Topic 10: 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 eye colour, hair colour
- **Discrete:** numerical data that only takes on set values (usually whole numbers) e.g.s shoe size, number of kids in house
- **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
- 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 and is sufficiently large

Methods of Collecting Data:

- **Phone Interview:**
  Advs: questions can be explained, can select from entire popn
  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, Qs can't be explained

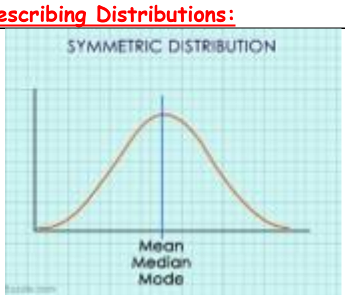
Tips for designing a questionnaire:

- Use clear & simple language
- Begin with simple questions and avoid personal questions
- Accommodate all possible answers & no leading question
- Be as brief as possible and be clear where answers go

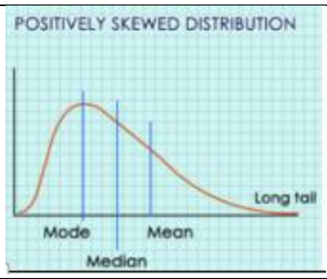
c) Types of Sampling:

1. **Simple Random Sample:** A sample of a certain size selected in such a way that each sample of that size has an equal chance of being selected. E.g. put names of the population being studied in a hat, and draw out the names.
2. **Stratified Random Sample:** - Population divided into two or more subgroups with similar characteristics - a proportional sample is drawn from each subgroup. E.g. to get the attitudes of students in a school to underage drinking - then a sample is selected from each group. If there are twice as many 2^{nd} years as 1^{st} years, then the sample of 2^{nd} years should be twice as big as the 1^{st} year sample.
3. **Cluster Sample:** Population divided into clusters and then the clusters selected randomly. E.g. political party wants to get opinions of citizens leaving polling stations on an election day - polling stations would be clusters of the population - a number of polling stations are selected and **everyone** coming out of that station is surveyed.
4. **Quota Sampling:** Person selecting the sample is given a quota to fill and selects it in the most convenient way. E.g. a company wants opinion of men under 25 yrs on an issue - person collecting the data stops random people in the street who are under 25 - not random and open to mistakes.
5. **Systematic Sampling**: The person selecting the sample chooses every n^{th} person from the population. E.g. if Tesco wanted to survey their customers, they could select every 20^{th} person that enters their stores on a particular day and survey them. A disadvantage would be the easy introduction of bias depending on who the n^{th} people are e.g. if every 20^{th} person was a pensioner than the results are not representative of the population of Tesco shoppers.
6. **Convenience Sampling**: You survey those that are easiest for you to reach. E.g. Surveying people from your workplace or school. One disadvantage of this method is the selection isn't random.

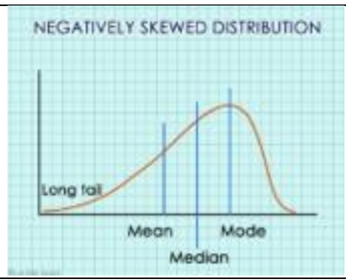
d) Describing Distributions:



Example: Surveyed the height of TY students in the country it would almost certainly be approximately symmetrical



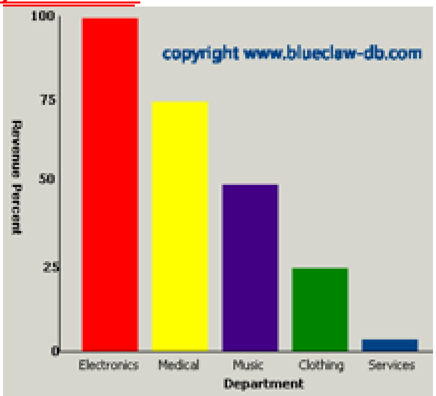
Example: Survey the ages at which people learn to drive in the country, it might be positively skewed as most people learn when they are young



Example: Surveyed the heights of players in the NBA, it would almost certainly be negatively skewed as the majority of them are very tall

2) Graphing Data from Junior Cert:

