



MAS1403

Quantitative Methods for Business Management

Semester 1, 2017–2018

Module leader: Dr. Lee Fawcett

Additional lecturers: Dr. Dave Walshaw and Dr. Ged Cowburn

Contact details

My name: Dr. Lee Fawcett

Office: Room 2.07 Herschel Building

Phone: 0191 208 7228

Email: `lee.fawcett@ncl.ac.uk`

www: `www.mas.ncl.ac.uk/~nlf8`

Classes

Weekly **lectures** take place on Mondays, at 5pm (sorry!), in the Curtis Auditorium in the Herschel Building.

Each student should attend one **tutorial** session later on in the week. Due to the size of the class, there are *three* sessions – A, B and C – all sessions take place on Thursdays:

Group	Day	Time	Room	Staff
A	Thurs	11–12	King George VI, LT1	Dr. Cowburn
B	Thurs	1–2	Armstrong, Spence Watson LT	Dr. Fawcett
C	Thurs	2–3	Armstrong, Spence Watson LT	Dr. Fawcett

All sessions are attendance-monitored!

Occasionally throughout the year, a **computer practical session** will take the place of the scheduled tutorial session.

This will happen in teaching weeks 2 and 7.

All sessions take place in the Herschel Building Computing Cluster on the first floor of the Herschel Building.

Group	Day	Time
A	Tues	10–11
B	Wed	10–11
C	Thurs	11–12

As well as seeing you for two hours a week in the **lecture** and the **tutorial/computer practical**, I have also timetabled two “drop-in” sessions each week.

Each Monday, right before the lecture, I will be available in my office to see students needing extra help with the work. This will be on a voluntary “drop-in” basis, from 4-5pm. I will also be available every Wednesday from 1-2pm.

Lecture notes and handouts

Lecture notes, together with solutions to tutorial exercises and any other handouts, will be posted on the course website. You should make the following link one of your "favourites":

www.mas.ncl.ac.uk/~nlf8

There will be a link to this website from the MAS1403 homepage in Blackboard.

You should print out and read (!) lecture notes before each lecture.

Some handouts will only be available to collect in class!

At the end of each lecture I will usually set you some work to do, from the lecture notes, before the tutorial.

The assessment for this module consists of a single two hour **exam** at the end of Semester 2 (next May/June), two written **assignments** (one in each Semester) and six **computer based assessments** (three in each Semester).

- **Exam (60%)**

The exam for MAS1403 will be “open–book”. This means you can take all your stuff in!

The exam is taken at a computer terminal, where you enter your responses on the screen. More information on the exam will be given later in the course – no need to think about this now!

- **Assignments (10% in each Semester)**

There are two assignments, one in each Semester. An announcement will be made in class when these are available to download. Assignment 1 is due in by *4pm, Thursday 14th December 2017*.

- **CBAs (10% in each Semester)**

There will be three computer based assessments in each Semester. You will be given a tutorial on how to use the system in the practical sessions next week.

- **Announcements**

Announcements relating to the course will be made via email. You should check your email at least twice a week!

- **Videos**

We are in the process of shooting some videos which will show detailed calculations relating to the more difficult parts of the course. These will be available to view from the course website.

- **Wine/beer/book questions**

Some weeks, in tutorials, I will set “prize questions”. You will have until the following lecture to submit solutions to these questions to me. These questions will be quite taxing, and as such will attract a prize for students submitting the best solutions.

- **Calculator**

All students should buy a **Casio fx-83** or a **Casio fx-115** calculator. Others might be allowed – you should check the University’s regulations (see course website).

False sense of security

The course starts *very easy*

- This week, we'll look at how data are collected
- Next week, we'll draw some graphs
- The week after, we'll look at simple summaries of data (mean, median, mode etc...)

That's when attendance usually drops off... students think the work is too easy, so why bother coming to lectures?

False sense of security

- You need the basics in order to get a handle on the more difficult stuff
- The course will get more difficult by about week 4 or 5
- By Christmas you will be tackling A–Level standard Statistics problems
- By Easter next year, you will be tackling degree–level Statistics problems!

Solution – regular attendance, regular testing!

Maths is too hard. I'll never pass, so what's the point?

- This is not a maths course. It's a Stats course – there is a difference!
- We start with stuff you did in Year 8 and build up gradually!
- Out of 454 students taking MAS1403 last year, only 5 students failed
- Regular lecture/tutorial attendance has a proven track record – even for those who are really scared of the subject! Practice makes perfect!
- Lots of support – and I don't think I'm scary!
- Out of 300 Marketing and Management students taking MAS1403 last year, 219 said this was their favourite module

Denial – Maths and Stats is unnecessary for me!

- Employers love graduates who are numerate
- You will need quite a few of the techniques this course provides in other courses and in your final year dissertation
- Real-life Accounting/Finance/Management/Marketing companies use the techniques you will learn in this course to help them make very big decisions – I know this for a fact
- We will use real-life data
- This course will make you a better person, and you will gain a massive sense of achievement when you pass next year!

Keep on top of the work!

Lecture 1

COLLECTING AND PRESENTING DATA

1. Collecting and presenting data

Data are the key to many important management decisions.

- Is a new product selling well?
- Do potential customers like the new advertising campaign?

These and many other questions can be answered with **data**.

In this lecture we will look at

- the **different kinds** of data that exist
- how we might **obtain** our data
- how we might **present** our data

Sizing clothes

- Most clothing now comes in essentially standard sizes...
- but where do these standards come from?
- By sampling from the population as a whole, standards can be set around the most common sizes
- We cannot say that an individual is *exactly* a standard size...
- however, we *can* say that they will probably fall within a range either side of a standard size

Car maintenance

- Just imagine you're buying a new car...
- it might be useful to know how much it will cost to run over the next few years
- Obviously this cannot be predicted exactly –
 - Each car will be slightly different
 - Each user will be slightly different
- Collecting data from people who have bought similar cars will give us some idea!

Some definitions...

The quantities measured in a study are called **random variables**.

A particular outcome is called an **observation**.

A collection of such observations is the **data**.

The collection of all possible outcomes is the **population**.

Example

Suppose we are interested in the height of students on Management courses at Newcastle...

- Our **random variable** is “the height of students on Management courses at Newcastle”
- If Joe Bloggs was a Management student, and we measured his height, then that value would be a single **observation**
- If we measured the height of every first year Management student, we would have a collection of such observations which would be our **data**
- This would be a **sample** from the **population** which consists of all students registered with the School of Management

Example

Ideally, to get a true idea of what is going on, we'd like to observe the whole population (take a **census**). However, this can be difficult:

- If the population is huge, then this would take ages!
- *And* it would be very costly!
- In reality, we usually observe a subset of the population... but how do we choose who to observe?

