Probability Theory and Statistics

Logo

Description automatically generated

ALY6010, WINTER 2022

Module-3- R Practice Output

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Prof. Tom Beur

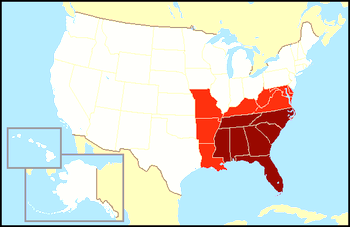
Date: 03/14/2022

**Introduction:**

In this study, we compare the average pay of different states in the United States to a one-sample t-test of the southern states. Essentially, we utilize a single sample test to see if our data is accessible from the population, but we do not gather data from the entire population. In this case, we will use a t-test to compute the average income in southern states.

Apart from that, we do hypothesis testing to see if the population's mean is equivalent to the population's value. On the other hand, if we do not know the population's standard deviation.

In this data set, the southeastern states of United States know as American southeast majorly eastern area of southern United states.



In this map, the area in dark red mostly are southeastern united states, and light red are southeastern few in from which with other regions of United states.

Task1: Import data in R- This task imported the data of the median income of Southern state, which has 2 variables and 52 observations. I have defined this with Med\_Incm

Text

Description automatically generated

Task2: Clean data with gsub- In this task after getting the structure found that in this table 2 variables are character type. To convert the variable name x2010 in numeric type with as. numeric with that removing the comma by using gsub().

A picture containing calendar

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Task3:Extract Southern State: This task is done through Wikipedia and the total number in the dataset.

Text

Description automatically generated with medium confidence

Task4: Summary: In this task get an understanding of the dataset. We have 52 states with character type and x2010 income with a minimum of 42996, 1st quartile is 45660, the median value is 55374, mean of 50936, 3rd quartile at 50594, and the maximum income 72336.

Text

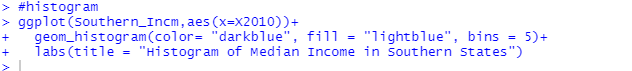
Description automatically generated

Task5: Str(): in this task define the data frame called southern\_incm

Text

Description automatically generated

Task5:Histogram: In this task first two bars are showing the average salary in the range of 40000 to 50000



Chart, histogram

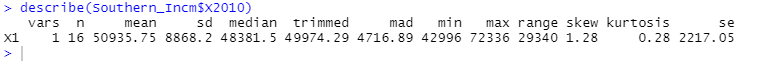
Description automatically generated

Task 6: Boxplot-In this task the given boxplot has no outlier while the average income in most states is within the range of the dataset. We can find some outliers in the boxplots above 60000 average income in southern states.

Chart

Description automatically generated

Task7: Describe(): In this task have to define the descriptive statistics with variance is equal to 1, number of observation is 52, mean is 50935, whereas standard deviation is 8868.20, median 48381.50, the maximum value is 72336, the range for the income is 29340



Task8: Mean()- In this task calculated the mean of the income in all states in R which will help in analyzing the hypothesis testing of the Median\_Income dataset. In this average income is 50935

Text

Description automatically generated

Task9: t-test- In this task function in R called t.test(), we can do the one-sample t-tests for the income in southern states. This output provides all the values about performing the one-sample t-test with the value of test statistics, the p-value, the degree of freedom, alternative hypothesis, confidence interval, sample estimated mean.

Graphical user interface, text, application, email

Description automatically generated

**Points of Sequence:** In this task, we can make the distribution curve

Text

Description automatically generated

**Distribution curve for the t-test**

Chart, line chart

Description automatically generated

**Summary:**

In this report, we have defined the one-sided t-test to get the significant value of the mean income in the southern states in the United States. Whereas the t-test is used for hypothesis testing in the field of statistics. The data given in the dataset provides the States and Income of the states in the United States.

While cleaning that data I found both the values are character types. By manipulation technique, cleaned the data and collected the southern states given in the data frame that link helped collect the data.

Through the process of finding the t-test plotted a boxplot which shows the outlier for those who have values above 60000.

However, a histogram shows that the income of the maximum state is between 40000 to 50000.

**T-test**

The average income of the southern states is mu = 50935, whereas the confidence interval is 95% i.e., 0.95.

While calculating the first t-test Null hypothesis of one sample t-test the degree of freedom is 15, the value of t is 0.00033829 and the p-value is 0.9997, moreover, the 95% confidence interval shows the average income is 46210.22 and 55661.28. The range proves that the value of t is the same are the value of range which indicates that it’s a two-sided test.

In the case of the alternative hypothesis for greater the p-value is 0.4999, whereas for less is 0.5001.

**References:**

[1] How to Use gsub() in R - With Examples

<https://www.programmingr.com/tutorial/gsub-in-r/>

[2] One-tailed Test or Two-tailed Test - Nipissing University

A one-tail test is applied. If words such as ^change, the same, different/difference, and so on \_ are used in the claim of the question (≠ is used in H 1), a two-tailed test is applied. In a one-tailed test, the critical region has just one part (the green area below). It can be a left-tailed test or a right-tailed test.

<https://www.nipissingu.ca/sites/default/files/One-tailed-Test-or-Two-tailed-Test.pdf>

[3] One-Sample T-test in R

<http://www.sthda.com/english/wiki/one-sample-t-test-in-r>

Appendix:

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated