STATISTICS GCSE 'HINTS'

Census	An official count/ survey of the whole population	**Advantage:** views of the entire population are represented. **Disadvantage:** expensive / time consuming/ difficult to do
Random	People are chosen at random using a random number generator.	**Advantage**: Every member of the population has equal chance **Disadvantage**: time consuming/ impractical
Systematic	People are chosen at regular intervals	**Advantage**: unlikely to be biased **Disadvantage**: not strictly random/ some members of the population won't be chosen
Stratified	When the population is made up of different groups of people. Our sample is chosen so it has the same proportion of each group	**Advantage**: best way to reflect population accurately **Disadvantage**: time consuming/ limit the variables for it to be practical
Interview V's Questionnaire	**Advantage**: explain questions/ responses are easier **Disadvantage**: interview may influence answers, causing bias	
Qualitative data	Gender/ types of transport/ colours	
Primary data	**Advantage**: Know how data was obtained/Reliability is known	
Secondary data	**Advantage**: quicker/ cheaper/ easier **Disadvantage**: unknown origin/ may be unreliable	
Quantitative data	Age/ numbers of something	
Pilot survey	**Advantage**: 1) Ensures relevant answers 2) That questions are understood 3) Allows for changes to questions 4) Check how long it takes 5) Identify ambiguity	
Continuous data	Weights / distance/ time	
Discrete data	Data that takes certain values: number of people in a class	
Nominal data	Categorised data e.g. male/ female	
Categorical data	Date that can be divided into groups. Race, gender, age group, educational level	
Numerical data	Weights	

Rank data	**Ranking** is used to recode the **data** into their **rank** ordering from smallest to largest or largest to smallest
Quota	Where there is a pre-determined number of customers (different ages and genders)
Population	Everyone
Comparing box plots	Comment on: **Median:** which one is higher, what does this mean in relation to question **IQR:** which has wider **Skew:** positive(median to the left)/ negative(median to the right)/ none(median in the middle)
Mean	Add them altogether and divide by how many there are.
Median	Put them in order, find the middle value
Mode	The most common
Range	Biggest - smallest
Index numbers	Index numbers are used to measure the changes in some quantity which we cannot observe directly. Base year has an index of 100 Index is a number not a %. Index number$= \frac{current\ price}{base\ year\ price} \times 100$ Chain Index number$=$ $\frac{current\ price}{previous\ year\ price} \times 100$

Part of the wage bills of a factory (in £1000s) for the years 1999 to 2003 are shown in the table.

Some index numbers are shown. The base year is the year 1999.

Year	1999	2000	2001	2002	2003
Wages (£1000s)	200	240		300	320
Index number	100	120	125	150	

(a) Work out the index number for 2003.

$$\frac{Current\ price}{Base\ year} \times 100 = \frac{320}{200} \times 100 = 160$$

(2)

(b) Work out the wage bill of the factory in 2001.

$$200 \times 1.25 = £250$$

(2)

The base year is changed to 2000.

Work out the new index number for 2002.

$$\frac{300}{240} \times 100 = 125$$

(2)

Binomial distribution

The tour company has just had 4 people who made provisional bookings.

(i) Calculate the probability that exactly 3 of these people will go on to confirm their booking.

You may use $(p+q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4$.

$p = p(\text{confirm}) = 0.8 \quad q = (\text{not}) = 0.2$

$4 \times (0.8^3) \times 0.2 = 0.41$

(ii) For these 4 provisional bookings, find which are the **two** most likely numbers of people who go on to confirm their bookings. Show your working.

$p^4 = 0.4096$ *

$4p^3q = 0.41$ *

$6p^2q^2 = 0.1536$

$4pq^3 = 0.0256$

$q^4 = 0.0016$

These two are biggest, representing 4 and 3 people.

3 and 4 confirmations

(5)

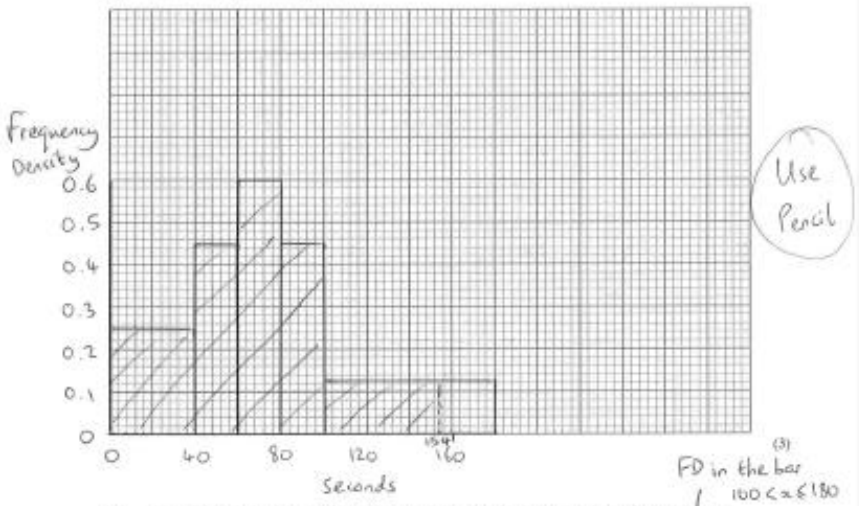
(Total 7 marks)

