

WHAT IS STATISTICS?

The study and use of *statistics* or *statistical methods* involve three related disciplines:

Data Analysis - the *collection, display* and *summary* of data

Probability - the laws of chance

Statistical Inference - using a knowledge of probability to draw *conclusions* or make *predictions* about the entire group or *population* from specific data for a smaller group or *sample*

Sometimes the word "statistics" is used in a narrower sense to refer to the actual data or information eg. accident statistics.

☺ Group discussion p 197, Ex 5.1 p 199

COLLECTING AND ORGANISING DATA

A *variable* is any quantity or characteristic whose value varies for different members of a population or sample eg. height, occupation.

Ordinal or *quantitative* variables (eg. height) take numerical values whereas *nominal* or *qualitative* variables (eg. occupation) take values which are not numbers.

A *discrete* variable is a quantitative variable which takes certain distinct or discrete values eg. a person's shoe size. A *continuous* variable is a quantitative variable which can take all values within a given interval eg. length of a person's foot.

A *frequency* is the number of times a value occurs eg. 12 males in a class.

Relative frequency or *experimental probability* is $\frac{\text{frequency}}{\text{total frequency}}$ eg. $\frac{12}{30}$ or 0.4 or 40% of a class are male.

A *frequency distribution* is a *table* showing the different values and the frequencies with which they occur. Frequency distributions may be *grouped* or *ungrouped*. A *tally* is often used to determine the frequencies.

UNGROUPED		GROUPED	
mark in test	frequency	mark in test	frequency
0	3	5 - 9	12
1	5	10 - 14	8
2	1	15 - 19	15
4	6	20 - 24	19
7	4	25 - 29	14
8	9	30 - 34	7
	<hr/>		<hr/>
	28		75

For grouped data:

- the *class intervals* or *classes* are the intervals into which the data is grouped
- the *class boundaries* or *class limits* are the lower and upper boundaries of the classes
- the *class width* or *class size* is the difference between the lower and upper boundary of a class
- the *class midpoint* or *class score* or *class mark* is the value halfway between the lower and upper boundary.

class interval	class boundaries	class width	class midpoint	frequency
5 - 9	4.5 - 9.5	5	7	12
10 - 14	9.5 - 14.5	5	12	8
15 - 19	14.5 - 19.5	5	17	15
20 - 24	19.5 - 24.5	5	22	19
25 - 29	24.5 - 29.5	5	27	14
30 - 34	29.5 - 34.5	5	32	7
				75

☺ Ex 5.4 p 210, Ex 5.5 p 212

STEM PLOT or STEM AND LEAF PLOT

A *stem plot* is a way of grouping the data. Each value is split into a *stem* and a *leaf* eg. a mark of 23 is represented by a stem of 2 and a leaf of 3. The stem plot for the example above is:

mark in a test

0	5 5 6 7 7 7 8 8 8 8 9 9
1	0 0 1 1 2 3 3 3
1	5 5 6 6 6 8 8 8 8 8 9 9 9 9 9
2	0 0 0 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4
2	6 6 6 7 7 7 7 8 8 8 8 8 9 9
3	0 0 1 1 2 2 4

The advantages of the stem plot are:

- it allows the individual values to be identified
- if the values are in order of increasing size, fractiles can be identified (see later)
- the shape of the distribution can be seen.

☺ Ex 5.6 p 218

GRAPHICAL REPRESENTATION OF DATA

Graphs are a powerful method of analysing data.

Bar Graph

A bar graph is appropriate in the case of an ungrouped frequency distribution of discrete data. This is because the variable takes certain values but not the values in between.

Dot Plot

In a dot plot, a dot is plotted for each value with repeated values resulting in the corresponding dots forming a vertical line. In the case of a grouped distribution, the dots should be plotted against the class midpoints (the graph on p 221 needs correction).

Histogram

For grouped data, rectangles are drawn against a continuous horizontal scale such that:

- the *area of each rectangle represents the class frequency*
- the positions of the bases of the rectangles correspond to the class boundaries ie. *no gaps between rectangles*
- the *horizontal scale* usually consists of the class midpoints (ie. *mark and label the midpoints*).

NB. If the class widths are equal, then the height of each rectangle is proportional to the frequency.

Frequency Polygon

For grouped data, a *straight line graph* is drawn with *class frequency plotted against class midpoint*.

- A *zero frequency should be plotted for the empty class at each end* of the distribution. This ensures that the total area under the frequency polygon is the same as the total area of the histogram ie. *area still represents frequency*.
- The horizontal scale usually consists of the class midpoints (ie. *mark and label the midpoints*).
- Avoid drawing a frequency polygon on top of a histogram ie. *draw one or the other but not both*.

☺ Why is a frequency polygon not appropriate when the class widths are not equal?

☺ Ex 5.7 p 223

CUMULATIVE FREQUENCY DISTRIBUTIONS

A *cumulative frequency* is the number of items that lie above (or below) a certain value. The *cumulative percentage* is the cumulative relative frequency shown as a percentage. A cumulative frequency distribution ("below" version) for the example above is:

mark in test	frequency	mark in test	cumulative frequency	cumulative percentage
		less than 4.5	0	0%
5 - 9	12	less than 9.5	12	16.0%
10 - 14	8	less than 14.5	20	26.7%
15 - 19	15	less than 19.5	35	46.7%
20 - 24	19	less than 24.5	54	72.0%
25 - 29	14	less than 29.5	68	90.7%
30 - 34	7	less than 34.5	75	100%
	<hr/> 75			

Cumulative frequency polygon

This is a straight line graph of the cumulative frequency or cumulative percentage plotted against the class boundary. If a smooth curve is drawn through the points, the graph is called an *ogive*.

Fractiles can be used to describe the data and can be read from the cumulative frequency polygon, ogive or stem plot.

Quartiles divide the data into four equal parts.

- 25% of the values lie below the *first or lower quartile* Q_1
- 50% of the values lie below the *second or middle quartile* Q_2 (also called the *median*)
- 75% of the values lie below the *third or upper quartile* Q_3

Percentiles divide the data into 100 equal parts.

- eg. 65% of the values lie below the 65th percentile P_{65}

Deciles divide the data into 10 equal parts

- eg. 80% of the values lie below D_8

☺ Ex 5.8 p 228

SUMMARISING DATA

Up till now, we have looked at the meaning of statistics, collecting and organising data, stem plots, graphical representation of data and cumulative frequency distributions (including reading fractiles from an ogive or cumulative frequency polygon). These ideas will now be extended to using *statistical measures* to summarise data ie. *measures of central tendency* and *measures of spread*.

MEASURES OF CENTRAL TENDENCY - A measure of central tendency is a single value used to represent a set of data. The most common measures of central tendency are the *mode*, *median* and *mean*.

MODE - *the value that occurs most often*

- The mode can be useful especially for non-numeric data eg. the most popular subject among year 11 students is Mathematics.
- In the case of a grouped frequency distribution, it is only possible to talk about the *modal class* ie. the class or group which occurs most often.
- Data may be *bimodal* ie. there are two values which occur with the highest frequency.

☺ Ex 13.2 p 480

MEDIAN - *the middle value when the data is arranged in order of size*

Median of a list of n values (or ungrouped frequency distribution) is the $\left(\frac{n+1}{2}\right)$ th value.

$$\begin{array}{ccccccc} 5 & 7 & 8 & 8 & \boxed{12} & 15 & 19 & 21 & 23 \\ \text{median} = \left(\frac{9+1}{2}\right) \text{th value} = 5 \text{th value} = 12 \end{array}$$

$$12 \quad 14 \quad 15 \quad 16 \quad \boxed{17} \quad \boxed{21} \quad 23 \quad 27 \quad 29 \quad 31$$

$$\text{median} = \left(\frac{10+1}{2} \right) \text{th value} = 5\frac{1}{2} \text{th value} = \frac{17+21}{2} = 19$$

