Stats1 Chapter 1 :: Data Collection

jfrost@tiffin.kingston.sch.uk

www.drfrostmaths.com

@DrFrostMaths

Last modified: 1st February 2018

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

Interestingly, most of this chapter is from the old S3 module (a Further Maths module!) with also some S2. There is little 'calculation' involved in this chapter; consider this a 'bookwork' one!

1:: 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.

4:: Types of data

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

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.

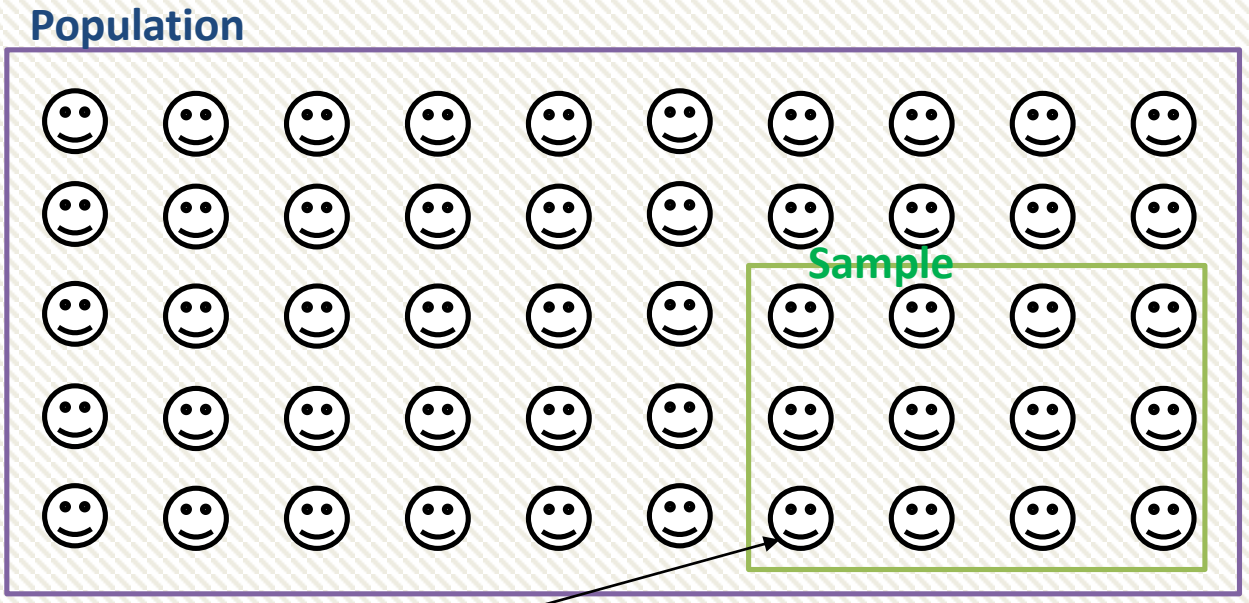
Populations and samples

A **population** is: ?

A **sample** is: ?

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 ?

Exercise 1A

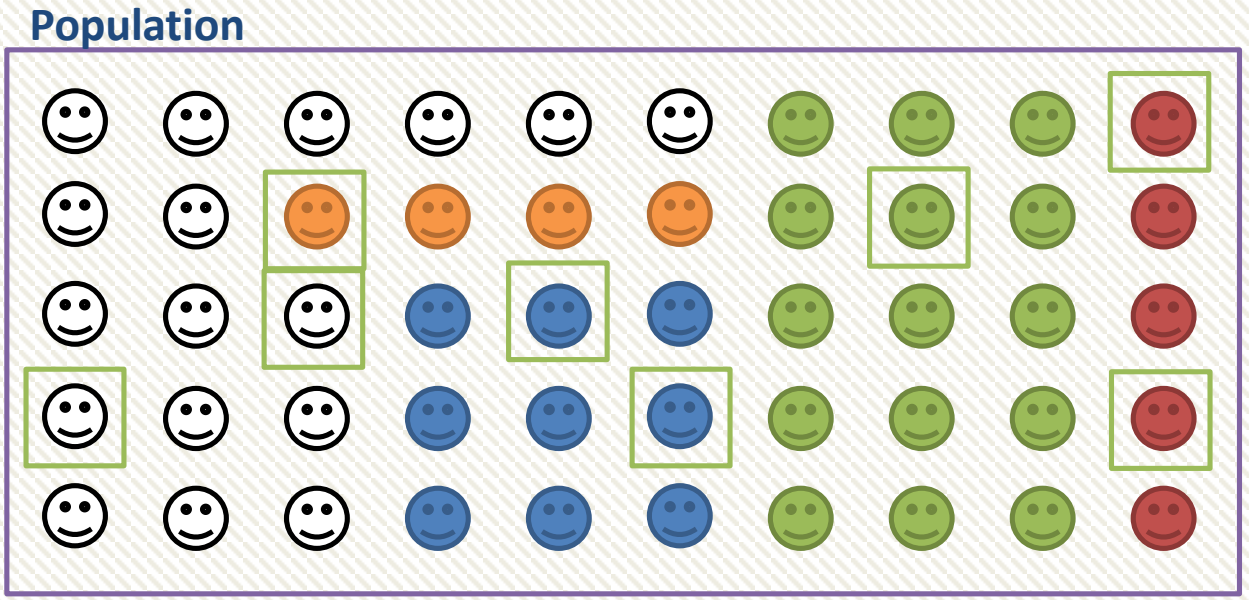
Pearson Statistics & Mechanics Year 1/AS
Page 3

