historical note: first derived by W.S. Gosset, while working for the Guiness brewery. Guiness did not allow him to use his own name in publishing, so he published under the pseudonym "Student".

- see table 4 in your book.

- this is arranged "backwards" from your normal tables. In other words, the numbers "in" the normal table are now along the top, and the numbers from the margins of the normal table are now inside the table.

- Why? because with the t-distribution we're hardly every interested in anything except the values for 90%, 95%, 99%, etc.

- also, note that the values are for the "upper tail", NOT for the area below the tail (opposite to how your normal table is tabulated).

- We should probably fix the *t*-tables to make them less confusing (though the reason we're doing this may not be obvious yet):

- To the bottom of your t-tables, add the following numbers (one for each column - note that each number is exactly double the number on the top):

.40 .20 .10 .08 .06 .05 .04 .02 .01 .001

Then add, "TWO TAILED PROBABILITY" to the bottom.

- **CAUTION**: the description in the text doesn't match the table (an error in the text that wasn't fixed in the 3rd edition).

IV. Now we're finally ready to construct confidence intervals.

- for example, if we want a 95 % CI, we compute the intervals as:

$$\bar{y} - t_{95\%,\nu} SE_{\bar{y}} \quad and \quad \bar{y} + t_{95\%,\nu} SE_{\bar{y}}$$
$$or, briefly,$$
$$\bar{y} \pm t_{95\%,\nu} \frac{s}{\sqrt{n}}$$

where the critical *t*-value is from table 4, with *n*-1 degrees of freedom (remember, *n* = sample size).

Concerning subscripts: *t* above has a subscript of 95%. This indicates that we want the value of t which puts 2.5 % of the area into each tail. *t* requires another subscript indicating the degrees of freedom. Remember, there are an infinite number of *t*-distributions, so knowing the degrees of freedom is important.

Another comment on subscripts: your text does things a little differently here. The way it's presented here in the notes is a bit easier to understand

(Your book would use $t_{.025,\nu}$ instead of $t_{95\%,\nu}$; this would indicate that you want 0.025 of the probability in each tail)

- here is a specific example, exercise 6.9, p. 198 **[6.10, p. 194]** *{6.3.3, p. 185}* ***{6.3.3, p. 191}***. We are given that the sample mean is 31.7, and the sample standard deviation is 8.7, and the sample size (*n*) is 5:

- (b) construct a 90% confidence interval.

$$\bar{y} \pm t_{4,90\%} \frac{s}{\sqrt{n}} = 31.7 \pm t_{4,90\%} \frac{8.7}{\sqrt{5}} = 31.7 \pm (2.132) 3.891 = 31.7 \pm 8.295$$

$$which \quad gives \quad us:$$

$$(23.4,\ 40.0)$$

- (c) now construct a 95% confidence interval (same problem):

follow the same procedure as in (b), but now our critical value for t becomes 2.776 ($t_{90\%,4} = 2.776$)), and we get:

(20.9, 42.5)

- see also examples 6.6 and 6.7 **[6.6 & 6.7]** *{6.3.1 and 6.3.2 (different in 4th edition, but the same type of examples)}* ***{6.3.1 and 6.3.2 (different in 5th edition, but the same type of examples)}*** in your text (don't pay attention to the graphs - we'll get to that stuff soon).

IV. Comments on confidence intervals.

1) note that our confidence interval gets bigger if we want to be more certain. Where would the ends be if we wanted to be 100% certain??

2) Exactly what does the CI tell us? It tells us how confident we are that the limits of our CI contains the true value of μ. Or, another way of looking at it, if we construct 100 CI's, and construct 95% CI's for each of these, 95 of our CI's (on average) will contain the true value of μ. You text has a nice picture on p. 195 **[p. 191]** *{p.182}* ***{p. 188}*** [go through this]

3) What does the CI *NOT* tell us?

$\Pr(20.9 < \mu < 42.5) = .95$

This is absolutely WRONG. μ is a constant, and so are obviously the endpoints of our interval. The statement above is then either correct or incorrect, never anything in between. The probability is either 0 or 1.

For instance, suppose we know that $\mu = 43.1$. What happens to the above statement??? [Stress -> μ is a constant, not a random variable (Y-bar, on the other hand, is a random variable)].

4) Go through example 6.9, p. 196 **[6.9, p. 192]** *{6.3.4, p. 183}* ***{6.3.4, p. 189}***:

Bone mineral density. 94 women took medication to prevent bone mineral loss (hormone replacement therapy (controversial these days)). After the study, the mean bone mineral density for this sample was .878 g/cm^3 (the book erroneously uses "^{2}" - this is not fixed in any of the editions!). The standard deviation is .126 gm/cm^3, and therefore the standard error is .013 gm/cm^3. We use *t* with 80 df (the table does not give 93, so we use the next

lower row) for 95%.

Constructing a CI, gives us .879 ± 1.990 (.013)

and this gives a CI of (.852, .904)

What does this mean? We are 95% confident that the average hip bone mineral density for women aged 45 to 64 taking this medication is between 0.852 gm/cm^3 and 0.904 gm/cm^3.

Be careful on how you interpret CI's

5) Section 6.4 (all editions) in book discusses how to get at sample size, but it's not a good method because you need to guess at the Standard Deviation (and sometimes you just don't have a clue).

- the basic idea is that you want to have your SE be as small as practically possible (after all, a smaller SE => a smaller CI)

- so you make a guess for the SD, and remember that

$$SE = \frac{SD}{\sqrt{n}}$$

- then plug in your numbers for SD, the SE you want, and solve for n:

$$n = \left(\frac{SD}{SE}\right)^2$$