Introduction to Statistics for Psychology

and

Quantitative Methods for Human Sciences

Jonathan Marchini

Course Information

There is website devoted to the course at

`http://www.stats.ox.ac.uk/∼marchini/phs.html`

This contains

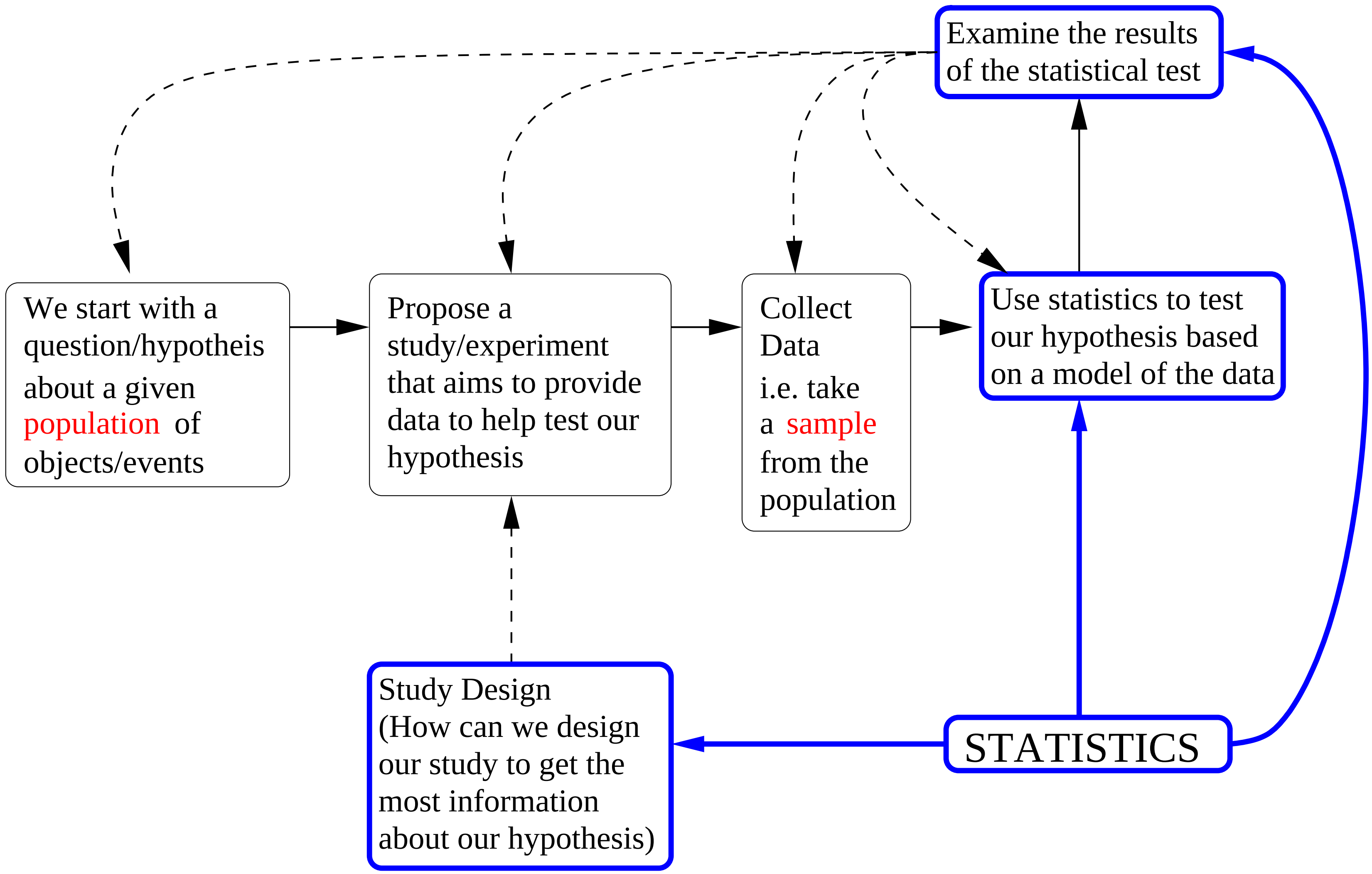
- Course timetable/information
- Lecture slides (on the website before each lecture)
- Lecture notes (a bit more detailed than the slides)
- Exercise sheets (tutors may or may not use these)
- Formulae booklet and definitions booklet
- Links to past exam papers on the web

PLEASE ENTER THE LECTURE THEATRE QUICKLY!

Lecture 1 : Outline

- Why we need Statistics?
  - The scientific process
- Different types of data
  - Discrete/Continuous
  - Quantitative/Qualitative
- Methods of looking at data
  - Bar charts, Histograms, Dot plots, Scatter plots, Box plots
- Calculating summary measures of data
  - Location - Mean, Median, Mode
  - Dispersion - SIQR, MAD, sample variance, sample standard deviation

The role of Statistics in the Scientific Process



An Example

Psychologists have long been interested in the relationship between stress and health.

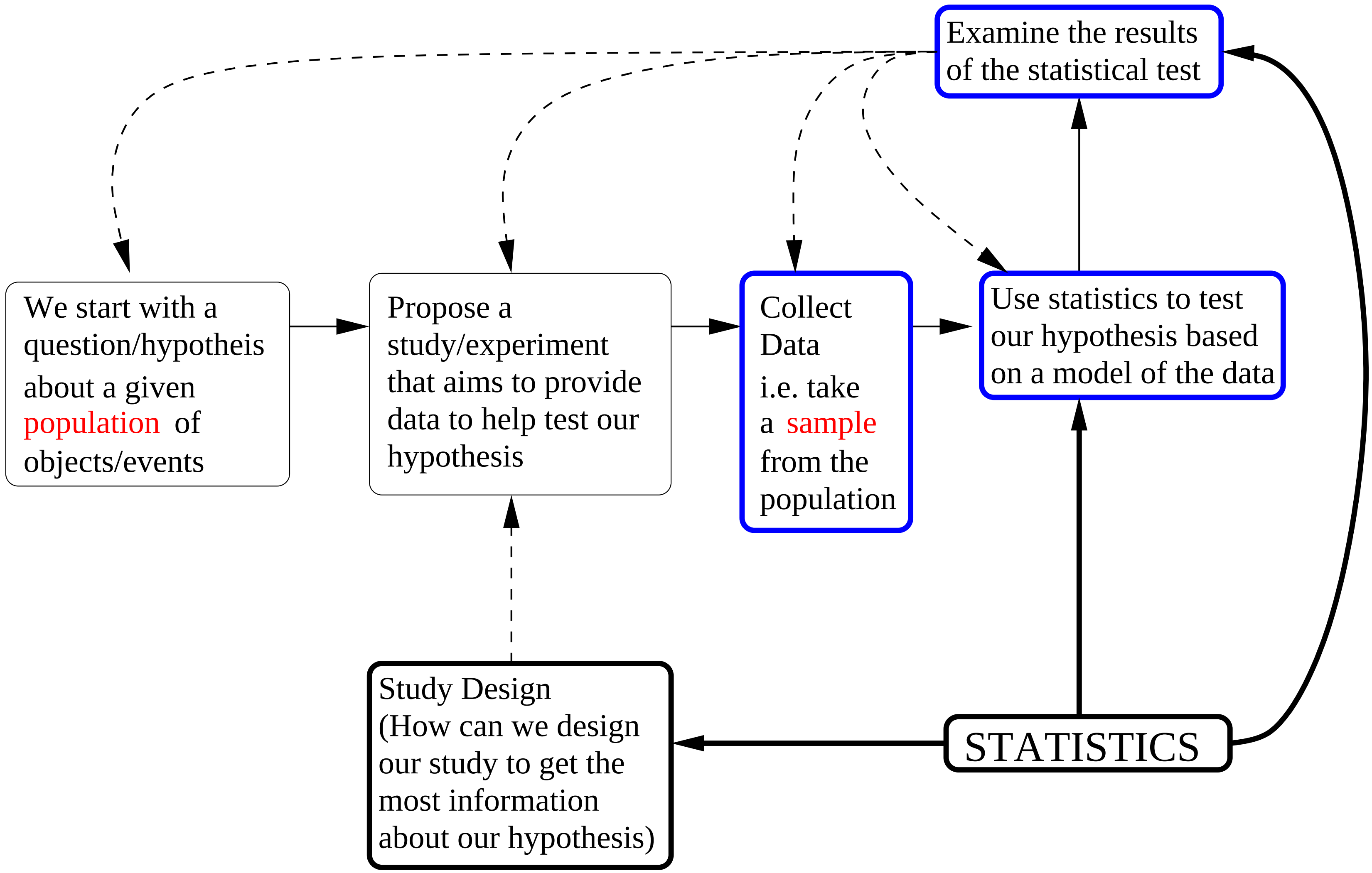
A focused question might involve the study of a specific psychological symptom and its impact on the health of the population.

