MAP 2112 – Introdução à Lógica de Programação e Modelagem Computacional

1º Semestre - 2020

Prof. Dr. Luis Carlos de Castro Santos

lsantos@ime.usp.br

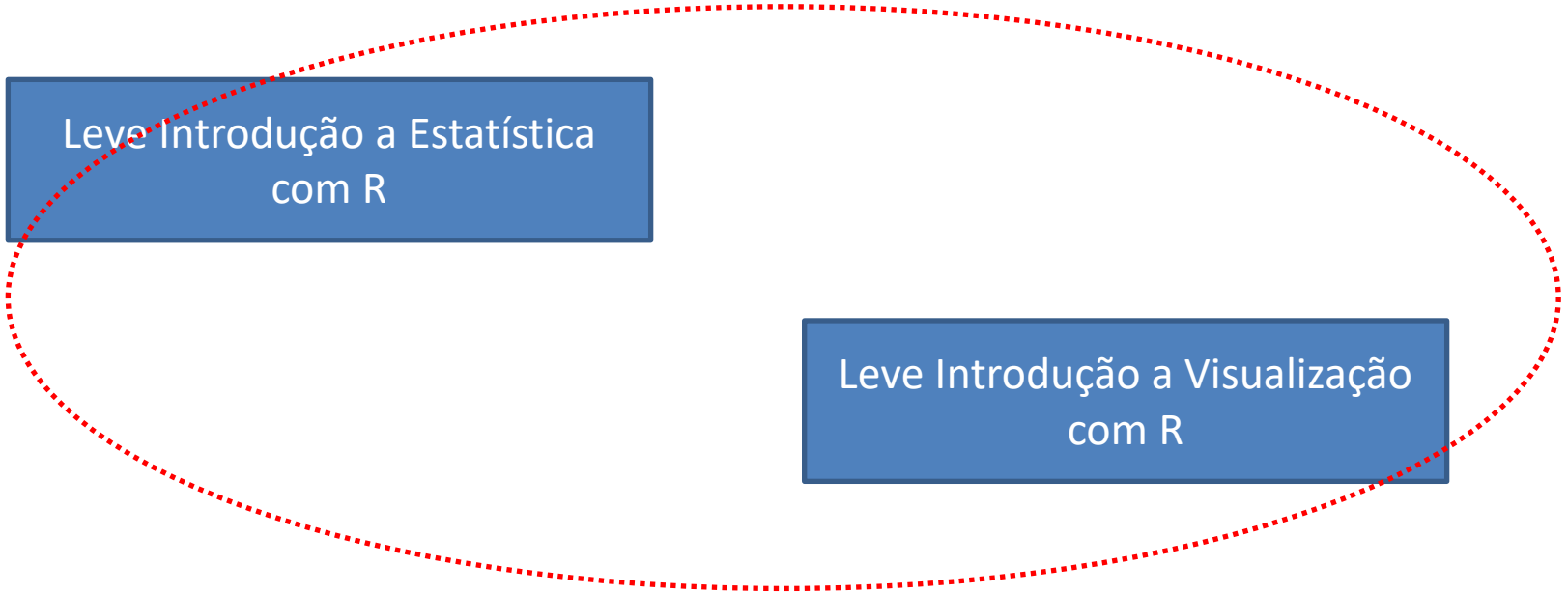
NÃO ESQUEÇA DE INICIAR A GRAVAÇÃO

MAP 2112 – Introdução à Lógica de Programação e Modelagem Computacional

1º Semestre - 2020

Prof. Dr. Luis Carlos de Castro Santos

lsantos@ime.usp.br



Projeto de Data Science Individual

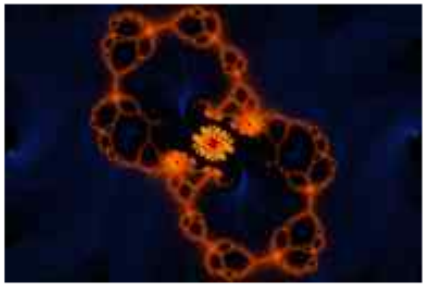
Projeto de [illegible]cience em [illegible]

R Tutorial

An R Introduction to Statistics

HOME DOWNLOAD SALES

Elementary Statistics with R



Ever wonder how to finish your statistics homework real fast? Or you just want a quick way to verify your tedious calculations in your statistics class assignment. We provide an answer here by solving statistics exercises with R.

Here, you will find statistics problems similar to those found in popular college textbooks. The R solutions are short, self-contained and requires minimal R skill. Most of them are just a few lines in length. With simple modifications, the code samples can be turned into homework answers. In additional to helping with your homework, the tutorials will give you a taste of working with statistics software in general, and it will prove invaluable in the success of your career.

We have included separate introductory tutorials for basic R concepts. The topics are by no means comprehensive. Nevertheless, even if you are not familiar with R, you can go through just the first *R Introduction* page. Then go straight to the statistics tutorials, and only come back for reference as needed.

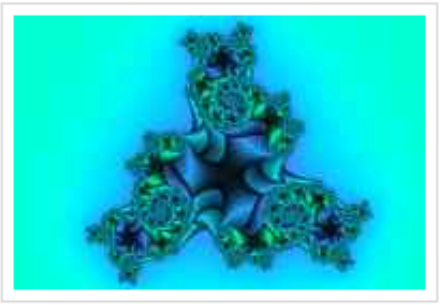
http://www.r-tutor.com/elementary-statistics

O objetivo desse material é auxiliar na realização dos trabalhos. Não haverá cobrança em prova.

Elementary Statistics with R

- Qualitative Data
- Quantitative Data
- Numerical Measures
- Probability Distributions
- Interval Estimation
- Hypothesis Testing
- Type II Error
- Inference About Two Populations
- Goodness of Fit
- Analysis of Variance
- Non-parametric Methods
- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression

Qualitative Data



A data sample is called **qualitative**, also known as **categorical**, if its values belong to a collection of known defined non-overlapping classes. Common examples include student letter grade (A, B, C, D or F), commercial bond rating (AAA, AAB, ...) and consumer clothing shoe sizes (1, 2, 3, ...).

The tutorials in this section are based on an R built-in data frame named **painters**. It is a compilation of technical information of a few eighteenth century classical painters. The data set belongs to the MASS package, and has to be pre-loaded into the R workspace prior to its use.

```
> library(MASS)      # load the MASS package
> painters
             Composition Drawing Colour Expression School
Da Udine              10       8     16          3      A
Da Vinci              15      16      4         14      A
Del Piombo             8      13     16          7      A
Del Sarto             12      16      9          8      A
Fr. Penni              0      15      8          0      A
Guilio Romano         15      16      4         14      A
                   ..................
```

The last School column contains the information of school classification of the painters. The schools are named as A, B, ..., *etc*, and the School variable is qualitative.

```
> painters$School
 [1] A A A A A A A A A A B B B B B B C C C C C C D D D D
[27] D D D D D D E E E E E E E F F F F G G G G G G G H H
[53] H H
Levels: A B C D E F G H
```

Frequency Distribution of Qualitative Data

The **frequency distribution** of a data variable is a summary of the data occurrence in a collection of non-overlapping categories.

Example

In the data set painters, the frequency distribution of the School variable is a summary of the number of painters in each school.

Problem

Find the frequency distribution of the painter schools in the data set painters.

Solution

We apply the table function to compute the frequency distribution of the School variable.

```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> school.freq = table(school)   # apply the table function
```

Answer

The frequency distribution of the schools is:

```
> school.freq
school
 A  B  C  D  E  F  G  H
10  6  6 10  7  4  7  4
```

Relative Frequency Distribution of Qualitative Data

The **relative frequency distribution** of a data variable is a summary of the frequency proportion in a collection of non-overlapping categories.

The relationship of frequency and relative frequency is:

$$Relative\ Frequency = \frac{Frequency}{Sample\ Size}$$

Example

In the data set painters, the relative frequency distribution of the School variable is a summary of the proportion of painters in each school.

Problem

Find the relative frequency distribution of the painter schools in the data set painters.

```
> school.relfreq = school.freq / nrow(painters)
```

Answer

The relative frequency distribution of the schools is:

```
> school.relfreq
school
       A        B        C        D        E        F
0.185185 0.111111 0.111111 0.185185 0.129630 0.074074
       G        H
0.129630 0.074074
```

Bar Graph

A **bar graph** of a qualitative data sample consists of vertical parallel bars that shows the frequency distribution graphically.

Example

In the data set painters, the bar graph of the School variable is a collection of vertical bars showing the number of painters in each school.

