Numerical Measures

We explain how to compute various statistical measures in R with examples. The tutorials are based on the previously discussed built-in data set faithful.

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:

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

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

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

Problem
Find the mean eruption duration in the data set faithful.

Solution
We apply the mean function to compute the mean value of eruptions.

```
> duration = faithful$eruptions     # the eruption durations
> mean(duration)                    # apply the mean function
[1] 3.4878
```

Answer
The mean eruption duration is 3.4878 minutes.

Exercise
Find the mean eruption waiting periods in faithful.

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.

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

Solution
We apply the median function to compute the median value of eruptions.

```
> duration = faithful$eruptions     # the eruption durations
> median(duration)                  # apply the median function
[1] 4

> median(1:4)
[1] 2.5
```

Answer
The median of the eruption duration is 4 minutes.

Exercise
Find the median of the eruption waiting periods in faithful.

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%.

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

Solution
We apply the *quantile* function to compute the quartiles of eruptions.

```
> duration = faithful$eruptions     # the eruption durations
> quantile(duration)                # apply the quantile function
    0%    25%    50%    75%   100%
1.6000 2.1627 4.0000 4.4543 5.1000
```

Answer
The first, second and third quartiles of the eruption duration are 2.1627, 4.0000 and 4.4543 minutes respectively.

Exercise
Find the quartiles of the eruption waiting periods in faithful.

Note
There are several algorithms for the computation of quartiles. Details can be found in the R documentation via help(*quantile*).

Percentile

The n^{th} 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.

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

Solution
We apply the *quantile* function to compute the percentiles of eruptions with the desired percentage ratios.

```
> duration = faithful$eruptions          # the eruption durations
> quantile(duration, c(.32, .57, .98))
   32%    57%    98%
2.3952 4.1330 4.9330
```

Answer
The 32^{nd}, 57^{th} and 98^{th} percentiles of the eruption duration are 2.3952, 3.3422 and 4.9330 minutes respectively.

Exercise
Find the 17^{th}, 43^{rd}, 67^{th} and 85^{th} percentiles of the eruption waiting periods in faithful.

Note
There are several algorithms for the computation of percentiles. Details can be found in the R documentation via help(*quantile*).

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.

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

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

Solution
We apply the *max* and *min* function to compute the largest and smallest values of eruptions, then take the difference.

```
> duration = faithful$eruptions       # the eruption durations
> max(duration) − min(duration)     # apply the max and min functions
[1] 3.5
```

Answer
The range of the eruption duration is 3.5 minutes.

Exercise
Find the range of the eruption waiting periods in faithful.

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.

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

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

Solution
We apply the *IQR* function to compute the interquartile range of eruptions.

```
> duration = faithful$eruptions     # the eruption durations
> IQR(duration)                     # apply the IQR function
[1] 2.2915
```

Answer
The interquartile range of eruption duration is 2.2915 minutes.

Exercise
Find the interquartile range of eruption waiting periods in faithful.

Box Plot

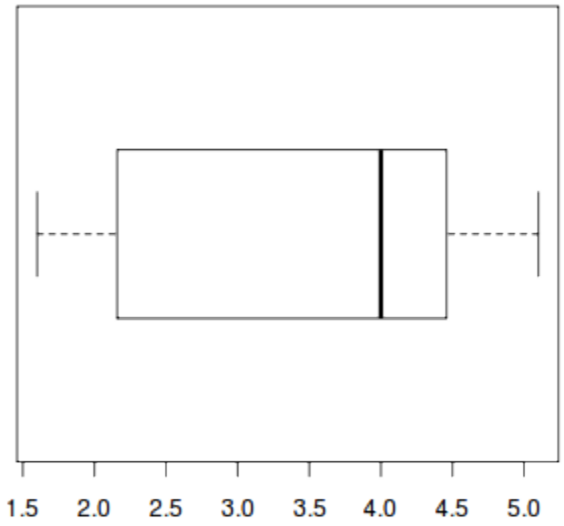
The box plot of an observation variable is a graphical representation based on its quartiles, as well as its smallest and largest values. It attempts to provide a visual shape of the data distribution.

