STAT5002: Introduction to Statistics

Outline

- STAT5002: Introduction to Statistics
  - Introduction to the course
  - Critical thinking with data
    - Population and samples
    - Sources of bias
    - Experimental design
  - Numerical and graphical summaries
    - Example: Australian Road Fatalities Jan-April 2016
    - Identifying Variables
    - Numerical and Graphical Summaries
    - Barplot
    - Frequency table and ordinate diagram
    - Frequency table and histograms
    - Notations
    - Summaries for Centre
    - Numerical Summaries for Spread
    - Boxplot
    - Outliers

Resources

- See `https://canvas.sydney.edu.au` for Unit Information Sheet, Lecture Notes, Tutorial Sheets, Quiz and assignment
- Consultation Hours: Wednesday 5-6pm
- Lecture Hours: Wednesday 6-8pm
- Tutorial Hours: Wednesday 8-9pm

Outline of the course

1. Summary statistics
2. Probability
3. Random variables
4. Proportion hypothesis tests
5. Mean hypothesis tests
6. Goodness-of-fit tests
7. Confidence intervals
8. Bivariate data
9. Multiple Linear Regression
10. Model selection
11. Logistic and non-parametric regression
12. Bayesian statistics

Extra Revision

Assessments

Objectives

- Fundamental statistical concepts.
- Methodologies related to statistical data analysis, data mining and data science.
- A number of useful statistical models.
- Computer oriented estimation procedures.
- Non-parametric concepts.
- Analysis of large data sets.
- The R computing language - for all computational aspects in the course.

What is a statistician?

A primary objective of the statistician is to answer a research question. Most of the time, the research question concerns with activities/behaviours/phenomena in a targeted **population**.

Measurements with respect to the population is always very difficult to obtain, so statisticians aim to answer the research question using a **sample**.

A majority of the statistics literature looks at how to ‘best’ infer population characteristics from a sample.

Population and samples

What is the voter turnout rate for the 2016 Australian Federal election?

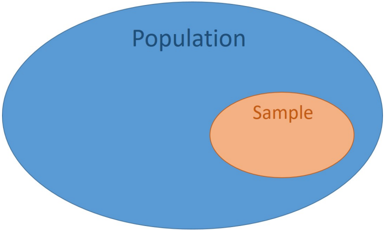
Population and samples

What is the voter turnout rate for the 2016 Australian Federal election?

1. Solution 1: Survey all Australians.
2. Solution 2: Choose model for voter turnout based on 2012 data. Make a prediction for the 2016 election.

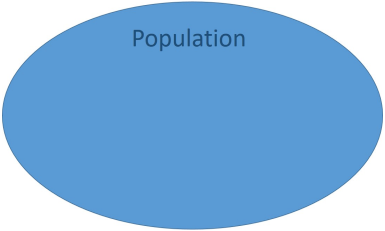
Population and samples

- The target population comprises **all** relevant subjects of interest.
- The sample is a manageable subset, selected to make the study feasible.



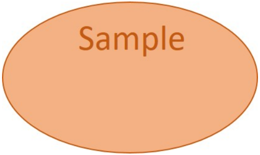
Population

- The target population is the collection of all things (or subjects) we are interested in.
- The target population should be well-defined in terms of
  - Who/What
  - Where
  - When



Sample

- A sample is a subset of the population.
- It should be representative of the target population (not biased).
- Large enough to give accurate information about the population.
- Ideally, the observations should be independent of each other.



A representative sample

Only a representative sample should be used to make inferences about the target population. One way to ensure that a sample is representative of the target population is to obtain a random sample.

A representative sample

Only a representative sample should be used to make inferences about the target population. One way to ensure that a sample is representative of the target population is to obtain a random sample.

Sources of bias

Bias may be defined as any systematic error (ie. not occurring randomly) which results in incorrect conclusions about the target population.

Some types of bias include

- Selection bias
- Measurement bias
- Response bias
- Confounding

Selection bias

Selection bias refers to any systematic differences occurring in the way that subjects are selected for a study.

E.g. In a height study, we accidentally selected a group of basketball players.

Measurement bias

Measurement bias refers to systematic differences in the measurement of variables.

E.g. In a human body temperature study, an in-ear thermometer is consistently higher than that of an oral thermometer.

Response bias

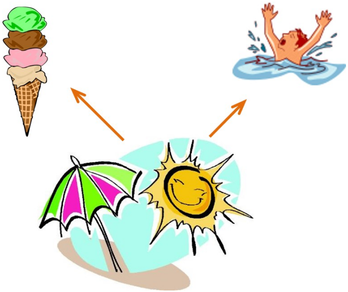
- Response bias can occur when the response rate to a survey is too low.
- This is because those who respond to a survey often have different characteristics or attitudes than those who don't respond.
- This is most common when sensitive issues are involved.

Acceptable response rates vary amongst researchers

- Some say the response rate should be at least 75% to ensure that a study is not significantly affected by response bias.
- Others are satisfied with a response rate of at least 50%.
- The response rate should always be reported.

Confounding

A confounder is a variable that distorts (increases or decreases) the apparent effect of one variable (determinant) on another (outcome).



Study design

There are two types of study designs.

An **observational study** is one in which there is no treatment imposed by the investigator.

- We simply observe.
- Data are observed and recorded based on responses from subjects.

An **experimental study** is one in which the investigator has some control over the subjects by giving some kind of treatment.

- **Explanatory variable** (determinant) is perturbed, behaviour of **dependant variable** (response) is noted.
- Data are observed and recorded based on responses from subjects.