Normal distribution

95% of data should be within 2 S.D's of the mean.

Normal distribution curve is symmetrical

(b) Draw a histogram to represent the data in the table.



(3)

FD in the bar $100 < x \leq 180$

$154 - 100$

(c) (i) Shade the region in your histogram that is within two standard deviations of the mean.

(ii) Find the proportion of people represented by this region.

i) Within two standard deviations:

$74 \pm (40 \times 2) = -6 \leq x \leq 154$

so shade all histogram below 154.

ii) 40 people up to 100 seconds

$+ (0.125 \times 54) = 40 + 6.75$

$= 46.75$

(4)

$\frac{46.75 \text{ people}}{50 \text{ people}} = 0.935$

It is claimed that the time to complete the puzzle is normally distributed.

(d) Comment on the validity of this claim.

- 95% of data should be within 2 sd of mean in a normal distribution but the proportion is only 93.5%.
- A normal distribution is symmetrical.

(2)

(Total 14 marks)

Spearman's rank

Close to 1: positive correlation

Close to 0: no correlation

Close to -1: negative correlation

6. The table gives information about the age, and the minimum stopping distance at 40 kilometres per hour, for each of 10 cars.

Car	Age of car (months)	Stopping distance (metres)	Age rank	Distance rank	d	d^2
A	9	28.4	10	10	0	0
B	15	29.3	9	9	0	0
C	24	37.6	8	4	4	16
D	30	36.2	7	7	0	0
E	38	36.5	6	5	1	1
F	46	35.4	5	8	-3	9
G	53	36.3	4	6	-2	4
H	60	44.1	3	3	0	0
I	64	44.8	2	2	0	0
J	76	47.2	1	1	0	0

(Source: www.bized.co.uk)

Total $d^2 = 30$

(a) Work out Spearman's rank correlation coefficient for these data.

You may use the blank columns in the table to help with your calculations.

$$1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 30}{10(100-1)}$$

$$= 1 - \frac{180}{990}$$

$$= 0.\dot{8}\dot{1} = 0.818181\ldots$$

0.82

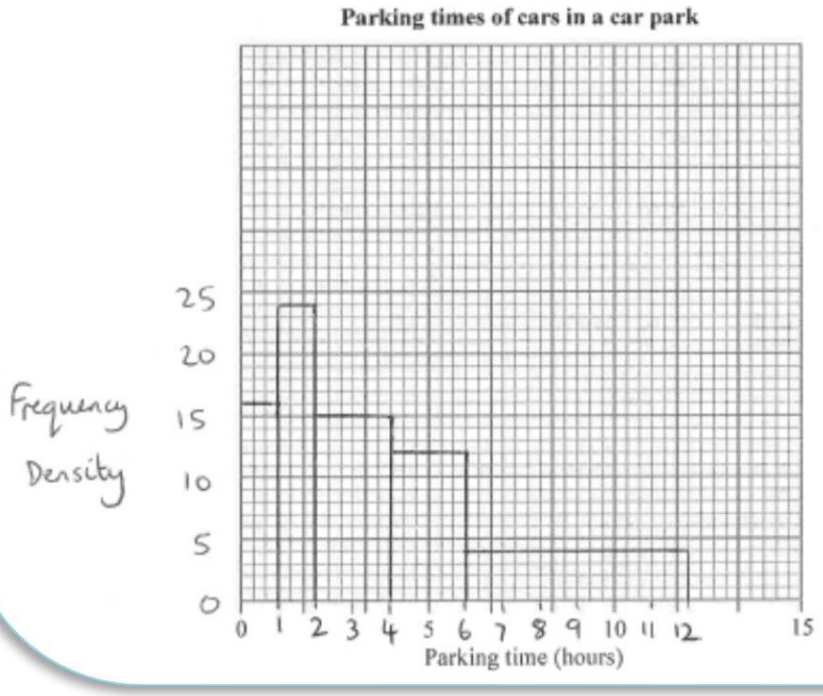
(4)

Histograms

The parking times in hours (p) for 118 cars in a car park are summarised in the table.

