CEG2002

Statistics and Numerical Methods for Civil Engineers

Part I: Statistics

Semester 2, 2011–2012

Lecturer: Jian Qing SHI

CEG2002

Statistics for Civil Engineers

Dr. Jian Qing SHI

Room 2.29, Level 2, Herschel Building, Tel: 222 7315

Email: j.q.shi@ncl.ac.uk

http://www.staff.ncl.ac.uk/j.q.shi/teaching/CEG2002/

- This course includes two parts:
  - Part I (weeks 1-5, Dr Jian Qing Shi): Statistics.
  - Part II (weeks 6-10, Dr David Swailes): Numerical Methods.
  - Revision (Week 11: Numerical Methods; Week 12: Statistics)
- Assessment

Assignment/project	20% (10% for each part).
Exam (2hrs)	80%

- Arrangement for part I (Statistics)
  - Lectures: Monday 11am-12pm (STB F13), Wednesday 10am-11am (STB. F16)
  - Tutorials: Tuesday 5pm-6pm (Weeks 3 and 5, STB F13)
  - Computer Practical: Monday 1pm-2pm (Weeks 2, 4 and 6, STB Tree, room 3.04C)
  - Office hours: Monday 12pm-1pm and Wednesday 11am-12pm (room 2.29, Herschel Building)
  - References
    - Montgomery, D. C. and Runger, G. C. (2006). Applied Statistics and Probability for Engineers (4th edition). Wiley.
    - Ross, S. M. (2004). Introduction to Probability and Statistics for Engineers and Scientists (3rd edition). Academic Press.

Contents

1 Collecting, Presenting and Summarising data

1.1 Introduction

Data are the key to many important engineering decisions:

- Is a sea–wall high enough to withstand extreme sea levels which are expected to occur once every fifty years?
- Or once every five hundred years?
- Is such a defence *strong enough*?
- What are the design requirements for a lift which should be able to carry, on average, 17 people?

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

We begin this course by looking at some basic methods of **collecting**, **representing** and **describing** data.

1.1.1 Definitions

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

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

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

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

Example

If we were interested in the height of people doing Civil Engineering courses at Newcastle,

- our **random variable** would be X: height of people doing Civil Engineering at Newcastle
- an **observation** would be a particular person's height
- if we measured everyone doing CEG2002, those would be our **data**

- these data would form a **sample** from the **population** of all students registered with the School of Civil Engineering.

In practice it is difficult to observe whole populations, unless we are interested in a very limited population, e.g. the students taking CEG2002. In reality we usually observe a subset of the population; we will come back to **sampling** shortly.

1.1.2 Types of data

Variables can be classified into two types: **qualitative** and **quantitative**.

- *Qualitative* variables have non–numeric outcomes. They are usually **categorical**.

  Examples: sex of a person, colour of a car, mode of transport, football team supported,...

- *Quantitative* variables have numeric outcomes with a natural ordering.

  Examples: people's height, time to failure of a component, the number of defective components in a batch,...

Quantitative variables are usually one of two types: **discrete** or **continuous**.

Discrete random variables

- can only take a sequence of distinct values which are usually integers
- are **countable**
- **Examples:** number of defective pieces in a manufacturing batch, the number of people in a tutorial group, a person's shoe size, ...
- There are other kinds of discrete data. **Ordinal** data (data which are ordered) usually take numeric values, but these values are not really numbers in the usual sense.

Continuous random variables

- can take any value over some continuous scale
- can be measured to differing degrees of accuracy using different equipment; but we can never say *absolutely*, *precisely*, how much someone weighs (for example)
- are often expressed up to a number of significant digits and could appear to be discrete
- **Examples:** height, weight, wind speed, compressive strength, or the fuel consumption of a car

It is the underlying variable which defines their status and not the form in which they are expressed.

1.2 Sampling (self-study)

We can rarely observe the whole **population**. Instead, we observe some sub–set of this called the **sample**. The difficulty lies in obtaining a **representative** sample.

For example, if you were to ask the 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 can not be stressed too highly.

There are three general forms of sampling technique:

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.

1.2.1 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.
- 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**.