Types of Sampling

I recommend laying out your notes like this for next bit of the chapter. Use a full page.

	Type	How to carry out	Advantages	Disadvantages
Random Sampling	Simple Random Sampling			
	Systematic Sampling			
	Stratified Sampling			
Non-Random	Quota Sampling			
	Opportunity Sampling			

Random 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…

Simple Random Sampling

Type	How to carry out	Advantages	Disadvantages
Simple Random Sampling	**What is it :** ? **Method:** ?	?	?

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

Type	How to carry out	Advantages	Disadvantages
Systematic Sampling	**What is it :** ?	?	?

Edexcel S3 June 2009 Q1a

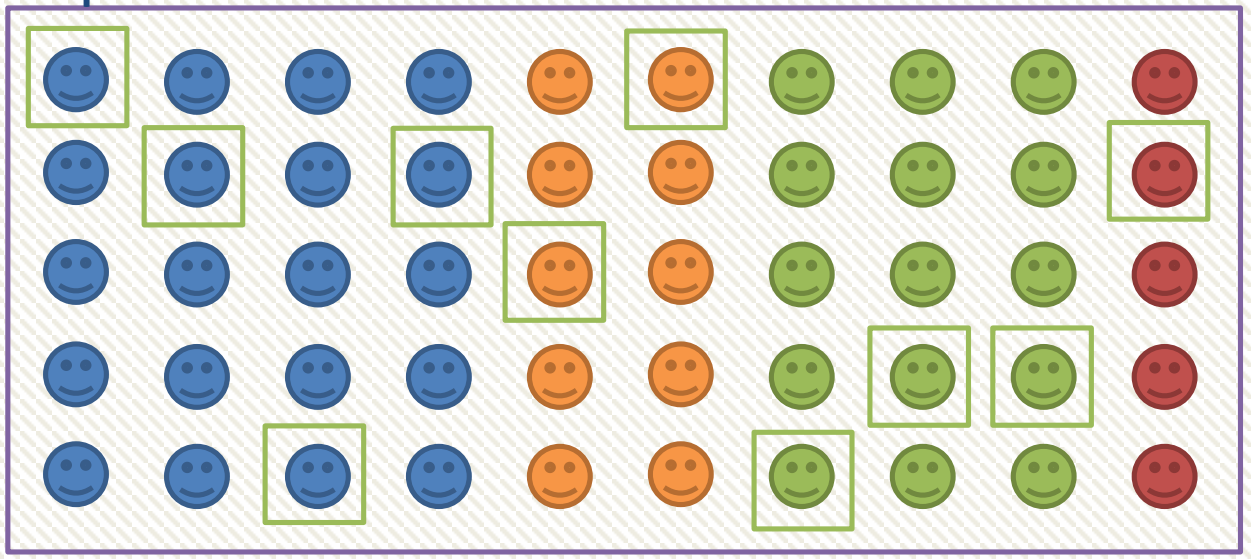
A telephone directory contains 50 000 names. A researcher wishes to select a systematic sample of 100 names from the directory. Explain in detail how the researcher should obtain such a sample. **(2)**

?

Stratified Sampling

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.

Population



Type	How to carry out	Advantages	Disadvantages
Stratified Sampling	**What is it :** ?	?	?

Example Question

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)**

?

You would certainly want to know your mark scheme on this one!

Exercise 1B

Pearson Statistics & Mechanics Year 1/AS
Pages 6-7

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?

?

Famous Lefties

OK, maybe not so famous.

For this scenario 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 **opportunity sampling**, where we find people **at the same time the survey is being carried out** (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.

Quota & Opportunity Sampling

Type	How to carry out	Advantages	Disadvantages
Quota Sampling	?	?	?
Opportunity/ Convenience Sampling	?	?	?

Example Question

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)**

(a) ?

(b) ?

(c) ?

(d) ?

