A PROJECT REPORT ON

PROJECT 2

EXPERIENTIAL DATA COLLECTION

AND

DESCRIPTIVE STATISTICAL ANALYSIS

BY

1002071263 ANUSHKA CHANDRAKANT PAWAR

UNDER THE GUIDANCE OF

PROF DR. EMMA YANG

OCTOBER 20, 2022

"I __**Anushka Chandrakant Pawar**___ did not give or receive any assistance on this project, and the report submitted is wholly my own."

INDEX

INTRODUCTION

In this Project, we implemented the concepts we learned in class. We worked on Real-world Data Sets and learned to solve real-world problems. The objective of this project was to perform Chi-square testing for the Goodness of fit test on the Data Sets we collected in Project 1. We used various formulas in Excel to perform this analysis, which are mentioned in the Appendix section.

DATA COLLECTION

Dataset 1:

- After a quick search via websites like the CERN open data portal, Datahub.io, and the UCI Machine Learning Repository, we came to a decision on the online dataset which we took from the Kaggle website.
- The references contain a link to the online dataset.
- The dataset provides data from 2013 for the temperature and humidity in Austin, Texas.
- From the full dataset, we have selected the column with the highest Temperature values.
- 116 values were used as our observations, the additional statistical analysis was finished and further Hypothesis testing was carried out on it.

Dataset 2:

- On Friday, October 7, 2022, from 2:00 PM to 3:30 PM, we had a data collection time window.
- Four members of the group observed people coming into the library and leaving in four different directions (e.g., up from the elevator, down from the elevator, left towards Studio and Tech Learning, or right toward FabLab/Einstein Bros. Bagels).
- We measured the precise lap intervals between people going on different sides using a stopwatch.
- I used the observations of people going in up from the Elevator.
- 99 values were used as our observations, the additional statistical analysis was finished and further Hypothesis testing was carried out on it.

STATISTICAL ANALYSIS

Dataset 1:

1. Sample Mean= 67.4741
2. Sample Median= 70
3. Quartiles: Q1= 59, Q2= 70 , Q3= 76.25
4. Class Range= 14.25 ~15
5. Standard Deviation= 12.771

Tabular Summary:

Class Interval	Frequency	Relative Frequency	Cumulative Relative Frequency
[30,45)	7	0.060344828	0.060344828
[45,60)	23	0.198275862	0.25862069
[60,75)	47	0.405172414	0.663793103
[75,90)	39	0.336206897	1

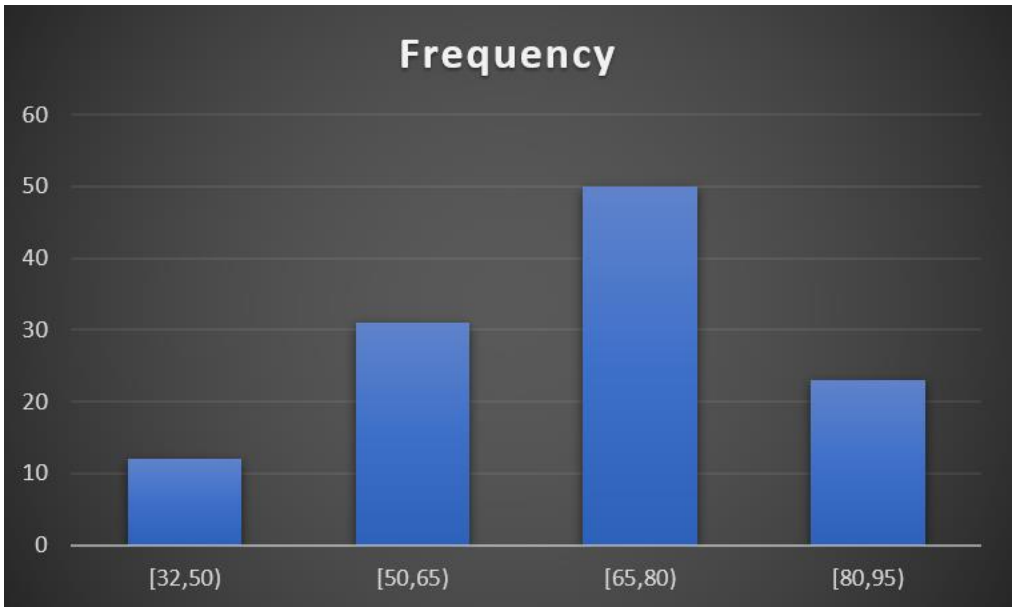
Data Set 2:

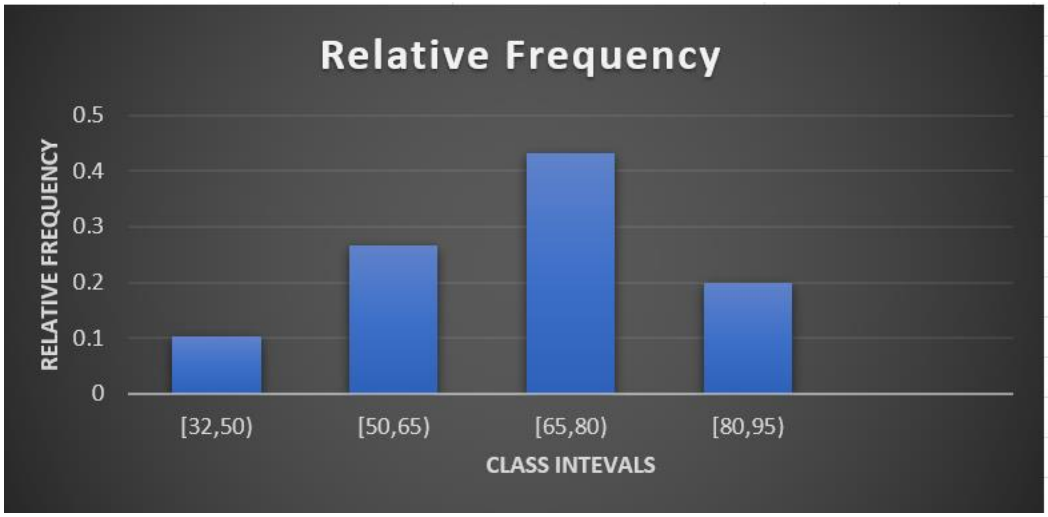
1. Sample Mean= 32.8830
2. Sample Median= 23.78
3. Q1= 6.095, Q2= 23.78, Q3= 51.025.
4. Class Range= 26
5. Standard Deviation= 31.6194

Class Interval	Frequency	Relative Frequency	Cummulaative Relative Frequen
[0,26)	50	0.505050505	0.505050505
[26,52)	25	0.252525253	0.757575758
[52,78)	15	0.151515152	0.909090909
[78,104)	7	0.070707071	0.97979798
[104,130)	1	0.01010101	0.98989899
[130,156)	0	0	0.98989899
[156,182)	1	0.01010101	1

HYPOTHESIS TESTING (Goodness of Fit) OF DATA SET 1

Histogram:





According to the Histograms above, I structured Data classes and merged or split the classes to ensure that each class had sufficient observations.

Excel Calculation process:

- Started by getting necessary data classes.
- Then I noted the Observed frequency from the observed data.
- After that, I calculated the class probability (By using NORM.DIST(X,Mean,Standard Deviation,TRUE)).
- Then, I calculated the expected frequency (Class Probability/n).
- Finally, I calculated the Chi-square values by the given formula [(Oi-ei)^2/ei].

Class Intervals	Obeserved Frequency	Class Probability	Class expected Frequency	Chi-Square Class Component
X<=50	13	0.085618074	9.931696618	0.9479232
50<X<=65	35	0.337577556	39.15899644	0.441718455
65<X<=80	48	0.413456886	47.96099877	3.17153E-05
X>80	20	0.163347484	18.94830817	0.058372267
Total	116	1	116	1.448045637

Class Probability(X<=50)= NORM.DIST(50,67.4741,12.7712,TRUE)

Class Probability(50<X<=65)= NORM.DIST(65,67.4741,12.7712,TRUE)-NORM.DIST(50,67.4741,12.7712,TRUE)

Class Probability(50<X<=65)= NORM.DIST(80,67.4741,12.7712,TRUE)-NORM.DIST(65,67.4741,12.7712,TRUE)

Class Probability(X>80)= 1-NORM.DIST(80,67.4741,12.7712,TRUE)

Total Interval: 4

Alpha: 0.05

Degree of Freedom: 3

HYPOTHESIS TESTING:

Ho= The Data follows Exponential Distribution

H1=The Data does not follow Exponential Distribution

Decision Rule: Reject Ho if Calculated Chi-square >Tabular Chi-square.

