Data Analytics

SESSION 2: GETTING STARTEDWITH STATISTICS

Assignment 1 Data Analytics

Table of Contents

1.Introduction ...............................................................................................................................................

2.Objective .................................................................................................................................................... 3.Prerequisites ..............................................................................................................................................

4.Associated Data Files .................................................................................................................................

6.Expected Output ........................................................................................................................................

7.Approximate Time to Complete Task ........................................................................................................

**1. Introduction**

This assignment will help you understand the concepts learnt in the session.

**2. Objective**

This assignment will test your skills on the concepts of statistics.

**3. Prerequisites**

Not applicable.

**4. Associated Data Files**

Not applicable.

**5. Problem Statement**

1. If I find Covariance between same variable what will be the output? What will be correlation coefficient?

Answer: First of all let us understand what is variable. It is a characteristic, number, or quantity that increases or decreases over time, or takes different values in different situations. Examples could be change in oil prices, change in share prices etc.

Covariance: it is basically measure of how much two random variable changed or related together. If the greater values of one variable mainly correspond with greater value of other variable, and the same hold for smaller values, the covariance is positive.

In opposite case when greater values of one variable mainly correspond with smaller value of other, the covariance is negative.

Covariance formula.

 cov(X,Y)=sum_(i=1)^N((x_i-x^_)(y_i-y^_))/N. 

Now let us understand this with an example. Let us consider economic growth of India (variable x) and its relation with BSE (variable y).

|  |  |
| --- | --- |
| Economic growth | BSE |
| 2.1 | 8 |
| 2.5 | 12 |
| 4.0 | 14 |
| 3.6 | 10 |

Using covariance formula we can calculate whether economic growth and BSE have positive or negative relationship.

But first of all we need to calculate mean for both of the variable.

Mean for economic growth is 2.1+2.5+4.0+3.6/4= 12.2/4=3.1

Mean for BSE is 8+12+14+10/4 = 44/4 =11

Now using covariance formula we can find out if both of these variable are positively related or negatively related.

Cov xy= (2.1-3.1)(8-11)+(2.5-3.1)(12-11)+ (4.0-3.1)(14-11)+(3.6-3.1)(10-11)/4-1

=3+ (-0.6)+2.7+(-0.5)/3

=4.6/3 = 1.53

1.53 is positive so relationship between both of this variable is positive.

It is also answer of our question that if covariance found between two same variable then output is positive.

Now let us understand what correlation coefficient is.

Correlation is another way of determining how two variables are related. In addition to telling you whether variables are positively or adversely related, correlation also tells you degree to which the variable tends to move together. You can say that correlation coefficient is used to interpret magnitude to covariance.

Correlation formula Rxy = cov(x,y)/SxSy where x and y are variable and S is the standard deviation.

Now for calculating correlation coefficient we must have the standard deviation with us.

Sx =0.90 and Sy =2.58

Now to calculate correlation Rxy= 1.53 / (0.90)(2.58) =0.66 which is positive. So we can say in most of the cases if covariance is positive then correlation coefficient will also be positive.

2. Assume I have a set of numbers. The mean, median and mode of the set of numbers are equal. If I draw a Frequency plot of individual distinct numbers, how would the plot look like?

A B C



Answer:

Suppose we have a set of number (weight ) for a population is 0,10,20,30,40,50,50,60,70,80,90,100

Mean of these numbers are 0+10+20+30+40+50+50+60+70+80+90+100/2= 50

Median of this number is also 50. Also, as 50 occurs most number of times so it’s going to be the mode.

Now let us understand how a frequency plot will look like if we draw a frequency plot for this data set.



This frequency plot indicates that the population average weight is around 50 kg and most of the people belong to this weight group only. Also, if we study the data we will find that 50 is the middle age group between 0-100.