Disadvantages

- we don't often have a complete list of the population (a **sampling frame**).
- For example, if you were surveying the market for some new PC software, the population would be everybody with a compatible computer. It would be almost impossible to find out this information.
- Not all elements of the population are **equally accessible** – you could waste time and money trying to obtain data from people who are unwilling to provide it.
- It is possible that, purely by chance, you pick an **unrepresentative sample**

1.2.2 Stratified Sampling

This is a form of random sampling where clearly defined groups, or **strata**, exist, for example:

- **Males and females**
- **Age groups**
- **Employment status**

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 according to these proportions.

Example
If we know the population is 55% female and 45% male, and we want a sample of 1000, we would

(i) decide to have 550 females and 450 males in our sample, and then

(ii) pick the members of our sample from their respective groups randomly.

Disadvantages

- Clear information on the size and composition of each group or stratum is needed – which can be difficult to obtain.
- We still need to have a list of the entire population so we can sample from it (as with simple random sampling).

1.2.3 Systematic Sampling

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

Example
If you were interested in obtaining a 10% sample from a batch of components being manufactured, you would:

(i) select the first component at **random**

(ii) after that you would pick **every tenth** item to come off the production line

The simplicity of selection here makes this a particularly easy sampling scheme to implement, especially in a production setting.

Disadvantages

- It is not (completely) random, and if there is a pattern in the process it may be possible to obtain a **biased sample**.
- It is only really applicable to **structured populations**.

1.2.4 Cluster Sampling

In cluster sampling, we divide the population into groups known as **clusters** (we could sub–divide a large geographical area into smaller areas, for example).

We then take a **simple random sample** of clusters, and take every member of these selected clusters to obtain our final sample.

The advantage of this method is that, because the sampling takes place in a concentrated area, it is relatively inexpensive to perform.

Disadvantages

- The very fact that small clusters are picked to allow the entire cluster to be surveyed introduces the strong possibility of **bias** within the sample.
- For example, If you were interested in the take up of organic foods and were sampling via the cluster method you could easily get biased results:
  - You could, by chance, pick an economically deprived area
  - The proportion of those surveyed that ate organically might thus be very low
  - If you picked a middle class suburb the proportion is likely to be higher than the overall population.

1.2.5 Judgemental sampling

Here, the person interested in obtaining the data decides whom they are going to ask. This can provide a coherent and focused sample by choosing people with experience and relevant knowledge to provide their opinions.

For example, the head of a service department might suggest particular clients to survey based on his judgement. They might be people he believes will be honest or have strong opinions.

This methodology is **non–random** and relies on the judgement of the person making the choice, and hence it cannot be guaranteed to be **representative**. It is prone to **bias**.

1.2.6 Sample Size

When considering collecting data, it is important to ensure that the sample contains a sufficient number of members of the population for adequate analysis to take place.

Larger samples will generally give **more precise information** about the population.

Unfortunately, in reality, questions of **expense** and **time** tend to limit the size of the sample it is possible to take.

For example, national opinion polls often rely on samples in the region of 1000.

1.3 Frequency tables

The following table presents the modes of transport used daily by 30 students to get to and from 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

The table obviously contains much information. However, it is difficult to see which method of transport is the most widely used.

Idea: count the number of students using each mode of transport!

Mode	Frequency	Relative Frequency (%)
Car	10	33.3
Walk	7	23.4
Bike	4	13.3
Bus	4	13.3
Metro	4	13.3
Train	1	3.4
Total	**30**	**100**

This gives us a much clearer picture of the methods of transport used.

Relative frequency

Also of interest might be the ***relative*** frequency of each of the modes of transport.

This is simply the frequency expressed as a proportion of the total number of students surveyed. If this is given as a percentage, as here, this is known as the ***percentage relative frequency***.

Example

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

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

Cars Sold	**Tally**	**Frequency**	**Relative Frequency %**
1		0	0
2		0	0
3		0	0
4	l	1	7.14
5	l	1	7.14
6	ll	2	14.29
7	ll	2	14.29
8	lll	3	21.43
9	ll	2	14.29
10	ll	2	14.29
11	l	1	7.14
Totals	**14**	**14**	**100**

Continuous data frequency tables

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

For example, the following data set represents rainfall totals (in mm) for 50 recording

stations across the Deep South over a 48 hour period during Hurricane Katrina:

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

This is what we do...

