**P.R. Government College (A), Kakinada**

**Department of Statistics**

**Semester-III Paper-III**

**Unit-II**

**Correlation**

**Introduction:**

 I n the previous semesters almost we familiar with single phenomenon and its analysis. But, in practical so many issues involve multiple phenomena.

For example,

* If we want to study the effect of CVID-19 on Indian people, we must consider simultaneously their age, gender, area, health issues, travelling history, etc.
* If we want to study the economic status of industries, the study should cover production capacity, investment, income, expenditure, etc.
* The study of weights of a group of children requires essentially a simultaneous study of their age, height, gender etc.

 Clearly there will be inter-relation between various phenomena. We need some measures to the study the relations. One of such measure is **correlation**.

**Definition:**

 The Correlation is a statistical tool which measures the linear relationship between two or more variables.

* the correlation is **Simple Correlation**, If there are only TWO variables in study,
* the correlation is **Multiple/Partial Correlation**, If there are more than TWO variables in study,

 In this unit, we start with simple correlation and at final we will study multiple and partial correlation with three variables.

**Simple Correlation**:

 The correlation between two variables is called a simple correlation or simply correlation.

Let X and Y be two variables. The relation between X and Y may be one of the following three types.

* Positive correlation
* Negative correlation
* Zero correlation

**Positive Correlation**: If the two variables X and Y deviate in the same direction, then it is a positive correlation.

That is, positive correlation is there if X increases, Y increases

 and X decreases, Y decreases

Examples:

* The relation between Height(X) and Weight(Y) of a group of children – a Positive correlation.
* The relation between industries establishment rate(X) and job rate(Y) in a group of cities- a positive correlation.
* The relation between sales(X) and profit(Y) of company over a period of time- a ‘+’ correlation.

**Negative Correlation**: If the two variables X and Y deviate in the oposite direction, then it is a negative correlation.

That is, negative correlation is there if X increases, Y decreases

 and X decreases, Y increases

Examples:

* The relation between Price(X) and Demand(Y) of a group of items – a Negative correlation.
* The relation between Security level(X) and crime rate(Y) in a group of villages- a negative correlation.
* The relation between Pressure(X) and volume(Y) of a gas at different intervals of time - a ‘-’ correlation.

**Zero correlation**: If the two variables X and Y are independent then it is zero correlation. That is the movement of one variable is independent of the other’s movement.

Here the variables are said to be uncorrelated.

Examples:

* The relation between amount of water drunk(X) and intelligence(Y) of a group of students – a Zero correlation.
* The relation between level of beauty(X) and job level(Y) of a group of women graduates- a Zero correlation.

**Methods of measuring correlation**:

 The following are different methods to measure correlation between X and Y.

1. Scatter diagram method
2. Graphical method
3. Karl Pearson’s Coefficient of correlation
4. Spearman’s coefficient of correlation
5. Method of concurrent deviations

**Scatter diagram method**:

**Scatter diagram**: The diagram the obtained by plotting the given sat of paired values of X and Y on a 2D plane is called a scatter diagram or scatter plot.

 Y

 X

 Scatter Diagram

In scatter diagram, if the points are scattered in the direction of positive slope, then the correlation is positive.

If the points are scattered in the direction of negative slope, then the correlation is negative.

If the points are rounded off a cluster, then the variables are uncorrelated (zero correlation).

Y Y Y

 X X X

 Positive correlation Negative correlation Zero correlation

Example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Child Name | A | B | C | D | E |
| Height (in cm) | 80 | 90 | 100 | 110 | 120 |
| Weight(in kg) | 20 | 23 | 25 | 27 | 28 |

 Weight

 Positive correlation

 Height

**Graphical Method**:

In graphical method, two separate graphs for X and Y are constructed by taking X and Y values along x-axis and reference numbers the pairs along y-axis.

If the curves of these two graphs are moving in the same way, correlation b/w X and Y is positive.

If the curves of these two graphs are moving in the opposite way, correlation b/w X and Y is negative.

