Statistical Computing
Week 5: 1. November 2019
Descriptive Statistics: Summarising two variables, correlation

Contents

So far we have only considered summarising a variable on it's own.
Today we consider describing the joint behaviour of two variables.

How you do this depends on what type of variables we have.

Two qualitative variables

For nominal or ordinal variables.

The standard method to describe nominal or ordinal variables individually is with a frequency table.

Example:

Blood group	0	A	B	AB	Total
Frequency	31	32	9	4	76

In R: `table(bloodgroup$ABO)`

Contingency table

The standard method to describe two nominal or ordinal variables at once is with a **contingency table** (aka cross-tabulation)

Blood group	0	A	B	AB	Total
Rhesus Factor					
Positive	26	27	7	4	64
Negative	5	5	2	0	12
Total	31	32	9	4	76

The totals are called the marginal frequencies.

in R: `table(bloodgroup$Rhesus,bloodgroup$ABO)`

Relative frequencies

We can now specify either the relative frequencies, the row frequencies or the column frequencies.

Relative frequencies; all entries add up to one

Blood group	0	A	B	AB
Positive	0.342	0.355	0.092	0.053
Negative	0.066	0.066	0.026	0.000

In R: `prop.table(table(bloodgroup$Rhesus,bloodgroup$ABO))`

Row frequencies; all row entries add up to one

Blood group	0	A	B	AB
Positive	0.406	0.422	0.109	0.062
Negative	0.417	0.417	0.167	0.000

Used to compare the column frequencies in each row group. Here ABO frequencies are similar for rhesus positive and negative.
In R: `prop.table(table(bloodgroup$Rhesus,bloodgroup$ABO),1)`

Column frequencies; all column entries add up to one

Blood group	0	A	B	AB
Positive	0.839	0.844	0.778	1.000
Negative	0.161	0.156	0.222	0.000

Used to compare the row frequencies in each column group.
In R: `prop.table(table(bloodgroup$Rhesus,bloodgroup$ABO),2)`

Graphics

A bar chart is the best way to display qualitative variables. For two variables together use a stacked or grouped barchart.

Examples in the workshop

One continuous and one qualitative variable

The usual approach is to obtain the summary statistics of the continuous variable, eg. mean, s.d., for each value of the qualitative variable

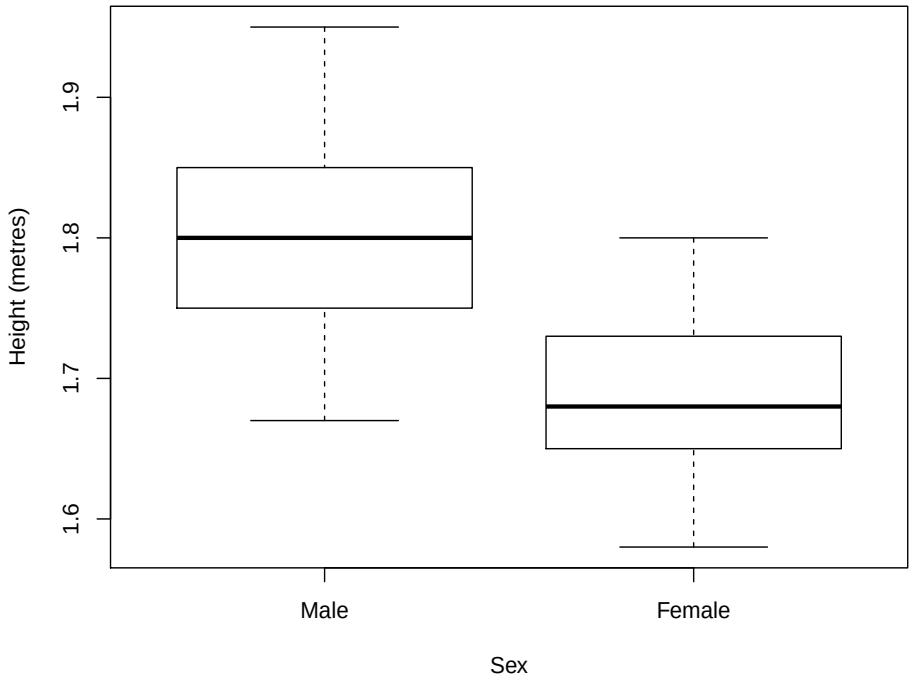
Height (metres)	Males	Females
Mean	1.802	1.689
Standard deviation	0.074	0.063

The command to do this in R is `tapply()`.

```
> tapply(height,sex,mean)
```

This command needs to be looked at in more detail: in workshop.

