

STU11002

Statistical Analysis I

Dr. Hannah Comiskey

Learning objectives

- ▶ Explain the terms population and samples.
- ▶ Describe the types of data used in sampling
- ▶ Name, construct and explain different data visualisation approaches
- ▶ Calculate the arithmetic and geometric mean for a given sample
- ▶ Define and calculate the median, mode and percentiles for a given sample of data.
- ▶ Interpret and explain a sample of data using boxplots.

Building blocks

Statistics

Statistics revolves around the collection and the analysis (and interpretation) of data. Depending on the focus, two main “approaches” are possible:

- ▶ **Descriptive statistics.** This involves the description and summarization of data with measures, indexes, figures, etc. In general, the aim is at reducing the complexity of the original, observed, data into a series of “simpler” metrics. These still capture the information contained in the original data, but in a more ‘bitesize’ way.
- ▶ **Inferential statistics.** This involves the use of data (and possibly models) to generalize the findings obtained on a sample to the population it refers to.

Population and sample

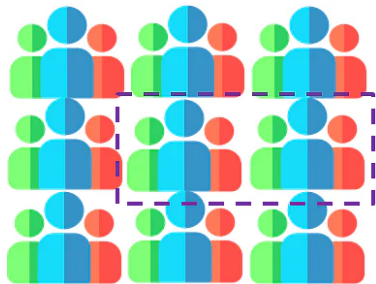
Population

A population coincides with a set of units which have at least one or more common characteristics. Populations are the “ultimate” object of interests in statistical analysis.

Sample

Often times, studying/observing/collecting data for a population in its entirety it's not possible. In such cases, samples are drawn from the populations of interest. The **size** (number of units contained) of the sample is lower than that of the population.

Population and sample



Population

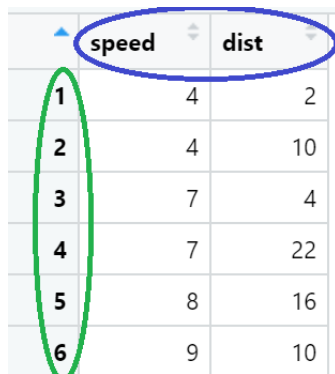


Sample

Talking data

Data, either collected for a sample or a population, are generally “structured” into two main “components”:

- ▶ The statistical **units** for which information (data) are collected;
- ▶ The **variables** recorded for the given sets of units. A variable is one of the character of interest that was recorded/collected in the data.



	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10

Types of variables

Variables are categorized by the type of values they can take:

QUANTITATIVE



DISCRETE



CONTINUOUS

QUALITATIVE



NOMINAL



ORDINAL

Quantitative variables

Generally speaking, quantitative variables are numerical variables. They can be separated in two groups:

- ▶ **Discrete** variables. These are numerical variables that can only take specific numeric values in a predefined interval. E.g. Shoe sizes.
- ▶ **Continuous** variables. These are numerical variables that can take any value in a predefined interval. E.g. Height of individuals.

Qualitative variables

Generally speaking, qualitative (or **categorical**) variables are variables that take values which can not be unequivocally quantified. They can be separated in two groups:

- ▶ **Nominal** variables. These are categorical variables whose values can not be ordered, they can only be distinguished from one another. E.g. Colours of cars.
- ▶ **Ordinal** variables. These are categorical variables whose values can be ordered. E.g. Education levels obtained.

Discrete variables

Examples

- ▶ Number of tries made in a rugby match
- ▶ Number of black socks in a drawer
- ▶ Number of votes received by a party in an election
- ▶ Number of phone calls received in a day in an office
- ▶ Number of times a sequence of coin tosses returns tail

Continuous variables

Examples

- ▶ The height of individuals
- ▶ The time it takes swimmers to swim 50 meters breaststroke
- ▶ House prices in Dublin
- ▶ The grams of fat contained in different foods
- ▶ The liters of water drunk by people in a day

Nominal variables

Examples

- ▶ M&Ms candies' colors
- ▶ People's hair colors
- ▶ People's nationalities
- ▶ Music genres
- ▶ Schools' names

Ordinal variables

Examples

- ▶ Levels of satisfaction
- ▶ Players' rankings in tennis
- ▶ Days of the week
- ▶ Months of the years

Data visualization

Exploring data with plots

- ▶ **Exploration:** there is a message in the data and the display helps us to learn what it is.
- ▶ **Communication:** we know something and the display helps us to effectively tell others.
- ▶ **Decoration:** the plot is engaging and enlivens the presentation.

Quantitative data

Old faithful geyser (Yellowstone):

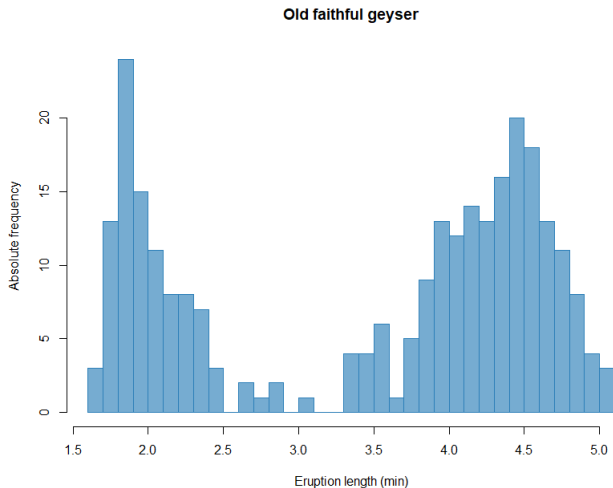
- ▶ **eruptions**: length of eruption, measured in minutes
- ▶ **waiting**: Waiting time to next eruption, measured in minutes



```
> 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
```

