EXPLORATORY DATA ANALYSIS

Chapter 1:
Data handling, descriptive statistics & graphical methods

- Descriptive Statistics
- Quantiles and the five number summary
- Distribution diagrams
- Bivariate data, scatter plots and correlation

General comments on Data Handling

Ordering

choose appropriate order for categorical data eg decreasing frequency (ordinal data have a specified order)

Coding

conveneient to code data to numerical values for categorical/ordinal data e.g. female=1, male=2 (only for convenience)

Checking

always necessary. Most important check is common sense; do results and conclusions agree with our common sense. If not, why not? Can we explain differences?

General comments on Data Handling

- Significant figures - two to three figures is usually best for the presentation of our results; in statistics want to round off to meaningful level.
- Transformations - if data contain widely differeing values e.g. 0.001 – 1000 then transform to same scale. Common are from log() $(0, \infty)$ to real line $(-\infty, \infty)$ and logit() from $(0, 1)$ to real line $(-\infty, \infty)$. Beware the interpretation of results!

Descriptive Statistics

Main roles of descriptive statistics are to:

- Detect anomalies;
- Examine and summarise the data;
- Communicate results.

Types of variables

$$\begin{aligned} \text{categorical variable} &\quad \leftrightarrow \quad \text{category} \\ \text{ordinal variable} &\quad \leftrightarrow \quad \text{category} + \text{order} \\ \text{numerical variable} &\quad \leftrightarrow \quad \text{category} + \text{order} + \text{scale} \end{aligned}$$

An ordinal variable contains more information than a categorical variable and a numerical variable contains more information than an ordinal variable. The more information in the variable, the more that can be done with it. Thus the treatment depends on the variable type:

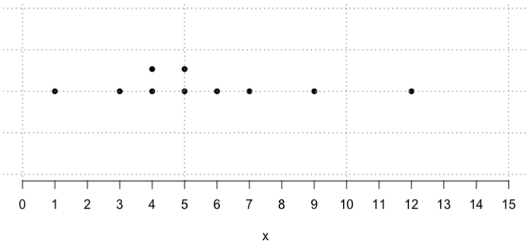
	variable type		
Data description	categorical	ordinal	numerical
frequency distribution	yes	yes	yes
cumulative/quantiles	$\times$	yes	yes
moment statistics ($\bar{x}, s$)	$\times$	$\times$	yes

Descriptive Statistics

We look at the more important and useful statistics; and mention a few others. To begin, let's look at an example:

x: 4 5 4 6 1 9 7 3 12 5

Perhaps the simplest, and one of the first, things to do with a data set is to construct a dotplot. This gives a quick idea of the



distribution.

R code: Dotplot

```
> x=c(4, 5, 4, 6, 1, 9, 7, 3, 12, 5)
> stripchart(x,            # name of object to plot
    method="stack",        # stack points
    xlab = "x",            # x-axis label
    pch=16,                # uses filled circles
    offset=1.5,            # vertical distance between stacked points
    frame.plot=0,          # no frame around plot
    axes = FALSE,          # to specify tick marks
    xlim=c(0,15))          # specify the limits on the axis
> axis(side=1,at = seq(0,15,by=1))       # specify tick marks
> grid(ny=NULL,nx=NULL,col="darkgray")      # gridlines on
x-axis and y-axis
```

Descriptive statistics: simple example

To begin, let's look at an example:

$$x: \quad 4 \quad 5 \quad 4 \quad 6 \quad 1 \quad 9 \quad 7 \quad 3 \quad 12 \quad 5$$

Descriptive statistics are numbers derived from the data to describe various features of the data. Here is an example:

Mean $= \bar{x} =$ sample mean $= \frac{1}{n}\sum_{i=1}^{n} x_i = 56/10$

Sample standard deviation
StDev = s, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

```
> mean(x)

[1] 5.6

> sd(x)

[1] 3.134042
```

Descriptive statistics: simple example

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	4.00	5.00	5.60	6.75	12.00

`Median` = $\hat{c}_{0.5}$ = sample median = 5;

middle observation: (1 3 4 4 5 5 6 7 9 12)

`TrMean` = trimmed mean = 43/8; *(10% trimmed mean)*

`Min` = sample minimum = 1, `Max` = sample maximum = 12;

```
> summary(x)
```

Output is appropriate no matter whether x is discrete or continuous (x could actually be either, if we didn't already know)

Descriptive statistics: simple example

x: 4 5 4 6 1 9 7 3 12 5

Q1 = lower (first) quartile = $\hat{c}_{0.25} = x_{(2.75)} = ??$
Q3 = upper (third) quartile = $\hat{c}_{0.75} = x_{(8.25)} = ??$

IQR (=Q3 - Q1) contains 50% of the sample.

