Sardar Patel Institute of Technology,Mumbai
Department of Electronics and Telecommunication Engineering
T.E. Sem-V (2018-2019)
ETL54-Statistical Computational Laboratory

Lab-1: Numerical/Statistical Measures

Name: Manish Dsilva **Roll No. 15**

Objective: How to compute various statistical measures in R with examples.

Outcomes:

1. To load and use built-in data sets in R
2. To install R library and packages in R
3. To compute the numerical measures and describe the significance of the measures.

System Requirements: Ubuntu OS with R and RStudio installed and e1071 library

Procedure:

1. Open RStudio
2. Go to RConsole (>)
3. To install e1071 package

```
>install.packages("e1071")
```

4. Load package e1071, the function kurtosis from the package and compute it.

```
> library(e1071)
>help(kurtosis)
```

5. Load data sets which are built-in R

```
> attach(faithful)
> attach(mtcars)
```

6. To know about the data sets

```
>?faithful or help(faithful)
>?mtcars or help(mtcars)
```

7. To find the mean:

```
>mean(faithful$eruptions)
```

Numerical Measures:

1.Mean

The mean of an observation variable is a numerical measure of the central location of the data values. It is the sum of its data values divided by data count.

Hence, for a data sample of size n, its sample mean is defined as follows:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Similarly, for a data population of size N, the population mean is:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Find the mean eruption duration in the data set faithful.

Mean Eruption Duration	3.487783 mins
Mean Waiting Duration	70.89706 mins

My Interpretation: This means that the average time taken per eruption (in Old Faithful geyser in Yellowstone National Park, Wyoming, USA) is 3.487783 mins and the average waiting duration is 70.89706 mins

2.Median

The median of an observation variable is the value at the middle when the data is sorted in ascending order. It is an ordinal measure of the central location of the data values.

Find the median of the eruption duration in the data set faithful.

Median Eruption Duration	4 mins
Median Waiting Duration	76 mins

My Interpretation:The median time of the eruption is 4 mins and the median interval between eruptions is 76 mins(most occurrence).Which means they are respective 50% of their data.

3.Quartile

There are several quartiles of an observation variable. The first quartile, or lower quartile, is the value that cuts off the first 25% of the data when it is sorted in ascending order. The second quartile, or median, is the value that cuts off the first 50%. The third quartile, or upper quartile, is the value that cuts off the first 75%.

Find the quartiles of the eruption durations in the data set faithful.

Faithful Attributes	25%	50%	75%
Eruption Duration	2.163 mins	4mins	4.454 mins
Waiting Duration	58 mins	76mins	82 mins

My Interpretation: 25% means 25% of the data below it. Therefore x percentile is x% of data below x. It is similar to CAT Percentile.

4.Percentile

The nth percentile of an observation variable is the value that cuts off the first n percent of the data values when it is sorted in ascending order.

Find the 32nd, 57th and 98th percentiles of the eruption durations in the data set faithful.

Faithful Attributes	32%	57%	98%
Eruption Duration	2.39524 mins	4.13300 mins	4.93300 mins
Waiting Duration	62.72 mins	77.47 mins	90.58 mins

My Interpretation: 32% means 32% of the data below it. Therefore x percentile is x% of data below x. It is similar to CAT Percentile. Where 100% would be the max value.

5. Range

The range of an observation variable is the difference of its largest and smallest data values. It is a measure of how far apart the entire data spreads in value.

$$Range = Largest\ Value - Smallest\ Value$$

Find the range of the eruption duration in the data set faithful.

Range of Eruption Duration	3.5 mins
Range of Waiting Duration	53 mins

My Interpretation: Range will tell us how wide-spread our data is.(i.e the difference between max value and min value)

6.Interquartile Range

The interquartile range of an observation variable is the difference of its upper and lower quartiles. It is a measure of how far apart the middle portion of data spreads in value.

$$Interquartile\ Range = Upper\ Quartile - Lower\ Quartile$$

Find the interquartile range of eruption duration in the data set faithful

Interquartile range of eruption duration	2.291mins
Interquartile range of waiting duration	24mins

My Interpretation : This is the range of data between 25% and 75% values of our Data.

7.Box Plot

The box plot of an observation variable is a graphical representation based on its quartiles, as well as its smallest and largest values. It attempts to provide a visual shape of the data distribution.

Find the box plot of the eruption duration in the data set faithful.

