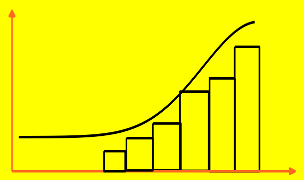
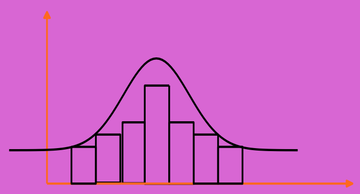
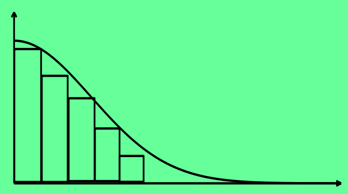
HISTOGRAMS

SHAPE

negatively skewed	symmetric	positively skewed
		
mean < median	median = mean	mean > median

OUTLIERS

- Data that lies away from the main body of values.
- Outliers do not determine the shape.

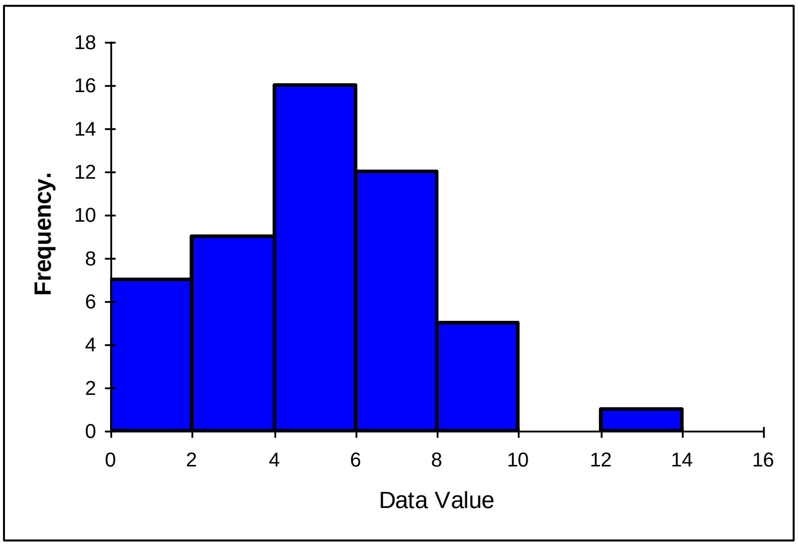
CENTRE

- Modal column is the highest column, observation that has the highest frequency
- Median approximated by finding the position of half (n)
- Mean cannot be obtained from the histogram graph.

SPREAD

- range is a measure of spread and is found by maximum value – minimum value, excluding outliers
- interquartile range is the preferred measure of spread, but when only given a histogram it is not used.
- standard deviation cannot be calculated from the histogram

HISTOGRAMS

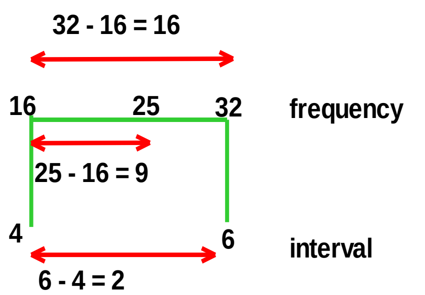


MEDIAN CALCULATIONS

MEDIAN INTERVAL

- Sum of the frequencies of all columns
7 + 9 + 16 + 12 + 5 + 1 = 50,
- approximate median location is found at 50 ÷ 2 = 25 value in the data set
- median interval column 1+column 2
7 + 9 = 16,
- column 1 + column 2 + column 3 7 + 9 + 16 = 32
- median location is the 3rd column, interval between 4 and 6

MEDIAN CALCULATION



Estimate median needs to around the middle but closer to 6 than 4 eg 5.1.

Calculation

$$\left(\frac{9}{16}\right) \times 2 + 4 = 5.125$$

INTERPRETATION

The data set is approximately symmetrical with an outlier between 12 and 14. The modal column is between 4 and 6. The median is approximately at 5.125. The spread is moderate with a range of 10 , excluding the outlier.