Order statistics: simple example

x:	4	5	4	6	1	9	7	3	12	5

For the above sample: $x_{(1)} = 1, x_{(2)} = 3, x_{(3)} = 4, \ldots, x_{(10)} = 12$. (sort)

So what is $Q_1 = \hat{c}_{0.25} = x_{(2.75)}$?

$x_{(2.75)}$ is 0.75 of the way from $x_{(2)} = 3$ to $x_{(3)} = 4$;

thus $x_{(2.75)} = 3.75$. (*linear interpolation*)

Note that $\hat{c}_{0.5} = x_{(5.5)}$. Check that $Q_3 = \hat{c}_{0.75} = x_{(8.25)} = 7.5$.

In R

```
> sort(x) #arranges x in increasing magnitude
   [1] 1 3 4 4 5 5 6 7 9 12
```

Order statistics definition

The median and quartiles are all examples of sample quantiles.
These are calculated by the order statistics.

Sample quantile

The sample quantile $\hat{c}_q$ is defined as the value such that a proportion q of the sample is less than $\hat{c}_q$.

Order statistics

Arrange the sample $x_1, x_2 \ldots, x_n$ in increasing magnitude:
$x_{(1)} \leqslant x_{(2)} \leqslant \ldots x_{(n)}$. Then $x_{(k)}$ is called the kth order statistic.

Then $\hat{c}_q = x_{(k)}$, where $k = (n+1) \times q$.

Example - find the quantiles

$$\texttt{x:}\quad 1\ 3\ 4\ 4\ 5\ 5\ 6\ 7\ 9\ 12, \qquad n = 10$$

What is the median? $\hat{m} = \hat{c}_{0.5} = x_{(k)} =?$
Find k:
$k = (n+1) \times q = 11/2 = 5.5$
Thus $\hat{m} = \hat{c} = x_{(5.5)}$, half way between $x_{(5)}$ and $x_{(6)}$
i.e. half way between 5 and 5: (5+5)/2 = 5!

What is Q3? $\hat{c}_{0.75} = x_{(k)} =?$
$k = 11 \times 0.75 = 8.25$
Thus,
$\hat{c}_{0.75} = x_{(8.25)} = x_{(8)} + 0.25 \times (x_{(9)} - x_{(8)})$
$\hat{c}_{0.75} = 7 + 0.25 \times (9 - 7) = 7 + 0.5$
Q3 = $\hat{c}_{0.75}$ = 7.5 (R: > quantile(x, probs=0.75,type=6))

UFC data

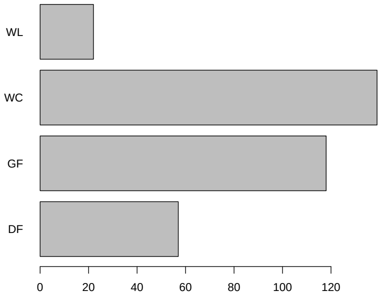
```
> ufc = read.csv("../data/ufc.csv")
> str(ufc) # structure of data ie of variables in the data
set
'data.frame':  336 obs.  of 5 variables:
  $ plot :  int 2 2 3 3 3 4 4 5 5 6 ...
  $ tree :  int 1 2 2 5 8 1 2 2 4 1 ...
  $ species :  Factor w/ 4 levels "DF", "GF", "WC", ...:
1 4 2 3
  $dbh.cm :num 39 48 52 36 38 46 25 54.9 51.8 40.9...
  $ height.m:  num 20.5 33 30 20.7 22.5 18 17 29.3 29 26
..

> summary(ufc$dbh.cm) # tree diameter

   Min.  1st Qu.  Median   Mean  3rd Qu.    Max.
  -5.00    24.73   35.00  37.28    47.15  101.50
```

Better clean that up!