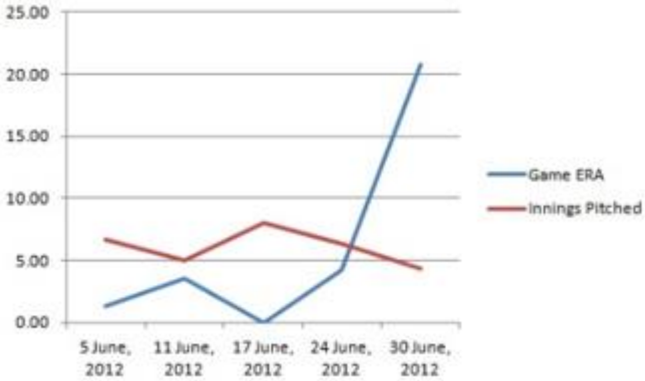
a) Bar Charts:



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

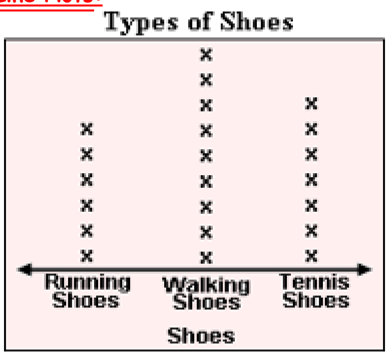
b) Trend Graphs:



Note:

- Axes labelled and scaled evenly

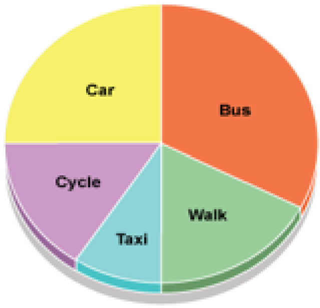
c) Line Plots:



Notes:

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

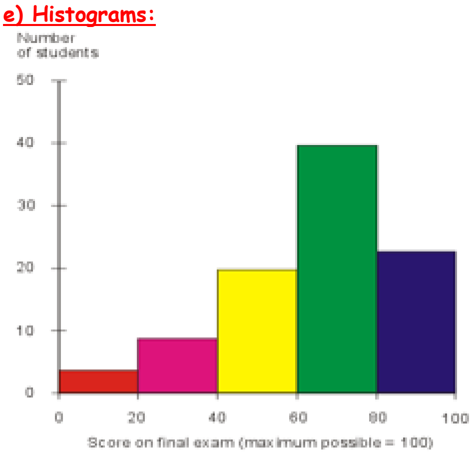
d) Pie Charts:



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:

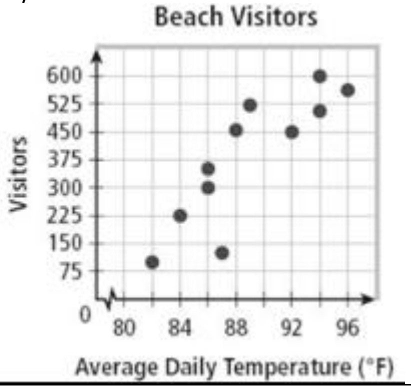
- Clear columns and rows of numbers (see diagram)
- Key MUST be included (see diagram)
- Can use comma separated list for leaves also
- Can use back-to-back plots to compare two sets of data (**bivariate data**)

3) Scatter Plots/Correlation:

a) Scatter Plots:

Notes:

- Also used to analyse **bivariate data** – See example below.
- Axes labelled and evenly scaled and drawn with ruler.
- Can only be used with numerical data.



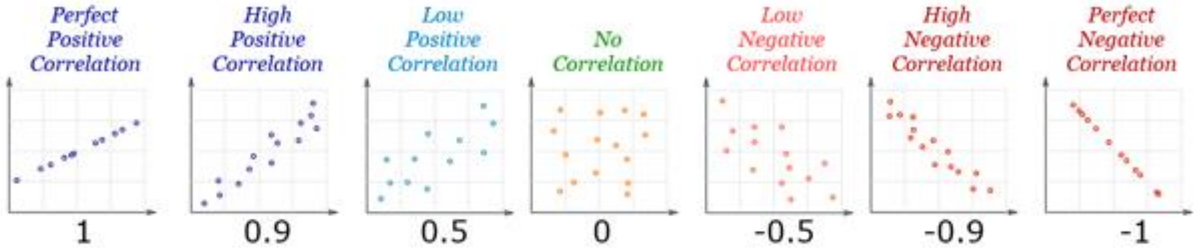
b) Correlation Coefficient:

Notes:

- The **Correlation Coefficient (r)** measures how strong a relationship is between two variables.
- It can have values between -1 and 1 i.e. $-1 \le r \le 1$
- If r = 1, then the correlation is said to be **strong** and **positive** and would be a straight line.
- If r = -1, then the correlation is said to be **strong** and **negative** and would also be a straight line.
- The further away the coefficient gets from 1 or -1, the **weaker** the correlation.
- 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.

