EXPERIMENTAL DATA COLLECTION AND DESCRIPTIVE STATISCTICS PART - I

Submitted by

Nelson Joseph (1002050500)

in partial fulfilment for the award of the degree of

M.S. IN DATA SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON

DASC 5302 INTRODUCTION TO PROBABILITY AND STATISTICS

INSTRUCTOR: Dr. Kendra L. Wallis

NOVEMBER 2022

Honor Code:

I Nelson Joseph did not give or receive any assistance on this project, and the report submitted is wholly on my own.

Introduction

The project’s main objective is to examine data from the real world. For this study, I investigated and analysed two different sets of real-world data. Dataset 1 contains the shoe sizes of a sample of 100 people. Dataset 2 was acquired at the University of Texas at Arlington Central Library and looked at the time difference between two students who entered the library through the main door. To better understand the trend of the population from the sample, the data is statistically aggregated and visualized.

Data Collection

The US shoe size of randomly chosen 100 individuals at The University of Texas at Arlington campus was taken for **Data set 1** on November 11, 2022. I asked the people I had chosen their shoe size and explained that the information was for a project. The 100 randomly chosen individuals represent a diverse group of people, including Americans, Asians, and individuals from other backgrounds. By correctly documenting the data collecting process, the data is carefully collected. The shoe size was noted along with the race and gender. The final Excel document had the following columns: "**Race**" (noted when inquiring about shoe size), "**Gender**" (gathered when asking about how each respondent identified themselves), and "**US shoe size**" (Collected by asking them). After the data was successfully collected, descriptive statistics were run to look for any patterns in the data. This includes, among other things, investigating the sample mean and standard deviation, compiling the data into a table, and making cumulative and relative frequency histograms.

Data for **Dataset 2** was gathered from The University of Texas at Arlington library on Monday, November 15, 2022, between 3 PM and 4 PM. One of the busiest areas at UTA is the library. The student typically visits this location for educational and recreational purposes. The library includes six stories and a basement that is exclusively used for gaming and leisure activities. Two of the six floors are designated for silent study, with the other four floors being used for group study. Data gathered for students accessing the library is used in this analysis. The event is defined as the student entering at a predetermined time, which is used as the start time. Every time a student enters the library, the time is recorded 101 times in succession.

The samples were taken using a timer on my personal phone and a clock app to record the time. Each lap represented a fresh interval between two students entering the library. The time intervals were then exported to an Excel file after the data gathering was finished. The Excel document that was created had the following columns: "**Time**" (different time intervals), "**Time Difference in Seconds**" (time difference found by deducting the first interval from the second interval), and "**Seconds**" (Seconds). After the data was successfully collected, descriptive statistics were run to look for any patterns in the data. Analysis of the patterns in students' interest in libraries can be done using this data. If we gather comparable information for additional structures, we can determine which structures draw students the most.

Descriptive Statistical Analysis

To extract useful insights from the data and showing them, descriptive analysis is crucial. Additionally, it aids in highlighting any potential connections between the variables. Performed descriptive statistics by using the datasets collected and processing in Excel: **AVERAGE (), MEDIAN (), STDEV.S (), QUARTILE (), VAR.S ()** and **STDEV.S ().**

Turning to the Relative Frequency Table, the data was divided into several classes, and the Online descriptive statistics calculator was used to determine the number of records that belong to each class. The proportion between the corresponding class count and total records count is used to calculate the relative frequency of each class. By using the class intervals and relative frequency columns from tables, created a relative frequency histogram, as well as a cumulative relative frequency histogram using the same class intervals and cumulative relative frequency columns. The built-in histogram function in Excel plots data that is provided as histograms.

Data set 1:

Statistics Value Units	Values	Units
Sample Mean	9.92	US
Sample Median	10.00	US
Sample Mode	10	US
Sample Variation	3.57	US
Sample Standard Deviation	1.89	US
Sample Range	9	US
Quartile: Q1	9	US
Quartile: Q2	10	US
Quartile: Q3	11	US

Table 1.1 Descriptive Statistics Exploring the US shoe size of randomly selected 100 people within UTA

Table 1.1 displays the measures of central tendency and measures of variability for the sample data. The relative frequency and cumulative relative frequency tables, as well as subsequent visualizations of the same, are made using this tabular data.

Tabular Summary of the Dataset 1			
Class Interval	Count/Frequency	Relative Frequency	Cumulative Relative Frequency
6 - 8	12	0.12	0.12
8 - 10	33	0.33	0.45
10 - 12	36	0.36	0.81
12 - 14	15	0.15	0.96
14 - 16	4	0.04	1.00
Sum	**100**	**1**	

Table 1.2 Tabular data representing the relative frequency and cumulative relative frequency of Dataset 1

The frequency count of students in each interval is shown in **Table 1.2**. Each frequency count is divided by the overall frequency to determine the relative frequency. The relative frequencies of the next periods are continually added to determine the cumulative frequency.