Barplot

A *bar chart* is often used to show the distribution of a categorical variable or an ordinal variable.
The number of trees of each species is plotted in

```
> barplot(table(ufc$species), horiz = TRUE)
```



Barplot

In a barchart or bargraph, vertical "bars" represent the observed frequency of certain values or categories.

This is suitable for categorical, ordinal and discrete numerical data.

Bars should be of equal width and separated from each other so as not to imply continuity.
Either use:
observed frequencies $freq(a < X < b)$ or relative frequencies $freq(a < X < b)/n$, for $a < x < b$.

Histogram

A histogram is a bargraph for continuous data.

Since each observation of a continuous variable is always distinct, we group the observations into "bins" which represent intervals of values.

If all intervals are of the same width, the heights of the bars can be observed frequencies $freq(a < X < b)$ or relative frequencies $freq(a < X < b)/n$, for $a < x < b$.

Histogram

If the intervals are not of the same width, it gets a little trickier: we want the area to be proportional to the relative frequencies. This keeps the "same" height as when we break up into smaller bins, so the "skyline" looks roughly the same.

Thus we set

height $= \frac{\text{relative frequency}}{\text{interval width}}$

Relative frequency correspond to the areas of the "bars".

Thus

height $= \hat{f}(x) = \frac{freq(a<X<b)/n}{b-a}$, for $a < x < b$.

UFC tree diameter

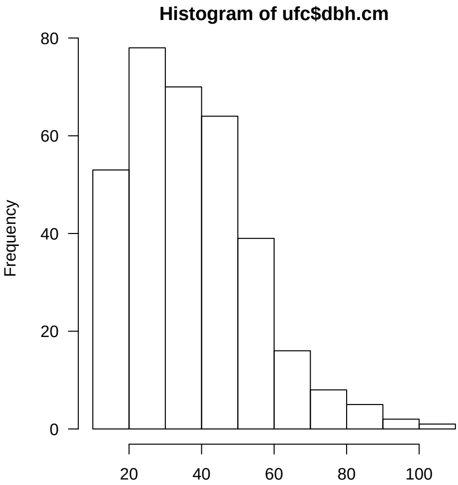


Figure: The number of trees of each diameter class from the ufc data.

Smoothed density

Another approach to represent the sampling distribution is to treat each point individually, but instead of placing a vertical bar (point mass) at each point, we place a small "kernel".

When we add up the kernels for all points, the result looks like a probability density fuction.

The smoothness of the density depends on the width of the kernel. A popular choice of kernel is the standard Gaussian function:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2}.$$

Example density plot- UFC tree diameter

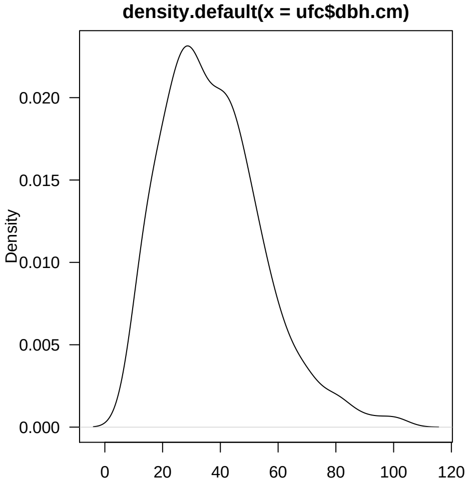


Figure: The number of trees of each diameter class from the ufc data.

Example density plot- UFC tree diameter

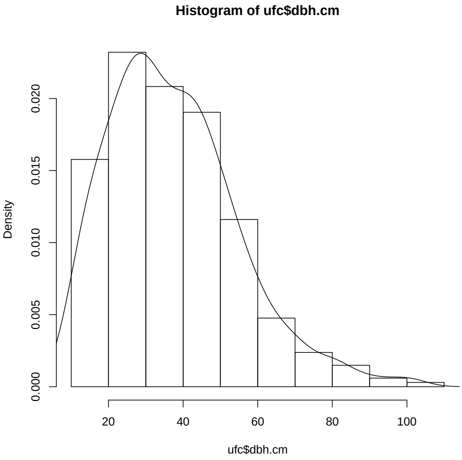
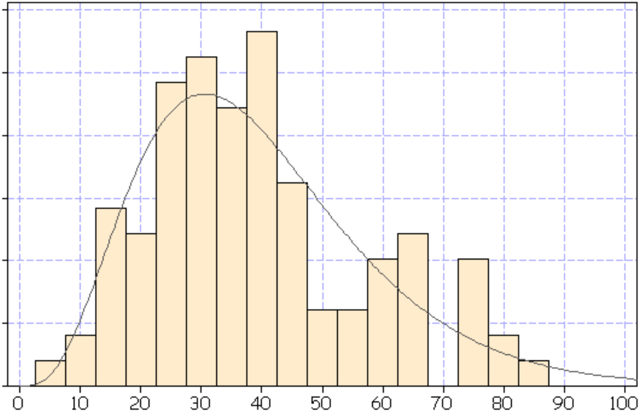


