# Lab 2: D7054E

Sebastian Morel (sebmor-8)
Patrick Pilipiec (patpil-9)

**Group** 7

February 26, 2021

## Introduction

In lab 2 is divided up into two parts part A and part B. Part A is again divided up into four parts, in part one to three in part A we are tasked to solve textbook problems. The first part of part A is about continuous distribution [1], the second part normal distribution [2], and the third part are questions centered around the central limit theorem [3]. In the fourth part we need to perform three different data analysis tasks, on a Melbourne housing dataset[4]. For part B we will once again work with the Melbourne housing dataset and use it to explore the gradient descent algorithm.

## 1 Methodology

Lab 2 was performed in Spyder and Jupyter notebook inside a virtual environment using the Anaconda navigator, in order to isolate things such as library versions from other virtual environments and on the system.

### 1.1 Techniques

For part A we mainly used pyplot to plot our graphs, and seaborn to plot our heatmaps. The data was mainly contained in a pandas dataframe for easier processing, and sometimes in a numpy array to for example when calculating percentiles for the second part and to help us plot a normal distribution curve where we used Scipy to create a probability density function from the numpy array.
For part B, we used Pandas, Numpy, and Matplotlib. Pandas was used to load data from CSV, clean data, and to work with dataframes. Numpy was used for performing vector operations, which are noticeable faster than running for loops in Python. Matplotlib was used for generating graphs.

### 1.2 The dataset

For lab 2 different types of datasets were used. In the first part of Part A our dataset was composed of 50 randomly generated values between zero and one, this dataset was generated through a random number generator. For the second part we were tasked to use an already existing dataset [5], this dataset is comprised of 20 rows each representing a unique race, and seven columns each representing the lap for the corresponding race. The third part has one column and 60 rows, each row represents what number a recipe has going from 0-60, and the column represents how long that recipe lasted at the Olmstead Homestead. The Melbourne housing dataset for the fourth part of Part A and Part B, contains 13 580 rows each representing a property, the dataset also contains 21 columns representing 21 different variables.

Table 1: Melbourne housing dataset name and type of each column.

| Column name | Data type |
|---|---|
| Suburb | string |
| Address | string |
| Rooms | int |
| Type | string |
| Price | float |
| Method | string |
| SellerG | string |
| Date | string |
| Distance | float |
| Postcode | float |
| Bedroom2 | float |
| Bathroom | float |
| Car | float |
| Landsize | float |
| BuildingArea | float |
| YearBuilt | float |
| CouncilArea | string |
| Lattitude | float |
| Longitude | float |
| Regionname | string |
| Propertycount | float |

### 1.3 The pre-processing of the dataset

To gather the data from the web pages for part A2 and A3 we used pandas' read_html function in order to convert the html data table to a dataframe. The Melbourne housing dataset was provided as a csv document we therefore used pandas' read_csv command to convert the csv datafile into a dataframe.

## 2 Results

### 2.1 Part A

#### 2.1.1 Continuous Distribution

**Collect the Data**

1. **Complete the table.**
   See the completed table in Table 2. This table shows 50 generated values between zero and one (inclusive). Numbers were rounded to four decimal places.

Table 2: 50 randomly generated number between 0 and 1 rounded to four decimal places.

| | | | | |
|---|---|---|---|---|
| 0.7674 | 0.9527 | 0.3973 | 0.7467 | 0.6425 |
| 0.5094 | 0.5767 | 0.6716 | 0.2754 | 0.2032 |
| 0.2113 | 0.6239 | 0.4760 | 0.5295 | 0.4359 |
| 0.9291 | 0.9112 | 0.8121 | 0.0577 | 0.7696 |
| 0.1986 | 0.5873 | 0.1148 | 0.0781 | 0.0990 |
| 0.6290 | 0.2410 | 0.8340 | 0.2268 | 0.5989 |
| 0.5338 | 0.3995 | 0.4968 | 0.7830 | 0.4956 |
| 0.0481 | 0.9763 | 0.4849 | 0.1970 | 0.0078 |
| 0.9676 | 0.7696 | 0.3223 | 0.5519 | 0.3817 |
| 0.1141 | 0.1573 | 0.2275 | 0.1061 | 0.5234 |

2. **Calculate the following:**
   $\bar{x} = 0.47302$
   $s = 0.278815184665398$

first quartile = 0.21517499999999998
third quartile = 0.664325
median = 0.4962

**Organize the Data**

1. **Construct a histogram of the empirical data. Make eight bars.**
   See Figure 1 for the histogram of the empirical data using eight bars.

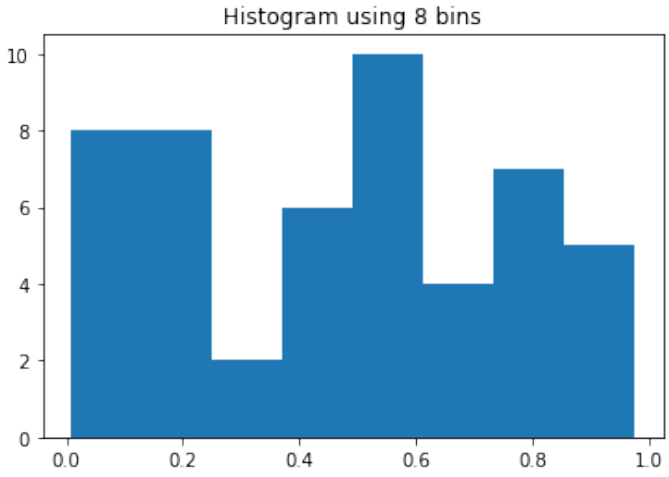


Figure 1: Histogram using Eight Bins

2. **Construct a histogram of the empirical data. Make five bars.**
   See Figure 2 for the histogram of the empirical data using five bars.

