

Topic 7: 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

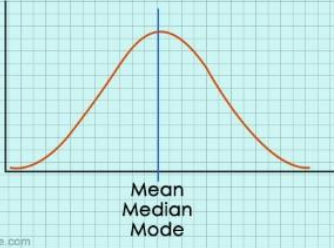
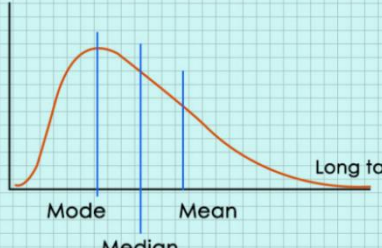
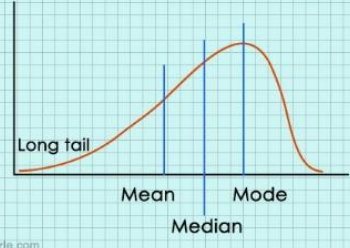
Methods of Collecting Data:

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

Tips for designing a questionnaire:

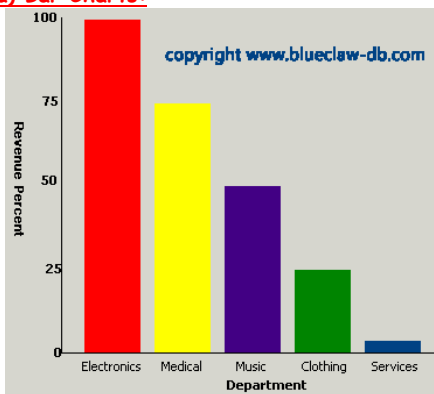
- Use clear & simple language
- Begin with simple questions
- Accommodate all possible answers
- Contain no leading questions
- Be as brief as possible
- Be clear where answers should be recorded
- Avoid personal questions

c) Describing Distributions

<p>SYMMETRIC DISTRIBUTION</p> 	<p>POSITIVELY SKEWED DISTRIBUTION</p> 	<p>NEGATIVELY SKEWED DISTRIBUTION</p> 
<p>Example: If we surveyed the height of TY students in the country it would almost certainly be approximately symmetrical</p>	<p>Example: If we surveyed the ages at which people learn to drive in the country, it would probably be positively skewed as most people learn how to drive when they are young</p>	<p>Example: If we surveyed the heights of players in the NBA, it would almost certainly be negatively skewed as the majority of them are very tall</p>

2) Graphing Data from Junior Cert:

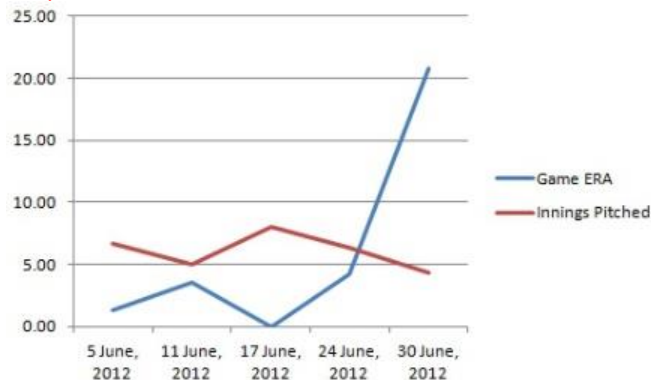
a) Bar Charts:



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

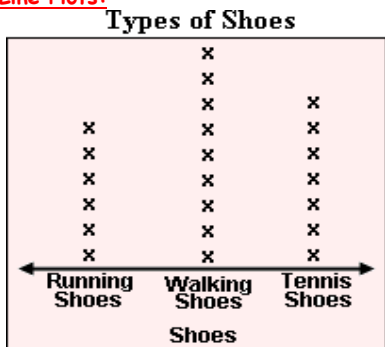
b) Trend Graphs:



Note:

- Axes labelled and scaled evenly

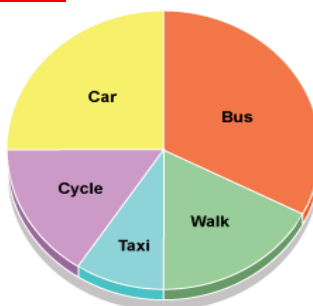
c) Line Plots:



Notes:

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

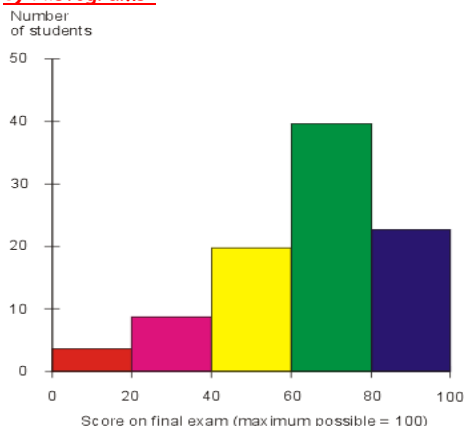
d) Pie Charts:



Notes:

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

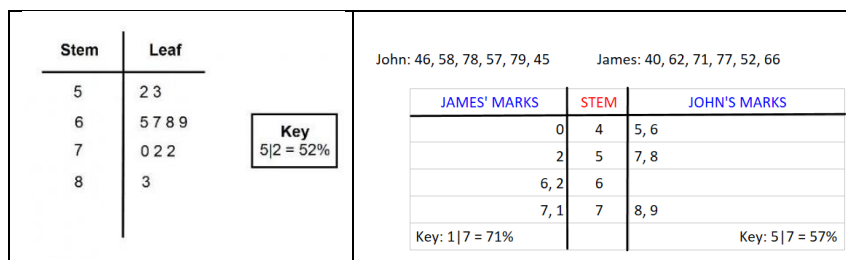
e) Histograms:



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:



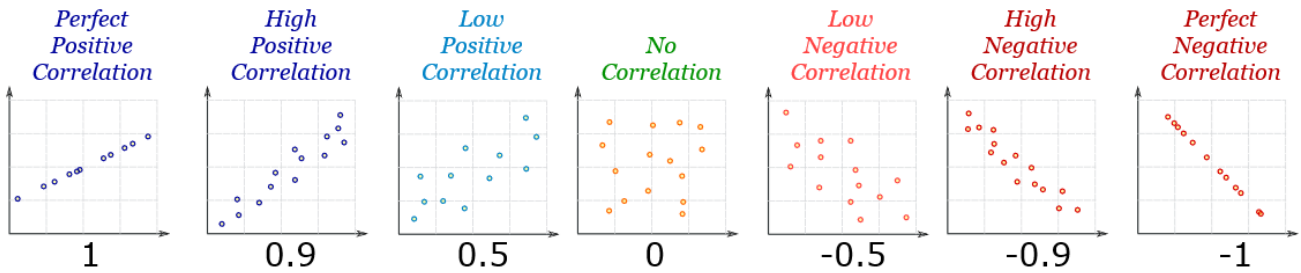
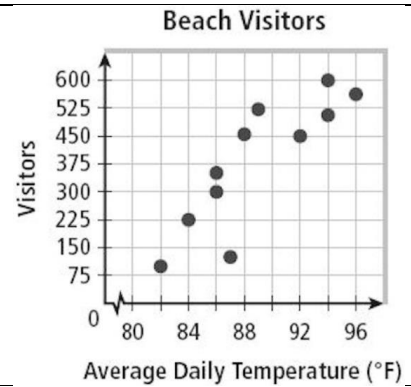
Notes:

- 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) Scatter Plots/Correlation:

Notes:

- Also used to analyse **bivariate data** - See example on the right
- Axes labelled and evenly scaled and drawn with ruler
- Can only be used with numerical data
- The **Correlation Coefficient 'r'** is a measure of how closely the two variables correlate i.e. how much a change in one affects the other
- Important to note that **correlation doesn't always mean causality** i.e. just because two variables fit one of the patterns below, doesn't mean that one necessarily affects the other.
- r can be between -1 and 1 (See diagrams below)
- Need to be able to estimate the correlation coefficient from a graph of data