Figure: The number of trees of each diameter class from the ufc data.

Example - smoother density



Cumulative frequency

For numerical data, the cumulative relative frequency function is defined as the relative frequency of observations less than or equal to the number x:

$$\hat{F}(x) = \frac{1}{n} freq(X \leqslant x) = \frac{\#(observations \leqslant x)}{\#(observations)}$$

This is the sample analogue of the cumulative distribution function (cdf), and so it is sometimes called the sample cdf, or the empirical cdf.

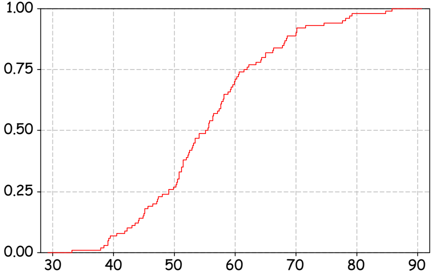
The cumulative frequency function is a step function. However, with more data it will get closer and closer to a continuous function.

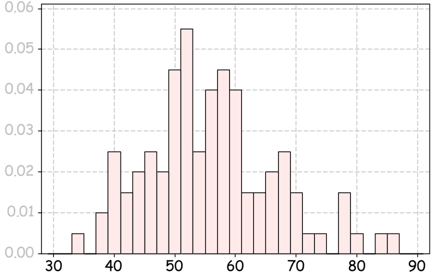
Sample quantiles

Sample quantiles are easy to find from the cumulative frequency.

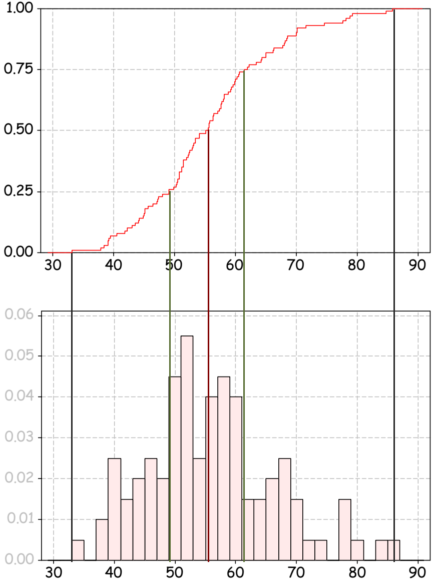
Remember that $\hat{c}_q$ is the number with a proportion q of the sample less than it.

It follows that $\hat{F}(\hat{c}_q) \approx q$:
the sample quantile function is the inverse of $\hat{F}$.





	N	Mean	StDev	Min	Q1	Med	Q3	Max
x	100	55.9	10.8	33.2	49.0	55.4	61.9	85.8

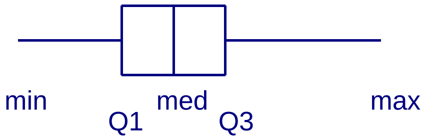


	N	Mean	StDev	Min	Q1	Med	Q3	Max
x	100	55.9	10.8	33.2	49.0	55.4	61.9	85.8

Box plots - construction & five-number summary

- ***boxplot***

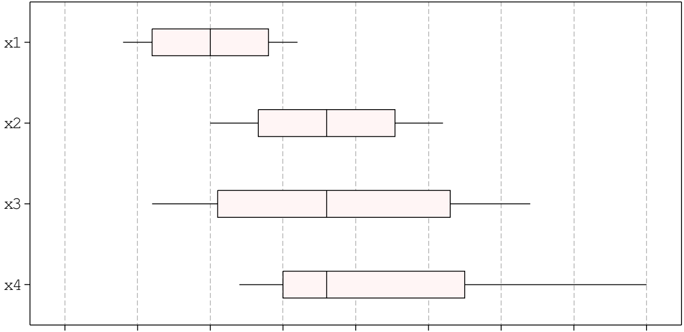
The boxplot is a graphical representation of the "five-number summary": (min, Q1, med, Q3, max).