Example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Child Name | A | B | C | D | E |
| Height (in cm) | 80 | 90 | 100 | 110 | 120 |
| Weight(in kg) | 20 | 23 | 25 | 27 | 28 |

 Height Weight

 Child Name Child Name

 Positive correlation

\*\* Scatter diagram and Graphical methods provide a rough measure of correlation b/w X and Y (positive or negative or zero). These are not appropriate to study the degree or intensity of correlation.

**Karl Pearson’s Coefficient of Correlation:**

 In order measure the degree of relationship between X and Y, Karl Pearson introduced a coefficient with covariance and standard deviations.

**Definition**: The Karl Pearson’s coefficient of correlation between X and Y is denoted by r(X,Y) or rXY or simply ‘r’ and it is defined as,

 r(X,Y) = $\frac{Cov(X,Y)}{σ\_{X}σ\_{Y}}$

 Where, Cov(X,Y) = Covariance b/w X and Y ( or Product moment µ11 )

 σX = Standard deviation of X = $\sqrt{Var(X)}$

 σY = Standard deviation of Y = $\sqrt{Var(Y)}$

 The Karl Pearson’s coefficient of correlation is also known as a product moment correlation coefficient.

For a simple bi-variate data of n pairs,

 X : x1, x2, x3, ........., xn

 Y : y1, y2, y3, .........., yn

 Cov(X,Y) = $\frac{1}{n}$∑(X-X̅)(Y-Y̅) or $\frac{1}{n}$∑XY – ($\frac{∑X}{n})(\frac{∑Y}{n})$

 σX = $\sqrt{\frac{1}{n}∑(X-X̅)^{2}}$ or $\sqrt{\frac{1}{n}∑X^{2}-(\frac{∑X}{n})^{2}}$

 σY = $\sqrt{\frac{1}{n}∑(Y-Y̅)^{2}}$ or $\sqrt{\frac{1}{n}∑Y^{2}-(\frac{∑Y}{n})^{2}}$

So, Simply

 r = $\frac{ ∑(X-X̅)(Y-Y̅)}{ \sqrt{∑(X-X̅)^{2}} X\sqrt{∑(Y-Y̅)^{2}} }$

 or r = $\frac{n∑XY –(∑X)(∑Y)}{\sqrt{n∑X^{2}-(∑X)^{2}} X\sqrt{n∑Y^{2}-(∑Y)^{2}}}$

**Note**: We can write,

r(X,Y) = $\frac{Cov(X,Y)}{σ\_{X}σ\_{Y}}$ = $\frac{∑(X-X̅)(Y-Y̅)}{σ\_{X}σ\_{Y}}$ = ∑($\frac{(X-X̅)}{σ\_{X}})(\frac{(Y-Y̅)}{σ\_{Y}}$) = ∑ZxZY  or Cov(Zx,ZY)

i.e., r(X,Y) is nothing but the covariance between their **standardized** variables.

**Properties of Karl Pearson’s Coefficient of Correlation:**

1. The range of ‘r’ is [-1, 1] . i.e., always -1 ≤ r ≤ 1
* r = +1 indicates, a perfect positive correlation
* r = -1 indicates, a perfect negative correlation
* r = 0 indicates, no correlation.
* the remaining values indicate the relation with a certain degree(high, moderate, low, etc).
1. The Karl Pearson’s correlation coefficient is independent of change of origin and scale.

i.e., r(X,Y) remains unchanged if there any changes in the values of X series or Y series or both arithmetically.

1. If the two variables of independent, then they are always uncorrelated.

But the converse may not be true, i.e., If the variables are uncorrelated, they may not be independent.

