UNIT 8 INTRA-CLASSES CORRELATION

Structure

8.1 INTRODUCTION

In Unit 5 of this block, you acquired the knowledge of fitting of various curves including straight line that shows the linear relationship between two variables. This linear relationship is measured by correlation coefficient that gives the strength of linear relationship. When characteristics are not measurable but ranks may be assigned to individuals according to their qualities, rank correlation is used to see the association between characteristics. Sometimes researchers are interested in studying the correlation among the members of the same family. In such cases intra-class correlation is used.

During the investigation more often it is observed that the two variables are not linearly related but they have some other type of curvilinear relationship. In this case correlation coefficient fails to give the strength of relationship and we use correlation ratio.

In this unit, you will study the coefficient of determination, correlation ratio and intra-class correlation. To understand the concept of correlation ratio and intra-class correlation, you are advised go through the concept and properties of correlation coefficient discussed in Unit 6 of this block.

Section 8.2 of this unit describes the coefficient of determination whereas correlation ratio is discussed with its properties in Section 8.3. Intra-class correlation coefficient is explained in Section 8.4. A derivation of limits of intra-class correlation coefficient is given in Sub-section 8.4.1.

Objectives

After reading this unit, you would be able to

- define the coefficient of determination;
- describe the correlation ratio;
- describe properties of correlation ratio;
- define the intra-class correlation coefficient; and
- describe the limits of intra-class correlation coefficient.

8.2 COEFFICIENT OF DETERMINATION

A measure which is used in statistical model analysis to assess how well a model explains and predicts future outcomes is known as coefficient of determination. The coefficient of determination is the measure of variation in the dependent variable that is explained by the regression function. In other words, the coefficient of determination indicates how much of the total variation in the dependent variable can be accounted by the regression function. This is the square of the correlation coefficient and denoted by r^2. It is expressed as a percentage. If coefficient of determination is 0.64, it implies that 64% of the variation in dependent variable Y is by the regression function or explained by the independent variable. Since it is the square of the correlation coefficient it ranges from 0 to 1.

The value of $r^2 = 0$ normal indicate that the dependent variable cannot be predicted from the independent variable, whereas the value of $r^2 = 1$ is the indication that the dependent variable can be predicted from the independent variable without error.

In analysis of variance, r^2 is the proportion of the total sum of squares which has been explained by linear regression. $(1 - r^2)$ is called the coefficient of non-determination.

Now solve the following exercises.

E1) If $r^2 = 0.75$, interpret the result.

E2) What would you conclude if $r^2 = 1$?

E3) What is the interpretation of $r^2 = 0$?

8.3 CORRELATION RATIO

Correlation coefficient measures the intensity or degree of linear relationship between two variables i.e. the extent of linear relationship can be explained by correlation coefficient if two variables are linearly related. If variables are not linearly related and show some curvilinear relationship then correlation coefficient is not a suitable measure to show the extent of relationship. In this type of cases, we study correlation ratio which is appropriate tool to know the degree of relationship between two variables i.e. concentration of points around the curve of best fit.

Correlation ratio is also used to measure the degree of association between a quantitative variable and another variable which may be qualitative or quantitative. Correlation ratio is determined only by the observations of dependent variable. Since quantitative scale for independent variable is not necessary so it is used only to classify dependent variable.

When regression is linear, the correlation coefficient and correlation ratio both produce the same results i.e. $r = \eta$, where η is correlation ratio.

So far we were dealing the situations where there was single value of Y corresponding to any value of X for example, data in the form

x	2	3	4	5
y	4	7	9	7

But in practice, there might be a situation where we have more than one values of y for each value of x. For example heights of 20 sons according to height of their fathers are given below:

Height of Fathers (in inches)	Height of sons (in inches)				
65	66	66	67	68	65
68	68	69	69	72	70
70	70	72	73	74	73
72	74	75	73	74	75

If we consider father's height by X and son's height by Y, in the above example more than one values of Y are available for each value of X.

