SAMPLING

POPULATIONS, SAMPLES AND CENSUS

- **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. Unprocessed information 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.**

RANDOM SAMPLING (SIMPLE, SYSTEMATIC AND STRATIFIED)

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

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.**

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.

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 as it is potentially time consuming, disruptive and expensive. • A sampling frame is needed

SYSTEMATIC SAMPLING

- **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 $100 \div 20 = 5$.

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.

Systematic sampling	
Advantages	**Disadvantages**
• Simple and quick to use • Suitable for large samples and large populations	• A sampling frame is needed • It can introduce bias if the sampling frame is not random

STRATIFIED SAMPLING

- **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:

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

Stratified sampling	
Advantages	**Disadvantages**
• Sample accurately reflects the population structure • Guarantees proportional representation of groups within a population	• Population must be clearly classified into distinct strata • Selection within each stratum suffers from the same disadvantages as simple random sampling

NON RANDOM SAMPLING (QUOTA, OPPORTUNITY)

QUOTA SAMPLING

- **In quota sampling, an interviewer or researcher selects a sample that reflects the characteristics of the whole population.**

The population is divided into groups according to a given characteristic. The size of each group determines the proportion of the sample that should have that characteristic.

As an interviewer, you would meet people, assess their group and then, after interview, allocate them into the appropriate quota.

This continues until all quotas have been filled. If a person refuses to be interviewed or the quota into which they fit is full, then you simply ignore them and move on to the next person.

Quota sampling	
Advantages	**Disadvantages**
• Allows a small sample to still be representative of the population • No sampling frame required • Quick, easy and inexpensive • Allows for 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 expense • Non-responses are not recorded as such

OPPORTUNITY SAMPLING

- **Opportunity sampling consists of taking the sample from people who are available at the time the study is carried out and who fit the criteria you are looking for.**

Notation Opportunity sampling is sometimes called **convenience sampling**.

This could be the first 20 people you meet outside a supermarket on a Monday morning who are carrying shopping bags, for example.

Opportunity sampling	
Advantages	**Disadvantages**
• Easy to carry out • Inexpensive	• Unlikely to provide a representative sample • Highly dependent on individual researcher

THE BIG DATA SET

You will need to answer questions based on real data in your exam. Some of these questions will be based on weather data from the **large data set** provided by Edexcel.

The data set consists of weather data samples provided by the Met Office for five UK weather stations and three overseas weather stations over two set periods of time: May to October 1987 and May to October 2015. The weather stations are labelled on the maps below.

The large data set contains data for a number of different variables at each weather station:

- **Daily mean temperature** in °C – this is the average of the hourly temperature readings during a 24-hour period.
- **Daily total rainfall** including solid precipitation such as snow and hail, which is melted before being included in any measurements – amounts less than 0.05 mm are recorded as 'tr' or 'trace'
- **Daily total sunshine** recorded to the nearest tenth of an hour
- **Daily mean wind direction and windspeed** in knots, averaged over 24 hours from midnight to midnight. Mean wind directions are given as bearings and as cardinal (compass) directions. The data for mean windspeed is also categorised according to the **Beaufort scale**

Beaufort scale	Descriptive term	Average speed at 10 metres above ground
0	Calm	Less than 1 knot
1–3	Light	1 to 10 knots
4	Moderate	11 to 16 knots
5	Fresh	17 to 21 knots

Notation A **knot** (kn) is a 'nautical mile per hour'. 1 kn = 1.15 mph.

- **Daily maximum gust** in knots – this is the highest instantaneous windspeed recorded. The direction from which the maximum gust was blowing is also recorded
- **Daily maximum relative humidity**, given as a percentage of air saturation with water vapour. Relative humidities above 95% give rise to misty and foggy conditions

Watch out For the overseas locations, the only data recorded are:

- Daily mean temperature
- Daily total rainfall
- Daily mean pressure
- Daily mean windspeed

- **Daily mean cloud cover** measured in 'oktas' or eighths of the sky covered by cloud
- **Daily mean visibility** measured in decametres (Dm). This is the greatest horizontal distance at which an object can be seen in daylight
- **Daily mean pressure** measured in hectopascals (hPa)

Any missing data values are indicated in the large data set as n/a or 'not available'.

Data from Hurn for the first days of June 1987 is shown to the right.

You are expected to be able to take a sample from the large data set, identify different types of data and calculate statistics from the data.

- **If you need to do calculations on the large data set in your exam, the relevant extract from the data set will be provided.**
- **You need to be familiar with the types and ranges of data in the large data set, and with the characteristics of each location. You may need to recall trends from within the data set, or identify a location based on given data.**

HURN						
Date	Daily mean temperature (°C)	Daily total rainfall (mm)	Daily total sunshine (hrs)	Daily mean windspeed (kn)	Daily mean windspeed (Beaufort conversion)	Daily maximum gust (kn)
01/6/1987	15.1	0.6	4.5	7	Light	19
02/6/1987	12.5	4.7	0	7	Light	22
03/6/1987	13.8	tr	5.6	11	Moderate	25
04/6/1987	15.5	5.3	7.8	7	Light	17
05/6/1987	13.1	19.0	0.5	10	Light	33
06/6/1987	13.8	0	8.9	19	Fresh	46
07/6/1987	13.2	tr	3.8	11	Moderate	27
08/6/1987	12.9	1	1.7	9	Light	19
09/6/1987	11.2	tr	5.4	6	Light	19
10/6/1987	9.2	1.3	9.7	4	Light	n/a
11/6/1987	12.6	0	12.5	6	Light	18
12/6/1987	10.4	0	11.9	5	Light	n/a
13/6/1987	9.6	0	8.6	5	Light	15
14/6/1987	10.2	0	13.1	5	Light	18
15/6/1987	9.2	3.7	7.1	4	Light	25
16/6/1987	10.4	5.6	8.3	6	Light	25
17/6/1987	12.8	0.1	5.3	10	Light	27
18/6/1987	13.0	7.4	3.2	9	Light	24
19/6/1987	14.0	tr	0.4	12	Moderate	33
20/6/1987	12.6	0	7.7	6	Light	17