Populations, Samples and Censuses

Definition. The population is the whole set of individuals or items we want information about. A census collects data from *every* member of the population. A sample is a subset of the population from which data are actually collected.

Definition. A sampling unit is an individual member of the population that can be selected. A sampling frame is a list of all sampling units — the thing you actually sample from (e.g. the school register, the electoral roll, a numbered list of components).

We use samples to make *informal inferences* about the population: the sample mean estimates the population mean, sample proportions estimate population proportions, and so on. Different samples lead to different estimates — and possibly different conclusions — which is why how the sample is chosen matters so much.

Example (Census or sample?)
Why might we choose to sample rather than carry out a census?

- ***Cost and time**: a census of a large population is expensive and slow (the UK decennial census is a ten-year, billion-pound operation).*
- ***Destructive testing**: testing the lifetime of a lightbulb or the breaking strain of a climbing rope destroys it — a census would destroy the entire stock.*
- ***Practicality**: some populations cannot be listed at all (all the fish in the North Sea).*
- ***Accuracy**: counter-intuitively, a well-run sample can be* more *accurate than a badly-run census — a huge census is hard to administer and suffers non-response and processing errors. (This is a serious argument, often made about the US census, for replacing or supplementing censuses with sampling.)*
- ***Ethics**: it may be unjustifiable to subject every member of a population to an intrusive or risky test when a sample would answer the question.*

A census, on the other hand, gives completely accurate information about the population (if administered perfectly) and is sensible for small populations.

Tip
If a question mentions testing items to destruction, the word **census** should ring an alarm bell: you cannot sell lightbulbs you have burnt out.

Random Sampling Methods

A sampling method is **random** if every member of the population has a known, non-zero probability of selection, with the selection made by a chance mechanism. The methods below require a sampling frame (except cluster sampling, which requires only a frame of clusters).

Simple random sampling

Definition. A **simple random sample** of size n is one chosen so that every possible subset of n members of the population is equally likely to be chosen. Equivalently: every member has equal probability of selection, independently of who else is selected.

How to carry it out: number every member of the population on the sampling frame (1 to N); generate n distinct random numbers in this range (using a calculator, random number tables or a computer); select the corresponding members. Unless told otherwise, assume the population is large enough that sampling without replacement is fine.

- *Advantages*: free of selection bias; every member equally likely; the basis of all the theory we develop later.
- *Disadvantages*: needs a complete sampling frame, which may not exist; can be slow and expensive for large or dispersed populations; may, by bad luck, miss small subgroups.

Systematic sampling

Definition. A **systematic sample** takes every kth member of the sampling frame, where $k = N/n$, starting from a randomly chosen member among the first k.

How to carry it out: to choose $n = 50$ from $N = 1000$, compute $k = 20$; pick a random starting point between 1 and 20 (say 7); select members $7, 27, 47, \ldots$

- *Advantages*: quick and simple to operate, especially for production lines; spreads the sample evenly through the frame.
- *Disadvantages*: needs a sampling frame; **biased if the frame has a periodic pattern** (e.g. every 20th item coming off a machine with 20 moulds would all come from the same mould). Only the starting point is random.

Stratified sampling

Definition. In a **stratified sample**, the population is divided into non-overlapping groups (**strata**) expected to differ in the quantity of interest (e.g. year groups, age bands, gender). A simple random sample is taken *within each stratum*, with size proportional to the size of the stratum:

$$\text{sample from stratum} = \frac{\text{stratum size}}{\text{population size}} \times \text{total sample size.}$$

- *Advantages*: guarantees the sample reflects the structure of the population across the strata; usually more precise than a simple random sample of the same size.
- *Disadvantages*: needs a sampling frame *with strata information*; strata must be sensible and non-overlapping; more administration.

Example

A sixth form has 360 students in Year 12 and 240 in Year 13. Describe how to take a stratified sample of 50 students.

Sample sizes proportional to strata sizes:

$$\textit{Year 12: } \frac{360}{600} \times 50 = 30, \qquad \textit{Year 13: } \frac{240}{600} \times 50 = 20.$$

Number the Year 12 students from 1 to 360 using the school roll; generate 30 distinct random numbers in this range and select the corresponding students. Similarly select 20 of the 240 Year 13 students using 20 distinct random numbers. (If proportional allocation gives non-integers, round sensibly.)

Example

A quality controller wishes to take a systematic sample of 25 light fittings from the 2000 produced by a machine in one day.

(a) Describe how the sample should be taken.

(b) The machine uses 8 moulds in rotation. Explain why a sample of every 80th fitting could give misleading results.

