

**STATISTICS(STA3030F) PROJECT1 REPORT**

**NAME: SIPHOSAKHE MAHLANGU**

**COURSE: INFERENTIAL STATISTICS**

**COURSE CODE: STA3030F**

**PROFESSOR NAME: DOMINIQUE KATSHUNGA**

**PAPER DUE DATE: 13 MARCH 2020**

**INSTITUTION NAME: UNIVERSITY OF CAPE TOWN**

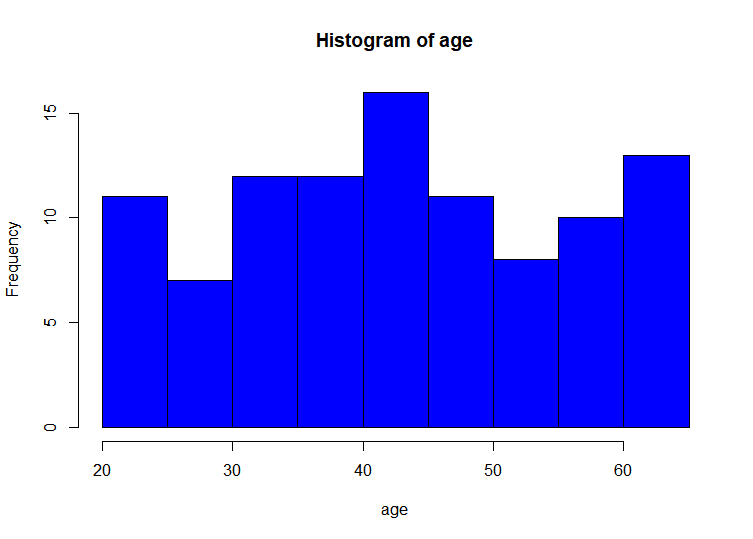
This report is about the study of the Income and Survey conducted by the Statistics South Africa (Sta SA) for a given year. The variables included in the study were PROVINCE( where the household is located), GENDER(the gender of the head of the household), RACE(the race of the group of the household), AGE(the age of the head of the household), AREATYPE(whether the household is in urban or rural), HHSIZE(total number of people living in the household), HHEXPENDITURE(total monthly expenditure of the household) and HHINCOME(total monthly income of the household).

**FULL DATA PLOT:**

A screenshot of a cell phone

Description automatically generated

The exploratory data analysis was done using the 100 samples from the **AGE** of the individuals and the following data representations (the histogram and box plot) were produced:



* The histogram produced showed a symmetric distribution, if the cut was done in the middle, the left and right side of the histogram will be merely the same.

A screenshot of a cell phone

Description automatically generated

* The box plot produced showed a symmetric distribution, if the cut was done in the middle, the left and right side will be merely be the same.
* Five number summary of box plot was (Minimum:12.00, 1st Q:33.00, 2nd, Q:42.50, 3rd Q:54.00, Maximum:65).
* The mean **AGE** is 43.66.

**THE BOOSTRAP WAS PERFORMED IN THE AGE DATA AND THESE WERE THE FOLLOWING RESULTS:**

**A screenshot of a video game

Description automatically generated**

* The histogram for the bootstrap sample means followed the normal distribution (it is bell shaped).
* The p-value and confidence interval of the bootstrapped means is 0.70675 and (41.34, 46.16) respectively.
* The alternative hypothesis was rejected due to higher p-value.

**HYPOTHESIS TEST WAS RUN AND YIELDED THE FOLLOWING RESULTS:**

* **H0(**Null hypothesis): mean <= 43 and **H1**(alternative): mean>43
* The alternative hypothesis was rejected and concluded that mean age is less or equals to 43 because of the higher p-value.
* The 90 percent confidence interval of errors for the median household income is (-1700.095, 260.505).

**THE DATA SETS WERE TAKEN FROM THE REALWORLD STUDIES COVERED IN THE BOOK “STATISTICS FOR MANAGEMENT ECONOMICS” AND THE HYPOTHESIS TESTS WERE PERFOMED:**

* **H0(**Null hypothesis): mean 1-mean2 = 0 and **H1**(alternative): mean1-mean2 is not equals to 0.
* The p-value of the test is 0.0034.
* The null hypothesis was rejected and concluded that means of the two groups are not equal due to the smaller p-value.
* The 95% confidence interval for this test is (-79.29, -51.09).