Histograms

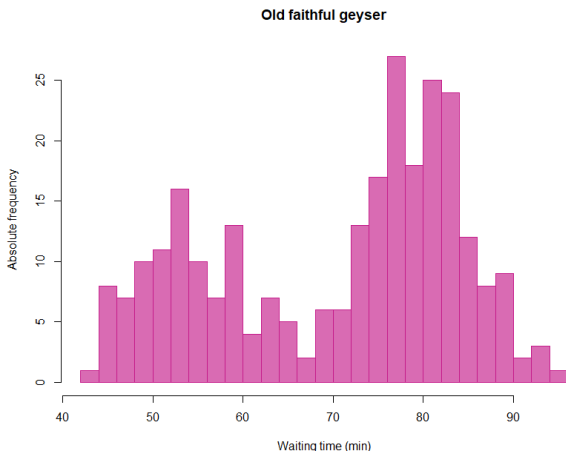
Both **eruptions** and **waiting** are quantitative, continuous, variables. We can represent them using **histograms**:



Histograms

Absolute frequencies

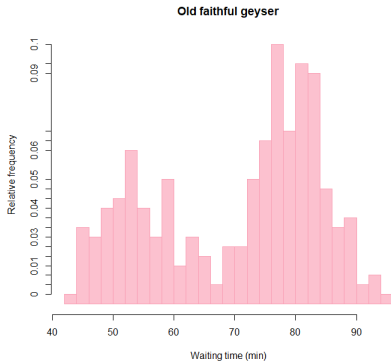
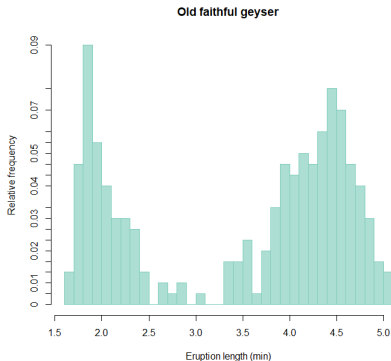
Computing the **absolute frequencies** for a variable means counting the number of times each value (observation) has occurred.



Histograms

Relative frequencies

Computing the **relative frequencies** for a variable means counting the relative number of times (relative to the total number of values observed) each value (observation) has occurred.



Histograms

Absolute frequencies - construction

1. Define B , the number of “classes” to be used to group our observations. Each class will contain an equal number of observations
2. Compute the **range** of the observations (**min** and **max** values)
3. Sub-divide the range into B classes, of equal width
4. Count how many observations fall in each of the B classes, this will be the absolute frequencies associated to each one of the classes. We can denote them with f_1, \dots, f_B

Histograms

Relative frequencies - construction

1. Define B , the number of “classes” to be used to group our observations. Each class will contain an equal number of observations
2. Compute the **range** of the observations (**min** and **max** values)
3. Sub-divide the range into B classes, of equal width
4. Count how many observations fall in each of the B classes, this will be the absolute frequencies associated to each one of the classes. We can denote them with:
 f_1, \dots, f_B
5. Compute the relative frequencies as: $\frac{f_1}{n}, \dots, \frac{f_B}{n}$, where n is the total number of observations in the data

Histograms

Construction example

1.19	4.40	14.07	16.15	7.69	2.33	9.70	0.23	7.65
------	------	-------	-------	------	------	------	------	------

Here, $n = 9$, the minimum is 0.23, and the maximum is 16.15. Let us set $B = 3$. We can compute $b = \frac{16.15 - 0.23}{B} = 5.31$, that is the width of the interval. The classes will then be defined as: $[0.23; 5.54)$, $[5.54; 10.84)$, $[10.84; 16.15]$.

Absolute frequencies:

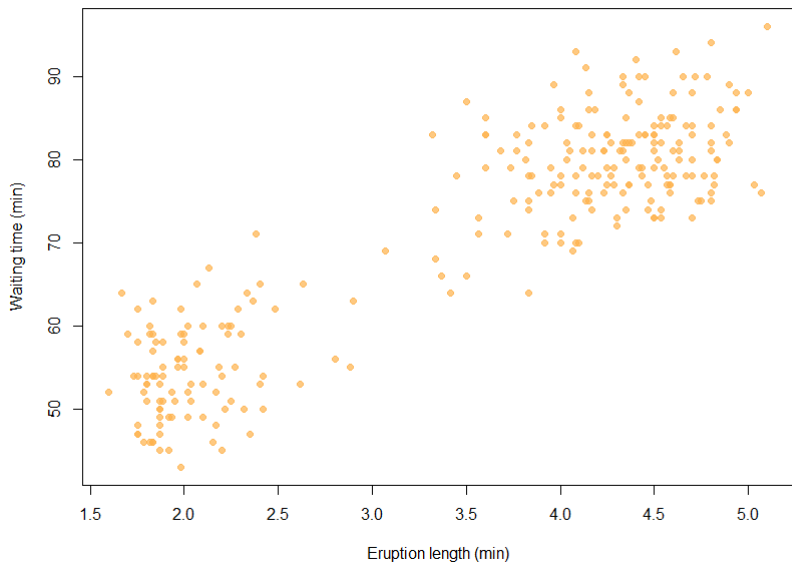
$[0.23; 5.54)$	$[5.54; 10.84)$	$[10.84; 16.15]$
4	3	2

Relative frequencies:

$[0.23; 5.54)$	$[5.54; 10.84)$	$[10.84; 16.15]$
0.44	0.33	0.22

Scatterplots

Old faithful geyser



Scatterplots

We can use **scatterplots** to represent any pair of **quantitative** variables (both discrete and continuous).