STEMPLOT

STEM	LEAF
30	1 6
40	2 8
50	0 | 2
60	2 4
70	5 5 | 8
80	2 7 7 8 | 9
90	0 1 7 9

KEY 90|9 = 99

STRUCTURE

- Data sets of less than 50 to 10 stems. All data can be found in the table. Calculations of statistics are simple.
- Split stems can be used to increase the number of required stems

SHAPE

- The shape is easily determined from a stemplot.
- Symmetric, Positively skewed, Negatively Skewed with or without outlier

OUTLIERS

- Outliers can be observed and must be stated
- Outliers have no effect on the shape of the distribution

QUARTILES

- The spread can be estimated by interquartile range. (IQR)
- The can be found after the median is determined. Split each half of the data set in half again.
- Q_1 is the first quartile and the shows the first 25% of values in the data set
- Q_3 is the third quartile and that shows where 75% of the data set sit below.
- Both values are found by counting in the stemplot
- **EG**

Q_1 = 51 found between the 5th and 6th values

Q_3 = 88.5 found 15th and 16th values

CENTRE

- Median can be found by locating the

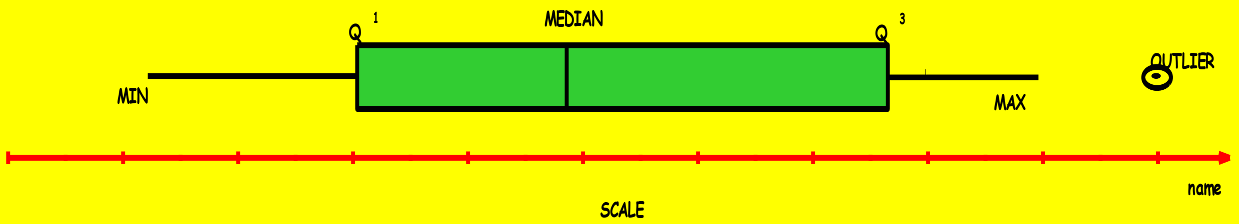
$$x_m = \frac{n+1}{2}^{th} \quad term$$

- This is found by counting on the stemplot
- Mean cannot be found from a stemplot , it must be calculated.

- **EG n = 20 the median is the 10.5 value in the above stemplot median = 77.5**

BOXPLOTS

A boxplot shows a summary of a data set. The median, 1st quartile, 3rd quartile, minimum maximum and outliers.