1. r(X,Y) = r(Y,X) -> r is **symmetric** ( since Cov(X,Y)=Cov(Y,X) -> symmetric)

**Proofs:**

1. By def.

 r = $\frac{ ∑(X-X̅)(Y-Y̅)}{ \sqrt{∑(X-X̅)^{2}} X\sqrt{∑(Y-Y̅)^{2}} }$

 Let a = $(X-X̅)$ and b= $(Y-Y̅)$

=> r = $\frac{ ∑ab}{ \sqrt{∑a^{2}} X\sqrt{∑b^{2}} }$ -------(1)

By the Cauchy-Schwartz inequality,

 ($∑ab)$2  ≤ $∑a^{2}$ X $∑b^{2}$

=> $\frac{(∑ab)2}{∑a^{2} X ∑b^{2}}$ ≤ 1

=> r2 ≤ 1

=> **-1 ≤ r ≤ 1**

1. Let X and Y be changed in new variables U and V respectively by shifting the origins to a and b and changing scales with h and k ( h, k > 0)as

 U = (X-a)/h and V = (Y-b)/k

Let us prove r(U,V) = r(X,Y)

By def, r(U,V) = $\frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}}$

 = $\frac{Cov(\frac{X-a}{h}, \frac{Y-b}{k})}{\sqrt{Var(\frac{X-a}{h})}\sqrt{Var(\frac{Y-b}{k})}}$

 = $\frac{\left(\frac{1}{hk}\right)Cov(X,Y)}{\sqrt{\left(\frac{1}{h}\right)^{2}Var(X)}\sqrt{\left(\frac{1}{k}\right)^{2}Var(Y)}}$ (applying properties of covariance and variance)

 = $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$

 = r(X,Y)

Hence , the correlation coefficient is independent of change of origin and scale.

1. Let X and Y be independent, then E(XY) =E(X)E(Y)

 => Cov(X, Y) = E(XY) - E(X)E(Y) = 0

 => r(X,Y) = 0 => X and Y are uncorrelated.

But the converse may not be true. Let us prove it with an example.

Consider the data,

 **Total**

 X : -2 -1 1 2 **0**

 Y : 4 1 1 4 **10**

 XY: -8 -1 1 8 **0**

=> Cov(X,Y) = $\frac{1}{n}$∑XY – ($\frac{∑X}{n})(\frac{∑Y}{n})$ = (1/4)(0) – (0/4)(10/4) = 0

=> r(X,Y) = 0 => X and Y uncorrelated.

 But, clearly X and Y related with functional relation Y=X2. Thus uncorrelated may not be independent.

**Problem:**

 Calculate the Karl Pearson’s Coefficient of correlation for the following data relating to the Capital employed(X)(in crores) and profit(Y) (in crores) of 6 farms in an industrial area.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | 10 | 20 | 30 | 40 | 50 | 60 |
| Y | 2 | 4 | 8 | 9 | 14 | 12 |

 Solution:

 r = $\frac{n∑XY –(∑X)(∑Y)}{\sqrt{n∑X^{2}-(∑X)^{2}} X\sqrt{n∑Y^{2}-(∑Y)^{2}}}$

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | XY | X2 | Y2 |
| 10 | 2 | 20 | 100 | 4 |
| 20 | 4 | 80 | 400 | 16 |
| 30 | 8 | 240 | 900 | 64 |
| 40 | 9 | 360 | 1600 | 81 |
| 50 | 14 | 700 | 2500 | 196 |
| 60 | 12 | 720 | 3600 | 144 |
| **∑X=210** | **∑Y=49** | **∑XY=2120** | **∑X2=9100** | **∑Y2=505** |

 Here, n= 6

 r = $\frac{6\left(2120\right)–\left(210\right)(49)}{\sqrt{6(9100)-(210)^{2}} X\sqrt{6(505)-(49)^{2}}}$

 = $\frac{12720–10290}{\sqrt{54600-44100} X\sqrt{3030-2401}}$

 = $\frac{2430}{\sqrt{10500} X\sqrt{629}}$

 **r = 0.9455**

 i.e., there is a **high positive correlation** between capital employed and profit. It means more capital gives more profit.

**Coefficient of Determination:**

The **Square of Correlation Coefficient (r2)** is called the Coefficient of Determination.

**Interpretation of r2**: r2 indicates the **percentage of data** of one variable that **explained** by the other variable.