Hours (p)	Frequency (f)	Frequency Density
$0 < p \leq 1$	16	16
$1 < p \leq 2$	24	24
$2 < p \leq 4$	30	15
$4 < p \leq 6$	24	12
$6 < p \leq 12$	24	4
$p > 12$	0	0

(a) Draw a histogram for the data.



Mean from a frequency table

p	Frequency (f)	Midpoint (x)	fx
$0 < p \leq 1$	16	0.5	8
$1 < p \leq 2$	24	1.5	36
$2 < p \leq 4$	30	3	90
$4 < p \leq 6$	24	5	120
$6 < p \leq 12$	24	9	216
Totals	118		Total = 470

(b) Work out an estimate for the mean parking time of the cars.

You may use the space in the table.

"total"

$$\text{Mean} = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{470}{118}$$

$$= 3.98$$

3.98 ..... hours

Standard deviation

(c) Work out an estimate of the standard deviation of these parking times.

Give your answer to 1 decimal place.

You may use $\Sigma fx^2 = 2872$.

$$sd = \sqrt{\frac{\Sigma fx^2}{n} - \bar{x}^2}$$

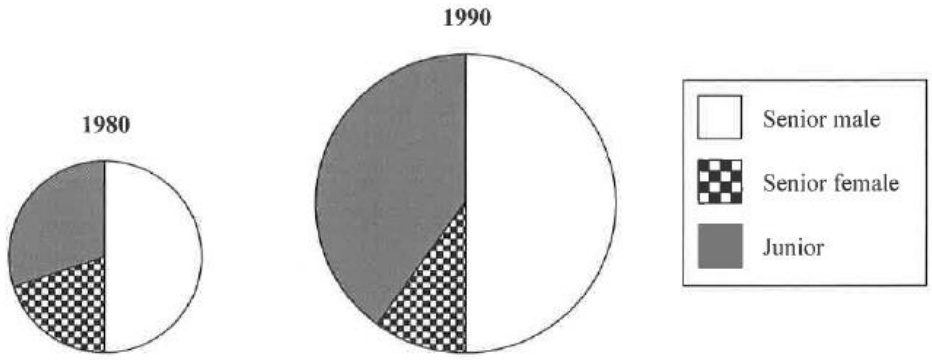
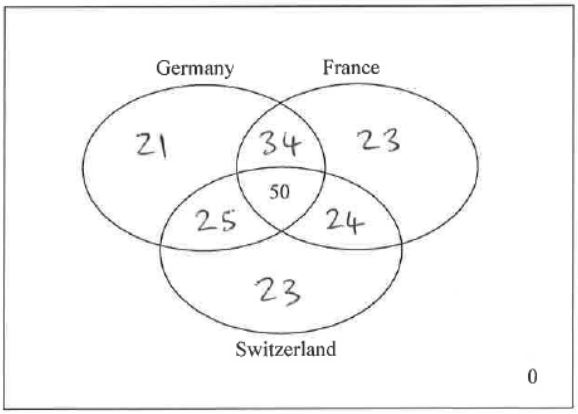
$$= \sqrt{\frac{2872}{118} - 3.98^2}$$