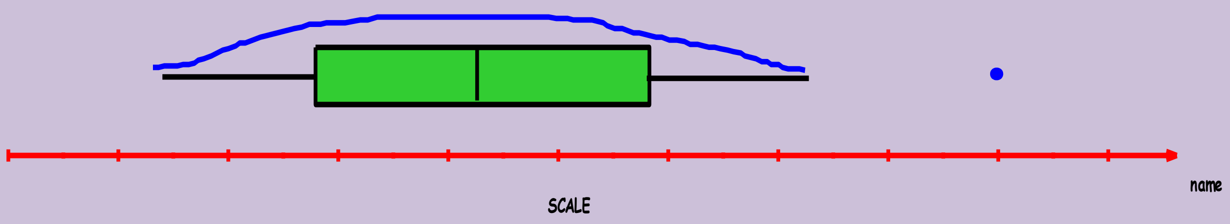
OUTLIER CALCULATIONS

LOWER FENCE Q1 - 1.5 X IQR

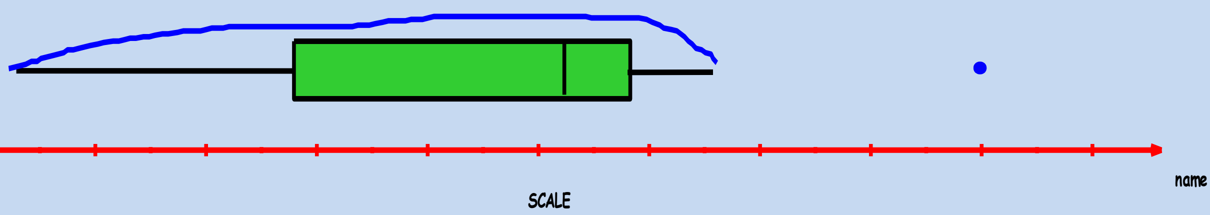
UPPER FENCE Q3 + 1.5 X IQR

SHAPE PROPERTIES OF BOXPLOT

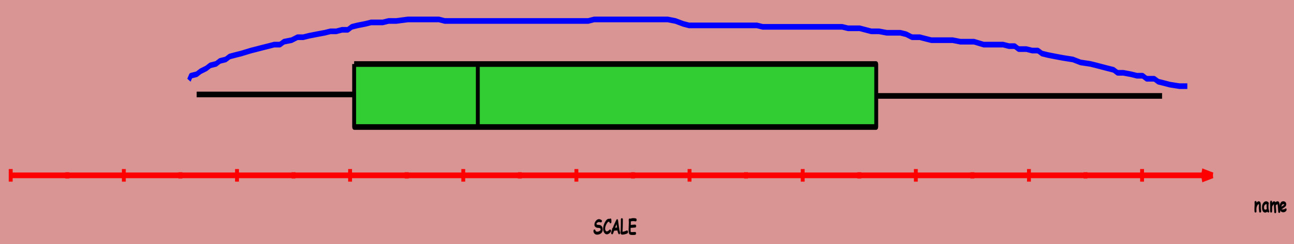
APPROIMATELY SYMMETRICAL WITH OUTLIER AT........



NEGATIVELY SKEWED WITH OUTLIER AT........



POSITIVELY SKEWED NO OUTLIERS



INTERPRETATION OF DATA

SHAPE AND OUTLIERS

SHAPE

- SYMMETRIC, (APPROXIMATELY)
- POSITIVELY/NEGATIVELY SKEWED (SLIGHTLY, CLEARLY)

OUTLIERS

- WITH OUTLIERS (GIVE VALUE OR INTERVAL)
- NO OUTLIERS

CENTRE AND SPREAD

CENTRE

- MEDIAN FOR SYMMETRICAL/SKEWED DATA GIVE VALUE
- MEAN FOR SYMMETRICAL DATA GIVE VALUE

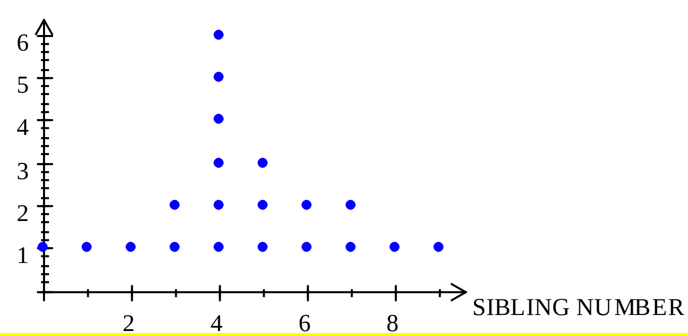
SPREAD

- IQR/RANGE FOR HISTOGRAM (MEDIAN)
- IQR FOR STEMPLOT AND BOXPLOT (MEDIAN)
- STANDARD DEVIATION FOR SYMMETRICAL DATA (MEAN)

INTERPRETATION OF DATA

DOT PLOT

NO OF SIBLINGS IN THE 1950'S



INTERPRETATION

The data set for the number siblings in the 1950's is slightly positively skewed with no outliers. The centre is given by the median at 4.5 siblings. The interquartile range is 9 which shows a large spread.

STEMPLOT

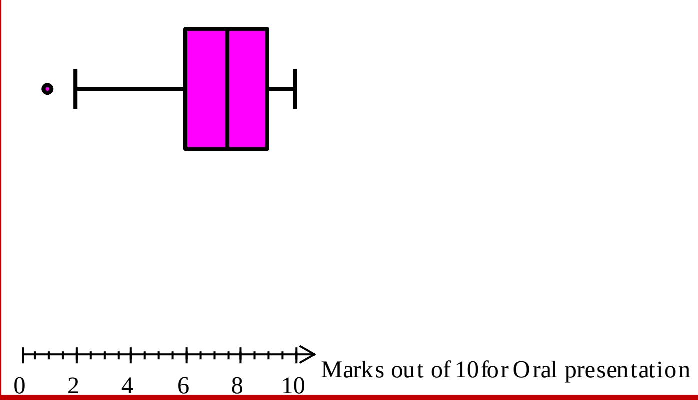
STEM	LEAF
30	1 6
40	2 8
50	0 | 2
60	2 4
70	5 5 | 8
80	2 7 7 8 | 9
90	0 1 7 9

90|9 = KEY 99

INTERPRETATION

The data set is negatively skewed with no observed outliers. The median is 76.5 with moderate spread of 37.5 as given by the IQR.

