

Title: Basic statistical commands on the dataset using R & data exploration.

Problem statement: To execute basic statistical commands on the given dataset and explore the data to obtain useful information.

Pre-Lab: A basic understanding of descriptive statistics will help in executing R commands on the given dataset.

Theory:

Statistics commands in R:

1. Mean:

In R, a mean can be calculated on an isolated variable. Alternatively, a mean can be calculated for each of the variables in a dataset by using the mean (DATAVAR) command where DATAVAR is the name of variable containing the data.

The syntax is:

`mean(x, trim=0, na.rm=FALSE, ...)`

- `x` is the input vector.
- `trim` is used to drop some observations from both end of the sorted vector.
- `na.rm` is used to remove the missing values from the input vector.

4 Standard deviation:

Within R, standard deviation are calculated in the same way as means. The standard deviation of a single variable can be computed with the `sd(VAR)` command, where VAR is the name of the variable whose standard deviation you wish to retrieve. The syntax is:

`sd(x, na.rm = FALSE)`

Following is the description of the parameters used =

- x is the input vector.
- na.rm is used to remove the missing values from input vector.

5 Range:

Minimum & Maximum

keeping with the pattern, a minimum can be computed on a single variable using the `min(VAR)` command.

The syntax is:

`min(x)`

Following is the description of the parameters used -

- x is input vector.

6.2 Percentiles from Values (Percentile Rank):

In the opposite situation, where a percentile rank corresponding to a given value is needed, one has to devise a custom method. To begin, consider the steps involved in calculating a percentile rank.

Formula for calculating percentile rank -
$$\text{percentile rank} = \frac{\text{length}(\text{VAR}[\text{VAR} \leq \text{VAL}])}{\text{length}(\text{VAR})} * 100$$

Where VAR is the name of the Variable and VAL is the given value.

7. 5- Number Summary

A 5-Number Summary is a set of 5 descriptive statistics for summarizing a continuous univariate data set. It consists the data set's

- minimum
- Median
- 1st quartile
- 3rd quartile
- Maximum

There are 2 functions that are commonly used to calculate the 5-number summary in R.

The syntax is:

`fivenum(X)`

`Summary(X)`

Following it is the descriptive of the parameters used -

- X is the input vector.

Past - Lab : Students will be able to execute statistical R commands on any given dataset & explore the dataset

Conclusion : Thus exercised various statistical & data exploration commands on the given dataset using R.