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: SHUBHAM PARULEKAR Roll No. 2017120044

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.

OUTPUT :

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

```
[1] 3.487783
```

```
mean(faithful$waiting)
```

```
[1] 70.89706
```

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.

OUTPUT :

```
median(faithful$eruptions)
```

```
[1] 4
```

```
median(faithful$waiting)
```

```
[1] 76
```

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.

OUTPUT :

```
> quantile(faithful$eruptions)

     0%     25%     50%     75%    100%
1.60000 2.16275 4.00000 4.45425 5.10000
> quantile(faithful$waiting)
 0%  25%  50%  75% 100%
 43   58   76   82   96
```

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.

OUTPUT :

```
> quantile(dur, c(.32, .57, .98))
    32%     57%     98%
2.39524 4.13300 4.93300
> quantile(wait, c(.32, .57, .98))
  32%   57%   98%
62.72 77.47 90.58
```

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.

OUTPUT :

```
> range(faithful$eruptions)

[1] 1.6 5.1
> range(faithful$waiting)
[1] 43 96
> range=faithful$eruptions
> max(range)-min(range)
[1] 3.5
> range=faithful$waiting
> max(range)-min(range)
[1] 53
```

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.

OUTPUT :

```
> range=faithful$eruptions

> IQR(range)
[1] 2.2915
> range=faithful$waiting
> IQR(range)
[1] 24
```

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.

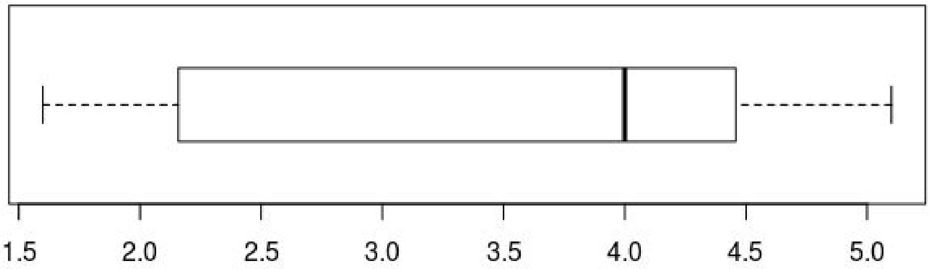
Example Solution:

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

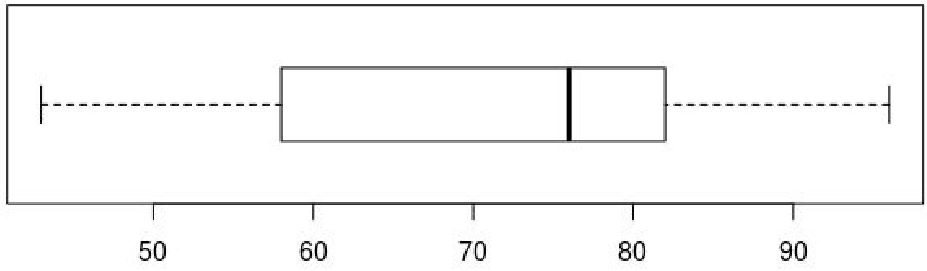
```
> duration = faithful$eruptions      # the eruption durations
> boxplot(duration, horizontal=TRUE)  # horizontal box plot
```

OUTPUT :

```
> range=faithful$eruptions
> boxplot(range, horizontal=TRUE)
```



```
> range=faithful$waiting
> boxplot(range, horizontal=TRUE)
```



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.

OUTPUT :

```
var(faithful$waiting)
[1] 184.8233
> var(faithful$eruptions)
[1] 1.302728
```

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.

OUTPUT :

```
> sd(faithful$eruptions)
[1] 1.141371
> sd(range)
[1] 13.59497
```

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.

OUTPUT :

```
> dur=faithful$eruptions

> wait=faithful$waiting
> cov(dur,wait)
[1] 13.97781
```

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.

OUTPUT :

```
> cor(dur,wait)

[1] 0.9008112
```

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

The third central moment of eruption duration is -0.6149.

OUTPUT :

```
> moment(dur, order=3, center=TRUE)

[1] -0.6149059
> moment(wait, order=3, center=TRUE)
[1] -1040.307
```

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.

OUTPUT :

```
> skewness(dur)

[1] -0.4135498
> skewness(wait)
[1] -0.414025
```

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.

OUTPUT :

```
> kurtosis(dur)

[1] -1.511605
> kurtosis(wait)
[1] -1.156263
```

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
```

Answer: 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.

Exercise

Find the kurtosis of eruption waiting period in faithful.

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 a statistical term used to describe the outlier (or rare, extreme data) characteristic of a probability distribution. A mesokurtic distribution has a similar extreme value character as a normal distribution. Kurtosis is a measure of tails, or extreme values, of a probability distribution. With greater kurtosis, extreme values (e.g., values five or more standard deviations from the mean) occasionally occur.

- **Platykurtic**
  Platykurtic distributions have negative kurtosis. The tails are very thin compared to the normal distribution, or — as in the case of the uniform distribution— non-existent.

- **Leptokurtic**
  Leptokurtic distributions are statistical distributions with occurrences plotted beyond three standard deviations. This results in more occurrences farther from the mean and a higher kurtosis.
  Leptokurtic distributions are known for having more than three standard deviations, which is beyond that of a normal distribution. They are also known

for having fatter tails because a higher number of occurrences are plotted beyond three standard deviations.

- **Left-skewed**
  A left skewed distribution is sometimes called a negatively skewed distribution because it's long tail is on the negative direction on a number line.A common misconception is that the *peak* of distribution is what defines "peakness." In other words, a peak that tends to the left is left skewed distribution. This is incorrect.

- **Right-skewed**
  A right skewed distribution is sometimes called a positive skew distribution. That's because the tail is longer on the positive direction of the number line.

- **Positively linearly related**
  If a straight line on a graph travels **upwards** from left to right, it has a **positive** linear relationship. It shows a steady rate of increase.

Conclusion:

1. We learnt various types of statistical functions, their explanations, their uses in a given dataset and how to formulate it in R
2. We performed around 14 different statistical operations on the given dataset "faithful"and found out its mean,median, covariance, skewness etc
3. Individual computation of various functions for each variable columns gives us the interrelations between different data values