Exploratory data analysis (EDA)

Part 4: Measures of location (percentiles and quartiles) and measures of variation (range and interquartile range)

By: Noureddin Sadawi, PhD

University of London

Measures of location

- In this section, we will consider the following measures of location:
  - percentiles.
  - quartiles.

- Then we will speak about the following measures of variation:
  - range.
  - Interquartile range.

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

a) Sort the data in ascending order

b) 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 step b is not an integer then round up to obtain the position of the *pth* percentile.
- If it is integer, then the percentile is the average of the corresponding value and its next value in the data.

Example

- Suppose we wish to find the *75th* percentile of 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

- The position of the *75th* percentile is: (75/100)*12 = 9.
- So the *75th* percentile is the average of the 9th and 10th observations: (15 + 16) / 2 = 15.5.

- The position of the *60th* percentile is: 7.2.
- Hence, 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 *Q1* is also the *25th* percentile.
- The second quartile *Q2* is also the *50th* percentile.
  - It is the median.
- The upper quartile *Q3* is also the *75th* percentile.

Example

- Let us refer to the data from the previous example:

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

- We have found the *75th* percentile: *Q3* = 15.5.
- The position of *Q2* = (50/100)x12 = 6 and therefore *Q2* is the value halfway between the *6th* and *7th* values. That is: *Q2* = (10+11)/2 = 10.5.
- The position of *Q1* = (25/100)x12 = 3 and therefore *Q1* is the value halfway between the *3rd* and *4th* values. That is: *Q2* = (8+8)/2 = 8.
- For a frequency distribution, we use the cumulative frequency to locate a class containing the quartiles.

Range

- Range is the simplest measure of variability.
- It is defined as:

> ‘The difference between the largest and the smallest value in a data set.’
>
> Bruce and Bruce 2020

Range = Largest value - Smallest value

Example

Consider the following dataset:

10, 12, 16, 20, 22, 25, 30, 35, 37, 40

The range is 30 which is obtained by R = 40 - 10.

Interquartile range (IQR)

- Interquartile range is a measure of spread which can help us eliminate outliers, that is, extremely high or low observations.
- It calculates the range of the middle 50% of the data. That is:

$$IQR = Q3 - Q1$$

Example

For this dataset:

10, 12, 16, 20, 22, 25, 30, 35, 37, 40

Q1 = 16

Q3 = 35

(see previous slides on how to find them)

IQR = 35 - 16 = 19