Lab 01: Inferential Statistics and Data Exploration

**Due date:** Wednesday, Feb 5, 2025 submitted as Word document to Canvas ***Lab01***  link

This lab counts 9 % toward your total grade.

**Objectives:** In this lab, you will practice your skills in

1. Inferential statistics
2. Confidence interval
3. Statistical graph
4. Data exploration

**Format of answer:** Submit your answers as a **Word document** with **graphs** and **verbal** descriptions, properly labeled in the task sequence, with answers in red text and only relevant content included. Provide R code for each task.

# Task 1: Confidence Interval (3 pts)

A survey was conducted on a sample of 1,000 university students to determine the proportion of students who regularly use public transportation. Out of the 1,000 students surveyed, 72.4% reported that they regularly use public transportation.

1. Calculate the 95% confidence interval for the proportion of students who use public transportation. (1 pts)

x\_95=propCI(n = 1000, p = 0.724, alpha = 0.05)

1. Calculate the 99% confidence interval for the same proportion. (1 pts)

x\_99=propCI(n = 1000, p = 0.724, alpha = 0.01)

1. Interpret the results. (1 pts)

With 95% confidence, the proportion of university students who regularly use public transportation falls between 69.6% and 75.2%. More specifically, 95% of the sample their proportion of students who regularly use public transportation falls between 0.696 and 0.752.

With 99% confidence, the proportion falls between 68.8% and 76.0%. More specifically, 99% of the sample their proportion of students who regularly use public transportation falls between 0.688 and 0.76.

As expected, the 99% confidence interval is wider, indicating greater certainty but less precision compared to the 95% confidence interval.

A graph with numbers and lines

Description automatically generated

# Task 2. Statistical Graph (3 pts)

The following graph is reproduction from the [R Graphics Cookbook](https://r-graphics.org/recipe-scatter-continuous-scatter). R has extensive documentation to introduce its functions. You should be able to learn how to understand the documentation and apply the function to your study. Study the code in the [link](https://r-graphics.org/recipe-scatter-continuous-scatter) and reproduce the graph with a different data (**Boston**).

Boston data in MASS package **(??MASS::Boston**) contain variables about the housing value in suburbs of Boston.

In this task, you will work with ggplot2 to create a scatter plot with additional layers to enhance the visualization. The output can show relationship between two variables, using different colors to distinguish the third variable. **Show your R code and graphs for this calculation.**

1. Reproduce the graph below using **Boston** data in **MASS** package**.** In **ggplot** function, set up parameters as: **aes(x = rm, y = medv, colour = indus) .** (1pts)

A graph with blue dots

Description automatically generated

library(MASS)

data('Boston')

ggplot(Boston, aes(x = rm, y = medv, colour = indus)) +

geom\_point()

1. Explain the data distribution pattern based on **Boston** data. (0.5pts)

The scatter plot shows the relationship between rm(average number of rooms per dwelling) on the x-axis and medv (median value of owner-occupied homes in $1000s) on the y-axis.  
There is a positive correlation between rm and medv. As the average number of rooms per dwelling increases, the median value of owner-occupied homes also increases. We can say that houses with larger number of rooms are more expensive. Furthermore, for the same average number of rooms per dwelling, the median value of owner-occupied homes tends to increase when the proportion of non-retail business areas per town is lower.

1. Based on the scatterplot, a smoothing line is added to the plot using the **geom\_smooth()** function. Your task is to reproduce the below graph using Boston data in **MASS** package. Before you produce the graph, convert the **chas** variable in **Boston** data to factor:

**Boston$chas = as.factor(Boston$chas).**

Set up the parameters in **ggplot()**

function as follows:

**aes(x = rm, y = medv, colour = chas)**. (1 pts)

A graph with red and blue dots

Description automatically generated

Boston$chas = as.factor(Boston$chas)

hw\_sp <- ggplot(Boston, aes(x = rm, y = medv, colour = chas)) + geom\_point() + scale\_colour\_brewer(palette = "Set1")

hw\_sp + geom\_smooth()

1. Explain the data distribution pattern from **Taks2. c**. (0.5 pts)

The visualization provides an analysis of the relationship between the number of rooms (rm) and the median home value (medv), distinguishing between two groups based on whether the region borders the Charles River (chas). The general trend indicates that as the number of rooms increases, the median home value also tends to rise.

Specifically, for properties where chas = 0 (not adjacent to the river), the median home value decreases within the mid-range and higher-range of rm. Conversely, for properties where chas = 1 (adjacent to the river), the median home value generally increases, although an outlier is observed at the upper end of the rm range.

It is important to note that this visualization considers only the impact of the number of rooms on median home value. Future analysis should incorporate additional variables to provide a more comprehensive understanding of the factors influencing home prices.

# Task 3: Data Exploration (3pts)

The **MplsDemo** Demographic Data 2015 in **carData** package include the demographic data from the 2015 American Community Survey. **Show your R code and plot for this calculation.**

1. Import the **MplsDemo** data use the function **data().** (1 pts)

library(carData)

data('MplsDemo')

1. Examine the histogram and pairwise relationships between variables using **car::scatterplotMatrix().** Identify any skewed pattern visually and provide a description. Please use following variables for **car::scatterplotMatrix()**: **~population + white + black + hhIncome**. (1 pts)

scatterplotMatrix(~population + white + black + hhIncome, data = MplsDemo)

A collage of graphs

Description automatically generated

Based on the histogram analysis, the variables representing population and Black residents exhibit positive skewness, while the variable representing White residents demonstrates negative skewness. The household income (hhincome) variable appears to follow a normal distribution. Furthermore, a positive relationship between the White population and household income. Conversely, a negative relationship exists between the White and Black populations.

1. Evaluate the skewness of the variables that you visually identify as skewed from the scatterplot matrix. Use the skewness function to confirm your observations. **e1071::skewness( )**. (1 pts) ()

skewness(MplsDemo$population)

skewness(MplsDemo$white)

skewness(MplsDemo$black)

skewness(MplsDemo$hhIncome)

A screenshot of a computer code

Description automatically generated

These values are reflecting on the histograms as shown in part b. Only White is left skewed. Rest of the indicators (population, black and hhincome) are right skewed. Value closer to 0 means, closer to the symmetric shape.