1. Divide the range of the variable into smaller ranges called **class intervals**

2. There should be **no gaps** between these intervals

3. The class interval width should be a convenient number (e.g. 5, 10, 100, depending on the data)

4. You should aim for no more than about **ten to fifteen classes**

For the Deep South data, we might get:

Class Interval	**Tally**	**Frequency**	**Relative Frequency %**
175–179.9999	|	1	2
180–184.9999	|||	3	6
185–189.9999	|||	3	6
220–224.9999			
Totals		**50**	**100**

1.4 Summarising data – Graphical summaries

We will consider six different methods of summarising data **graphically**:

- Stem–and–leaf plots

- Bar charts

- Histograms
- Time series plots
- Scatterplots
- Box–and–whisker plots

We will look at the first five of these now, and consider the last of these – the **box–and–whisker plot** – towards the end of the chapter.

1.4.1 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... (in lecture)

...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
- These interval widths must be **equal**
- You should use **sensible** values for the interval widths

Example

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

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 $= 1$.

Remarks.

- "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: Cuban rainfall data The following numbers show the day–to–day percentage change in rainfall totals for Havana, Cuba, for 23 consecutive days in August:

–0.2	**–2.1**	**1.0**	**0.1**	**–0.5**	**2.4**	**–2.3**	**1.5**
1.2	**–0.6**	**2.4**	**–1.2**	**1.7**	**–1.3**	**–1.2**	**0.9**
0.5	**0.1**	**–0.1**	**0.3**	**–0.4**	**0.5**	**0.9**	

- The largest value is 2.4 and the smallest –2.3, and we have lots of decimal values in between
- It seems sensible to have a **stem unit** of 1 and a **leaf unit** of 0.1

A stem and leaf diagram for this set of returns might look like:

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

Stem Leaf

$n = 23$, stem unit $= 1$, leaf unit $= 0.1$.

Remark: We should select stem unit carefully.

Example 2: Dam flows It all looks pretty easy... so what can go wrong?

Consider the following data, which are the total monthly flows (October 2006), in cubic metres per second, through the spillways of ten dams in southern India:

17 18 15 14 12 19 20 21 24 15

If you were to choose 10 as the interval width (i.e. go up in 10s), the stem and leaf plot would look like

Stem	**Leaf**
1	2 4 5 5 7 8 9
2	0 1 4

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

The interval width is **too large**! Two intervals is not enough to reveal any **patterns** in the data

Idea: Use an interval width of 5!

Stem	**Leaf**
1	2 4
1	5 5 7 8 9
2	0 1 4

$n = 10$, stem unit = 5, leaf unit = 1.

We can see the pattern more clearly now!

1.4.2 Bar charts

Bar charts are a commonly–used and clear way of presenting categorical 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 Recall the example on students' modes of transport:

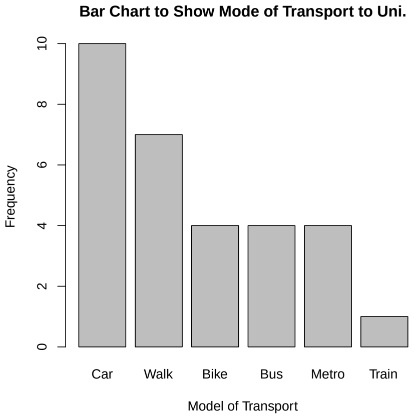
Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

Draw a bar chart to represent these data

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 present this information as a bar chart.



1.4.3 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*
- The *area* of the rectangle is proportional to the frequency – not the height

Initially we will only consider histograms with equal class intervals.

Example Consider the following data, which show rainfall totals (in mm) for fifty recording stations across the Deep South over a 48 hour period during Hurricane Katrina in August 2005.

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

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.