To assess whether the symptom is a good indicator of stress we need to measure the symptom and stress levels in a sample of individuals from the population.

It is not immediately clear how we should go about collecting this sample, i.e. how we should design the study.

We haven’t got very far before we need Statistics!

The general focus of this course



Datasets consist of measured variables

The datasets that Psychologists and Human Scientists collect will usually consist of one more observations on one or more “variables”.

A **variable** is a property of an object or event that can take on different values.

Example Suppose we collect a dataset by measuring the hair colour, resting heart rate and score on an IQ test of every student in a class. The variables in this dataset would then simply be hair colour, resting heart rate and score on an IQ test, i.e. the variables are the properties that we measured/observed.

2 main types of variable

1 Measurement (Quantitative) Data occur when we ‘measure’ things e.g. height or weight.

2 Categorical (Qualitative) Data occur when we assign objects into labelled groups or categories e.g. when we group people according to hair colour or race.

(i) **Ordinal variables** have a natural ordering e.g. gold/silver/bronze medal

(i) **Nominal variables** do not have a natural ordering e.g. gender

Discrete and Continuous Variables

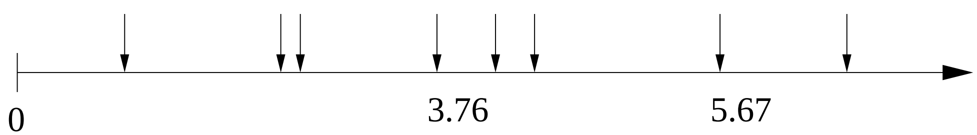
Discrete Data

No. of students late for a lecture

There are only a limited set of distinct values/categories
i.e. we can’t have exactly 2.23 students late, only integer values are allowed.

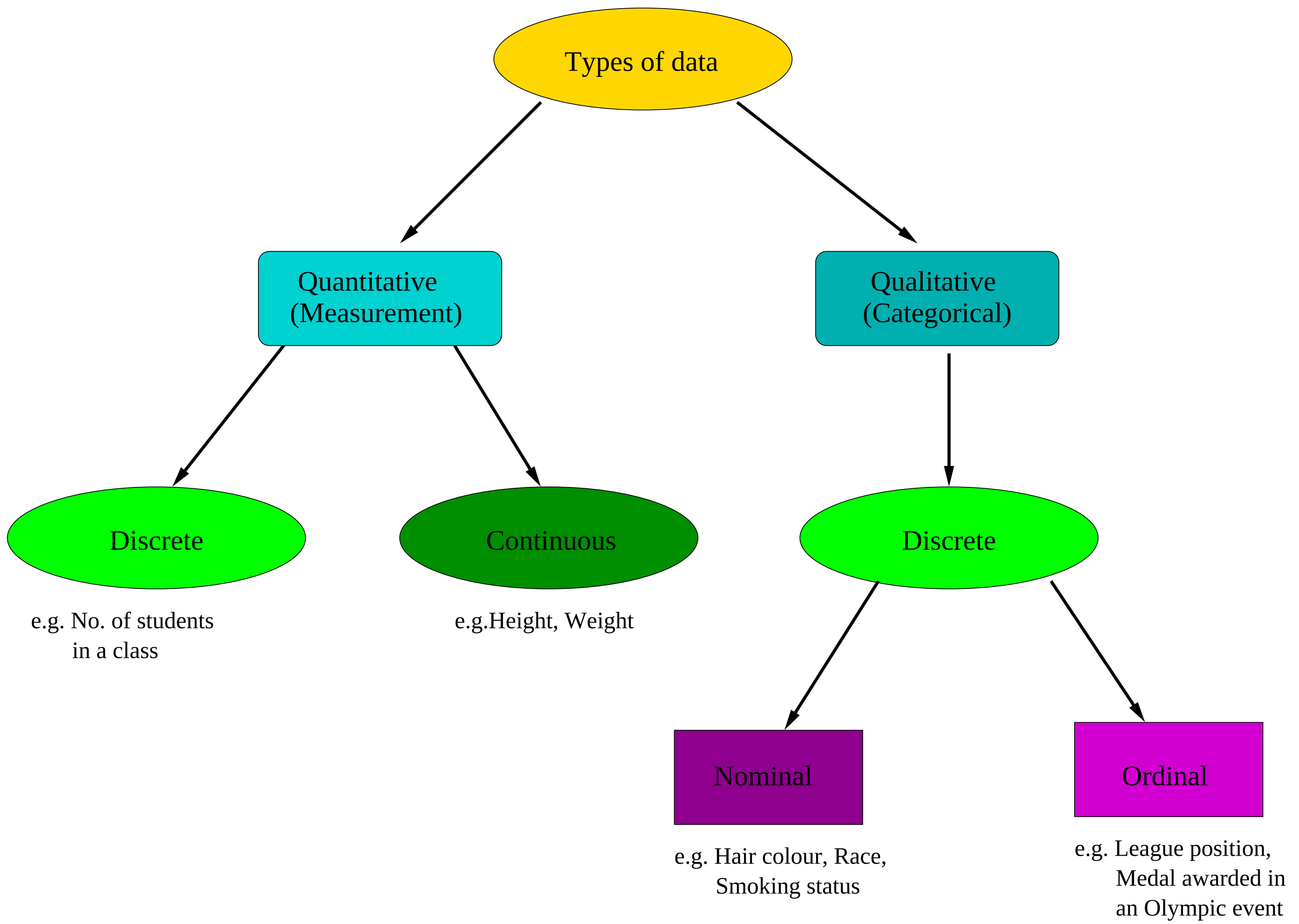
Continuous Data

Time spent studying statistics (hrs)



In theory there are an unlimited set of possible values!
There are no discrete jumps between possible values.

Summary of Data Types



Plotting Data

One of the most important stages in a statistical analysis can be simply to look at your data right at the start.

By doing so you will be able to spot characteristic features, trends and outlying observations that enable you to carry out an appropriate statistical analysis.

Also, it is a good idea to look at the results of your analysis using a plot. This can help identify if you did something that wasn't a good idea!

REMEMBER Data is messy! No two datasets are the same

ALWAYS LOOK AT YOUR DATA

The Baby-Boom dataset

Forty-four babies (a new record) were born in one 24-hour period at the Mater Mothers' Hospital in Brisbane, Queensland, Australia, on December 18, 1997. For each of the 44 babies, The Sunday Mail recorded the time of birth, the sex of the child, and the birth weight in grams.