Study design

Conclusions of a study depends on the design. Roughly speaking:

- Observational studies allow us to infer **association**.
  - But one should be very careful about the implications!
  - Confounding variables are always a possible cause of ridiculous conclusions.
  - E.g. 3 glasses of water can cure flu. Problem: I take my flu medicine with a glass of water, 3 times day.
- Experimental studies allow us to infer **causation**.
  - Usually more informative on the underlying mechanisms, since the researchers can eliminate external factors in their experiments.
  - One must be very familiar with **experimental designs** and take into account of all sources of variations.
  - E.g. Crops growth is better in soil with high nitrogen and locations with good sunlight. If an experiment only contains high nitrogen soil and good sunlight, then it is not possible to separate out these two effects.

Example: Australian Road Fatalities Jan-April 2016

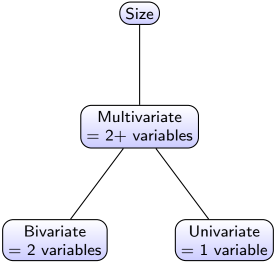
The number of road fatalities in Australia continues to rise, given the ever increasing volume of vehicles on the road, despite preventative measures as compulsory seat belts and school zones. Last year in Australia, 1,209 died on our roads.

Data from the Australian Bureau of Statistics (ABS) from the first four months of 2016, gives the following variables:
Crash ID, State, Date, Day, Month, Year, Dayweek, Time, Hour, Min, Crash Type, Bus Involvement, Rigid Truck Invovement, Articulated Truck Involvement, Speed Limit, Road User, Gender, Age.

What questions do you have?

The 1st step in EDA is to identify the variables, in terms of form and type.

(i) Size of Variables
How many bits of information have been recorded?
How many variables, p, have we observed?
How many observations, n have we observed?



Size

Multivariate
$=$ 2$+$ variables

Bivariate
$=$ 2 variables

Univariate
$=$ 1 variable

In 'big data' we commonly have 'large p, small n' meaning that we have stacks of variables (eg gene data) relative to the data size.

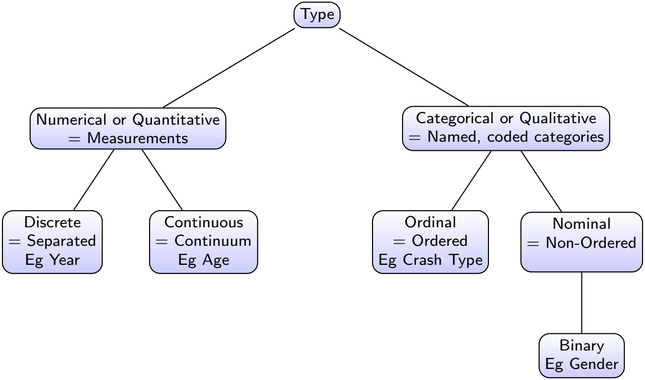
Identifying Variables

See live coding demo.

```
#read in data
data = read.csv("../datasets/2016Fatalities.csv",header=T)

#get dimension of data
dim(data)

## [1] 442  18
```

(ii) Type of Variables

What is the nature of the variables – i.e. what process or situation ‘produced’ the data? This will determine how we should later analyse our data.


Type
Numerical or Quantitative = Measurements
Categorical or Qualitative = Named, coded categories
Discrete = Separated Eg Year
Continuous = Continuum Eg Age
Ordinal = Ordered Eg Crash Type
Nominal = Non-Ordered
Binary Eg Gender

Identifying Variables

```
names(data) #Lists all the variables
```

```
##  [1] "Crash.ID"                        "State"
##  [3] "Date"                            "Day"
##  [5] "Month"                           "Year"
##  [7] "Dayweek"                         "Time"
##  [9] "Hour"                            "Min"
## [11] "Crash.Type"                      "BusInvolvement"
## [13] "RigidTruck..Involvement"         "Articulated.Truck..Involvement."
## [15] "SpeedLimit"                      "RoadUser"
## [17] "Gender"                          "Age"
```

```
colnames(data)  #Lists all the variables
```

```
##  [1] "Crash.ID"                        "State"
##  [3] "Date"                            "Day"
##  [5] "Month"                           "Year"
##  [7] "Dayweek"                         "Time"
##  [9] "Hour"                            "Min"
## [11] "Crash.Type"                      "BusInvolvement"
## [13] "RigidTruck..Involvement"         "Articulated.Truck..Involvement."
## [15] "SpeedLimit"                      "RoadUser"
## [17] "Gender"                          "Age"
```

Identifying Variables

```
data[1,]  #Extracts the 1st row
```

```
##     Crash.ID State     Date Day   Month Year Dayweek  Time Hour Min
## 1 2.2016e+12   VIC 1-Jan-16   1 January 2016  Friday 20:30   20  30
##       Crash.Type BusInvolvement RigidTruck..Involvement
## 1 Single vehicle             No                      No
##   Articulated.Truck..Involvement. SpeedLimit         RoadUser Gender Age
## 1                              No         80 Motorcycle rider   Male  25
```

```
head(data, 3)  #List the first 3 rows of data
```

```
##     Crash.ID State     Date Day   Month Year  Dayweek  Time Hour Min
## 1 2.2016e+12   VIC 1-Jan-16   1 January 2016   Friday 20:30   20  30
## 2 4.2016e+12    SA 1-Jan-16   1 January 2016   Friday  1:00    1   0
## 3 1.2016e+12   NSW 2-Jan-16   2 January 2016 Saturday  0:30    0  30
##       Crash.Type BusInvolvement RigidTruck..Involvement
## 1 Single vehicle             No                      No
## 2 Single vehicle             No                      No
## 3 Single vehicle             No                      No
##   Articulated.Truck..Involvement. SpeedLimit         RoadUser Gender Age
## 1                              No         80 Motorcycle rider   Male  25
## 2                              No        110           Driver   Male  40
## 3                              No        100        Passenger   Male  18
```