The boxplot gives an immediate impression of not only the location and spread of the data, but also of the symmetry (or otherwise) of the distribution.

Box plots and skewness

Boxplot indicates location, spread and skewness.



Compared with the top boxplot, the second has greater location measure; the third has a greater spread measure.

The first three are symmetrical, but the bottom boxplot shows positive skewness, i.e. a longer tail at the positive end.

Box plots

One problem with this representation is that one or two outlying data values could give a misleading impression of the spread of the distribution.

For this reason, we limit the length of the “whiskers" (the lines at either end of the box) to $1.5\mathrm{IQR}$, i.e., 1.5 times the interquartile range.

The line extends to the most extreme data value within these limits, i.e. ($\mathrm{Q1}-1.5\,\mathrm{IQR}$ for the lower end, and $\mathrm{Q3}+1.5\,\mathrm{IQR}$ at the upper end). (These are sometimes called the *‘inner fences’*).

Any data value outside this interval is indicated separately.

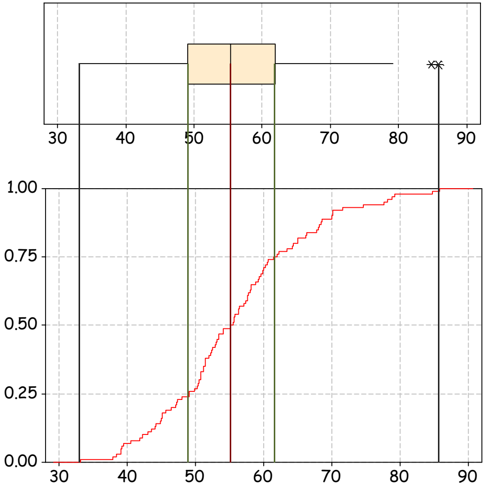
Some boxplots also define ‘outer fences’ $(\mathrm{Q1}-3\,\mathrm{IQR}, \mathrm{Q3}+3\,\mathrm{IQR})$, *and label points outside these limits as “extreme outliers”.*

Box plots

Extreme values are often indicated separately on a boxplot.

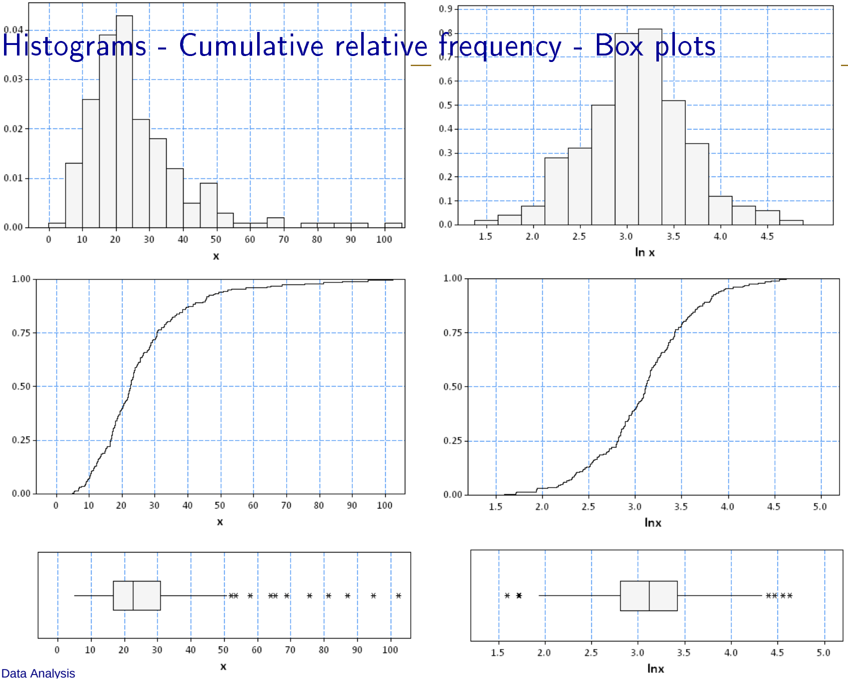
It is common to label these outlying data values by individual name or case number or some other identification. There may be some explanation of their oddity — in any case, the outlying data values are often of interest.