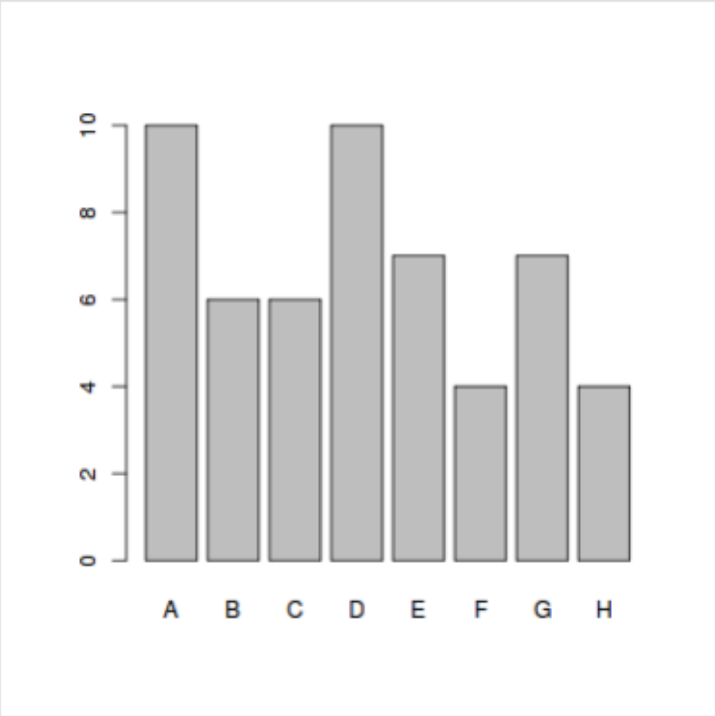
Problem

Find the bar graph of the painter schools in the data set painters.

```
> barplot(school.freq)          # apply the barplot function
```

Answer

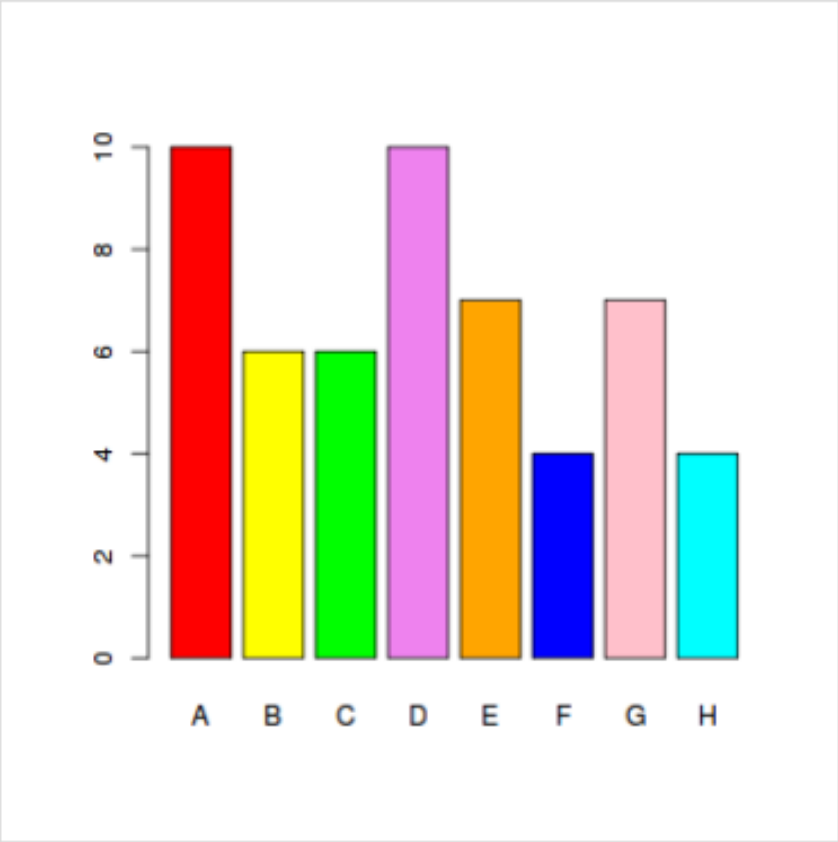
The bar graph of the school variable is:



Enhanced Solution

To colorize the bar graph, we select a color palette and set it in the col argument of barplot.

```
> colors = c("red", "yellow", "green", "violet",
+   "orange", "blue", "pink", "cyan")
> barplot(school.freq,          # apply the barplot function
+   col=colors)                 # set the color palette
```



Exercise

Find the bar graph of the composition scores in painters.

Pie Chart

A **pie chart** of a qualitative data sample consists of pizza wedges that shows the frequency distribution graphically.

Example

In the data set painters, the pie chart of the School variable is a collection of pizza wedges showing the proportion of painters in each school.

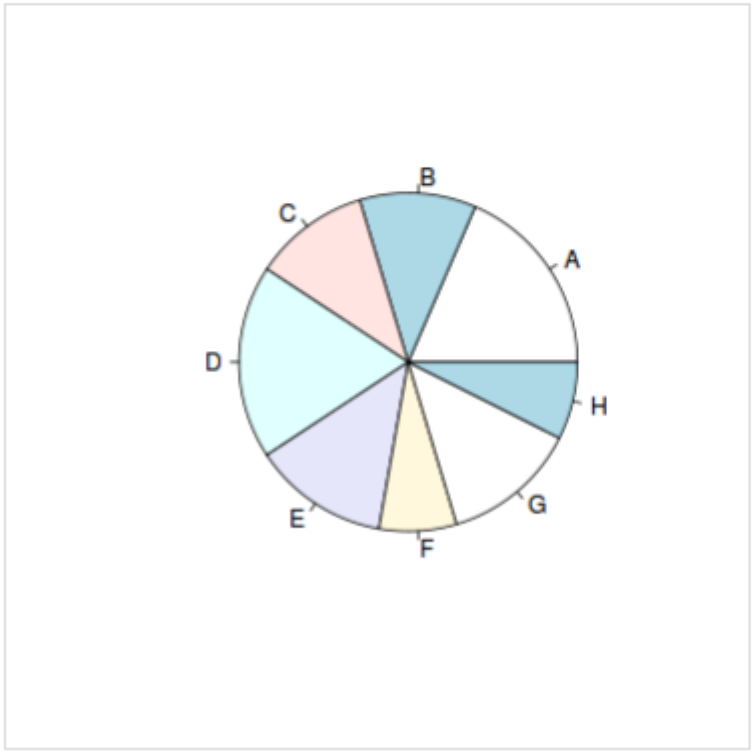
Problem

Find the pie chart of the painter schools in the data set painters.

```
> pie(school.freq)              # apply the pie function
```

Answer

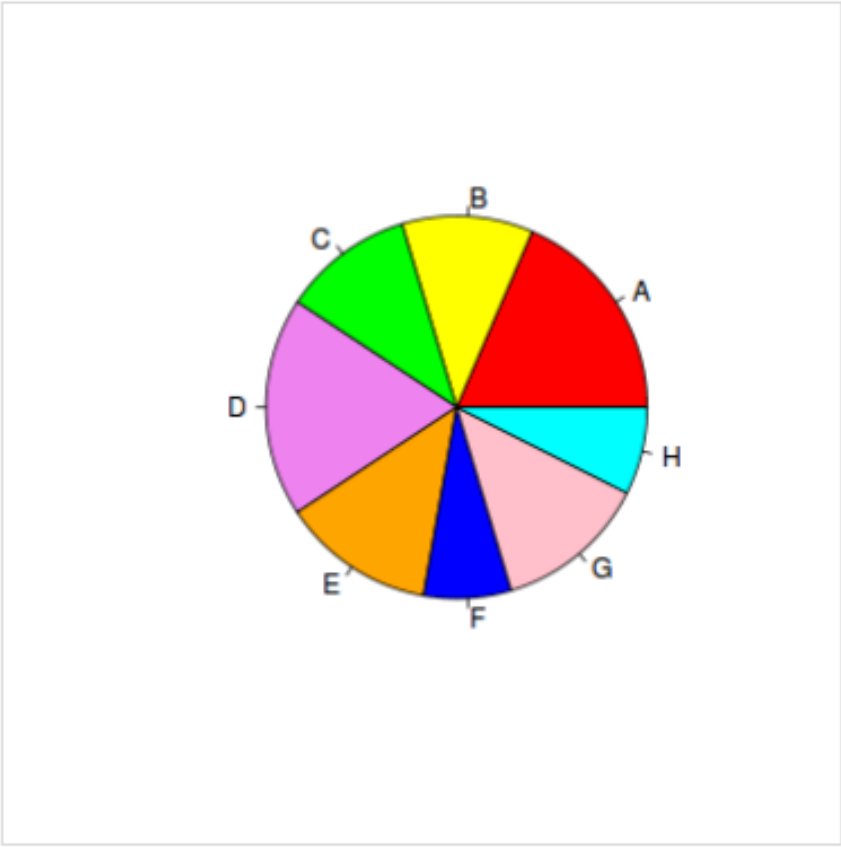
The pie chart of the school variable is:



Enhanced Solution

To colorize the pie chart, we select a color palette and set it in the col argument of pie.

```
> colors = c("red", "yellow", "green", "violet",
+   "orange", "blue", "pink", "cyan")
> pie(school.freq,              # apply the pie function
+   col=colors)                 # set the color palette
```



Exercise

Find the pie chart of the composition scores in painters.

Category Statistics

In the built-in data set painters, the painters are classified according to the schools they belong. Each school can be characterized by its various statistics, such as mean composition, drawing, coloring and expression scores.

Suppose we would like to know which school has the highest mean composition score. We would have to first find out the mean composition score of each school. The following shows how to find the mean composition score of an arbitrarily chosen school.

Problem

Find out the mean composition score of school C in the data set painters.

Solution

The solution consists of a few steps:

1. Create a logical index vector for school C.

```
> library(MASS)                 # load the MASS package
> school = painters$School      # the painter schools
> c_school = school == "C"      # the logical index vector
```

2. Find the child data set of painters for school C. For explanation, please consult the tutorial of *Data Frame Row Slice*.

```
> c_painters = painters[c_school, ]  # child data set
```

3. Find the mean composition score of school C.

```
> mean(c_painters$Composition)
[1] 13.167
```

Answer

The mean composition score of school C is 13.167.