To display one continuous and one qualitative variable graphically, use a box plot or strip chart (covered in Week 3)



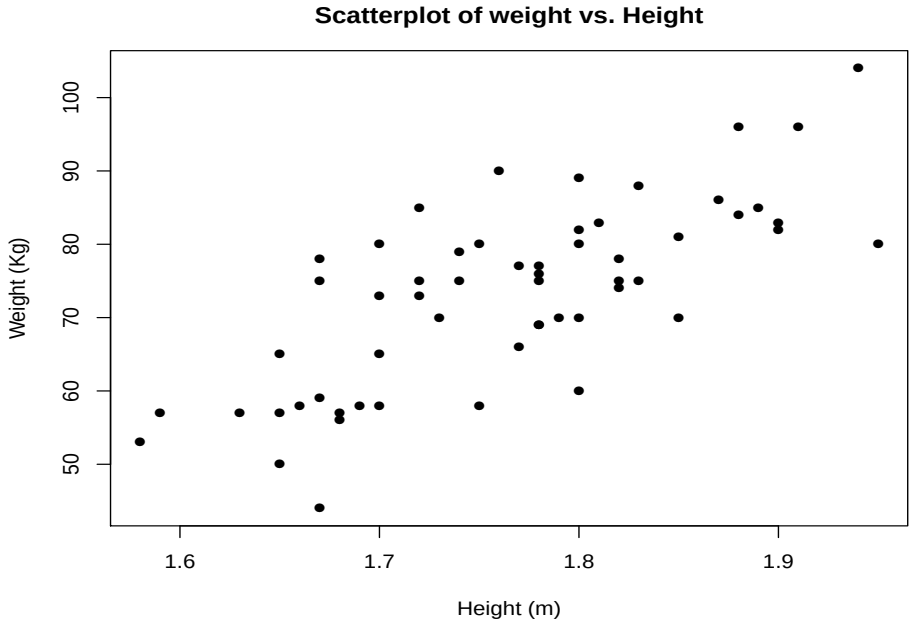
Two continuous variables

For two continuous variables, we compare the codependency between the two variables.

If two variables have a **linear codependency** then they have large **covariance** and **correlation**.

Scatterplot

It is usually a good idea to plot the data in a scatter plot. This usually gives us an idea of what type and how strong the codependency is.
Here we see a fairly strong linear and positive codependency between height and weight.



Covariance and Correlation

Variance for variable X is:

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

The **covariance** is a similar measure using two variables X and Y and is defined as:

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

Note: if you replace y with x you obtain the same formula as the variance of X.

The covariance is mathematically useful, but is difficult to interpret.

The **correlation coefficient** is a standardised version of the covariance. The correlation coefficient of X and Y is the covariance divided by the standard deviation of X and Y.

$$r_{x,y} = \frac{s_{xy}}{s_x . s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

The result is a number between -1 and 1.

Properties of correlation:

- $-1 \leqslant r \leqslant 1$
- $r_{x,y}$ measures the linear relationship between X and Y
- $r_{x,y} = r_{y,x}$
- If $r_{x,y} = 1$ or $r_{x,y} = -1$ then all the points lie on a straight line.
- $r_{x,y}$ is called Pearson's* correlation coefficient.

*Karl Pearson was influential in developing statistics as a scientific subject WikiLink.

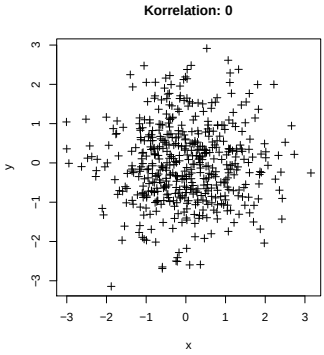
+ large X implies large Y and vice versa then the *correlation coefficient* of X and Y is near to 1 (roughly greater than 0.7) .

- large X implies small Y and vice versa then the *correlation coefficient* of X and Y is near to -1 (roughly less than -0.7).

