Introduction to Statistics Part 1

Natasha Stanbridge, Tiangang Cui, Jaslene Lin

Table of contents

1 Graphical Summary

Throughout this chapter, we will explore the types of data that we can have and how we can visualise it. We will learn about initial data analysis, identifying variables, graphical summaries (barplot, histogram & more) and logical operators to explore a data story.

1.1 Data story: What causes Australian road fatalities?

We are going to investigate data from the Australian Bureau of Statistics (ABS) (last updated Nov 2023). This dataset contains information about all road crash fatalities in Australia from 1989 to 2023. You can check this ABC Animation out for a visualisation of this data set.

Variables

A variable is a characteristic which changes from person to person in a study. Interviewers for the survey use a battery of questions: How old are you? How many people are there in your family? What is your family's total income? Are you married? Do you have a job? The corresponding variables would be: age, family size, family income, marital status, and employment status.

The following are the variables contained in the road fatalities data set. This includes information about the time and date of the crash, the vehicles involved, and the individuals involved. You can find more information about the variables in the data set by looking at the data dictionary.

```
## Read in data
data = read.csv("data/2023fatalities.csv", header = TRUE)
## Names of Variables
names(data)
```

```
 [1] "Crash.ID"                      "State"
 [3] "Month"                         "Year"
 [5] "Dayweek"                       "Time"
 [7] "Crash.Type"                    "Bus.Involvement"
 [9] "Heavy.Rigid.Truck.Involvement" "Articulated.Truck.Involvement"
[11] "Speed.Limit"                   "Road.User"
[13] "Gender"                        "Age"
[15] "National.Remoteness.Areas"     "SA4.Name.2021"
[17] "National.LGA.Name.2021"        "National.Road.Type"
[19] "Christmas.Period"              "Easter.Period"
[21] "Age.Group"                     "Day.of.week"
[23] "Time.of.day"                   "X"
```

Possible research questions:

- How many road fatalities have there been so far this year, and how does it compare to last year?
- What is the most common day and time for a crash?
- Does gender affect the type of road fatality?
- What is the chance that a motorcycle rider is involved in a road fatality?
- How many people wear seatbelts

1.2 Initial data analysis

Data is **information** about the set of **subjects** being studied (like road fatalities). Most commonly, data refers to the **sample**, not the population.

Initial data analysis is a first general look at the data, without formally answering the research questions.

- IDA helps you to see whether the data can answer your research questions.
- IDA may lead to new research questions.
- IDA can
  - identify the data's main qualities;
  - suggest the population from which a sample derives.

What's involved in IDA?

Initial Data Analysis commonly involves:

- data background: checking the quality and integrity of the data
- data structure: what information has been collected?
- data wrangling: scraping, cleaning, tidying, reshaping, splitting, combining
- data summaries: graphical and numerical

Here we focus on **structure** & **graphical summaries** for qualitative and quantitative data.

1.3 Structure of the data

There are many different types of data, in different formats.

For example, Survey data, Spreadsheet type data, and MRI image data

1.3.1 Variables

A **variable** measures or describes some attribute of the subjects.

- Data with p variables is said to have **dimension** p.

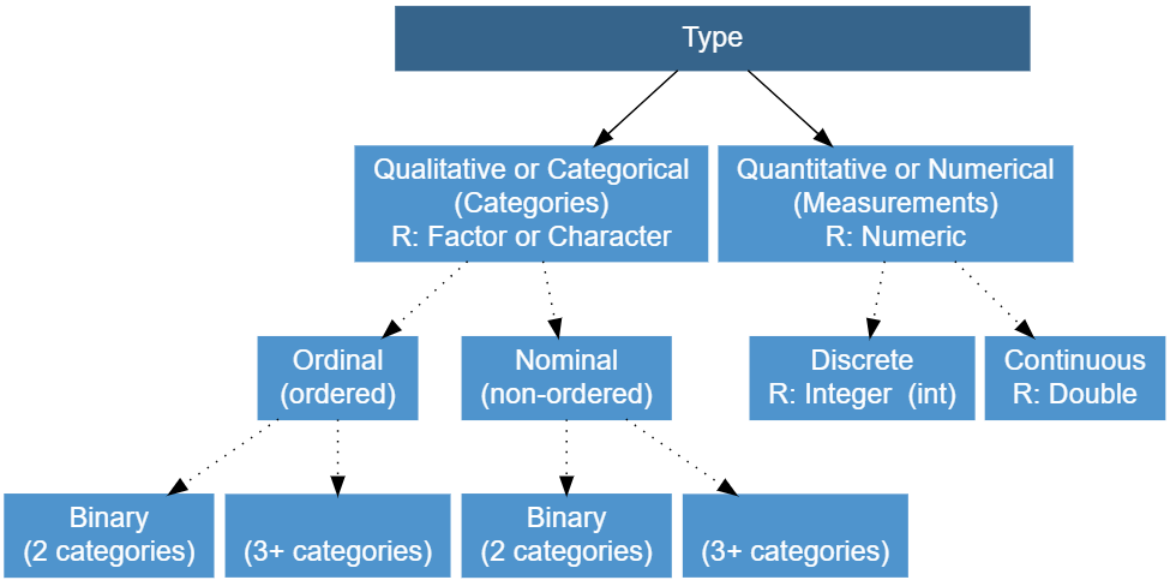
How many variables does the road fatality data have?

- The road fatality data has dimension $p = 23$, as the CrashID serves as an anonymous identifier.

```
## Size of Data (rows and columns)
dim(data)
```

```
[1] 55360    24
```

The following diagram shows the classification of variables



Often the classification of a variable depends on the context.

Example

Classify the variable `Age` in the Road Fatality Data.

- Technically Age is a quantitative, continuous variable, but here the ages have been reported as discrete 'integer' (by rounding down to the nearest year).
- Age may be also be recorded as a qualitative variable in a survey, as respondents may be more willing to give their age category. (e.g. 18-24)
- However, it is more precise to record quantitative data if possible.

1.3.2 Change variable types in R

The following function shows the types of variables in a data frame. Note that some of the variable types may not be correct.

```
## Structure of Data (tells us how each variable is stored in R)
str(data, vec.len = 2)
```

```
'data.frame':   55360 obs. of  24 variables:
 $ Crash.ID                     : int  20237008 20234009 20233087 20233149 20233190 ...
 $ State                        : chr  "NT" "SA" ...
 $ Month                        : int  10 10 10 10 10 ...
 $ Year                         : int  2023 2023 2023 2023 2023 ...
 $ Dayweek                      : chr  "Friday" "Saturday" ...
 $ Time                         : chr  "" "03:00" ...
 $ Crash.Type                   : chr  "Single" "Single" ...
 $ Bus.Involvement              : chr  "No" "No" ...
 $ Heavy.Rigid.Truck.Involvement: chr  "No" "No" ...
 $ Articulated.Truck.Involvement: chr  "No" "No" ...
 $ Speed.Limit                  : chr  "-9" "100" ...
 $ Road.User                    : chr  "Driver" "Driver" ...
 $ Gender                       : chr  "Female" "Male" ...
 $ Age                          : int  24 22 19 37 35 ...
 $ National.Remoteness.Areas    : chr  "" "Outer Regional Australia" ...
 $ SA4.Name.2021                : chr  "" "Barossa - Yorke - Mid North" ...
 $ National.LGA.Name.2021       : chr  "" "Yorke Peninsula" ...
 $ National.Road.Type           : chr  "" "Local Road" ...
 $ Christmas.Period             : chr  "No" "No" ...
 $ Easter.Period                : chr  "No" "No" ...
 $ Age.Group                    : chr  "17_to_25" "17_to_25" ...
 $ Day.of.week                  : chr  "Weekend" "Weekend" ...
 $ Time.of.day                  : chr  "Night" "Night" ...
 $ X                            : logi  NA NA NA ...
```

Importance of storing variables correctly

We need to reclassify some of the variables within the data set to ensure they are the correct type. It is important that variables are stored correctly to prevent errors when creating graphical or numerical summaries.

Qualitative variables should be recorded as factors (if they have an order) or characters (if they don't). Quantitative variables should be recorded as numeric or integers (depending on if they are integers or could have decimal points).

```
## Change qualitative variables stored as 'numeric' to 'factors'
data$Crash.ID = as.factor(data$Crash.ID)
data$Month = as.factor(data$Month)
```

```
## New structure of Data Display the first 5 variables using list.len=5
str(data, list.len = 5)
```

```
'data.frame':   55360 obs. of  24 variables:
 $ Crash.ID                    : Factor w/ 49903 levels "19891001","19891002",..: 49880 4964
 $ State                       : chr  "NT" "SA" "Qld" "Qld" ...
 $ Month                       : Factor w/ 12 levels "1","2","3","4",..: 10 10 10 10 10 10 1
 $ Year                        : int  2023 2023 2023 2023 2023 2023 2023 2023 2023 2023 ...
 $ Dayweek                     : chr  "Friday" "Saturday" "Saturday" "Sunday" ...
  [list output truncated]
```

```
## Change quantitative variables stored as 'characters' to 'numeric'
data$Speed.Limit = as.numeric(data$Speed.Limit)
```

```
## New structure of Data Display variables 11 to 15
str(data[c(11, 12, 13, 14, 15)])
```

```
'data.frame':   55360 obs. of  5 variables:
 $ Speed.Limit              : num  -9 100 80 60 100 70 60 80 60 60 ...
 $ Road.User                : chr  "Driver" "Driver" "Driver" "Passenger" ...
 $ Gender                   : chr  "Female" "Male" "Male" "Male" ...
 $ Age                      : int  24 22 19 37 35 32 29 51 39 33 ...
 $ National.Remoteness.Areas: chr  "" "Outer Regional Australia" "Inner Regional Australia" "
```

1.4 Graphical summaries

Once we've identified the variables, we can summarise the data, both graphically and numerically, in order to identify and highlight the main features of interest. We often start with graphical summaries because 'A (well-designed) picture is worth a thousand words.'

E.g. I didn't finish reading the "Lord of the Ring" books, but the movies are graphical summary the contents of the books. Yes, the specific details are omitted, but the movies told the same meaningful story in lesser time (11 hours vs 455,000 words.)

When choosing an appropriate graphical summary, there are several factors to consider:

- The critical question is: 'What plot is the more informative?' or 'What plot will best highlight features of the data?' or 'What plot will best guide the next analysis?'.
- To some extent we use trial and error. We try some standard forms and see what is revealed about the data. One graphical summary can suggest another, and often a combination will highlight different features of the data
- In practice we use computer packages like R to construct summaries.
- However, it is important to understand how to construct graphical summaries 'by hand', so that you understand how to interpret computer output and for your final exam.

We will explore two common types of graphical summary, bar plots and histograms. These are both important ways of summarising quantitative data so that it is easy to understand.

1.4.1 Barplot (qualitative data)

A bar plot is a common type of plot that shows the relationship between a numeric and a categorical variable, with the size of each bar representing the numeric value. In the above bar plot, we plot the frequency of fatalities for each day of the week being a separate bar.

Question: What was the most common day of road fatality?

Step 1: Build a frequency table

```
## Select the DayWeek variable from the whole data frame
Dayweek = data$Dayweek
## Produce a frequency table of fatalities per day of the week
table(Dayweek)
```

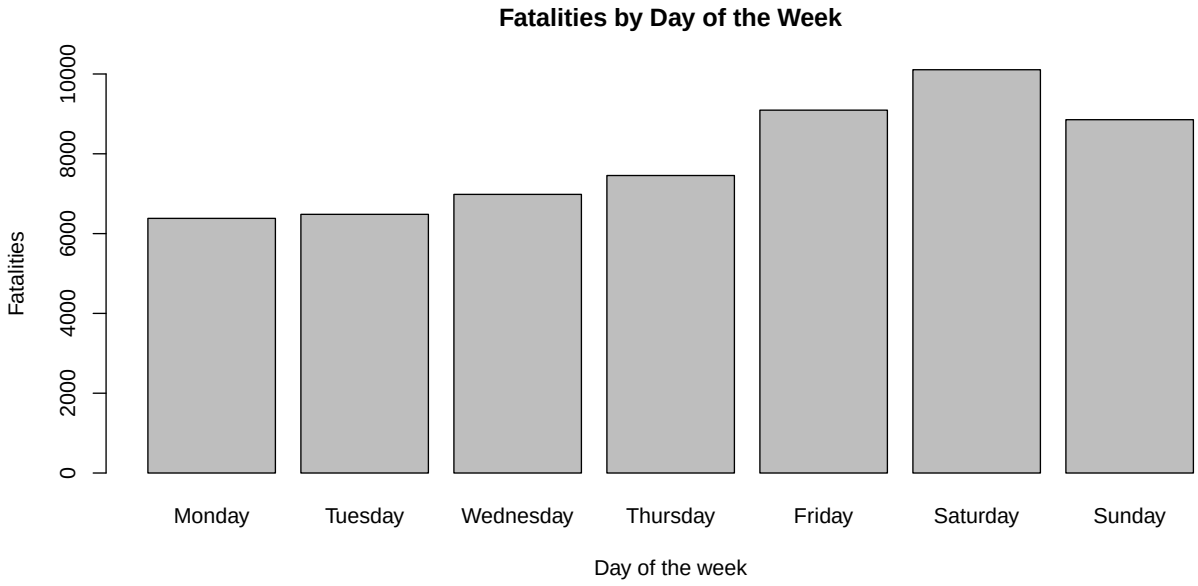
```
Dayweek
   Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
     9094      6382     10107      8855      7456      6483      6983
```

```
## Order days
ordered = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday",
 ↪  "Sunday")
Dayweek = factor(Dayweek, levels = ordered)
table(Dayweek)
```

```
Dayweek
   Monday   Tuesday Wednesday  Thursday    Friday  Saturday    Sunday
     6382      6483      6983      7456      9094     10107      8855
```

Step 2: Produce a barplot of the frequency table

```
## Produce a barplot
barplot(table(Dayweek), main = "Fatalities by Day of the Week",
        xlab = "Day of the week", ylab = "Fatalities")
```



Statistical Thinking

What was the most common day of road fatality?

- Saturday

Why might that be the case?

- More volume of cars on the road, or people driving faster?

What data would you need to check your hypotheses?

- Data on volume and speed of cars on the road each day.

Two-way frequency table

Things get more interesting when we consider 2 qualitative variables. Note: Here Gender refers to biological sex as it was historically recorded in this dataset. Read more.

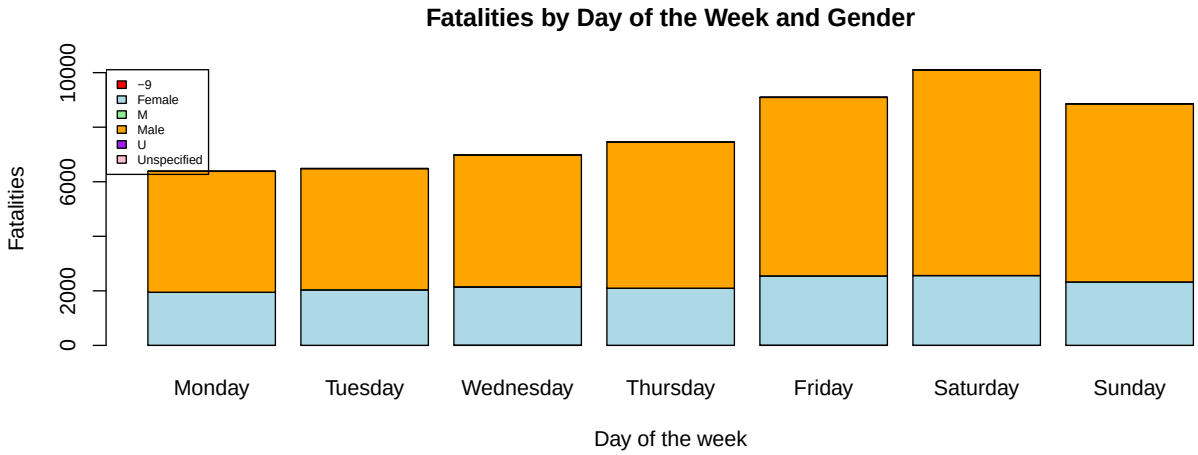
```
## Select Gender variable
Gender = data$Gender

## Produce a double frequency table (contingency table)
data1 = table(Gender, Dayweek)

## display the table
data1
```

```
             Dayweek
Gender        Monday Tuesday Wednesday Thursday Friday Saturday Sunday
  -9               3       2        12        3     10        6      2
  Female        1945    2034      2135     2094   2538     2555   2325
  M                0       0         1        0      0        0      0
  Male          4433    4447      4835     5359   6545     7541   6528
  U                1       0         0        0      1        4      0
  Unspecified      0       0         0        0      0        1      0
```

There are several ways of making barplots for the two-way frequency table.

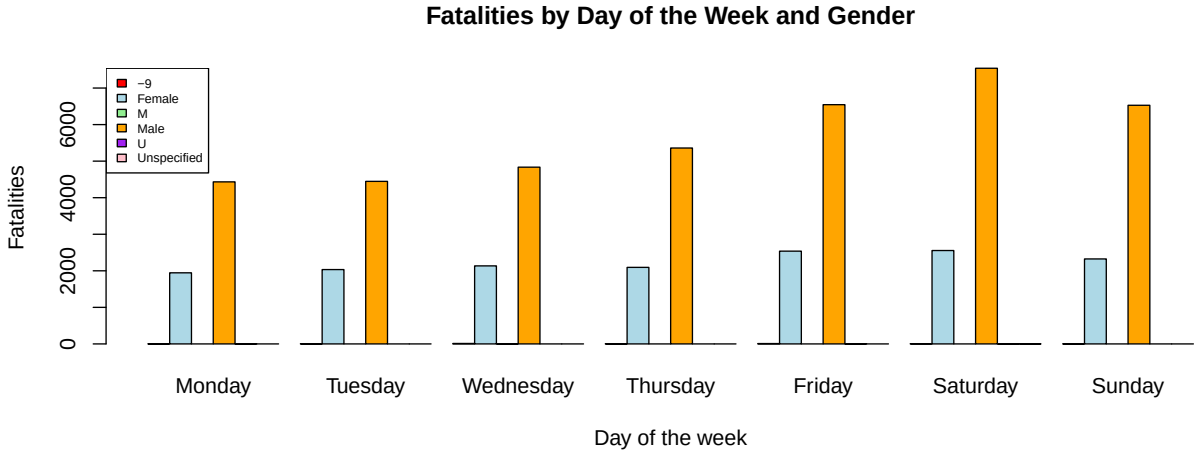
Stacked barplot

```
barplot(data1, main="Fatalities by Day of the Week and Gender",
        xlab="Day of the week", ylab = "Fatalities",
        col=c('red', "lightblue", "lightgreen", 'orange', "purple", "pink"))
## Add a legend
legend( "topleft", legend = rownames(data1), fill =
 ↪  c('red',"lightblue","lightgreen",'orange', "purple", "pink"), cex = 0.58)
```



Side-by-side barplot

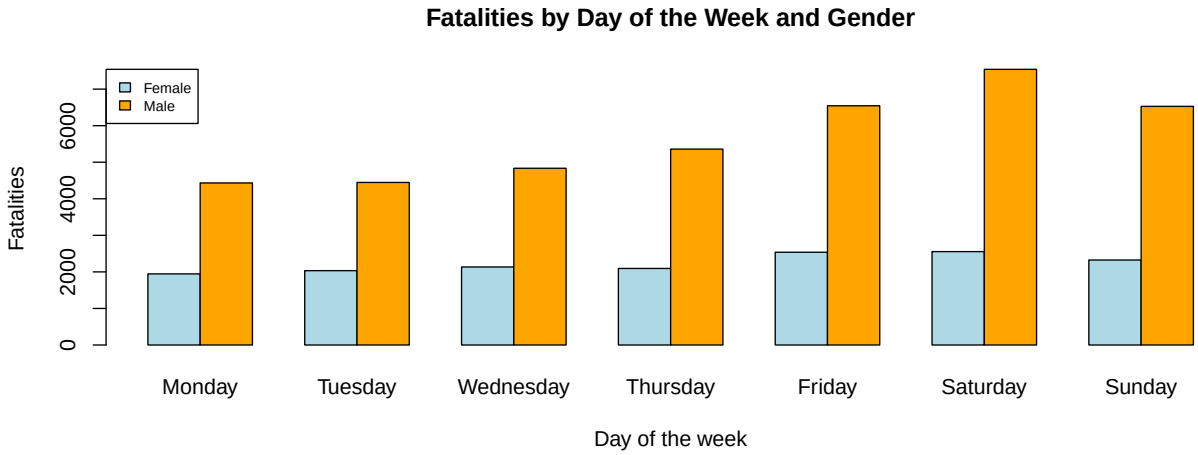
```
barplot(data1, main="Fatalities by Day of the Week and Gender",
        xlab="Day of the week", ylab = "Fatalities",
        col=c('red',"lightblue","lightgreen",'orange', "purple", "pink"),
        beside=TRUE)
legend( "topleft", legend = rownames(data1), cex = 0.58,
        fill = c('red',"lightblue","lightgreen",'orange', "purple", "pink"))
```



We can use the following to ignore '-9', 'M', 'U' and 'Unspecified' in the plot. The command `c(1,3,5,6)` creates a list of indices to be ignored, then adding a minus sign `-c(1,3,5,6)` to ignore those rows of the two-way frequency table.

```
barplot(data1[-c(1,3,5,6),],
        main="Fatalities by Day of the Week and Gender",
        xlab="Day of the week", ylab = "Fatalities",
        col=c("lightblue","orange"),
```

```
        legend = rownames(data1[-c(1,3,5,6),]),
        beside=TRUE, args.legend = list(x = "topleft", cex = 0.7))
```



Statistical Thinking

Are these plots telling us anything useful? How could they be misread?

- There seems to be a similar proportion of gender fatalities across each day.
- We could posit that men are more likely to be involved in fatal accidents than women. However, perhaps there are more men on the road than women. More data is needed.

1.4.2 Histogram

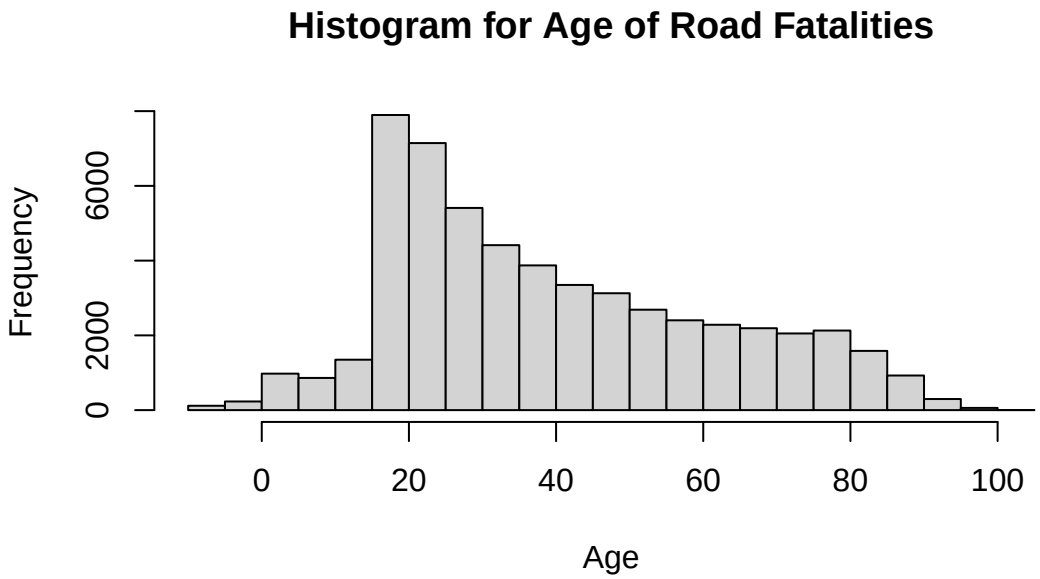
The frequency table can also be used to summarise a set of **quantitative** data, by collecting the data into **class intervals** (or 'bins'). A histogram highlights the frequency of data in one class interval compared to another.

A histogram is used to summarise data that is measured on an interval scale. It is often used to illustrate major features of the data's distribution in a convenient form. In a histogram, the height of a block represents crowding — percentage per horizontal unit.

Q: What were the most common age groups at which a road fatality occurred?

This is the default histogram generated by R for Age of Road Fatalities .

```
hist(data$Age, xlab = "Age", ylab = "Frequency", main = "Histogram for Age of Road
 ↪  Fatalities")
```



We can also provide user-defined class intervals and, more importantly, use the **density scale**.

Data cleaning

Before moving into density scale histogram, we noticed the 1st block start below 0, why?

- Data Dictionary: missing values are coded as '-9'.
- It is better to replace the "-9" by "NA".

```
## Replacing the '-9' entries
data$Age[data$Age == -9] = NA
```

User defined class intervals

```
## Select the variable Age
Age = data$Age

## Define end points for class intervals
breaks = c(0, 18, 25, 70, 101)

## Build frequency table
table(cut(Age, breaks, right = F))
```

```
  [0,18)  [18,25)  [25,70) [70,101)
    5631    11541    30566     7504
```

Q: What proportion of total road fatalities does each age group account for?

```
## Convert the frequency table into densities
table(cut(Age, breaks, right = F))/length(Age)
```

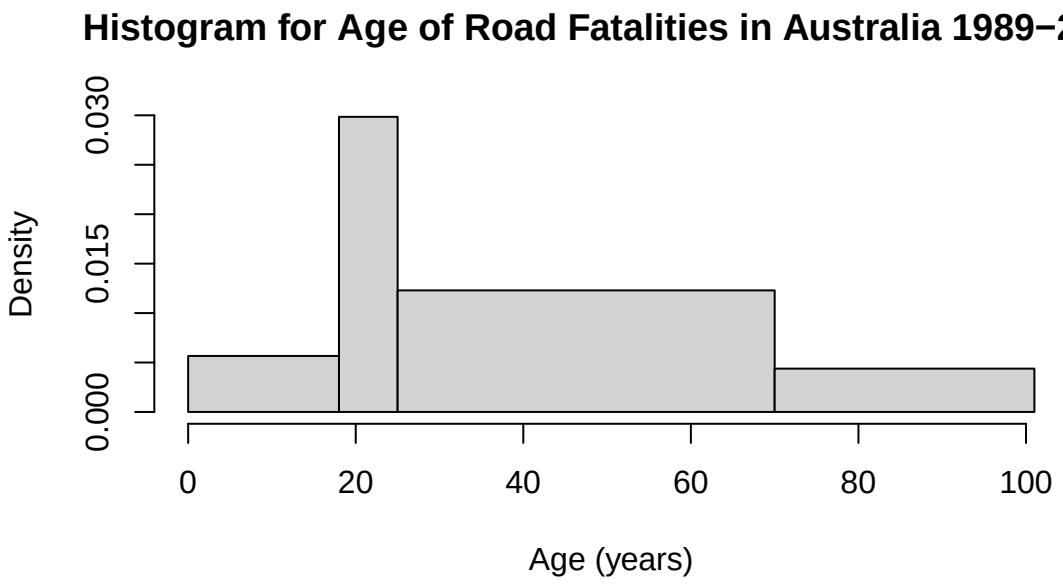
```
   [0,18)   [18,25)   [25,70)  [70,101)
0.1017160 0.2084718 0.5521315 0.1355491
```

Density scale histogram

With the density scale on the vertical axis, the areas of the blocks represent percentage. The area under the histogram over an interval equals the percentage of cases in that interval. The total area under the histogram is 100%.

The following summarises the proportions of each age group graphically. This makes it easy to understand which age groups account for a higher proportion of road fatalities.

```r
hist(Age, br = breaks, right = F, freq = F, xlab = "Age (years)", ylab = "Density",
    main = "Histogram for Age of Road Fatalities in Australia 1989-2023")
```



- The horizontal scale is divided into **class intervals** with potentially unequal sizes.
- The **area of each block** represents the **proportion** of subjects in that particular class interval.

How can we interpret this histogram?

Why is the histogram tallest above [18,25)?

- It is not the interval with the largest number of fatalities, but it has the higheset density because the width of the interval is small.

Which age group has overall most fatalities?

- [25,70), as it has the largest area, but not the highest density, as the size of the interval is large

Details of density-scale histograms

1. The area of the whole histogram on the density scale is one (or, in percentage, 100%).

The area and height of each block are calculated in the following ways.

$$\text{area (proportion) of each block} = \frac{\text{number of subjects in the class interval}}{\text{total number of subjects}}$$

$$\text{height (density) of each block} = \frac{\text{proportion of the block}}{\text{length of the class interval}}$$

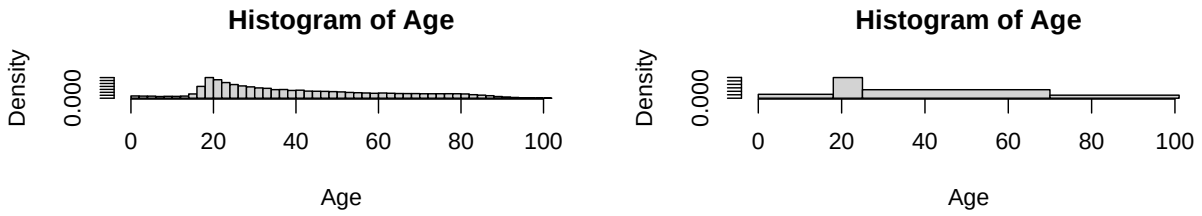
2. For continuous (quantitative) data, we need an **endpoint convention** for data points that fall on the border of two class intervals.

- If an interval contains the left endpoint but excludes the right endpoint, then an 18 year old would be counted in [18,25) not [0,18).
- We call this left-closed and right-open.
- Similarly, we can also have left-open and right-closed, e.g., (18,25].

3. Number of class intervals

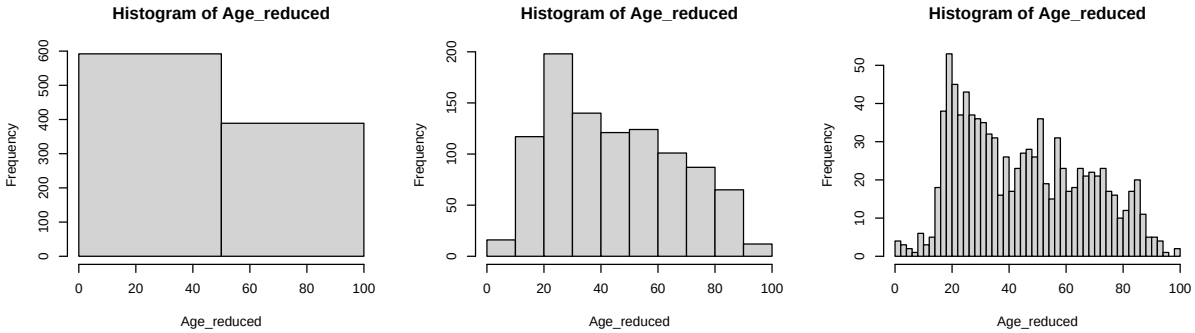
Think about how many class intervals (or the sizes of class intervals) you want to have.

```
par(mfrow = c(1, 2))  # This puts the graphic output in 1 row with 2 columns
breaks = seq(0, 102, 2)  # This sets the breaks to be every 2 numbers
hist(Age, br = breaks, freq = F, right = F, xlab = "Age", ylab = "Density")
breaks = c(0, 18, 25, 70, 101)
hist(Age, br = breaks, freq = F, right = F, xlab = "Age", ylab = "Density")
```



Using too many or too few class intervals can hide the true pattern in the data. As a rule of thumb, use between 10-15 class intervals and make sure you consider the size of the data.

```
Age_reduced = Age[1:1000]  # only look at subset of data
par(mfrow = c(1, 3))
hist(Age_reduced, breaks = 3)
hist(Age_reduced, breaks = 10)
hist(Age_reduced, breaks = 50)
```



Produce a histogram by hand

Class intervals	Number of subjects in the interval	%	Height of block
[0,18)	5747	10.4	0.0058
[18,25)	11541	20.8	0.0298
[25,70)	30566	55.2	0.0123
[70,101)	7504	13.6	0.0044
	55360	100	

Summary in R

```
#Read in data
data = read.csv("data/2023Fatalities.csv",header=T)
## Cleaning
data$Age[data$Age==-9] = NA
## Choose a variable
Age = data$Age
## Choose the class intervals
breaks=c(0,18,25,70,101)
## Produce a histogram
hist(Age,br=breaks,freq=F,right=F, xlab="Age (in years)", ylab="Density",
     main="Histogram for Age of Road Fatalities in Australia 1989-2020")
```

Note

- `freq=F` produces the histogram on the density scale.
- `right=F` makes the intervals right-open.