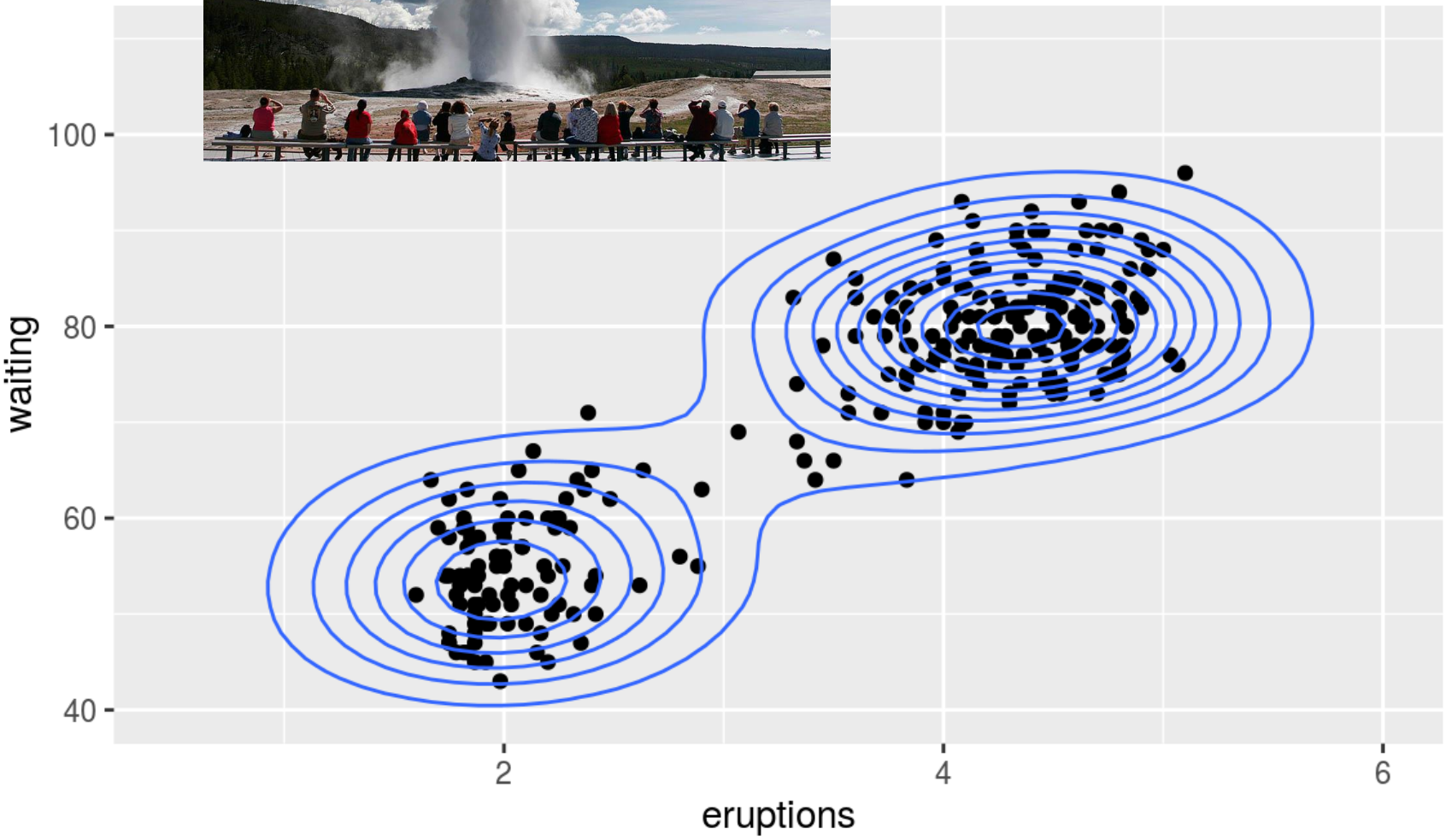
Quantitative Data

Quantitative data, also known as **continuous** data, consists of numeric data that support arithmetic operations. This is in contrast with qualitative data, whose values belong to pre-defined classes with no arithmetic operation allowed. We will explain how to apply some of the R tools for quantitative data analysis with examples.

The tutorials in this section are based on a built-in data frame named **faithful**. It consists of a collection of observations of the Old Faithful geyser in the USA Yellowstone National Park. The following is a preview via the head function.

```
> head(faithful)
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

There are two observation variables in the data set. The first one, called eruptions, is the duration of the geyser eruptions. The second one, called waiting, is the length of waiting period until the next eruption. It turns out there is a correlation between the two variables, as shown in the *Scatter Plot* tutorial.



https://www.ggplot2-exts.org/ggalt.html

Frequency Distribution of Quantitative Data

The **frequency distribution** of a data variable is a summary of the data occurrence in a collection of non-overlapping categories.

Example

In the data set faithful, the frequency distribution of the eruptions variable is the summary of eruptions according to some classification of the eruption durations.

Problem

Find the frequency distribution of the eruption durations in faithful.

Solution

The solution consists of the following steps:

1. We first find the range of eruption durations with the range function. It shows that the observed eruptions are between 1.6 and 5.1 minutes in duration.

```
> duration = faithful$eruptions
> range(duration)
[1] 1.6 5.1
```

2. Break the range into non-overlapping sub-intervals by defining a sequence of equal distance break points. If we round the endpoints of the interval [1.6, 5.1] to the closest half-integers, we come up with the interval [1.5, 5.5]. Hence we set the break points to be the half-integer sequence { 1.5, 2.0, 2.5, ... }.

```
> breaks = seq(1.5, 5.5, by=0.5)    # half-integer sequence
> breaks
[1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

3. Classify the eruption durations according to the half-unit-length sub-intervals with cut. As the intervals are to be closed on the left, and open on the right, we set the right argument as FALSE.

```
> duration.cut = cut(duration, breaks, right=FALSE)
```

4. Compute the frequency of eruptions in each sub-interval with the table function.

```
> duration.freq = table(duration.cut)
```

Answer

The frequency distribution of the eruption duration is:

```
> duration.freq
duration.cut
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
     51      41       5       7      30      73      61
[5,5.5)
      4
```

Enhanced Solution

We apply the cbind function to print the result in column format.

```
> cbind(duration.freq)
        duration.freq
[1.5,2)            51
[2,2.5)            41
[2.5,3)             5
[3,3.5)             7
[3.5,4)            30
[4,4.5)            73
[4.5,5)            61
[5,5.5)             4
```

Note

Per R documentation, you are advised to use the hist function to find the frequency distribution for performance reasons.

Exercise

1. Find the frequency distribution of the eruption waiting periods in faithful.
2. Find programmatically the duration sub-interval that has the most eruptions.

Histogram

A **histogram** consists of parallel vertical bars that graphically shows the frequency distribution of a quantitative variable. The area of each bar is equal to the frequency of items found in each class.

Example

In the data set faithful, the histogram of the eruptions variable is a collection of parallel vertical bars showing the number of eruptions classified according to their durations.

Problem

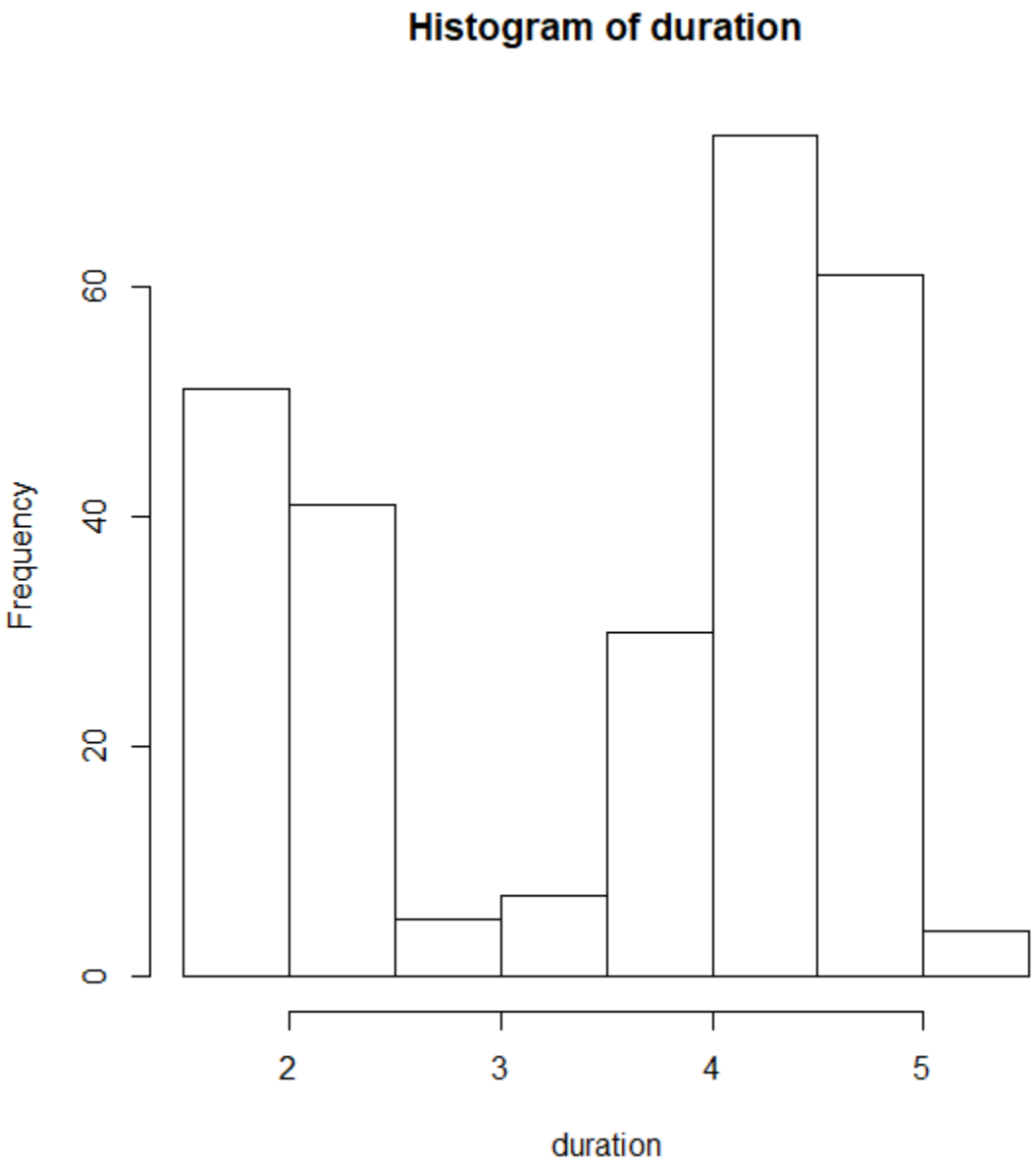
Find the histogram of the eruption durations in faithful.