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

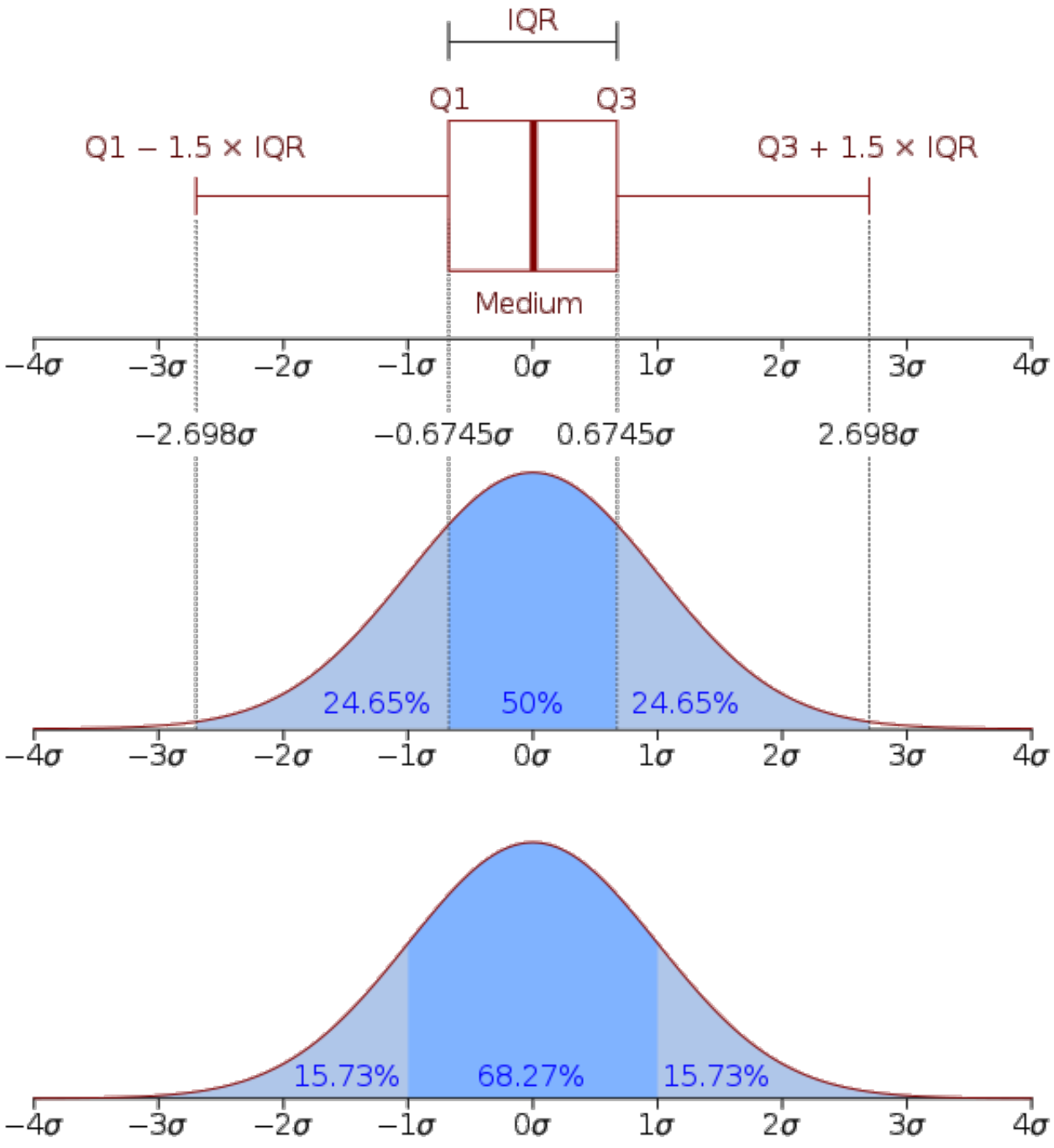
Solution
We apply the *boxplot* function to produce the box plot of eruptions.

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

Answer
The box plot of the eruption duration is:



Exercise
Find the box plot of the eruption waiting periods in faithful.


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Medium
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

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:

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

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

Solution
We apply the var function to compute the variance of eruptions.

```
> duration = faithful$eruptions   # the eruption durations
> var(duration)                   # apply the var function
[1] 1.3027
```

Answer
The variance of the eruption duration is 1.3027.

Exercise
Find the variance of the eruption waiting periods in faithful.

Standard Deviation

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

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

Solution
We apply the *sd* function to compute the standard deviation of eruptions.

```
> duration = faithful$eruptions    # the eruption durations
> sd(duration)                     # apply the sd function
[1] 1.1414
```

Answer
The standard deviation of the eruption duration is 1.1414.