1.5 Other graphical summaries

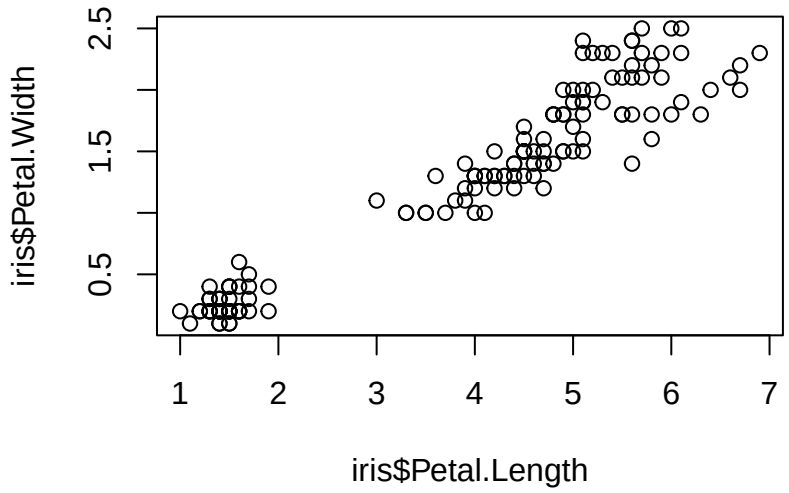
1.5.1 Scatter plot

Scatter plots examine the relationship between two quantitative variables.

Scatter plots can easily show a reader what happens to one variable when another variable is changed, and can test the relation between the two variables. If there is a strong association between two variables, then knowing one helps a lot in predicting the other. But when there is a weak association, information about one variable does not help much in guessing the other.

The Iris dataset contains measurements of 150 iris flowers from three species: Iris setosa, Iris versicolor, and Iris virginica. The dataset includes the following measurements for each flower: sepal length, sepal width, petal length, and petal width.

```
plot(iris$Petal.Length, iris$Petal.Width)
```



We can see that there is a relationship between petal length and petal width for these iris flowers. As the petal length increases, we can see that the petal width also increases.

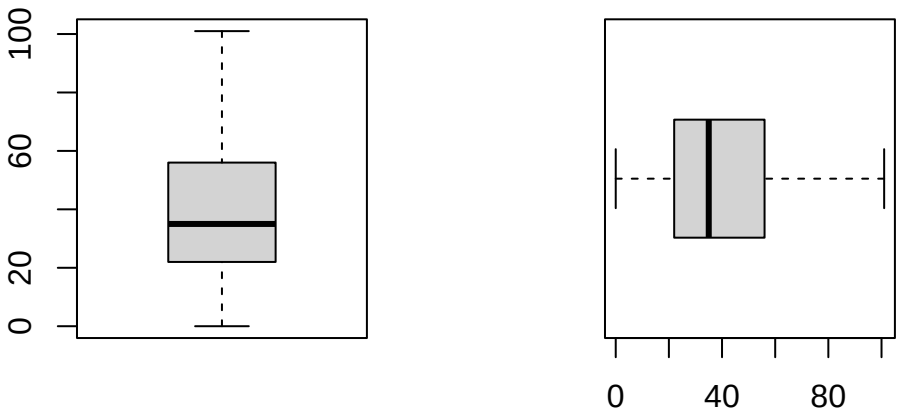
1.5.2 Boxplot

Boxplots are a graphical representation of the numerical five number summary of a data set. It summarises centre, spread and outliers through plotting the median ('middle' data point), the middle 50% of the data in a box, the expected maximum and minimum in the whiskers, and any outliers.

We will consider how to draw the box plot when we learn about the interquartile range (IQR) in numerical summaries.

Boxplot in R

```
par(mfrow = c(1, 2))
boxplot(Age)  # We can plot the box plot vertically
boxplot(Age, horizontal = T)  # Or horizontally
```



What does the simple boxplot reveal about the age of fatalities?

- The box plot is fairly symmetric with no outliers.
- There does not seem to be any extreme ages for fatalities.

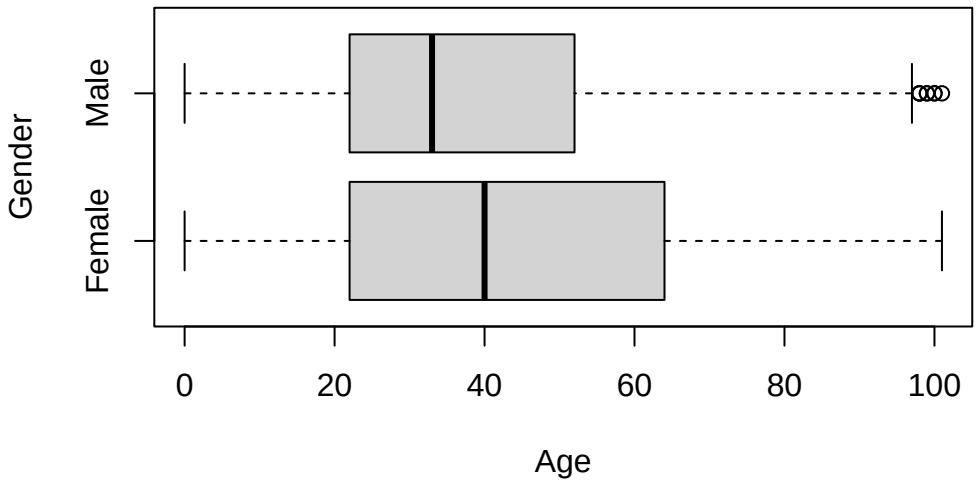
Comparative box plots

A comparative boxplot splits up a quantitative variable by a qualitative variable.

The purpose of a box plot is to show distributions of numeric data values, especially when you want to compare them between multiple groups. They are built to provide high-level information at a glance, overing general information about the data's symmetry, skew, variance, and outliers.

Comparing Age by Gender

```
## Filtering the gender variable to only include 'female' and 'male' values
data$Gender = factor(data$Gender, levels = c("Female", "Male"))

## Selecting the gender variable
Gender = data$Gender

boxplot(Age ~ Gender, horizontal = T)
```



The median ('middle') age is fairly similar but higher for women than for men. We also observe some outliers in men.

1.6 Logical operators

Basics of logical operators

The basic logical values in R are `TRUE` (or just `T`) and `FALSE` (or just `F`). These come up very often in R when you are checking an object, or comparing an object to a value or another object, as in $x > 5$ or $x > y$.

Some commonly used logical operators:

`>`	greater than	`>=`	greater than or equal to
`<`	less than	`<=`	less than or equal to
`==`	equal to	`!=`	not equal to

Many of these are exactly what you would expect (like `>`) but remember to use **two** equal signs rather than one when assessing equality (`==` not `=`). If you use just one equal sign, R thinks you are trying to assign a value to an object.

```
x = 5  # This assigns the value 5 to x
x == 5  # This checks to see if x equals 5
```

```
[1] TRUE
```

Combining logical conditions

You can combine logical conditions using `&` (and), `|` (or), and `!` (not).

The evaluation of `&` (and): both conditions need to be `TRUE` to have a `TRUE`

`&`	True	False
True	True	False
False	False	False

Examples:

```
x = 10
is.numeric(x) & x < 20  # True and True
```

```
[1] TRUE
```

```
x = 10
is.numeric(x) & x < 0  # True and False
```

```
[1] FALSE
```

The evaluation of | (or): need to have at least one of the conditions to be TRUE to give a TRUE evaluation

|	True	False
True	True	True
False	True	False

Examples:

```
x = 10
!is.numeric(x) | x < 20  # False and True
```

```
[1] TRUE
```

```
x = 10
is.character(x) & x < 0  # False and False
```

```
[1] FALSE
```

Data selection and counting

You can apply logical operators elementwise to vectors or matrices. This can be particularly useful for data selection and counting.

```
x = c(-1, 0, 1)
## Check each element of x against the condition (elementwise)
x <= 0
```

```
[1]  TRUE  TRUE FALSE
```

`TRUE` and `FALSE` in R also correspond to integers `1` (`TRUE`) and `0` (`FALSE`). This way, they are also useful for counting. For example, how many data points of x in the following case are less than 5?

```
x = 1:10
## Check each element of x against the condition (elementwise)
x <= 5
```

```
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
```

```
sum(x <= 5)  # sum over those TRUEs (data points <= 5)
```

```
[1] 5
```

Example on data selection

```
## creating a data frame
rating = 1:5
animal = c("koala", "hedgehog", "sloth", "panda", "alligator")
country = c("Australia", "Italy", "Peru", "China", "USA")
avg_sleep_hours = c(21, 18, 17, 10, 15)
sleepers = data.frame(rating, animal, country, avg_sleep_hours)
str(sleepers)
```

```
'data.frame':   5 obs. of  4 variables:
 $ rating         : int  1 2 3 4 5
 $ animal         : chr  "koala" "hedgehog" "sloth" "panda" ...
 $ country        : chr  "Australia" "Italy" "Peru" "China" ...
 $ avg_sleep_hours: num  21 18 17 10 15
```

Q1: Filter `sleepers` data with rating less than 3

```
sleepers1 = sleepers[sleepers$rating < 3, ]
dim(sleepers1)
```

```
[1] 2 4
```

```
str(sleepers1)
```

```
'data.frame':   2 obs. of  4 variables:
 $ rating         : int  1 2
 $ animal         : chr  "koala" "hedgehog"
 $ country        : chr  "Australia" "Italy"
 $ avg_sleep_hours: num  21 18
```

Q2: Filter `sleepers` data with rating more than 3 and sleeping hour more than 15

```
sleepers2 = sleepers[sleepers$rating > 3 & sleepers$avg_sleep_hours > 15, ]
dim(sleepers2)
```

```
[1] 0 4
```

```
str(sleepers2)
```

```
'data.frame':   0 obs. of  4 variables:
 $ rating         : int
 $ animal         : chr
 $ country        : chr
 $ avg_sleep_hours: num
```

2 Numerical Summary

In this chapter, we will learn how to explore the main features of data through numerical summaries. We will learn about the sample mean, sample median, robustness, standard deviation and interquartile range to explore a data story. Finally, we will learn how to write functions in R.

2.1 Data story: How much does a property in Newtown cost?

We are going to investigate data taken from domain.com.au. This dataset contains information about all properties sold in Newtown (NSW 2042) between April-June 2017. The variable `Sold` has prices in $1000s.

```
data <- read.csv("data/NewtownJune2017.csv", header = T)
head(data, n = 2)
```

```
                  Property  Type    Agent Bedrooms Bathrooms Carspots Sold
1 19 Watkin Street Newtown House RayWhite        4         1         1 1975
2  30 Pearl Street Newtown House RayWhite        2         1         0 1250
     Date
1 23/6/17
2 23/6/17
```

```
dim(data)
```

```
[1] 56  8
```

```
str(data)
```

```
'data.frame':   56 obs. of  8 variables:
 $ Property : chr  "19 Watkin Street Newtown" "30 Pearl Street Newtown" "26 John Street Newto
 $ Type     : chr  "House" "House" "House" "Apartment" ...
 $ Agent    : chr  "RayWhite" "RayWhite" "Belle" "RayWhite" ...
 $ Bedrooms : int  4 2 2 1 1 5 1 1 1 3 ...
 $ Bathrooms: int  1 1 1 1 1 1 1 1 1 2 ...
 $ Carspots : int  1 0 0 1 1 1 0 1 1 0 ...
 $ Sold     : int  1975 1250 1280 780 650 2100 675 740 625 1950 ...
 $ Date     : chr  "23/6/17" "23/6/17" "17/6/17" "17/6/17" ...
```

> **Note**
>
> - How can we analyse this data set?
> - Can we learn enough useful information just from graphical summaries?

2.2 Basics of Numerical summaries

A numerical summary reduces all the data to one simple number ("statistic"). - This loses a lot of information. - However it allows easy communication and comparisons.

Major features that we can summarise numerically are: - Maximum - Minimum
- Centre [sample mean, median] - Spread [standard deviation, range, interquartile range]

> **Note**
>
> Which summaries might be useful for talking about Newtown house prices?

- It depends!
- Reporting the centre without the spread can be misleading!

Useful notation for data

- Observations of a single variable of size n can be represented by

$$x_1, x_2, \dots, x_n$$

- The ranked observations (ordered from smallest to largest) are

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

  such that $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

- The sum of the observations are

$$\sum_{i=1}^{n} x_i$$

2.3 Sample mean

The sample mean is the average of the data.

$$\text{sample mean} = \frac{\text{sum of data}}{\text{size of data}}$$

or

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

Note that the sample mean involves **all** of the data.

- The sample mean of all the properties sold in Newtown is:

```
mean(data$Sold)
```

```
[1] 1407.143
```

- Focusing specifically on houses with 4 bedrooms (large), the sample mean is:

```
mean(data$Sold[data$Type == "House" & data$Bedrooms == "4"])
```

```
[1] 2198.857
```

2.3.1 Deviation from the mean

Given a data point x_i, its deviation from the sample mean $\bar{x}$ is

$$D_i = x_i - \bar{x}$$

For example,

- 19 Watkin St sold for \$1950 (thousands).
  - This gives a gap of (\$1950-\$1407.143) = \$542.857 (thousands)
  - \$542.857 (thousands) **above** the sample mean
- 30 Pearl St sold for \$1250 (thousands).
  - This gives a gap of (\$1250-\$1407.143) = -\$157.143 (thousands)
  - \$157.143 (thousands) **below** the sample mean

2.3.2 Sample mean as a balancing point

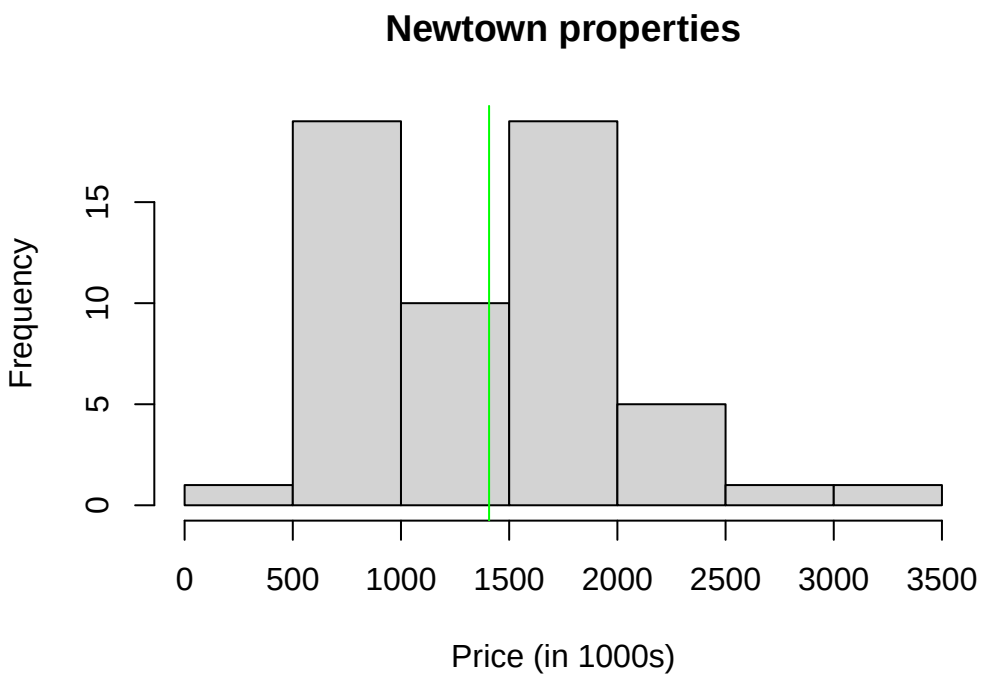
The sample mean is the point at which the data is **balanced** in the sense the sum of the **absolute deviations** for values to the left of the mean is the same as the sum of absolute deviations to the right of the mean.

$$\sum_{x_i<\bar{x}} |x_i - \bar{x}| = \sum_{x_i>\bar{x}} |x_i - \bar{x}|$$

2.3.3 Sample mean on the histogram

However, sample mean may **not** be balancing point of a histogram (as it does not report values), the area to the left of the mean may not be the same as the area to the right of the mean.

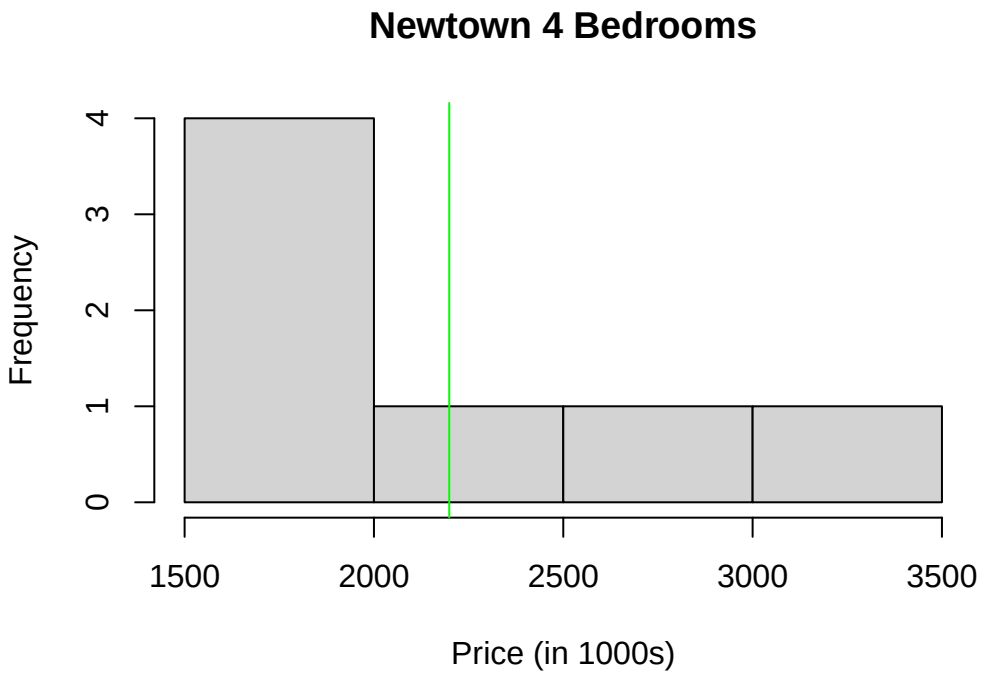
```
hist(data$Sold, main = "Newtown properties", xlab = "Price (in 1000s)")
abline(v = mean(data$Sold), col = "green")
```



Skewed data

When the data is skewed, this effect is more significant.

```
hist(data$Sold[data$Type=="House" & data$Bedrooms=="4"], main="Newtown 4 Bedrooms",
 ↪  xlab="Price (in 1000s)")
abline(v=mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="green")
```



The mean is affected by outliers that do not influence the median, this means that when the data is skewed the mean and median's values have a greater difference.

In the example, we can see that there are a some very expensive houses which influence the mean a lot due to their magnitude.

2.4 Sample median

The sample median $\tilde{x}$ is the **middle data point**, when the observations are ordered from smallest to largest.

- For an odd sized number of observations:

$$\text{sample median} = \text{the unique middle point} = x_{(\frac{n+1}{2})}$$

- For an even sized number of observations:

$$\text{sample median} = \text{average of the 2 middle points} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

2.4.1 Ordering observations

The ordered observations are:

```
sort(data$Sold)
```

```
 [1]  370  625  645  650  675  692  720  740  740  755  770  780  812  860  861
[16]  920  935  955  955  999 1100 1240 1250 1280 1309 1315 1370 1375 1400 1460
[31] 1553 1575 1590 1600 1600 1600 1605 1662 1701 1710 1750 1780 1790 1806 1850
[46] 1940 1950 1975 2000 2100 2200 2235 2300 2410 2810 3150
```

```
length(data$Sold)
```

```
[1] 56
```

As we have n = 56 observations (even), the sample median is found between the $(\frac{n}{2}) = 28$th and $(\frac{n}{2} + 1) = 29$th prices, or $\frac{1375+1400}{2} = 1387.5$.

- The sample median of all the properties sold in Newtown is:

```
median(data$Sold)
```

```
[1] 1387.5
```

- Focusing specifically on houses with 4 bedrooms (large), the sample median is:

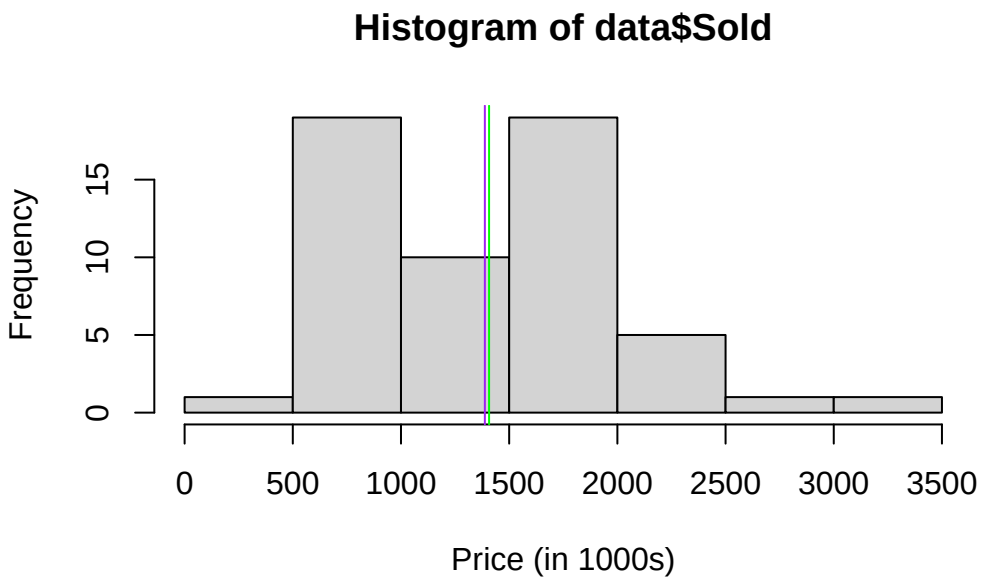
```
median(data$Sold[data$Type == "House" & data$Bedrooms == "4"])
```

```
[1] 1975
```

2.4.2 Sample median on the histogram

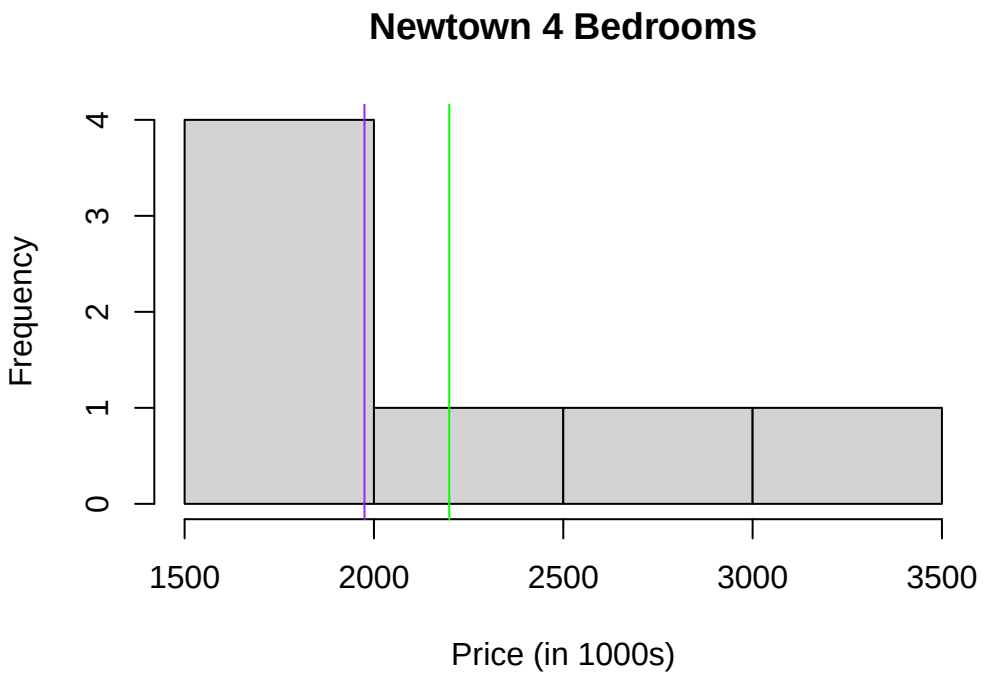
- The sample median is the **half way point** on the histogram - i.e., 50% of the houses sold are below and above $1.3875 million.

```
hist(data$Sold, xlab = "Price (in 1000s)")
abline(v = mean(data$Sold), col = "green")  # create a green line for the mean
abline(v = median(data$Sold), col = "purple")  # create a purple line for the median
```



Histogram for 4 Bedroom Houses

```
hist(data$Sold[data$Type == "House" & data$Bedrooms == "4"], main = "Newtown 4
 ↪  Bedrooms",
    xlab = "Price (in 1000s)")
abline(v = mean(data$Sold[data$Type == "House" & data$Bedrooms == "4"]), col =
 ↪  "green")
abline(v = median(data$Sold[data$Type == "House" & data$Bedrooms == "4"]), col =
 ↪  "purple")
```



Comparison between sample mean and median

If you had to choose between reporting the sample mean or sample median for Newtown properties, which would you choose and why?

- For the full property portfolio, the sample mean and the sample median are fairly similar.
- For the 4 bedroom houses, the sample mean is higher than the sample median because it is being "pulled up" by some very expensive houses.
- For the average buyer, the sample median would be more useful as an indication of the sort of price needed to get into the market.
- For any agent selling houses in the area, the sample mean might be more useful in order to predict their average commissions!
- In practice, we can report both!

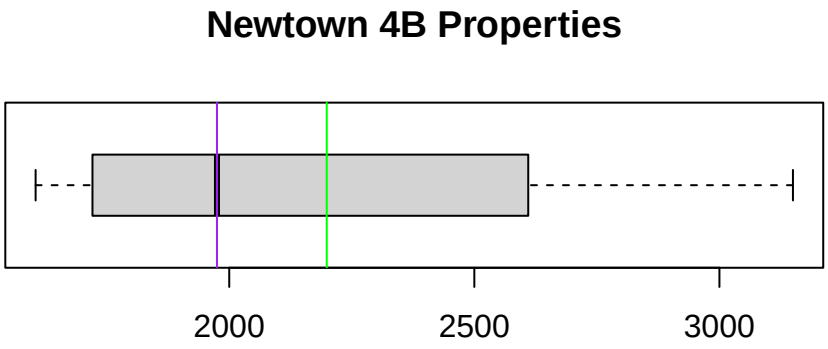
2.4.3 Sample mean and median on the boxplot

The sample median is the centre line on the boxplot.

```
boxplot(data$Sold, main = "Newtown properties", horizontal=T)
abline(v=mean(data$Sold),col="green")
abline(v=median(data$Sold),col="purple")
```

Boxplot for 4 Bedroom Houses

```
boxplot(data$Sold[data$Type=="House" & data$Bedrooms=="4"], main = "Newtown 4B
↪  Properties", horizontal=T)
abline(v=mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="green")
abline(v=median(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="purple")
```



2.5 Robustness and comparisons

2.5.1 Robustness

The sample median is said to be **robust** and is a good summary for skewed data as it is not affected by **outliers** (extreme data values).

A robust statistic is resistant to errors in the dataset. Robust statistical analyses can produce valid results even when the ideal conditions do not exist with real-world data.

Example

Recently a heritage building was sold for 13 million in Newtown.

> **Note**
>
> How would the sample mean and sample median change if it was added to the data?

- The sample mean would be a lot higher.
- The sample median would be a bit higher: it moves from the average of the 28th and 29th points to the 29th point.

```
data2 = c(data$Sold, 13000)
sort(data2)
```

```
 [1]   370   625   645   650   675   692   720   740   740   755   770   780
[13]   812   860   861   920   935   955   955   999  1100  1240  1250  1280
[25]  1309  1315  1370  1375  1400  1460  1553  1575  1590  1600  1600  1600
[37]  1605  1662  1701  1710  1750  1780  1790  1806  1850  1940  1950  1975
[49]  2000  2100  2200  2235  2300  2410  2810  3150 13000
```

```
mean(data2)
```

```
[1] 1610.526
```

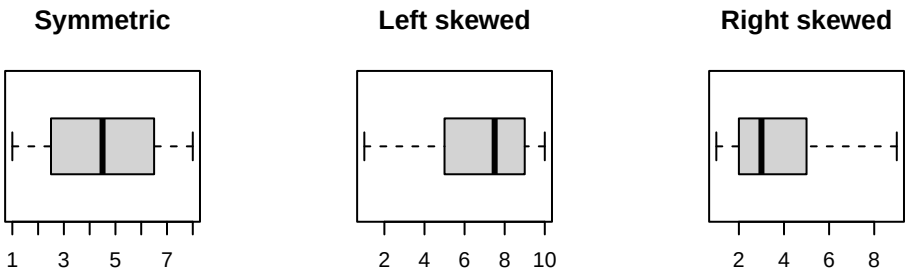
```
median(data2)
```

```
[1] 1400
```

2.5.2 Skewness

The difference between the sample mean and the sample median can be an indication of the **shape** of the data.

- For symmetric data, the sample mean and sample median are the same: $\bar{x} = \tilde{x}$.
- For left skewed data (the most frequent data are concentrated on the right, with a left tail), the sample mean is smaller than the sample median: $\bar{x} < \tilde{x}$.
- For right skewed data (the most frequent data are concentrated on the left, with a right tail), the sample mean is larger than the sample median: $\bar{x} > \tilde{x}$.



2.5.3 Which is optimal for describing centre?

- Both have strengths and weaknesses depending on the nature of the data.
- Sometimes neither gives a sensible sense of location, for example if the data is **bimodal**.
- As the **sample median is robust**, it is preferable for data which is skewed or has many outliers, like Sydney house prices.
- The **sample mean** is helpful for data which is **basically symmetric**, with not too many outliers, and for theoretical analysis.

2.6 Standard deviation

For each property sold, we could calculate the **deviation** (or the gap) from the sample mean, $D_i = x_i - \bar{x}$, between the house and the sample mean \$1407 (thousands).

Property	Sold	Gap	Conclusion
19 Watkin Street	\$1950 (thousands)	1950-1407=543	More than half a million dollars more expensive than the average house price
30 Pearl St	\$1250 (thousands)	1250-1407=-157	Cheaper than the average house price

```
## Deviations of Newtown data
gaps = data$Sold - mean(data$Sold)
gaps
```

```
 [1]   567.857143  -157.142857  -127.142857  -627.142857  -757.142857
 [6]   692.857143  -732.142857  -667.142857  -782.142857   542.857143
[11]   -32.142857   167.857143  -408.142857  -452.142857  -547.142857
[16]   197.857143   182.857143  -167.142857 -1037.142857   532.857143
[21]  -687.142857  -452.142857  -487.142857   442.857143   192.857143
[26]  -652.142857    -7.142857   145.857143  1402.857143   192.857143
[31]   792.857143   372.857143   398.857143   293.857143   -98.142857
[36]  -307.142857  -472.142857  -762.142857    52.857143   -37.142857
[41]  -715.142857  1742.857143  1002.857143  -637.142857   254.857143
[46]   827.857143   592.857143   382.857143   342.857143   302.857143
[51]   192.857143  -546.142857  -667.142857   -92.142857   892.857143
[56]  -595.142857
```

```
max(gaps)
```

```
[1] 1742.857
```

How do we **summarise** all the deviations into **1 number** ("spread")?

2.6.1 1st attempt: The mean gap

We could calculate the **average** of the deviations.

$$\text{mean deviation} = \text{sample mean(data - sample mean(data))}$$

```
round(mean(gaps))
```

```
[1] 0
```

> **Note**
>
> What's the problem?

Note: It will always be 0.

- From the definition, the mean deviation must be 0, as the mean is the **balancing point** of the deviations.
- The mean deviation is

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

2.6.2 Better option: Standard deviation

Standard deviation is a measure of spread that is based on the average.

First define the **root mean square** (RMS).

- The RMS measures the **average** of a set of numbers, regardless of the signs.
- The steps are: *Square* the numbers, then *Mean* the result, then *Root* the result.

$$\text{RMS(numbers)} = \sqrt{\text{sample mean}\left(\text{numbers}^2\right)}$$

- So effectively, the *Square* and *Root* operations "reverse" each other.
- RMS retain the same unit as the unit of the sample mean.
- Applying RMS to the deviations, we get

$$\text{RMS of deviations} = \sqrt{\text{sample mean}\left(\text{deviations}^2\right)} = \sqrt{\frac{\sum_{i=1}^{n} D_i^2}{n}}$$

- To avoid the cancellation of the deviations, another possible method is to consider the average of the absolute values of the deviations:

$$\text{mean absolute deviation (MAD)} = \frac{\sum_{i=1}^{n} |D_i|}{n}.$$

  However, MAD is much harder to analyse.