c) Estimation of Correlation Coefficient:

- Need to be able to estimate the correlation coefficient from a graph of data.



d) Calculation of Correlation Coefficient:

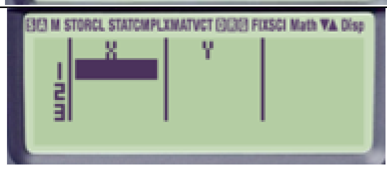
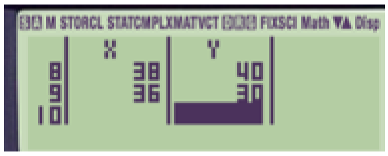
Variable 1	35	42	51	38	44	37	48	38	36
Variable 2	31	33	46	32	53	37	32	40	30

Step 1: Putting the calculator in the correct mode.

- Press 'MODE' and then '2' to enter STAT mode.
- The menu shown on the right should now be shown. Press '2' to select the "A+BX" mode

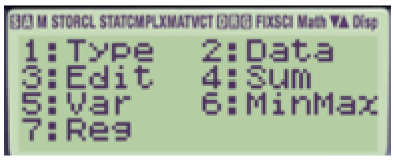
Step 2: Entering the data.

- The screen should now look like the screen on the right. There might be an extra "FREQ" column but that will have 1s in it that won't affect anything.
- The data for "Variable 1" in the table above goes into Column X and the data for "Variable 2" goes into Column Y.
- Type the first data value '35' and then press '='
- Enter the rest of the data for Variable 1 by typing in the value and pressing '=' each time.
- Repeat the steps above for Variable 2.
- The screen should now look like the screen on the right.
- The data is now entered so press 'AC' to clear the screen

Step 3: Calculating the correlation coefficient 'r'.

- Press 'SHIFT' and '1' to enter the STAT menu. It looks like the screen on the right.
- Press '7' to access the REG menu.
- Now press '3' to access the correlation coefficient 'r'.
- The screen should now have an 'r' on the top of it. Press '=' to get the answer of **0.495**.



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

   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

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 4^{th} and 5^{th} values, so the lower quartile will be the median of the lower 4 values:

   => LQ = $\frac{4+1}{2} = 2.5^{th}$ value => 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

   => UQ = $\frac{4+1}{2} = 2.5^{th}$ value => 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) Measures of Relative Standing:

a) Percentiles:

Notes:

- Percentiles divide a data set up into 100 equal parts.
- P_{50} would be the 50^{th} percentile, which means 50% of the data is **lower** than this value i.e. the median

Steps to find the k^{th} percentile P_k:

1. Rank the data
2. Count the number of values and add 1 to it.
3. Find k% of the number from step 2, and call it 'c'
   1) If 'c' is a whole number, then this represents the value in the data set
   2) If 'c' is not a whole number, then find the mean of the c^{th} and $(c + 1)^{th}$ value

Example: Find i) P_{40} and ii) P_{75} for the following set of Maths results:

97, 94, 88, 95, 96, 81, 83, 92, 80, 87, 93, 92, 89, 83, 95

i) To find P_{40}

- Rank the data first:

  80, 81, 83, 83, 87, 88, 89, 92, 92, 93, 95, 95, 95, 96, 97

- There are 15 values here so we add 1 to 15 in a similar way as we did when finding quartiles and the median. To find P_{40} we need to find 40% of 16:

  40% of 16 = 6.4

- 6.4 is between 6 and 7 so that means we want the average of the 6th and 7th value i.e.

  $P_{40} = \frac{88+89}{2} = 88.5$ => 40% of the data is below 88.5

ii) To find P_{75}

- We need to find 75% of 16 this time

  75% of 16 = 12 so this is the 12th value

- So that means

  P_{75} = 95 => 75% of the data is below 95

b) Z-Scores:

Notes:

- Another way of comparing values in a data set.
- Z-scores tell us how many standard deviations a particular value is from the mean.
- To calculate the z-score for a particular data value we use the formula:

$$z = \frac{x - \mu}{\sigma}$$

See pg34 Tables book

where μ is the mean and σ is the standard deviation.

- If a z-score is bigger than 2 or less than -2, then the data value is said to be unusual.

Example: A particular Maths class had a mean result of 63% and a standard deviation of 7%. If a certain student got a result of 48%, comment on his performance.

