KING EDWARD VI
HANDSWORTH GRAMMAR
SCHOOL FOR BOYS

KING EDWARD VI
ACADEMY TRUST
BIRMINGHAM

Year 12
Statistics
1 Data Collection
Booklet

HGS Maths

Dr Frost Course

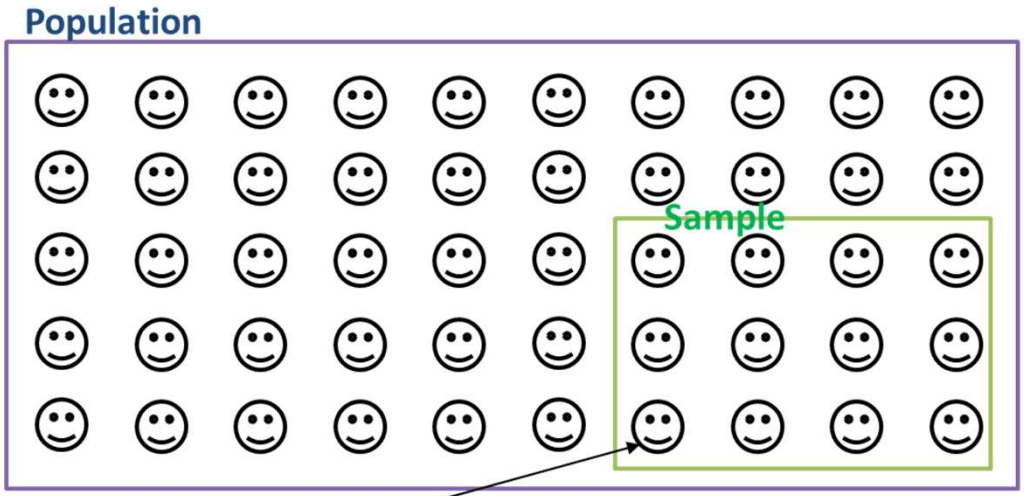
Name: _____

Class: _____

Contents

1.1 Populations and Samples

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



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**.

Notes

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.

Notes

Worked Example

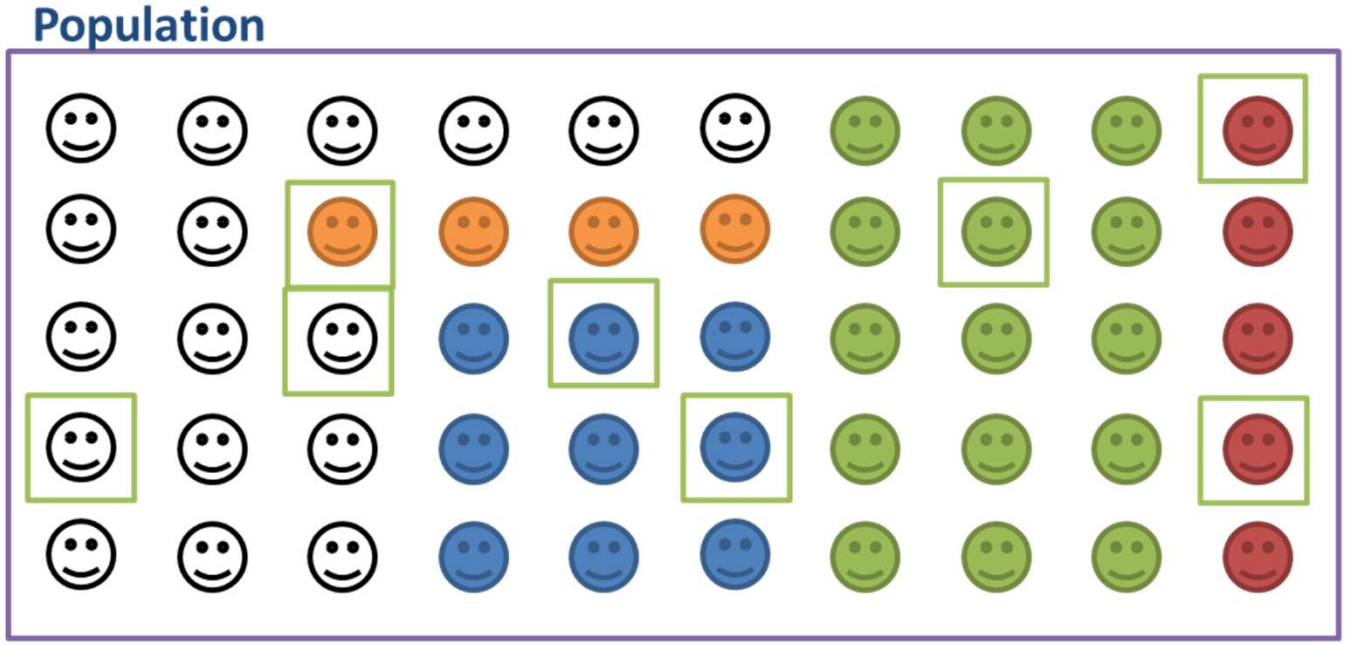
A supermarket wants to test a delivery of avocados for ripeness by cutting them in half.