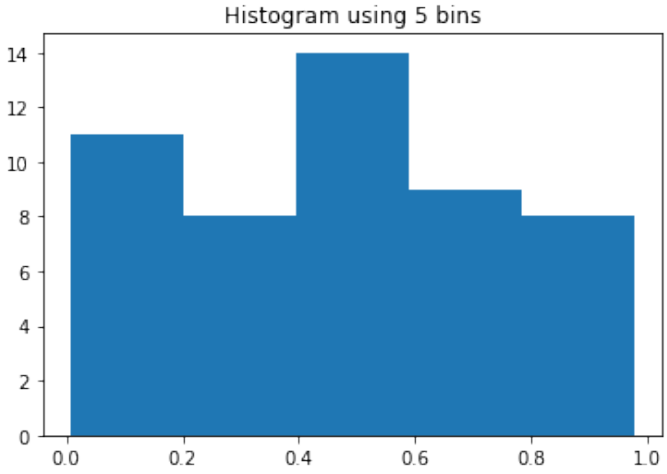


Figure 2: Histogram using Five Bins

**Describe the Data**

1. **In two to three complete sentences, describe the shape of each graph. (Keep it simple. Does the graph go straight across, does it have a V shape, does it have a hump in the middle or at either end, and so on. One way to help you determine a shape is to draw a smooth curve roughly through the top of the bars.)**
   The histogram in Figure 1 with eight bins shows that values between 0.5 and 0.6 are most common. The histogram has very occurrences of values between 0.2 and 0.4. We start to observe a V shape this is flipped 180 degrees. The histogram in Figure 2 with five bins that values between 0.4 and 0.6 are most common. There are slightly more values between 0.0 and 0.2, compared to the remaining three bars in this graph. This graph starts to look more like a continuous distribution. We start to observe a V shape this is flipped 180 degrees.

2. **Describe how changing the number of bars might change the shape.**
   As the histogram illustrates, a normal distribution can be better recognized in the dataset when the number of bins is increased. More bars gives more shape to the distribution, as compared to fewer bars. Although this normal distribution is far from perfect, the pattern is certainly an improvement compared to when the number of bars is low. However, when dealing with few observations, we concurrently find that bars may be shorter and that there exist gaps between bars. Therefore, to approach the normal distribution, it is key to have many bars and many observations. For comparative purposes, see also the accompanied code for other visualizations of histograms that use 10, 15, 20, and 25 bars.

**Theoretical Distribution**

1. **In words, X =**
   X = a random variable that is distributed uniformly

2. **The theoretical distribution of X is X U(0,1).**
   There is no question here that can be answered.

3. **In theory, based upon the distribution X U(0,1), complete the following.**
   **For the theoretical distribution:**
   $\mu = 0.5472998576534653$
   $\sigma = 0.30062706459044447$
   first quartile = 0.28381132732732905
   third quartile = 0.830700148937587
   median = 0.5595067094062611

   **For the empirical distribution:**
   $\mu = 0.47302$
   $\sigma = 0.278815184665398$
   first quartile = 0.21517499999999998
   third quartile = 0.664325
   median = 0.4962

4. **Are the empirical values (the data) in the section titled Collect the Data close to the corresponding theoretical values? Why or why not?**
   Overall, we find that there are differences between statistics for the empirical and theoretical distribution. The theoretical distribution produces statistics that we would expect. Therefore, in the theoretical distribution, we would expect a first quartile of 0.25, mean and median of 0.50, and a third quartile of of 0.75. We, however, observe that the theoretical distribution is still different from these theoretical values.
   By comparing the empirical distribution against the theoretical distribution, we find that the former has a lower mean value, a slightly lower standard deviation, the first quartile is roughly 0.07 points lower, while the third quartile is 0.17 points lower. Both distributions also have a different median. The middle value in the empirical distribution lies at 0.50, while it is 0.56 in the theoretical distribution.
   In this sense, the empirical distribution is closer to the theoretical median value of 0.50 compared to the theoretical distribution!

**Plot the data**

1. **Construct a box plot of the data. Be sure to use a ruler to scale accurately and draw straight edges.**
   We used Matplotlib the generate the boxplots in 3 and 4.