Calculated Chi-square value=1.4480

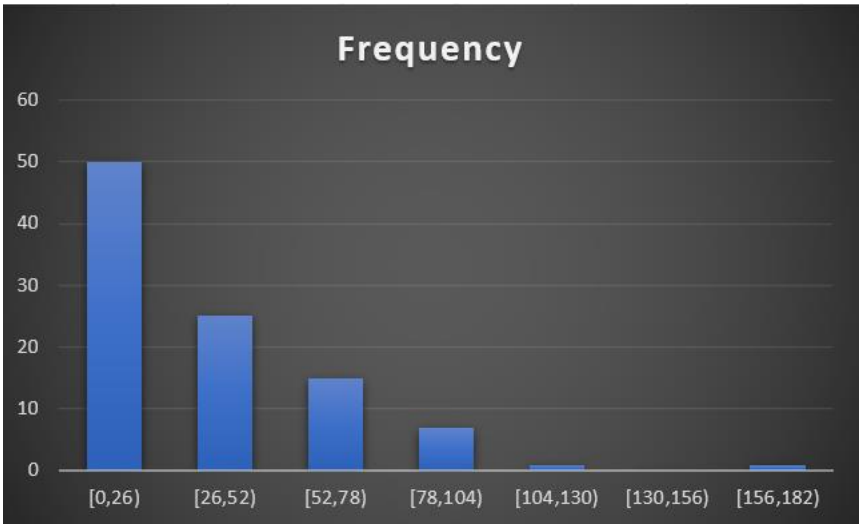
Tabular Chi-square value= 7.8147

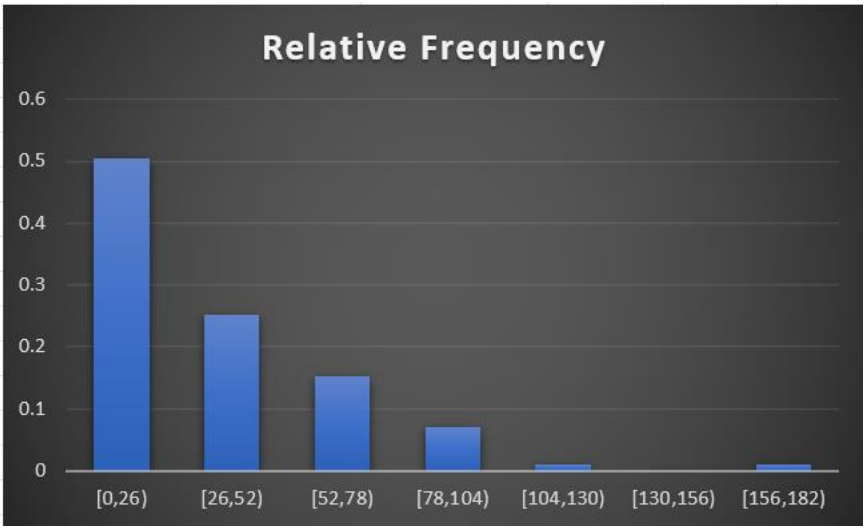
CONCLUSION:

As the calculated Chi-square is less than the tabular Chi-square, we fail to reject hypothesis Ho. Hence the Data follows Normal Distribution.

HYPOTHESIS TESTING (Goodness of Fit) OF DATA SET 2

Histogram:





According to the Histograms above, I structured Data classes and merged or split the classes to ensure that each class had sufficient observations.

Excel Calculation Process

- Started by getting necessary data classes.
- Then I noted the Observed frequency from the observed data.
- After that, I calculated the class probability (By using GAMMADIST(X,1, mean,1)).
- Then, I calculated the expected frequency (Class Probability/n).
- Finally, I calculated the Chi-square values by the given formula [(Oi-ei)^2/ei].

Class Intervals	Obeserved Frequency	Class Probability	Class expected Frequency	Chi-Square Class Component
X<26	50	0.546464674	54.10000269	0.31072128
26<X<=52	25	0.247841034	24.53626238	0.008764684
52<X<=78	16	0.112404664	11.12806176	2.132966431
X>78	8	0.093289628	9.235673167	0.165325055
Total	99	1	99	2.61777745

Class Probability(X<26) = GAMMADIST(26, 1, 32.883, 1)

Class Probability(26<X<=52) = GAMMADIST(52, 1, 32.883, 1)-GAMMADIST(26, 1, 32.883, 1)

Class Probability(52<X<=78)=GAMMADIST(78, 1, 32.883, 1)-GAMMADIST(52, 1, 32.883,1)

Class Probability(X>78) = 1-GAMMADIST(78,1,32.883,1)

Total Interval: 4

Alpha: 0.05

Degree of Freedom: 3

HYPOTHESIS TESTING:

Ho= The Data follows Exponential Distribution

H1=The Data does not follow Exponential Distribution

Decision Rule: Reject Ho if Calculated Chi-square > Tabular Chi-square.

Calculated Chi-square value=1.318166847

Tabular Chi-square value= 7.8147

CONCLUSION:

As the calculated Chi-square is less than the tabular Chi-square, we fail to reject hypothesis Ho. Hence, the Data follows Exponential Distribution.

Appendix

- Data Set 1 is taken from the website: https://www.kaggle.com/datasets/grubenm/austin-weather
- The link for Hypothesis Testing Calculations and the Data Set:

  https://docs.google.com/spreadsheets/d/1XY-UfjhVnU_d99gElT9pZrqD1FusIwE0qe1UG64_mok/edit?usp=sharing

- Formulas used for calculations:
  Observed Frequency= COUNTIFS(Datapoints,"<=X")
  Expected Frequency= Class probability*n
  Chi-square Component= (observed frequency- expected frequency) ^2/expected freq
  Degree of Freedom= Total intervals-1
  Alpha=0.05
  Class Probability of Exponential Distribution= GAMMADIST(X,1, mean,1)).
  Class Probability of Normal Distribution=NORM.DIST(X,Mean,Standard Deviation,TRUE).

REFERENCES:

- https://www.wunderground.com/history/weekly/us/tx/austin
- https://libraries.uta.edu/services