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.

T: p.3 all Qs, P: p.1 all Qs

1.2 Sampling



Ordinarily, we would want each thing in our sampling frame to have an **equal chance of being chosen**, in order to **avoid bias**.

This is known as **random sampling**.
There are a few ways of doing this...

Notes

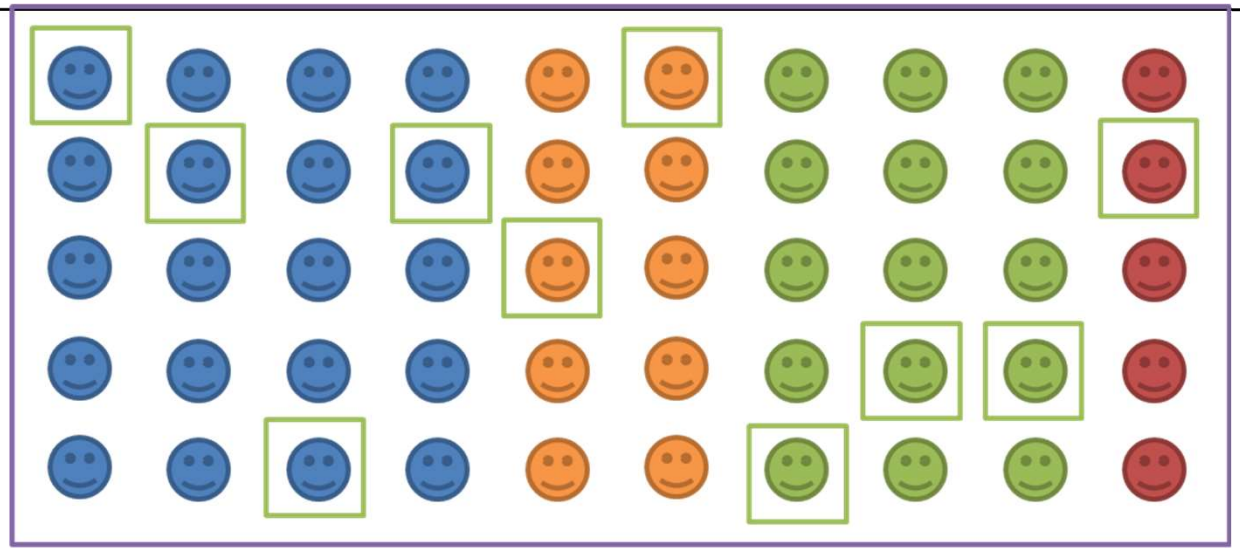
Type	How to carry out	Advantages	Disadvantages
Simple Random Sampling	**What is it :** Every sample has an equal chance of being selected. **Method:** In sampling frame each item has identifying number. Use random number generator, or 'lottery sampling' (names in a hat).	• Bias free. • Easy and cheap to implement. • Each number has a known equal chance of being selected.	• Not suitable when population size is large. • Sampling frame needed.