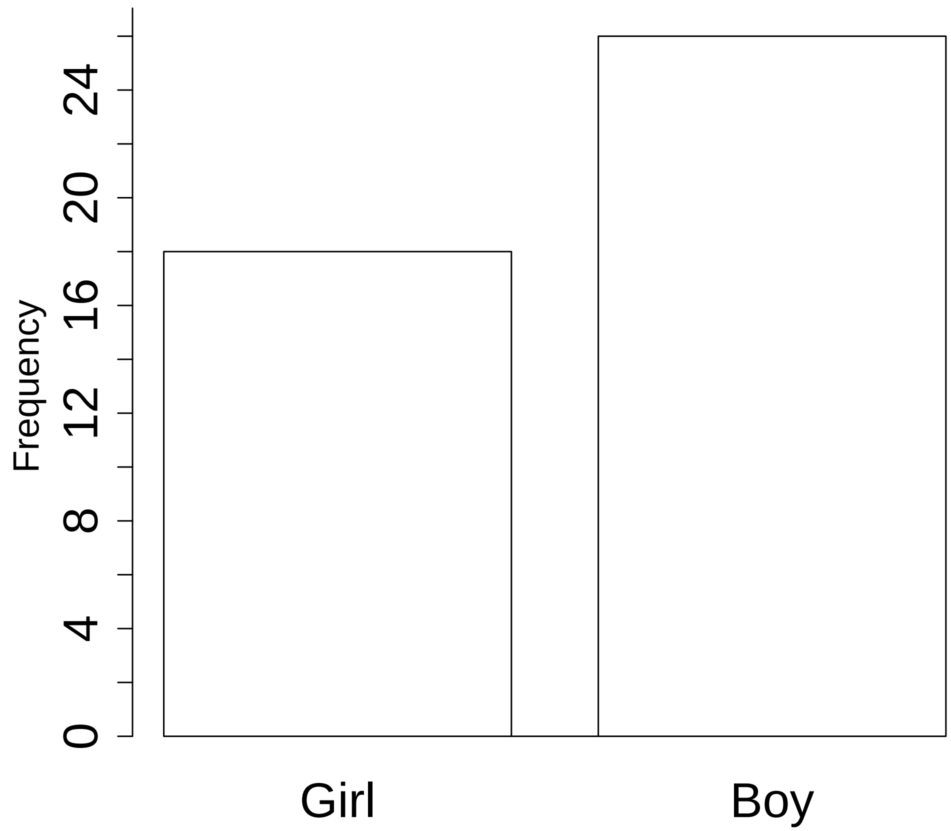
Whilst, we did not collect this dataset based on a specific hypothesis, if we wished we could use it to answer several questions of interest.

- Do girls weigh more than boys at birth?
- What is the distribution of the number of births per hour?
- Is birth weight related to the time of birth?
- Is gender related to the time of birth?
- Is there an equal chance of being born a girl or boy?

Time	Gender	Weight
5	1	3837
64	1	3334
78	2	3554
115	2	3838
177	2	3625
245	1	2208
247	1	1745
262	2	2846
271	2	3166
428	2	3520
455	2	3380
492	2	3294
494	1	2576
549	1	3208
635	2	3521
649	1	3746
653	1	3523
693	2	2902
729	2	2635
776	2	3920
785	2	3690
846	1	3430
847	1	3480
873	1	3116
886	1	3428
914	2	3783
991	2	3345
1017	2	3034
1062	1	2184
1087	2	3300
1105	1	2383
1134	2	3428
1149	2	4162
1187	2	3630
1189	2	3406
1191	2	3402
1210	1	3500
1237	2	3736
1251	2	3370
1264	2	2121
1283	2	3150
1337	1	3866
1407	1	3542
1435	1	3278

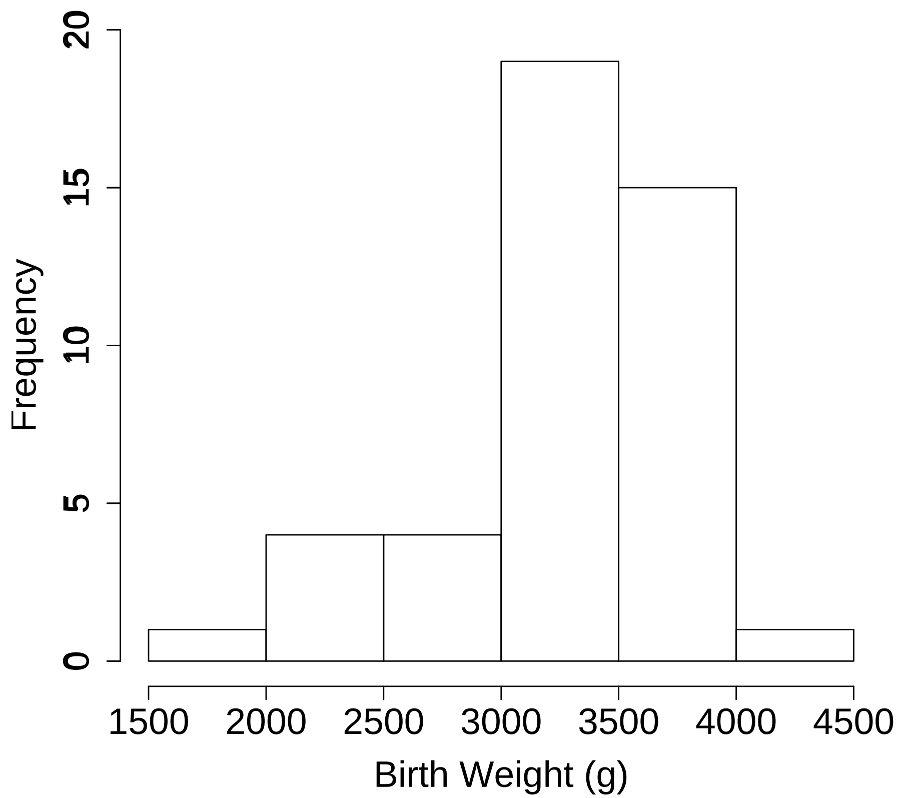
Bar Charts

A Bar Chart is a useful method of summarising Categorical Data. We represent the counts/frequencies/percentages in each category by a bar.



Histograms

'A Bar Chart is to Categorical Data as a Histogram is to Measurement Data'



Constructing Histograms (an example)

For the baby-boom dataset we can draw a histogram of the birth weights.

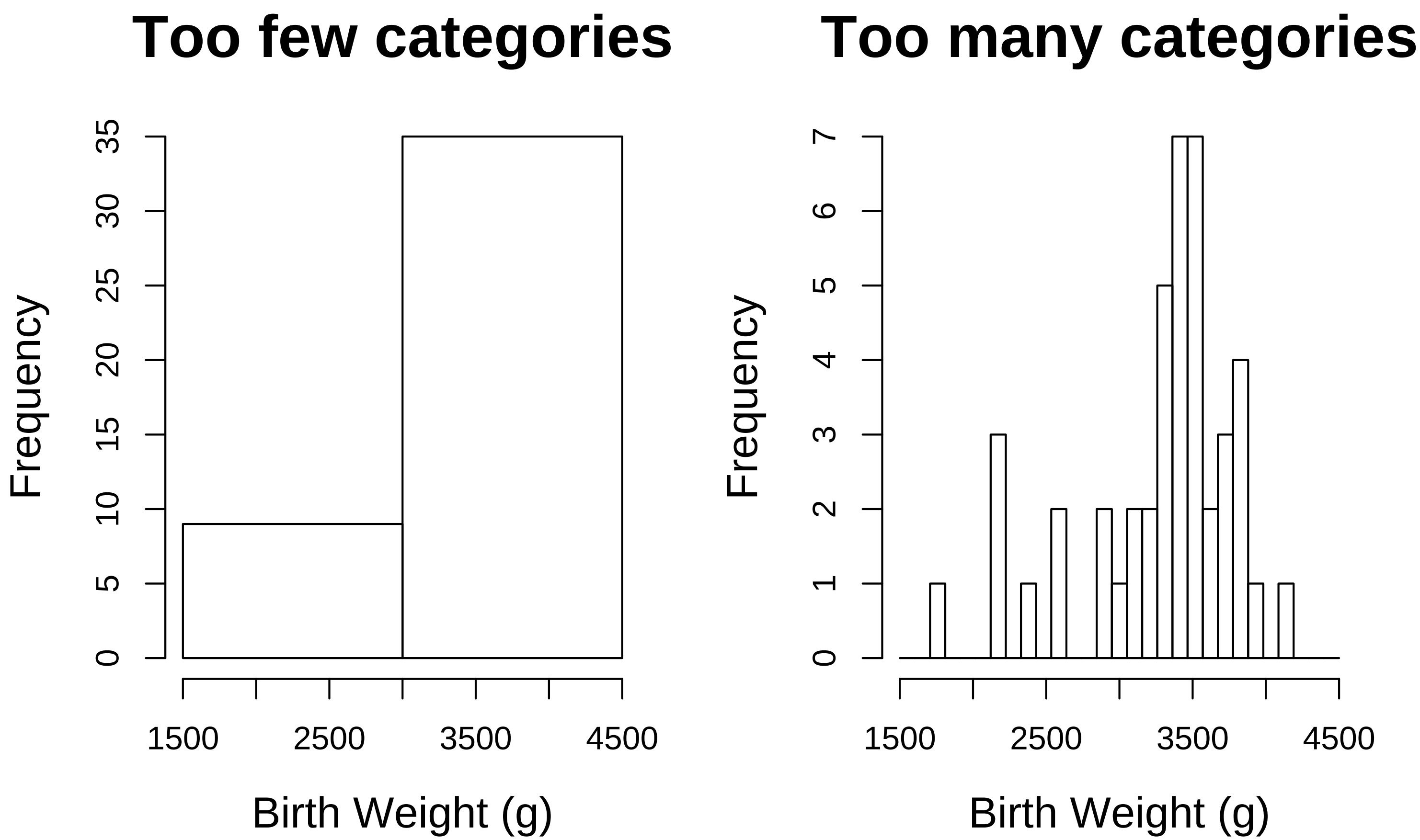
To draw the histogram I found the smallest and largest values