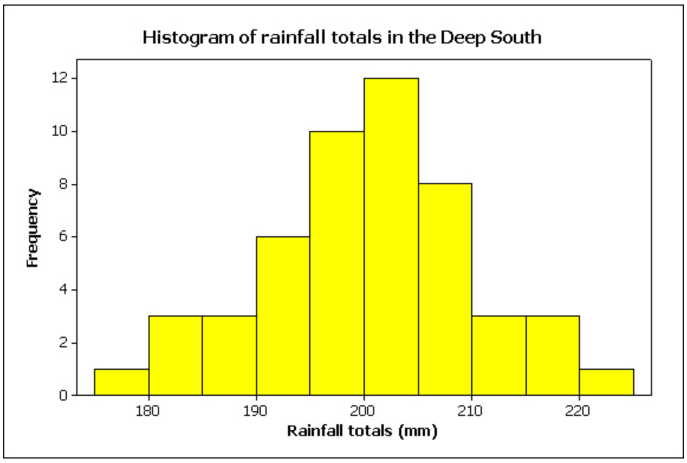
Rainfall totals x	**Frequency**
$175 \leq x < 180$	1
$180 \leq x < 185$	3
$185 \leq x < 190$	3
$190 \leq x < 195$	6
$195 \leq x < 200$	10
$200 \leq x < 205$	12
$205 \leq x < 210$	8
$210 \leq x < 215$	3
$215 \leq x < 220$	3
$220 \leq x < 225$	1
Total	**50**

Once we have the frequency table, the process is very similar to before...

– Find the maximum frequency and draw the **vertical** (y–axis) from zero to this value, including a sensible numeric scale

– The range of the **horizontal** (x–axis) needs to include not only the full range of observations but also the full range of the class intervals from the frequency table.

– Draw a bar for each group in your frequency table. These should be the same width and touch each other



Percentage relative frequency histograms

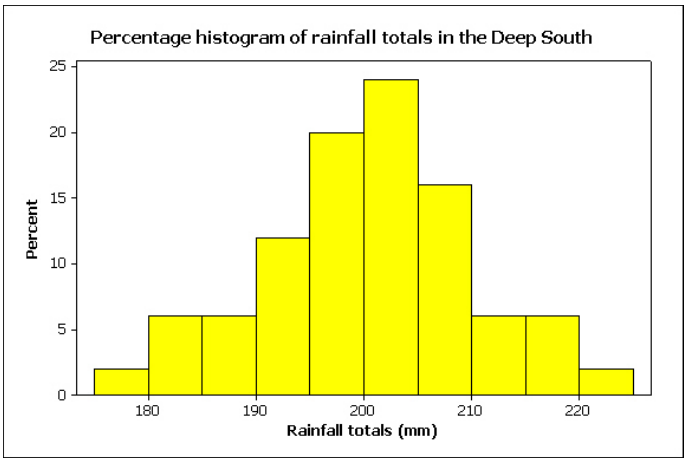
These are an extension to the **percentage relative frequency tables** we have already considered.

To produce a percentage relative frequency histogram, we follow the same procedure as for a regular histogram, but use *percentage frequencies* instead of just *frequencies*.

A percentage relative frequency table for the Deep South data is:

***Rainfall totals** x*	***Frequency***	***Relative Frequency (%)***
$175 \leq x < 180$	1	2
$180 \leq x < 185$	3	6
$185 \leq x < 190$	3	6
$190 \leq x < 195$	6	12
$195 \leq x < 200$	10	20
$200 \leq x < 205$	12	24
$205 \leq x < 210$	8	16
$210 \leq x < 215$	3	6
$215 \leq x < 220$	3	6
$220 \leq x < 225$	1	2
Totals	**50**	**100**

Instead of using **frequency** on the vertical axis (y-axis), you *could* use the **percentage relative frequency**.



Why use percentage relative frequency? **It is useful for comparing two or more histograms**

1.4.4 Time Series Plots

So far we have only considered data where we can **ignore the order** in which the data come.