Scatterplots simply represent the values observed for one quantitative variable against those of another quantitative variable.

As histograms, they can be quite useful to help detecting patterns and characteristics of data.

Qualitative data

Effectiveness of Insect Sprays:

- ▶ **count**: count of dead insects
- ▶ **spray**: type of spray (A, B, C, D, E, F)



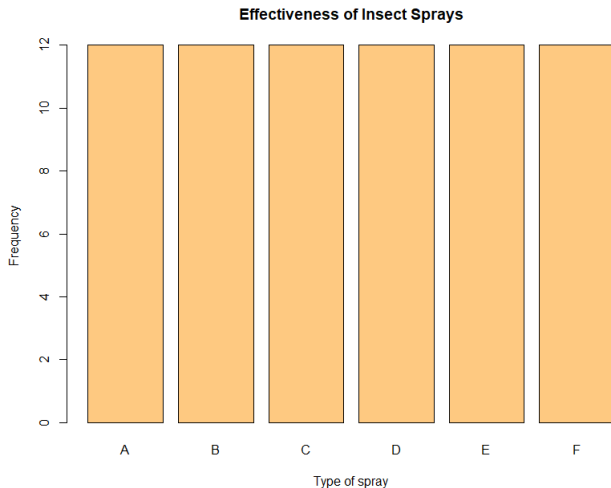
```
> InsectSprays
```

```
count spray
```

1	10	A
2	7	A
3	20	A
4	14	A
5	14	A
6	12	A
7	10	A
8	23	A
9	17	A
10	20	A
11	14	A

Bar plots

Spray is a categorical variable. We can represent it using a bar plot:



Categorical data

Frequencies and contingency tables

As in the case of continuous data, frequencies can be computed also in the presence of categorical data.

Example

Consider the two categorical variables below ($n = 9$):

var 1	A	A	B	C	C	A	B	C	A
var 2	X	X	Y	Y	X	X	Y	X	Y

The absolute frequencies of **var 1** are:

Value	A	B	C
f_j	4	2	3
\tilde{f}_j	0.444	0.222	0.333

where f_j and \tilde{f}_j are the absolute and the relative frequencies, respectively.

Categorical data

Frequencies and contingency tables

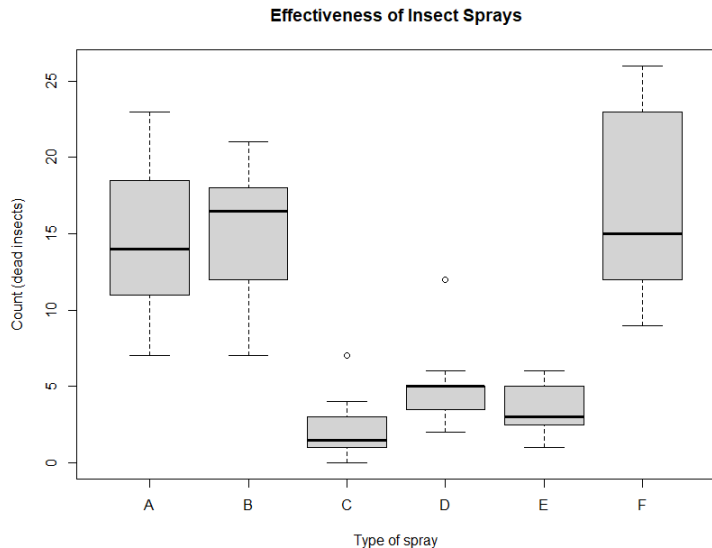
We can visualize the **joint** frequency distribution of two categorical variables using a **contingency table** (or a cross tabulation).

Example

The contingency table for **var 1** and **var 2** is given by:

Value	A	B	C	Total
X	3	0	2	5
Y	1	2	1	4
Total	4	2	3	9

Box plots



Box plots

We can use **box plots** to represent any pair of variables made up of a **quantitative** variable and a **qualitative** variable.

Box plots simply represent the values observed for the quantitative variable against those of the qualitative variable.

They can be quite useful to help detecting patterns and characteristics of data. For example, does the quantitative variable “behave” differently depending on the level of the categorical variable?

Summary statistics

A bit of notation

- ▶ Let us denote with \mathbf{X} a given dataset (matrix)
- ▶ \mathbf{X} has dimensions $n \times P$, where:
 - ▶ n is the number of observations
 - ▶ P is the number of variables
- ▶ \mathbf{X}_p denotes the p^{th} variable (vector), $p = 1, \dots, P$
- ▶ \mathbf{X}_i denotes the i^{th} observation (vector), $i = 1, \dots, n$
- ▶ x_{ip} denotes the value of the p^{th} variable for the i^{th} unit

Unit	\mathbf{X}_1	\mathbf{X}_2	...	\mathbf{X}_p	...	$\mathbf{X}_{(P-1)}$	\mathbf{X}_P
1	x_{11}	x_{12}	...	x_{1p}	...	$x_{1(P-1)}$	x_{1P}
2	x_{21}	x_{22}	...	x_{2p}	...	$x_{2(P-1)}$	x_{2P}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
(n-1)	$x_{(n-1)1}$	$x_{(n-1)2}$...	$x_{(n-1)p}$...	$x_{(n-1)(P-1)}$	$x_{(n-1)P}$
n	x_{n1}	x_{n2}	...	x_{np}	...	$x_{n(P-1)}$	x_{nP}

Summary statistics

In order to describe the values x_{ip} , $i = 1, \dots, n$, $p = 1, \dots, P$, one option is to use **summary statistics**.