Median of a grouped frequency distribution

mark in test	frequency	mark in test	cumulative frequency
		less than 4.5	0
5 - 9	12	less than 9.5	12
10 - 14	8	less than 14.5	20
15 - 19	15	less than 19.5	35
20 - 24	19	less than 24.5	54
25 - 29	14	less than 29.5	68
30 - 34	7	less than 34.5	75
	75		

Previously, the median for a grouped frequency distribution was read from an ogive or a cumulative frequency polygon. This method of *interpolation* can be performed by *calculation without drawing the graph*. Consider that part of the cumulative frequency polygon which includes the point with a cumulative frequency of 37.5 (ie. 75×0.5).

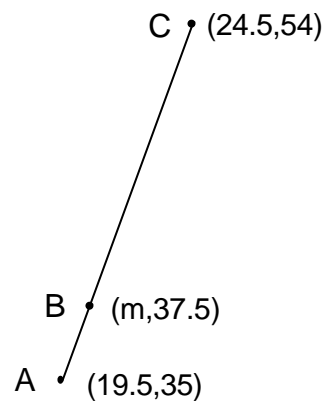
$$\text{gradient of AB} = \text{gradient of AC}$$

$$\frac{37.5 - 35}{m - 19.5} = \frac{54 - 35}{24.5 - 19.5}$$

$$\frac{2.5}{m - 19.5} = \frac{19}{5}$$

$$m = 19.5 + \frac{2.5}{19} \times 5$$

$$m = 20.2$$



The median is 20.2 marks.

An equivalent way of looking at the median for a grouped frequency distribution is to say that it is the value which divides the area of the histogram into two equal parts.

☺ Ex 13.3 p 483

MEAN

arithmetic mean (or \bar{x}) = $\frac{\text{sum of values}}{\text{number of values}}$

list of values	frequency distribution
$\bar{x} = \frac{\sum x}{n}$	$\bar{x} = \frac{\sum fx}{\sum f}$
	(for a grouped distribution, the class midpoint is used for x)

Ex 13.4 p 487

- ☺ If the same number is subtracted from each value in a set of data, what is the effect on the mean?
- ☺ If each value in a set of data is divided by the same number, what is the effect on the mean?
- ☺ If the 189.5 is subtracted from each value in a set of data and the result divided by 20, what is the effect on the mean? Use your answer to calculate the mean for Ex 13.5 p 490 qu 5 without using a calculator and by making the calculations as simple as possible.

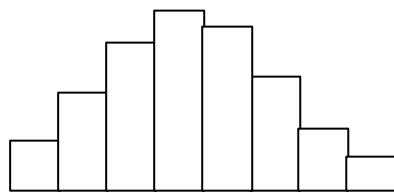
COMPARISON OF MODE, MEDIAN & MEAN

- The mode is useful for non-numeric data. It provides little information about the rest of the values in the data.
- The mean can be seriously affected by the presence of outliers (very small or large values) but the median is not.

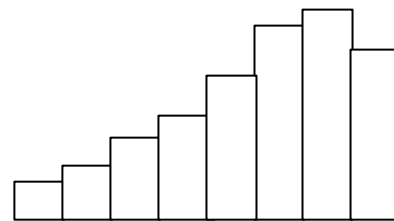
5 7 8 8 12 15 19 21 23
median = 12, mean = 13.1

5 7 8 8 12 15 19 21 23 200
median = 13.5, mean = 31.8

- Half of the data lies below the median and half of the data lies above it. This will be approximately true for the mean when the data is *symmetric*. If the data is *skewed*, then the median may differ significantly from the mean and usually the median would be used.



symmetric data



skewed data

- The mean can be used to calculate the sum of the values. Eg. if the mean sales price achieved during May for 18 sales in a car yard was \$12500, then the total value of sales was \$225000 (ie. $18 \times \$12500$). The use of the mean in this case would enable comparisons to be made between the total sales of different car yards.

