Basic Statistics and Probability

Chapter 2:
Methods for Describing Sets of Data

- Qualitative data
- Quantitative data. Graphical description
- Quantitative data. Numerical description

Qualitative data

Qualitative data . . .

. . . are non-numerical (although may be numerically codified)
. . . are classified into categories (called classes)

Example 2.1: Aphasia. Source: McClave and Sincich (2017)

Aphasia is the "impairment or loss of the faculty of using or understanding spoken or written language." There are three types of aphasia: Broca's, conduction, and anomic. To determine if one type of aphasia occurs more often than any other, some researchers measured the type of aphasia for a sample of 22 adult aphasics.

Example 2.1: Aphasia

Subject	Type of Aphasia	Subject	Type of Aphasia
1	Brocas	12	Brocas
2	Anomic	13	Anomic
3	Anomic	14	Brocas
4	Conduction	15	Anomic
5	Brocas	16	Anomic
6	Conduction	17	Anomic
7	Conduction	18	Conduction
8	Anomic	19	Brocas
9	Conduction	20	Anomic
10	Anomic	21	Conduction
11	Conduction	22	Anomic

Numerical summaries of qualitative data:

Frequency of a class Number of cases or observations falling into that class.

Relative frequency of a class Frequency of that class divided by the sample size (the total number of observations in the sample).

Class percentage of a class Relative frequency of that class multiplied by 100.

Example 2.1: Aphasia

```
DataSet = read.table("APHASIA.txt",header=TRUE)
summary(DataSet)
          TYPE
 Anomic    :10
 Brocas    : 5
 Conduction: 7
table(DataSet)
DataSet
    Anomic     Brocas Conduction
        10          5          7
n = nrow(DataSet)
n
[1] 22
table(DataSet)/n
DataSet
    Anomic     Brocas Conduction
 0.4545455  0.2272727  0.3181818
```

Example 2.1: Aphasia

Number of cases: 22

Classes: Anomic, Broca's, Conduction

Frequencies: Anomic: 10; Broca's: 5; Conduction: 7

Relative frequencies: Anomic: 0.45; Broca's: 0.23 ; Conduction: 0.32

Class percentage: Anomic: 45%; Broca's: 23% ; Conduction: 32%

In table form:

Classes	*Anomic*	*Broca's*	*Conduction*	Total
Frequencies	10	5	7	22
Relative frequencies	0.45	0.23	0.32	1.00
Class percentage	45%	23%	32%	100%

Graphical summaries of qualitative data:

Bar graphs The height of the bar may represent

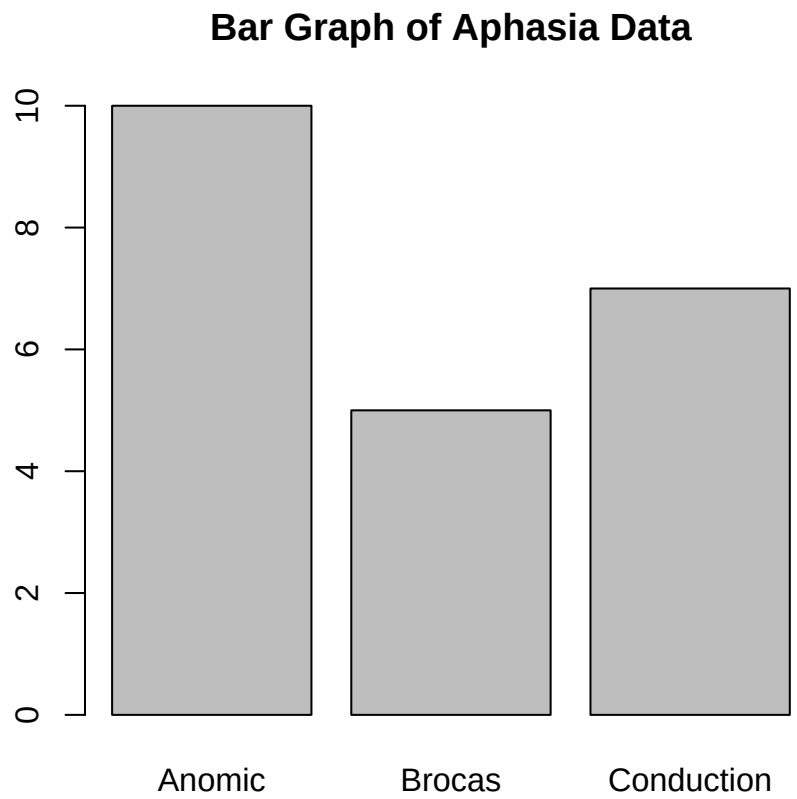
- the frequency
- the relative frequency
- the percentage

Pie charts Relative frequencies are represented by fraction of total area of a pie.

Pareto Diagrams Bar graphs with classes arranged by height in descending order from left to right.

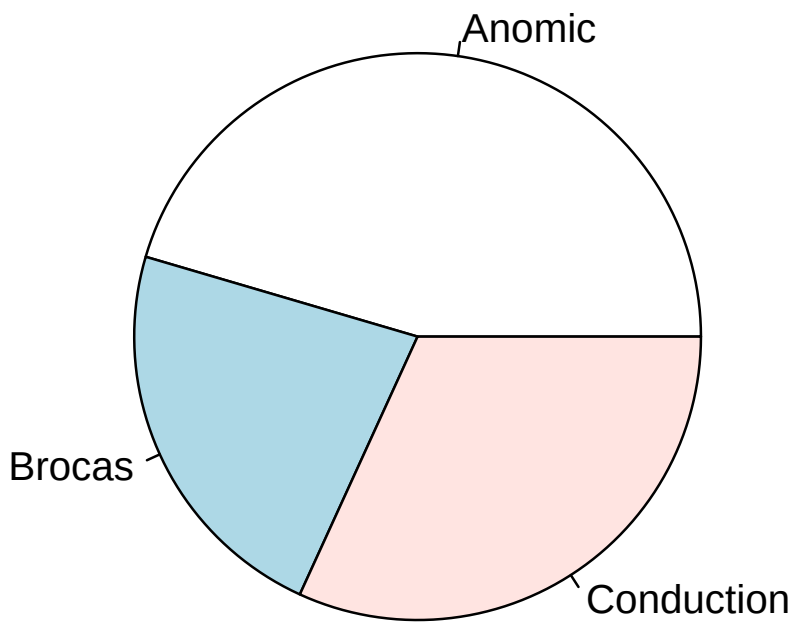
Example 2.1: Aphasia

```
barplot(table(DataSet),main="Bar Graph of Aphasia Data")
```



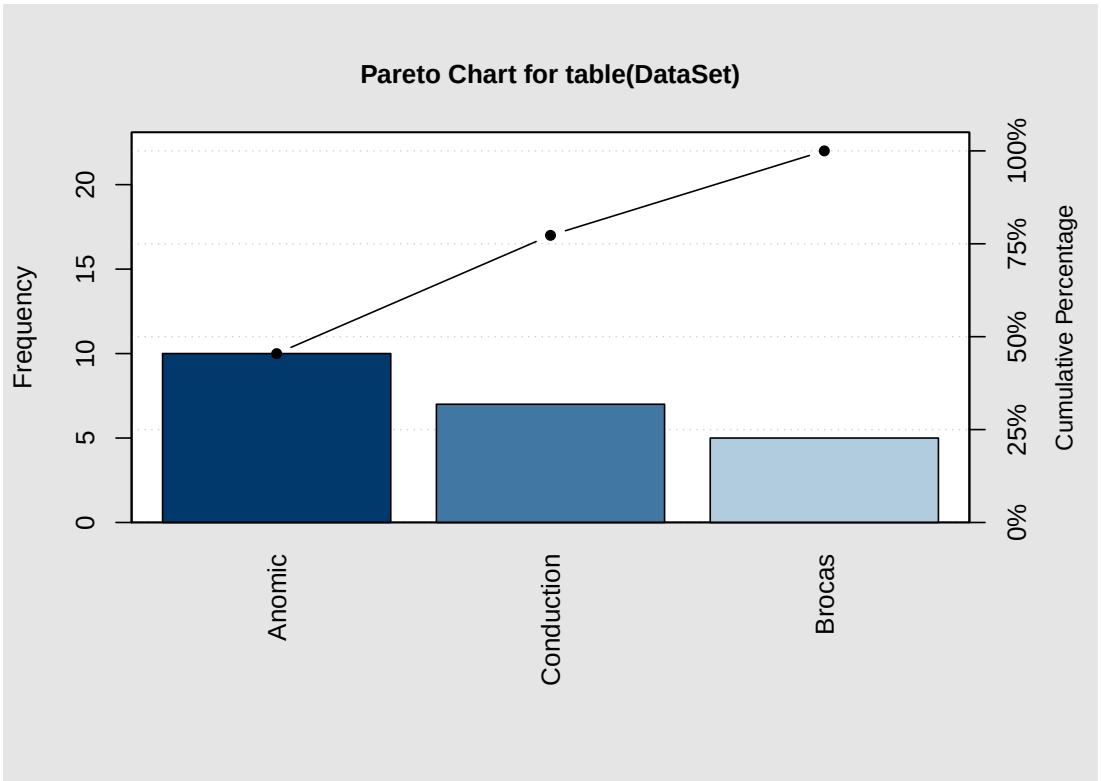
Example 2.1: Aphasia

```
pie(table(DataSet),main="Pie Chart of Aphasia Data")
```



Example 2.1: Aphasia

```
library("qcc")
pareto.chart(table(DataSet))
```



Quantitative data. Graphical description

A variable is quantitative ... . . .

. . . if it represents a measure, given in a meaningful numerical scale: age, height, time, length, concentration, pressure, ...

Graphical summaries of quantitative data:

Their aim is usually to gain insight into the "location" of the data and their "dispersion" around the location.

Dot plots For discrete data and small samples

Stem-and-leaf displays For small samples. Obsolete.

Bar plots For discrete data

Histograms For continuous data and moderate/large samples

Boxplots For continuous data and moderate/large samples

Example 2.2: Siblings

Students in Mr. Smith's class were surveyed as to the number of brothers and sisters in their families (not counting themselves).