In general, a summary statistic is any statistic that summarizes the information coming from a given variable, sets of variables, unit, or sets of units.

Different types of data allow for the construction and use of different types of summary statistics.

Different summary statistics may incorporate differently the information present in the data.

Means

Arithmetic mean

When in the presence of **quantitative** variables, the **arithmetic mean** of n values $\{x_1, \dots, x_n\}$ can be computed as as:

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

If the variable is discrete and its frequency distribution is known, the arithmetic mean can be equivalently computed as:

$$\bar{x}_a = \frac{1}{n} \sum_{j=1}^J x_j f_j = \sum_{j=1}^J x_j \tilde{f}_j$$

where f_j is the absolute frequency of the j^{th} value, \tilde{f}_j its relative frequency, J is the total number of different values for that variable.

Means

Weighted arithmetic mean

When in the presence of **quantitative** variables, the **weighted arithmetic mean** of n values $\{x_1, \dots, x_n\}$ can be computed as as:

$$\bar{x}_{\tilde{a}} = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}$$

where $\{w_1, \dots, w_i, \dots, w_n\}$ is a sets of observation-specific weights, with $\sum_{i=1}^n w_i = 1$.

Arithmetic means

Example

Type	Time (min)	Number of items
A	45	400
B	15	2500
C	18	1500

Let us consider the variable Time. Its arithmetic mean is 26 minutes:

$$\bar{x}_a = \frac{45 + 15 + 18}{3} = 26$$

Its weighted arithmetic mean (with respect to the numbers of items) is 18.75 minutes:

$$\bar{x}_a = \frac{45 \times 400 + 15 \times 2500 + 18 \times 1500}{400 + 2500 + 1500} = 18.75$$

Arithmetic means

Properties

- a The sum of n values is equal to its arithmetic mean value, multiplied by n :

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

- b The sum of the differences between n values and their arithmetic mean value is 0:

$$\sum_{i=1}^n (x_i - \bar{x}_a) = 0$$

- c The sum of the squared differences between n values and a constant value c is minimal when $c = \bar{x}_a$:

$$\arg \min_c \left[\sum_{i=1}^n (x_i - c)^2 \right] = \sum_{i=1}^n (x_i - \bar{x}_a)^2$$

Arithmetic means

Properties

- d If we divide n values $\{x_1, \dots, x_n\}$ into L non-overlapping groups, $L < n$, $\{n_1, \dots, n_l, \dots, n_L\}$, such that $\sum_{l=1}^L n_l = n$, and such that $\bar{x}_{a,l}$ is the arithmetic mean of group l , for $l = 1, \dots, L$, we have that:

$$\bar{x}_a = \frac{1}{n} \sum_{l=1}^L \bar{x}_{a,l} n_l$$

- e The arithmetic mean of a linear transformation of the values $\{x_1, \dots, x_n\}$ is equal to the same linear transformation applied to the arithmetic mean of the values:

$$\bar{y}_a = \frac{1}{n} \sum_{i=1}^n (\alpha + \beta x_i) = \alpha + \beta \bar{x}_a$$

Means

Geometric mean

The **geometric** mean is a type of mean that can be quite useful when the values are ratios (hence also all positive). It can be computed as:

$$\bar{x}_g = \sqrt[n]{x_1 \times \cdots \times x_n}$$

If the variable at hand is discrete, we can compute its geometric mean as:

$$\bar{x}_g = \sqrt[n]{x_1^{f_1} \times \cdots \times x_j^{f_j} \times \cdots \times x_J^{f_J}}$$

where J is the total number of different values and f_j the absolute frequencies, $j = 1, \dots, J$.

Geometric mean

Properties

- a The product of n values is equal to the n^{th} power of their geometric mean:

$$x_1 \times \cdots \times x_n = [\bar{x}_g]^n$$

- b The logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of the values:

$$\log(\bar{x}_g) = \frac{1}{n} \sum_{i=1}^n \log(x_i)$$

Geometric mean

Example

A person has invested 10000 euro in a stock with variable interest rate. Below are the data regarding the interest rate values per year and the capital values at the end of each year:

Year	Interest rate	Capital (at end of the year)
1	0.015	10150
2	0.02	10353
3	0.072	11098.4

The capital at the end of year 3 is computed as:

$$10000(1 + 0.015)(1 + 0.02)(1 + 0.072) = 11098.4$$

Geometric mean

Example

We can compute the average coefficient for the increment in interest rate over the 3 years using the geometric average as:

$$\bar{x}_g = \sqrt[3]{1.015 \times 1.02 \times 1.072} = 1.03535$$

This correspond to an average yearly interest rate of 0.03535. We can use this average interest rate to alternatively compute the capital at the end of year 3:

$$10000(1.03535)^3 = 11098.43$$

Generally, it is useful to use the geometric mean when the values are not independent or if they present large fluctuations.

