THE GC SCHOOL OF CAREERS

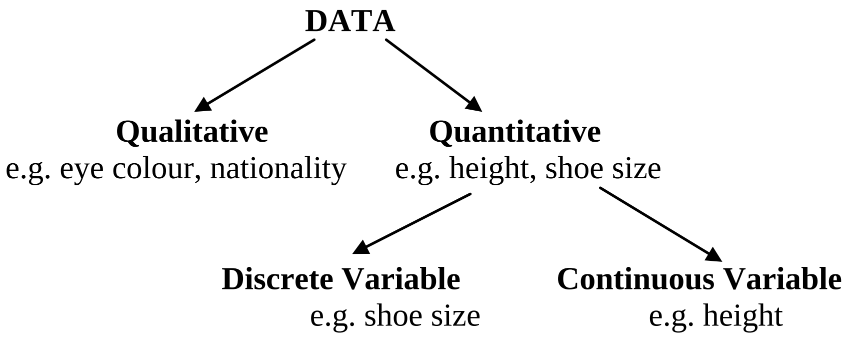
MATHEMATICS DEPARTMENT
FORM 7

S1

STATISTICS 1

CHAPTERS 2 – 4

METHODS OF SUMMARIZING SAMPLE DATA



DATA

Qualitative
e.g. eye colour, nationality

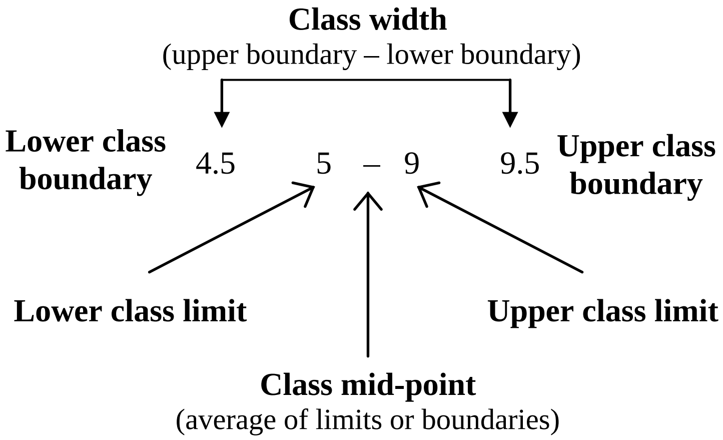
Quantitative
e.g. height, shoe size

Discrete Variable
e.g. shoe size

Continuous Variable
e.g. height

Grouped Frequency Distribution (with gaps between the classes)

CLASS



Class width
(upper boundary – lower boundary)

Lower class boundary 4.5

5 – 9

9.5 **Upper class boundary**

Lower class limit

Upper class limit

Class mid-point
(average of limits or boundaries)

Measures of location can be used as a single value to represent the whole data set.
Measures of dispersion represent the spread or variation within the data set.

[A] Measures of Location

Mode: This is the value that occurs most often (highest frequency)
Advantages: It is easy to calculate
It can be used for both qualitative and quantitative data
It is not affected by extreme values in the data set
Disadvantage: It has no useful mathematical properties
It is not very informative if each value in the data set appears only once

Median: It is the middle value when the data is put in order (ascending/ descending).
Advantages: It is relatively easy to calculate or estimate
It is not affected by extreme values in the data set
Disadvantage: It has no useful mathematical properties

Mean: This is the arithmetic average. It is the sum of all the observations divided by the total number of observations.
Advantages: All the values are used directly,
It has very important mathematical properties.
Disadvantage: It is influenced by extreme values in the data set.
It is not as easy to calculate as the mode or the median.

	Raw Data	**Frequency Distribution**	**Grouped Frequency Distribution**
MODE / MODAL CLASS	Most frequent value	Value with the highest frequency	Class with the highest frequency
MEDIAN Q_2	**1.** Order the data **2.** Find $\frac{n}{2}$ **3.** If an integer: Q_2 is the average of that term and the next one If not an integer: Q_2 is the value of the next term	**1.** Find the cumulative frequency. **2.** Find $\frac{n}{2}$ **3.** If an integer: Q_2 is the average of that term and the next one If not an integer: Q_2 is the value of the next term	**1.** Find the cumulative frequency **2.** Find $\frac{n}{2}$ **3.** Find the median class **4.** Use *interpolation* to find an estimate of the median value
MEAN $\bar{x}$	$\bar{x} = \frac{\sum x}{n}$	$\bar{x} = \frac{\sum fx}{\sum f}$	$\bar{x} = \frac{\sum fx}{\sum f}$ x is the mid-point of the class.

Combined Means

If set A, of size n_1 has mean $\bar{x}_1$ and set B of size n_2, has mean $\bar{x}_2$, then the mean of the combined set of A and B is:

$$\boxed{\bar{x} = \frac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2}{n_1 + n_2}}$$

Quartiles, Deciles, Percentiles

Quartiles divide the distribution into four equal parts (quarters).

- **Q_1 - lower quartile**: 25% of the data lies to the left of Q_1
- **Q_2 - median**: 50% of the data lies to the left of Q_2
- **Q_3 - upper quartile**: 75% of the data lies to the left of Q_3

Deciles divide the distribution into 10 equal parts.

Percentiles divide the distribution into 100 equal parts.

To find any quartile/ decile/ percentile we use the same method of calculation as the median.

e.g. for Q_1 we use $\frac{n}{4}$, for P_{15} we use $\frac{15n}{100}$

[B] Measures of Dispersion

- $Range = Highest\ Value - Lowest\ Value$
- $Interquartile\ \ Range\ (IQR) = Q_3 - Q_1$
- $Semi - Interquartile\ \ Range\ (SIQR) = \frac{Q_3 - Q_1}{2}$
- $\boxed{Variance = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$
- $\boxed{Standard\ Deviation\ = \sqrt{Variance}}$

Also, to work out the variance and standard deviation for a frequency table and grouped frequency distribution, where x is the midpoint, use

$$Variance = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

Coding

When the data values are large, you can use coding to make the numbers easier to work with.

