A LEVEL 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 (sample)

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 $S_{xx} = \sum fx^2 - n\bar{x}^2$

s² = $\frac{S_{xx}}{n-1}$

STANDARD DEVIATION (population)

Standard deviation $\sigma = \sqrt{\frac{S_{xx}}{n}}$ **Variance** = $\sigma^2 = \frac{S_{xx}}{n}$

Check with your syllabus/exam board to see if you are expected to divide by n or n-1 when calculating the standard deviation

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) $-1 \le r \le 1$

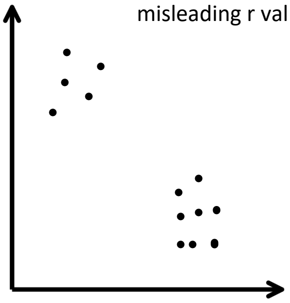
  -1 perfect negative correlation

  0 no correlation

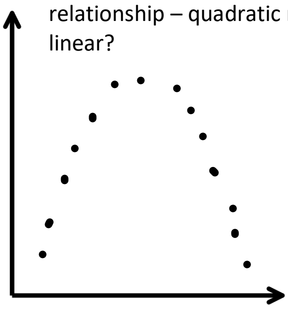
  1 perfect positive correlation.

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

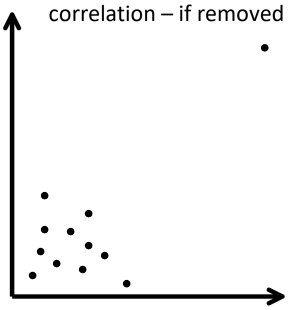
2 distinct groups misleading r value



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 Anomalies'

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

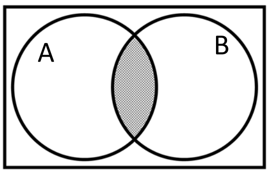
Remove values which are **2 ×** **Standard Deviation** above or below the mean.

5 PROBABILITY

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

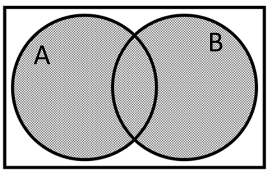
<u>**Notation**</u>

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



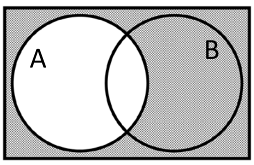
For independent events
P($A \cap B$) = P(A)×P(B)

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



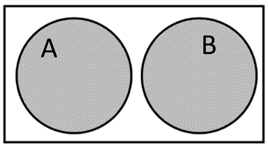
P($A \cup B$) = P(A) + P(B) - P($A \cap B$)

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 $\frac{35}{100} \times \frac{65}{99} \times 2 = 0.460$

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 × 0.4 + 0.3 × 0.6

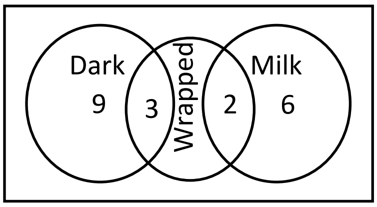
= 0.46

Conditional Probability

When the outcome of the first event effects the outcome of a second event the probability of the second event happening is conditional on the probability of the first event happening

- P(B|A) means that the probability of B given that A has occurred
- P(B|A) = $\frac{P(A \cap B)}{P(A)}$ so $P(A \cap B) = P(A)$P(B/A)
- If the probabilities needed are not stated clearly a tree diagram or venn diagram may help

In a box of dark and milk chocolates there are 20 chocolates. 12 of the chocolates are dark and 3 of these dark chocolates are wrapped. There are 5 wrapped chocolates in the box. Given that a chocolate chosen is a milk chocolate, what is the probability that it is not wrapped.



P(Not Wrapped/Milk)

$= \frac{P(Not\ wrapped \cap Milk)}{P(Milk)} = \frac{6}{20} \div \frac{8}{20} = \frac{3}{4}$

6 PROBABILITY DISTRIBUTIONS

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

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 × 0.5 + 1 × 0.3 + 2 × 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) = $_n\mathrm{C}_r \times p^r \times (1-p)^{n-r}$

> 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   1 – p = 0.9   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)

