Statistics Sector 1: Sampling

After studying this section you should be able to:

1. Define random, stratified, systematic, quota, cluster and opportunity samples.
2. Select an appropriate method of sampling for use in particular circumstances.
3. Understand the advantages and disadvantages of the different methods of sampling.

The purpose of sampling is to obtain information about a population by examining only a part of it (a sample).
This may be done for reasons of economy (it's easier and cheaper to test a sample rather than the whole batch), because the test could be destructive (such as testing the working life of a large batch of electric light bulbs) or because it would be impossible, in practice, to examine the whole population.

If a sample is to be used to make inferences about the population or to estimate population parameters, it is essential that the sample is chosen, as far as possible, to be representative of the population and to avoid bias.

Random Sample

A random sample of size n is a sample selected in such a way that all possible samples of size n have an equal chance of being chosen

Each member of the population is equally likely to be chosen. In addition each possible sample of size n is equally likely to be chosen.

A **simple random sample** is a random sample taken without replacement.

- For example, if you decide to choose 5 students out of a population of 100 at random, you usually intend to choose 5 different students and would not consider choosing the same student twice. This is known as sampling without replacement

An **unrestricted random sample** is a random sample taken with replacement

- If you do allow a member of the population to be chosen more than once then this would be sampling with replacement.

Methods for choosing a random sample include:

1. Selecting via lottery methods (such as choosing numbers/names at random from a bag)
2. Random numbers (numbers which have been generated so that each digit from 0 to 9 has an equal chance of appearing in each position). These may be obtained from your calculator or from tables.

When describing how a random sample is taken, you need to mention the following points:

- Assign each person/item in the population with a number. You need to state the range of numbers needed and each number needs to be the **same number of digits**
  (so if there were 100 people in the population, you could assign each person a number from 00-99 or 001-100)
- If you use a random number table to obtain each number, you need to move along each row by the number of digits chosen in the previous step, writing down the number selected.
- Note that you could also use a graphical calculator to generate random numbers instead of using a random number table. For this example, you can use the first three digits of each number generated.
- Make sure that you mention that you ignore any numbers outside of the interval assigned to each person in the population, stating which numbers need to be ignored.
  (for example, if you chose 001-100 in the previous example, you would select 3-digit numbers from the random number table, rejecting 000 and any number > 100).
- You also need to state that you'll ignore repeats.
- This process of selecting numbers is continued until the correct amount for the sample size is collated (so if the sample size is 10, you would continue the process until 10 numbers had been found).
- The numbers found must then be used to select the corresponding individuals/objects that were assigned those numbers in the sample.

Example 1

Describe how a simple random sample of size 4 would be taken from a class of 20 students using a random number table

Number students 01 to 20 (or 00 to 19)
Move along random number table in pairs of digits
Ignore 00 and numbers > 20 (or numbers > 19)
Ignore repeats. Continue until you have four random numbers. Select corresponding students

ple 2

Example 2

On a particular day there are 2125 books on the shelves in the fiction section of a library. Describe how random numbers could be used to select a random sample of size 20 (without replacement) from the 2125 books.

Allocate random numbers 0001 to 2125 to the books
Move along random number table in sets of 4 digits
Ignore 0000 numbers > 2125 and repeats
Continue until you have twenty random numbers
Select corresponding books

Example 3

A market researcher wants to take a random sample of 2 houses from a street containing houses numbered 1 to 48. Use the following extract from a table of random numbers to choose the two houses

68236 35335 71329 96803 24413

(number houses 01 to 48 – already done)
~~68~~ 23 ~~63~~ ~~53~~ 35 House 23 and 35.
✓ ✓

How would you use the random number function on a calculator to do the same exercise?

ran# on calculator
look at different ways of using this to obtain a random sample.

Stratified Random Sample

A stratified sample requires prior knowledge to be used to divide the population into strata. Random samples are then taken from each of these strata, usually in proportion to the size of each strata.

Advantage
Ensures that the sample contains same proportion of each of the strata as the population.

