



MAS1403

Quantitative Methods for Business Management

Semester 1

Dr. Lee Fawcett

School of Mathematics, Statistics & Physics

MAS1403: Quantitative Methods for Business Management

2017/18

Lecturer: Dr. Lee Fawcett, Room 2.07 Herschel Building.

Email: lee.fawcett@ncl.ac.uk

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

Lectures:	Mondays at 5pm	In the Curtis Auditorium, Herschel Building
Tutorials:	One per week	There are 3 groups – check the module webpage to see which tutorial to attend.
Practicals:	Occasionally	Check the full schedule overleaf for dates. These will take place instead of the tutorials.
Drop-in:	Mon 4-5pm, Wed 1-2pm	Optional “office hours” where I will be available in my office for any help with the work.

Lecture notes and handouts

You will be provided with a booklet containing lecture notes and tutorial exercises.

You should bring your booklet to every class!

There will often be gaps in the lecture notes for you to complete during the lecture, so make sure you’ve got them with you!

All lecture notes, slides and solutions to tutorial exercises will be available to download from the course website (see above). There should be a link to this website from within Blackboard. Some additional handouts may only be available in lectures and tutorials.

You will notice that my lecture slides are colour-coded: **Green** for announcements, **blue** for “listen and learn” and **red** for “write”!

Assessment

Assessment for this course is via examination (**60% at end of Semester 2**), assignments (**10% each semester**) and computer-based assessments (**10% each semester**). **Ordinarily, if you fail this module you cannot proceed to Stage 2 of your degree!**

Exam:	May/June 2018	A two hour, open-book , computer-based exam based on whole course: Answer all questions.
Assignments:	Dec 2017, May 2018	About three big questions in each, some of which will use your own personal datasets and some of which will require you to use the computer package Minitab.
CBAs:	Throughout the year	Three CBAs in each Semester. Available in “practice mode” for one week and then “exam mode” the next week. Some multiple choice questions, but mainly data response/calculations. Every student will get a different set of questions from a bank of hundreds! Must be done in your own time.

Late Work Policy:

It is not possible to extend submission deadlines for coursework in this module and no late work can be accepted. For details of the policy (including procedures in the event of illness etc.) please look at the School web site:

<http://www.ncl.ac.uk/maths/students/resources/late-missed/>

Other Stuff

Email:	Check your University email every day – announcements about the course will be made regularly!
Calculator:	There is no way around it, you must have a scientific calculator for this course, and it must be on the University’s approved list! I recommend the Casio fx-85GT PLUS (about £10). You can get advice on how to use the Statistics mode of your calculator in tutorials, and some video presentations on use of the calculator will be available from the module webpage. You should bring your calculator to every class. You will be stuck without one!

MAS1403 - Provisional Schedule for Semester 1

Week 1 (week commencing 2/10/17)

Topic 1: Data collection, display and summaries

Mon	2nd October	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	5th October	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	5th October	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	5th October	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Week 2 (week commencing 9/10/17)

Mon	9th October	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Tue	10th October	Practical	10 - 11	Herschel Building PC cluster
Wed	11th October	Practical	10 - 11	Herschel Building PC cluster
Thu	12th October	Practical	11 - 12	Herschel Building PC cluster

Week 3 (week commencing 16/10/17)

CBA1 opens in "practice mode"

Mon	16th October	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	19th October	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	19th October	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	19th October	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Week 4 (week commencing 23/10/17)

Topic 2: Probability and decision making

CBA1 opens in "assessed mode" – deadline: midnight Friday 27th October

Mon	23rd October	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	26th October	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	26th October	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	26th October	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Week 5 (week commencing 30/10/17)

Mon	30th October	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	2nd November	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	2nd November	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	2nd November	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Week 6 (week commencing 6/11/17)

CBA2 opens in "practice mode"

Mon	6th November	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	9th November	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	9th November	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	9th November	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Week 7 (week commencing 13/11/17)

Topic 3: Probability models

CBA2 opens in "assessed mode" – deadline: midnight Friday 17th November

Assignment 1 available

Mon	13th November	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Tue	14th November	Practical	10 - 11	Herschel Building PC cluster
Wed	15th November	Practical	10 - 11	Herschel Building PC cluster
Thu	16th November	Practical	11 - 12	Herschel Building PC cluster

Week 8 (week commencing 20/11/17)

