

SPSS Lab 4

Lab Demo: Estimation of the population mean and confidence intervals

We will work with two data sets. Students will analyze on their own a data set containing birthweights of newborns from alcoholic mothers. They will follow the same approach as in the demo section of this lab. Students have to answer all questions, including questions from the demo section and send answers in the body of their emails.

Open the data file gssft.sav, which includes data for full-time workers.

1. You must select people with college degrees to run the analysis! How? Go to: “Data>Select Cases\If condition is satisfied” and choose the criteria that will only leave college graduates (degree ≥ 3 & degree ≤ 4). Make sure the option “Filter out unselected cases” is selected.
2. Run Analyze\Descriptive statistics\Explore and select “number of hours worked last week” as the dependant variable. Check all options.
3. Describe the histogram: What seems more likely, that people will work a long week or a short week?
4. When looking at the histogram, you see that the distribution of the values is not quite normal, but you can count on the Central Limit Theorem to ensure that the sampling distribution of the means is approximately normal.
5. Look at the descriptive stats: the average work week is approx. 47 hours with a stdev of 10.6 hours.
6. Is the observed sample mean of 47 hours unlikely if the population mean for number of hours worked is 40? . A wrong approach would be to evaluate the probability via standardization from a normal distribution (Chapter 5= $\Pr(X>47)=1-\Pr(Z<47-40/10.6)$). Why? ... Because the data is not known to be normally distributed!!! ***You know that means calculated from samples from the same population vary. But we know that the “sample means” are normally distributed.*** Therefore we will use a different formula (Chapter 6, page 187:
$$Z = \frac{\bar{X} - \mu}{(\sigma/\sqrt{n})}$$
 - a. First, find the difference between the observed mean and the hypothetical population mean: $47-40=7$ hours.
 - b. Then calculate the standard error of the mean, which is the population stdev σ divided by the square root of the sample mean: $\sigma/(\sqrt{n})$. Let’s assume that you know that the ***population stdev σ*** is 11.66 and $n=568$ (this may be different depending on your choice of “graduates”). Then the $SEM=0.48$.
 - c. Next, you have to figure out the standard score for your observed mean. You do this by dividing the difference between the observed and hypothetical mean by the SEM: $z=7/0.48=14.58$. Since 99% of the cases in a normal distribution have *standardized values* between -2.6 and +2.6, you know that a standard score of 14.58 is extremely unusual.
7. The problem with the approach shown in 6 is that we “assumed” we known the population variance, but in reality we don’t know it! To test the hypothesis that a sample comes from a population with known mean ***but an unknown standard***

deviation σ , you calculate what's called a *t statistic*. The calculations are *exactly the same* as for the standard score, except that the value of the *sample stddev* is used in place of the population value and therefore we obtain: SEM estimate=0.43 and $t=47-40/0.43=16.28$. We will learn about t-tests next week.

8. Now take a look at the descriptive statistics. Can you find the 95% confidence interval? What are the lower and upper bounds? Is the assumed population mean of 40 hours a week within that interval? How does this answer compare to the answer obtained in step (6) above?

Sampling

We will do some random sampling from a larger set of values and compare the statistics we obtain from the smaller sample with the larger sample.

1. Open file birthweights.sav
2. We will study the variable birthweight: "birthwt". Let's start by deleting all rows where the birthweight is missing:
3. Remove all missing birthweight values: Data>Select Cases>If Condition is satisfied. Write or add into the white textbox: ~MISSING(birthwt)
4. Click Continue and select "Delete unselected cases"
5. Now click Data>Select Cases\Random Sample of Cases\Sample. Select option "Exactly" and type in 15 cases from the first 462. Click Continue and the select "Copy selected cases to a new dataset" and write the new dataset name: for example "b1"
6. As you can see, SPSS picked 15 rows/values at random from the larger table. Now calculate descriptive statistics for the larger data set "birthweights" and for the smaller data set "b1". (Use Analyze\Descriptives Statistics\Explore as in previous labs and this time select only the "descriptive" statistics and the "histogram" plot). Describe in your email how the mean, the standard deviation and the distributions differ between the two data sets. Describe the range of the 95% confidence interval (CI) from the large and the small data set.
7. Run the "Analyze\Descriptive Statistics\Explore" analysis again but this time select a confidence interval of 99%. Describe the range of the 99% CI from the large and the small data set. What is the difference to the 95% CI?

Lab Assignment

More on Normality and Confidence Intervals

In a previous lab you clicked all checkboxes and got many results that we did not analyze in detail.

1. Click analyze\Descriptive statistics\Explore. Select Dependant List=mwt0, birthwt, and gest. In Plots, click all checkboxes. Look at your output data.

2. Find the table with the tests of normality by Kolmogorov-Smirnov and Shapiro-Wilk statistics. The tests of normality overlay a normal curve on actual data, to assess the fit. A significant test with Sig. < 0.05 means the fit is poor. Otherwise, they fit the normal curve well. Report in the body of your email the results of your analysis of the birthweight data.
3. Locate the Stem-and-leaf plots, which use the original data values to display the distribution's shape. Describe in the body of your email how the values cluster uniformly or not in a certain range of values, then disperse gradually at the higher\lower values.
4. Repeat all the calculations done in the Demo section and answer the same questions but for the birthweight data, taking into account that the average weight for a baby at birth is 7.5 pounds or 3400 grams. Why is it necessary in this case to use the standard score using the SEM (chapter 6) instead of the standardized values (chapter 5)? What are your conclusions given that you know the newborn mothers suffer from alcoholism?