smallest = 1745 largest = 4162

There are only 44 weights so I decided on 6 equal sized categories

Interval	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500
Frequency	1	4	4	19	15	1

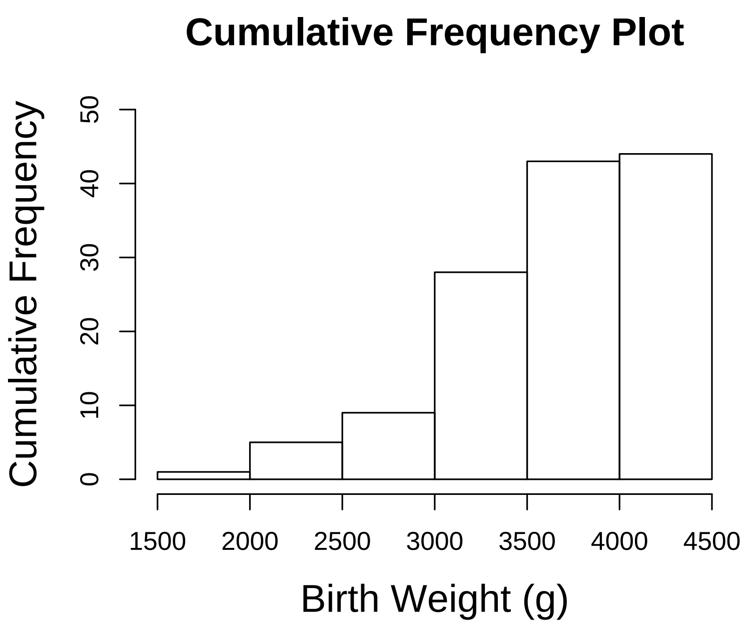
Using these categories works well, the histogram shows us the shape of the distribution and we notice that distribution has an extended left 'tail'.

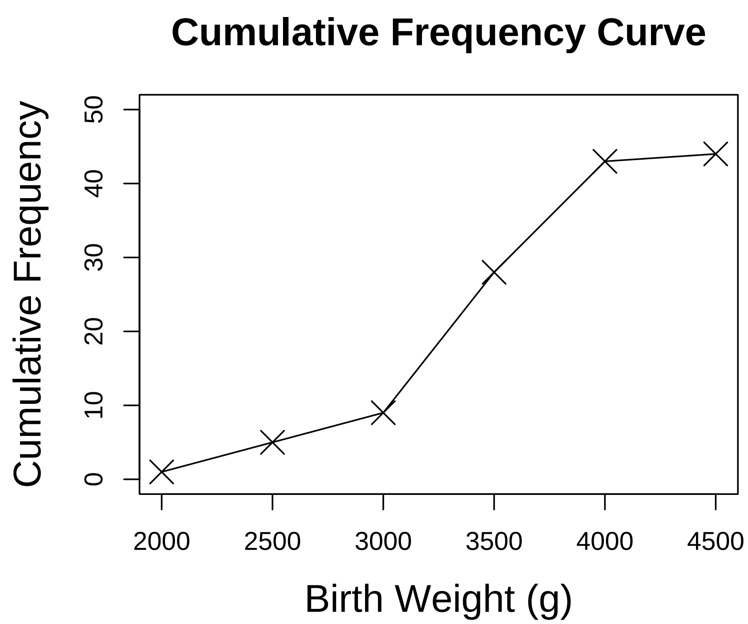


Too few categories and the details are lost. Too many categories and the overall shape is obscured by too many details

Cumulative Frequency Plots and Curves

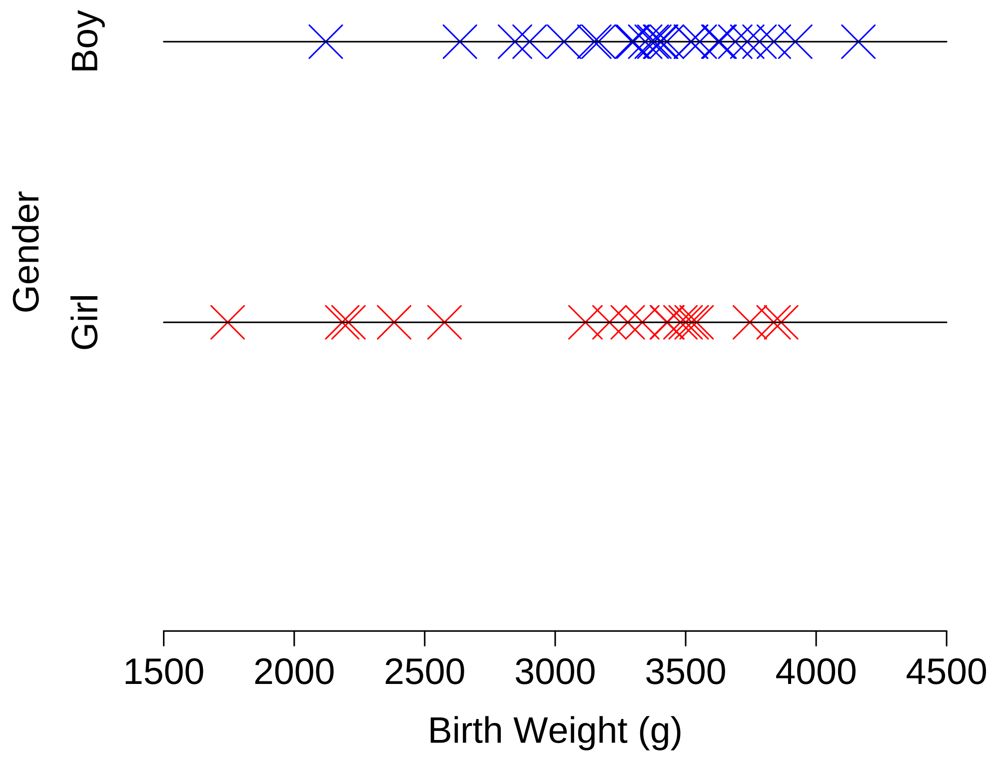
Interval	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500
Frequency	1	4	4	19	15	1
Cumulative Frequency	1	5	9	28	43	44