Notes

Type	How to carry out	Advantages	Disadvantages
Systematic Sampling	**What is it :** Required elements are chosen at regular intervals in ordered list. i.e. Take every kth elements where: $k = \frac{pop\ size\ (N)}{samp\ size\ (n)}$ starting at random item between 1 and k.	• Simple and quick to use. • Suitable for large samples/ populations.	• Sampling frame again needed. • Can introduce bias if sampling frame not random.

Notes

We want to sample 20% of the population. If the population were divided into distinct groups (e.g. age ranges), known as '**strata**', we could randomly sample 20% from each group, ensuring each group is equally represented.



Type	How to carry out	Advantages	Disadvantages
Stratified Sampling	**What is it :** Population divided into groups (strata) and a <u>simple random sample carried out in each group</u>. Same proportion $\frac{samp\ size\ (n)}{pop\ size\ (N)}$ sampled from each strata. Used when sample is large and population naturally divides into groups.	• Reflects population structure. • Guarantees proportional representation of groups within population.	• Population must be clearly classified into distinct strata. • Selection within each stratum suffers from same disadvantages as simple random sampling.

Worked Example

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.

Worked Example

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 using lottery sampling.

Worked Example

A telephone directory contains 50000 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.

Worked Example

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.

Worked Example

A company wants to survey the opinions of workers.
The manager decides to give a questionnaire to a sample of 80 workers.
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.
Explain how the manager could obtain a stratified sample of worker opinions.

T: p.6 all Qs, P: p.2 all Qs + DFM Key skill 531

1.3 Non-Random Sampling

Consider the following scenario: You wish to conduct a survey in the UK **on whether being left-handed affects IQ**. We need to choose people to assess.

Why would random sampling be problematic?
Because <u>we don't know the sampling frame</u>, i.e. don't have a list of all left-handed (and non-left-handed) people in the UK.

For this we'd likely use **quota sampling**, i.e.

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**.

A variant of this is **<u>opportunity sampling</u>**, where we find people **<u>at the same time the survey is being carried out</u>** (e.g. exit polls at polling stations). This is not a suitable method for the left-handed example, because giving the likely time-consuming nature of assessment coupled with resources required, we'd likely arrange with the people taking part before the actual assessment tasks took place.

Notes

Type	How to carry out	Advantages	Disadvantages
Quota Sampling	**What is it :** Population divided into groups according to characteristic. A quota of items/people in each group is set to try and reflect the group's proportion in the whole population. Interviewer selects the actual sampling units.	• Allows small sample to still be representative of population. • No sampling frame required. • Quick, easy, inexpensive. • Allows for easy comparison between different groups in 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, adding time/expense. • Non-responses are not recorded.
Opportunity/ Convenience Sampling	Sample taken from people who are available at time of study, who meet criteria.	• Easy to carry out. • Inexpensive.	• Unlikely to provide a representative sample. • Highly dependent on individual researcher.

