Statistical sampling

AQA A-level Mathematics 7357, section K: populations, samples, sampling methods, bias, the large data set

The words of sampling

Term	Meaning
Population	the whole group the claim is about
Census	data from every member of the population
Sample	the subset actually measured
Sampling unit	one member that can be selected
Sampling frame	the list the sample is drawn from

Census or sample?

Census

Accurate for the whole population.

Slow and expensive for large groups.

Impossible if testing destroys the item.

Sample

Cheaper, quicker, less data to handle.

Results vary from sample to sample.

Can be biased if badly chosen.

> Testing the life of light bulbs destroys them: only a sample makes sense.

Simple random sampling

1. **Frame** number the 500 students 001 to 500
2. **Select** generate random numbers from 001 to 500
3. **Repeats** ignore repeats until 25 different numbers are chosen
4. **Measure** survey those 25 students

Decision

Free of selection bias, but needs a full list and may miss small groups.

Systematic sampling

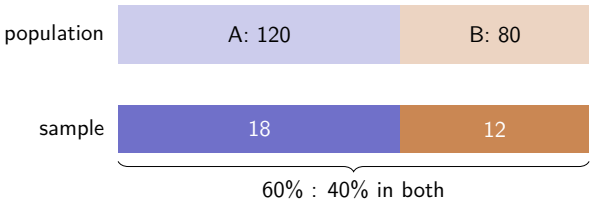


- $N = 40$, $n = 8$: every $k = 5$th.
- Random start in 1–5, here 3.
- Units 3, 8, 13, 18, 23, 28, 33, 38.

$$k = \frac{N}{n}$$

take every kth unit after a random start

Stratified sampling, drawn



- 120 in group A, 80 in group B; sample 30.
- $30 \times \frac{120}{200} = 18$, $30 \times \frac{80}{200} = 12$.
- Then select randomly within each group.

$$n_{\text{group}} = n \times \frac{N_{\text{group}}}{N}$$

Stratified allocation with rounding

Group	**Size**	$30 \times \frac{\textbf{size}}{200}$	**Take**
Year 12	47	7.05	7
Year 13	85	12.75	13
Staff	68	10.2	10
Total	200	30	30

If rounding gives 29 or 31, adjust the group whose decimal was closest to .5.

Non-random methods

Quota sampling

Fill a quota per group, such as 18 from A and 12 from B.

No frame needed; the interviewer picks who to ask.

Opportunity sampling

Take whoever is available, such as the first 30 people at the gate.

Easiest of all, and the most likely to be unrepresentative.

Cluster sampling

1. **Clusters** a city of 40 schools; each school is a cluster
2. **Select** choose 4 schools at random
3. **Measure** survey every student in those 4 schools

Decision

Cheap to run, but clusters that differ from each other make results vary.

The methods compared

Method	Random?	Frame?	Main weakness
Simple random	yes	yes	small groups can be missed
Systematic	start only	yes	a periodic list biases it
Stratified	within groups	yes	needs group sizes known
Quota	no	no	interviewer choice
Opportunity	no	no	rarely representative

Size cannot fix bias

1. **Survey** a voluntary online poll on a new bus route gets 5000 replies
2. **Who replies** those with strong opinions, and those online
3. **Effect** the extremes are over-represented, whatever the count
4. **Fix** a random sample of 200 from the electoral roll

Decision

A small random sample beats a large biased one.

Working with the large data set

Check	Why
Variable definitions and units	a value means nothing without its unit
Missing or blank entries	missing is not zero: exclude, do not count as 0
Dates and regions covered	a sample from one year or region may not generalise
How the data were collected	the collection method decides possible bias

Which method?

Your turn

A school wants the views of its 600 students, with the year groups represented in proportion. It has a full list of students by year group. Which method fits best?

A. Opportunity sampling at lunch

B. A stratified sample by year group, random within each year

C. A quota sample chosen by the interviewer

D. A census of all 600

Check yourself

Question	**Answer**
Sampling frame for a school survey?	the list of students
1200 people, sample 60: systematic k?	20
Stratified: 150 and 150, sample 40?	20 and 20
Why not test every battery?	testing destroys them
Does a larger biased sample help?	no: same bias