S1 Revision Notes

Definitions to learn

Statistical Experiment - A method for collecting data to test against a hypothesis/prediction

Event - A set of possible outcomes in a Statistical Experiment

Statistical Model

- Real world problem is observed
- A model is devised
- Predictions are made
- Data is collected (from a Statistical experiment)
- Comparisons are made to expected outcomes
- Model is refined changing parameters to improve model
- Model is used to make predicted outcomes about the real world problem

Advantages

Cheaper/Quicker - Easy to produce
Help understanding of a real world problem
Help make predictions about a real world problem

Disadvantages

Not be completely accurate as it can never completely replicate problem

Histograms

Use histograms when the data is continuous
(grouped data normally with different class widths)

Bars must go from the boundary points - no gaps

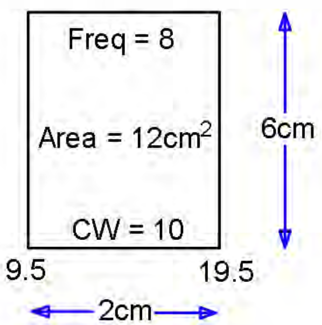
Label AXES!

Key feature - Area α Frequency   Area = k x Frequency

Frequency density = Frequency ÷ class width

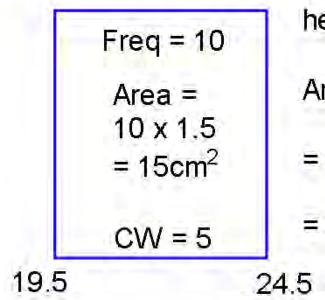
Class	Freq
10-19	8
20-24	10

The 10-19 bar has a width of 2cm and a height of 6cm.
What would the height and width of the 20-24 bar be.



width = CW ÷ 5

Area = Freq x 1.5   k = 1.5



height =

Area ÷ width

= 15 ÷ 1

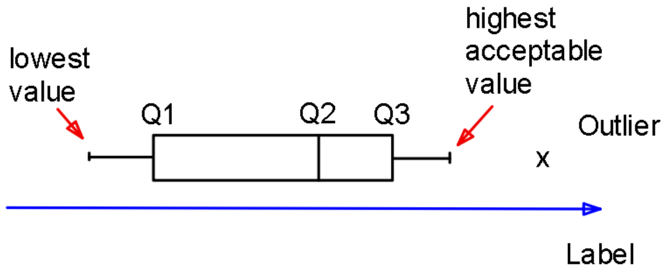
= 15cm

width = 5 ÷ 5 = 1cm

Quartiles

¼ n = 2.3 (ROUND UP) The lower quartile would be the 3rd peice of ordered data

¼ n = 4 The lower quartile would be half-way between the 4th and 5th peice of ordered data



25% of data is below Q1

25% of data is above Q3

Q2 is the median

IQR = Q3 - Q1

Outlier limits are normally given by

Lower Limit Q1 - 1.5 x IQR
Upper Limit Q3 + 1.5 x IQR

Show limits in working AND write down Outliers

Anything below lower limit or above upper limit is an outlier

Mean and Standard Deviation

Mean = x or $\mu = \frac{\sum fx}{n}$

If data is in a list ignore f

Variance = $\frac{\sum fx^2}{n} - \bar{x}^2$

If data is grouped x is the MIDDLE VALUE of the classes

$\sum fx^2$ = sum of each frequency multiplied by the square of its middle value

Standard deviation = σ = $\sqrt{Variance}$

Class	Freq
1 - 5	8
6 - 8	7
9 - 12	4
13 - 20	1

Find

a) Mean
b) Standard Deviation
c) Median

give answers to 2dp where appropriate

Mid-point	Freq
3	8
7	7
10.5	4
16.5	1

mid-point is x

Σfx = 8 x 3 + 7 x 7 + 4 x 10.5 + 1 x 16.5 = 131.5

$\frac{\Sigma fx}{n}$ = 131.5 ÷ 20 = 6.575 = 6.58 (2dp)

$\Sigma fx^2 = 8 \times 3^2 + 7 \times 7^2 + 4 \times 10.5^2 + 1 \times 16.5^2 = 1128.25$

$\frac{\Sigma fx^2}{n} - \bar{x}^2 = \frac{1128.25}{20} - 6.575^2 = 13.181875$

NEVER USED ROUNDED ANSWERS IN CALCULATIONS

$\sigma = \sqrt{13.181875} = 3.63$ (2dp)

To find the median of grouped (continuous) data you must INTERPOLATE

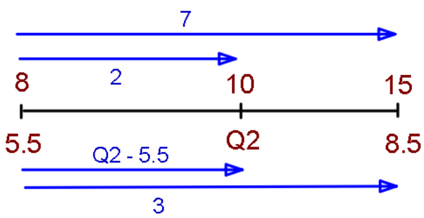
Class Boundaries	Freq	CF
		0
0.5 - 5.5	8	8
5.5 - 8.5	7	15
8.5 - 12.5	4	19
12.5 - 20.5	1	20 = n