Notes

Worked Example

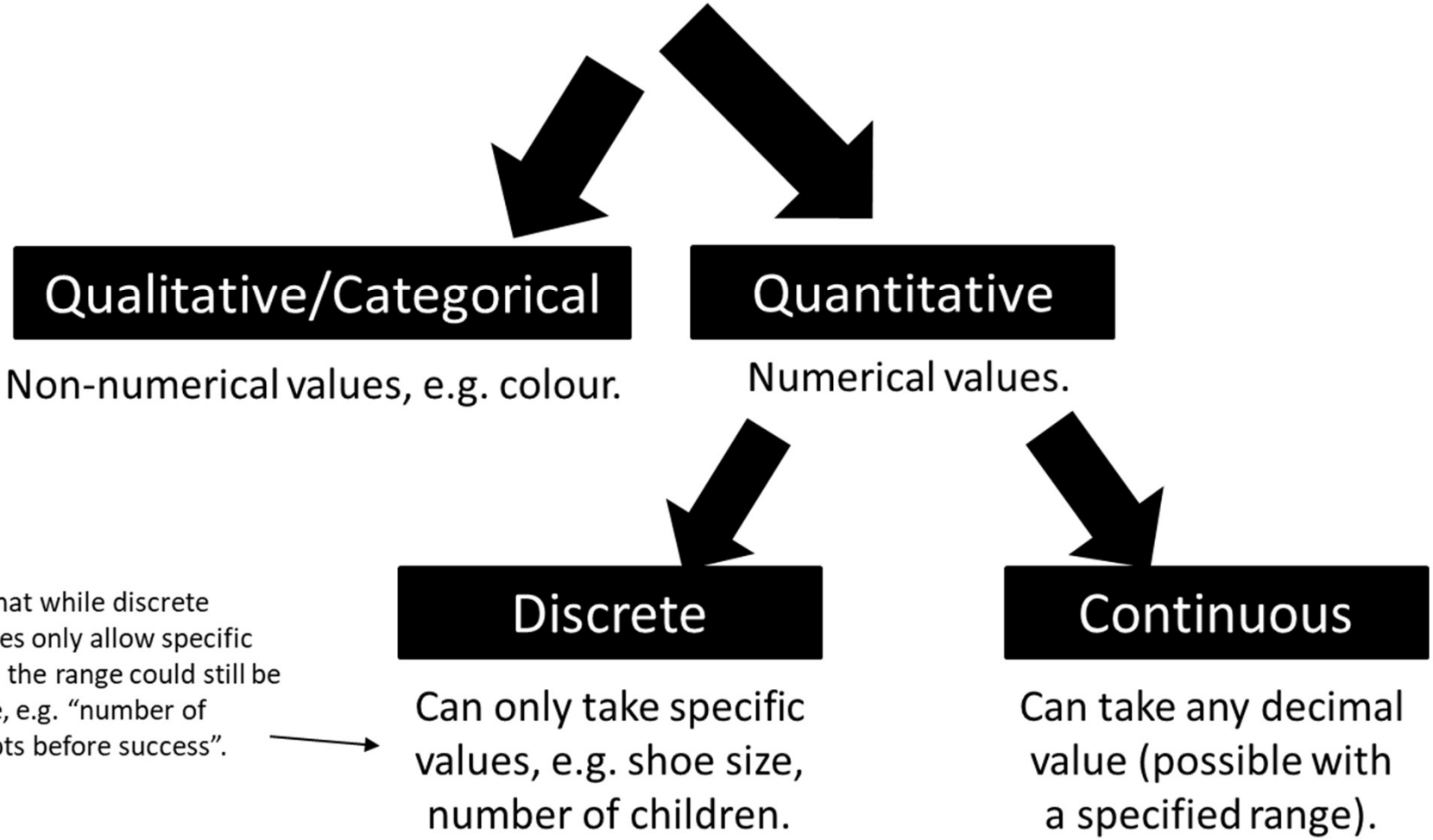
Explain how you would use opportunity sampling to survey 50 supermarket shoppers.

Worked Example

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 chosen.
Explain how you would use an appropriate sampling method.

T: p.8 all Qs, P: p.3 all Qs

1.4 Types of Data



Notes

Worked Example

State the type of data:

a) Human shoe size measured as 1, 2 or 3 etc.

b) Height of a tree

c) Favourite colour

Worked Example

The lengths, x mm, to the nearest mm, of the forewings of a random sample of male adult butterflies are measures and shown in the table.

