Statistics

One of the major questions in Section B is almost certain to be dominated by statistics. Such a question will be of a practical nature, and may involve a lot of reading. It may also require a number of definition or opinion responses. For this reason, you should make sure you fully understand the concepts in statistics and are able to put into words precise answers to questions involving verbal answers.

Statistics may also be the basis of one of the 25 mark questions near the start of Section A. This question, if it occurs, is likely to be of a more mathematical nature.

Statistics can be divided into a number of sections: collecting reliable data, presenting this data in a graphical way, analysing the data to obtain summaries and using the data to draw inferences about the population. This last topic has been left to the next section.

For collecting data, you have to understand the different methods of sampling a population, as well as the importance of a random sample. You also need to be able to describe different types of study, including their advantages, disadvantages and potential problems, e.g. bias.

Data can be represented in a number of different forms: pie charts, bar charts, histograms, stemplots, scatterplots, etc. At higher level, it is unlikely that you will be asked to construct one of these from scratch. Instead, we may have to interpret a given graph.

Analysing data involves calculating summary statistics such as an average: mean, median, mode, or a measure of dispersion: IQR, standard deviation. It is important not to limit your study of these topics to the calculation of the statistic, you also need to know the advantages and disadvantages of each, as well as when they are suitable for use.

Finally, we come to the idea of investigating correlation, and the distinction between correlation and causality. You must be able to draw and interpret scatterplots, calculate the correlation coefficient by calculator and understand exactly what it represents. You must also be able to draw the line of best fit by eye, and find its equation.

1. Sampling and surveys

e.g. To gauge how its employees felt about proposed higher college fees, a university divided its employees into three categories: teaching staff, non-teaching staff and student employees. A random sample was selected from each group and they were telephoned and asked for their opinion.

(i) Describe the type of sampling being used by the university.

This is Stratified Sampling. Each member of the population is placed in one of the non-overlapping sub-groups or 'strata.' Random samples are then selected from each strata.

(iii) Mention any possible bias that might exist in the sampling plan.

If the samples taken from each strata are not in proportion to their presence in the population it would lead to the over or under representation of members of a particular strata.

(ii) Give a reason why the university might have chosen this type of sampling.

Done properly it fair and representitive of different categories of employee that might have particular different views on this issue.

2. Controlled experiments and observational studies

e.g. A suggestion has been made that eating fast-food on a regular basis increases the incidence of requiring an appendectomy (surgical removal of the appendix). You want to conduct a study to test this suggestion.

(i) What is the precise goal of the study?

To determine if there is a correlation between eating fast-food on a regular basis and appendectomies.

(ii) What is the target population?

The target population would be people who have had their appendix removed.

(iii) Explain why a controlled experiment is not appropriate in this case.

It would be ethically suspect to support the regular eating of fast-food for the purpose of the experiment aware that there may be unhealthy side-effects.

(iv) Describe what type of observational study you would conduct.

A questionnaire.

(v) How would you gather your data?

Randomly select a number of hospitals that perform appendectomies and survey all appendectomy patients over period in selected hospitals. (cluster sampling)

3. Graphs

e.g. Twelve sample employees were chosen from each of two large companies, A and B. On a given day, the time taken (in minutes) by each employee to arrive at their place of work, measured from the moment they left their accommodation, was recorded.

A: 10 2 17 25 12 18 14 12 10 6 32 11

B: 8 17 25 16 24 33 32 7 29 36 45 22

(i) Draw a back to back stem plot to represent this data.

A		B
6,2	0	7,8
8,7,4,2,2,1,0,0	1	6,7
5	2	2,4,5,9
2	3	2,3,6
	4	5

Key 4|1 = 14 mins Key 1|6 = 16 mins

(ii) If one of these companies is based in a county town and the other in a large city such as Cork or Dublin, could you guess which is which? Give a reason.

Employees in company B on average take longer to arrive at work. So this is more likely to be in the large city due to traffic problems.

4. Mean, median and mode

e.g. Consider the number of fingers that everybody in the world has on their right hand.

(i) What would you say the median number is?