Mon	20th November	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	23rd November	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	23rd November	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	23rd November	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Week 9 (week commencing 27/11/17)

Mon	27th November	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	30th November	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	30th November	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	30th November	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Week 10 (week commencing 4/12/17)

CBA3 opens in “practice mode” and “assessed mode”

Mon	4th December	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	7th December	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	7th December	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	7th December	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Week 11 (week commencing 11/12/17)

Assignment 1 deadline: 4pm, Thursday 14th December

CBA3 deadline: midnight, Friday 15th December

Mon	11th December	Lecture	5 - 6	Herschel Building, Curtis Auditorium
Thu	14th December	Tutorial	11 - 12	King George VI Building, Lecture Theatre 1
Thu	14th December	Tutorial	1 - 2	Armstrong Building, Spence Watson Lecture Theatre
Thu	14th December	Tutorial	2 - 3	Armstrong Building, Spence Watson Lecture Theatre

Christmas vacation!

Week 12 (week commencing 8/1/18) – Revision week

Mon	8th January	Lecture	5 - 6	Herschel Building, Curtis Auditorium
-----	-------------	---------	-------	--------------------------------------

1 Collecting and presenting data

1.1 Definitions

The quantities measured in a study are called *random variables* and a particular outcome is called an *observation*. A collection of observations is the *data*. The collection of all possible outcomes is the *population*.

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.

Data/random variables are of different types:

- **Qualitative** (i.e. non-numerical)
 - **Categorical**
 - * Outcomes take values from a set of categories, e.g. mode of transport to Uni: {car, metro, bus, walk, other}.
- **Quantitative** (i.e. numerical)
 - **Discrete**
 - * Things that are countable, e.g. number of people taking this module.
 - * **Ordinal**, e.g. response to questionnaire; 1 (strongly disagree) to 5 (strongly agree)
 - **Continuous**
 - * Things that we measure rather than count, e.g. height, weight, time.

Example 1



Identify the type of data described in each of the following examples:

- (a) The time between emails arriving in your inbox is recorded.
- (b) An opinion poll was taken asking people what is their favourite chocolate bar.
- (c) The number of students attending a MAS1403 tutorial is recorded.

1.2 Sampling techniques

We typically aim for the *sample* to be *representative* of the *population*. The larger the sample size the more precise information we have about the population.

There are three main types of sampling: *random*, *quasi-random*, *non-random*.

- **Simple random sampling** (random)

- Each element in the population is equally likely to be drawn into the sample.
- All elements are “put in a hat” and the sample is drawn from the “hat” at random.
- Advantages – easy to implement; each element has an equal chance of being selected.
- Disadvantages – often don’t have a complete list of the population; not all elements might be equally accessible; it is possible, purely by chance, to pick an unrepresentative sample.

- **Stratified sampling** (random)

- We take a simple random sample from each “strata”, or group, within the population. The sample sizes are usually proportional to the population sizes.
- Advantages – sampling within each stratum ensures that that stratum is properly represented in the sample; simple random sampling within each stratum has the advantages listed under simple random sampling above.
- Disadvantages – need information on the size and composition of each group; as with simple random sampling, we need a list of all elements within each strata.

- **Systematic sampling** (quasi-random)

- The first element from the population is selected at random, and then every k th item is chosen after this. This type of sampling is often used in a production line setting.
- Advantages – its simplicity! – and so it’s easy to implement.
- Disadvantages – not completely random; if there is a pattern in the production process it is easy to obtain a biased sample; only really suited to structured populations.

- **Judgemental sampling** (non-random)

- The person interested in obtaining the data decides who should be surveyed; for example, the head of a service department might suggest particular clients to survey based on his judgement, and they might be people who he thinks will give him the responses he wants!
- Advantages – very focussed and aimed at the target population.
- Disadvantages – relies on the judgement of the person conducting the questionnaire/survey, and so cannot be guaranteed to be representative; is prone to bias.

- **Accessibility sampling** (non-random)
 - Here, the most easily accessible elements are sampled.
 - Advantages – easy to implement.
 - Disadvantages – prone to bias.
- **Quota sampling** (non-random)
 - Similar to stratified sampling, but uses judgemental sampling within each strata instead of random sampling. We sample within each strata until our quotas have been reached.
 - Advantages – results can be very accurate as this technique is very targeted.
 - Disadvantages – the identification of appropriate quotas can be problematic; this sampling technique relies heavily on the judgement of the interviewer.