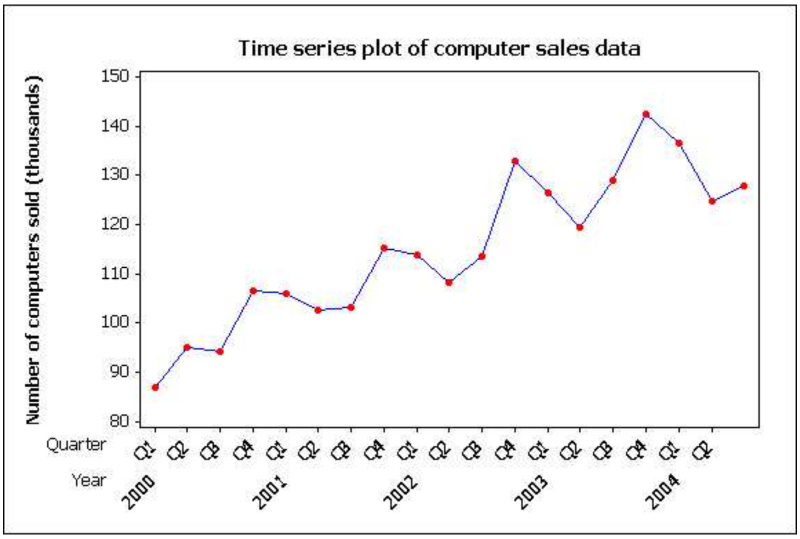
Not all data are like this: one exception is data which have been collected over **time**.

- Monthly sales of a product
- The price of a share at the end of each day
- The air temperature at midday each day

Such data can be plotted simply using **time** as the x-axis.

Example Consider the following data on the number of computers sold (in thousands) by quarter at a large warehouse outlet.

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



What information you can find from the above time series plot? It will be discussed in the lecture.

1.4.5 Scatter plots

The final type of graph we are going to look at (for now!) is the **scatter plot**.

Such graphs are used to plot two variables which you believe might be **related** – for example:

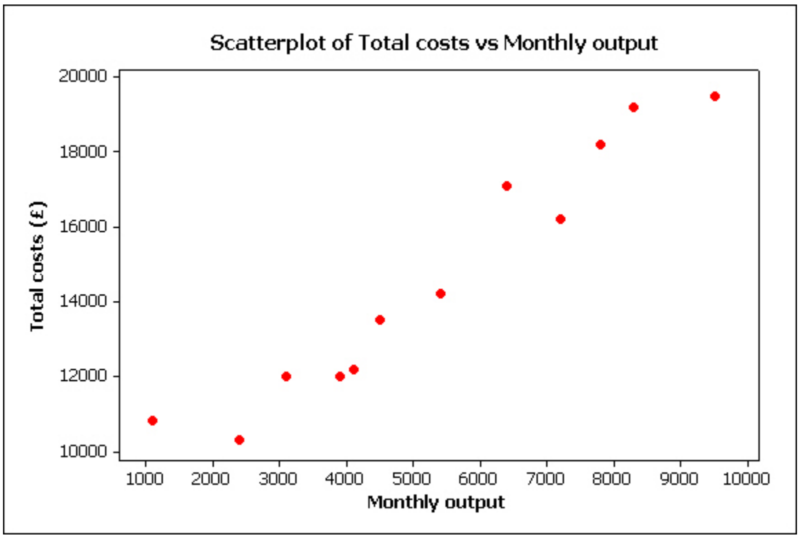
- height and weight
- advertising expenditure and sales
- age of machinery and maintenance costs

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

Total costs (£)	**Monthly Output**
10300	2400
12000	3900
12000	3100
13500	4500
12200	4100
14200	5400
10800	1100
18200	7800
16200	7200
19500	9500
17100	6400
19200	8300

If you were interested in the relationship between the cost of production and the number of units produced you could easily plot this by hand.

1. The “ **response**” variable is placed on the y-axis. Here the response variable is “total costs”
2. The variable that is used to try to explain the response variable (here, monthly output) is placed on the x-axis – this is the **explanatory** variable
3. Plot the pairs of points on the graph



What can we find from the above scatter plot?

1.5 Summarising data – Numerical summaries

If you summaries data numerically, it is essential that you calculate both

- a measure of **location** (or "average"), and
- a measure of **spread**

There are three measures of **location** which are commonly used:

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

We will consider these in turn.