BOXPLOT



INTERPRETATION

The distribution for the marks out of ten for an oral presentation are negatively skewed with one outlier of 1 mark. The centre is at 7 as given by the median. The interquartile range of 3 shows a moderate spread.

MEAN

The mean is a measure of the centre, where all values of the data set are added and then divided by the size (n) of the data set.
The mean will evenly distribute the total data set to each member of the data set.
The mean is affected by extreme values in the data set, that is very low or very high values. The mean is not used when outliers are present or the data set is skewed.

$$mean = \frac{sum \ \ of \ \ data \ \ values}{total \ \ number \ \ of \ \ data \ \ values}$$

MATHEMATICALLY THE MEAN IS DEFINED AS

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

$\bar{x}$ mean

$\sum$ the sum of all values

x_i the first data value to last data value

CAS INSTRUCTIONS

enter data in list and spreadsheets, (frequency) tables may be used if required→cntrl i →calculator page→menu→6 stats→1 stat calc →1 one variable stats → num of lists→x1 list use your data set name→ all statistical data is given.

Ex 1 Find the mean and median of the following set of data, correct to one decimal place.

10 12 11 15 18 12 27 14 15 9 16 17 11

answer mean is 14.4 median is 14 (+ve skewed as mean>median)

STANDARD DEVIATION

STANDARD DEVIATION

The standard deviation is a measure of **spread** that uses every data value in the set. It is used with the mean as a summary statistic.
The standard deviation is influenced by outliers and should not be used to calculate the spread of skewed data or data containing outliers.
Standard deviation is never negative
The greater the spread the higher the standard deviation.
To interpret the standard deviation in needs to be compared with the size of the mean.

symbol is s_x

THE MATHEMATICAL FORMULA IS

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

$(x-\bar{x})$ is the difference between an observation and the mean

CAS – Standard deviation will be shown when you calculate the mean.

ESTIMATING THE STANDARD DEVIATION

$$S_x \approx \frac{range}{6}$$

STANDARD DEVIATION

Ex 1 Use the formula to calculate the standard deviation of the following number of traffic lights found in neighbouring suburbs.

1 3 5 7 9

Calculate the mean

$$\begin{aligned}\bar{x} &= \frac{25}{5}\\ &= 5\end{aligned}$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
1	-4	16
3	-2	4
5	0	0
7	2	4
9	4	16
		$\Sigma(x-\bar{x})^2 = 40$

$$\begin{aligned}s_x &= \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}\\ &= \sqrt{\frac{40}{4}}\\ &= 3.16\end{aligned}$$

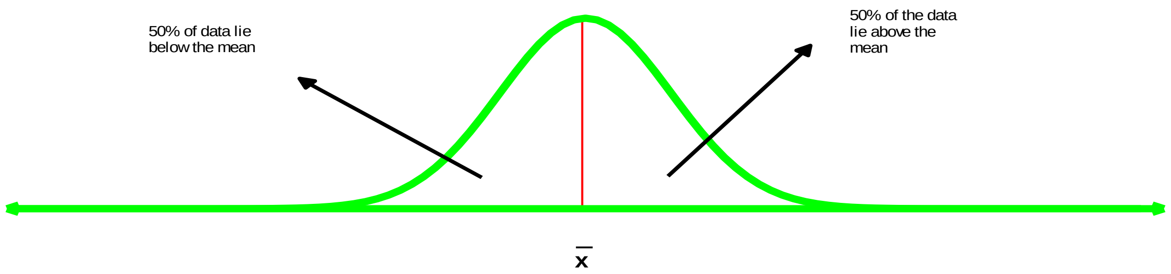
Ex 2 Use the CAS to calculate the mean and standard deviations of the number of aces served in 8 men's tennis matches at the Australian open.