- The RMS measures the **average** of the data, regardless of the signs.
- The steps are: *Square* the data, then *Mean* the result, then *Root* the result.

$$\text{RMS} = \sqrt{\text{Mean of squared data}} \approx \text{Mean of absolute data}$$

- Formally, $\text{RMS} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \approx \frac{\sum_{i=1}^{n} |x_i|}{n}$

```
## RMS
sqrt(mean(data$Sold^2))
```

```
[1] 1527.269
```

```
## RMS approximation
mean(abs(data$Sold))
```

```
[1] 1407.143
```

2.6.3 Standard deviation in terms of RMS

Population Standard deviation

- The standard deviation measures the **spread** of the data.
- A small standard deviation value indicates data points are clustered tightly around the mean, while a large standard deviation value indicates data points are more spread out.

$$\text{SD}_{pop} = \text{RMS of (deviations from the mean)}$$

- Formally, $\text{SD}_{pop} = \sqrt{\text{Mean of (deviations from the mean)}^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}}$

```
sqrt(mean(gaps^2))
```

```
[1] 593.7166
```

An equivalent calculation is:

$$\text{s} = \sqrt{\text{Mean of squared data} - \text{Mean of data}^2}$$

or

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

```
sqrt(mean(data$Sold^2) - mean(data$Sold)^2)
```

```
[1] 593.7166
```

2.6.4 Standard deviation in R?

It is easy to calculate in R.

```
sd(data$Sold)
```

```
[1] 599.0897
```

> **Note**
>
> But why is this slightly different?

2.6.5 Adjusting the standard deviation

- There are **two** different formulas for the standard deviation, depending on whether the data is the **population** or a **sample**.
- The `sd` command in R always gives the **sample** version, as we most commonly have samples.
- Formally, $\text{SD}_{pop} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} D_i^2}$ and $\text{SD}_{sample} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} D_i^2}$, where $D_i = x_i - \bar{x}$ is the deviation.

```
sd(data$Sold) * sqrt(55/56)  # adjust by sqrt((n-1)/n), it calculates the population
↪  SD.
```

```
[1] 593.7166
```

```r
gaps = data$Sold - mean(data$Sold)  # calculate the gaps
sqrt(mean(gaps^2))  # calculates the population SD.
```

```
[1] 593.7166
```

Why does the sample SD use the adjustment $\sqrt{(n-1)/n}$?

- It is an **unbiased estimator** of the standard deviation (beyond the scope of this unit, will be covered in Year 2)
- Estimating the sample mean uses all of the n data points. The sum (or the mean) of n deviations is zero
$$\sum_{i=1}^{n} D_i = \sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$
This means, given the first $n-1$ deviations, we know the n-th deviation, because
$$\left(\sum_{i=1}^{n-1} D_i \right) + D_n = 0 \quad \Longrightarrow \quad D_n = -\sum_{i=1}^{n-1} D_i.$$
Hence, there are only $n-1$ effective pieces of information in the deviations.

2.6.6 Summary: population and sample

Type of data	Formula	In R
Population mean	Average $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$	`mean(data)`
Sample mean	Average $\overline{(x)} = \frac{1}{n}\sum_{i=1}^{n} x_i$	`mean(data)`
Population standard deviation	RMS of gaps from the mean $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(gaps)^2}$	`sd(data)*sqrt((n-1)/n))`
Sample standard deviation	Adjusted RMS of gaps from the mean $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(gaps)^2}$	`sd(data)`

- The population standard deviation is always smaller than a sample standard deviation, (SD*{pop}* *SD*{sample}), why? Extra variability due to sampling.
- Note for large sample sizes, the difference becomes negligible.

How to tell the difference?

- It can be tricky to work out whether your data is a population or sample!

- Look at the information about the data story and the research questions.
  - If we are just interested in the Newtown property prices during April-June 2017, then the `data` is the whole **population**.
  - If we are studying the property prices during April-June 2017 as a window into more general property prices (for the rest of the year or for the Inner West area) , then the `data` could be considered a **sample**.
- Population SD and sample SD get closer with increasing sample size n.

2.6.7 Variance

The squared standard deviation is called the **variance**. Similar to the sample SD and the population SD, there are two versions of the variance

$$\text{Var}_{sample} = \text{SD}^2_{sample} \quad \text{and} \quad \text{Var}_{pop} = \text{SD}^2_{pop}.$$

- For summarising spread, we often prefer SD, as it has the same unit as the data points and the mean. - In some situations, e.g., dealing with random variables (Part III) and understanding the property of sample mean, using the variance can be much simpler.

2.6.8 Standard units ("Z score")

Standard units of a data point = how many standard deviations is it below or above the mean

$$\text{standard units} = \frac{\text{data point - mean}}{\text{SD}}$$

This means that

$$\text{data point} = \text{mean} + \text{SD} \times \text{standard units}$$

It gives the relative location of a data point in the data set. It also have other benefits in data modelling (see later lectures).

We can compare data points usinsg standard units.

Property	Sold	Standard units	Conclusion
19 Watkin Street	\$1950 (thousands)	$\frac{1950-1407}{599} = 0.91$	Almost 1 SD higher than the average house price
30 Pearl St	\$1250 (thousands)	$\frac{1250-1407}{599} = -0.26$	0.26 SDs cheaper than the average house price

So 19 Watkin is a more unusual purchase than 30 Pearl St, relative to the mean.

2.7 Interquartile range

2.7.1 Quantile, quartile, percentile

The set of q-**quantiles** divides the **ordered** data into q equal size sets (in terms of percentage of data).

Percentile is 100-quantile, so the set of percentiles divides the data into 100 equal parts.

The set of **quartiles** divides the data into four quarters.

2.7.2 Interquartile range (IQR)

The IQR is another measure of spread by **ordering** the data.

$$\text{IQR} = \text{range of the middle 50\% of the data}$$

More formally, $\text{IQR} = Q_3 - Q_1$, where

- Q_1 is the 25-th percentile (1st quartile) and Q_3 is the 75-th percentile (3rd quartile).
- The median is the 50-th percentile, or 2nd quartile $\tilde{x} = Q_2$.
- p-th percentile: there are p% of **ordered** data below the value of p-th percentile.

```
summary(data$Sold)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  370.0   860.8  1387.5  1407.1  1782.5  3150.0 
```

```
summary(data$Sold)[5] - summary(data$Sold)[2]  # one way to calculate IQR
```

```
3rd Qu. 
 921.75 
```

```
IQR(data$Sold)  # use the built-in function
```

```
[1] 921.75
```

So the range of the middle 50% of properties sold is almost a million dollars!

2.7.3 Reporting

- Like the median, the IQR is **robust**, so it's suitable as a summary of spread for skewed data.
- We report in pairs: (mean,SD) or (median,IQR).

2.7.4 IQR on the boxplot and outliers

- The IQR is the length of the box in the boxplot. It represents the span of the middle 50% of the houses sold.
- The **lower** and **upper thresholds** (expected minimum and maximum) are a distance of 1.5IQR from the 1st and 3rd quartiles (by Tukey's convention).

$$LT = Q_1 - 1.5IQR$$

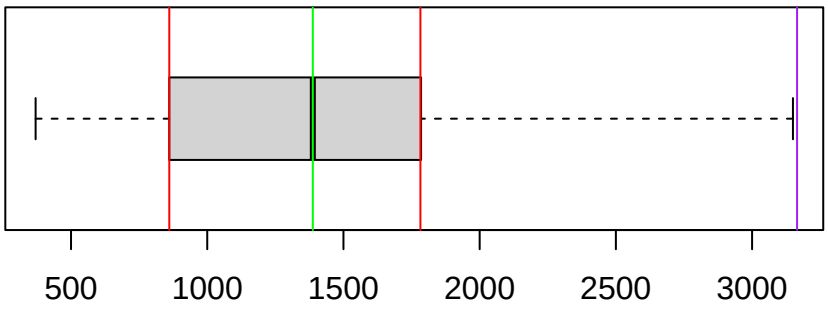
and

$$UT = Q_3 + 1.5IQR$$

- Data outside these thresholds is considered an **outlier** ("extreme reading").

2.7.5 Lower and Upper Thresholds on the Boxplot

```
boxplot(data$Sold, horizontal = T)
iqr = quantile(data$Sold)[4] - quantile(data$Sold)[2]
abline(v = median(data$Sold), col = "green")
abline(v = quantile(data$Sold)[2], col = "red")
abline(v = quantile(data$Sold)[4], col = "red")
abline(v = quantile(data$Sold)[2] - 1.5 * iqr, col = "purple")
abline(v = quantile(data$Sold)[4] + 1.5 * iqr, col = "purple")
```

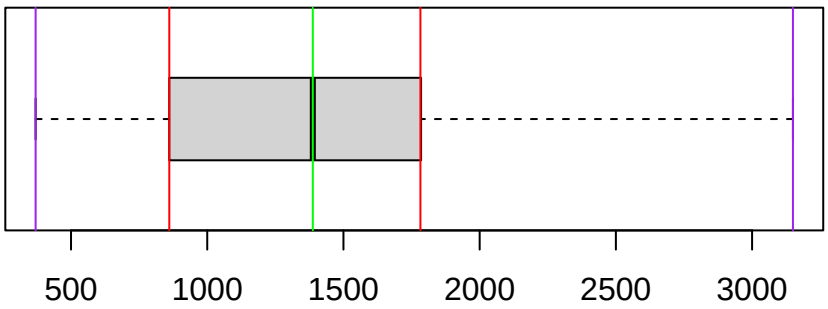


> **Note**
>
> Note the lower threshold is not shown...why?

- The lower threshold should be the same difference from Q1 as the upper threshold is from Q3.
- Therefore the lower threshold's value would be below 0, which is not possible in our data set as we are dealing with house prices.

2.7.6 Thresholds can be outside of the data's range

```
boxplot(data$Sold, horizontal = T)
abline(v = median(data$Sold), col = "green")
abline(v = quantile(data$Sold)[2], col = "red")
abline(v = quantile(data$Sold)[4], col = "red")
abline(v = max(min(data$Sold), quantile(data$Sold)[2] - 1.5 * iqr), col = "purple")
abline(v = min(max(data$Sold), quantile(data$Sold)[4] + 1.5 * iqr), col = "purple")
```



To make the LT and UT staying within the range of data, R uses the convention

$$LT = max(min(x), Q_1 - 1.5IQR)$$

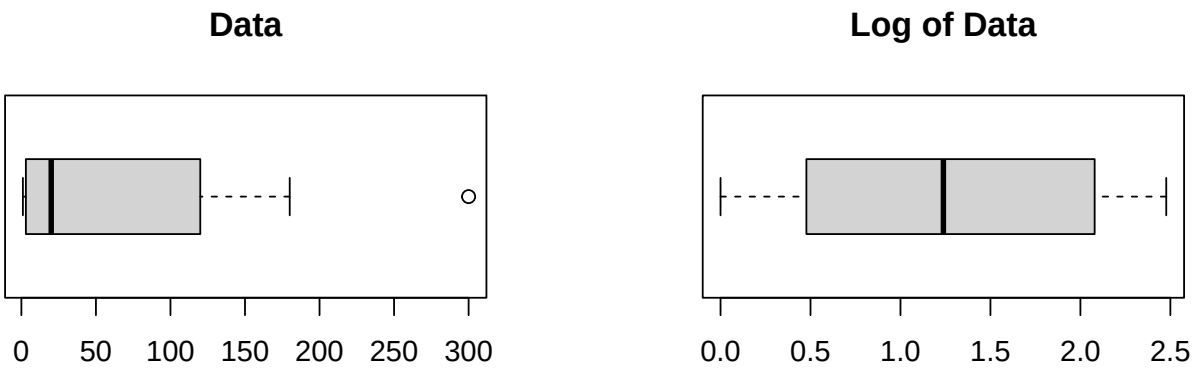
and

$$UT = min(max(x), Q_3 + 1.5IQR)$$

2.7.7 Dealing with outliers (not for examination)

Sometimes outliers indicate that a better model is needed. We may remove outliers by transforming the data. For example, a right skewed data set with outliers can be transformed into the logarithmic scale.

```
w = c(1, 2, 3, 4, 10, 30, 60, 120, 180, 300)
w1 = log(w, 10)
par(mfrow = c(1, 2))
boxplot(w, main = "Data", horizontal = T)
boxplot(w1, main = "Log of Data", horizontal = T)
```



2.7.8 Coefficient of variation (not examinable)

- The **Coefficient of Variation (CV)** combines the mean and standard deviation into one summary:

$$\text{CV} = \frac{\text{SD}}{\text{mean}}$$

- It is the standard deviation relative to the mean (which measures the relative spread) - The CV is used in: - analytical chemistry to express the precision and repeatability of an assay; - engineering and physics for quality assurance studies; - economics for determining the volatility of a security.

2.8 Write a function in R

A function in R is one of the most used objects. For example, `mean`, `median`, `sd` are all R functions. It is very important to understand the purpose and syntax of R functions and knowing how to create or use them.

To declare a user-defined function in R, we use the keyword function.

```
function_name <- function(parameter1, parameter2) {
    # function body
    c = parameter1 + parameter2
    # return the outputs
    return(c)
}
```

Here we declared a function with name `function_name`, the function takes inputs `parameter1`, `parameter2` and returns an output `c`. It can take any number of inputs but **only one** outputs.

Example

Here we want to write a function in R that calculates the sample mean and sample standard deviation

```
my_summary <- function(X) {
    # Write operations within the curly brackets
    m = sum(X)/length(X)
    s = sqrt(sum((X - m)^2)/(length(X) - 1))
    # put mean and sd in a vector, then return the vector as a single output
    return(c(m, s))
}
```

Then we can reuse all the operations defined in the function.

```
w = c(1, 2, 3, 4, 10, 30, 60, 120, 180, 300)  # a data vector
my_summary(w)  # our function
```

```
[1]  71.0000 100.5651
```

```
c(mean(w), sd(w))  # R built-in function
```

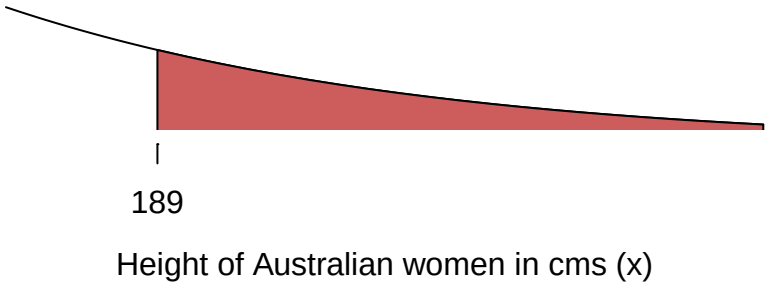
```
[1]  71.0000 100.5651
```

3 Normal Curve

This section introduces the normal curve, which is the first modelling tool.

3.1 Data story: How likely is it to find an elite netball goal player in Australia?

We are going to investigate how likely it is to find an elite netball goal played in Australia. This ABC News article explains that being over 189cm tall is a competitive advantage in Australian netball.

> **Note**
>
> 
>
>
> "A total of 10 goal players (goal shooters, goal keepers, goal attacks and goal defenders) … were **all over 189cm** in height".

Q: However, this is well above the average Australian woman's height, how likely is it to find someone above this height? How could you investigate the proportion of Australian women who are over 189cm in height (potential elite goal players)?

- Collect the heights of Female students in the unit. For example, we have the data collected from "Statistical Thinking with Data" (MATH1005) in 2022 S2

Then we have two options:

- Use the data to represent the population
- Use the data to create a model for the population

Investigation: Data from MATH1005, 2022 S2

```
math1005 = read.csv("data/math1005_cleaned.csv", header = T)
FemaleHeights = math1005$Height[math1005$Gender == "Female"]
FemaleHeights = na.omit(FemaleHeights)
length(FemaleHeights)  # There were 109 female students
```

```
[1] 109
```

```r
hist(FemaleHeights, main = "Histogram of Female students", xlab = "Heights (cm)",
    freq = F)
```

Numerical Investigation

```r
mean(FemaleHeights)
```

```
[1] 164.8633
```

```r
sd(FemaleHeights)
```

```
[1] 7.516324
```

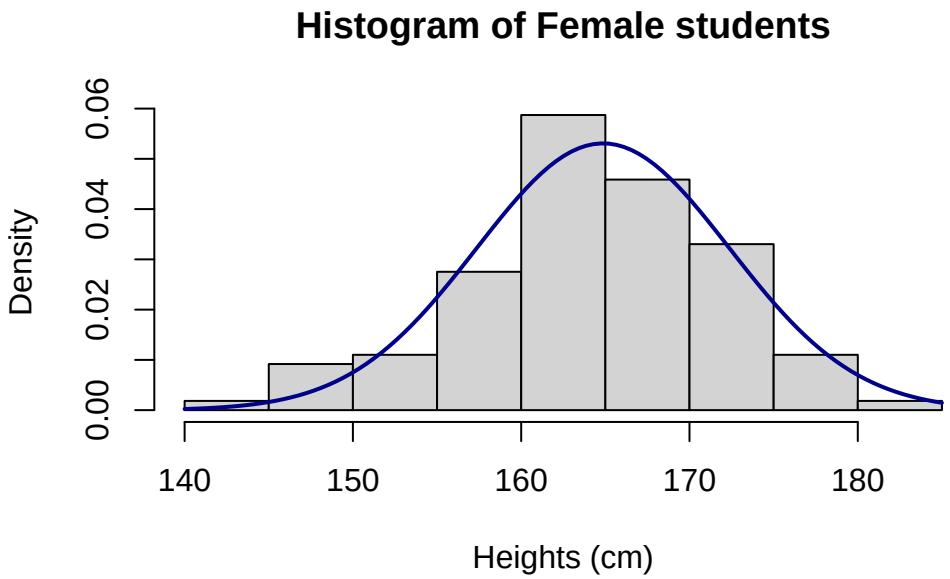
```r
## sum(...) counts the number of FemaleHeights > 189
sum(FemaleHeights > 189)/length(FemaleHeights)
```

```
[1] 0
```

> **Note**
>
> How many students could be elite goal players?

- In this sample, none! But we know there are women in Australia taller than 189cm...

Approximate density-scale histogram



We can draw a smooth curve to approximate the **density-scale** histogram. This curve may extend beyond the range of observed data in the sample to allow us to answer the research question.

How would you describe its shape?

- Fairly symmetric and bell-shaped. Is there something special about this curve?

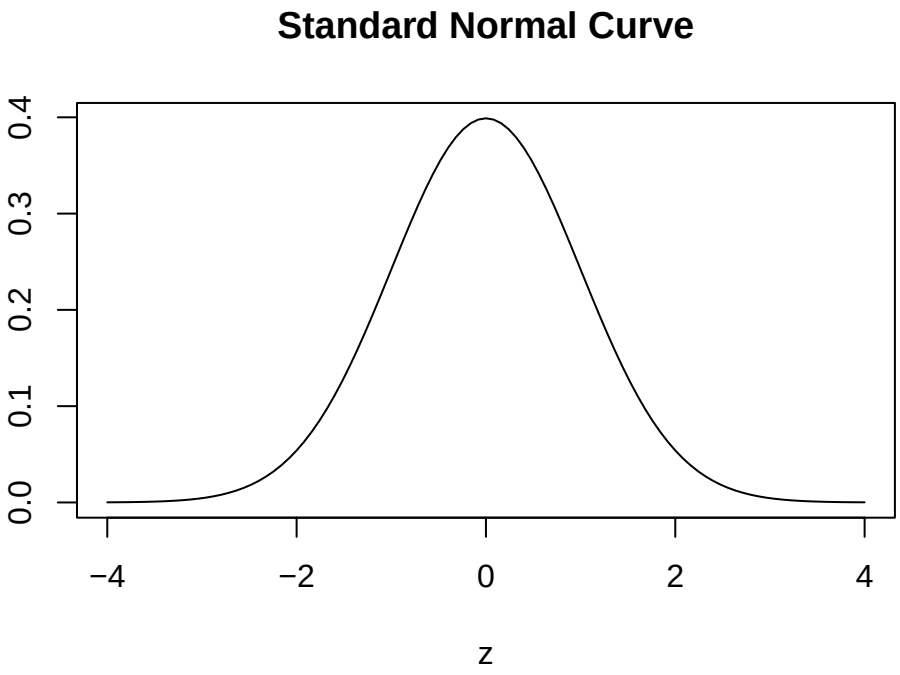
3.2 Normal curve

The Normal curve was defined around 1720 by Abraham de Moivre, also famous for the beautiful de Moivre's formula.

- The Normal curve approximates many **natural phenomena**.
- The Normal curve can model data caused by combining a **large number of independent observations**. (Coming up in a future lecture after introducing probability)
- Many of its properties can be obtained using elementary single variable calculus.
- Defined by mean (center) and standard deviation (spread) of the data, both of which can be estimated from a sample.

3.2.1 General & Standard Normal curves

- The **General** Normal Curve (X) has any mean and SD. Caution: It is denoted by N(mean, **Variance**), where **Variance = SD**^{2}.
- The **Standard** Normal Curve (Z) has mean 0 and SD 1. Short: $N(0, 1)$



3.2.2 The Normal curve formula

The **general normal curve** can be described by the formula

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in (-\infty, \infty)$$

where we can control the shape by (μ, σ):

- μ is the mean, or the centre of the curve
- σ is the standard deviation, or the spread of the curve.

3.3 Area under normal curves

The area under any general normal curve $N(\mu, \sigma^2)$, bounded by some interval (a, b), is given by

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

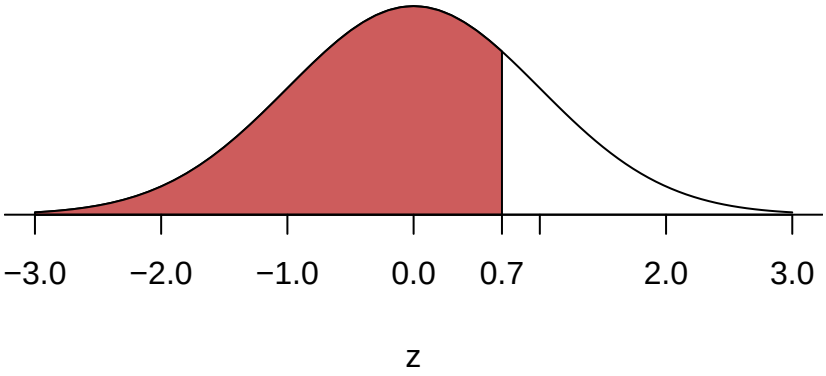
- The total area under the normal curve (between $a = -\infty$ and $b = \infty$) is 1.
- X denotes data following a general normal curve with mean μ and SD σ.
- $P(a < X < b)$ denotes the proportion of data falling into the interval (a, b).
- We will later use this notation also for probability and random variables.

3.3.1 Simplification: Area under the standard normal curve

We start with some data Z modeled by the standard normal curve $N(0, 1)$. As $\mu = 0$ and $\sigma = 1$, the proportion of data falling into the interval (a, b) is

$$P(a < Z < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$$

For example, the proportion of data that is 0.7 or lower is given by the area up to 0.7.



But how to calculate this?

Method 1: Integration

By its definition, we could use integration:

$$P(Z < 0.7) = \int_{-\infty}^{0.7} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

But this does not have a closed form.

Method 2: Normal tables (not for assessment)

This is the old way. We table the values of the integral.

TABLE 1. **Lower tail areas of the Standard Normal distribution (CDF)** The point tabulated is $\Phi(z) = P(Z \leq z)$, where $Z \sim N(0, 1)$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Method 3: Use R

- The `pnorm(x)` command works out the **lower tail** area, it gives

$$P(Z < x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

- `pnorm(x,lower.tail=F)` works out the **upper tail** area, it gives $P(Z > x)$
- We also have

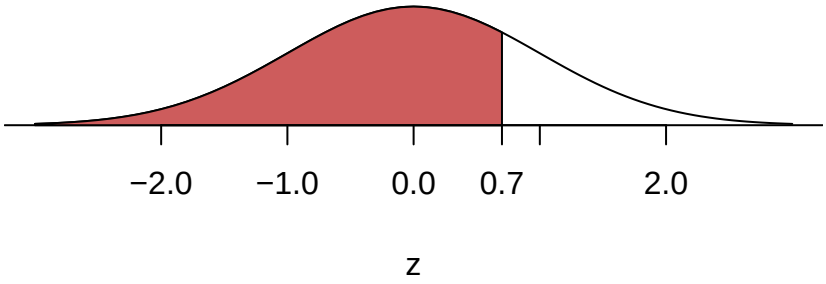
$$P(Z > x) = 1 - P(Z < x) \quad \text{or} \quad \text{upper tail area} = 1 - \text{lower tail area}$$

- It is useful to sketch the normal curve and the relevant area ... and then use R.

Lower tail

What proportion of data is 0.7 or lower?

$$P(Z < 0.7) \approx 0.76$$



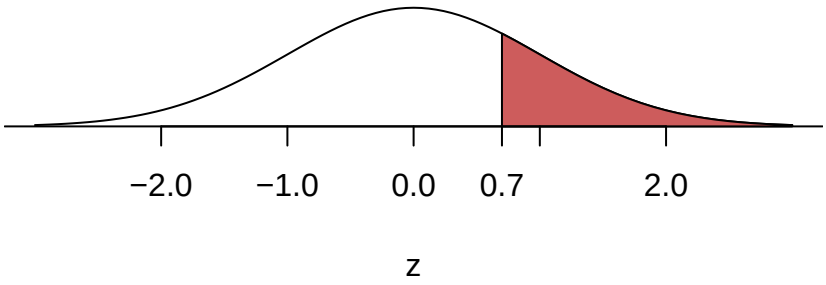

```
pnorm(0.7)
```

```
[1] 0.7580363
```

Upper tail

What proportion of data is 0.7 or higher?

$$P(Z > 0.7) \approx 0.24$$




```
pnorm(0.7, lower.tail = F)
```

```
[1] 0.2419637
```

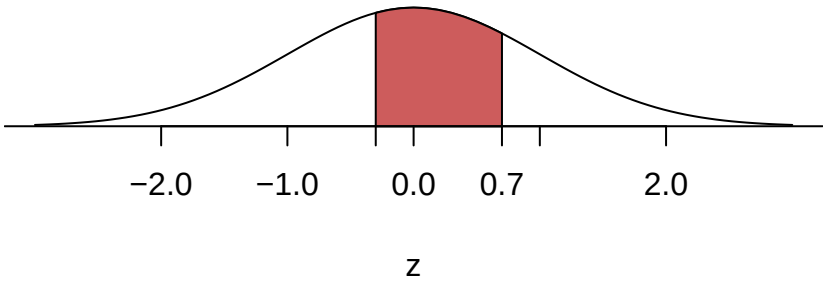
```
1 - pnorm(0.7)  # alternative way
```

```
[1] 0.2419637
```

Interval

What proportion of data is between -0.3 and 0.7?

$$P(-0.3 < Z < 0.7) = \underbrace{P(Z < 0.7)}_{\int_{-\infty}^{0.7} f(z)dz} - \underbrace{P(Z < -0.3)}_{\int_{-\infty}^{-0.3} f(z)dz} \approx 0.38$$




```
pnorm(0.7) - pnorm(-0.3)
```

```
[1] 0.3759478
```

3.3.2 Area under general normal curves

```
mean(FemaleHeights)
```

```
[1] 164.8633
```

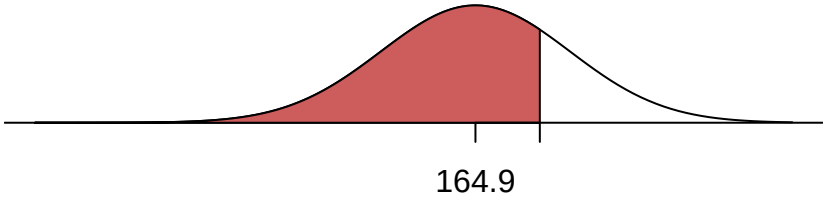
```
sd(FemaleHeights)
```

```
[1] 7.516324
```

- The heights of female students in MATH1005 has a mean of 164.9cm and a standard deviation of 7.52cm.
- Now we can model the heights of all Australian women with a normal curve with mean 164.9cm and standard deviation of 7.5cm.

Lower tail

Suppose the heights of Australian women follow a normal distribution with mean 164.9cm and sd 7.5cm. What proportion of women will have height less than 170cm?

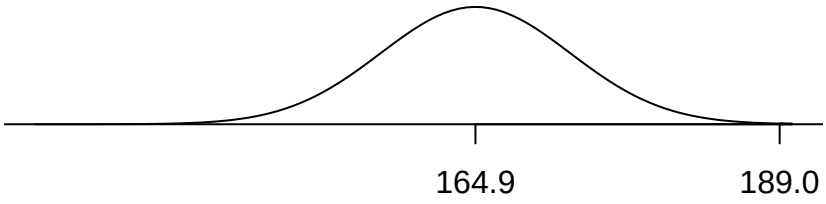



```
m = mean(FemaleHeights)
s = sd(FemaleHeights)
pnorm(170, m, s)  #pnorm(x,mean,sd)
```

```
[1] 0.7528247
```

Upper tail

What proportion of women will have height greater than 189cm? How likely is to find an elite netball goal player in Australia?




```
m = mean(FemaleHeights)
s = sd(FemaleHeights)
pnorm(189, m, s, lower.tail = FALSE)  #upper tail, pnorm(x,mean,sd)
```

```
[1] 0.0006608243
```

```
1 - pnorm(189, m, s)  # 1 - lower tail
```

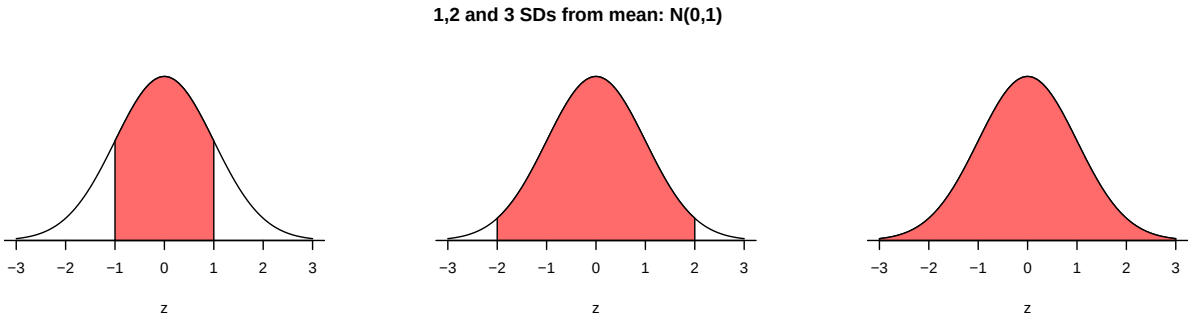
```
[1] 0.0006608243
```

3.4 Properties of the normal curve

3.4.1 68% 95% 99.7% Rule

All normal curves satisfy the "68%-95%-99.7% rule":

- The area **1** SD out from the mean in both directions is **0.68** (68%).
- The area **2** SDs out from the mean in both directions is **0.95** (95%).
- The area **3** SDs out from the mean in both directions is **0.997** (99.7%).



Under a normal curve, it has a low chance (0.3%) to have data points that fall more than 3 SD away from the mean.

3.4.2 Rescaling

Any general normal curve can be rescaled into the standard normal curve.

Consider data X following a general normal curve $N(\mu, \sigma^2)$. For any point on this normal curve, recall that the standard unit (or z score) is how many standard deviations that point is above (+) or below (-) the mean.

$$\text{standard unit} = \frac{\text{data point - sample mean}}{\text{sample SD}} \quad \text{or} \quad z = \frac{x - \mu}{\sigma}$$

- The standard units give the relative location of a data point on the standard normal curve. - The proportion under a general normal curve $P(X < a)$ is equivalent to the proportion under the standard normal curve $P(Z < \frac{a-\mu}{\sigma})$

The following derivation is not for assessment.