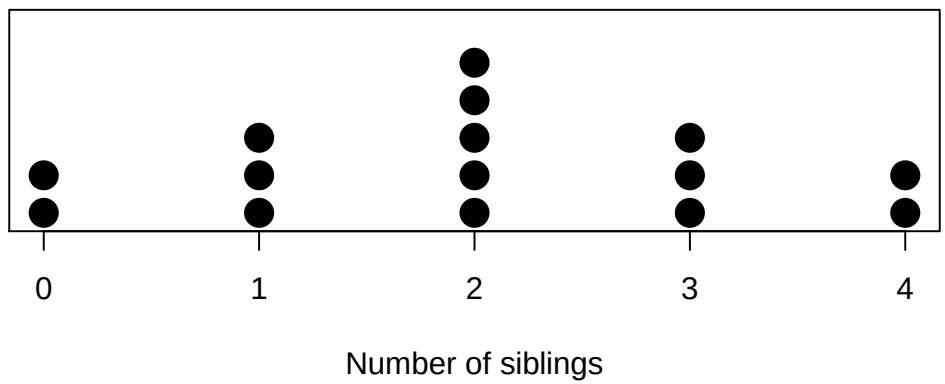
Student	Number of Siblings	Student	Number of Siblings
Allison	2	Bernard	4
Carlos	3	Catherine	2
Delia	2	Dion	1
Emma	0	Fiona	2
Harley	3	Ian	2
Justin	1	Paul	1
Rhianna	3	Stanley	0
Vincent	4		

Dot plots (discrete data and small samples)

- Each observation is displayed using one filled-in circle (dot).
- Repeated observations = dots stacked in a column.
- Column heights represent count (frequency).

Example 2.2: Siblings

```
NSiblings = c(2,3,2,0,3,1,3,4,4,2,1,2,2,1,0)
stripchart(NSiblings, method = "stack", offset = .5,
           at = .1, pch = 19, cex=2,
           xlab = "Number of siblings")
```



Location: The center of the data is at 2 siblings.
Spread: The range of the number of siblings is from 0 to 4. There are no outliers.
Shape: The graph is symmetric.

Example 1.6: Cavendish

The data on the density of Earth were ($n = 29$)

5.50	5.57	5.42	5.61	5.53	5.47	4.88	5.62	5.63	4.07
5.29	5.34	5.26	5.44	5.46	5.55	5.34	5.30	5.36	5.79
5.75	5.29	5.10	5.86	5.58	5.27	5.85	5.65	5.39	

Stem-and-leaf displays (Small samples. Obsolete)

- Group the numbers "in stems" by all-but-last equal digits

  Example 1.6: Cavendish

5.5	5.4	5.6	4.8	4.0	5.2	5.3	5.7	5.1	5.8

- List vertically only once the group digits
- Write last digits (the "leaves") ordered within each group.

Example 1.6: Cavendish

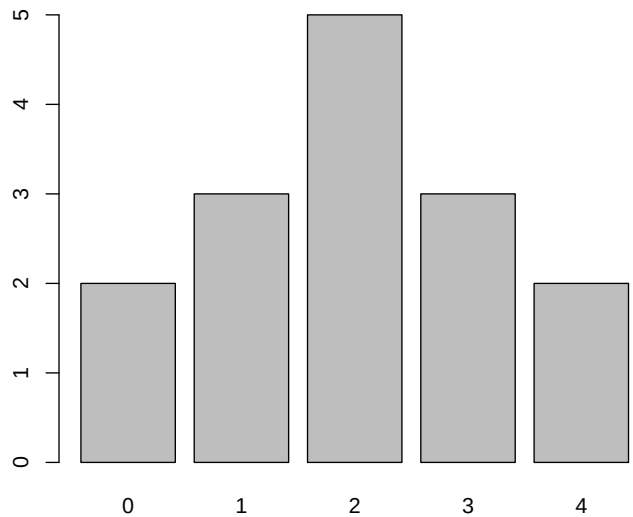
```
EarthDens = scan("Cavendish.txt")
stem(EarthDens)

The decimal point is 1 digit(s) to the left of the |

40 | 7
42 |
44 |
46 |
48 | 8
50 | 0
52 | 679904469
54 | 246703578
56 | 123559
58 | 56
```

Bar plots (discrete data)

- On the horizontal axis we represent the variable possible values.
- We plot a bar on each variable value.
- Column height of the bar equals sample count (frequency).

Example 2.2: Siblings

```
TNSiblings = table(NSiblings)
TNSiblings
NSiblings
0 1 2 3 4
2 3 5 3 2
barplot(TNSiblings)
```



Histograms (continuous data and moderate/large samples)

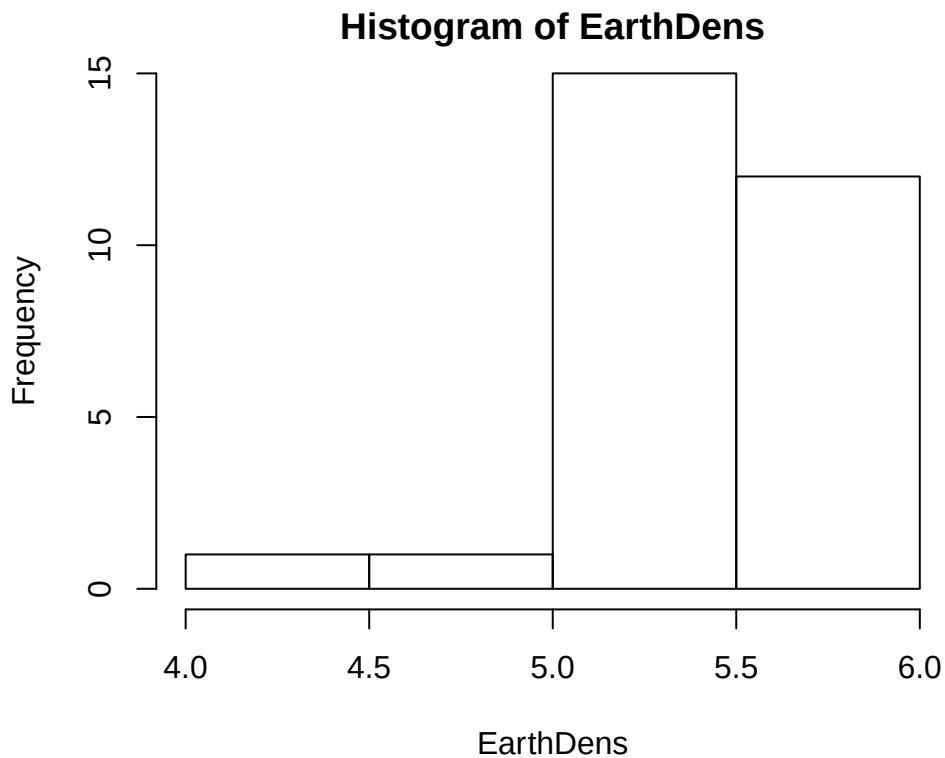
- Data are grouped into intervals $I_1, \ldots, I_k$ (generally of the same length). Each datum belongs to one and only one interval I_i.
- We compute the frequency n_i of each interval I_i (n_i = number of sample observations falling in I_i).
- We divide the horizontal or x-axis into the intervals $I_1, \ldots, I_k$.
- On each interval I_i we plot a bar whose height is the frequency n_i.
- Let n be the sample size and $f_i = n_i/n$ the relative frequency on I_i. Then the proportion or ratio of the area of the bar on I_i over the total area of the histogram area must equal f_i.
- Alternatively (and, as we'll see later, a more convenient choice), we can plot a histogram with total area equal to 1. Then the area of the bar on I_i must be equal to the relative frequency f_i on I_i.

- In practice, if we choose the intervals (although, generally, statistical software has built-in automatic choices), they should contain all the observations and be easy to handle.
- Some rules of thumb for the number k of intervals are $k \simeq \sqrt{n}$ or

n	k
< 25	5–6
25–50	7–14
> 50	15–20

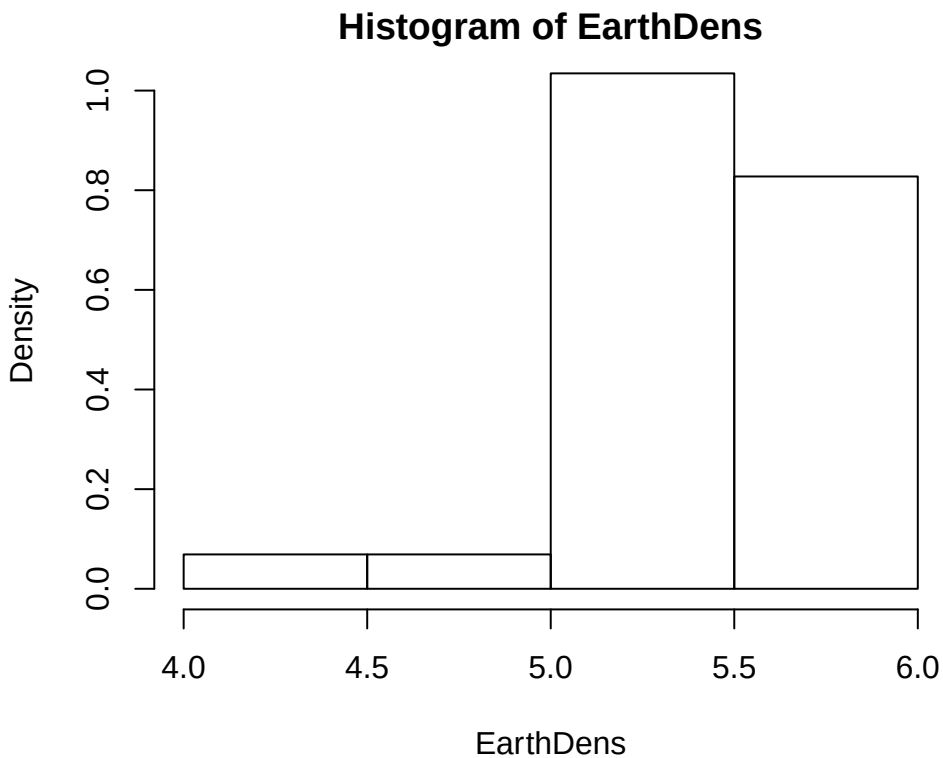
Example 1.6: Cavendish

```
hist(EarthDens)
```



