University of Texas at Arlington

**IE 3301 Spring 2022 Project**

An Analysis of movie runtimes and crosswalk wait periods

Landon Moon

IE-3301

Dr. LeBoulluec

Due. April 25th

# Introduction:

The goal of this project is to collect data from a normally distributed data set an from a inter-interval dataset and apply multiple probability and stats concepts to each. The first data set (normal) will be the runtime of movies to help determine whether they follow a pattern, have a most common length, are a good fit for a normal distribution. The second data set (inter-interval) will be the time between Crossings times made by the crosswalk across Cooper Street near UTA. By applying probability and stats concepts we can determine the busiest times during the day, the average amount of traffic, and determine whether the data matches with an exponential distribution.

This data can then be tested through a chi-squared test to determine whether they match their respective distributions. Variations in the data can show us trends in the two data sets that might have otherwise gone unnoticed.

Table of Contents

[Project Part I 3](#_Toc101377497)

[Data Set 1 (Normal): 3](#_Toc101377498)

[Data Collection: 3](#_Toc101377499)

[Descriptive Statistics: 3](#_Toc101377500)

[Report: 4](#_Toc101377501)

[Data Set 2 (Inter-Interval): 6](#_Toc101377502)

[Data Collection: 6](#_Toc101377503)

[Descriptive Statistics: 6](#_Toc101377504)

[Report: 7](#_Toc101377505)

[Project Part II 9](#_Toc101377506)

[Data Set 1 (Normal): 9](#_Toc101377507)

[Chi-Square Goodness-of-Fit Test: 9](#_Toc101377508)

[Report: 9](#_Toc101377509)

[Data Set 2 (Inter-Interval): 10](#_Toc101377510)

[Chi-Square Goodness-of-Fit Test: 10](#_Toc101377511)

[Report: 10](#_Toc101377512)

[Appendices 11](#_Toc101377513)

[Data Set 1: 11](#_Toc101377514)

[Data Set 2: 11](#_Toc101377515)

[Formulas: 11](#_Toc101377516)

# Project Part I

## Data Set 1 (Normal):

### Data Collection:

Definition of Variable:

The runtime of the top grossing movies of all time in minutes.

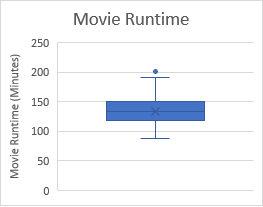
Data Collection:

Collected from the list on boxofficemojo.com and runtimes are collected from imdb.com.

Data Set:

Data contained in the “Data Set 1.xlsx” excel sheet. This file Is contained in the zip file submitted. Data can also be found in the appendix.

### Descriptive Statistics:

Sample Mean: 132.8952 (Minutes) Box-and-whisker plot:

Sample Standard Deviation: 24.9229 (Minutes)

Quartiles:

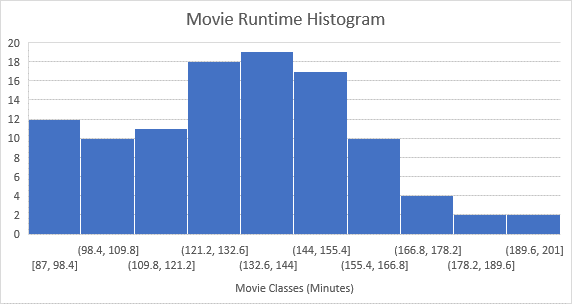
1. 87-118 (Minutes)
2. 119-133 (Minutes)
3. 134-149 (Minutes)
4. 150-201 (Minutes)

Frequency table:

|  |  |  |  |
| --- | --- | --- | --- |
| Movie Runtime | Frequency | Relative Frequency | Class Midpoint |
| 87-98 | 12 | 11.42% | 92.5 |
| 99-109 | 10 | 9.52% | 104 |
| 110-121 | 11 | 10.48% | 115.5 |
| 122-132 | 18 | 17.14% | 127 |
| 133-144 | 19 | 18.10% | 138 |
| 145-155 | 17 | 16.19% | 150 |
| 156-166 | 10 | 9.52% | 161 |
| 167-178 | 4 | 3.81% | 172.5 |
| 179-189 | 2 | 1.90% | 184 |
| 190-201 | 2 | 1.90% | 195.5 |

N=105

Frequency Histogram:



### Report:

Data:

Going to www.boxofficemojo.com/chart/top\_lifetime\_gross/?area=XWW gives a list of the top grossing movies of all time. By using this list and cross referencing the imdb.com page for each movie, I could determine the runtime of the top ~100 movies. This data was collected and put into an excel sheet/

Descriptive Statistics:

Overall, the set of movies seem to be normally distributed that is skewed right. That means that there are more movies that have shorter run times than a normal distribution would predict. This seems to be because of two major factors: animation, and attention span.

Animated movies such as Frozen, Minions, Zootopia, Despicable Me 2, Etc. all have runtimes less than 2 hours. This is likely because animation requires a fair amount of work even for extremely simple scenes. For example, A real life movie is able to have a panning shot of a city or landscape adding multiple seconds to the movie without any need to model the city or landscape itself. A animation on the other hand would most likely not have this scene unless it was already going to be used. This barrier to entry gets rid of multiple scenes that would otherwise probably exist. Overall, shorter movies are easier to make whether they’re animated or not.