The proportion of data modelled by $N(\mu, \sigma^2)$ falling below a is

$$P(X < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Apply the change of variable (standardisation) $z = \frac{x-\mu}{\sigma}$

$$\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = \int_{-\infty}^{\frac{a-\mu}{\sigma}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}z^2} \frac{dx}{dz} dz$$

where

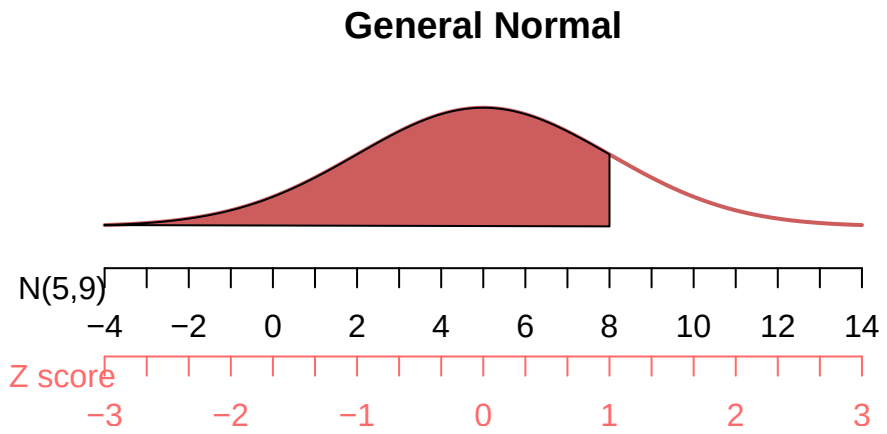
$$z = \frac{x-\mu}{\sigma} \quad \Longrightarrow \quad x = \sigma z + \mu \quad \Longrightarrow \quad \frac{dx}{dz} = \sigma$$

so the proportion simplifies to

$$P(X < a) = \int_{-\infty}^{\frac{a-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = P\left(Z < \frac{a-\mu}{\sigma}\right)$$

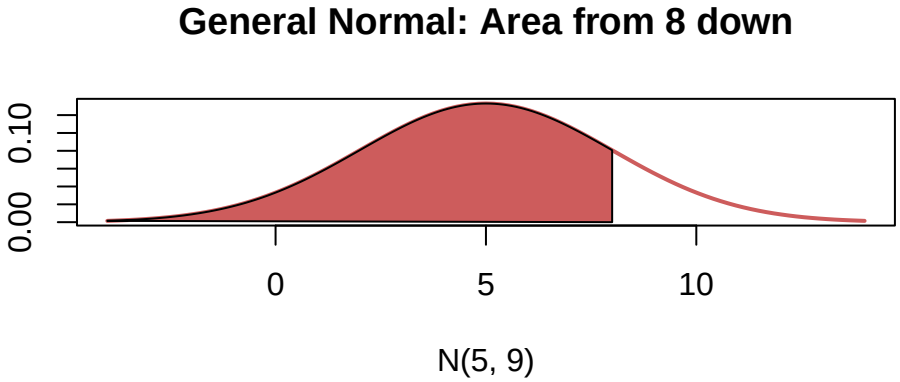
which is the proportion of data modelled by $N(0,1)$ falling below $\frac{a-\mu}{\sigma}$.

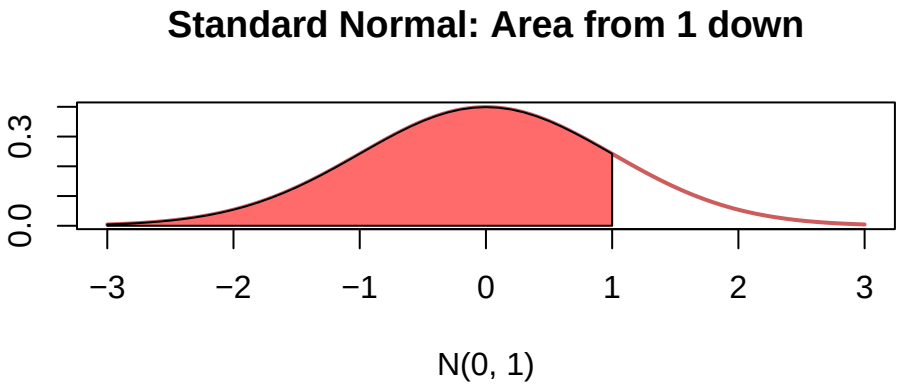
Example 1



- Consider the point $= 8$.
- So the z score of the point is $\frac{8-5}{3} = 1$.

The following 2 areas are of the same size.





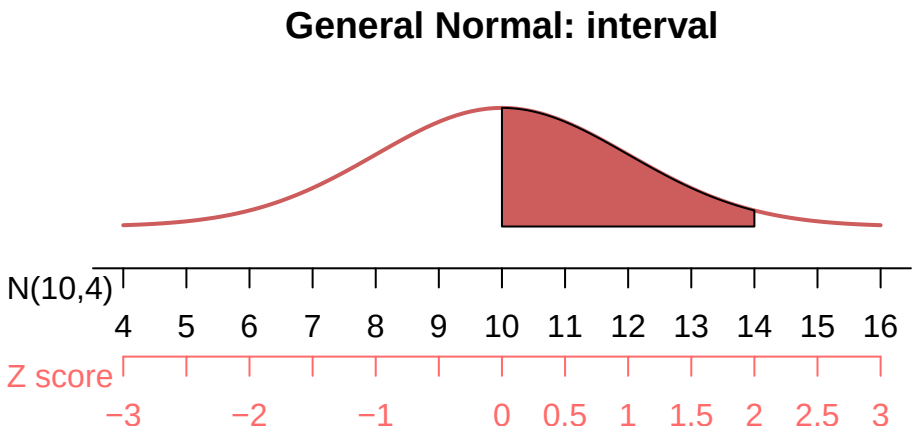

```
pnorm(8, 5, 3)
```

```
[1] 0.8413447
```

```
pnorm(1)
```

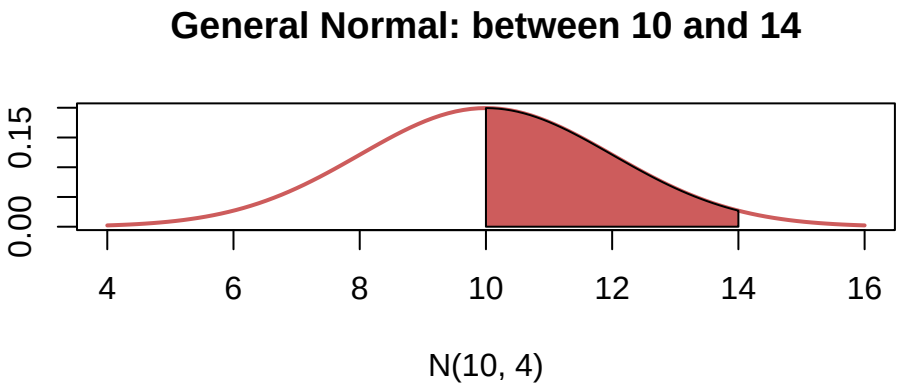
```
[1] 0.8413447
```

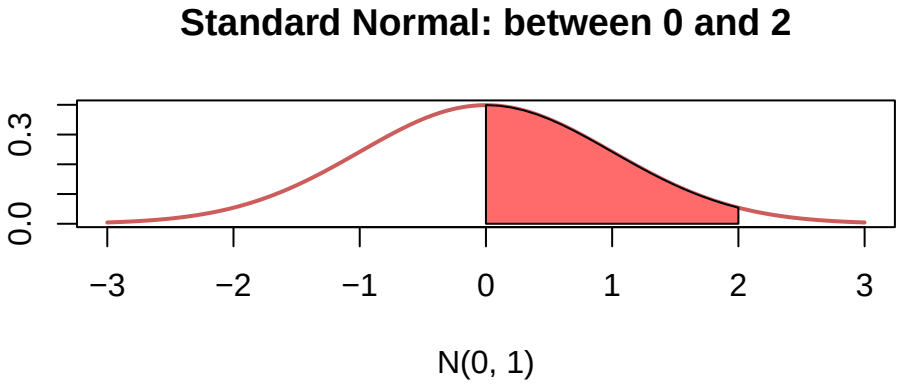
Example 2



- Here the lower point is 10 and the upper point is 14.
- So the z scores are $z_1 = \frac{10-10}{2} = 0$ and $z_2 = \frac{14-10}{2} = 2$.

The following 2 areas are of the same size.






```
pnorm(14, 10, 2) - pnorm(10, 10, 2)
```

```
[1] 0.4772499
```

```
pnorm(2) - pnorm(0)
```

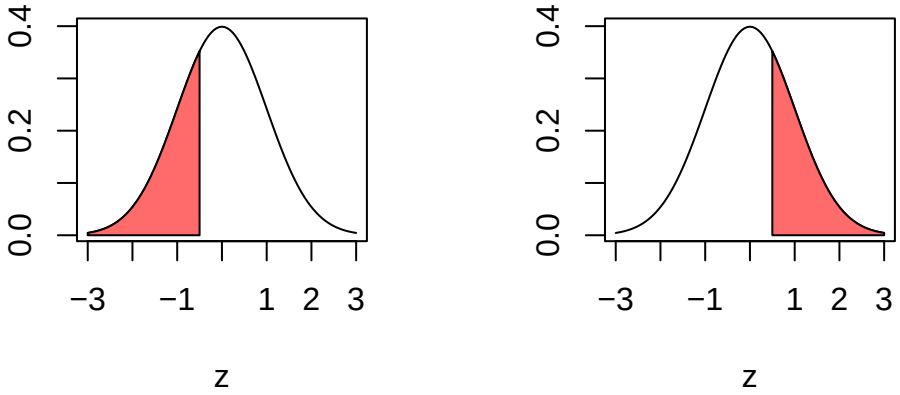
```
[1] 0.4772499
```

3.4.3 The normal curve is symmetric about the mean

If Z follows the standard normal curve $N(0, 1)$, then

$$P(Z < -a) = P(Z > a)$$

The red areas below are of the same size (where $a = 0.5$).



More generally, if X follows a general normal curve $N(\mu, \sigma^2)$, then

$$P(X < \mu - a) = P(X > \mu + a)$$

```
mu = 10
sigma = 2
a = 2
pnorm(mu - a, mu, sigma)  # lower tail
```

```
[1] 0.1586553
```

```
pnorm(mu + a, mu, sigma, lower.tail = F)  # upper tail
```

```
[1] 0.1586553
```

3.5 Calculate the quantiles of normal curves using R

The function `pnorm()` finds "the proportion of data X following a normal curve falling below the value a", we are also interested in

- What is the quantile Q such that p% of the data X falling below the value Q?

Similar to the proportion, there is no close-form solution for the quantiles of normal curves. We can calculate the quantiles using `qnorm(x, mu, sigma)` in R.

```
mu = 10
sigma = 2
qnorm(0.7, mu, sigma)  # 70-percentile of N(10, 4)
```

```
[1] 11.0488
```

```
qnorm(0.5, mu, sigma)  # 50-percentile (or the median) of N(10, 4)
```

```
[1] 10
```

```
qnorm(0.7)  # 70-percentile of the standard normal N(0, 1)
```

```
[1] 0.5244005
```

```
qnorm(0.5)  # 50-percentile (or the median) of N(0, 1)
```

```
[1] 0
```

3.6 Summary

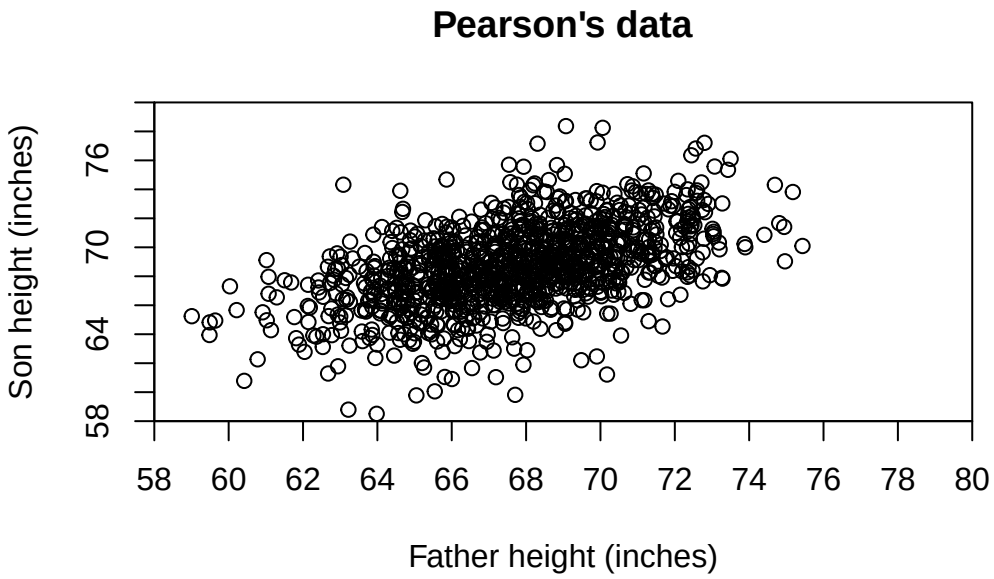
- The Normal curve naturally describes many histograms, and so can be used in modelling data.
- It can be described by the mean and the variance (SD^2).
- Area under normal curves and using `pnorm` to find the area below x.
- It has many useful properties, including the 68/95/99.7% rule.
- Any general normal curve can be rescaled into a standard normal curve.
- The normal curve is symmetric about the mean.
- Quantiles and using `qnorm` to find Q for which p% of values fall below.

4 Linear Models

In this section, we will explore correlation coefficient of bivariate data and its properties. We will then introduce the linear model, learning about predictions, and analysing residuals to assess model accuracy. Finally, we will discuss the coefficient of determination and how to check if a model fits the data well.

4.1 Scatter plots and Pearson's data

Sir Francis Galton (England, 1822–1911) studied the degree to which children resemble their parents. Galton's work was continued by his student Karl Pearson (England, 1857–1936). Pearson measured the heights of 1,078 fathers and their sons at maturity.



Generally, taller fathers tend to have taller sons. **Code for plotting Pearson's data:**

```
# install.packages('UsingR')
suppressMessages(library(UsingR))
library(UsingR)  # Loads another collection of datasets
data(father.son)  # This is Pearson's data.
data = father.son
x = data$fheight  # fathers' heights
y = data$sheight  # sons' heights
## scatter plot
plot(x, y, xlim = c(58, 80), ylim = c(58, 80), xaxt = "n", yaxt = "n", xaxs = "i",
    yaxs = "i", main = "Pearson's data", xlab = "Father height (inches)", ylab = "Son
     ↪ height (inches)")
axp = seq(58, 80, by = 2)  # Adjust the gap between label and plot
```

```r
axis(1, at = axp, labels = axp)
axis(2, at = axp, labels = axp)
```

Why do we care the association between two variables (here: height of father and son)?

- The association is interesting on its own.
- Association between two variables can be used for prediction, i.e, use outcome in one variable to predict the outcome in another variable.
- How can we quantify a possible association?

4.2 Correlation coefficient

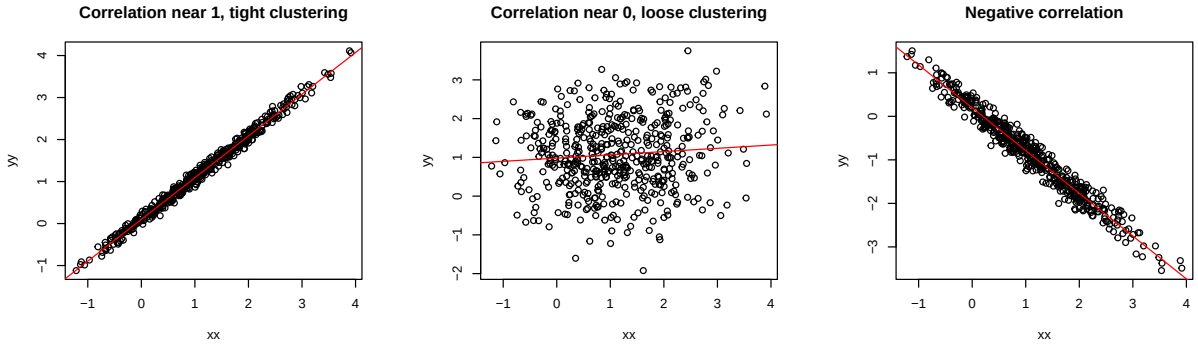
Pearson's data is a typical example of **bivariate data** involves a **pair** of variables. We are interested in the relationship between the two variables. Can one variable be used to predict the other?

- Formally, we have (x_i, y_i) for $i = 1, 2, \ldots, n$.
- X and Y can have the same role
- X and Y may have different roles: for example, X can be an **independent** variable (or explanatory variable, predictor or regressor) which we use to explain or predict Y, the **dependent** variable (or response variable).

Bivariate data can be summarised by the following **five** numerical summaries:

- Sample mean and sample SD of X ($\bar{x}$, SD_x)
- Sample mean and sample SD of Y ($\bar{y}$, SD_y)
- Correlation coefficient (r).

The following clouds have the **same centre and horizontal and vertical spread**. However they have **different spread** around a line (linear association). We use the correlation coefficient (r) to measure this.

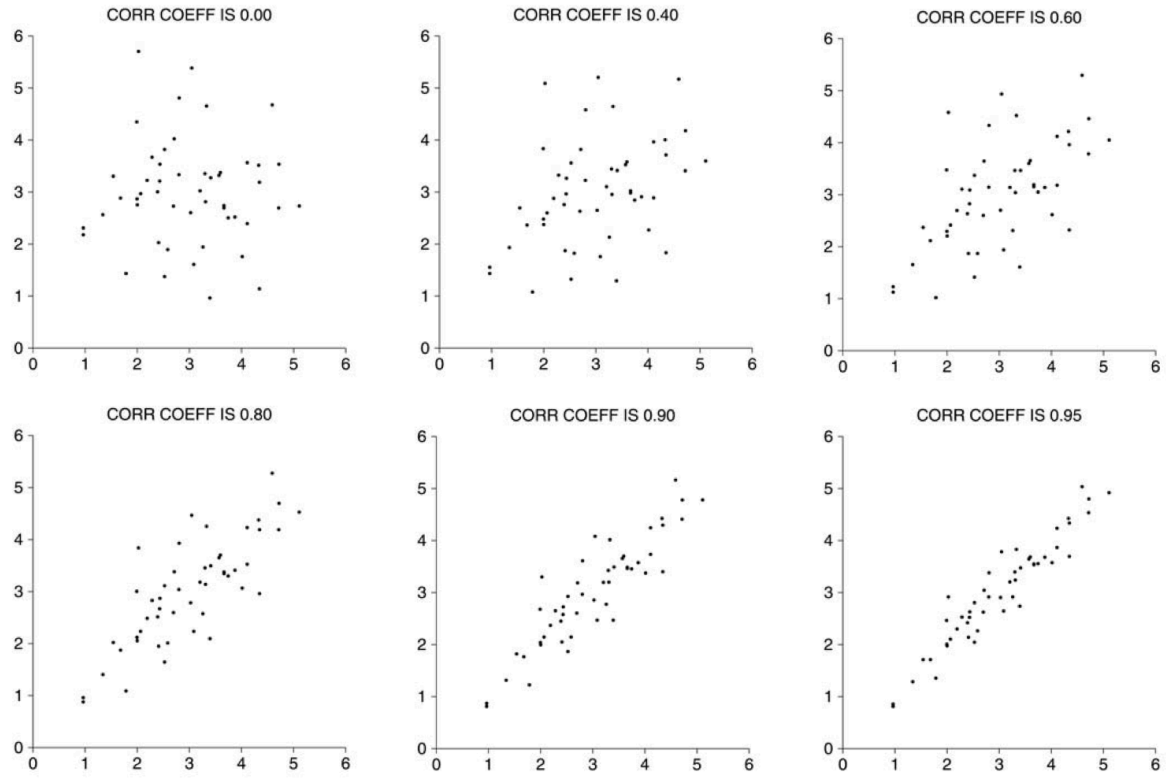


4.2.1 The correlation coefficient

The (Pearson's) **correlation coefficient** r is a numerical summary which measures of how points are spread around the line. It indicates both the sign and strength of the **linear association**. The correlation coefficient is between -1 and 1.

- If r is positive: the cloud slopes up.
- If r is negative: the cloud slopes down.
- As r gets closer to ± 1: the points cluster more tightly around the line.
- $r = 0$ implies no linear dependency between two variables.

Examples (Source: Freedman et al, Statistics p127)



The (Pearson) correlation coefficient (r)

- A numerical summary measures of how points are spread around the line.
- It indicates both the sign and strength of the **linear association**.
- It is defined as the mean of the product of the variables in **standard units**.

Recall that

$$\text{standard unit} = \frac{\text{data point} - \text{mean}}{SD}$$

Using sample SD, we divide by $n-1$ in the average:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{SD_{sample}(X)} \frac{(y_i-\bar{y})}{SD_{sample}(Y)} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

which simplifies to $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$.

Obtaining r using the population SD

The same correlation coefficient r can be obtained using the population SD as well (dividing by n in the average).

$$r = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{SD_{pop}(X)} \frac{(y_i-\bar{y})}{SD_{pop}(Y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

which also simplifies to $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$.

Quick calculation in R using `cor()`.

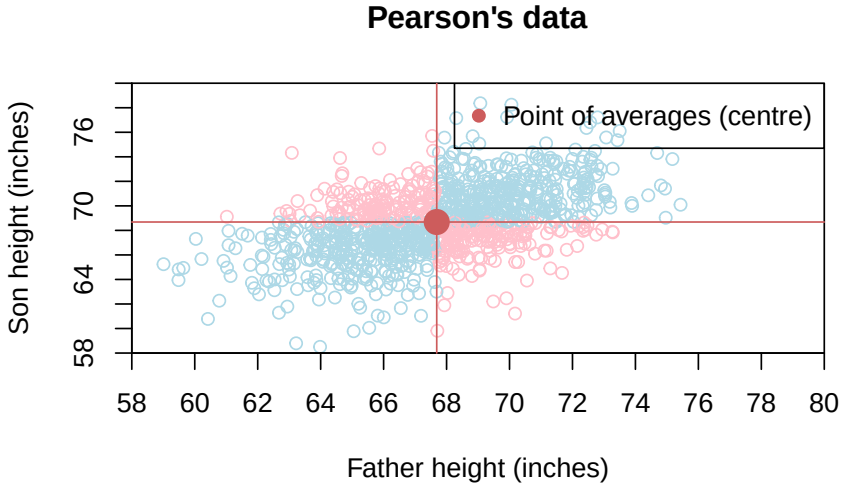
```
cor(x, y)
```

```
[1] 0.5013383
```

How does r measure association?

We divide the scatter plot into 4 quadrants, at the point of averages (centre).

- In the upper right and lower left quadrants, products of standard units are (+)
- In the upper left and lower right quadrants, products of standard units are (-)



- A majority of points in the upper right (+) and lower left quadrants (+) will be indicated by a positive r
- A majority of points in the upper left (-) and lower right quadrants (-) will be indicated by a negative r

4.3 Properties and warnings

4.3.1 Interpretations of r values

- The correlation coefficient r always takes values between -1 and 1 (inclusive).
  - This can be shown using the definition of r and the Cauchy-Schwarz inequality (only for your information).
- If r is positive: the cloud slopes up.
- If r is negative: the cloud slopes down.
- $r = 0$ implies no linear dependency between two variables.
- As r gets closer to ± 1: the points cluster more tightly around the line.

4.3.2 Invariant properties

Shift and scale invariant

The correlation coefficient is shift and scale invariant. Why? **Shifting and scaling do not change the standard unit.**

```
cor(x, y)
```

```
[1] 0.5013383
```

```
cor(0.2 * x + 3, 3 * y - 1)
```

```
[1] 0.5013383
```

Symmetry (commutative)

The correlation coefficient is not affected by interchanging the variables.

```
cor(x, y)
```

```
[1] 0.5013383
```

```
cor(y, x)
```

```
[1] 0.5013383
```

4.3.3 Warnings

Warning 1: Wrong interpretations of correlation coefficient

Mistakes:

- $r = 0.8$ means that 80% of the points are tightly closed around the line.
- $r = 0.8$ means that the points are twice as tightly closed as $r = 0.4$.

> **Note**
>
> $r = 0.8$ suggests a stronger association between variables compared to the case $r = 0.4$ BUT does not suggest the data points are twice as tight.

Warning 2: Outliers can overly influence the correlation coefficient

Suppose there was an extra unusual reading of (100,50).

```
f1 = c(data$fheight, 100)  # Add an extra point to data
s1 = c(data$sheight, 50)
```

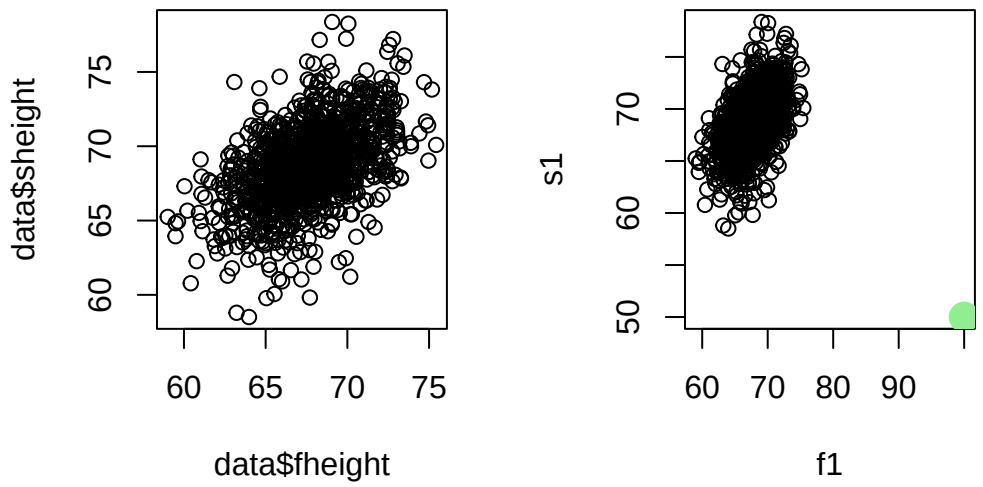
```
cor(data$fheight, data$sheight)
```

```
[1] 0.5013383
```

```
cor(f1, s1)
```

```
[1] 0.3956794
```

```
par(mfrow = c(1, 2))
plot(data$fheight, data$sheight)
plot(f1, s1)
points(100, 50, col = "lightgreen", pch = 19, cex = 2)
```



Warning 3: Nonlinear association can't be detected by the correlation coefficient

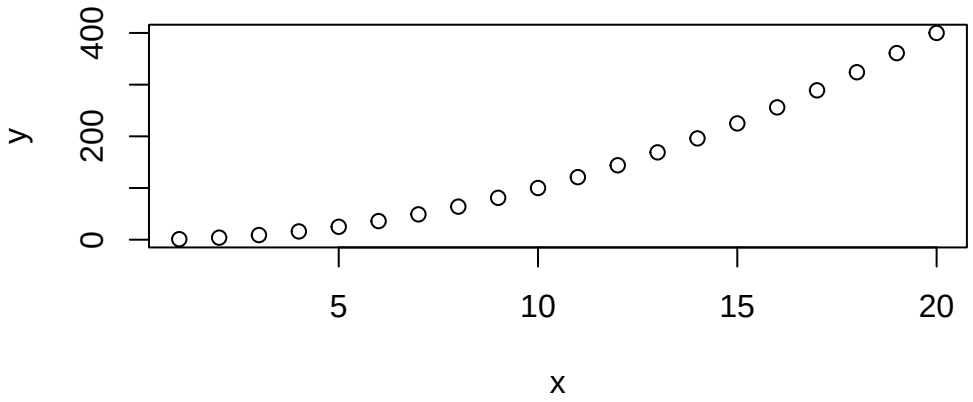
What interpretation mistake could be made in the following data set?

```
x = c(1:20)
y = x^2
cor(x, y)
```

```
[1] 0.9713482
```

Based on the correlation coefficient, the points should cluster very tightly around the line sloping up.

```
plot(x, y)
```



This data should be modelled by a quadratic curve, not a line.

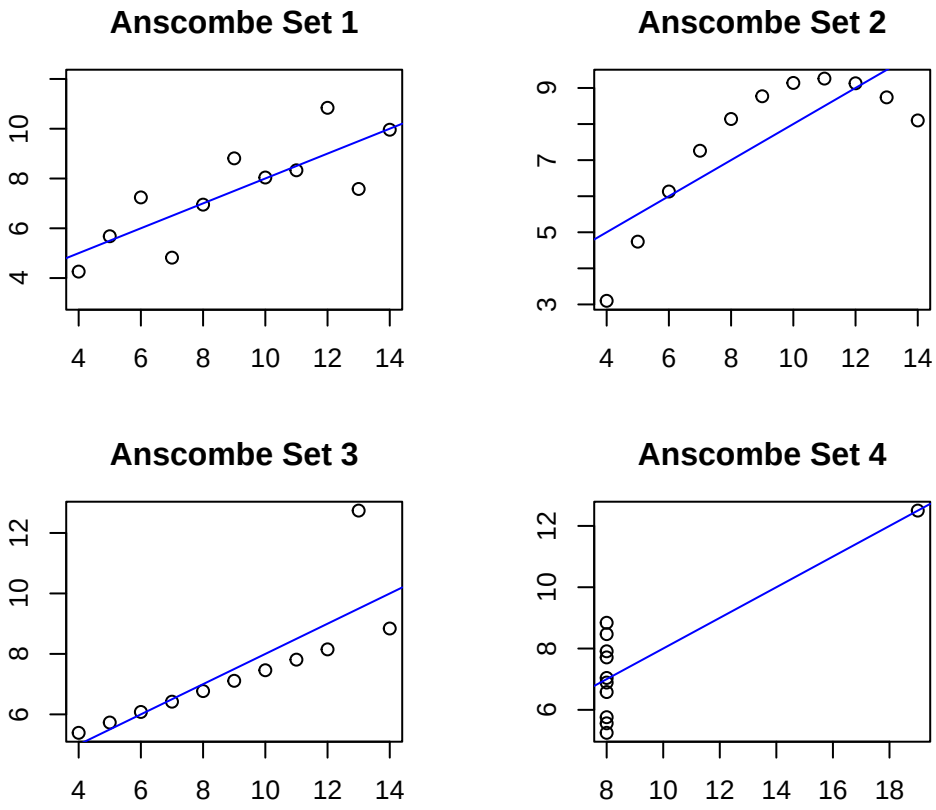
We should always use correlation coefficient together with the scatter plot.

Warning 4: The same correlation coefficient can arise from very different data

The following 4 data sets (Anscombes Quartet) have the **same** $\bar{x}, SD_x, \bar{y}, SD_y$, and also the **same** value of r.

```
x_mean: 9 9 9 9

x_sd: 3.316625 3.316625 3.316625 3.316625

y_mean: 7.500909 7.500909 7.5 7.500909

y_sd: 2.031568 2.031657 2.030424 2.030579

r: 0.8164205 0.8162365 0.8162867 0.8165214
```

But look at the scatter plots.



4.3.4 Association and causation

It is rather easy to establish association (that one thing is linked to another).

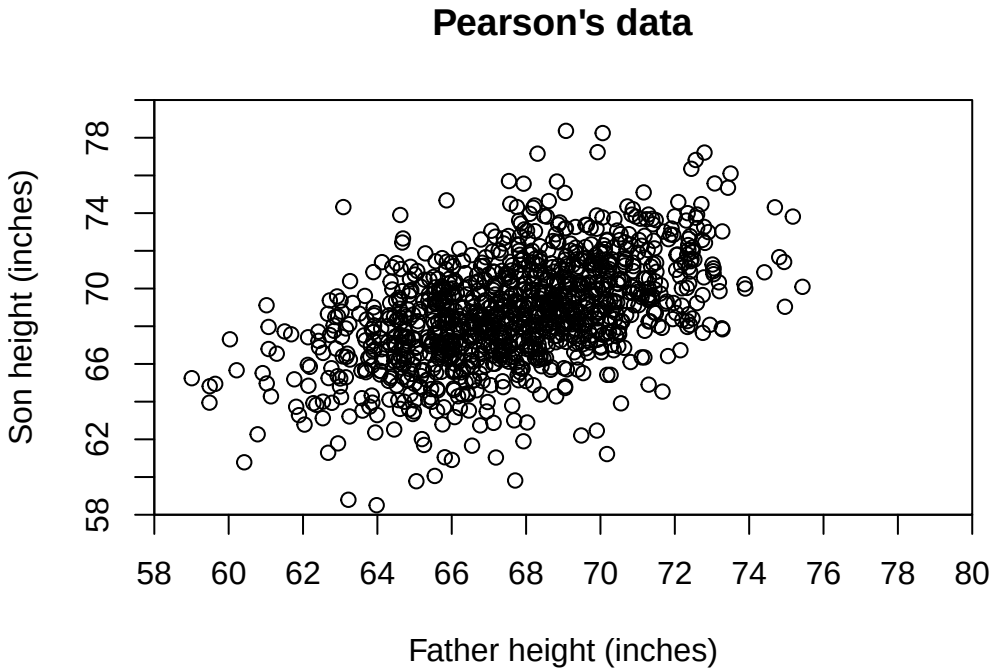
- Association may **suggest** causation. But association does not **prove** causation.
- We need to take **confounding** variables into account. They can mislead about a cause and effect relationship.

What could explain the fact that smokers have a higher rate of liver cancer?

