Lab 6: The Normal Distribution

Lab Objectives:

- Use an online simulation to visualize sampling distributions, properties of the normal distribution, and the central limit theorem.
- Practice using Z-standardization and statistical tables.
- Use SPSS to conduct one-sample t-tests and to calculate 95% confidence intervals.

Exercise 1: Visualizing the normal distribution & central limit theorem

The normal distribution is the "bell-shaped curve." It approximates the distribution of many biological variables. It has a single mode equivalent to its mean, and is symmetric around the mean. The normal distribution is fully described by two parameters: the mean and the standard deviation.

The Central Limit Theorem states that the sum or mean of a set of independent and equally distributed values will have a normal distribution, as long as enough values are added together. This means that sample means will be normally distributed if the sample size is large enough, even if the distribution of the variable itself is not normal.

Simulation of sampling distributions

Open the following link: http://www.zoology.ubc.ca/~whitlock/kingfisher/SamplingNormal.htm

This page contains an applet that lets you visualize the process of collecting samples and calculating means. The default settings use a population of fish that are on average 106 mm long, with a standard deviation of 30. These values are the true population parameters, which we would not normally know in real life. The variable "length" has a normal distribution. The default sample size is 10 fish, meaning that you are going to collect 10 fish, measure them, and determine mean fish length.

- Click on the button "Sample 1 individual". This process measures one fish, and begins preparing a histogram using this fish's measurement in the first graph below the ruler. Click "sample 1 individual again", and make sure you understand the process. Now click the button "complete sample of 10".
- 10 fish have now been sampled and measured. Click "calculate mean" to calculate the mean length of fish from this sample of n = 10.
- Now click "complete sample of 10" another time. This process imagines that you catch and measure a new sample of 10 fish. Click "calculate mean" to calculate the mean length from this second sample.
- Repeat the last step at least two more times.
- Notice that the graph lower graph shows the distribution not of individuals, but of **sample means**. The bar added in the distribution of sample means always corresponds to the mean of the sample collected in the second graph. This is what is known as a **sampling distribution**.

- Click on the button “calculate many means”.
- Click on the buttons “show population” and “show sampling distribution”. The curve illustrated in the top graph shows the distribution of the fish length variable in the population. The curve in the bottom graph shows the sampling distribution – the distribution of mean length when many different samples of n=10 are collected and mean length is determined for each sample.
- Change the sample size to n = 75. Click “complete sample of 75”, then “calculate many means”. Notice how the graphs differ from the previous example with n = 10.
- Use the slider to change the standard deviation to 40, and then to 20. Notice how the graphs change with these adjustments.
- Finally, make note of how the two curves change when you change the mean to a smaller mean value, and a larger mean value.

➢ Answer questions 1 – 2 on Sakai.

Simulation of the central limit theorem

- At the bottom left of your screen, select “topics” then central limit theorem.
- Click on the tutorial button at the bottom left. Work through the steps of the tutorial, which illustrates important concepts about the central limit theorem.

➢ Answer questions 3 – 5 on Sakai.

Exercise 2: Z-standardization

Z-standardization converts values from a normal distribution with a known mean and standard deviation into **standard normal deviates**, using the equation:

$$Z = \frac{Y - \mu}{\sigma}$$

A standard normal deviate, or Z, tells us how many standard deviations a particular value is from the mean. Z standardization can be used to determine the probability that a randomly selected individual from a normal distribution falls above or below a given value, or within a certain range. It also tells us the proportion of individuals from a normal distribution who fall within a given range. After you have calculated a standard normal deviate, you can use a statistical table for the standard normal (Z) distribution to determine the probability of sampling a value greater than or equal to a given value of Z.

SPSS can be used to compute standard normal deviates (Z scores) for a given variable. Let’s try this now.

Baby Birth Weight

Z standardization can only be used for numerical variables that have a normal distribution. Weight and height tend to be good examples of these. Other continuous numerical variables such as wing length, antenna length, growth rate, and temperature tend to also be a good fit to the normal distribution. The data in the file birthweight.sav contains information on the birth weight of 189 babies, as well as data on various demographic parameters and indicators of the mother's health.