Types of data

There are two types of variables: **qualitative** and **quantitative**.

- *Qualitative* variables have non-numeric outcomes, and are usually **categorical**.
 - sex of a person (categories: male or female)
 - colour of a car (categories: red, black, silver, blue – could code this to 1, 2, 3 and 4)
 - mode of transport
- *Quantitative* variables have numeric outcomes with a natural ordering
 - people's height
 - number of defective components in a batch

Quantitative variables

We are most interested in **quantitative** variables.

These variables can be subdivided into two types:

discrete and **continuous**.

Discrete random variables

- are usually integers
- are *countable* – e.g. the number of red cars going past a junction in an hour
- can be **ordinal** – where the outcomes are ordered

Continuous random variables

- can take any value over some continuous scale – e.g. height or weight
- can be measured to a very high degree of accuracy (provided we have the equipment to do so)...
- ...however, we can never say *precisely* how much someone weighs, for example
- might be measured to the nearest whole number – and so could “look” discrete – be careful!

Example 1

- (a) Time between emails arriving in your inbox:
- (b) Favourite chocolate bar:
- (c) Number of students attending a tutorial:

Example 1

- (a) Time between emails arriving in your inbox: **Continuous**
- (b) Favourite chocolate bar: **Qualitative / Categorical**
- (c) Number of students attending a tutorial: **Discrete**

1.2 Sampling techniques

surveys or **questionnaires** are often used to gain insight into the impact of many management decisions – for example, customer views on the impact and potential of new technologies.

When preparing a survey, you should consider the following key questions:

- what is the **purpose** of the survey?
- what is the **target population**?
- is there a list of the target population?
- how can **bias** be avoided?
- how **accurate** does the survey have to be?
- what **resources** are available for conducting the survey?
- how are the data to be **collected**?

How can we collect survey information?

There are many ways of collecting survey information. Each has its advantages and disadvantages.

1. *Postal questionnaire* – low cost, low response rate, slow turn around time, low quality information
2. *Telephone interviewing* – moderate cost, moderate response rate, fast turn around time, good quality information (?)
3. *Face-to-face interviewing* – high cost, high response rate, fast turn around time, high quality information

We can rarely observe the whole population. Instead we observe some sub-set of this (called the **sample**).

The difficulty is in obtaining a **representative** sample.

For example, if you were to ask people leaving a gym if they took exercise this would produce a **biased** sample and would not be representative of the population as a whole.

The importance of obtaining a representative sample cannot be stressed too highly.

Sampling techniques

There are three general forms of sampling techniques:

1. **Random sampling** – where the members of the sample are chosen by some random mechanism
2. **Quasi-random sampling** – where the mechanism for choosing the sample is only partly random
3. **Non-random sampling** – where the sample is specifically selected rather than randomly selected

Simple random sampling...

This method is the **simplest to understand**.

If we had a population of 200 students we could put all their names into a hat and draw out 20 names as our sample.

Each name has an **equally likely chance** of being drawn and so the sample is completely random.

Furthermore each possible sample of 20 has an equal chance of being selected.

In reality the drawing of the names would be done by a computer and the population and samples would be considerably larger.

... and some disadvantages

We often don't have a **complete list** of the population.

For example, if you were surveying the market for some new software, the population would be everybody with a compatible computer!

Not all elements of the population are **equally accessible**.

Purely by chance, you could pick an **unrepresentative sample**!

Stratified sampling...

This is a form of random sample where clearly defined groups, or **strata**, exist within the population.

If we know the overall proportion of the population that falls into each of these groups, we can randomly sample from each of the groups and then adjust the results according to the known proportions.

For example, if we assume that the population is 55% female and 45% male and we wanted a sample of 1000, we would

- decide to have 550 females and 450 males in our sample, and then
- pick the members of our sample from their respective groups randomly

... and some disadvantages

We need clear information on the **size** and **composition** of each stratum which can be difficult to obtain.

We still need a **list of the entire population** to sample from.

Systematic sampling...

This is a form of **quasi-random sampling** which can be used when the population is clearly structured.

For example, if you were interested in obtaining a 10% sample from a batch of components being manufactured, you would

- select the first component at random, and then
- pick every 10th item to come off the production line after that

This scheme is **very easy to implement!**

... and some disadvantages

This method is **not entirely random!**

If there is a pattern in the process, it is very easy to obtain a **biased sample**.

Other methods of sampling

Judgemental sampling

- Here, the person interested in obtaining the data decides whom they are going to ask
- This scheme is **non-random**
- It is also prone to **bias**

Accessibility sampling

- Here, only the **most easily accessible** individuals are sampled
- This is obviously prone to **bias**
- It is, however, **convenient** and **cheap**

Quota sampling

- This method is similar to stratified sampling but uses judgemental sampling within groups
- The population is broken down into groups (could be age groups) and we sample within these groups until our **quota** is reached
- One benefit is the accuracy of the results obtained as a result of specific targeting.

Example 2

- (a) Answer: **Systematic sampling**
- (b) Answer: **Stratified sampling**, where the sample sizes for the iPhone/Samsung phones are in proportion to the population sizes (e.g. 10/20, or 5/10,...)

1.3 Frequency tables

Once we have collected our data, we need a convenient way to *view* it.

The simplest way is via a **frequency table**.

Example 3

The following table shows the raw data for car sales at a new car showroom over a two week period in July.

Date	Cars Sold	Date	Cars Sold
01/07/04	9	08/07/04	10
02/07/04	8	09/07/04	5
03/07/04	6	10/07/04	8
04/07/04	7	11/07/04	4
05/07/04	7	12/07/04	6
06/07/04	10	13/07/04	8
07/07/04	11	14/07/04	9

Example 3

Presenting these data in a relative frequency table by number of days on which different numbers of cars were sold, we get the following table:

Cars Sold	Tally	Frequency	Relative Frequency %
1		0	0
2		0	0
3		0	0
4	I	1	7.14
5	I	1	7.14
6	II	2	14.29
7	II	2	14.29
8	III	3	21.43
9	II	2	14.29
10	II	2	14.29
11	I	1	7.14
Totals	14	14	100

1.3 Frequency tables

With discrete data it is easy to count the quantities in the defined categories. With **continuous data** this is not possible.

Example 4

The following data set represents the service time in seconds for callers to a credit card call centre.

