Name: Temitayo Aderounmu

UTA ID: 1001568524

IE 3301 Fall 2021 Project

Project Report

Due Date: 11/29/2021

“I __Temitayo Aderounmu__ did not give or receive any assistance on this project, and the report submitted is wholly my own.”

X ____

Contents

Name: Temitayo Aderounmu

UTA ID: 1001568524

IE 3301 Fall 2021 Project

Part I

Due Date: 11/29/2021

“I __Temitayo Aderounmu__ did not give or receive any assistance on this project, and the report submitted is wholly my own.”

X ____

IE3301: Set 1-- Descriptive Statistic

The following data set refers to the height in centimeters of a group of adult men, living in the region of the western edge of the Kalahari Desert.

Data Set:
152 157 164 169 165 151 163 157 122 161
130 146 137 167 148 162 161 152 163 171
157 154 147 157 137 166 164 161 152 164
140 167 149 166 161 162 161 159 158 157
179 170 158 144 160 167 159 133 133 163
132 136 170 159 161 170 159 98 147 162
163 100 163 162 163 142 171 157 149 134
157 136 167 159 161 160 161 155 160 167
159 162 159 162 159 173 166 142 128 129
157 161 124 121 163 159 159 158 148 157
157 108 168 149 155 163 161 166 157 164

n = COUNT =110

$\sum x_i$ = SUM = 16978

Measure of Location:

Sample Mean = Average = 154.3455
Geometric Mean = GEOMEAN = 153.5518
Sample Median = MEDIAN = 159
Sample Mode = MODE = 157

Measure of Variability:

Sample Range = MAX-MIN = 81
Sample Variance = VAR = 217.2007
Sample Standard Deviation=STDEV= 14.74
Coefficient of Variation (%) = s/x = 9.55

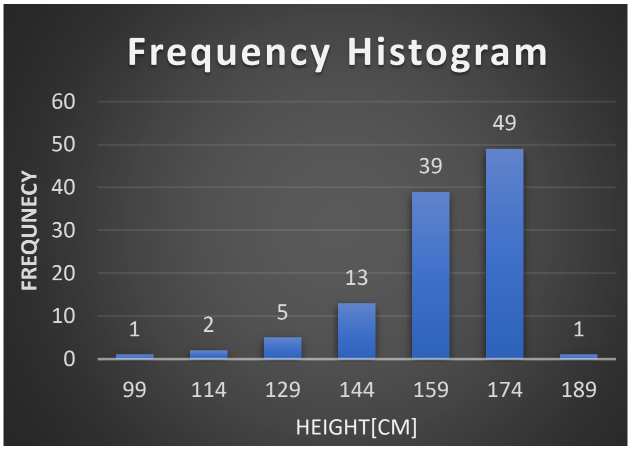
Percentiles:

21st Percentile = PERCENTILE (case,0.21) = 147
First Decile = 10th Percentile = PERCENTILE (case,0.10) = 133
Q1 = First Quartile = 25th Percentile = PERCENTILE (case,0.25) = 149
Q2 = Second Quartile = 50th Percentile = Sample Median = PERCENTILE (case,0.50) = 159
Q3 = Third Quartile = 75th Percentile = PERCENTILE (case,0.75) = 163

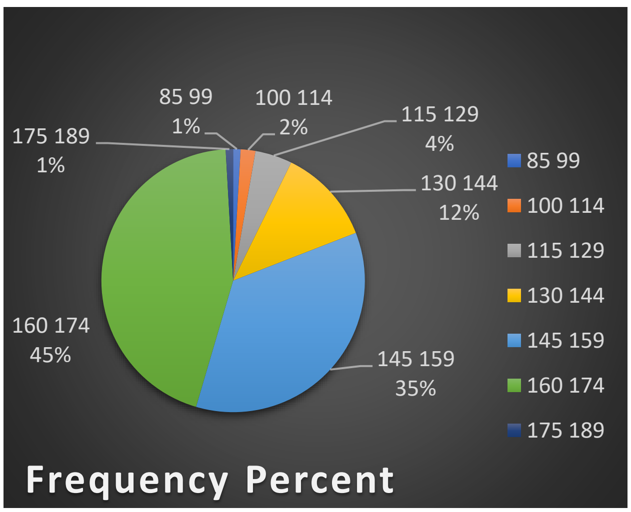
Tabular Summary:

Class Interval Tally	**Count or Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
$85 \leq X < 99$	1	0.009091	0.009091
$100 \leq X < 114$	2	0.018182	0.027273
$115 \leq X < 129$	5	0.045455	0.072727
$130 \leq X < 144$	13	0.118182	0.190909
$145 \leq X < 159$	39	0.354545	0.545455
$160 \leq X < 174$	49	0.445455	0.990909
$175 \leq X < 189$	1	0.009091	1

Histogram (Bar Chart):



Pie Chart:



Boxplot:

Q1 = 149
Q2 = 159
Q3 = 163
IQR = Q3 - Q1= 14

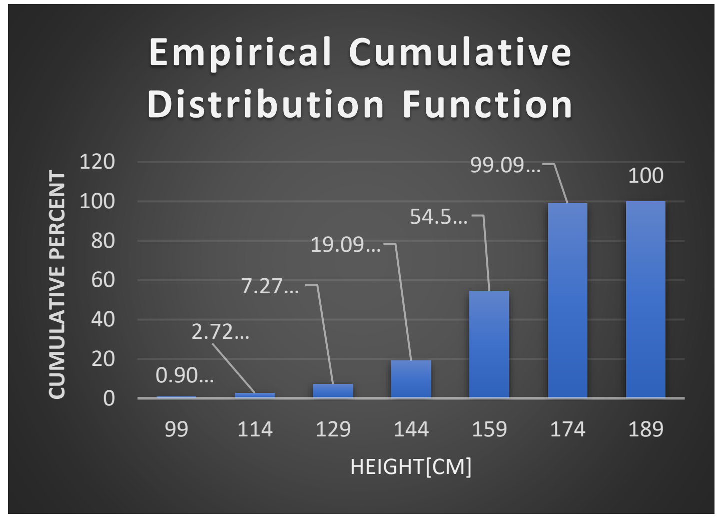
<u>Lower Whisker:</u>

Q1 – 1.5*IQR = 128

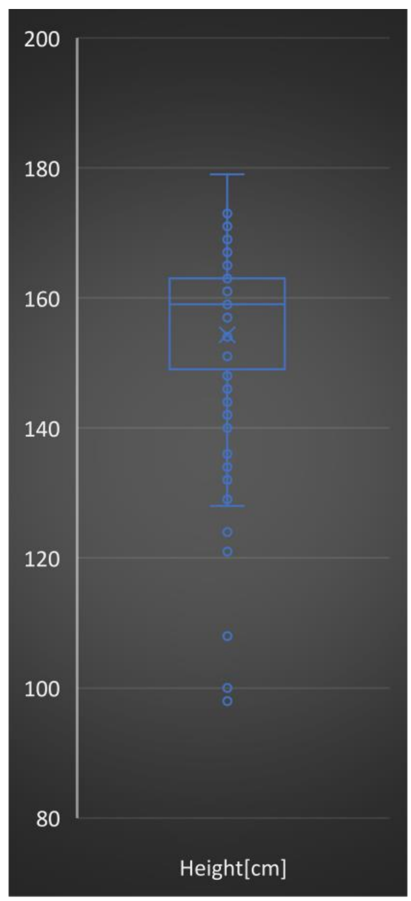
<u>Upper Whisker:</u>

Q3 + 1.5*IQR = 184

Empirical Cumulative Distribution Function:



Boxplot:



IE3301: Set 2-- Descriptive Statistic

The following data set includes the arrival time and inter arrival time of customers at an ATM service point in minutes. The data was collected from the *ATM service point of the Fidelity Bank Plc, North Central on 21st May 2019.*

n = COUNT =100

$\sum x_i$ = SUM = 217

Measure of Location:

Sample Mean = Average = 2.1919
Geometric Mean = GEOMEAN = 1.9064
Sample Median = MEDIAN = 2
Sample Mode = MODE = 2

Measure of Variability:

Sample Range = MAX-MIN = 7
Sample Variance = VAR = 1.56483
Sample Standard Deviation=STDEV= 1.251
Coefficient of Variation (%) = s/x = 57.07

Data Sets:

ART	INT	ART	INT	ART	INT	ART	INT	ART	INT
1	-	44	2	91	3	134	3	176	1
3	2	45	1	92	1	136	2	178	2
4	1	48	3	94	2	138	2	180	2
5	1	51	3	97	3	140	2	184	4
6	1	55	4	99	2	141	1	186	2
11	5	58	3	101	2	142	1	189	3
12	1	59	1	102	1	145	3	191	2
15	3	60	1	104	2	149	4	193	2
17	2	61	1	107	3	151	2	195	2
19	2	69	8	111	4	152	1	196	1
24	5	72	3	116	5	154	2	201	5
25	1	73	1	117	1	157	3	204	3
26	1	74	1	118	1	163	6	205	1
28	2	76	2	121	3	164	1	207	2
31	3	79	3	123	2	166	3	209	2
34	3	81	2	125	2	169	2	211	2
36	2	84	3	127	2	171	1	214	3
39	3	85	1	128	1	172	1	216	2
40	1	86	1	130	2	173	2	218	2
42	2	88	2	131	1	175	1	219	1

ART = Arrival Time in minutes, INT = Inter Arrival Time in minutes

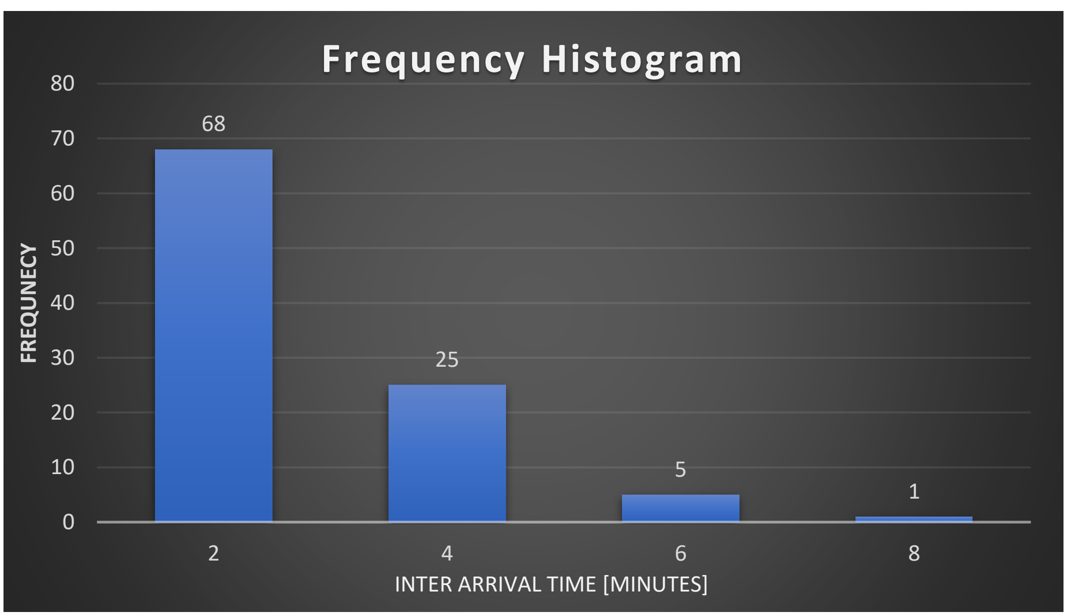
Percentiles:

21^{st} Percentile = PERCENTILE (case,0.21) = 1
First Decile = 10^{th} Percentile = PERCENTILE (case,0.10) = 1
Q1 = First Quartile = 25^{th} Percentile = PERCENTILE (case,0.25) = 1
Q2 = Second Quartile = 50^{th} Percentile = Sample Median = PERCENTILE (case,0.50) = 2
Q3 = Third Quartile = 75^{th} Percentile = PERCENTILE (case,0.75) = 3

Tabular Summary:

Class Interval Tally	Count or Frequency	Relative Frequency	Cumulative Relative Frequency
$1 \leq X < 3$	68	0.68	0.68
$3 \leq X < 5$	25	0.25	0.93
$5 \leq X < 7$	5	0.05	0.98
$7 \leq X < 9$	1	0.01	0.99

Histogram (Bar Chart):



Boxplot:

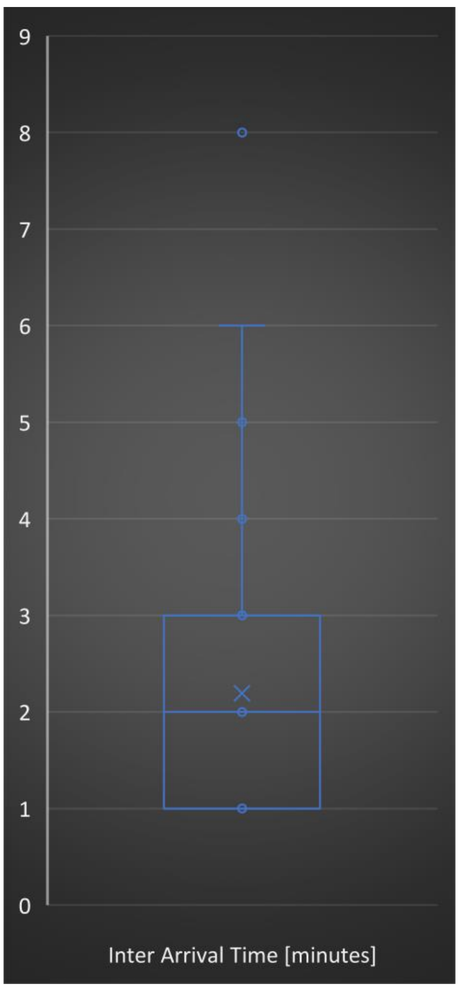
Q1 = 1
Q2 = 2
Q3 = 3
IQR = Q3 - Q1= 2

<u>Lower Whisker:</u>

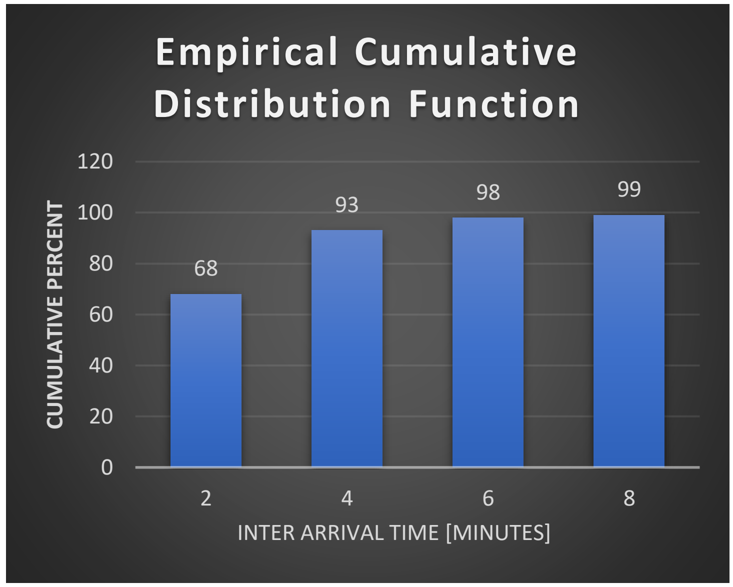
Q1 – 1.5*IQR = 2

<u>Upper Whisker:</u>

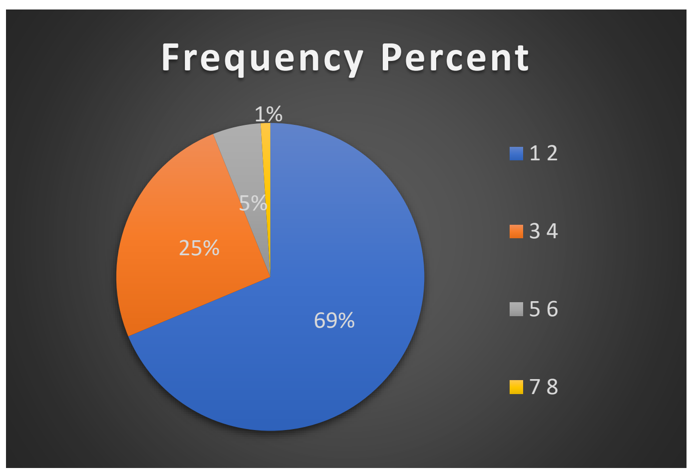
Q3 + 1.5*IQR = 6



Empirical Cumulative Distribution Function:



Pie Chart:



Name: Temitayo Aderounmu

UTA ID: 1001568524

IE 3301 Fall 2021 Project

Part II

Due Date: 11/29/2021

“I __Temitayo Aderounmu__ did not give or receive any assistance on this project, and the report submitted is wholly my own.”

X

IE3301: Set 1— Chi-square Goodness-of-Fit Test

Significance level of 0.05

Hypothesis: Set 1 is sampled from a Normal Distribution with a population mean equal to the sample mean, and a population standard deviation equal to the sample standard deviation.