- To find the mean of the original data, find the mean of the coded data, equate this to the coding used and solve.

- To find the standard deviation of the original data, find the standard deviation of the coded data and either multiply this by what you divided by or divide this by what you multiplied by.

REPRESENTATION OF SAMPLE DATA

Stem-and-Leaf Diagram

A stem and leaf diagram is used to order and present data. The advantage of a stem and leaf diagram is that is reveals the shape of the distribution and enables quartiles to be found. Also, it enables the comparison of two data sets using back-to-back stem and leaf diagrams.

Note:

- Remember to use a key (e.g. 2|3 means 23).
- Never leave a stem behind. Write it down and leave the leaf empty.

Histogram

We can represent *continuous data* summarized in a grouped frequency distribution by a histogram. In a histogram Frequency Density is plotted against Class Boundaries.

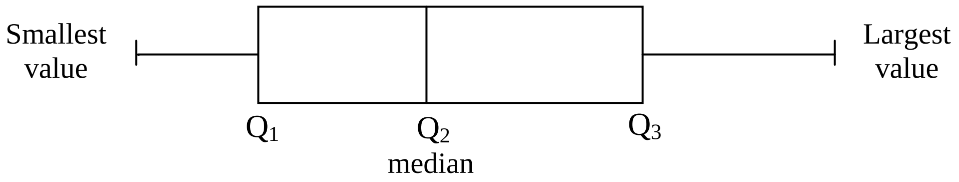
$$\boxed{Frequency\ \ Density = \frac{Frequency}{Class\ \ Width}}$$

$$Frequency \propto Area$$

Remember to use the boundaries of the interval in order to find the area of each bar.

Box-and-Whisker Plot

A box plot represents important features of the data. It shows quartiles (Q_1, Q_2 and Q_3), the maximum and minimum values and any outliers (extreme values) in the data set. Box plots can also be used to compare two sets of data by showing two box plots on the same scale.



Note: Remember to clearly label the axis on the plot.

Outliers

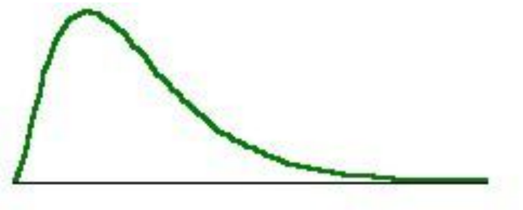
An outlier is an extreme value that lies outside the overall pattern of the data.
You will be told what rule to apply to identify outliers.
Usually, an outlier is any value that is:

- greater than $Q_3 + 1.5 \cdot IQR$ or
- smaller than $Q_1 - 1.5 \cdot IQR$.

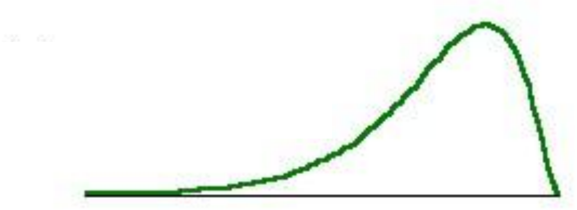
Skewness

The shape (skewness) of a data set can be described using diagrams, measures of location and measures of spread.

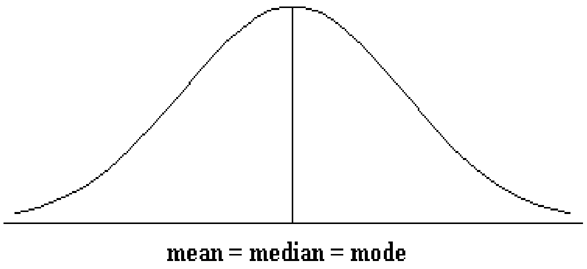
Positively Skewed Distribution
(skewed to the right)



Negatively Skewed Distribution
(skewed to the left)



Symmetrical Distribution
(Bell shaped)



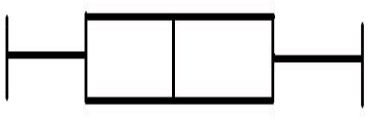
Ways to Describe Skewness

- Using the Quartiles:
  If $Q_2 - Q_1 = Q_3 - Q_2$ then the distribution is **symmetrical**.
  If $Q_2 - Q_1 < Q_3 - Q_2$ then the distribution is **positively skewed**.
  If $Q_2 - Q_1 > Q_3 - Q_2$ then the distribution is **negatively skewed**.

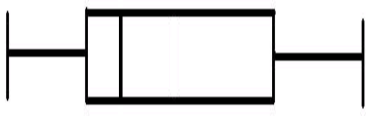
- Using the measures of location:
  Mode = Median = Mean describes a distribution which is **symmetrica**l.
  Mode < Median < Mean describes a distribution with a **positive skew**.
  Mode > Median > Mean describes a distribution with a **negative skew**.

- By calculating $\dfrac{3(mean - median)}{s\tan dard\ deviation}$
  The closer the number is to zero the more **symmetrical** the data.
  A negative number means the data is **negatively skewed**.
  A positive number means the data is **positively skewed**.
  The larger the number is, the greater the skew.

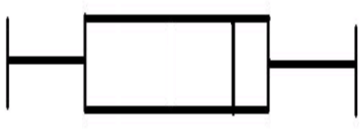
- Using Box Plots:



Symmetrical Distribution



Positively Skewed Distribution



Negatively Skewed Distribution