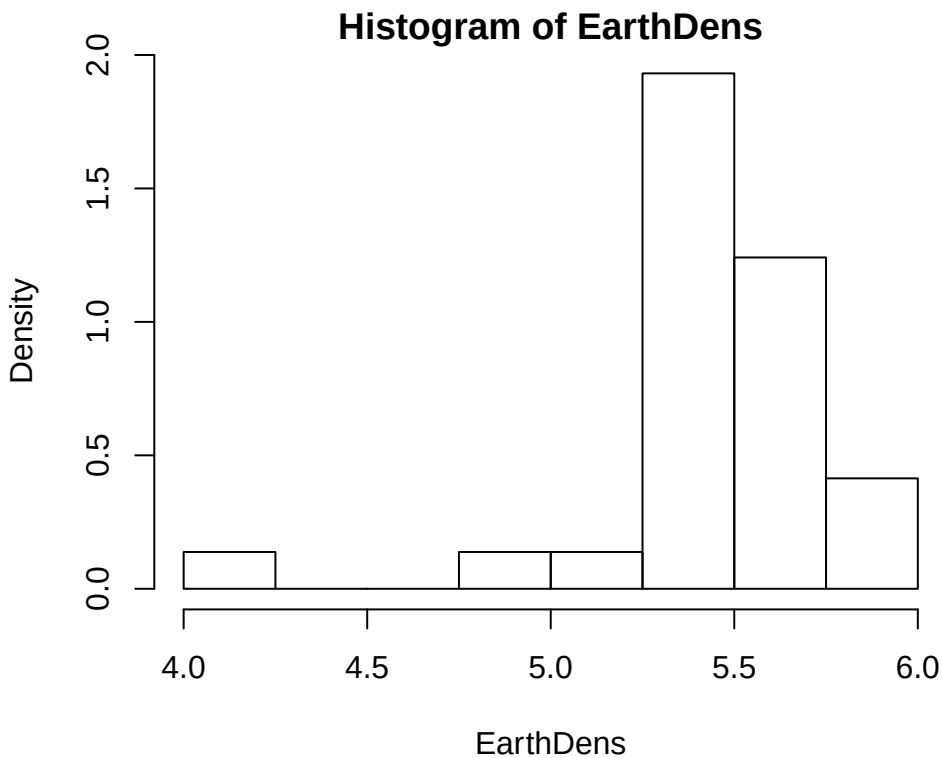
Example 1.6: Cavendish

```
hist(EarthDens,freq=F)
```



Example 1.6: Cavendish

```
hist(EarthDens,freq=F,breaks=seq(4,6,0.25))
```



Example 2.3: Cuckoo's eggs

Latter (1902) investigated the behavior of female cuckoos (*Cuculus canorus*), that lay their eggs on the ground and then move them to the nests of other birds. In particular, Latter gathered data on the lengths of the cuckoo eggs found in these foster-nests, classified by the species of the nest where they where found.

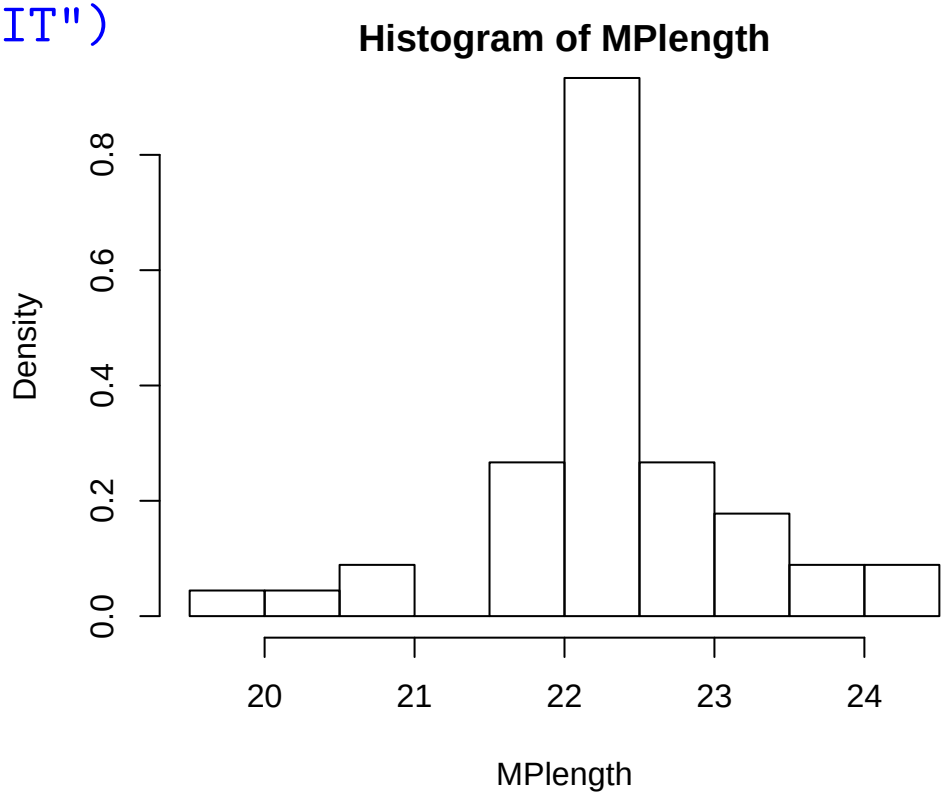
The whole set of data is in file `cuckoodat.txt`.

The following measures, in mm, correspond to the length of the eggs found in Meadow Pipit's (*Anthus pratensis*) nests.

19.65	20.05	20.65	20.85	21.65	21.65	21.65	21.85	21.85
21.85	22.05	22.05	22.05	22.05	22.05	22.05	22.05	22.05
22.05	22.05	22.25	22.25	22.25	22.25	22.25	22.25	22.25
22.25	22.45	22.45	22.45	22.65	22.65	22.85	22.85	22.85
22.85	23.05	23.25	23.25	23.45	23.65	23.85	24.25	24.45

Example 2.3: Cuckoo's eggs

```
OurSample = read.table("cuckoodat.txt",sep=";")
Length = OurSample$V1
Host = OurSample$V2
summary(Host)
 HDGESPRW  MDWPIPIT PIEDWTAIL   ROBIN TREEPIPIT     WREN
       14        45        15      16        15       15
MPindex = (Host == "MDWPIPIT")
MPlength = Length[MPindex]
hist(MPlength,freq=F)
```



Numerical measures of location or central tendency

They describe the tendency of the data to cluster or center about certain numerical values.

The sample mean is the arithmetic mean of the observations in the sample $x_1, \ldots, x_n$:

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

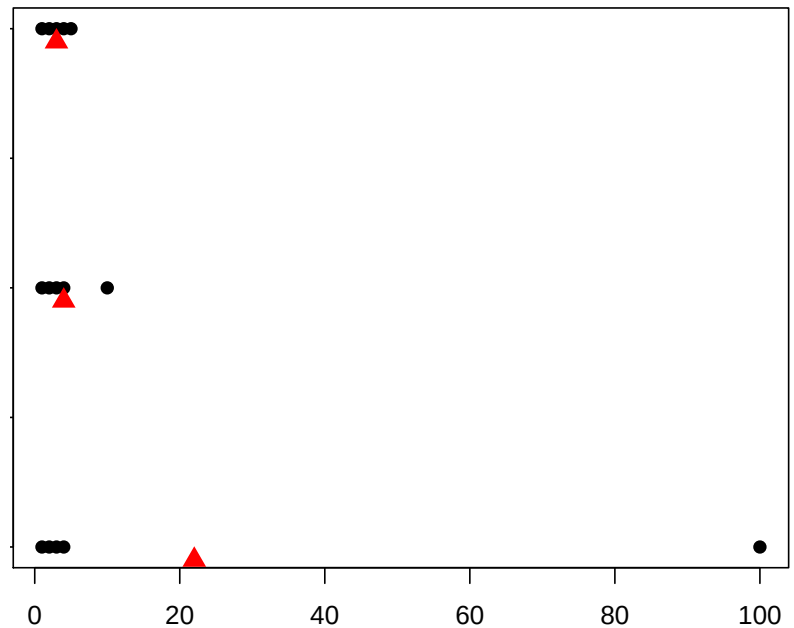
Example 1.6: Cavendish

```
mean(EarthDens)
[1] 5.419655
```

The mean is very sensitive to extremely large or extremely small observations in the sample, by being shifted towards them. We say that the mean is not robust to outliers.

Example 2.4: Naive example

```
v1 = 1:5
v1
[1] 1 2 3 4 5
mean(v1)
[1] 3

v2 = c(1:4,10)
v2
[1]  1  2  3  4 10
mean(v2)
[1] 4

v3 = c(1:4,100)
v3
[1]   1   2   3   4 100
mean(v3)
[1] 22
```



The median M is the point in the center of the sample, when it is arranged in ascending (or descending) order $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Example 1.6: Cavendish

```
EarthDensSort = sort(EarthDens)
EarthDensSort
 [1] 4.07 4.88 5.10 5.26 5.27 5.29 5.29 5.30 5.34
[10] 5.34 5.36 5.39 5.42 5.44 5.46 5.47 5.50 5.53
[19] 5.55 5.57 5.58 5.61 5.62 5.63 5.65 5.75 5.79
[28] 5.85 5.86
```

If the sample size n is an odd number $(n = 2m + 1)$, then $M = x_{(m+1)}$, the central observation in the sorted sample.

Example 1.6: Cavendish

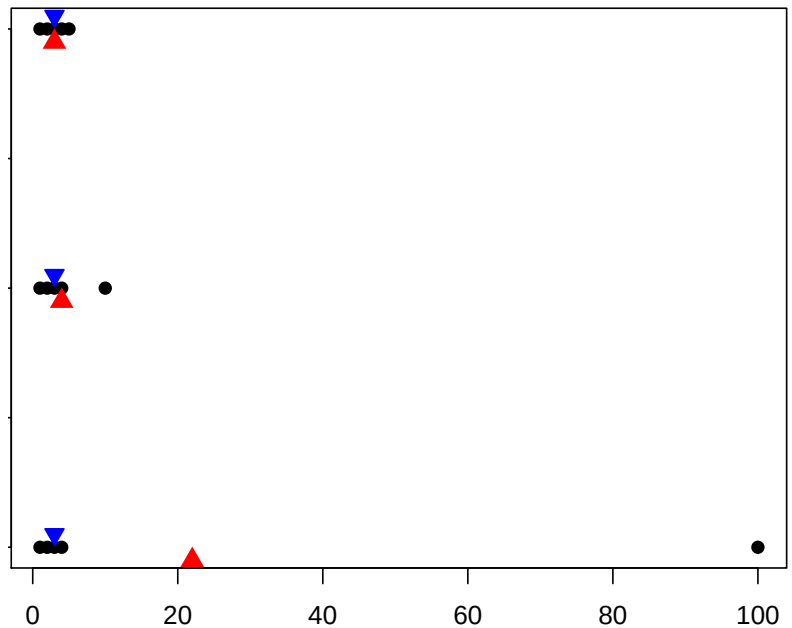
$$n = 29 = 2 \cdot 14 + 1 \longrightarrow M = x_{(14+1)} = x_{(15)} = 5.46$$

```
median(EarthDens)
[1] 5.46
```

The median is not affected directly by extreme measurements, since only the middle measurement (or two of them) is explicitly used to compute M. The median is a robust location descriptor.

