Topic2: Numerical Summaries

Outline

Example: Australian Road Fatalities Jan-April 2016

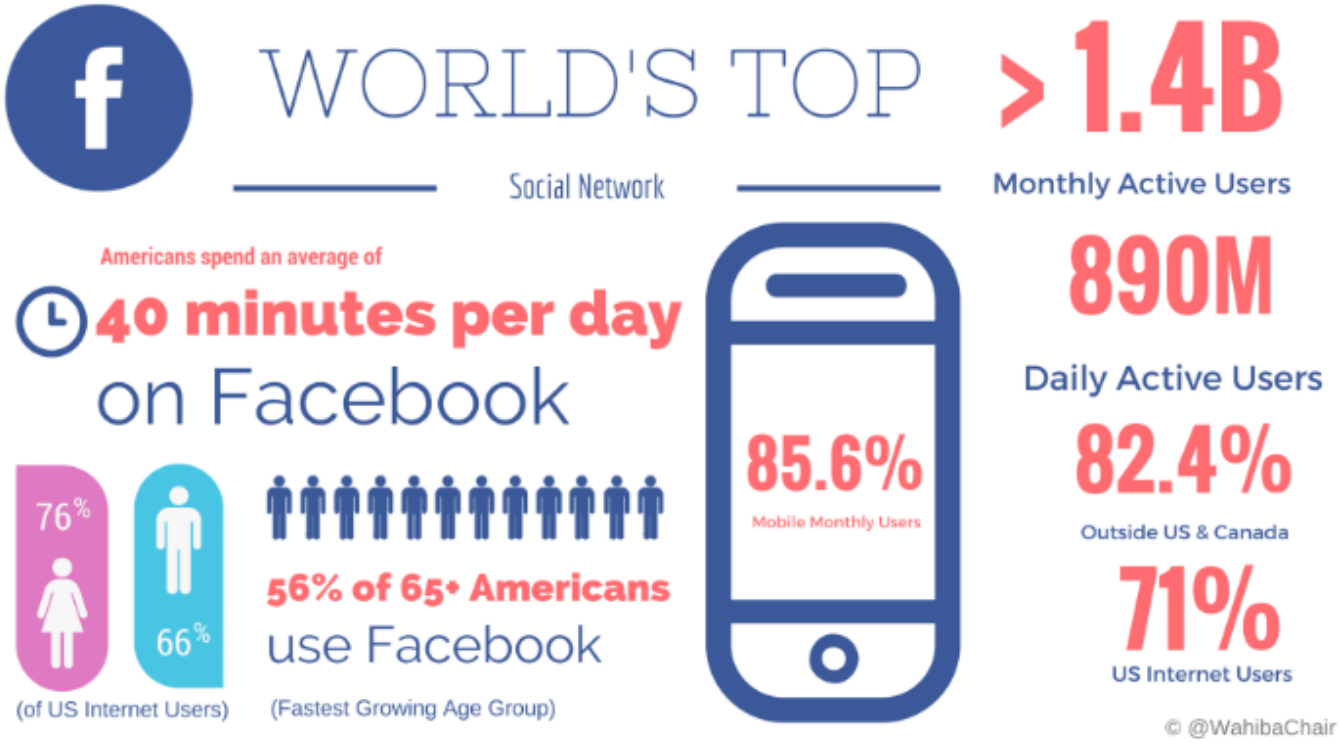
The Australian Road Deaths Database provides basic details of road transport crash fatalities in Australia as reported by the police each month to the State and Territory road safety authorities.

Details provided in the database fall into two groups: (1) the circumstances of the crash, for example, date, location, crash type (2) some details regarding the persons killed, for example, age, gender and road user group.

What is the most common profile of person killed on Australian roads (eg gender, age, time of accident)?

Australian Road Deaths Database

Numerical Summaries



A numerical summary describes a certain characteristic of the data in a single value. The notion of a numerical summary is very nice – what is lost in subtlety is gained in simplicity. For example, in the Australian census data, 1 mean can summarise 21.5 million data points. ▸ SocialMediaStats ▸ AusCensusData

From the Australian Road Fatalities summaries, what do you learn about the variables?

```
summary(data[1:8])
```

```
##     Crash.ID              State             Date           Day
##  Min.   :1.202e+12   NSW    :138   17-Apr-16: 10   Min.   : 1.0
##  1st Qu.:1.202e+12   VIC    :100   20-Feb-16:  9   1st Qu.: 9.0
##  Median :2.202e+12   QLD    : 76   27-Jan-16:  9   Median :15.0
##  Mean   :2.973e+12   WA     : 64   5-Mar-16 :  9   Mean   :15.5
##  3rd Qu.:4.202e+12   SA     : 31   9-Jan-16 :  9   3rd Qu.:22.0
##  Max.   :8.202e+12   TAS    : 18   16-Feb-16:  8   Max.   :31.0
##                      (Other): 15   (Other)  :388
##       Month          Year           Dayweek         Time
##  April   :116   Min.   :2016   Friday   :68   13:00  : 10
##  February:105   1st Qu.:2016   Monday   :50   14:00  : 10
##  January :106   Median :2016   Saturday :85   21:00  : 10
##  March   :115   Mean   :2016   Sunday   :56   8:00   : 10
##                 3rd Qu.:2016   Thursday :58   15:00  :  9
##                 Max.   :2016   Tuesday  :67   17:00  :  8
##                                Wednesday:58   (Other):385
```

```r
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
```

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.