☺ Ex 13.6 p 492 qu 4,6-12

MEASURES OF SPREAD

Measures of central tendency give no indication as to how the values vary ie. the spread of the data. It is possible to have two sets of data with the same measures of central tendency but the two sets of data are very different. Measures of spread are used to measure the spread or variation and include *range*, *interquartile range*, *five-number summary*, *variance* and *standard deviation*. *Box plots* or *box and whisker diagrams* represent the five-number summary in pictorial form.

Consider the marks of 15 students in two tests A and B.

A 7 8 11 14 15 16 17 20 21 22 25 27 31 32 34
B 11 12 14 16 17 18 19 20 21 21 22 24 26 29 30

The median and mean mark for both tests are 20 but data A is more spread out than data B.

Range = largest value - smallest value

range for A	range for B
$34 - 7 = 27$ marks	$30 - 11 = 19$ marks

The range is a poor measure of spread as it only involves two of the values and is therefore sensitive to outliers.

Interquartile range = $Q_3 - Q_1$ (ie. upper quartile - lower quartile)

- To find Q_1 and Q_3 for a list of values (or ungrouped frequency distribution), first use the median to divide the data into two equal parts. Q_1 is the median of the smaller values and Q_3 is the median of the larger values.

interquartile range for A	interquartile range for B
$27 - 14 = 13$ marks	$24 - 16 = 8$ marks

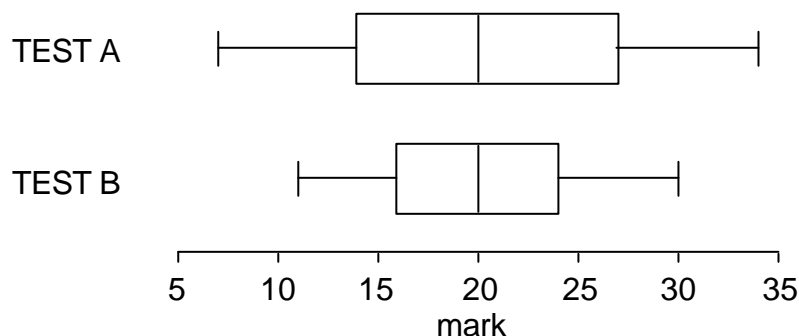
- To find Q_1 and Q_3 for a grouped frequency distribution, use interpolation either by reading the values from an ogive or cumulative frequency polygon or by calculation. (Multiply the total frequency by 0.25 and 0.75.)

Five-number Summary and Box Plot

- The spread of data is sometimes indicated by the five-number summary:
smallest value, lower quartile, median, upper quartile, largest value
- The range and interquartile range can easily be found from the five-number summary.

five-number summary for A	five-number summary for B
7, 14, 20, 27, 34	11, 16, 20, 24, 30

- The box plot or box and whisker diagram gives a picture of the five-number summary. The most important use of box plots is to draw them side by side so that sets of data can be compared. The interquartile range is represented by the length of the box and the median by the line drawn inside the box. One method of drawing the whiskers is to extend them as far as the smallest and largest values so that the range is represented by the total length of the diagram.



- The above method of drawing the whiskers for a box plot is open to the serious criticism that distortion would be caused by the existence of outliers. There is a method which is widely used to overcome this problem; in choosing the value at the end of a whisker, the length of the whisker cannot exceed $1.5 \times \text{IQR}$ and each outlier beyond the whisker is shown as a separate point.