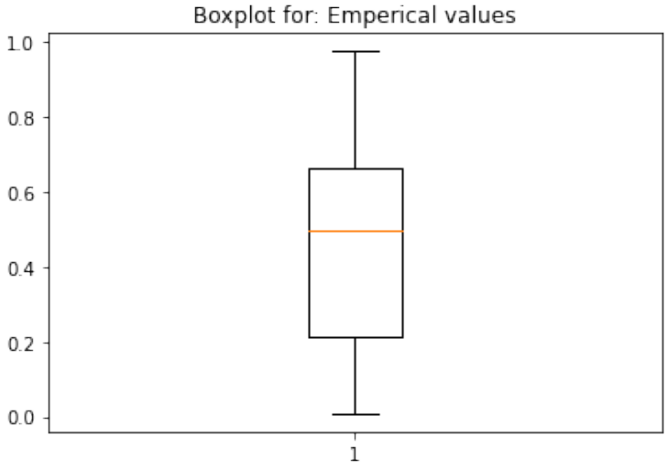


Figure 3: Boxplot for Empirical Values

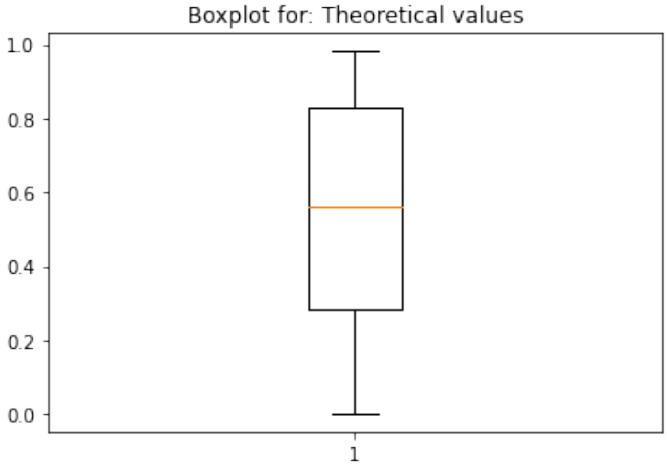


Figure 4: Boxplot for Theoretical Values

2. **Do you notice any potential outliers? If so, which values are they? Either way, justify your answer numerically. (Recall that any DATA that are less than Q1 − 1.5(IQR) or more than Q3 + 1.5(IQR) are potential outliers. IQR means interquartile range.)**
   In the plots, we do not observe dots that represent outliers. However, it is notable that, see Figure 3, the upper whisker is longer than the bottom whisker, which may indicate that more observations. Also, we find that the IQR does range from 0.2 to 0.65, instead of the expected range from 0.25 to 0.75. Therefore, it is possible that there exist outliers. Also, in the boxplot in Figure 4, we find that the IQR starts at 0.25 and goes until 0.80, which is slightly longer than the expected 0.50, which starts at 0.25 and goes until 0.75. Again, we do not visually see outliers here.

**Compare the data**

1. **For each of the following parts, use a complete sentence to comment on how the value obtained from the data compares to the theoretical value you expected from the distribution in the section titled Theoretical Distribution.**
   See Figure 5 for an overview of the statistics for the theoretical and empirical distribution. These statistics will be intepreted and described below.

```
Uniform
μ = 0.49453837041949206
σ = 0.2891982909162114
min = 0.030723393834400037
first quartile = 0.22616532524348035
median = 0.5214021222437473
third quartile = 0.7262130998032923
ICQ = 0.500047774559812
max = 0.953853953954548

Random
μ = 0.43176999999999993
σ = 0.25596833183032625
min = 0.0089
first quartile = 0.21802500000000002
median = 0.46504999999999996
third quartile = 0.59565
ICQ = 0.377625
max = 0.9531
```

Figure 5: Statistics for Uniform (Theoretical) and Random (Empirical) Distribution

(a) minimum value: The minimum value in the uniform distribution is 0.031, while it is 0.009 in the random distribution. These values lie very close. Interestingly, no value is exactly 0.

(b) first quartile: The first quartile in the uniform distribution is 0.227 and 0.218 in the random distribution. These values are very comparable.

(c) median: The median value in the uniform distribution is 0.512 and 0.465 in the random distribution. These values are comparable.

(d) third quartile: The third quartile in the uniform distribution is 0.726 compared to 0.596 in the random distribution. Both values are quite different, meaning that the values in the uniform distribution are as expected (0.726 is very close to the expected 0.750), while in the random distribution, there are more values that are closer to 0, which shifts the third quartile from the expected 0.750 to 0.596.

(e) maximum value: The maximum value is comparable in both distributions, namely 0.954 and 0.953. Interestingly, no value is precisely 1.

(f) width of IQR: The ICQ in the uniform distribution, 0.500 is the same as the expected ICQ of 0.500. The ICQ is however far smaller in the random distribution (0.378), which means that there are likely outliers. These outliers were however not visualized on the box plot above.

(g) overall shape: while the boxplot constructed using the theoretical values is exemplary for what we expect, we find that the boxplot that was created using the empirical values has a smaller IQR and more values lie above the third quartile. This is an indication that there may be outliers.

2. **Based on your comments in the section titled Collect the Data, how does the box plot fit or not fit what you would expect of the distribution in the section titled Theoretical Distribution?**
The boxplot does not very well fit the normal distribution that we expected. This may be explained by the small number of observations. Therefore, increasing the number of observations, preferably as close as possible to infinity, makes that a normal distribution will emerge.

**Discussion question**

1. **Suppose that the number of values generated was 500, not 50. How would that affect what you would expect the empirical data to be and the shape of its graph to look like?**
As the number of values increases, e.g., from 50 to 500, this also means that the overall distribution of these values is less likely to be distorted by bias, and instead, the distribution of these values is expected to be more likely as the normal distribution.

### 2.1.2 Normal Distribution

**Collect the Data**

1. **Use the data from Appendix C. Use a stratified sampling method by lap (races 1 to 20) and a random number generator to pick six lap times from each stratum. Record the lap times below for laps two to seven.**

Table 3: Stratified sampled times from laps two to seven.

| | | | | | |
|---|---|---|---|---|---|
| 130 | 133 | 128 | 124 | 127 | 132 |
| 130 | 130 | 127 | 129 | 130 | 130 |
| 129 | 131 | 128 | 126 | 130 | 129 |
| 128 | 126 | 130 | 131 | 126 | 128 |
| 131 | 132 | 128 | 128 | 129 | 128 |
| 131 | 131 | 129 | 130 | 128 | 124 |

2. **Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.**

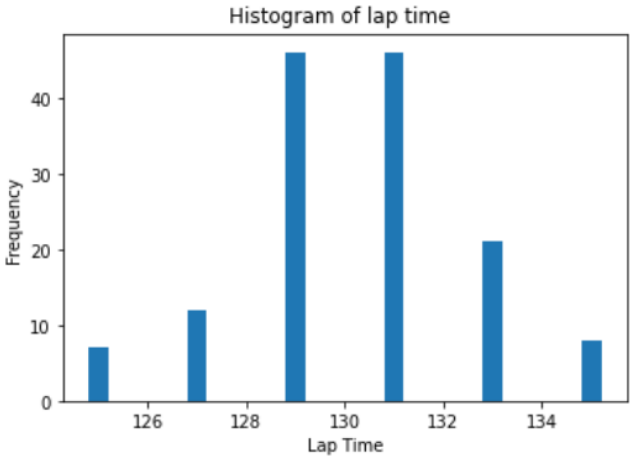


Figure 6: Histogram of laps in 6 intervals

3. **Calculate the following:**
   (a) $\bar{x} = 129.74$
   (b) Because this is stratified sampling we need to take into account the bias, therefore instead of taking the standard deviation of the whole dataframe we first need to take the sum of the mean variance for each stratum and then we take the square root of that to get the standard deviation. $s = 2.2$

