**Statistics Revision Sheet**

**Averages and Spread**

**Samples**

**Mean** $\overbar{x}$ = $\frac{\sum\_{}^{}x}{n}$ raw data

$\overbar{ x}$ = $\frac{\sum\_{}^{}fx}{\sum\_{}^{}f}$ grouped or tabled data *(where x is the midpoint)*

**Variance** s² = $\frac{1}{n-1}\left\{\sum\_{}^{}x²- \frac{(\sum\_{}^{}x)²}{n}\right\}$ raw data

s² = $\frac{1}{n-1}\left\{\sum\_{}^{}fx²- \frac{(\sum\_{}^{}fx)²}{\sum\_{}^{}f}\right\}$ grouped or tabled data *(where x is the midpoint)*

**Standard Deviation** s = $\sqrt{variance}$

**Populations**

**Mean** μ = $\frac{\sum\_{}^{}x}{n}$ raw data

μ = $\frac{\sum\_{}^{}fx}{\sum\_{}^{}f}$ grouped data *(where x is the midpoint)*

**Variance** σ² = $\frac{1}{n}\left\{\sum\_{}^{}x²\right\}- μ²$ raw data

 σ² = $\frac{1}{n}\left\{\sum\_{}^{}fx²- \frac{(\sum\_{}^{}fx)²}{\sum\_{}^{}f}\right\}$ grouped data *(where x is the midpoint)*

**Standard Deviation** σ = $\sqrt{variance}$

*Remember the alternative formula s² =* $\frac{\sum\_{}^{}(x- \overbar{x})²}{n-1}$ *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:**

$\sum\_{}^{}x$*,* $\sum\_{}^{}x²$*, 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($A^{'})=1-P(A)$ **(complementary events)**

P(A **or** B) = P(A $\bigcup\_{}^{}B)$ = P(A) + P(B) – P(A $\bigcap\_{}^{}B)$ **if not mutually exclusive**

P(A **or** B) = P(A $\bigcup\_{}^{}B$) = 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 $\bigcap\_{}^{}B$) = P(A) x P(B) **if independent events**

**Conditional Probability**

P(A **given** B) = P(A|B) = $\frac{P(A\bigcap\_{}^{}B)}{P(B)}$ P(B **given** A) = P(B|A) = $\frac{P(B\bigcap\_{}^{}A)}{P(A)}$

P(A **and** B) = P(A $\bigcap\_{}^{}B$) = P(A) x P(B|A)

P(B **and** A) = P(B $\bigcap\_{}^{}A)$ = P(B) x P(A|B)

**If events are independent:**

P(A|B) = P(A|$B^{'}$) = P(A)

P(B|A) = P(B|$A^{'}$) = 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

$\left(\frac{n}{x}\right)$ = $\frac{n!}{x!\left(n-x\right)!}$ = $^{n}∁\_{r}$

Example

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

P(X = 4) = $^{15}∁\_{4}$ x $0.42^{4}$ x $0.58^{11}$

 = 0.106

**Using the tables (pg 15 – 20)**

X $\~$ B(40, 0.30)

Example

a) Find P(X$ \leq 7)$

There are too many combinations to be able to use $^{n}∁\_{r}$ method. Therefore, the binomial tables should be used.

P(X$ \leq 7)= $0.0553

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

 = P(X$ \leq $ 12) – P(X$ \leq 8$)

= 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 = $\sqrt{variance}$

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 = $\sqrt{10.5}$ = 3.24 (3 s.f.)

**The Normal Distribution**

X $\~$ N( $μ, σ²)$ where $μ $is the mean and $σ²$ 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 $\leq 4.5)$

P(X $\leq $3.5) = P(Z $\leq $ $\frac{4.5-4}{0.5}$) standardise first using Z = $\frac{X- μ}{σ}$

 = P(Z $\leq $ 1.0) = $Φ$ (1.0) look up z value in tables on pg 24

 = 0.84134

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

 P($\frac{3.25-4}{0.5}$ $\leq $ Z $\leq $ $\frac{5.0-4.0}{0.5}$) = P(-1.5 $\leq Z \leq 2.0)$

 = $C \left(2.0\right)- Φ$ (-1.5)

 = $Φ \left(2.0\right)$ – (1 - $Φ (1.5)$)

 = 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

 $Φ$ (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($μ$, 4²) P (Y $\leq $ 23) = 0.75 find $μ$.

P (Z $\leq $ $\frac{23 - μ}{4}$ ) = 0.75

 $\frac{23 - μ}{4}$ = 0.6745

 $μ $ = 20.302

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

X $\~$ N( $μ, σ²)$

 P ( X $\leq $ 1.83) = 0.3 P ( X $\leq $ 2.31) = 0.7

 P ( Z $\leq $ $\frac{1.83 - μ}{σ}$) = 0.3 P ( Z $\leq $ $\frac{2.31 - μ}{σ}$) = 0.7

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

 $\frac{1.83 - μ}{σ}$ = -0.5244 $\frac{2.31 - μ}{σ}$ = 0.5244

 1.83 - $μ$ = -0.5244$σ$ **(1)**  2.31 - $μ$ = 0.5244$σ$ **(2)**

 Eliminateeither $μ$ or $σ$ $μ$ = 2.07m $σ$ = 0.458

**Estimation**

This uses samples to draw conclusions about the population.

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

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

If X$\~$ N($μ$, $σ$ ²) then $\overbar{X}\~$ N($μ$, $\frac{σ ²}{n}$) where n is the sample size

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

Z = $\frac{\overbar{x}-μ }{\sqrt{\frac{σ ²}{n}}}$ $\sqrt{\frac{σ ²}{n}}$ is known as the *standard error*

X$\~$ N(1, 0.1 ²) n=25 $\overbar{X}\~$ N($1$, $\frac{0.1 ²}{25}$) = $\overbar{X}\~$ N($1$, 0.0004)

P ($\overbar{X}$ > 1.04) = P (Z >$\frac{1.04-1}{\sqrt{0.0004}}$) = P ( Z < 2)

 = 1 – $Φ$ (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:

**(**$\overbar{x}$ **– z x** $\frac{σ}{\sqrt{n}}$ **,** $\overbar{x}$ **+ z x** $\frac{σ}{\sqrt{n}}$ **)**

where $\overbar{x}$ is the sample mean and will be given or can be found,

 n is the sample size and will be given

$σ²$ is the population variance will be given or can be found

 (*NB sometimes ‘s² ’ will have to be used instead of* $σ²$ *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 = $\frac{s\_{xy}}{\sqrt{s\_{xxs\_{yy}}}}$ $s\_{xy}$ = $∑xy- \frac{∑x∑y}{n}$ $s\_{xx}$ = ∑x² - $\frac{(∑x)²}{n}$ $s\_{yy}$ = ∑y² - $\frac{(∑y)²}{n}$

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**$ \leq r \leq 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 ($\overbar{x}$ , $\overbar{y}) $

$\overbar{x}$ = $\frac{∑x}{n}$ $\overbar{y}$ = $\frac{∑y}{n}$

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

**b =** $\frac{s\_{xy}}{s\_{xx}}$ $s\_{xy}$ = $∑xy- \frac{∑x∑y}{n}$ $s\_{xx}$ = ∑x² - $\frac{(∑x)²}{n}$ these are on pg 13 of the formula book

**a =** $\overbar{y}$ **- b**$\overbar{x}$

*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)*