**Lab 4**

**44/49**

**Name: \_\_\_\_Udayraj Suthapalli\_\_\_\_\_\_\_\_\_\_ (Also, click on the header and enter your full name)**

**Note:** The word document of this assignment can be downloaded from CANVAS course page.

Please first read instructions and/or examples below and then **answer questions** **in the highlighted box**.

In this lab, we will practice sampling using data birth.sas7bdat saved in our course folder in SAS Studio. Also, we will construct CI and obtain hypothesis testing in SAS Studio using the same dataset.

Description of data: Babies with low birth weights (defined to be less than 2500 grams) are a concern because of their potential medical problems. Health researchers used this data to identify possible contributing factors to low birth weight and recommend strategies to reduce the number of low-birth-weight babies. There are 189 cases with 11 variables including

* ID: Identification Code
* LOW: Low Birth Weight (0: birth weight>=2500g, 1: birth weight<2500g)
* AGE: Age of the Mother in Years
* LWT: Weight in Pounds at the Last Menstrual Period
* ETH: Ethnicity Category (1, 2, 3)
* SMOKE: Smoking Status During Pregnancy (1: yes, 0: no)
* PTL: History of Premature Labor (0: none, 1: one, and so on)
* HT: History of Hypertension (1: yes, 0: no)
* UI: Presence of Uterine Irritability (1: yes, 0: no)
* FTV: Number of Physician Visits During the First Trimester (0: none, 1: one, and so on)
* BWT: Birth Weight in Grams

The data is stored in the shared course folder Lab4.

Graphical user interface, application

Description automatically generated

You can double click the data to open it and it will appear in a temporary library (\_TEMP6 in my case) under Libraries since it is a SAS data, i.e., no importing is needed for the data. Create a new library called Lab4\_youriniitals and then simply drag the birth data from your temporary library to this newly created library.

Table

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

**Part 1: Sampling**

Usually, data entered SAS have already been sampled from some larger population. However, SAS has a sampling procedure that can take random samples. When doing preliminary analysis of a very large data set, sampling using SAS saves time. Although our birth data is originally small and so sampling is not so beneficial, I will draw a simple random sample from it just for illustration purpose.

To sample from the data, we will use Select Random Sample task. Expand Data menu under Tasks and click on Select Random Sample. Select the birth data you just saved in the new library. The ‘OUTPUT DATA SET’ allows users to determine the target library and the data name for the sampled data. By default, the sample will be stored in work library. I will save to Lab4\_JD library with the name ‘birthsample10’.

Graphical user interface, text, application

Description automatically generated

Move to OPTION tab and set sample size to 10. Rows option is selected by default meaning that we will select 10 cases. If Percent of rows is selected, then you will sample 10% from the birth data. Let’s keep the default setting Rows as below. Check Specify the random seed and input 12345. Sampling is a random process so it can generate different samples. However, a same random seed will allow us to reproduce the same results. Click on Run icon.

Graphical user interface, application, Word

Description automatically generated

Change the sample size to 50 and 100, respectively, and name the output sample as birthsample50 and birthsample100, accordingly. Locate the samples in your library, Lab4\_yourinitial (Lab4\_JD in my example). Besides the original birth data, the three samples are all saved here.

Graphical user interface

Description automatically generated with medium confidence

Now, let’s investigate means and SEs of these three samples that have different sample sizes. The variables of interest are HT, UI, FTV, and BWT. Under Tasks, select Statistics and then Summary Statistics. We will analyze HT and BWT in the first sample--birthsample10.

Graphical user interface, text, application

Description automatically generated

Move to OPTION tab, deselect Minimum value and Maximum value (not our focus here), and expand Additional Statistics to select Standard error as below.

Graphical user interface, text, application

Description automatically generated

Click on Run to obtain the output as below.

Table

Description automatically generated

Repeat the procedure for birthsample50, birthsample100, and original birth data. I obtained the following tables.