Disadvantage
You need prior knowledge of the population.
More complicated to administer.
When should you use a stratified random sample instead of a simple random sample?
When you believe there to be a difference in "opinion" between the strata. E.g. Music taste: Young, Middle Aged, Old

Is each member of the population equally likely to be chosen?

Yes if the random samples are taken in proportion to the size of each strata.
When describing stratified sampling:

- Calculate how many people/objects from each strata have been selected in the sample (the proportion of the strata in the sample needs to equal the proportion of the population that is in that particular strata).
  For example if a strata contains 20% of a population, a sample of 20 would need 4 people from that strata.
- Describe how to select the number of people from the population within that strata using the simple random sampling method.
  So, if you've calculated that 4 people are to be selected from a strata that has a population of 50, you need to describe how to select a simple random sample of 4 people from a population of 50 using the steps in the previous section (all the details previously described need to be stated).
- Once you've described how one strata is randomly sampled, you can then state that the people/items other strata(s) are similarly selected.

Example 4

A German class consists of 12 male students and 18 female students. Describe how you would use random numbers to select a stratified sample of size 10 from the class

Population 30 — 4 males
Sample 10 — 33.3% sample — 6 females.

Number males from 01 to 12
Move along random number table in pairs of digits ignore 00, >12 and repeats continue until you have 4 numbers. Select corresponding males
Similarly select 6 females.

Systematic Sampling

In a systematic sample members of the population are chosen at regular intervals.

Random numbers are usually used to decide on the interval.

The sample can be set up so that each person is equally likely.

It is not a random sample though because not all possible samples of size n are equally likely. E.g. People next to each other have no chance of being in the same sample.

When describing systematic sampling:

- Divide the population size by the sample size to calculate the interval required
  So if 5 people are required from a population of 20, the interval would be 4.
- Assign the people/items with in population a number, using the same method as in simple random sampling
- Then describe how a number from the numbers within the interval is selected
- The number selected and every number that is one interval on from the previous selected number is then selected
  So if the interval of 4 means that we're selecting a number from 1-4 and the number 2 is selected, you would then chose every 4th number after until the correct amount of numbers had been selected for the sample. In this case it's be 02, 06, 10, 14 and 18
- The corresponding people/items to these numbers are then selected (this needs to be stated).

Example 5

Describe how a systematic sample of size 3 can be taken from a population of 18

$\frac{18}{3} = 6$

Number population from 1 to 18

Choose a random digit between 1 and 6

Select every 6th thereafter.

Example 6

An alphabetical list of the 2700 employees of a distribution company is available. Describe how a systematic sample of size 50 could be selected.

$\frac{2700}{50} = 54$

choose a random number between 00 and 54. This is the starting place in the list. Choose every 54th thereafter.

Quota Sampling

A quota sample is a stratified sample where the samples for each strata are chosen for convenience. How many are selected for each strata is calculate in proportion to the population in the same way as in stratified sampling. No attempt is made to select the sample within each strata at random. The sampler is given a quota of people to locate from a given strata but is left to choose for themselves the particular individuals to include in the sample

Use when it is impossible to take a random sample and prior knowledge is available to divide the population into strata.
It is often used in street interviews. The interviewer is often given a series of targets e.g. Equal number of men and women of whom ¼ should be aged over 60.

It is not as good as stratified random sampling because the interviewer will inevitably interview people who are easily accessible and willing to be interviewed.

Cluster Sampling

The population is divided into clusters. These are often chosen just for geographical convenience. A cluster sample uses randomly chosen clusters of the population. It is not a random sample but contains an element of randomness.

Advantage – easier to administer
Disadvantage – people within a cluster may be likely to have the same views

<u>Is each member of the population equally likely to be chosen?</u>
Generally No, people in small clusters are more likely to be chosen than people in large clusters.
If the clusters are of equal size then yes the clusters have an equal chance of being chosen.

You may have to describe how the clusters are selected randomly and/or a sample is randomly selected from within a cluster. Each of these are the same method as mentioned in simple random sampling.