n = observations = 110

Sample Mean: 154.345
Sample Standard Deviation: 14.74
Chi-square formula:

$$X^2 = \sum_{i=0}^{i} \frac{(o_i - e_i)^2}{e_i}$$

Z formula:

$$z = \frac{x - \mu}{s}$$

P(X<=X) formulas<excel>:
-NORM.S.DIST(z,cumulative)
-FREQUENCY(data_arrays,bins_arrays)

Expected Frequency Table:

Interval	**Observed Frequency**	**Class Probability**	**Expected Frequency**	$\boldsymbol{X^2}$ **Class Component**
$130 \leq X < 144$	21	0.241349	26.54836	1.159555
$145 \leq X < 159$	39	0.382584	42.0842	0.226031
$160 \leq X < 174$	49	0.284904	31.33939	9.952237
$175 \leq X < 189$	1	0.081813	8.999432	7.11055
Total	110	0.990648966	108.9714	18.44837

Goodness of Fit Test for Normal Distribution:

H_0: The observed distribution of Set 1 may not be fitted by the normal distribution.

H_1: The observed distribution of Set 1 may be fitted by the normal distribution.

Degree of Freedom: df = n – 1 = 4 – 1 = 3

Statement: Using the X^2 distribution tables, at 5% level of significance for 3 degrees of freedom, the critical value is 7.815

Conclusion: if X > 7.815 and X< -7.815, then the null hypothesis should be rejected.

It is observed that the calculated value (18.45) is greater than the critical value (7.815). The null hypothesis should be rejected, as there is insufficient evidence to conclude that the observed distribution does not fit normal distribution.