4) 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

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

- Only used with numerical data
 - **Adv:** uses all the data
 - **Disadv:** 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
 - **Adv:** Not affected by outliers, can be used for any data
 - **Disadv:** 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
 - **Adv:** Easy to calculate, not heavily affected by outliers
 - **Disadv:** Does not use all the data

b) Measures of Spread:

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

1. **Range:** the difference between the max and the min value
 e.g. Data: 20, 40, 40, 45, 60 => Range = 60 - 20 = 40
2. **Lower Quartile:** the quarter mark (Remember: Find the median, and then find the median of the lower half of the data)
 e.g. Data: 20, 30, 35, 50, 55, 60, 70, 75
 8 values => $\frac{8+1}{2} = 4.5$, which is between 4th and 5th values, so the lower quartile will be the median of the lower 4 values:

$$\Rightarrow \text{LQ} = \frac{4+1}{2} = 2.5^{\text{th}} \text{ value} \Rightarrow \text{LQ} = \frac{30+35}{2} = 32.5$$
3. **Upper Quartile:** the three-quarter mark (Remember: Find the median, and then find the median of the upper half of the data)
 e.g. Data: 20, 30, 35, 50, 55, 60, 70, 75
 Using median above, Upper Quartile will be the median of the upper 4 values


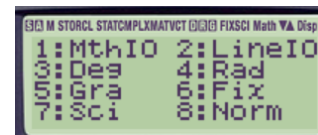

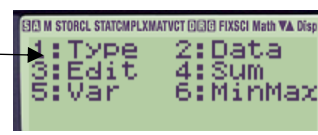
$$\Rightarrow \text{UQ} = \frac{4+1}{2} = 2.5^{\text{th}} \text{ value} \Rightarrow \text{UQ} = \frac{60+70}{2} = 65$$
4. **Interquartile Range:** the interquartile range of a set of values is the difference between the upper quartile and the lower quartile
 e.g. Data: 20, 30, 35, 50, 55, 60, 70, 75
 IQ Range = UQ - LQ = 65 - 32.5 = 32.5

5) Frequency Distributions:

<p>a) Frequency Distributions:</p> <ul style="list-style-type: none"> A frequency distribution is a way of grouping together a large amount of data into a table. E.g. <table border="1" style="margin-left: 20px; border-collapse: collapse; text-align: center;"> <tr> <td>No. in Household</td> <td>2</td> <td>3</td> <td>4</td> <td>5</td> <td>6</td> <td>7</td> </tr> <tr> <td>No. of People</td> <td>6</td> <td>8</td> <td>14</td> <td>11</td> <td>4</td> <td>1</td> </tr> </table> <ul style="list-style-type: none"> 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, 3, 3, 4, 4..... <p>c) Grouped Frequency Distributions:</p> <ul style="list-style-type: none"> 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. <table border="1" style="margin-left: 20px; border-collapse: collapse; text-align: center;"> <tr> <td>Age</td> <td>0-10</td> <td>10-20</td> <td>20-30</td> <td>30-40</td> </tr> <tr> <td>Freq</td> <td>2</td> <td>5</td> <td>4</td> <td>8</td> </tr> </table> <ul style="list-style-type: none"> 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). 	No. in Household	2	3	4	5	6	7	No. of People	6	8	14	11	4	1	Age	0-10	10-20	20-30	30-40	Freq	2	5	4	8	<p>b) Mean, Mode and Median of a Frequency Distribution:</p> <p>Mode: Can be read straight away from the table on the left \Rightarrow Mode = 4 as it appears the most often (14 times)</p> <p>Mean:</p> <ul style="list-style-type: none"> 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. $(2 \times 6) + (3 \times 8) + (4 \times 14) + (5 \times 11) + (6 \times 4) + (7 \times 1)$ We then divide this by 44 to get a mean of 4.04 <p>Median:</p> <ul style="list-style-type: none"> 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 \Rightarrow Median = $\frac{4+4}{2} = 4$
No. in Household	2	3	4	5	6	7																			
No. of People	6	8	14	11	4	1																			
Age	0-10	10-20	20-30	30-40																					
Freq	2	5	4	8																					