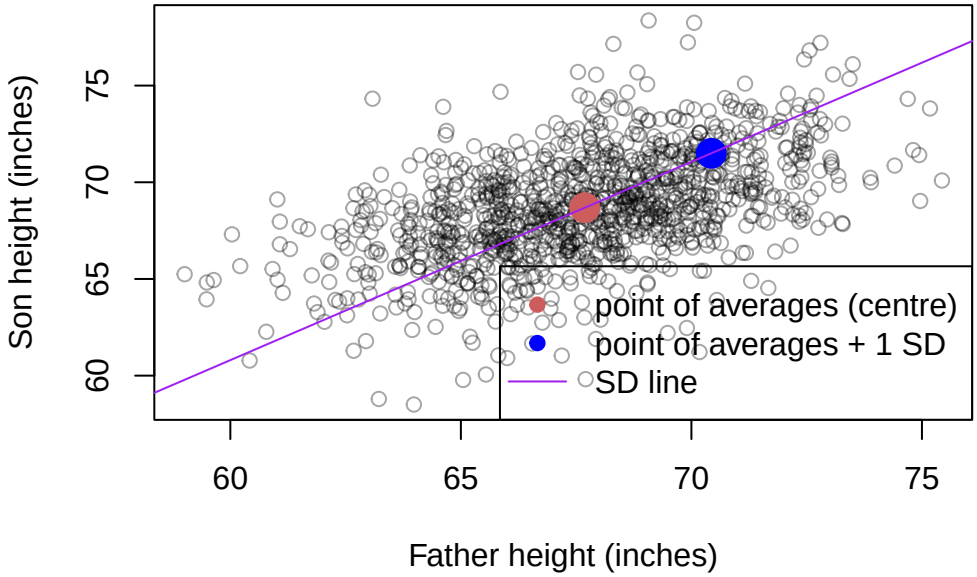
- Smokers tend to drink more alcohol than non-smokers, and excessive alcohol consumption causes liver cancer.
- So the effect of smoking is confounded (mixed-up) with the effect of alcohol consumption.
- Here alcohol consumption is a confounding factor.

4.4 Regression line

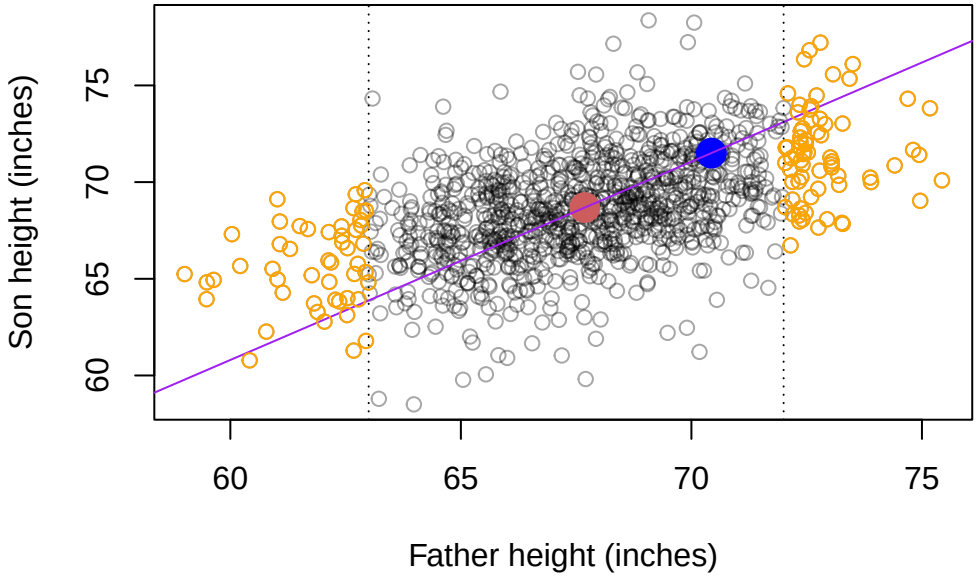
How can we summarise the data with a line? What is the **optimal** line?



1st option: SD line (not so good): The SD line might look like a good candidate as it connects the point of averages $(\bar{x}, \bar{y})$ to $(\bar{x} + \text{SD}_x, \bar{y} + \text{SD}_y)$ (for this data with positive correlation).



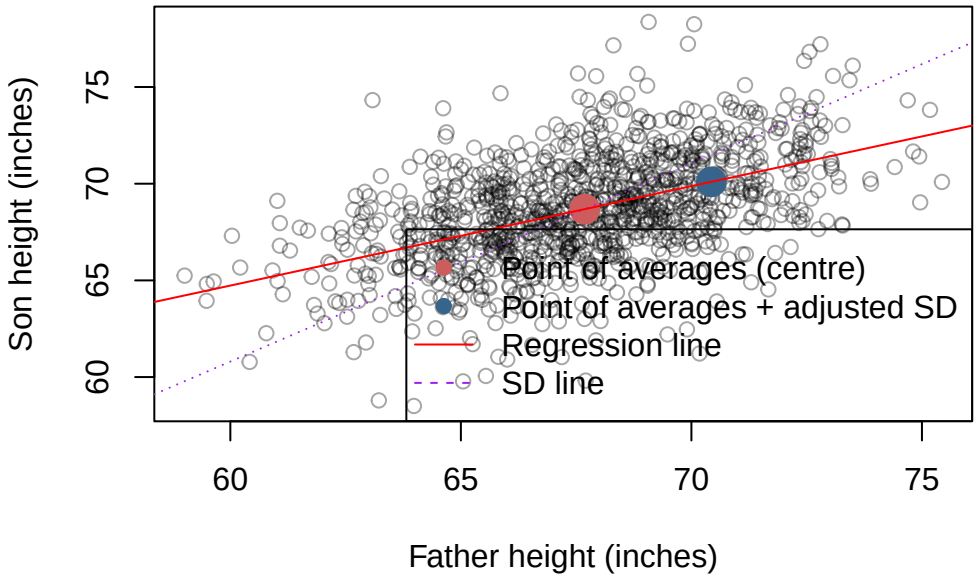
Note how it underestimates (LHS) and overestimates (RHS) at the extremes.



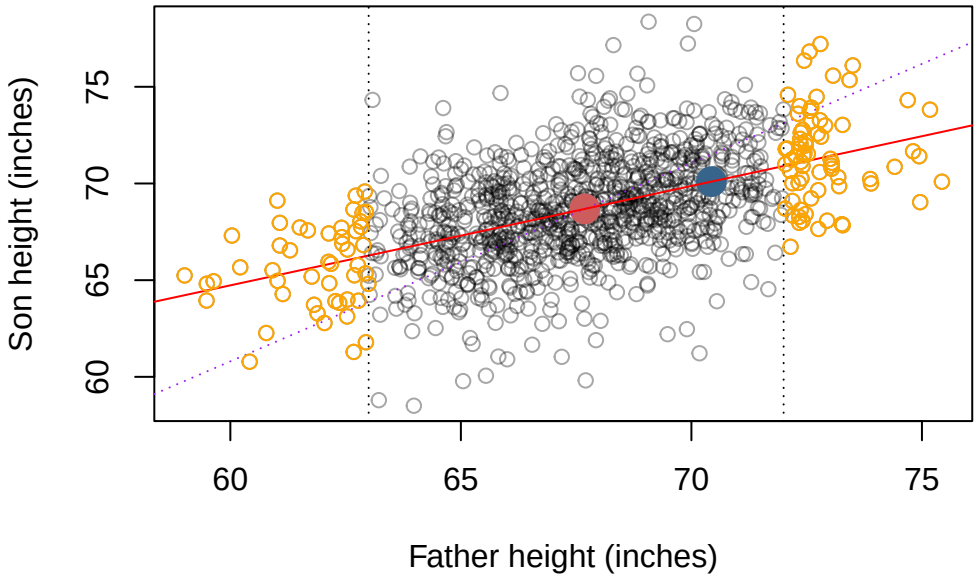
- Recall that X, Y can have the same mean and SD but very different correlation coefficient.
- The above model does not use the correlation coefficient, so it is insensitive to the amount of clustering around the line.

Best option: regression line

- To describe the scatter plot, we need to use **all five** summaries: $\bar{x}$, $\bar{y}$, SD_x, SD_y **and** r.
- The **regression line** connects $(\bar{x}, \bar{y})$ to $(\bar{x} + \text{SD}_x, \bar{y} + r\text{SD}_y)$



Note the improvement at the extremes.



Summary of regression line

Feature	Regression Line $y \sim x$ $(y = a + bx)$
Connects	$(\bar{x}, \bar{y})$ to $(\bar{x} + \text{SD}_x, \bar{y} + r\text{SD}_y)$
Slope (b)	$r\frac{\text{SD}_y}{\text{SD}_x}$
Intercept (a)	$\bar{y} - b\bar{x}$

Optimality: We can derive the regression line using calculus, by minimising the **sum of squares** of the **residuals**.

To calculate the regression line in R, we need to first create a linear model using the function `lm(y ~ x)`. This will give us the y-intercept and gradient of the optimal regression line.

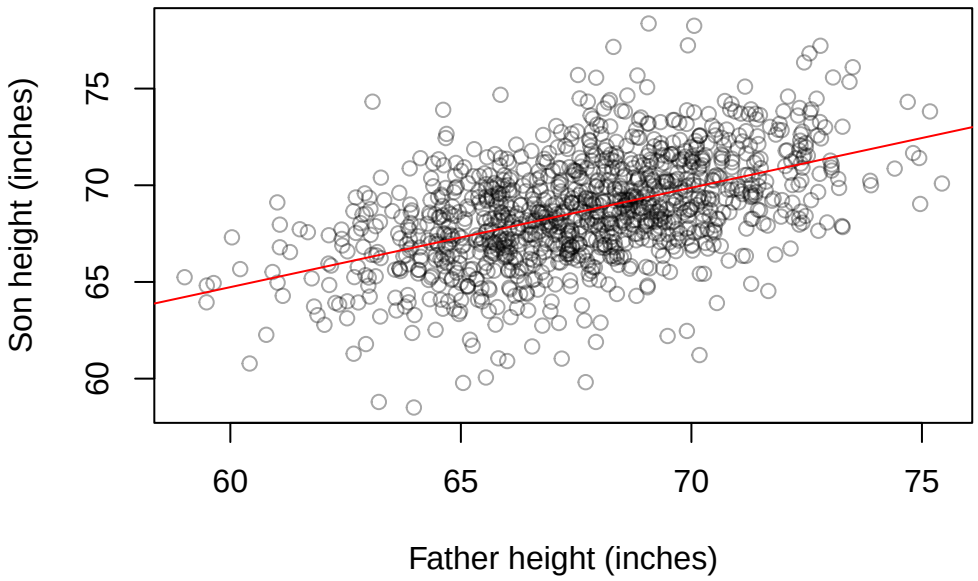
```
model = lm(y ~ x)
model$coeff
```

```
(Intercept)           x
  33.886604    0.514093
```

So for x = father height and y = son height, the regression line is

$$y = 33.886604 + 0.514093x$$

```r
plot(x, y, xlab = "Father height (inches)", ylab = "Son height (inches)", col =
 ↪  adjustcolor("black",
    alpha.f = 0.35))
abline(lm(y ~ x), col = "red")
```



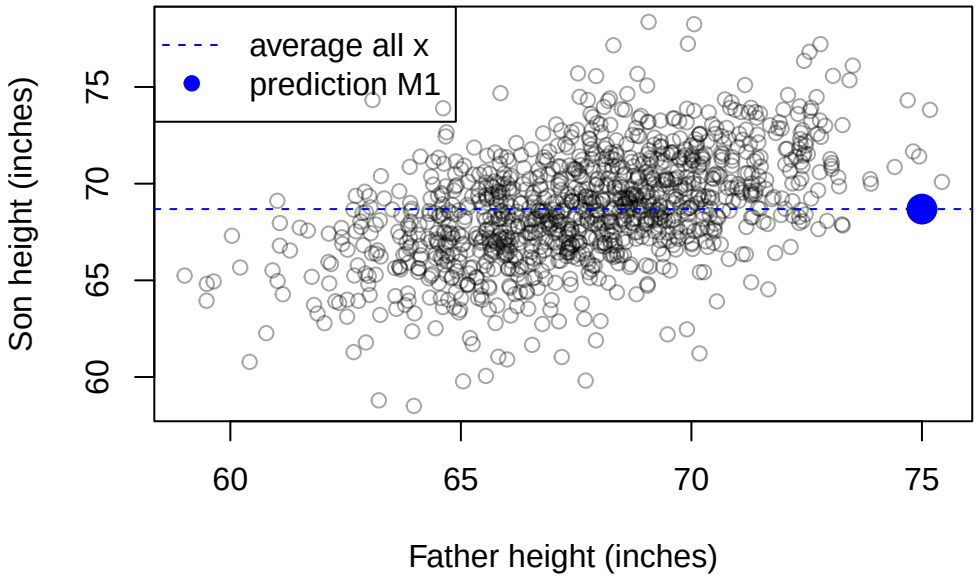
4.5 Prediction

4.5.1 Baseline prediction

- For new born (son), the father is 75 inches tall, how can we predict the son's height?
- If you don't use the information of the independent variable x at all, a basic prediction of y would be the **average** of y for **all** the x values in the data.
- So for any father's height, we could predict the son's height to be 68.68.

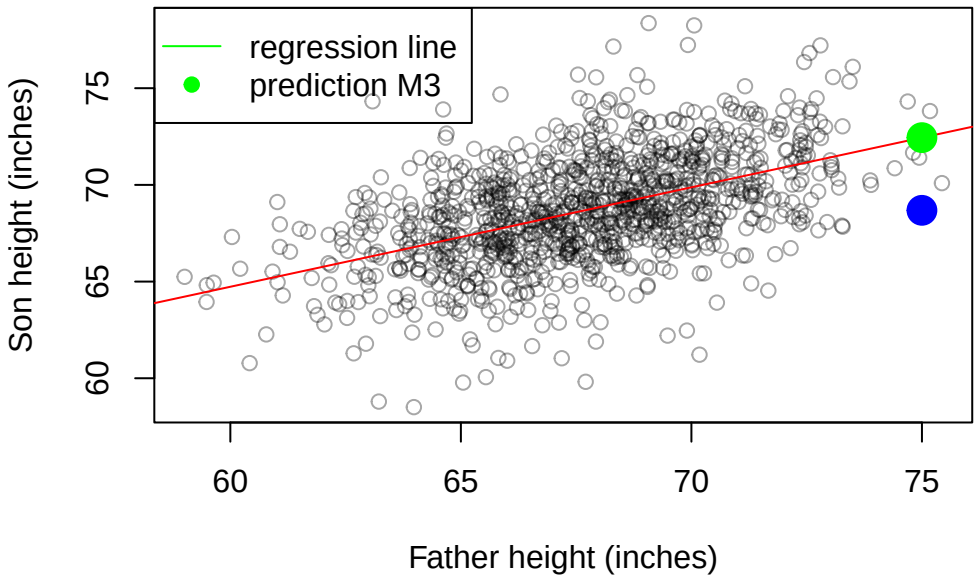
```
mean(y)
```

```
[1] 68.68407
```



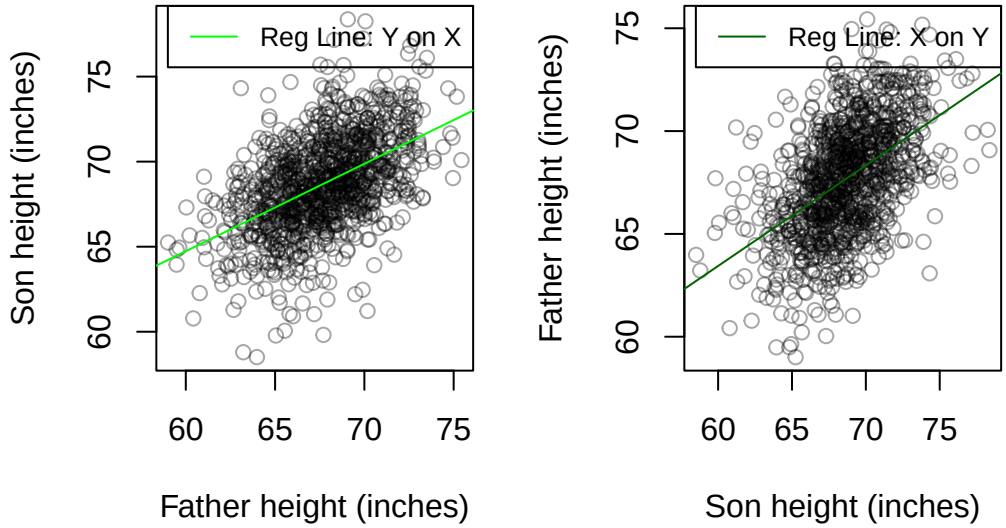
4.5.2 The Regression line

- A better prediction is based on the regression line $y = \text{slope} \times x + \text{intercept}$
- For the height data: $y = 33.886604 + 0.514093x$
- So for any father's height 75, we could predict the son's height to be 72.44.



Can we also use Y to predict X?

We can predict Y from X or X from Y, depending on what fits the context.



Beware!

- Can we just simply rearrange the equation?

$(y = a + bx) \implies (x = -\frac{a}{b} + \frac{1}{b}y)$

- The answer is NO unless $r = \pm 1$ (data clustered along the line).
- We need to **refit** the model.

Feature	Regression Line $y \sim x$ $(y = a + bx)$	Regression Line $x \sim y$ $(x = \tilde{a} + \tilde{b}y)$
Connects	$(\bar{x}, \bar{y})$ to $(\bar{x} + \text{SD}_x, \bar{y} + r\text{SD}_y)$	$(\bar{y}, \bar{x})$ to $(\bar{y} + \text{SD}_y, \bar{x} + r\text{SD}_x)$
Slope	$b = r\frac{\text{SD}_y}{\text{SD}_x}$	$\tilde{b} = r\frac{\text{SD}_x}{\text{SD}_y}$
Intercept	$a = \bar{y} - b\bar{x}$	$\tilde{a} = \bar{x} - \tilde{b}\bar{y}$

Rearranging the equation leads to different coefficients

```
lm(y ~ x)
```

```

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    33.8866       0.5141
```

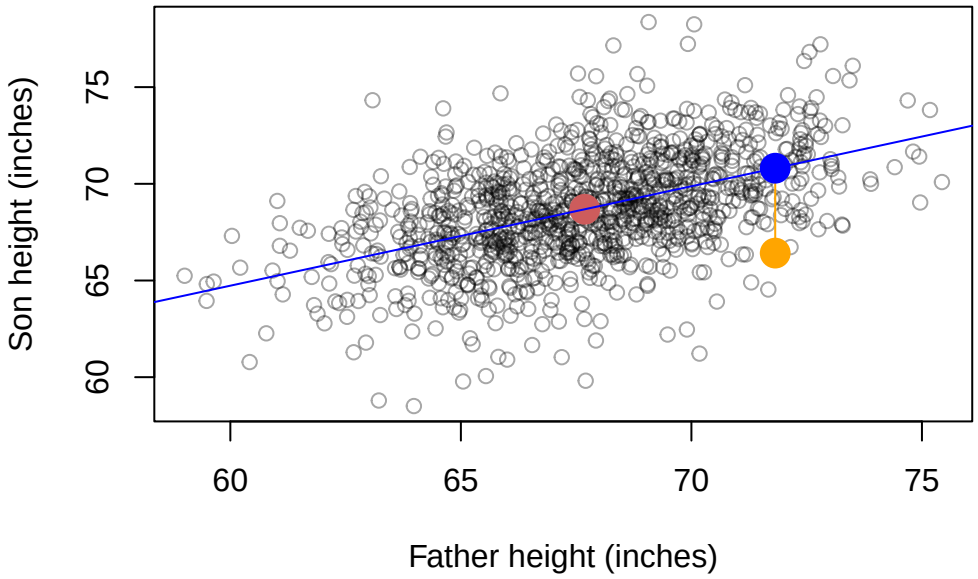
```
lm(x ~ y)
```

```

Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)            y
    34.1075       0.4889
```

4.6 Residuals and properties

We can now make predictions using the regression line. But we have some prediction **error**.

- A **residual** is the vertical distance of a point above or below the regression line.
- A residual represents the error between the actual value and the prediction.



When the father's height is 71.82, the **actual value** of the son's height is 66.42 with **predicted value** 70.81, so the residual is -4.39.

Formally, given the actual value (y_i) and the prediction ($\hat{y}_i$), a residual is

$$e_i(a,b) = y_i - \hat{y}_i = y_i - (\underbrace{a}_{\text{intercept}} + \underbrace{b}_{\text{slope}} x_i).$$

```
l = lm(y ~ x)
y[39] - l$fitted.values[39]
```

```
       39
-4.390582
```

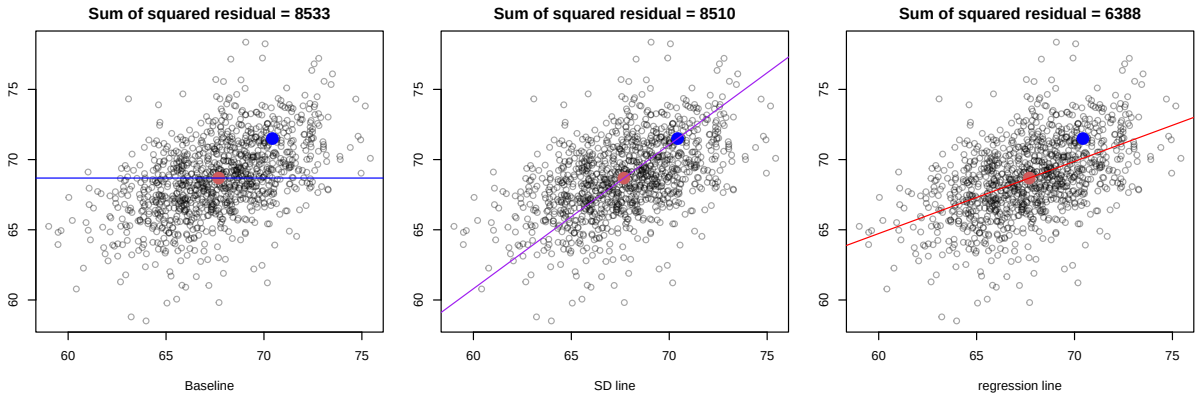
```
l$residuals[39]  # Or directly
```

```
       39
-4.390582
```

The regression line is the **best** (optimal) linear model - it provides the best fit to the data as the sum of the squared residuals $\sum_{i=1}^{n} e_i(a,b)^2$ is as small as it can be.

4.6.1 Optimality of regression line (not for assessment)

- We first consider a general line $y = \alpha + \beta x$ with intercept α and slope β.
- Given the data set $\{x_i, y_i\}, i = 1, \ldots, n$, a pair of variables (α, β) for defining a line, the residual is
$$e_i(\alpha, \beta) = y_i - (\alpha + \beta x_i).$$
so that the sum of squared residuals becomes
$$f(\alpha, \beta) = \sum_{i=1}^{n} e_i(\alpha, \beta)^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$



- Our goal is to find the intercept a and the slope b that minimises $f(\alpha, \beta)$:
$$f(a,b) \leq f(\alpha, \beta) \quad \text{for all} \quad \alpha, \beta$$

Derivation of optimality

How to find such a minimiser (a, b)? It needs to be a stationary point of the function f such that
$$\frac{\partial f}{\partial \alpha}(a,b) = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) \ = 0$$
and
$$\frac{\partial f}{\partial \beta}(a,b) = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = 0.$$
We use the **first equation** to find the **intercept**, $\frac{\partial f}{\partial \alpha}(a,b) = 0$ is equivalent to
$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(a + bx_i) = na + b\sum_{i=1}^{n} x_i$$

Dividing both sides by n, this gives

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = a + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + b\bar{x},$$

which leads to $a = \bar{y} - b\bar{x}$.

We can find the **slope** by substituting $a = \bar{y} - b\bar{x}$ into the **second equation**. This way, $\frac{\partial f}{\partial \beta}(a, b) = 0$ becomes

$$\sum_{i=1}^{n} [y_i - (\bar{y} - b\bar{x}) - bx_i]x_i = 0.$$

After rearrangement,

$$\sum_{i=1}^{n} (y_i - \bar{y})x_i = b\sum_{i=1}^{n} (x_i - \bar{x})x_i.$$

Because the sum of deviations is zero (topic 3 in week 3), we have $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, and hence

$$\sum_{i=1}^{n} (y_i - \bar{y})\bar{x} = 0 \quad \text{and} \quad \sum_{i=1}^{n} (x_i - \bar{x})\bar{x} = 0.$$

as $\bar{x}$ is a constant for all i.

$$LHS = \left(\sum_{i=1}^{n} (y_i - \bar{y})x_i\right) - \left(\sum_{i=1}^{n} (y_i - \bar{y})\bar{x}\right) = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$RHS = b\left(\sum_{i=1}^{n} (x_i - \bar{x})x_i\right) - b\left(\sum_{i=1}^{n} (x_i - \bar{x})\bar{x}\right) = b\sum_{i=1}^{n} (x_i - \bar{x})^2$$

By solving the second equation $\frac{\partial f}{\partial \beta}(a, b) = 0$, the slope is

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Recall that

- $SD_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}$
- $SD_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2}$
- $r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$

This gives exactly $b = r\frac{SD_y}{SD_x}$ as we claimed in the definition of the regression line. So that we know the regression line is indeed the best among all lines (linear functions) in the sense of sum of squared residuals.

4.6.2 Average of residual is zero

Given the regression line $y = a + bx$, where $a = \bar{y} - b\bar{x}$, the sum of residual

$$\sum_{i=1}^{n} e_i(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i) = \sum_{i=1}^{n}(y_i - (\bar{y} - b\bar{x}) - bx_i)$$

can be expressed as

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

Thus, **the mean (average) of residual is zero.**

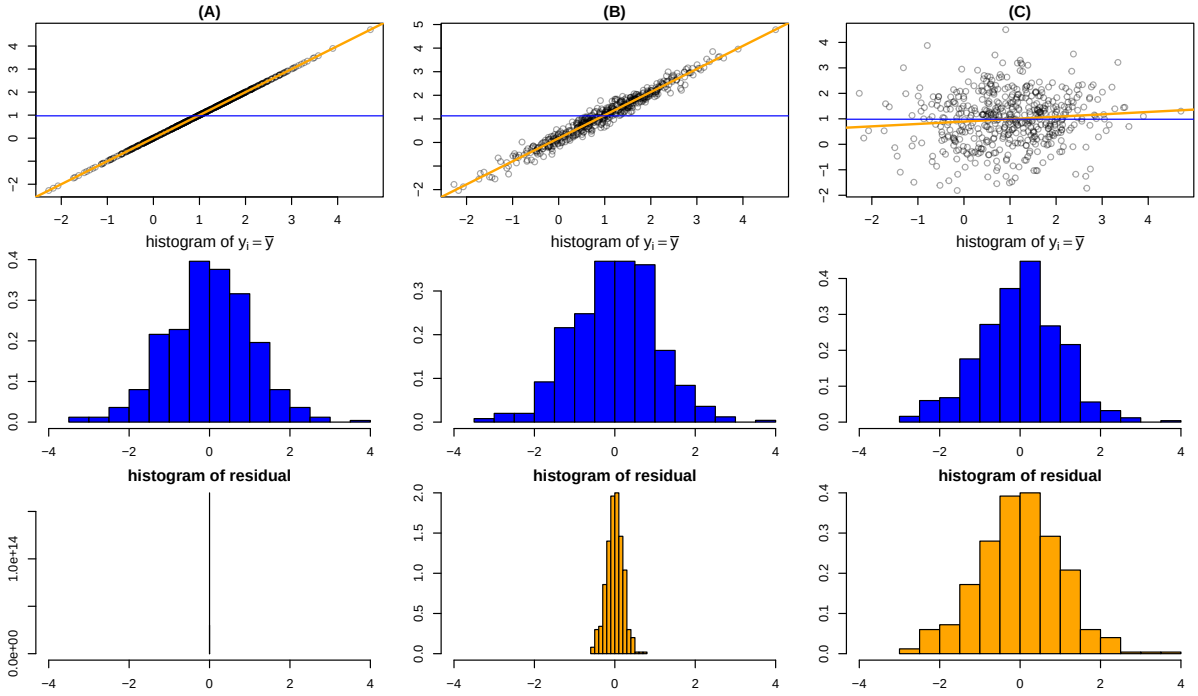
4.6.3 Summary of residual

Feature	Regression Line $y \sim x$ $(y = a + bx)$
Connects	$(\bar{x}, \bar{y})$ to $(\bar{x} + \text{SD}_x, \bar{y} + r\text{SD}_y)$
Slope (b)	$r\frac{\text{SD}_y}{\text{SD}_x}$
Intercept (a)	$\bar{y} - b\bar{x}$
Residual	$e_i = y_i - a - bx_i$

- $y = a + bx$ is the best line that minimises the sum of squared residuals $\sum_{i=1}^{n} e_i^2$.
- The average residual of the regression line is zero: $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$.

4.7 Coefficient of determination

How much variability of data y can be explained by the linear model?



Blue: baseline prediction/deviations in y, Orange: regression line/residuals

The degree to which variation in y is explained by a linear model depends on how closely the data points align with the line. When all variation is explained, the points fall exactly on the line, meaning 100% of y's variation is attributed to its linear relationship with x. If most variation is explained, the points deviate slightly from the line, but the residuals are small compared to the overall variability in y, indicating the model captures much of the y variation. However, when little variation is explained, the points scatter widely around the line, showing that the model fails to account for y's variability in relation to x. You can see how these relationships look in the different columns above.

4.7.1 Explaining variations

- The sum of squared residuals (or SSE for sum of squared errors)

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

  measures **variation in y left unexplained by the regression line**.

- Note that since $\sum_{i=1}^{n} e_i = 0$ (the sum or average of residual is zero), $\frac{1}{n-1}$SSE is the sample variance of the residual.
- In (A) SSE = 0, and there is no unexplained variation, whereas unexplained variation is small for (B), and large for (C).
- A quantitative measure of **the total amount of variation in observed y values** is given by the total sum of squares (sum of squared deviations about sample mean)

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

  measures variation in y left unexplained by the baseline prediction.
- Note that since $\sum_{i=1}^{n} y_i - \bar{y} = 0$ (the sum or average of deviation is zero), $\frac{1}{n-1}$SST is the sample variance of the dependent variable y.
- **SST $\geq$ SSE**. Why? The regression is optimal for sum of squared errors, so SSE (regression line) cannot be worse than SST (baseline).

4.7.2 Coefficient of determination

The ratio $\frac{\text{SSE}}{\text{SST}}$ is the proportion of total variation that cannot be explained by the simple linear regression model, and the coefficient of determination is

$$1 - \frac{\text{SSE}}{\text{SST}} = r^2$$

which is the **squared correlation coefficient** (a number between 0 and 1) giving the proportion of observed y variation explained by the model.

- The higher the value of r^2, the more successful is the simple linear regression model in explaining y variation.
- Note that if SSE = 0 as in case (A), then $r^2 = 1$.
- This can be verified using a, b, SDs, and r (see next)

4.7.3 Derivation of the coefficient of determination (not for assessment)

Recall that the coefficient of determination of the regression line is

$$1 - \frac{\text{SSE}}{\text{SST}}, \quad \text{SSE} = \sum_{i=1}^{n}(y_i - a - bx_i)^2, \quad \text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

where a and b are the intercept and the slope of the regression line, respectively. We want to show that the coefficient of determination can be indeed be written as squared correlation coefficient (r^2). The following steps demonstrate one of many ways (with a statistical interpretation) to show this.

Sum of squared differences between regression and baseline

We consider the sum of squared differences between the regression line and the baseline prediction:

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

We want to verify $\text{SST} = \text{SSR} + \text{SSE}$, so that **SSR represents the variability explained by the regression model**.

For each y_i, we can first consider the decomposition:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

where $\hat{y}_i$ is the predicted value from the regression, and $y_i - \hat{y}_i$ is the residual (error) for that observation. Taking the sum of squares of both sides, we have

$$(y_i - \bar{y})^2 = \left((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)\right)^2$$

and taking the summation, we get:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

which leads to

$$\text{SST} = \text{SSR} + \text{SSE} + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})\hat{e}_i$$

where the last term is zero, so we have

$$\text{SST} = \text{SSR} + \text{SSE}.$$

Correlation between the prediction of the regression line and the response variable

Then, we want to verify that

$$\text{cor}(\hat{y}, y)^2 = \frac{SSR}{SST} = 1 - \frac{\text{SSE}}{\text{SST}}$$

so that the correlation coefficient between the predicted values of the regression line $\hat{y}$ and the response variable y gives the coefficient of determination.

Since $\bar{y}$ is also the mean of $\hat{y}$, we have

$$\text{cor}(\hat{y}, y)^2 = \frac{\left(\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^n (\hat{y}_i - \bar{y})^2\right)\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)}$$

We also have

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \bar{y}) = \sum_{i=1}^n (\hat{y}_i - \bar{y})(\hat{y}_i - \hat{y}_i + y_i - \bar{y}) = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y}) e_i$$

Again, the last term is zero by step 1. Thus, we have

$$\text{cor}(\hat{y}, y)^2 = \frac{SSR^2}{SSR\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)} = \frac{SSR}{SST}.$$

Putting the results together

Since correlation coefficient is shift and scale invariant, we have

$$\text{cor}(x, y)^2 = \text{cor}(a + bx, y)^2 = \text{cor}(\hat{y}, y)^2,$$

which gives the result.