196.3	199.7	206.7	203.8	203.1
200.8	201.3	205.6	181.6	201.7
180.2	193.3	188.2	199.9	204.7



MAS1403

Quantitative Methods for Business Management

Semester 1, 2017–2018

Module leader: Dr. Lee Fawcett

Additional lecturers: Dr. Dave Walshaw and Dr. Ged Cowburn

Tutorials

- This week tutorials are replaced with **computer practicals**
- Days/times:

Group	Day	Time
A	Tues	10–11
B	Wed	10–11
C	Thurs	11–12

- All sessions take place in the Herschel PC cluster
- You will be taught how to use the “NUMBAS” software to access the online **Computer Based Assessments**
- Your first assessed assignment, using NUMBAS, goes live next week – so attendance at this week’s practical session is vital!

Lecture 2

GRAPHICAL METHODS FOR PRESENTING DATA

So far, we have looked at

- how **data** can differ
- how to take **surveys**
- how to **sample** from the **population**
- how to display data in a **tabular form**
 - frequency tables
 - *relative* frequency tables

We now look at how to display data **graphically**

2.1 Stem-and-leaf plots

Stem-and-leaf plots are a quick and easy way of representing data graphically.

They can be used with both **discrete** and **continuous** data.

The easiest way to describe how such a plot is constructed is via demonstration...

...but first, you need to think about the following...

- You need to decide on a reasonable number of **intervals** which span the range of the data (same as for frequency tables last week)
- These interval widths must be **equal**
- You should use **sensible** values for the interval widths

2.1 Stem-and-leaf plots

Construct a stem-and-leaf plot for the following data:

11 12 9 15 21 25 19 8

We need to decide on an **interval width** – **idea: try 10** – i.e.

- 0–9
- 10–19
- 20–29

2.1 Stem-and-leaf plots

This gives a **stem unit** of 10 and a **leaf unit** of 1:

0		8	9		
1		1	2	5	9
2		1	5		

Stem Leaf

$n = 8$ stem unit = 10, leaf unit = 1.

2.1 Stem-and-leaf plots

Some notes...

- “Stem” units are to the left of the line, “leaves” to the right
- For example, for the first observation – **11** – we put a “1” in the stem (one ten) and a “1” as the first leaf
- Each leaf must be equally spaced along the row
- It’s like a **bar chart** sideways! (see later)
- But better – it contains all the **raw observations**
- Put the data in **ascending order** first!

Example 1

The observations in the table below are the recorded time it takes to get through to an operator at a telephone call centre (in seconds).

54	56	50	67	55	38	49	45	39	50
45	51	47	53	29	42	44	61	51	50
30	39	65	54	44	54	72	65	58	62

2		9											
3		0	8	9	9								
4		2	4	4	5	5	7	9					
5		0	0	0	1	1	3	4	4	4	5	6	8
6		1	2	5	5	7							
7		2											

Stem Leaf

$n = 30$, stem unit = 10, leaf unit = 1

Extra example

If there is more than one significant figure in the data, the extra digits are **cut**, not **rounded** – e.g. 2.97 would be “cut” to 2.9.

Consider the following data on lengths of items on a production line (in cm):

2.97	3.81	2.54	2.01	3.49
3.09	1.99	2.64	2.31	2.22

The stem and leaf plot for this is as follows:

1		9		
2		0	2	3
2		5	6	9
3		0	4	
3		8		

$n = 10$, stem unit = 1 cm, leaf unit = 0.1 cm.

2.2 Bar charts

Bar charts are a commonly-used and clear way of presenting categorical data or any ungrouped discrete data

- As with stem and leaf plots, various computer packages allow you to produce these
- First, let us work through the process of producing these by *hand*.

Example 2

The following frequency table represents the modes of transport used daily by 30 students to get to University.

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

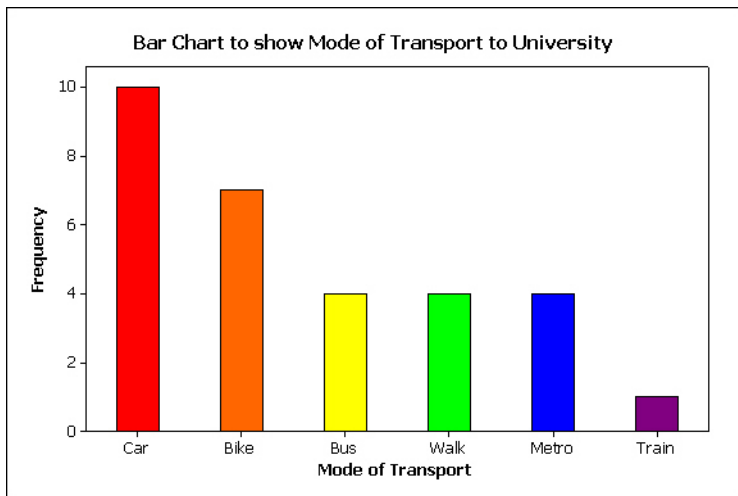
Example 2

The first logical step is to put these into a frequency table, giving:

Mode	Frequency
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
Total	30

We can then easily present this information as a bar chart:

Example 2



2.2 Extension: Multiple bar charts

The data below gives the daily sales of CDs (in £) by music type for an independent retailer.

Day	Chart	Dance	Rest	Total
Monday	12000	10000	2700	24700
Tuesday	11000	8000	3000	22000
Wednesday	9000	6000	2000	17000
Thursday	10000	5000	2500	17500
Friday	12000	11000	3000	26000
Saturday	19000	12000	4000	35000
Sunday	10000	8000	2000	20000
Total	83000	60000	19200	162200

2.2 Extension: Multiple bar charts

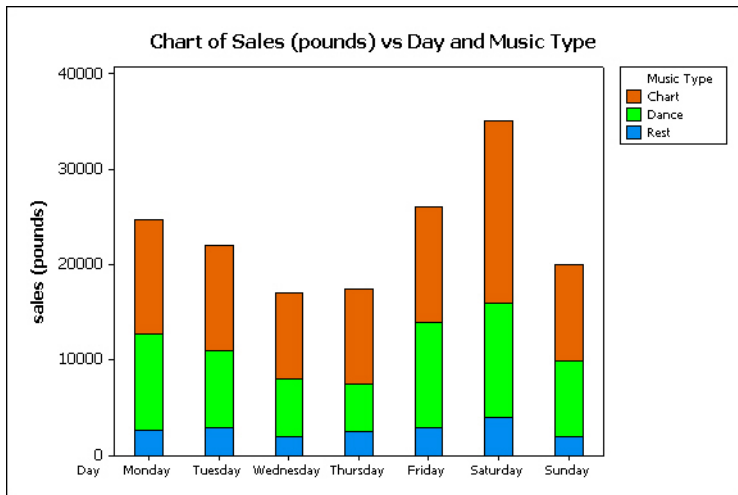
Bar charts could be drawn of **total sales per music type** in the week...