Median

Both the arithmetic and the geometric mean can be quite sensitive to extreme values and “outliers”.

An alternative centrality index, which is more “robust”, is the **median**.

The median can be computed for both quantitative and qualitative ordinal variables.

Definition: Given a set of *ordered* values $\{x_1, \dots, x_n\}$, the median is defined as the value of the central value. That is, the median correspond to the value of the units that splits into two equal parts $\{x_1, \dots, x_n\}$.

Median

There can be two different situations:

- ▶ n is an **odd** number. In this case, the median is the $\frac{n+1}{2}$ value in the ordered sequence $\{x_1, \dots, x_n\}$. That is:

$$M_e = x_{\frac{n+1}{2}}$$

- ▶ n is an **even** number. In this case, one must find the $\frac{n}{2}$ and $\frac{n}{2} + 1$ values in the ordered sequence $\{x_1, \dots, x_n\}$. The median is then computed as:

$$M_e = \frac{1}{2} (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$$

Median

Properties

- a Given quantitative values $\{x_1, \dots, x_n\}$, the sum of their absolute differences from a constant c is minimal when c is the median:

$$\arg \min_c \left[\sum_{i=1}^n |x_i - c| \right] = \sum_{i=1}^n |x_i - M_e|$$

Median

Classes and frequencies

If we are given the frequency distribution of a variable whose observed values have been grouped in classes, we can compute the median using frequencies.

Example

class	0 – 3	3 – 6	6 – 9	9 – 12	12 – 15
f_j	10	5	30	21	2
f_{Cj}	$\frac{10}{68}$	$\frac{10+5}{68}$	$\frac{15+30}{68}$	$\frac{45+21}{68}$	$\frac{66+2}{68}$
fc_j	0.147	0.221	0.662	0.971	1

where $n = 68$ is the total number of observations, and fc_j is the **cumulative relative frequency** for a class. We can compute it as:

$$fc_j = \frac{1}{n} \sum_{k=1}^j f_k; \quad j = 1, \dots, J$$

Median

Classes and frequencies

We can use a formula based on cumulative relative frequencies to compute the median:

$$M_e = m_{M_e} + \left[\frac{0.5 - fc_{(M_e-1)}}{fc_{M_e} - fc_{(M_e-1)}} \right] \Delta_{M_e}$$

where:

- ▶ m_{M_e} is the lower bound of the class containing the median value
- ▶ fc_{M_e} and $fc_{(M_e-1)}$ are, respectively, the cumulative relative frequencies of the class containing the median and the class before this one
- ▶ Δ_{M_e} is the width of the class containing the median

Median

Classes and frequencies

Example

class	0 – 3	3 – 6	6 – 9	9 – 12	12 – 15
f_j	10	5	30	21	2
f_{Cj}	$\frac{10}{68}$	$\frac{10+5}{68}$	$\frac{15+30}{68}$	$\frac{45+21}{68}$	$\frac{66+2}{68}$
f_{Cj}	0.147	0.221	0.662	0.971	1

$$M_e = 6 + \left[\frac{0.5 - 0.221}{0.662 - 0.221} \right] 3 = 7.898$$

NOTE: we are making the assumption that values are equally distributed within the classes (and especially in the class containing the median!)

Mode

The **Mode** is the distinct value that is associated to the largest frequency (absolute or relative).

We can compute the mode for all different types of variables.

Example 1

x_j	A	B	C	D	E
f_j	4	15	3	18	4

Here the mode is D ($f_D = 18$).

Example 2

class	0 – 3	3 – 6	6 – 9	9 – 12	12 – 15
f_j	10	5	30	21	2

Here the mode is the class 6 – 9.

Mode

Example 3

class	0 – 3	3 – 10	10 – 20	20 – 22	22 – 25
f_j	10	5	30	21	2

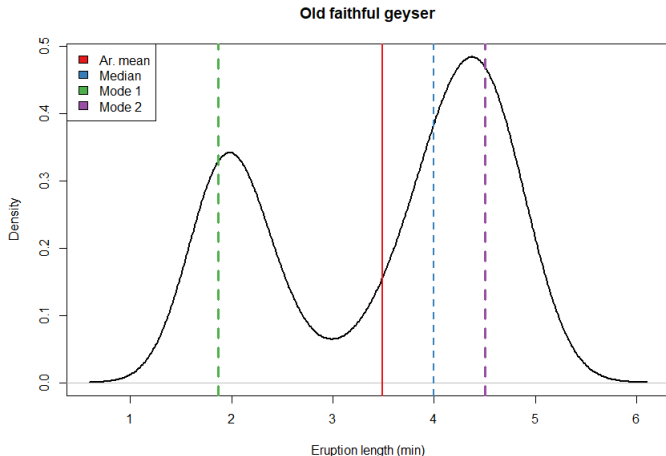
Here the classes have different widths. In order to compute the mode, we need to adjust for this. We can do it by computing absolute frequencies adjusted for class-width:

class	0 – 3	3 – 10	10 – 20	20 – 22	22 – 25
adj. freq	$\frac{10}{3}$	$\frac{5}{7}$	$\frac{30}{10}$	$\frac{21}{2}$	$\frac{2}{3}$
adj. freq	3.333	0.714	3	10.5	0.667

The modal class is 20 – 22.

