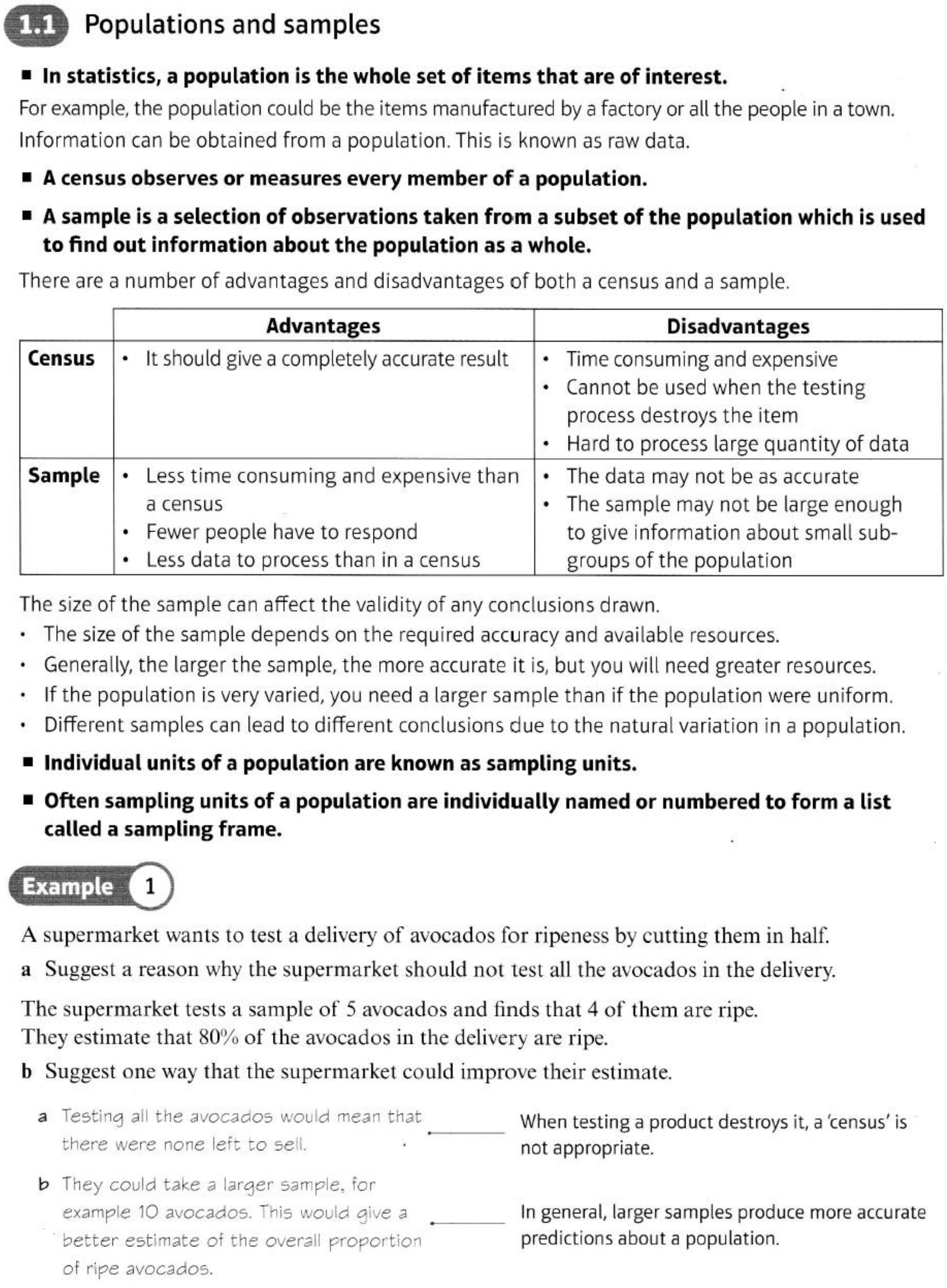
Data Collection



1.1 Populations and samples

- **In statistics, a population is the whole set of items that are of interest.**

For example, the population could be the items manufactured by a factory or all the people in a town. Information can be obtained from a population. This is known as raw data.

