Jairik "JJ" McCauley
Dr. Hetzler
Math385 Portfolio
May 14th, 2024

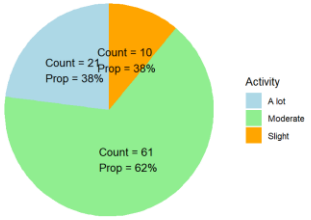
Utilizing R as a Tool for Statistical Thinking

JJ McCauley's MATH385 Portfolio

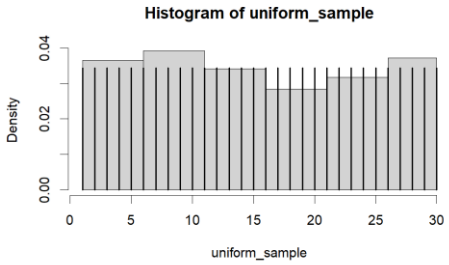
Throughout the semester, I have been given the privilege to learn numerous important functionalities of R as it relates to statistical thinking. Below, I will be diving into each skill I have developed, as well as how they appeared in the labs.

Data Visualization

One of the fundamental skills throughout this class has been visualizing data through numerous methods. In lab zero, we covered the basics of making a table to cleanly represent the data and important characteristics of it, specifically showcasing the summary of the data. Although these tables did not tell us everything we needed to know about the data, they proved to be essential steppingstones in us learning more about the data. Building off these ideas, we created pie-charts and bar charts in lab 2, proving to be an easily digestible way to examine proportions within a given population.



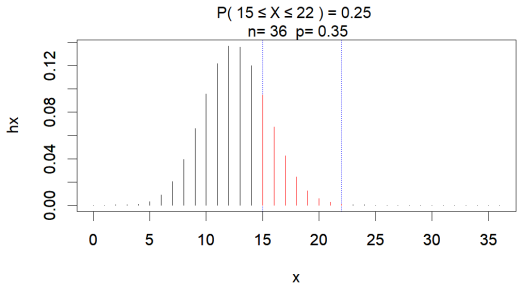
Although tables and pie charts are important for understanding basic characteristics, they are not quite able to tell us everything about a population/sample. For example, how could we examine the distribution of a population/sample? In lab 3, we solved this by utilizing histograms. These visualizations can show us the distribution of data, which is crucial for understanding the data's tendencies, shape, and variability. Histograms are additionally used throughout lab 4, lab 5, and lab 6/7, aiding us in understanding other key concepts. In addition to histograms, lab 5 also introduced QQ-plots, which are essential for understanding the normality of data. Throughout lab 5, these visualizations were primarily used to demonstrate how the distribution of samples gets closer to normal as the sample sizes get larger (Central Limit Theorem), which proved to be clear in the QQ-plots. In addition to histograms and QQ-plots, basic visualizations such as plot() or lines(x,y) (line graphs), as seen in lab 4, aided us in understanding the basic structures of datasets. This visualization also proved to be increasingly useful with histograms, as it aided in understanding the true distribution within ranges.



Probabilities

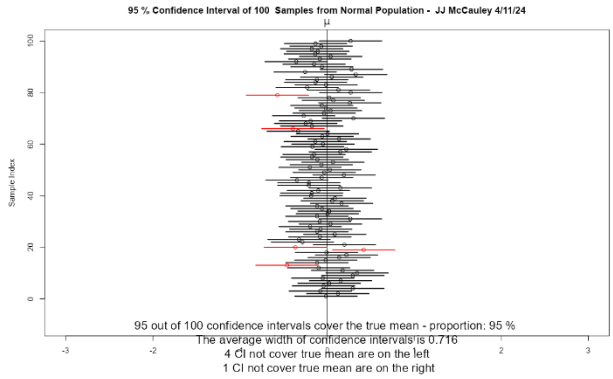
Probabilities was first introduced in lab 2, where the cross-tabulation function was used to create functions, then manual probabilities were calculated. This allowed for a refresher of calculating and interpreting probabilities, as well as learning the syntax for typing important statistical symbols (union,

intersection, power) in a markdown file. In lab 3, we expanded these concepts through various commands and through the Binom_plotter() command, which allows us to visualize the distribution of probabilities.



These ideas were expanded in lab 4, where we used the probability density function to determine the relative likelihood of a random variable falling within a given value. The functions that we used to find this were pnorm() for normal distributions and pexp()/qexp(), which we later tied into real-world applications to ensure that we can apply these concepts.

Confidence Intervals



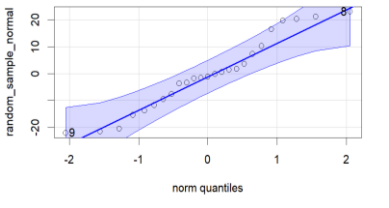
Lab 6 focuses primarily on Confidence Intervals, which allows us to estimate certain parameters of an unknown population, given a sample and measure of reliability. For instance, we could look at various samples and their means to draw conclusions about the population mean within a given confidence level. This lab focused on plotting confidence intervals for different intervals, populations, and sample sizes. One of the primary takeaways of this lab was the effect of different confidence levels, in which it was concluded that higher confidence levels would output a wider spread of results, while lower confidence levels tend to be more precise. Understanding these distinctions is very important, as different scenarios may warrant different confidence intervals to be analyzed.

Hypothesis Testing

Lab 7 focused primarily on hypothesis testing for large and small samples, specifically using the appropriate test, and understanding its results. In this lab, we practiced the 8-step hypothesis test procedure, which included determining the appropriate test. For this, we would examine the sample size of the samples being tested and the sample's distribution. If we are observing means, and the sample size is sufficiently large (≥ 30) or the data is normally distributed, we would run a one-sample t-test. If we are observing medians, the sample size is not sufficiently large, and the data is not normally distributed, then we would run a Wilcoxon test. Lab 8 expanded this to observe hypothesis testing for data with two samples, running tests on the differences or running two-sample t-tests.

```
        One Sample t-test

data:  x1
t = -1.9453, df = 9, p-value = 0.0418
alternative hypothesis: true mean is less than 450
95 percent confidence interval:
     -Inf 443.9602
sample estimates:
mean of x
    345.3
```



By using hypothesis testing with the appropriate test, claims can be either supported or denied when given sample data. This is extremely important when validating your own hypothesis, or when attempting to disprove somebody else's.

My Future with R

Throughout these series of labs, I have learned an immense amount about both the R language and Statistics as a whole. As a Data Science major, I find the powerful tool of R to be vastly interesting and would love to explore more of the following concepts:

1. Data Preparation: A necessary step when working with any type of data is data manipulation and preprocessing, which is something that R excels at. In the coming months, I plan on building on top of the knowledge built upon this course to eventually clean data to train a machine learning model.
2. Data Visualization: On top of the methods already discussed in class, I plan to learn more ways to visualize data.
3. Feature Engineering: Building on top of the methods discussed in this course, I plan to learn how to utilize R to transform raw data into trainable data for a ML model.
4. Machine Learning Models: Although this is indefinitely outside the scope of this introductory course, I would love to explore the capabilities of creating Machine Learning Models in R.

These concepts will only be complemented by the rich understanding of the fundamentals developed in this course.

Conclusion

Throughout this semester, this course has given me a comprehensive understanding of R as a means of statistical thinking. This portfolio not only showcases the skills and experiences I have accumulated through lab sessions but also marks a significant steppingstone in my journey towards a career in data science/ AI engineering.