# Exploratory data analysis (EDA)

## Part 3: Measures of location

By: Noureddin Sadawi, PhD

University of London

# Measures of location

- In this section, we will consider the three main measures of location:
  - mean
  - median
  - mode.

- Then we will speak about:
  - percentiles
  - quartiles.

# Mean

- ‘The most basic estimate of location is the mean, or *average* value.
- The mean is the sum of all values divided by the number of values.
- Consider the following set of numbers:
  - {3 5 1 2}. The mean is (3 + 5 + 1 + 2) / 4 = 11 / 4 = 2.75.
- You will encounter the symbol $\overline{x}$ pronounced “x-bar”) being used to represent the mean of a sample from a population.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

# Mean

- ‘The formula to compute the mean for a set of n values $x_1$, $x_2$, ..., $x_n$ is:

$$\text{Mean} = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- N (or n) refers to the total number of records or observations.
- In statistics it is capitalised if it is referring to a population, and lowercase if it refers to a sample from a population.
- In data science, that distinction is not vital, so you may see it both ways.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

# Example

- Suppose we wish to find the mean of the following data set

  10, 11, 13, 14, 16, 18, 23

- Then, the mean is given by

$$\overline{x} = \frac{\sum x_i}{n} = \frac{10+11+13+14+16+18+23}{7} = \frac{105}{7} = 15$$

# Trimmed mean and weighted mean

- Trimmed mean and weighted mean are two variations of the mean.
- Please check the Bruce & Bruce textbook pp.9 and 10 for more details.

# Median

- ‘The median is the middle number on a sorted list of the data.
- If there is an even number of data values, the middle value is one that is not actually in the data set, but rather the average of the two values that divide the sorted data into upper and lower halves.
- Compared to the mean, which uses all observations, the median depends only on the values in the center of the sorted data.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

# Median

- ‘While this might seem to be a disadvantage, since the mean is much more sensitive to the data, there are many instances in which the median is a better metric for location.
- Let us say we want to look at typical household incomes in neighborhoods around Lake Washington in Seattle.
- In comparing the Medina neighborhood to the Windermere neighborhood, using the mean would produce very different results because Bill Gates lives in Medina.
- If we use the median, it won’t matter how rich Bill Gates is – the position of the middle observation will remain the same.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

# Example

- Referring to the data in the previous example,

10, 11, 13, 14, 16, 18, 23

- We see that there are 7 observations, so the median position is

$$\frac{7}{2} = 3.5 \approx 4$$

- So the median is 14.

# Example

- If our data was the following

10, 11, 13, 14, 16, 18

- then the median position would be

$$\frac{6}{2} = 3$$

# Mode

- ‘The mode is the value—or values in case of a tie—that appears most often in the data.
- The mode is a simple summary statistic for categorical data, and it is generally not used for numeric data.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

- For example: the mode of the following data

  6,5,7,5,6,2,3,4,5,76,9,7

  is 5 because more than any other number.

# Example

- Suppose we are given the following frequency distribution showing sales of car brands.

| Car Brand | Frequency |
|---|---|
| Audi | 4 |
| Ford | 4 |
| Jaguar | 3 |
| Tesla | 3 |
| Vauxhall | 6 |

- The mode of this data set is Vauxhall because it has the highest frequency.
- So the most frequently purchased car was **Vauxhall**.

# Percentiles

- The *pth* percentile is a value which divides the data into two parts such that at least *p* percent of the observations are less than or equal to this value and at least (*100-p*) percent of the observations are greater than or equal to this value.
- We can calculate the *pth* percentile using the following steps.

# How to calculate percentiles

- Sort the data in ascending order

- Compute the position of the pth percentile by doing the following calculation

$$\frac{p}{100} \times n$$

Where p denotes the percentile of interest and n is the number of observations

- If the value obtained in b) is not an integer, then round up to get the position of the pth percentile

# Example

- As an illustration, suppose we wish to find the 75th percentile for the following data

10,20,25,15,11,13,16,8,9,8,7,6

- Here, the sample size n=12
- We begin by sorting the data

# Example

- The sorted data

6,7,8,8,9,10,11,13,15,16,20,25

- then we compute the position of the 75th percentile by doing the following calculation and we obtain 9, which is an integer.

$$\frac{75}{100} \times 12 = 9$$

# Example

- So, the 75th percentile is the average of the 9th and 10th observations

$$\frac{15+16}{2} = 15.5$$

For the 60th percentile, we compute its position by doing the following calculation

$$\frac{60}{100} \times 12 = 7.2$$

# Example

- So, the 75th percentile is the average of the 9th and 10th observations

$$\frac{15+16}{2}=15.5$$

- So the 60th percentile is the 8th observation which is 13.

# Quartiles

- Quartiles divide the ranked data into four parts with each part containing approximately 25% of the data.

- The lower (or first) quartile $Q_1$ is also the 25th percentile.

- The second quartile $Q_2$ is also the 50th percentile. It is the median.

- The upper quartile $Q_3$ is also the 75th percentile.

# Example

- Let us refer to the data for the previous example

6,7,8,8,9,10,11,13,15,16,20,25

- We found Q_3, the 75th percentile to be

$$Q_3 = 15.5$$

# Example

- For $Q_2$ position we do the following calculation

$$\frac{50}{100} \times 12 = 6$$

- So, $Q_2$ is the value halfway between the 6th and 7th values. That is,

$$Q_2 = \frac{10 + 11}{2} = 10.5$$

# Example

- For $Q_1$ position we do the following calculation

$$\frac{25}{100} \times 12 = 3$$

- So, $Q_1$ is the value halfway between the 3rd and 4th values. That is,

$$Q_1 = \frac{8+8}{2} = 8$$

- For a frequency distribution, we use the cumulative frequency to locate a class containing the quartiles.