USING CUMULATIVE TABLES

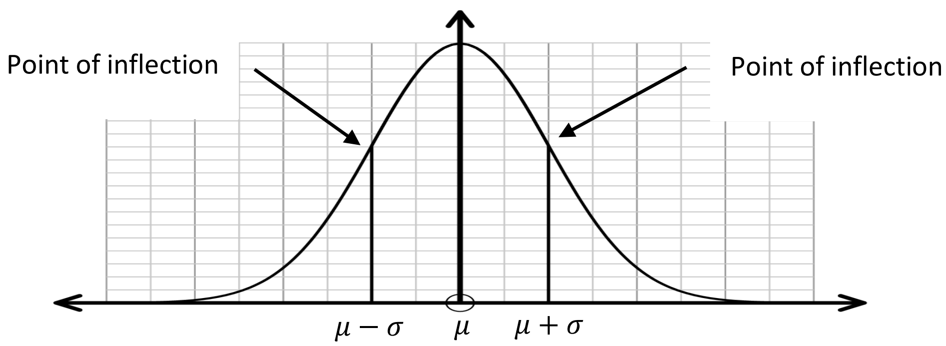
- Check if you can use your calculator for this
- Remember the tables give you less than or equal to the lookup value
- List the possible outcomes and identify the ones you need to include

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 ≥ 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 THE NORMAL DISTRIBUTION

- Defined as X~N(μ, σ^2) where μ is the mean of the population and σ^2 is the variance
- Symmetrical distribution about the mean such at
  - two-thirds of the data is within 1 standard deviation of the mean
  - 95% of the data is within 2 standard deviations of the mean
  - 99.7% of the data is within 3 standard deviations of the mean
  - points of inflection of the Normal curve lie one standard deviation either side of the mean



- X ~ N(μ, σ^2) can be transformed to the standard normal distribution Z ~N(0,1) using

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

Calculating probabilities

Probabilities can be calculated by either using the function on a calculator or by transforming the distribution to the standard normal distribution

A sketch graph shading the required region is a good idea.

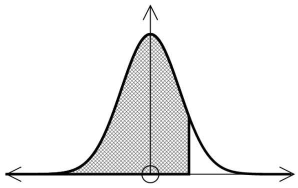
IQs are normally distributed with mean 100 and standard deviation 15. What percent of the population have an IQ of less than 120?

X ~ N($(100, 15^2)$)

P(X <120)   P(z < $\frac{120-100}{15}$)

P(z < 1.333) = 0.909

90.9 % of the population have an IQ less than 120

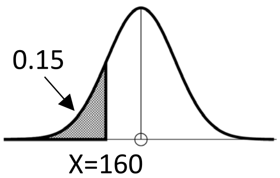


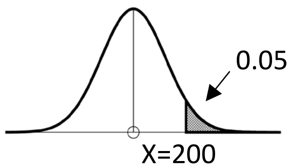
Calculating the mean, standard deviation or missing value (Using Inverse Normal)

If the probability is given then you need to work backwards to find the missing value(s)

The time, X minutes to install an alarm system may also be assumed to be a normal random variable such that P(X<160) = 0.15 and P(X>200) = 0.05

Determine to the nearest minute, the values for the mean and standard deviation of X





Use the tables or the calculator function to find the z values corresponding to the probabilities given

P(z < -1.0364) = 0.15

P(z > 1.6449) = 0.05

$\frac{160-\mu}{\sigma} = -1.0364 \qquad 160 - \mu = -1.0364\sigma$

$\frac{200-\mu}{\sigma} = 1.6449 \qquad 200 - \mu = 1.6449\sigma$

Solving simultaneously gives $\mu = 175\ minutes \quad \sigma = 15\ minutes$

Using the normal distribution to approximate a binomial distribution

For a valid result the following conditions are suggested

X ~ B(n,p) np > 5 and n(1-p) > 5   (ie p is close to ½ or n is large)

If the conditions are true then

X~B(n,p) can be approximated using X ~ N(np, np(1-p))

(NB As the binomial distribution is discrete and the Normal distribution is continuous some exam boards specify that a continuity correction is used. If you are calculating P(X < 80) you use P(X < 79.5) in your normal distribution calculation)

A dice is rolled 180 times. The random variable X is the number of times three is scored. Use the normal distribution to calculate P(X < 27)

X ~ B(180, $\frac{1}{6}$) can be approximated by X ~ N(30, 25)

Without continuity correction	With continuity correction
P(X < 27) = 0.274 (3 s.f.)	P(X < 26.5) = 0.242 (3 s.f.)

9 SAMPLING

If you are working with the mean of a sample of several observations from a population (eg calculating the probability that the mean ($\bar{x}$) is less than a specified value) then the following distribution must be used

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ where n is the sample size, μ is the population mean and σ^2 is the population variance

> Alex spends X minutes each day looking at social media websites. X is a random variable which can be modelled by a normal distribution with mean 70 minutes and standard deviation 15 minutes. Calculate the probability that on 5 randomly selected days the mean time Alex spends on social media is greater than 85 minutes.
>
> n = 5 $\bar{X} \sim N(70, \frac{15^2}{5})$ P($\bar{X} > 85$) = 0.0127 (3 s.f.)