...or of **total daily sales**...

It might be interesting to see **daily sales broken down into music types!**

This can be done using a **multiple bar chart**.

2.2 Extension: Multiple bar charts



2.3 Histograms

Bar charts have their limitations, one of which is that they cannot be used to present **continuous data**.

When dealing with continuous random variables a different kind of graph is required – one such graph is the **histogram**.

At first sight these look similar to bar charts. There are, however, some critical differences:

- The horizontal (x -axis) is a *continuous scale*
- As a result there are *no gaps between the bars*
- Strictly speaking, the *area* of the rectangle is proportional to the frequency – not the height...
- ... however, most computer packages will automatically use equal class widths, meaning the resulting histogram would look the same on a “frequency” or “frequency density” scale anyway

Example 3

Consider the following data, which show service times (in seconds) for a telephone call centre.

214.8412	220.6484	216.7294	195.1217	211.4795
195.8980	201.1724	185.8529	183.4600	178.8625
196.3321	199.7596	206.7053	203.8093	203.1321
200.8080	201.3215	205.6930	181.6718	201.7461
180.2062	193.3125	188.2127	199.9597	204.7813
198.3838	193.1742	204.0352	197.2206	193.5201
205.5048	217.5945	208.8684	197.7658	212.3491
209.9000	197.6215	204.9101	203.1654	192.9706
208.9901	202.0090	195.0241	192.7098	219.8277
208.8920	200.7965	191.9784	188.8587	206.8912

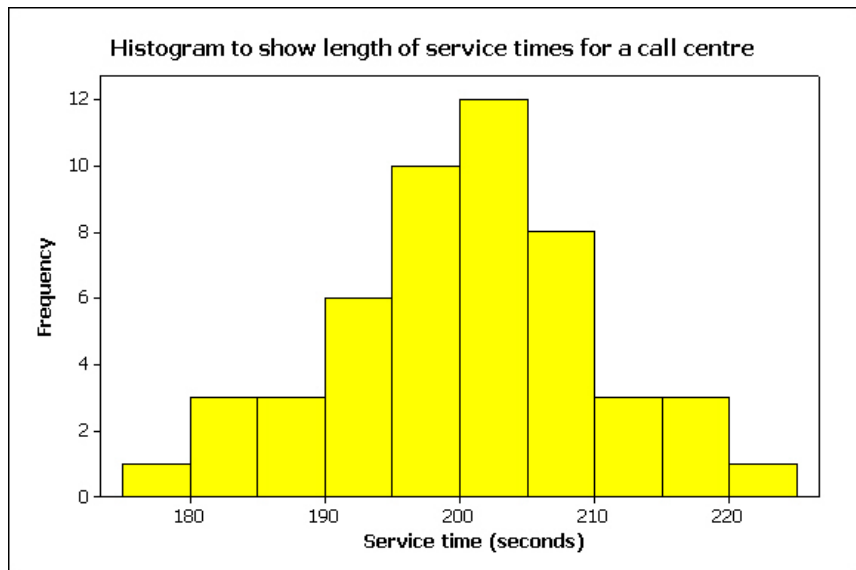
Example 3

Producing a histogram is much like producing a bar chart.

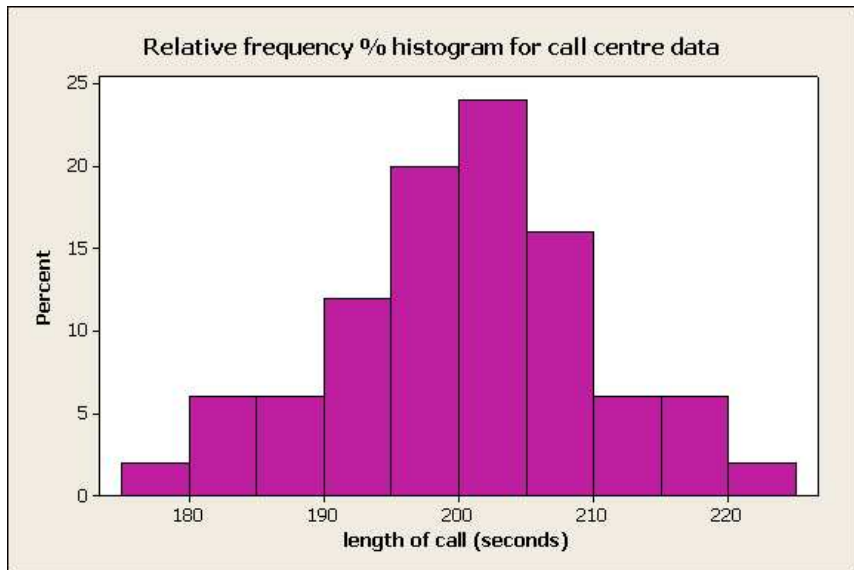
It is often best to produce a frequency table first which collects all the data together in an ordered format.

Service time	Frequency	Rel. Freq. (%)
$175 \leq \text{time} < 180$	1	2
$180 \leq \text{time} < 185$	3	6
$185 \leq \text{time} < 190$	3	6
$190 \leq \text{time} < 195$	6	12
$195 \leq \text{time} < 200$	10	20
$200 \leq \text{time} < 205$	12	24
$205 \leq \text{time} < 210$	8	16
$210 \leq \text{time} < 215$	3	6
$215 \leq \text{time} < 220$	3	6
$220 \leq \text{time} < 225$	1	2
Total	50	100

Example 3



Example 3



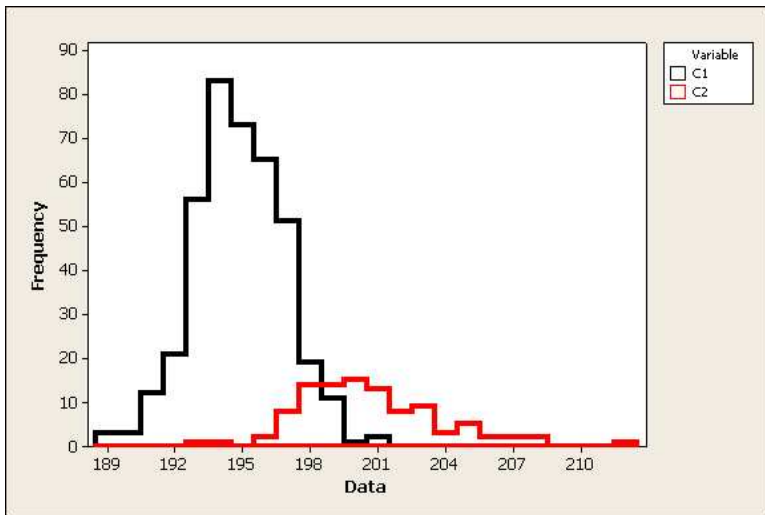
Why use percentage relative frequency?

