



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

- 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

- 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 .

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 $x_2 = 0$.

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.

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 Σ is the upper case of the Greek letter **sigma**. It is used to represent the phrase “**sum the values**”.

For example:

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:

$$\sum_{i=1}^n x_i =$$

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:

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

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:

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

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 (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 (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:

Example 2

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

0 0 6 10 16 21 52

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

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

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

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

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

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

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

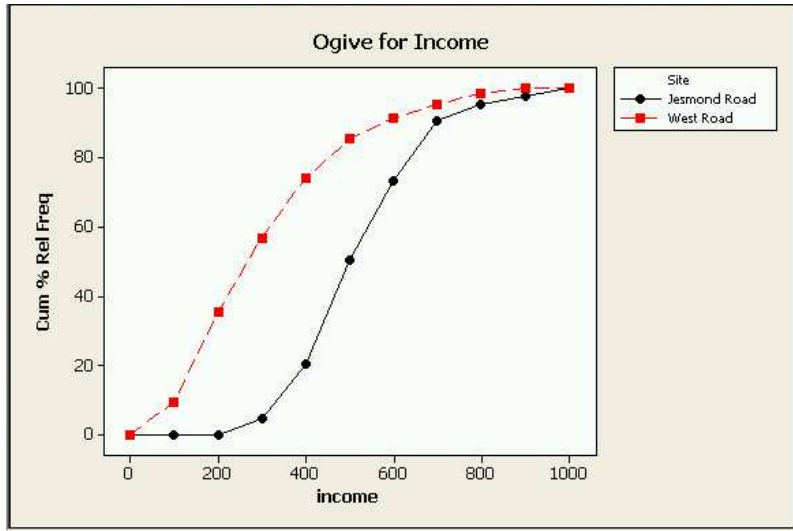
Median = 14 units per week

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

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,

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

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. $Q_1 = \mathbf{9.275}$.

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 :

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

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. $Q_3 = \mathbf{10.325}$.

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

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

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”:

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”:

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

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:

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

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

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

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