Box plots



	N	Mean	StDev	Min	Q1	Med	Q3	Max
x	100	55.9	10.8	33.2	49.0	55.4	61.9	85.8

Histograms - Cumulative relative frequency - Box plots



Communicate Results - Example UFC tree counts

The tree counts of the four species are different.

```
> tapply(ufc$dbh.cm, ufc$species, length)
 DF GF WC WL
 57 118 139 22
```

The average diameters of the four species are different.

```
> tapply(ufc$dbh.cm, ufc$species, mean)
 DF GF WC WL
 39.90526 35.21186 38.84460 33.72727
```

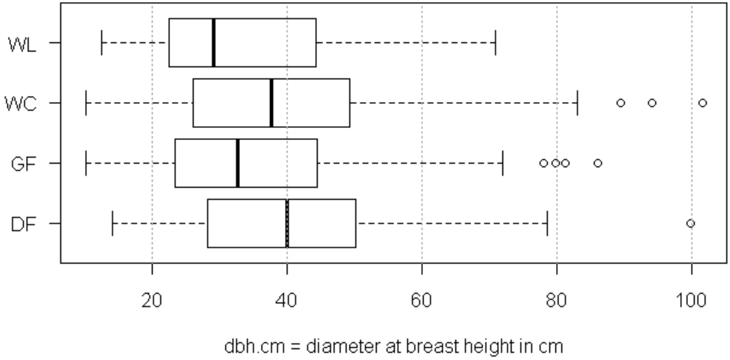
Communicate Results: Boxplot tree diameter by species

Visual display of the five-number summary. Useful for summarising a large data set into a few values which show its location, shape and spread.

- boxplots good for 1 categorical and 1 numeric variable
- shows outliers
- robust to wild values, i.e. outliers
- width of box is proportional to the number of species
- symmetric distributions?
- length of box is Q3-Q1 ...

Communicate Results: Tree diameter by species

```
> boxplot(dbh.cm ~ species, data=ufc, horizontal=T, las=1,
+   xlab="dbh.cm = diameter at breast height in cm")
> grid(ny=NA,nx=NULL,col="darkgray")
```



Useful (ray-approved) options

```
options(scipen=999999, digits=4)
```

penalises scientific notation out of existence

restricts number of significant figures in output to 4.

In plot functions:

`las=1`	ensures that all tick marks are horizontal
`grid(col="darkgray")`	overprints a standard grid in "darkgray" colour

also:

`grid=T`	can sometimes be used as part of the plot function This is equivalent to using `grid()` after the plot command
`grid(ny=NA, nx=NULL)`	gridlines on x-axis, not on y-axis

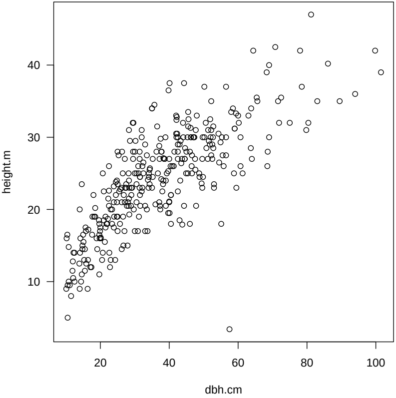
Scatterplots

- We may also wish to use graphical methods to represent the way two numeric variables relate.
- *Scatterplots* are used to graphically present the relationship between pairs of variables.

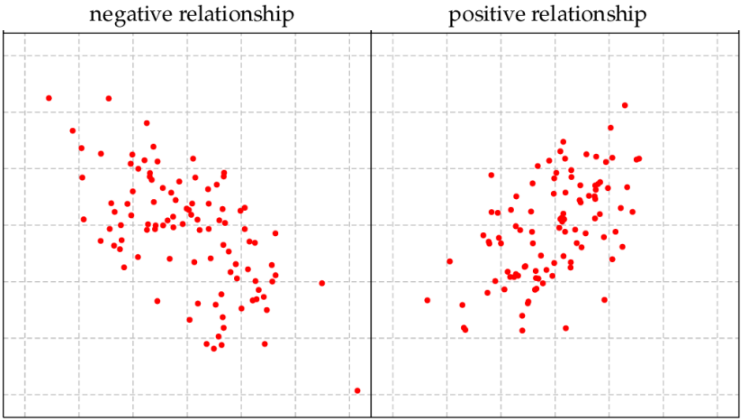
Scatterplots

The dbh of trees of each species is plotted against their height

```
> plot(height.m ~ dbh.cm, data = ufc)
```



Scatterplots - relationships



If $(-)$ve relationship then large/small x to small/large y occur together. If $(+)$ve relationship then small/small x and large/large y occur together.

