Sampling methods and the large data set

S1.1 · Statistics · revision notes

Populations, samples and the main sampling methods (simple random, systematic, stratified, quota and opportunity), with their advantages and disadvantages, plus what you need to know about the large data set of weather data. Questions are short: describe a method, give an advantage or disadvantage in context, or use your knowledge of the large data set.

Key facts

- Census: data from every member of the population. Sample: data from a selection of them
- Sampling frame: a list of all the sampling units in the population
- Stratified: number from a stratum $= \dfrac{\text{stratum size}}{\text{population size}} \times \text{sample size}$
- Systematic: every kth item, $k = \dfrac{\text{population size}}{\text{sample size}}$, with a random start among the first k
- Need a sampling frame: simple random, systematic, stratified
- Not random, no sampling frame needed: quota, opportunity
- Large data set: 5 UK and 3 overseas stations; 1 May to 31 October (184 days); 1987 and 2015
- tr = trace of rain (less than 0.05 mm); n/a = not available; 1 knot = 1.15 mph

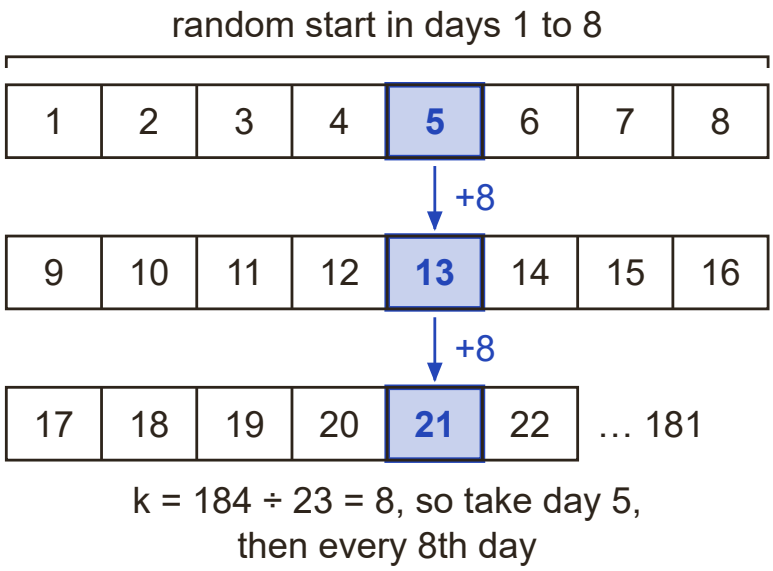
Notes

Populations and samples

- The **population** is the whole set of items of interest. A **census** collects data from every member; a **sample** from a selection of them.
- A census should be fully accurate, but it is slow and expensive, gives a lot of data to process, and is impossible when testing destroys the item (e.g. testing how long batteries last).
- A sample is quicker and cheaper but may not represent the population. Different samples can lead to different conclusions (natural variation); larger samples usually give more reliable results.
- A **sampling unit** is one member that can be sampled; a **sampling frame** is a numbered list of the sampling units.

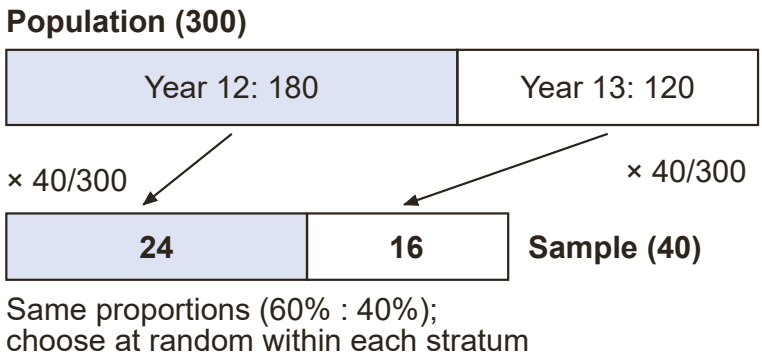
Random sampling methods (all need a sampling frame)

- **Simple random**: every sample of size n is equally likely. Number the sampling frame, generate random numbers and ignore repeats. Free of bias, but impractical for a large population.
- **Systematic**: take every kth item, where $k = \frac{\text{population size}}{\text{sample size}}$, starting at a random item among the first k. Quick for large populations, but biased if the list has a pattern.



Systematic sampling: a random start among the first k, then every kth item

- **Stratified**: split the population into groups (strata), sample each in proportion to its size, and choose randomly within each stratum. It reflects the population's structure, but the strata must be clearly defined.



Stratified sampling: each stratum is sampled in proportion to its size

Non-random sampling methods

- **Quota**: an interviewer fills a quota for each group (e.g. 20 men and 30 women) with people they choose. No sampling frame is needed and it is quick and cheap, but the interviewer's choices can cause bias, and people who refuse are simply replaced.
- **Opportunity (convenience)**: use whoever is available at the time and fits the criteria. Easy and cheap, but unlikely to be representative.

The large data set

- Met Office weather data from five UK stations (Camborne, Heathrow, Hurn, Leeming, Leuchars) and three overseas stations (Beijing, Jacksonville, Perth), for every day from 1 May to 31 October in 1987 and in 2015.
- Variables include daily mean temperature (°C), total rainfall (mm) and total sunshine (hours), mean windspeed and maximum gust (knots), cloud cover (oktas: eighths of the sky), visibility, pressure (hPa) and relative humidity (%). The overseas stations record only temperature, rainfall, pressure and windspeed.
- **tr** means a trace of rain (less than 0.05 mm); **n/a** means not available. 1 knot = 1.15 mph.
- Perth is in the southern hemisphere, so May to October includes its winter.

How to answer each type of question

M = method mark, A = answer mark, B = a mark on its own.