(a) $k = 2000/25 = 80$. *Choose a random integer between 1 and 80, say by random number generator — suppose it is 34. Select the 34th fitting produced, then every 80th thereafter: the 34th, 114th, 194th, . . .*

(b) Since 80 is a multiple of 8, every fitting in the sample comes from the same mould. *If one mould is faulty (or unusually good), the sample will be completely unrepresentative of the day's production: the sampling interval resonates with the periodic pattern in the frame.*

Example (OCR MEI Paper 2, June 2024)

A teacher is investigating how pupils travel to and from school each day. Pupils can travel by bus, train, car, bicycle, or on foot. The teacher decides to collect a sample of size 60 for the investigation.

(a) The teacher lives in a village 10 miles away from the school. Explain how collecting a sample which consists only of pupils who live in the same village as the teacher might introduce bias.

The table shows how many pupils there are in each year.

Year 7	Year 8	Year 9	Year 10	Year 11
86	105	107	101	101

(b) The teacher decides to use proportional stratified sampling. Calculate the number of pupils in the sample who are in Year 9.

(c) The teacher generates a sample of 10 pupils from the 86 in Year 7 by listing them in alphabetical order and selecting the first name on the list and every ninth name thereafter. Explain whether this method will generate a simple random sample of the Year 7 pupils.

(a) Pupils living 10 miles from the school are unlikely to walk or cycle, so the sample would over-represent bus, train and car travel and under-represent walking and cycling — it is biased towards certain modes of transport.

(b) Total number of pupils $= 86 + 105 + 107 + 101 + 101 = 500$*, so Year 9 should contribute*

$$\frac{107}{500} \times 60 = 12.84 \approx 13 \text{ pupils.}$$

(c) No. In a simple random sample every possible sample of 10 pupils must be equally likely; here, once the first name is fixed the whole sample is determined, so most subsets of 10 pupils (e.g. two adjacent names on the list) can never be selected. (Saying "because it is systematic" is not an explanation — the reason is that not all samples are equally likely. Note also that the start is not even random here.)

Cluster sampling

Definition. In a **cluster sample**, the population is divided into naturally occurring groups (**clusters**), e.g. schools within a county or boxes within a warehouse. A random sample of *clusters* is chosen, and then every member of the chosen clusters is surveyed (or a random sample within each).

- *Advantages*: cheap and practical when the population is geographically spread — you only travel to the chosen clusters; no frame of individuals needed, only a frame of clusters.
- *Disadvantages*: clusters tend to be internally similar (a school's students share a catchment area), so the sample can be unrepresentative; less precise than the other random methods for the same sample size.

Remark. Stratified vs cluster — easily confused. Stratified: sample from **every** stratum, where strata are deliberately chosen to differ from each other. Cluster: sample only **some** clusters, ideally each cluster being a miniature of the whole population.

Non-Random Sampling Methods

Quota sampling

Definition. In a quota sample, the interviewer is given quotas to fill from specified subgroups (e.g. "interview 20 men and 20 women, 10 of each aged under 30"), but chooses *which* individuals fill each quota — typically whoever is available.

- *Advantages*: no sampling frame needed; fast and cheap; the sample matches the population's structure across the quota characteristics — this is why market researchers use it.
- *Disadvantages*: **non-random**: within each quota the interviewer's choice introduces selection bias (they approach people who look approachable, at one time and place); no valid measure of sampling error.

Example (OCR MEI AS Paper 2, November 2020 (parts))

A researcher is conducting an investigation into the number of portions of fruit adults consume each day. The researcher decides to ask 50 men and 50 women to complete a simple questionnaire.

(a) State the type of sampling procedure the researcher is using.

(b) Write down one disadvantage of this sampling procedure.

(c) A second researcher chooses a proportional stratified sample of 100 children from years 5 and 6 in a certain primary school. There are 220 children to choose from. In year 5 there are 125 children, of whom 81 are boys. How many year 5 girls should be included in the sample?

(a) Quota sampling (50 men and 50 women are quotas; the researcher chooses who fills them).

(b) It is non-random: the researcher's choice of who to ask introduces selection bias.

(c) Year 5 girls: $125 - 81 = 44$, *and each child is sampled with proportion* $\frac{100}{220}$, *so*

$$44 \times \frac{100}{220} = 20 \text{ girls.}$$

Opportunity (convenience) sampling

Definition. An opportunity sample (or convenience sample) simply takes whichever members of the population are conveniently available — the first 30 people you meet, your own class, volunteers who reply to a poster.

- *Advantages*: the cheapest and quickest method of all; sometimes the only practical option (e.g. a pilot study).
- *Disadvantages*: **non-random** and highly likely to be biased — those who are available (or who volunteer) often differ systematically from the population; conclusions cannot safely be generalised.

Fact (Summary: random or not?) —

Method	Random?	Sampling frame needed?
Simple random	Yes	Yes
Systematic	Yes (only the start)	Yes
Stratified	Yes	Yes, with strata information
Cluster	Yes (clusters chosen at random)	Only a list of clusters
Quota	No	No
Opportunity	No	No

Bias, Representativeness and Exam Technique

Definition. A sampling method is **biased** if it systematically over- or under-represents some part of the population, so that estimates from the sample tend to differ from the true population values in a consistent direction. A sample is **representative** if its composition reflects that of the population in the respects that matter for the question being asked.

