ST. LAWRENCE HIGH SCHOOL

A JESUIT CHRISTIAN MINORITY INSTITUTION

FOR GOD AND COUNTRY

STUDY MATERIAL-1
SUBJECT – STATISTICS

Pre-test

Chapter: BIVARIATE ANALYSIS **Class: XII**

Topic:CORRELATION **Date: 09.05.2020**

CORRELATION

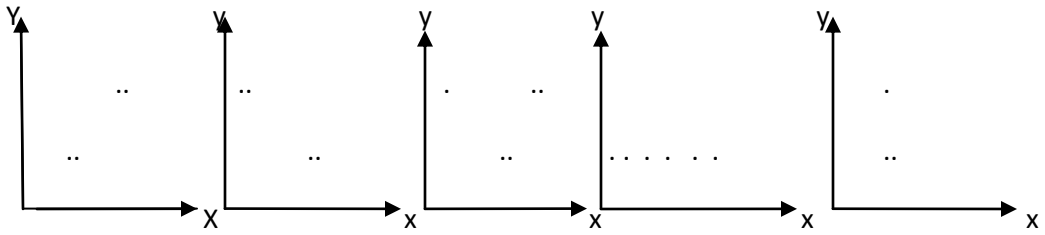
PART 1

CORRELATION

Correlation defines the qualitative associationship between two variables in terms of nature of change.

To check the correlation the bivariate points are being plotted on a plane with two axes. The diagram which we get is known as scatter diagram.

The following are some commonly observed scatter diagrams.



Natue of change:	same	opposite	can't be said	one variable doesn't change	
x ↑or ↓	y ↑or ↓	y ↓ or ↑	y can'I be inferred	y remains same	x remains same for y
Correlation:	positive	negative	zero	zero	zero

It is not always feasible to plot too many values in a scatter diagram and check the correlation. So we need to have an analytic measure which is known as correlation coefficient. It is denoted as r_{xy} .

Formula for $\boldsymbol{r_{xy} = \frac{cov\,(x,y)}{s_x s_y}}$.

Where cov(x , y) = covariance between x and y

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \quad \text{.....(*)}$$

- ✓ <u>Note</u>: From the formula cov (x, y) = cov (y, x), hence $r_{xy} = r_{yx}$

If we see the formula of variance

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

$$= \frac{1}{n} \sum_{i=1}^{n} x_i x_i - \bar{x}\bar{x} \quad \text{.....(**)}$$

- ✓ Note :So we can notice that in the formula of variance if one factor of x is being replaced by y, then it becomes the formula of covariance.

Hence comparing (*) and (**) it can be said that cov(x, x) = v(x)

Now if the variable x has unit as 'A' and y has unit as 'B' , then the unit of covariance becomes 'AB'.

For those variables the unit of standard deviation remains same as the unit of those variables. Hence the unit of the product of the standard deviations becomes 'AB'.

So correlation coefficient becomes an unit free measure.

Hence we can compare the correlations of (x and y), (y and z), (z and u),……..

PROPERTIES OF CORRELATION COEFFICIENT

- Change of base or origin and scale

  If $u_i = a + b\,x_i \;\; and \; v_i = c + d\,y_i \;\; for\; all\; i = 1(1)n,$

  then cov (u, v) = bd cov (x, y)

  And $r_{uv} \;=\; \pm\, r_{xy}$

Pf: cov (u, v) $= \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})$

= bd $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

= bd cov (x, y)

Now $r_{uv} = \frac{cov\ (u,v)}{s_u s_v} = \frac{bd\ cov(x,y\ )}{\mathrm{I}b\mathrm{I}\ s_x\ \mathrm{I}d\mathrm{I}\ s_y} = \frac{bd}{\mathrm{I}b\mathrm{I}\ .\ \mathrm{I}d\mathrm{I}}\ r_{xy}$

So r_{uv} = + r_{xy} when b and d are of same sign

= $-\, r_{xy}$ when b and d are of opposite sign.