**THE TEST FOR THE EQUALITY OF VARIANCES WERE TESTED USING THE BOOTSTRAP METHOD AND THE YIELDED THE FOLLOWING RESULTS:**

* **H0**(Null hypotheses): varA = varB and **H1**(alternative): varA is not equals to varB.
* The P-value was found to be 0.0034.
* The null hypothesis was rejected and concluded that variance of both samples were not equal.

**THE TEST FOR THE EQUALITY OF VARIANCES WERE TESTED USING THE NORMAL THEORY AND YIELDED THE FOLLOWING RESULTS:**

* **H0**(Null hypotheses): varA = varB and **H1**(alternative): varA is not equals to varB.
* P-value was found to be 0.54408.
* The F-stat 1.39214.
* By the results above the alternative is rejected due to higher p-value and concludes that the variances are not equal.

**COMPARISONS BETWEEN THE BOOTSTRAPPING AND NORMAL THEORY:**

In the bootstrap method the p-value was found to be 0.0034 which led to the rejection of the null hypotheses whereas in the normal theory p-value was found to be 0.54408 which led to the rejection of alternative hypotheses.

**TEST THE SIGNIFICANCE MEAN DIFFERENCE BETWEEN THE GROUPS BY SING THE BOOTSTRAP APPROACH:**

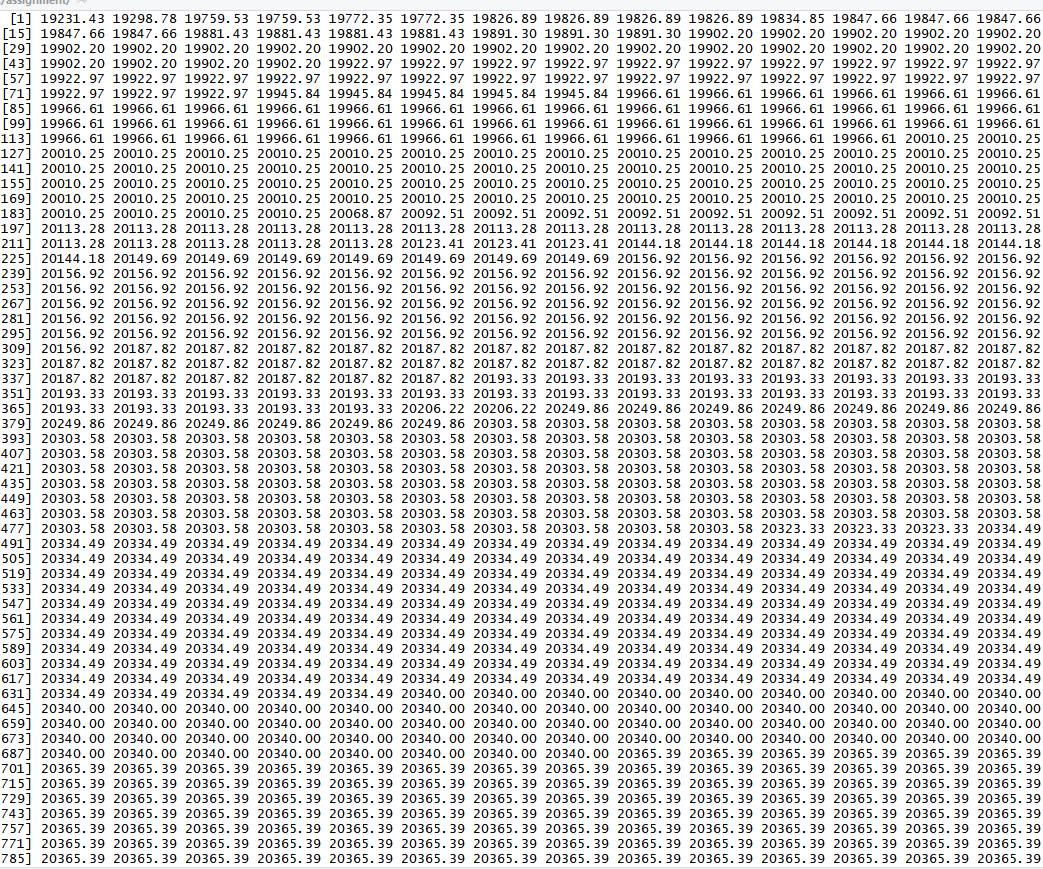
* **H0**(null hypotheses) and **H1**(alternative hypotheses).
* The p-value was found to be 0.8372.
* Which led to the rejection of the alternative hypotheses due to larger p-value.

**Appendix:**

For Question 1

**A screen shot of a social media post

Description automatically generated**

****

For Question 2