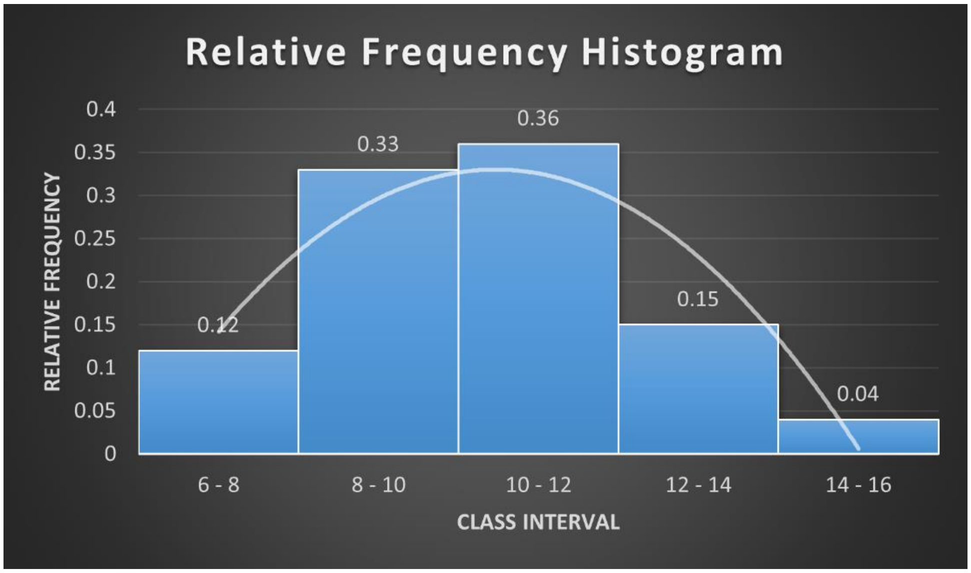

Figure 1.1 Relative Frequency Histogram of Data set 1

The percentage of students who belong to each group is revealed by the relative frequency histogram. Here, majority of students' shoe sizes fall between 10 and 12, which explains why the median and mean values of the data are 10 and 9.92 respectively. Only 4% of the sample data, or relatively few individuals, have shoes in the size range of 14 to 16.

The sample data displays like a normal distribution, but the mean is not located in the exact middle of the distribution. The distribution must be symmetrical with the mean centre to be considered normal. However, the mean value is 9.92 while the maximum values range from 10 to 12. As a result, **it can be approximated to a normal distribution** but is not a true normal distribution.

The US shoe size cumulative frequency histogram is shown below **Figure 1. 2.** Each bar height in cumulative frequency histograms displays the number of values in that interval as well as all lower intervals.

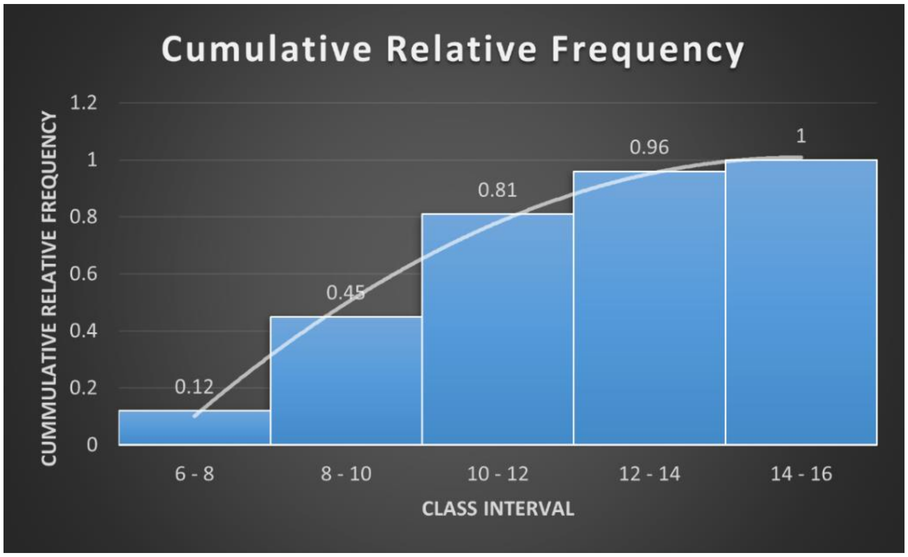

Figure 1.2 Cumulative Relative Frequency histogram of US shoe size for Data set 1

To better effectively understand the entire sample of data, a box and whisker plot is created in **Figure 1.3**. A five-number summary of a collection of data is given. The minimal US shoe size is 6; the first quartile is 9, the median is 10, the third quartile is 11, and the maximum is 15. As they highlight outliers within a data collection, box charts are helpful. An observation that deviates numerically from the rest of the data is known as an outlier. The observation of a 15 US shoe size is unusual, hence it is marked as an outlier.