Example 2

- (a) A toy company, *Toys 4 U*, is to be inspected for the quality and safety of the toys it produces. The inspection team takes a sample of toys from the production line by choosing the first toy at random, and then selecting every 100th toy thereafter. What form of sampling are the team using?
- (b) Another inspection team is to investigate the quality of the smartphone covers made by a local factory. In a typical working day the factory produces 100 covers for the new i-Phone and 200 covers for the latest Samsung phone. Suggest a suitable form of sampling to check the quality of the smartphone covers produced.

Solution



1.3 Frequency tables

Once we have collected our data, often the first stage of any analysis is to present them in a simple and easily understood way. Tables are perhaps the simplest means of presenting data. The way we construct the table depends on the type of data.

Example 3 (discrete data)

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
1st July	9	8th July	10
2nd July	8	9th July	5
3rd July	6	10th July	8
4th July	7	11th July	4
5th July	7	12th July	6
6th July	10	13th July	8
7th July	11	14th July	9

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 %
Totals			

Example 4 (continuous data)

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

We can present these data in a relative frequency as follows:



Class Interval	Tally	Frequency	Relative Frequency %
$180 \leq \text{time} < 185$			
$185 \leq \text{time} < 190$			
$190 \leq \text{time} < 195$			
$195 \leq \text{time} < 200$			
$200 \leq \text{time} < 205$			
$205 \leq \text{time} < 210$			
Totals			

1.4 Exercises

1. Identify the type of data described in each of the following examples:
 - (a) An opinion poll was taken asking people which party they would vote for in a general election.
 - (b) In a steel production process the temperature of the molten steel is measured and recorded every 60 seconds.
 - (c) A market researcher stops you in Northumberland Street and asks you to rate between 1 (disagree strongly) and 5 (agree strongly) your response to opinions presented to you.
 - (d) The hourly number of units produced by a beer bottling plant is recorded.
2. A credit card company wants to investigate the spending habits of its customers. From its lists, the first customer is selected at random; thereafter, every 30th customer is selected.
 - (a) Is this an example of simple random sampling, stratified sampling, systematic sampling, or judgemental sampling?
 - (b) Is this form of sampling random, quasi-random or non-random?
3. The number of telephone calls made by 20 students in a day is shown below.

3 5 1 0 0 2 1 0 3 1 4 3 2 0 1 1 1 2 0 4

Put these data into a relative frequency table.

4. The following data are the recorded length (in seconds) of 25 mobile phone calls made by one student.

281.4	293.4	306.5	286.6	298.4
312.7	327.7	311.5	314.8	303.3
270.7	293.9	310.9	346.4	304.6
304.1	320.7	283.6	337.5	259.6
305.4	317.9	289.5	286.9	300.5

Complete the following percentage relative frequency table for these data.

Class Interval	Tally	Frequency	Relative Frequency %
$250 \leq \text{time} < 270$			
$270 \leq \text{time} < 290$			
$290 \leq \text{time} < 310$			
$310 \leq \text{time} < 330$			
$330 \leq \text{time} < 350$			
Totals			

2 Graphical methods for presenting data

Once we have collected our data, often the best way to summarise this data is through an appropriate graph. Graphs are more eye-catching than tables, and give us an “at-a-glance” picture of the main features of our data: its distribution, location, spread, outliers etc.

2.1 Stem-and-leaf plots

Example 1

The observations 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

Represent the data in a stem-and leaf plot.



Stem		Leaf
$n =$	stem unit =	leaf unit =

Some notes on stem-and-leaf plots.

- Always show the stem units and the leaf units.
- The stem unit will usually be either 10 or 1; the corresponding unit for the leaves is usually 1 and 0.1.
- Order the leaves from smallest to largest.
- If you have observations recorded to 2 d.p., always round *down*, e.g. 2.97 would become 2.9 rather than 3.0.