It is useful for comparing two or more histograms

- If one sample were larger than the other, the standard histograms would look different just because of the different sample sizes
- Looking at percentages “**puts both samples on the same scale**” and removes this difference
- This enables us to look at **relative** differences

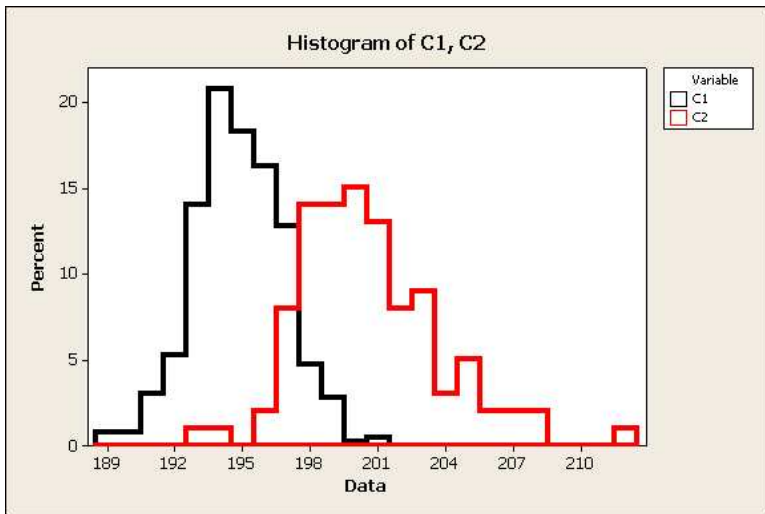
Why use percentage relative frequency?

The following (frequency) histograms were created from two samples, one of size 100 and one of size 400 ...



Why use percentage relative frequency?

... and this is what happens when we use **% relative frequency**:



2.4 Relative frequency polygons

So percentage relative frequency histograms are useful for comparing two groups.

However ...

... can you imagine how **messy** these “superimposed” histograms would get if we had three or more groups?

To get round this **messy problem** we use **relative frequency polygons**.

2.4 Relative frequency polygons

These are a **natural extension** to the relative frequency histogram.

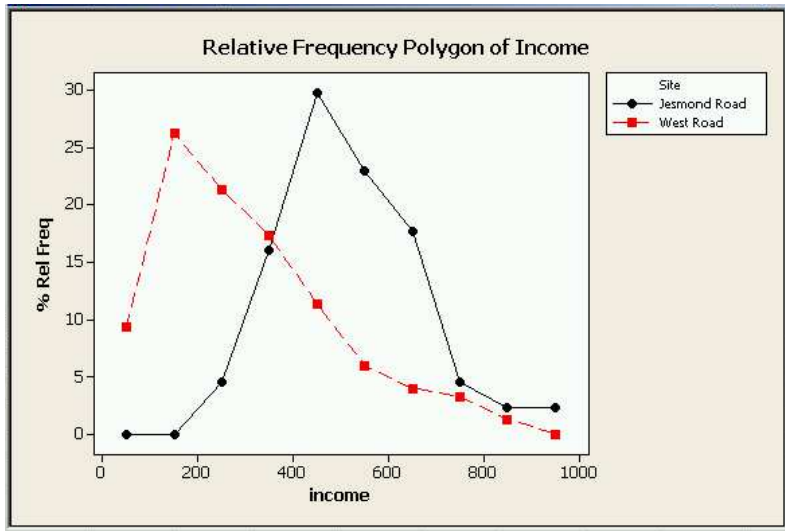
They differ in that, rather than drawing bars, each class is represented by one point (**class interval mid-point**) and these are joined together by straight lines.

Example 4

Consider the following data on gross weekly income (in £) collected from two sites in Newcastle.

Weekly Income (£)	West Road (%)	Jesmond Road (%)
$0 \leq \text{income} < 100$	9.3	0.0
$100 \leq \text{income} < 200$	26.2	0.0
$200 \leq \text{income} < 300$	21.3	4.5
$300 \leq \text{income} < 400$	17.3	16.0
$400 \leq \text{income} < 500$	11.3	29.7
$500 \leq \text{income} < 600$	6.0	22.9
$600 \leq \text{income} < 700$	4.0	17.7
$700 \leq \text{income} < 800$	3.3	4.6
$800 \leq \text{income} < 900$	1.3	2.3
$900 \leq \text{income} < 1000$	0.0	2.3

Example 4



Example 4

Example comments

- Distribution of incomes on West Road is **skewed towards lower values**; those on Jesmond Road are more **symmetric**
- There is evidence of higher incomes in the Jesmond Road area than in the West Road area
- The **spread** of incomes is roughly the same in the two areas
- There are no obvious **outliers**

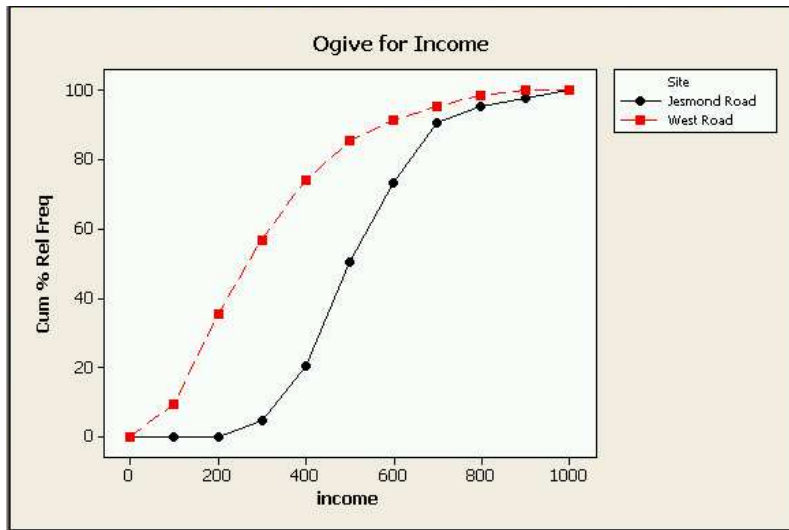
2.5 Cumulative frequency polygons

These are very useful for comparing datasets

- Construct a percentage relative frequency table
- Add a “cumulative” column by adding up the percentages as you go along
- Plot the **upper end-point** of each class interval against the cumulative value