Example : Naive example

```
median(v1)
[1] 3

median(v2)
[1] 3

median(v3)
[1] 3
```



The median separates the sorted sample into two subsamples with equal number of individuals ($n/2$).

If the sample size n is even ($n = 2m$), then $M = \dfrac{x_{(m)} + x_{(m+1)}}{2}$, the average of the two most central observations in the sorted sample.

Example 2.5: Smokers Abstinence

A study conducted by researchers investigated whether time perception, an indication of a person's ability to concentrate, is impaired during nicotine withdrawal. After a 24-hour smoking abstinence, 20 smokers were asked to estimate how much time had passed during a 45-second period. The resulting data on perceived elapsed time (in seconds) were as follows:

69	65	72	73	59	55	39	52	67	57
56	50	70	47	56	45	70	64	67	53

```
X = c(69, 65, 72, 73, 59, 55, 39, 52, 67, 57, 56, 50,
   70, 47, 56, 45, 70, 64, 67, 53)
median(X)
[1] 58
```

A data set is called skewed if one tail of the distribution has more extreme observations than the other tail.

In rightward or positevely skewed data, the right tail has more extreme observations.

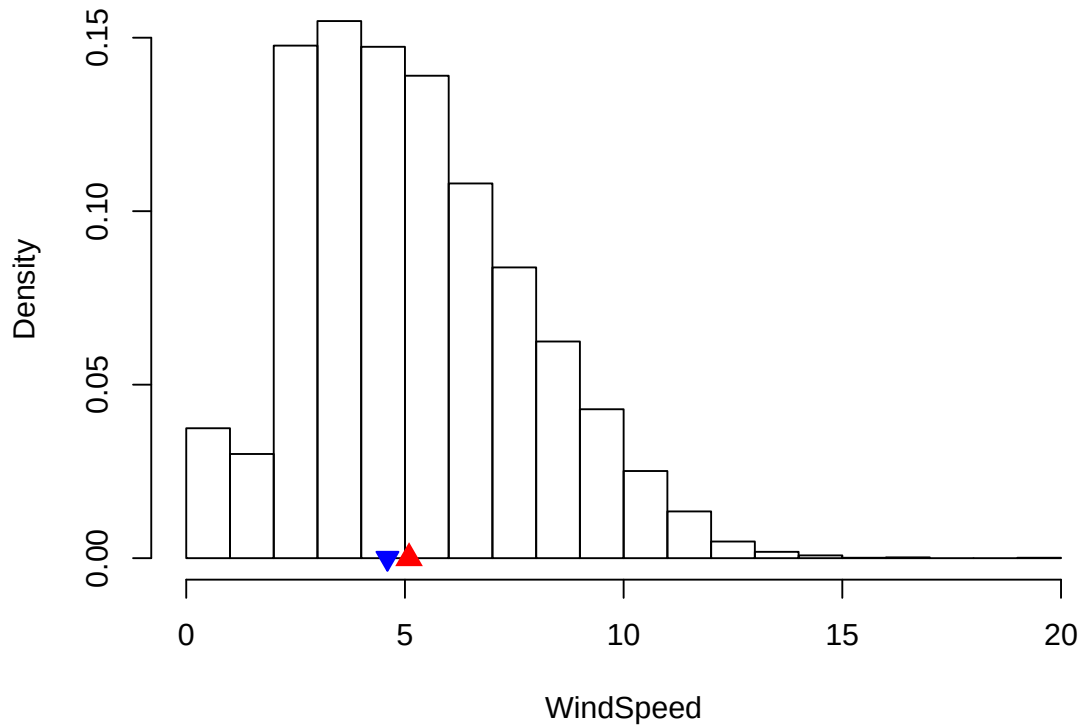
Example 2.6: Wind speed. Source: `www.oz-energy-analysis.org`

In the context of wind energy production, it is necessary to characterize the wind in locations surrounding a wind turbine. This is used to calculate the optimal cut-in and cut-out speed of the turbine and its likely power output.

The Green Grid Report studies the financial viability of installing a wind farm on the Eyre Peninsula (Australia). As part of the study, wind speed data were recorded at several stations of the Australian Bureau of Meteorology (BoM) on the peninsula. We have considered the 2009 hourly wind speeds (in m/s) of the Whyalla Aero BoM station (`WhyallaWindSpeed.txt`).

Example 2.6: Wind speed

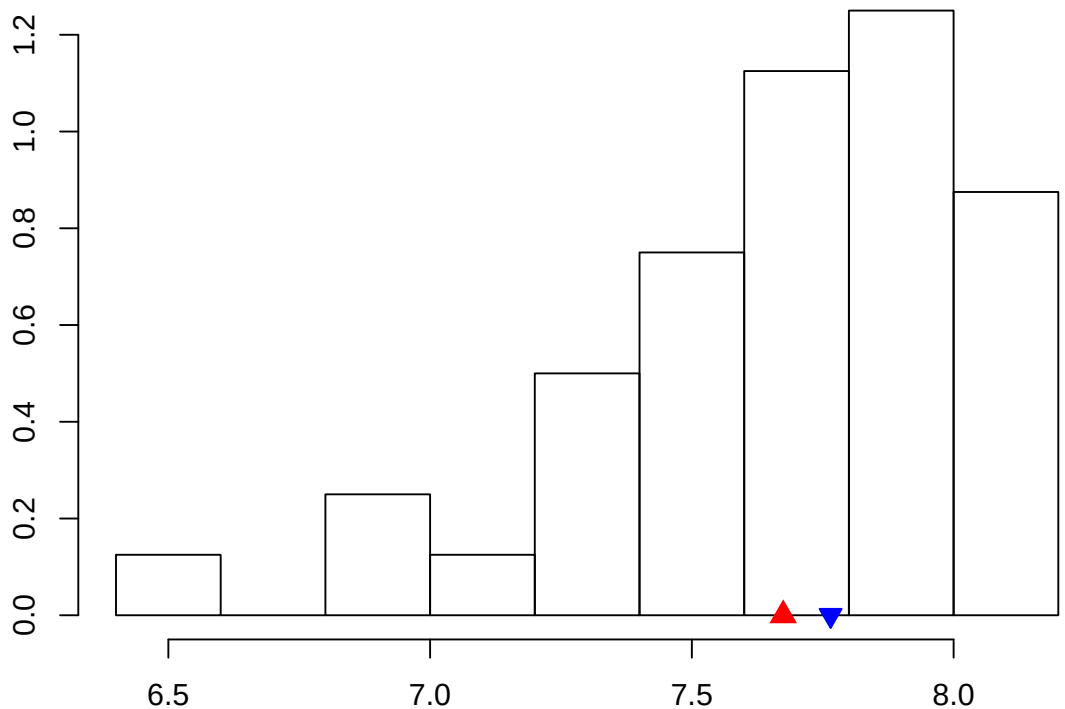
```
WindSpeed = scan("WhyallaWindSpeed.txt")
hist(WindSpeed,freq=F)
points(mean(WindSpeed),0,pch=17,cex=1.5,col="red")
points(median(WindSpeed),0,pch=25,cex=1.2,bg="blue",col="blue")
```



In leftward or negatively skewed data, the left tail has more extreme observations.

Example 2.7: Athletes

Results for the forty athletes who successfully completed a legal jump in the qualifying round of the 2012 Olympic men's long jump.



A third location descriptor is the mode.

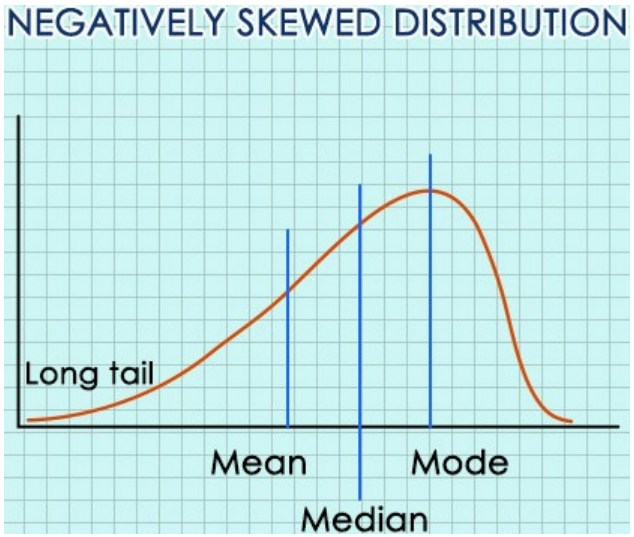
For discrete data, the mode is the observation that occurs most frequently in the data set.