4. **Draw a smooth curve through the tops of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a v-shape, does it have a hump in the middle or at either end, and so on?)**

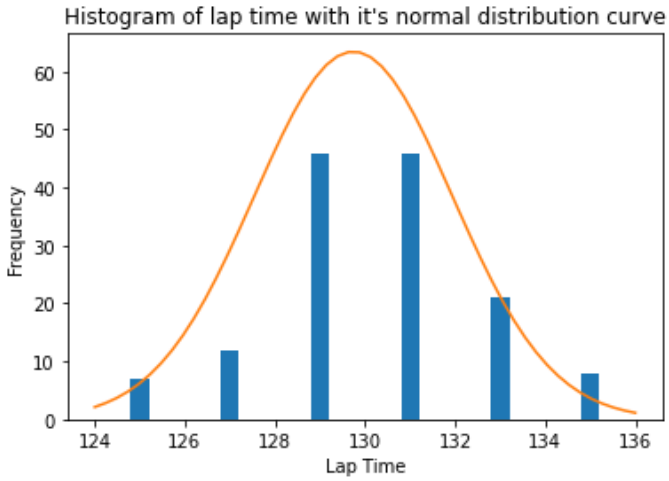


Figure 7: Histogram of laps in 6 intervals with normal distribution curve.

Description: The shape of the curve is a bell curve, the mu value is at the average 129.74, by looking at it it looks a bit too much to the left, this is because of the intervals. There is actually alot more values at 124 than at 136, which is why it is shifted more to the left.

**Analyze the Distribution**
**Using your sample mean, sample standard deviation, and histogram to help, what is the approximate theoretical distribution of the data?**
$X \sim N(129.74, 2.2)$
**How does the histogram help you arrive at the approximate distribution?**
The histogram didn't help me at arriving to the approximate distribution. The histogram only shows how the data is distributed, not the specific values. Although if we wanted a rough estimate of the distribution the histogram could be used to find an approximate mean which would be at 130 in this case, and because we only have 6 intervals we could easily with the help of the mean, calculate the standard deviation with a formula.

**Describe the Data**
**Use the data you collected to complete the following statements.**
The IQR goes from 128.0 to 131.0
The IQR is 3.0
The 15th percentile is 128.0
The 85th percentile is 132.0
The median is 130.0
The empirical probability that a randomly chosen lap time is more than 130 seconds is 35.0%
The 85th percentile shows the time you need to have a better time than 85% the other.

**Theoretical Distribution**
**Using the theoretical distribution, complete the following statements. You should use a normal approximation based on your sample data.**
The IQR goes from 128.26 to 131.22
The IQR is 2.96
The 15th percentile is 127.47
The 85th percentile is 132.03
The median is 129.74
The probability that a randomly chosen lap time is more than 130 seconds is 45.22 %
The 85th percentile in the normal distribution shows what time you need to have a better time than 85% of the other times.

**Discussion Questions**
**Do the data from the section titled Collect the Data give a close approximation to the theoretical distribution in the section titled Analyze the Distribution? In complete sentences and comparing the result in the sections titled Describe the Data and Theoretical Distribution, explain why or why not.**
Yes, because in the "Collect the Data" section we calculated the mu and sigma for the histogram, which equates to

the theoretical distribution of the histogram. For the "Describe the Data" and "Theoretical distribution" the values are not very close, one of the reasons is because the theoretical distribution takes into account float numbers while the empirical data only has integers. This is also why "The probability that a randomly chosen lap time is more than 130 seconds" is so different between the two. While the collected data only takes values that are 131 seconds or greater, the theoretical distribution will start at whatever float is the closest to 130 seconds. Which is why there is a bigger probability for the theoretical distribution.

### 2.1.3 Central Limit Theorem

**Given**
**Calculate the following:**
$\mu_x = 3.566667$
$\sigma_x = 2.165732$
**Collect the Data**

1. **Complete the table:**
   See Figure 8 for the data, sample means, and standard deviations.

| Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample means from other groups: |
|---|---|---|---|---|
| 6.0 | 4.0 | 3.0 | 2.0 | 1.0 |
| 6.0 | 5.0 | 2.0 | 1.0 | 5.0 |
| 2.0 | 1.0 | 11.0 | 2.0 | 6.0 |
| 5.0 | 6.0 | 2.0 | 1.0 | 5.0 |
| 4.0 | 1.0 | 6.0 | 6.0 | 2.0 |
| | | | | 5.0 |
| | | | | 1.0 |
| | | | | 1.0 |
| | | | | 3.0 |
| | | | | 2.0 |
| | | | | 2.0 |
| | | | | 2.0 |
| | | | | 4.0 |
| | | | | 6.0 |
| | | | | 1.0 |
| | | | | 6.0 |
| | | | | 5.0 |
| | | | | 2.0 |
| | | | | 5.0 |
| | | | | 4.0 |
| | | | | 6.0 |
| | | | | 2.0 |
| | | | | 5.0 |
| | | | | 6.0 |
| | | | | 6.0 |
| | | | | 1.0 |
| | | | | 1.0 |
| | | | | 1.0 |
| | | | | 6.0 |
| | | | | 2.0 |
| | | | | 2.0 |
| | | | | 5.0 |
| | | | | 6.0 |
| | | | | 5.0 |
| | | | | 1.0 |
| | | | | 1.0 |
| | | | | 2.0 |
| | | | | 4.0 |
| | | | | 3.0 |
| | | | | 5.0 |
| mean = 4.6 | mean = 3.4 | mean = 4.8 | mean = 2.4 | mean = 3.45 |
| S = 1.6733 | S = 2.3022 | S = 3.8341 | S = 2.0736 | S = 1.9474 |

