1. Introduction

Welcome to the series of E-learning modules on correlation ratio and intra class correlation coefficient. In this module we are going to study about the correlation ratio, range for the correlation ratio, examples on correlation ratio, comparison of correlation ratio given by Pearson and Fisher, correlation ratio and regression, intra class correlation, expression for intra class correlation, limits for it and the interpretation.

By the end of this session, you will be able to

- Explain about correlation ratio
- Explain range for correlation ratio
- Understand comparison of correlation ratio given by Pearson and Fisher
- Distinguish between correlation ratio and regression
- Explain about intra-class correlation
- Explain the limits for intra-class correlation coefficient

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.

The measure is defined as the *ratio* of two standard deviations representing these types of variation. The context here is the same as that of the intra-class correlation coefficient, whose value is the square of the correlation ratio.

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.

The basic idea behind the correlation ratio is to partition the sample feature vectors into classes according to the observed outcome. If a feature is significant, then it should be possible to identify at least one outcome class where the feature's average value is significantly different from the average on all classes, otherwise that component would not be useful to discriminate any outcome.

2. Correlation Ratio and Range of Correlation Ratio

Now let us see the range of correlation ratio.
The correlation ratio eta takes values between 0 and 1. The limit eta is equal to zero represents the special case of no dispersion among the means of the different categories, while η is equal to 1 refers to no dispersion within the respective categories. Note further, that eta is undefined when all data points of the complete population take the same value.

Let us define correlation ratio.
Suppose each observation is *yx i* 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
Y bar x is equal to summation over i y x I divided by n x and y bar is equal to summation over x n x into y bar x divided by summation over x n x, where y bar x is the mean of the category x and y bar is the mean of the whole population.

Correlation ratio eta is defined as to satisfy
Eta square is equal to summation over x n x into y bar x minus y bar whole square divided by summation over x and i y x i minus y bar whole square, which can be written as
Eta square is equal to sigma y bar square divided by sigma y square, where sigma y bar square is equal to summation over x n x into y bar x minus y bar whole square divided by summation over x, n x
And sigma y square is equal to summation over x and i y x i minus y bar whole square divided by n.
That is the weighted variance of the category means divided by the variance of all samples.

It is worth noting that if the relationship between values of x and values of y bar x is linear (which is certainly true when there are only two possibilities for *x*) this will give the same result as the square of the correlation coefficient; otherwise the correlation ratio will be larger in magnitude. It can therefore be used for judging non-linear relationships.

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

- Algebra: forty five, seventy, twenty nine, fifteen and twenty one (that is five scores)
- Geometry: forty, twenty, thirty and forty two (that is four scores)
- Statistics: sixty five, ninety five, eighty, seventy, eighty five and seventy three (that is six scores).

Then the subject averages,
For algebra is, 45 plus 70 plus 29 plus 15 plus 21 whole divided by 5 is equal to 36.
For geometry is, 40 plus 20 plus 30 plus 42 whole divided by 4 is equal to 33
And For statistics is, 65 plus 95 plus 80 plus 70 plus 85 plus 73 whole divided by 6 is equal to 78.
The overall average is given by,

45 plus 70 plus 29 plus 15 plus 21 plus 40 plus 20 plus 30 plus 42 plus 65 plus 95 plus 80 plus 70 plus 85 plus 73 whole divided by 15 is equal to 52.

Now let us find the sum of square of different subjects as follows.
For algebra,
Forty five minus thirty six square plus seventy minus thirty six square plus twenty nine minus thirty six square plus fifteen minus thirty six square plus twenty one minus thirty six square is equal to one thousand nine hundred and fifty two.
For geometry,
Forty minus thirty three square plus twenty minus thirty three square plus thirty minus thirty three square plus forty minus thirty three square is equal to three hundred and eight
For Statistics
Sixty five minus seventy eight square plus ninety five minus seventy eight square plus eighty minus seventy eight square plus seventy minus seventy eight square plus eighty five minus seventy eight square plus seventy three minus seventy eight square is equal to six hundred.
Adding the above three we get two thousand eight hundred and sixty.