- **A census observes or measures every member of a population.**
- **A sample is a selection of observations taken from a subset of the population which is used to find out information about the population as a whole.**

There are a number of advantages and disadvantages of both a census and a sample.

	Advantages	Disadvantages
Census	• It should give a completely accurate result	• Time consuming and expensive • Cannot be used when the testing process destroys the item • Hard to process large quantity of data
Sample	• Less time consuming and expensive than a census • Fewer people have to respond • Less data to process than in a census	• The data may not be as accurate • The sample may not be large enough to give information about small sub-groups of the population

The size of the sample can affect the validity of any conclusions drawn.

- The size of the sample depends on the required accuracy and available resources.
- Generally, the larger the sample, the more accurate it is, but you will need greater resources.
- If the population is very varied, you need a larger sample than if the population were uniform.
- Different samples can lead to different conclusions due to the natural variation in a population.

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

Example 1

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.

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 delivery are ripe.

b Suggest one way that the supermarket could improve their estimate.

a Testing all the avocados would mean that there were none left to sell.

When testing a product destroys it, a 'census' is not appropriate.

b They could take a larger sample, for example 10 avocados. This would give a better estimate of the overall proportion of ripe avocados.

In general, larger samples produce more accurate predictions about a population.

1.2 Sampling

In random sampling, every member of the population has an equal chance of being selected. The sample should therefore be **representative** of the population. Random sampling also helps to remove **bias** from a sample.

There are three methods of random sampling:

- Simple random sampling
- Systematic sampling
- Stratified sampling

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

To carry out a simple random sample, you need a sampling frame, usually a list of people or things. Each person or thing is allocated a unique number and a selection of these numbers is chosen at random.

There are two methods of choosing the numbers: generating random numbers (using a calculator, computer or random number table) and **lottery** sampling.

In lottery sampling, the members of the sampling frame could be written on tickets and placed into a 'hat'. The required number of tickets would then be drawn out.

Example 2

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.

a Explain how the committee could use a calculator or random number generator to take a simple random sample of the members.

b Explain how the committee could use a lottery sample to take a simple random sample of the members.

a Allocate a number from 1 to 100 to each member of the yacht club. Use your calculator or a random number generator to generate 12 random numbers between 1 and 100.
Go back to the original population and select the people corresponding to these numbers.

If your calculator generates a number that has already been selected, ignore that number and generate an extra random number.

b Write all the names of the members on (identical) cards and place them into a hat. Draw out 12 names to make up the sample of members.

■ **In systematic sampling, the required elements are chosen at regular intervals from an ordered list.**

For example, if a sample of size 20 was required from a population of 100, you would take every fifth person since 100 ÷ 20 = 5.

The first person to be chosen should be chosen at random. So, for example, if the first person chosen is number 2 in the list, the remaining sample would be persons 7, 12, 17 etc.

- **In stratified sampling, the population is divided into mutually exclusive strata (males and females, for example) and a random sample is taken from each.**

The proportion of each strata sampled should be the same. A simple formula can be used to calculate the number of people we should sample from each stratum:

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

Example 3

A factory manager wants to find out what his workers think about the factory canteen facilities.

The manager 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.

There are 140 workers between 33 and 47.

There are 85 workers between 48 and 62.

a Write down the name of the method of sampling the manager should use.

b Explain how he could use this method to select a sample of workers' opinions.

a Stratified sampling.

b There are: 75 + 140 + 85 = 300 workers altogether.

Find the total number of workers.

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.

For each age group find the number of workers needed for the sample.

Where the required number of workers is not a whole number, round to the nearest whole number.

Number the workers in each age group. Use a random number table (or generator) to produce the required quantity of random numbers. Give the questionnaire to the workers corresponding to these numbers.

Each method of random sampling has advantages and disadvantages.

Simple random sampling	
Advantages	**Disadvantages**
• Free of bias • Easy and cheap to implement for small populations and small samples • Each sampling unit has a known and equal chance of selection	• Not suitable when the population size or the sample size is large • A sampling frame is needed