BANA 7031 – Probability Models

Final Project

Group:

Hardik Gupta – M13474715

Kevin Gilmore – M08331816

**Brief Description**

The dataset we selected is from the package “Lock5Data.” This dataset is open to anyone to install the package and then load “CommuteStLouis” data. This dataset examines the commute times of people in St. Louis.

The dataset consists of 500 observations and contains 5 variables:

* **City** – Factor with 1 level, everyone in study is from St. Louis
* **Distance** – distance traveled to get to destination
* **Age** – age of person in dataset
* **Time** – the time taken to reach the destination
* **Sex** – Factor with 2 levels, either “M” for male of “F” for female

A screenshot of a cell phone

Description automatically generated

The above screenshot depicts the summary of the dataset.

**Packages Required**

Following are the packages required to analyze the given datasets:

**library**(Lock5Data)  
**library**(tidyverse)  
**library**(ggplot2)  
**library**(Ecdat)

**library** (bootstrap)

**library** (stats4)

**Histograms (distribution):**

We wanted to see the distribution of mean time to reach the destination. To do this, we created a matrix with 1000 rows:

In this instance, sample () was used to sample the values in Commute$StLouis. The replace = TRUE function fills all the B multiplied by n elements in the matrix created. Furthermore, using the apply() function learned in Data Wrangling, we use this function to apply () on three arguments:

We then found the standard deviation and plotted a density chart and histogram simultaneously.

A close up of a map

Description automatically generated

The data seems to be more than less bell-shaped.

## **Methodologies:**

We have tried to incorporate the following analytical methodologies we have learned in the class:

## Empirical CDF

We selected the variables of time and distance and then created two separate Empirical CDFs. From the plots below, we can find the probability of time of commute and distance of the commute. Graphs on the left depict the ECDF plot, while the plots on the right depict the ECDF plots with confidence bands.

A screenshot of a cell phone

Description automatically generatedA screenshot of a map

Description automatically generated

Above are the time ECDFs, confidence bands are on the right.

A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generated

Above are the distance ECDFs, confidence bands are on the right.

1. Bootstrap

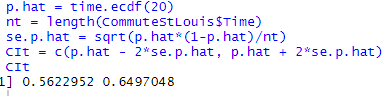
First, we have performed bootstrap confidence interval calculations on the time variable as a whole. We based our bootstrap on 1000 bootstrap replications:

A screenshot of a cell phone

Description automatically generated

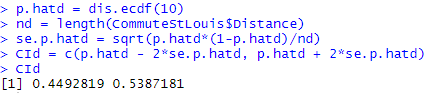
## Empirical CDF Testing and Confidence Intervals

For our project, we also wanted to estimate the probability that the time taken by a random person to travel to work is less than 20 minutes.



In the confidence interval we have values of .56 and .65

We then examined the variable distance and wanted to estimate the probability of distance commuted to destination being less than 10 miles, f(10).



In this confidence interval we have values of .45 and .54

1. Adding gender (histograms)

For the next portion of the project, we wanted to examine the mean travel time and distance for males and females individually. The first set of histograms are the times.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

After examining the two histograms, the datapoints seem to be relatively similar. The mean time for females to commute is 21.6 minutes, while males take 22.3 minutes.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

After examining the two distance histograms, the datapoints seem to be a bit more dispersed, however the majority are still in the same intervals. The mean distance for females to commute is 13.8, while males travel 14.5.

1. Hypothesis Testing and MLE

We concluded that our hypothesis at alpha = 0.05, both genders commute the same average distance. After finding our SE and using a t-test, we find a p-value of 0.4616. Since this p-value is significantly greater than the alpha = 0.05, we cannot reject the null hypothesis, which determines males and females commute equal distance on average to their destination. This is a two tailed test because the mean distance for both genders are the same.

We have taken the MLE for means of distance travelled by both males and females.

* Males = 14.496
* Females = 13.7875

1. Density Plots

We have also examined density plots for both males and females. These plots are checking density with the variable “distance.”

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

After examining the density plots, the trend is similar for distance for both males and females.

Appendix

[Combined Code \_ probability.R](Combined%20Code%20_%20probability.R)