DP Analysis & Approaches (SL/HL)

Maths with Lemon

Topic 4.1 – Sampling and Data Analysis

Theory and Notes

1. Population, Sample and Variables (SL 4.1)

A **population** is the entire group we want to investigate. A **sample** is a subset of the population used to make estimates. A **random sample** is selected using a chance process rather than the investigator's personal preference.

Example: To estimate nightly homework time at a school of 1500 students, the population is all 1500 students; the sample is the students surveyed; the sampling variable is nightly homework time (minutes).

Target population	The group about which conclusions are intended.
Sampling frame	The list or source from which sampling units can be selected.
Sampling unit	One member or item selected from the sampling frame.
Sampling variable	The characteristic measured for each sampling unit.
Sampling values	The possible outcomes of that measurement.

2. Five Sampling Techniques

Method	How it works	Typical strength / limitation
Simple random	Choose individuals at random from a complete list.	Reduces selection bias; requires a suitable frame.
Convenience	Choose people who are easiest to reach.	Fast, but often unrepresentative.
Systematic	Choose a random starting position and then every kth item.	Easy to organize; ordering may introduce bias.
Quota	Fill predetermined numbers in specified categories without necessarily sampling randomly.	Ensures specified categories are included; within-group bias remains.
Stratified	Split population into groups and randomly sample in population proportions.	Represents group proportions; needs accurate group sizes.

Example (stratified): A school is 60% female and 40% male. For a proportional stratified sample of 100, randomly select 60 girls and 40 boys.

Example (systematic): Number 1500 students and select every 20th after a random start from 1 to 20. This gives $1500/20 = 75$ students.

3. Types of Data

- **Categorical (qualitative):** labels such as eye colour or transport method.
- **Numerical discrete:** countable values, such as the number of children in a family.
- **Numerical continuous:** measurements such as height, rainfall, mass and temperature.

A continuous quantity may be *recorded* to the nearest whole unit without becoming discrete in nature. For example, a bottle's fill volume is continuous even when reported to the nearest millilitre.

4. Reliability, Bias and Missing Data

Ask: Who collected the data? How were participants chosen? Was the question neutral? Were some groups excluded? Are observations missing or incorrectly recorded?

Selection bias: friends may not represent the whole school. **Non-response bias:** students who do not answer may differ from those who do. **Measurement error:** inaccurate timing or recording can distort results.

A larger sample is not automatically representative: a large convenience sample can remain biased. Missing data should be reported and handled consistently, rather than silently replaced with zero.

5. Outliers and the Interquartile Range

Let Q_1 and Q_3 be the lower and upper quartiles. Then

$$\mathrm{IQR} = Q_3 - Q_1.$$

A value is classified as an outlier if it is below $Q_1 - 1.5\mathrm{IQR}$ or above $Q_3 + 1.5\mathrm{IQR}$.

Worked example: If $Q_1 = 12$ and $Q_3 = 20$, then $\mathrm{IQR} = 8$. The fences are 0 and 32. A recorded value of 35 is an outlier. Investigate whether it is a valid unusual observation or an error before deciding what to do with it.

6. Sample Means and the Population Mean

For values $x_1, \ldots, x_n$, the sample mean is

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

For the population $\{0, 1, 2, 3, 4\}$,

$$\mu = \frac{0 + 1 + 2 + 3 + 4}{5} = 2.$$

There are $\binom{5}{3} = 10$ unordered samples of three distinct digits. Across all ten possible samples, each digit appears in $\binom{4}{2} = 6$ samples. Therefore the mean of the ten sample means is

$$\frac{6(0 + 1 + 2 + 3 + 4)}{10 \cdot 3} = \boxed{2}.$$

This illustrates that the mean of all equally likely simple random sample means equals the population mean; individual sample means may be above or below it.

7. Applying the Theory to the Pencil Investigation

Measure all 50 pencils to obtain the population mean, then draw samples by simple random, systematic, stratified and quota methods. For each method, record the selected pencils, calculate $\bar{x}$, and compare it with the population mean μ using $\bar{x} - \mu$ or $|\bar{x} - \mu|$. Different samples may give different results.

Important: One sample cannot prove a method is always more accurate. Repeated samples are needed to compare typical performance.



Maths with Lemon

Mathematics made clear.