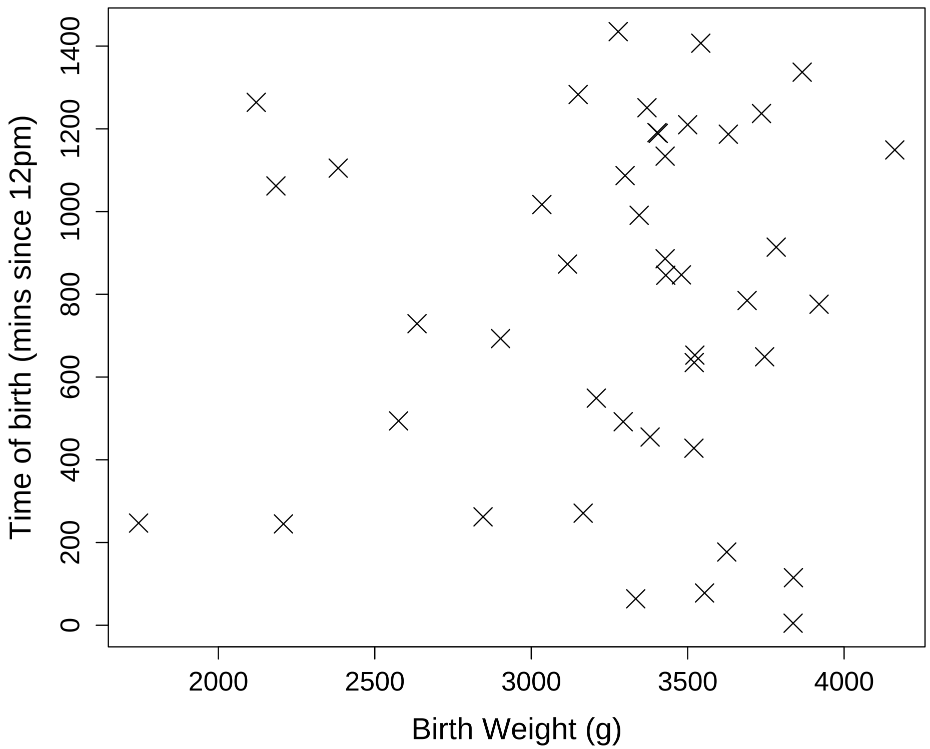
Dot plots

A Dot Plot is a simple and quick way of visualising a dataset. This type of plot is especially useful if data occur in groups and you wish to quickly visualise the differences between the groups.

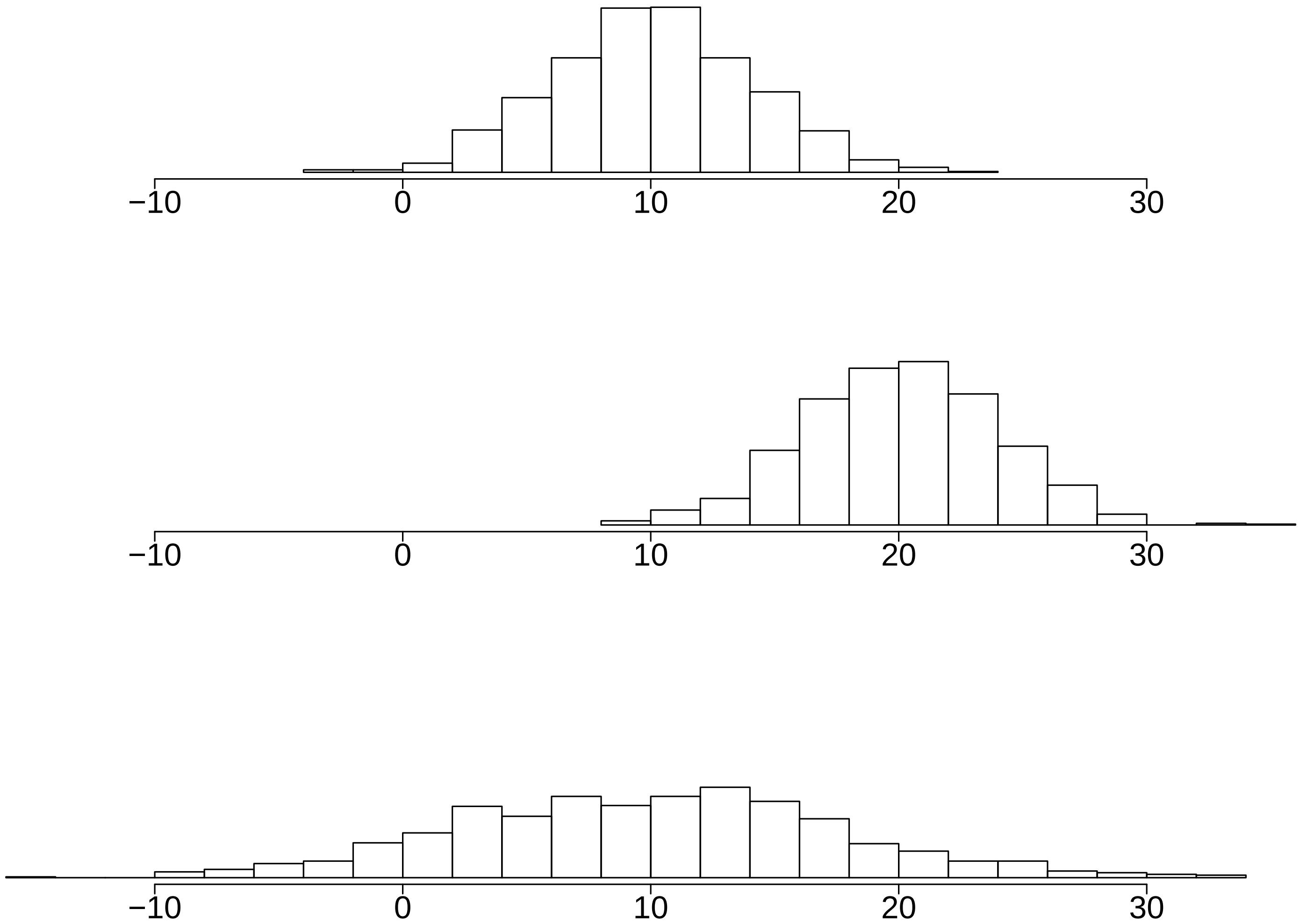


Scatter Plots

Scatter plots are useful when we wish to visualise the relationship between two measurement variables.



Comparing Histograms



Summary Measures

There are 3 main measures of location

- The Mode
- The Median
- The Mean

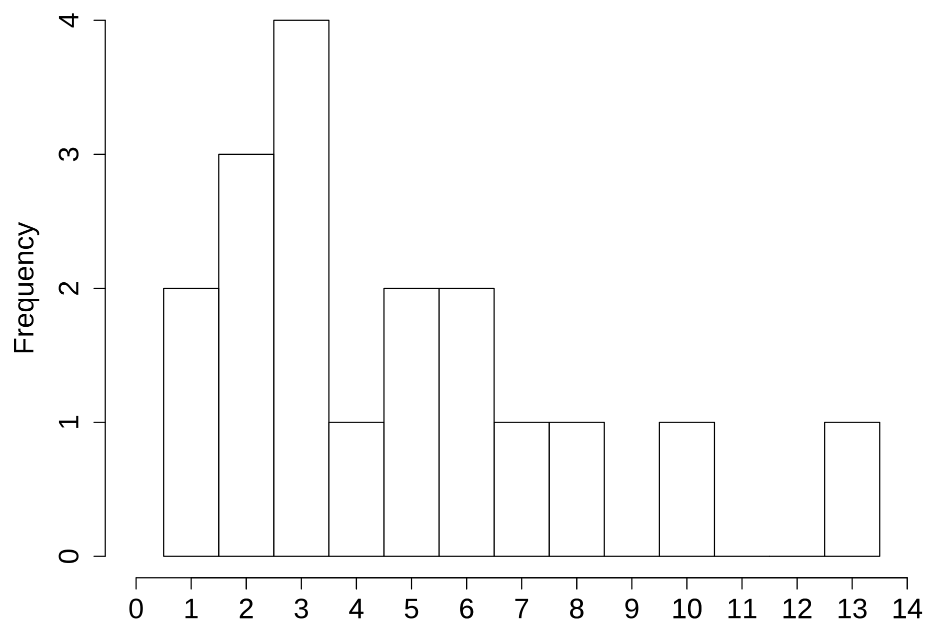
There are 5 main measures of dispersion

- The Interquartile Range (IQR) and Semi-Interquartile Range (SIQR)
- The Mean Deviation
- The Mean Absolute Deviation (MAD)
- The Sample Variance (s^2) and Population Variance (σ^2)
- The Sample Standard Deviation (s) and Population Standard Deviation (σ)