Figure 8: Means for Five Randomly Selected Samples

2. **Complete the table:**
   See Figure 9 for the data, sample means, and standard deviations.

| Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample means from other groups: |
|---|---|---|---|---|
| 6.0 | 2.0 | 1.0 | 5.0 | 1.0 |
| 6.0 | 2.0 | 2.0 | 2.0 | 5.0 |
| 2.0 | 5.0 | 2.0 | 2.0 | 6.0 |
| 5.0 | 2.0 | 1.0 | 6.0 | 2.0 |
| 4.0 | 5.0 | 1.0 | 1.0 | 5.0 |
| 2.0 | 6.0 | 1.0 | 3.0 | 1.0 |
| 6.0 | 6.0 | 6.0 | 1.0 | 1.0 |
| 1.0 | 4.0 | 5.0 | 5.0 | 3.0 |
| 1.0 | 4.0 | 2.0 | 6.0 | 2.0 |
| 1.0 | 6.0 | 5.0 | 6.0 | 2.0 |
| | | | | 4.0 |
| | | | | 6.0 |
| | | | | 6.0 |
| | | | | 5.0 |
| | | | | 4.0 |
| | | | | 11.0 |
| | | | | 5.0 |
| | | | | 1.0 |
| | | | | 2.0 |
| | | | | 3.0 |
| mean = 3.4 | mean = 4.2 | mean = 2.6 | mean = 3.7 | mean = 3.75 |
| S = 2.2211 | S = 1.6865 | S = 1.9551 | S = 2.1108 | S = 2.4895 |

Figure 9: Means for 10 Randomly Selected Samples

3. **For the original population, construct a histogram. Make intervals with a bar width of one day. Sketch the graph using a ruler and pencil. Scale the axes.**
   See Figure 10

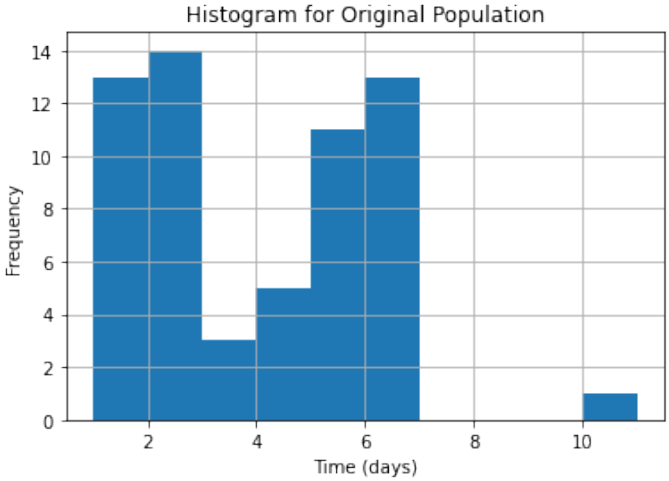


Figure 10: Histogram for Original Population

4. **Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.**
   See Figure 11. Overall, we find that there is no clear normal distribution visible in this graph. Also, around 10 days, there seems to be an outlier.

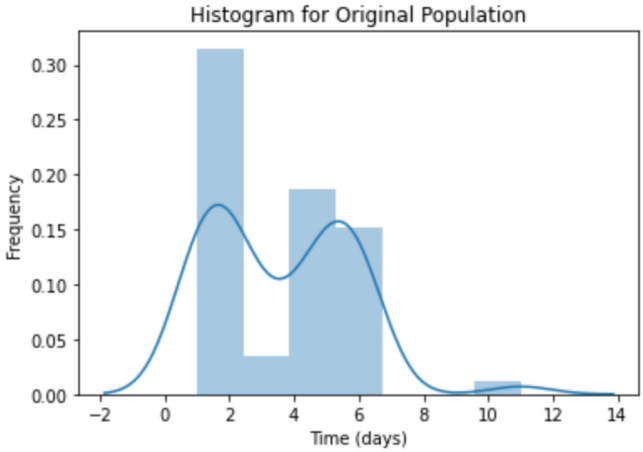


Figure 11: Histogram for Original Population with Curve Line

**Repeat the Procedure for n = 5**

1. **For the sample of n = 5 days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths of 1/2 a day. Sketch the graph using a ruler and pencil. Scale the axes.**
   See Figure 12.

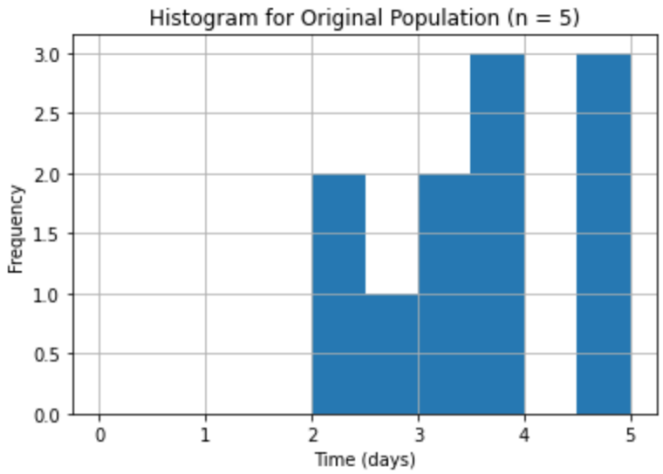


Figure 12: Histogram for Original Population (n = 5)

2. **Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.**
   Figure 13 displays the histogram including curve line where $n = 5$. In this figure, we indeed find a curve that is close to a normal distribution. However, the curve is fairly flattened, thereby, it does not resemble a theoretical normal distribution.