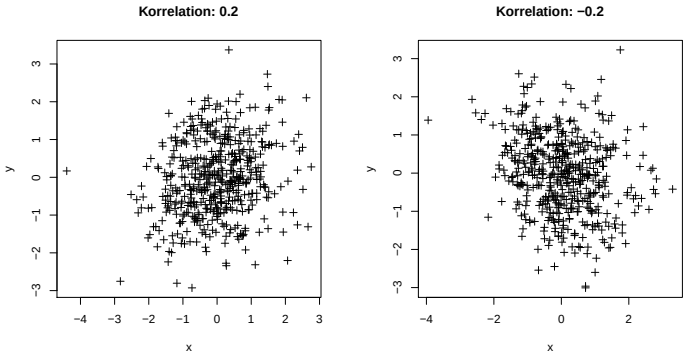
0 If there is no clear association then the covariance of X and Y will be small (roughly between -0.2 and +0.2).

Graphical examples

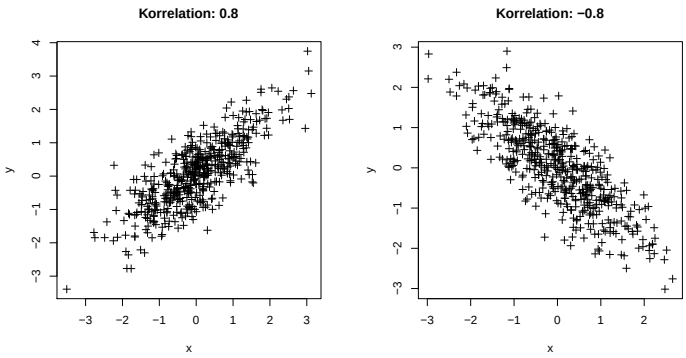
No correlation (maybe no co-dependency)



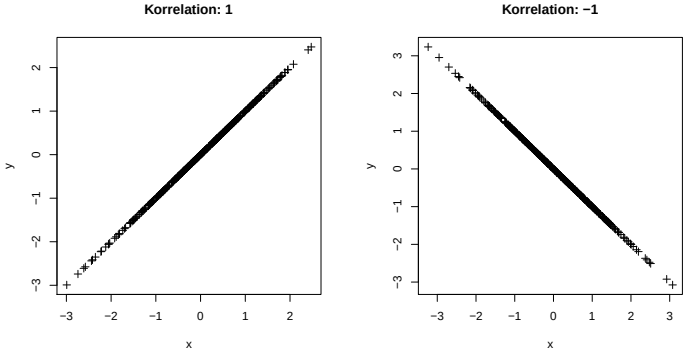
Weak correlation



Strong correlation



Perfect correlation



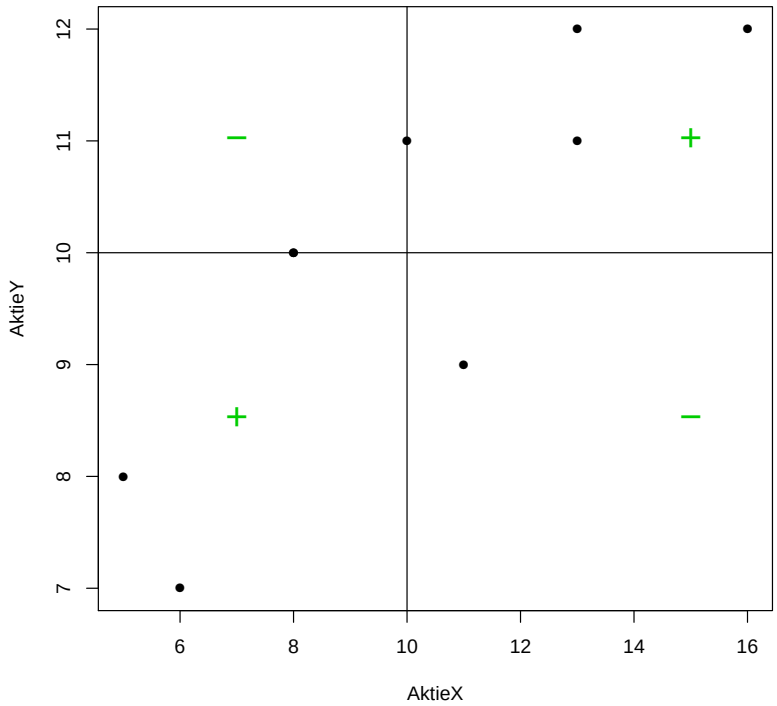
Small worked example

Two share prices X and Y are recorded on nine successive days.

