Lab 5.3CI: Estimation

Objectives

To use sampling from a known population to illustrate confidence intervals.

Lab Procedure

I. The Population

The SAS data set SASDATA.STATOPOLIS contains information on 100,000 households. For the purposes of this lab, these 100,000 households will constitute the population.

A. Open SASDATA.STATOPOLIS in SAS/INSIGHT now (Recall that to get into SAS/INSIGHT you choose Solutions: Analysis: Interactive Data Analysis from any of the main SAS windows). You will see that there are four variables in the data set:
   - HHSIZE: household size.
   - VALUEH: the value of the house.
   - HIINCOME: household income.
   - HI.GENDER: gender of the head of the household (0=female, 1=male)

B. Do a distribution analysis on HIINCOME (by choosing Analyze: Distribution ( Y )).

Notice that the density histogram has many bars. HIINCOME takes so many different values, it is easier to model its distribution using a density curve. To see what such a curve might look like, select Curves: Kernel Density then click OK. Print or save this histogram with the density curve for your lab report.

By using some statistical trickery, we have managed to come up with a standard density curve that models the population closely. It's called a gamma distribution with parameters $\alpha = 2.3$ and $\beta = 25000$. The density curve for this gamma distribution is

$$p(y) = \begin{cases} 
(6.57 \times 10^{-11}) y^{1.3} e^{-y/25000}, & y > 0, \\
0, & y \leq 0. 
\end{cases}$$

The gamma distribution is common in probability and statistics, and probabilities involving it may be computed using the SAS macro NPROBS, which you will do in Part II of this lab. In the rest of the lab, we will assume this gamma distribution is the population distribution.

II. Selecting Samples and Obtaining Data

In this part of the lab, you will take two random samples from the population: one of size 5, and one of size 50, which you will use to estimate the population mean household income using a confidence interval.

After computing these quantities on the data you sampled, you will pool your results with those of others in the class. This pooled data will be used in this lab next term to evaluate the performance of the confidence intervals you calculated. Since this is a new lab, we have created a pooled data set (under the name SASDATA.LAB5_3CI) for you to analyze in Part III of this lab.

A. Select the samples by running the SAS macro LAB5_3CI. The samples of size 5 and 50 will be written to the SAS data sets WORK.SAMP5 and WORK.SAMP50, respectively.

---

1 The data were taken from an online program called StatVillage, and then expanded and modified for our use. The StatVillage data were taken from Canadian census databases.

2 Recall that you access macros by selecting Solutions: EIS/OLAP Application Builder: Applications: Run Private Applications.
B. Open each of the samples in SAS/INSIGHT. For the samples of size 5 and 50, compute the mean, \( \bar{y} \), and a 95% confidence interval for the population mean \( \mu = 57500 \). To do this, choose Analyze: Distribution (Y) and input H\_INCOME as the the Y variable. From the resulting analysis window, select Tables: Basic Confidence Intervals: 95%. The first row of the 95% Confidence Intervals table contains \( \bar{y} \) (under Estimate) and the confidence interval endpoints (LCL and UCL). Now evaluate whether this interval contains the true population mean \( \mu = 57500 \). Write down these four quantities for both the SAMP5 and SAMP50 data sets, and submit the results to the TA. The values for the entire class will be input to a SAS data set for use next term. Because this is a new lab, we have created a data set of 100 observations for you. You will find it in the SAS data set SASDATA.LAB5.3CI.

III. Analysis

Open the SAS data set SASDATA.LAB5.3CI in SAS/INSIGHT now (Recall that to get into SAS/INSIGHT you choose Solutions: Analysis: Interactive Data Analysis from any of the main SAS windows). The data set has the following variables:

- LCL5: The lower confidence limit from the sample of size 5.
- UCL5: The upper confidence limit from the sample of size 5.
- INCL5: 1 if the confidence interval from the sample of size 5 includes the population mean; 0 otherwise.
- LCL50: The lower confidence limit from the sample of size 50.
- UCL50: The upper confidence limit from the sample of size 50.
- INCL50: 1 if the confidence interval from the sample of size 50 includes the population mean; 0 otherwise.

Have a look at these to familiarize yourself with them.

A. Run the SAS Macro LAB5.3CI. This will produce two plots of the confidence intervals in the SASDATA.LAB5.3CI data set: one for sample size 5 and the other for sample size 50. The plots are color-coded: green indicates the population mean, \( \mu \), is contained in the interval, and red indicates it is not. The macro also computes the mean width of the confidence intervals. Print the plots and write down the values of the mean widths for submission with your lab report.3

Two issues in the performance of confidence intervals are coverage and precision.

1. Coverage refers to the proportion of intervals that contain the true parameter value. Calculate the coverage from the confidence interval plots for sample sizes 5 and 50 for submission with your lab report. Are they both close to the nominal coverage of 0.95? To each other?

2. Precision refers to interval width. Compare the mean interval widths for both sample sizes. Theory says that the width should be proportional to \( 1/\sqrt{n} \) (since the standard error of the mean is \( \sigma/\sqrt{n} \)). Is this the case here? Justify your answer.

B. Based on what you have seen in part III. A., summarize how sample size affects coverage and precision of confidence intervals.

The population distribution of H\_INCOME is nonnormal. In fact, it's pretty heavily right skewed. Sometimes this can have an adverse effect on the coverage of confidence intervals. Do you think the skewness affected the coverage of the confidence intervals you evaluated? Explain.

IV. Lab Report Checklist

In your lab report, be sure to include the following:

- Histogram of population values with density curve (I.B.).
- For the confidence intervals you compute by hand: (1) The sample size (5 or 50) (2) The sample mean, \( \bar{y} \) (3) The interval (4) Whether it contains the population mean, \( \mu \) (II.B.).
- Two plots, mean widths for confidence intervals and comparison with theoretical, and coverage (III.A.).
- Overall summary of findings (III.B.).

---

3 Recall that printing from the SAS/Graph window works differently than printing from SAS/INSIGHT. In the window that appears when you click on Print, select Use SAS/Graph drivers. For portrait mode, choose the PS driver, and for landscape, PSLL.