2.2 Bar charts

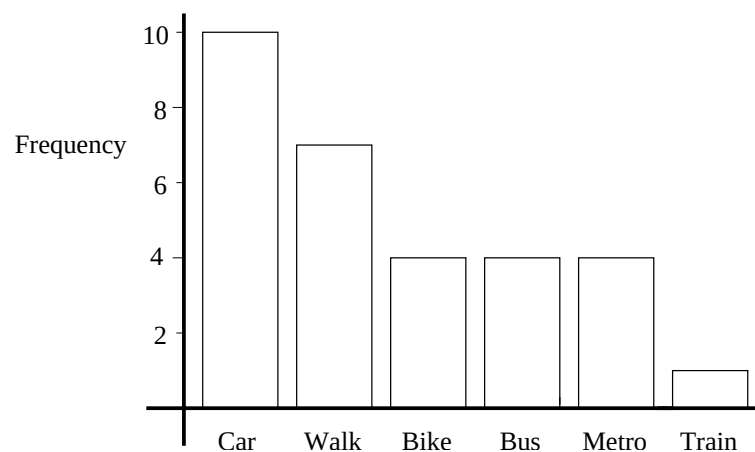
A commonly-used and clear way of presenting *categorical* data or any ungrouped *discrete* data.

Example 2

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

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

This gives the following bar chart:



This bar chart clearly shows that the most popular mode of transport is the car and the least popular is the train (in our small sample).

2.3 Histograms

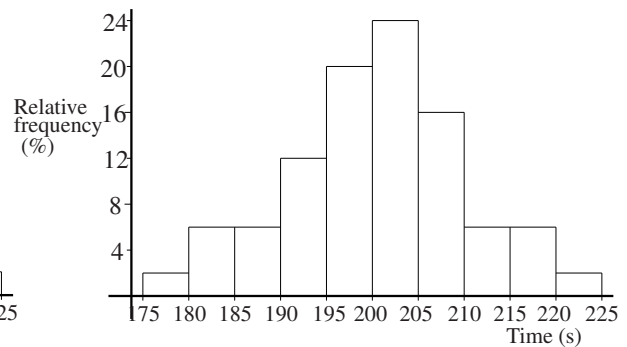
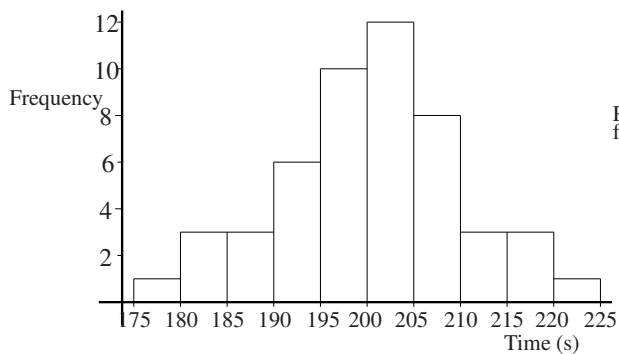
Histograms can be thought of as “bar charts for continuous data”. First construct a grouped frequency table then draw a bar for each *class interval*. Important point: unlike bar charts, there are no gaps between the bars in a histogram.

Example 3

The following frequency table summarises the service times (in seconds) at a telephone call centre.

Service time	Frequency	Relative Frequency (%)
$175 \leq \text{time} < 180$	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
Totals	50	100

The histogram for these data is:



We can also plot relative frequency (%) on the vertical axis: this gives a **percentage relative frequency histogram**. These are useful for comparing datasets of different sizes.

2.4 Relative frequency polygons

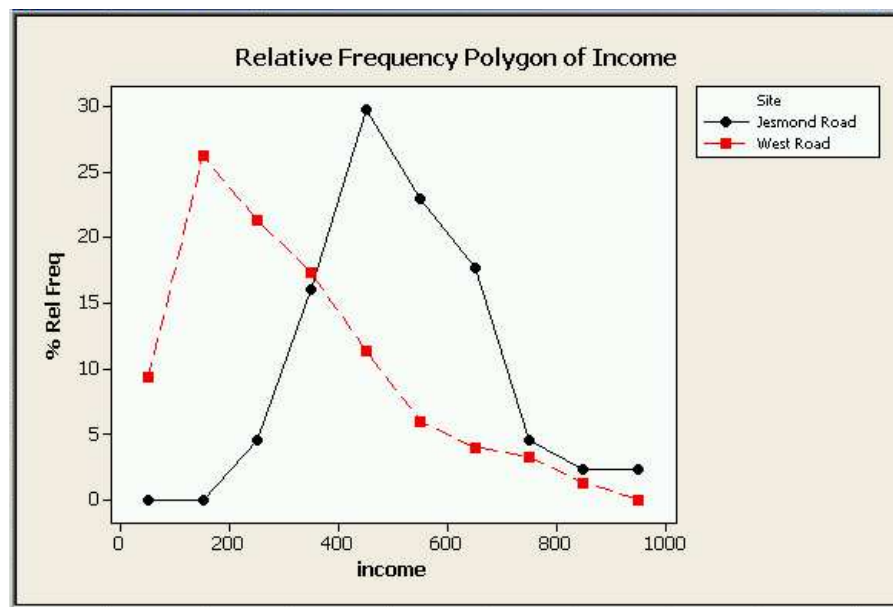
The relative frequency polygon is exactly the same as the relative frequency histogram, but instead of having bars we join the mid-points of the top of each bar with a straight line. These are useful for illustrating the relative differences between two or more groups.