$$= 2.9$$

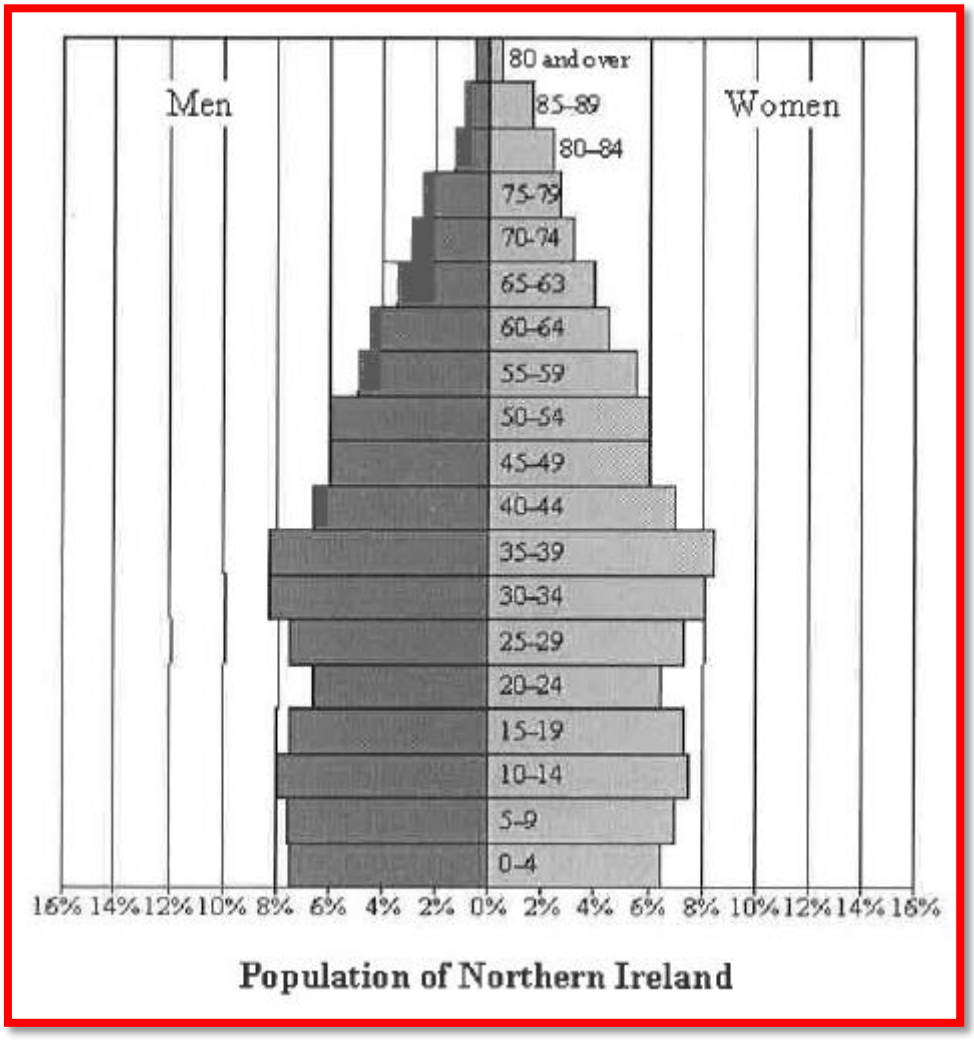
2.9 ..... hours

(3)

(Total 9 marks)

Cumulative frequency	○ Cumulative means to add. ○ Add the frequency's together line by line. ○ Cumulative frequency always goes on the y axis. ○ Use the upper bound to plot with.
Comparative pie charts	The comparative pie charts show some information about the players at Seaton squash club in 1980 and in 1990. The three types of players at Seaton squash club are Senior male, Senior female and Junior.  (Data source: *Seaton squash club*) (a) What has happened to the number of Senior male players at Seaton squash club between 1980 and 1990? Give a reason for your answer. · Increased. · This is because the area has got bigger (2)
Venn diagrams	A total of 200 people went on these tours. Of these 130 people went on a tour to Germany, → 130 - 34 - 50 - 25 = 21 131 people went on a tour to France, → 131 - 34 - 50 - 24 = 23 122 people went on a tour to Switzerland, → 122 - 25 - 50 - 24 = 23 74 people went on a tour to Switzerland and France, → 74 - 50 = 24 84 people went on a tour to France and Germany, → 84 - 50 = 34 75 people went on a tour to Germany and Switzerland, → 75 - 50 = 25 50 people went on a tour to all three countries. Last calculation. Work from the inside out Start here: First calculation. (a) Complete the Venn diagram for this information.  0

Population pyramid



Time series

Seasonal trends are plotted.
Moving averages: 4 point (add 4 numbers then divide by four), 2 point moving average (add two numbers, divide by two).
Trend lines are plotted using the moving average.

Year	1996				1997				1998			
Quarter	1	2	3	4	1	2	3	4	1	2	3	4
Sales	189	244	365	262	190	266	359	250	201	259	401	265

265
265.25
270.75
269.25

4 point Moving Average data

Quarters	1 - 4	2 - 5	3 - 6	4 - 7	5 - 8	6 - 9	7 - 10	8 - 11	9 - 12
Moving Average	265	265.25	270.75	269.25	266.25	269	267.25	277.75	281.5

Frequency Polygons

- A frequency polygon shows the trend of the data
- You plot the midpoint against the frequency

- The weight of 100 dogs at a dogs home are shown in the table below.

Weight	$0 < w \leq 5$	$5 < w \leq 10$	$10 < w \leq 15$	$15 < w \leq 20$	$20 < w \leq 25$	$25 < w \leq 30$
Frequency	4	13	25	32	17	9
Midpoints	2.5	7.5	12.5	17.5	22.5	27.5