SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

III Year / VI Semester

Unit 2 – DESCRIPTIVE ANALYTICS USING STATISTICS

Topic 2: Mean, Mode, Standard Deviation, Variance

Mean, Median and Mode

Mean

The mean, as you probably know, is just another name for the average.

To calculate the mean of a dataset, all you have to do is sum up all the values and divide it by the number of values that you have.

Sum of samples/Number of samples.

Median

The way you compute the median of the dataset is by sorting all the values (in either ascending or descending order), and taking the one that ends up in the middle.

So, for example, let's use the same dataset of children in my neighborhood
0, 2, 3, 2, 1, 0, 0, 2, 0
I would sort it numerically, and I can take the number that's slap dab in the middle of the data, which turns out to be 1.
0, 0, 0, 0, 1, 2, 2, 2, 3

Mean, Median and Mode

Mode
All mode means, is the most common value in a dataset.

Let's go back to my example of the number of kids in each house.
0, 2, 3, 2, 1, 0, 0, 2, 0

How many of each value are there:
0: 4, 1: 1, 2: 3, 3: 1

The MODE is 0

Standard deviation, Variance

Standard deviation and variance are two fundamental quantities for a data distribution.

Variance

Variance measures how spread-out the data is. A variance is the average of the squared differences from the mean.

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

X denotes each data point
μ denotes the mean
N denotes the number of data points

Standard deviation, Variance

Standard deviation and variance are two fundamental quantities for a data distribution.

Variance

Variance measures how spread-out the data is. A variance is the average of the squared differences from the mean.

To compute the variance of a dataset, first figure out the mean. Lets say our data set has five values (1,4,5,4,8)

1. The first step in computing the variance is just to find the mean, or the average, of that data.
   Mean of the above dataset is (1+4+5+4+8)/5 = 4.4

2. Now the next step is to find the differences from the mean for each data point.
   1-4.4 = -3.4,
   4-4.4 = -0.4,
   5-4.4 = 0.6,
   4-4.4 = -0.4,
   8-4.4 = 3.6

-3.4, -0.4, 0.6, -0.4, 3.6

Standard deviation, Variance

Variance

3. Next is to do is find the square of these differences.

$(-3.4)^2 = 11.56$

$(-0.4)^2 = 0.16$

$(0.6)^2 = 0.36$

$(-0.4)^2 = 0.16$

$(3.6)^2 = 12.36$

4. To find the actual variance value, we just take the average of all those squared differences.

$\sigma^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.86) / 5 = 5.04$

Standard deviation, Variance

Standard Deviation

Standard deviation is just the square root of the variance.

Variance is $\sigma^2 = 5.04$

Standard Deviation is $\sqrt{5.04} = 2.24$

Population variance versus sample variance

If variance is calculated for complete data set then this is called population variance.

For example $\sigma^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.86) / 5 = 5.04$

Sample Variance

If we calculate for a subset of the data then that is called sample variance.

Instead of dividing by the number of samples, you divide by the number of samples minus 1.

Sample variance, which is designated by S^2, it is found by the sum of the squared variances divided by 4, that is (n - 1).

$S^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.86) / 4 = 6.3$

THANK YOU