Centrality measures

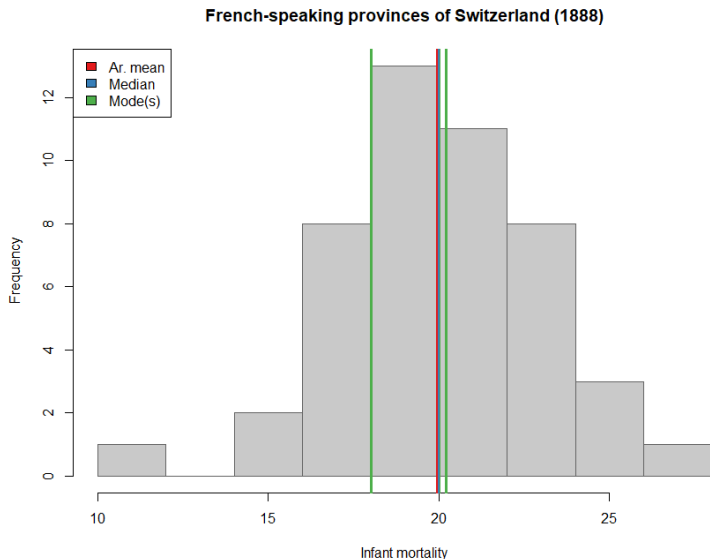
A comparison



NOTE: One mode \rightarrow *unimodal*; Two modes \rightarrow *bimodal*; Two or more modes \rightarrow *multimodal*

Centrality measures

A comparison



Quantiles

The median is the value that splits in half the ordered distribution of observations. This means that 50% of the observations will fall before the median value, while the other 50% after.

In general, we can split in “different chunks” the ordered distribution of observations, such that each “chunks” contains the **same number of observations**. We can do so using **quantiles**.

Quantiles are values that split the distribution of the observations in K parts, each one containing the same number of observations.

Special case: Percentiles are values that split the distribution of the observations in 100 parts, each one containing the same number of observations.

Percentiles

We can compute the k^{th} percentile using cumulative relative frequencies and the formula below:

$$q_k = m_{q_k} + \left[\frac{\frac{k}{100} - fc_{(q_k-1)}}{fc_{q_k} - fc_{(q_k-1)}} \right] \Delta_{q_k}$$

where:

- ▶ m_{q_k} is the lower bound of the class containing the k^{th} percentile, q_k
- ▶ fc_{q_k} and $fc_{(q_k-1)}$ are, respectively, the cumulative relative frequencies of the class containing q_k and the class before this one
- ▶ Δ_{q_k} is the width of the class containing q_k

Percentiles

Example

Consider the cumulative relative frequencies in the table below:

class	0 – 3	3 – 6	6 – 9	9 – 12	12 – 15
f_j	10	5	30	21	2
f_{Cj}	$\frac{10}{68}$	$\frac{10+5}{68}$	$\frac{15+30}{68}$	$\frac{45+21}{68}$	$\frac{66+2}{68}$
fc_j	0.147	0.221	0.662	0.971	1

Let us compute the 10% percentile:

$$q_{10} = 0 + \left[\frac{\frac{10}{100} - 0}{0.147 - 0} \right] 3 = 2.041$$

Let us compute the 88% percentile:

$$q_{88} = 9 + \left[\frac{\frac{88}{100} - 0.662}{0.971 - 0.662} \right] 3 = 11.116$$

Box plots

A **Boxplot** is a visualization method, which allows to summarize some of the main characteristics of the **empirical** distribution of a variable (that is, of the observations).

Given the observations $\{x_1, \dots, x_n\}$. it is constructed using:

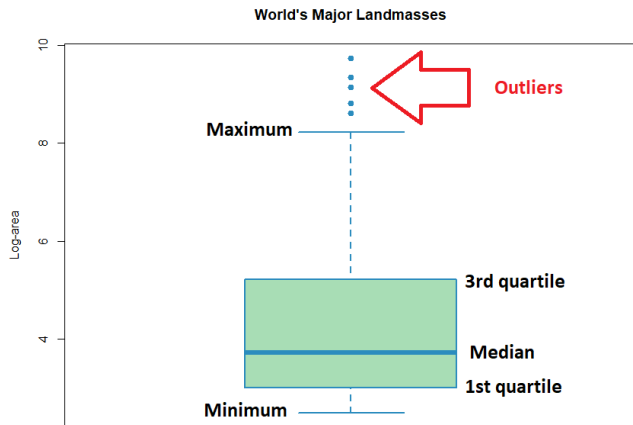
- ▶ q_0 : the minimum value of $\{x_1, \dots, x_n\}$
- ▶ q_{25} : the 25th percentile of $\{x_1, \dots, x_n\}$ (1st quartile)
- ▶ q_{50} : the 50th percentile of $\{x_1, \dots, x_n\}$ (M_e)
- ▶ q_{75} : the 75th percentile of $\{x_1, \dots, x_n\}$ (3rd quartile)
- ▶ q_{100} : the maximum value of $\{x_1, \dots, x_n\}$

NOTE: many software will compute the above quantities after having “removed” **outliers**, which will then be represented separately in the box plot.