- Open the file *birthweight.sav*
- Verify that the birthweight variable is a relatively good approximation of the normal distribution by plotting a histogram of this variable.
- Select analyze, descriptive statistics, descriptives.
- Add birth weight in g to the variable list.
- Check the box "save standardized values as variables".
- Select OK. In the output window you should see a table listing the N, minimum, maximum, mean, and standard deviation.
- Return to data view. Note that you now have a new variable listed called "ZBirthWeight".
- Select analyze, descriptive statistics, descriptives again. This time add both Birthweight and ZBirthweight to the variable list. Uncheck the box "save standardized values as variables".
- Consider the mean and standard deviation of the variable ZBirthweight. Examine the ZBirthWeight variable more closely. What do the positive and negative signs on the Zscores indicate?
- Select analyze, reports, summarize cases. Uncheck the box that says "limit cases to first 100". Add birth weight in g and Zscore birth weight to the variables list. Click OK.
- Consider a situation where you have only the data found in the case summaries table, and do not know the mean or standard deviation. How can you estimate the mean? How can you estimate the standard deviation?

➢ Answer questions 6 – 8 on Sakai.

Death Valley temperature

Answer the following question as you would in class, or on a test – that is, using the appropriate formulas and the standard normal distribution table.

The highest recorded temperature during the month of July for a given year in Death Valley, California, has an approximately normal distribution with a mean of 123.8°F and a standard deviation of 3.1°F.

A. What is the probability for a given year that the temperature never exceeds 120°F in July in Death Valley?
B. What is the probability that the temperature goes above 128°F during July in a randomly chosen year?
C. What is the probability that the highest recorded temperature will be between 128°F and 130°F?

➢ Answer questions 9 – 10 on Sakai.

Exercise 3: One-sample t-tests

If a variable Y is normally distributed in the population with a mean μ, and we have a random sample of *n* individuals, then the sample means are also normally distributed with a mean equal to μ and a standard error of σ/√n. Z-standardization can be used to calculate the probability of observing a mean of $\bar{Y}$ using the formula:

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

However, if we do not know the true standard deviation in the population (σ), we must estimate it using SE = s/√n. This leads to a related quantity called Student's *t*:

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$$

The sampling distribution for *t* is not a normal distribution, but rather a *t*-distribution. A *t*-distribution is fatter in the tails than the standard normal distribution. The sample size determines the number of degrees of freedom of the *t*-distribution (df = n – 1), and which particular version of the t-distribution we need to use. As sample size increases, *t* becomes more like Z.

The confidence interval for a mean assumes that the variable has a normal distribution in the population and that the sample is a random sample. The 95% confidence interval for the population mean is *approximately* 2 standard errors above and below the sample mean. However, it can be more precisely calculated by multiplying the critical t-value from the appropriate t-distribution by the standard error.

A one-sample t-test compares the observed sample mean with μ_0, a specific value for the population mean proposed in the null hypothesis. It calculates the test statistic *t*, and determines the probability of observing a *t*-value as extreme as, or more extreme than, the calculated *t*-value, if the null hypothesis is true, using the appropriate *t*-distribution with n – 1 degrees of freedom.

Age on the titanic

- Open the file *titanic.sav*. This file contains information on the passengers of the *Titanic*. We'll use a one-sample t-test to ask whether the mean age of passengers was significantly different from 18 years old.

- Select analyze, compare means, one sample t-test.
- Select age as the test variable.
- For the test value, enter 18. Click OK to run the test.
- The first table provides the N, mean, SD and SE.
- The second table provides the t-statistic, the degrees of freedom (df), the significance value (P-value) for a two-tailed test, the mean difference and the 95% confidence interval of the difference (note that this is **not** the 95% confidence interval of the mean).
- If there is a significant difference between the data mean and the test value (which in this case we input as 18), the significance value will be less than 0.05.
- Determine the 95% confidence interval of the mean age on the titanic by selecting Analyze, descriptive statistics, explore.
- As a refresher of what you learned in the last lab, conduct a statistical test to determine whether passenger class and survival on the Titanic are associated. Conduct a second test to determine whether gender and survival are associated.

