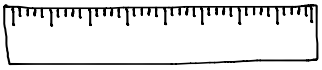
S1 - Chapters 2, 3 and 4 - Representation and summary of data - Summary

* In statistics we collect observations or measurements of data

Quantitative variables
numerical observations

- **Discrete**: can only take certain values in a given range, eg Number of accidents, shoe size
- **Continuous**: can take any value in a given range, e.g. Height, weight

Qualitative variables
Non-numerical observations
eg. Hair colour, blood type

* Hint: Imagine you have a ruler in front of you 

If what you are measuring can take values on this ruler then it is a quantitative variable, otherwise it is a qualitative variable
If it can take any value on the ruler then it is continuous, if it can only take certain values then it is discrete.

* When you collect a small number of observations then you can just list them.
If you have a large number of observations then you can present them in a frequency table or as grouped data

Individual observations
eg Age (to the nearest year) of 5 students

16, 17, 17, 15, 18

Frequency table
eg. Number of cars owned by 30 families

Number of cars	Number of families
0	1
1	13
2	12
3	4

Grouped data
eg Speed of 50 cars in the motorway

Speed (km/h)	Number of cars
80 - 100	7
100 - 120	31
120 - 140	10
140 - 150	2

MAKE SURE YOU UNDERSTAND THE DIFFERENCE BETWEEN THESE

* With grouped data, groups are known as classes. You need to know how to determine class boundaries, mid-points and class width.

* Measures of location
  1. Mode: This is the value that occurs most often
     In case we are working with grouped data, we can only talk about the modal class (the class with the highest frequency).
  2. Median: The middle value when the data is put in ascending order
     To find the median evaluate $n/2$ where n is the number of observations

Then, if all the observations are available:
- If $n/2$ is an integer, the median is the mid-point of that and the next observation
- If $n/2$ is not an integer then round UP and the median is given by the corresponding observation

In case we are working with grouped data you need to use INTERPOLATION (Remember not to do any rounding).

③ Mean: It is the sum of all the observations divided by the total number of observations. If $x_1, x_2, x_3, \ldots, x_n$ are the observations then

$$\bar{x} = \frac{\Sigma x_i}{n}$$

In case we are working with grouped data, we estimate each category by its midpoint

$$\bar{x} = \frac{\Sigma f x}{\Sigma f}$$

* Using the correct measure of location
  - Mode: It is used with qualitative data, or with quantitative data with a single mode or bimodal.
  - Median: It is used with quantitative data, in the presence of extreme values.
  - Mean: It is used with quantitative data and uses all the observations. It is susceptible to the presence of extreme values.

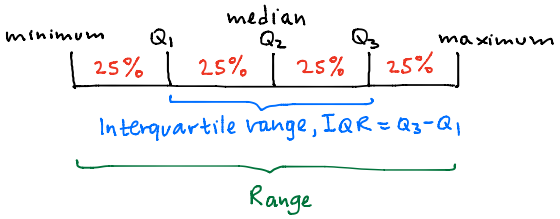
* Combining means: Suppose that dataset A consists of n_A observations and has mean $\bar{x}_A$, while dataset B consists of n_B observations and has mean $\bar{x}_B$

$$\text{Then, } \bar{x} = \frac{n_A \cdot \bar{x}_A + n_B \cdot \bar{x}_B}{n_A + n_B}$$

(Remember: Never average the averages)

* Measures of dispersion
  ① Range = Highest value - lowest value (You should never report this as an interval)

  ② Quartiles: They split the data into four parts



  - To find the lower quartile, Q_1, evaluate $n/4$ and proceed just like we do with the median.
  - To find the upper quartile, Q_3, evaluate $n/4$ and proceed just like we do with the median.

  ③ Percentiles: The idea is the same as quartiles, the only difference being that the data is split into 100 parts. Similarly to the IQR, in this case we refer to the interpercentile range.