Describe how to take a simple random or systematic sample 2 to 3 marks **Grade E**

1. Say how the sampling frame is numbered (e.g. days 1 to 184).
2. Simple random: generate random numbers in the range, ignoring repeats, until you have enough.
3. Systematic: find k = population size ÷ sample size, choose a random start from 1 to k, then take every kth item.

Example. Priya wants a systematic sample of 23 days from the 184 days of data for Leeming in 2015. Explain how she could take this sample.

Model answer.

Number the days 1 to 184 in date order and find $184 \div 23 = 8$ (B1)

Choose a random starting day from days 1 to 8, e.g. using the random number function on a calculator (B1)

Take that day and every 8th day after it, e.g. days 5, 13, 21, …, 181 (B1)

Calculate and describe a stratified sample

3 to 4 marks | Grade C

1. Find each stratum's share: stratum size ÷ population size × sample size.
2. Round sensibly and check that the numbers add up to the sample size.
3. Describe a simple random sample within each stratum (numbered list, random numbers, ignore repeats).

Example. A sixth form has 180 students in Year 12 and 120 students in Year 13. A sample of 40 students, stratified by year group, is to be taken.
(a) Find the number of students to be sampled from each year group.
(b) Describe how the Year 12 students in the sample could be chosen.
(c) Give one advantage of stratified sampling in this situation.

Model answer.

(a) Year 12: $\frac{180}{300} \times 40 = 24$ (M1); Year 13: $\frac{120}{300} \times 40 = 16$ (A1)

(b) Number the Year 12 students 1 to 180 from a list of names, then use random numbers to choose 24 different students, ignoring repeats (B1)

(c) The sample has the same proportions of Year 12 and Year 13 students as the sixth form, so both year groups are fairly represented (B1)

Critique a sampling method and suggest an improvement

2 to 4 marks | Grade B

1. Name the method.
2. Say who is likely to be left out or over-represented, and why, in context.
3. Suggest a specific improvement, e.g. a simple random sample from a named sampling frame.

Example. A council wants residents' views on a new cycle lane. An officer stands outside one supermarket on a Monday morning and asks the first 50 adults who pass.
(a) Name this sampling method.
(b) Give two reasons why the sample may not represent the residents.
(c) Suggest an improvement.

Model answer.

(a) Opportunity (convenience) sampling (B1)

(b) Any two, e.g. people who work on Monday mornings are unlikely to be included (B1); only people near one supermarket are asked, and some of them may not live in the area (B1)

(c) Use a list of residents, such as the electoral register, as a sampling frame and take a simple random sample from it (B1)

Use your knowledge of the large data set 1 to 2 marks each **Grade A**

1. Recall the relevant fact: station location, dates, units or abbreviation.
2. Apply it to the situation: explain the pattern or problem, not just the fact.

Example. (a) In a sample of daily total rainfall from Camborne in 2015, some values are recorded as 'tr'. State what 'tr' means.
(b) A student draws graphs of daily mean temperature against date for Perth and for Leuchars, using the 1987 data. Using your knowledge of the large data set, explain why the two graphs have different shapes.

Model answer.

(a) A trace of rain: less than 0.05 mm fell that day (B1)

(b) Perth is in the southern hemisphere, so May to October includes its winter: temperatures tend to fall to a minimum in the middle of the period, then rise (B1). Leuchars is in the UK, where the period includes summer, so temperatures tend to rise to a maximum in the middle of the period, then fall (B1)

Don't lose marks

- Describing simple random sampling without saying how the items are numbered, how the random numbers are generated, or that repeats are ignored.
- Forgetting the random start in systematic sampling: 'take every 8th day' on its own is not enough.
- Calling opportunity or quota sampling 'random'.
- Giving a generic answer such as 'it is biased' without saying who is left out and why.
- Rounding stratum sizes so that they no longer add up to the sample size.

More tips

Memory tricks

- Only the random methods (simple random, systematic, stratified) need a sampling frame; quota and opportunity sampling do not.
- Stratified and quota sampling both split the population into groups: stratified then chooses randomly within each group, quota lets the interviewer choose.
- Check a stratified sample: the numbers from the strata must add up to the total sample size.
- The large data set has 184 days per station per year (31 + 30 + 31 + 31 + 30 + 31), which is handy for systematic sampling questions.

Exam technique

- Advantages and disadvantages must fit this situation: use the words of the question (the shoppers, the days, the students).
- You don't need to remember values from the large data set, but you do need to know its structure: stations, dates, years, variables, units and abbreviations.
- If asked why two samples lead to different conclusions, mention natural variation between samples (and a small sample size).

What each grade needs

What you need to be able to do, from the first marks up to the top grade.

Grade E — **Define population, sample, census and sampling frame**
A census measures every member of the population; a sample is a selection, chosen from a sampling frame (a list of the sampling units).

Grade E — **Describe how to take a simple random sample**
Number every item in the sampling frame, then use random numbers to choose the required number of different items.

Grade D — **Describe systematic, quota and opportunity sampling**
Say how each method is carried out and whether it needs a sampling frame.

Grade C — **Calculate the numbers for a stratified sample**
Sample each stratum in proportion to its size, then choose randomly within each stratum.

Grade B — **Give advantages and disadvantages in context**
Link each point to the situation, e.g. no list of shoppers exists, so a quota sample is practical.

Grade A — **Critique a sampling method and suggest improvements**
Identify who is left out or over-represented, and why, and propose a random method that fixes it.

Grade A* — **Reason with knowledge of the large data set**
Use facts about the stations, dates, variables and abbreviations to explain patterns or problems in a sample.

Free notes, Full packs, practice papers and mock exams at chhetriacademy.com. Written for this site; not taken from past papers. Stuck? 1-to-1 tutoring with a free 30-minute trial lesson.