Frequently Asked Questions

1. What is Correlation ratio?

Answer:

In statistics, the correlation ratio is a measure of the relationship between the statistical dispersion within individual categories and the dispersion across the whole population or sample.

2. Define the measure of correlation ratio.

Answer:

The measure is defined as the *ratio* of two standard deviations representing these types of variation.

3. Why we use correlation ratio?

Answer:

In many cases, the desired outcome of our learning algorithm is categorical (a ``yes/no'' answer or a limited set of choices). The correlation coefficient assumes that the outcome is quantitative, thus it is not applicable to the categorical case. In order to sort out general dependencies, the *correlation ratio* method can be used.

4. What is range for correlation ratio?

Answer:

The correlation ratio η takes values between 0 and 1.

5. How to interpret the extreme values taken by the correlation ratio?

Answer:

The limit η = 0 represents the special case of no dispersion among the means of the different categories, while η=1 refers to no dispersion within the respective categories. Note further, that η is undefined when all data points of the complete population take the same value.

6. Obtain an expression for the correlation ratio.

Answer:

Suppose each observation is y_{xi} where x indicates the category that observation is in and i is the label of the particular observation. Let n_x be the number of observations in category x.

Then $\overline{y}_x = \frac{\sum_i y_{xi}}{n_x}$ and $\overline{y} = \frac{\sum_x n_x \overline{y}_x}{\sum_x n_x},$

Where $\bar{y}_x$ is the mean of the category x and $\bar{y}$ is the mean of the whole population. The correlation ratio η is defined as to satisfy

$$\eta^2 = \frac{\sum_x n_x (\overline{y}_x - \overline{y})^2}{\sum_{x,i} (y_{xi} - \overline{y})^2}$$

This can be written as,

$$\eta^2 = \frac{{\sigma_{\overline{y}}}^2}{{\sigma_y}^2}, \text{ where } {\sigma_{\overline{y}}}^2 = \frac{\sum_x n_x(\overline{y}_x - \overline{y})^2}{\sum_x n_x} \text{ and } {\sigma_y}^2 = \frac{\sum_{x,i}(y_{xi} - \overline{y})^2}{n},$$

I.e. the weighted variance of the category means divided by the variance of all samples.

7. Find correlation ratio for the following data regarding the distribution of test scores in three topics:

   Algebra: 45, 70, 29, 15 and 21
   Geometry: 40, 20, 30 and 42
   Statistics: 65, 95, 80, 70, 85 and 73

Answer:

Suppose there is a distribution of test scores in three topics (categories):

- Algebra: 45, 70, 29, 15 and 21 (5 scores)
- Geometry: 40, 20, 30 and 42 (4 scores)
- Statistics: 65, 95, 80, 70, 85 and 73 (6 scores).

Then the subject averages are
For algebra, (45+70+29+15+21)/5=36,
For geometry (40+20+30+42)/4= 33 and
For statistics, (65+95+80+70+85+73) / 6=78
The overall average is (45+70+29+15+21+40+20+30+42+65+95+80+70+85+73)/15= 52
The sums of squares of the differences from the subject averages are given below.
For Algebra,
$(45-36)^2+ (70-36)^2+ (29-36)^2+ (15-36)^2+ (21-36)^2=1952$
For Geometry,
$(40-33)^2+ (20-33)^2+ (30-33)^2+ (42-33)^2= 308$ and
For Statistics
$(65-78)^2+ (95-78)^2+ (80-78)^2+ (70-78)^2+ (85-78)^2 + (73-78)^2= 600$ adding above three we get, 2860.
The overall sum of squares of the differences from the overall average is given by,
$(45-52)^2+(70-52)^2+(29-52)^2+(15-52)^2+(21-52)^2+(40-52)^2+(20-52)^2+(30-52)^2+(42-52)^2+(65-52)^2+(95-52)^2+(80-52)^2+(70-52)^2+(85-52)^2+(73-78)^2 =9640.$
The difference of 6780 between these is also the weighted sum of the square of the differences between the subject averages and the overall average:

$$5(36-52)^2 + 4(33-52)^2 + 6(78-52)^2 = 6780$$

This gives

$$\eta^2 = \frac{6780}{9640} = 0.7033\ldots$$

suggesting that most of the overall dispersion is a result of differences between topics, rather than within topics. Taking the square root

$$\eta = \sqrt{\frac{6780}{9640}} = 0.8386\ldots$$

8. Write a note on comments given by Pearson and Fisher.

Answer:

The correlation ratio was introduced by Karl Pearson as part of analysis of variance. Ronald A Fisher commented:

- *As a descriptive statistic the utility of the correlation ratio is extremely limited. It will be noticed that the number of degrees of freedom in the numerator of depends on the number of the arrays*

To which Egon Pearson (Karl's son) responded by saying

- *Again, a long-established method such as the use of the correlation ratio [The "Correlation Ratio" η] is passed over in a few words without adequate description, which is perhaps hardly fair to the student who is given no opportunity of judging its scope for him.*

9. How correlation ratio is used in regression?

Answer:

The correlation coefficient r is a good measure of correlation only when the regression is linear. It is necessary, therefore, before placing any reliance upon the computed r to examine the data for linearity of the regression lines. One common test of linearity is Blakeman's; compare the value of $\eta^2 - r^2$ with its probable error, η being a correlation ratio. In applying this test it is then necessary to calculate along with r the correlation ratios.