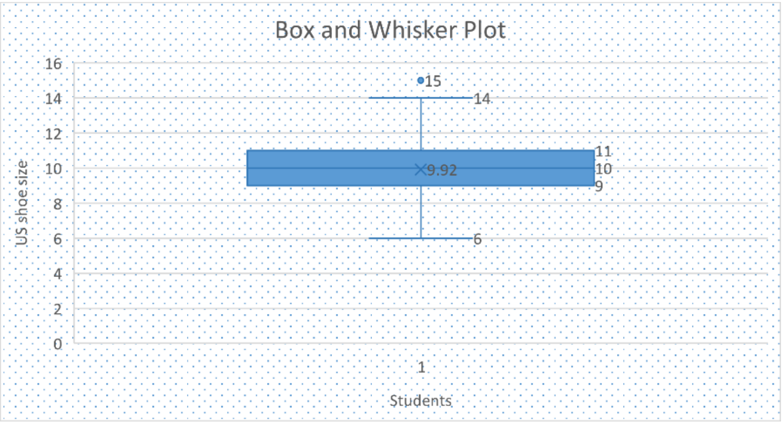

Figure 1.3 Box and whisker plot for Data set 1

Data set 2:

Statistics Value Units	Values	Units
Sample Mean	**7.23**	*Seconds*
Sample Median	**6.00**	*Seconds*
Sample Mode	**4**	*Seconds*
Sample Variation	**17.82**	*Seconds*
Sample Standard Deviation	**4.22**	*Seconds*
Sample Range	**21**	*Seconds*
Quartile: Q1	**4**	*Seconds*
Quartile: Q2	**6**	*Seconds*
Quartile: Q3	**9**	*Seconds*

Table 2.1 Descriptive Statistics Exploring the time interval between students entering the UTA Central Library

The descriptive statistics values for the intervals between students entering the library are shown in **Table 2.1**. Here, average time it takes a student to enter the library is 7.23 seconds. While 43% of the students enter within a 6-second window.

Tabular Summary of the Dataset 2			
Class Interval	**Count/Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0 <= x < 6	43	0.43	0.43
6 <= x < 12	34	0.34	0.77
12 <= x < 18	18	0.18	0.95
18<= x < 24	5	0.05	1.00
Sum	**100**	**1.00**	**1.00**

Table 2.1 Tabular Summary of the Data set 2 combined into different intervals.

In **Table 2.1,** categorized the data from Data Set 2 into 5 classes for the tabular overview of data. In comparison to all other intervals, the number of students entering the library between 0 and 6 seconds is the highest.