Example 5

For the income data on West Road and Jesmond Road, we get:



2.6 Scatter plots

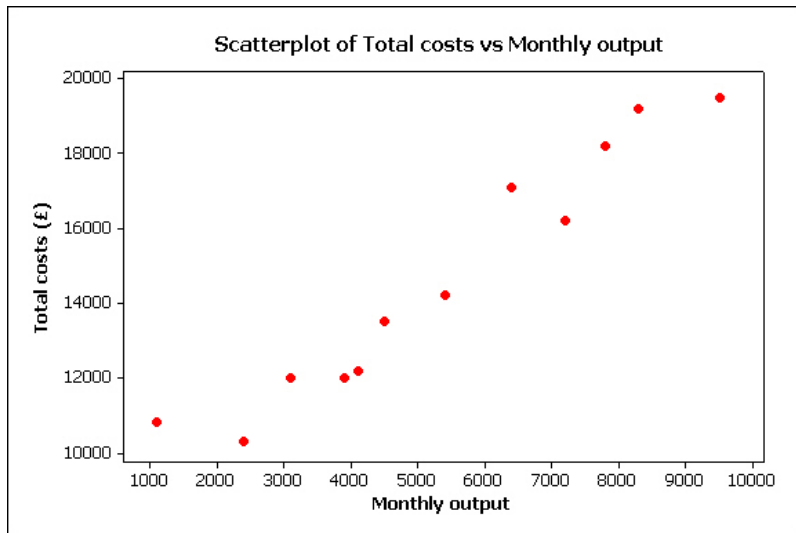
Scatter plots are used to plot two variables which you think might be related – for example: height versus weight; advertising expenditure versus sales...

Example 6

Consider the following data for monthly output and total costs at a factory.

Total costs (£)	Monthly Output (units)
10,300	2,400
12,000	3,900
12,000	3,100
13,500	4,500
12,200	4,100
14,200	5,400
10,800	1,100
18,200	7,800
16,200	7,200
19,500	9,500
17,100	6,400
19,200	8,300

Example 6



Example 6

Example comments

- Clear **positive**, roughly **linear**, relationship between the two variables
- Strong, positive **correlation**
- The more units that are made, the higher the total costs

2.7 Time series plots

Data collected over time can be plotted by using a scatter plot, but with **time** as the (horizontal) x -axis, and where the points are connected by lines.

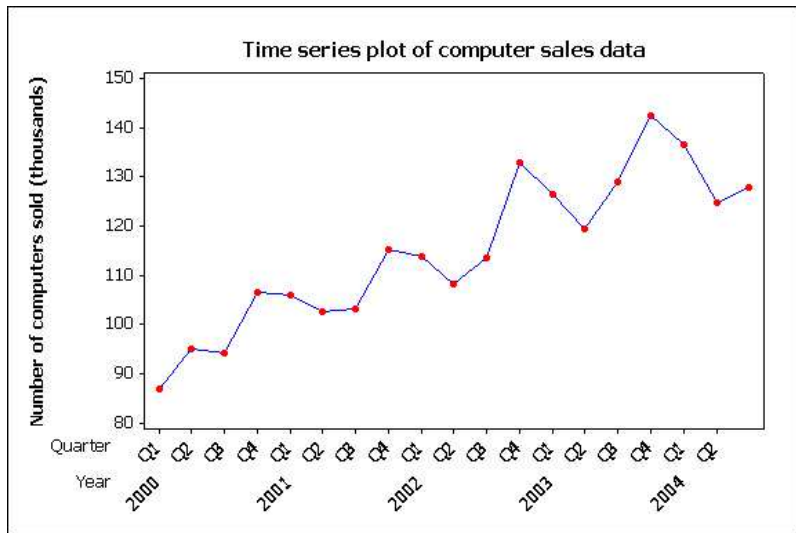
Such a plot is usually called a time series plot (or sometimes simply a “line plot”).

Example 7

Consider the following data on the number of computers sold (in thousands) by quarter (January–March, April–June, July–September, October–December) at a large warehouse outlet.

Quarter	Units Sold	Quarter	Units Sold
Q1 2000	86.7	Q1 2003	126.3
Q2 2000	94.9	Q2 2003	119.4
Q3 2000	94.2	Q3 2003	128.9
Q4 2000	106.5	Q4 2003	142.3
Q1 2001	105.9	Q1 2004	136.4
Q2 2001	102.4	Q2 2004	124.6
Q3 2001	103.1	Q3 2004	127.9
Q4 2001	115.2		
Q1 2002	113.7		
Q2 2002	108.0		
Q3 2002	113.5		
Q4 2002	132.9		

Example 7



Example 7

Example comments

- There is a clear **positive trend** in computer sales through time
- There appear to be **seasonal cycles** (highest sales in Q4 in each year, lowest sales in Q1)
- Seasonal differences appear to be getting more defined?
- No obvious **outliers**



MAS1403

Quantitative Methods for Business Management

Semester 1, 2017–2018

Module leader: Dr. Lee Fawcett

Additional lecturers: Dr. Dave Walshaw and Dr. Ged Cowburn

- This week tutorials are back! Remember:

Group	Day	Time	Place	Who?
A	Thurs	11–12	KGVI, LT1	Ged
B	Thurs	1–2	ARMB, Spence Watson LT	Lee
C	Thurs	2–3	ARMB, Spence Watson LT	Lee

- The first CBA goes live in **practice mode** this week
- Keep an eye on things on the course webpage and on BB:

`www.mas.ncl.ac.uk/~nlf8/teaching/mas1403`

Lecture 3

NUMERICAL SUMMARIES

3. Numerical summaries for data

So far we have only considered **graphical** methods for presenting data.

These are always useful starting points.

As we shall see, however, for many purposes we might also require **numerical** methods for summarising data.

Before we introduce some ways of summarising data numerically, let us first think about some **notation**.

Mathematical notation

In statistics we often replace numbers with letters in order to write **general formulae**.

We generally use a **single letter** to represent sample data and use **subscripts** to distinguish individual observations in the sample.

The most common letter to use is x , although y and z are frequently used as well.

Mathematical notation

For example, suppose we ask a random sample of three people:

“How many mobile phone calls did you make yesterday?”