6) Standard Deviation:

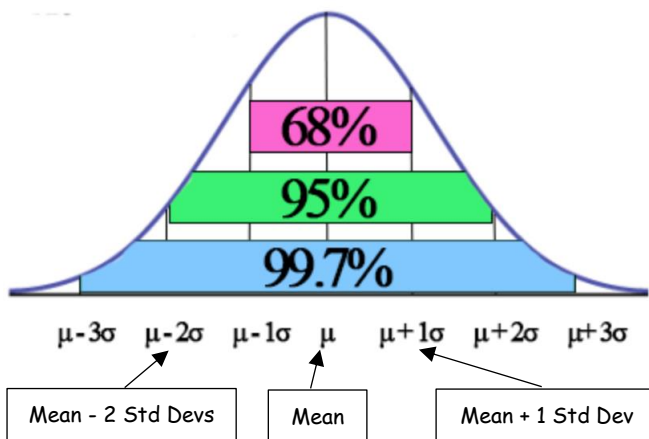
<p>a) Standard Deviation:</p> <p>Notes:</p> <ul style="list-style-type: none"> Used to measure spread using all the data. Symbol: σ The higher the value, the more spread out the data is. <p>Steps:</p> <ol style="list-style-type: none"> Calculate the mean. Subtract the mean from all data values. Square all the values from step 2. Add up all the answers from step 3. Divide answer to step 4 by the total number of values. Take the square root of the answer to step 5. 	<p>Example: Find the standard deviation of 4, 8, 3, 2, 7, 6.</p> $\text{Mean} = \frac{4+8+3+2+7+6}{6} = \frac{30}{6} = 5$ $\sigma = \sqrt{\frac{(4-5)^2 + (8-5)^2 + (3-5)^2 + (2-5)^2 + (7-5)^2 + (6-5)^2}{6}}$ $\sigma = \sqrt{\frac{1+9+4+9+4+1}{6}}$ $\sigma = \sqrt{4.66666666} = 2.16$
---	--

	<table border="1" style="border-collapse: collapse; text-align: center;"> <tr> <td>x</td> <td>5</td> <td>8</td> <td>11</td> <td>14</td> <td>17</td> <td>20</td> </tr> <tr> <td>f</td> <td>8</td> <td>5</td> <td>3</td> <td>9</td> <td>5</td> <td>2</td> </tr> </table>	x	5	8	11	14	17	20	f	8	5	3	9	5	2	
x	5	8	11	14	17	20										
f	8	5	3	9	5	2										
<p>b) Calculator Use</p> <p>Step 1:</p> <ul style="list-style-type: none"> Turn on STAT mode by pressing 'MODE' and then '2'. Press '1' then to select "1-VAR" mode. The screen should now look like the screen on the right. <p>Note: If the "FREQ" column is not visible, then follow the following steps:</p> <ul style="list-style-type: none"> Press 'SHIFT' and then 'MODE' to enter the screen shown on the right. Press the Down Arrow and then press '3' for "STAT". Now press '1' to turn the frequency setting ON. 	 															
<p>Step 2:</p> <ul style="list-style-type: none"> Enter the data from the table above by typing in the value and pressing '=' every time. Use the arrows to navigate between the columns. When all the data has been entered, the screen should look like the screen on the right. Now press 'AC' to clear the screen. 																
<p>Step 3:</p> <ul style="list-style-type: none"> Press 'SHIFT' and '1' to enter the STAT menu, which looks something like the screen on the right. Press the number corresponding to the "VAR" option. Now press the number that corresponds to the "Sigma x" and then press '=' to get the standard deviation of 4.85. <p>Note: We could also find the mean of the distribution by selecting \bar{x} from the menu instead. Verify that the mean of this distribution is 11.375.</p>																