Common sources of bias: an incomplete sampling frame; non-random selection (interviewer choice, self-selection by volunteers); non-response; sampling at a particular time or place (a survey outside a gym at 6am does not represent the town); periodic patterns interacting with systematic sampling.

> **Tip**
> Exam answers about sampling must be **in context**. "It might be biased" scores nothing; "students who arrive early are more likely to be sampled, and they may also be more conscientious, so the sample is likely to overestimate average homework time" scores everything. Name the group that is over/under-represented and say what effect this has on the conclusion.

Example (Describe how)

A factory produces 5000 jars of jam per day. Describe how to take a simple random sample of 40 jars from one day's production to check filled mass.

Number the jars from 1 to 5000 (the day's production list is the sampling frame). Use a calculator or random number generator to produce 40 distinct random integers between 1 and 5000, ignoring repeats. Select the jars with those numbers and weigh them.

Example (Criticise)

To estimate how long students at a school spend on homework, a teacher questions the first 30 students who arrive at the library on Monday morning.

(a) Name the sampling method used.

(b) Give two criticisms of this method in context.

(c) Suggest an improvement.

(a) Opportunity (convenience) sampling.

(b) Students at the library are likely to be more studious than average, so the sample will tend to overestimate *homework time for the school. Also, only students who arrive early on a Monday can be selected — the method is non-random, and most of the school population has no chance of being included, so the sample is unrepresentative.*

(c) Take a random sample using the school roll as a sampling frame: e.g. a stratified sample by year group (since homework time plausibly varies by year), with students within each year group selected using random numbers.

Example

A county council wants to survey primary school pupils about school dinners. It selects 5 of the county's 80 primary schools at random and surveys every pupil in those schools.

(a) Name the sampling method.

(b) Give one advantage and one disadvantage of this method in context.

(c) Explain how a stratified element could improve the design.

(a) Cluster sampling (the schools are the clusters, chosen at random).

(b) Advantage*: the council only needs to visit* 5 *schools rather than all* 80*, saving substantial time and cost, and no county-wide list of pupils is needed.* Disadvantage*: pupils within one school share the same kitchen, menus and catchment area, so the chosen schools may not represent the county — e.g. if all* 5 *happen to be rural schools, urban pupils' views are missed entirely.*

(c) Stratify the schools *before selecting: divide the* 80 *schools into groups expected to differ (e.g. urban/rural, or by size), and randomly select schools from each group in proportion, ensuring all types of school are represented.*

Example (OCR MEI S3, June 2006 (parts))

An employer has commissioned an opinion polling organisation to undertake a survey of the attitudes of staff to proposed changes in the pension scheme. The staff are categorised as management, professional and administrative, and it is thought that there might be considerable differences of opinion between the categories. There are 60, 140 and 300 staff respectively in the categories. The budget for the survey allows for a sample of 40 members of staff to be selected for in-depth interviews.

(i) Explain why it would be unwise to select a simple random sample from all the staff.

(ii) Discuss whether it would be sensible to consider systematic sampling.

(iii) What are the advantages of stratified sampling in this situation?

(iv) State the sample sizes in each category if stratified sampling with as nearly as possible proportional allocation is used.

(i) A simple random sample of 40 *might, by chance, be very unrepresentative of the three categories — it could contain hardly any (or even no) managers, whose opinions are thought to differ considerably from the rest.*

(ii) There is presumably a list of all 500 *staff, so systematic sampling (every* $\frac{500}{40} = 12.5$*th, i.e. alternately every* 12*th and* 13*th member) is possible. If the list is alphabetical, the sample may still misrepresent the categories; but if the list is arranged* by category*, a systematic sample would spread proportionally across the categories and could work well.*

(iii) Every category is guaranteed to be covered, in proportion; and the survey can also report results within *each category separately — exactly what is wanted when the categories are expected to differ.*

(iv) Proportional allocation with $N = 60 + 140 + 300 = 500$ and $n = 40$:

$$40 \times \tfrac{60}{500} = 4.8, \qquad 40 \times \tfrac{140}{500} = 11.2, \qquad 40 \times \tfrac{300}{500} = 24,$$

so as nearly as possible: 5 *management*, 11 *professional*, 24 *administrative (total* 40*).*

Remark. Different samples lead to different conclusions. Two simple random samples of 30 students from the same school will give different mean homework times — this is *sampling variability*, not bias. The hypothesis-testing machinery in the rest of this course exists precisely to decide whether an observed difference is too large to be explained by sampling variability alone.

Example (In class)

A market researcher stands outside a supermarket on a weekday morning and interviews shoppers until she has spoken to 25 men and 25 women. Name the method, and discuss whether the results will fairly represent the views of the town's residents.

Textbook Exercises: [CUP.1] Ch 18 §1; [S2] Ch 4