The overall sum of squares of the differences from the overall average is given by,
45 minus 52 square plus 70 minus 52 square plus 29 minus 52 square plus 15 minus 52 square plus 21 minus 52 square plus 40 minus 52 square plus 20 minus 52 square plus 30 minus 52 square plus 42 minus 52 square plus 65 minus 52 square plus 95 minus 52 square plus 80 minus 52 square plus 70 minus 52 square plus 85 minus 52 square plus 73 minus 52 square is equal to 9 thousand 640.

The difference between 9 thousand 640 and 2 thousand 860 is equal to 6 thousand 780, which is also the weighted sum of the square of the differences between the subject averages and the overall average.
That is 5 into 36 minus 52 square plus 4 into 33 minus 52 square plus 6 into 78 minus 52 square is equal to 6 thousand 780.

This gives eta square is equal to 6 thousand 780 divided by 9 thousand 640 is equal to zero point 7 zero 3 three, suggesting that most of the overall dispersion is a result of difference between topics, rather than within topics. Taking the square root,
Eta is equal to square root of thousand 780 divided by 9 thousand 640 is equal to zero point 8 three 8 six.

3. Comments and Role of Correlation Ratio in Regression

Now let us see the comments given by Ronal A Fisher and Egon Pearson.
The correlation ratio was introduced by Karl Pearson as part of analysis of variance. Ronal 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 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" eta] 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 himself.*

Now let us discuss the role of correlation ratio in regression.
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 square minus r spare with its probable error, eta being a correlation ratio. In applying this test it is then necessary to calculate along with r the correlation ratios.

The correlation ratio of a random variable Y relative to a random variable X is given by the expression,
Eta y given x square is equal to 1 minus E of D y given X divided by D y
Where D Y is the variance of Y and D of Y given Y is the conditional variance of Y given X which characterizes the spread of Y about its conditional mathematical expectation E of Y given X for a given value of X

Invariably, zero less than or equal to eta y given x square less than or equal to one. The equality eat Y given X square is equal to zero corresponds to non correlated variables; eta Y given X square is equal to one, if and only if there is an exact functional relationship between Y and X.

4. Intra Class Correlation

Now let us discuss about intra-class correlation.
Intra-class correlation means within class correlation. It is distinguishable from product moment correlation in as much as here both the variables measure the same characteristics. 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.

For example, we may require the correlation between the heights of brothers of a family or between yields of plots of an experimental block. In such cases both the variables measure the same characteristic, example, height and height or weight or weight. There is nothing to distinguish one from the other so that one may be treated as X variable and the other as the Y variable.

Suppose we have A_1, A_2, etc., A_n families with k_1, k_2, etc. till k_n members, each of which may be represented as shown below.
X one one, x two one, etc x i one, etc ,x n one
X one two, x two two, etc, x i two, etc x n two, etc
X one i, x two j, etc x i j, etc x n j etc.,
x one k one, x two j two, etc x i k, i, etc x n, k, n

And let x i, j (where i is equal to one, 2, etc., n; and j is equal to one, 2, etc., k i) denote the measurement on the j^{th} member in the i^{th} family.

We shall have k i into k i minus one pairs for each i^{th} family or group like (x i j and x i l), j not equal to l. there will be summation over i from one to n k i into k i minus one is equal to N pairs for all the n families or groups.

If we prepare a correlation table there will be k i into k i minus one entries for the i^{th} group or family and summation over i from one to n k i into k i minus one is equal to 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 i one occurs k i minus one times, x i two occurs k i minus one times etc., x i k i occurs k i minus one times. That is from the i^{th} family we have k i minus one into summation over j, x i j and hence for all the n families we have summation over i, k i minus one into summation over j, x i j as the marginal frequency, the table being symmetrical about the principal diagonal.