We might get the following data:

1 5 7

If we take another sample we will most likely get different data:

2 0 3

Mathematical notation

Using algebra we can represent the general case as x_1, x_2, x_3 :

1st sample	1	5	7
2nd sample	2	0	3
typical sample	x_1	x_2	x_3

The i th observation in the samples above is denoted as x_i .

In the first sample above, the second observation is $x_2 = 5$.

In the second sample above, the second observation is $x_2 = 0$.

The letters i and j are most commonly used as the **index numbers** for the subscripts.

The **total** number of observations in a sample is usually referred to by the letter n .

Mathematical notation

The symbol $\sum$ is the upper case of the Greek letter **sigma**. It is used to represent the phrase “**sum the values**”.

For example:

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n.$$

This notation is used to represent the sum of all the values in our data (from the **first** $i = 1$ to the **last** $i = n$), and is often abbreviated to $\sum x$.

Mathematical notation

Consider the following three cases:

$$3 + 4^2 = 19$$

$$3^2 + 4^2 = 25$$

$$(3 + 4)^2 = 49.$$

- In the first case, we simply square 4 and then add this to 3
- In the second case, we square both numbers and then add them together
- In the third case, we add the numbers together and then square the result
- In general terms the second case can be represented as $\sum x^2$
- The third equation can be represented as $(\sum x)^2$

3.1 Measures of location

These are also referred to as measures of **centrality** or, more commonly, **averages**.

In general terms, they tell us the value of a “**typical**” observation in our sample.

There are three measures which are commonly used:

- 1 the **mean**,
- 2 the **median** and
- 3 the **mode**.

We will consider these in turn.

The (arithmetic) mean

The **arithmetic mean** is the most commonly used measure of location.

We often refer to it as the **average** or just the **mean**.

The arithmetic mean is calculated by simply adding all our data together and dividing by the number of data we have.

So if our data were **10**, **12**, and **14**, then our mean would be

$$\frac{10 + 12 + 14}{3} = \frac{36}{3} = 12.$$

The (arithmetic) mean

We denote the **mean** of our **sample**, or **sample mean**, using the notation $\bar{x}$ (“x bar”).

In general, the mean is calculated using the formula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

or equivalently as

$$\bar{x} = \frac{\sum x}{n}.$$

For small data sets this is easy to calculate by hand, but is simplified by using the statistics (**SD**) mode on a University approved calculator.

Example 1

Suppose we ask 7 Stage 2 Business Management students how many units of alcohol they drank last week and get:

16 52 0 6 10 0 21

The sample mean alcohol consumption for these $n = 7$ students is:

$$\bar{x} = \frac{16 + 52 + 0 + 6 + 10 + 0 + 21}{7} = \frac{105}{7} = 15.$$

The (arithmetic) mean

Sometimes we might not have the **raw data**; instead, the data might be available in the form of a **table**.

Previously we have seen the data:

Date	Cars Sold	Date	Cars Sold
01/07/04	9	08/07/04	10
02/07/04	8	09/07/04	5
03/07/04	6	10/07/04	8
04/07/04	7	11/07/04	4
05/07/04	7	12/07/04	6
06/07/04	10	13/07/04	8
07/07/04	11	14/07/04	9

The mean number of cars sold per day is

$$\bar{x} = \frac{9 + 8 + \dots + 8 + 9}{14} = 7.71.$$

The (arithmetic) mean

These data can be presented as the frequency table

Cars Sold ($x_{(j)}$)	Frequency (f_j)
4	1
5	1
6	2
7	2
8	3
9	2
10	2
11	1
Total (n)	14

The sample mean can be calculated from these data as

$$\bar{x} = \frac{4 \times 1 + 5 \times 1 + 6 \times 2 + \dots + 11 \times 1}{14} = 7.71.$$

The (arithmetic) mean

We can express this calculation of the sample mean from the tabulated data as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k x_{(j)} f_j.$$

Here the different values of X which occur in the data are $x_{(1)}, x_{(2)}, \dots, x_{(k)}$.

In this example, $x_{(1)} = 4$, $x_{(2)} = 5$, $\dots$, $x_{(k)} = 11$ and $k = 8$.

The (arithmetic) mean

If we only have **grouped** frequency data, it is still possible to approximate the value of the sample mean. For example:

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

The sample mean of these data is 9.73.

The (arithmetic) mean

Grouping these data into a frequency table gives

Class Interval	Mid-point (m_j)	Frequency (f_j)
$8.0 \leq x < 8.5$	8.25	1
$8.5 \leq x < 9.0$	8.75	1
$9.0 \leq x < 9.5$	9.25	5
$9.5 \leq x < 10.0$	9.75	7
$10.0 \leq x < 10.5$	10.25	2
$10.5 \leq x < 11.0$	10.75	3
$11.0 \leq x < 11.5$	11.25	1
Total (n)		20

What if only this grouped frequency table was given?

The (arithmetic) mean

When the raw data are not available, we don't know where each observation lies in each interval.

The best we can do is to assume that all the values in each interval lie at the **central value of the interval**, that is, at its mid-point.

Therefore, the (approximate) sample mean is calculated using the the frequencies (f_j) and the mid-points (m_j) as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k f_j m_j.$$

For the grouped data above, we obtain

$$\bar{x} = \frac{1}{20} (1 \times 8.25 + 1 \times 8.75 + \cdots + 3 \times 10.75 + 1 \times 11.25) = 9.775.$$

The Median

The median is occasionally used instead of the mean, particularly when the data are **asymmetric**.

The median is the **middle value** of the observations when they are listed in ascending order.

The median is the value that has half the observations above it and half below.

If the sample size (n) is an **odd** number, we have:

$$\text{median} = \left(\frac{n+1}{2} \right)^{th} \text{ smallest observation.}$$

Example 2

For example, ordering the student alcohol data from smallest to largest, we get:

0 0 6 10 16 21 52

The middle value is 10, so the median is 10 units per week.

Example 3

Suppose we also asked Stage 2 Marketing and Management students how many units of alcohol they drank last week, and we got:

21 0 12 14

Rearranging into ascending numerical order, we get:

0 12 14 21

Now we have two values in the middle! So the median is the average of these two:

$$\text{Median} = \frac{12 + 14}{2} = 13 \text{ units per week.}$$

The median

The median is often used if the dataset have an asymmetric profile, since it is not distorted by outliers.