Notation for Numerical Summaries

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

Numerical Data - 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):

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

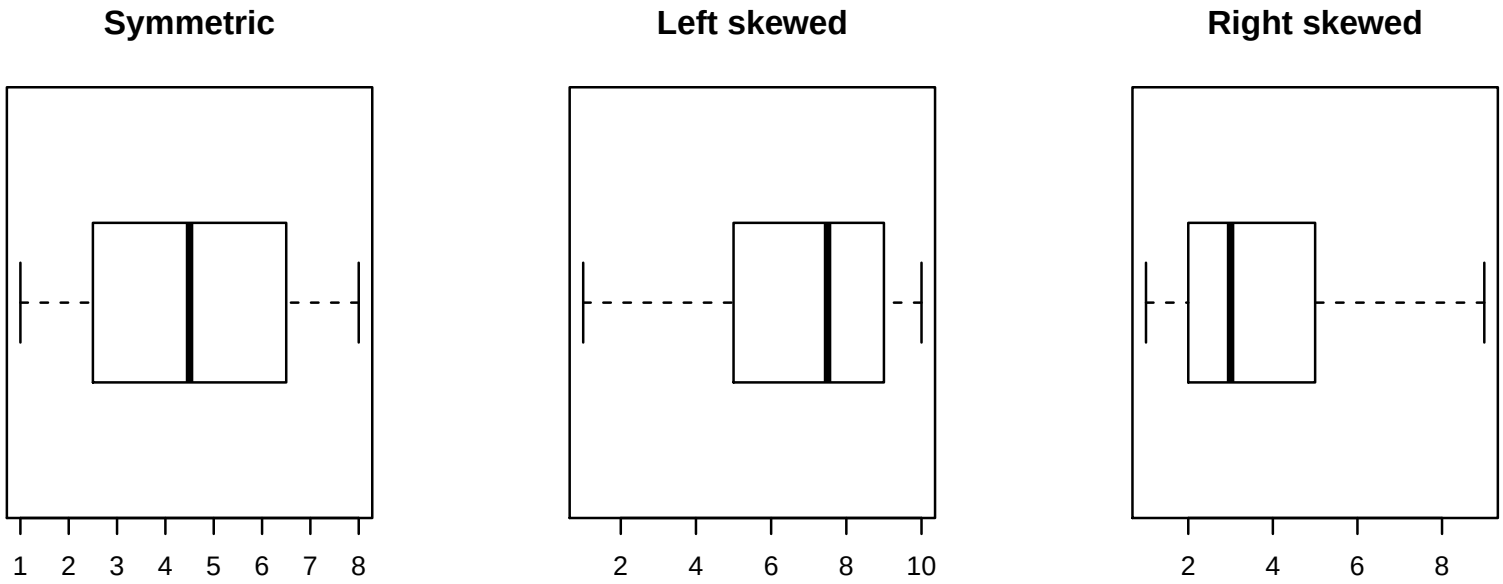
Have a try

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, $\bar{x} > \tilde{x}$.

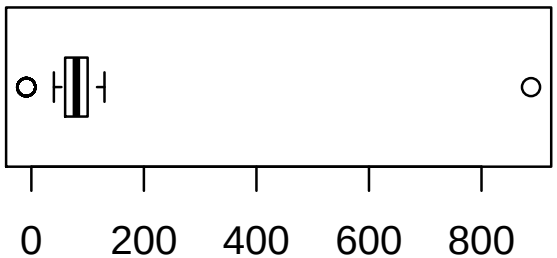


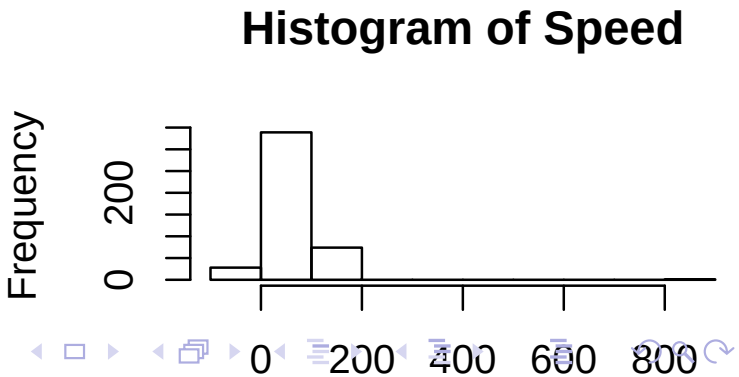
- Which is 'optimal' for describing the centre of the data?