In general X and Y may be in the following form

x_1: $y_{11}, y_{12},..., y_{1j},..., y_{1n}$

x_2: $y_{21}, y_{22},..., y_{2j},..., y_{2n}$

.

x_i: $y_{i1}, y_{i2},..., y_{ij},..., y_{in}$

.

x_m: $y_{m1}, y_{m2},..., y_{mj},..., y_{mn}$

Let us suppose that for each value of $x_i (i = 1, 2,..., m)$, variable Y has n values $y_{ij} (i = 1, 2,..., m; j = 1, 2,..., n)$ then mean of variable Y for i^{th} array is defined as

$$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

Then correlation ratio η is obtained by

$$\eta_{yx}^2 = \frac{\sum_{i=1}^{m} n_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} (y_{ij} - \bar{y})^2}$$

Now we present the above distribution with frequencies i.e.

x_1 $y_{11} f_{11}, y_{12} f_{12},..., y_{1j} f_{1j},..., y_{1n} f_{1n}$

x_2 $y_{21} f_{21}, y_{22} f_{22},..., y_{2j} f_{2j},..., y_{2n} f_{2n}$

.

x_i $y_{i1} f_{i1}, y_{i2} f_{i2},..., y_{ij} f_{ij},..., y_{in} f_{in}$

.

x_m $y_{m1} f_{m1}, y_{m2} f_{m2},..., y_{mj} f_{mj},..., y_{mn} f_{mn}$

It means for $x_i (i = 1, 2,..., m)$, Y takes values $y_{ij} (j = 1, 2,..., n)$ and frequency f_{ij} is attached with y_{ij}.

Note: You might have studied the frequency distribution earlier. Frequency is the number of repetitions of a value. If in a series of data, 2 is repeated 5 times then we say frequency of 2 is 5. And frequency distribution is the arrangement of values of variable with its frequencies.

In above frequency distribution $\sum_{j=1}^{n} f_{ij} = n_i$ and $\sum_{i=1}^{m} f_{ij} = n_j$

Total frequency $N = \sum_{i}^{m} n_i = \sum_{j}^{n} n_j$ i^{th} row can be considered as i^{th}array.

Then, mean of the i^{th} array can be defined as

$$\bar{y}_i = \frac{\sum_{j=1}^{n} f_{ij} y_{ij}}{\sum_{j=1}^{n} f_{ij}} = \frac{\sum_{j=1}^{n} f_{ij} y_{ij}}{n_i} = \frac{T_i}{n_i}$$

and over all mean

$$\bar{y} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} y_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} = \frac{\sum_{i=1}^{m} \bar{y}_i n_i}{\sum_{i=1}^{m} n_i} = \frac{T}{N}$$

Then, correlation ratio of Y on X is denoted by η_{yx} and defined as:

$$\eta_{yx}^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$$

where,

$$\sigma_e^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij} - \bar{y}_i)^2$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij} - \bar{y})^2$$

On simplifying, we get

$$\eta_{yx}^2 = \frac{\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_i - \bar{y})^2}$$

Some more simplification gives

$$\eta_{yx}^2 = \left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] \Big/ N\sigma_y^2$$

Example 1: Compute η_{yx}^2 for the following table:

y \ x	47	52	57	62	67
57	4	4	2	0	0
62	4	8	8	1	0
67	0	7	12	1	4
72	0	3	1	8	5
77	0	0	3	5	6

Solution: It is known that

$$\eta_{yx}^2 = \left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] / N\sigma_y^2$$

X / Y	47	52	57	62	67	$f_{i.}$	$f_{ij}y_{ij}^2$
57	4	4	2	0	0	10	32490
62	4	8	8	1	0	21	80724
67	0	7	12	1	4	24	107736
72	0	3	1	8	5	17	88128
77	0	0	3	5	6	14	83006
n_i	8	22	26	15	15	N = 86	$\sum_{i=1}^{m} f_{ij}y_{ij}^2 =$ 392084
T_i	476	1406	1717	1090	1090	$T = \sum_{i=1}^{m} T_i$ = 5782	$\bar{y} = \frac{T}{N} = \frac{5782}{86} = 67.23$
$\frac{T_i^2}{n_i}$	15705.07	48215.51	60165.08	44003.70	44003.70	$\sum_{i=1}^{m} \frac{T_i^2}{n_i} =$ 211493.07	
$\bar{y}_i$	59.50	64.05	66.04	72.67	72.67		