Secondly, the attention span of the average viewer prefers shorter movies with a denser story versus a longer movie that takes multiple hours to watch. For almost all of the movies, people would have to go to the movie theater. People don’t want to stay out of the house all day/night and would rather have a movie they can enjoy then get going with the rest of their day. The addition of at-home movies with streaming services might make longer movies more popular.

## Data Set 2 (Inter-Interval):

### Data Collection:

Definition of Variable:

The time between available times to cross the cross walk at the intersection of Cooper Street and UTA Blvd. Specifically the south crossing going from east to west UTA.

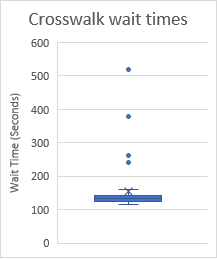
Data Collection:

Sat near the crosswalk recording times between crossing times using the stopwatch app on my phone. This only gives lap time and total time. Clock time is back-calculated from the final clock time. Laps are recorded within the beginning and ending walk periods giving some variance to recorded values.

Data Set:

Data contained in the “Data Set 2.xlsx” excel sheet. This file Is contained in the zip file submitted. Data can also be found in the appendix.

### Descriptive Statistics:

Sample Mean: 154.44 (Seconds) Box-and-whisker plot:

Sample Standard Deviation: 66.42 (Seconds)

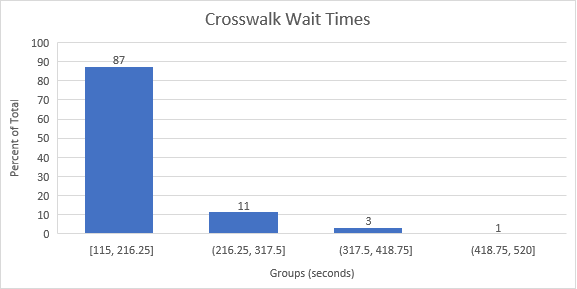
Quartiles:

1. 115-124
2. 125-132
3. 133-143
4. 144-520

Frequency table:

|  |  |  |  |
| --- | --- | --- | --- |
| Wait Time (Seconds) | Frequency | Relative Frequency | Class Midpoint |
| 115-216 | 87 | 87% | 108 |
| 217-317 | 11 | 11% | 267 |
| 318-418 | 3 | 3% | 368 |
| 419-520 | 1 | 1% | 469.5 |

Frequency Histogram:



### Report:

Data:

I sat right behind the UT Arlington sign at the intersection of Cooper Street and UTA Blvd so I could have a good view of the cross walk without being in view of all the cars. Each time the crosswalk allowed people to cross I would press the lap button on my phone’s stopwatch app. Because of the long time needed to collect all the data, I couldn’t collect all the data right as the crosswalk started but instead some data points are a couple seconds behind. Not all data was collected in the same timespan because of the lengthy amount of time needed to gather the data.

Descriptive Statistics:

The Crosswalk seems to be set up to go in a loop giving multiple times around 2 minutes. If there were no people present the crosswalk’s light wouldn’t change and instead let cars go early. This causes multiple spikes in interval times around every 2 minutes such as 2, 4, 6, and 8 minutes. The frequency of the longer wait times decreases because most of the time there are people waiting to cross. This pattern generally follows an exponential distribution because each time the loop comes around the probability that it gets doesn’t get reset is multiplied by the current probability. For example, the probability of the time taking 3 crosswalk loops would be P\*P\*P or P^3 while the probability for 4 crosswalk loops would be P\*P\*P\*P or P^4.

An odd part of the data is that crosswalk times that take one loop have a slight deviation in times. While sometimes it takes 115 seconds to allow people to cross, it can also take up to ~150 seconds. This deviation is most likely because of two reasons: safety and efficiency. First allowing people to cross the crosswalk while some late cars are still in the intersection would be dangerous, while waiting a few seconds would make is safer to cross. Secondly, sometimes during high congestion the intersection could get more cars through by allowing longer green lights. This speeds up the number of cars that can get through the intersection but makes longer wait times for cars and more importantly, people waiting on the crosswalk. This is probably done on the fly with pressure sensors or cameras to detect where cars are, so that no one must wait longer than they need to.

# Project Part II

## Data Set 1 (Normal):

### Chi-Square Goodness-of-Fit Test:

Expected Frequency Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Movie Runtime | Observed Frequency | Class Probability | Expected Frequency | Difference | Class Component |
| 87-98 | 12 | 11.42% | 5.03 | 6.97 | 9.658 |
| 99-109 | 10 | 9.52% | 8.60 | 1.4 | 0.227 |
| 110-121 | 11 | 10.48% | 14.43 | -3.43 | 0.815 |
| 122-132 | 18 | 17.14% | 16.24 | 1.76 | 0.191 |
| 133-144 | 19 | 18.10% | 17.89 | 1.11 | 0.069 |
| 145-155 | 17 | 16.19% | 13.23 | 3.77 | 1.074 |
| 156-166 | 10 | 9.52% | 8.91 | 1.09 | 0.133 |
| 167-178 | 4 | 3.81% | 5.29 | -1.29 | 0.314 |
| 179-189 | 2 | 1.90% | 2.10 | -0.10 | 0.005 |
| 190-201 | 2 | 1.90% | .82 | 1.18 | 1.698 |

Sample Mean: 132.8952 (Minutes)

Degree of Freedom: 10-1 = 9

Test Statistic: = 14.186

Chi-Squared Value: a=.05, v=9 -> 16.919

14.186 < 16.919 so the data is withing the level of significance.