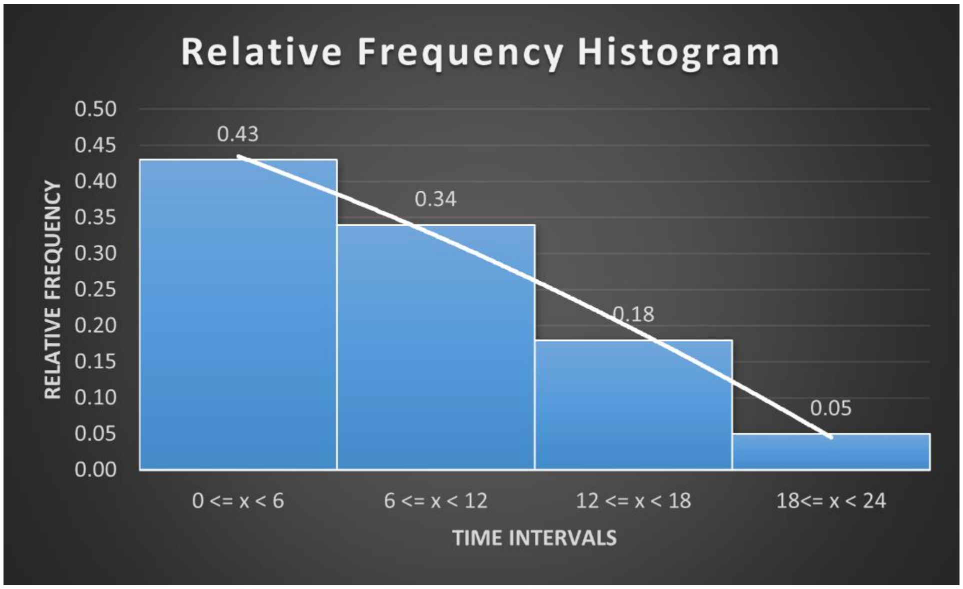

Figure 2.1 Relative Frequency Histogram of the Dataset 2

Figure 2.1 shows that when the time interval lengthens, fewer students are entering the library on a relative basis. Given that students enter the library on average in 7.73 seconds, it makes sense that it is quite packed. The relative histogram tells us that, after one student enters the library, 77% of students arrive there in less than 12 seconds from our sample data.

The time of arrival of students at the library is taken from the sample Data set 2. **For the selected intervals, it follows an exponential distribution** in this case. As the time passes, fewer students are entering the library. Most students arrive with a shorter gap, typically between 0 and 6 seconds. But when I reduced the interval gap it does not follow an exponential distribution.

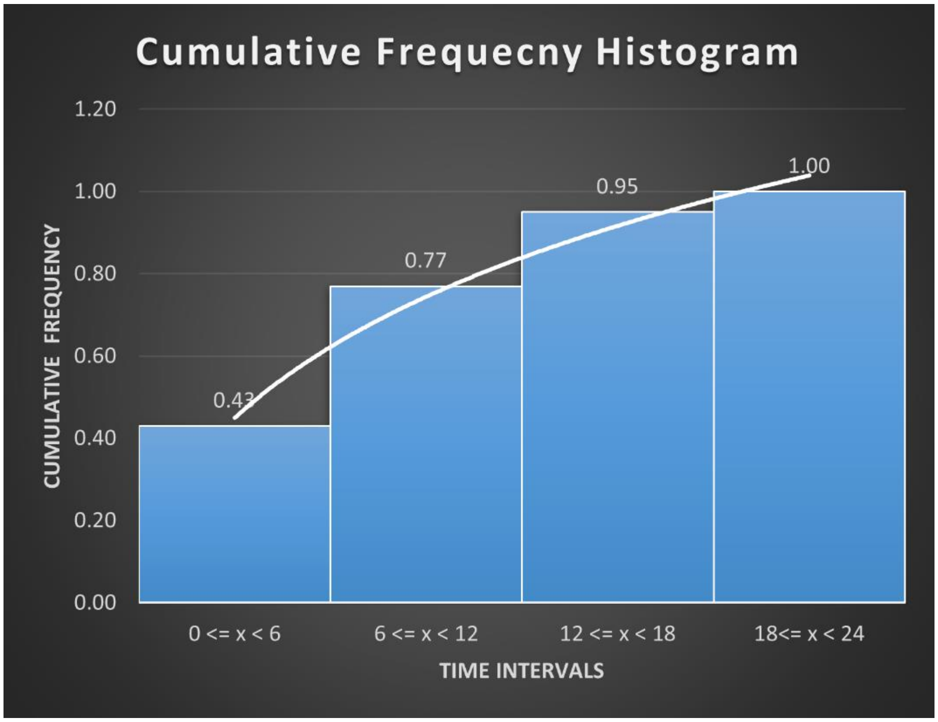

Figure 2.2 Cumulative Relative Frequency Histogram of Data set 2

According to **Figure 2.2**, every student in our sample set enters the library between 0 and 24 seconds after another student.