Exercise 1C

Pearson Statistics & Mechanics Year 1/AS
Pages 8-9

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 ?

? ? ?

Midpoint = ?

Class width = ?

Exercise 1D

Pearson Statistics & Mechanics Year 1/AS
Page 10

(This exercise could probably be skipped)

Name That Sampling Method!

Simple Random Sampling	Systematic Sampling	Stratified Sampling	Quota Sampling	Opportunity Sampling

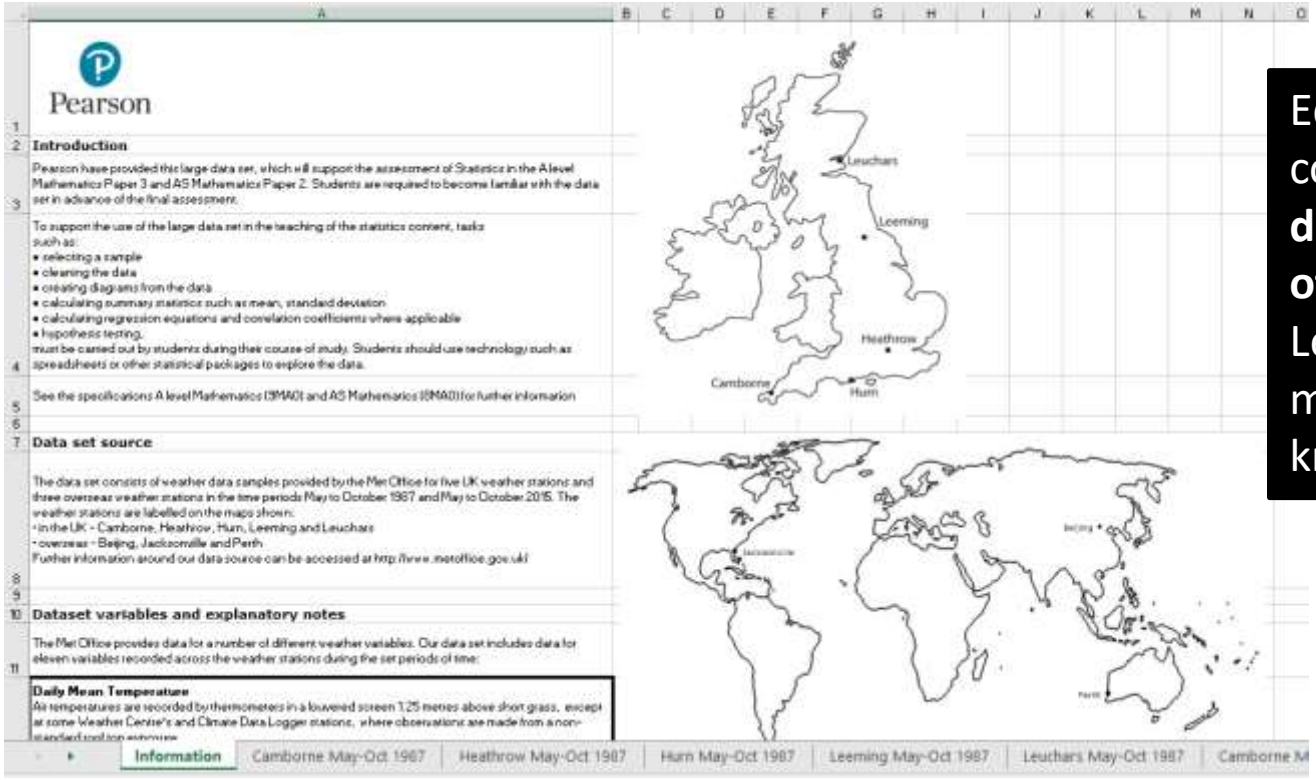
Suggest a suitable sampling method.

"You wish to test lightbulbs produced by a factory in a daily batch."	?
"You wish to survey consumer opinion on your new drink *FizzGuzz* released in the UK."	?
"You wish to determine students' favourite TV programmes in your school, that is fairly representative of each year group."	?

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.