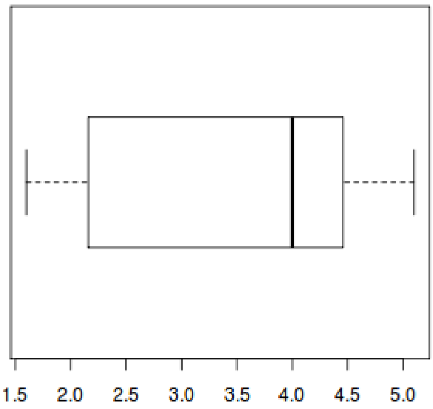
Example Solution:

Apply the boxplot function to produce the box plot of eruptions.

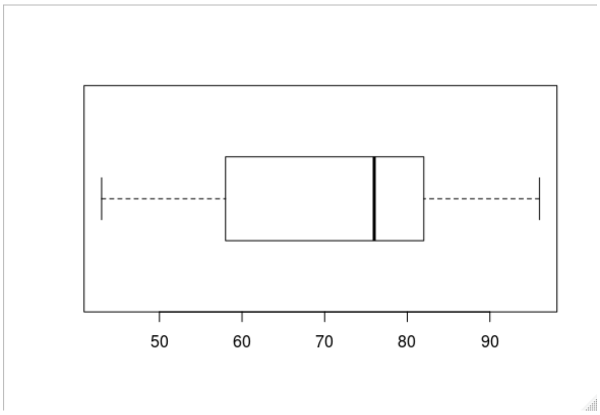
```
> duration = faithful$eruptions      # the eruption durations
> boxplot(duration, horizontal=TRUE)  # horizontal box plot
```

Answer

The box plot of the eruption duration is:



The box plot of the waiting duration is:



My Interpretation : The Black Line Indicates the median of our data.The box indicates data between 25% and 75%.

8.Variance

The variance is a numerical measure of how the data values is dispersed around the mean. In particular, the sample variance is defined as:

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

Similarly, the population variance is defined in terms of the population mean μ and population size N:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Find the variance of the eruption duration in the data set faithful.

Variance of eruption duration	1.302728
Variance of waiting duration	184.8233

My Interpretation : The variance of eruption duration is pretty less which means our data is very close to the mean. But in case of waiting duration it is more wide-spread.

9. Standard Deviation

The standard deviation of an observation variable is the square root of its variance

Find the standard deviation of the eruption duration in the data set faithful.

Standard Deviation of Eruption Duration	1.141371
Standard Deviation of Waiting Duration	13.59497

My Interpretation :Again it would mean our eruption data is close to each other but waiting duration is more wide spread.

10. Covariance

The covariance of two variables x and y in a data set measures how the two are linearly related. A positive covariance would indicate a positive linear relationship between the variables, and a negative covariance would indicate the opposite.

The sample covariance is defined in terms of the sample means as:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Similarly, the population covariance is defined in terms of the populations means μ_x, μ_y as:

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

Find the covariance of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the two variables.

Covariance is 13.97781

My Interpretation :This means more the eruption time more the waiting time. It means if the eruption occurs for a long time then the next eruption will happen after a long time.(Positive Linear Relationship)

11. Correlation Coefficient

The correlation coefficient of two variables in a data set equals to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related.

Formally, the sample correlation coefficient is defined by the following formula, where s_x and s_y are the sample standard deviations, and s_{xy} is the sample covariance.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Similarly, the population correlation coefficient is defined as follows, where σ_x and σ_y are the population standard deviations, and σ_{xy} is the population covariance.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

If the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

Find the correlation coefficient of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the variables.

Correlation coefficient is 0.9008112

My Interpretation : Therefore we can say they are positively linearly related (positive correlation coefficient) .Increasing eruption time would increase waiting time in general.

12. Central Moment

The k^th^ central moment (or moment about the mean) of a data population is:

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^k$$

Similarly, the k^th^ central moment of a data sample is:

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k$$

In particular, the second central moment of a population is its variance.

Find the third central moment of eruption duration in the data set faithful.

Example Solution:

Apply the function moment from the e1071 package. As it is not in the core R library, the package has to be installed and loaded into the R workspace.

```
> library(e1071)                  # load e1071
> duration = faithful$eruptions      # eruption durations
> moment(duration, order=3, center=TRUE)
[1] -0.6149
```

Answer

3rd Central Moment of Eruption Duration	-0.6149
3rd Central Moment of Waiting Duration	-1040.307

My Interpretation : This tells us that the 3rd Central Moment is -0.6149 and -1040.307 this helps us calculate other parameters like skewness of the data manually.

