Matters of Discussion

Refresh Basic statistics:

mean, median, standard deviation, variance, correlation, covariance

R-Implements

Compiled By: Dr. Nilamadhab Mishra [(PhD- CSIE) Taiwan]

Statistical analysis in R

- Statistical analysis in R is performed by using many in-built functions.
- Most of these functions are part of the R base package.
- These functions take R vector as an input along with the arguments and give the result.

Mean

It is calculated by taking the sum of the values and dividing with the number of values in a data series.

The function mean() is used to calculate this in R.

Syntax

The basic syntax for calculating mean in R is –

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

Following is the description of the parameters used –

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.

Example

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Find Mean.
result.mean <- mean(x)
print(result.mean)
```

O/P—

[1] 8.22

Applying Trim Option

- When trim parameter is supplied, the values in the vector get sorted and then the required numbers of observations are dropped from calculating the mean.
- When trim = 0.3, 3 values from each end will be dropped from the calculations to find mean.
- In this case the sorted vector is (−21, −5, 2, 3, 4.2, 7, 8, 12, 18, 54) and the values removed from the vector for calculating mean are (−21,−5,2) from left and (12,18,54) from right.

Example

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Find Mean.
result.mean <-  mean(x, trim = 0.3)
print(result.mean)
```

O/P-

5.55

Applying NA Option

- If there are missing values, then the mean function returns NA.
- To drop the missing values from the calculation use na.rm = TRUE. which means remove the NA values.

Example; O/P–

[1] NA

[1] 8.22

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5,NA)

# Find mean.
result.mean <-  mean(x)
print(result.mean)

# Find mean dropping NA values.
result.mean <-  mean(x,na.rm = TRUE)
print(result.mean)
```

****

Median

Median

The middle most value in a data series is called the median. The median() function is used in R to calculate this value.

Syntax

The basic syntax for calculating median in R is –

median(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 the input vector.

Example

```
# Create the vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Find the median.
median.result <- median(x)
print(median.result)
```

O/P—

[1] 5.6

****

Mode

Mode

- The mode is the value that has highest number of occurrences in a set of data.
- Unike mean and median, mode can have both numeric and character data.
- R does not have a standard in-built function to calculate mode.
- So we create a user function to calculate mode of a data set in R.
- This function takes the vector as input and gives the mode value as output.

Example

```
# Create the function.
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Create the vector with numbers.
v <- c(2,1,2,3,1,2,3,4,1,5,5,3,2,3)

# Calculate the mode using the user function.
result <- getmode(v)
print(result)

# Create the vector with characters.
charv <- c("o","it","the","it","it")

# Calculate the mode using the user function.
result <- getmode(charv)
print(result)
```

O/P
[1] 2
[1] “it”

****

standard deviation and variance

standard deviation

- ‘Standard deviation is the measure of the dispersion of the values’.
- The higher the standard deviation, the wider the spread of values.
- The lower the standard deviation, the narrower the spread of values.
- In simple words the formula is defined as – Standard deviation is the square root of the ‘variance’.

Variance – It is defined as the squared differences between the observed value and expected value.

$$Variance, \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$Standard\ Deviation, \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Where $x_i = data\ set\ values$

$\bar{x} = mean\ of\ the\ data\ set$

Standard deviation in R

x <- c(34,56,87,65,34,56,89) #creates list 'x' with some values in it.

sd(x) #calculates the standard deviation of the values in the list 'x‘

---

create a list ‘x’ and add some value to it. Then we can find the standard deviation of those values in the list.

Computing variance of a vector

```
# # enter data
y=c(445, 530, 540, 510, 570, 530, 545, 545, 505, 535, 450, 500, 520, 460, 430, 520, 520, 430, 535, 535, 475, 545, 420, 495, 485, 570, 480, 495, 470, 490)
# # calculate
var(y)
sd(y)
```

****

Covariance and Correlation in R

Covariance and Correlation in R Programming

- Covariance and Correlation are terms used in statistics to measure relationships between two random variables.
- Both of these terms measure linear dependency between a pair of random variables or bivariate data.
- Y is the response variable(dependent); X is the predictor variable(Independent)
- $Y = a X + \varepsilon$

Covariance

- In R programming, covariance can be measured using cov() function.
- Covariance is a statistical term used to measures the direction of the linear relationship between the data vectors.
- Mathematically,

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

where,

A represents the A data vector
B represents the B data vector
$\bar{A}$ mean of A data vector
$\bar{B}$ mean of B data vector

Cont..

