Data Collection Cheat Sheet

Edexcel Stats/Mech Year 1

Population and sample

In statistics, population is the whole set of items that are of interest. Information obtained from a population is known as raw data. A census measures or observes every member of a population. A sample is a selection of observations taken from a subset of population and used to find out more information about the population as a whole.

	Advantages	Disadvantages
Census	• Results should be completely accurate	• Time consuming and expensive • Cannot be used when testing destroys process • Hard to process large quantity of data
Sample	• Less time consuming and cheaper • Fewer people have to respond • Less data needs to be processed	• Data may not be as accurate • Sample may not be large enough to give information about small subgroups of the population

Individual units of a population are known as sampling units. Sampling units are named and numbered to form a list called a sampling frame.

Random sampling

Each member of the population has an equal chance of being selected. The sample should be representative of the population and bias should be removed. There are 3 types of random sampling.

- Simple random sampling

A simple random sample of size n is one where every sample of size n has an equal chance of being selected.

Example 1: The 100 members of a yacht club are listed alphabetically in the club's membership book. The committee wants to select a sample of 12 members to fill in a questionnaire. Explain how a simple random sample can be taken using:

A) Calculator or random number generator:

Number each member from 1-100. Use a calculator or random number generator to generate 12 random numbers between 1-100. Select the members who correspond to the numbers.

B) Lottery sampling:

Write the name of members on identical cards and place them in the hat. Draw up 12 cards and select these members.

Advantages	Disadvantages
• Free of bias • Easy and cheap for small samples and populations • Each sampling unit has a known and equal chance of selection	• Not suitable for large samples and populations • Sampling frame needed

- Systematic sampling

The required elements are chosen at regular intervals from an ordered list.

Example 2: A sample of size 20 is required from a population of 100.

$100 \div 20 = 5$ so every fifth person is chosen.
The first person is chosen at random.
If the first person chosen is 2, the remaining samples will be 7, 12, 17 etc.

Advantages	Disadvantages
• Simple and quick to use • Suitable for large samples and large populations	• A sampling frame is needed • Bias introduced if sampling frame is not random

- Stratified sampling

The population is divided into mutually exclusive strata and a random sample is taken from each.

$$\text{Number sampled in a stratum} = \frac{\text{number in stratum}}{\text{number in population}} \text{ x overall sample size}$$

Example 3: A factory manager wants to find out about what his workers think about the factory canteen facilities. He decides to give a questionnaire to a sample of 80 workers. It is thought that different age groups will have different opinions.

There are 75 workers between ages 18 and 32, 140 workers between ages 33 and 47, and 85 workers between ages 48 and 62.

Explain how he can use stratified sampling to select the sample.

1. Total number of workers: $75 + 140 + 85 = 300$
2. Finding the number of workers needed from each age group:
   18-32: $\frac{75}{300} \times 80 = 20$ workers
   33-47: $\frac{140}{300} \times 80 = 37\frac{1}{3} \approx 37$ workers
   48-62: $\frac{85}{300} \times 80 = 22\frac{2}{3} \approx 23$ workers
   If the number of workers required is not a whole number, it is rounded off to the nearest whole number.
3. Number the workers in each group.
4. Use a random number generator or table to produce the required quantity of random numbers.

Advantages	Disadvantages
• Sample accurately reflects population structure • Proportional representation of group within population	• Population must be clearly classified into distinct strata • Same disadvantages as simple random sampling within each stratum

Non-random sampling

There are two types of non-random sampling that you need to know:

- Quota sampling

An interviewer or researcher selects a sample that reflects the characteristics of the whole population.

Advantages	Disadvantages
• Allows a small sample to still be representative of the population • No sampling frame required • Quick, easy and inexpensive • Easy comparison between different groups within a population	• Non-random sampling can introduce bias • Population must be divided into groups, which can be costly or inaccurate • Increasing scope of study increases number of groups, which adds time and expenses • Non-responses not recorded

- Opportunity sampling or convenience sampling

Sample is taken from people who are available at the time of study and who fits the criteria you are looking for.

Advantages	Disadvantages
• Easy and inexpensive	• Unlikely to provide a representative result • Highly dependent on individual researcher

Types of data

Variables or data associated with numerical observations are called quantitative variables or quantitative data.

Variables associated with non-numerical observations are qualitative variables or qualitative data.

A variable that can take any value in a given range is a continuous variable. A variable that can only take specific values is a discrete variable.

In a grouped frequency table, the specific data values are not shown.

- Class boundaries show the maximum and minimum values in each group or class
- The midpoint is the average of class boundaries
- The class width is the difference between upper and lower class boundaries

Large data set

If you need to do calculations on large data sets in your exam, the relevant extract will be provided.