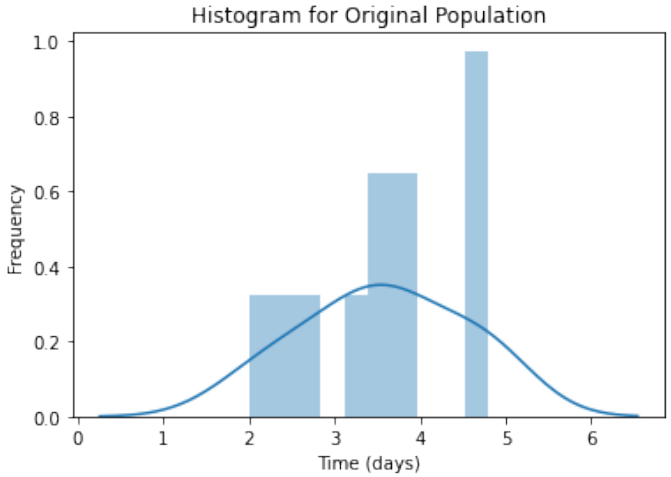


Figure 13: Histogram for Original Population (n = 5) with Curve Line

**Repeat the Procedure for n = 10**

1. **For the sample of n = 10 days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths of 1/2 a day. Sketch the graph using a ruler and pencil. Scale the axes.**
   See Figure 14.

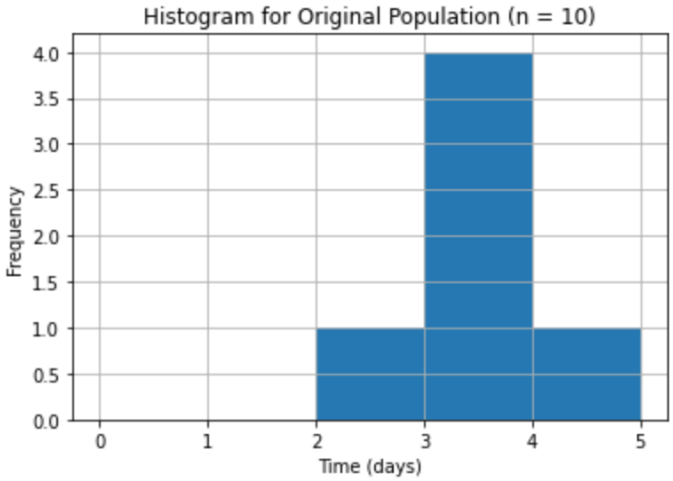


Figure 14: Histogram for Original Population (n = 10)

2. **Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.**
Figure 15 displays the histogram including curve line where $n = 10$. In this figure, we indeed find a curve that is very close to a normal distribution. However, the curve is still not perfect. Among others, we find that the left tail does not nicely descent, which may be explained by the values between 2.5 and 3.0 days. Note that the right tail descents very nicely.

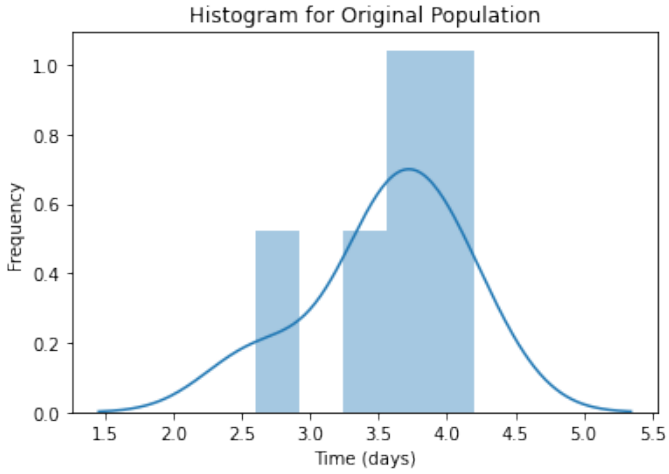


Figure 15: Histogram for Original Population (n = 10) with Curve Line

**Discussion Questions**

1. **Compare the three histograms you have made, the one for the population and the two for the sample means. In three to five sentences, describe the similarities and differences.**
A comparison of Figures 10, 12, and 14 shows that the Figure 10 is strongly skewed the the right. We find a binomial pattern, where the duration of days between 2 and 3, and between 6 and 7, is represented most. Only one observation is longer than 7 days, namely 10 days.
Figure 12, which uses five 10 samples of five days to compare their means, also shows a binomial pattern, where a duration between 3.5 and 4.0 days, and between 4.5 and 5.0 days, is represented most. This figure is strongly skewed to the left.
However, Figure 14, where we plot the sample means using sample groups of 10 observations, we find a nice pattern that looks similar to the normal distribution.
Therefore, we find that the distribution of a dataset can be manipulated by grouping observations into groups, and by the comparison of these groups.

2. **State the theoretical (according to the clt) distributions for the sample means.** a. When $n = 5$, we

find that, for four sample groups, the means are 4.6, 3.4, 4.8, 2.4, and the other group of observations has a mean of 3.45.
b. When $n = 10$, we find that, for four sample groups, the means are 3.4, 4.2, 2.6, 3.7, and for the other group of observations, the mean is 3.75.

3. **Are the sample means for $n = 5$ and $n = 10$ "close" to the theoretical mean, $\mu_x$? Explain why or why not.** For the overall population, we find that the theoretical mean is 3.57. For the mean of sample group 1, when $n = 5$, which is 4.6, and for the mean of sample group 1 when $n = 10$, which is 3.4, we thus find that, following Bernoulli's law of large numbers, the mean of a sample becomes closer to the theoretical mean when there are more observations.