Identifying Variables

```
class(data)  #Shows the way R has stored the data
```

```
## [1] "data.frame"
```

```
str(data)
```

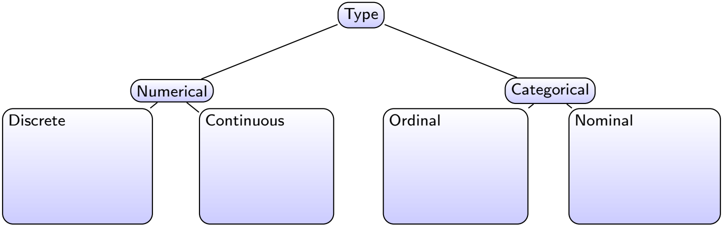
```
## 'data.frame': 442 obs. of  18 variables:
##  $ Crash.ID                      : num  2.2e+12 4.2e+12 1.2e+12 5.2e+12 6.2e+12 ...
##  $ State                         : Factor w/ 8 levels "ACT","NSW","NT",..: 7 5 2 8 6 6 4 6 2 2 ...
##  $ Date                          : Factor w/ 113 levels "1-Apr-16","1-Feb-16",..: 3 3 44 44 44 44 86 8
##  $ Day                           : int  1 1 2 2 2 2 3 3 4 4 ...
##  $ Month                         : Factor w/ 4 levels "April","February",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Year                          : int  2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...
##  $ Dayweek                       : Factor w/ 7 levels "Friday","Monday",..: 1 1 3 3 3 3 4 4 2 2 ...
##  $ Time                          : Factor w/ 225 levels "0:00","0:12",..: 141 128 5 101 127 127 56 29
##  $ Hour                          : int  20 1 0 17 19 19 14 11 20 21 ...
##  $ Min                           : int  30 0 30 20 58 58 0 55 25 45 ...
##  $ Crash.Type                    : Factor w/ 3 levels "Multiple vehicle",..: 3 3 3 1 1 1 2 1 3 3 ...
##  $ BusInvolvement                : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RigidTruck..Involvement       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Articulated.Truck..Involvement.: Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ SpeedLimit                    : int  80 110 100 110 80 80 60 100 100 60 ...
##  $ RoadUser                      : Factor w/ 6 levels "Bicyclist (includes pillion passengers)",..: 4
##  $ Gender                        : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ Age                           : int  25 40 18 53 17 31 70 51 59 17 ...
```

Note:

- In practice, continuous data is often reported as discrete data (by rounding), but the underlying quantity represented is still continuous (eg Age and Time).
- A helpful diagnostic for determining continuous data is to ask: "Could this data have been recorded to higher accuracy, given a more precise 'instrument'?"
- Quantitative data can be simplified to qualitative data. For example, in a survey, a respondent may feel more comfortable giving a general answer to a question about their personal income.

Have a try

Identify all the variables for Australian Road Fatalities.



Don't worry if your answer is slightly different to mine! Depending on the context of the data, it is possible to come up with different classifications of variables.

Numerical summaries

Note: We use different summaries for different types of variables.

- **Categorial Data**
  Categorical data is essentially already summarised by category. We note the most common category or any trend within the categories.

- **Numerical Data**
  Numerical summaries focus on a feature of interest, like the centre and spread.

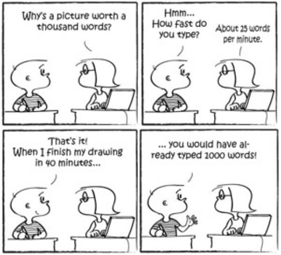
Data summary

- 99% of the time, we cannot look at our data directly because of its complexity and size.
- Summaries of our data are therefore useful in reflecting the aspect of the data we care the most about.
- The main two types of summaries: numerical and graphical summary.
- E.g. I have never read 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.)

Graphical Summaries

Once we identify the variables, we can summarise the data, both graphically and numerically, in order to identify and highlight the main features of interest. A careful choice of graphical and numerical summaries can give a quick, transparent, perceptive snapshot of the data.

We often start with graphical summaries because 'A (well-designed) picture is worth a thousand words.' (Similar idea: Arthur Brisbane, Syracuse Advertising Men's Club, 1911)


Why's a picture worth a thousand words?
Hmm... How fast do you type?
About 25 words per minute.
That's it! When I finish my drawing in 40 minutes...
... you would have already typed 1000 words!

© 2009 cartoosh.com

How to choose an appropriate graphical summary?

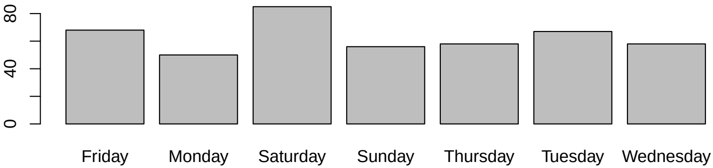
- The critical question is: ‘What plot is the more informative?’ or ‘What plot will best highlight features of the data?’ or ‘What plot will best guide the next analysis?’.
- To some extent we use trial and error. We try some standard forms and see what is revealed about the data. One graphical summary can suggest another, and often a combination will highlight different features of the data
- In practise we use computer packages like `R` [1] 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.

[1] Some computer packages vary slightly in construction. For example, in the calculation of the quartiles or the length of the whiskers in the boxplot.

