

YEAR 1 MATHS - STATISTICS KNOWLEDGE ORGANISER

PLANNING AND DATA COLLECTION

- **PROBLEM SPECIFICATION AND ANALYSIS**
What is the purpose of the investigation?
What data is needed?
How will the data be used?
- **DATA COLLECTION**
How will the data be collected?
How will bias be avoided?
What sample size is needed?
- **PROCESSING AND REPRESENTING**
How will the data be 'cleaned'?
Which measures will be calculated?
How will the data be represented?
- **INTERPRETING AND DISCUSSING**

1 DATA COLLECTION

Types of data Categorical/Qualitative data – descriptive
 Numerical/ Quantitative data

Sampling Techniques

Simple random Sampling - each member of the population has an equal chance of being selected for the sample

Systematic – choosing from a **sampling frame** - if the data is numbered 1, 2, 3, 4....randomly select the starting point and then select every nth item in the list

Stratified - A stratified sample is one that ensures that subgroups (strata) of a given population are each adequately represented within the whole sample population of a research study.

Sample size from each subgroup = $\frac{\text{size of whole sample}}{\text{size of whole population}} \times \text{population of the subgroup}$

Quota Sampling - sample selected based on specific criteria e.g age group

Convenience / opportunity sampling – e.g the first 5 people who enter a Leisure Centre or teachers in single primary school surveyed to find information about working in primary education across the UK

Self Selecting Sample – people volunteer to take part in a survey either remotely (internet) or in person

2 PROCESSING AND REPRESENTATION

Categorical/Qualitative data Pie Charts
 Bar charts (with spaces between the bars)
 Compound/Multiple Bar charts
 Dot charts
 Pictograms

Modal Class – used as a summary measure

Numerical/ Quantitative data

Represented using – Frequency diagrams
Histograms
Cumulative Frequency diagrams
Box and Whisker Plots

Measures of central tendency

- Mode (can have more than one mode)
- Median – middle value of ordered data
- Mean $\frac{\sum x}{n}$ or $\frac{\sum fx}{\sum f}$

If the mean is calculated from grouped data it will be an **estimated mean**

Measures of Spread

- Range (largest – smallest value)
- Inter Quartile Range : Upper Quartile – Lower Quartile (not influenced by extreme values)
- Standard Deviation (includes all the sample)

Finding the quartiles (sample size = n)

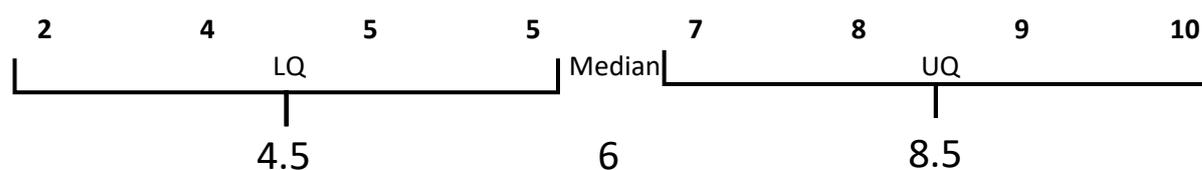
n is odd Data : 2, 4, 5, 7, 8, 9, 9



Lower Quartile : middle value of data less than the median

Upper Quartile : middle value of data greater than the median

n is even Data : 2, 4, 5, 5, 7, 8, 9, 10



Lower Quartile : middle value of the lower half of the data

Upper Quartile : middle value of the upper half of the data

STANDARD DEVIATION

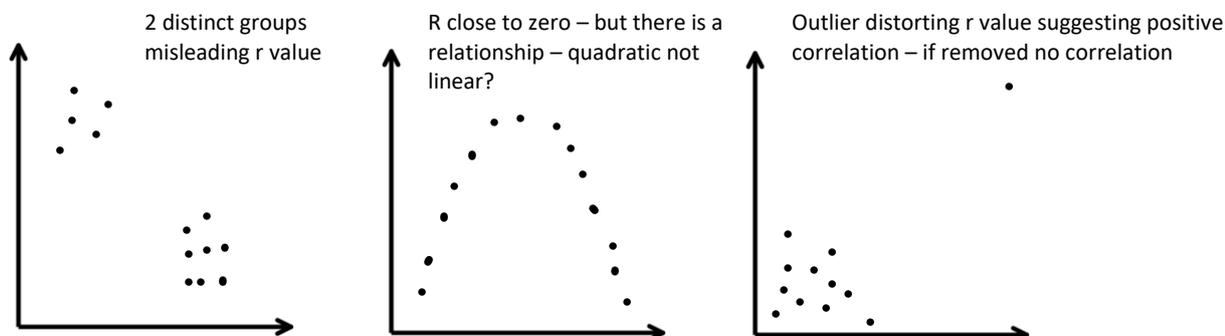
Standard deviation $s = \sqrt{\frac{S_{xx}}{n-1}}$ where $S_{xx} = \sum(x - \bar{x})^2$ or $S_{xx} = \sum x^2 - n\bar{x}^2$
or $S_{xx} = \sum fx^2 - n\bar{x}^2$

Variance = $\frac{S_{xx}}{n-1}$

3 BIVARIATE DATA – investigating the ‘association/ correlation’ between 2 variables

- The explanatory/control/independent variable is usually plotted on the horizontal axis
- A numerical measure of correlation can be calculated (Spearman’s Rank, Product Moment correlation coefficient) $-1 < r < 1$
 - 1 perfect negative correlation
 - 0 no correlation
 - 1 perfect positive correlation.