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

The following plot shows percentage relative frequency polygons for the two groups.



Example comments: The distribution of incomes on West Road is skewed towards lower values, whilst those on Jesmond Road are more symmetric. The graph clearly shows that income in the Jesmond Road area is higher than that in the West Road area. The spread of incomes is roughly the same in the two areas. There are no obvious outliers.

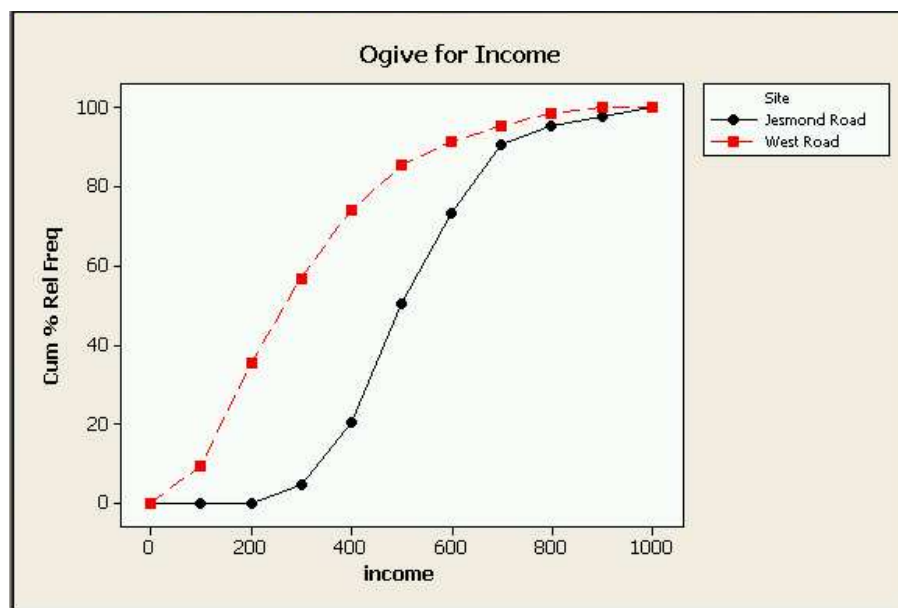
2.5 Cumulative frequency polygons

These are very useful for comparing datasets.

- Construct a percentage relative frequency table for your data.
- 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

The following plot contains the cumulative frequency polygons for the income data at both the West Road and Jesmond Road sites.



It clearly shows the line for Jesmond Road is shifted to the right of that for West Road. This tells us that the surveyed incomes are higher on Jesmond Road. We can compare the percentages of people earning different income levels between the two sites quickly and easily.