Length of forewing (mm)	Number of butterflies, f
$30 - 31$	2
$32 - 33$	25
$34 - 36$	30
$37 - 39$	13

a) State whether length is
   i) quantitative or qualitative
   ii) discrete or continuous

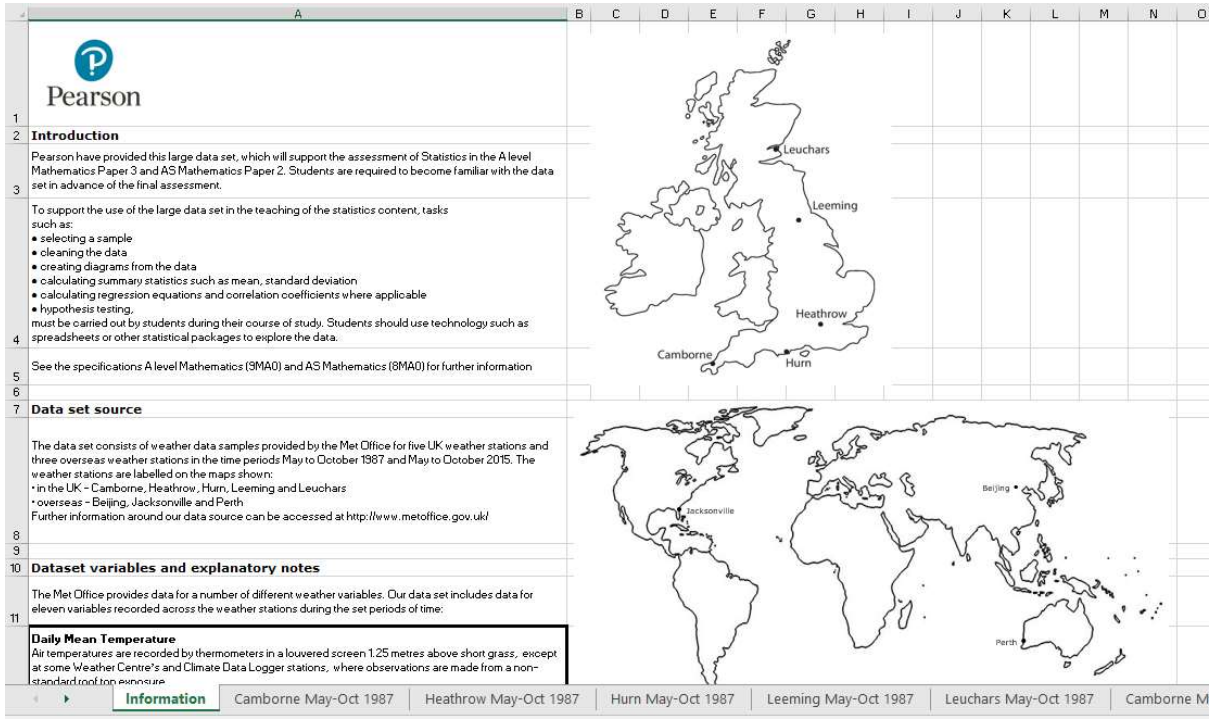
b) Write down the class boundaries, midpoint and class width for the class $34 - 36$.