Solution

We apply the hist function to produce the histogram of the eruptions variable.

```
> duration = faithful$eruptions
> hist(duration,    # apply the hist function
+   right=FALSE)    # intervals closed on the left
```

```
> hist(duration,right=FALSE)
```

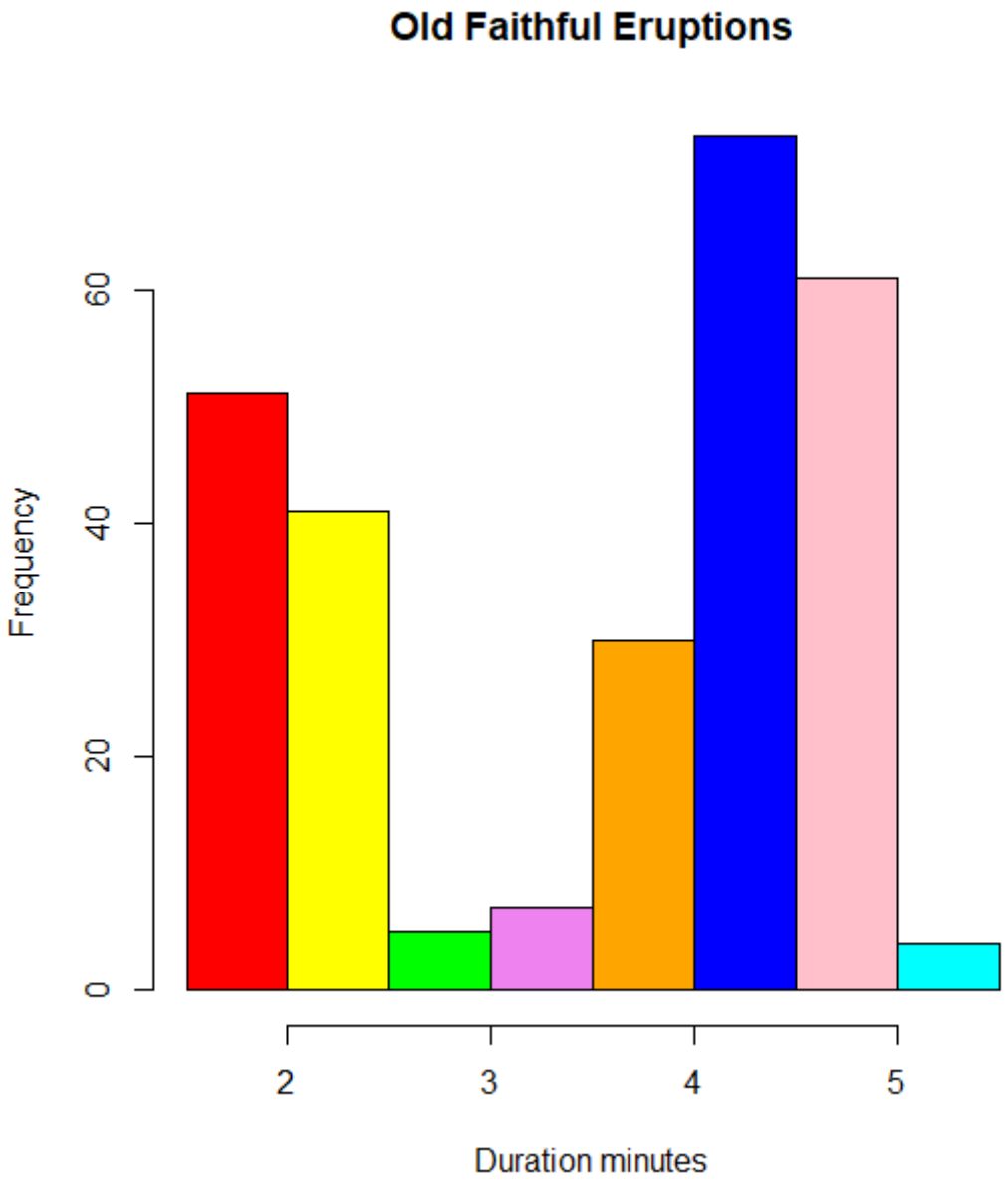


Enhanced Solution

To colorize the histogram, we select a color palette and set it in the col argument of hist. In addition, we update the titles for readability.

```
> colors = c("red", "yellow", "green", "violet", "orange",
+   "blue", "pink", "cyan")
> hist(duration,    # apply the hist function
+   right=FALSE,    # intervals closed on the left
+   col=colors,     # set the color palette
+   main="Old Faithful Eruptions", # the main title
+   xlab="Duration minutes")       # x-axis label
```

```
> hist(duration,right=FALSE,col=colors,main="Old Faithful Eruptions",
 xlab="Duration minutes")
```



Relative Frequency Distribution of Quantitative Data

The **relative frequency distribution** of a data variable is a summary of the frequency proportion in a collection of non-overlapping categories.

The relationship of frequency and relative frequency is:

$$Relative\ Frequency = \frac{Frequency}{Sample\ Size}$$

Example

In the data set faithful, the relative frequency distribution of the eruptions variable shows the frequency proportion of the eruptions according to a duration classification.

Problem

Find the relative frequency distribution of the eruption durations in faithful.

```
> duration.relfreq = duration.freq / nrow(faithful)
```

Answer

The frequency distribution of the eruption variable is:

```
> duration.relfreq
duration.cut
 [1.5,2)  [2,2.5)  [2.5,3)  [3,3.5)  [3.5,4)  [4,4.5)
0.187500 0.150735 0.018382 0.025735 0.110294 0.268382
 [4.5,5)  [5,5.5)
0.224265 0.014706
```

Enhanced Solution

We can print with fewer digits and make it more readable by setting the digits option.

```
> old = options(digits=1)
> duration.relfreq
duration.cut
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
   0.19    0.15    0.02    0.03    0.11    0.27    0.22
[5,5.5)
   0.01
> options(old)    # restore the old option
```

We then apply the cbind function to print both the frequency distribution and relative frequency distribution in parallel columns.

```
> old = options(digits=1)
> cbind(duration.freq, duration.relfreq)
        duration.freq duration.relfreq
[1.5,2)            51              0.19
[2,2.5)            41              0.15
[2.5,3)             5              0.02
[3,3.5)             7              0.03
[3.5,4)            30              0.11
[4,4.5)            73              0.27
[4.5,5)            61              0.22
[5,5.5)             4              0.01
> options(old)    # restore the old option
```

Exercise

Find the relative frequency distribution of the eruption waiting periods in faithful.

Cumulative Frequency Distribution

The **cumulative frequency distribution** of a quantitative variable is a summary of data frequency below a given level.

Example

In the data set faithful, the cumulative frequency distribution of the eruptions variable shows the *total* number of eruptions whose durations are less than or equal to a set of chosen levels.

Problem

Find the cumulative frequency distribution of the eruption durations in faithful.

```
> duration.cumfreq = cumsum(duration.freq)
```

Answer

The cumulative distribution of the eruption duration is:

```
> duration.cumfreq
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
     51      92      97     104     134     207     268
[5,5.5)
    272
```

Cumulative Frequency Graph

A **cumulative frequency graph** or **ogive** of a quantitative variable is a curve graphically showing the cumulative frequency distribution.

Example

In the data set faithful, a point in the cumulative frequency graph of the eruptions variable shows the *total* number of eruptions whose durations are less than or equal to a given level.

Problem

Find the cumulative frequency graph of the eruption durations in faithful.