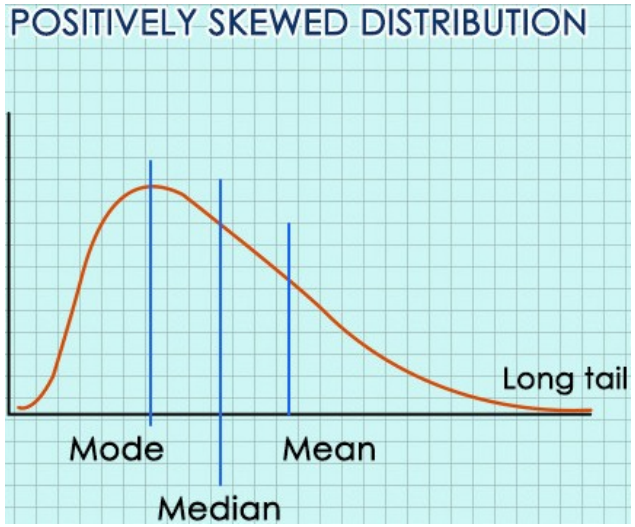
Example 2.2: Siblings

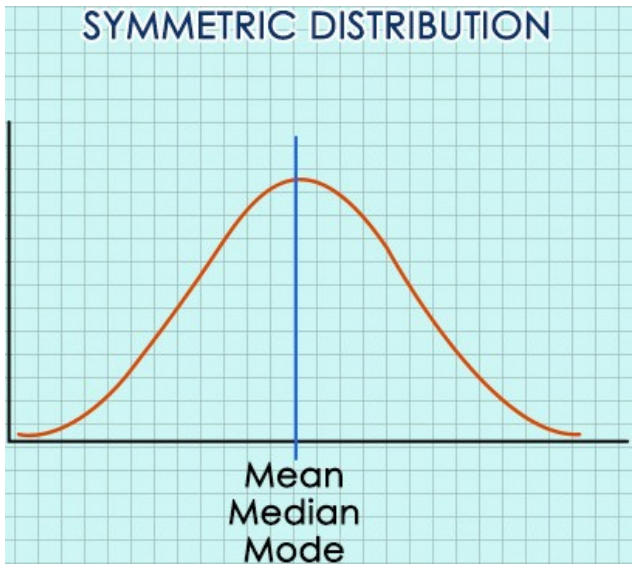
For continuous data, we first draw the histogram. The interval containing the largest frequency is called the modal class. The mode is the midpoint of this interval.

Example 2.3: Cuckoo's eggs

Skewness and the location measures





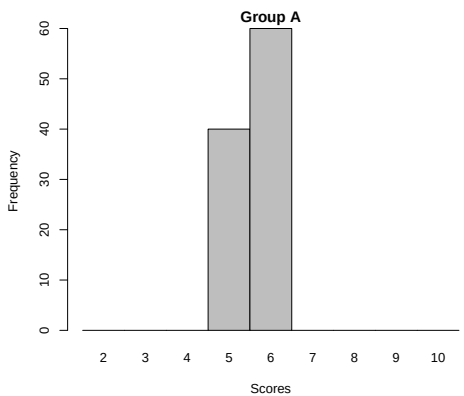


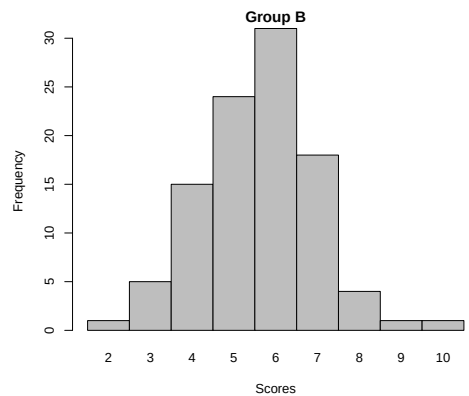
Numerical measures of variability or spread

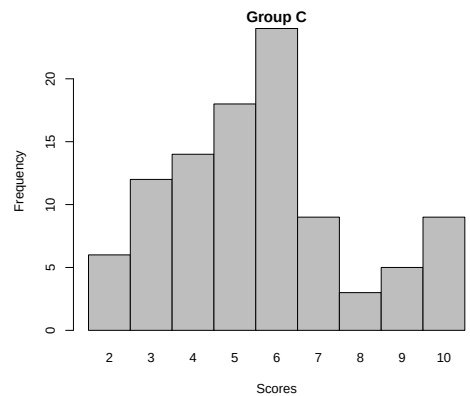
They quantify the variability of the sample with respect to the measure of location. The center descriptor summarizes better the information in the data when the spread is small.

Example 2.8: Scores

	Score obtained								
	2	3	4	5	6	7	8	9	10
Frequency in group A	0	0	0	40	60	0	0	0	0
Frequency in group B	1	5	15	24	31	18	4	1	1
Frequency in group C	6	12	14	18	24	9	3	5	9







Example 2.8: Scores

The sample mean in all the three groups is the same: 5.6.

The simplest measure of variability of a quantitative sample is its range, the difference of the sample maximum and minimum:

$$\text{range} = x_{(n)} - x_{(1)}.$$

Example 2.8: Scores

$$\begin{aligned}\text{Range of Group A} &= 6\text{-}5{=}1\\ \text{Range of Group B} &= 10\text{-}2{=}8\\ \text{Range of Group C} &= 10\text{-}2{=}8\end{aligned}$$

The range is easy to compute and understand. But two large data sets might have the same range and have a very different spread and shape.

Example 1.6: Cavendish

```
range(EarthDens)
[1] 4.07 5.86
range(EarthDens)[1]
[1] 4.07
range(EarthDens)[2]
[1] 5.86
range(EarthDens)[2]-range(EarthDens)[1]
[1] 1.79
```

Alternative, more sensitive, measures of data spread, are the deviations of the observations with respect to the sample mean $\bar{x}$:

$$x_1 - \bar{x},\ x_2 - \bar{x},\ \ldots, x_n - \bar{x}.$$

If they are large, the data are spread out. If the deviations are small, then the data are clustered around $\bar{x}$.

Disadvantage: The sum of these discrepancies is 0:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0.$$

So we define the discrepancies of the observations with respect to the sample mean as the squared differences

$$(x_1 - \bar{x})^2,\ (x_2 - \bar{x})^2,\ \ldots, (x_n - \bar{x})^2.$$

We summarize this information defining the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Example 2.8: Scores

$$s_A^2 = \frac{1}{99}\left(40(5-5.6)^2 + 60(6-5.6)^2\right) =$$

$$s_B^2 =$$

$$s_C^2 =$$

Example 1.6: Cavendish

```
var(EarthDens)
[1] 0.1148392
```

The unit of measurement of the sample mean is that of the data.

The unit of measurement of the sample variance is the square of the unit in the data.

To have an idea of the spread in terms of the unit of measure in the variable, we can take the sample standard deviation (sd), which is the square root of the variance:

$$s = \sqrt{s^2}.$$

Example 2.8: Scores

$s_A =$

$s_B =$

$s_C =$

Example 1.6: Cavendish

```
sqrt(var(EarthDens))
[1] 0.3388793
sd(EarthDens)
[1] 0.3388793
```

Extracting info from the standard deviation

The Chebishev rule, very conservative and valid for any data set, states that the fraction of points within k standard deviations from the mean is at least

$$1 - \frac{1}{k^2} = \frac{k^2 - 1}{k^2}$$

Thus, at least 3/4 of the data will fall within 2 sd's of the mean, i.e., within the interval $(\bar{x} - 2s, \bar{x} + 2s)$.

At least 8/9 of the data will fall within 3 sd's of the mean, i.e., within the interval $(\bar{x} - 3s, \bar{x} + 3s)$.

The normal rule, valid for special types of symmetric data (bell-shaped with light tails), states that

- 68% of the data are within one sd from the mean
- 95% of the data are within two sd's from the mean
- More than 99% of the data are within three sd's from the mean

Numerical measures of relative standing

They describe the relative quantitative location of a particular measurement within a data set, that is, relative to the rest of the data.

For $p \in (0, 1)$ we define quantile p, q_p, o $100p$-th percentile as the value leaving a proportion p of the sample data (arranged in increasing order) "to the left" (i.e., a proportion p of the data is smaller than q_p) and the rest of the data "to the right" (i.e., a proportion $1 - p$ of the data is larger than q_p).

The method for computing percentiles with small data sets varies according to the software used. As the sample size increases, the percentiles from the different software packages will converge to a single number.

In general, there are various (similar) formulas for computing the sample $100p$-th percentile.

They all compute a weighted average of two consecutive observations, $x_{(j)}$ and $x_{(j+1)}$, in the arranged sample leaving approximately a proportion p of the data "to the left".

For a large sample size n, the results of all the methods are similar. R offers a large number (9) of methods to compute the quantiles.

R's default quantile is Type 7. To compute it, use the following decomposition

$$p(n-1)+1 = k + r \quad \text{with } k \text{ an integer and } \; 0 \leq r < 1.$$

Then

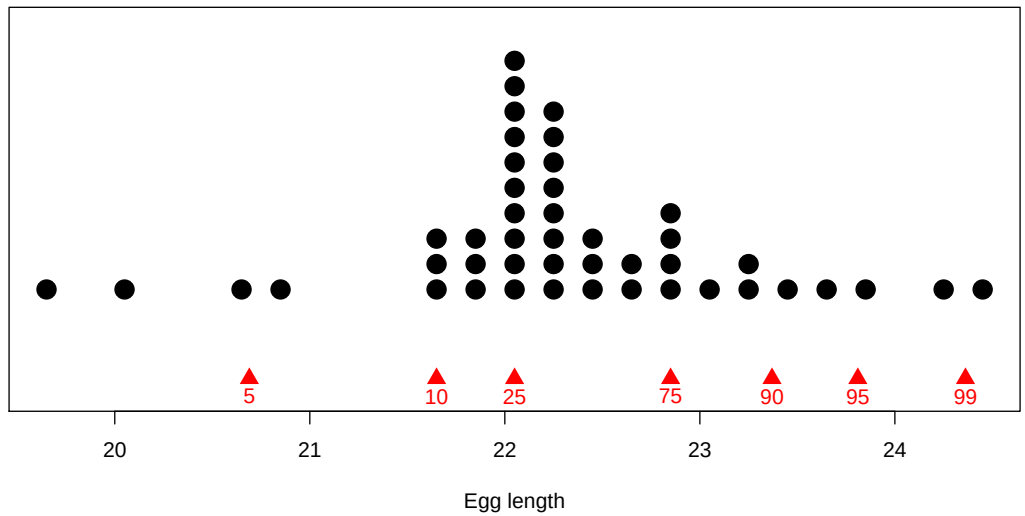
$$q_p = (1-r)x_{(k)} + r\, x_{(k+1)}.$$

Example 2.3: Cuckoo's eggs

Length of cuckoos' eggs found in Meadow Pipit's nests

19.65	20.05	20.65	20.85	21.65	21.65	21.65	21.85	21.85
21.85	22.05	22.05	22.05	22.05	22.05	22.05	22.05	22.05
22.05	22.05	22.25	22.25	22.25	22.25	22.25	22.25	22.25
22.25	22.45	22.45	22.45	22.65	22.65	22.85	22.85	22.85
22.85	23.05	23.25	23.25	23.45	23.65	23.85	24.25	24.45

```
quantile(MPlength,0.9)
  90%
23.37
quantile(MPlength,0.95)
  95%
23.81
```




```
quantile(MPlength,probs=c(0.05,0.1,0.25,0.75,0.9,0.95,0.99))
    5%    10%    25%    75%    90%    95%    99%
20.690 21.650 22.050 22.850 23.370 23.810 24.362
```

Quartiles are the percentiles partitioning the arranged sample into quarters:

- The first or lower quartile $Q_L = q_{0.25}$ is the value leaving 25%=1/4 of the sorted data "to the left" and 75% of the sorted data "to the right".