<u>Example 7</u>

The membership secretary of a football supporters club, which has 120 branches, wishes to contact a representative sample of members. She has a complete list of members classified by branches. She selects four branches at random and then ten members at random from each of the chosen branches.

a. What name is given to this method of sampling? cluster
b. Would all members be equally likely to be included in the sample? Explain your answer.
c. Under what (unlikely circumstances) would all members be equally likely to be included in the sample?

b.) no those in small (branches) clusters would have a greater chance
c) if all 120 branches were the same size

Opportunity Sampling

This is when the sample is chosen from a section of the population that is convenient for the sampler
This is different to quota sampling, where the proportions of each strata are calculated in the same way as stratified sampling but then the people within each strata are selected from convenience.

Example 8

Mel thinks that most people watch her favourite TV programme. She asks 20 friends whether they watch the television programme.

a. Name the sampling method Mel used — opportunity
b. Give a reason why Mel's sample may be biased.

Asking friends - likely to watch the same sort of programme

Summary Table – Fill In The Gaps

Sample Type	Description	Advantages/Disadvantages	Are all subjects equally likely?	Is this a random sample?
Simple Random Sample	A random sample is taken from the population without replacement. Random numbers are usually used to select the sample.	<u>Advantages</u> Simple and easy to use where small populations involved Sample is representative as it is randomly chosen <u>Disadvantages</u> More time consuming and costly to carry out Does not ensure all the strata are fairly represented	Yes	Yes
Systematic Sample	In a systematic sample members of the population are chosen at regular intervals. Random numbers are usually used to decide on the interval.	<u>Advantages</u> Easier, quicker and cheaper to carry out Can give useful results so long as the pattern of sampling does not follow a pattern in the population <u>Disadvantages</u> Method may interact with a hidden pattern in the population leading to bias.	They are equally likely if the allocation of random numbers has been done correctly	No. Not random as not all subsets are possible e.g two consecutive numbers cannot be in the same sample.
Stratified Random Sample	Population is split into strata. Random samples are then taken from each of these strata, usually in proportion to the size of each strata.	<u>Advantages</u> Ensures that all strata are fairly represented. Sample is representative as it is randomly chosen <u>Disadvantages</u> You need prior knowledge. More complicated to administer.	Yes	No. Not all possible samples can be chosen. E.g You can't have a sample all from one strata
Cluster Sample	The population is divided into clusters. Clusters are chosen at random. The subjects are then chosen from those clusters.	<u>Advantages</u> Is less costly in terms of travelling, time and money. Contains an element of randomness. <u>Disadvantages</u> Subjects within a cluster may have similar views	Each cluster has an equal chance of being chosen No. Those in small clusters have a greater chance of being chosen	No

Quota Sample	The population is divided into strata. The samples from each strata are chosen for convenience. There is no attempt at random sampling.	<u>Advantages</u> Easier, quicker and cheaper to carry out. Can give useful results when the stratification is skilfully carried out <u>Disadvantages</u> Interviewers will choose people easily accessible to themselves which may introduce bias.	No this depends on the convenient method used to choose subjects.	No
Opportunity Sampling				

Exam style questions

1.

The government of a European country wishes to survey the opinions of its headteachers. The country is divided into 14 regions. It is proposed to select two of the regions at random and then to select 40 headteachers at random from each of these two regions. These 80 headteachers would then be asked to agree to a face-to-face interview.

(a) (i) Write down the name given to this method of sampling. *(1 mark)*

(ii) Give **one** advantage and **one** disadvantage of this method of sampling. *(2 marks)*

(b) It is decided to undertake a pilot study before carrying out the survey. For this pilot study a sample of 20 headteachers will be contacted by telephone. A list of the names of the 1934 headteachers in the country is obtained and the names are numbered from 0000 to 1933. It is proposed to select a number between 0000 and 0033 by a random process and to include the corresponding headteacher and every 100th headteacher thereafter in the sample (ie if the number 0017 is selected, then headteachers numbered 0017, 0117, 0217, … , 1917 would be included in the sample).