Correlation - Linear relationships

We measure the strength of the linear relationship with the correlation coefficient r.

It is a number in the interval $[-1, 1]$ which reflects the strength of the linear relationship between two variables.

The larger the magnitude, the stronger the relationship.

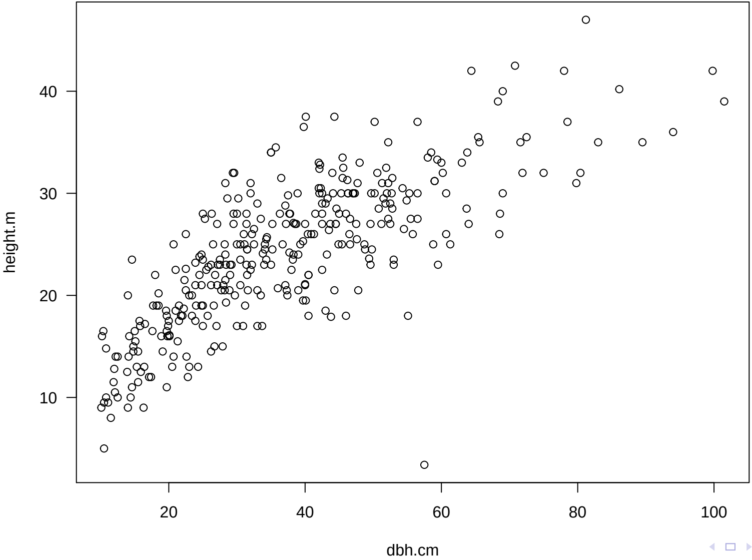
The sign reflects the type of relationship (positive or negative).

An r of ± 1 means the two variables are directly linearly related: $y = a + bx$.

We will see this again in linear models.

Correlation: example tree diameter by height

What is the correlation?



Correlation

```
> cor(ufc$dbh.cm, ufc$height.m)
[1] 0.7699552

> cor.test(ufc$dbh.cm,ufc$height.m)
        Pearson's product-moment correlation
data:  ufc$dbh.cm and ufc$height.m
t = 22.0522, df = 334, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7224794 0.8102042
sample estimates:
      cor
0.7699552
```

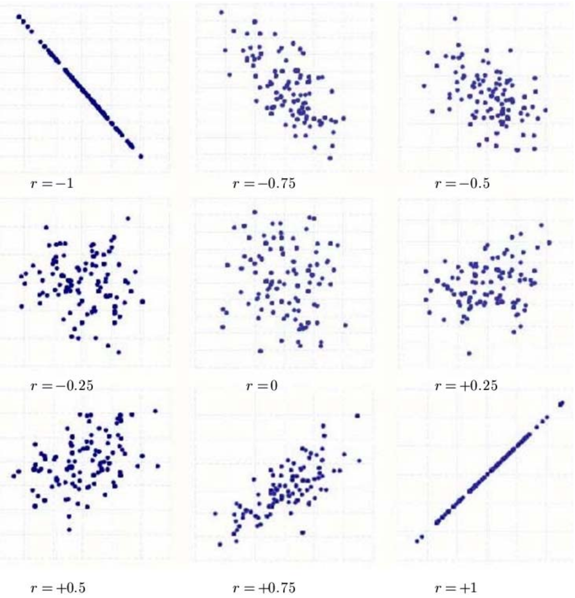
\# Test $H_0 : \rho = 0$

```
> options(scipen=999999,digits=4)
```