4. **Which of the two distributions of sample means has the smaller standard deviation? Why?** The theoretical distribution has smaller standard deviation. This is because when the number of observations increase toward unlimited, the central limit theorem dictates that the samples very closely follow the normal distribution. Therefore, the mean of the samples becomes very close to the theoretical mean. In addition, the variance will decrease, which also results in a lower standard deviation.

5. **As $n$ changed, why did the shape of the distribution of the data change? Use one to two complete sentences to explain what happened.** When $n$ increased, there were more observations in the dataset. As there were more observations, following Bernoulli's law of large numbers, we were more likely to find a normal distribution. Therefore, it became more likely that the shape of the dataset was normally distributed. In other words, more observations makes that bias and variance in the dataset are reduced.

### 2.1.4 Explore the Melbourne Real-Estate Dataset

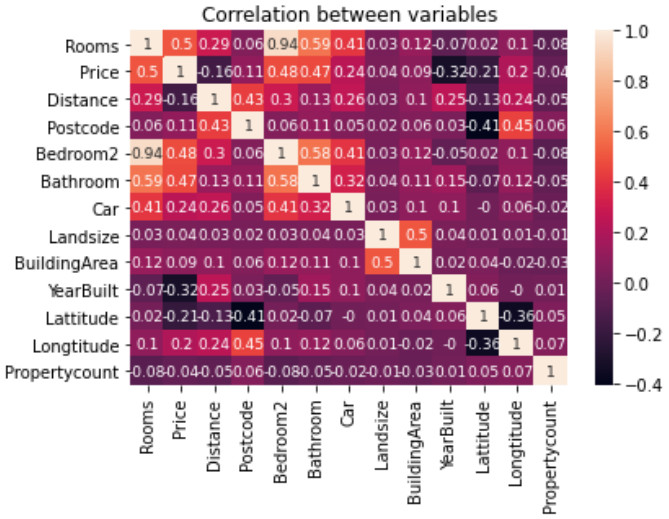


Figure 16: Heatmap showing the correlation between variables.

Table 4: Price of average housing per suburb ranked.

| Rank | Suburb | Avg. Price |
|---|---|---|
| 1 | Kooyong | 2185000 |
| 2 | Eaglemont | 1901000 |
| 3 | Albert Park | 1900000 |
| 4 | Canterbury | 1890000 |
| 5 | Middle Park | 1880000 |
| ... | ... | ... |
| 310 | Melton South | 390000 |
| 311 | Wallan | 366000 |
| 312 | Kurunjang | 353500 |
| 313 | Rockbank | 340000 |
| 314 | Bacchus Marsh | 285000 |

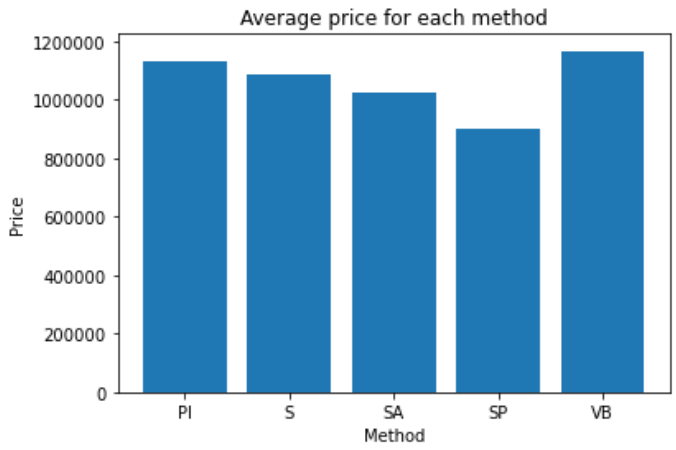


Figure 17: Bargraph showing the average price for each method.

## 2.2 Part B

First, we plot the data to get a visual understanding of the relation between property prices and the number of bedrooms. See Figure 18

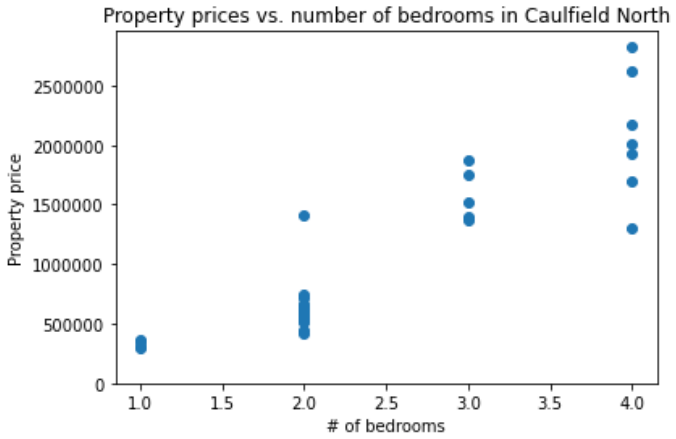


Figure 18: Scatterplot of property vs  of bedrooms

Subsequently, we run the Gradient Descent on the linear regression algorithm, and store the loss function, which is the Mean Squared Error, such that it can be plotted. In case the MSE no longer is optimized, after three times, we stop the Gradient Descent early. The history of minimizing the MSE is visualized in 19

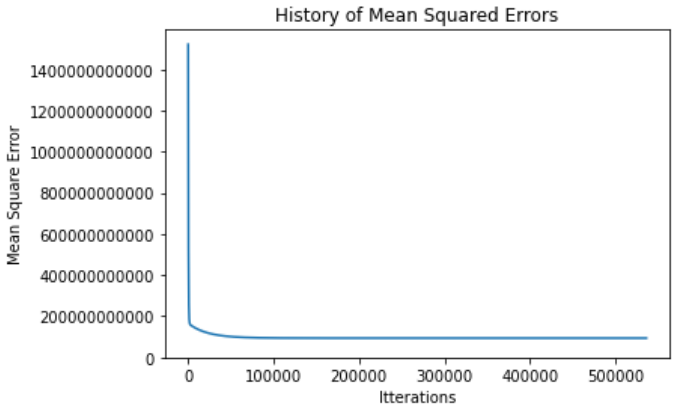


Figure 19: History of Minimizing Mean Squared Error

Consequently, at iteration 536.401, using a learning rate of 0.0001, we have minimized the Mean Squared Error.
Optimized Gradient Descent at:
The gradient (m) is: 668248.68
The intercept (c) is: -604246.37
MSE: 93777953573.12823

