Stats1 Chapter 1: Data Collection

1.1 Populations and Samples

The chapters of Stats Year 1 could be broadly organised as follows:

Experimental

i.e. Dealing with collected data.

Chp1: Data Collection

Methods of sampling, types of data, and populations vs samples.

Chp2: Measures of Location/Spread

Statistics used to summarise data, including mean, standard deviation, quartiles, percentiles. Use of linear interpolation for estimating medians/quartiles.

Chp3: Representation of Data

Producing and interpreting visual representations of data, including box plots and histograms.

Chp4: Correlation

Measuring how related two variables are, and using linear regression to predict values.

Theoretical

Deal with probabilities and modelling to make inferences about what we 'expect' to see or make predictions, often using this to reason about/contrast with experimentally collected data.

Chp5: Probability

Venn Diagrams, mutually exclusive + independent events, tree diagrams.

Chp6: Statistical Distributions

Common distributions used to easily find probabilities under certain modelling conditions, e.g. binomial distribution.

Chp7: Hypothesis Testing

Determining how likely observed data would have happened 'by chance', and making subsequent deductions.

This Chapter Overview

There is little 'calculation' involved in this chapter; consider this a 'bookwork' one!

1a: Types of data

Continuous vs discrete, terms such as class intervals, class boundaries, class width.

1b: Populations vs samples

"Suggest why we would not test all the light bulbs."
"Identify the sampling frame."

2:: Random Sampling

Describe the disadvantages of systematic sampling.

3:: Non-Random Sampling

Describe how a stratified sample would be conducted, including strata sizes.

5:: Edexcel's 'Large Data Set'

What you're expected to know about the 'large data set' of weather data, and how to use it.

Types of Data

Qualitative/Categorical

Non-numerical values, e.g. colour.

Quantitative

Numerical values.

Discrete

Can only take specific values, e.g. shoe size, number of children.

Note that while discrete variables only allow specific values, the range could still be infinite, e.g. "number of attempts before success".

Continuous

Can take any decimal value (possible with a specified range).

Weight w (kg)	Frequency
$0 \leq w < 20$	3
$20 \leq w < 70$	4

Data can be **grouped** for conciseness, at the expense of losing the exact original values.

$20 \leq w < 70$

This is known as a **class interval**.

Lower class boundary

Midpoint **= 45**

Upper class boundary

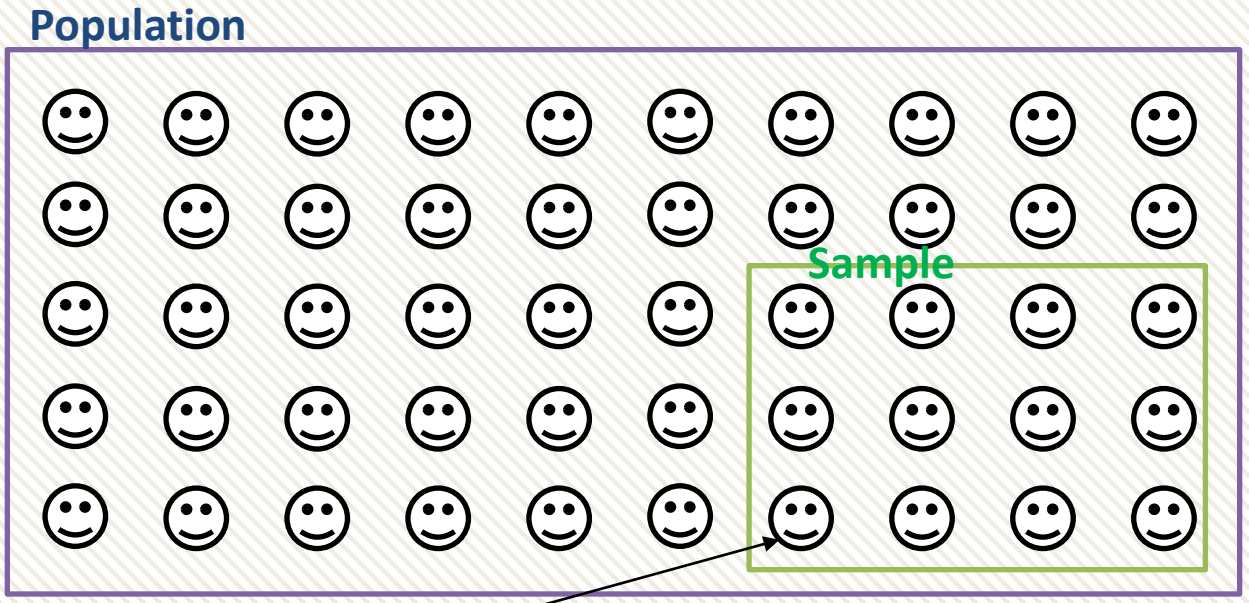
Class width = $70 - 20 = \mathbf{50}$

Populations and samples

A **population** is:	the <u>whole</u> set of items that are of interest.
A **sample** is:	some <u>subset of the population</u> intended to <u>represent the population</u>.

You're probably used to a 'population' meaning all humans/animals within a country/ecosystem. But a population could be "*all the lightbulbs in a factory*" or "*all the cars in the UK*".

Sampling key terms



- Each individual thing in the population that can be sampled is known as a **sampling unit**.

- Often sampling units of a population are individually named or numbered **to form a list** called the **sampling frame**.

Populations vs Samples

We could collect data either from a sample, or from the entire population.
Data collected from the entire population is known as a ?

	Advantages	Disadvantages
Census	?	?
Sample	?	?