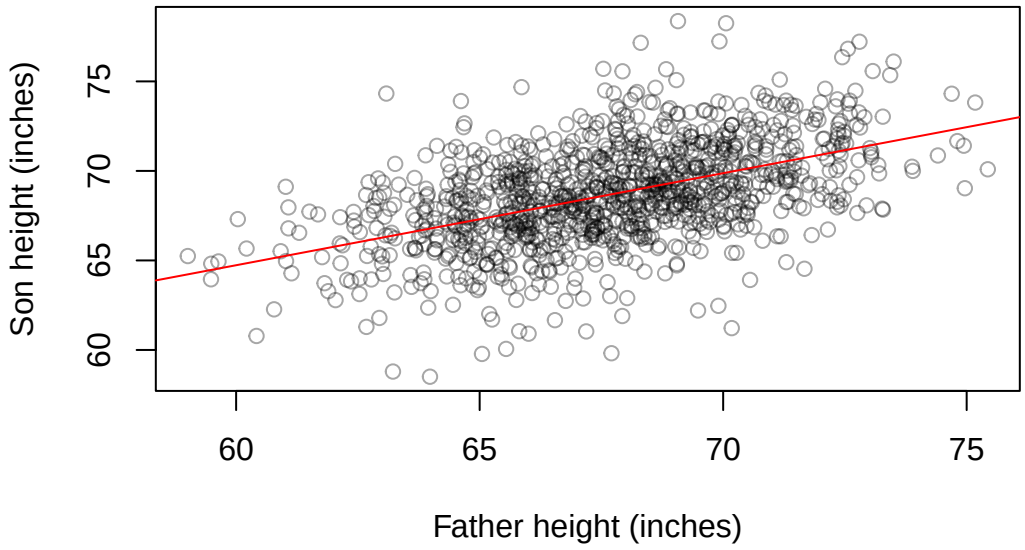
Example

```
cor(x, y)^2  # quick way
```

```
[1] 0.2513401
```

```
lm.fit <- lm(y ~ x)
SSE = sum(lm.fit$residuals^2)
SST = sum((y - mean(y))^2)
r2 = 1 - SSE/SST
```

```
[1] 0.2513401
```



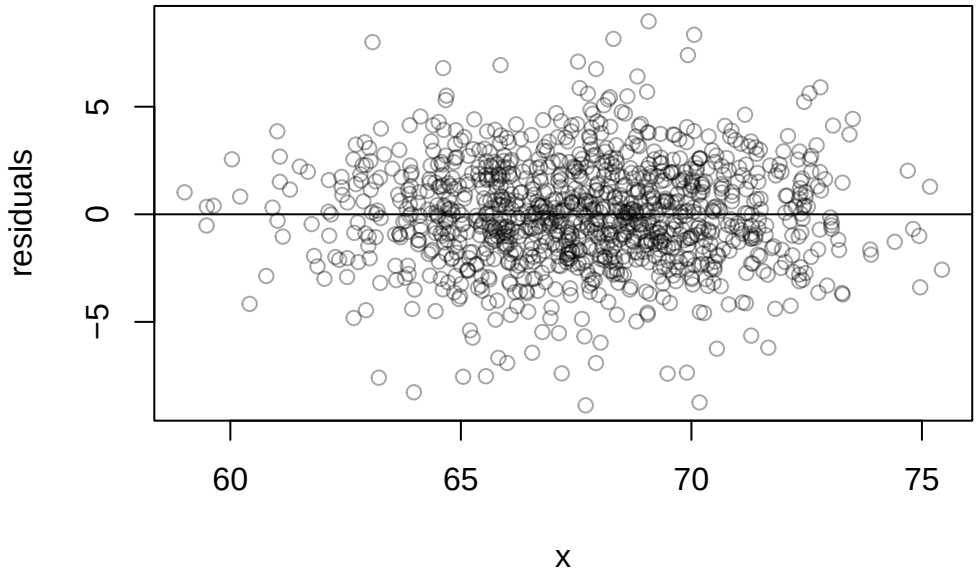
The coefficient of determination for Pearson's height data is 0.25, about 25% of the variations in son's height can be explained by the regression line.

4.8 Diagnostics

4.8.1 Residual Plot

- A residual plot graphs the residuals vs x.
- If the linear fit is appropriate for the data, it should show no pattern (random points around 0).
- By checking the patterns of the residuals, the residual plot is a diagnostic plot to check the appropriateness of a linear model.

```
plot(x, l$residuals, ylab = "residuals", col = adjustcolor("black", alpha.f = 0.35))
abline(h = 0)
```



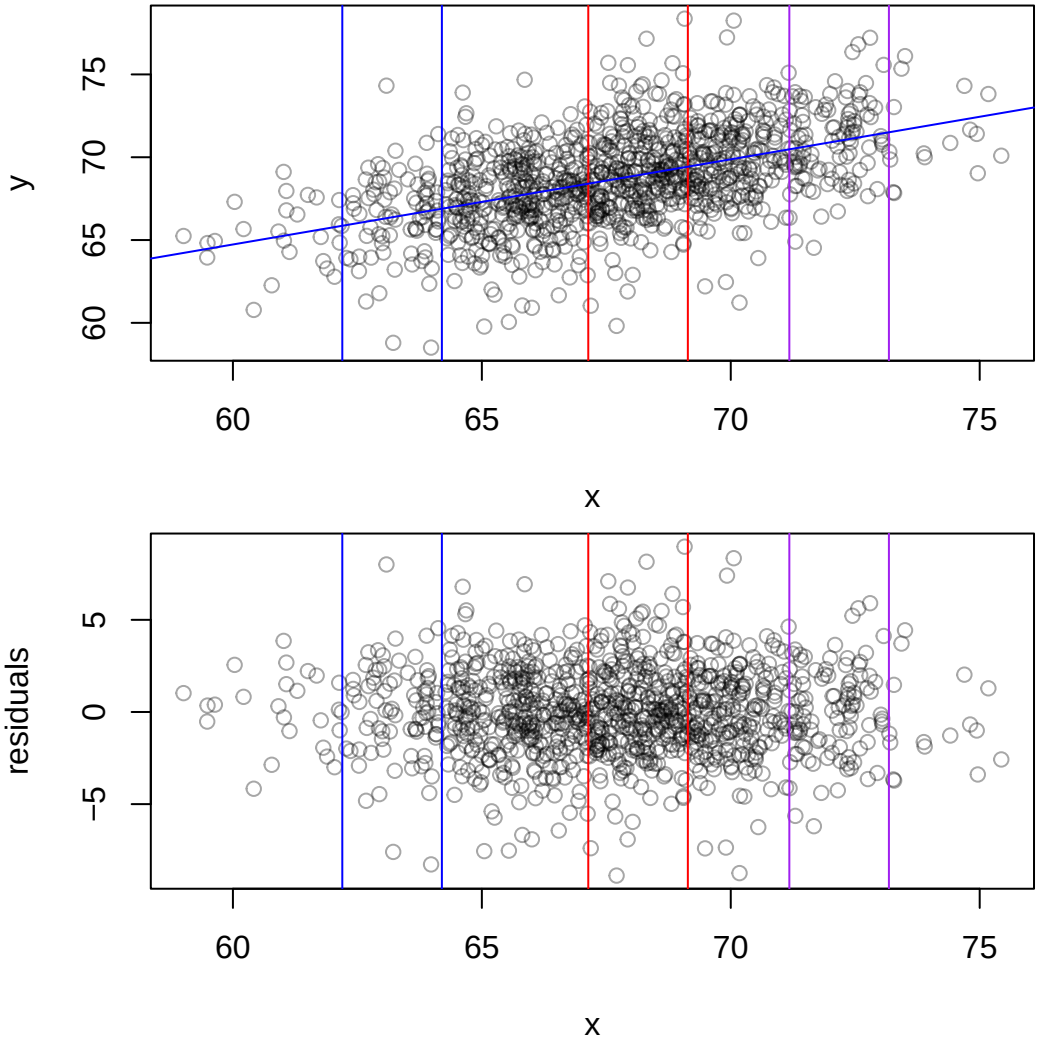
> **Note**
>
> Does this residual plot look random?

It looks reasonably random.

4.8.2 Homoscedasticity and Heteroscedasticity

In linear models and regression analysis generally, we need to check the homogeneity of the spread of the response variable (or the residuals). We can divide the scatter plot or the residual plot into vertical strips.

- If the vertical strips on the scatter plot show equal spread in the y direction, then the data is **homoscedastic**.
  - The regression line could be used for predictions.
- If the vertical strips don't show equal spread in the y direction, then the data is **heteroscedastic**.
  - The regression line should not be used for predictions.



> **Note**
>
> Is the Pearson's height data homoscedastic?

5 Probability

The probability of an event is a measure of the likelihood of that event occurring. Probability Theory is a set of mathematical tools which dates back centuries to casino type games. A modern mathematical theory was developed in the 1930s by the Russian mathematician A. N. Kolmogorov. (History video)

Probability Theory offers important modelling tools that make many problems easy. For example, the normal curve is dedrived from probability theory. Its interpretation such as the p-value is crucial for hypothesis testing, which is essential to scientific research.

Example: Why did the Chevalier lose money?

The Chevalier de Méré was a 17th century gambler, who played 2 games:

- Game A: Roll a die 4 times. Win = at least 1 "ace".
- Game B: Roll a pair of dice 24 times: Win = at least 1 "double-ace".
- Note: an "ace" means "1".

He reasoned:

Game	1 roll	# rolls	Win
A	P(1 Ace) = 1/6	4	P(at least 1 Ace) = 4x1/6 = 2/3
B	P(1 Double-Ace) = 1/36	24	P(at least 1 Double-Ace) = 24*1/36 = 2/3

Q: But he lost consistently in Game B. Why?

Example: Coin tossing during WWII

John Edmund Kerrich (1903–1985) was a mathematician noted for a series of experiments in probability. With a fellow internee Eric Christensen, Kerrich set up a sequence of experiments demonstrating the empirical validity of a number of fundamental laws of probability.

- They tossed a coin 10,000 times and counted the number of heads.
- They investigated tosses of a "biased coin", made from a wooden disk partly coated in lead.

In 1946 Kerrich published his finding in a monograph, An Experimental Introduction to the Theory of Probability.

Q: How many heads do you think he counted? What is the probability of getting a head on a fair coin?

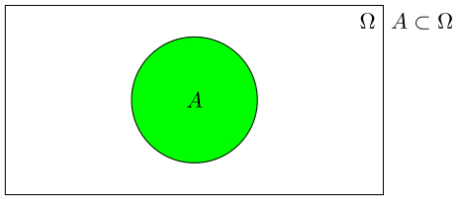
5.1 Definitions

The frequentist definition of **probability** (or chance) is the percentage of time a certain event is expected to happen, if the same process is repeated long-term (infinitely often). This differs from the Bayesian definition of probability which relates to the degree of belief that an event will occur (extension).

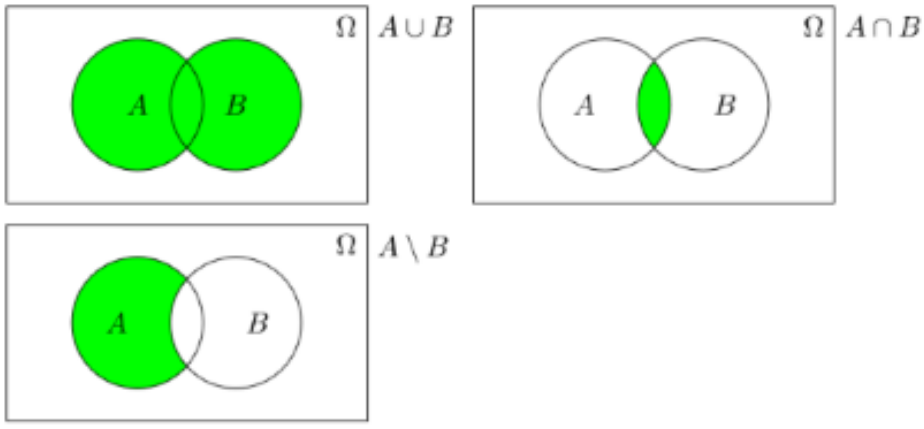
We will learn some of the definitions and properties of probability (mostly the frequentist definition) and a useful tool (box model) to understand the basics of probability for our modelling purpose.

5.1.1 Describing simple probability models

We can describe probability models using set notation and the Venn diagram.



Symbol	Name	Meaning
Ω	Sample Space	Everything that can occur
$A, B, ...$	Events	A subset of the sample space
$A \subset \Omega$	Belongs to	Event A belongs to sample space Ω
$\emptyset$	Empty Set	An event which cannot occur
A^c or A'	Complement	Everything not in A



If A and $B \subset \Omega$:

Symbol	Name	Meaning
$A \cup B$	Union	A or B or both occur
$A \cap B$	Intersection	Both A and B occur.
$A \setminus B$	Minus or Relative Complement	In A but not in B

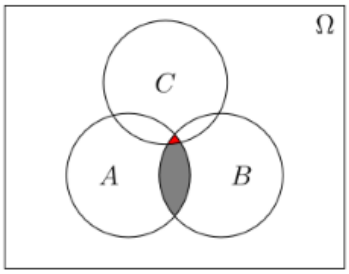
Probabilities are between 0 (impossible) and 1 (certain)

$$P(\text{Impossible event}) = 0$$

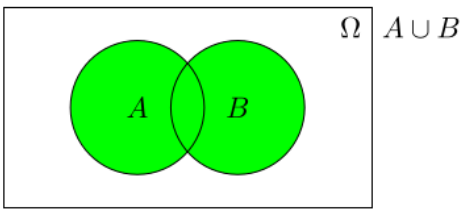
$$P(\text{Certain event}) = 1$$

For example, we randomly toss a coin (two possible events: Head or Tail), P(either a Head or a Tail) = 1. We randomly roll a die, P(with an outcome 0) = 0.

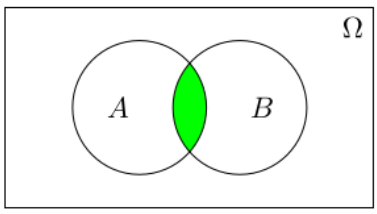
Exercise: Describe the grey and red regions in set notation.



5.1.2 De Morgan's law (not for assessment)



$A \cup B$



$A \cap B$

For any two events A and B:

- The complement of their union is the intersection of their complements:

$$(A \cup B)^c = A^c \cap B^c$$

  - not (A or B) = (not A) and (not B) = "white area" on the left

- The complement of their intersection is the union of their complements:

$$(A \cap B)^c = A^c \cup B^c$$

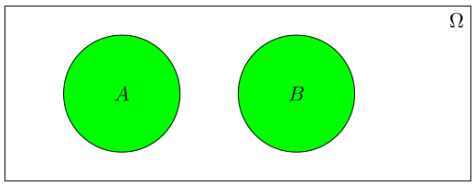
  - not (A and B) = (not A) or (not B) = "any white area" on the right

5.2 Properties

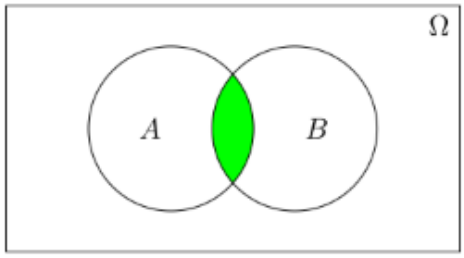
5.2.1 Complement, Mutual Exclusivity, and Independence

Complement: The probability that an event A does not occur is $\boxed{P(A) = 1 - P(A^c)}$. For example, we randomly toss a coin, P(Head) = 1 - P(Tail).

Mutually exclusive: Two events A and B are mutually exclusive if $\boxed{P(A \cap B) = 0}$.



Independence: Two events A and B are independent (the result of one does not affect the result of the other) if and only if $\boxed{P(A \cap B) = P(A)P(B)}$.



For example, $P(A) = 0.2, \quad P(B) = 0.5, \quad P(A \cap B) = P(A)P(B) = 0.1$.

Many of our assumptions are based on independence. It will simplify the analysis.

What's the difference between mutually exclusive and independence?

Term	Definition
Mutually exclusive	The occurrence of Event B prevents Event A occuring
Independence	The occurrence of Event B does not change the chance of Event A

5.2.2 Conditional probability

For events A and $B \subset \Omega$ where $P(B) \neq 0$, the conditional probability of A given B is

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

This leads to a second definition of independence: Two events A and B are independent if and only if

$$\boxed{P(A|B) = P(A)}$$

Conditional probability is the chance that a certain event (1) occurs, *given* another event (2) has occurred.

$$\text{P(Event 1|Event 2)}$$

5.2.3 Multiplication and Addition Rule

Consider two events A and B.

Multiplication rule: The probability that A and B occur is $P(A)$ **multiplied** by the conditional probability $P(B|A)$.

$$\boxed{P(A \cap B) = P(A)P(B|A)}$$

Addition rule (Union rule): The probability at least one of A and B occurs is $P(A)$ **plus** $P(B)$ **minus** the probability that both events occur.

$$\boxed{P(A \cup B) = P(A) + P(B) – P(A \cap B)}$$

Imagine the area of the union of A and B.

What	When	Formula	Condition
Addition Rule	P(At least one of two events occurs)	$P(A) + P(B) – P(A \cap B)$	Always
		$P(A) + P(B)$	If mutually exclusive
—	—	—	—
Multiplication Rule	P(Both events occur)	$P(A)P(B\|A)$	Always
		$P(A)P(B)$	If independent

5.3 The prosecutor's fallacy (reading material, not for assessment)

The **prosecutor's fallacy** is a mistake in statistical thinking, whereby it is assumed that the probability of a random match is equal to the probability that the defendant is innocent. It has been used by the prosecution to argue for the guilt of a defendant during famous criminal trials. It can also be used by defense lawyers to argue for the innocence of their client.

Here is an example. Suppose there are about 5 million people living in Sydney. A murder occurs with DNA left on the weapon. A person matching the DNA is arrested. The chance of a DNA match is 1 in 500,000 (very small). Hence, the chance that the arrested person is guilty is very high. We fill out the following table:

	DNA Match	DNA doesn't match
Guilty	1	0
Innocent	9	4,999,990

Note:

- Only 1 person is guilty and has a DNA match.
- No-one (0) is guilty and doesn't match DNA.
- If 1 in 500,000 people matches DNA, then for a city size about 5 million, we expect 10 people to match (which is 1 guilty person and 9 innocent people).
- This leaves almost 5 million innocent people (4,999,990) that don't match.

Hence, the chance that DNA matches, given innocent person is

$$\text{P(DNA Match|Innocent)} = \frac{9}{4,999,999}$$

which is tiny. But the chance that the person is innocent, given a DNA match is

$$\text{P(Innocent|DNA Match)} = \frac{9}{10}$$

which is very high.

Note P(DNA Match|Innocent) $\neq$ P(Innocent|DNA Match). So for any person with DNA match, we can't say P(Guilty|DNA Match) is high. In conclusion,

$$\text{P(Guilty|DNA Match)} = 1-\text{P(Innocent|DNA Match)} \neq 1-\text{P(DNA Match|Innocent)} = 1-\frac{9}{4,999,999}$$

5.3.1 OJ Simpson

Orenthal James Simpson ("OJ") (July 1947 - April 2024) was a National Football League (NFL) player and actor.

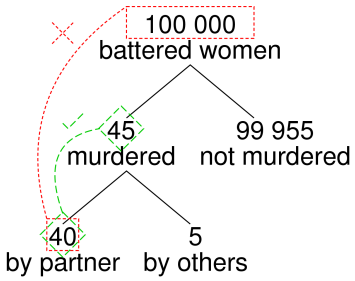
- In 1994, OJ was tried for the murders of his former wife Nicole Brown Simpson and her friend Ron Goldman.
- In 1995, he was **acquitted**.
- In 1997, a civil court awarded a $33.5 million judgment **against** Simpson for the victims' wrongful deaths.

Mistake 1 in the OJ case (article)

- **Fact**: The crime scene blood matched Simpson's blood, with characteristics shared by 1 in 400 people.
- **Mistake**: The prosecutor tried to imply that the probability that OJ was innocent given that the blood type matched was 1/400. But in LA there were more than 10,000 people whose blood type matched the one in the crime scene.

Mistake 2 in the OJ case (article)

- **Fact**: The prosecutor presented evidence that Simpson had been violent toward his wife.
- **Mistake**: The defense argued that only 1 of 2500 of men who beat their domestic partners go on to murder them, so that any history of Simpson being violent toward his wife was irrelevant. It was ignored that Simpson's wife had not only been subjected to domestic violence, but was also murdered. See the following figure.
  - **Defence lawyer:** P(murder|domestic violence) = $\frac{1}{2500}$
  - **Prosecution:** P(murder|death & previous domestic violence) = $\frac{8}{9}$



5.3.2 Sally Clark

Sally Clark (August 1964 15 March 2007) was an English solicitor.

- In December 1996, her 1st baby son Christopher died suddenly at home.

- In January 1998, her 2nd son Harry died in a similar way.
- In February 1998, Sally was arrested, and in November 1999 was **convicted** of both their murders.
- Paediatrician Professor Sir Roy Meadow testified that the chance of two children from an affluent family suffering from sudden infant death syndrome (SIDS) was 1 in 73 million.

In January 2003, a 2nd appeal was successful in **overturning** the conviction. It was discovered that the prosecution forensic pathologist Dr Alan Williams, who had examined both of the babies, had incompetently failed to disclose microbiological reports that suggested the second of her sons had died of natural causes.

Mistakes in the Sally Clark case

- **Mistake 1 (Dependent probability)**: The chance of 2 SIDS deaths in an affluent family was claimed to be 1 in 73 million (tiny). This figure was improperly derived by ignoring the association between events. So The Royal Statistical Society in the UK issued a public statement pointing out the statistical invalidity of this number.

  $$\text{P(2 child deaths)} = \text{P(1st child death)} \times \text{P(2nd child death} \mid \text{1st child death)} \neq \frac{1}{8500} \times \frac{1}{8500} \approx \frac{1}{73M}$$

- **Mistake 2 (Prosecutor's Fallacy)**: The chance Sally Clark was guilty was said to be very high.

  $$\text{P(Guilty}|\text{2 deaths)} = 1 - \text{P(Innocent}|\text{2 deaths)} \neq 1 - \text{P(2 deaths}|\text{Innocent)} = 1 - \text{P(2 SIDS)} = 1 - \frac{1}{73M}$$

5.4 Simulation and Sample with/without replacement

We consider the **"classical" probability**, where the sample space Ω consist of a finite, known number of equally likely outcomes (e.g., coins, dice, cards). The probability of an event $A \subset \Omega$ occurring is

$$\boxed{P(A) = \frac{\text{Number of ways } A \text{ can occur}}{\text{Total number of possible outcomes in } \Omega}}$$

For example, suppose we want to know the probability of getting an even number when we roll a fair die. There are 6 equally likely possible outcomes,

$$\Omega = \{\boxdot \boxdot \boxdot \boxdot \boxdot \boxdot\}$$

of which 3 are even. Therefore, the probability of rolling an even number is

$$P(\text{even number}) = \frac{P(\text{number of outcomes that are even})}{P(\text{number of possible outcomes})} = \frac{3}{6}.$$

For simple problems, a good start is to enumerate all the possible outcomes by either writing a list of all outcomes and count the outcomes of interest, or drawing a tree. Here we will focus on simulation techniques and advanced counting techniques.

5.4.1 A simple box model

> **Note**
>
> **The box model:** Classical probability can be explained using the box model. The box model represents the sample space Ω as a collection of tickets, where each ticket corresponds to a possible outcome.

In a box model, there are N tickets in a box, and we draw m tickets from the box.

- For example, three rolls of a fair die can be modeled as $m = 3$ draws from the box

1	2	3	4	5	6

  we have to place the ticket back in the box after each draw, so the outcome of one die roll does not affect the outcome of another. In other words, the $m = 3$ draws are made with replacement.

- In other situations, the draws are made **without replacement**. For example, consider drawing four cards from a standard deck of 52 cards (without putting the drawn cards back).

5.4.2 Simulation (in R)

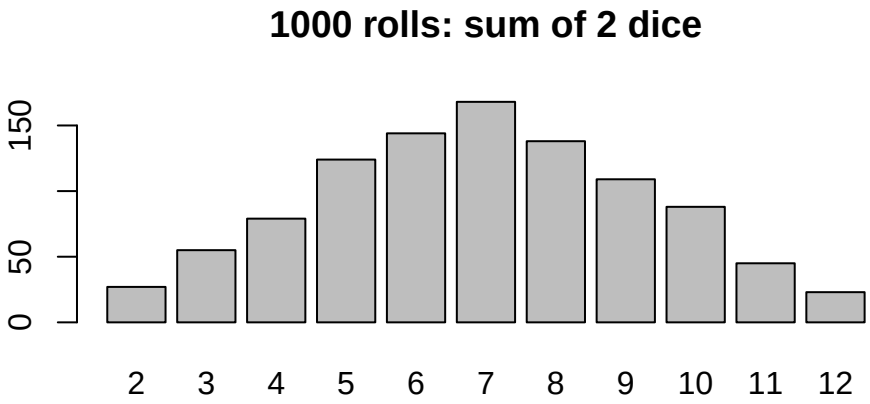
Use R and simulate throwing 2 dice x times and record the findings.

```
set.seed(23)  # set the random seed
totals = sample(1:6, 1000, rep = T) + sample(1:6, 1000, rep = T)
table(totals)
```

```
totals
  2   3   4   5   6   7   8   9  10  11  12
 27  55  79 124 144 168 138 109  88  45  23
```

- We set the random seed so this can be reproduced
- Sample from 1,2,3,4,5,6 (a die) with equal probability
- Sample 1000 times
- Sample **with replacement** using `rep=T` (independent experiments)

```
barplot(table(totals), main = "1000 rolls: sum of 2 dice")
```



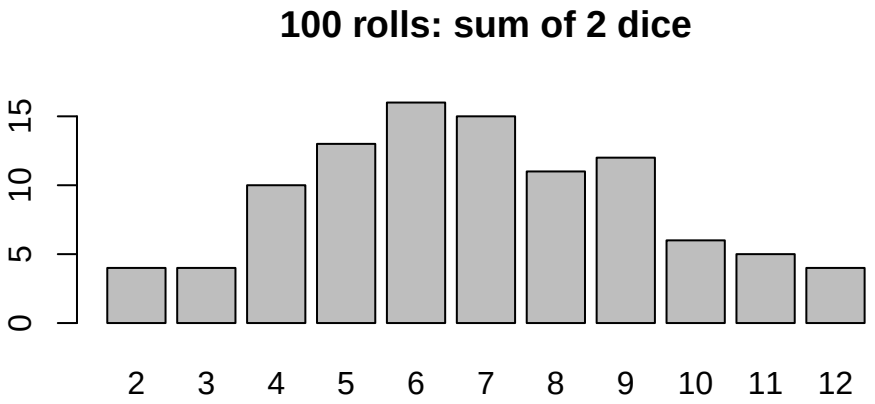
So the (simulated) chance of getting a total of 6 is $144/1000 = 0.144$, which is very close to the exact answer of $5/36 = 0.139$.

What will happen if we sample **without replacement** (without `rep=T`)?

- This implies dependent experiments - the next outcome depends on previous ones
- See code demo

More advanced code

```
set.seed(1)
Roll1Die = function(n) sample(1:6, n, rep = T)  #Creates a function to roll die
roll1 = NULL  #Initialise variable.
roll2 = NULL
for (i in 1:100) {
    roll1[i] = Roll1Die(1)
    roll2[i] = Roll1Die(1)
}
barplot(table(roll1 + roll2), main = "100 rolls: sum of 2 dice")
```

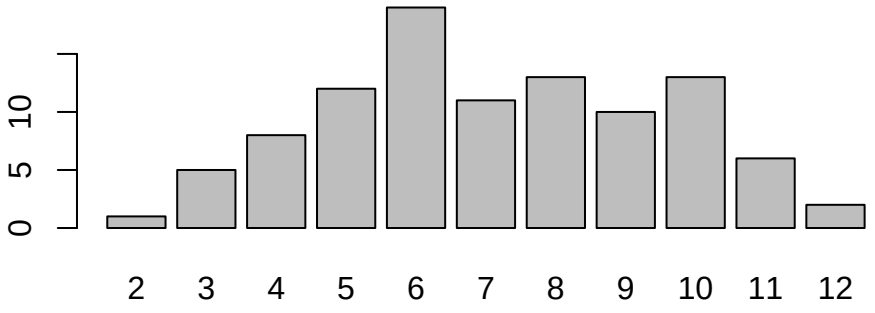


More advanced code [simulating totals directly]

```
set.seed(1)
s1 = sample(2:12, size = 100, replace = TRUE, prob = table(outer(1:6, 1:6, "+"))/36)
table(s1)
```

```
s1
 2  3  4  5  6  7  8  9 10 11 12
 1  5  8 12 19 11 13 10 13  6  2
```

```
barplot(table(s1))
```

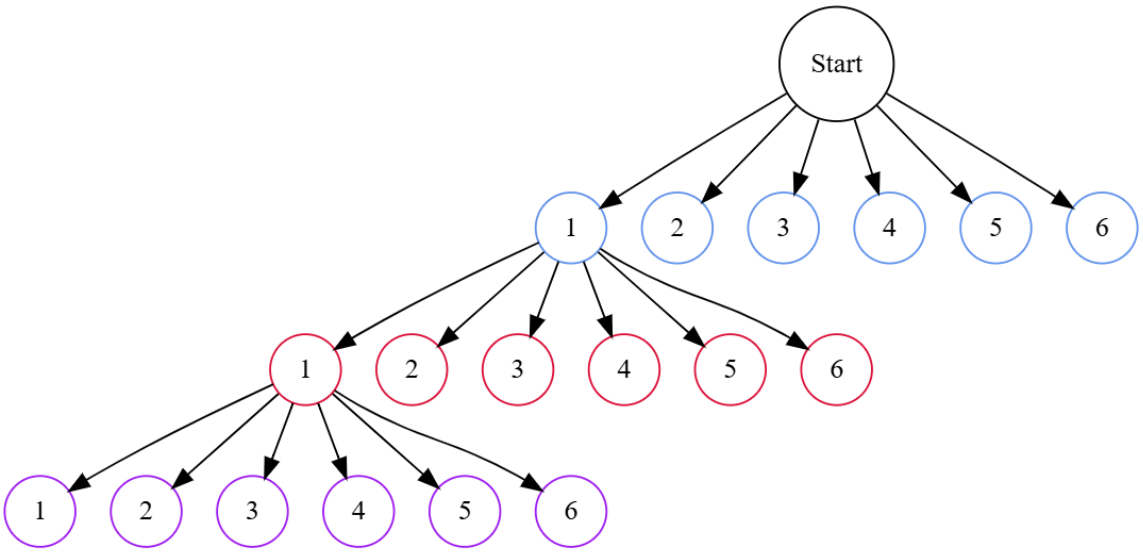


Example: Three dice are thrown. What is the chance of getting a total equal to 6?

Method 1: Write a list manually

- Total number of outcomes are 6x6x6 = 216
- The outcomes where the total is equal to 6 are: (1,1,4) (1,2,3) (1,3,2) (1,4,1) (2,1,3) (2,2,2) (2,3,1) (3,1,2) (3,2,1), (4,1,1)
- So exact chance of getting total of 6 is 10/216 (approx 0.046).

Method 2: Summarise in a tree diagram

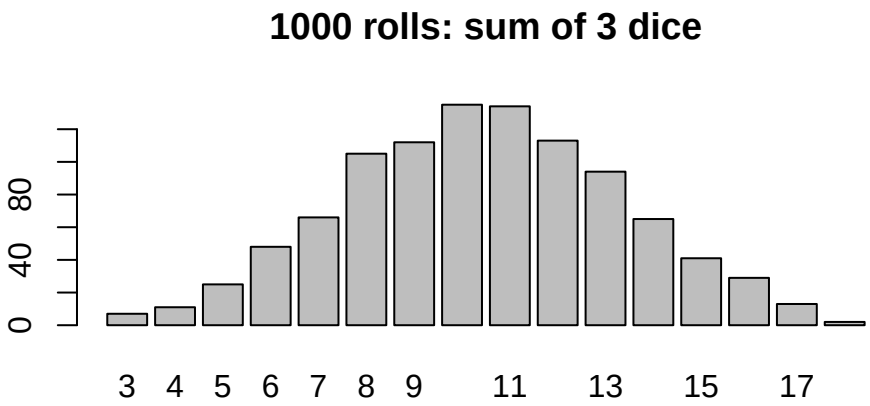


Method 3: Simulate in R

```
set.seed(23)
totals = sample(1:6, 1000, rep = T) + sample(1:6, 1000, rep = T) + sample(1:6, 1000,
    rep = T)
table(totals)/1000
```

```
totals
    3     4     5     6     7     8     9    10    11    12    13    14    15
0.007 0.011 0.025 0.048 0.066 0.105 0.112 0.135 0.134 0.113 0.094 0.065 0.041
   16    17    18
0.029 0.013 0.002
```

```r
barplot(table(totals), main = "1000 rolls: sum of 3 dice")
```



Example 2: Why did the Chevalier lose money? What is the chance of winning?