  Example 2.3: Cuckoo's eggs

  25% of the eggs' lengths are smaller than 22.050 mm and 75% are larger.

- The second or middle quartile $Q_M = q_{0.5}$ is the median M.
- The third or upper quartile $Q_U = q_{0.75}$ leaves 75%=3/4 of the sorted data "to the left" and 25% "to the right".

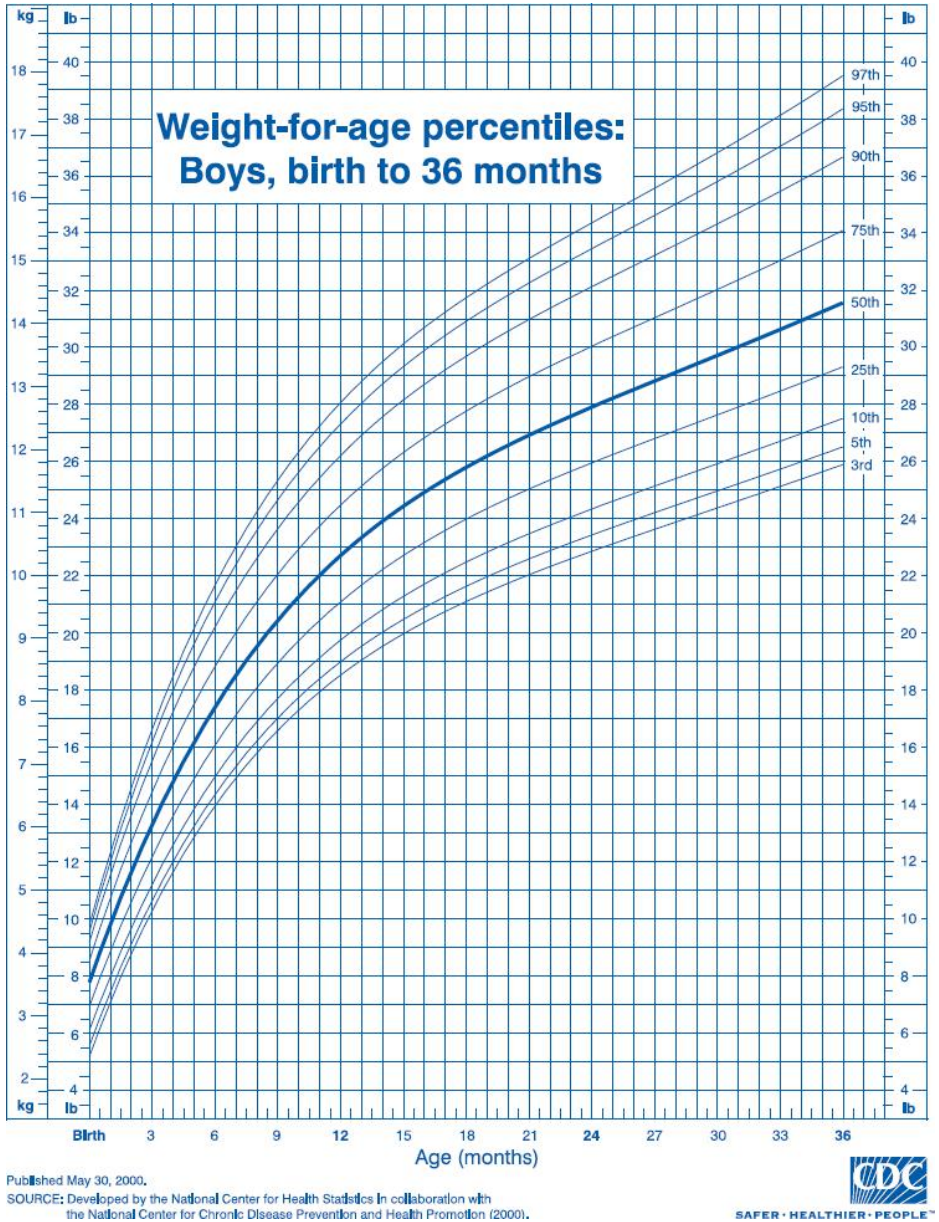
  Example 2.3: Cuckoo's eggs

  $Q_U = 22.850$

  50% of the eggs' lengths are between 22.050 and 22.850 mm.

Example 2.9: Weight-for-age or growth chart

3rd, 5th, 10th, 25th, 50th, 75th, 90th, 95th, 97th percentiles of boys weight in terms of age, from birth to 36 months



Another measure of relative standing of x with respect to the sample is the z-score

$$z = \frac{x - \bar{x}}{s}.$$

It represents the distance between x and the mean $\bar{x}$, expressed in standard deviations.

When we compute the z-score of x we say that we standardize x.

Example 2.10: SAT scores

Suppose a sample of 2000 high school seniors' verbal SAT scores is selected. The mean and standard deviation (s.d.) are $\bar{x} = 550$ and $s = 75$. Suppose Joe Smith's score is 475. What is his z-score?

$$z = \frac{475 - 550}{75} = -1.0$$

tells us that Joe Smith's score is 1.0 s.d.'s below the sample mean.

Boxplots or box-and-whiskers plots

It is a diagram that makes the interpretation of quartiles easier, aids in the detection of outliers and gives insight into the symmetry of the data set.

Outliers are unusually large or small observations relative to the other observations in the data set.

Outliers may be due to one of the following causes:

1. The measurement is incorrect.
2. The measurement comes from a different population.
3. The measurement is correct but corresponds to an odd event in the population.

To construct the boxplot, we compute Q_L, M, Q_U, the interquartile range

$$\text{IQR} = Q_U - Q_L,$$

the inner fences

$$\begin{aligned} \text{Lower inner fence} &= Q_L - 1.5\ \text{IQR}, \\ \text{Upper inner fence} &= Q_U + 1.5\ \text{IQR}, \end{aligned}$$

the limits of the whiskers

$$\begin{aligned} \text{Lower limit} &= \text{Smallest observation inside the inner fences} \\ \text{Upper limit} &= \text{Largest observation inside the inner fences} \end{aligned}$$

and the outer fences

$$\begin{aligned} \text{Lower outer fence} &= Q_L - 3\ \text{IQR}, \\ \text{Upper outer fence} &= Q_U + 3\ \text{IQR}. \end{aligned}$$

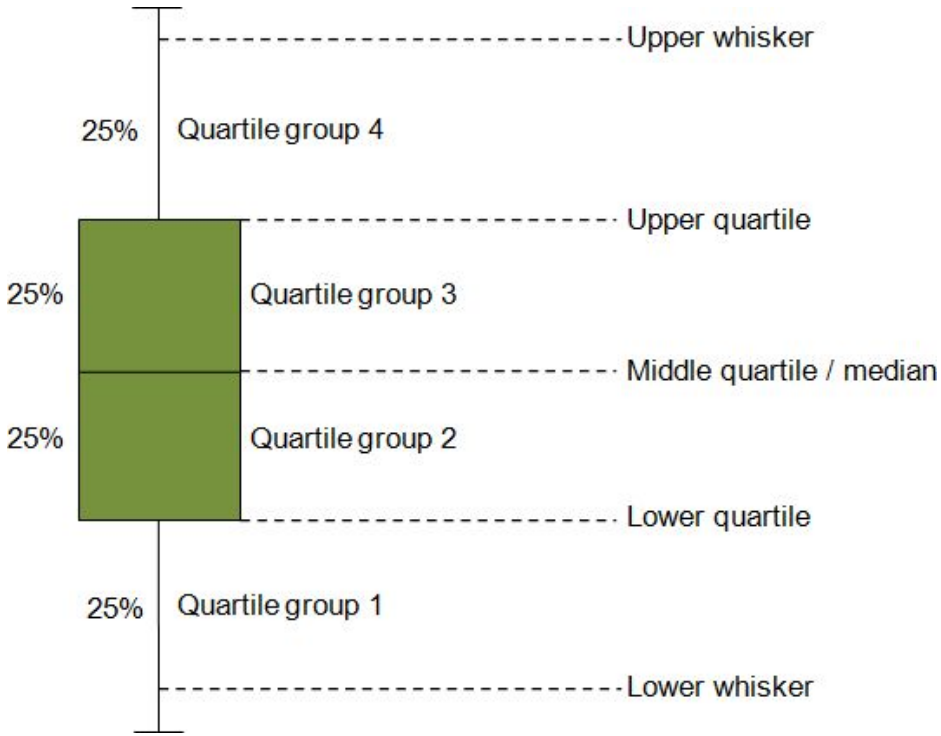
The hinges of the box are Q_L and Q_U.