Summary0: Barplot (Categorical data)

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

```
DayWeek <- data$Dayweek
table(DayWeek)


## DayWeek
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##        68        50        85        56        58        67        58


barplot(table(DayWeek))
```

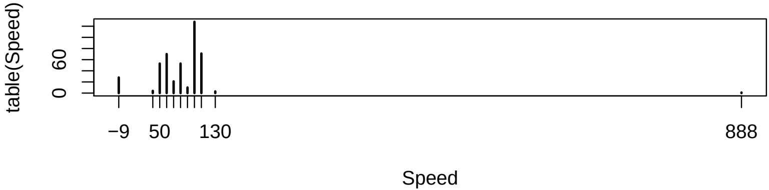


Frequency table and ordinate diagram (discrete data)

Q: What was the most common speed at which a road fatality occurred?

The frequency table is a very simple way to summarise a set of discrete data and when plotted gives an ordinate diagram.

```
Speed <- data$SpeedLimit  #Extracts SpeedLimit
table(Speed)


## Speed
##  -9  40  50  60  70  80  90 100 110 130 888
##  28   4  53  70  21  53  10 128  71   3   1


plot(table(Speed))
```



Frequency table and histograms (continuous data)

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

The frequency table can also be used to summarise a set of continuous data, by collecting it into intervals (or ‘bins’). What is lost?

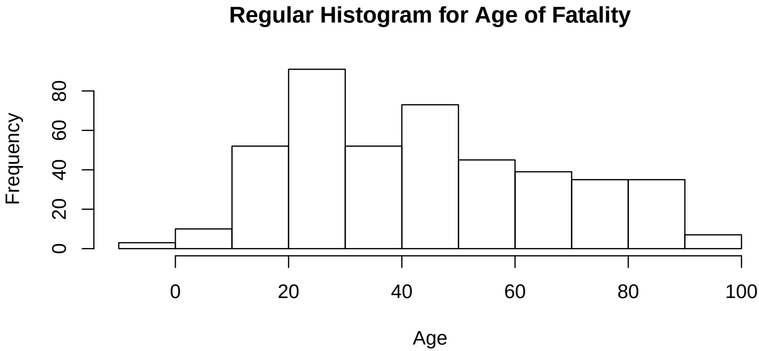
- For equal bin lengths, we can simply sort the data into the bins, and then plot the frequency against each bin. This is called a ‘regular’ histogram.
- For unequal bin lengths, we need to sort the data into the bins, then work out the relative frequency (=frequency/sample size) and the height (=relative frequency/interval length). Plotting the height against each bin is called a 'probability' histogram.

(i) Using equal bins: Regular Histogram

Bin	Frequency
[-10,0)	?
[0,10)	11
[10,20)	
. . .	
[90,100)	9

```
Age <- data$Age
min(Age)

## [1] -9

max(Age)

## [1] 96
```

```r
Age <- data$Age
hist(Age,xlab="Age",
     main="Regular Histogram for Age of Fatality")
```



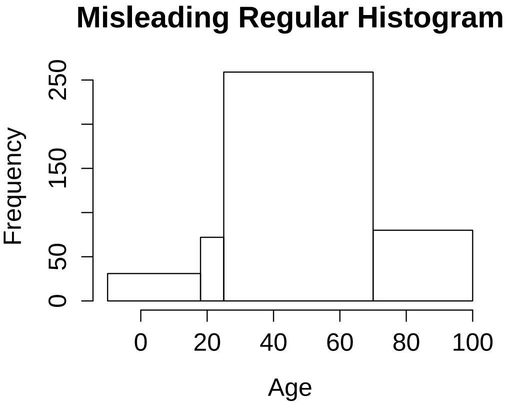
(ii) Using unequal bins: Probability Histogram

Bin	Frequency	Relative Frequency	Height
[-10,18)	31	$31/442 = 0.07$	0.0025
[18,25)	72	$72/442 = 0.16$	0.0232
[25,70)	259	$259/442 = 0.59$	0.0130
[70,100)	80	$80/442 = 0.18$	0.0060
Total	442	1	

where:
Relative Frequency = Frequency/442
Height = Relative Frequency/Bin length
Eg For bin [-10,18): height $= 0.07/28 =$3.6.

```
breaks=c(-10,18,25,70,100)
table(cut(Age,breaks,right=F))

##
## [-10,18)  [18,25)  [25,70) [70,100)
##       31       72      259       80

# hist(Age,br=breaks,freq=F,right=F,
#      xlab="Age",
#      main="Probability Histogram for Age of Fatality")
```

Note how the 'regular' histogram is misleading for unequal bin lengths, as it suggests that [25,70) is the most likely bin.



Half-time break (5 mins)

Numerical summary

- Most of the time, data come in the form of numbers. Summarising the information embedded in a large dataset should be quite straight-forward, right?
- In fact, a ‘statistic’ means a transformation/summary of data. Different statistics (plural of a statistic) have different properties and will summarise a data in different ways.
- While it is important to know how to calculate these summary statistics, it is also important to know what kind of informations are captured by these statistics.
- In other words, you need to **interpret** these statistics in context of your data.

