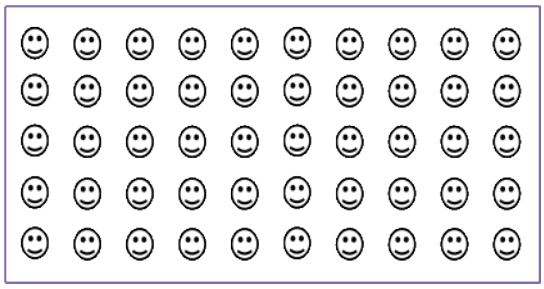
Chapter 1 - Data Collection

Personal notes:

1.1 - Populations and Samples

Key Terms



- A *population* is
- A *sample* is
- A *sampling unit* is
- A *sampling frame*

Notes

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

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

1.2, 1.3 - Sampling & Non-random Sampling

- There are two types of sampling: Random and Non-Random.
- Each type of sampling has a few methods.

1.2 - Random Sampling

A *random sampling* is such that each thing in the sampling frame to have an equal chance of being chosen, in order to **avoid bias**.

Simple Random Sampling (Random)

Edexcel S3 June 2004 Q1a

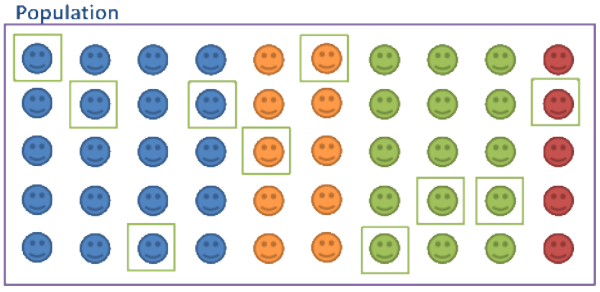
There are 64 girls and 56 boys in a school. Explain briefly how you could take a random sample of 15 pupils using a simple random sample. **(3)**

Systematic Sampling (Random)

Edexcel S3 June 2009 Q1a

A telephone directory contains 50 000 names. A researcher wishes to select a systematic sample of 100 names from the directory. Explain in detail how the researcher should obtain such a sample. **(2)**

Stratified Sampling (Random)



For example, we want to sample 20% of the population (shown as the diagram). If the population were divided into distinct group (4 colours = 4 groups, say age), known as "**strata**".
We could randomly sample 20% from each group, ensuring each group is equally represented.

Edexcel S3 Jan 2006 Q1

A school has 15 classes and a sixth form. In each class there are 30 students. In the sixth form there are 150 students. There are equal numbers of boys and girls in each class. There are equal numbers of boys and girls in the sixth form. The head teacher wishes to obtain the opinions of the students about school uniforms. Explain how the head teacher would take a stratified sample of size 40. **(7)**

1.3 - Non-Random Sampling

Consider this scenario:
You are going to conduct a survey on whether ability in mathematics correlates the person's innate art talent.
We need to choose some mathematicians to assess.

<u>Why would random sampling be problematic?</u>

Solution: For this scenario, _____ should be deployed.

1. As with stratified sampling, divide population into groups according to characteristic of interest, then determine size of each group in sample to reflect proportions within the population.
2. But instead of random sampling within each group, we actively choose people within each group via suitable means (e.g. advertising), until the 'quota' for each group is filled.

Another type of non-random sampling is _____ (aka convenience sampling).

- It 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.
- E.g. Survey about relationship between age and likeliness to use re-usable bags
  -> Could collect data from the first 20 people coming out from a supermarket

<u>Example</u>

Edexcel S3 June 2010 Q2

A lake contains 3 species of fish. There are estimated to be 1400 trout, 600 bass and 450 pike in the lake. A survey of the health of the fish in the lake is carried out and a sample of 30 fish is chosen.

(a) Give a reason why stratified random sampling cannot be used. **(1)**
(b) State an appropriate sampling method for the survey. **(1)**
(c) Give one advantage and one disadvantage of this sampling method. **(2)**
(d) Explain how this sampling method could be used to select the sample of 30 fish. You must show your working. **(4)**

1.4 - Types of Data

Notes

- Data can be presented in a _____ for conciseness.

Shoe size, w	Frequency
$2 \leq w < 5$	3
$6 \leq w < 11$	4

$6 \leq w < 11$

1.5 - The large data set

Starter: Suggest a suitable sampling method for each of the following scenarios.

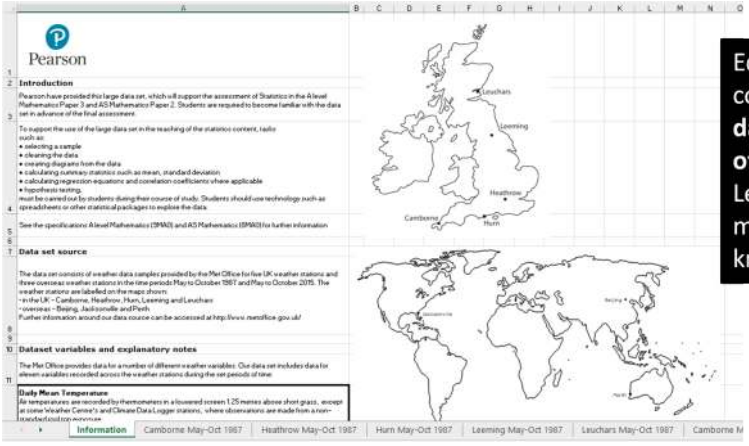
Simple Random Sampling	Systematic Sampling	Stratified Sampling	Quota Sampling	Opportunity Sampling

"You wish to test lightbulbs produced by a factory in a daily batch."