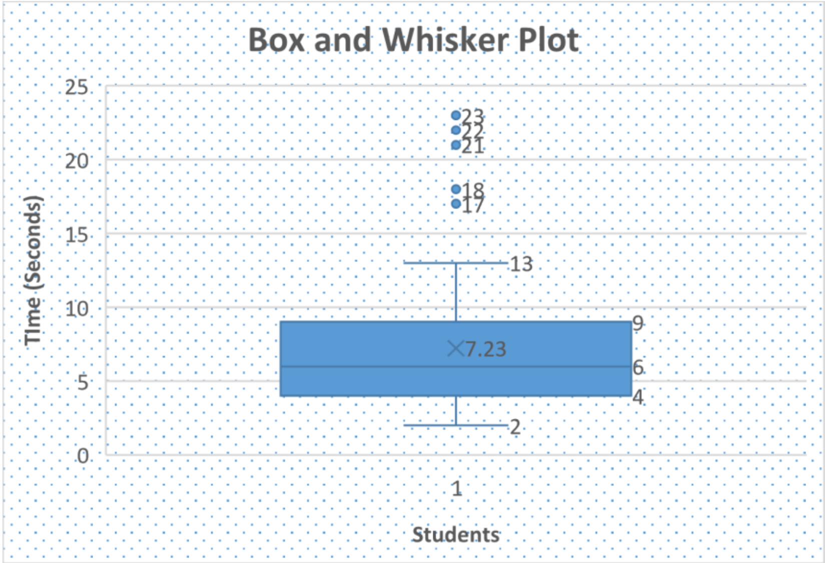

Figure 3.3 Box and Whisker Plot of Data set 2

Outliers can be identified using Data Set 2's box and whisker plot. The sample data outliers in this case are 17, 19, 21, 22, and 23.

Conclusion

Given that it can be challenging to glean useful information from raw data, descriptive analysis uses statistical approaches to interpret insights from data by describing it. Descriptive statistics are crucial for facilitating data visualization and presenting data in a relevant way. By removing less important data, it makes it possible to highlight any trends in the dataset and ignore outliers. On two datasets, I ran the statistical analysis on first data set, which examined the US shoe sizes on November 15, 2022, among 100 randomly chosen students. The goal of this analysis was to look for any trends in the shoe sizes of the 100 students who were chosen at random. Data Set 2 was gathered from the library on November 15, 2022, between 3 - 4 PM. I gathered information on the student's arrival times at the library. Finding trends in students' library interest is the goal of this investigation. Comparable data can be gathered for additional structures, allowing us to identify which ones entice pupils the most.

Data set 1 was subjected to descriptive statistical analysis, and it was discovered that the average shoe size is **9.92, or roughly 10**. According to Table 1, many students likely wear shoes with a size between **10 and 12**. In accordance with Table I.2, **81%** of students have shoes that are smaller than size 12. Additionally, Figure 1.1 shows that just **4%** of the sample data, or only a small number of people, use shoes that are between sizes **14 and 16**. These findings are since most of the data was collected from graduate students, many of whom had large feet due to their age range of 23 to 28. Additionally, it has been shown that US citizens tend to have larger shoe sizes than other people. **Although not a true normal distribution, the Data Set 1 can be approximated to one.** To have a normal distribution, the mean, median, and mode must all be equal; in this case, they are 9.92, 10 and 10, respectively.

For **Data set 2**, According to descriptive statistics, the average amount of time it takes a student to enter the library is **7.23 seconds**. While **43%** of the students enter in the **first 6 seconds**. The biggest number of students enter the library between **0 and 6 seconds**, according to Table 2.1, when compared to all other intervals. According to the relative histogram, based on our sample data, 77% of students enter the library within 12 seconds of the first student doing so. The median student arrival time at the library was 6 seconds, which is shorter than the mean of 7 minutes and 7 seconds. As a result, it meets one of the requirements for an exponential distribution. **It follows an exponential distribution for the chosen intervals**. A decreasing number of students are using the library as time goes on. Many students come with a gap that is shorter, usually between 0 and 6 seconds. However, the interval gap does not follow an exponential distribution in a histogram when I reduced the time intervals.

Appendix

You can view the original data by visiting the following link. This sheet includes calculations for descriptive statistics as well as data for both Data Set 1 and Data Set 2. This file consists of 3 sheets each, including sheets with the data and descriptive statistics for Data Set 1 and Data Set 2 respectively.

Data set 1.xlsx

Data set 2.xlsx

References

Joos Korstanje. (2019, December 13). 6 ways to test for a Normal Distribution
https://towardsdatascience.com/6-ways-to-test-for-a-normal-distribution-which-one-to-use-9dcf47d8fa93

Calculator Soup | Descriptive Statistics Calculator
https://www.calculatorsoup.com/

The University of Utah. The Normal Distribution and Z Scores
https://soc.utah.edu/sociology3112/normal-distribution.php#:~:text=The%20normal%20distribution%20is%20a%20symmetrical%2C%20bell%2Dshaped%20distribution%20in,central%20component%20of%20inferential%20statistics.

Andy Polk. (2021, April 14). Americans' feet getting bigger, but shoe choices slim
https://fdra.org/latest-news/americans-feet-getting-bigger-but-shoe-choices-slim/#:~:text=Shoe%20manufacturers%20and%20retailers%2C%20parents,9%C2%BD%20to%2010%C2%BD%20for%20men.

Courtney Taylor. (2019, January 02). Exponential Distribution Medians
https://www.thoughtco.com/calculate-the-median-of-exponential-distribution-3126442#:~:text=Median%2DMean%20Inequality%20in%20Statistics,is%20less%20than%20the%20mean.