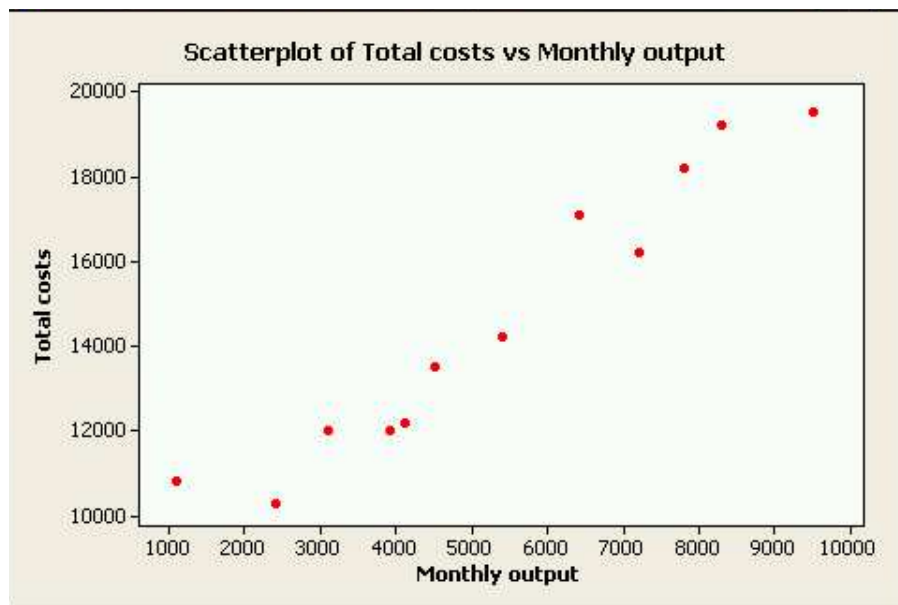
2.6 Scatter plots

Scatter plots are used to plot two variables which you believe might be related, for example, advertising expenditure and sales.

Example 6

The following data represents 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



For scatter plots, we comment on whether there is a *linear association* between the two variables? If so, is this *positive* (“uphill”) or *negative* (“downhill”)? Is the association *strong*? Or maybe *moderate* or *weak*?

The plot above shows a clear positive, roughly linear, relationship between the two variables: the more units made, the more it costs in total.

2.7 Time Series Plots

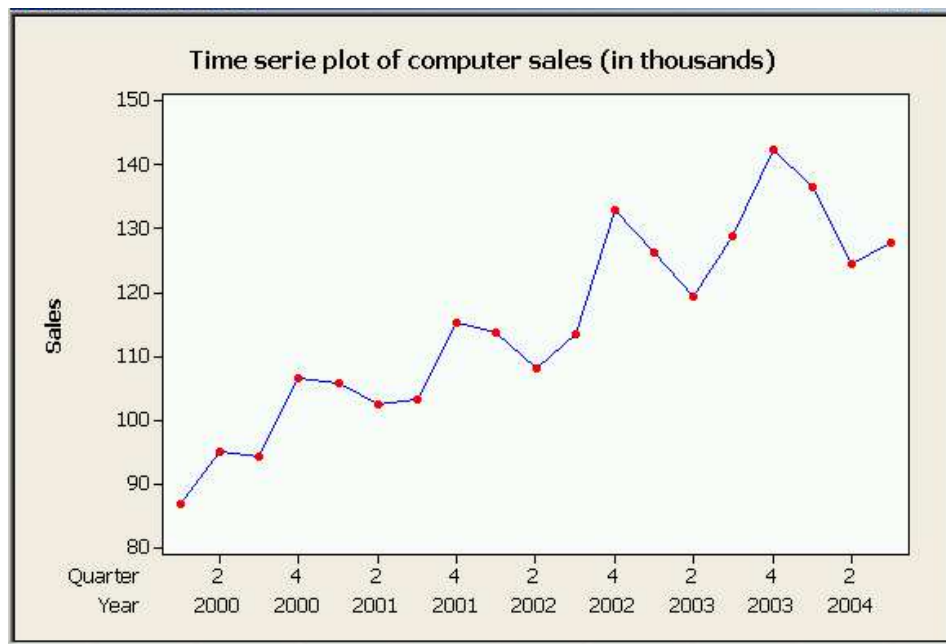
Data collected over time can be plotted by using a **scatter plot**, but with *time* as the (horizontal) *x*-axis, and where the points are connected by lines: a *time series plot*.

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, starting in quarter 1 2000.

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

The time series plot is:



For time series plots, look out for *trend* and *seasonal cycles* in the data. Also look out for any *outliers*.

The above plot clearly shows us two things: firstly, that there is an upwards trend to the data (sales increase over time), and secondly that there is some regular variation around this trend (sales are usually higher in quarters 1 and 4 than quarters 2 and 3).

