1. Introduction

Welcome to the series of E-learning modules on Various measures for association of two-way. In this module we are going to discuss about the different methods of studying association and manifold classification.

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

- Explain different measure of studying association of two way classified attributes, five different methods
- Explain manifold Classification

The study of association can be done by any of the following methods:

1. Comparison of actual and observed frequencies
2. Comparison of various proportions and products
3. Calculation of Yule's Coefficient of association
4. Calculation of coefficient of Collignation
5. Calculation of coefficient of Contingency

In the above mentioned methods, the last one that is calculation of coefficient of contingency is generally used to study manifold classification.
We shall now examine each one of the above mentioned methods.

2. Comparison of Actual and Observed Frequencies

Let us consider the first method, Comparison of actual and observed frequencies:
Whenever we want to study the association between two attributes A and B we try to find out whether attribute A is more commonly found with attribute B than in ordinarily expected. Thus, in a study of association the first thing to be calculated is expected value of (AB). This value is calculated on the basis of simple rules of probability.

Thus, if two attributes A and B are studied in a universe and if the frequency of A is represented by (A) and of B by (B), then
The probability of (A) is equal to (A) by N and
The probability of (B)is equal to (B) by N
The combined probability of two independent events is equal to the product of their individual probabilities. Thus, the combined probability of (A) and (B) would be (A) by N into (B) by N and the expectation is obtained by multiplying the probability by N.

Hence the expectation of (A) and (B) combined would be, (A) into (B) divided by N.
From the above it is clear that ordinarily if attributes A and B are independent the expected frequency of (AB) would be equal to (A) into (B) divided by N.

We can give criterion for independence of any two attributes as follows.

If there is no kind of relationship between the attributes A and B we may expect to find the same proportion of A's in B's as in betas. In other words, attribute A must be equally popular in B's and in not Beta's. If for example, blindness and deafness are not associated, the proportion of blind people amongst the deaf and amongst the hearing must be equal. If however, it is found that the proportion of blind people amongst the deaf is more than their proportion amongst hearing, it indicates that blindness and deafness have association.

Two attributes A and B are said to be independent if the observed frequency of (AB) is equal to its expected frequency that is, (A) into (B) divided by N.

The main limitation of studying association by a comparison of actual and expected values of AB is that it only determines the nature of association between A and B that is, whether the association (if any) is positive or negative. It does not tell us about the degree of association that is, whether it is high or low.

At this stage it is necessary to point out that if the value of (AB) is found to be greater than the value of (A) into (B) divided by N, it should not be at once concluded that there is positive association between the two attributes. It is quite possible, particularly when the difference between observed and expected values is not much, that the association may be the result of sampling fluctuations and the true association may be zero. As such, unless the difference between the observed and expected values is very significant we should not conclude that there is any association or disassociation between the two attributes.

The question which naturally arises here is how much divergence between the observed and actual values can be termed as significant. We shall discuss this question in detail in the coming modules of sampling and chi-square tests.

3. Comparison of Various Proportions and Products

Now let us consider the second method, comparison of various proportions and products.

As pointed out earlier the main limitation in the method of comparing actual and expected value of attributes and B lies in the fact that it does not give any idea about the degree of association. A slightly better method would be the comparison of proportions between various classes.

Thus attributes A and B are:

i. Independent if (AB) by B is equal to (A beta) by (Beta)
ii. Positively associated if (AB) by B is greater than (A beta) by (Beta)
iii. Negatively associated (AB) by B is less than (A beta) by (Beta)

If the relation (i) holds good, the corresponding relationship, alpha B by B is equal to alpha beta by beta; (AB) by (A) is equal to alpha B by alpha; A Beta by A is equal to alpha beta by alpha would also hold good.

Further it can also be concluded that A and B would be independent if,
(AB) by (A) is equal to (B) by B; (AB) by B is equal to (A) by (N) and (AB) by B is equal to (A) by (N) is equal to (A) by (N) into (B) by (N).

It can also be found out easily that A and B would be independent if
(AB) into (alpha beta) is equal to (A beta) into (alpha beta).

Now let us consider the Yule's coefficient of Association.
So far we have discussed a rough idea about the extent of association or disassociation between two attributes by finding out the extent of the difference between their observed and expected frequencies or the difference in various proportions. For practical purpose it is enough to take a decision about whether the two attributes in question are associated, disassociated or independent. But in some cases the difference between observed and expected frequencies may be due to fluctuations of sampling. Under such circumstances it becomes necessary to obtain an idea about the extent to which the difference between the observed and expected frequencies can be due to chance fluctuations.