X bar is equal to y bar is equal to one by N into summation over i, k i minus one into summation over j, x i j
Similarly, sigma x square is equal to sigma y square is equal to one by N into summation over i, k i minus one into summation over j, x i, j minus x bar whole square

Further covariance of X, Y is equal to one by into summation over i, summation over j and l, x i, j minus x bar into x i l minus x bar, where j is not equal to l

Is equal to summation over I, summation over j is equal to one to k i, summation over l is equal to one to k i, x i,l minus x bar into x i j minus x bar minus summation over j is equal to one to k i, x i, j minus x bar whole square.

If we write x i bar is equal to summation over j, x i, j divided by k i, then
Summation over i, summation over j is equal to one to k i, summation over i is equal to one to k i, x i j minus x bar into x i,l minus x bar
Is equal to summation over i, summation over j x i, j minus x bar into summation over l x i, l minus x bar
Is equal to summation over i, k i into x i bar minus x bar into k i into x i bar minus x bar
Is equal to summation over i, k i square into x i bar minus x bar whole square.

Therefore intra-class correlation coefficient is given by,
r X Y is equal to covariance of X, Y divided by square root of variance of X into variance of Y
is equal to summation over i, k i square into x i bar minus x bar whole square minus summation over i, summation over j, x i j minus x bar square divided by summation over i, summation over j, k i minus one into x i j minus x bar square

If we say k_i is equal to k, that is if all families have equal members then,
R is equal to k square into summation over i, x i bar minus x bar whole square minus summation over i, summation over j, x i j minus x bar the whole square divided by k minus one into summation over i, summation over j, into x i j minus x bar the whole square.
Is equal to n into k square into sigma m square minus n into k into sigma square divided by k minus one into n into k into sigma square

Is equal to one by k minus one into k into sigma m square divided by sigma square minus one
Where sigma square denotes the variance of X and sigma m square the variance of means of families.

5. Limit of Intra Class Correlation Coefficient

Now let us find the limit of the intra-class correlation coefficient.
We have from above r is equal to one by k minus one into k into sigma m square divided by sigma square minus one.
Or
one plus k minus one into r is equal to k into sigma m square divided by sigma square, which is greater than or equal to zero.
Implies, r is greater than or equal to minus one divided by k minus one

Also one plus k minus one into r is less than or equal to k as the ratio, sigma m square divided by sigma square is less than or equal to one
Implies r is less than or equal to one
Therefore, minus one divided by k minus one is less than or equal to r is less than or equal to one

Now let us discuss how to interpret the data
Intra-class correlation cannot be less than minus one divided by k minus one 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.

Interclass correlation, or Pearson's correlation, is a correlation in the ordinary sense. It is used to estimate the correlation between two different variables or
between two groups. The intra-class correlation is not the correlation between a predictor variable and the dependent variable but it reflects the extent to which members of the same group or class tend to act alike. It is the proportion of the total variability in the measured factor that is due to the variability between individuals.

The intra- class Correlation Coefficient is similar to the more familiar Pearson correlation coefficient, except that it is sensitive to absolute size of the values in two arrays. Thus whereas the Pearson correlation coefficient would be unchanged by adding a constant to all the values in one of two correlated arrays, the intra- class Correlation Coefficient would decrease as the mean of the two arrays became more discrepant.

The intra- class Correlation Coefficient thus gives a global index of similarity between two waveforms; a low intra- class Correlation Coefficient can arise if there is either amplitude or latency differences in peaks and troughs. The intra- class Correlation Coefficient agrees well with ratings of waveform similarity made by untrained observers on the basis of visual inspection

Here's a summary of our learning in this session where we have:

- Understood about Correlation ratio
- Explained Range for correlation ratio

- Understood the comparison of correlation ratio given by Pearson and Fisher
- Explained Correlation ratio and regression
- Understood Intra-class correlation and
- Limits for intra-class correlation coefficient