### Report:

Even though there is a bias for movies to have a shorter runtime, the Chi-Squared test still says that the distribution of movies runtimes is normally distributed within a significance level of .05. Most of the test statistic comes from the shortest runtime movies. The first class of movie runtimes have a class component of 9.658 while the total is 14.186. this means that the shortest movies account for more than half of the variance while only accounting for eleven percent of the total.

In conclusion, The data set matches a normal distribution with a significance level of .05 but is close to not being valid. Almost all of this observed variance is due to a bias for movies to have a shorter runtime.

## Data Set 2 (Inter-Interval):

### Chi-Square Goodness-of-Fit Test:

Expected Frequency Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Wait Time (Seconds) | Observed Frequency | Class Probability | Expected Frequency | Difference | Class Component |
| 0-216 | 87 | 87% | 75.31 | 11.69 | 1.815 |
| 217-317 | 11 | 11% | 11.69 | -0.69 | 0.041 |
| 318-418 | 3 | 3% | 6.08 | -3.08 | 1.560 |
| 419-520 | 1 | 1% | 3.23 | -2.23 | 1.539 |

Sample Mean: 154.44 (Seconds)

Degree of Freedom: 4-1 = 3

Test Statistic: = 4.96

Chi-Squared Value: a=.05, v=3 -> 7.815

4.96 < 7.815 so the data is within the level of significance.

### Report:

This result is expected because the more time one waits for a person to press the button on the crosswalk, it is more likely for someone to arrive. This is representative of a negative binomial distribution which is in the exponential family of distributions. In my data set I changed the wait time for the quickest interval to include the time between 0 and the minimum value. I believe this is an appropriate change because my data isn’t perfectly continuous. This is because the timer for the crosswalk occurs approximately every two minutes with gaps before and after these times.

In conclusion, the data set is mostly exponential due to it following a negative binomial distribution. It is within a significance level of .05 of the expected exponential distribution but would be better described by a negative binomial distribution. There is no outstanding bias left or right in the data.

# Appendices

## Data Set 1:

Runtime of movies (minutes):

Total=105

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 162 | 181 | 191 | 138 | 149 | 148 | 124 | 118 | 143 | 137 |
| 103 | 141 | 134 | 130 | 152 | 128 | 102 | 129 | 118 | 136 |
| 130 | 91 | 157 | 153 | 201 | 129 | 123 | 154 | 143 | 165 |
| 127 | 164 | 122 | 141 | 100 | 103 | 151 | 88 | 133 | 128 |
| 136 | 89 | 97 | 136 | 108 | 108 | 152 | 169 | 152 | 146 |
| 98 | 106 | 119 | 144 | 169 | 161 | 178 | 138 | 100 | 153 |
| 93 | 134 | 176 | 178 | 157 | 139 | 94 | 148 | 133 | 161 |
| 88 | 87 | 152 | 123 | 140 | 146 | 121 | 95 | 112 | 130 |
| 148 | 149 | 115 | 121 | 141 | 87 | 145 | 132 | 93 | 105 |
| 123 | 142 | 129 | 115 | 147 | 158 | 122 | 127 | 130 | 119 |
| 108 | 121 | 163 | 121 | 159 |  |  |  |  |  |

## Data Set 2:

Wait Times for Crosswalk (Seconds):

Total=100

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 116 | 140 | 116 | 137 | 262 | 138 | 136 | 382 | 127 | 135 |
| 129 | 130 | 124 | 135 | 520 | 148 | 124 | 122 | 146 | 121 |
| 124 | 127 | 146 | 115 | 143 | 127 | 129 | 133 | 134 | 140 |
| 262 | 125 | 120 | 132 | 142 | 119 | 159 | 142 | 242 | 132 |
| 123 | 141 | 126 | 525 | 128 | 132 | 131 | 130 | 124 | 135 |
| 250 | 378 | 124 | 122 | 146 | 116 | 124 | 253 | 146 | 115 |
| 143 | 127 | 120 | 143 | 119 | 132 | 133 | 132 | 242 | 118 |
| 128 | 127 | 159 | 142 | 131 | 132 | 123 | 141 | 126 | 252 |
| 128 | 132 | 253 | 130 | 124 | 135 | 131 | 378 | 124 | 122 |
| 146 | 116 | 124 | 253 | 146 | 115 | 143 | 133 | 131 | 143 |

## Formulas:

All formulas used can be seen in the provided excel sheets included in submission. Mean uses AVERAGE(), SD uses STDEV.S(), quartiles use QUARTILE() along with using MIN() and MAX().