- Game A: Roll a die 4 times. Win = at least 1 "ace".
- Game B: Roll a pair of dice 24 times: Win = at least 1 "double-ace".
- Note: an "ace" means "1".

Method 3: Simulate in R (using a function)

```r
gameA <- function() {
    rolls <- sample(1:6, size = 4, replace = TRUE)
    condition <- sum(rolls == 1) > 0
    return(condition)
}
simsA <- replicate(1e+05, gameA())
sum(simsA)/length(simsA)
```

```
[1] 0.51543
```

```r
gameB <- function() {
    first.die <- sample(1:6, size = 24, replace = TRUE)
    second.die <- sample(1:6, size = 24, replace = TRUE)
    condition <- sum((first.die == second.die) & (first.die == 1)) > 0
    return(condition)
}
simsB <- replicate(1e+05, gameB())
sum(simsB)/length(simsB)
```

```
[1] 0.48979
```

Indeed, Game A is better.

5.4.3 Sample without replacement (using R)

Example: A company has 10,000 male employees and 11,000 female employess. A representative committee is created by randomly picking 10 employees. What is chance that more than 75% in the committee are male?

```
set.seed(1)
committee <- function() {
    committee <- sample(c(rep(1, 10000), rep(0, 11000)), size = 10, replace = FALSE)
    condition <- mean(committee) > 0.75
    return(condition)
}
sim <- replicate(10000, committee())
mean(sim)
```

```
[1] 0.0418
```

5.5 Factorial and combination (reading material, not for assessment)

5.5.1 Multiplication Principle of Counting

If a task can be performed in n_1 ways, and for each of these ways, a second task can be performed in n_2 ways, then the two tasks can be performed together in a total of $n_1 \times n_2$ ways. If there are k tasks in such a sequence, then k tasks can be performed together in a total of $n_1 \times n_2 \times \cdots n_k$ ways.

For example, there are a total number $6 \times 6 \times 6 = 216$ outcomes in rolling three dice.

5.5.2 Factorial

How many ways to arrange a deck of 52 cards?

We can use the multiplication principle to determine the number of ways to arrange the deck.

- The first card can be any one of 52 cards.
- No matter which one it is, the second card can be any one of the remaining 51 cards. So there are 52×51 ways to choose the first 2 cards.
- For every one of these 52×51 ways, there are 50 remaining cards to choose as the third card, which makes $52 \times 51 \times 50$ ways to choose the first 3 cards.
- And so on. By the time we get to the last card in the deck, there is only 1 card left. So there are $52 \times 51 \times 50 \times \cdots \times 1$ ways to arrange the 52 cards in a deck.

The quantity $n!$ (pronounced: "n factorial") is defined as

$$n! = n \times (n-1) \times \cdots \times 1.$$

It represents the number of ways to arrange n objects.

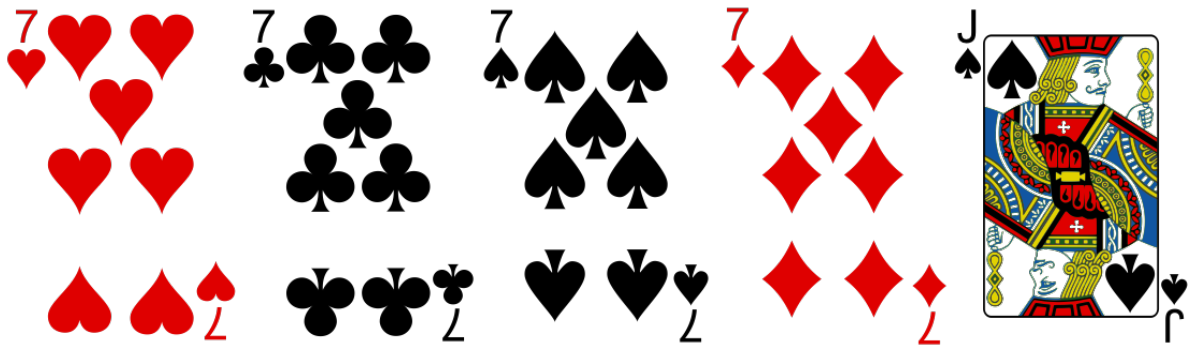
Example

A deck of 52 cards is shuffled thoroughly. What is the probability that the four aces are all next to each other? (Hint: First, count the number of positions that the block of four aces can go, then multiply this by the number of ways of ordering the four aces.)

There are 52! ways to order a deck of 52 cards (total number of outcomes in Ω)

How many possible ways to have four aces next to each other?

Note: A deck of cards has 13 ranks (ace, king, queen, jack, 10, ..., 2) and 4 suits (spades, clubs, hearts, and diamonds).



- Step 1: Consider the block of four aces (next to each other) as a single block
  - then we have 48 other cards plus this one block, making a total of $48+1 = 49$ units to arrange.
- Step 2: there are 49! ways to arrange these 49 units.
- Step 3: Calculate the number of ways to arrange the aces (with different suits) within the block.
  - the four aces can be arranged among themselves in 4! ways.
- Step 4: Calculate the probability

$$\frac{\text{Number of ways } A \text{ can occur}}{\text{Total number of possible outcomes in } \Omega} = \frac{49! \cdot 4!}{52!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{52 \cdot 51 \cdot 50} = \frac{1}{5525}$$

Example

If a five-letter word (in English) is formed at random (meaning that all sequences of five letters are equally likely), what is the probability that no letter occurs more than once?

Total number of outcomes in Ω: 26^5

Number of ways no letter occurs more than once: $26 \cdot 25 \cdot 24 \cdot 23 \cdot 22$ (taking out a letter from the box once it's been used)

So the probability is

$$P = \frac{26 \cdot 25 \cdot 24 \cdot 23 \cdot 22}{26^5} \approx 0.6588$$

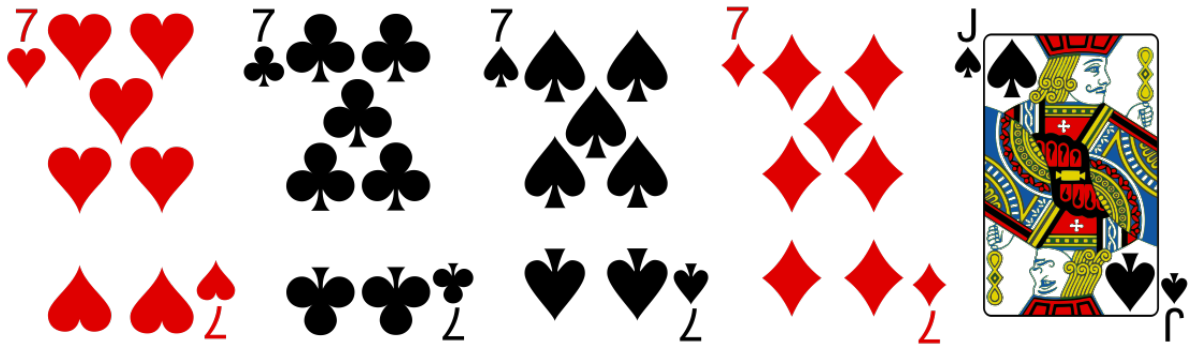
5.5.3 Combination

Factorials help us count how many ways we can arrange a group when order matters. However, when order doesn't matter, we use combinations to find the number of ways to choose a subset of specific outcomes from a larger group.

Example

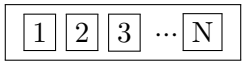
One of the most coveted hands in poker is a four-of-a-kind, which is when the hand contains all four cards of a particular rank. For example, the hand below is an example of a four-of-a-kind, since it contains all four 7s in the deck. (The last card, called the "kicker", can be any other card.)

Note: the order of the cards in the hand does not matter.



What is the probability of a four-of-a-kind?

If drawing **without replacement**, the number of ways to draw k tickets from the box



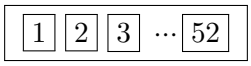
is

$$\underbrace{N \times (N-1) \times \cdots \times (N-k+1)}_{k \text{ terms}} = \frac{N!}{(N-k)!}$$

since the number of tickets remaining in the box decreases by 1 on each draw.

How many possible poker hands are there?

If we assign a number 1 to 52 to each card in a standard playing deck, then a poker hand can be modeled as $k = 5$ draws, without replacement, from the box



Number of possible **ordered** poker hands: $\frac{52!}{(52-5)!} = 52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 > 300 \times 10^6$.

Note that the order of cards in a hand matters here. We count not only how many hands of cards, but also how many **ordered** hands.

Two hands formed by the same set of cards, but with different orders, for example

$$\{2\spadesuit, 3\clubsuit, \text{Ace}\clubsuit, 9\heartsuit, 5\diamondsuit\}$$

and

$$\{9\heartsuit, 3\clubsuit, 5\diamondsuit, 2\spadesuit, \text{Ace}\clubsuit\}$$

are considered as two different hands by directly applying factorials.

The factorial considers the different orders in which the cards might be drawn.

How many of these possible ordered outcomes result in a four-of-a-kind?

Let's start by assuming that the first four cards in the hand are the four-of-a-kind and the last card is the kicker.

- The first card can be any one of the 52 cards.
- Once we have chosen the first card, the rank of the four-of-a-kind is determined. The second card must be one of the 3 remaining cards of the same rank.
- The third card must be one of the 2 remaining cards of that rank.
- The fourth card must be the 1 remaining card of that rank.
- The last card, the kicker, is one of the other 48 cards in the deck.

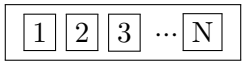
We assume the kicker is the last card in the hand. But the kicker can be in any one of 5 positions. So we need to multiply everything by 5 in the end.

There are $(52 \cdot 3 \cdot 2 \cdot 1 \cdot 48) \cdot 5$ outcomes, the chance of getting a four-of-a-kind is

$$\frac{(52 \cdot 3 \cdot 2 \cdot 1 \cdot 48) \cdot 5}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} \approx 0.00024.$$

Combinations (order doesn't matter)

The previous calculation was complicated because we had to consider the different orders in which the cards might be drawn. It is often easier to ignore the order when counting outcomes. The number of ways to draw k tickets from the box



when the order doesn't matter, is symbolized $\binom{N}{k}$ (pronounced: "N choose k") and is equal to

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

$k!$ is the number of ways of ordering the same set of k objects.

Revisit the probability of a four-of-a-kind using combinations.

If we ignore the order of the cards in the hand, there are $\binom{52}{5} = 2,598,960$ possible poker hands.

```
choose(52, 5)
```

```
[1] 2598960
```

Notice this number is much smaller than the 300+ million ordered poker hands. That is because when order matters, each distinct (unordered) poker hand gets counted 5! = 120 times, once for each possible way of reordering the 5 cards in the hand.

How many "unordered" four-of-a-kind hands are there?

- Any one of the 13 ranks (Ace through King) could be the rank for the four-of-a-kind.
- Once we have chosen the rank, it completely determines 4 of the 5 cards in the four-of-a-kind. There are only 13 ways to include all 4 cards of a given rank.
- All that's left is the kicker, which can be any one of the remaining 48 cards.

So when we ignore order, there are $13 \times 48 = 624$ ways to get a four-of-a-kind.

The chance is $624/2,598,960 \approx 0.00024$

6 The Box Model

A **model** is a representation of something which is **simpler** but at the same time captures the **key features** of the original. Data obtained "in real life" is generated (in general) by quite complicated processes. **Statistical models** are models for data-generating processes. They are much simpler than the "real" data-generating process but hopefully capture the key features, at least in terms of the **random variability** of the data.

The **box model** (which was introduced before) is a very simple statistical model for data generation. Recall the key features of a box model:

- A collection of N objects, e.g. tickets, balls is imagined "in a box".
- Each object bears a number.
- A **random sample** of a certain number n of the objects is taken.
- The sampling may be **with** or **without** replacement.

Random samples

In this context, when we say "a random sample is taken", it means a sample of the appropriate size is taken from a box model in such a way that **each possible sample** is equally likely.

We will use box model to build intuiation about random samples and some key properties. Those properties will be used in this unit and also generalised to random variables in later studies.

6.1 Random draws

6.1.1 Single random draws (samples of size $n = 1$)

If a single draw is taken, then each object in the "box" has an equal chance of being picked. If we *completely know* the contents of the box, we can write down the chance of each possible value.

We let X denote the **random draw**:

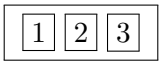
- This represents the "value we might get"
- X can take different values with different probabilities/chances.

The **distribution** of X is a **table** with two "columns":

- Each possible value x that X can take (note the capitalisation!) *and*
- The corresponding probability/chance of that value.

Simple examples (Box 1)

For example, suppose X is a random draw from the following box (box 1):



There are then three possible tickets: $\boxed{1}$, $\boxed{2}$ and $\boxed{3}$ and each has (equal) chance of $\frac{1}{3}$ of being picked, so:

$$P(X = 1) = P(X = 2) = P(X = 3) = \frac{1}{3}\,.$$

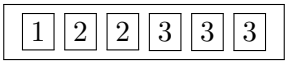
Here we write $P(\cdot)$ to denote the "probability" or "chance" of each event. The distribution of X is

x	1	2	3
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

6.1.2 Non-equal chances (Box 2)

We can have box models where the different possible *values* are not necessarily equally likely.

For the box (box 2)



if each "ticket" is equally likely, we have

$$P(X = 1) = \frac{1}{6}\,, \quad P(X = 2) = \frac{2}{6} = \frac{1}{3}\,, \quad P(X = 3) = \frac{3}{6} = \frac{1}{2}\,.$$

X then has distribution

x	1	2	3
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

Larger box example

Consider the box defined by the file `y.dat` in the R code below:

```
y = scan("y.dat")
y
```

```
  [1]  3  4  5  6  7  8  4  5  6  7  8  9  5  6  7  8  9 10  6  7  8  9 10 11  7  8
↪  9 10 11 12
 [31]  8  9 10 11 12 13  4  5  6  7  8  9  5  6  7  8  9 10  6  7  8  9 10 11  7  8
↪  9 10 11 12
 [61]  8  9 10 11 12 13  9 10 11 12 13 14  5  6  7  8  9 10  6  7  8  9 10 11  7  8
↪  9 10 11 12
 [91]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15  6  7  8  9 10 11  7  8
↪  9 10 11 12
[121]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12 13 14 15 16  7  8
↪  9 10 11 12
[151]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12 13 14 15 16 12 13
↪  14 15 16 17
[181]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12 13 14 15 16 12 13
↪  14 15 16 17
[211] 13 14 15 16 17 18
```

Q: What is the chance that a single draw from this is less than 8?

```
table(y)  # note: first two rows below are only labels: the 'real' output is the
↪  third line
```

```
y
 3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
 1  3  6 10 15 21 25 27 27 25 21 15 10  6  3  1
```

```
sum(table(y))  # gives total freq, i.e. size of the box
```

```
[1] 216
```

```
length(y)
```

```
[1] 216
```

```
sum(y < 8)  # the vector 'y<8' is of length 216, with TRUE=1 and FALSE=0 if each
↪  value <8 or >=8
```

```
[1] 35
```

```
sum(y < 8)/length(y)
```

```
[1] 0.162037
```

```
mean(y < 8)  # mean of a vector of 0's and 1's is the *proportion* of 1's
```

```
[1] 0.162037
```

- The chance of drawing a value less than 8 is $\frac{35}{216} \approx 16\%$.
- Note: $35 = 1 + 3 + 6 + 10 + 15$ (the frequencies of 3, 4, 5, 6 and 7 respectively).

6.1.3 Histogram, normal curve

In some situations, we may not know the *exact* contents of the box, but we might have access to some approximation. For example, what if the histogram of the box has a normal shape? In that case, knowing only the mean and SD of the box, we can approximate *proportions*, and hence chances of getting different values.

Firstly note the mean and SD for our example `y`:

```
mn.y = mean(y)
mn.y
```

```
[1] 10.5
```

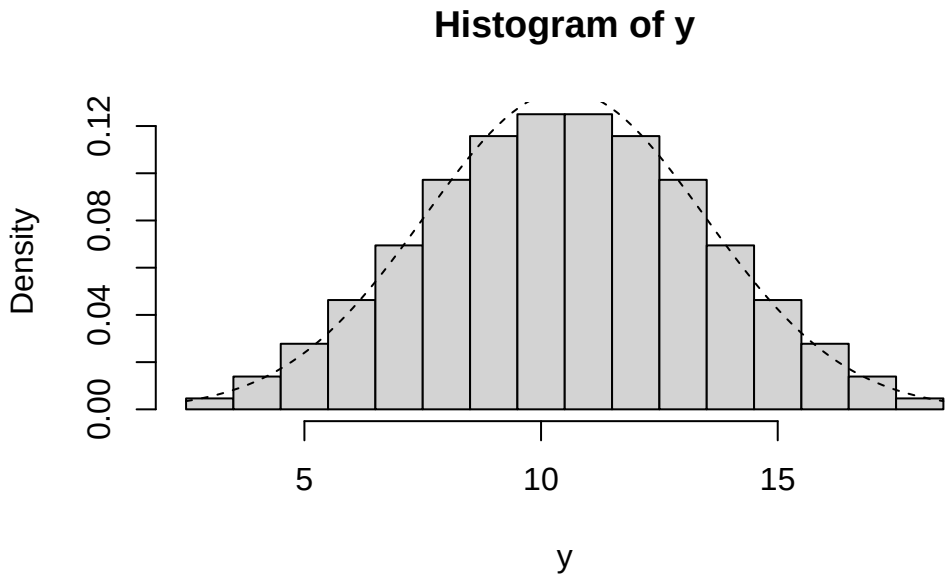
```
SD.y = sqrt(mean((y - mn.y)^2))
SD.y
```

```
[1] 2.95804
```

```
br = (2:18) + 0.5
br  # this gives rectangles centred on each integer 3,4,...,18
```

```
 [1]  2.5  3.5  4.5  5.5  6.5  7.5  8.5  9.5 10.5 11.5 12.5 13.5 14.5 15.5 16.5 17.5 18.5
```

```
hist(y, breaks = br, pr = T)
curve(dnorm(x, mn.y, SD.y), add = T, lty = 2)  # lty=2 gives a dashed line
```



6.1.4 Normal approximation

We can find the "area" to the left of 8, for a normal curve with the same mean and SD:

```
pnorm(8, mn.y, SD.y)  # not a bad approximation, but a bit big
```

```
[1] 0.1990124
```

Compare this to the "true" value of 16%

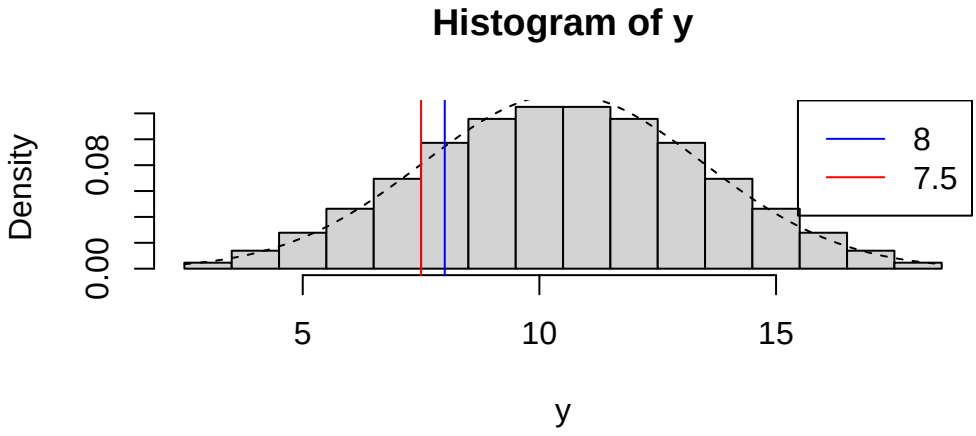
Non-examinable: Note that a *better* approximation can be obtained if we get the area to the left of 7.5:

```
pnorm(7.5, mn.y, SD.y)  # much closer to the true value!
```

```
[1] 0.1552472
```

This works because the values in the box are whole numbers, and the area under the rectangles we want is actually to the left of 7.5 (see the histogram repeated on the next slide):

```
hist(y, breaks = br, pr = T)
curve(dnorm(x, mn.y, SD.y), add = T, lty = 2)  # lty=2 gives a dashed line
abline(v = 8, col = "blue")
abline(v = 7.5, col = "red")
legend("topright", leg = c("8", "7.5"), lty = c(1, 1), col = c("blue", "red"))
```



6.1.5 New interpretation of mean and SD of box

When we are taking a random draw X from a box, we see that the mean and SD of the box have a new, special interpretation.

We call the mean of the box the **expected value** of the draw:

- We write this as $E(X)$.

We call the SD of the box the **standard error** of the random draw:

- We write this as $SE(X)$.

We can treat the box that we are drawing from as a population, which includes every possible outcome and we will draw a sample from it.

Random draw = Expected value + Chance error

This way, the random draw may be "decomposed" into two pieces:

$$X = E(X) + [X - E(X)] = E(X) + \varepsilon\,.$$

- The first part $E(X)$ is *not random*.
- All randomness is included in the chance error ε, which is itself a random draw from an **error box** (a box with mean zero).
- **Example**: a random draw X from the box (box 1)

1	2	3

  (which has mean 2) may instead be thought of as $X = 2 + \varepsilon$ where the chance error ε is a random draw from the error box

−1	0	+1

  .

Standard error is the "root-mean-square" of the error box.

- It measures the "size" of the errors in some sense.
- It is a measure of "random variation".
- For two different random draws, one with the larger SE is likely to differ from its expected value by a larger amount.

6.2 Sums of two random draws

We have introduced the concepts of

- A random draw X from a box;
- Its expected value $E(X)$;
- Its standard error $SE(X)$.

The expected value and standard error are not "new" things. Rather, they are new interpretations of old things. They become very useful when we have **more than one draw**.

6.2.1 Sum of two random draws (an example)

Consider the two boxes (box 1)

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}} \text{ and } \boxed{\boxed{2}\,\boxed{4}\,\boxed{6}\,\boxed{8}}.$$

- The first box has mean 2 and SD $\sqrt{\frac{1}{3}\left[(-1)^2+0^2+1^2\right]} = \sqrt{\frac{2}{3}} \approx 0.816$.
- The second box has mean 5 and SD

$$\sqrt{\frac{1}{4}\left[(-3)^2+(-1)^2+1^2+3^2\right]} = \sqrt{5} \approx 2.236\,.$$

Suppose we are going to take a random draw from each, X from the first box, Y from the second box, in such a way that **each possible pair of values is equally likely**. What is the behaviour of the (random) **sum** $S = X + Y$?

All possible pairs/sums

- There are 12 possible pairs:

$$\left(\boxed{1},\boxed{2}\right),\left(\boxed{1},\boxed{4}\right),\left(\boxed{1},\boxed{6}\right),\left(\boxed{1},\boxed{8}\right),$$
$$\left(\boxed{2},\boxed{2}\right),\left(\boxed{2},\boxed{4}\right),\left(\boxed{2},\boxed{6}\right),\left(\boxed{2},\boxed{8}\right),$$
$$\left(\boxed{3},\boxed{2}\right),\left(\boxed{3},\boxed{4}\right),\left(\boxed{3},\boxed{6}\right),\left(\boxed{3},\boxed{8}\right).$$

- Table of all possible pairs and their sums

Sample	Sum
(1,2)	3
(1,4)	5
(1,6)	7
(1,8)	9
(2,2)	4
(2,4)	6
(2,6)	8
(2,8)	10
(3,2)	5
(3,4)	7
(3,6)	9
(3,8)	11

Single random draw from a "bigger" box

Thus getting a random pair (X, Y) and forming the sum $S = X + Y$ is **equivalent** to a *single random draw* from the bigger box (box 1)

3	4	5	5	6	7	7	8	9	9	10	11

What are the mean and SD of this "bigger" box?

The R function `outer()` forms a two-way array by applying an operation to each pair of elements from two vectors:

```
bx = c(1, 2, 3)
by = c(2, 4, 6, 8)
bs = outer(bx, by, "+")
bs
```

```
     [,1] [,2] [,3] [,4]
[1,]    3    5    7    9
[2,]    4    6    8   10
[3,]    5    7    9   11
```

```
mean(bs)
```

```
[1] 7
```

```
mean((bs - mean(bs))^2)
```

```
[1] 5.666667
```

Expected value and standard error of the sum

So we have that $E(S) = 7$ and $SE(S) = \sqrt{5 + \frac{2}{3}} \approx 2.38$. Note that we have

$$7 = E(S) = E(X + Y) = E(X) + E(Y) = 2 + 5\,.$$

and

$$5 + \frac{2}{3} = SE(S)^2 = SE(X + Y)^2 = SE(X)^2 + SE(Y)^2 = \frac{2}{3} + 5\,.$$

So in this case we have

- expected value of sum is sum of expected values;
- *squared* SE of the sum is the sum of the *squared* SEs

These results hold quite generally.

6.2.2 Sum of two random draws (general case).

Consider two boxes (box 1)

$$\boxed{x_1}\ \boxed{x_2}\ \cdots\ \boxed{x_M} \quad \text{and} \quad \boxed{y_1}\ \boxed{y_2}\ \cdots\ \boxed{y_N}$$

Suppose we are going to take a random draw from each: X from the first box, Y from the second box, in such a way that **each possible pair of values is equally likely**.

All possible sums

There are MN possible sums, we may arrange them in a two-way array with M (horizontal) rows and N (vertical) columns. Noting that $\sum_{i=1}^{M} x_i = M\bar{x}$, we may write the column sums below the line:

$$\begin{array}{cccc}
x_1 + y_1 & x_1 + y_2 & \cdots & x_1 + y_N \\
x_2 + y_1 & x_2 + y_2 & \cdots & x_2 + y_N \\
\vdots & \vdots & \ddots & \vdots \\
x_M + y_1 & x_M + y_2 & \cdots & x_M + y_N \\
\hline
M\bar{x} + My_1 & M\bar{x} + My_2 & \cdots & M\bar{x} + My_N
\end{array}$$

The sum of column sums is

$$\underbrace{M\bar{x} + \cdots + M\bar{x}}_{N \text{ terms}} + M(y_1 + \cdots + y_N) = NM\bar{x} + MN\bar{y}\,.$$

Thus the *average* of all possible sums is

$$\frac{\text{sum of all possible sums}}{\text{no. of all possible sums}} = \frac{NM\bar{x} + MN\bar{y}}{MN} = \bar{x} + \bar{y} = E(X) + E(Y)\,.$$

That is, $E(S) = E(X + Y)$.

6.2.3 Aside: Computing formula for SD

To work out the general formula of the SE of sum of random draws, we first recall the formula for the (population) SD. For a list of numbers $x_1, x_2, \ldots, x_M$, the square of the SD may be written as

$$SD^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \bar{x})^2 = \left(\frac{1}{M}\sum_{i=1}^{M} x_i^2\right) - \bar{x}^2$$

which is the "mean square minus the square of the mean".

To see why, recall that $\sum_{i=1}^{M} x_i = M\bar{x}$ and so:

$$\begin{aligned}\sum_{i=1}^{M}(x_i - \bar{x})^2 &= (x_1^2 - 2\bar{x}x_1 + \bar{x}^2) + \cdots + (x_M^2 - 2\bar{x}x_M + \bar{x}^2) \\ &= (x_1^2 + \cdots + x_M^2) - 2\bar{x}(x_1 + \cdots + x_M) + \underbrace{\bar{x}^2 + \cdots + \bar{x}^2}_{M \text{ terms}} \\ &= \sum_{i=1}^{M} x_i^2 - 2\bar{x}M\bar{x} + M\bar{x}^2 = \sum_{i=1}^{M} x_i^2 - M\bar{x}^2\end{aligned}$$

Thus, we have $SD^2 = E(X^2) - E(X)^2$. This computing formula can be used to write a quick-and-easy R function to compute the (population) SD of a list of numbers.

```
popsd = function(x) sqrt(mean(x^2) - (mean(x)^2))
```

Let's try it out:

```
x = 1:10
x  # this list has mean 5.5
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```
sqrt(mean((x - 5.5)^2))
```

```
[1] 2.872281
```

```
popsd(x)
```

```
[1] 2.872281
```

6.2.4 SE of a sum

It is possible to deduce the SE of our general sum $S = X + Y$. We do so by first working out the mean-square of the bigger box of all possible sums ($E[(X+Y)^2]$), and then minus the square of the mean of the sum ($[E(X+Y)]^2$).

Write each squared sum $(x_i + y_j)^2 = x_i^2 + 2x_i y_j + y_j^2$ in an array and add over columns:

$$\begin{array}{ccc} x_1^2 + 2x_1y_1 + y_1^2 & \cdots & x_1^2 + 2x_1y_N + y_N^2 \\ x_2^2 + 2x_2y_1 + y_1^2 & \cdots & x_2^2 + 2x_2y_N + y_N^2 \\ \vdots & \ddots & \vdots \\ x_M^2 + 2x_My_1 + y_1^2 & \cdots & x_M^2 + 2x_My_N + y_N^2 \\ \hline \sum_i x_i^2 + 2M\bar{x}y_1 + My_1^2 & \cdots & \sum_i x_i^2 + 2M\bar{x}y_N + My_N^2 \end{array}$$

The sum of squares (of all possible sums) is then

$$\begin{aligned} E[(X+Y)^2] =& N\sum_i x_i^2 + 2M\bar{x}(y_1 + \cdots + y_N) + M(y_1^2 + \cdots + y_N^2) \\ =& N\sum_i x_i^2 + 2MN\bar{x}\bar{y} + M\sum_j y_j^2 \,. \end{aligned}$$

Since there are MN possible sums, the mean square is

$$E[(X+Y)^2] = \frac{1}{M}\sum_i x_i^2 + 2\bar{x}\bar{y} + \frac{1}{N}\sum_j y_j^2 \,.$$

Since mean of all possible sums is $\bar{x} + \bar{y}$, the squared SD of all possible sums is

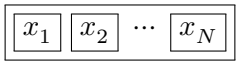
$$
\begin{aligned}
&E[(X+Y)^2] - [E(X+Y)]^2 \\
&= \underbrace{\frac{1}{M}\sum_i x_i^2 + 2\bar{x}\bar{y} + \frac{1}{N}\sum_j y_j^2}_{mean\ sq.} - \underbrace{(\bar{x}^2 + 2\bar{x}\bar{y} + \bar{y}^2)}_{sq.\ of\ mean} \\
&= \frac{1}{M}\sum_i x_i^2 - \bar{x}^2 + \frac{1}{N}\sum_j y_j^2 - \bar{y}^2 \\
&= SE(X)^2 + SE(Y)^2\,.
\end{aligned}
$$

That is, $SE(S)^2 = SE(X)^2 + SE(Y)^2$.

6.3 Sums and averages of random samples of size n

6.3.1 Random samples with replacement of size $n = 2$

A special case of our general sum is where we have a **single** box (box 1)



but take two random draws with replacement. This means each of the N^2 possible pairs $(x_1, x_1), \ldots, (x_1, x_n), \ldots, (x_n, x_1), \ldots, (x_n, x_n)$ is **equally likely**.

The "with replacement" part can be considered as we have two identican boxes, so $E(X) = E(Y)$ and $SE(X) = SE(Y)$. If we write the mean of the box as μ and the SD of the box as σ, then the sum S of the two random draws has

- $E(S) = 2\mu$
- $SE(S) = \sqrt{2}\sigma$.

6.3.2 Random samples of size n

We may easily extend the results to any $n \geq 2$.

Suppose - We have a box with mean μ and SD σ; - We are going to take a random sample of size n from the box **with replacement**; - So each possible sample of size n is equally likely.

Let us write - The random draws as $X_1, X_2, \ldots, X_n$; - The sum as $S = X_1 + \cdots + X_n$; - The *sample average* as $\bar{X} = \frac{S}{n} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^n X_i$.

What are the expected value and standard error of both S and $\bar{X}$?

The sum S

We may extend our results from $n = 2$ easily as each single draw has the same behaviour.

- X_1 (the first draw) is a single random draw and so has
  - $E(X_1) = \mu$
  - $SE(X_1) = \sigma$.
- The same is true for each other draw.

Expected value of sum is sum of expected values:

$$E(S) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = \underbrace{\mu + \cdots + \mu}_{n \text{ terms}} = n\mu\,.$$

Also, $SE(S)^2 = SE(X_1)^2 + \cdots + SE(X_n)^2 = n\sigma^2$, so $SE(S) = \sigma\sqrt{n}\,.$

Going from the sum to the average $\bar{X}$

So the "box of all possible sums" has mean $n\mu$ and SD $\sigma\sqrt{n}$. Then, the box of all possible averages is obtained by taking each possible sum and dividing it by n. This has the effect of

- dividing the mean and the SD by n;

as the values of each element in the box is divided by n. We thus obtain immediately that for the average $\bar{X} = \frac{S}{n} = \frac{X_1 + \cdots + X_n}{n}$,

$$E(\bar{X}) = \frac{E(S)}{n} = \frac{n\mu}{n} = \mu\,;$$

So the "bigger box" of all possible sample means has average equal to the "population mean" μ, which is not surprising.