-

13. Skewness

The skewness of a data population is defined by the following formula, where μ_2 and μ_3 are the second and third central moments.

$$\gamma_1 = \mu_3 / \mu_2^{3/2}$$

Intuitively, the skewness is a measure of symmetry. As a rule, negative skewness indicates that the mean of the data values is less than the median, and the data distribution is left-skewed. Positive skewness would indicate that the mean of the data values is larger than the median, and the data distribution is right-skewed.

Find the skewness of eruption duration in the data set faithful.

Skewness of Eruption Duration	-0.4135498
Skewness of Waiting Duration	-0.414025

My Interpretation : Mean of eruption time is less than Median of eruption time and Mean of waiting time is less than Median of waiting time. And both the parameters are negatively skewed

14. Kurtosis

The kurtosis of a univariate population is defined by the following formula, where μ_2 and μ_4 are respectively the second and fourth central moments

$$\gamma_2 = \mu_4 / \mu_2^2 - 3$$

Intuitively, the kurtosis describes the tail shape of the data distribution. The normal distribution has zero kurtosis and thus the standard tail shape. It is said to be mesokurtic. Negative kurtosis would indicate a thin-tailed data distribution, and is said to be platykurtic. Positive kurtosis would indicate a fat-tailed distribution, and is said to be leptokurtic.

Find the kurtosis of eruption duration in the data set faithful.

Example Solution: Apply the function kurtosis from the e1071 package to compute the kurtosis of eruptions. As the package is not in the core R library, it has to be installed and loaded into the R workspace.

```
> library(e1071)                  # load e1071
> duration = faithful$eruptions     # eruption durations
> kurtosis(duration)              # apply the kurtosis function
[1] -1.5116
```

The kurtosis of eruption duration is -1.5116, which indicates that eruption duration distribution is platykurtic. This is consistent with the fact that its histogram is not bell-shaped.

Answer

Kurtosis of Eruption Duration	-1.5116
Kurtosis of Waiting Duration	-1.156263

My Interpretation : Since the Kurtosis is negative , therefore both the distributions are Platykurtic(Data Distribution is thin tailed)

Note

The default algorithm of the function kurtosis in e1071 is based on the formula $g_2 = m_4 / s^4 - 3$, where m_4 and s are the fourth central moment and sample standard deviation respectively. See the R documentation for selecting other types of kurtosis algorithm.

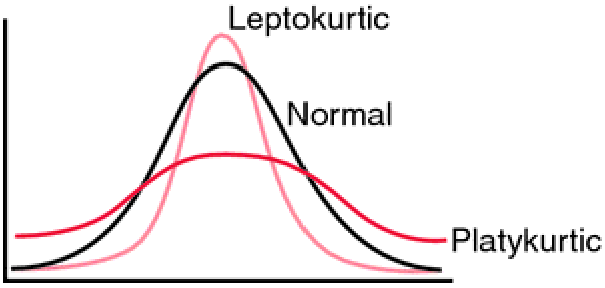
```
> library(e1071)                 # load e1071
> help(kurtosis)
```

Describe the following terms with respect to statistical measures:

Mesokurtic
Mesokurtic is having the same kurtosis as the normal distribution.

Platykurtic
Platykurtic distributions have negative kurtosis. The tails are very thin compared to the normal distribution

Leptokurtic
Leptokurtic is having greater kurtosis than the normal distribution; more concentrated about the mean.



Left-Skewed

For a left skewed (negatively skewed) distribution, the median is typically greater than the mean

Right-Skewed

For a right skewed(positively skewed) distribution, the mean is typically greater than the median

Positively Linearly Related

A relationship of direct proportionality that, when plotted on a graph, traces a straight line. In linear relationships, any given change in an independent variable will always produce a corresponding change in the dependent variable.

Conclusion:

- In this experiment we have analyzed the data of eruptions (faithful dataset) using R and e1071 library
- By our analysis we have found out that there is a linear relation between eruption duration and waiting duration.
- We have also found that both the data is widespread and not concentrated about mean
- We have also found that mean is less than median for both our data
- In the same way we can calculate the statistical parameters of any dataset and analyze the data.