Solution

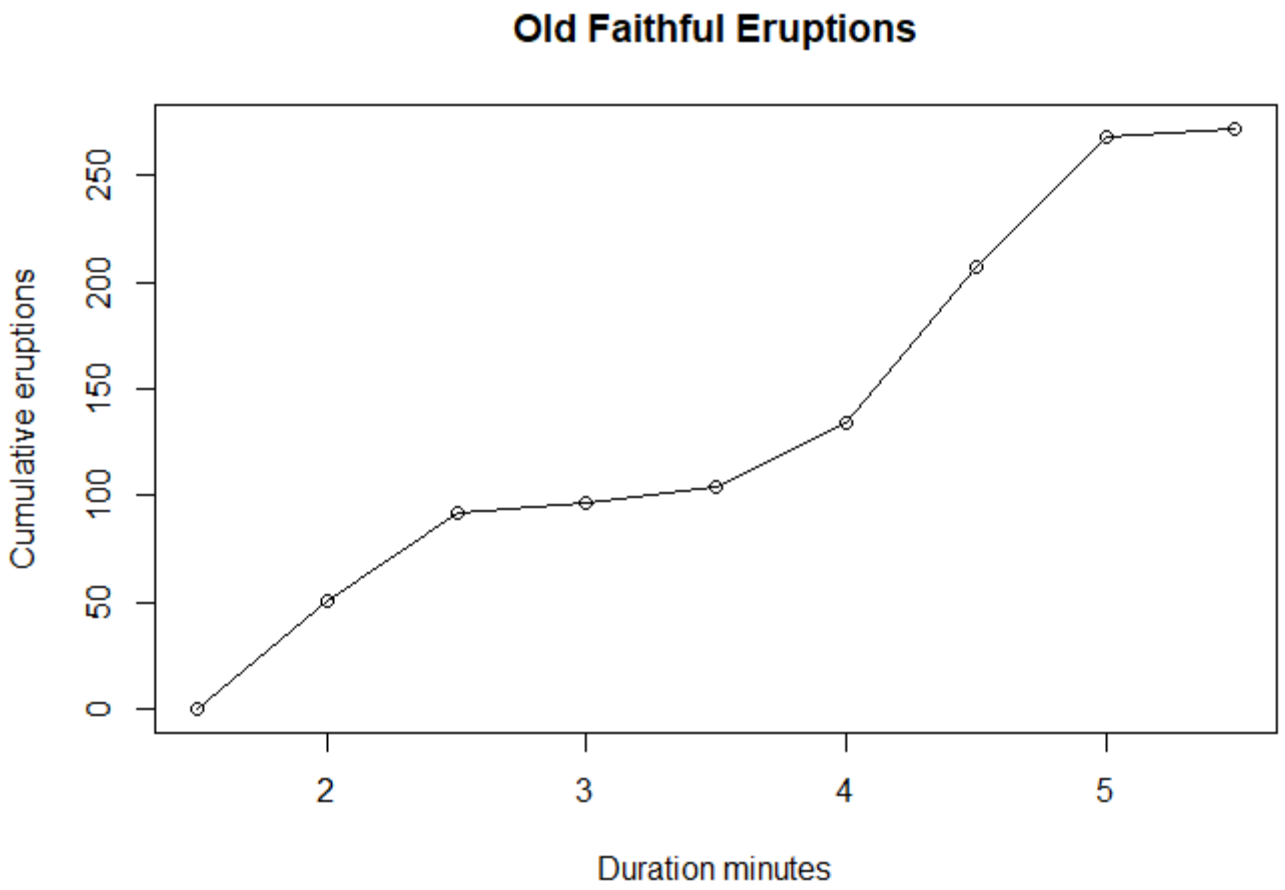
We first find the frequency distribution of the eruption durations as follows. Check the previous tutorial on *Frequency Distribution* for details.

```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
```

We then compute its cumulative frequency with cumsum, add a starting zero element, and plot the graph.

```
> cumfreq0 = c(0, cumsum(duration.freq))
> plot(breaks, cumfreq0,             # plot the data
+   main="Old Faithful Eruptions",  # main title
+   xlab="Duration minutes",        # x-axis label
+   ylab="Cumulative eruptions")   # y-axis label
> lines(breaks, cumfreq0)           # join the points
```

```
> plot(breaks, cumfreq0,main="Old Faithful Eruptions", xlab="Duration minutes",
ylab="Cumulative eruptions")
>
> lines(breaks, cumfreq0)
>
```



Exercise

Find the cumulative frequency graph of the eruption waiting periods in faithful.

Cumulative Relative Frequency Distribution

The **cumulative relative frequency distribution** of a quantitative variable is a summary of frequency proportion below a given level.

The relationship between cumulative frequency and relative cumulative frequency is:

$$Cumulative\ Relative\ Frequency = \frac{Cumulative\ Frequency}{Sample\ Size}$$

Example

In the data set faithful, the cumulative relative frequency distribution of the eruptions variable shows the frequency proportion of eruptions whose durations are less than or equal to a set of chosen levels.

Problem

Find the cumulative relative frequency distribution of the eruption durations in faithful.

Answer

The cumulative relative frequency distribution of the eruption variable is:

```
> duration.cumrelfreq
[1.5,2) [2,2.5) [2.5,3) [3,3.5) [3.5,4) [4,4.5) [4.5,5)
0.18750 0.33824 0.35662 0.38235 0.49265 0.76103 0.98529
[5,5.5)
1.00000
```

Cumulative Relative Frequency Graph

A **cumulative relative frequency graph** of a quantitative variable is a curve graphically showing the cumulative relative frequency distribution.

Example

In the data set faithful, a point in the cumulative relative frequency graph of the eruptions variable shows the frequency proportion of eruptions whose durations are less than or equal to a given level.

Problem

Find the cumulative relative frequency graph of the eruption durations in faithful.

Solution

We first find the cumulative relative frequency distribution of the eruption durations as follows. Check the previous tutorial on *Cumulative Relative Frequency Distribution* for details.

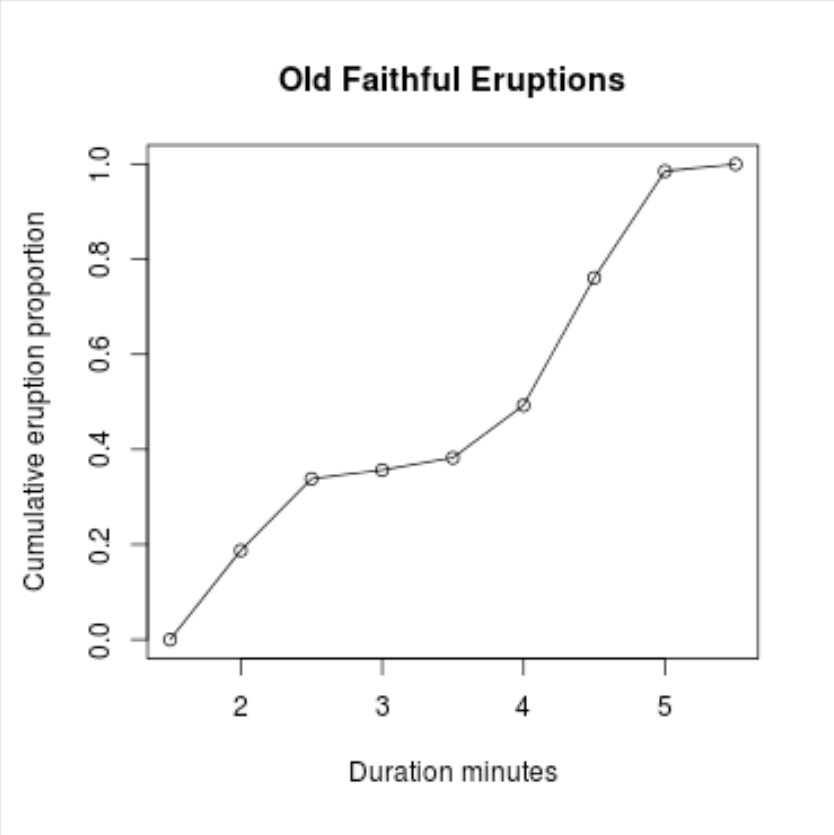
```
> duration = faithful$eruptions
> breaks = seq(1.5, 5.5, by=0.5)
> duration.cut = cut(duration, breaks, right=FALSE)
> duration.freq = table(duration.cut)
> duration.cumfreq = cumsum(duration.freq)
> duration.cumrelfreq = duration.cumfreq / nrow(faithful)
```

We then plot it along with the starting zero element.

```
> cumrelfreq0 = c(0, duration.cumrelfreq)
> plot(breaks, cumrelfreq0,
+   main="Old Faithful Eruptions",  # main title
+   xlab="Duration minutes",
+   ylab="Cumulative eruption proportion")
> lines(breaks, cumrelfreq0)        # join the points
```

Answer

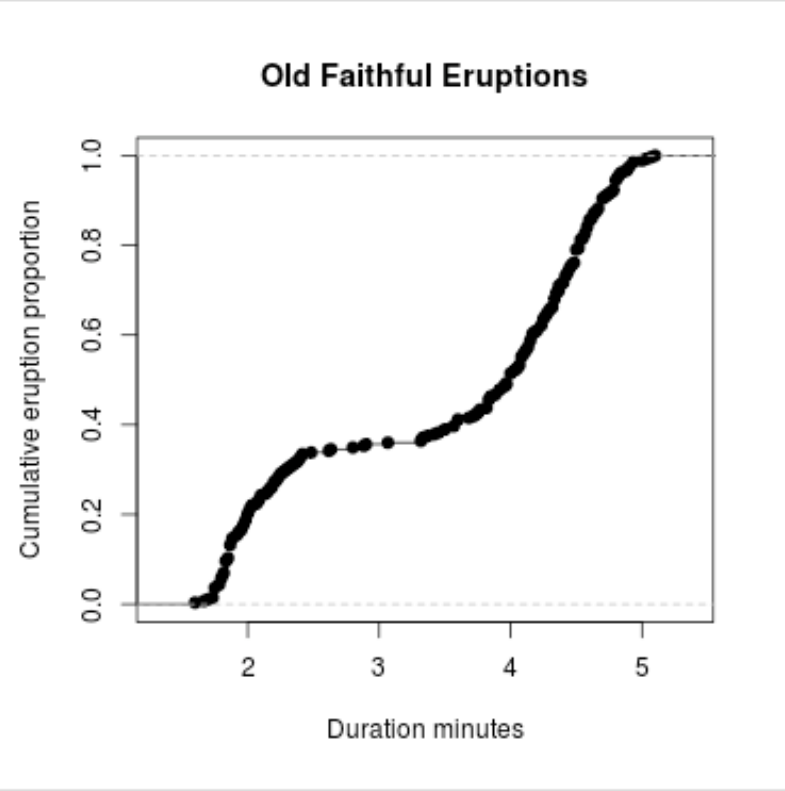
The cumulative relative frequency graph of the eruption duration is:



Alternative Solution

We create an interpolate function Fn with the built-in function ecdf. Then we plot Fn right away. There is no need to compute the cumulative frequency distribution *a priori*.

```
> Fn = ecdf(duration)
> plot(Fn,
+   main="Old Faithful Eruptions",
+   xlab="Duration minutes",
+   ylab="Cumulative eruption proportion")
```



Exercise

Find the cumulative relative frequency graph of the eruption waiting periods in faithful.