It would be convenient if the coefficient of association is such that its value is zero when the two attributes are independent, plus 1 when they perfectly associated and minus 1 when they are perfectly disassociated. Many such coefficients of association have been worked out by different authors but the one given by Yule is very easy and simple.
Yule's coefficient of association,
Q is equal to (AB) into (alpha beta) minus (A beta) into (alpha beta) whole divided by (AB) into (alpha beta) plus (A beta) into (alpha beta).

We know that when two attributes A and B are independent the value of (AB) into (alpha beta)

is equal to (A beta) into (alpha B). As such, if two attributes are independent, the value of the numerator in the above formula would be zero and the value of the coefficient of association would also be zero. Similarly, if there is perfect association between the two attributes A and B the value of (A beta) into (alpha B) would be zero and since it will be so both in the numerator and the denominator. It is evident that the value of the coefficient of association would be plus 1. Similarly, if there is perfect disassociation between the two attributes A and B the value of (AB) into (alpha beta) would be zero and it will be both in the numerator and denominator, the coefficient of association would be minus 1.

With above argument we know that the value of Q lies between -1 and 1. Now let us prove this using notation.
Consider (AB) into (alpha beta) is equal to a and (A beta) into (alpha B) is equal to B. Then a is greater than or equal to zero and b is also greater than or equal to zero.

Therefore, modulus of a minus b is less than or equal to modulus of a plus b
Implies modulus of a minus b divided by a plus b is less than or equal to 1.

But a minus b divided by a plus b gives the expression of Q
Hence, modulus of Q is equal to modulus of a minus b divided by a plus is less than or equal to 1
Implies minus 1 less than or equal to Q less than or equal to 1.

Following is an advantage of Yule's coefficient of Association.
An important property of Q is that it is independent of the relative proportion of A's and alpha's in the data. Thus, if all the terms containing A in Q are multiplied by a constant k, say, its value remains unaltered. Similarly, we have to do for B and beta and alpha. This property renders it especially useful to situations where the proportions are arbitrary, example, experiments.

Another important coefficient given by Yule is the Coefficient of Collignation. This is also independent of the relative proportions of A's and alpha's (like Yule's coefficient of association). This coefficient is denoted by gamma. The formula of calculation is as follows:
Gamma is equal to 1 minus square root of (A beta) into (alpha B) divided by (AB) into (alpha beta) whole divided by
1 plus square root of (A beta) into (alpha B) divided by (AB) into (alpha beta).

4. Calculation of Coefficient of Collignation

Now let us consider some remarks on coefficient of Collignation.

1. If Q is equal to zero then (AB) into (alpha beta) is equal to A beta into alpha B. Implies gamma is equal to 1 minus 1 divided by 1 plus 1 is equal to zero
   Q is equal to minus 1 implies, gamma is equal to minus 1 and
   Q is equal to 1 implies gamma is equal to 1.
2. Let A beta into alpha B divided by (AB) into (alpha beta) is equal to k so that,
   Gamma is equal to 1 minus square root of k divided by 1 plus square root of k
   Implies gamma square is equal to 1 plus k minus 2 into square root of k whole divided by 1 plus k plus 2 into square root of k
   Therefore 1 plus gamma square is equal to 2 into 1 plus k divided by 1 plus k plus 2 into square root of k
   Is equal to 2 into 1 plus k divided by 1 plus square root of k square.

Therefore, 2 into gamma divided by 1 plus gamma square is equal to 2 into 1 minus square root of k into 1 plus square root k whole divided by 2 into 1 plus k
Is equal to 1 minus k divided by 1 plus k.
Now for substituting k, we get
1 minus (A beta) into (alpha B) divided by (AB) into (alpha beta) whole divided by 1 plus (A beta) into (alpha B) divided by (AB) into (alpha beta)

Is equal to (AB) into (alpha beta) minus (A beta) into (alpha B) divided by (AB) into (alpha beta) plus (A beta) into (alpha B) which is equal to Q
Hence, Q is equal to 2 into gamma divided by 1 plus gamma square.

5. Calculation of Coefficient of Contingency and Manifold Classification

Before we start with coefficient of contingency, let us discuss the manifold classification.
We have discussed in previous module that classification of data can be either dichotomous or manifold. For example instead of dividing the universe in two parts, tall and not tall, we may divide it in a larger number of parts, very tall, tall, medium sized, short and very short. Here the attribute tall and its counterpart not tall have been further divided into a number of sub-divisions. Similarly, the two classes, heavy and not heavy may be subdivided as very heavy, heavy, normal light and very light.