Exercise
Find the standard deviation of the eruption waiting periods in faithful.

Median Absolute Deviation

Compute the median absolute deviation, i.e., the (lo-/hi-) median of the absolute deviations from the median, and (by default) adjust by a factor for asymptotically normal consistency.

$$constat * cMedian(.abs.(x - center))$$

where constant = 1.4826, with the default value of ‘center’ being ‘median(x)’, and ‘cMedian’ being the usual, the ‘low’ or ‘high’ median.

Problem
Find the *mad* of the eruption duration in the data set faithful.

Solution
We apply the *mad* function to compute the median absolute deviation of eruptions.

```
> duration = faithful$eruptions      # the eruption durations
> mad(duration)                      # apply the mad function
[1] 0.9510879
```

Answer
The median absolute deviation of the eruption duration is 0.9510879

Exercise
Find the median absolute deviation of the eruption waiting periods in faithful.

Covariance

The covariance of two variables *x* and *y* in a data sample measures how the two are linearly related. A positive covariance would indicates 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 population means μ_x, μ_y as:

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

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

Solution
We apply the *cov* function to compute the covariance of eruptions and waiting.

```
> duration = faithful$eruptions   # the eruption durations
> waiting = faithful$waiting      # the waiting period
> cov(duration, waiting)          # apply the cov function
[1] 13.978
```

Answer
The covariance of the eruption duration and waiting time is 13.978. It indicates a positive linear relationship between the two variables.

Correlation Coefficient

The correlation coefficient of two variables in a data sample is 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 indicates 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 indicates a weak linear relationship between the variables.

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

Solution
We apply the cor function to compute the correlation coefficient of eruptions and waiting.

```
> duration = faithful$eruptions   # the eruption durations
> waiting = faithful$waiting      # the waiting period
> cor(duration, waiting)          # apply the cor function
[1] 0.90081
```

Answer
The correlation coefficient of the eruption duration and waiting time is 0.90081. Since it is close to 1, we can conclude that the variables are positively linearly related.

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:

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

For example, the second central moment of a population is its variance.

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

Solution
We apply the moment function in the moments package. As it is not in the core R library, the package has to be installed and loaded into the R workspace.

```
> library(moments)                    # load the moments package
> duration = faithful$eruptions       # the eruption durations
> moment(duration, order=3, central=TRUE)
[1] −0.6149
```

Answer
The third central moment of the eruption duration is -0.6149.

```
> moment(duration, order=1)
[1] 3.487783
```

Exercise
Find the third central moment of the eruption waiting periods in faithful.

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.

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

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 indicates that the mean of the data values is larger than the median, and the data distribution is right-skewed. Of course, this rule applies only to unimodal distributions whose histograms have a single peak.

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

Solution
We apply the skewness function in the moments package to compute the skewness coefficient of eruptions. As the package is not in the core R library, it has to be installed and loaded into the R workspace.

```
> library(moments)                # load the moments package
> duration = faithful$eruptions   # the eruption durations
> skewness(duration)              # apply the skewness function
[1] -0.41584
```

Answer
The skewness of the eruption duration is -0.41584. It indicates that the eruption duration distribution is skewed towards the left.

Exercise
Find the skewness of the eruption waiting periods in faithful.

Note
The skewness function in the moments package is based on the formula $g_1 = m_3/m_2^{3/2}$, where m_2 and m_3 are the second and third sample central moments. Besides being prone to rounding errors, the function provides only a biased estimate of the corresponding population statistics γ_1.

Kurtosis

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

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$$

Intuitively, the kurtosis is a measure of the peakedness of the data distribution. Negative kurtosis would indicates a flat distribution, which is said to be platykurtic. Positive kurtosis would indicates a peaked distribution, which is said to be leptokurtic. Finally, the normal distribution has zero kurtosis, and is said to be mesokurtic.

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

Solution
We apply the kurtosis function in the moments 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(moments)                # load the moments package
> duration = faithful$eruptions   # the eruption durations
> kurtosis(duration) - 3          # apply the kurtosis function
[1] -1.5006
```

Answer
The kurtosis of the eruption duration is -1.5006. It suggests that the eruption duration distribution is platykurtic, in consistent with the fact that its histogram is not bell-shaped.

Exercise
Find the kurtosis of the eruption waiting periods in faithful.

Note
The kurtosis function in the moments package is based on the formula $g_2 = m_4/m_2^2$, where m_2 and m_4 are the second and fourth sample central moments.