Can we summarise data numerically?

```
summary(data[9:18])
```

```
##       Hour            Min                   Crash.Type  BusInvolvement
##  Min.   : 0.00   Min.   : 0.00   Multiple vehicle:195   No :434
##  1st Qu.: 8.00   1st Qu.: 0.00   Pedestrian      : 51   Yes:  8
##  Median :13.00   Median :20.00   Single vehicle  :196
##  Mean   :12.52   Mean   :20.84
##  3rd Qu.:17.00   3rd Qu.:35.00
##  Max.   :23.00   Max.   :59.00
##  RigidTruck..Involvement Articulated.Truck..Involvement.   SpeedLimit
##  No :412                 No :408                         Min.   : -9.00
##  Yes: 30                 Yes: 34                         1st Qu.: 60.00
##                                                          Median : 80.00
##                                                          Mean   : 79.76
##                                                          3rd Qu.:100.00
##                                                          Max.   :888.00
##                                     RoadUser      Gender        Age
##  Bicyclist (includes pillion passengers): 16   Female:116   Min.   :-9.0
##  Driver                                 :212   Male  :326   1st Qu.:25.0
##  Motorcycle pillion passenger           :  4                Median :42.5
##  Motorcycle rider                       : 89                Mean   :44.6
##  Passenger                              : 66                3rd Qu.:61.0
##  Pedestrian                             : 55                Max.   :96.0
```

Some basic notations

Given a univariate data set of sample size n:

- The data is $\{x_i\}, i = 1, 2, \ldots, n$ or $\{x_1, x_2 \ldots, , x_n\}$.
- The ordered (ascending) data set is $\{x_{(i)}\}, i = 1, 2, \ldots, n$ or $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$.
- The sum of the data is $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots x_n$.

Have a try

Given a data set $\{1, 4, 6, 2, 3, 7\}$, find

$$\sum_{i=1}^{6} x_i, \sum_{i=2}^{5} x_i^2, \sum_{i=1}^{6} i x_i, \sum_{i=1}^{6} (x_{(i)} - 1)$$

```
#Check your answers
x=c(1,4,6,2,3,7)
y=c(sum(x), sum(x[2:5]^2), sum(c(1:6)*x), sum(sort(x)-1))
y

## [1] 23 65 92 17
```

▸ More practise here

Summaries for Centre

There are 2 main measures of the centre (or location) of the data:

- **Mean** $\bar{x}$

  The mean is the average of the data.

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

- **Median** $\tilde{x}$

  The median is the centre of the data, also called the 50% percentile or the 2nd quartile. It splits the data into 2 equal groups.

  - If n is odd, the unique median is the middle value:

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

  - If n is even, the median is the average of the 2 middle values (by convention):

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

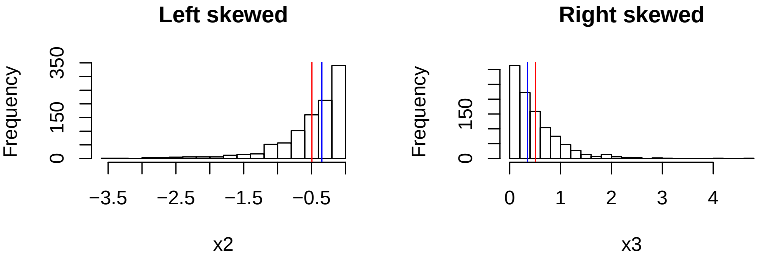
Have a try

For the data: $\{1, 4, 6, 2, 3, 7\}$, show that the mean is 3.83 and the median is 3.5.

```
#Check your answers
x=c(1,4,6,2,3,7)
mean(x)

## [1] 3.833333

median(x)

## [1] 3.5
```

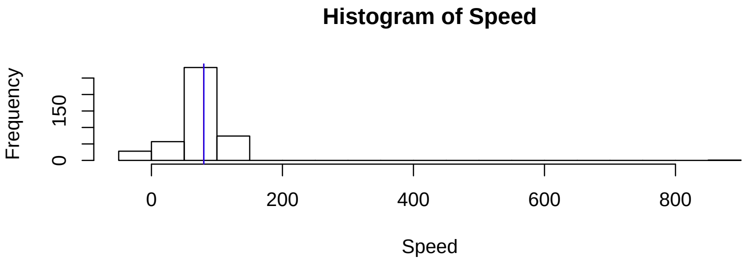
Comparing the Mean and Median

- For symmetric data, we expect $\bar{x} = \tilde{x}$. For left skewed data, we expect $\bar{x} < \tilde{x}$ and for right skewed data, $\bar{x} > \tilde{x}$.



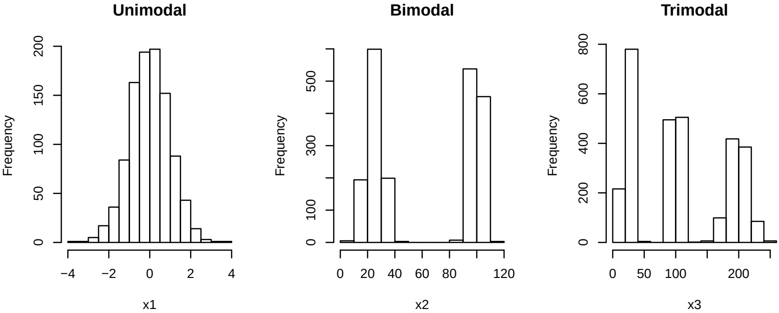
red = mean, blue = median

```
Speed <- data$SpeedLimit
mean(Speed)


## [1] 79.76471


median(Speed)


## [1] 80


hist(Speed,breaks = 20)
abline(v = c(mean(Speed),median(Speed)), col = c("red","blue"))
```



- Which is ‘optimal’ for describing the centre of the data? E.g. Why don’t we always just use the median?
- Because sometimes the mean has superior statistical properties compare to the median, and the use of any statistics should always be data and context dependent.
- Both have strengths and weaknesses depending on the nature of the data.
  - The mean is helpful for data which is basically symmetric and does not have too many outliers. As we will see later in the course, it also have many desirable theoretical properties.
    - In the calculation of the mean: all data points were used. Thus, if data is contaminated by a large outlier, the value of the mean would change dramatically.
  - The median is **robust** which means it is not affected by some extreme readings. This makes the median preferable for data which is skewed or has many outliers (e.g. Sydney house prices).
    - The median calculation used the “middle numbers” after sorting. A small set of large number contamination will not affect the median dramatically.