In the above example, r = 0.9455 => r2 = 0.8940 => 89% of data are explained by each other.

If r2 is high, we can fit a good model for prediction of one from other(s)

 ( we will discuss it in the lesson “Regression”)

**Probable error of Correlation coefficient**:

Since the data are a sample data, there is possibility of some error in ‘r’.

The Probable Error of ‘r’ can be calculated by the formula,

 **P.E(r) = 0.67 x (1-r2)/√n**

In the above example, n = 6, r = 0.9455,

 P.E(r) = 0.67 x (1-0.94552)/√6 = 0.029.

i.e., most probably ‘r’ lies between 0.9455-0.029 and 0.9455+0.029

 i.e, 0.9426 and 0.9484

**Spearman’s Rank Correlation:**

 If the data are in ordinal scale, Spearman’s Rank correlation coefficient is very simple tool to study the relation b/w two variables. Let us discuss it.

**Definition**:

 The Spearman’s Coefficient of Correlation between the two variables X and Y is the Karl Pearson’s coefficient of correlation between the ranks of X and Y.

It is denoted by rs(X,Y) or ρ(X,Y)

i.e., If Rx = Ranks of X and Ry = Ranks of Y, then rs (X,Y) = r(Rx,Ry)

\*\*Ranks may assign either from large value or from small value, but opt the same in both series X & Y

**Derivation of Rank Correlation coefficient:**

**Case (i) : If values are not repeated : (untied ranks)**

 By def.

 rs(X,Y) = r(Rx,Ry)

 = Cov(Rx,Ry) / σRx σRy ....... (1)

 If the ‘n’ values are not repeated in X series and in Y series, obviously the ranks are Rx : 1, 2, 3, ..., n and Ry: 1, 2, 3, ..., n in some order.

 => ∑Rx = ∑Ry = ∑i = 1+2+3+...+n = n(n+1)/2

 => R̅x = R̅y = [n(n+1)/2]/n = (n+1)/2

 and ∑Rx2 = ∑Ry2 =∑i2 = 12+22+32+...+n2 = n(n+1)(2n+1)/6

 => Var(Rx) = Var(Ry) = 1/n ∑R2 - R̅2 = [n(n+1)(2n+1)/6]/n – [(n+1)/2]2

 = (n+1)[(2n+1)/6 – (n+1)/4]

 = [(n+1)(n-1) ]/12

 Var(Rx) = Var(Ry) = (n2 -1)/12 ........(2)

Let the difference, D = Rx – Ry for each pair

Now we can write, D = (Rx - R̅x) – (Ry - R̅y) [ since R̅x = R̅y ]

Now taking square and then ∑ over i= 1, 2 ..., n on both sides

 => ∑D2 = ∑[(Rx - R̅x) – (Ry - R̅y)]2

 = ∑[(Rx - R̅x)2 + ∑(Ry - R̅y)2 - 2∑(Rx - R̅x) (Ry - R̅y)

 = n Var(Rx) + n Var(Ry) – 2n Cov(Rx,Ry)

 = n Var(Rx) + n Var(Ry) – 2n rs(X,Y) σRx σRy ( from 1)

 = n (n2-1)/12 + n (n2-1)/12 –2 n rs(X,Y) (n2-1)/12 (from 2)

 ∑D2 = 2[n (n2-1)/12] [ 1 - rs(X,Y)]

 => 1 - rs(X,Y) = 6∑D2 /n(n2-1)

 => **rs(X,Y) = 1 – [6∑D2 /n(n2-1)]**

 **Or**

 **rs(X,Y) = 1 – [6∑D2 /(n3-n)]**

 This is the Spearman’s Rank correlation formula for untied ranks.

 Here, D=Rx – Ry and n = no. of pairs

**Case (ii) : If Values are repeated: ( Tied Ranks)**

 If any set of values are repeated either in X series or in Y series, in order to use the difference D as in case (i) and satisfying the condition

 R̅x = R̅y=n(n+1)/2 an **average rank is to be assigned** as a common to the repeated values. Here average rank is nothing but the average of the ranks of the values if actually they are not tied.