➢ Answer questions 12 – 14 on Sakai.

Malaria & mosquitos

Malaria is spread by mosquitoes. To properly understand how the disease spreads it is essential to understand the biting behavior of mosquitoes. A study in Kenya measured the relative attractiveness of people infected with malaria, compared to the same people after they had been treated with an effective anti-malaria drug. Mosquitoes were given a choice between the target infected person and two others, and the proportion of mosquitoes going to the infected person was recorded. The same experiment was done with the same three people after the infected person was cured. The data in the "Difference" column are the changes in the proportion of mosquitoes biting the infected person as a result of the treatment. Positive numbers indicate that the target person was bit more when they had malaria.

- Open the file *malaria mosquito.sav*.
- Consider what mean value you would expect to observe for “difference” if mosquitos bit a person equally before and after malaria treatment.
- Conduct a one-sample t-test to determine if the difference observed was significantly different from this value.

➢ Answer question 15 on Sakai.

Hurricanes & soil lead

Hurricanes Katrina and Rita caused flooding of large parts of New Orleans, leaving behind large amounts of sediment. Forty-six sites were monitored for soil lead content before and after the hurricanes. While the ratio of soil lead content is not normally distributed, the log ratio has an approximately normal distribution and can be used in statistical analysis. A ratio of 1 is equivalent to a log ratio of 0. Therefore, a negative log ratio indicates a reduction in soil lead content, while a positive log ratio indicates an increase in soil lead content.

- Open the file *hurricanes.sav*.
- Determine the mean and 95% confidence interval of the log ratio.
- Consider what mean value you would expect to observe for "log ratio" if the mean soil lead content was unchanged by the hurricanes.
- Conduct a one-sample t-test to determine if the difference observed was significantly different from this value.

➢ Answer questions 16 – 17 on Sakai.

Hurricanes & blood lead

In the same study, the concentration of lead in the blood of children living in 46 areas was measured before and after the hurricanes. The ratio of blood lead concentrations has an approximately normal distribution.

- Open the file *hurricanes blood lead.sav*.
- Use the transform, compute variable function to calculate the ratio variable, which should be equal to the blood lead concentration after the hurricanes, divided by the concentration before the hurricanes.
- Determine the mean and 95% confidence interval of the ratio.
- Consider what mean value you would expect to observe for "ratio" if the concentration of blood lead was unchanged by the hurricanes.
- Conduct a one-sample t-test to determine if the observed ratio was significantly different from this value.

➢ Answer question 18 on Sakai.

Exercise 4: Additional practice

Back to Birthweight

The birthweight data in the file *birthweight.sav* were collected at Baystate Medical Center in Springfield Massachusetts during 1986. Conduct an appropriate statistical test to determine whether mean birth weight in this data set is significantly different from the current mean birth weight in Canada, as reported by Statistics Canada (http://www.statcan.gc.ca/pub/84f0210x/2009000/t015-eng.htm)

Use the appropriate function in SPSS to calculate the mean and 95% confidence interval for birth weight comparing mothers who smoked to mothers who did not smoke.

Use the appropriate function in SPSS to determine whether there is an association between low birth weight (<2500 g) and the presence of uterine irritability in the mother.

Cat body weight

The file *cats.sav* contains data on brain weight and body weight of 137 cats in lbs. Body weight is approximately normally distributed. Use SPSS to compute standard normal deviates for the body weight variable. Use these values to answer the question on Sakai.

US Cereal Nutrition

The file *UScereal General Mills.sav* contains data on the nutritional information of 22 different breakfast cereals manufactured by General Mills. A larger survey of 150 US breakfast cereals estimates the average values for certain nutritional variables to be as follows:

Variable	Average value
Calories	149 kcal
Protein	3.7 g
Fat	1.4 g
Fibre	3.8 g
Complex carbohydrates	19.9 g
Sugars	10.1 g

Assume that these continuous numerical variables are approximately normally distributed in the full "population" of breakfast cereals. Determine whether the observed mean values in the General Mills data set are significantly different from the expected mean values according to the US breakfast cereal data.

- Answer the remaining questions on Sakai.