MSCI152: Introduction to Business Intelligence and Analytics

Lecture 6: Measures of Location

Lancaster University Management School

Overview

- Descriptive Measures: Location

Summary Statistics

Location:

- Mean
- Median
- Mode

Spread:

- Standard deviation
- Range
- Percentiles and Quartiles

Each measures a slightly different characteristic and so tells us something different about the data

Measures of Location: Averages

People **like** averages, they are all around

- Reduces all the variability into a single number
- We usually think of an average as a typical value, middle value, normal value, central value

Examples of use:

- What is the average first-job salary?
- What is the average goals scored by my favourite team?
- Am I taller than the average?
- What is the average mark for the course?
- What is the average time to wait for a bus?
- What is the average temperature in Lancaster in February?

Averages: Terminology

People sometimes use "*average*" to refer specifically to the mean, e.g.:

- Excel
- The average temperature last month was . . .

People sometimes use "*average*" to refer more generally to a measure of the central or typical value (i.e., mean, median or mode). Sometimes it is vague:

- Average teenager uses their phone for x hours per day

The dictionary definitions include both meanings

Averages: A word of caution. . .

Averages are useful but do not tell the whole story:

- The mean daily rainfall in Lancaster is 2.9mm per day
- Therefore we should build flood defences to cope with this level of rainfall

Why is this incorrect?

Averages: A word of caution...

Averages are useful but do not tell the whole story:

- In a campus shop the mean number of customers per hour is 30
- The mean number of customers that a shop assistant can serve per hour is 30
- Therefore we need 1 shop assistant on duty

Why is this incorrect?

Some Notation: Sums

I have a sample of 5 data points:

$$2, \quad 3, \quad 7, \quad 3, \quad 5.$$

If I want to add these number together, I get

$$2 + 3 + 7 + 3 + 5 = 20.$$

What if we had a 1,000 data points and wanted to add them together and write down the calculation?

- That would be a lot of effort for little gain!
- Notation allows us to express the same information more simply, but retain meaning
- We just need to learn this new language and alphabet!

I'm sure that we all know what the following symbol means

$$\text{“} = \text{”}$$

Did you know it was invented in 1557 by Robert Recorde?

Some Notation: Sums

I have a sample of n data points

- So the sample can be of any size "n"

I need way to refer to any data point in our sample, and I use $x_i, \; i = 1, \ldots, n$

- Hence, value of the i^{th} data point is x_i

Using the data on the previous slide our data are

$$x_1 = 2, \quad x_2 = 3, \quad x_3 = 7, \quad x_4 = 3, \quad x_5 = 5.$$

So x_1, represents the first sampled measurement from the first element from your sample, and similarly for the other observed values, $x_i, \; i = 2, \ldots, 5$.

Some Notation: Sums

If I want to add together these 5 data points I can now write

$$x_1 + x_2 + x_3 + x_4 + x_5 = 20.$$

Still not useful with a lot of data! We use the symbol "Σ" to donote a **sum** in the following way:

$$\sum_{i=1}^{5} x_i = 20.$$

In words, the L.H.S. reads "*the sum of the data points x_1 to x_5*".

More generally, we can write

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n.$$

Measures of Location: Mean

The most commonly used measure of the centre of the data

Add up all the values and divide by the number of elements

We use "$\bar{x}$" to denote the mean:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Example:

$$\begin{aligned}\bar{x} &= \frac{29 + 12 + 14 + 21 + 12 + 42 + 10}{7} \\ &= \frac{140}{7} \\ &= \mathbf{20}.\end{aligned}$$

Excel: AVERAGE() function

Measures of Location: Median

The median is the **middle value when the values are arranged in numerical order**:

- Arrange the sample, of size n, in **increasing** numerical order
- Let x_i be the i^{th} value in the ordered list
- The middle value is given by

$$x_{\frac{(n+1)}{2}}$$

Excel: MEDIAN() function

Median: Odd number of elements

Our sample is of size $n = 7$ – an odd number:

$$29, \quad 12, \quad 14, \quad 21, \quad 12, \quad 42, \quad 10$$

Step 1: Re-arrange in increasing numerical order

$$10, \quad 12, \quad 12, \quad 14, \quad 21, \quad 29, \quad 42$$

Step 2: For 7 elements, select the

$$\left(\frac{(n+1)}{2}\right)^{\text{th}} \text{ value } = \left(\frac{(7+1)}{2}\right)^{\text{th}} \text{ value } = 4^{\text{th}} \text{ value}$$

Median $= \mathbf{14}$

[Can you see why we use the $\{(n+1)/2\}^{\text{th}}$ value?]