IE3301: Set 2— Chi-square Goodness-of-Fit Test

Significance level of 0.05

Hypothesis: Set 2 is sampled from an Exponential Distribution with a population mean equal to the sample mean, and a population standard deviation equal to the sample standard deviation.

n = observations = 99

Sample Mean: 2.1919
Sample Standard Deviation: 1.251
Chi-square formula:

$$X^2 = \sum_{i=0}^{i} \frac{(o_i - e_i)^2}{e_i}$$

Z formula:

$$z = \frac{x - \mu}{s}$$

P(X<=X) formulas<excel>:
-NORM.S.DIST(z,cumulative)
-FREQUENCY(data_arrays,bins_arrays)

Expected Frequency Table:

Interval	**Observed Frequency**	**Class Probability**	**Expected Frequency**	$\boldsymbol{X^2}$ **Class Component**
$1 \leq X < 3$	68	0.5	49.5	6.914141
$3 \leq X < 5$	25	0.433	42.867	7.44698
$5 \leq X < 7$	6	0.0668	6.6132	0.056858
Total	99	0.9998	98.9802	14.41798

Goodness of Fit Test for Exponential Distribution:

H_0: The observed distribution of Set 2 may not be fitted by the exponential distribution.

H_1: The observed distribution of Set 2 may be fitted by the exponential distribution.

Degree of Freedom: df = n – 1 = 3 – 1 = 2

Statement: Using the X^2 distribution tables, at 5% level of significance for 2 degrees of freedom, the critical value is 5.991

Conclusion: if X > 5.991 and X< -5.991, then the null hypothesis should be rejected.

It is observed that the calculated value (14.417) is greater than the critical value (5.991). The null hypothesis should be rejected, as there is insufficient evidence to conclude that the observed distribution does not fit exponential distribution.

Name: Temitayo Aderounmu

UTA ID: 1001568524

IE 3301 Fall 2021 Project

Part III

Due Date: 11/29/2021

“I __Temitayo Aderounmu__ did not give or receive any assistance on this project, and the report submitted is wholly my own.”

X ______________________________

IE3301: Data Collection

First Sample:

The following data set refers to the height in centimeters of a group of adult men, living in the region of the western edge of the Kalahari Desert.

n_1 = COUNT = 7

Data Set 1:
152 157 151 157 161 162 161

Measure of Location:
Sample Mean = Average = 156
Geometric Mean = GEOMEAN = 155.945
Sample Median = MEDIAN = 157
Sample Mode = MODE = 152
Measure of Variability:
Sample Range = MAX-MIN = 11
Sample Variance = VAR = 20
Sample Standard Deviation=STDEV= 4.47
Coefficient of Variation (%) = s/x = 2.867

Second Sample:

The following data set refers to the height in centimeters of a group of adult women, living in the region of the western edge of the Kalahari Desert.

n_2 = COUNT = 7

Data Set 2:
140 137 145 149 148 145 150

Measure of Location:
Sample Mean = Average = 144.857
Geometric Mean = GEOMEAN = 144.788
Sample Median = MEDIAN = 145
Sample Mode = MODE = 145
Measure of Variability:
Sample Range = MAX-MIN = 13
Sample Variance = VAR = 23.1429
Sample Standard Deviation=STDEV= 4.812
Coefficient of Variation (%) = s/x = 3.32

IE3301: Hypothesis Testing

One-Sided lower-tail [Form B]:

$\alpha = 0.05$; $n = 7$; $\sigma^2 = 23.1429$; $s = 4.812$
H_0: $\sigma^2 = 18.08$
H_1: $\sigma^2 < 18.08$

Conclusion: if $T < X^2_{\alpha,n\text{-}1}$, then the null hypothesis should be rejected. With a 5% level of significance, there is insufficient evidence that the variance is the specified value (18.08).

Test statistic: $T = (n-1)(\frac{s}{\sigma_0})^2 = (7-1)(\frac{4.812}{18.08})^2 = 0.425$
Degrees of freedom: $n - 1 = 6$
Significance level: $\alpha = 0.05$
Critical values: $X^2_{\alpha,n\text{-}1} = 12.592$
Critical region: Reject H_0 if $T < 12.59$

One-Sided upper-tail [Form C]:
α = 0.05; n = 7; μ = 144.857; s = 4.812

H_0: μ = 142
H_1: μ > 142

Decision: if Z ≥ 1.645, then the null hypothesis should be rejected.