Box plots

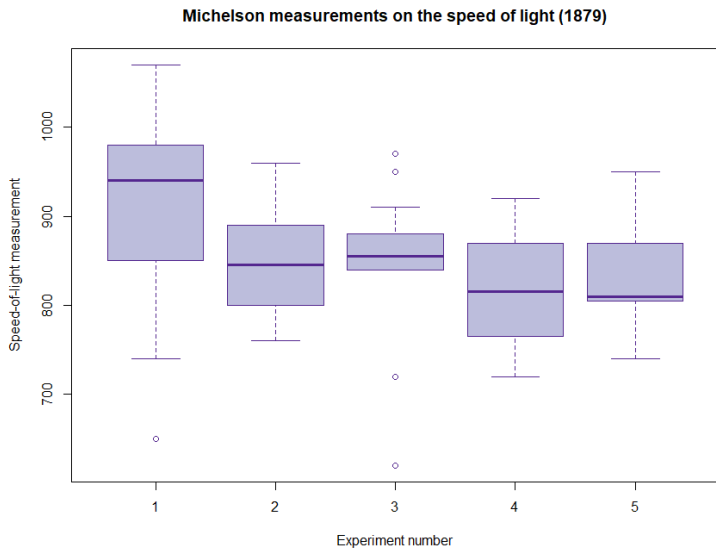
Example

NOTE: The difference between the 3rd and the 1st quartiles is called **interquartile range (IQR)**



Box plots

Example



Extreme values

Sometimes, some values may be quite different from the majority of others. When this difference is substantial, such values are referred to as **extreme values** or **outliers**.

The definition of outlier can not be objective, it is **context-driven**, and should be done with extreme care!

If an observation (or some of them) presents an extreme value, it does not mean that we need to remove it (or them) from the data. However, we can always investigate what would happen if we were to remove it (or them).

Extreme values

IQR criterion

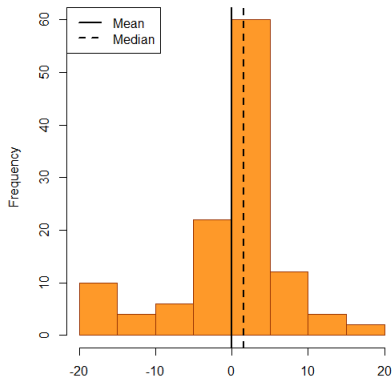
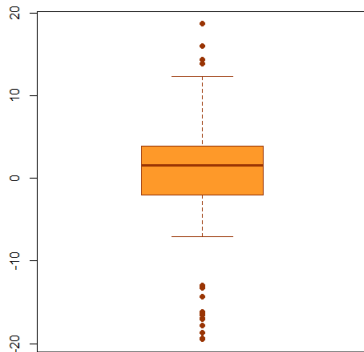
One criterion that may be used to find extreme values is one based on the IQR. This criterion postulates that the observations falling outside the interval

$$[q_{25} - 1.5IQR; q_{75} + 1.5IQR]$$

can be regarded as potential outliers. This is the criterion used in R when drawing box plots.

Extreme values

Example - IQR criterion

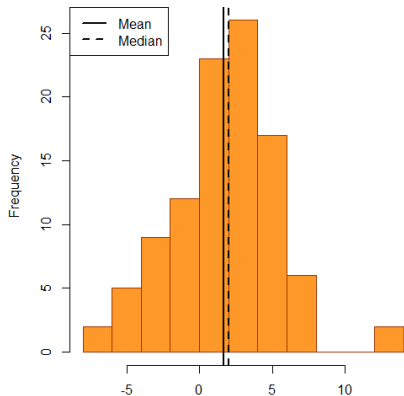
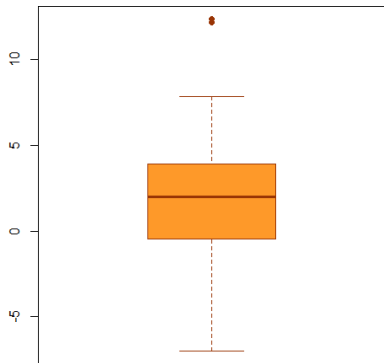


Here, the mean value is 0.003, while the median value is 1.590.

What if we removed the outliers?

Extreme values

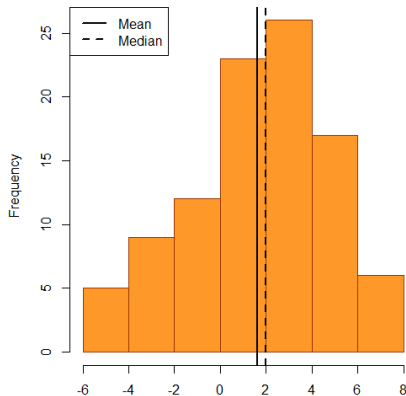
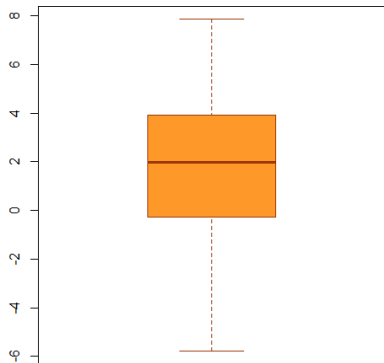
Example - IQR criterion



Here, the mean value is 1.648, while the median value is 1.998.