10 HYPOTHESIS TESTING

Binomial

Set up the hypothesis

	$H_1 : p < a$	one sided test
$H_o : p = a$	$H_1 : p \neq a$	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 we cannot reject the null hypothesis 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 ≤ 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 (20, 0.7)

$H_o : p = 0.7$

$H_1 : p \neq 0.7$

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 $X \ge 18$

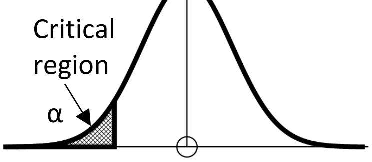
Normal Distribution: testing for changes in the mean

1. Set up the hypothesis

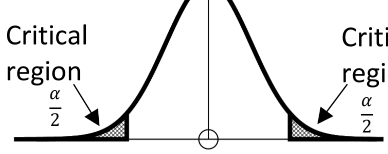
$$H_o : \mu = \mu_0$$

$H_1 : \mu < \mu_0$ one sided test mean has decreased
$H_1 : \mu \neq \mu_0$ two sided test $H_1 : \mu \neq \mu_0$ two sided test
$H_1 : \mu > \mu_0$ one sided test mean has increased

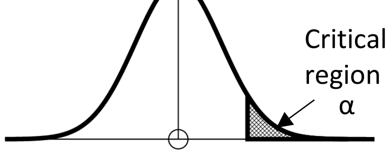
$H_1 : \mu < \mu_0$ one sided test
mean has decreased



$H_1 : \mu \neq \mu_0$ two sided test
mean has changed



$H_1 : \mu > \mu_0$ one sided test
mean has increased



2. Investigate the value you are working with by either
Method 1: See if your observed value lies in the critical region – reject H_0 if it does
or
Method 2: Calculate the probability (p value) of getting the observed value (or greater if testing for increase) if H_0 is true and reject H_0 if the probability is less than the significance level

3. Write a conclusion DO NOT just state 'Accept/Reject H_0'
Accept H_o

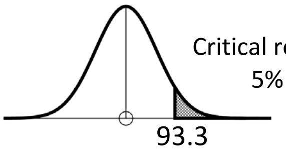
"There is insufficient evidence to suggest that the mean of …… therefore we cannot reject the null hypothesis that $\mu = \mu_0$.

Reject H$_o$
"There is sufficient evidence to suggest that the mean has changed and based on the results conclude that the mean of……has increased/decreased/does not equal μ_0"

The test results of a large group of students are thought to follow a normal distribution with mean 90 points and variance 80 points. A random sample of 20 students is found to have a mean of 94 points. Test at the 5% significance level to investigate the claim that the mean has increased.

H$_o$: $\mu = 90$ H$_1$: $\mu > 90$ $\bar{X} \sim N(90, \frac{80}{20})$

METHOD 1



(93.3 from calculator)

Using tables:

z = 1.6449 (for 5% significance)

1.6449 = $\frac{x-90}{\sqrt{\frac{80}{20}}}$ rearrange to give x = 93.3

As 94 >93.3 the observed value is in the critical Region indicating that

METHOD 2

P(x >94) $\frac{94-90}{\sqrt{\frac{80}{20}}}$ = 2

p = P(z > 2) = 0.02275

Significance level 5% = 0.05

As 0.02275 < 0.05

there is sufficient evidence to suggest that the mean has increased indicating an improved performance in the test

CORRELATION COEFFICIENT: testing to investigate whether the linear relationship represented by r (calculated from the sample) is strong enough to use the model the relationship in the population

r = correlation coefficient calculated using sample size n
ρ = unknown population correlation coefficient

The test checks whether ρ is 'close to 0' or 'significantly different from 0'

H$_o$: $\rho = 0$ there is no correlation between the 2 variables

H$_1$: $\rho \neq 0$ the two variables are correlated (2 tailed test)
H$_1$: $\rho > 0$ the two variables are positively correlated (one tailed test)
H$_1$: $\rho < 0$ the two variables are negatively correlated (one tailed test)

The length of service and current salary is recorded for 30 employees in a large company. The product-moment correlation coefficient r, of the 30 employees is 0.35. Test the hypothesis that there is no correlation between an employees length of service and current salary at the 5% significance level.

H$_o$: $\rho = 0$ H$_1$: $\rho \neq 0$ (2 tailed test) n = 30

To be significant at the 5% level the probability of r being in the critical regions must be < 0.025

Critical value from tables = 0.3610 leading to a critical region r < -0.361 and r > 0.361

r = 0.35 is not in the critical region so there is insufficient evidence to show that correlation is significantly different from zero