Day i	Share X x_i	Share Y y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	5	8	-5	-2	10
2	6	7	-4	-3	12
3	11	9	1	-1	-1
4	8	10	-2	0	0
5	13	11	3	1	3
6	8	10	-2	0	0
7	10	11	0	1	0
8	16	12	6	2	12
9	13	12	3	2	6
Sum	90	90	0	0	42

$S_x = 3.6056$ $S_y = 1.7320$ $\sum_{i=1}^{9}(x_i - \bar{x})(y_i - \bar{y}) = 42$

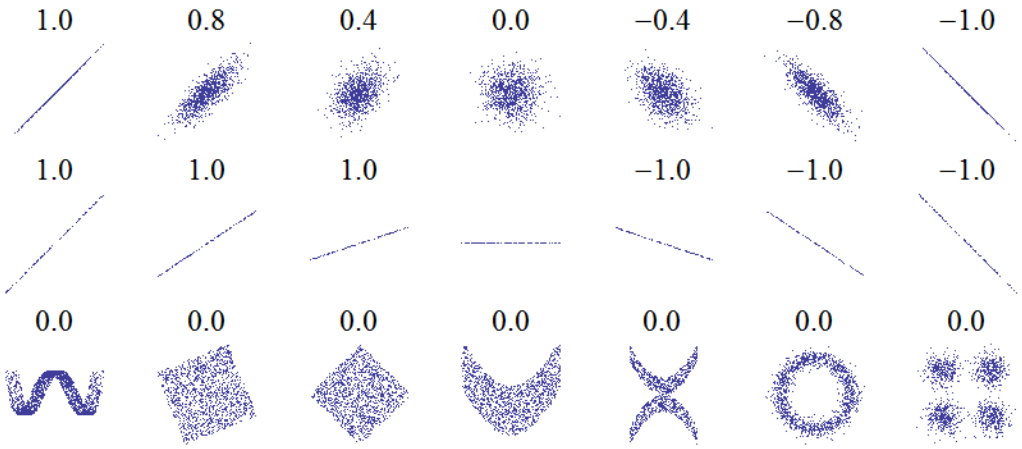
$S_{xy} = 42/8 = 5.25$ $r_{xy} = 5.25/(3.6056 \cdot 1.7320) = 0.841$



There are lots of points in the + Quadrants and only one point in the - Quadrants, leading to a large positive correlation.

When X and Y are statistically independent, then $r_{xy} = 0$.

Important The converse is not true! If $r_{xy} \approx 0$, X and Y can have a codependency ie. not statistically independent, because r_{xy} measures linear dependency.



Source:
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Warning!

Just because a correlationhas been found, it does not mean that one variable causes an effect on the other.

Correlation is not causation

Example: the number of (human) births per year in villages in rural Poland between 1918 and 1939 is highly correlated with the number of stork nests in that village.

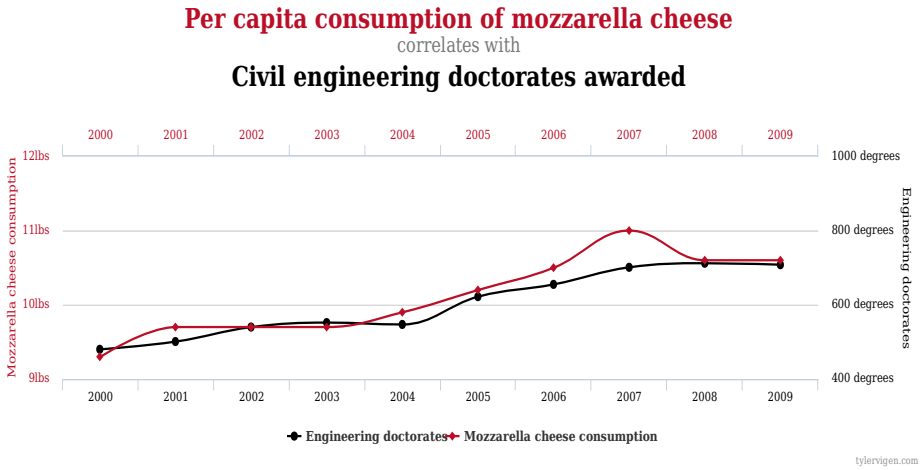
Storks are not causing the births, but larger villages have more stork nests and more human births![1]

[1] A website on the “stork problem” in German: http://www.storchproblem.de/

Another example: Consumption of ice cream is correlated with the incidence of sun burn. Sun burn neither causes nor is caused by eating ice cream!

Tyler Vigen has collected many absurd examples of spurious correlations in his website while he was a PhD Student at Harvard University: Web site http://tylervigen.com/spurious-correlations

My favourite example



Rank correlation

There is another measure of correlation: Spearman's rank correlation.