The Mode

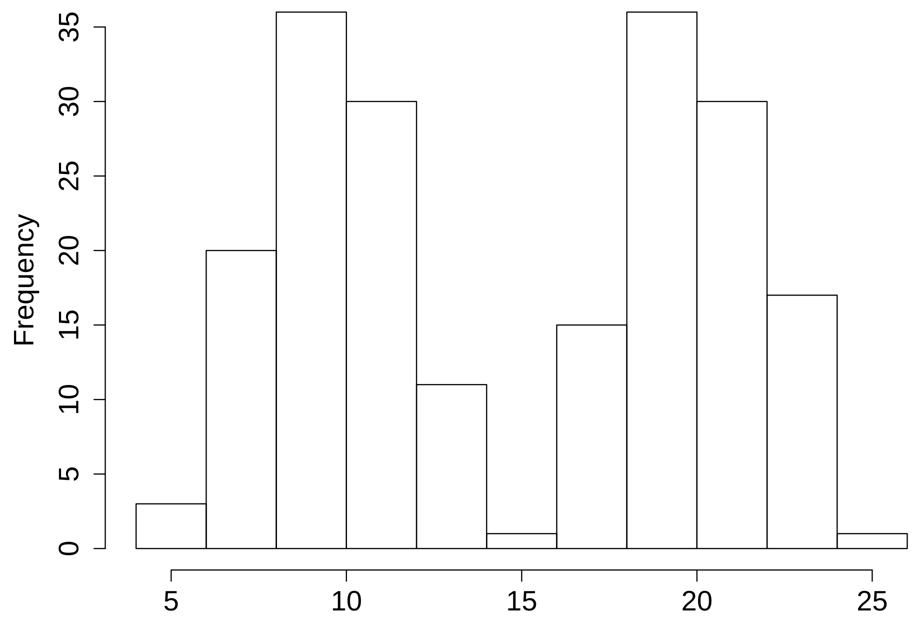
The Mode of a set of numbers is simply the most common value e.g. the mode of the following set of numbers

$$1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6, 7, 8, 10, 13 \qquad \text{is} \quad 3.$$



Advantage The Mode has the advantage that it is always a score that actually occurred and can be applied to nominal data.

Disadvantage There may be two or more values that share the largest frequency. In the case of two modes we would report both and refer to the distribution as *bimodal*.



The Median

The Median can be thought of as the 'middle' value i.e. the value for which 50% of the data fall below when arranged in numerical order.

For example, consider the numbers

$$15, 3, 9, 21, 1, 8, 4,$$

When arranged in numerical order

$$1, 3, 4, \boxed{8}, 9, 15, 21$$

we see that the median value is 8.

If there were an even number of scores e.g.

$$1, 3, \boxed{4,8}, 9, 15$$

then we take the midpoint of the two middle values. In this case the median is $(4 + 8)/2 = 6$.

In general, if we have N data points then the **median location** is defined as follows:

$$\text{Median Location} = \frac{(N+1)}{2}$$

For example, the median location of 7 numbers is $(7 + 1)/2 = 4$ and the median of 8 numbers is $(8 + 1)/2 = 4.5$ i.e. between observation 4 and 5 (when the numbers are arranged in order).

Advantage The median is unaffected by extreme scores (a point it shares with the mode). We say the median is **resistant** to outliers. For example, the median of the numbers

$$1, 3, 4, \boxed{8}, 9, 15, 99999$$

is still 8. This property is very useful in practice as *outlying* observations can and do occur (Data is messy remember!).

The Mean

The Mean of a set of scores is the sum[1] of the scores divided by the number of scores. For example, the mean of

$$1, 3, 4, 8, 9, 15 \quad \text{is} \quad \frac{1+3+4+8+9+15}{6} = 6.667 \quad \text{(to 3 dp)}$$

In mathematical notation, the mean of a set of n numbers $x_1, \ldots, x_n$ is denoted by $\bar{x}$ where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{or} \qquad \bar{x} = \frac{\sum x}{n} \qquad \text{(in the formula book)}$$

See the appendix of the notes for a brief description of the summation notation $(\sum)$

---

[1] The total when we add them all up

Advantage The mean is the most widely used measure of location. Historically, this is because statisticians can write down equations for the mean and derive nice theoretical properties for the mean, which are much harder for the mode and median.

Disadvantage The mean is *not resistant* to outlying observations. For example, the mean of

$$1, 3, 4, 8, 9, 15, 99999$$

is 14323.57, whereas the median (from above) is 8.

Sometimes discrete measurement data are presented in the form of a frequency table in which the frequencies of each value are given.

Data (x)	1	2	3	4	5	6
Frequency (f)	2	4	6	7	4	1

We calculate the sum of the data as

$$(2 \times 1) + (4 \times 2) + (6 \times 3) + (7 \times 4) + (4 \times 5) + (1 \times 6) = 82$$