(i) Write down the name given to this method of sampling. *(1 mark)*

(ii) State, giving a reason, whether every headteacher would be equally likely to be selected. *(2 marks)*

(c) One change was made to the method proposed in part (b). The process was to be started by selecting a number between 0000 and 0099, instead of between 0000 and 0033.

For this changed method:

(i) state, giving a reason, whether every headteacher would be equally likely to be selected; *(2 marks)*

(ii) explain why the sample obtained would **not** be a random sample; *(2 marks)*

(iii) comment on the size of the sample which would be obtained. *(2 marks)*

7(a)(i)	cluster sampling	B1		cluster
(ii)	reduces travelling time/expense	E1		less travelling/expense
	head teachers in same region may be more homogenous than all head teachers/sample not representative/random	E1	3	more homogenous/not representative/random
(b)(i)	systematic sampling	B1		systematic
(ii)	no – many head teachers have no chance of being selected e.g. 0034	B1 E1	3	no reason
(c)(i)	yes – there is one number between 00 and 99 corresponding to each head teacher.	B1		yes
	Probability 0.01	E1	2	explanation or 0.01
(ii)	not all combinations possible e.g. numbers 0000 and 0001 can not both be included in the sample	E2(1)	2	two marks for clear explanation
(iii)	sample size would depend on number picked	E1		sample size variable
	00 - 33 → sample of 20 34 - 99 → sample of 19	E1	2	explanation or statement that size may be 19 or 20
	Total		**12**	

2.

A bus company has 950 employees who are divided into four employment categories as follows:

620 drivers;
120 mechanics;
130 clerical staff;
and 80 managers.

For each category, a list of the names of the employees is available.

The company hopes to improve morale among the employees by providing better and healthier meals in its canteen. As a first step, it plans to distribute a questionnaire to a sample of employees on the meals currently provided in the canteen.

(a) Describe how random numbers could be used to select a sample of 95 employees stratified by employment category. *(6 marks)*

(b) Describe how a systematic sample of 95 employees could be selected. *(3 marks)*

(c) Data from a similar questionnaire, used at another company, had been analysed. The results suggested that there was little difference in the opinions on canteen meals between the different employment categories but that there was a difference in the opinions on canteen meals between males and females.

The bus company is to decide whether to use the stratified sample from part (a), a differently stratified sample or the systematic sample from part (b). Make **three** points to be considered before this decision is made. *(3 marks)*

7(a)	Number drivers 000 to 619	E1		number drivers 000 to 619 or mechanics/clerical staff/managers
	Select 3 digit random numbers	E1		select 3 digit random numbers
	Ignore repeats and >619 Continue until 62	E1		ignore repeats
	numbers obtained	E1		ignore >619 (must be consistent with numbering)
	Select corresponding drivers	E1		idea of stratified sample
	Similarly select 12 mechanics,13 clerical staff, 80 managers	E1	6	12,13,8 or explanation why not necessarily so allow max 3 for random sample
(b)	Number all employees 000 to 949	E1		number 000 to 949
	Choose a random digit between 0 and 9	E1		choose a random digit between 0 and 9
	Select every 10th employee e.g. if 7 picked select 007,017.....947	E1	3	select every 10th employee
(c)	No point in stratifying by employment categories if no difference between categories	E1		
	Would be worth stratifying by sex	E1		
	Systematic sample would not ensure a fair representation of sexes (unless men numbered together and women numbered together)	E1	3	a mark for any sensible point - max 3

A company, with 9320 employees, provides refuse collection services for 47 councils in the United Kingdom. The company asks a market research firm to carry out an opinion poll of its employees concerning union membership.

(a) Describe how the market research firm could obtain a simple random sample of size 120 from the 9320 employees. *(4 marks)*

(b) The market research firm selects 4 of the 47 councils at random.