As for the standard error we have

$$SE(\bar{X}) = \frac{SE(S)}{n} = \frac{\sigma\sqrt{n}}{n} = \frac{\sigma}{\sqrt{n}}$$

which gets smaller as n increases.

6.3.3 Example: 6-sided die

Consider rolling a fair 6-sided die. In this case each of the numbers 1,2,3,4,5,6 are equally likely. This is equivalent to a random draw from the box (box 1)

1	2	3	4	5	6

which has mean $\mu = 3.5 = \frac{7}{2}$, mean-square $\frac{1+4+9+16+25+36}{6} = \frac{91}{6}$ and thus SD

$$\sigma = \sqrt{\frac{91}{6} - \frac{49}{4}} = \sqrt{\frac{182-147}{12}} = \sqrt{\frac{35}{12}} \approx 1.71\,.$$

Rolling the die 3 times: Sum of rolls

Suppose we roll the die ("independently") 3 times. Let X_1, X_2, X_3 denote 3 random draws with replacement from the box

1	2	3	4	5	6

Then the sum of the 3 rolls $S = X_1 + X_2 + X_3$ has $E(S) = 3\mu = \frac{21}{2} = 10.5$ and

$$SE(S) = \sigma\sqrt{3} = \sqrt{\frac{35}{12} \times 3} = \sqrt{\frac{35}{4}} = \frac{\sqrt{35}}{2} \approx 2.958\,.$$

The box of all possible sums here is exactly the dataset `y.dat` from earlier in the lecture!

Rolling the die 3 times: Average of rolls

Writing $\bar{X} = \frac{X_1+X_2+X_3}{3} = \frac{S}{3}$, we have

$$E(\bar{X}) = \frac{E(S)}{3} = \frac{3\mu}{3} = \mu = 3.5$$

and

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{3}} = \sqrt{\frac{35}{12} \times \frac{1}{3}} = \sqrt{\frac{35}{36}} = \frac{\sqrt{35}}{6} \approx 0.956\,.$$

Demonstration

Let us simulate 3 rolls of a 6-sided die 1000 times, and look at the corresponding 1000 sums and averages of each triplet.

```
d = 1:6
S = 0  # empty vector to catch the sums
for (i in 1:1000) {
    rolls = sample(d, size = 3, replace = T)
    S[i] = sum(rolls)
}
mean(S)
```

```
[1] 10.491
```

```
sd(S)
```

```
[1] 2.953412
```

```
popsd(S)
```

```
[1] 2.951935
```

```
hist(S, pr = T, breaks = br)
```

Note these proportions are *close* to (but not *exactly* equal to) the corresponding proportions in `y.dat`.

```
Xbar = S/3
mean(Xbar)
```

```
[1] 3.497
```

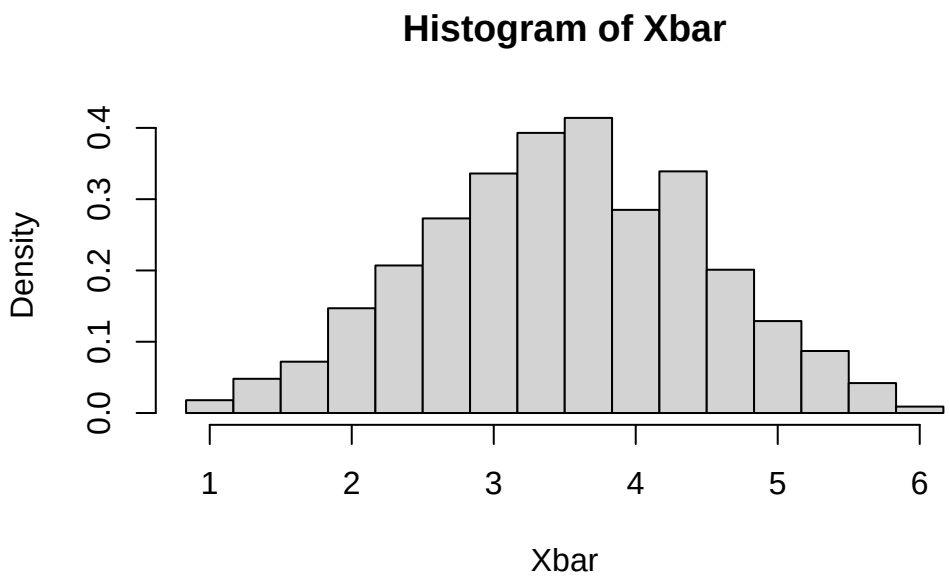
```
sd(Xbar)
```

```
[1] 0.9844706
```

```r
popsd(Xbar)
```

```
[1] 0.9839783
```

```r
hist(Xbar, pr = T, breaks = br/3)
```



Same shape as for the sums, but centred on 3.5 and less spread-out.

6.4 Summary of box models

6.4.1 Single draws from box models

Suppose we have a "box" containing tickets each bearing a number: $\{x_1, \ldots, x_N\}$. The probability a random draw X from the box takes a value in any given range is just the *proportion of x_i values* in that range.

If we know

1. $\mu = \frac{1}{N}(x_1 + \cdots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$ (the average in the box, also called $E(X)$, the **expected value** of X);
2. $\sigma = \sqrt{\frac{1}{N}\left[(x_1 - \mu)^2 + \cdots + (x_N - \mu)^2\right]}$ (the SD of the box, also called $SE(X)$, the **standard error** of X);
3. that the histogram of the x_i values has a normal shape,

we can approximate the probability/chance X takes a value in any given range using the normal curve (i.e. with `pnorm()`).

6.4.2 Chance error

We may *decompose* such a random draw X into two parts:

$$X = E(X) + [X - E(X)] = E(X) + \varepsilon\,,$$

where ε is a random draw from the **error box** $\{x_1 - \mu, \ldots, x_N - \mu\}$.

The SD σ of the original box is the *root-mean-square* of the error box, and describes the "size" of the errors. We may interpret $\sigma = SE(X)$ as the "likely size" of the chance error ε, i.e. the likely size of the deviation of X from its expected value $E(X)$.

6.4.3 Random samples from a box

The sum $S = X_1 + \cdots + X_n$ of a random sample (with replacement) of size n from the box has

- $E(S) = n\mu$;
- $SE(S) = \sigma\sqrt{n}$.

The box of all possible sums has average $n\mu$ and SD $\sigma\sqrt{n}$.

The sample mean $\bar{X} = S/n = \frac{X_1 + \cdots + X_n}{n}$ has

- $E(\bar{X}) = \mu$
- $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

The box of all possible sample means $\bar{X}$ has average μ and SD $\sigma/\sqrt{n}$, i.e., the random variability about $E(\bar{X}) = \mu$ gets less as the sample size n increases.

The two boxes (all possible sums, all possible sample means) have the same shape, and more importantly, **Surprisingly often** they have a normal shape! See next chapter.

7 Central Limit Theorem

In this chapter, we will re-explored the box model and looked at what happens when take larger samples from a box. We then learnt about the central limit theorem as one of the most important results in statistics.

7.1 Kerrich's experiments

We start with Kerrich's experiments in Chapter 5:

- They tossed a (fair) coin 10,000 times and counted the number of heads (5,067).
- They investigated tosses of a "biased coin", made from a wooden disk partly coated in lead.

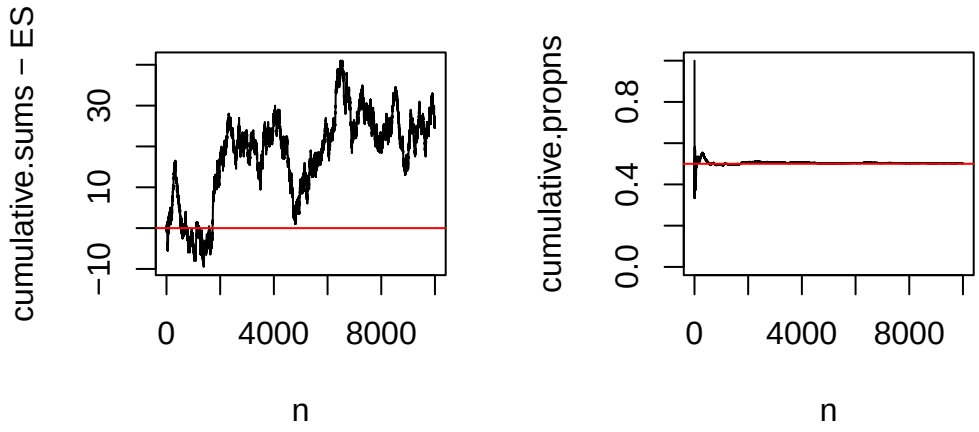
We can simulate the (1st) experiment:

- Each coin flip (assuming the coin is fair) is like a random draw from the "box" $\boxed{0}\,\boxed{1}$
- This box has average $\mu = \frac{1}{2}$ and also SD

$$\sigma = \sqrt{\text{mn.sq.} - (\text{mean})^2} = \sqrt{\frac{1}{2} - \left(\frac{1}{2}\right)^2} = \sqrt{\frac{1}{4}} = \frac{1}{2}\,.$$

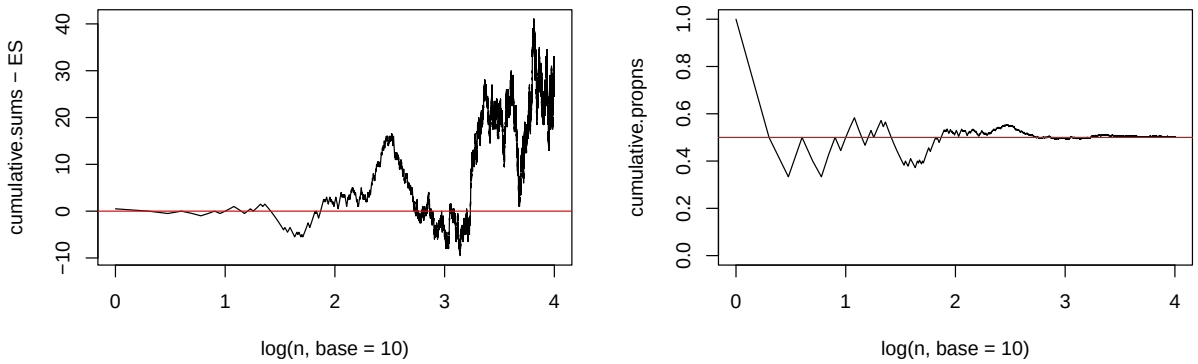
- We may then model n "independent" flips $X_1, \ldots, X_n$ as a random sample with replacement of size n from this box.
- The sum $S = X_1 + \cdots + X_n$ is the **number** of heads.
- The average $\bar{X} = S/n$ is the **proportion** of heads.

```
par(mfrow = c(1, 2))
flips = sample(c(0, 1), size = 10000, replace = T)  # 'box' is c(0,1)
cumulative.sums = cumsum(flips)
n = 1:10000
ES = n/2
plot(n, cumulative.sums - ES, type = "l")
abline(h = 0, col = "red")
cumulative.propns = cumulative.sums/n  # remember n = 1:10000  is a vector!
plot(n, cumulative.propns, type = "l", ylim = c(0, 1))
abline(h = 0.5, col = "red")
```



Looking at the **logarithmic scale**, we are able to see that as we flip the coin more times, the number of heads compared to tails becomes more even. We can also see that there is a lot of fluctuation at the start.

```
par(mfrow = c(1, 2))
plot(log(n, base = 10), cumulative.sums - ES, type = "l")
abline(h = 0, col = "red")
plot(log(n, base = 10), cumulative.propns, type = "l", ylim = c(0, 1))
abline(h = 0.5, col = "red")
```



Size of chance errors as n increases

It seems that

- The size of the chance error in the **sums increases**;
- The size of the chance error in the **proportion decreases**;

This makes perfect sense, because

- The "likely size" of the chance error for the **sum**, i.e.

$$SE(S) = \sigma\sqrt{n} \to \infty$$

  as $n \to \infty$
- The "likely size" of the chance error for the **proportion**, i.e.

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \to 0$$

  as $n \to \infty$.

7.2 Law of Averages

For the sample mean $\bar{X}$ from *any* box model,

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \to 0$$

as $n \to \infty$. So the likely size of the chance error between $\bar{X}$ and $E(\bar{X}) = \mu$ gets smaller and smaller as n increases. In other words, as the sample size n increases, the distribution of a sample mean $\bar{X}$ gets "more concentrated" about the "population mean" μ. This "phenomenon" is (loosely) known as the "Law of Averages" or the "Law of Large Numbers".

7.2.1 Demonstration

We can determine the box of all possible sums for small values of n. Starting with $n = 2$:

```
box = c(0, 1)
s2 = outer(box, box, "+")  # forms two-way array of all possible sums for n=2
s2
```

```
     [,1] [,2]
[1,]    0    1
[2,]    1    2
```

```
as.vector(s2)  # converts matrix to a vector
```

```
[1] 0 1 1 2
```

We can iterate this procedure to get all sums for $n = 3$:

```
s3 = as.vector(outer(box, s2, "+"))  # each sum for n=3 adds 0 or 1 to each sum in s2
s3
```

```
[1] 0 1 1 2 1 2 2 3
```

And to $n = 4, 5, 6$:

```
s4 = as.vector(outer(box, s3, "+"))  # each sum for n=5 adds 0 or 1 to each sum in s4
s4
```

```
 [1] 0 1 1 2 1 2 2 3 1 2 2 3 2 3 3 4
```

```
s5 = as.vector(outer(box, s4, "+"))  # each sum for n=5 adds 0 or 1 to each sum in s4
s5
```

```
 [1] 0 1 1 2 1 2 2 3 1 2 2 3 2 3 3 4 1 2 2 3 2 3 3 4 2 3 3 4 3 4 4 5
```

```
s6 = as.vector(outer(box, s5, "+"))  # each sum for n=6 adds 0 or 1 to each sum in s5
s6
```

```
 [1] 0 1 1 2 1 2 2 3 1 2 2 3 2 3 3 4 1 2 2 3 2 3 3 4 2 3 3 4 3 4 4 5 1 2 2 3 2 3 3 4 2 3 3 4
[46] 4 4 5 2 3 3 4 3 4 4 5 3 4 4 5 4 5 5 6
```

From all possible sums to all possible averages

```
m2 = as.vector(s2)/2
m2
```

```
[1] 0.0 0.5 0.5 1.0
```

```
m3 = s3/3
m3
```

```
[1] 0.0000000 0.3333333 0.3333333 0.6666667 0.3333333 0.6666667 0.6666667 1.0000000
```

```
m4 = s4/4
m5 = s5/5
m6 = s6/6
m6
```

```
 [1] 0.0000000 0.1666667 0.1666667 0.3333333 0.1666667 0.3333333 0.3333333 0.5000000 0.16666
[10] 0.3333333 0.3333333 0.5000000 0.3333333 0.5000000 0.5000000 0.6666667 0.1666667 0.33333
[19] 0.3333333 0.5000000 0.3333333 0.5000000 0.5000000 0.6666667 0.3333333 0.5000000 0.50000
[28] 0.6666667 0.5000000 0.6666667 0.6666667 0.8333333 0.1666667 0.3333333 0.3333333 0.50000
[37] 0.3333333 0.5000000 0.5000000 0.6666667 0.3333333 0.5000000 0.5000000 0.6666667 0.50000
[46] 0.6666667 0.6666667 0.8333333 0.3333333 0.5000000 0.5000000 0.6666667 0.5000000 0.66666
[55] 0.6666667 0.8333333 0.5000000 0.6666667 0.6666667 0.8333333 0.6666667 0.8333333 0.83333
[64] 1.0000000
```

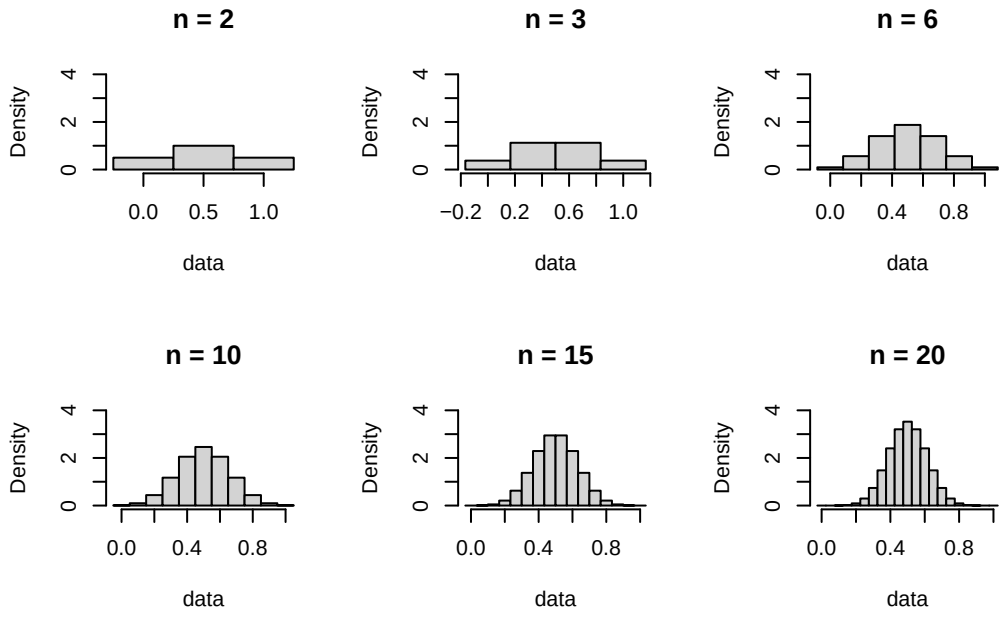
```
s7 = as.vector(outer(box, s6, "+"))
m7 = s7/7
means = list(`n=2` = m2, `n=3` = m3, `n=4` = m4, `n=5` = m5, `n=6` = m6, `n=7` = m7)
```

Create a list to $n = 20$

```
s8 = as.vector(outer(box, s7, "+"))
s9 = as.vector(outer(box, s8, "+"))
s10 = as.vector(outer(box, s9, "+"))
s11 = as.vector(outer(box, s10, "+"))
s12 = as.vector(outer(box, s11, "+"))
s13 = as.vector(outer(box, s12, "+"))
s14 = as.vector(outer(box, s13, "+"))
s15 = as.vector(outer(box, s14, "+"))
s16 = as.vector(outer(box, s15, "+"))
s17 = as.vector(outer(box, s16, "+"))
s18 = as.vector(outer(box, s17, "+"))
s19 = as.vector(outer(box, s18, "+"))
s20 = as.vector(outer(box, s19, "+"))

m8 = s8/8
m9 = s9/9
m10 = s10/10
m11 = s11/11
m12 = s12/12
m13 = s13/13
m14 = s14/14
m15 = s15/15
m16 = s16/16
m17 = s17/17
```

```
m18 = s18/18
m19 = s19/19
m20 = s20/20


means <- list(`n=2` = m2, `n=3` = m3, `n=4` = m4, `n=5` = m5, `n=6` = m6, `n=7` = m7,
 ↪  `n=8` = m8,
    `n=9` = m9, `n=10` = m10, `n=11` = m11, `n=12` = m12, `n=13` = m13, `n=14` = m14,
     ↪  `n=15` = m15,
    `n=16` = m16, `n=17` = m17, `n=18` = m18, `n=19` = m19, `n=20` = m20)
```

```
n_values <- c(2, 3, 6, 10, 15, 20)
par(mfrow = c(2, 3))

for (n in n_values) {
    ((n - 1) <= length(means))
    br <- (0:(n + 1) - 0.5)/n
    data <- means[[n - 1]]
    hist(data, pr = TRUE, breaks = br, main = paste("n =", n), ylim = c(0, 4))
}
```



...and so on...

In this example it is very clear that **TWO** important things are happening:

1. The spread of the distribution of all possible averages/proportions is getting **more concentrated about** $\mu = 0.5$ **as** n **increases**;

2. The shape of the histogram of all possible averages/proportions is becoming "normal-shaped".

The normal shape means we can approximate probabilities, knowing only $E(\bar{X}) = \mu$ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Is the "normal shape" due to something special about this particular simple box?

- **Not really!**

7.2.2 Example: Rolling a 6-sided die

Suppose we are interested in rolling a 6-sided die n times. This is like taking a random sample of size n from the box

1	2	3	4	5	6

This box has

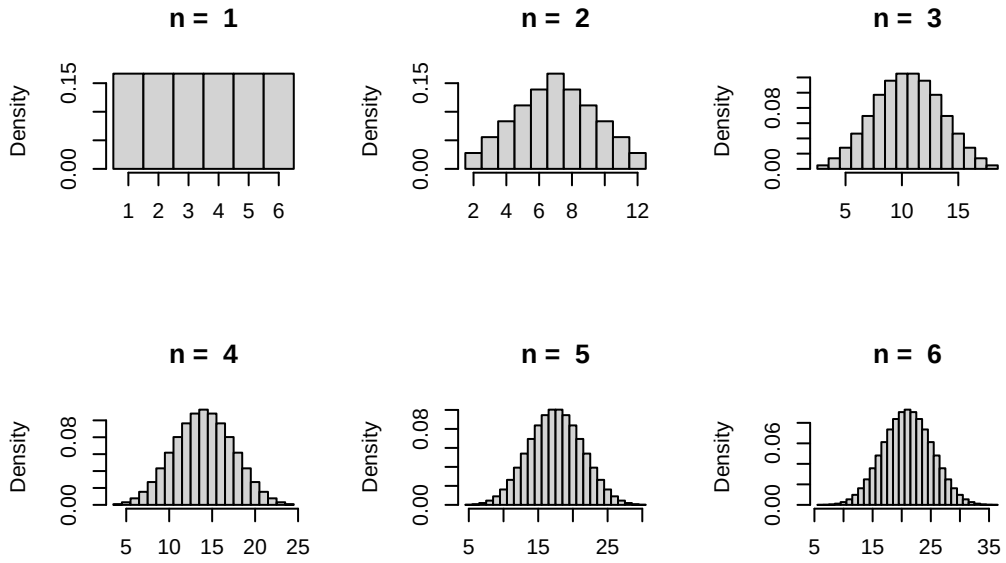
- mean $\mu = 3.5 = \frac{7}{2}$
- mean square $\frac{1+4+9+16+25+36}{6} = \frac{91}{6}$
- SD $\sigma = \sqrt{\frac{91}{6} - \left(\frac{7}{2}\right)^2} = \sqrt{\frac{182-(3\times 49)}{12}} = \sqrt{\frac{35}{12}} \approx 1.708.$

```
box = 1:6
box
```

```
[1] 1 2 3 4 5 6
```

```
s2 = as.vector(outer(box, box, "+"))
s3 = as.vector(outer(s2, box, "+"))
s4 = as.vector(outer(s3, box, "+"))
s5 = as.vector(outer(s4, box, "+"))
s6 = as.vector(outer(s5, box, "+"))
sums.rolls = list(box, s2, s3, s4, s5, s6)
```

Histograms of all possible sums-of-n-rolls



For $n = 6$ this is certainly normal-shaped too!

Asymmetric example

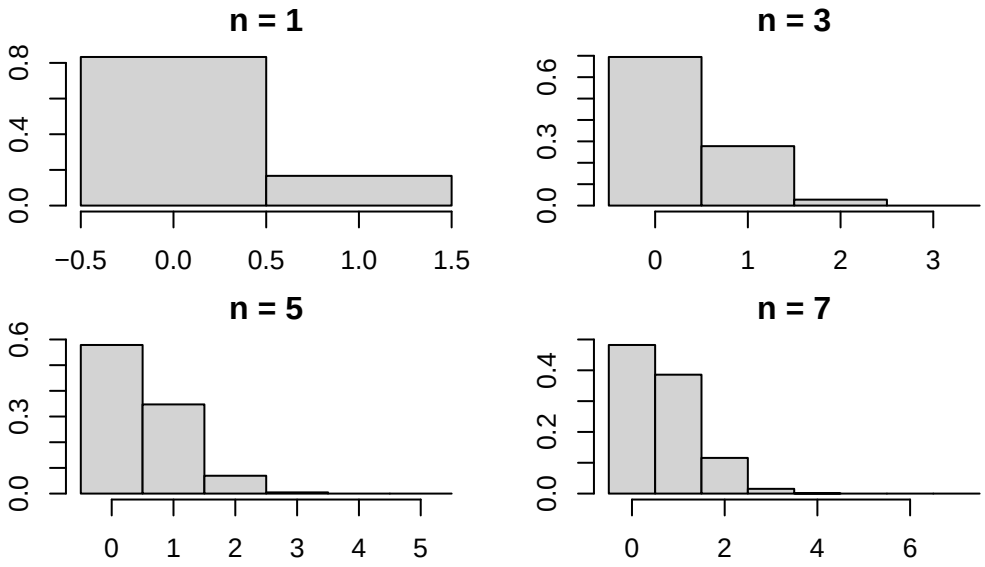
Instead of the sum of the rolls, how about the number of times we roll a $\boxed{6}$? We can instead use the box

$$\boxed{\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{1}}$$

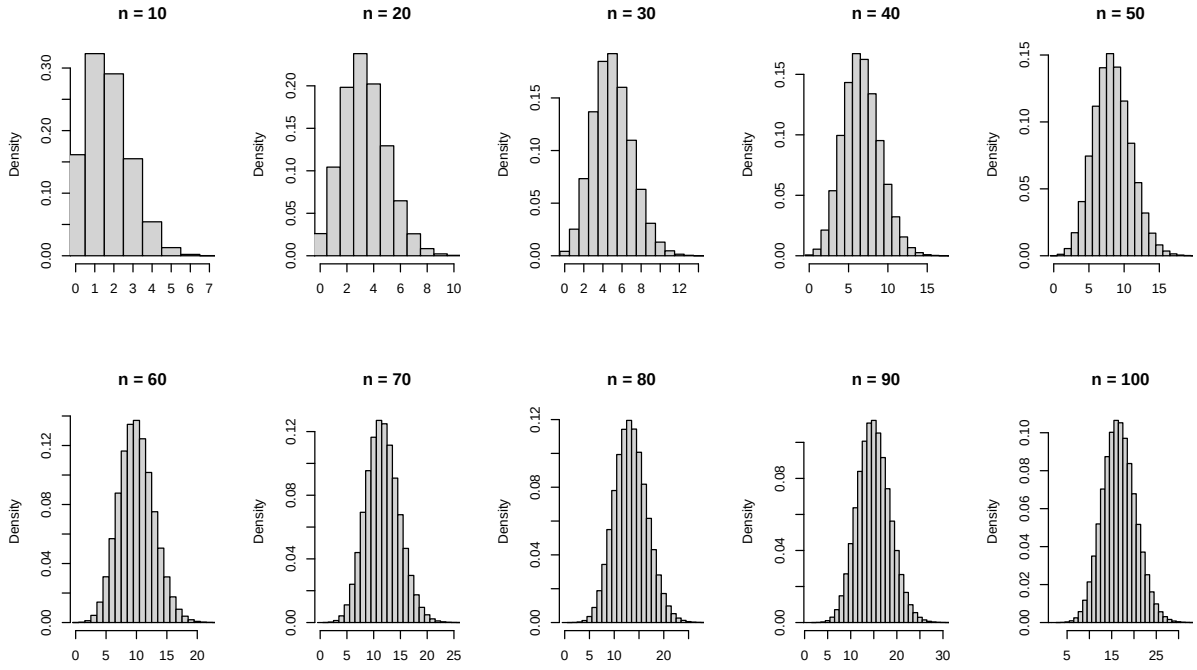
The number of times we get $\boxed{6}$ in n rolls is just like the sum S when we take a random sample of size n from this new box. This new box has

- mean $\mu = \frac{1}{6}$
- mean square $\frac{1}{6}$
- SD $\sigma = \sqrt{\frac{1}{6} - \left(\frac{1}{6}\right)^2} = \sqrt{\frac{6-1}{36}} = \frac{\sqrt{5}}{6} \approx 0.373$.

Histograms of all possible no.s-of-6s



Not looking very normal-shaped…what about if we let n get larger?



We get a normal shape, but only for larger n

So although the histograms of all possible sums ("nnumbers-of-times-we-roll-six") are not normal-shaped for smaller n, as n increases the shape gets closer to a normal. By the time $n > 100$, the shape is quite symmetric. It turns out that for essentially any box, we get the

same phenomenon occuring as n gets larger and larger, the box of all possible sums gets a "more normal" shape.

7.3 The Central Limit Theorem

Most important result in Statistics

This phenomenon can be *mathematically proven* to hold for any fixed (finite) box. This result is a special case of the **Central Limit Theorem**.

- It is a "limit theorem" because it describes what happens "in the limit" as $n \to \infty$.
- "Central" here means "most important".

The "standard normal CDF" $\Phi(z)$ is the function given in R by `pnorm(`z`)`.

> If $S = X_1 + \cdots + X_n$ is the sum of random sample (with replacement) of size n from a box with mean μ and SD σ, then for **large** n,
>
> $$P(S \le s) = P\left(\frac{S - n\mu}{\sigma\sqrt{n}} \le \frac{s - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{s - n\mu}{\sigma\sqrt{n}}\right)$$

Deconstructing the Central Limit Theorem

Note that the desired sum value s being considered here, when converted into standard units is

$$z_s = \frac{s - E(S)}{SE(S)} = \frac{s - n\mu}{\sigma\sqrt{n}}\,,$$

which is the ratio inside the $\Phi(\cdot)$. Therefore, converting to R code, we have

$$P(S \le s) \approx \texttt{pnorm}((s - n\mu)/(\sigma\sqrt{n})) = \texttt{pnorm}(s, \texttt{m} = n\mu, \texttt{s} = \sigma\sqrt{n})\,.$$

Example: Roulette

A roulette wheel has slots numbered 1 to 36, plus 1 (or more) slots marked 0.

- half the positive numbers are coloured black;
- the remaining positive numbers are coloured red;
- the zero slots are coloured green.

If you bet on either "red" or "black",

- you double your money if the ball lands in a slot of your colour
- you lose your money otherwise.

Suppose a wheel has two green slots ("0" and "00"), each slot is equally likely and a player bets \$1 on "red" for n consecutive spins.

Let S denote the total winnings after n spins. We want to approximate $P(S > 0)$ for $n = 5, 25, 125, 625$.

The Roulette Box

- There are 38 slots in total, 18 of which are red.
- If the ball
  - lands in a red slot the player wins \$1;
  - does **not** land in a red slot, the player loses \$1, i.e. they win –\$1.
- Use the following box:

$$\underbrace{\boxed{-1}\ \cdots\ \boxed{-1}}_{\text{20 of these}}\quad\underbrace{\boxed{+1}\ \cdots\ \boxed{+1}}_{\text{18 of these}}$$

  which has
  - mean $\mu = \frac{-2}{38} = -\frac{1}{19}$;
  - mean square 1
  - SD $\sigma = \sqrt{1 - \left(\frac{1}{19}\right)^2} = \sqrt{\frac{360}{361}} \approx 0.9986$.

Exact answers

- It is possible to work out the exact probabilities (using the "binomial distribution", more on this later).
- These are

```
n = c(5, 25, 125, 625)
prob.win = 1 - pbinom(n/2, n, 18/38)
rbind(n, prob.win)
```

```
              [,1]       [,2]        [,3]         [,4]
n        5.0000000 25.0000000 125.0000000 625.00000000
prob.win 0.4507489  0.3951246   0.2775865   0.09388094
```

7.3.1 Normal approximation

- According to the Central Limit Theorem, for "large n",

$$P(S > 0) = 1 - P(S \leq 0) \approx 1 - \Phi\left(\frac{0 - \left(-\frac{n}{19}\right)}{\sqrt{\frac{360n}{361}}}\right) = 1 - \texttt{pnorm}\left(\frac{\sqrt{361n}}{19\sqrt{360}}\right)$$

- This gives

```
1 - pnorm(sqrt(361 * n)/(19 * sqrt(360)))
```

```
[1] 0.45309281 0.39607370 0.27784490 0.09381616
```

- These are quite good approximations (even for $n = 5$)!
- Makes sense, because the box is reasonably symmetric (not that different in shape to Kerrich's box).

Final comments

When we take a random sample of size n (with replacement) from a box with mean μ and SD σ, the box of all possible sums

- Has mean equal to $E(S) = n\mu$;
- Has SD equal to $SE(S) = \sigma\sqrt{n}$;
- Is (approx.) normal-shaped for "large enough n".

For such n we can approximate probabilities for the random sum S or average $\bar{X} = S/n$, using `pnorm()`.

How large is "large enough n"? **It depends**, on how "non-normal" the original box is. If the original box is

- Reasonably symmetric (without too many outliers), $n = 5$ or 10 may do;
- Very skewed, we may need $n > 100$ before the box of all possible sums has a nice, symmetric normal shape.