and the number of observations as

$$2 + 4 + 6 + 7 + 4 + 1 = 24$$

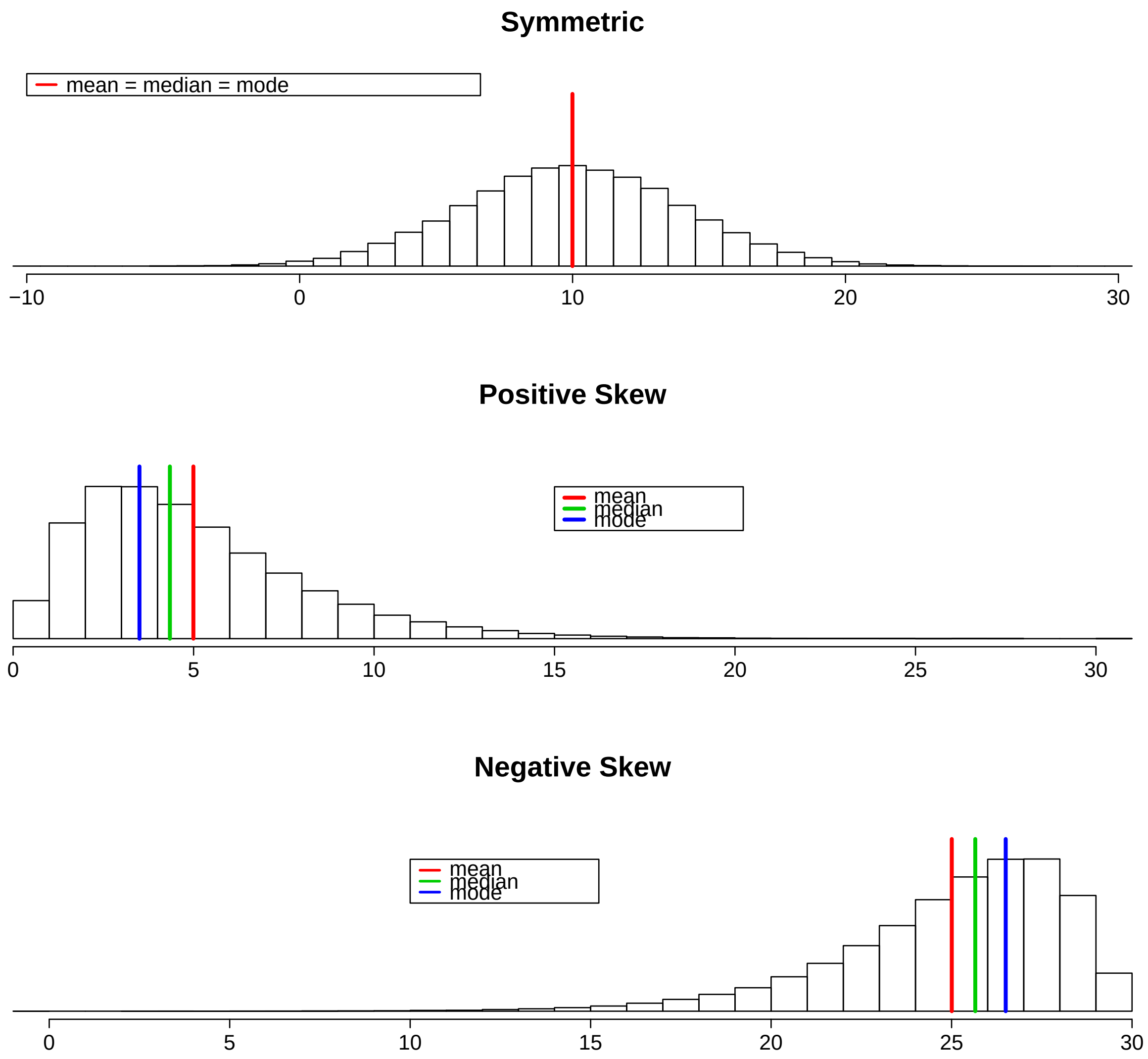
The the mean is given by

$$\bar{x} = \frac{82}{24} = 3.42 \quad \text{(2 dp)}$$

In mathematical notation the formula for the mean of frequency data is given by

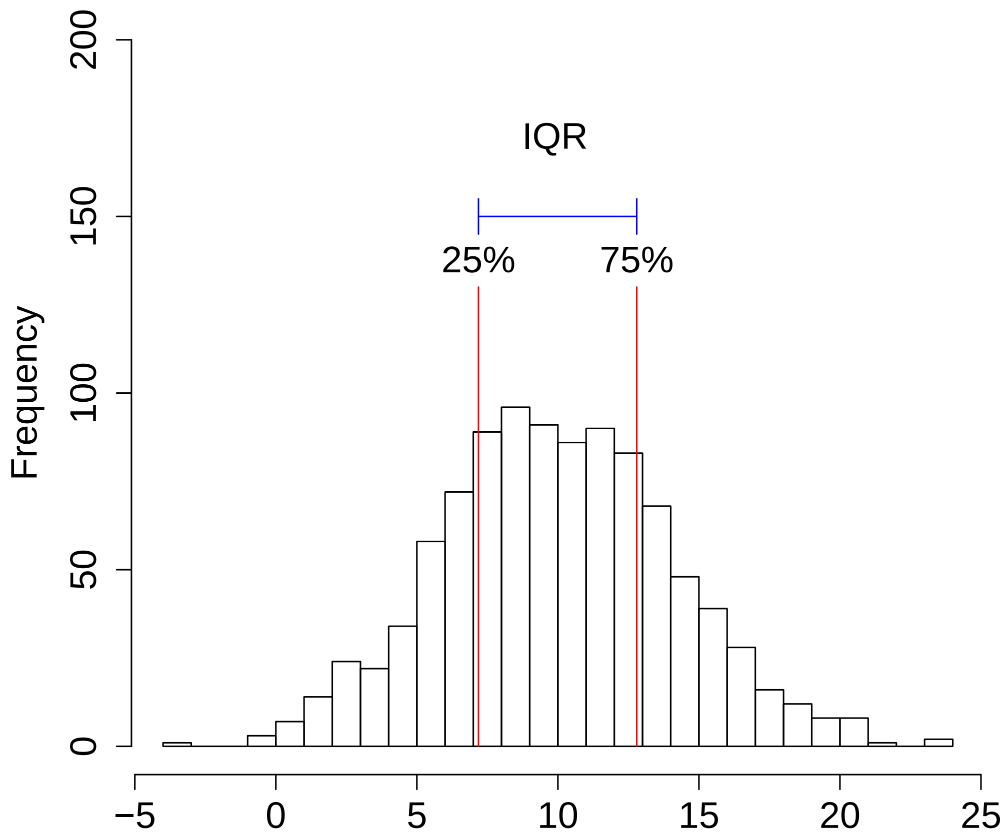
$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \qquad \text{or} \qquad \bar{x} = \frac{\sum fx}{\sum f}$$

The relationship between the mean, median and mode



IQR and SIQR

The IQR is the range of the middle 50% of the data.
The SIQR is simply half the IQR.



We calculate the IQR in the following way:

- Calculate the 25% point (**1st quartile**) of the dataset. The location of the 1st quartile is defined to be the $\left(\frac{N+1}{4}\right)$th data point.
- Calculate the 75% point (**3rd quartile**) of the dataset. The location of the 3rd quartile is defined to be the $\left(\frac{3(N+1)}{4}\right)$th data point.
- Calculate the IQR as

$$\text{IQR = 3rd quartile - 1st quartile}$$

Example 1 Consider the set of 11 numbers (which have been arranged in order)

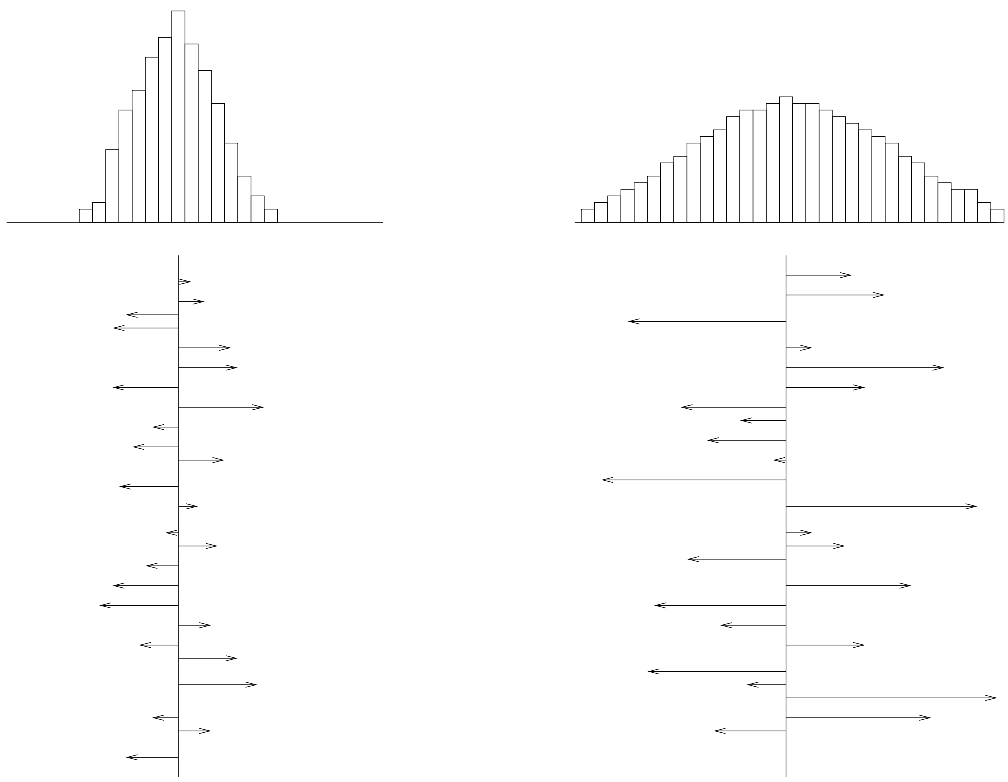
$$10, 15, 18, 33, 34, 36, 51, 73, 80, 86, 92$$

The 1st quartile is the $\frac{(11+1)}{4} = 3$rd data point = 18
The 3rd quartile is the $\frac{3(11+1)}{4} = 9$th data point = 80

$$\Rightarrow \text{IQR = 80 - 18 = 62}$$
$$\Rightarrow \text{SIQR = 62 / 2 = 31.}$$

The Mean Deviation

To measure the spread of a dataset it seems sensible to use the 'deviation' of each data point from the mean of the distribution. The deviation of each data point from the mean is simply the data point minus the mean.



small spread = small deviations   large spread = large deviations

Data x	Deviations $x - \bar{x}$
10	10 - 48 = -38
15	15 - 48 = -33
18	18 - 48 = - 30
33	33 - 48 = -15
34	34 - 48 = -14
36	36 - 48 = -12
51	51 - 48 = 3
73	73 - 48 = 25
80	80 - 48 = 32
86	86 - 48 = 38
92	92 - 48 = 44
Sum = 528 $\sum x = 528$	Sum = 0 $\sum(x - \bar{x}) = 0$

The mean is $\bar{x} = \frac{528}{11} = 48$

The Mean Deviation of a set of numbers is simply mean of deviations.

