9. Confidence Intervals

In chapter 8, we discussed some of the properties of the sample mean, and decided that it is a "good" estimate of the population mean. But, we should not be satisfied with just a point estimate for the population mean (or other parameter). Just reporting the sample mean (one number) as the estimate of the population mean doesn't really tell us much since we know that if we were to obtain a different sample, the mean of this new sample would be different. So, we need to report some information on how variable the sample mean is. A confidence interval can be used for this purpose. Rather than a point estimate, it is an interval of values that contains the parameter of interest (e.g. population mean) with some level of "confidence."

9.1. Two-Sided CIs

E.g. Sea lion blood toxicity, continued.

- Let X_i be the RV for the heavy metal concentration in the ith randomly measured sea lion (μg/l) out of n=100 sea lions from a population with mean, μ (unknown), and variance, $\sigma^2 \approx s^2 = 1.5^2$ (assume s^2 is a good estimate of σ^2 for "large" n)
  - i.e., $X_i \sim (\mu, \sigma^2)$
- From the CLT, we know $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$, at least approximately.
- Let $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

- $P(-1.96 \le Z \le 1.96) = 0.95$
- Draw a picture or two

- In terms of the sample mean we have

$$P(-1.96 \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96) = 0.95$$

- After doing some algebra (see page 215 POB), we can write

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

- Thus, before we sample (i.e., before we actually calculate a value for the sample mean), we can say that the population mean, μ, has a 0.95 probability of being in the (random) interval

$$[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}]$$

- After we sample (i.e., after we calculate the sample mean) the (fixed) interval $[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ either contains μ or not.

- Because the interval is fixed, we do not speak of probability, but conventionally speak of "confidence". Thus, for a calculated interval we say "we are 95% confident" that the true population mean falls in the reported interval (note that 95 can be replaced by other values associated with different z-scores: typical values are 80, 90, 95, 99 associated with z-scores of 1.28, 1.65, 1.96, and 2.58, respectively)
- For the sea lion population, assume we calculated a sample mean of $\bar{x} = 6.25$. Calculate a 95% CI:

- Interpretation:

Generally, for a (1-α)100% (2-sided) CI we calculate the interval, $[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$, where $z_{\alpha/2}$ is the z-score that "cuts off" $\alpha/2$ probability in the upper tail of the N(0,1) distribution. Again, since the interval is fixed once we calculate it, we don't speak of probability of the interval containing the population mean, μ, but of "confidence". The typical mantra is "we are (1-α)100% confident that the true population mean falls in the interval $[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$." The interpretation is that, ***IF*** we obtained many sample of the same size, n, and, for each of these samples we calculated a (1-α)100% CI, then approximately (1-α)100% of the intervals would contain the true population mean, μ. Note that (1-α) is often called the confidence coefficient and (1-α)100% the confidence level (although many people casually use "coefficient" and "level" interchangeably to refer to the decimal or percent)

Draw some pictures

Check out this CI simulator:

http://www.stat.sc.edu/%7Ewest/javahtml/ConfidenceInterval.html

E.g. Sea lions again. (n=100, $\bar{x} = 6.25$ μg/l, $\sigma \approx s = 1.5$ μg/l, assuming n=100 is "large") Calculate confidence intervals with various levels of confidence:

A Diversion: Sample Size Determination (from chapter 8)

How large should we choose the sample size, n, to estimate the population mean, μ, within ±d with (1-α)100% confidence? That is, how do we choose n so that, before the sample is obtained, we have

$$P(\bar{X} - d \le \mu \le \bar{X} + d) = (1 - \alpha)$$

or

$$P(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 0.95?$$

- Choose your desired confidence level (equivalently, choose α)

- Choose your desired level of precision, d
- Need an estimate of σ from pilot study or from other similar studies.
- Then solve for sample size, n

From above 2 probability statements we see that

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = d$$

so that

$$n = \left(z_{\frac{\alpha}{2}} \frac{\sigma}{d} \right)^2$$

Back to CIs:

9.2. One-Sided CIs

In the sea lion example, it may be sensible to consider only how large the levels of heavy metal concentrations since these levels are, presumably, the most detrimental to sea lion health. Thus, you may want to estimate a one-sided CI, giving an upper bound below which the true population mean concentration can be stated to lie with some level of confidence. (You may also calculate a lower bound for the population mean, but an upper bound seems more interesting here.)

E.g. Sea lions again (n=100, $\bar{x} = 6.25$ µg/l, $\sigma \approx s = 1.5$ µg/l, assuming n=100 is "large" and confidence level of 95%)

- Calculate a (1-α)100% upper confidence bound (lower CI) for the true population mean heavy metal concentration.
- $P(\mu \leq \overline{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}) = (1-\alpha)$ leads to an upper bound of

$$\overline{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}} =$$

and corresponding interval of

Interpretation (same as before):

How does a CI change with the following?:

- Confidence level
- Sample size
- Variance

9.3. Student's t Distribution

If the sample size, n, is not large, then s^2 may not be a good estimate of σ^2. In this case, we may still calculate "scores" and calculate probabilities and construct CIs, but the N(0,1) is no longer a good approximation for the distribution of

$$t_{n-1} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Notice, when we use s in the denominator, we call this a “t statistic” and it follows a Student's t distribution. Looks like a normal distribution with fatter tails. Notice that the distribution now depends on n (n-1 "degrees of freedom" or df). Note, as n "gets big", the distribution of t_{n-1} approaches that of Z (i.e., N(0,1)).

Using the same argument as for the z-based CIs (page 215, POB) we can create t-based CIs. A (1-α)100% (2-sided) CI for the population mean, μ, is

$$\left[\overline{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \overline{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right],$$

where $t_{\alpha/2,n-1}$ is the value of the t distribution with n-1 df such that it "cuts off" an upper tail probability of $\alpha/2$ (Table A.4, page A-10, POB). In practice, most CIs for the mean are t-based CIs. One-sided CIs are analogous to the z-based 1-sided CIs.

E.g. Sea lions again. Construct a 95% (2-sided) CI using n=25, s=1.5 μg/l and $\bar{x}$=6.2.

- CI:

- Interpretation (same as before):

- How does this compare to a 95% z-based CI?

Some S-Plus

Low birth weight infants (data set lowbwt) 95% CI for systolic blood pressure (sbp; mm Hg) by sex (male=1, female=0).

Statistics > Data Summaries > Summary Statistics… (etc.)

```
    ***  Summary Statistics for data in:  lowbwt ***

sex:0
                 sbp
    Mean: 46.464286
 Total N: 56.000000
Std Dev.: 11.145263
 SE Mean:  1.489348
LCL Mean: 43.479565
UCL Mean: 49.449007
-------------------------------------------------------
sex:1
                 sbp
    Mean: 47.863636
 Total N: 44.000000
Std Dev.: 11.805775
 SE Mean:  1.779788
LCL Mean: 44.274353
UCL Mean: 51.452920
```