For Eg: If X values : 90, 85, 66, 66, 50, 43, 32, 32, 32, 28,........

 Ranks are, 90 – 1st rank, 85 – 2nd rank ,

 66 ( repeated 2 times) – Avg( actual ranks if they not tied)

 = Avg(3rd,4th ranks) = (3+4)/2=7/2

 = 3.5th rank each,

 50 – 5th rank, 43 – 6th rank,

 32(repeated 3 times)– (7th +8th +9th )/3 = 24/3 = 8th rank each,

 28 – 9th rank , and so on.

 Spearman’s Rank Correlation coefficient in case of tied ranks is given by

 **rs(X,Y) = 1 – {[6∑D2 + (m3-m)/12 + (m3-m)/12+.....]/(n3-n)}**

 where, n = no. of pairs

 D = Rx – Ry

 term (m3-m)/12 is repeated one for each tied set( in X and Y)

 m = the no. of tied ranks in the set.

**\*\* The Rank correlation lies between -1 and +1.**

 **Proof:** Since rank correlation is Pearson’s correlation b/w the ranks, so the limits are -1 and +1.

(Or)

-> by formula, rs is maximum, when **∑D2** is minimum. The minimum value of

 **∑D2** is zero when all D’s are zero (i.e., Rx=Ry, for each pair)

 => **Maximum value of rs is +1**.

-> Again by formula, rs is miniimum, when **∑D2** is maximum. The maximum value of **∑D2** is attained when all D’s are maximum (i.e., Rx~Ry, is max)

It will happen when the ranks of X are reverse in order to the ranks of Y.

 i.e., Rx : 1, 2, 3, ........, n-1, n

 Ry : n, n-1, n-2, ....., 2, 1

 Clearly Rx +Ry = n+1 and we know Rx-Ry = D

Solving these two equations, we can get, **∑D2** = (n3-n)/3

 => rs = -1

=> **Minimum value of rs is -1**.

Therefore, Rank correlation lies b/w -1 and +1.

**Problem 1:**

Calculate the rank correlation between the percentage of increased covid-19 active cases(X) and percentage of increased deaths(Y) from previous day in India for a period of 10 days(from 25-4-2020 to 3-4-2020) as given below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 24 | 10 | 20 | 11 | 3 | 23 | 12 | 45 | 27 | 22 |
| Y | 20 | 67 | 0 | 21 | 13 | 19  | 9 | 66 | 24 | 18 |

Solution:

Since values are repeated in either series,

 rs(X,Y) = 1 – [6∑D2 /(n3-n)]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | Y | Rx | Ry | D=Rx-Ry | D2 |
| 24 | 20 | 3 | 5 | -2 | 4 |
| 10 | 67 | 9 | 1 | 8 | 64 |
| 20 | 0 | 6 | 10 | -4 | 16 |
| 11 | 21 | 8 | 4 | 4 | 16 |
| 3 | 13 | 10 | 8 | 2 | 4 |
| 23 | 19 | 4 | 6 | -2 | 4 |
| 12 | 9 | 7 | 9 | -2 | 4 |
| 45 | 66 | 1 | 2 | -1 | 1 |
| 27 | 24 | 2 | 3 | -1 | 1 |
| 22 | 18 | 5 | 7 | -2 | 4 |
|  |  |  | **Total** | **0** | **118** |

 **n = 10**

 => rs(X,Y) = 1 – [6∑D2 /(n3-n)]

 = 1 – [6x118 /(103-10)]

 = 1- 0.7151 = **0.2849**

**Problem 2:**

Calculate the rank correlation between the percentage of increased covid-19 active cases(X) and percentage of increased deaths(Y) from previous day in India for a period of 8 days(from 4-4-2020 to 11-4-2020) as given below.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 17 | 18 | 11 | 11 | 11 | 12 | 12 | 9 |
| Y | 15 | 19 | 15 | 18 | 11 | 28 | 10 | 16 |

