

Lecture 4

NUMERICAL SUMMARIES FOR DATA

Don't forget...

CBA1 goes live in exam mode today!

Deadline: Midnight this Friday!

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.

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

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 .

Σ 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:

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

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 Σx .

Other “mathsy” stuff

Powers

- “3 squared”, or 3^2 , means 3×3 , which is 9
- “3 to the power 7”, or 3^7 , means $3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3$, which is 2187
- So “ x squared”, or x^2 , means $x \times x$
- “ x to the power 4” means $x \times x \times x \times x$
- More generally, “ x to the power k ” means you multiply x by itself k times

Brackets

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 $(\sum x)^2$

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.

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

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

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.

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

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

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

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

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 $x_{(1)}, x_{(2)}, \dots, x_{(k)}$.

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

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

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

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:

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

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

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

that is, the median is the fourth largest (or smallest) ranked observation.

For these data the median = 5.

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

median = average of the $\left(\frac{n}{2}\right)^{th}$
and the $\left(\frac{n}{2} + 1\right)^{th}$ largest observations.

For example, if our data were:

2 3 3 5 6 7 9 10

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

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

It is possible to estimate the median value from an **ogive** 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 ogive drawn.

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.

For example, consider the following data:

2 2 2 3 3 4 5

Quite obviously the mode is 2 as it occurs most often.

We often talk about modes in terms of categorical data. Recalling the newspaper example, the mode was The Sun, as it was the most popular paper.

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 $30 \leq x < 40$.

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 **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.

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**, 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

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.

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

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

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

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 $IQR = 10.375 - 9.225 = 1.15$.

The interquartile range can also be **estimated** from the ogives in a similar manner to the median:

- Draw the ogive
- Read off the values for 75% and 25%
- Calculate the difference between these two values

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

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

which can be simplified to

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

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 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.

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

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

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

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

$$\sum x^2 = 8.4^2 + 8.7^2 + \cdots + 11.4^2 = 1904.38$$

$$n(\bar{x})^2 = 1893.458$$

and so the sample variance is

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

and the sample standard deviation is

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

First the mean...

$$n = 10$$

$$\sum x = 45$$

$$\bar{x} = \frac{\sum x}{n} = 4.5$$

Now draw up a table...

x_i	$(x_i - \bar{x})$ (method 1)	$(x_i - \bar{x})^2$ (method 1)	x_i^2 (method 2)
1	-3.5	12.25	
1	-3.5	12.25	
2	-2.5	6.25	
3	-1.5	2.25	
5	0.5	0.25	
6	1.5	2.25	
6	1.5	2.25	
6	1.5	2.25	
7	2.5	6.25	
8	3.5	12.25	
Totals	0	58.5	

Now the variance (method 1)...

$$\sum (x - \bar{x})^2 = 58.5$$

$$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$$

$$= \frac{1}{9} \times 58.5$$

$$= 6.5$$

Or alternatively...

x_i	$(x_i - \bar{x})$ (method 1)	$(x_i - \bar{x})^2$ (method 1)	x_i^2 (method 2)
1	-3.5	12.25	1
1	-3.5	12.25	1
2	-2.5	6.25	4
3	-1.5	2.25	9
5	0.5	0.25	25
6	1.5	2.25	36
6	1.5	2.25	36
6	1.5	2.25	36
7	2.5	6.25	49
8	3.5	12.25	64
Totals	0	58.5	261

So using the alternative formula (method 2)...

$$\sum x^2 = 261$$

$$s^2 = \frac{1}{n-1} \left\{ \sum x^2 - n(\bar{x})^2 \right\}$$

$$= \frac{1}{9} \{ 261 - 10 \times 4.5^2 \}$$

$$= 6.5$$

Now the standard deviation...

$$s = \sqrt{6.5} = 2.55$$

which is in the original units of the data!

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”)

Draw a stem-and-leaf plot for data with the following summaries:

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