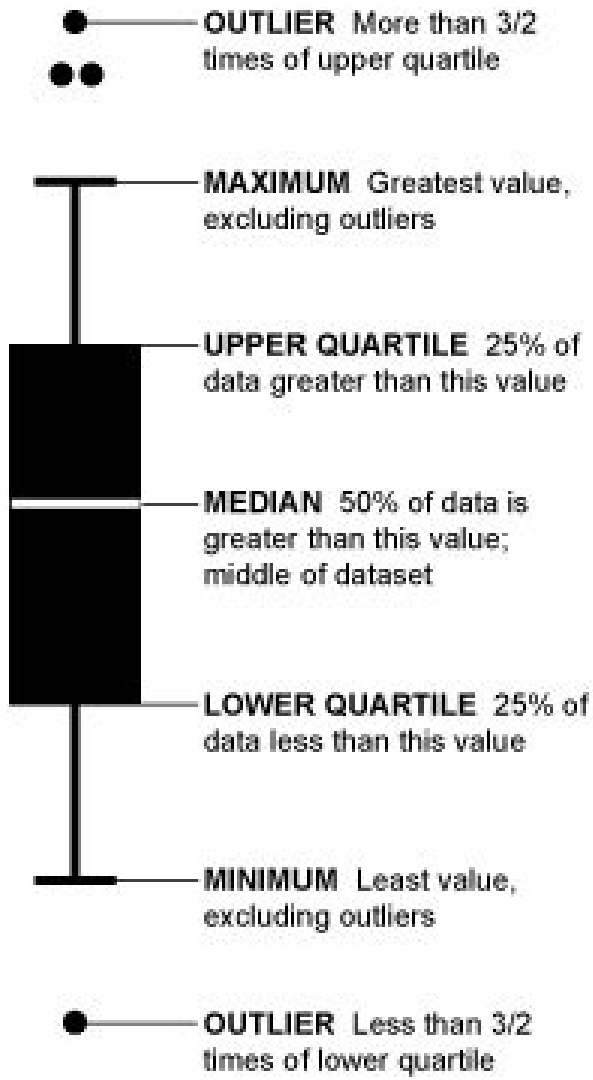
Values beyond the inner fences are marked as outliers. Values outside the outer fences are sometimes marked differently.

Asymmetry of the sample shape causes observations to be erroneously detected as outliers. We have to be cautious when interpreting boxplots of skewed data.

Tips for interpreting boxplots:

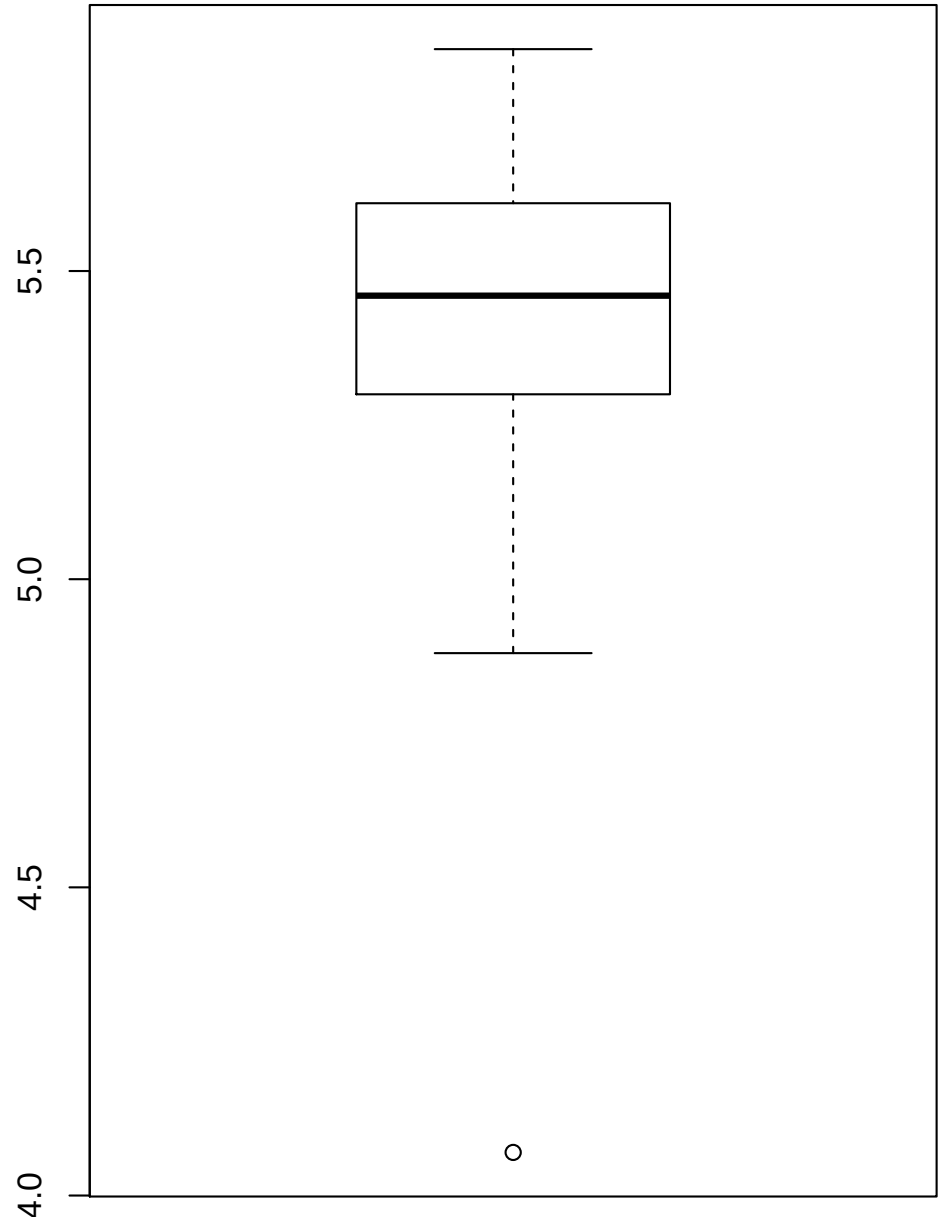
- The line (M) inside the box represents the "center" of the sample.
- The IQR is a measure of sample variability.
- If the lengths of the whiskers are very different, the data are probably skewed in the direction of the longest whisker.
- Check for outliers.


Upper whisker
25%
Quartile group 4
Upper quartile
25%
Quartile group 3
Middle quartile / median
25%
Quartile group 2
Lower quartile
25%
Quartile group 1
Lower whisker


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

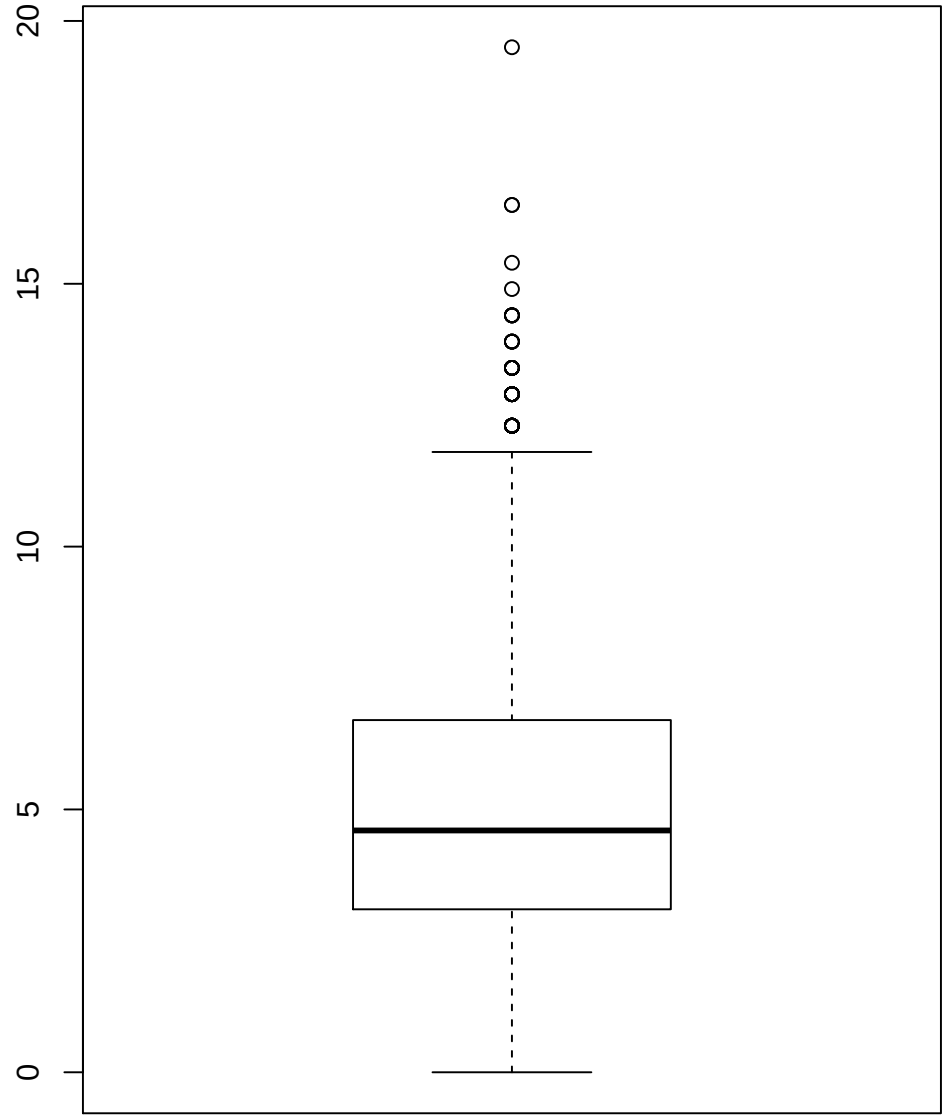
Example 1.6: Cavendish

```
boxplot(EarthDens)
```



