Stat 123 Midterm 2

Monday, March 20, 2023

Duration: 4:30 pm to 5:20 pm + 10 minutes to upload submissions due by 5:20 pm.

- The test duration is 40 minutes + 10 minutes for uploading. If you have a CAL time accommodation, you may add it to your 40-minute time limit.
- You can access all the required materials in the "Midterm 2" dropbox on Brightspace. Please follow the path: Brightspace-> Course Tools-> Assignments -> Midterm 2.
- Download the "Midterm2-Questions" file in PDF format and the "Houses.CSV" dataset.
- The lecture notes can be found in the "Midterm 2" dropbox.

0. Open a new R Markdown file.

Note: Your worksheet should be submitted as an R Markdown file (You MUST knit it to PDF on your computer, or you can knit it to Word and then convert to a PDF). You must have uploaded Your solutions in PDF format to the Brightspace drop box named “Midterm2” by no later than 5:20 pm unless you have a CAL time accommodation.

1. You are given a dataset containing the number of hours spent by a group of students studying different subjects. The dataset is as follows:

Subject	**Hours**
Math	20
English	15
Science	10
History	5
Geography	5

(a) Create a data frame in R named "data" using the provided dataset.

(b) Calculate the percentage of total study hours for each subject, rounded to one decimal place.

(c) Create a pie chart titled "Study Hours by Subject" and assign colors to each slice: red, orange, yellow, green, and blue. Set the x-axis and y-axis scales to be between -1 and 1.The label of each slice should be the name of subjects along with the percentage of total study for each subject.

Hint: You can use the percentile() function to calculate the percentage of total study hours for each subject.

2. Use the built-in Titanic dataset. Please refer to the Titanic dataset description.

(a) If we want to know the mean of children passengers on the Titanic, what is the parameter of interest?

(b) Identify the variables in the dataset and describe their types.

(c) Create a variable in R called "totalChildren" that contains the total number of children on the Titanic (across all genders, classes, and survival outcomes).

(d) Create a variable in R called "totalSurvival" that contains the total number of survivors in our sample (children who survived the Titanic).

(e) What is the observed value of the statistic that we should use to estimate the population mean of interest (survived children on the Titanic)?

(f) What is the estimated standard error for the population mean?

(g) What is critical value for a 90% confidence interval for the population mean?

(h) What is the margin of error for our estimate?

(i) Determine a 90% confidence interval for the true value of the population mean.

3. You have been given a dataset containing information about house prices in a particular neighborhood. Please download the "Houses.csv" dataset from the Midterm 2 folder on Brightspace. The dataset includes the following variables:

Price: The price of the house in dollars

SquareFeet: The size of the house in square feet

Bedrooms: The number of bedrooms in the house

Bathrooms: The number of bathrooms in the house

YearBuilt: The year the house was built

(a) Create a new variable called "TotalRooms" that is the sum of the "Bedrooms" and "Bathrooms" variables using the for() statement.

(b) Build a linear regression model using the lm() function that predicts the price of a house based on its SquareFeet and TotalRooms.

(c) Which independent variable should be removed from the model?

(d) Why should the variable be removed?

4. Consider the built-in "esoph" dataset in R. Please refer to the "esoph" dataset description. The dataset contains data from a case-control study of esophageal cancer. Use the column named "ncontrols" in the "esoph" dataset and answer the following questions:

(a) Bootstrap 10,000 samples to find the 95th percentile and save the bootstrapped 95th percentiles to a vector called "Bootstrap."

(b) Plot the sampling distribution of the 95th percentile. Your plot should include a title and label for the x-axis.

(c) Compute a 92% confidence interval for the 95th percentile.