1.5.1 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

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

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.

Calculate a mean from a frequency table

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

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

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

We can express this calculation of the sample mean from **discrete** 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 $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$.

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

Calculate a mean from a grouped frequency table

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.

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?

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

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

1.5.2 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:

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

For example, if our data were:

2 3 3 5 6 7 9

then the sample size ($n = 7$) is an odd number. Thus, the median is the

$$\frac{7+1}{2} = 4^{th} \text{ smallest observation,}$$

that is, the median is the fourth smallest ranked observation.

For these data the median = 5.

If the sample size (n) is an **even** number the process is slightly more complicated:

$$\begin{aligned}\text{median} &= \text{average of the } \left(\frac{n}{2}\right)^{th} \\ &\quad \text{and the } \left(\frac{n}{2}+1\right)^{th} \text{ smallest observations.}\end{aligned}$$

For example, if our data were:

2 3 3 5 6 7 9 10

then the sample size ($n = 8$) is an **even** number and therefore

$$\begin{aligned}\text{median} &= \text{average of the } \left(\frac{8}{2}\right)^{th} \\ &\quad \text{and the } \left(\frac{8}{2}+1\right)^{th} \text{ smallest observations} \\ &= \frac{5+6}{2} \\ &= 5.5.\end{aligned}$$

1.5.3 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.

1.5.4 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!

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

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

1. the **sample variance**
2. the **range**
3. the **inter–quartile range**

The Sample 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

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

which can be expressed to

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

Or equivalently as

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

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 unit as the original data and so is perhaps easier to understand.

Remark. Most calculators will give the sample standard deviation when in **SD** mode.

Example. Consider the following 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

We have already calculated the sample mean as $\bar{x} = 9.73$. Now

$$\begin{aligned} \sum x^2 &= 8.4^2 + 8.7^2 + \cdots + 11.4^2 = 1904.38 \\ n(\bar{x})^2 &= 1893.458 \end{aligned}$$

and so the sample variance is

$$s^2 = \frac{1}{19}(1904.38 - 1893.458) = 0.57484$$

and the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{0.57484} = 0.75818.$$

Remark: A different calculation is needed when the data are given in the form of a grouped frequency table with frequencies (f_i) in intervals with mid–points (m_i).

First the sample mean $\bar{x}$ is approximated (as described earlier) and then the sample variance is approximated as

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

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.

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** 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 inter–quartile range (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**, $Q1$) has 25% of the data below it

- The second quartile (**median**, $Q2$) has 50% of the data below it
- The third quartile (**upper quartile**, $Q3$) has 75% of the data below it

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

$$\begin{aligned} Q1 &= \frac{(n+1)}{4}\text{th smallest observation} \\ Q3 &= \frac{3(n+1)}{4}\text{th smallest observation.} \end{aligned}$$

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

For example, in a dataset with $n = 20$ values:

- the lower quartile is the $5\,\frac{1}{4}$th smallest observation
- the upper quartile is the $15\,\frac{3}{4}$th smallest observation

Example Consider again the data

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

Here the 5th and 6th smallest observations are 9.2 and 9.3.

Therefore, the **lower quartile** is $Q1 = 9.225$.

Similarly, the **upper quartile** is the $15\,\frac{3}{4}$ smallest observation, that is, three quarters of the way between 10.3 and 10.4; so $Q3 = 10.375$.

The **inter–quartile range** is simply the difference between the upper and lower quartiles, that is

$$IQR = Q3 - Q1$$

which for these data is $IQR = 10.375 - 9.225 = 1.15$.

1.6 Box and Whisker Plots

Box and whisker 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**.

The plot is constructed as follows:

- Lay out an x–axis for the full range of the data
- Draw a rectangle with ends at the the **upper** and **lower quartiles** (the "box")
- Split the rectangle in two using the **median**
- Draw lines from the "box" to the **minimum** and **maximum** values (the "whiskers")

Example Draw a Box and Whisker plot for data with the following summaries:

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

This example will be discussed in the lecture.