2.8 Exercises

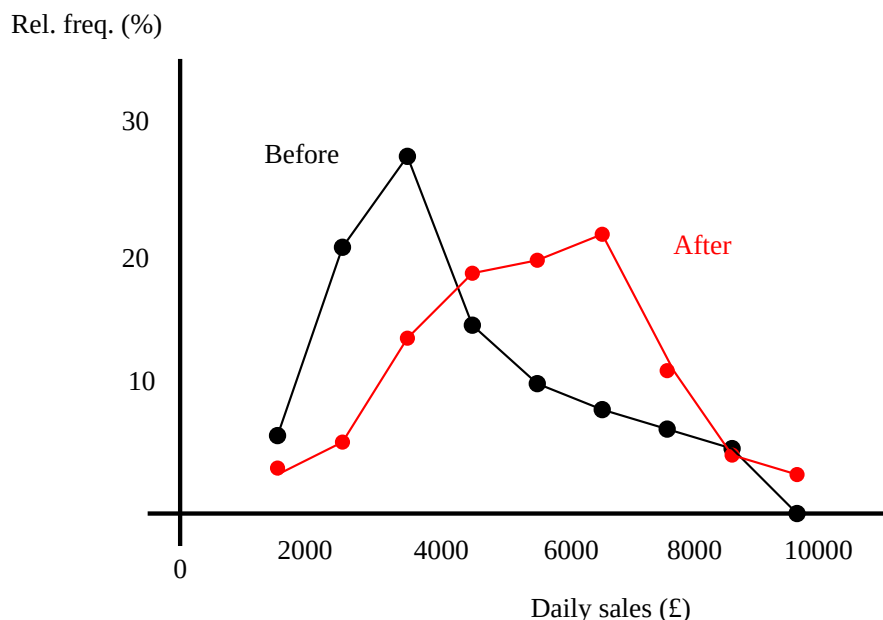
- The following table shows the weight (in kilograms) of 50 sacks of potatoes leaving a farm shop (the data have been ordered from smallest to largest).

8.1	8.2	8.5	8.7	8.8
8.9	9.2	9.3	9.3	9.4
9.5	9.5	9.6	9.6	9.6
9.7	9.7	9.9	9.9	10.0
10.0	10.0	10.0	10.0	10.1
10.2	10.2	10.2	10.3	10.3
10.4	10.4	10.4	10.5	10.6
10.6	10.6	10.6	10.6	10.7
10.8	10.9	11.0	11.2	11.3
11.3	11.3	11.5	11.6	12.8

Display these data in a stem and leaf plot. State clearly both the stem and the leaf units. Comment on the distribution of the data.

- Which is more suitable for representing the data from Question 1 (above), a bar chart or a histogram? Justify your answer.
- A small clothes shop have records of daily sales both before and after a local radio advertising campaign. Relative frequency polygons of the sales data are shown below.

Relative frequency polygons of sales (before and after)



Comment, with justification, on the success, or otherwise, of the advertising campaign.

3 Numerical summaries for data

Numerical summaries are numbers which summarise the main features of your data. You should use both a measure of *location* and a measure of *spread* to summarise your dataset.

3.1 Measures of location

A measure of location is a value which is “typical” of the observations in our sample

1. The mean

The sample mean is the “average” of our data: the total divided by the sample size. It’s given by the formula

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

which, put more simply, means “add them up and divide by how many you’ve got”.

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 of these $n = 7$ students is



If your data are given in the form of a frequency table, then you “multiply each observation by its frequency, add these numbers together and then divide by how many you’ve got”. If you have a grouped frequency table, then you don’t know the value of each observation and so just use the midpoint of the class interval.

2. The median

This is just the observation “in the middle”, when the data are put into order from smallest to largest:

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

Example 2

Ordering the student alcohol data from the previous example gives 0, 0, 6, 10, 16, 21, 52. Clearly the middle value is 10, so the median is 10 units per week.

Example 3

Suppose we also asked four Stage 2 Marketing and Management students how many units of alcohol they drank last week, and got: 21, 0, 12, 14. Calculate the median.

Solution



The median is often used if the dataset has an asymmetric profile, since it is not distorted by extreme observations (“outliers”).

3. The mode

The mode is simply the most frequently occurring observation. For example, consider the following data: 2, 2, 2, 3, 3, 4, 5. The mode is 2 as it occurs most often. The modal class is easily obtained from a grouped frequency table or a histogram; it's the class with the highest frequency.

3.2 Measures of spread

A measure of spread quantifies how “spread out” (or how “variable”) our data are.

1. The range

Range = largest value – smallest value. For example, the range of the data: 2, 2, 2, 3, 3, 4, 5 is $5 - 2 = 3$.

- Advantage: very simple to calculate.
- Disadvantages: sensitive to extreme observations; only suitable for comparing (roughly) equally sized samples.

