EMPL4005

INFORMATION, DATA AND ANALYSIS

Descriptive Statistics:
1. Central Tendency

PRIFYSGOL CYMRU
Y Drindod Dewi Sant
UNIVERSITY OF WALES
Trinity Saint David

There are 2 broad categories of descriptive statistics

- Measures of central tendency (This presentation) describe the centre point of a data set with a single value
  - Mean (i.e. average)
  - Median
  - Mode
- Dispersion measures how far individual data values have strayed from the mean. (Pres 9)
  - Range
  - Quartiles
  - Standard deviation (and variance)

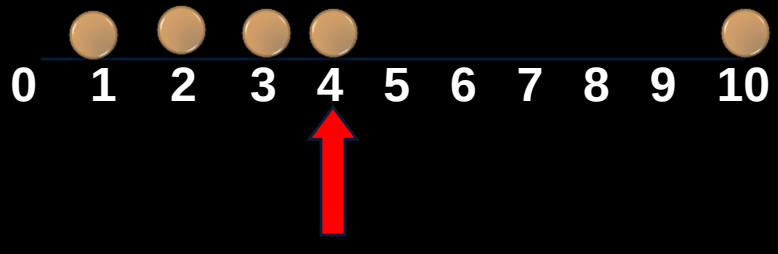
The Mean

- The most common measure of central tendency
- Mean = sum of all values divided by the number of values
- Affected by extreme values (outliers)



Mean = 3

$$\frac{1+2+3+4+5}{5}=\frac{15}{5}=3$$



Mean = 4

$$\frac{1+2+3+4+10}{5}=\frac{20}{5}=4$$

The Mean

- 10 students were asked how many hours they spent online in a week
- The answers were: 3 7 4 9 5 4 6 17 4 7
- 3+7+4+9+5+4+6+17+4+7=66
- 66 = Sum of answers
- This is divided by the number of students (10) to arrive at the Mean (average) = 6.6
-

The Weighted Mean

- A weighted mean allows you to assign more importance to certain values and less importance to others
- E.g. If a student's module grade was made up of: 50% exam, 35% project and 15% homework

Type	Score	Weight (Percent)
Exam	94	50
Project	89	35
Homework	83	15

THE WEIGHTED MEAN

Type i	Score x_i	Weight w_i	Weight × Score $(w_i x_i)$
Exam 1	94	0.50	47.0
Project 2	89	0.35	31.2
Homework 3	83	0.15	12.4

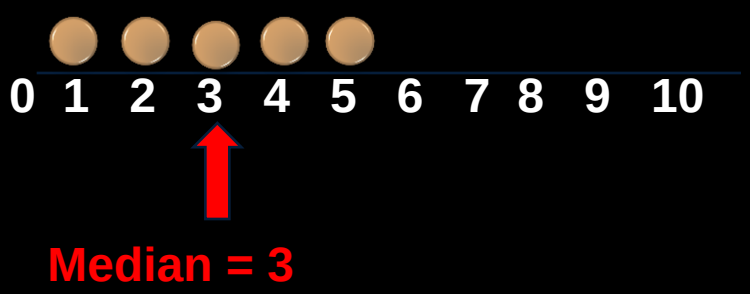
$$\sum_{i=1}^{3} w_i = 1.0 \qquad \sum_{i=1}^{3} (w_i x_i) = 90.6$$

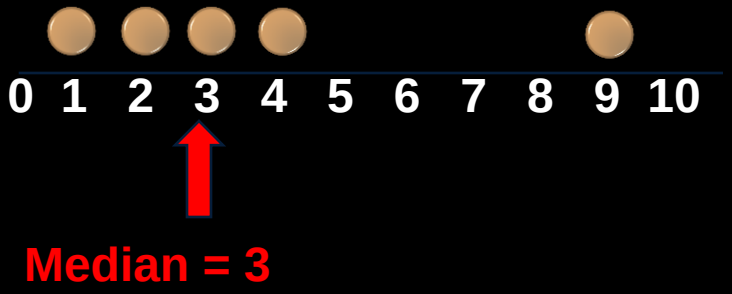
We can obtain the same result by plugging the numbers directly in to the formula for a weighted average:

$$\bar{x} = \frac{(0.50)(94)+(0.35)(89)+(0.15)(83)}{0.50+0.35+0.15} = \frac{47.0+31.2+121.4}{1.0} = 90.6$$

The Median

In an ordered array, the median is the "middle" number (50% above, 50% below)





Notice that the result is not affected by extreme values.

The Median

- This is another way to measure central tendency.
- The median is the value in the data set for which half the observations are higher and half are lower
- Using the previous example: 10 students were asked how many hours they spent online in a week
- The answers were: 3 7 4 9 5 4 6 17 4 7
- Rearrange these into ascending order:
- 3 4 4 4 5 6 7 7 9 17

The Median

- Because we have an even number of data, the median is the average of the two centre values (the values 5 and 6) which is 5.5 hours per week.

- If we had an odd number of data, the median would simply be the middle value of the set.

- It's the value in the data set for which half the observations are higher and half are lower

The Mode

- The mode is the number that occurs most frequently in the data.
- In our online example:
- 3 4 4 4 5 6 7 7 9 17
- The mode is 4 hours per week
-

The Mode

- Not affected by extreme values
- Used for either numerical or categorical (nominal) data
- There may be no mode
- There may be several modes



The Median & Mean

- Mean was 6.6
- Median was 5.5
- Mode was 4
- Quite a difference
- Which is best?

The Median & Mean

- Very small and very large numbers (outliers)can have significant impact on the mean
  - In our data set example the number 17
  - If you think 17 is not a typical value choose the median
- Mode has limited applications. Primarily used for data in the nominal scale.

Example

- Five houses on a hill by the beach

House Prices:

£2,000,000
500,000
300,000
100,000
100,000

Summary Statistics

House Prices:

£2,000,000
500,000
300,000
100,000
100,000

Sum £3,000,000

- **Mean:** (£3,000,000/5)
- = **£600,000**

- **Median:** middle value of ranked data
  = **£300,000**

- **Mode:** most frequent value
  = **£100,000**

Using Excel

- Excel will calculate the mean, median and mode with just a few mouse clicks:
- Enter a column of data and click
- Data Analysis>Descriptive Stats>Input Range

Summary

- Calculate the mean of a data set by summing the values and dividing this sum by the number of values
- The median of a data set is the midpoint of the set so long as the set is arranged in rank order.
- The median is the single centre value from an odd numbered data set; or the average of the two centre values from an even numbered set.
- The mode is the value that appears most often.