Thus attribute A can be divided into a number of groups A_1, A_2, A_3 …,A_5 and similarly the attribute B can be sub-divided as B_1, B_2, B_3 …,B_5. It will be observed that each one of the classes A_1, A_2, A_3 etc., of the first attribute would be divided into a number of heads like B_1, B_2, B_3 etc., when a second attribute B is taken into account. Such classification is called Manifold Classification.

Now let us see how many classes we get in the manifold classification.

Figure 1

Attribute	A_1	A_2	A_3	... ...	A_s	**Total**
B_1	$(A_1 B_1)$	$(A_2 B_1)$	$(A_3 B_1)$	... ...	$(A_s B_1)$	(B_1)
B_2	$(A_1 B_2)$	$(A_2 B_2)$	$(A_3 B_2)$	... ...	$(A_s B_2)$	(B_2)
B_3	$(A_1 B_3)$	$(A_2 B_3)$	$(A_3 B_3)$	... ...	$(A_s B_3)$	(B_3)
...						...
...						...
B_t	(A_1B_s)	(A_2B_s)	(A_3B_s)	... ...	(A_sB_s)	(B_t)
Total	(A_1)	(A_2)	(A_3)	... ...	(A_s)	N

Let us consider only two attributes A and B. If attribute A is sub-divided in s classes and attribute B in t classes, we shall have a table of the following type.

The first row indicates the different classes of attribute A and the first column indicates the different classes of attribute B. Observe that there are t rows and s columns. Hence, the number of entries in the table is s into t.

In this table the totals of various columns A_1, A_2 etc., and the totals of various rows B_1, B_2, etc., would give the first order frequencies and the frequencies in various cells would be second order frequencies. The total of either A_1 A_2 etc., or B_1, B_2, etc., would give the grand total N. Such a table is called Contingency Table.

Now let us find the coefficient of contingency.

If A and B are completely independent of each other in the universe at large, then the actual values A_1 B_1, A_2 B_2, etc., must be equal to their expected values which are in turn equal to (A1) into (B1) divided by N and (A2) into (B2) divided by N respectively. In other words, if the observed frequency in each of the cells of a contingency table is equal to the expected frequency of that cell, A and B would be completely independent of each other.

If these values are not equal in all the cells it is an indication of association between the attributes A and B. In order to test the intensity of association, the difference between the actual and expected frequencies of various cells is calculated.
If these values are not equal in all the cells it is an indication of association between the attributes A and B. In order to test the intensity of association, the difference between the actual and expected frequencies of various cells is calculated. With these differences the value of Chi-square is obtained. The value of chi-square is represented by
Chi square is equal to summation differences of actual and expected frequencies square divided by expected frequencies.

If O stands for actual or observed frequency of a class and E for expected frequency the value of chi square would be
Chi square is equal to summation O minus E whole square divided by E
This value is called “Square Contingency” and if the mean of the square contingency is calculated, it is called “Mean Square Contingency”.
Thus, Square contingency is equal to chi square.
Mean Square Contingency
Phi square is equal to chi square divided by N.

Chi square can also be calculated by the following formula.
Chi square is equal to summation O square divided by E minus N
It is obvious that chi square and phi square which are the sums of square cannot have negative values. If, however the actual and expected values are equal in all cases the values of chi square and phi square would be zero. The limits of chi square and phi square vary in different cases and as such they are not suitable for studying the association in contingency tables.

Karl Pearson has given the following formula for the calculation of “coefficient of mean square contingency”. According to it the coefficient of mean square contingency is given by,
C is equal to square root of chi square divided by N plus chi square is equal to square root of phi square divided by 1 plus chi square.
If chi square is calculated by formula,
Chi square is equal to summation o square divided by E minus N and summation O square divided by E is represented by S.

Then the coefficient of mean square contingency is given by,

C is equal to square root of S minus N divided by N plus S minus N is equal to square root of S minus N divided by S
The above coefficient has a drawback and it is that it never reaches the limit of 1. The limit of 1 is reached by it only if the number classes are infinite. Ordinarily its maximum value depends on the values of s and t. That is the number of sub-divisions of the two attributes A and B.

In a t by t contingency table, the maximum value of C is given by square root of t minus 1 divided by t.
But the coefficients calculated from different types of classification are not comparable with each other.

Now let us consider, Tschuprow's Coefficient:
Since the Pearsonian coefficient of mean square contingency does not reach the maximum limit of 1 and since this a drawback, Tschurprow has suggested the coefficient T. It is calculated as follows.
T is equal to square root of C square divided by 1 minus C square into square root of s minus 1 into t minus 1, where C stands for coefficient of attribute A and t stands for sub-divisions of attribute B.

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

- The different measure of studying association of two way classified attributes, five different methods
- Manifold Classification