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij} - \bar{y})^2$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}y_{ij}^2 - \bar{y}^2$$

$$= \frac{392084}{86} - (67.23)^2 = 38.90$$

So,

$$\eta_{yx}^2 = \left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] / N\sigma_y^2$$

$$= \frac{211493.07 - (5782)^2/86}{86 \times 38.90}$$

$$= \frac{1624.76}{86 \times 38.90} = 0.485$$

E1) Compute η_{yx}^2 for the following table:

x / y	5	10	15	20	25
5	8	8	4	0	0
10	7	15	15	1	0
20	0	6	1	15	11
25	0	0	5	10	8

8.3.1 Characteristics of Correlation Ratio

1. Absolute value of correlation ratio is always less than or equal to one.
2. Absolute value of correlation ratio can not be less than the value of correlation coefficient.
3. η_{YX} is independent of change of origin and scale.
4. Correlation coefficient of X on Y i.e. r_{XY} and correlation coefficient of Y on X, r_{YX} are equal but correlation ratio of Y on X, η_{yx} and correlation ratio of X on Y, η_{xy} are different.
5. Difference $(\eta_{yx}^2 - r^2)$ measures the extent to which the true regression of Y on X departs from linearity.
6. It is based on only the values of dependent variable. Auxiliary variable is used just for classify the observations of independent variable.
7. When correlation is linear and forming a straight line then $\eta_{yx}^2 = r^2$ are same.
8. When scatter diagram does not show any trend then both η_{yx} and r are zero.
9. If scatter diagram shows straight line and dots lie precisely on a line, then the correlation coefficient and the correlation ratio, both are 1, i.e. $\eta_{yx} = r = 1$.
10. If $\eta_{yx} > r$ then, scatter diagram shows curved trend line.

8.4 INTRA-CLASS CORRELATION COEFFICIENT

In product moment correlation, both variables measure different characteristics e.g. one variable is price and another is demand. Similarly one variable is expenditure on advertisement and another may be sales revenue. But in many practical situations specially in medical, agriculture and biological field one may be interested to know the correlation among the units of a group or a family. For example, in agricultural experiment scientist may be interested to know the correlation among the yields of the plots of the block given same fertilizer.

In the study of heights of brothers, one may be interested in correlation between the heights of brothers of a family. In such correlation both variables measure the same characteristics, i.e. yield and height. By this correlation, we mean the extent to which the members of the same family resemble each other with respect to the considered characteristic. This correlation is called intra-class correlation.

Let us consider n families $F_1, F_2, ..., F_n$ with number of members $k_1, k_2, ..., k_n$ respectively. Let $x_{ij}\ (i = 1, 2, ..., n; j = 1, 2, ..., k_i)$ be the value of j^{th} member of the i^{th} family.

In i^{th} family there would be $k_i(k_i-1)$ pairs of observation. So the total number of pairs are $\sum_{i=1}^{n} k_i(k_i-1) = N$.

Note: If in a family there are three members 1, 2 and 3 then there would be (1, 2) (2, 1) (1, 3) (3, 1) (2, 3) and (3, 2) i.e. 6 pairs of observation, here number k = 3 so k (k - 1) = 3(3 - 1) =3×2 = 6 pairs.