- Take care when interpreting the correlation coefficient (look at the scatter graph)



4 ‘CLEANING THE DATA’ removing ‘Outliers or Anomaly’s’

Remove values which are $1.5 \times$ **Inter Quartile range** above or below the U/L Quartile

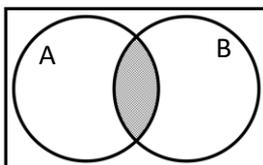
Remove values which are $2 \times$ **Standard Deviation** above or below the mean.

5 PROBABILITY

- **Outcome** : an event that can happen in an experiment
- **Sample Space** : list of all the possible outcomes for an experiment

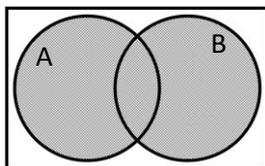
Notation

$A \cap B$ A and B **both** happen



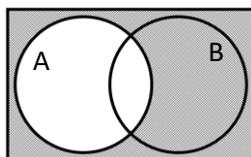
For independent events
 $P(A \cap B) = P(A) \times P(B)$

$A \cup B$ A or B or **both** happen



$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

A' A does **not** happen



$P(A') = 1 - P(A)$

Research has shown that approximately 10% of the population are left handed. A group of 8 students are selected at random.

What is the probability that less than 2 of them are left handed?

X : number of left handed students

$$p = 0.1 \quad 1 - p = 0.9 \quad n = 8$$

Less than 2 : $P(0) + P(1)$

$$P(0) = 0.9^8$$

$$P(1) = {}_8C_1 \times 0.1 \times 0.9^7$$

$$P(x < 2) = 0.813$$

(this can be found using tables or using a calculator function)

USING CUMULATIVE TABLES

- Check if you can use your calculator for this
- Remember the tables give you less than or equal to the lookup value
- List the possible outcomes and identify the ones you need to include

$$P(X < 5) \quad \boxed{0 \quad 1 \quad 2 \quad 3 \quad 4} \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad \text{Look up } x \leq 4$$

$$P(X \geq 4) \quad 0 \quad 1 \quad 2 \quad 3 \quad \boxed{4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10} \quad 1 - \text{Look up } x \leq 3$$

8 HYPOTHESIS TESTING – for binomial

- Set up the hypothesis

$H_0 : p = a$	$H_1 : p < a$ one sided test
	$H_1 : p = a$ two sided test
	$H_1 : p > a$ one sided test

- State the significance level (as a percentage) – the lower the value the more stringent the test.
- State the distribution/model used in the test Binomial (n, p)
- Calculate the probability of the observed results occurring using the assumed model
- Compare the calculated probability to the significance level – Accept or reject H_0
- Write a conclusion (in context)

Reject H_0

“There is sufficient evidence to suggest thatis underestimation/overestimating.....”

Accept H_0

“There is insufficient evidence to suggest thatincrease/decrease.....therefore conclude that $p = a$.”

The probability that patients have to wait more than 10 minutes at a GP surgery is 0.3. One of the doctors claims that there is a decrease in the number of patients having to wait more than 10 minutes. She records the waiting times for the next 20 patients and 3 wait more than 10 minutes. Is there evidence at the 5% level to support the doctors claim?

$$H_0 : p = 0.3$$

$$H_1 : p < 0.3$$

5% Significance level

X = number of patients waiting more than 20 minutes

X Binomial (20, 0.3)

Using tables $P(X \leq 3) = 0.107$ (10.7%)

$$10.7\% > 5\%$$

There is insufficient evidence to suggest that the waiting times have reduced therefore accept H_0 and conclude that $p = 0.3$

CRITICAL VALUES AND REGIONS

For the above example

Binomial (20, 0.3) 5% Significance Level

$$P(X \leq 0) = 0.000798 \quad (0.01\%)$$

$$P(X \leq 1) = 0.00764 \quad (0.08\%)$$

$$\underline{P(X \leq 2) = 0.0355 \quad (3.55\%) \quad < 5\%}$$

$$P(X \leq 3) = 0.107 \quad (10.7\%) \quad > 5\%$$

Critical Values : 0, 1, and 2

Critical Region: $X \leq 2$

A sweet manufacturer packs sweets with 70% fruit and the rest mint flavoured. They want to test if there has been a change in the ratio of fruit to mint flavours at the 10% significance level. To do this they take a sample of 20 sweets. What are the critical regions?

X = number of fruit sweets Binomial (20, 0.7)

$$H_0 : p = 0.7$$

$$H_1 : p \neq 0.7$$

10% Significance level (**2 tailed – 5% at each tail**)

Lower tail	$\underline{P(X \leq 10) = 0.0480 \quad 4.8\%}$	Critical Region $X \leq 10$	(Critical Value = 10)
	$P(X \leq 11) = 0.113 \quad 11.3\%$		

Upper tail	$P(X \geq 17) = 0.107 \quad 10.7\%$	Critical Region $X \geq 18$	(Critical value = 18)
	$\underline{P(X \geq 18) = 0.035 \quad 3.5\%}$		

Critical Regions Critical Region $X \leq 10$ or $X \geq 18$