"You wish to survey students' opinion on Mr. Fan's new drawing *Geometry*."

"You wish to determine Londoners' favourite TV programmes.

<u>Basic Information</u>

- All A Level exam boards are obligated to provide a 'large data set'. Data in exam questions will often be from this set, and you are encouraged to explore this data (which is publicly available) in Microsoft Excel.
- **It is important to note that you are expected to be familiar with this data set before you go into your exam, including some basic geographic knowledge**
- 

Edexcel's data set concerns **weather data from a number of weather stations.** Let's explore what you might be expected to know...

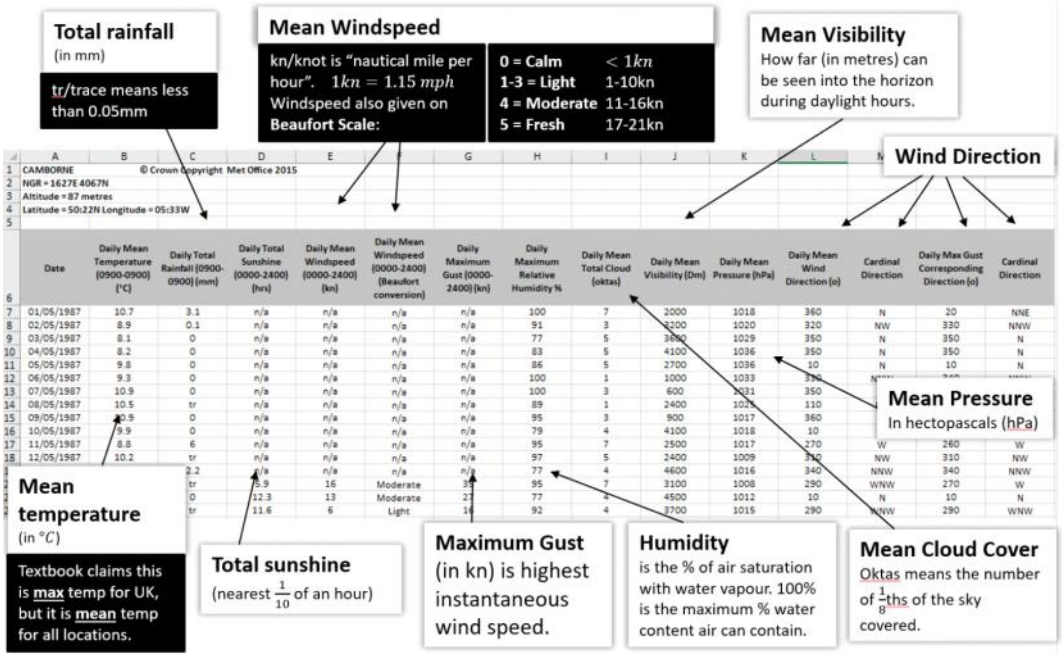
1.5 - The large data set

<u>What you'll need to be familiar with</u>

1. Names and rough locations of the 5 UK weather stations, as well as the 3 international weather stations.

The data was recorded for:
- May to Oct 1987
- May to Oct 2015

2. Variables involved and their respective units.



3. A vague idea of the range of values for each location.
   Example: Mean wind speed in UK across full period was roughly 9 nm. But 4 nm in Beijing (i.e. lower), 5 in Jacksonville (again lower), 8 in Perth (similar to UK).

4. A vague idea of the range of values of each variable.

Variable	Typical value(s)
Gust (UK only)	8 – 52 nm
Rainfall	0 – 60 mm in UK, but more extreme maximums elsewhere (e.g. 102mm in Perth)
Pressure	988 – 1038 hPa
Wind Speed on Beaufort scale	Max is 'fresh' (5). Most Light or Moderate.
Sunshine (UK only)	0 – 16 hrs
Cloud Cover	0 – 8 ocktas (i.e. full spread)

**Note:
- 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 will need to recall trends from within the data set, or identify a location based on the given data.

1.5 - The large data set

Example

Answer the following questions using the extract provided from the large data set

a) Describe the type of data represented by daily total rainfall.

Alison is investigating daily maximum gust. She wants to select a sample of size 5 from the first 20 days in Hurn in June 1987. She uses the first two digits of the date as a sampling frame and generates five random numbers between 1 and 20

b) State the type of sample selected by Alison.

c) Explain why Alison's process might not generate a sample of size 5

HURN	© Crown Copyright Met Office 1987					
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

Example

Calculate (using the same extract):

a) The mean daily maximum temperature for the first five days of June in Hurn in 1987.

b) The median daily total rainfall for the week of 14th June to 20th June inclusive.

c) The median daily total rainfall for the same week in Perth was 19.00mm. Karl states that more southerly countries experience higher rainfall during June. State with a reason whether your answer to part (b) supports this statement.