T: p.10 all Qs, P: p.4 all Qs

1.5 The Large Data Set

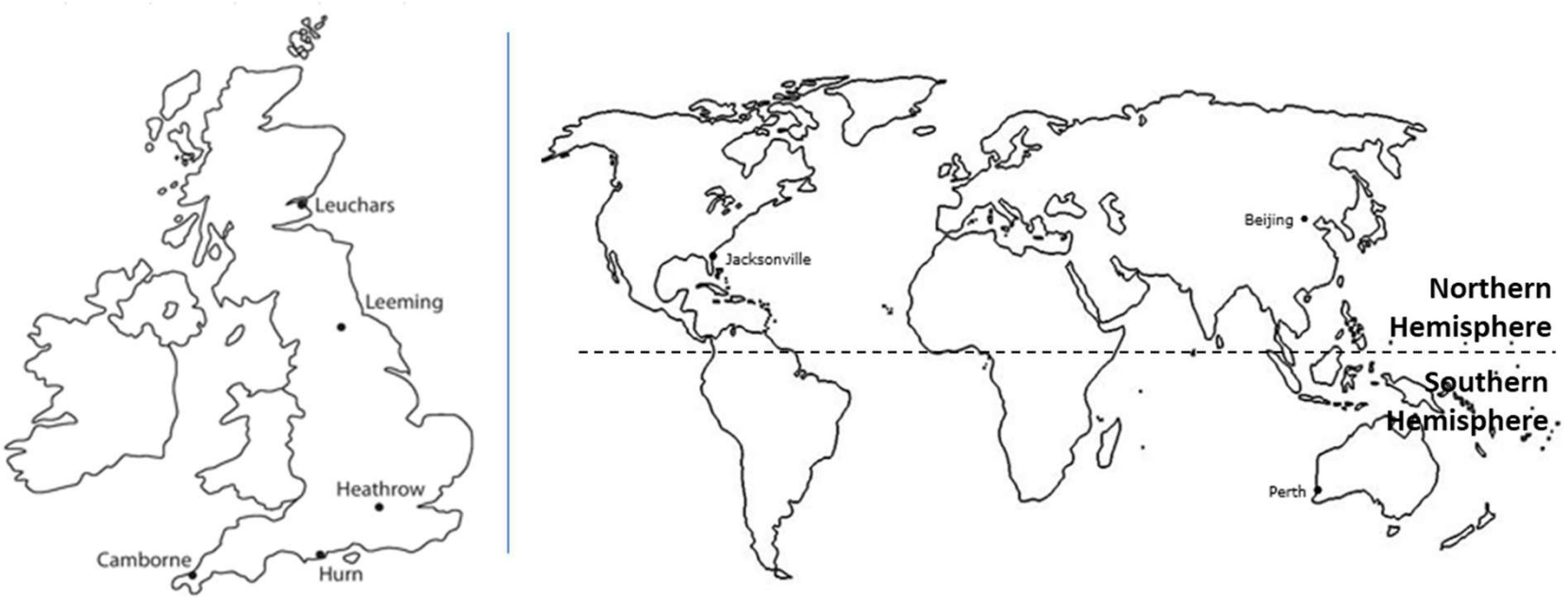
All A Level exam boards are obligated to provide a 'large data set'. Data in exam questions will often be from this set, and you are encouraged to explore this data (which is publicly available) in Microsoft Excel.

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…

Notes



1 You should know the names and rough locations of the 5 UK weather stations, as well as the 3 international weather stations.

The data was recorded for:

- May-Oct 1987
- May-Oct 2015

All the following are daily…

2 You should be familiar with the variables involved and their respective units.

Total rainfall
(in mm)

tr/trace means less than 0.05mm

Mean Windspeed

kn/knot is "nautical mile per hour". $1kn = 1.15\ mph$
Windspeed also given on **Beaufort Scale**:

0 = Calm	$< 1kn$
1-3 = Light	1-10kn
4 = Moderate	11-16kn
5 = Fresh	17-21kn

Mean Visibility
How far (in metres) can be seen into the horizon during daylight hours.