35 41 32 48 21 25 27 29

answer: mean 32.25 std dev 8.86 aces

NORMAL DISTRIBUTION

Many data sets follow a normal distribution. In particular, data sets pertaining to biological statistics. A normal distribution shows a symmetrical bell shape curve. The mean and median are approximately equal.



The following results can be generalised with complex mathematical processes beyond the scope of the Further Maths course.

FACTS

1. 50% of observations lie either side of the mean.

2. 68% of observations lie within the interval of the values $\bar{x} \pm s_x$

3. 95% of observations lie within the interval of values $\bar{x} \pm 2s_x$

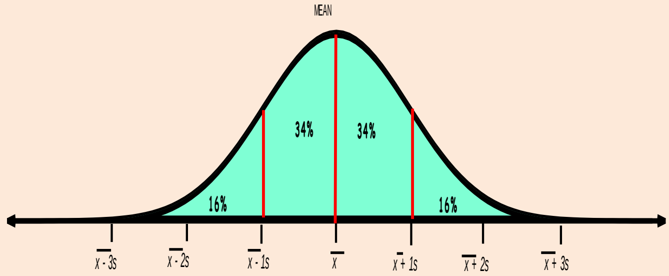
4. 99.7% of observations lie within the interval of values $\bar{x} \pm 3s_x$

This is known as the 68-95-99.7% Rule.

SUMMARY OF NORMAL CURVE

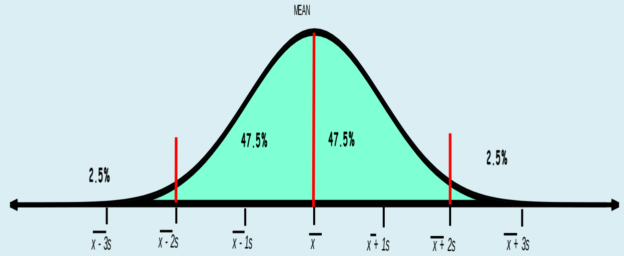
$$\bar{x} \pm 1s_x$$

Inside the interval is 68%, 34 % either side of the mean, outside the interval is 32%, 16% either side of the mean



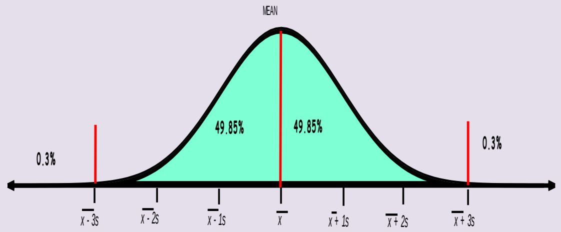
$$\bar{x} \pm 2s_x$$

Inside the interval 95%, 47.5% either side of the mean, outside 5%, 2.5% either side of the mean



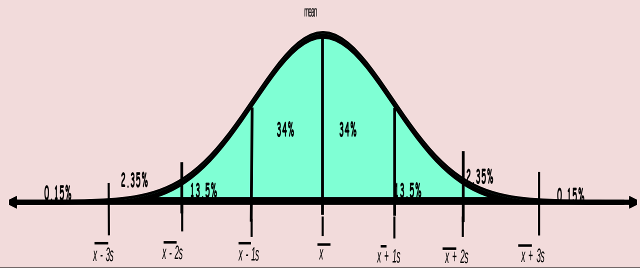
$$\bar{x} \pm 3s_x$$

Inside the interval 99.7%, 49.85% either side of the mean, outside 0.3% or either side of the mean