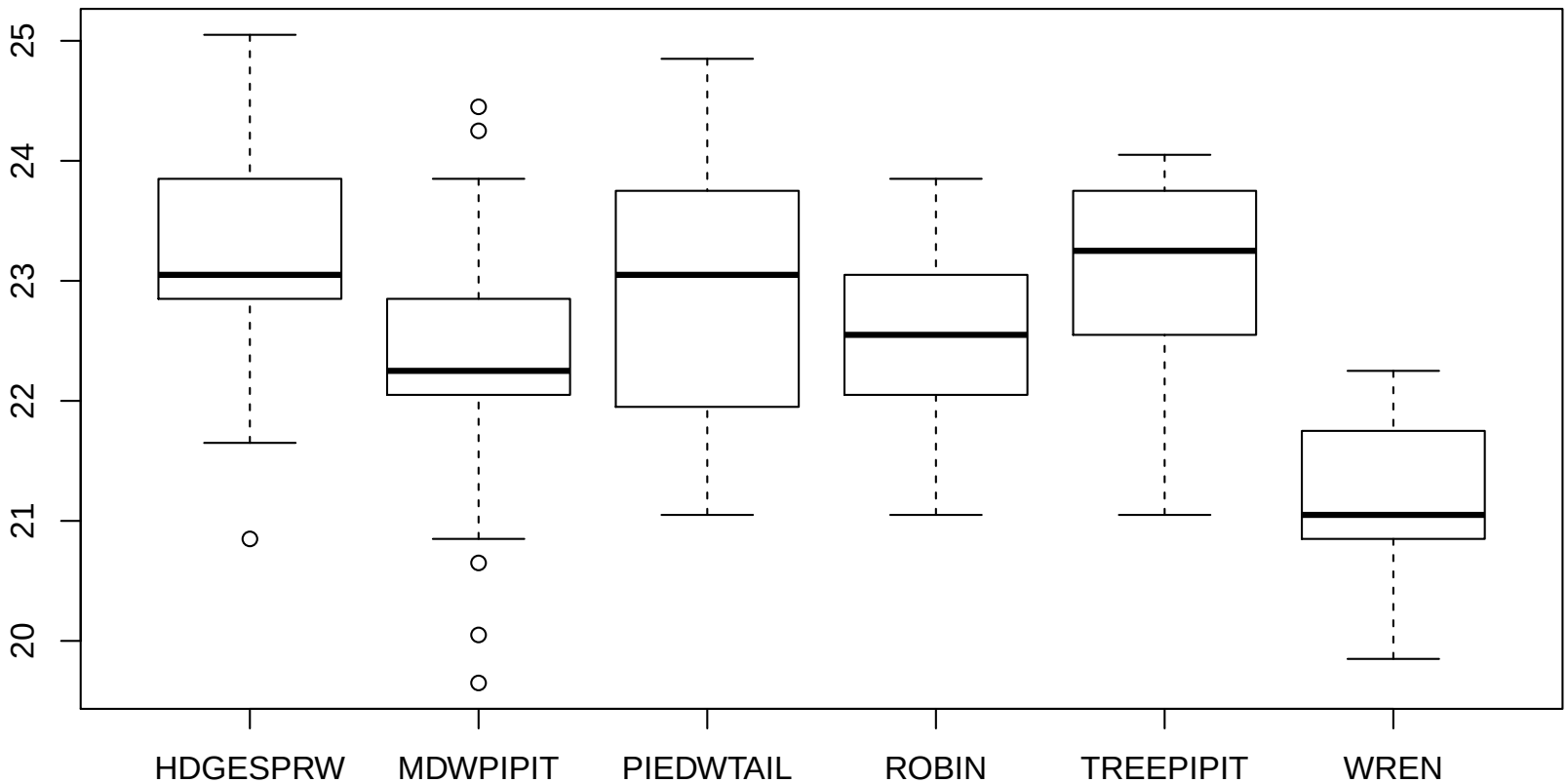
Example 2.6: Wind speed

```
boxplot(WindSpeed)
```



Example 2.3: Cuckoo's eggs

```
boxplot(Length~Host)
```



Observations with z-scores greater than 3 in absolute value are considered possible outliers and should be checked.

Example 1.6: Cavendish

The z-score of the minimum of the sample, 4.07, with respect to the rest of the sample is

$$\frac{4.07 - 5.4679}{0.2219} = -6.3.$$

Graphing bivariate data

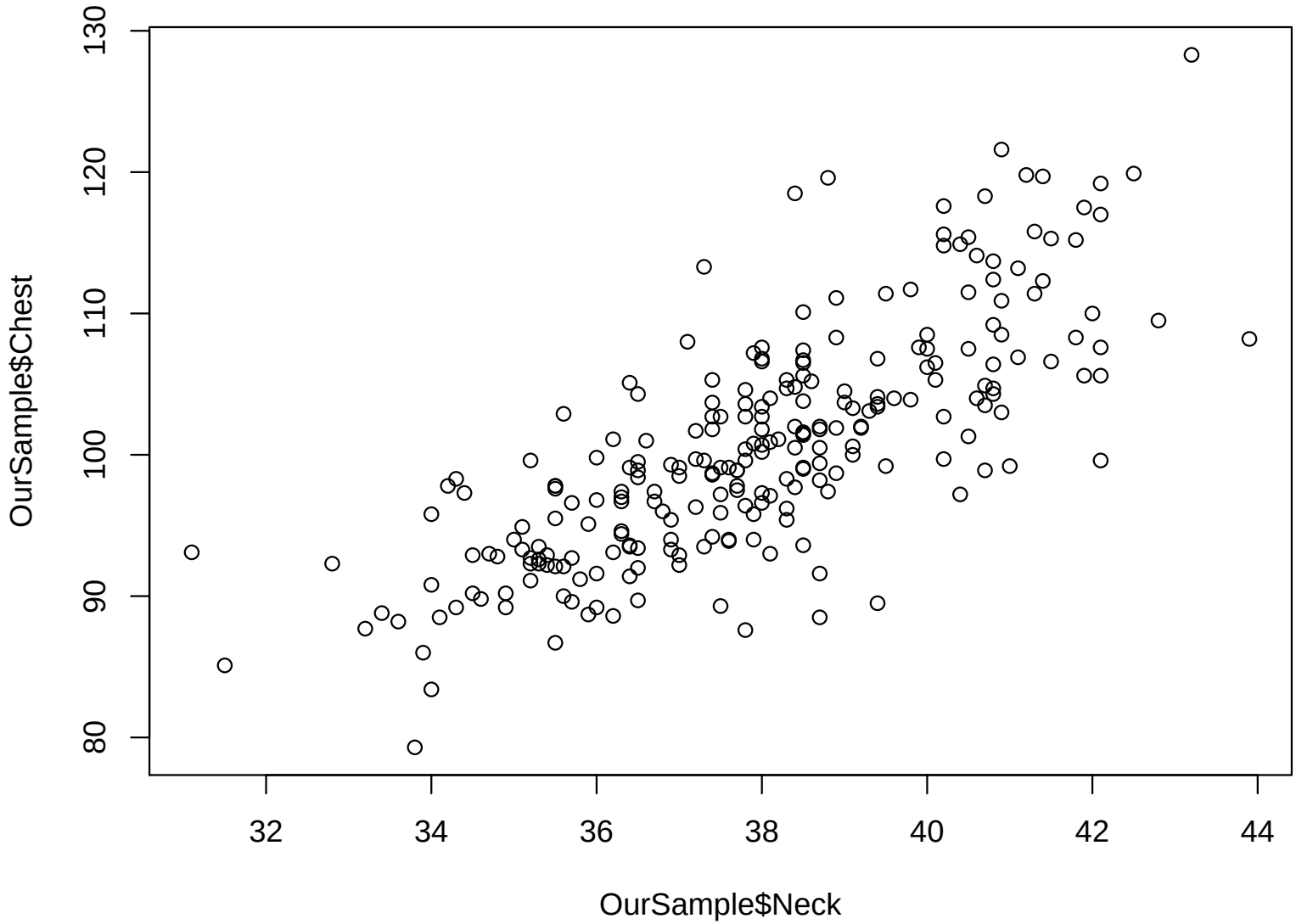
The words correlated and associated imply a relationship between two quantitative variables.

The first way to check on the relationship between two quantitative variables (a bivariate relationship) is to draw a scatterplot.

In a scatterplot we draw two perpendicular axes and plot each sample individual as a point on the plane choosing one of the variables for the horizontal axis and the other variable for the vertical axis.

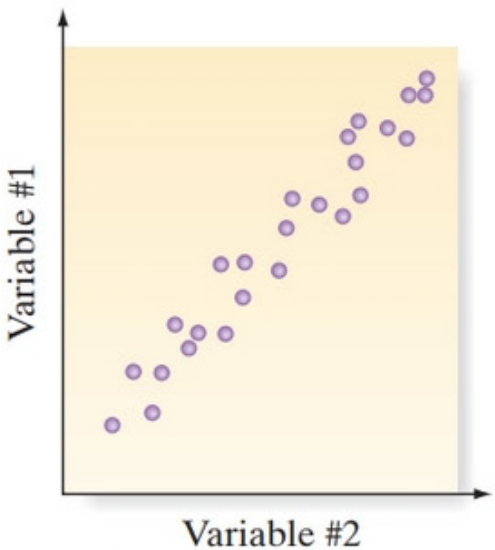
Example 1.1: Bodyfat

```
OurSample = read.table("bodyfat.txt",header=TRUE)
plot(OurSample$Neck,OurSample$Chest)
```

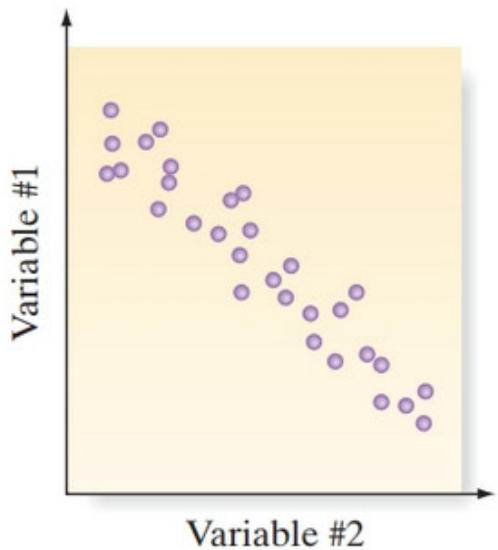


When an increase in one variable is usually associated with an increase in the second variable, we say that the variables are "positively related". If the relationship is linear and positive, then the variables are positively correlated.

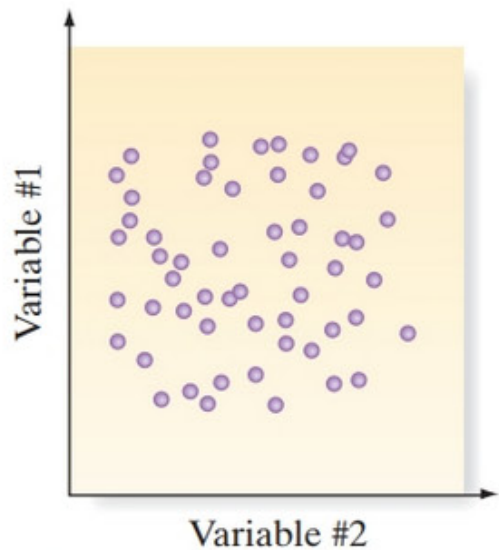
If the relationship is linear and one variable tends to decrease as the other increases, then the variables are "negatively correlated".



a. Positive relationship



b. Negative relationship



c. No relationship