Wind Direction

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CAMBORNE		© Crown Copyright Met Office 2015										
2	NGR = 1627E 4067N												
3	Altitude = 87 metres												
4	Latitude = 50:22N Longitude = 05:33W												
5													
6	Date	Daily Mean Temperature (0900-0900) (°C)	Daily Total Rainfall (0900-0900) (mm)	Daily Total Sunshine (0000-2400) (hrs)	Daily Mean Windspeed (0000-2400) (kn)	Daily Mean Windspeed (0000-2400) (Beaufort conversion)	Daily Maximum Gust (0000-2400) (kn)	Daily Maximum Relative Humidity %	Daily Mean Total Cloud (oktas)	Daily Mean Visibility (Dm)	Daily Mean Pressure (hPa)	Daily Mean Wind Direction (o)	Cardinal Direction	Daily Max Gust Corresponding Direction (o)	Cardinal Direction
7	01/05/1987	10.7	3.1	n/a	n/a	n/a	n/a	100	7	2000	1018	360	N	20	NNE
8	02/05/1987	8.9	0.1	n/a	n/a	n/a	n/a	91	3	3200	1020	320	NW	330	NNW
9	03/05/1987	8.1	0	n/a	n/a	n/a	n/a	77	5	3650	1029	350	N	350	N
10	04/05/1987	8.2	0	n/a	n/a	n/a	n/a	83	5	4100	1036	350	N	350	N
11	05/05/1987	9.8	0	n/a	n/a	n/a	n/a	86	5	2700	1036	10	N	10	N
12	06/05/1987	9.3	0	n/a	n/a	n/a	n/a	100	1	1000	1033	330	NNW	340	NNW
13	07/05/1987	10.9	0	n/a	n/a	n/a	n/a	100	3	600	1031	350			
14	08/05/1987	10.5	tr	n/a	n/a	n/a	n/a	89	1	2400	1025	110			
15	09/05/1987	10.9	0	n/a	n/a	n/a	n/a	95	3	900	1017	360			
16	10/05/1987	9.9	0	n/a	n/a	n/a	n/a	79	4	4100	1018	10			
17	11/05/1987	8.8	6	n/a	n/a	n/a	n/a	95	7	2500	1017	270	W	260	W
18	12/05/1987	10.2	tr	n/a	n/a	n/a	n/a	97	5	2400	1009	310	NW	310	NW
19			2.2	n/a	n/a	n/a	n/a	77	4	4600	1016	340	NNW	340	NNW
20			tr	5.9	16	Moderate	35	95	7	3100	1008	290	WNW	270	W
21			0	12.3	13	Moderate	27	77	4	4500	1012	10	N	10	N
22			tr	11.6	6	Light	16	92	4	3700	1015	290	WNW	290	WNW

Mean Pressure
In hectopascals (hPa)

Mean temperature
(in $°C$)

Textbook claims this is **max** temp for UK, but it is **mean** temp for all locations.

Total sunshine
(nearest $\frac{1}{10}$ of an hour)

Maximum Gust
(in kn) is highest instantaneous wind speed.

Humidity
is the % of air saturation with water vapour. 100% is the maximum % water content air can contain.

Mean Cloud Cover
Oktas means the number of $\frac{1}{8}$ths of the sky covered.

3 You should have a vague idea of the range of values for each location.

UK Location (2015)	Temp Range	Wind Speed Range
Camborne	10-20	3-18
Heathrow	8-**29**	3-19
Hurn	6-24	2-19
Leeming	4-23	3-17
Leuchars	4-19	3-23

World Location (2015)	Temp Range	Wind Speed Range
Beijing	8-33	2-9
Jacksonville	15-31	1-12
Perth	8-25	4-14

Beijing temp range relatively large.
Min Jacksonville temp high.
Perth similar to UK.

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).

From new A Level sample assessment materials:

"A meteorologist believes that there is a relationship between the daily mean windspeed, w kn, and the daily mean temperature, t °C. A random sample of 9 consecutive days is taken from past records from a town in the UK in July and the relevant data is given in the table below. …

Using the same 9 days, a location from the large data set gave $\bar{t} = \mathbf{27.2}$ and $\bar{w} = \mathbf{3.5}$.

(d) Using your knowledge of the large data set, suggest, giving your reason, the location that gave rise to these statistics."

4 You should have a vague idea of the range of values for each variable for the data set as a whole.

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)

Notes

Worked Example

Suggest a suitable sampling method:

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

b) You wish to survey consumer opinion on a new product your company have released.

c) You wish to determine students' favourite TV programmes in your school. That is fairly representative of each year group.

Worked Example

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

Worked Example

Calculate:

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.