In practice, the mean deviation is *always* zero.

Mean Absolute Deviation (MAD)

We solve the problem of the deviations summing to zero by considering the **absolute values** of the deviations.

The absolute value of a number is the value of that number with any minus sign removed, e.g. $|-3| = 3$.

We then measure spread using the mean of the absolute deviations, denoted (MAD).

This can be written in mathematical notation as

$$\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad \text{or} \qquad \frac{\sum |x - \bar{x}|}{n}$$

Data x	Deviations $x - \bar{x}$	|Deviations| $\lvert x - \bar{x} \rvert$
10	10 - 48 = -38	38
15	15 - 48 = -33	33
18	18 - 48 = - 30	30
33	33 - 48 = -15	15
34	34 - 48 = -14	14
36	36 - 48 = -12	12
51	51 - 48 = 3	3
73	73 - 48 = 25	25
80	80 - 48 = 32	32
86	86 - 48 = 38	38
92	92 - 48 = 44	44
Sum = 528 $\sum x = 528$	Sum = 0 $\sum (x - \bar{x}) = 0$	Sum = 284 $\sum \lvert x - \bar{x} \rvert = 284$

$$\text{MAD} = \frac{284}{11} = 25.818 \quad \text{(to 3dp)}$$

The Sample Variance (s^2) and Population Variance (σ^2)

Another way to ensure the deviations don't sum to zero is to look at the *squared* deviations. Thus another way of measuring the spread is to consider the mean of the squared deviations, called the **Variance**

If our dataset consists of the whole population (a rare occurrence) then we can calculate the population variance σ^2 (said 'sigma squared') as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad \text{or} \qquad \sigma^2 = \frac{\sum(x - \bar{x})^2}{n}$$

When we just have a sample from the population (most of the time) we can calculate the sample variance s^2 as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad \text{or} \qquad s^2 = \frac{\sum(x - \bar{x})^2}{n-1}$$

NB We divide by $n-1$ when calculating the sample variance as then s^2 is a 'better estimate' of the population variance σ^2 than if we had divided by n.

Data x	Deviations $x - \bar{x}$	Deviations^2 $(x - \bar{x})^2$
10	10 - 48 = -38	1444
15	15 - 48 = -33	1089
18	18 - 48 = - 30	900
33	33 - 48 = -15	225
34	34 - 48 = -14	196
36	36 - 48 = -12	144
51	51 - 48 = 3	9
73	73 - 48 = 25	625
80	80 - 48 = 32	1024
86	86 - 48 = 38	1444
92	92 - 48 = 44	1936
Sum = 528 $\sum x = 528$	Sum = 0 $\sum(x - \bar{x}) = 0$	Sum = 9036 $\sum(x - \bar{x})^2 = 9036$

$$s^2 = \frac{9036}{11 - 1} = 903.6$$

The Sample and Population Standard Deviation (s and σ)

Notice how the sample variance in our example is much higher than the SIQR and the MAD.

$$\text{SIQR} = 31 \qquad \text{MAD} = 25.818 \qquad s^2 = 903.6$$

This happens because we have squared the deviations transforming them to a quite different scale. We can recover the scale of the original data by simply taking the square root of the sample (population) variance.

Thus we define the sample standard deviation s as

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

and we define the population standard deviation σ as

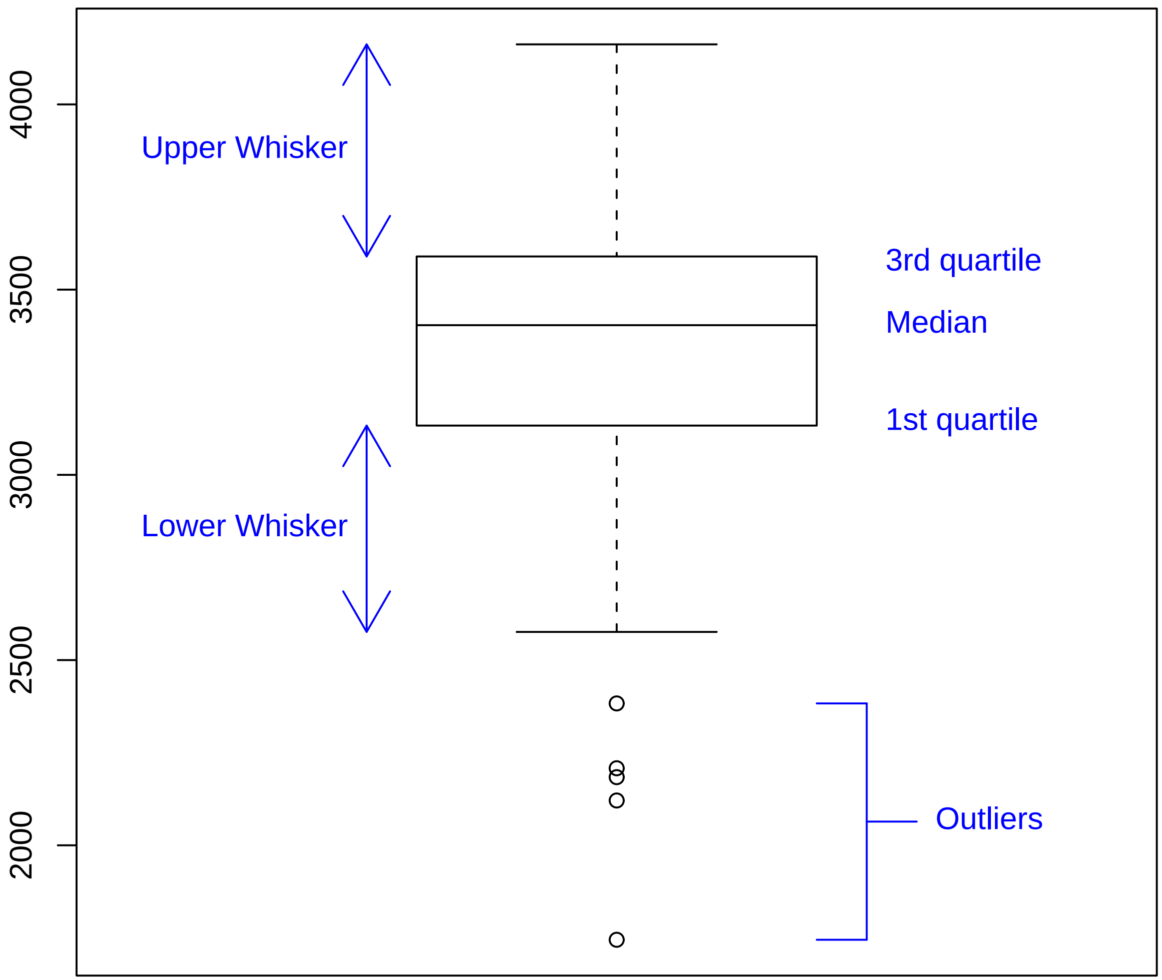
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Returning to Example 1 the sample standard deviation is

$$s = \sqrt{903.6} = 30.05 \quad \text{(to 2dp)}$$

which is comparable with the SIQR and the MAD.

Box Plots



4000
3500
3000
2500
2000
Upper Whisker
Lower Whisker
3rd quartile
Median
1st quartile
Outliers

A box plot consists of three main parts

1. A box that covers the middle 50% of the data. The edges of the box are the 1st and 3rd quartiles. A line is drawn in the box at the median value.
2. Whiskers that extend out from the box to indicate how far the data extend either side of the box. The whiskers should extend no further than 1.5 times the length of the box, i.e. the maximum length of a whisker is 1.5 times the IQR.
3. All points that lie outside the whiskers are plotted individually as outlying observations.

Plotting box plots of measurements in different groups side by side can be illustrative. For example, box plots of birth weight for each gender side by side and indicates that the distributions have quite different shapes.