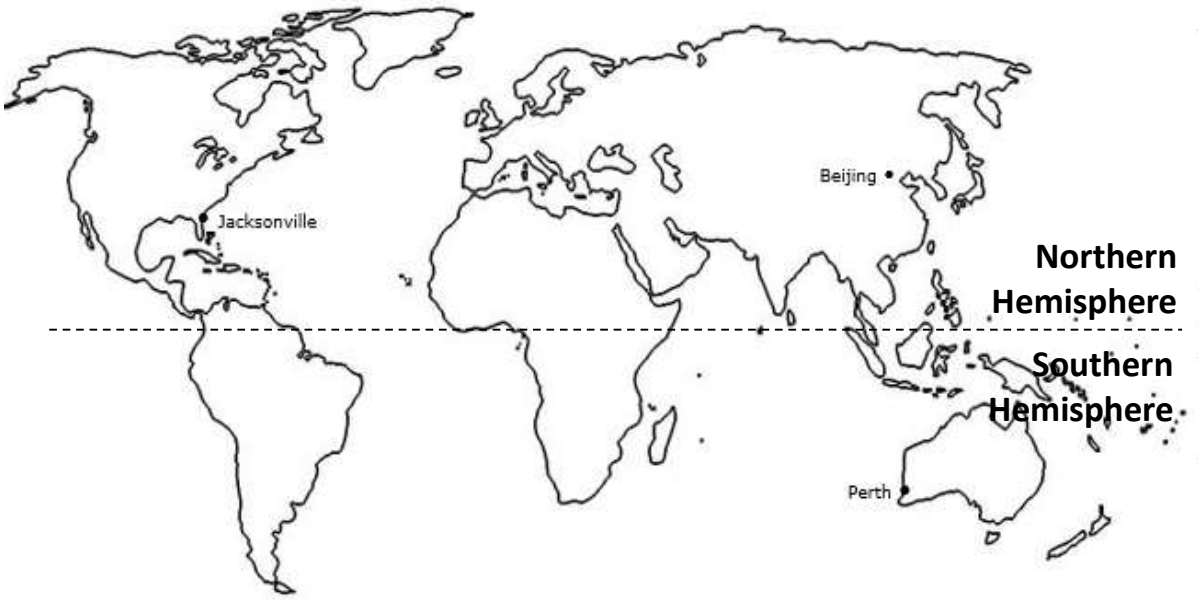
It is important to note that you are expected to be familiar with this data set before you go into your exam, including some basic geographic knowledge!



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

https://qualifications.pearson.com/content/dam/pdf/A%20Level/Mathematics/2017/specification-and-sample-assesment/Pearson%20Edexcel%20GCE%20AS%20and%20AL%20Mathematics%20data%20set%20-%20Issue%201%20(1).xls

What You Need To Be Familiar With…



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

CAMBORNE © Crown Copyright Met Office 2015
NGR = 1627E 4067N
Altitude = 87 metres
Latitude = 50:22N Longitude = 05:33W

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
01/05/1987	10.7	3.1	n/a	n/a	n/a	n/a	100	7	2000	1018	360	N	20	NNE
02/05/1987	8.9	0.1	n/a	n/a	n/a	n/a	91	3	3200	1020	320	NW	330	NNW
03/05/1987	8.1	0	n/a	n/a	n/a	n/a	77	5	3600	1029	350	N	350	N
04/05/1987	8.2	0	n/a	n/a	n/a	n/a	83	5	4100	1036	350	N	350	N
05/05/1987	9.8	0	n/a	n/a	n/a	n/a	86	5	2700	1036	10	N	10	N
06/05/1987	9.3	0	n/a	n/a	n/a	n/a	100	1	1000	1033	330	NNW		
07/05/1987	10.9	0	n/a	n/a	n/a	n/a	100	3	600	1031	350			
08/05/1987	10.5	tr	n/a	n/a	n/a	n/a	89	1	2400	1025	110			
09/05/1987	10.9	0	n/a	n/a	n/a	n/a	95	3	900	1017	360			
10/05/1987	9.9	0	n/a	n/a	n/a	n/a	79	4	4100	1018	10			
11/05/1987	8.8	6	n/a	n/a	n/a	n/a	95	7	2500	1017	270	W	260	W
12/05/1987	10.2	tr	n/a	n/a	n/a	n/a	97	5	2400	1009	310	NW	310	NW
		2.2	n/a	n/a	n/a	n/a	77	4	4600	1016	340	NNW	340	NNW
		tr	5.9	16	Moderate	35	95	7	3100	1008	290	WNW	270	W
		0	12.3	13	Moderate	27	77	4	4500	1012	10	N	10	N
		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."

(Note to teachers: I will not otherwise use SAM questions in these slides. I made one exception here!)

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)

Example Questions

Hurn

© Crown Copyright Met Office 1987

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

As previously noted, the **actual** data set has mean temperature for <u>all</u> locations. I changed to maximum temperature for this example for consistency with the textbook.

[Textbook]

(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.

a ?

b ?

c ?

Example Questions

Hurn

© Crown Copyright Met Office 1987

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

[Textbook] 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.

a ?

b ?

c ?

Example Questions

Hurn

© Crown Copyright Met Office 1987

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

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.

a

b

c

Exercise 1E

Pearson Statistics & Mechanics Year 1/AS
Pages 13-15