7) Empirical Rule:

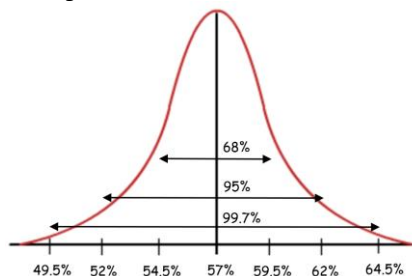
Notes:

- Rule is used to make some estimates of populations that are **normally distributed**.
- Need to calculate the mean of the data to write in the value for μ in the diagram on the right.
- Need to calculate the standard deviation σ to work out the other values along the bottom of the diagram.



Example: A 6th year Maths class results have a mean of 57% and a standard deviation of 2.5%. There are 25 in the class. Estimate the following:

- the percentage of the class that scored between 52% and 62%
- the number of students who scored between 54.5% and 57%
- The percentage of the class that scored above 64.5%



- From the diagram above the answer is 95%.
- If 68% are between 54.5% and 59.5%, then there must be half that between 54.5% and 57%
=> 34% of the class => 34% of 25 = 8.5 students
- The percentage outside of the range 49.5% and 64.5% is 100% - 99.7% = 0.3%
=> Half of this percentage must be above 64.5% => 0.15%

8) Hypothesis Testing:

a) Population/Sample Proportions:

Notes:

- The **population proportion** p is the percentage of the entire population.
E.g. if the number of Leaving Cert students who get an A in Ordinary Level Maths in the country is 500 out of 12000, then the population proportion is $\frac{500}{12000} = 0.04$
- The **sample proportion** \hat{p} is the percentage of the sample.
E.g. if the number of Leaving Cert students who get an A in a sample of 10 Maths classes across the country is 12 out of 250 then the sample proportion is $\frac{12}{250} = 0.05$
- The **Margin of Error** can be found using:

$$E = \frac{1}{\sqrt{n}}$$

where n is the size of the sample.

b) Confidence Intervals:

Steps:

- Calculate the sample proportion \hat{p} .
- Find the Margin of Error $E = \frac{1}{\sqrt{n}}$.
- Construct the Confidence Interval using:

$$\hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$

Example: A company wants to check an item that it's producing for defects. A random sample of 30 products is taken. 6 of the sample were defective. Construct a 95% interval for p .

$$\text{Sample Proportion } \hat{p} = \frac{6}{30} = 0.2$$

$$\text{Margin of Error } = E = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{30}} = 0.18$$

Confidence Interval:

$$\hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$

$$0.2 - 0.18 < p < 0.2 + 0.18$$

$$0.02 < p < 0.38$$

c) Hypothesis Testing:

Notes:

- A **hypothesis** is a statement or claim about a population.
- A **hypothesis test** is a method of testing a claim.
- The **null hypothesis** is a statement that describes the population proportion.

Steps:

- State the null hypothesis H_0 and the alternative H_1 .
- Calculate the sample proportion \hat{p} .
- Calculate the Margin of Error E .
- Construct a Confidence Interval for p .
- If the value of p stated in the null hypothesis is:
 - inside the Confidence Interval, then "Fail to Reject H_0 ".
 - outside the Confidence Interval, then "Reject the Null Hypothesis H_0 in favour of the alternative H_1 ".

Example: A make-up company advertises that 75% of its customers liked a new product they released. In a random sample of 300 people, 230 said they liked the product. Can we reject the company's claim that 75% are satisfied?

H_0 : 75% of customers do like the product

H_1 : The % that like the product is **not** 75%

$$\text{Sample Proportion} = \frac{230}{300} = 0.77$$

$$\text{Margin of Error} = E = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{230}} = 0.066$$

$$\text{Confidence Interval: } \hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$

$$0.77 - 0.066 < p < 0.77 + 0.066$$

$$0.704 < p < 0.84$$

The population proportion of 75% (0.75) is within this confidence interval so we fail to reject the hypothesis => we are not rejecting the claim.