Central Tendency and Spread in R Programming

The central tendency of a dataset refers to the central or typical value around which the data points tend to cluster. There are several measures of central tendency, including mean, median, and mode.

For this section mtcars dataset will be used. To get that data, install ggplot2 package and load the package if you didn’t do it till now. Then load the data.

Use the below code do that.

install.packages("ggplot2")

library(ggplot2)

data(mtcars)

Now you can access the mtcars data by using ‘mtcars’. Explore the data little bit using names(), str(), summary(), dim() functions etc.

str(mtcars)

Output:

'data.frame': 32 obs. of 11 variables:

$ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...

$ cyl : num 6 6 4 6 8 6 8 4 4 6 ...

$ disp: num 160 160 108 258 360 ...

$ hp : num 110 110 93 110 175 105 245 62 95 123 ...

$ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...

$ wt : num 2.62 2.88 2.32 3.21 3.44 ...

$ qsec: num 16.5 17 18.6 19.4 17 ...

$ vs : num 0 0 1 1 0 1 0 1 1 1 ...

$ am : num 1 1 1 0 0 0 0 0 0 0 ...

$ gear: num 4 4 4 3 3 3 3 4 4 4 ...

$ carb: num 4 4 1 1 2 1 4 2 2 4 ...

dim(mtcars)

Measure Central Tendency in R

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. The mean and median are the most likely measure of central tendency for numerical data.

Mean

The mean is the average value of a set of data points. In R, the mean() function can be used to calculate the mean.

Median

The median is the middle value in a set of data points when they are arranged in order. In R, the median() function can be used to calculate the median.

Mode

The mode is the value that occurs most frequently in a set of data points. In R, there is no built-in function to calculate the mode. However, you can create a function to calculate it.

There is a variable in this dataset called ‘mpg’ Miles/(US) gallon. If you want to know mean and median mpg miles/gallon then type the below code.

mean(mtcars$mpg)

median(mtcars$mpg)

Output:

> mean(mtcars$mpg)

[1] 20.09062

> median(mtcars$mpg)

[1] 19.2

Measures of Spread in R

The spread of a dataset refers to the amount of variability or dispersion among the data points. There are several measures of spread, including range, variance, and standard deviation.

To get the measures of spread you can use variance, standard deviation, interquartile range (IQR), minimum value, maximum value, range etc.

Range

The range is the difference between the maximum and minimum values in a set of data points. In R, you can calculate the range using the range() function.

Variance

The variance is a measure of how much the data points vary from the mean. In R, you can calculate the variance using the var() function.

Standard deviation

The standard deviation is the square root of the variance and is also a measure of the amount of variability among the data points. In R, you can calculate the standard deviation using the sd() function.

Calculate Variance in R

var(mtcars$mpg)

Output:

[1] 36.3241

Calculate Standard deviation in R

sd(mtcars$mpg)

Output:

[1] 6.026948

Calculate Interquartile range (IQR) in R

IQR(mtcars$mpg)

Output:

[1] 7.375

The min(), max(), and range() functions in R

In R, you can use the min(), max(), and range() functions to find the minimum value, maximum value, and range of a vector or a set of values.

min(mtcars$mpg)

max(mtcars$mpg)

range(mtcars$mpg)

Output:

> min(mtcars$mpg)

[1] 10.4

> max(mtcars$mpg)

[1] 33.9

> range(mtcars$mpg)

[1] 10.4 33.9

Categorical Variable

In statistics and data analysis, a categorical variable is a variable that can take on one of a limited, finite set of values, known as categories or levels. These categories can be either nominal or ordinal.

Categorical variables are often represented in R using factors, which are variables that have a fixed set of categories, or levels. Factors are useful in R for representing categorical variables because they can be ordered, labeled, and used in statistical models.

For categorical variables, counts and percentages can be used for summary.

table(mtcars$cyl)

table(mtcars$cyl)/nrow(mtcars)

Output:

> table(mtcars$cyl)

4 6 8

11 7 14

>

> table(mtcars$cyl)/nrow(mtcars)

4 6 8

0.34375 0.21875 0.43750

The unique() function in

If you want to know how many unique values are there in a column then use unique() function.

unique(mtcars$cyl)

Output:

[1] 6 4 8

The table() function in R

In R, the table() function is used to create a frequency table of the counts of values in a vector or a data frame column. The output of the table() function is a table object that shows the number of occurrences of each unique value in the vector or column.

If you want to get a frequency table for Number of cylinders vs Number of carburetors then use the below code.

table(mtcars$cyl, mtcars$carb)

Output:

>table(mtcars$cyl, mtcars$carb)

1 2 3 4 6 8

4 5 6 0 0 0 0

6 2 0 0 4 1 0

8 0 4 3 6 0 1

Percentiles are measures of central tendency, which depict that out of the total data about certain percent data lies below it. In R, we can use quantile() function to get the job done.

Syntax: quantile( data, probs)

Parameter:

data: data whose percentiles are to be calculated

probs: percentile value

1: Calculate percentile

To calculate the percentile we simply pass the data and the value of the required percentile.

x<-c(2,13,5,36,12,50)

res<-quantile(x,probs=0.5)

res

Output:

50%

12.5

Calculate percentiles of vector

We can calculate multiple percentiles at once. For that, we have to pass the vector of percentiles instead of a single value to probs parameter.

x<-c(2,13,5,36,12,50)

res<-quantile(x,probs=c(0.5,0.75))

res

Variability (also known as Statistical Dispersion) is another feature of descriptive statistics. Measures of central tendency and variability together comprise of descriptive statistics. Variability shows the spread of a data set around a point. Example: Suppose, there exist 2 data sets with the same mean value:

A = 4, 4, 5, 6, 6 Mean(A) = 5 B = 1, 1, 5, 9, 9 Mean(B) = 5

So, to differentiate among the two data sets, R offers various measures of variability.

Measures of Variability

Following are some of the measures of variability that R offers to differentiate between data sets:



Example:

# Defining vector

x <- c(5, 5, 8, 12, 15, 16)

# Print variance of x

print(var(x))





# Defining vector

x <- c(5, 5, 8, 12, 15, 16)

# Standard deviation

d <- sqrt(var(x))

# Print standard deviation of x

print(d)

Output:

[1] 4.875107

Range

Range is the difference between the maximum and minimum value of a data set. In R language, max() and min() is used to find the same, unlike range() function that returns the minimum and maximum value of the data set.

Example:

# Defining vector

x <- c(5, 5, 8, 12, 15, 16)

# range() function output

print(range(x))

# Using max() and min() function

# to calculate the range of data set

print(max(x)-min(x))