CS5701 Quantitative Data Analysis (QDA) Suggested Pre-reads

Isabel Sassoon

2024

Here is a list of items to read and be familiar before the first lecture:

- It may be helpful to set up R and R Studio on your PC or laptop. You will have access to R and R studio in the Labs.
- There are some week 0 worksheets to help you with the install and an example analysis in R studio - we will show you more in lab 1
- The material in this document.

1 Measures of Location

We will be using measures of location to describe data in the lecture and labs this week so below are some definitions and an opportunity to practice.

Numerical Data - Descriptive Statistics - Measures of Location

There are

- **Mode** the most frequent observation in the data set, there can be more than one
- **Mean** (average) this is calculated as the sum of the observations in the variable divided by the number of the observations.

  $$\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$$

  assuming we have an attribute X which has the following observations $x_1, x_2, \ldots, x_n$
- **Median** when the set of observations is ordered from smallest to largest, the **median** is the value in the middle, or if there are an even number of observations then the mean of the two middle observations is used.

Note: the **mean** is affected by extreme values, the **median** is not.

Exercise - Numerical Data - Descriptive Statistics - Measures of Location

Assuming this is our data set:

$$4, 7, 2, 5, 6, 2$$

Compute the following:

- **Mode**
- **Mean**
- **Median**

Exercise - Numerical Data - Descriptive Statistics - Measures of Location

Computing the measures of location this data set:

$$4, 7, 2, 5, 6, 2$$

- **Mode** The most frequent observation is: 2
- **Mean**
$$\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i = \frac{1}{6} \times (4 + 7 + 2 + 5 + 6 + 2) = 4.333$$
- **Median** The values arranged from smallest to largest are: $2, 2, 4, 5, 6, 7$, there is an even number of values so the median is the mean of the two middle values $4, 5$. This results in 4.5

Numerical Data - Descriptive Statistics - Measures of Location

- There are other measures of location such as *k-th* **percentile**. This measure finds the value that divides the data such that $k\%$ of the values are below it. *75th percentile* and *90th percentile* are commonly used.
- The *50th quantile* is the median.
- If we have data with n values and want to find the $k\%$ **percentile**:
  1. sort the n data points in ascending order
  2. compute $k\% \times n$ and find the value in the sorted data that is in that position
- Looking at our small data set: $4, 7, 2, 5, 6, 2$, where $n = 6$
  - The *75th percentile* is 5.75
  - The *25th percentile* is 2.5

*see p67 Crawley for more

Weighted Mean - another measure of location

There is another measure related to the mean which is useful when there are weights involved. A good example of this is when computing the overall grade point when different module grades have different weights as they account for a different number of credits (if you attended the induction talks this should sound familiar) .

$$\bar{x}_w = \frac{\Sigma_{i=1}^{n} w_i \times x_i}{\Sigma_{i=1}^{n} w_i}$$

Lets assume that you have 3 module grade points $13, 16, 12$ and that the respective weights (credits) are $15, 15, 30$ - what is the weighted grade point mean?

$$\bar{x}_w = \frac{13 \times 15 + 16 \times 15 + 12 \times 30}{15 + 15 + 30} = 13.25$$

2 Measures of Variation

Numerical Data - Descriptive Statistics - Measures of Variation

There are different ways of quantifying the spread of a variable of data:

- The **range**: defined as *maximum value in the variable - minimum value*
- The **Inter Quartile Range - IQR** defined as the *75th Percentile - 25th Percentile*, this will measure the spread of the central 25% of the data
- The **Standard Deviation** - measures the spread of the observations from the mean

Numerical Data - Descriptive Statistics - Standard Deviation

The definition of **Standard Deviation** is:

$$\sigma = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

- The **standard deviation** is the average distance between each observation value and the mean
- The square of the **standard deviation** is known as the **Variance** $Var = \sigma^2$
- When computing the **standard deviation** for a *sample* then the following is used:

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

  (this takes into account that one parameter (the mean) is estimated from the same data before computing this)

In Lecture and lab 1 we will use these concepts and see how they fit in Exploratory Data Analysis.