**Statistics Revision Sheet**

**Averages and Spread**

**Samples**

**Mean** = raw data

= grouped or tabled data *(where x is the midpoint)*

**Variance** s² = raw data

s² = grouped or tabled data *(where x is the midpoint)*

**Standard Deviation** s =

**Populations**

**Mean** μ = raw data

μ = grouped data *(where x is the midpoint)*

**Variance** σ² = raw data

σ² = grouped data *(where x is the midpoint)*

**Standard Deviation** σ =

*Remember the alternative formula s² = is also in your formula book on page 12*

**Before you begin a question in which you are asked to find mean, standard deviation or variance, always make a note of the following values first:**

*, , n these are known as the* ***summary statistics***

**Linear scaling**

If data values have been increased or decreased by a constant amount, the same will happen to the mean. The standard deviation and variance will be unaffected.

If data values have been multiplied by a constant value, the same will happen to the mean and the standard deviation.

**Probability**

**Mutually Exclusive Events** - events that cannot happen at the same time e.g. passing and failing an exam

**Exhaustive Events** - events where all possible outcomes are included e.g. throwing a head or a tail on a fair coin

**Independent Events** - one event has no effect on another event occurring e.g. throwing a 1 or a 2 on a fair dice

**Formulae to learn**:

P(**not** A) = P( **(complementary events)**

P(A **or** B) = P(A = P(A) + P(B) – P(A **if not mutually exclusive**

P(A **or** B) = P(A ) = P(A) + P(B) **if mutually exclusive**

*If P(A or B) = 1 -* ***the events A and B are exhaustive***

P(A **and** B) = P(A ) = P(A) x P(B) **if independent events**

**Conditional Probability**

P(A **given** B) = P(A|B) = P(B **given** A) = P(B|A) =

P(A **and** B) = P(A ) = P(A) x P(B|A)

P(B **and** A) = P(B = P(B) x P(A|B)

**If events are independent:**

P(A|B) = P(A|) = P(A)

P(B|A) = P(B|) = P(B)

**Venn Diagrams**

Always start from the middle and work out

**The Binomial Distribution**

X ~ B(n, p) where n is the number of trials, and p is the probability of success

= =

Example

X ~ B(15, 0.42) Find P(X=4)

P(X = 4) = x x

= 0.106

**Using the tables (pg 15 – 20)**

X B(40, 0.30)

Example

a) Find P(X

There are too many combinations to be able to use method. Therefore, the binomial tables should be used.

P(X0.0553

b) Find P(8 < X < 13) - we want to include 9, 10, 11 & 12

= P(X 12) – P(X)

= 0.5772 – 0.1110 = 0.4662

**Finding the Mean and Variance**

Mean = np

Variance = np (1 – p) *these formulae are in your formula book in the table on pg 11*

Standard deviation =

Example

Find the mean and standard deviation of X ~ B(50, 0.30)

n = 50 and p = 0.30

mean = np = 50 x 0.30 = 15

variance = np (1 – p) = 15 (0.7) = 10.5 standard deviation = = 3.24 (3 s.f.)

**The Normal Distribution**

X N( where is the mean and is the variance

**Finding Probabilities**

*Always use a sketch to help highlight the shaded area you are finding.*

Example

X N(4, 0.5²)

a) Find P(X

P(X 3.5) = P(Z ) standardise first using Z =

= P(Z 1.0) = (1.0) look up z value in tables on pg 24

= 0.84134

b) P(3.25 < X < 5.0) = P(3.25 X 5.0) **Remember P(X = a) = 0**

P( Z ) = P(-1.5

= (-1.5)

= – (1 - )

= 0.97725 – (1 – 0.93319)

= 0.91044

**Finding z values – Using the table of percentage points**

You will find these tables on pg 25 of the formula book.

Example

X N(10, 0.5²)

Find z such that P(X z) = 0.96 - look up 0.96

(z) = 0.96 z = 1.7507

**Finding mean (µ) and variance (σ²)**

* To find one of these, you will need to form an equation as follows:

Y N(, 4²) P (Y 23) = 0.75 find .