Variance and standard deviation

The *deviation* of a value x is defined as $x - \bar{x}$ (ie. value - mean) and indicates how far the value is from the mean and whether it is smaller or larger than the mean. The *mean of the squares of the deviations* is a measure of spread (the larger the answer, the greater the spread) and is called the variance. The standard deviation is the square root of the variance; taking the square root gives a measure with the same units as the original values.

variance for A	variance for B	standard deviation for A	standard deviation for B
s^2 $= \frac{\sum (x - \bar{x})^2}{n}$ $= \frac{1020}{15}$ $= 68$	s^2 $= \frac{\sum (x - \bar{x})^2}{n}$ $= \frac{450}{15}$ $= 30$	s $= \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$ $= \sqrt{\frac{1020}{15}}$ $= 8.2$ marks	s $= \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$ $= \sqrt{\frac{450}{15}}$ $= 5.5$ marks

DATA A			DATA B			DATA A		DATA B	
x	$x - \bar{x}$	$(x - \bar{x})^2$	x	$x - \bar{x}$	$(x - \bar{x})^2$	x	x^2	x	x^2
7	-13	169	11	-9	81	7	49	11	121
8	-12	144	12	-8	64	8	64	12	144
11	-9	81	14	-6	36	11	121	14	196
14	-6	36	16	-4	16	14	196	16	256
15	-5	25	17	-3	9	15	225	17	289
16	-4	16	18	-2	4	16	256	18	324
17	-3	9	19	-1	1	17	289	19	361
20	0	0	20	0	0	20	400	20	400
21	1	1	21	1	1	21	441	21	441
22	2	4	21	1	1	22	484	21	441
25	5	25	22	2	4	25	625	22	484
27	7	49	24	4	16	27	729	24	576
31	11	121	26	6	36	31	961	26	676
32	12	144	29	9	81	32	1024	29	841
34	14	196	30	10	100	34	1156	30	900
		1020			450		7020		6450

Manual calculations are usually performed with alternative formulas:

variance for A	variance for B	standard deviation for A	standard deviation for B
s^2 $= \frac{\sum x^2 - n\bar{x}^2}{n}$ $= \frac{7020 - 15 \times 20^2}{15}$ $= 68$	s^2 $= \frac{\sum x^2 - n\bar{x}^2}{n}$ $= \frac{6450 - 15 \times 20^2}{15}$ $= 30$	s $= \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n}}$ $= \sqrt{\frac{7020 - 15 \times 20^2}{15}}$ $= 8.2$ marks	s $= \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n}}$ $= \sqrt{\frac{6450 - 15 \times 20^2}{15}}$ $= 5.5$ marks

The formulas for a frequency distribution are:

$$s^2 = \frac{\sum f(x - \bar{x})^2}{\sum f} = \frac{\sum f x^2 - (\sum f)\bar{x}^2}{\sum f} \qquad s = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}} = \sqrt{\frac{\sum f x^2 - (\sum f)\bar{x}^2}{\sum f}}$$

In practice, calculations are best performed using the statistics functions on your calculator. Make sure that you can find the mean, variance and standard deviation for a set of data both with and without using the statistics functions on your calculator.

Sample mean, sample variance and sample standard deviation

Suppose calculations are performed for a random sample from a larger population. To obtain the *best estimates of the population variance and population standard deviation*, the formulas need to be amended so that the sum of the squares of the deviations is divided by a number 1 less than the number of values in the sample (rather than dividing by the number of values in the sample).

For a list of values:

$$s_{n-1}^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{\sum x^2 - n\bar{x}^2}{n - 1} \qquad s_{n-1} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n - 1}}$$

For a frequency distribution:

$$s_{n-1}^2 = \frac{\sum f(x - \bar{x})^2}{\sum f - 1} = \frac{\sum f x^2 - (\sum f)\bar{x}^2}{\sum f - 1} \qquad s_{n-1} = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f - 1}} = \sqrt{\frac{\sum f x^2 - (\sum f)\bar{x}^2}{\sum f - 1}}$$

These alternative measures are also calculated by the statistics functions on your calculator.

- ☺ Ex 13.9 p 504 qu 3-5, 10-14
- ☺ If the same number is subtracted from each value in a set of data, what is the effect on the standard deviation?
- ☺ If each value in a set of data is divided by the same number, what is the effect on the standard deviation?
- ☺ Ex 13.10 p 506