Scatter Plot

A **scatter plot** pairs up values of two quantitative variables in a data set and display them as geometric points inside a Cartesian diagram.

Example

In the data set faithful, we pair up the eruptions and waiting values in the same observation as (x,y) coordinates. Then we plot the points in the Cartesian plane. Here is a preview of the eruption data value pairs with the help of the cbind function.

```
> duration = faithful$eruptions      # the eruption durations
> waiting = faithful$waiting         # the waiting interval
> head(cbind(duration, waiting))
     duration waiting
[1,]    3.600      79
[2,]    1.800      54
[3,]    3.333      74
[4,]    2.283      62
[5,]    4.533      85
[6,]    2.883      55
```

Problem

Find the scatter plot of the eruption durations and waiting intervals in faithful. Does it reveal any relationship between the variables?

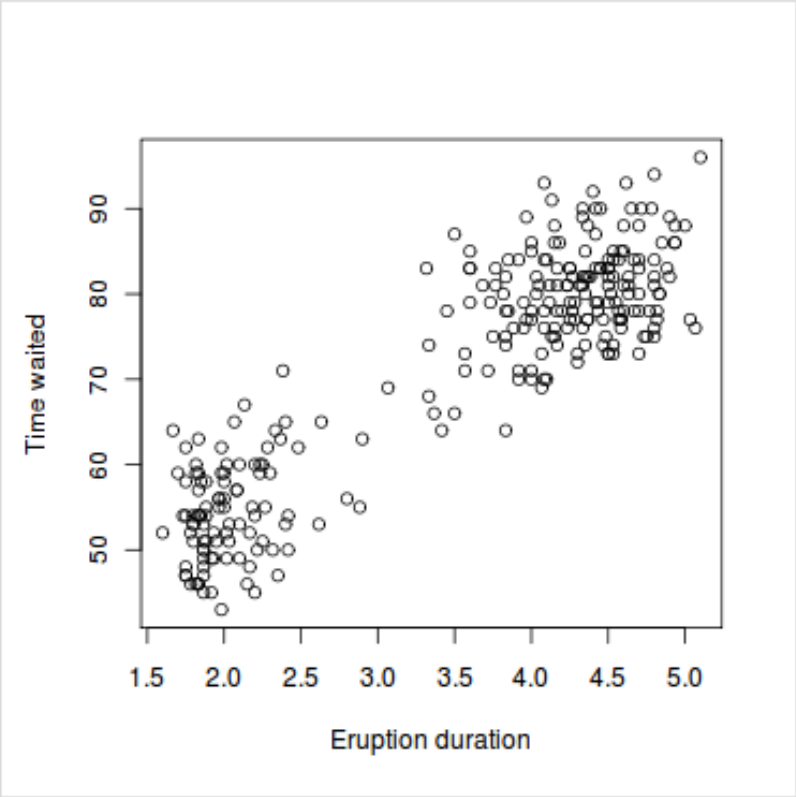
Solution

We apply the plot function to compute the scatter plot of eruptions and waiting.

```
> duration = faithful$eruptions      # the eruption durations
> waiting = faithful$waiting         # the waiting interval
> plot(duration, waiting,            # plot the variables
+   xlab="Eruption duration",        # x-axis label
+   ylab="Time waited")              # y-axis label
```

Answer

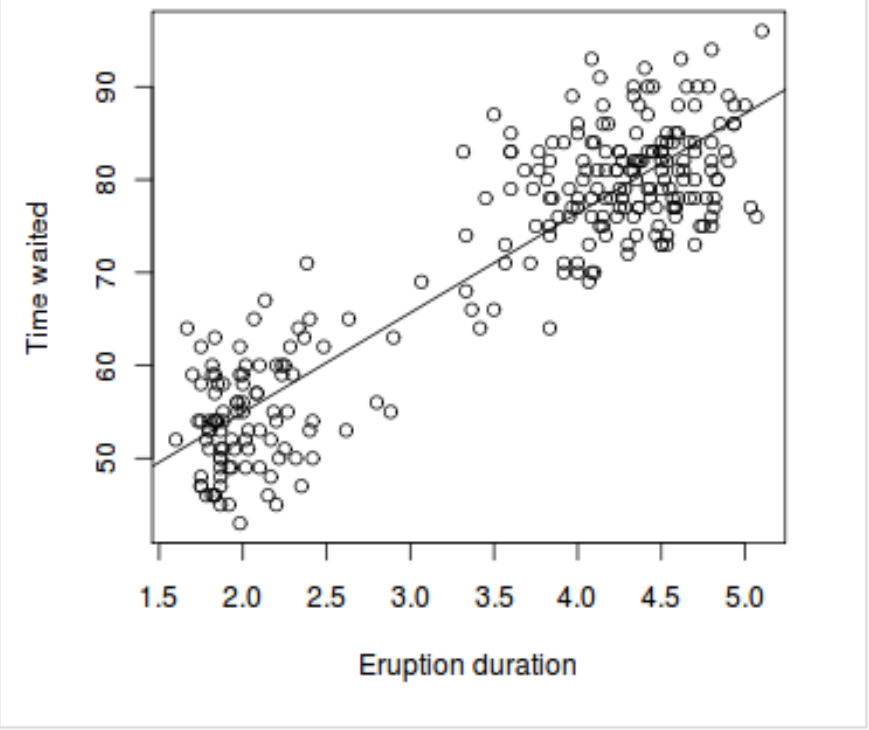
The scatter plot of the eruption durations and waiting intervals is as follows. It reveals a *positive linear relationship* between them.



Enhanced Solution

We can generate a linear regression model of the two variables with the lm function, and then draw a trend line with abline.

```
> abline(lm(waiting ~ duration))
```



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.

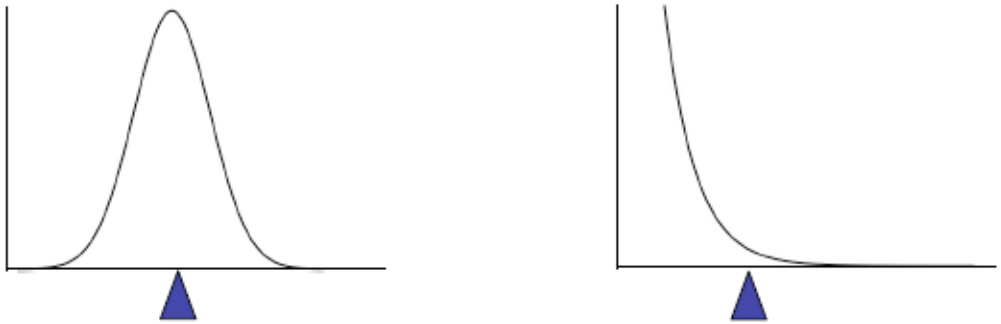
- Mean
- Median
- Quartile
- Percentile
- Range
- Interquartile Range
- Box Plot
- Variance
- Standard Deviation
- Covariance
- Correlation Coefficient
- Central Moment
- Skewness
- Kurtosis

Location: Mean

I. The Mean

To calculate the average $\overline{x}$ of a set of observations, add their value and divide by the number of observations:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf

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:

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

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.

Location: Median

- **Median** – the exact middle value

ímpar

- **Calculation**:

  - If there are an odd number of observations, find the middle value

par

  - If there are an even number of observations, find the middle two values and average them

- **Example**

Some data:

Age of participants: 17 19 21 22 23 23 23 38

Median = (22+23)/2 = 22.5

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
```

Answer

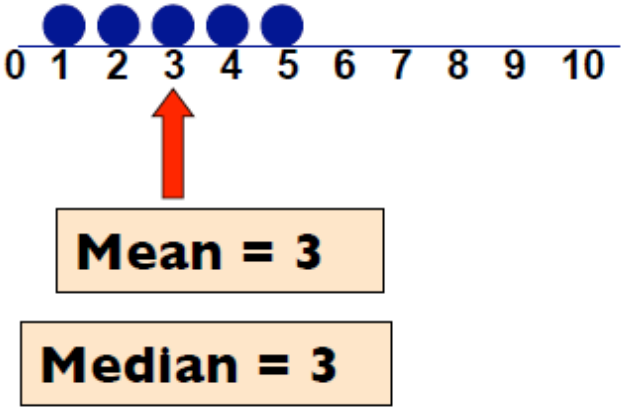
The median of the eruption duration is 4 minutes.

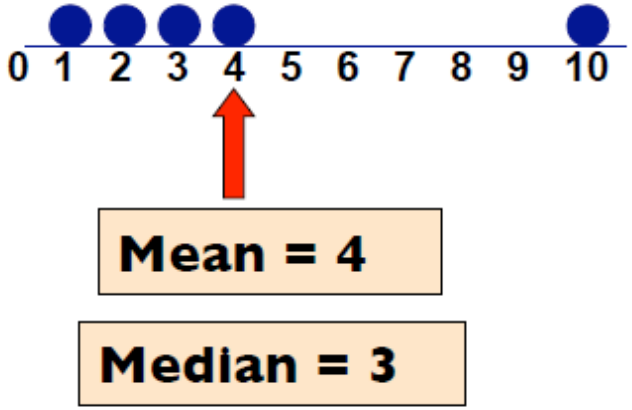
Exercise

Find the median of the eruption waiting periods in faithful.

Which Location Measure Is Best?

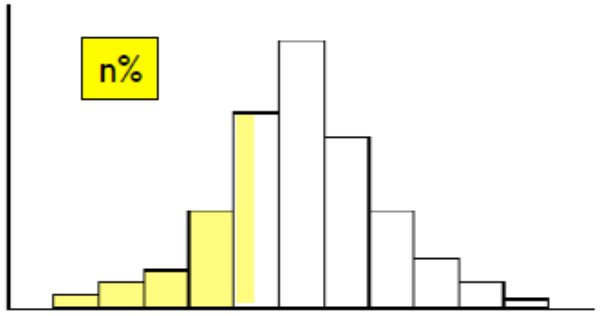
- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers





Percentiles (aka Quantiles)

In general the **n^{th} percentile** is a value such that n% of the observations fall at or below or it



Q_1 = 25th percentile

Median = 50th percentile

Q_2 = 75th percentile

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, 4.1330 and 4.9330 minutes respectively.

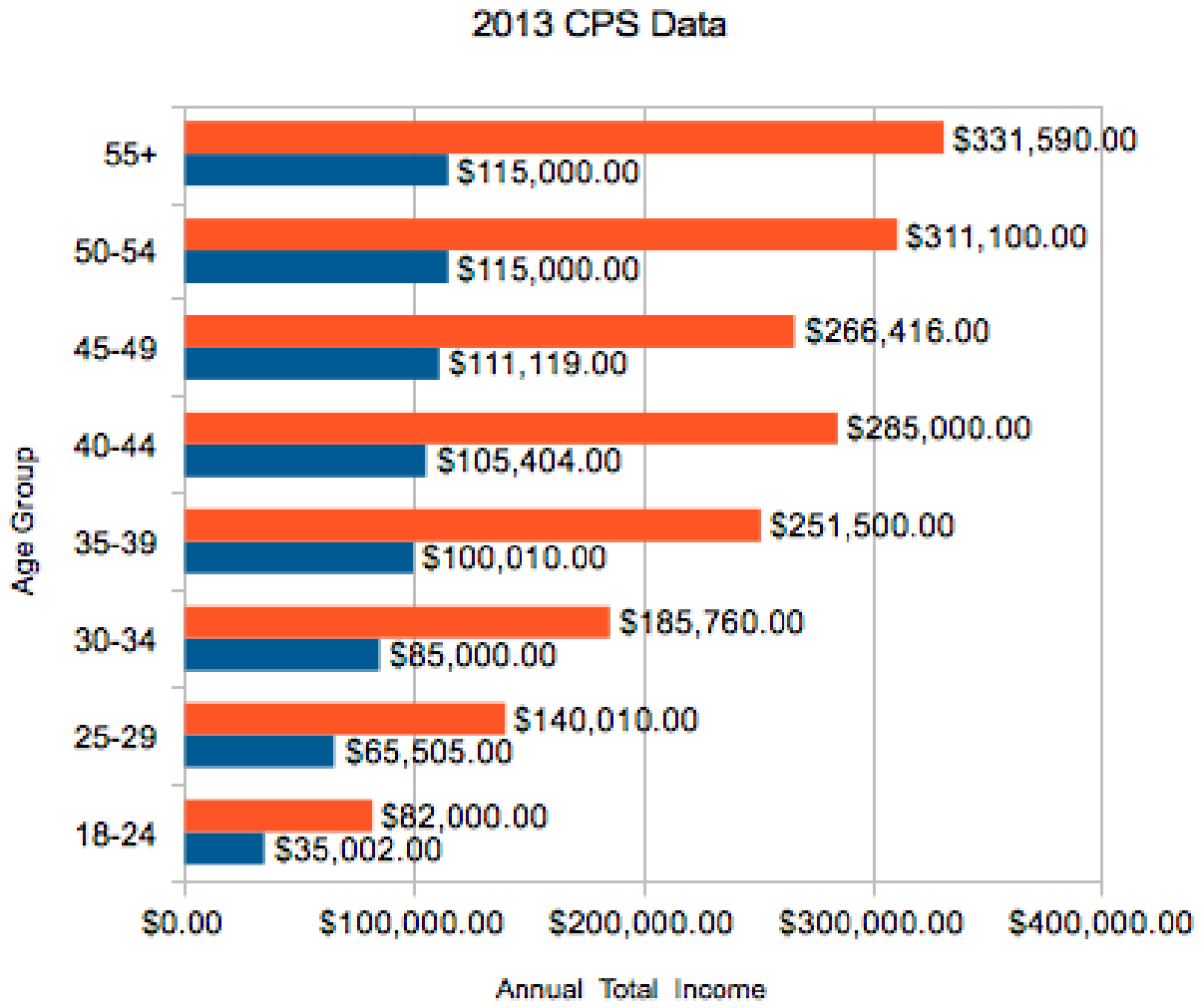
Exercise

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

Note

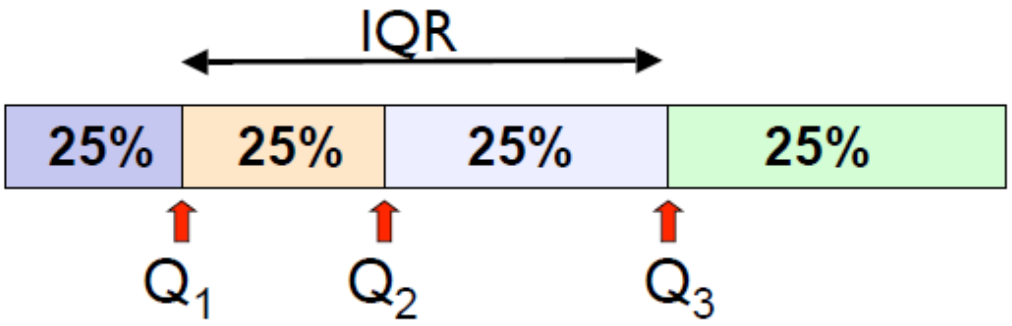
There are several algorithms for the computation of percentiles. Details can be found in the R documentation via help(quantile).

Incomes by Age - The Top Decile



Orange: The 1%
Blue: The 10%
(99th quantile & 90th quantile)

Scale: Quartiles and IQR



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

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.

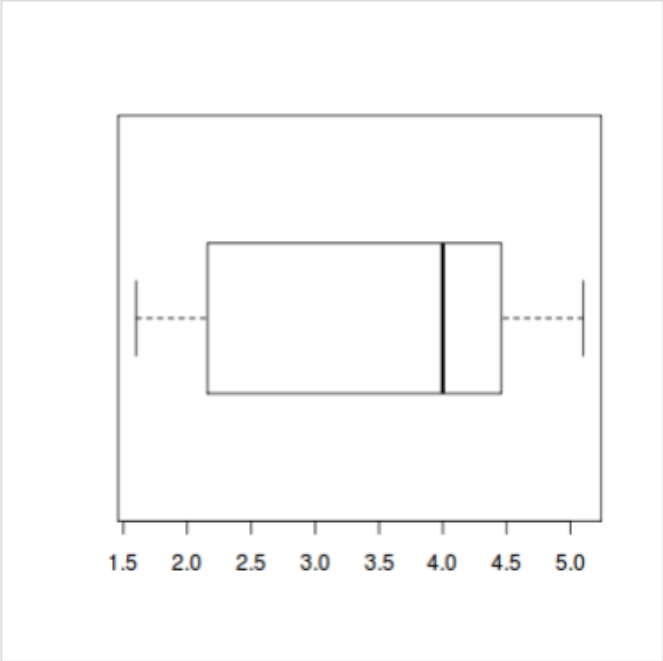
Solution

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
```

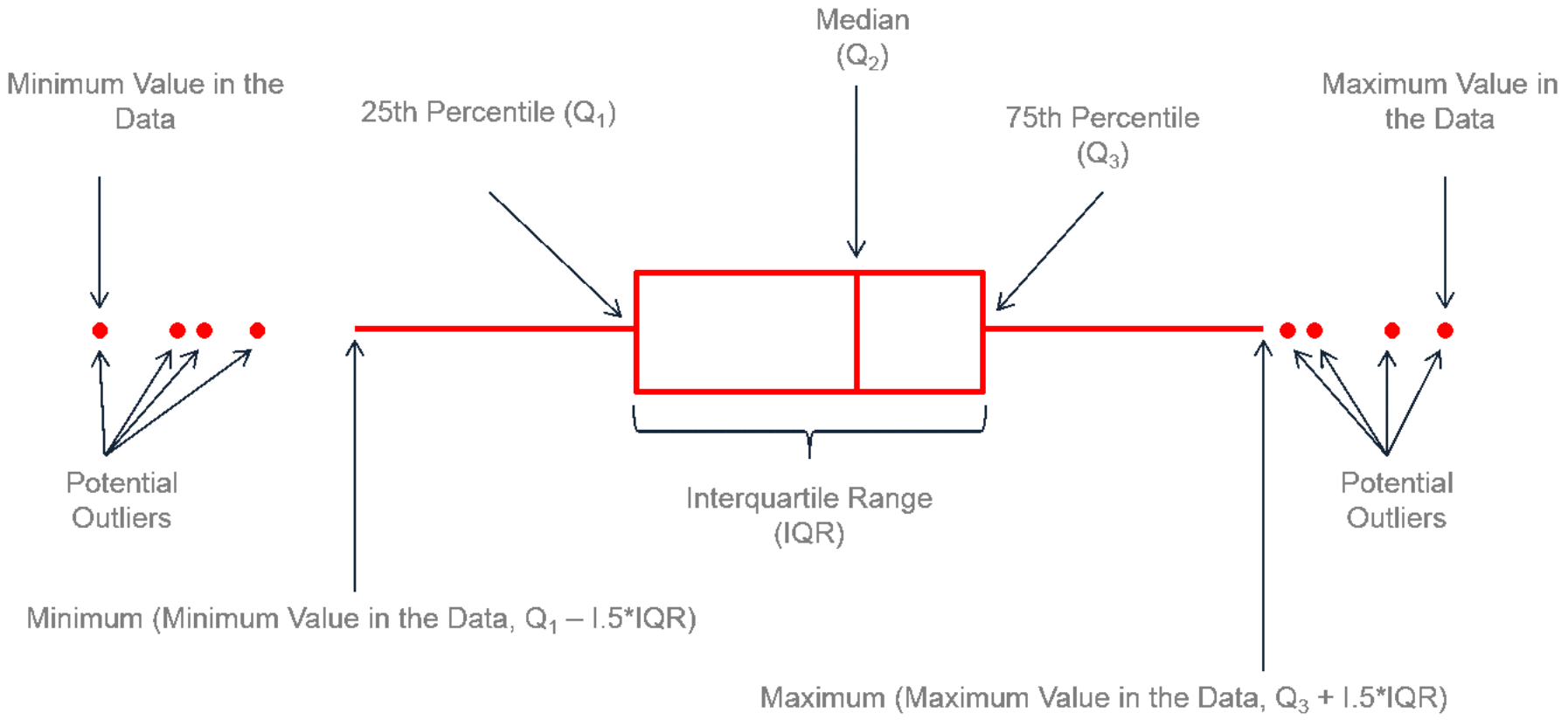
Answer

The box plot of the eruption duration is:



Exercise

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


Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile
(Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range
(IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

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.

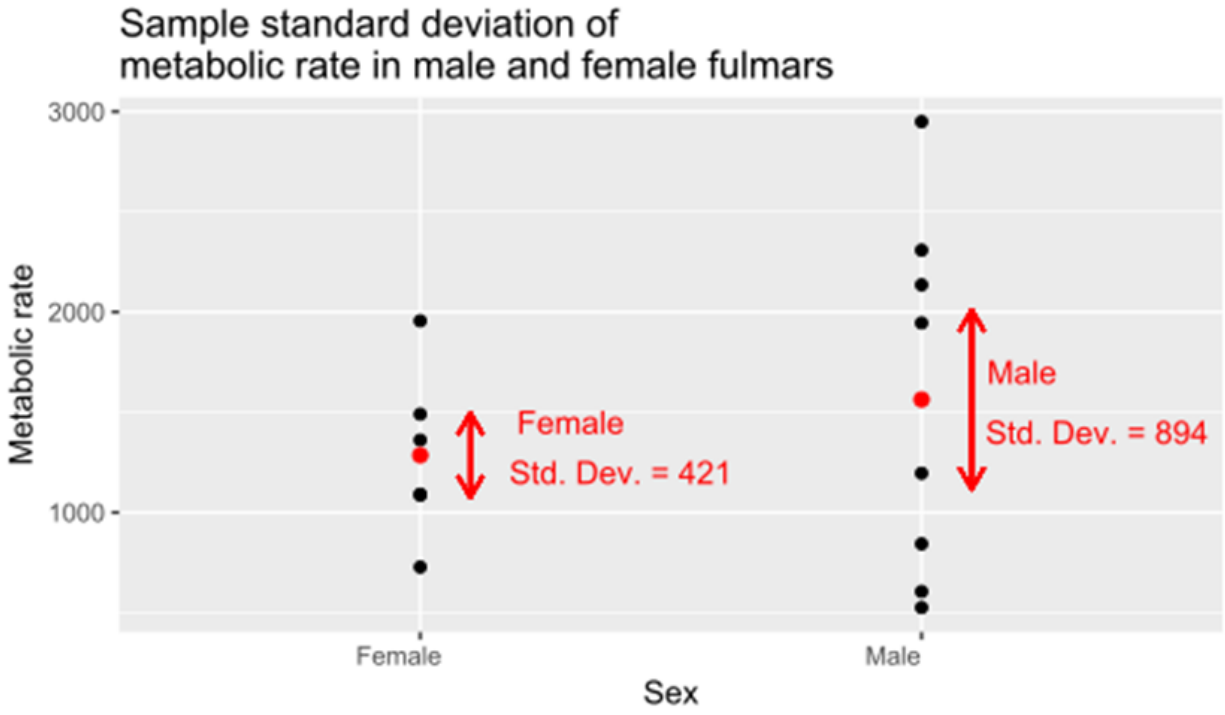
```
> duration = faithful$eruptions    # the eruption durations
> sd(duration)                     # apply the sd function
[1] 1.1414
```

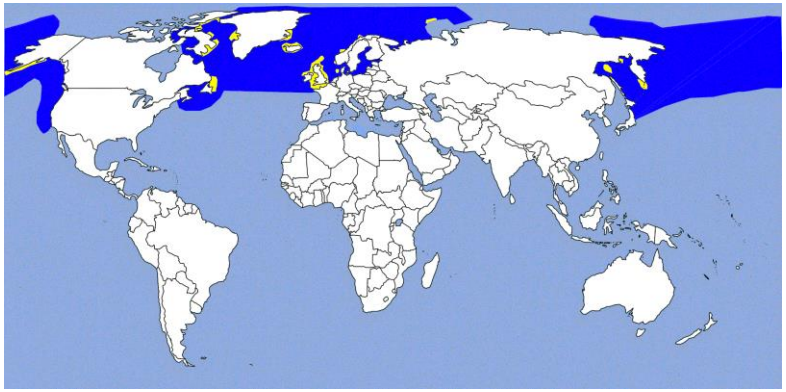
Answer

The standard deviation of the eruption duration is 1.1414.

Exercise

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





https://pt.wikipedia.org/wiki/Fulmar

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

Answer

The covariance of eruption duration and waiting time is about 14. It indicates a positive linear relationship between the two variables.

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.

Problem

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.

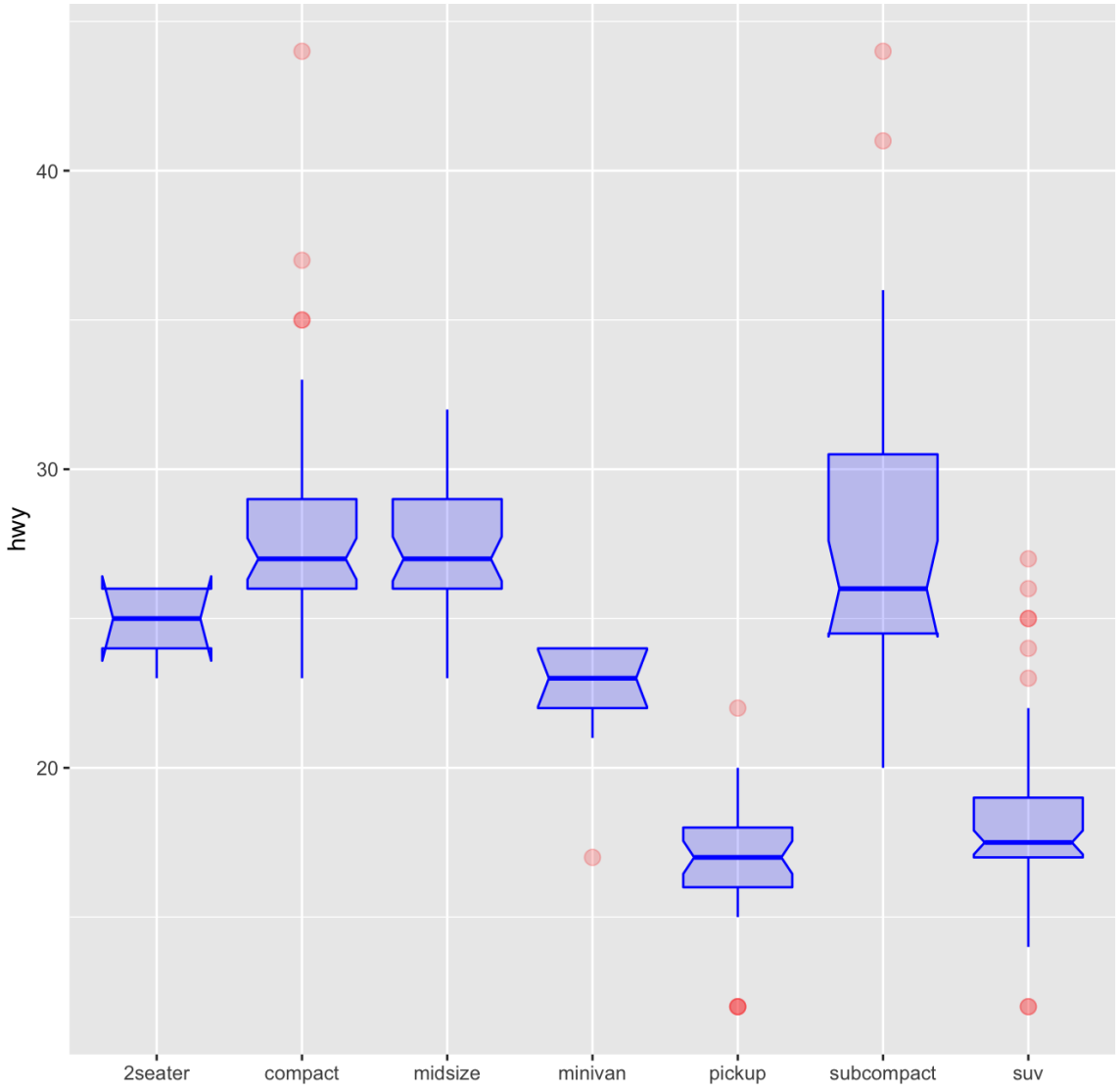
Solution

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

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

Answer

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



https://www.r-graph-gallery.com/263-ggplot2-boxplot-parameters.html