Extreme values

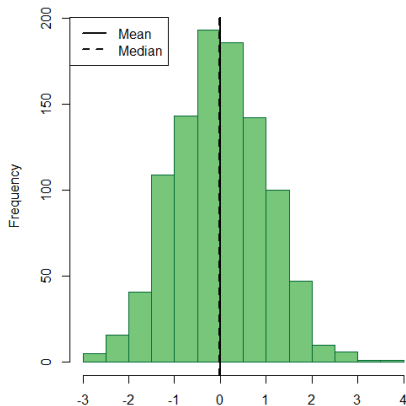
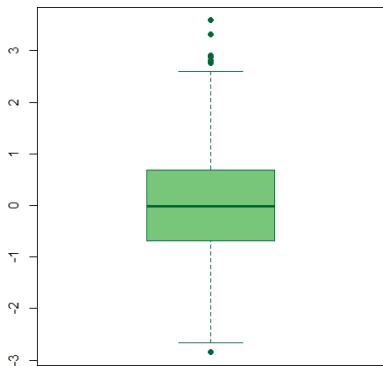
Example - IQR criterion



Here, the mean value is 1.608, while the median value is 1.998.

Shapes

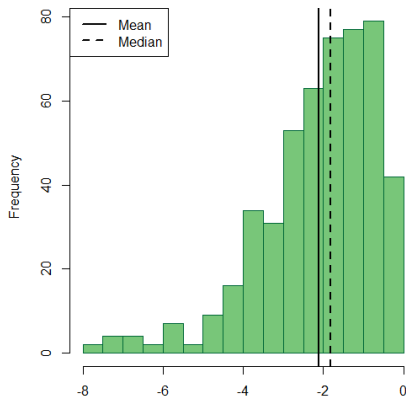
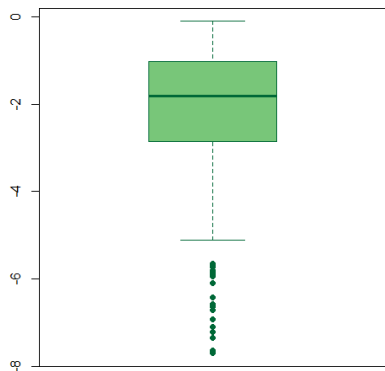
Symmetric - bell shaped



The median and the mean are quite close. The **tails** behave similarly.

Shapes

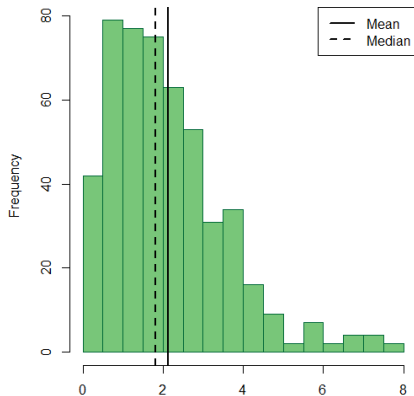
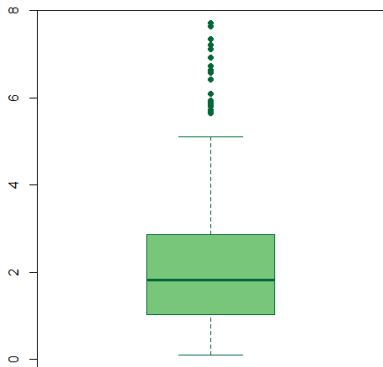
Left skewed



The mean has a lower value than the median. The **left tails** is “heavier” than the right one.

Shapes

Right skewed



The mean has a larger value than the median. The **right tails** is “heavier” than the left one.

Exercises

Exercises

- ▶ Consider the following set of observed values: $X = \{7.02, 1.57, 3.40, 2.13, -0.12, 4.99, 1.96, 5.41, 3.66, 1.04, 2.21, 1.55, 8.51, 2.15, -3.32\}$. Construct the histogram for X , setting $B = 5$. You should compute both the absolute and relative frequencies, and choose which one of the two to use when drawing your histogram.
- ▶ Consider the following set of observed values: $X = \{A, B, B, B, C, C, D, E, F, G, C, D, D, E, A, B\}$, and $Y = \{a, a, a, e, e, c, b, b, c, e, a, a, b, e, c, a\}$. Construct the contingency table for X and Y . Further, draw the barplot for X , using relative frequencies.

Exercises

- ▶ Consider the following set of observed values: $X = \{7.02, 1.57, 3.40, 2.13, -0.12, 4.99, 1.96, 5.41, 3.66, 1.04, 2.21, 1.55, 8.51, 2.15, -3.32\}$.
 - ▶ Compute the arithmetic mean of X .
 - ▶ Consider the following weights: $w = \{0.09, 0.13, 0.07, 0.07, 0.06, 0.02, 0.08, 0.09, 0.01, 0.08, 0.11, 0.00, 0.04, 0.05, 0.10\}$. Compute the weighted arithmetic mean of X , using the weights w .
 - ▶ Compute the geometric mean of X .

Exercises

- ▶ Consider the data in the table below:

class	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25
f_j	7	15	10	2	22
fc_j	-	-	-	-	-

- ▶ Compute the fc_j values (the cumulative relative frequencies).
- ▶ Compute the median.
- ▶ Which one is the modal class?
- ▶ Compute the 75% and the 35% percentile.

Exercises

- ▶ Show that: $\sum_{i=1}^n x_i = n\bar{x}_a$
- ▶ Show that: $\sum_{i=1}^n (x_i - \bar{x}_a) = 0$
- ▶ Show that $\log(\bar{x}_g) = \frac{1}{n} \sum_{i=1}^n \log(x_i)$