Example: A supermarket wants to test a delivery of avocados for ripeness by cutting them in half.

a. Suggest a reason why the supermarket should not test all the avocados in the delivery.

b. The supermarket tests a sample of 5 avocados and finds that 4 of them are ripe. They estimate that 80% of the avocados in the deliver are ripe. Suggest one way that the supermarket could improve their estimate.

a ?

b ?

Populations vs Samples

We could collect data either from a sample, or from the entire population.
Data collected from the entire population is known as a **census**.

	Advantages	Disadvantages
Census	Should give completely accurate result.	• Time consuming and expensive. • Can not be used when testing involves destruction. • Large volume of data to process.
Sample	• Cheaper. • Quicker. • Less data to process.	• Data may not be accurate. • Data may not be large enough to represent small sub-groups.

Example: A supermarket wants to test a delivery of avocados for ripeness by cutting them in half.

a. Suggest a reason why the supermarket should not test all the avocados in the delivery.
b. The supermarket tests a sample of 5 avocados and finds that 4 of them are ripe. They estimate that 80% of the avocados in the deliver are ripe. Suggest one way that the supermarket could improve their estimate.

a Testing the avocados <u>destroys</u> them (and thus can't be sold).

b Use a <u>larger sample size</u> (as this would be better estimate of the proportion of ripe avocados).

Exercise 1.1

Pearson Statistics & Mechanics Year 1/AS
Page 1-2

Homework Exercise

1 State whether each of the following variables is qualitative or quantitative.

a Height of a tree

b Colour of car

c Time waiting in a queue

d Shoe size

e Names of pupils in a class

2 State whether each of the following quantitative variables is continuous or discrete.

a Shoe size

b Length of leaf

c Number of people on a bus

d Weight of sugar

e Time required to run 100 m

f Lifetime in hours of torch batteries

3 Explain why:

a 'Type of tree' is a qualitative variable

b 'The number of pupils in a class' is a discrete quantitative variable

c 'The weight of a collie dog' is a continuous quantitative variable.

4 The distribution of the masses of two-month-old lambs is shown in the grouped frequency table.

Mass, m (kg)	Frequency
$1.2 \leqslant m < 1.3$	8
$1.3 \leqslant m < 1.4$	28
$1.4 \leqslant m < 1.5$	32
$1.5 \leqslant m < 1.6$	22

Hint The class boundaries are given using inequalities, so the values given in the table are the actual class boundaries.

a Write down the class boundaries for the third group.

b Work out the midpoint of the second group.

c Work out the class width of the first group.

Homework Exercise

5 A school uses a census to investigate the dietary requirements of its students.

a Explain what is meant by a census.

b Give one advantage and one disadvantage to the school of using a census.

6 A factory makes safety harnesses for climbers and has an order to supply 3000 harnesses. The buyer wishes to know that the load at which the harness breaks exceeds a certain figure.

a Suggest a reason why a census would not be used for this purpose.

The factory tests four harnesses and the load for breaking is recorded:

320 kg 260 kg 240 kg 180 kg

b The factory claims that the harnesses are safe for loads up to 250 kg. Use the sample data to comment on this claim.

c Suggest one way in which the company can improve their prediction.

7 A city council wants to know what people think about its recycling centre.
The council decides to carry out a sample survey to learn the opinion of residents.

a Write down one reason why the council should not take a census.

b Suggest a suitable sampling frame.

c Identify the sampling units.

Homework Exercise

8 A manufacturer of microswitches is testing the reliability of its switches. It uses a special machine to switch them on and off until they break.

a Give one reason why the manufacturer should use a sample rather than a census.

The company tests a sample of 10 switches, and obtains the following results:

23 150 25 071 19 480 22 921 7455

b The company claims that its switches can be operated an average of 20 000 times without breaking. Use the sample data above to comment on this claim.

c Suggest one way the company could improve its prediction.

9 A manager of a garage wants to know what their mechanics think about a new pension scheme designed for them. The manager decides to ask all the mechanics in the garage.

a Describe the population the manager will use.

b Write down the main advantage in asking all of their mechanics.

Homework Answers

1 **a** Quantitative **b** Qualitative
c Quantitative **d** Quantitative
e Qualitative

2 **a** Discrete **b** Continuous
c Discrete **d** Continuous
e Continuous **f** Continuous

3 **a** It is descriptive rather than numerical.
b It is quantitative because it is numerical. It is discrete because its value must be an integer; you cannot have fractions of a pupil.
c It is quantitative because it is numerical. It is continuous because weight can take any value in a given range.

4 **a** 1.4 kg and 1.5 kg **b** 1.35 kg
c 0.1 kg

5 **a** A census observes or measures every member of a population.
b Advantage: will give a completely accurate result. Disadvantage: ANY ONE FROM: time consuming, expensive.

6 **a** The testing process will destroy the harness, so a census would destroy *all* the harnesses.
b 250 kg is the median load at which the harnesses in the sample break. This means that half of the harnesses will break at a load less than 250 kg.
c Test a larger number of harnesses.

7 **a** ANY ONE FROM:
It would be expensive.
It would be time consuming.
It would be difficult.
b A list of residents. **c** A resident.

8 **a** The testing process will destroy the microswitches, so a census would destroy *all* the switches.
b The mean is less than the stated average but one of the switches lasted a significantly lower number of operations which suggests the median might be a better average to take – not affected by outliers. The data supports the company claim.
c Test a larger number of microswitches.

9 **a** All the mechanics in the garage.
b Everyone's views will be known.