Median: Even number of elements

Our (ordered) sample is of size $n = 8$ – an even number:

$$10, \quad 12, \quad 12, \quad 14, \quad 21, \quad 29, \quad 42, \quad 49$$

We require the

$$\left(\frac{(n+1)}{2}\right)^{\text{th}} \text{ value } = \left(\frac{(8+1)}{2}\right)^{\text{th}} \text{ value } = 4.5^{\text{th}} \text{ value}$$

We have to average the $(n/2)^{\text{th}}$ and the $\{(n+2)/2\}^{\text{th}}$ (i.e., the 4^{th} and 5^{th} values)

$$\textbf{Median } = \frac{x_4 + x_5}{2} = \frac{14 + 21}{2} = \mathbf{17.5}.$$

[Can you see why we use the $\{(n+1)/2\}^{\text{th}}$ value?]

Median: Ordinal Data

To calculate the median for ordinal data, order data by category and return the category where $x_{\{(n+1)/2\}}$ lies

Example: 85 people rated the service at a hotel as follows

Rating	Frequency	Elements
Very Good	37	1–37
Good	22	38–59
Average	18	60–77
Poor	6	78–83
Very Poor	2	84–85

Middle element:

$$\frac{(n+1)}{2} = \frac{86}{2} = 43^{\text{rd}} \text{ element}$$

Median Rating = Good

Comparing the Mean and Median

Mean

- Uses all of the data
- Weights all of the data equally
- Affected by **extreme values**

Median

- Essentially uses the ranking of the data, not the values
- Not affected by **extreme values**

Mean and Median: Interpret Carefully!

A small shop employs 3 part-time staff and 3 full-time staff. Part-time staff work 4 hours per day. Full-time staff work 8 hours per day. What is the mean and median hours per day?

- **Data values:** 4, 4, 4, 8, 8, 8
- **Mean** = 6, **Median** = 6
- These values are not in the data set

Note:

- If there was 1 extra part-time staff the median would be 4
- If there was 1 extra full-time staff the median would be 8.

Measures of Location: Mode

Mode is the value that **occurs most frequently**

It is used less often, just where it is **meaningful** to think about the **most common value**. Usually fine for discrete data or grouped data, but not continuous data.

One example is in elections (if **first past the post** system).

Mode: MODE() function

Measures of Location: Mode

Example1:

$$10, \quad 12, \quad 12, \quad 14, \quad 21, \quad 29, \quad 42.$$

Mode $=$ **12**.

Example2:

$$10, \quad 12, \quad 12, \quad 14, \quad 21, \quad 21, \quad 29, \quad 42.$$

Mode $=$ **12** and **21**

We call this **bimodal**, more modes is **multi-modal**

Mode: Love it or hate it?

Survey data about a product:

Hate it 37%, Don't like 8%, Neutral 2%, Like 22%, Love it 31%.

Mode = Hate it"

Combine the categories:

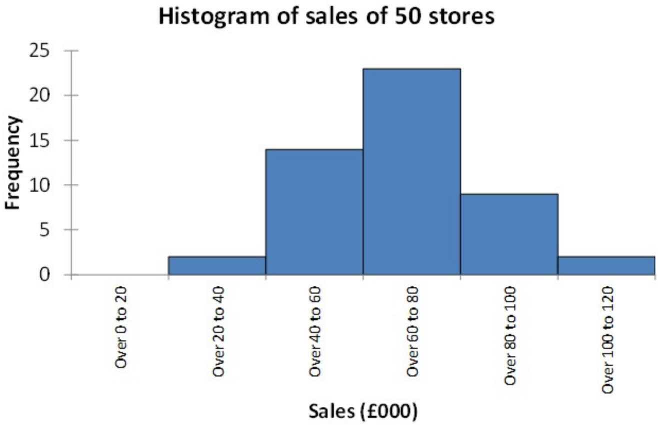
- Dislike (hate & don't like) 45%
- Neutral 2%,
- Like (like & love) 53%

Mode = Like it

Mode can depend on how we group the data!

