AS MATHS - STATISTICS REVISION NOTES

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 Categorial/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 = $\frac{size\ of\ whole\ sample}{size\ of\ whole\ population} \times population\ of\ the\ subgroup$

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

Categorial/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

2 $\boxed{4}$ 5 $\boxed{7}$ 8 $\boxed{9}$ 9

Median

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

2 4 5 5 7 8 9 10

LQ Median UQ

4.5 6 8.5

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 horizonal axis
- A numerical measure of correlation can be calculated (Spearman's Rank, Product Moment correlation coefficient) $-1 < r < 1$

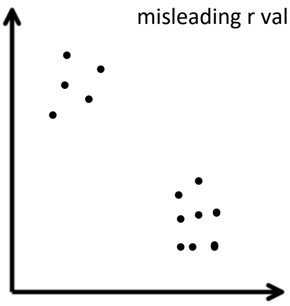
  -1 perfect negative correlation

  0 no correlation

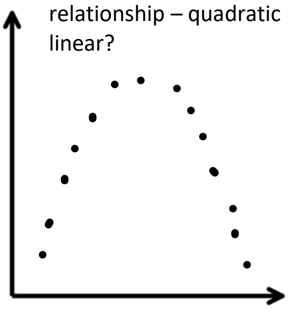
  1 perfect positive correlation.

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

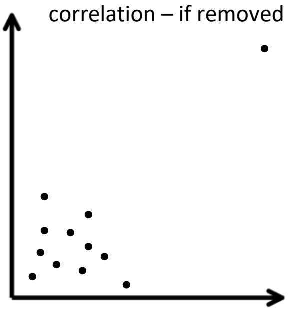
2 distinct groups misleading r value



R close to zero – but there is a relationship – quadratic not linear?



Outlier distorting r value suggesting positive correlation – if removed no correlation



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

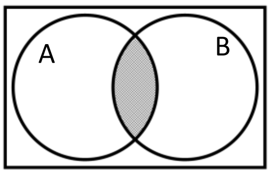
Remove values which are **2** $\times$ **Standard Deviation** above or below the mean.

5 PROBABILITY

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

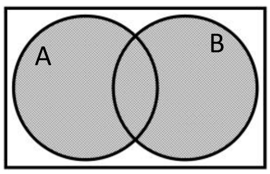
<u>**Notation**</u>

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



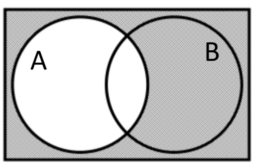
For independent events
P($A \cap B$) = P(A)$\times$P(B)

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



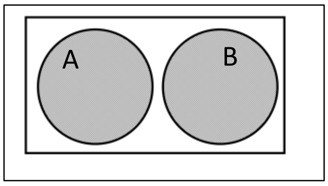
P($A \cup B$) = P(A) + P(B) - P($A \cap B$)

A' A does **not** happen



P(A')= 1 – P(A)

Mutually Exclusive events – two or more events which cannot happen at the same time



$P(A \cap B)=0$

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

	Male	Female	TOTAL
Junior	15	20	35
Senior	32	33	65
TOTAL	47	53	100

Find the probability of

a) picking a female = 0.53

b) pickling a junior male = 0.15

c) not picking a junior male = 1 – 0.15 = 0.85

d) picking a junior and a senior when 2 members are selected at random $\frac{47}{100} \times \frac{53}{99} \times 2 = 0.503$

On his way to work Josh goes through 2 sets of traffic lights. The probability that he has to stop at the 1st set is 0.7 and the probability for the 2nd set is 0.6 (assume independence)

Find the probability that he has to stop at only one of the traffic lights.

Stop and Not Stop or Not Stop and Stop

$0.7 \times 0.4 + 0.3 \times 0.6$

$= 0.46$

6 PROBABILITY DISTRIBUTIONS

A probability distribution shows the probabilities of the possible outcomes $\sum P(X = x) = 1$

It can be used to calculate the EXPECTATION (mean) of the distribution E(X) = $\sum x \times P(X = x)$

x	0	1	2
P(X = x)	0.5	3y	2y

Calculate the value of y $\sum P(X = x) = 1$

0.5 + 3y + 2y = 1 5y = 0.5 y = 0.1

Calculate E(X)

$0 \times 0.5 + 1 \times 0.3 + 2 \times 0.2 = 0.7$

7 BINOMIAL DISTRIBUTION B(n,p)

- 2 possible outcomes — probability of success = p; Probability of failure = (1 - p)
- fixed number of trials n
- The trials are independent
- E(x) = np

P(getting r successes out of n trials) = $_nC_r \times p^r \times (1-p)^{n-r}$

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

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

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

8 HYPOTHESIS TESTING – for binomial

- Set up the hypothesis

	$H_1 : p < a$	one sided test
$H_o : p = a$	$H_1 : p = a$	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_o
- Write a conclusion (in context)

Reject H_o
"There is sufficient evidence to suggest that ………is underestimation/overestimating……."

Accept H_o
"There is insufficient evidence to suggest that ……increase/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?
$H_o : p = 0.3$
$H_1 : p < 0.3$
5% Significance level

X = number of patients waiting more than 20 minutes
X Binomial (20, 0.3)
Using tables $P(X \le 3) = 0.107$ (10.7%)

10.7% > 5%
There is insufficient evidence to suggest that the waiting times have reduced therefore accept Ho and conclude that p = 0.3

CRITICAL VALUES AND REGIONS

For the above example
Binomial (20, 0.3) 5% Significance Level

$P(X \le 0) = 0.000798$ (0.01%)
$P(X \le 1) = 0.00764$ (0.08%)
$P(X \le 2) = 0.0355$ (3.55%) < 5%
$P(X \le 3) = 0.107$ (10.7%) > 5 %

Critical Values : 0, 1, and 2
Critical Region: $X \le 2$

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 (40, 0.7)
$H_o : p = 0.7$
$H_1 : p \ne 0.7$
10% Significance level (**2 tailed – 5% at each tail**)

Lower tail $P(X \le 10) = 0.0480$ 4.8 % Critical Region $X \le 10$ (Critical Value = 10)
$P(X \le 11) = 0.113$ 11.3%

Upper tail $P(X \ge 17) = 0.107$ 10.7%
$P(X \ge 18) = 0.035$ 3.5% Critical Region $X \ge 18$ (Critical value = 18)

Critical Regions Critical Region $X \le 10$ or $X \ge 18$