Assume: thumb is a finger

The median would be 5 as those with more or less than 5 would be outliers leaving most, including the middle person with 5 fingers.

(ii) What would you say the mode is?

5 would be most common

(iii) Give a description of roughly what value the mean would be. Do you think this value is useful?

I imagine that there are more people with less than 5 fingers than with more than 5. The mean is likely to be slightly less than 5. (perhaps 4.99)

More trivia than useful in my opinion

5. Range, percentiles and IQR

e.g. At a Garda speed checkpoint, the speeds of a number of vehicles, in km/h, were recorded as they passed a certain point. The data is given below.

~~57~~ ~~59~~ ~~64~~ ~~48~~ ~~52~~ ~~58~~ ~~67~~ ~~54~~
~~53~~ ~~58~~ ~~72~~ ~~61~~ ~~55~~ ~~60~~ ~~57~~ ~~58~~
~~73~~ ~~66~~ ~~59~~ ~~56~~ ~~53~~ 61 ~~54~~ ~~59~~
~~68~~ ~~85~~ ~~57~~ ~~84~~ ~~58~~ ~~60~~.

(i) Show this data on a stem plot.

Stem & leaf plot

4	8					Q1									Q2				
5	2	3	3	4	4	5	6	7	7	7	8	8	8	8	9	9	9		
6	0	0	1	1	4	6	7	8											
7	2	3			Q3														
8	4	5																	

Key: 6|4 = 64 km/h

(ii) Calculate the lower quartile and the upper quartile.

30 vehicles

Quartile position

$Q_2 = 31/2 = 15.5 = (15+16)/2$

Q_1 = ½ way between start & Q_2

Q_3 = ½ way between end & Q_2

median: $Q_2 = 58.5$ km/h

lower quartile: $Q_1 = 56$ km/h

upper quartile: $Q_3 = 64$ km/h

(iii) Calculate the interquartile range.

$IQR = Q_3 - Q_1$

$Q_3 - Q_1 = 64 - 56 = 18$ km/h

6. Standard deviation

e.g. The mean of the numbers 4, x, 13, 17, 27 is 14. Find the value of x.
Calculate the standard deviation of these numbers.

$(\bar{x})$ mean = ?

$$\bar{x} = \frac{4 + x + 13 + 17 + 27}{5} = 14$$

$$\Rightarrow \quad x + 61 = 14(5)$$
$$x = 70 - 61$$
$$x = 9$$

(σ) standard deviation

X : 4, 9, 13, 17, 27, 14

use calculator :

$\sigma \approx 7.12$

7. Analysis of graphs

e.g. A frequency curve has a mode of 15 and a median of 13.

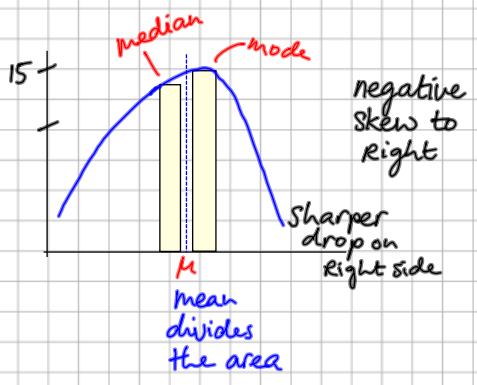
(i) Give a rough estimate of where the mean should lie.

mean is between mode and median

estimate : $m = 14$

(ii) Would you describe the frequency curve as symmetric, skewed left or skewed right? Give a reason.

(iii) Draw a frequency curve that fits the information given.



8. Scatterplots and correlation coefficient

e.g. For a number of patients undergoing kidney dialysis, measurements of heart rate (X) and blood pressure (Y) were taken. The data is given below in the form of couples (x, y).

(83,141), (86,162), (88,161), (92,154), (94,171), (98,174), (101,184), (114,190), (117,187), (121,191)

(i) Represent the data on a scatterplot.

(ii) Use the scatterplot to estimate the correlation coefficient.

estimate $r = 0.8$

(iii) Calculate the correlation coefficient.

using calculator

(iv) How accurate was your estimate from the scatterplot?

underestimates positive correlation

(v) Draw by eye the line of best fit.