10. Give an expression for correlation ratio of Y relative to X

Answer:

The correlation ratio of a random variable Y relative to a random variable X is given by the expression,

$$\eta^2_{Y|X} = 1 - \mathsf{E}\left[\frac{\mathsf{D}(Y|X)}{\mathsf{D}Y}\right]$$

Where DY is the variance of Y and D(Y/X) is the conditional variance of Y given X which characterizes the spread of Y about its conditional mathematical expectation E(Y/X) for a given value of X.

11. What is the range for correlation ratio of Y relative to X and how to interpret the extreme values?

Answer:

Invariably, $0 \le {\eta_{Y/X}}^2 \le 1$

The equality ${\eta_{Y/X}}^2 = 0$ corresponds to non-correlated variables; ${\eta_{Y/X}}^2 = 1$ if there is an exact functional relationship between Y and X.

12. What do you mean by intra-class correlation?

Answer:

Intra-class correlation means within class correlation. Sometimes especially in biological and agricultural study, it is of interest to know how the members of a family or group are correlated among themselves with respect to some one of their common characteristic.

13. Derive an expression of intra-class correlation coefficient.

Answer:

Suppose we have $A_1, A_2 \ldots A_n$ families with k_1, k_2, k_n members.

And let x_{ij} (i=1, 2 … n; j=1, 2,k_i) denote the measurement on the j^{th} member in the i^{th} family.

We shall have k_i (k_i -1) pairs for each i^{th} family or group like (x_{ij} and x_{il}), j not equal to l.

There will be $\sum_{i=1}^{n} k_i(k_i - 1) = N$ pairs for all the n families or groups. If we prepare a

correlation table there will be k_i (k_i -1) 1 entries for the i^{th} group or family and $\sum_{i=1}^{n} k_i(k_i-1) = N$ entries for all the n families or groups.

The table is symmetrical about the principal diagonal. Such a table is called intra-class correlation table and the correlation is called intra-class correlation.

In a bivariate table, x_{i1} occurs ($k_i - 1$) times, x_{i2} occurs ($k_i - 1$) times . . . x_{iki} occurs ($k_i - 1$) times. i.e., from the i^{th} family we have ($k_i - 1$) Σx_{ij} and hence for all the n families we have $\sum_i \left[(k_i - 1) \sum_j x_{ij} \right]$ as the marginal frequency, the table being symmetrical about the principal diagonal.

$$\bar{x} = \bar{y} = \frac{1}{N}\left[\sum_i \left\{ (k_i - 1) \sum_j x_{ij} \right\} \right]$$

Similarly, $\sigma_x = \sigma_y = \frac{1}{N}\left[\sum_i \left\{ (k_i - 1) \sum_j \left(x_{ij} - \bar{x}\right)^2 \right\} \right]$

Further $Cov(X,Y) = \frac{1}{N} \sum_i \left[\sum_{i,l} (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right], j \neq l$

$$= \frac{1}{N} \sum_i \left[\sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{ij} - \bar{x})(x_{il} - \bar{x}) - \sum_{j=1}^{k_i} (x_{ij} - \bar{x})^2 \right]$$

If we write $\bar{x}_i = \sum_j x_{ij} / k_i$, then

$$\sum_i \left[\sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right] = \sum_i \left[\sum_j (x_{ij} - \bar{x}) \sum_l (x_{il} - \bar{x}) \right]$$

$$= \sum_i \left[k_i(\bar{x}_i - \bar{x}) k_i(\bar{x}_i - \bar{x}) \right] = \sum_i \left[k_i^2 (\bar{x}_i - \bar{x})^2) \right]$$

Therefore intra-class correlation coefficient is given by,

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sum_i k_i^2 (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{\sum_i \sum_j (k_i - 1)(x_{ij} - \bar{x})^2}$$

If we put k i is equal to k, that is if all families have equal members then,

$$r = \frac{k^2 \sum_i (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{(k-1) \sum_i \sum_j (x_{ij} - \bar{x})^2} = \frac{nk^2 \sigma_m^2 - nk\sigma^2}{(k-1)nk\sigma^2} = \frac{1}{(k-1)} \left\{ \frac{k\sigma_m^2}{\sigma^2} - 1 \right\}$$

Where σ^2 denotes the variance of X and σ^2_m the variance of means of families.

14. Find the limit for intra-class correlation coefficient

Answer:

We have from above r= $= \frac{1}{(k-1)} \left\{ \frac{k\sigma_m^2}{\sigma^2} - 1 \right\}$

Or $1 + (k-1)r = \frac{k\sigma_m^2}{\sigma^2} \geq 0 \Rightarrow r \geq -\frac{1}{k-1}$

Also, 1+ (k-1) r ≤ k as the ratio, $\frac{\sigma_m^2}{\sigma^2} \leq 1 \Rightarrow r \leq 1$

$-\frac{1}{k-1} \leq r \leq 1$

15. Comment on the coefficient of intra-class correlation.

Answer:

Intra-class correlation cannot be less than -1/(k – 1) though it may attain the value plus 1 on the positive side, so that it is a skew coefficient and a negative value has not the same significance as a departure from independence as an equivalent positive value.