Solution:

- First, let's calculate the z-score:

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{48 - 63}{7}$$

$$z = \frac{-15}{7} = -2.14$$

- This z-score tells us that the student's result was unusual in the context of this particular test. The majority of the class would have scored between 49% (2 standard deviations below the mean) and 77% (2 standard deviations above the mean)

6) 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

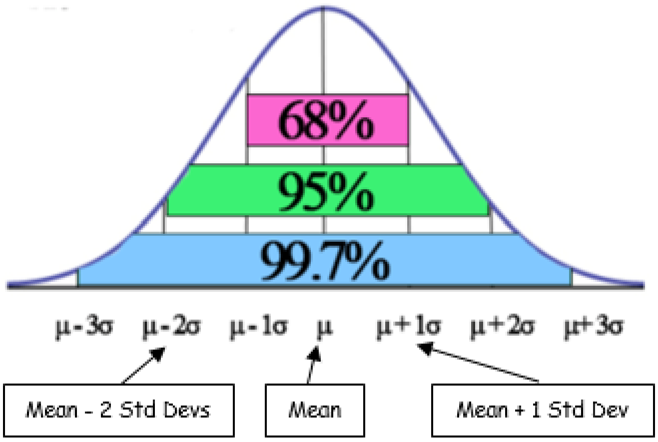
Median:

- 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

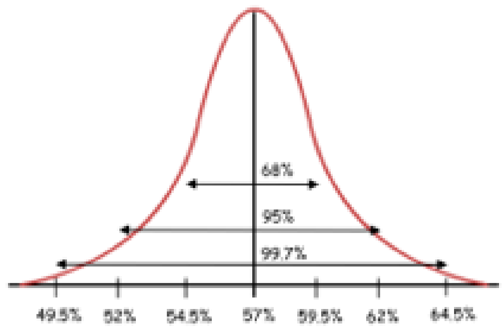
7) Empirical Rule:

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:
i) the percentage of the class that scored between 52% and 62%
ii) the number of students who scored between 54.5% and 57%
iii) The percentage of the class that scored above 64.5%



i) From the diagram above the answer is 95%.
ii) 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
iii) 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) Standard Deviation:

a) Standard Deviation:

Notes:

- Used to measure **spread** using **all the data**.
- Symbol: σ
- The higher the value, the more spread out the data is.

Steps to calculate by hand:

1. Calculate the mean.
2. Subtract the mean from all data values.
3. Square all the values from step 2.
4. Add up all the answers from step 3.
5. Divide answer to step 4 by the total number of values.
6. Take the square root of the answer to step 5.

Example: Find the standard deviation of 4, 8, 3, 2, 7, 6.

Mean = $\frac{4+8+3+2+7+6}{6} = \frac{30}{6} = 5$

$$\sigma = \sqrt{\frac{(4-5)^2 + (8-5)^2 + (3-5)^2 + (2-5)^2 + (7-5)^2 + (6-5)^2}{6}}$$

$$\sigma = \sqrt{\frac{1 + 9 + 4 + 9 + 4 + 1}{6}}$$

$$\sigma = \sqrt{4.6666666} = 2.16$$

b) Calculator Use:

x	5	8	11	14	17	20
f	8	5	3	9	5	2

Step 1:

- 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.

Note: If the "FREQ" column is not visible, then follow the following steps:

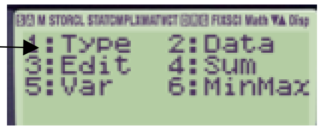
- Press 'SHIFT' and then 'MODE' to enter the screen shown on the right.
- Press the Down Arrow and then press '3' for "STAT".
- Now press '1' to turn the frequency setting ON.

Step 2:

- 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.

Step 3:

- 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 "σ x" and then press '=' to get the standard deviation of 4.85.



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.

c) Standard Deviation of Frequency Distribution by Hand:

Notes:

- Generally easier to use the calculator but need to be familiar with how to calculate by hand also.

Example: Find the standard deviation of the frequency distribution below by hand.

x	5	8	11	14	17	20
f	8	5	3	9	5	2

- Before calculating the standard deviation of a set of data, we have to calculate the mean.
- We will make out a table:

x	f	xf	d = $x - \bar{x}$	d²	d²f
5	8	40	-6.4	40.96	327.68
8	5	40	-3.4	11.56	57.8
11	3	33	-0.4	0.16	0.48
14	9	126	2.6	6.76	60.84
17	5	85	5.6	31.36	156.8
20	2	40	8.6	73.96	147.92
$\sum$	32	364			751.52

