M.K. HOME TUITION

Mathematics Revision Guides
Level: GCSE Higher Tier

SURVEYS AND SAMPLING

Spain	Portugal	Italy	Greece	Croatia
48	30	23	14	10

Spain	Portugal	Italy	Greece	Croatia
9	6	5	3	2

$$M = 80,\ m = 7,\ n = 50,$$

$$N = \frac{Mn}{m} = \frac{80 \times 50}{7} = 571$$

$$M = 80,\ m = 6,\ n = 50,$$

$$N = \frac{Mn}{m} = \frac{80 \times 50}{6} = 667$$

Manufacturer / Age	New (0-9yrs)	Older (10-19 yrs)	Oldest (20+ yrs)	**Total**
Boeing	19	42	25	**86**
Airbus	32	19	13	**64**
Others	2	4	4	**10**
Total	**53**	**65**	**42**	**160**

Version: 1.1 Date: 18-07-2016

HANDLING AND COLLECTING DATA

A few definitions.

The collection, set or group of objects being studied is termed the **population**.
Examples would include:

- All the pupils in Year 11 in a particular school
- All the people on the electoral register
- All people who watch soaps on television
- All the cars produced by a factory in one week
- All the hotels owned by a certain holiday chain

A **sample** is a small part of a population and can be selected using various ways.

A sample is said to be **biased** if there are any underlying factors distorting the data in such a way that it will not give a representative picture of the population.

Data Collecting.

There are several ways in which data can be collected directly.
Data can either be 'primary' or 'secondary'.

Primary data is information that is obtained directly from first-hand sources by means of surveys, observations or experiments.

Two common methods of collecting primary data are the tally chart and the questionnaire, although the latter is no studied in detail at Higher GCSE.

The Tally Chart.

This is typically used to record data which can be easily observed and counted, usually over a time interval, although it can also be used for static data.

A few examples:

- The number of items paid for at a "10 items or less" supermarket checkout over one day
- The number of cyclists using a stretch of road over 15-minute intervals
- The number of cars, by year of registration, using a stretch of road over one hour

Example (1): A group of Year 11 pupils have been tallying the number of cyclists using a section of the A56 into Manchester during the morning rush-hour on a Tuesday morning.

They grouped the times into 15-minute sections beginning at 0700, 0715, 0730 .... up to the quarter-hour beginning at 0915.

The completed tally chart looks like this, with the 'gate' graphics :

This tally chart can then be re-displayed as a bar chart, frequency polygon or histogram.

Biases in primary data collecting.

Biases in data collecting are usually centred about location or time, and examples are easy to find or suggest.

Example (2) : A group of Year 11 pupils distributed a questionnaire outside the leisure centre about how much time a week people spent on physical exercise. A disproportionately large number of respondents ticked the "Over 5 hours per week" box.
(Users of the leisure centre were more likely to engage in physical exercise than non-users.)

Example (3) : The results of a questionnaire about whether the widening of the M60 motorway was a good idea to improve traffic flow. The response was a very strong 'No'.
(The people questioned all lived within 200 metres of the proposed motorway.)

Example (4) : The queues of daytime traffic into Mill Gate car park in Bury, as counted on three successive Mondays and Tuesdays. The apparent results showed that long queuing (over 20 cars waiting) was not an issue, yet many shoppers complained on Wednesdays and Fridays about difficulty in finding parking spaces.
(Wednesdays and Fridays are full market days in Bury; Mondays and Tuesdays are not.)

Avoiding bias in primary data collecting.

Each of the questionnaires in the last three examples could have produced improved results in various ways. Here are some suggested corrections.

Example (2) Correction : Instead of distributing the questionnaire solely outside the leisure centre, the pupils should have chosen about four or five different town centre locations.

Example (3) Correction : Extend the area for distributing the questionnaire to include properties up to. say, 500 metres of the proposed motorway, rather than just the 'inner zone' of 200 metres.

Example (4) Correction : Count the traffic entering the car park for three complete weeks, Sundays to Saturdays, and not just Mondays and Tuesdays.

Secondary data is data collected by someone other than the user.

A few examples are :

- Using data from the Met Office website to compare Manchester's rainfall statistics for January 2015 and January 2016.
- Using data from the Department of Transport and Greater Manchester Police to analyse road accident statistics.
- Looking up the database of Olympic Games statistics to measure trends in performance for particular athletic events.

