Roll No :- 307

(5)

Assignment 1 (c)

Date : / /20

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

Problem statement :- To execute basic statistical commands on given dataset & explore 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 calculate 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 the variable containing the data.

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

Pooja

Date : / / 20

2) Median :-
The middle most value in a data series is called the madien.
The syntax is :
median (x, na.rm = FALSE)
- x. is the input vector.
- na.rm is used to remove missing values from input vector.

3) Mode :
The mode is value that has highest no of occurances in a set of data. unlike mean & median, mode can have both numeric & character data.
mode by frequencies.
table(mydata $country) # gives no of occurances of each value in vector.
calculation of mode
max (table (mydata $country)) # gives Count of maximum occurance of a particular value.

names (sort (table (mydata $country)))
gives value which has maximum occurances

4) Standard Deviation :
Within R, standard deviations are calculated in same way as means.
The syntax is :
sd (x, na.rm = FALSE)

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

5) Range :
minimum & maximum
keeping with pattern, a minimum can be computed on a single variable using min(VAR) command.
The syntax is :
min (x)
- x is input vector.

The maximum, via max (VAR)
syntax -
max (x)
- x is input vector.

Range can be computed on a single variable using range (VAR) command which gives maximum & minimum value from single variable.
The Syntax is :
range (x)
- x is input vector

6. percentiles
6.1 values from percentile (quantiles)
quantile (VAR, c(PROB1, PROB2, ---))

Date : / / 20

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 steps involved in calculating a percentile rank.

1) count number of data points that are at or below given value.
2) devide by total number of data points
3) multiply by 100.

Percentile rank = length (VAR[VAR <= VAL])/ length (VAR) * 100.

Where,
VAR is the name of variable & VAL is the given value.
This formula makes use of length function in two variations.

7 5-Number Summary

A 5-number summary is a set of 5 descriptive statistics for summarizing a continuous univariate data set. It consists of data set's

- minimum
- 1st quartile
- median
- 3rd quartile
- maximum

Pooja

Date : / /20

this is a simple but very useful way of summarizing your data for several reasons.

- the medium gives a measure of the centre of the data
- the minimum & maximum give range of data.
- the 1st & 3rd quartiles, give a sense of spread of data, especially when compared to minimum, maximum & median.

The syntax is -

fivenum(x)

- x is input vector

summary(x)

- x is input vector

perform above statistical functions on dataset given below:

NO	SEX	AGE	NOOFCHILDREN	WEIGHT	HEIGHT
1	0	57	1	65	158
2	1	70	3	100	175
3	0	45	0	71	162
4	0	38	2	58	164
5	0	25	1	81	170
6	1	50	4	68	172
7	1	61	0	85	179

Pooja

Date : / / 20

Exploring data in R:

summary(mydata) # provides basic descriptive statistics & frequencies
edit(mydata) # open data editor
str(mydata) # provides structure of dataset
names(mydata) # lists variables in dataset
head(mydata) # first 10 rows of dataset.
head(mydata, n=10) # first 10 rows of dataset
head(mydata, n=-10) # all rows but last 10
tail(mydata) # last 6 rows tail(mydata, n=10) # last 10 rows.
tail(mydata, n=-10) # All rows but first 10
mydata[1:10] # first 10 rows.
mydata[1:10, 1:3] # first 10 rows of data of first 3 variables.
mydata[,] # all rows of data.

Post 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 given dataset using R.