

**Lab 10: Confidence intervals**  
**STT 421: Summer, 2004**  
**Vince Melfi**

Confidence intervals provide a way to report an estimate of a population parameter (such as a mean  $\mu$  or a proportion  $p$ ) along with some information about the estimate's precision. Although different settings lead to different formulas for computing confidence intervals, the basic interpretation is always the same. And understanding the interpretation is most important.

In this lab we will investigate confidence intervals for the population mean  $\mu$  in the most basic setting, where the population standard deviation  $\sigma$  is known. Our main focus will be understanding what we can and cannot claim based on a confidence interval.

## Computing a confidence interval

The file `u:\msu\course\stt\421\summer04\cipop.dat` contains three variables, `pop1`, `pop2`, and `pop3`. We'll treat each of these as the population of interest. Our goal is estimation of the population mean. Populations `pop1` and `pop2` are normally distributed, so the confidence interval procedures should work well for any sample size  $n$ . Population `pop3` is skewed, so it will take a large sample size  $n$  to get accurate confidence levels.

First we'll read in the data and compute 95% and a 99% confidence intervals for the means of `pop1` and `pop2` based on a sample of size  $n = 10$ .

```
data cipop;
  infile 'u:\msu\course\stt\421\summer04\cipop.dat';
  input pop1 pop2 pop3;

proc surveysselect data=cipop n=10 rep=1 out=ci1;
  id pop1;

proc surveysselect data = cipop n=10 rep=1 out=ci2;
  id pop2;

proc means data = ci1;

proc means data = ci2;

run;
```

Remember that the formula for a confidence interval for  $\mu$  when  $\sigma$  is known is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

where  $z^*$  is the appropriate percentile of a standard normal distribution. For a 95% confidence interval  $z^* = 1.96$ , since 95% of the area under a standard normal density is between

-1.96 and 1.96, while for a 99% confidence interval  $z^* = 2.576$ , because 99% of the area under a standard normal density is between -2.576 and 2.576.

For `pop1` we'll use  $\sigma = 12$ , while for `pop2` we'll use  $\sigma = 3$ .

The output of the two `proc means` statements gives you  $\bar{x}$ , so you have all the ingredients necessary to compute the confidence intervals. Please do this, and put the resulting confidence limits in Table 1.

Confidence interval	lower limit	upper limit
95%, <code>pop1</code>		
95%, <code>pop2</code>		
99%, <code>pop1</code>		
99%, <code>pop2</code>		

Table 1: Confidence intervals for the means of `pop1` and `pop2`.

### Questions

1. For a given confidence level, the interval for `pop2` is narrower than the interval for `pop1`. What component in the formula makes the interval narrower? How would you explain to someone who knows no statistics why it makes sense that the interval for `pop2` should be narrower?
2. For a given population, the 95% confidence interval is narrower than the 99% confidence interval. What component in the formula makes this true? How would you explain to someone who knows no statistics why it makes sense that the 95% interval should be narrower?
3. Without knowing the population mean  $\mu$ , do you know whether your 99% interval contains this value? How would you explain, in nontechnical terms, the meaning of the confidence level 99%?

## Repeating the procedure 50 times

Now we'll repeat the confidence interval procedure 50 times and determine how many of the intervals contain the population mean  $\mu$ .<sup>1</sup> We'll stick with `pop1` in this section.

### Questions

1. If the confidence level is 95%, about how many of the 50 confidence intervals do you expect to contain the population mean  $\mu$ ?
2. If the confidence level is 99%, about how many of the 50 confidence intervals do you expect to contain the population mean  $\mu$ ?

The program below takes 50 independent random samples of size  $n = 10$  from `pop1`, computes confidence intervals from each sample, and plots the lower limits and upper limits using "1" to denote the lower limits and "2" to denote the upper limits. In addition, it places a vertical line on the plot at the population mean value  $\mu = 53$ , so we can easily see which intervals contain the mean and which do not.

```
proc surveysselect data=cipop n=10 rep=50 out=ci1;
  id pop1;

proc univariate data = ci1 noprint;
  output out = cilmeans mean=Mean;
  var pop1;
  by replicate;

data cillimits;
  set cilmeans;
  lowlim95 = mean - 1.96 * (12/sqrt(10));
  uplim95 = mean + 1.96 * (12/sqrt(10));
  lowlim99 = mean - 2.576 * (12/sqrt(10));
  uplim99 = mean + 2.576 * (12/sqrt(10));
  obsnum = _N_;
  drop mean;

proc gplot data = cillimits;
  plot obsnum * lowlim95 = "1" obsnum * uplim95 = "2"
  /overlay href = 53;

proc gplot data = cillimits;
  plot obsnum * lowlim99 = "1" obsnum * uplim99 = "2"
  /overlay href = 53;

run;
```

---

<sup>1</sup>We can do that in our simulation because we have access to  $\mu$ . In a true statistical study, of course, we wouldn't know  $\mu$ .

## Explanation

There are a few new SAS features used in this program.

1. The **proc surveyselect** and **proc univariate** statements should be painfully familiar by now.
2. The block of statements beginning with **data ci1limits** creates a dataset that will contain the lower and upper confidence limits for 95% and 99% confidence levels. The only unfamiliar part is the line `obsnum = _N_`. SAS uses the symbol `_N_` to store the observation number, and the line creates a variable `obsnum` with these values that we'll use in the **plot** statements below.
3. The **proc gplot** statements tell SAS to plot both the lower and upper limits on the same plot (this is what `/overlay` means) and to use the symbols "1" and "2" for these points. Also, the code `href = 53` tells SAS to draw a vertical line at 53, which happens to be the true mean  $\mu$  for `pop1`.

## Questions

1. From the plot, what percentage of the 95% intervals contain the population mean  $\mu = 53$ ?
2. From the plot, what percentage of the 95% intervals contain the population mean  $\mu = 53$ ?
3. When a new sample is taken and a new interval is computed, does  $\mu$  change? What does change?

## A larger-scale simulation

If many people (independently) collected data and computed 95% confidence intervals, we'd expect about 95% of the intervals to contain the true mean  $\mu$ . We'll perform a simulation like the one above to check this, but this time with 1000 repetitions rather than 50. We'll let SAS count the number of intervals that contain the true mean rather than trying to read the number off a plot.

```
proc surveyselect data=cipop n=10 rep=1000 out=ci1;  
  id pop1;
```

```
proc univariate data = ci1 noprint;  
  output out = cilmeans mean=Mean;  
  var pop1;  
  by replicate;
```

```
data ci1limits;  
  set cilmeans;  
  lowlim95 = mean - 1.96 * (12/sqrt(10));
```

```

uplim95 = mean + 1.96 * (12/sqrt(10));
lowlim99 = mean - 2.576 * (12/sqrt(10));
uplim99 = mean + 2.576 * (12/sqrt(10));
obsnum = _N_;
drop mean;

data cilworks;
set cilimits;
if(lowlim95 < 53 AND uplim95 > 53) THEN yes95 = 1;
  ELSE yes95 = 0;
if(lowlim99 < 53 AND uplim99 > 53) THEN yes99 = 1;
  ELSE yes99 = 0;
drop lowlim95 uplim95 lowlim99 uplim99 obsnum replicate;

proc freq data = cilworks;

run;

```

## Questions

1. What proportion of the 95% confidence intervals contained the population mean?
2. What proportion of the 99% confidence intervals contained the population mean?

## A skewed population

If the population is not normally distributed, then the confidence level may be inaccurate for small sample sizes, because the distribution of the sample mean will not be well-approximated by a normal distribution. In other words, what we report as a 95% confidence interval may really have a coverage probability of only 87%.

We'll investigate this in the context of `pop3`, which is skewed and hence not normal, and we'll make the sample size very small ( $n = 3$ ). Note that the mean of this population is  $\mu = 1$  and the variance is  $\sigma = 1$ . We'll also add some other confidence levels.

```

proc surveyselect data=cipop n=3 rep=1000 out=ci3;
  id pop3;

proc univariate data = ci3 noprint;
  output out = ci3means mean=Mean;
  var pop3;
  by replicate;

data ci3limits;
set ci3means;
lowlim95 = mean - 1.96 * (1/sqrt(3));
uplim95 = mean + 1.96 * (1/sqrt(3));

```

```

lowlim98 = mean - 2.326 * (1/sqrt(3));
uplim98 = mean + 2.326 * (1/sqrt(3));
lowlim99 = mean - 2.576 * (1/sqrt(3));
uplim99 = mean + 2.576 * (1/sqrt(3));
lowlim995 = mean - 2.807 * (1/sqrt(3));
uplim995 = mean + 2.807 * (1/sqrt(3));
obsnum = _N_;
drop mean;

data ci3works;
set ci3limits;
if(lowlim95 < 1 AND uplim95 > 1) THEN yes95 = 1;
  ELSE yes95 = 0;
if(lowlim98 < 1 AND uplim98 > 1) THEN yes98 = 1;
  ELSE yes98 = 0;
if(lowlim99 < 1 AND uplim99 > 1) THEN yes99 = 1;
  ELSE yes99 = 0;
if(lowlim995 < 1 AND uplim995 > 1) THEN yes995 = 1;
  ELSE yes995 = 0;

drop lowlim95 uplim95 lowlim98 uplim98
lowlim99 uplim99 lowlim995 uplim995 obsnum replicate;

proc freq data = ci3works;

run;

```

## Questions

1. What proportion of the 95% confidence intervals contained the mean  $\mu = 1$ ?
2. What proportion of the 98% confidence intervals contained the mean  $\mu = 1$ ?
3. What proportion of the 99% confidence intervals contained the mean  $\mu = 1$ ?
4. What proportion of the 99.5% confidence intervals contained the mean  $\mu = 1$ ?
5. Comment generally on whether the confidence levels we observed were close to the stated (95%, 98%, 99%, 99.5%) confidence levels. If they differed, were the observed confidence levels smaller (this would be a worse mistake, since we'd be claiming more than the data supported) or larger?