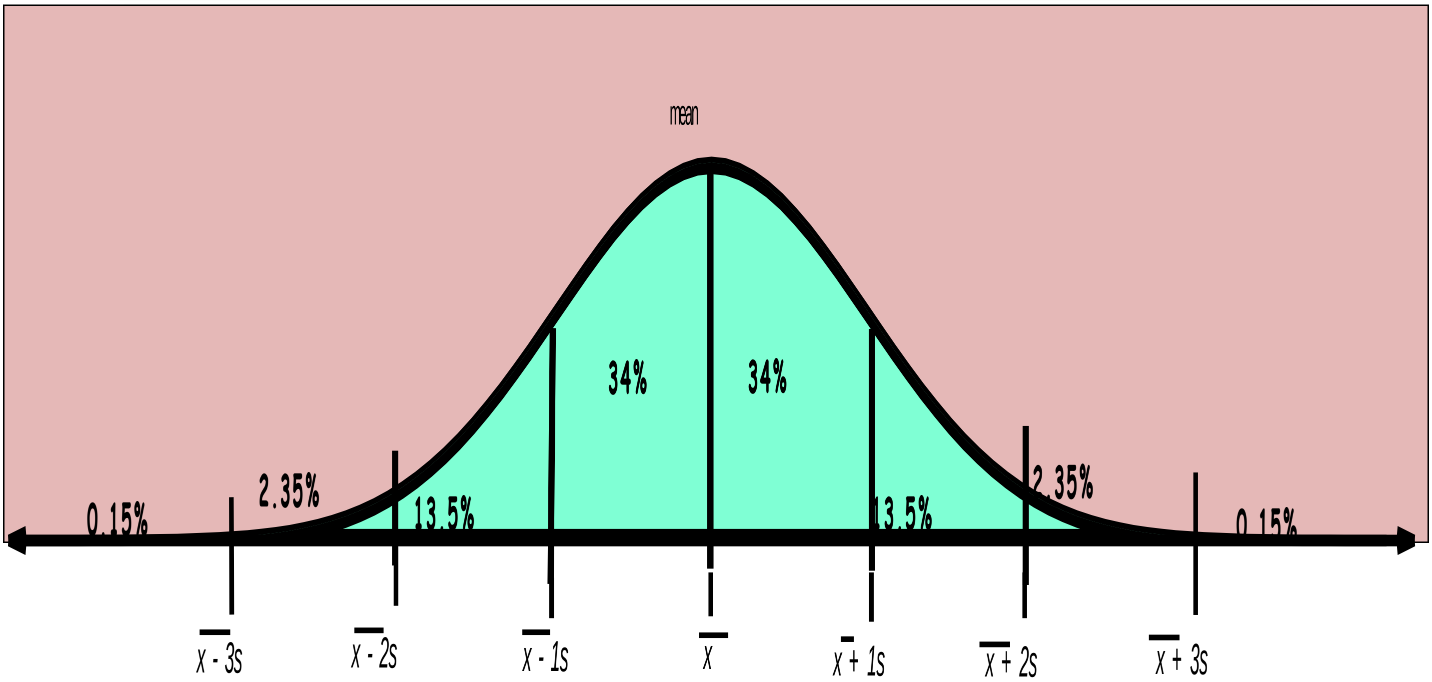
SUMMARY

ALL INTERVALS SUM TO 100%



1. $\bar{x} \pm 1s_x$ **Inside the interval is 68%, 34 % either side of the mean, outside the interval is 32%, 16% either side of the mean**
2. $\bar{x} \pm 2s_x$ **Inside the interval 95%, 47.5% either side of the mean, outside 5%, 2.5% either side of the mean**
3. $\bar{x} \pm 3s_x$ **Inside the interval 99.7%, 49.85% either side of the mean, outside 0.3% or either side of the mean**

NORMAL CURVE - INTERVALS



1. 50% of the population lie either side of the mean.

2. $\bar{x} \pm 1s_x$ Inside the interval is 68% , 34 % either side of the mean, outside the interval is 32%, 16% either side of the mean.

3. $\bar{x} \pm 2s_x$ Inside the interval 95%, 47.5% either side of the mean, outside 5%, 2.5% either side of the mean.

4. $\bar{x} \pm 3s_x$ Inside the interval 99.7%, 49.85% either side of the mean, outside 0.3% or either side of the mean

Z-SCORES

z-scores are used to classify all observations in a normal shaped distribution into percentage range of the population.

Z-SCORES CALCULATION

$$Z = \frac{x - \bar{x}}{s_x}$$

Z-SCORE VALUE	LOCATION IN POPULATION
0 to 1	TOP 50%
1 to 2	TOP 16%
2 to3	TOP 2.5%
3 →	TOP 0.15%
-1 to 0	BOTTOM 50%
-2 to -1	BOTTOM 16%
-3 to -2	BOTTOM 2.5%
→ -3	BOTTOM 0.15%