Table giving the values of N pairs of observations is called intra-class correlation table and the product moment correlation coefficient calculated from $\sum_{i=1}^{n} k_i(k_i-1) = N$ pairs of observations is called intra-class correlation coefficient. Since the value of each member of the i^{th} family occurs (k_i-1) times as a value of the X variable as well as a value of the Y variable. Then mean of variable X and Y are same and

$$\bar{x} = \bar{y} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(k_i-1)x_{ij}}{\sum_{i=1}^{n}k_i(k_i-1)} = \frac{1}{N}\left[\sum_{i=1}^{n}(k_i-1)\sum_{j=1}^{k_i}x_{ij}\right]$$

Similarly, both variable X and Y have same variance which is

$$\sigma_x^2 = \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{n}(k_i-1)\sum_{j=1}^{k_i}(x_{ij}-\bar{x})^2 ,$$

and covariance between X and Y is

$$Cov(x,y) = \frac{1}{N}\sum_{i}\sum_{j\neq 1}(x_{ij}-\bar{x})(x_{il}-\bar{x})$$

Finally, formula for intra-class correlation coefficient is

$$r_{ic} = \frac{\sum_{i=1}^{n}k_i^2(\bar{x}_i-\bar{x})^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2}{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(k_i-1)(x_{ij}-\bar{x})^2}$$

If $k_i = k$ i.e. each family have equal number of members then

$$r_{ic} = \frac{nk^2\sigma_m^2 - nk\sigma_x^{\ 2}}{(k-1)nk\sigma_x^{\ 2}}$$

$$r_{ic} = \frac{1}{(k-1)}\left\{\frac{k\sigma_m^2}{\sigma_x^{\ 2}} - 1\right\}$$

where, $\sigma_x^{\ 2}$ is the variance of X and σ_m^2 is the variance of means of families.

8.4.1 Limits of Intra-class Correlation Coefficient

Intra-class correlation coefficient is

$$r_c = \frac{1}{(k-1)}\left\{\frac{k\sigma_m^2}{\sigma_x^{\ 2}} - 1\right\}$$

$$\Rightarrow r_c(k-1) = \left\{\frac{k\sigma_m^2}{\sigma_x^{\ 2}} - 1\right\}$$

$$\Rightarrow r_c(k-1)+1=\frac{k\sigma_m^2}{\sigma_x^2}\geq 0$$

$$\Rightarrow r_c \geq -\frac{1}{k-1}$$

Thus, lower limit of intra-class correlation coefficient is $-\frac{1}{k-1}$

also

$$1+(k-1)r_c \leq k \quad \text{as} \quad \frac{\sigma_m^2}{\sigma_x^2}\leq 1$$

$$r_c \leq 1$$

Thus, $-\frac{1}{k-1} \leq r_c \leq 1$

8.5 SUMMARY

In this unit, we have discussed:

1. The coefficient of determination;
2. Properties of the coefficient of determination;
3. Concept of correlation ratio;
4. Properties of correlation ratio; and
5. The intra-class correlation coefficient.

8.6 SOLUTIONS / ANSWERS

E1) It is known that

$$\eta_{yx}^2=\left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right)-\frac{T^2}{N}\right]/N\sigma_y^2$$

X / Y	5	10	15	20	25	$f_{i.}$	$f_{ij}y_{ij}^2$
5	8	8	4	0	0	20	500
10	7	15	15	1	0	38	3800
15	0	6	1	15	11	45	10125
20	0	0	5	10	8	33	13200
25	5	10	15	20	25	23	14375
n_i	15	41	49	27	27	$\sum_{i=1}^{m} n_i = N = 159$	$\sum_{i=1}^{m} f_{ij}y_{ij}^2 = 42000$
T_i	110	490	675	575	540	$T=\sum_{i=1}^{m} T_i = 2390$	$\bar{y}=\frac{T}{N}=\frac{2390}{159}$ $= 264.1509$
$\frac{T_i^2}{n_i}$	806.66	5856.1	9298.47	12245.37	10800	$\sum_{i=1}^{m}\frac{T_i^2}{n_i}=3900660$	
$\bar{y}_i$	7.33	11.9	13.77	21.296	20.00		

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij} - \bar{y})^2$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}y_{ij}^2 - \bar{y}^{\,2}$$

$$= \frac{42000}{159} - (246.1509)^2$$

$$= 38.2066$$

So,

$$\eta_{yx}^2 = \left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] / N\sigma_y^2$$

$$= \frac{39006.6040 - (2390)^2 / 159}{159 \times 38.2066}$$

$$= 0.9211$$