P (Z ) = 0.75

= 0.6745

= 20.302

* To find both of them, you will need to form a pair of simultaneous equations:

X N(

P ( X 1.83) = 0.3 P ( X 2.31) = 0.7

P ( Z ) = 0.3 P ( Z ) = 0.7

This will be a negative ‘z’ value This will be a positive value

= -0.5244 = 0.5244

1.83 - = -0.5244 **(1)**  2.31 - = 0.5244 **(2)**

Eliminateeither or = 2.07m = 0.458

**Estimation**

This uses samples to draw conclusions about the population.

population parameters = population characteristics e.g. mean (µ) and variance (

statistics = any number calculated from, or summarising, the data in a sample e.g. mode, median, range and standard deviation

If X N(, ²) then N(, ) where n is the sample size

To work out probabilities – standardise as with the normal distribution as follows

Z = is known as the *standard error*

X N(1, 0.1 ²) n=25 N(, ) = N(, 0.0004)

P ( > 1.04) = P (Z >) = P ( Z < 2)

= 1 – (2) = 0.023

**Central Limit Theorem**

**Used if the original population is not normally distributed or is unknown**. Assumes the sampling distribution is normally distributed provided the value of ‘n’ is large enough i.e. greater than about 25-30.

**Confidence Intervals (CI)**

The z values for the given confidence intervals will either have to be learnt or can be worked out using the % point table on pg 25

The formula is:

**( – z x , + z x )**

where is the sample mean and will be given or can be found,

n is the sample size and will be given

is the population variance will be given or can be found

(*NB sometimes ‘s² ’ will have to be used instead of which can be worked out using the formulae in the first section of this revision sheet*)

You may be asked to reject or accept a given value for µ. If the value lies outside the range of the CI there are grounds to **reject** the claim.

**Correlation and Regression**

**Pearson’s Product Moment Correlation Coefficient (PMCC) ‘r’**

r = = = ∑x² - = ∑y² -

These formulae are all on pg 13 of your formula book

*NB you should get used to using your calculator in order to find the value for ‘r’ as it will save you a lot of time.*

Interpreting the value of r:

**-1**

where -1 is perfect negative linear correlation, 0 is no linear correlation and 1 is perfect positive correlation

between ±0.2 and 0 weak (linear) correlation

between ±2 and ±0.7 moderate (linear) correlation

between ±0.7 and ±0.9 strong (linear) correlation

You should plot the scatter graph to help describe the correlation along with the value for ‘r’ as this will give a clearer picture of any linear relationship between the two variables. Remember to make reference to the variables in your description e.g. there is some positive (linear) correlation **between** the height and the weight of a person.

*The value of ‘r’ is unaffected by linear scaling*

**Explanatory variable** = independent variable, which changes independently of the other variable (usually denoted by ‘x’)

**Response variable** = dependent variable, which changes as the value of the other variable changes (usually denoted by ‘y’)

**Calculation of least squares regression lines**

This should always go through the point ( ,

= =

The equation of a least square regression line is given by **y = a + bx**

**b =**  = = ∑x² - these are on pg 13 of the formula book

**a = - b**

*your calculator can work these out for you*

Before plotting the line, you must work out at least two coordinates that lie on the line by substituting in values for x.

**Prediction** estimates a future value of Y by substituting in a value of x within the given range

**Extrapolation** using the equation to predict value of Y by substituting a value of x *not* within the given range

**Residuals** an observed y-value subtract the value given by the regression line.

Large residuals indicates points not well ‘explained’ by the regression line:

**Outliers** a data point with a relatively large residual

If you identify an outlier, you should try to explain it, e.g. it could be an error that you are told to correct

**Influential data points** have an x-value much greater or less than the other x-values.

Data points can be both influential and an outlier but don’t have to be.

*When asked to comment on the least squares regression line, always make reference to the relationship between the two variables and the residuals (e.g. any outliers or influential points)*