For example, suppose we added one more Marketing and Management student to the last sample:

0 12 14 21 150

$$\bar{x} = \frac{0 + 12 + 14 + 21 + 150}{5} = 39.4 \text{ units per week}$$

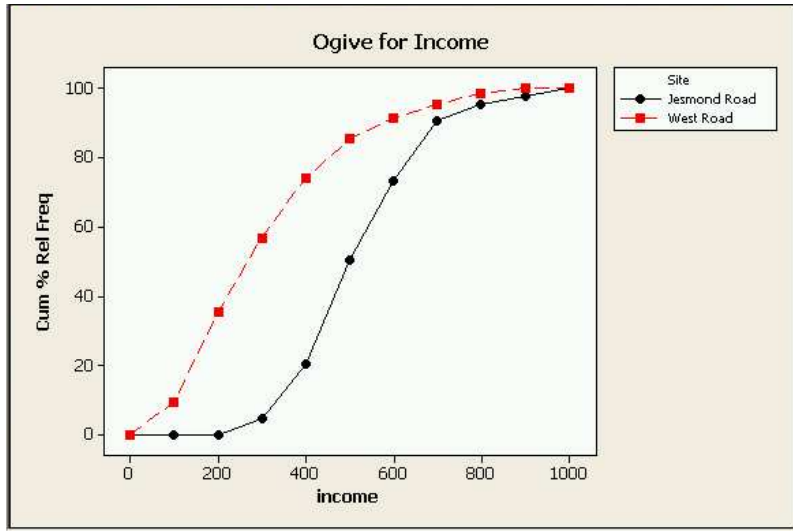
$$\text{Median} = 14 \text{ units per week}$$

The median

It is possible to estimate the median value from an **cumulative relative frequency polygon** as it is half way through the ordered data and hence is at the 50% level of the cumulative frequency.

The accuracy of this estimate will depend on the accuracy of the graph drawn.

The median



The Mode

This is the final measure of location we will look at. It is the value of the random variable in the sample which occurs with the **highest frequency**.

It is usually found by **inspection**.

For **discrete** data this is easy. The mode is simply the most common value. On a bar chart, it would be the category with the highest bar.

The Mode

It is possible to refer to **modal classes** with grouped data.

This is simply the class with the greatest frequency of observations.

For example, the modal class of

Class	Frequency
$10 \leq x < 20$	10
$20 \leq x < 30$	15
$30 \leq x < 40$	30

is obviously $30 \leq x < 40$.

3.2 Measures of spread

A measure of **location** is insufficient in itself to summarise data as it only describes the value of a typical outcome.

For example:

Sample 1	6	22	38	$\bar{x} = 22$	median = 22
Sample 2	21	22	23	$\bar{x} = 22$	median = 22

Both samples have the same measures of average. But they are clearly very different samples!

3.2 Measures of spread

The mean or the median does not fully represent the data.

There are three basic **measures of spread** which we will consider:

- 1 the **range**
- 2 the **inter-quartile range**
- 3 the **sample variance**

The range

This is the **simplest** measure of spread.

It is simply the difference between the largest and smallest observations.

In our simple examples from the previous slides the range for the first set of numbers is $38 - 6 = 32$...

...and for the second set it is $23 - 21 = 2$.

These clearly describe very different data sets.

We say the first set has a **wider range** than the second.

The range

There are two **problems** with the range as a measure of spread:

- It is unduly influenced by extreme observations or (**outliers**)
- It is only suitable for comparing (roughly) equally sized samples

The Inter-Quartile Range

The **inter-quartile range** (IQR) describes the range of the middle half of the data and so is less prone to the influence of any extreme values.

To calculate the IQR we simply divide the the ordered data into four quarters.

The three values that split the data into these quarters are called the **quartiles**

- The first quartile (**lower quartile**, Q_1) has 25% of the data below it
- The second quartile (**median**, Q_2) has 50% of the data below it
- The third quartile (**upper quartile**, Q_3) has 75% of the data below it

The Inter-Quartile Range

We already know how to find the median; the other quartiles are calculated as follows:

$$Q1 = \frac{(n+1)}{4} \text{th smallest observation}$$

$$Q3 = \frac{3(n+1)}{4} \text{th smallest observation.}$$

Just as with the median, these quartiles might not correspond to actual observations.

Example 4

Calculate the IQR for the following data:

8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8

Here,

$$\text{Position of } Q_1 = \frac{19}{4} = 4.75.$$

So Q_1 itself is three-quarters of the way between the 4th and the 5th observations; i.e. $Q_1 = \mathbf{9.275}$.

Next, Q_3 :

$$\text{Position of } Q_3 = 3 \times 4.75 = 14.25.$$

So Q_3 itself is one-quarter of the way between the 14th and 15th observations; i.e. $Q_3 = \mathbf{10.325}$.

Thus $IQR = 10.325 - 9.275 = 1.05$.

The variance and standard deviation

The **sample variance** is the standard measure of spread used in statistics.

It is usually denoted by s^2 and is the **average of the squared distances** of the observations from the sample mean.

We use the formula

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}.$$

which can be simplified to

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The variance and standard deviation

Or equivalently, written as

$$s^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^n x_i^2 - (n \times \bar{x}^2) \right\}$$

Standard deviation

The **sample standard deviation** s is the square root of the sample variance.

This quantity is often used in preference to the sample variance as it has the same units as the original data and so is perhaps easier to understand.

If this appears complicated, don't worry, as most calculators will give the sample standard deviation when in **SD** mode.

Example 5

Consider again the data on the number of units of alcohol consumed by a sample of 7 students last week:

16 52 0 6 10 0 21

Find the sample variance and standard deviation.

Example 5

Recall we have

$$s^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^n x_i^2 - (n \times \bar{x}^2) \right\}$$

First, the “sum of squares”:

$$\sum x^2 = 16^2 + 52^2 + 6^2 + 10^2 + 21^2 = 3537.$$

Next:

$$n \times \bar{x}^2 = 7 \times 15^2 = 1575.$$

Then

$$s^2 = \frac{3537 - 1575}{6} = 327.$$

Therefore, $s = \sqrt{327} = 18.08$ units per week.

Box and Whisker Plots

Box and whisker plots (box plots) are another graphical method for displaying data.

They are particularly useful in highlighting differences between groups.

These plots use some of the key summary statistics we have looked at earlier, the **quartiles**, as well as the **maximum** and **minimum**.

Example 6

Draw a box plot for data with the following summaries:

Minimum	$\text{min} = 10$
Lower quartile	$Q1 = 40$
Median	$Q2 = 43$
Upper quartile	$Q3 = 45$
Maximum	$\text{max} = 50$

Example 7