Date	Daily Max Temp (09-00-0900 C)	Daily Total Rainfall (0900-0900) (mm)	Daily Total Sunshine (0000-2400) (hrs)	Daily Mean Windspeed (0000-2400) (kn)	Daily Mean Windspeed (0000-2400) (Beaufort conversion)	Daily Maximum Gust (0000-2400) (kn)
01/06/1987	15.1	0.6	4.5	7	Light	19
02/06/1987	12.5	4.7	0	7	Light	22
03/06/1987	13.8	tr	5.6	11	Moderate	25
04/06/1987	15.5	5.3	7.8	7	Light	17
05/06/1987	13.1	19	0.5	10	Light	33
06/06/1987	13.8	0	8.9	19	Fresh	46
07/06/1987	13.2	tr	3.8	11	Moderate	27
08/06/1987	12.9	1	1.7	9	Light	19
09/06/1987	11.2	tr	5.4	6	Light	19
10/06/1987	9.2	1.3	9.7	4	Light	n/a
11/06/1987	12.6	0	12.5	6	Light	18
12/06/1987	10.4	0	11.9	5	Light	n/a
13/06/1987	9.6	0	8.6	5	Light	15
14/06/1987	10.2	0	13.1	5	Light	18
15/06/1987	9.2	3.7	7.1	4	Light	25
16/06/1987	10.4	5.6	8.3	6	Light	25
17/06/1987	12.8	0.1	5.3	10	Light	27
18/06/1987	13.0	7.4	3.2	9	Light	24
19/06/1987	14.0	tr	0.4	12	Moderate	33
20/06/1987	12.6	0	7.7	6	Light	17

T: p.13 all Qs, P: p.5 all Qs

[EdExcel Statistics 2 June 2006]

1. Before introducing a new rule the secretary of a golf club decided to find out how members might react to this rule.

(a) Explain why the secretary decided to take a random sample of club members rather than ask all the members.

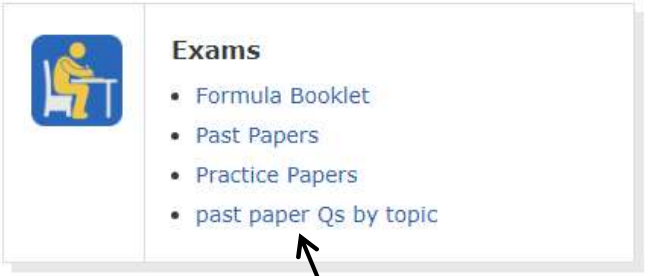
(1)

(b) Suggest a suitable sampling frame.

(1)

(c) Identify the sampling units.

(1)



Past paper practice by topic. Both new and old specification can be found via this link on hgsmaths.com

1.(a)	Saves time / cheaper / easier or A census/asking all members takes a long time or is expensive or difficult to carry out	any one	B1	(1)
(b)	List, register or database of all club members/golfers or Full membership list		B1	(1)
(c)	Club member(s)		B1	(1)

Summary of Key Points

1 • In statistics, a **population** is the whole set of items that are of interest.
- A **census** observes or measures every member of a population.

2 • 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.
- 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**.

3 • A **simple random sample** of size n is one where every sample of size n has an equal chance of being selected.
- In **systematic sampling**, the required elements are chosen at regular intervals from an ordered list.
- In **stratified sampling**, the population is divided into mutually exclusive strata (males and females, for example) and a random sample is taken from each.
- In **quota sampling**, an interviewer or researcher selects a sample that reflects the characteristics of the whole population.
- **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.

4 • Variables or data associated with numerical observations are called **quantitative variables** or **quantitative data**.
- Variables or data associated with non-numerical observations are called **qualitative variables** or **qualitative data**.

5 • A variable that can take any value in a given range is a **continuous variable**.
- A variable that can take only specific values in a given range is a **discrete variable**.

6 • When data is presented in a grouped frequency table, the specific data values are not shown. The groups are more commonly known as **classes**.
- Class boundaries tell you the maximum and minimum values that belong in each class.
- The midpoint is the average of the class boundaries.
- The class width is the difference between the upper and lower class boundaries.

7 If you need to do calculations on the large data set in your exam, the relevant extract from the data set will be provided.

8 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.

T: p.1 mixed ex, P: BSG p.6