Conclusion: Fail to reject H_0, therefore, there is enough evidence that the mean could be the specified value (142).

Test statistic:
$z = \frac{x-\mu}{s} = \frac{144.857-142}{4.812}$ = 0.594

Degrees of freedom: *n* - 1 = 6
Significance level: *α* = 0.05
Critical region: Fail to Reject H_0, *0.59* < 1.65

F-test [Form A]:
α = 0.10; $n_1 = n_2 = 7$; $\sigma_1{}^2 = 20$; $\sigma_2{}^2 = 23.1429$

H_0: $\sigma_1{}^2 = \sigma_2{}^2$
H_1: $\sigma_1{}^2 \neq \sigma_2{}^2$

Decision: if F < *F* (1-*α*/2, N_1-1, N_2-1) or F > *F* (*α*/2, N_1-1, N_2-1), then the null hypothesis should be rejected.

Conclusion: Fail to reject H_0, therefore, we conclude that there is enough evidence that the two population variances might be equal.

Test statistic:
$\boldsymbol{F} = \sigma_2{}^2 / \sigma_1{}^2$ = 20/23.1429 = 0.864

Degrees of freedom: $n_1 - 1$, $n_2 - 1$ = 6,6
Significance level: *α* = 0.05
Critical values:
$f_{0.05}(6,6) = 4.28$
$f_{0.95}(6,6) = 1/(f_{0.05}(6,6)) = 0.234$

Critical region: Reject H_0 if *F* < 0.234 or F > 4.28

Two-sample t-test:
α = 0.05; $n_1 = n_2 = 7$; $\mu_2 = 144.857$; $\mu_1 = 156$; $s_1 = 4.47$; $s_2 = 4.812$

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

Critical region: t > 1.812, with v = 10 degrees of freedom; v = $n_1 + n_2$ -2 = 6 + 6 - 2 = 10

$$S_p = \sqrt{\frac{s_1{}^2(n_1-1)+s_2{}^2(n_2-1)}{n_1+n_2-2}} = \sqrt{\frac{(19.98)(6)+(23.16)(6)}{7+7-2}} = 4.64$$

$$t = \frac{(\mu_1-\mu_2)-(n_1-n_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(156-144.857)-(7-7)}{(4.64)\sqrt{\frac{1}{7}+\frac{1}{7}}} = 4.488$$

P = P (T > 4.488) ≈ 0.0006

Decision: Fail to reject H_0. We are not able to conclude the mean of the first and second sample are not equal to each other. Form A and Form C, respectively the F-test and upper-tail test, makes good comparison for the two-sample t-test.

APPENDIX

I. Data:

Through a rigorous search online, I came across a raw data file that contains the height, weight, and gender of people living in the western edge of the Kalahari Desert in Namib. Carefully going through the file, I collected the data I needed, which was the height of the adult male in Namib. Set 1 comprises of this collected data, which is then observed for various components.

Set 2 data set comprises of the inter-arrival time of the arrival of customers at an ATM service point of the Fidelity Bank Plc, North Central. The actual time of the customers were recorded in the ART column of the Set 2 data set on page 6. Then the interval between each occurrence was recorded by taking the difference of the present arrival time and the previous arrival time. Set 1 is then observed and several other components were derived from it.

II. Descriptive Statistics:

Set 1, measure of location includes a sample mean of 154.345, a sample median of 157, and a sample mode of 159. Its measure of variability includes a range of 81, a sample variance of 217.2, and a sample standard deviation of 14.74. Referring to the frequency histogram for Set 1, it illustrates that as the height increases, the frequency also increased till a certain point before it decreased. Before the height reached that certain point where it started decreasing, it can be inferred that the frequency is in direct proportion to the height data. If the outliers in the data are to be neglected, it appears that the Set does follows a Normal distribution, however if the outliers were taken into considered, the set will more likely follow a left-skewed distribution.