  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.
  - 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 (eg Sydney house prices).
  - The mean is helpful for data which is basically symmetric which not too many outliers, and for theoretical analysis.

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

## [1] 79.76471

median(Speed)

## [1] 80

par(mfrow = c(1, 2))
boxplot(Speed, horizontal=T)
hist(Speed)
```





Numerical Data - Summaries for Spread

Having summarised the centre of the data, we now want to couple this with a summary of the spread of the data: how far is the data from the centre? The combination of both a numerical centre of centre and spread is a surprisingly helpful snapshot of the data.

We can base our summaries for spread on the mean or the quantiles.

Spread - based on the Mean (Standard Deviation)

There are 3 main measures of spread based on the mean:

- **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.

- **Standard deviation (SD)**

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

This is called the definition formula: it represents the average of the squared deviations from the mean $\{(x_i - \bar{x})\}$. We need squared deviations as $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

The calculation formula is

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

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

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.

Calculating Standard Deviation

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

Definition formula:

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

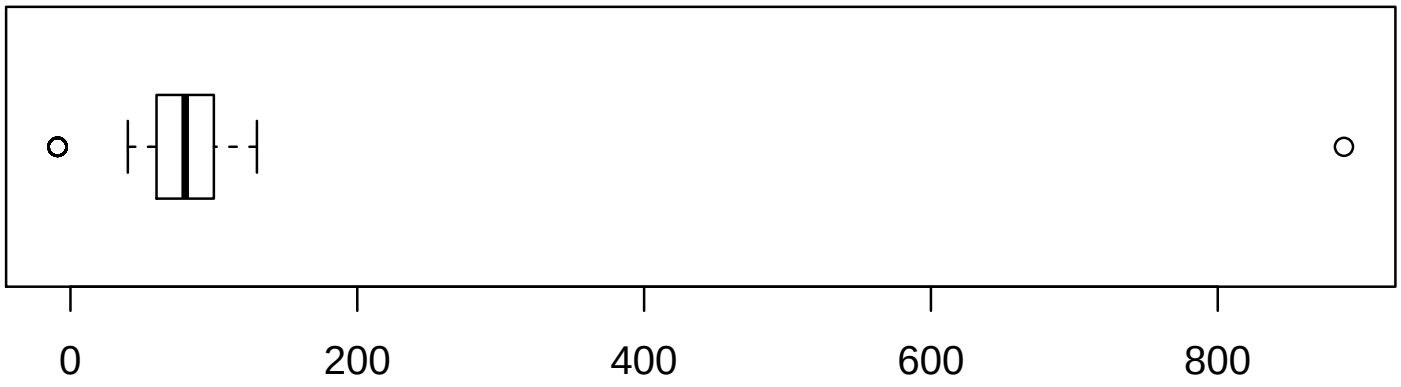
Calculation formula:

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

```
sd(x)

## [1] 2.316607
```

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

## [1] 79.76471

median(Speed)

## [1] 80

sd(Speed)

## [1] 49.54275

iqr=fivenum(Speed)[4]-fivenum(Speed)[2]
iqr

## [1] 40
```

```r
fivenum(Speed)

## [1]  -9  60  80 100 888

boxplot(Speed, horizontal=T)
```



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.

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 $Q_3 = 6$.

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

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

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 $Q_3 = 6$.

```
# 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 $Q_3 = 6.5$.

```
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]    # Don't use iqr()

## [1] 4
```

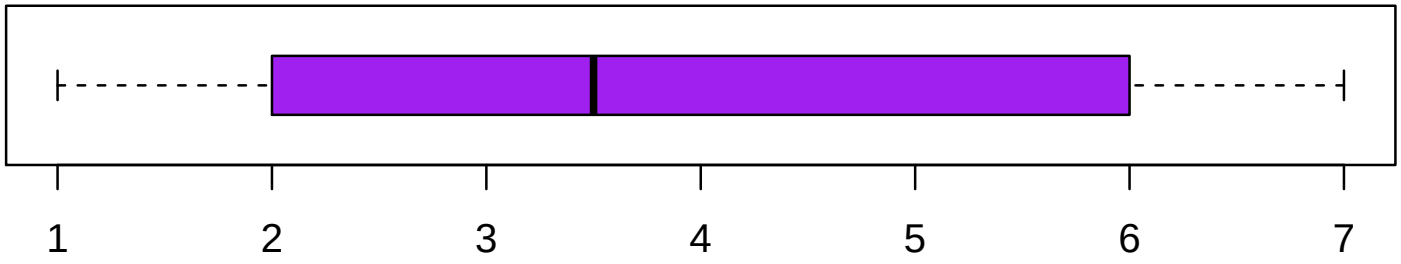
The Five Number Summary

The five number summary is a neat way to summarise the data

$$(x_{(1)}, Q_1, Q_2, Q_3, x_{(n)})$$

and is essentially drawn by the boxplot.

```
fivenum(x)

## [1] 1.0 2.0 3.5 6.0 7.0

boxplot(x, horizontal=T, col="purple")
```



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
```

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.

There are 2 main 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("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
```

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")
```