2. Is it unimodal, bimodal, trimodal, multimodal or other?



Note: Bimodality can be an indication of interesting behaviour to explore. However, it can also arise from 2 populations mistakenly put together.

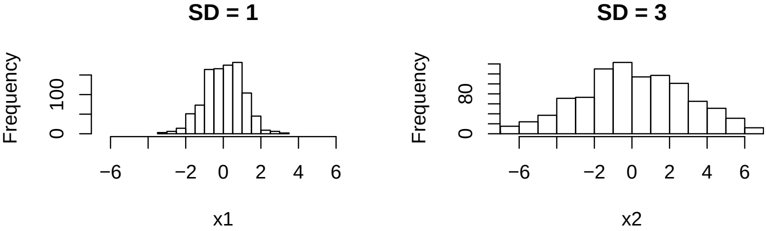
Spread in the data

- What makes the data $x = \{-1, 0, 1\}$ and $y = \{-100, 0, 100\}$ different?
- Using both mean and median, we can conclude both data has the same centre. But how far is the data from this centre? That is, what is the **spread** of the data.
- Just like locations, there are many different ways of measuring spread of the data.
- E.g. The range of the data, defined as $\max(x) - \min(x)$ is one very simple way of measuring spread.
- Another very popular measure is **variance**. It represents the average of the squared deviation from the mean:

$$
\begin{aligned}
Var(x) &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \\
&= \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.
\end{aligned}
$$

- But variance is not on the same scale as the mean, so we often report the square root of variance instead.
- **Standard deviation (SD)**

$$sd(x) = \sqrt{Var(x)}$$



Using the Standard Deviation

Note that s has the same units as $\bar{x}$, so we can couple $(\bar{x}, s)$ as a summary of centre and spread. Thus, we often report this pair to give a sense of variabilities in the data.

Calculating Standard Deviation

Given $\{1, 4, 6, 2, 3, 7\}$ with $\bar{x} = 23/6$, what is the standard deviation?

Definition formula:

$$s = \sqrt{\frac{1}{5}[(1-23/6)^2 + (4-23/6)^2 + \ldots (7-23/6)^2]} \approx 2.32$$

Calculation formula:

$$s = \sqrt{\frac{1}{5}[1^2 + 4^2 + 6^2 + 2^2 + 3^2 + 7^2 - 6(23/6)^2]} \approx 2.32$$

```
x =  c(1,4,6,2,3,7)
sd(x)

## [1] 2.316607
```

Spread - based on the Quartiles (IQR)

The quartiles are a set of 3 values $\{Q_1, Q_2 = \tilde{x}, Q_3\}$ that roughly split the data into quarters.

There is no universal way to define quartiles. We use the following convention: we divide the data into 2 sets at the median (including the median for an odd sized data set), and then find the median of each half set of data.

Once we have found Q_1, we can find Q_3 by symmetry, by counting back from the end of sorted data set.

See more definitions: ▸ Quartiles

Calculating the Quartiles (even sized sample)

Given $\{1, 4, 6, 2, 3, 7\}$, the sorted data is $\{1, 2, 3, 4, 6, 7\}$ and the median $Q_2 = 3.5$ splits the data into $\{1, 2, 3\}$ and $\{4, 6, 7\}$, hence $Q_1 = 2$ and $Q_3 = 6$.

```r
# Finds min, Q1, Q2, Q3, max
fivenum(x)

## [1] 1.0 2.0 3.5 6.0 7.0
```

Calculating the Quartiles (odd sized sample)

Given $\{1, 4, 6, 2, 3, 7, 8\}$, the sorted data is $\{1, 2, 3, 4, 6, 7, 8\}$ and the median $Q_2 = 4$ splits the data into $\{1, 2, 3, 4\}$ and $\{4, 6, 7, 8\}$, hence $Q_1 = 2.5$ and $Q_3 = 6.5$.

```r
x2=c(1,4,6,2,3,7,8)
fivenum(x2)
```

```
## [1] 1.0 2.5 4.0 6.5 8.0
```

Interquartile Range (IQR)

The full range of the data is $x_{(n)} - x_{(1)}$, but this ignores $n-2$ data points.

The Interquartile Range is defined as

$$IQR = Q_3 - Q_1$$

and represents the range of the middle 50% of the data.

We couple $(\tilde{x}, IQR)$ as a summary of centre and spread.

```
fivenum(x)[4]-fivenum(x)[2]

## [1] 4
```

Comparing the SD and the IQR

Like the mean and median, the IQR is **robust** and so preferable for data which is skewed or has many outliers. However, the standard deviation is good for theoretical analysis.

Comparing the SD and the IQR

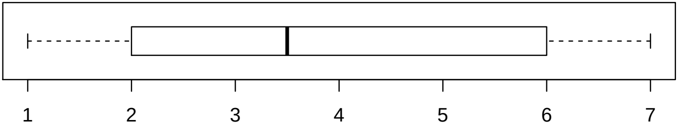
Given $\{1, 4, 6, 2, 3, 7, 100\}$, what is the sd and IQR?

```
x1=c(1,4,6,2,3,7,100)
sd(x1)

## [1] 36.40905

fivenum(x1)[4] - fivenum(x1)[2]

## [1] 4
```

The Five Number Summary

The five number summary $(x_{(1)}, Q_1, Q_2, Q_3, x_{(n)})$ is a neat way to summarise the data and it can be graphically summarized by the **boxplot**.

```
fivenum(x)

## [1] 1.0 2.0 3.5 6.0 7.0

boxplot(x, horizontal=T)
```

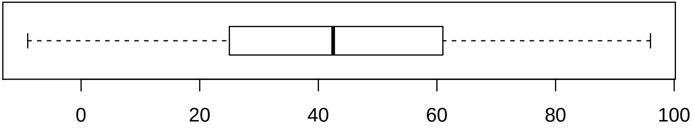


Boxplot

Q: Were there any unusual ages at which a road fatality occurred? Is there any difference between the ages of male and female fatalities?