Sampling Techniques.

Sampling is a statistical method where a sample of the population is chosen to represent the whole, and is usually chosen because it would be otherwise too time-consuming and expensive to question the whole population.

Example (5) : A university wishes to select a sample of 400 students out of a total of 8000 to take part in an open question-and-answer session. The details of those students are stored on a database.
How can we select those 400 without showing a bias ?

This brings us to three main sampling techniques:

Systematic Sampling : A sample of 400 out of 8000 students is also 1 out of 20 or 5%. We could list all the students by alphabetical order, and select every 20^{th} one in the list.

Random Sampling : We can use a database query to select 400 random rows from the student database using randomising query commands.

Attribute Sampling: Each student has an ID containing a seven-digit numerical part. Since we are selecting one student in every 20, we could for example select only those students whose numeric IDs end in 20,40, 60, 80 or 00.

There are several other techniques available, not covered in the previous simple example.

Quota sampling is used by market research companies to target sections of the population by age, social class or other criteria. Further sub-sampling of the quota could be either random or systematic.

Cluster sampling is when the population is targeted by clusters, such as a set of 'key' political constituencies as used by opinion pollsters. Again, further sub-sampling of the quota could be either random or systematic in order to reduce the sample size to a few hundred voters.

Stratified Sampling is when the population is proportionally divided up into sub-groups or strata, and that the proportions in the sample reflect those in the population as a whole. The sampling itself can be random or systematic.

This technique is the one most commonly featured in examination questions and will be studied in more detail than the others.

Example (6): A holiday company owns several hotels in Europe as shown below:

Spain	Portugal	Italy	Greece	Croatia
48	30	23	14	10

An inspector wants to select a sample of 25 hotels stratified by country.
How many hotels by country should be included in the stratified sample ?

Firstly, we count the total number of hotels, which is (48 + 30 + 23 + 14 + 10) = 125 hotels.

The stratified sample consists of 25 hotels out of the 125.

The required fractional multiplier is thus $\frac{25}{125}$ or one fifth of the total.

There are 48 hotels in Spain, so the stratified sample would include $\frac{1}{5} \times 48 = 9.6$
or 10 hotels in Spain to the nearest integer.

The calculations for the other countries are similar. For Portugal, we have $\frac{1}{5} \times 30 = 6$
hotels ; for Italy, $\frac{1}{5} \times 23 = 4.6$ or 5 hotels, for Greece, $\frac{1}{5} \times 14 = 2.8.$ or 3 hotels,
and finally for Croatia, $\frac{1}{5} \times 10 = 2$ hotels.

Looking at the numbers of hotels in Portugal and Croatia, we can see that they are in the ratio 30:10 or 3:1. This ratio is preserved as 6:2, or 3:1 in the stratified sample as well.

This example highlights a possible snag due to rounding errors. Adding the results so obtained gives a total of 10 + 6 + 5 + 3 + 2 = 26 rather than 25. The inspector could either choose to inspect 26 hotels in total, or to inspect only (say) 9 hotels in Spain instead of 10. Choosing the second option gives the stratified sample as :

Spain	Portugal	Italy	Greece	Croatia
9	6	5	3	2

Example (7) : An airline has a fleet of 160 planes, of which 86 are manufactured by Boeing, 64 are manufactured by Airbus, and 10 are manufactured by other companies. They are also classified by age.

Manufacturer / Age	New (0-9yrs)	Older (10-19 yrs)	Oldest (20+ yrs)	**Total**
Boeing	19	42	25	**86**
Airbus	32	19	13	**64**
Others	2	4	4	**10**
Total	**53**	**65**	**42**	**160**

i) The airline owners want to carry out a quick random inspection of 15 new planes, stratified by manufacturer. How many planes, by manufacturer, should be included in the stratified sample ?

ii) The same airline wants a more thorough inspection of a random sample of 40 planes in the 'older' and 'oldest' categories, again stratified by manufacturer. How many planes, by manufacturer, should be included in the stratified sample this time ?

i) The total number of planes in the 'new' category is 53, and the stratified sample contains 15.
The required multiplier is therefore $\frac{15}{53}$.