④ Variance: It is a measure of total dispersion as it takes into account all of the observations

$$\sigma^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$$

In case we are working with grouped data, $\sigma^2 = \frac{\sum f(x-\bar{x})^2}{\sum f} = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$

* Coding

This is of the form $y = \frac{x-a}{b}$

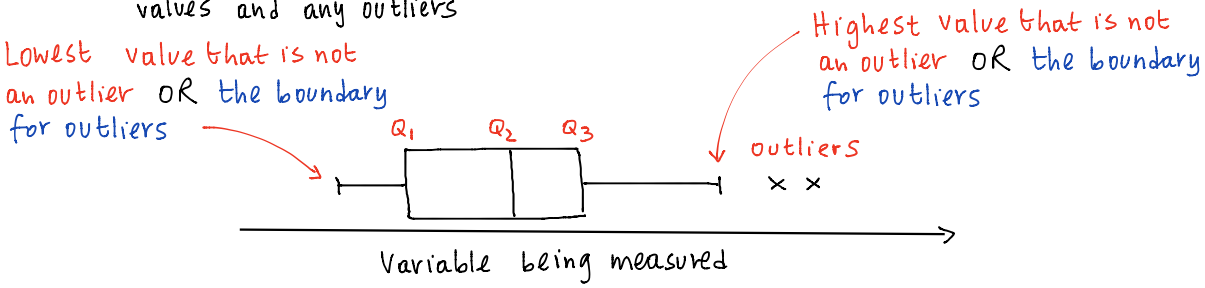
where x is the orrginal data
y is the coded data
and a and b are constants to be chosen

- All measures of location are affected by coding just like the original data is.
  Hence, $\bar{y} = \frac{\bar{x}-a}{b}$
- All measures of spread are affected only by multiplication or division
  Hence, $\sigma_y = \frac{\sigma_x}{b}$ Note that for variance $\sigma_y^2 = \frac{\sigma_x^2}{b^2}$

* Presenting data

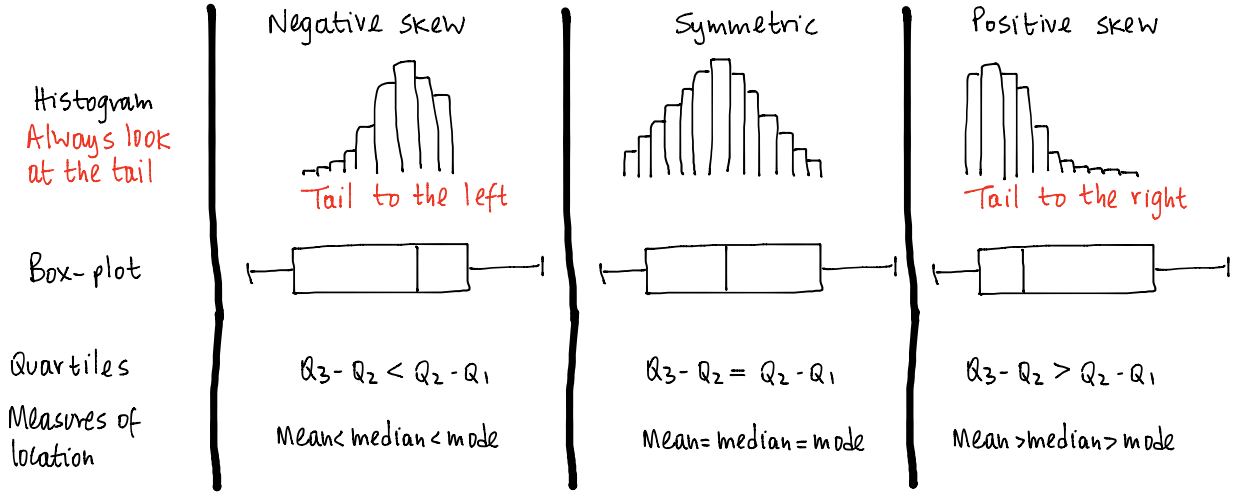
- stem and leaf diagram
  ① A stem and leaf diagram keeps the detail of the data.
  ② It enables the shape of the distribution to be revealed
  ③ Back-to-back stem and leaf diagrams can be used to compare two sets of data.

- Box-plot
  A box plot is a graph showing the quartiles, maximum and minimum values and any outliers



- Histogram
  ① Used to represent continuous data
  ② Different from bar charts in that there are no gaps between the bars and the area under the graph is proportional to the frequency.

* Shape of a distribution



All of the above can only classify a distribution as symmetric, positively skewed or negatively skewed. They do not allow comparisons between distributions like "this is more negatively skewed than the other".

One may calculate $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$

This not only tells us if there is skew or not, it also tells us how skewed the data is.

* If you are asked to compare two datasets you need to comment upon
  1. A measure of location
  2. A measure of spread
  3. Skewness