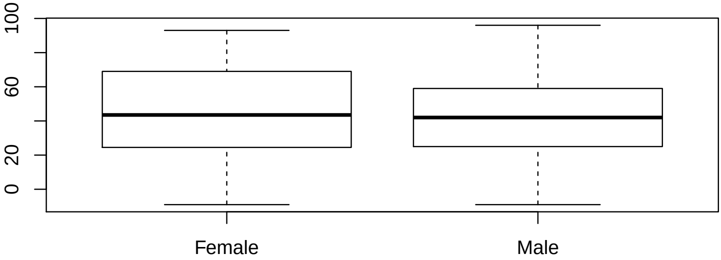
Boxplots are useful for comparing data sets and identifying outliers.

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

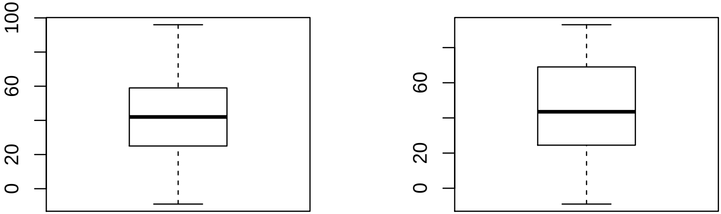


In R, we can provide a data and then asks to split the Age variable (numeric) by the Gender variable (categorical).

```
boxplot(Age~Gender, data = data)
```



```
## Alternative way to plot. We first subset the data
AgeM <- data$Age[ data$Gender == "Male"]
AgeF <- data$Age[ data$Gender == "Female"]
par(mfrow = c(1, 2))  #Puts 2 boxplots in a row
boxplot(AgeM)
boxplot(AgeF)
```



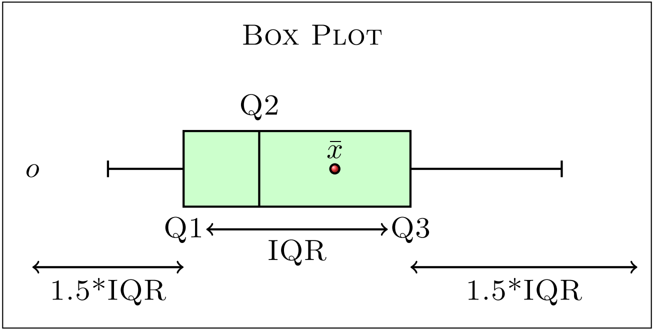
The boxplots show that the ages of road fatalities for men and women is similar.
However, what do we learn about the data from these simple outputs?

```
length(AgeM)

## [1] 326

length(AgeF)

## [1] 116
```

There are different conventions for boxplots. We will use the convention that the whiskers extend to the minimum and maximum observations within the thresholds [LT,UT], where

- Lower Threshold $LT = Q_1 - 1.5IQR$;
- Upper Threshold $UT = Q_3 + 1.5IQR$;
- Interquartile range is $IQR = Q_3 - Q_1$.

An outlier is any observation lying outside of [LT,UT].



Note: Here we have indicated the mean $\bar{x}$ in red for comparision with the median Q_2, but normally that is not shown on the boxplot.

Steps for Constructing a Boxplot by Hand

1. Calculate the quartiles Q_1, Q_2 and Q_3 and the interquartile range IQR.
2. Draw a box from Q_1 to Q_3, with a line within the box for the median$= Q_2$.
3. Calculate the upper and lower thresholds.
4. Draw a whisker from the box to the nearest points within the thresholds.
5. Any points outside the thresholds are outliers, designated by circles.

Extra reading

- I will leave some materials for take-home reading. They are usually light and for your interest only.
- You should run these codes where possible to get a better understanding of R.

1. Other measures of spread

- **Mean Absolute Deviation (MAD)**

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

This is messy algebraically.

- **Mean Square Error (MSE)**

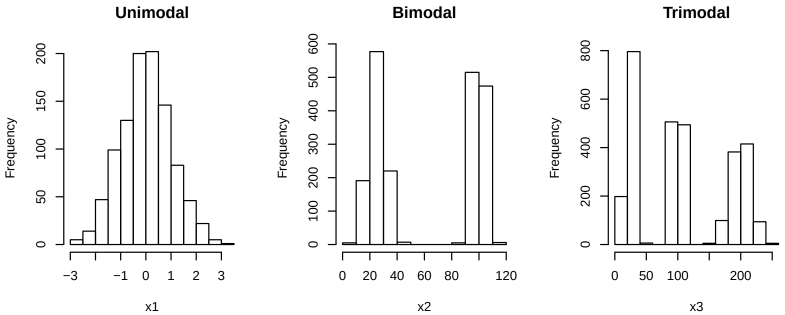
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

This requires that we sample $\bar{x}$ from the sample before calculating the MSE: only $n-1$ of the observations are independent of each other.

2. Examples in multi-modality

unimodal, bimodal, trimodal....