Table

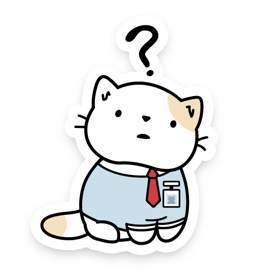
Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

According to the above 3 tables of summary statistics, answer questions in the highlighted Box 1.

|  |
| --- |
| Box 1 |
| 1. Review the data description in the beginning of this lab. What are the variable types for HT, UI, FTV, and BWT? Dichotomous, discrete, continuous, etc.? 4/4 points   **Ans: -**  HT - Dichotomous  UI - Dichotomous  FTV – Ordinal (Discrete variables)  BWT - Continuous   1. The means and standard deviations of the variables other than HT are similar for samples of sizes 50 and 100. These values are also close to that of the original birth data. Can you give a short explanation? 5/5 points   **Ans: -**The mean and SD of variables UI, FTV and BWT are very similar for samples 50 and 100 to the original values because of the size of the samples we have selected. Selecting a large sample data will create more chance of covering all the possible types of values in the original dataset.   1. What trend do you find in standard error for the variables other than HT when sample size increases from 10, 50, to 100? Please explain the reason. 5/5 points   **Ans: -**There is a decrease in the standard error for the variables of UI, FTV, BWT for the random samples 10,50,100. Standard error is the variation between the mean of the estimated random sample generated and the original known value. When we have very low random sample data there will be a lot of variation between mean and known value of original data. So, for low sampling data we have high SE and it decreases as the random sample value goes up.   1. The mean, standard deviation and standard error of HT are all 0 for the sample of size 10. What do you think is the reason? Does this make your reasoning in above questions 2 and 3 invalid? Please explain. 8/8 points   **Ans: -**We have a total of 12 1’s and 177 0’s for the HT value in the original data. When we draw a very low sample of 10 data the probability that we get more 0 is very higher which is 189/12 = 15.75 means that there will be only one row with value as “1” for every 15.75 rows and we have just drawn a sample of 10 which means the probability to have 1 in the random sample that we have drawn is 0. So, all the 10 values are 0 and we got 0’s for all mean, SD and standard error as 0.  No, the reason I have stated is directly promotional to the above 2 answers given which is the SE will have lesser value with large dataset. |

**Part 2: build a confidence interval**

We will use the original birth data (189 observations) for illustration. Suppose the researchers are interested in the proportion of normal birth weight newborns using the sample. Their research hypothesis is that the proportion of babies of normal birth weight is not 70%. Point estimate and confidence interval of the proportion will be used to answer their questions. Find One-Way Frequencies from Statistics Task. Locate your birth data and in the box of Roles, select the variable LOW.

Graphical user interface, text, application

Description automatically generated

Move to OPTION, under STATISTICS, select Asymptotic test. Input 0.7 into Null hypothesis proportion. We will let the confidence level to remain at 95%.

Graphical user interface, text, application

Description automatically generated

Click on Run icon to generate the following output.

Table

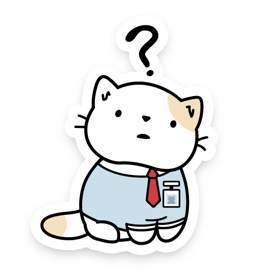
Description automatically generated

Chart, bar chart

Description automatically generated

Table

Description automatically generated

Read above output and answer questions in Box 2.

|  |
| --- |
| Box 2 |
| 1. Please write the hypotheses for the question brought up by the researchers. What are the point estimate and CI (focus on the Wald CI) for the proportion of babies of not low birth weight (BWT=0)? Interpret the CI in context and explain whether it leads us to reject the null hypothesis. 2+1+1+4+6 = 14 / 17   **Ans: -** The hypothesis of the researchers is that the proportion of babies of normal birth weight is not 70%. The Wald for the Confidence interval ranges between 0.6218 and 0.7539 when we consider a 95% confidence level which will be 5% alpha level. The z-value of the null hypothesis for the proportion 0.7 is -0.3651. The Z-value is the one that decides the null hypothesis which states we fail to prove that null hypothesis is false from the above conclusions drawn.   1. Prior to constructing the CI, can you predict whether 99% CI will be more precise or less precise than the 95% one? Why? Validate your answer with an implementation in SAS studio and paste the output here. 2+6 = 8/10 points   **Ans: -** The difference between 95% and 99% interval is that we are stating that researcher is confident about the hypothesis he stated with 95% confidence and 99% confidence. So, 99% will be more precise when compared to 95%. The 99% confidence interval will have the much wider range of values to be considered so will have more random sampling and more accuracy. After executing the 99% confidence level also the we fail to prove that null hypothesis is false.      We can clearly observe the wider values and more checking for the 99% confidence interval than the 95%. |