The same formula for the correlation is used as for Pearson's correlation, but instead of using the observations $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ the ranks of X and Y are used.

Rank: Definition The smallest value of X has rank 1
the second smallest value of X has rank 2
... the k-th smallest value of X has rank k
... The largest value of X has rank n.

In an athletics race, the second placed runner has rank 2 (second place) no matter if he was 10.1 second or 10 minutes behind the winner.

Dead Heats (ties):
If two or more values in a variable are equal then they are called a *tie*. We take the mean of those tied positions.

Eg. In a pub quiz there are 10 teams. After 3 rounds the teams have the following points:

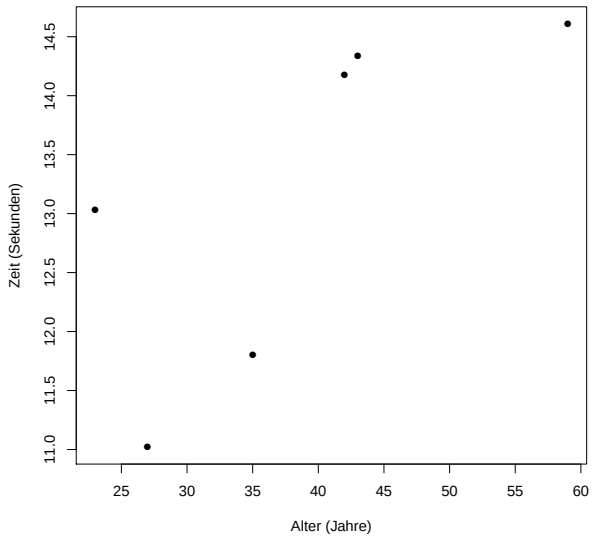
Team	A	B	C	D	E	F	G	H	I	J
Points	11	20	19	23	20	22	17	20	22	21
Rank	1	**5**	3	10	**5**	**8.5**	2	**5**	**8.5**	7

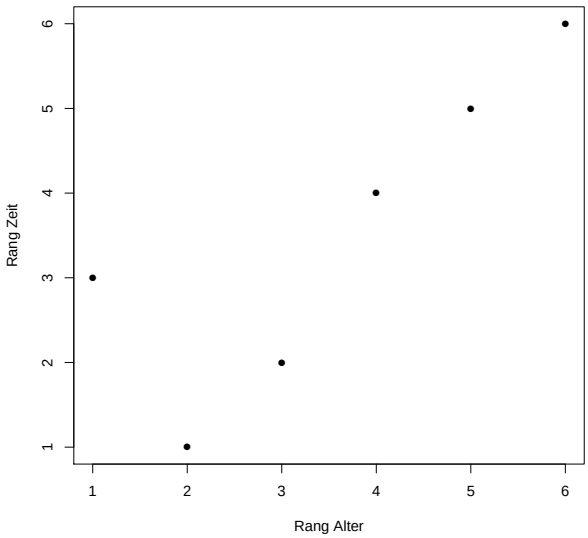
Note: Ranks always start with 1 as the smallest value.

The rank correlation is more robust against outliers. It can also be used when one or both of the variables are ordinal, because the ranks can be assigned to ordinal variables.

Simple worked example Age against vs. race times in 100m-sprint.

Person i	Age x_i	Rank of Age Rx_i	Time in Seconds y_i	Position Ry_i
A	59	6	14.61	6
B	35	3	11.80	2
C	43	5	14.34	5
D	23	1	13.03	3
E	42	4	14.18	4
F	27	2	11.02	1





$\overline{Rx} = 3.5 \qquad \overline{Ry} = 3.5$

Person	$Rx_i - \overline{Rx}$	$(Rx_i - \overline{Rx})^2$	$Ry_i - \overline{Ry}$	$(Ry_i - \overline{Ry})^2$	$(Rx_i - \overline{Rx})(Ry_i - \overline{Ry})$
A	2.5	6.25	2.5	6.25	6.25
B	-0.5	0.25	-1.5	2.25	0.75
C	1.5	2.25	1.5	2.25	2.25
D	-2.5	6.25	-0.5	0.25	1.25
E	0.5	0.25	0.5	0.25	0.25
F	-1.5	2.25	-2.5	6.25	3.75
Summe	-	17.5	-	17.5	14.5

$$r_S = \frac{\sum(Rx_i - \overline{Rx})(Ry_i - \overline{Ry})}{\sqrt{\sum(Rx_i - \overline{Rx})^2 \sum(Ry_i - \overline{Ry})^2}} = \frac{14.5}{\sqrt{17.5 \cdot 17.5}} = 0.828$$

Interpretation

For the 100m sprint example:

Pearson's correlation coefficient is	$r = 0.730$
Spearman's Rank correlation coefficient is	$r_S = 0.828$

The two coefficients are similar, we can conlude that there is a sron correlation between these ages and 100m sprint times.