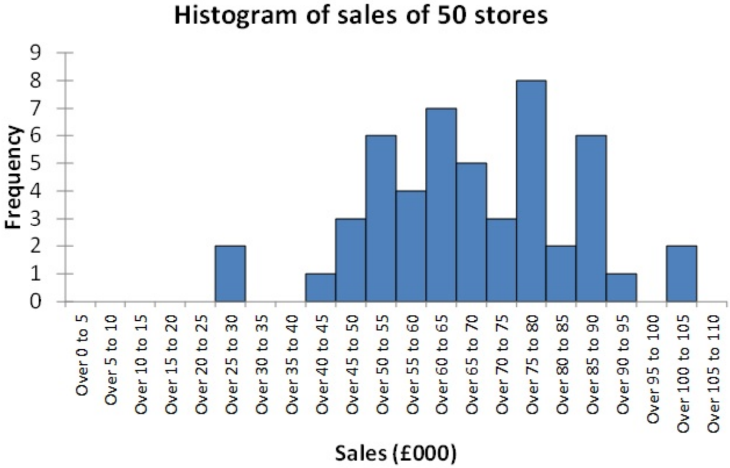
Mode: Peaks in the data

Mode: Highest interval



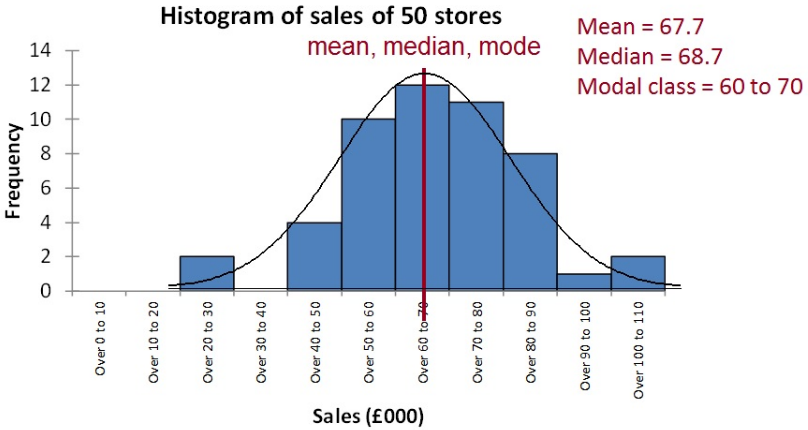
Depends on way histogram is drawn: This graph has 1 mode

Mode: Peaks in the data



Depends on way histogram is drawn: One mode or multi-mode?

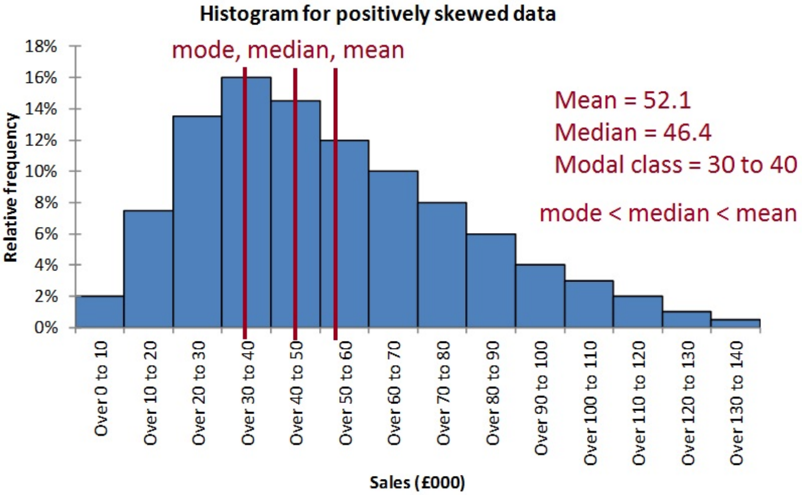
Distributions: Approximately Normal



Symmetrical and **bell-shaped**:

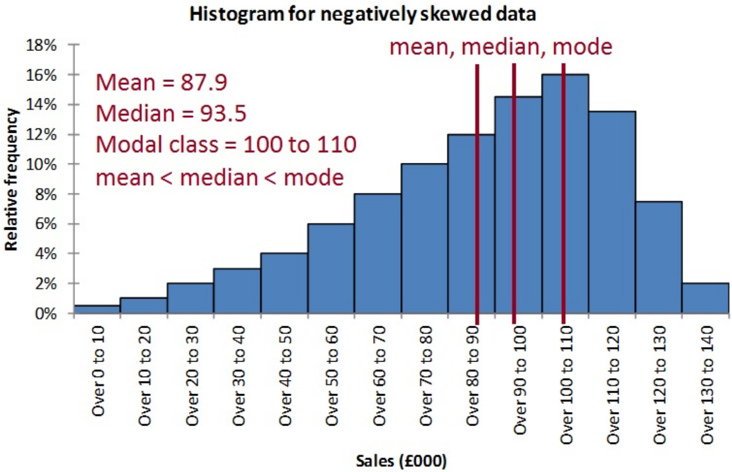
- Mean, Median, Mode are about the same.

Distributions: Positive Skewness



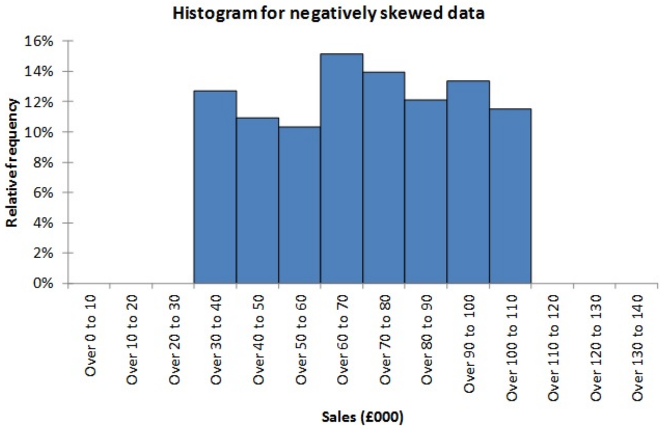
Positive Skewness: Longer "tail" to the right

Distributions: Negative Skewness



Negative Skewness: Longer "tail" to the left

Distributions: Approximately Uniform

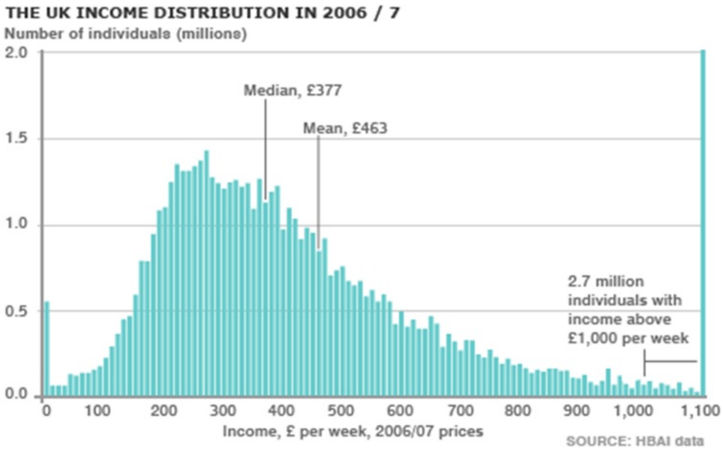


Symmetrical, roughly **equal frequencies** in each interval within the range:

- Mean and Median about the same
- Mode could be any interval, by chance

UK Income (£per week)

Source: http://news.bbc.co.uk/1/hi/magazine/7581120.stm



Why is the median often used in discussions about income?

- About 2/3 of people earn less than the mean

Wrap up

Here we:

- Looked at statistical measures of **location**

Next time:

- Statistical measures of **spread**