PROJECT - Grey Matter

A collection of Notes

Divyansh Chahar

https://www.linkedin.com/in/divyanshchahar/

https://github.com/divyanshchahar

Tuesday 15th September, 2020

1 | Introduction to Statistics and Measures used in Descriptive Statistics

1.1 Measure of Central Tendency

1.1.1 Mean

Mean

Mean is simply the average of the data.

It can be mathematically expressed as follows

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

1.1.2 Median

Median

Median is the central value of the data when the values are arranged in increasing or decreasing order

It can be mathematically expressed as

$$\widetilde{x} = \begin{cases} x_{(n+1)/2}, \text{if n is odd} \\ \frac{1}{2}\left(x_{(n)/2} + x_{(n+1)/2}\right), \text{if n is even} \end{cases}$$

1.1.3 Mode

Mode

Mode is the most frequently occurring value in the dataset

1.2 Measure of Dispersion

If data points are represented on number line than measure of dispersion will be an indication of how far are data points from each other.

1.2.1 Range

Range

It is the interval over which the data is spread. It is the difference between the highest and lowest value of the data

It is mathematically expressed as

$$R = Max(x_i) - Min(x_i)$$

1.2.2 Quartiles and Inter-Quartile Range

Quartiles

Quartiles divide the ordered dataset into 4 parts with equal members

- The **second quartile** or $\boldsymbol{Q_2}$ is the value at the middle position when the data is arranged in ascending order, it is also the **median value**.
- The first quartile or $\boldsymbol{Q_1}$ divides the first half of the data into 2 parts, we can say that the **first quartile is the median of the first half of the data**.
- Similarly the **third quartile** or $\boldsymbol{Q_3}$ is the **median of the second half of the data**.

Interquartile Range

Interquartile range is the differance between the median of the first and second half of the data, or we can say that inter quartile range is the differance between Q_1 and Q_3.

$$IQR = Q_3 - Q_1$$

1.2.3 Variance

Variance

Variance is a measure of how far the data points are from the mean.

Variance is one of the methods of measuring how spread out the values are in a dataset. Variance can be mathematically calculated as

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

Due to squaring of difference, the figure obtained could be pretty large and insignificant as the unit of measurement is squared.

1.2.4 Standard Deviation

Standard Deviation

Standard Deviation is the square root of Variance

Standard Deviation also measure the spread of data. Since the units are not squared in Standard Deviation and the values are not too large either, it is much more used than variance. Standard Deviation can be calculated as :

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

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

1.2.5 Coefficient of Variation or Relative Standard Deviation

Standard Deviation is the most common measure of the variability of a single data-set. However when we want to compare the standard deviation of two or more data-sets the issue becomes a little more complicated, we need a relative measure such as Coefficient of Variation. To better understand the concept of Variance, Standard Deviation and Coefficient of Variance, let us consider the following example.

Example

Lets us consider stock price of a random company over a week. Let us consider two investors, one from U.S.A. and another one from India. although the price of Stock is same but the Indian Investor has to pay in Indian Rupee at an exchange rate of 73.51 INR per USD.

	USD	INR
Day 1	514.79	37842.213
Day 2	498.79	36666.053
Day 3	504.72	37101.968
Day 4	508.57	37384.981
Day 5	504.05	37052.7155
Mean	506.184	37209.586
Variance	28.225	152519.7
Standard Deviation	5.313	390.538

- Mean, variance and standard deviation for USD are **506.184 USD**, **28.225 USD^2** and **5.313 USD** respectively.
- Mean, variance and standard deviation for INR are **37209.586 INR**, **152519.7 INR^2** and **390.538 INR** respectively.

From the above data we can observe:

- Variance values are comparatively larger than Standard Deviation

- Even though we have same variability for INR and USD values, Standard Deviation varies a lot.

To compare the variability of two different datasets we use Cofficient of Variance which can be calculated as

$$C_v = \frac{\sigma}{\mu}$$

$$C_v = \frac{S}{\bar{x}}$$

Example

Let us calculate the cofficient of variance for the above data using the population formulae

- Cofficient of variance for USD is $C_v = \frac{5.313}{506.184} = 0.010$
- Cofficient of variance for INR is $C_v = \frac{390.538}{32709.586} = 0.010$

1.3 Measure of Position

Measure of Position is the measure of the location of a data point in the population or sample.

1.3.1 Percentile

Percentile

N^{th} percentile will represent a datapoint which is larger than N percent of values in the dataset i.e. 99^th percentile represents a value which is greater than 99 percent of the values. Percentile divide the ordered data into 100 equal parts in terms of number of members.

Rank of Percentile can be mathematically measured as

$$R_p = \frac{P}{100}N$$

The above formulae works fine if R_p is a positive integer. But the situation can become bit complicated if R_p also has decimal value.

Example

For example, let us consider a dataset with following values: $P = 35$, $N = 90$. Therefore R_p can be calculate as follows:

$$R_p = \frac{35}{90}200 = 31.5$$

As we can observe that 31.5 rank could be difficult to calculate. However this can be achieved as follows, Let us consider the $\mathbf{30^{th}}$ value is **60** and $\mathbf{31^{st}}$ value is **65**. Thus we will proceed as follows.

- **STEP 1:** Break the rank i.e. R_p into integer and decimal part. In this case the integer part is 31 and decimal part is 0.5.
- **STEP 2:** Find the Value at integer rank and integer rank + 1. In this case the values are 60 and 65 respectively
- **STEP 3:** Calculate the difference between the value at integer rank and integer rank +1

$$65 - 60 = 5$$

- **STEP 4:** Multiply the difference with the decimal part.

$$0.5 \times 5 = 2.5$$

- **STEP 5:** Add the two values to get the final value

$$60 + 2.5 = 62.5$$

1.3.2 Quartile

As mentioned in the above section quartiles divide the dataset into 4 equal parts in terms of number of members. Rank of the three quartiles can be calculated as:

$$Q_1 = \frac{25}{100}N$$

$$Q_2 = \frac{50}{100}N$$

$$Q_3 = \frac{75}{100}N$$

- Second Quartile is the median of the dataset.
- If the Rank of a quartile is a not a poitive integer, than it can be resolved as show in the example above.

1.3.3 Standard Score

Standard Score

Standard Score measure the differance between the datatpoints and the mean in terms of standard deviation

It can be mathematically expressed as

$$Z = \frac{x - \bar{x}}{\sigma}$$