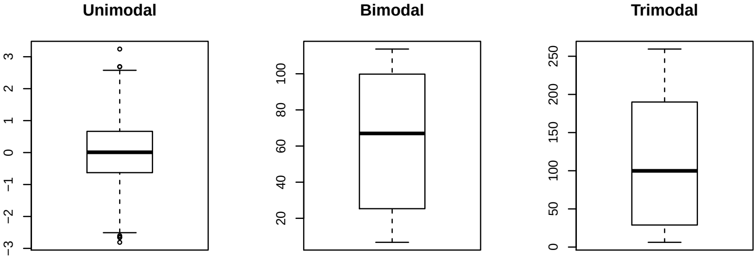
```r
set.seed(123)
x1=rnorm(1000)
x2=c(rnorm(1000,25,6),rnorm(1000,100,4))
x3=c(rnorm(1000,25,6),rnorm(1000,100,4), rnorm(1000,200,16))
par(mfrow = c(1, 3))
hist(x1,main="Unimodal")
hist(x2,main="Bimodal")
hist(x3,main="Trimodal")
```



2. Examples in multi-modality

unimodal, bimodal, trimodal....

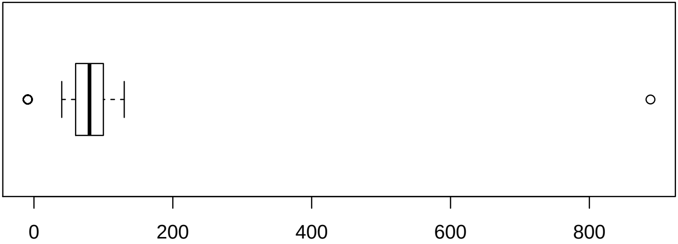
```r
set.seed(123)
x1=rnorm(1000)
x2=c(rnorm(1000,25,6),rnorm(1000,100,4))
x3=c(rnorm(1000,25,6),rnorm(1000,100,4), rnorm(1000,200,16))
par(mfrow = c(1, 3))
boxplot(x1,main="Unimodal")
boxplot(x2,main="Bimodal")
boxplot(x3,main="Trimodal")
```



3. Outliers

Q: What were there any unusual speeds at which fatalities occurred?

```
boxplot(data$SpeedLimit, horizontal = T)
```



Identifying Outliers

Outliers are 'unusual values' that do not fit the model. They can either indicate interesting values that need futher investigation or a transformation of the model, or they can indicate a possible mistake in your data.

Ways to identify outliers:

- **The IQR method (Tukey)**
  As outlined in the boxplot, we calculate the lower and upper thresholds

  $$LT = Q_1 - 1.5IQR \text{ and } UT = Q_3 + 1.5IQR$$

  Any data point lying outside these thresholds is deemed an outlier.
  Disadvantages: No outliers detected for $n \leq 4$ and for large samples wrongly identifies outliers.

Identifying Outliers

- **(Extension: The 3-σ method)**
  Any data point lying more than 3 standard deviations away from the mean is deemed to be an outlier.

  $$x_i \text{ is an outlier iff } |x_i - \bar{x}| > 3\sigma$$

  Disadvantages: No outliers detected for $n \leq 7$ and for large samples wrongly identifies outliers.

  Note: The 3-sigma edit rule is popular in economics, but it should be avoided in practice due to the following inflexibility, which will make more sense after Part2 of the course.

The 3-σ rule assumes that the underlying distribution is the Normal, and is based on both the sample mean and standard deviation. Problems can occur when either:

1) The data is sufficiently skewed. In this case, the mean is no longer a ‘good’ measure of central tendency, and defining outliers as points outside of some symmetric neighbourhood of the mean is not appropriate. The risk is that the ‘outliers’ are detected near the mode rather than the longer tail. Tukey’s five number approach is less likely to suffer from this.

2) The underlying population has heavy tails. The principle behind the 3-σ rule is that $P(|x_i - \bar{x}| > 3\sigma)$ occurs with small probability, for example when the population is Normal this probability is 0.0027.

```
2*(1-pnorm(3))

## [1] 0.002699796
```

If there are heavy tails then this probability can be substantially larger. For example, the probability is 0.029 when the population is t_3, or 0.10 when the population is t_1. In the latter case 10% of the observations will be deemed 'outliers'!

Detecting Outlier in Data with Mistake

Suppose we made a mistake in the data entry with heights: 1.68 1.58 1.64 1.73 1.60 1.62 1.78 1.69 1.80 1.74 1.71 1.59 1.63 1.77 1.70 1.77 1.63 1.62 1.80 1.70 1.60 1.77 1.79 1.65 1.66 1.60 1.71 **178**

IQR method

```
Usyd <- read.csv("../datasets/USyd.csv")
heights1=c(Usyd$Heights[1:27],178)
iqr=fivenum(heights1)[4]-fivenum(heights1)[2]
lt=fivenum(heights1)[2]-1.5*iqr
ut=fivenum(heights1)[4]+1.5*iqr
heights1[(heights1<lt) | (heights1 > ut)]   # | = ’or’

## [1] 178
```

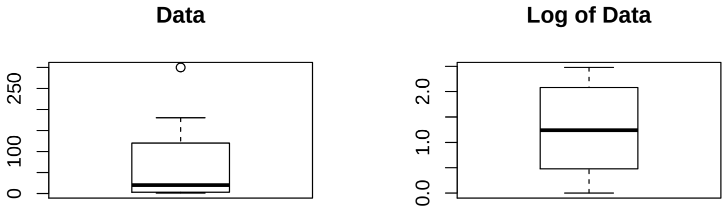
3-σ method

```
3*sd(heights1)

## [1] 99.95986

heights1[abs(heights1-mean(heights1))>3*sd(heights1)]

## [1] 178
```

4. Dealing with Outliers by Transformation

Sometimes an outlier indicates that a better model is needed.

```
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")
boxplot(w1, main="Log of Data")
```


Data
Log of Data