Set 2, measure of location includes a sample mean of 2.19, a sample median of 2, and a sample mode of 2. Its measure of variability includes a range of 7, a sample variance of 1.565, and a sample standard deviation of 1.251. Referring to the frequency histogram for Set 2, it illustrates that as the inter-arrival time of the customers increases, the frequency decreased. With this, it can be inferred that the frequency is indirectly proportional to the inter-arrival time data. While analyzing the sample, it appears that the Set does follows an Exponential distribution.

The sample mean for both set 1 and 2, is the average of the data collected, the sample median is a location measure that represents the middle value across the set, and the sample mode is the most frequent value in the sets respectively. The sample variance, range, and sample standard deviation help to measure the variability across both sets.

III. Goodness of Fit Tests:

For Set 1, the chi-square goodness of fit test which is a statistical hypothesis test was performed to determine whether the data set is a likely to follow the normal distribution. The formulas used, calculated values and tables are in the Set 1- Part II page of the report [PG 10]. For Set 2, the chi-square goodness of fit test which is a statistical hypothesis test was performed to determine whether the data set is a likely to follow the exponential distribution. The formulas used, calculated values and tables are in the Set 2- Part II page of the report [PG 11].

IV. Data:

For the first sample, 7 data points were randomly collected from Set 1. And for the second sample, I used the same raw data file that contains the height, weight, and gender of people living in the western edge of the Kalahari Desert in Namib. Carefully going through the file, I collected the data I needed [7], which was the height of the adult women in Namib. The second sample was chosen because it is a very relevant set of data good enough for comparison with the first sample.

V. Hypothesis Tests:

A. Single Population Variance:

For this, we used the one-sided lower-tailed chi-square test to test the population variance of the second sample, that comprises of the height of women in the Namib region. It is a test where we hypothesized that there is a decrease in the specified value to be tested on. The null hypothesis is that the variance is equal to 18.08, and the alternative hypothesis is that the variance will be less than the specified value. The null hypothesis is rejected as there was insufficient evidence that the variance is the specified value tested on. Therefore, there is a 5% probability that we commit a Type I error.

B. Single Population Mean

For this, we used the one-sided upper-tailed chi-square test to test the population mean of the second sample, that comprises of the height of women in the Namib region. It is a test where we hypothesized that there is an increase in the specified value to be tested on. The null hypothesis is that the mean is equal to 142, and the alternative hypothesis is that the mean will be greater than the specified value. The null hypothesis is failed to be rejected as there was sufficient evidence that the mean is the specified value tested on. Therefore, there is a 5% probability that we commit a Type II error.

C. Comparison of Two Variances

For this, we performed a F-test to test the population variance of the first and second sample, that comprises of the height of men and women in the Namib region. It is a test where we hypothesized that there is equality in the two variances being tested. The null hypothesis is that the variance of the first sample is equal to the variance of the second sample, and the alternative hypothesis is that both variances are not equal to each other. The null hypothesis is failed to be rejected as there was sufficient evidence that the variance of both variances is equal. Therefore, there is a 10% probability that we commit a Type II error.

D. Comparison of Two Means

For this, we performed a two-sample two-tailed test to test the population mean of the first and second sample, that comprises of the height of men and women in the Namib region. It is a test where we hypothesized that there is difference in the two mean being tested. The null hypothesis is that the mean of the first sample is equal to the mean of the second sample, and the alternative hypothesis is that both mean are not equal to each other. The null hypothesis is failed to be rejected as there was sufficient evidence that the mean of both samples is equal. Therefore, there is a 5% probability that we commit a Type II error. Using excel analysis function, the confidence interval at 95% for the first sample is 4.136 and for the second sample is 4.449.