```
# Data vectors
x <- c(1, 3, 5, 10)
  y <- c(2, 4, 6, 20)
 # Print covariance using different methods
print(cov(x, y))
print(cov(x, y, method = "pearson"))
```

****

Output:

```
[1] 30.66667
[1] 30.66667
```

Correlation

- Correlation is a relationship term in statistics that uses the covariance method to measure how strong the vectors are related.
- cor(x, y, method)
- x and y represents the data vectors
- method defines the type of method to be used to compute covariance. Default is "pearson".
- **Covariance** indicates the direction of the linear relationship between variables while **correlation** measures both the strength and direction of the linear relationship between two variables.

Correlation(cont..)

- Correlation means association - more precisely it is a measure of the extent to which two variables are related.
- There are three possible results of a correlational study:

✓ a positive correlation,

✓ a negative correlation,

✓ no correlation.

Correlation(cont..)

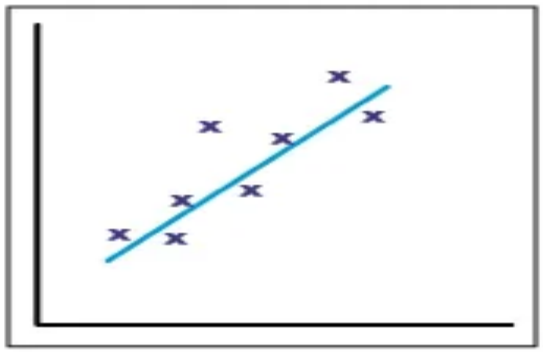
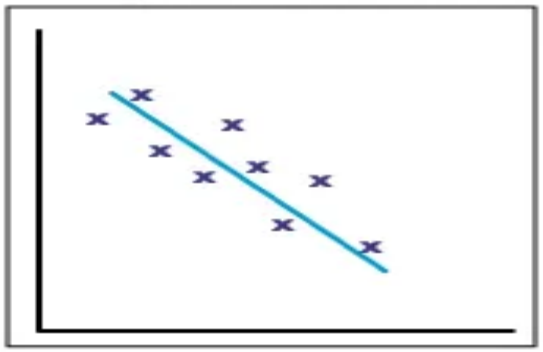
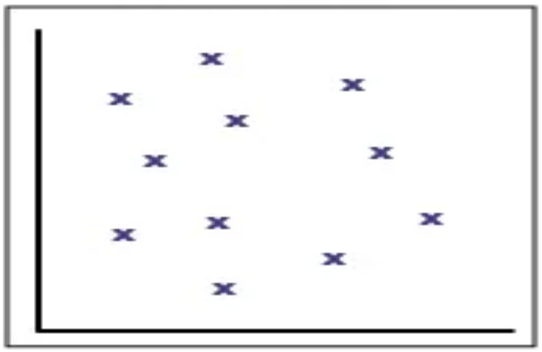
- A positive correlation is a relationship between two variables in which both variables move in the same direction.
- when one variable increases as the other variable increases, or one variable decreases while the other decreases.
- An example of positive correlation would be height and weight.
- Taller people tend to be heavier.

Correlation(cont..)

- A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other.
- An example of negative correlation would be height above sea level and temperature.
- As you climb the mountain (increase in height) it gets colder (decrease in temperature).

Correlation(cont..)

- A zero correlation exists when there is no relationship between two variables.
- For example there is no relationship between the amount of tea drunk and level of intelligence.

Positive correlation	Negative correlation	No correlation
		
The points lie close to a straight line, which has a positive gradient. This shows that as one variable **increases** the other **increases**.	The points lie close to a straight line, which has a negative gradient. This shows that as one variable **increases**, the other **decreases**.	There is no pattern to the points. This shows that there is **no connection** between the two variables.

Guidelines to interpreting Pearson's correlation coefficient

- measure of the strength of a linear association between two variables
- Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables.
- for **example,** that r = .67. That is, as height increases so does basketball performance.

	Coefficient, r	
Strength of Association	**Positive**	**Negative**
Small	**.1 to .3**	**-0.1 to -0.3**
Medium	**.3 to .5**	**-0.3 to -0.5**
Large	**.5 to 1.0**	**-0.5 to -1.0**

Cont..

```
# Data vectors
x <- c(1, 3, 5, 10)

y <- c(2, 4, 6, 20)

# Print correlation using different methods
print(cor(x, y))

print(cor(x, y, method = "pearson"))
```

o/p

```
 [1] 0.9724702
[1] 0.9724702
```

Example

```
# R program to illustrate
# pearson Correlation Testing
# Using cor()

# Taking two numeric
# Vectors with same length
x = c(1, 2, 3, 4, 5, 6, 7)
y = c(1, 3, 6, 2, 7, 4, 5)

# Calculating
# Correlation coefficient
# Using cor() method
result = cor(x, y, method = "pearson")

# Print the result
print("Pearson correlation coefficient is:", result)
```

Output:

Pearson correlation coefficient is: 0.5357143

Correlation measures the linear relationship between objects

ACTIVITY-4(LAB-01)

Investigate the R implements of mean, median, standard deviation, variance, correlation, and covariance.

Please practice those above statistical computations in R- Studio, prepare a report by taking all practice screen sorts along with relevant analysis, and finally, upload to the respective Google classroom assignment section.

Cheers For the Great Patience!

Query Please?

Compiled By: Dr. Nilamadhab Mishra [(PhD- CSIE) Taiwan]