 Solution:

 Since values are repeated ( 11 – 3 times, 12 –2 times in X and 15 -2 times in Y), the formula becomes,

 rs(X,Y) = 1 – {[6∑D2 + (m13-m1)/12 + (m23-m2)/12+(m33-m3)/12]/(n3-n)}

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | Y | Rx | Ry | D = Rx-Ry | D2 |
| 17 | 15 | 2 | [(5+6)/2] 5.5 | -3.5 | 12.25 |
| 18 | 19 | 1 | 2 | -1 | 1 |
| 11 | 15 | [(5+6+7)/3] 6 | 5.5 | 0.5 | 0.25 |
| 11 | 18 | 6 | 3 | 3 | 9 |
| 11 | 11 | 6 | 7 | -1 | 1 |
| 12 | 28 | [(3+4)/2] 3.5 | 1 | 2.5 | 6.25 |
| 12 | 10 | 3.5 | 8 | -4.5 | 20.25 |
| 9 | 16 | 8 | 4 | 4 | 16 |
|  |  |  | **Total** | **0** | **66** |

n= 8, m1 = 2 (case of 12) , m2 = 3 (case of 11) and m3 = 2 (case of 15)

=> rs(X,Y) = 1 – {[6x66 + (23-2)/12 + (33-3)/12+(23-2)/12]/(83-8)}

 = 1 – {[396 + 0.5 + 2+0.5]/504}

 = 1 – 399/504 = **0.2083**

**Method of Concurrent Deviations**:

 It is another method to measure correlation in which the simultaneous deviations in X and Y are considered. The deviation of X and Y in the same direction is called concurrent deviation.

The method consists in finding

* current deviations ( +,-,=) in X and Y separately from the previous (except for first values)
* product of current deviations of X and Y

(Note: i) +x+ and –x – are ‘**+’** , ii) +x- , +x= and -x= are ‘**-**‘ )

* C = no. of concurrent product deviations (+)

The formula of method of concurrent deviations is,

 **rc** = $\pm $$\sqrt{\pm \frac{(2C-m)}{m}}$

 where, m = n-1

 C = no. of concurrent deviations

\*\* If 2C-m > 0, then the formula takes ‘+’ sign inside and outside of the root.

\*\* If 2C-m < 0, then the formula takes ‘-’ sign inside and outside of the root.

**Problem:**

Calculate the correlation for the data given below using concurrent deviation method .

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 23 | 26 | 28 | 20 | 25 | 25 | 28 | 30 |
| Y | 12 | 15 | 18 | 23 | 15 | 16 | 20 | 30 |

Solution:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Current deviation in X | Current deviation in Y | Product of deviations |
| 23 | 12 |  |  |  |
| 26 | 15 | + | + | **+** |
| 28 | 18 | + | + | **+** |
| 20 | 23 | - | + | - |
| 25 | 15 | + | - | - |
| 25 | 16 | = | + | - |
| 28 | 20 | + | + | **+** |
| 30 | 30 | + | + | **+** |

Here, n= 8 => m= n-1 = 7 and C = no. of concurrent (+) deviations = 4

 So, **rc** = $\pm $$\sqrt{\pm \frac{(2C-m)}{m}}$

 = $\pm $$\sqrt{\pm \frac{(2x4-7)}{7}}$ **=** $\pm $$\sqrt{\pm \frac{1}{7}}$ **= + 0.3780**

Self Assessment Questions:

1. Define correlation and explain its types
2. Discuss about Scatter diagram
3. Define Karl Pearson’s coefficient of correlation. Show that it lies b/w -1 and +1
4. Show that correlation coefficient is independent of change of origin and scale.
5. Explain Rank correlation
6. Calculate the correlation coefficient b/w X and Y for the following data using i) Pearson’s method ii) Spearman rank method

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 24 | 28 | 34 | 45 | 56 | 56 | 80 | 90 |
| Y | 12 | 26 | 22 | 30 | 35 | 20 | 35 | 50 |