Medical Statistics – Lab 2

SPSS version

Welcome to lab 2 in the medical statistics course. For today's exercises, we will continue exploring the `lowbwt.sav` dataset. As a reminder, the dataset includes the following variables (see the previous lab for more details):

Variable	Abbreviation
Identification Code	ID
Low Birth Weight (0 = Birth Weight 2500g, 1 = Birth Weight < 2500g)	low
Age of the Mother in Years	age
Weight in Pounds at the Last Menstrual Period	lwt
Ethnicity (1 = Caucasian, 2 = Afro-American, 3 = Asian)	ethnicity
Smoking Status During Pregnancy (1 = Yes, 0 = No)	smoke
History of Premature Labor (0 = None, 1 = One, etc.)	ptl
History of Hypertension (1 = Yes, 0 = No)	ht
Presence of Uterine Irritability (1 = Yes, 0 = No)	urirr
Number of Physician Visits During the First Trimester (0 = None, 1 = One, 2 = Two, etc.)	pvft
Birth Weight in Grams	bwt

Point Estimates and 95% Confidence Intervals for Population Means

We will start by analyzing the variable 'birth weight in grams' (bwt), which is the main outcome of this study.

Go to `Analyze => Descriptive Statistics => Frequencies`. Select the variable 'birth weight in grams' from the list on the left. Then, uncheck 'Display frequency tables.' A frequency table shows how many individuals have a particular score, but due to the large number of different scores in this case, it would be quite overwhelming (you can keep it checked if you wish to see it).

Next, click on the 'Statistics' button. Here you can specify which statistics you want for the selected variable(s). Select mean, standard deviation, and S.E. mean. Press 'Continue'.

> Question 1
>
> Based on these summary statistics, what is the estimated mean birth weight for the population?

> Question 2
>
> Calculate the corresponding 95% confidence interval based on the normal approximation.

You can also use SPSS to calculate the 95% confidence interval. To do this, go to `Analyze => Descriptive Statistics => Explore`. Click on 'Statistics,' where you can specify the confidence interval you want to calculate. The default is set to 95%, so no changes are needed. SPSS uses the t-distribution to calculate this confidence interval, which provides a more accurate estimate when the population standard deviation is unknown and the sample size is small. Afterward, select 'Statistics' under 'Display' to ensure the output contains only the desired descriptive statistics. Press 'OK' to generate the output."

> Question 3
>
> How does the 95% confidence interval based on the t-distribution compare to the 95% confidence interval based on the normal approximation that you manually computed?

One-Sample t-Test

To determine whether the population mean birth weight differs significantly from a hypothesized value of 3000 grams, we conduct a one-sample t-test. To do this, go to `Analyze => Compare Means => One-Sample t-Test`. Select the variable 'birth weight in grams' and place it under 'Test Variables'. We want to determine if the population mean significantly differs from the threshold value of 3000 grams. Enter '3000' as the test value. Press 'OK'. Now you will see the result of the t-test in your output.

Question 4

You see that the test has 188 degrees of freedom. Why?

Question 5

Based on the results of the test, does the population mean significantly differ from 3000?

One of the assumptions underlying the one-sample t-test is that the data are normally distributed. We can check this assumption by creating a histogram. To do this. go to `Graphs => Legacy Dialoges => Histrogram...`. Select the variable 'birth weight in grams' and place it under 'Variable'. Check the 'Diplay normal curve' checkbox and press 'OK'.

Question 6

Looking at the histogram, would you say that the data are normally distributed?

Point Estimates and 95% Confidence Intervals for Population Proportions

Next, we will explore the variable 'low birth weight' (low), which is a dichotomous variable that takes a value 1 if the baby had a low birth weight (defined as a birth weight < 2500g) and a value of 0 otherwise.

We start by making a frequency table to calculate the frequency of each category of the 'low' variable:

- Go to `Analyze => Descriptive Statistics => Frequencies...`
- Select the variable 'low' and move it to 'Variable(s)'
- Press 'OK' to obtain the frequency table

Question 7

Based on these frequencies, what is the estimated proportion of low birth weight babies in the population?

Question 8

Calculate the corresponding 95% confidence interval based on the Normal approximation.

Binomial Test

Subsequently, we perform an exact binomial test to assess whether the proportion of low birth weight babies in the population differs significantly from a hypothesized value of 30%:

- Go to `Analyze => Compare Means => One-Sample Proportions....`
- Select the variable 'low' and move it to the 'Test Variable List'.
- Under 'Define Success', select 'Value(s)' and enter the number `1` to indicate that having a low birth weight baby is the event of interest.
- Click on the 'Test' button to open the dialog box, select 'Exact Binomial' as the test and enter `0.3` as the test value. Press the 'Continue' button.
- Click 'OK' to run the test.

Question 9

Does the proportion of low birth weight babies differ significantly from 30%?

Question 10

The Dutch government intends to start a campaign against drinking alcoholic beverages if over 50% of the adolescents drink alcoholic beverages regularly (at least once a week). A random sample of 200 adolescents is taken and 128 admit that they drink alcohol regularly (we assume all 200 speak the truth). Test the null hypothesis that 50% of the Dutch adolescents drink alcohol, using a significance level of 5%. Use the exact binomial test for this question.
Hint: to perform this test in SPSS, we need to create a new dataset containing the above observations. To achieve this, construct a new data file containing two variables, like this:

Alcohol	number
1	128
0	72

Then, instruct SPSS to weight the categories (Alcohol) by "number" via Data → Weight cases.

Question 11

Rather than using an exact binomial test, we can also use the normal approximation of the binomial distribution to obtain an approximate p-value for the above hypothesis test. Manually calculate this approximate p-value and compare it to the p-value obtained from the binomial test. Is the use of the normal approximation appropriate in this case?

Differences in Two-Sided P-Value Calculation Between SPSS and R

When conducting statistical tests, it is important to understand that different software packages can calculate two-sided p-values in slightly different ways, which may lead to variations in results. A key difference exists between how SPSS and base R handle this calculation:

- SPSS often calculates two-sided p-values by doubling the one-sided p-value. Specifically, SPSS determines the probability of the observed outcome in one direction (greater or less than a given value) and then multiplies this value by 2. This approach assumes that the distribution of the test statistic is symmetric under the null hypothesis. While this method is straightforward, it can be misleading if the distribution is skewed or the sample size is small, as it may not fully account for the asymmetry in the data.
- Base R (e.g., the `binom.test()` function) uses a more exact method for calculating two-sided p-values. R's approach sums the probabilities of observing outcomes that are as extreme as, or more extreme than, the observed value in both directions (both tails of the distribution). This method does not assume symmetry and provides a more accurate p-value, particularly for small samples or skewed distributions.