2. The inter-quartile range (IQR)

The IQR measures the range of the middle half of the data, and so is less affected by extreme observations. It is given by $Q3 - Q1$, where

$$Q1 = \frac{(n+1)}{4} \text{th smallest observation} \quad (\text{“lower quartile”})$$
$$Q3 = \frac{3(n+1)}{4} \text{th smallest observation} \quad (\text{“upper quartile”}).$$

Example 4

Calculate the inter-quartile range 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

Solution



3. The variance and standard deviation

The *sample variance* is the standard measure of spread used in statistics. It can be thought of as “the average squared deviation from the mean”, and is given by

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

The following formula is easier for calculations

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

In practice most people simply use the Statistics mode on their calculator (mode SD or Stat).

The sample *standard deviation* is just the square root of the variance, and is often preferred as it is in the “original units of the data”.

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. Calculate the sample variance and the sample standard deviation.

Solution



3.3 Box plots

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

Example 6

Suppose that, from our data, we obtain the following summary statistics:

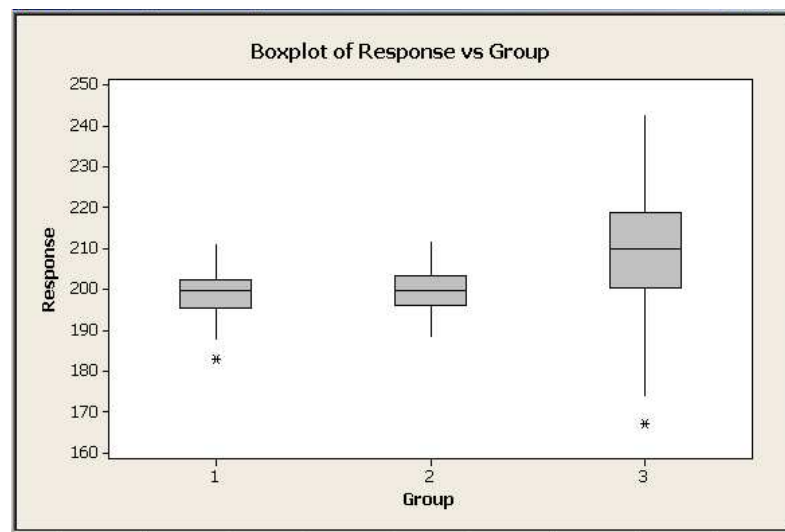
Minimum	Lower Quartile (Q1)	Median (Q2)	Upper Quartile (Q3)	Maximum
10	40	43	45	50

A box plot is constructed as follows.



Box plots are *particularly useful for highlighting differences between groups*.

Example 7



It clearly shows that although there is overlap between the three sets of data, the first and second datasets contain roughly similar responses and that these are quite different from those in the third set. Note that the asterisks (*) at the ends of the whiskers is the way Minitab highlights outlying values.

3.4 Exercises

- Recall the following data from Exercise 1 in Chapter 2 on the weight (in *kg*) of 50 sacks of potatoes leaving a farm shop.

8.1	8.2	8.5	8.7	8.8
8.9	9.2	9.3	9.3	9.4
9.5	9.5	9.6	9.6	9.6
9.7	9.7	9.9	9.9	10.0
10.0	10.0	10.0	10.0	10.1
10.2	10.2	10.2	10.3	10.3
10.4	10.4	10.4	10.5	10.6
10.6	10.6	10.6	10.6	10.7
10.8	10.9	11.0	11.2	11.3
11.3	11.3	11.5	11.6	12.8

- Calculate the mean of the data.
 - Calculate the median of the data.
 - Calculate the range of the data.
 - Calculate the inter-quartile range.
 - Calculate the sample standard deviation.
 - Draw a box plot for these data and comment on it.
 - Put the data in a grouped frequency table.
 - Find the modal class.
- Chloe collected the following data on the weight, in grams, of “large” chocolate chip cookies produced by Millie’s Cookie Company.

27.1 22.4 26.5 23.4 25.6 26.3 51.3 24.9 26.0 25.4

To summarise, Chloe was going to calculate the mean and standard deviation for this sample. However, her friend Mark warned her that the mean and standard deviation might be inappropriate measures of location and spread for these data.

- Do you agree with Mark? If so, why?
 - Calculate measures of location and spread that you feel are more suitable.
- An internet marketing firm was interested in the amount of time customers spend on their website. They recorded the lengths of visits to the website for a sample of 100 customers and whether the customer was male or female. The standard deviations of the lengths of visits were 12.2 seconds for males and 18.5 seconds for females. Which group has the more variable visit lengths, based on this sample, males or females?