(i) What further step(s) would be necessary to obtain a cluster sample of size 120 from the 9320 employees? *(2 marks)*

(ii) Give a reason why the market research firm might prefer a cluster sample to a random sample. *(2 marks)*

(c) It is proposed that a stratified sample be used.

(i) Suggest two factors which could be used to stratify the sample. *(2 marks)*

(ii) Suggest a reason why a stratified sample might be preferred to a cluster sample. *(1 mark)*

Q	Solution	Marks	Total	Comments
5(a)	number employees 0000 to 9319 select 4-digit random numbers ignore repeats and >9319 continue until 120 numbers obtained select corresponding employees	E1 E1 E1 E1	4	any valid numbering select 4-digit random numbers ignore repeats and >9319 (must be consistent in numbering) continue until 120 numbers obtained
(b)(i)	from each of the 4 chosen councils select a random sample of 30 employees	E1 E1	2	select a sample from each of the 4 councils of size 30
(b)(ii)	employees to be interviewed would be geographically localised / easier / cheaper	E2,1	2	reason – easier/cheaper without further explanation gets E1
(c)(i)	council / age / sex / length of service	B1B1	2	any sensible suggestion; B1 for each
(c)(ii)	More representative of population	E1	1	more representative allow all have equal chance
	Total		**11**	

4.

The Goodwell Medical Practice has a total of 3200 registered patients. The local health authority has asked this practice to complete questionnaires about a sample of 40 patients, but has not said how this sample should be chosen.

The practice has available a list of patients ordered alphabetically by their family name. The patients are numbered in the list from **1** to **3200**. The details of **20** patients are printed on each page of the list.

(a) Dr Dobry suggests picking two pages at random from the list and using the patients on those pages as the sample.

(i) Name this method of sampling.

(ii) Give a reason why Dr Dobry's method would be unlikely to give a representative sample.

[2 marks]

(b) Dr Kalos suggests obtaining the sample of 40 from the list by a systematic method. Explain briefly how this could be done.

[2 marks]

(c) Dr Bueno says the sample should be chosen at random in such a way as to proportionately represent the age distribution of the patients.

(i) Name this method of sampling.

(ii) Given that there are 737 registered patients aged over 60 years, how many of these should be chosen in a sample of 40 chosen using Dr Bueno's method?

[3 marks]

(d) Dr Mabuti suggests obtaining a simple random sample by the following method.

- Obtain a four-digit random number from tables, rejecting any number above 9599.
- Divide the number by 3200 and find the remainder.
- Add 1 to this remainder.
- Select the patient in the list corresponding to this number.
- Carry out this procedure 40 times.

(i) Using this method, which number of the patient in the alphabetical list would be generated by the random number 5817?

(ii) Explain why it is necessary to reject any random number above 9599.

(iii) Explain why it is necessary to add 1.

(iv) Dr Mabuti has omitted one instruction which is needed to make sure this method gives a simple random sample of patients. What instruction needs to be added?

[5 marks]

Q6	Solution	Marks	Total	Comments
(a) (i)	Cluster sampling	B1		
(ii)	Because the list is alphabetical by family name so the sample may contain several members of the same family	E1		Any indication of the problems arising because the list is alphabetic
			2	
(b)	Use random numbers to select a patient between 1 and 80 Select every 80th patient after that.	M1 A1		Allow even if method of random selection is not given SC If M0 then "every 80th" gains B1
			2	
(c)(i)	Stratified sampling	B1		
(ii)	$737 \div 3200 \times 40$ (= 9.2125) = 9	M1 A1		Must be integer
			3	
(d)(i)	2617 + 1 = 2618	M1 A1		
(ii)	Otherwise those numbered 1 to 400 would have a greater chance of being chosen than other numbers	E1		OE, being generous on details throughout part (d)
(iii)	Otherwise remainder 0 would not have a corresponding patient, Or otherwise patient 3200 could not be chosen	E1		Or random number 0000, 3200, 6400 stated For either of these
(iv)	Rejecting/ignoring any repeats	E1		
			5	

5.