The mean $= \frac{\sum xf}{\sum f} = \frac{364}{32} = 11.4$

And so, the standard deviation is $= \sqrt{\frac{\sum d^2 f}{\sum f}} = \sqrt{\frac{751.52}{32}} = 4.85$

9) Inferential Statistics:

a) Population/Sample Proportions:

Notes:

- We use the following symbols:
  - p = **population proportion** (the % of the entire popn)

  E.g. if 500 students out of 12000 in the country get an A in Maths, then the population proportion is $\frac{500}{12000} = 0.04$
  - $\hat{p}$ = **sample proportion** (the % of the sample)

  E.g. in a sample of 10 Maths classes across the country 12 out of 250 get an A in Maths then the sample proportion is $\frac{12}{250}$
- The **standard error** of the proportion is given by:

$$E = 1.96\sqrt{\frac{p(1-p)}{n}}$$

See Tables pg34 **

- Generally, we don't know p and we have to use $\hat{p}$ in the formula.
- A 95% confidence interval for the population proportion is:

$$\hat{p} - E < p < \hat{p} + E$$

** The formula in the tables doesn't contain the 1.96, as that figure is added when looking for 95% confidence.

Example: Company wants to get an estimate of the proportion of employees on sick leave. In a sample of 20, 9 reported that they had taken sick leave. Construct a 95% confidence interval for p.

- The sample proportion is : $\hat{p} = \frac{9}{20} = 0.45 \text{ or } 45\%$
- The margin of error is : $E = 1.96\sqrt{\frac{0.45(1-0.45)}{20}} = 0.218$
- And now, we can set up our confidence interval:

$$0.45 - 0.218 < p < 0.45 + 0.218$$
$$= 0.232 < p < 0.668$$

b) Population Mean:

Notes:

- We use the following symbols:

  μ = population mean, σ = population standard deviation

  $\bar{x}$ = mean of the sample, s = standard deviation of the sample

  $\sigma_{\bar{x}}$ = standard error of the mean
- The Central Limit Theorem is used in these problems:

Once the sample size ≥ 30, then 1) $\mu = \bar{x}$ 2) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ 3) Shape of sampling distribution is symmetric

- Central Limit Theorem holds for sample sizes of < 30, if the original population is normally distributed.
- The **Margin of Error** can be found using:

$$E = 1.96\frac{\sigma}{\sqrt{n}}$$

where n is the size of the sample, and σ is the standard deviation.

- A 95% confidence interval for the mean of the population is given by:

$$\bar{x} - E < \mu < \bar{x} + E$$

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:

1. State the null hypothesis H_0 and the alternative H_1.

Method 1: Using 95% Confidence Interval (can be used for either population proportion or population mean type questions)

2. Work out the 95% confidence interval and see if the value in the null hypothesis is in the interval or not
3. If the value is outside the interval, then we "reject the null hypothesis and accept the alternative".
4. If the value is inside the interval, then we "fail to reject the null hypothesis".

Method 2: Using z-scores and critical regions (can be used for population mean questions only)

2. Calculate the z-score for the claimed value
3. For 95% confidence, we check if the z-score is **above or below ±1.96** i.e. in the "critical regions".

- If the value is above/below ±1.96, then we "reject H_0, and accept H_1"
- If the value is within ±1.96 of the mean, then we "fail to reject H_0"

Method 3: Using p-values (See part (d))

d) P-Values:

Notes:

- The p-Value is the probability of getting a result as extreme as the actual value of the z-score, for that data value.

Steps:

1. State the null hypothesis H_0 and the alternative H_1.
2. Calculate the z-score for the claimed data value.
3. Use the tables to calculate the probability of a value being as extreme, or more extreme than that data value.
4. Double the probability you get from step 3.
5. If the p-value from step 4, is **less than** the level of significance (usually **5% = 0.05**), then we "reject the H_0 and accept H_1".
4. If p-value is from step 4 is **> 0.05**, then we "fail to reject H_0".

Example: Food company claims mean weight of packets of museli is 500g, with a std dev of 15g. Random sample of 64 shows a mean weight of 496g.

- State the null hypothesis and the alternative:

$$H_0: \mu = 500 \qquad H_1: \mu \neq 500$$

- For the sample $\bar{x} = 496$.
- So, the test statistic T is then:

$$T = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{496 - 500}{\left(\frac{15}{\sqrt{64}}\right)} = -2.13$$

=> From tables the p-Value = $2(1 - 0.9834)$ = 0.0332

- At a 5% level of significance, $\alpha = 0.05$. As the p-Value is less than α, the result is significant, and we reject H_0.