Change to class boundaries and add cumulative frequency

½ n = 10 so the median is the 10th piece of data

The 10th piece of data falls between cumulative frequencies 8 and 15

Median is somewhere between 5.5 to 8.5



$\frac{Q2 - 5.5}{3} = \frac{2}{7}$

$Q2 = \frac{2}{7} \times 3 + 5.5 = 6.35714... = 6.36$ (2dp)

Skew

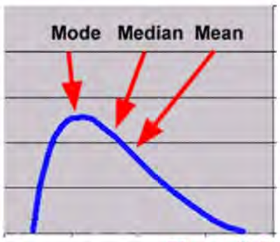
You can describe the shape of data, its skew, in a number of ways

Box Plot

POSITIVE SKEW

Q2 - Q1 < Q3 - Q2

if you are to the left
you are a happy hippy



The tail follows the positive x-axis

Mode < Median < Mean

NEGATIVE SKEW

Q2 - Q1 > Q3 - Q2

Mean < Median < Mode

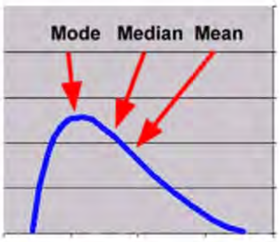
$$\text{Formula} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

if > 0 it is positive skew
if < 0 it is negative skew

The bigger (or smaller) the number the more skewed data is

Class	Freq
1 - 5	8
6 - 8	7
9 - 12	4
13 - 20	1

Describe the skew



Positive Skew

Median (6.36) < Mean (6.58)

(we do not include the mode in grouped data)

$$\text{Formula} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}} \qquad \frac{3\,(6.575 - 6.35714)}{3.63069} = 0.18$$

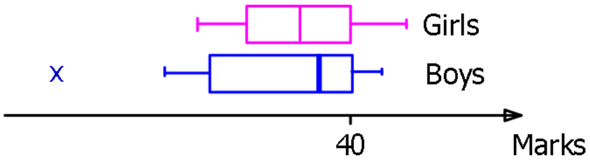
Slight Positive Skew

DO NOT USE ROUNDED
ANSWERS IN CALCULATIONS

Comparing Distributions

Make 3 statements

Location - compare specific values, median, quartiles
Dispersion - compare spread IQR
Shape - compare skew



Location -

Boys have a slightly higher average mark (boys median (38) > girls median (37))
75% of boys and 75% of girls scored 40 marks or less (boys Q3 = girls Q3)

Dispersion -

Boys marks are more spread out (boys IQR (8) > girls IQR (10))

Shape

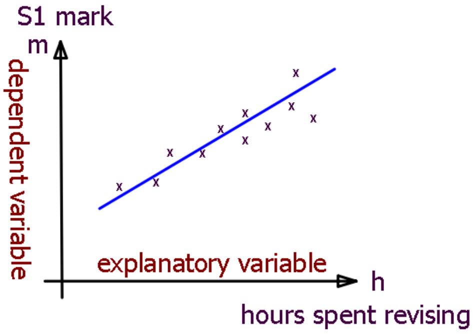
Boys marks are negatively skewed and contain an outlier
Girls marks are symmetrically skewed.

Median/IQR vs Mean / s.d.

Use median/IQR if the data is skewed as this will ignore extreme values

Use Mean/s.d. if the data is reasonably symmetrical (little skew)

Scatter Diagrams



x-axis - explanatory variable (choice variable) set independently of the other variable

y-axis - dependent variable, values are determined by the other variable

match the variables to x and y

S1 mark depends on the hours spent revising

$y = m$ and $x = h$

$y = a + bx$ becomes $m = a + bh$

$b = Sxy \div Sxx$ becomes $b = Shm \div Shh$

$a = \bar{y} - b\bar{x}$ becomes $a = \bar{m} - b\bar{h}$

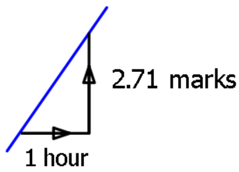
A regression line is suitable when correlation exists - the closer PMCC is to 1 (or -1) the more reliable the equation sho

For exmple if $m = 31.5 + 2.71h$ (should be given to 3sf unless stated otherwise)

a is the value of y when $x = 0$ this MUST be put into CONTEXT

$a = 31.5$ - A student would score 31.5 marks with no revision

b is the gradient, change in y divided by change in x



$b = 2.71$ - A student achieves an EXTRA 2.71 marks for each additional hour of revision.

Coding

Data has been coded using $p = x - 3$ $q = 10(y - 5)$

p and q are found to have a PMCC of 0.965 and a regression line with equation $q = 2 + 3p$

What is the PMCC of x and y and what would the equation of its regression line be

PMCC = 0.965 CODING DOES NOT AFFECT PMCC

$q = 2 + 3p$ SUB IN CODES AND MAKE Y THE SUBJECT

$10(y - 5) = 2 + 3(x - 3)$

$10y - 50 = 2 + 3x - 9$

$10y = 43 + 3x$

$y = 4.3 + 0.3x$

If $t = 10x + 3$ CODE is $\times 10$ $+ 3$

so DECODE a mean we -3 $\div 10$

$x = (t - 3) \div 10$

$\sigma x = \sigma y \div 10$ do not + or - when decoding a spread