Change of base or origin and scale

If $u_i = ax_i + b\,y_i \;\; and \; v_i = cx_i + d\,y_i \;\; for\; all\; i = 1(1)n,$

then $cov\ (u,v) = ac\ s_x^2 + \ bd\ s_y^2 + (\ ad + bc\ )cov\ (x,y)$

pf: $cov(u,v\ ) \;= \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})$

$= \frac{1}{n}\sum_{i=1}^{n}(\mathrm{a}(x_i - \bar{x}) + \ \mathrm{b}\,(\mathrm{y_i} - \bar{y}))(\mathrm{c}(x_i - \bar{x}) + \mathrm{d}\,(\mathrm{y_i} - \bar{y}))$

$$= ac\ \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + bd\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$+(ad+bc)\frac{1}{n}\sum(x_i - \bar{x})\,(\mathrm{y_i} - \bar{y})$$

$$= ac\ s_x^2 + bd\ s_y^2 + (ad+bc)cov(x,y)$$

$$s_u^2 = \frac{1}{n}\sum(\mathrm{u_i} - \bar{\mathrm{u}})^2 = \frac{1}{\mathrm{n}}\sum\left(\mathrm{a}(x_i - \bar{x}) + \mathrm{b}\left(\mathrm{y_i} - \overline{y}\right)\right)^2$$

$$= a^2\ s_x^2 + b^2\ s_y^2 + 2ab\ cov(x,y)$$

Similarly $s_v^2 = c^2\ s_x^2 + d^2\ s_y^2 + 2cd\ cov(x,y)$

Hence $s_u s_v r_{uv} = ac\ s_x^2 + bd\ s_y^2 + (ad+bc)\ s_x s_y r_{xy}$

By substituting the values we get the relation between $r_{uv}\ and\ r_{xy}$.

- ✓ Note: When we calculate the covariance of two linear functions we take the product of those two functions and then replace the square of the variable by its variance and product of the variables by its covariance.
  Similarly to calculate we take the square of the linear function the again replace the square of the variable by its variance and product of the variables by its covariance. So variance ⟷ square and cov ⟷product.

  (ax +by).(cx +dy) = ac x^2 + bd y^2 + (ad +bc) xy

  ↓↑ ↑↓ ↑↓ ↑↓ ↑↓

  Cov(ax+by,cx+dy) = ac s_x^2 + bds_y^2 + (ad+bc) cov(x, y)

  $(ax + by)^2 = a^2x^2 + b^2y^2 + 2ab\ xy$

  ↓↑ ↑↓ ↑↓ ↑↓

  Var(ax+ by) = $a^2 s_x^2 + b^2 s_y^2$ + 2ab cov(x, y)

- ➢ $-1 \le r \le 1$

Pf: Take $u_i = \frac{x_i - \bar{x}}{s_x}\ and\ v_i = \frac{y_i - \bar{y}}{s_y}$ for all I = 1(1)n.

$$\sum u_i^2 = n\ ,\ \sum v_i^2 = n\ and\ \sum u_i v_i = \frac{n.cov(x,y)}{s_x s_y} = n\ r_{xy}$$

Using Cauchy shewartz's inequality

$$\left(\sum u_i^2\right)\left(\sum v_i^2\right) \ge \left(\sum u_i v_i\right)^2 \Rightarrow n^2 \ge n^2 r_{xy}^2 \Longrightarrow -1 \le r_{xy} \le 1$$

- If x and y are independent then $r_{xy} = 0$ but the converse may not be true.

Pf: If x and y are independent then with the change of one variable the change of the other can not be inferred. Hence by the definition $r_{xy} = 0$

Let us consider the following example

X:	-3	-2	-1	0	1	2	3
Y:	9	4	1	0	1	4	9
xy:	-27	-8	-1	0	1	8	27

Cov (x, y) = $\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = 0 - 0.\bar{y} = 0$

But from the given data we get $y = x^2$. Hence x and y are not independent.

But x and y are linearly independent.

- ✓ Note: In correlation we can only determine the linear independence of the variables. This is the **limitation** of simple correlation.

Some important points to remember:

- Correlation is a qualitative measure.
- Correlation can be measured only between two variables only. Eg, correlation between the income and religion can not be measured as religion is not a variable. It is an attribute.
- Correlation coefficient does not depend upon the change of origin and magnitude of scale. It only depend on the sign of the scale.
- The value of correlation coefficient is zero only indicates that the variables are linearly independent.
- To calculate the covariance - two linear functions should be multiplied and in the result replace the product of the variables by the covariance and the square of the variable by the variance of the respective variables.

Prepared by

Sanjay Bhattacharya