A fan club has 2076 members who are divided geographically into 8 branches. The club's committee wishes to seek members' views on where to hold the next annual meeting. A sample of 100 members is to be obtained and their views sought.

The following suggestions are made as to how to choose the sample.

Suggestion A Members are selected from all 8 branches. The number of members from each branch is proportional to the size of the branch. The branch secretaries are asked to choose the appropriate number of members from their branches in any convenient way.

Suggestion B Two of the branches are selected at random. Fifty members from each of these branches are selected at random.

Suggestion C Members are selected by a random process from each branch. The number of members from each branch is proportional to the size of the branch.

Suggestion D The names of all members are to be listed and numbered from 0000 to 2075. One hundred different four-digit random numbers between 0000 and 2075 are taken from random number tables and the corresponding members are included in the sample.

(a) For **each** of the four suggestions:

(i) name the method of sampling; *(4 marks)*

(ii) either state that each member is equally likely to be included in the sample, or explain why this is not the case. *(6 marks)*

(b) (i) State, giving a reason, which of the four methods is preferable from a statistical point of view. *(2 marks)*

(ii) Give a reason why Suggestion A might be preferred to Suggestion C. *(1 mark)*

Q	Solution	Marks	Total	Comments
6(a)	**A** quota	B1		quota
	Not equally likely – those who are easy to contact most likely to be chosen	B1 E1		not equally likely explanation – allow – depends how secretaries choose samples
	B cluster	B1		cluster
	Not equally likely – those in small branches most likely to be chosen	B1 E1		not equally likely explanation allow - equally likely if branches of equal size
	C stratified (random)	B1		stratified
	Equally likely	B1		equally likely
	D random	B1		random
	Equally likely	B1	10	equally likely
(b)(i)	**C** ensures all branches fairly represented and all members equally likely to be chosen	B1 E1	2	**C** all branches fairly represented
(ii)	Easier to carry out	E1	1	reason

6.

A university employs 820 staff at a city centre site.

Natasha, an administrator in charge of car parking, wishes to survey members of staff as to their views on the present parking arrangements. A list of the 820 members of staff is available.

(a) Describe how Natasha could select a simple random sample of size 25 for the survey. *(4 marks)*

(b) The university has a car park with 200 parking spaces. The spaces are numbered from 1 to 200. There are 300 members of staff with permits to park in this car park and each morning the spaces are allocated on a first come, first served basis. Permit holders arriving when the car park is full have to park elsewhere. There are 220 members of staff on a waiting list for a permit and the remaining 300 members of staff do not wish to have a permit.

(i) It is suggested that, instead of a simple random sample, Natasha should take a stratified sample. Suggest two relevant factors that Natasha could use in the stratification of the staff. *(2 marks)*

(ii) A second suggestion is as follows:

Select a systematic sample of 25 parking spaces. Natasha and her assistant will wait in the car park in the morning and when a car parks in one of the selected spaces the driver will be given a questionnaire and asked to complete it.

(A) Describe how the systematic sample of 25 parking spaces could be selected.

(B) State one advantage of this method of obtaining completed questionnaires.

(C) State two sources of bias in this method of data collection. *(5 marks)*

8(a)	Number staff 000 to 819 Select 3-digit random numbers Ignore >819 Ignore repeats Continue until 25 selected and choose corresponding staff	E1 E1 E1 E1	4	Valid numbering 3-digit random numbers ignore >819 and repeats continue until 25 selected
(b)(i)	Permit holder/waiting list/other male/female full-time/part-time etc	B1 B1	2	permit holder status any other sensible strata
(ii)	(A) Choose a digit between 1 and 8 at random. Pick this space and every 8th thereafter. e.g. 3,11,19.....187,195	E1 E1		idea of systematic sampling correct method including "every 8th"
	(B) Easy and quick (C) Excludes anyone without a permit, favours those who usually arrive early etc	B1 E1 E1	5	easy - or other valid advantage any reasonable source of possible bias any different reasonable source of possible bias
	Total		**11**	