We also explored a learning rate of 0.0002, after 276900 iterations we get:
The gradient (m): 668249.11
The intercept (c): -604247.57
MSE: 93777953572.19083

Finally we explored a learning rate of 0.0003, after 187000 iterations we get:
The gradient (m): 668249.25
The intercept (c): -604247.94
MSE: 93777953571.97917

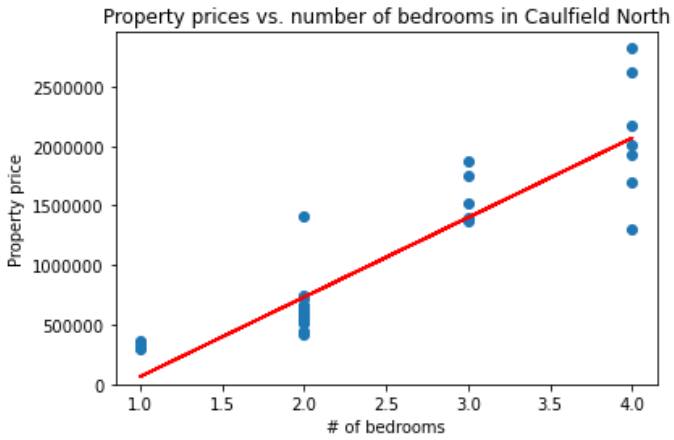


Figure 20: Scatterplot of property vs  of bedrooms with a plot representing the best model parameters.

# 3 Conclusion

## 3.1 Part A

### 3.1.1 Continuous Distribution

We find that for the continuous distribution, plotting a histogram using many bars and based on many records is most likely to result in a distribution that resembles the normal distribution.

### 3.1.2 Normal Distribution

Normal distribution is a great tool to use when wanting to, for this case, finding how some athletes compare to the rest of their peers.

### 3.1.3 Central Limit Theorem

Overall, we find that the central limit theorem shows that, when the number of observations in the dataset increase, that the shape of the dataset better resembles the normal distribution. This is consistent with Bernoulli's law of large numbers. Therefore, the mean of a large dataset is expected to be more closely to the theoretical mean, and the standard deviation in a large dataset will decrease.

### 3.1.4 Explore the Melbourne Real-Estate Dataset

To analyse the dataset we first decided to look at the correlation between all the variables. By creating a heatmap we can easily make correlations between variables stick out more. In figure 17 there is a lot of purple which indicates little to no correlation in most cases. Disregarding the purple, the biggest correlation we find is the number of rooms and

the number of bedrooms, which is pretty obvious. The most interesting correlation in my opinion is the one between the amount of rooms and the distance from Central Business District (CBD), which makes sense considering there should be more houses and thus more rooms in the outer parameters of the city. We can also see that the biggest variable to affect the price is the amount of rooms, bedrooms and bathrooms, but more interestingly the year the house was build which has a surprisingly high correlation. We can also see that houses with a high longitude and low latitude, which means houses situated in the south-west are the more expensive ones, and the ones in the north-east are the cheaper ones.
For the next data analysis we chose to rank the average price of the household for each suburb from most expensive to least expensive. There is not a lot to analyze from Table 3, but it would be helpful to know if for example someone is buying a home and wants to know where on the list they would be based on their budget.
For the last data analysis we decided to see how much the prices differed for the houses depending on what method were used. Here we find that the vendor bid (VB) has on average the highest selling price. This makes sense because a vendor bid is when the person who sells the house bids at an auction with not intention of buying it, to artificially increase the highest bid. When the property is sold prior (SP) to the auction it has the lowest price, which also makes sense because then it is more often a single buyer who will try to lowball the price before the auction directly with the seller without any competition in the form of other bidders who can make a counter offer. Therefore the property sold (S) at auction bar being situated between the two aforementioned methods on the y axis also makes sense. Overall we find this graph very informative.

### 3.2 Part B

We have illustrated that Gradient Descent can be used to minimize the loss function, Mean Squared Error, for linear regression. We find that using a learning rate of 0.0001, after 536 401 iterations, the lowest Mean Squared Error was found. We also explored different learning rates, and found that the lower the learning rate the more iterations were made which led to a lower MSE. We therefore decided to plot the cost function with a learning rate of 0.0001, which returned:

$$PropertyPrice = numberOfBedrooms * 668248.68 + (-604246.37)$$

With a MSE of 93777953573.12823, which is where the gradient descent function stops at when the MSE no longer reduces.

## References


[1] "Continuous Distribution.". https://openstax.org/books/introductory-statistics/pages/5-4-continuous-distribution. Accessed: 2021-01-30.

[2] "Normal Distribution.". https://openstax.org/books/introductory-statistics/pages/6-3-normal-distribution-lap-times. Accessed: 2021-02-02.

[3] "Central Limit Theorem.". https://openstax.org/books/introductory-statistics/pages/7-5-central-limit-theorem-cookie-recipes. Accessed: 2021-02-02.

[4] "Melbourne Realestate Dataset.". https://www.kaggle.com/dansbecker/melbourne-housing-snapshot. Accessed: 2021-02-02.

[5] "Normal Distribution Lap Times Table.". https://openstax.org/books/introductory-statistics/pages/c-data-sets. Accessed: 2021-02-21.