Topic 6: Statistics

1) The Basics:

a) Terminology:

- **Numerical:** data is numbers
  e.g.s shoe size, height, rainfall, number of kids in a family
- **Categorical:** data is text
  e.g.s favourite phone brand, tv programme, hair colour
- **Discrete:** numerical data that can only take on set values (generally whole numbers)
  e.g.s shoe size, number of kids in family
- **Continuous:** numerical data that can take on a range of values (can be decimals)
  e.g.s rainfall in mm, weight, height
- **Ordinal:** categorical data that can be put into order
  e.g. grades in an exam A, B, C….
- **Nominal:** categorical data that cannot be put into order
  e.g. phone brand
- **Primary Data**: data collected by person who's going to use it
- **Secondary Data:** data that's already available e.g. internet, magazines
- The **population** is the entire group being studied.
- A **sample** is a group that is selected from the population.
- A **census** is a survey of the whole population.
- A **sampling frame** is a list of all those within a population who can be sampled.
- An **outlier** is an extreme value that is not typical of other values in the data set.
- **Bias** can mean something which sways a respondent in a particular way or another, in a survey/questionnaire. The term bias can also be used if a sample doesn't reflect the population. E.g. selecting people coming out of Lidl and asking them their opinion on shopping in non-Irish owned retailers.

b) Collecting Data:

Notes: When selecting people to survey it is important that:

- the sample is selected randomly to avoid bias
- the sample represent the population
- the sample is sufficiently large

Methods of Collecting Data:

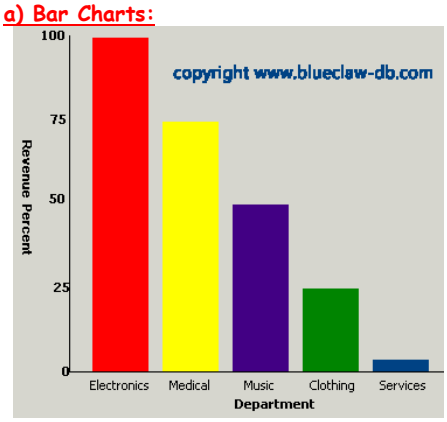
- **Phone Interview:**
  Advs: questions can be explained can select sample from entire population
  Disadvs: expensive compared to post or online
- **Online Questionnaire:**
  Advs: cheap, anonymous so answers are more honest
  Disadvs: people may not respond, not representative of entire population…only those that are online
- **Face to Face Interview:**
  Advs: questions can be explained
  Disadvs: people might not answer honestly when asked in person, expensive and not random
- **Postal Questionnaire:**
  Advs: not expensive
  Disadvs: people don't always respond
- **Observation:**
  Advs: low cost, easy to carry out
  Disadvs: not suitable for some surveys, questions can't be explained

Tips for designing a questionnaire:

- Use clear & simple language
- Begin with simple questions
- Accommodate all possible answers
- Contain no leading questions
- Be as brief as possible
- Be clear where answers should be recorded
- Avoid personal questions

2) Graphing Data from Junior Cert:

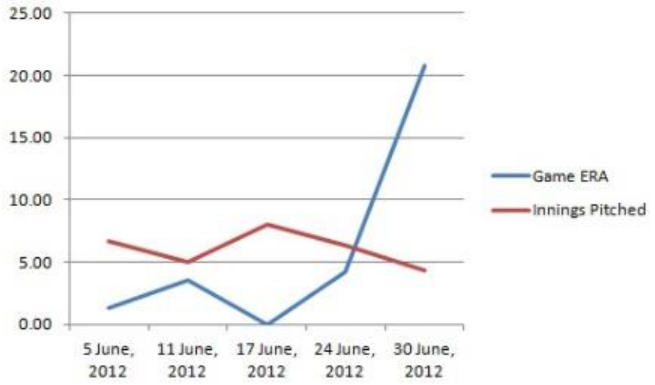
a) Bar Charts:



Notes:

- Individual bars must be labelled and axes labelled
- Must be an even scale on vertical (e.g. going up in 25s in example above)
- Bars and axes drawn with ruler
- Can be used for categorical data

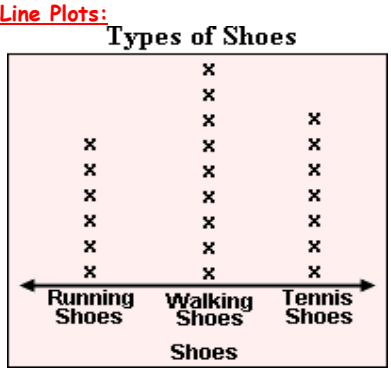
b) Trend Graphs:



Note:

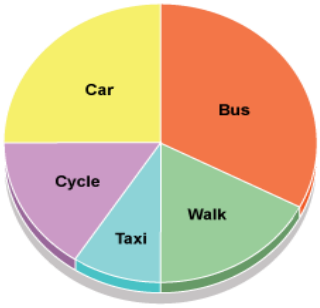
- Axes labelled and scaled evenly

c) Line Plots:



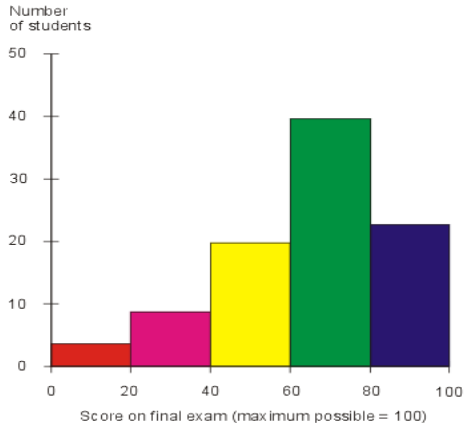
Notes:
- Clear columns and rows of 'x' (See Diagram)
- Each column labelled
- Can be used for categorical data

d) Pie Charts:



Notes:
- Circle drawn with compass
- Angles measured with protractor
- Label sectors and angles of the pie chart
- Can be used for categorical data

e) Histograms:



Notes:
- Different to bar chart as there is **a scale along the bottom** as well
- Axes labelled and evenly scaled
- Axes and bars drawn with ruler
- Can be used for continuous numerical data

f) Stem & Leaf Plots:

Stem	Leaf
5	2 3
6	5 7 8 9
7	0 2 2
8	3

Key: 5|2 = 52%

John: 46, 58, 78, 57, 79, 45 James: 40, 62, 71, 77, 52, 66

JAMES' MARKS	STEM	JOHN'S MARKS
0	4	5, 6
2	5	7, 8
6, 2	6	
7, 1	7	8, 9
Key: 1|7 = 71%		Key: 5|7 = 57%

Notes:
- Clear columns and rows of numbers (see diagram)
- Key MUST be included (see diagram)
- Can use comma separated list for leaves also
- Can use back-to-back plots to compare two sets of data (**bivariate data**)

3) Analysing Data:

a) Measures of Centre:

1. **Mean:** the sum of all the values divided by the number of values

e.g. Data: 1, 4, 3, 5, 4, 2, 1

Mean = $\frac{1+4+3+5+4+2+1}{7} = 2.86$

- Only used with numerical data
- **Advs:** uses all the data
- **Disadvs:** affected by outliers

2. **Mode:** the value that appears the most often

e.g. Data: 2, 3, 1, 2, 5, 4, 2, 1, 2

Mode = 2 (as it appears 4 times)

- Can be used for numerical but the only one that can be used for categorical data
- **Advs:** Not affected by outliers, can be used for any data
- **Disadvs:** There is not always a mode, does not use all the data

3. **Median:** the middle value (list must be in ascending order)

e.g. Data: 2, 1, 3, 3, 2, 5, 3, 2, 1

Rearrange in order first: 1, 1, 2, 2, 2, 3, 3, 3, 5

=> Median = 2

- Used only with numerical data
- **Advs:** Easy to calculate, not heavily affected by outliers
- **Disadvs:** Does not use all the data

b) Measures of Spread:

Note: For the following, the list of values should be in ascending order

Range: the difference between the max and the min value

e.g. Data: 20, 40, 40, 45, 60 => Range = 60 – 20 = 40

4) Frequency Distributions:

a) Frequency Distributions:

- A **frequency distribution** is a way of grouping together a large amount of data into a table. E.g.

No. in Household	2	3	4	5	6	7
No. of People	6	8	14	11	4	1

- Always remember what this table represents…..i.e. a full list of data: 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4……

c) Grouped Frequency Distributions:

- If the frequency distribution is a **grouped frequency distribution**, all the calculations shown above are the same except we use **mid-interval values** instead. E.g.

Age	0-10	10-20	20-30	30-40
Freq	2	5	4	8

- The **mid-interval values** for the age row are 5, 15, 25 and 35.
- We now proceed to find mean, median and mode as in (b).

b) Mean, Mode and Median of a Frequency Distribution:

Mode: Can be read straight away from the table on the left
=> Mode = 4 as it appears the most often (14 times)

Mean:

- We could add up all the values in the full list, shown below the table above, and then divide by 44
- Quicker way is to multiply the columns together from the table i.e. (2x6)+(3x8)+(4x14)+(5x11)+(6x4)+(7x1)
- We then divide this by 44 to get a mean of 4.04

Median:

- Count up how many values we have in total by adding the bottom row i.e. 6 + 8 + 14 + 11 + 4 + 1 = 44
- This means that the median here will be the average of the 22nd and 23rd values.
- We can find the 22nd and 23rd values from the table above i.e. the first 14 values are '2' and '3' and the next 14 values are '4', which would include the 22nd and 23rd values
  => Median = $\frac{4+4}{2}$ = 4