The sample would therefore include $\frac{15}{53}\times 19 = 5.377...$ or 5 Boeings (to nearest integer),

$\frac{15}{53}\times 32 = 9.056...$ or 9 Airbuses, and finally $\frac{15}{53}\times 2 = 0.566...$ or 1 'other'.

$\therefore$ There are 5 Boeings, 9 Airbuses and 1 'other' in the stratified sample of 15 planes.

ii) Adding together the numbers of aircraft in the 'older' and 'oldest' categories, there are 67 Boeings, 32 Airbuses, and 8 'others' – a total of 107. There are 40 planes in the stratified sample, and so the required multiplier is $\frac{40}{107}$.

$\therefore$ There are $\frac{40}{107}\times 67$ or 25 Boeings, in the sample, $\frac{40}{107}\times 32$ or 12 Airbuses

and $\frac{40}{107}\times 8$ or 3 'others'.

The Peterson Capture-Recapture method.

This is sometimes known as the "mark and recapture" method, and is a useful way of estimating a population in a well-defined area, most usually animals in a nature reserve.

In this method, a proportion of the population is captured, labelled, and re-released into the wild. After a period of time, another proportion is captured, and some of the population will have the labels relating to the first capture. We can then use the number of recaptures to estimate the size of the whole population.

There are some assumptions to be borne in mind – here are a few :

- The eco-system is closed, without migration
- There are no births or deaths in between captures
- The markings do not become illegible
- The capture does not harm the specimen
- The marked specimens redistribute themselves uniformly within the eco-system

Example (8): A rare fish species was reintroduced into Lake Windermere ten years ago. Bill captured and marked 40 specimens, and then released them into the lake. Three months later, he recaptured another 40 specimens and found that five of them had been marked.

i) Calculate the estimated number of the rare fish in the lake.

ii) Bill noticed that the markings on a few of the recaptured fish had become difficult to detect, and was concerned about how the estimate would be affected if some of the marked fish had 'lost' their markings over time. Explain how that would affect his estimate of the fish population in part (i).

i) When Bill carried out the first capture, he marked 40 specimens.
Those 40 marked fish were then released into the lake.
Bill then recaptured another group of 40 fish, five of them were found to have been marked.

Now $\frac{5}{40} = \frac{1}{8}$, which infers that the original sample of 40 fish was one eighth of the lake's total.

Hence, the estimated population of the rare fish in the lake is $\frac{40}{\left(\frac{1}{8}\right)}$ or 320.

We could have used the formula $N = \frac{Mn}{m}$, where

M = the number of fish marked and released thereafter
n = number of fish in the recaptured sample
m = number of marked fish in the recaptured sample
N = estimated total population of fish in the lake

Hence, with $M = 40$, $m = 5$, $n = 40$, we have $N = \frac{Mn}{m} = \frac{40 \times 40}{5} = 320$.

ii) If some of Bill's marked fish were to have lost their markings, then the value of m would have been an underestimate.

Since the formula $N = \frac{Mn}{m}$ involves division by m, an underestimate in m would result in an **overestimate** in N, the total population of fish in the lake.

Example (9):

The population of natterjack toads on a reserve near Southport was estimated at 600 adults in 2012.

Kate carried out a survey as part of a school project in 2016, to determine if the population was endangered, in other words, whether it had declined by over 50%.

She labelled 80 adult toads and then released them back into the wild, and then recaptured 50 of them four weeks later, in two separate stages over several days, each time releasing them into the wild after counting.

In one recapture, she recorded seven labelled toads, and in another, she recorded six.
Do her results show that the population might be endangered ? Explain showing the full working.

We use the formula $N = \frac{Mn}{m}$ here ;

M = the number of toads labelled and released
n = number of toads in the recaptured sample
m = number of marked toads in the recaptured sample
N = estimated total population of toads on the reserve

With Kate's first recapture, we have $M = 80$, $m = 7$, $n = 50$, we have $N = \frac{Mn}{m} = \frac{80 \times 50}{7} = 571.$

With her second recapture, we have $m = 6$, so $N = \frac{80 \times 50}{6} = 667.$

Although the estimates are quite far apart, they are both close to the 2012 figure of 600 adults, and hence Kate's results suggest a stable population.