\# scipen: don't use sci notation eg 1.2 e^2,
\# digits: restricts sig figures

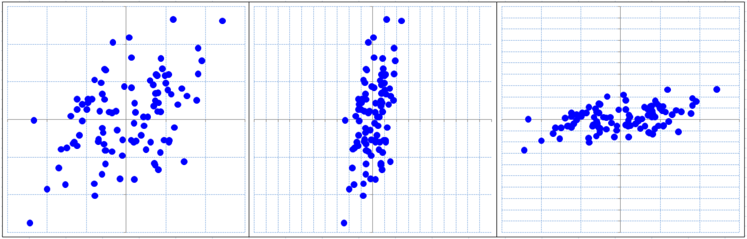
```
> cor.test(ufc$dbh.cm, ufc$height.m)
        Pearson's product-moment correlation
data:  ufc$dbh.cm and ufc$height.m
t = 22.05, df = 334, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7225 0.8102
sample estimates:
 cor
0.77
```

Correlation



Correlation: scale change, beware!

Scales should be similar, otherwise plot is distorted, giving a false impression of the (linear) relationship.



Data in the above plots are identical, $r = 0.45$. Only the scale has changed!

Sample (Pearson) Correlation: strength of linear association

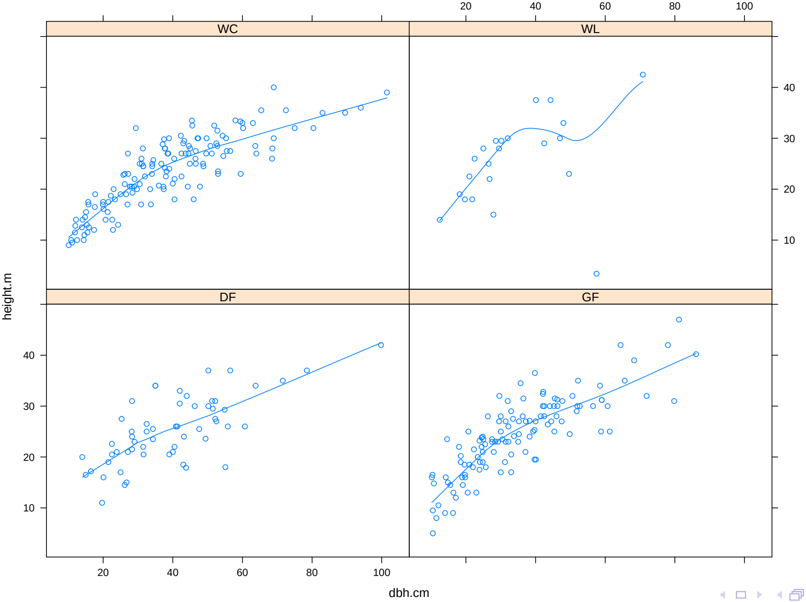
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}}$$

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{s_x s_y} = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y})$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} x_{si} y_{si}$$

- x_{si} & y_{si} are standardised scores, s_x & s_y are std dev., s_{xy} =cov(x,y)
- larger scores x_{si} & y_{si} contribute most to r
- r is not affected by location and scale but
- r is affected by outliers (just like $\bar{x}$ and s)
- $-1 \leqslant r \leqslant 1$

Communicate Results: Lattice plots



Communicate Results: Tree diameter by height for different species

Lattice plots

- lattice plots good for 2 numeric variables and 2 (or more) categorical variables
- positive/negative relationship between diameter and height?
- is there an effect of diameter on height?
- interaction of diameter with height and with species?

Summary

Number of variables		
Numerical	Categorical	Types of plot to consider
1	0	histogram, boxplot, cumulative frequency, bar chart
0	1	bar chart, dot chart
2	0	scatterplot
1	1	parallel boxplots, parallel histograms
0	2	clustered barchart

These may be extended to 3 variables eg lattice plots, clustered boxplots/barchats etc

Further Reading

- R
  (see resources on LMS)
  - icebreakeR chapters 1–6;
  - An Introduction to R (Kuhnert & Venables, CSIRO);
  - Using R (Maindonald);
  - Venables and Ripley (2002) Modern Applied Statistics with S.
- Descriptive Statistics
  - Introductory texts, e.g. Triola & Triola "Biostatistics for the Biological and Health Sciences"; **Utts and Heckard "Mind on Statistics"**; Moore and McCabe "Introduction to the Practice of Statistics".
- Graphics
  - Anything by Edward Tufte, e.g. "The Visual Display of Quantitative information".