What do we do if we get noticeably different results?

Look at the scatter plot!

This will usually tell us if the difference is due to:

- Outliers $\Rightarrow$ Rank correlation is better.
- The codependancy is non-linear $\Rightarrow$ Pearson's correlation is better.

Empirical cumulative distribution function (ECDF)

Two weeks ago you learnt about box plots as a means of plotting a continuous variable.

Another type of diagram which shows all values in that variable and is not very sensitive to the sample size is called the empirical cumulative distribution function

Definition Die ECDF $F_n(x)$ is the cumulative relative frequency, considered as a function of x.

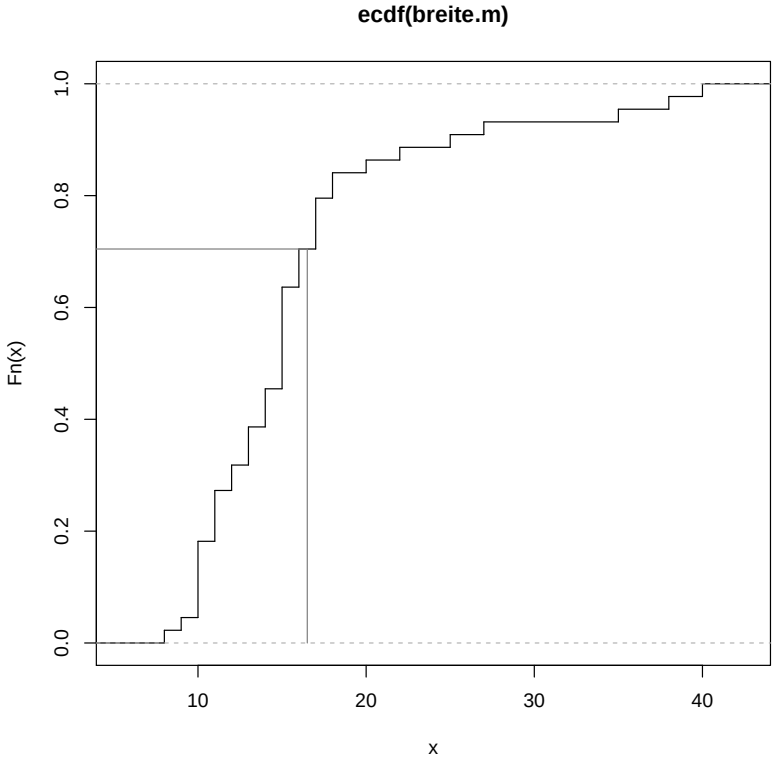
For a given x, $F_n(x)$ is the proportion of data values x_i, which are less than or equal to x:

$$F_n(x) := \frac{\#\left\{x_{(i)} | x_{(i)} \leqslant x\right\}}{n}$$

(# means *number of*).

The ECDF is a step function.

Example In Worksheet 3 Exercise 2 you plotted a box plot for the estimates of the size of the lecture theatre. The ECDF of these data is:



The x-Axis is is the width of the lecture theatre in metres. The y-Axis goes from 0 to 1. The ECDF function $F_{44}(x)$ is the proportion of x_i values less than or equal to x:

The example in the diagram is $x = 16.5$. There are 31 values $\leqslant$16.5:

$$F_{44}(16.5) = \frac{31}{44} = 0.704$$

Relation between quantiles and ECDF

ECDF:

- x is a value on the same scale as the variable
- $F_n(x)$ is a value between 0 und 1: $F_n(x) \in [0, 1]$.

p-quantile:

- p is a value between 0 and 1 $p \in [0, 1]$
- the p-quantile x_p is a value on the same scale as the variable,

Quantiles are the inverse of the ECDF.

Example

$p = 0.4$-quantile, $n = 12$, $np = 4.8$, $x_{0.4} = x_{(\lceil 4.8 \rceil)} = x_{(5)} = 178$ (green)

$p = 0.75$-quantile, $n = 12$, $np = 9$, $x_{0.75} = (x_{(9)} + x_{(10)})/2 = 184$ (blue)