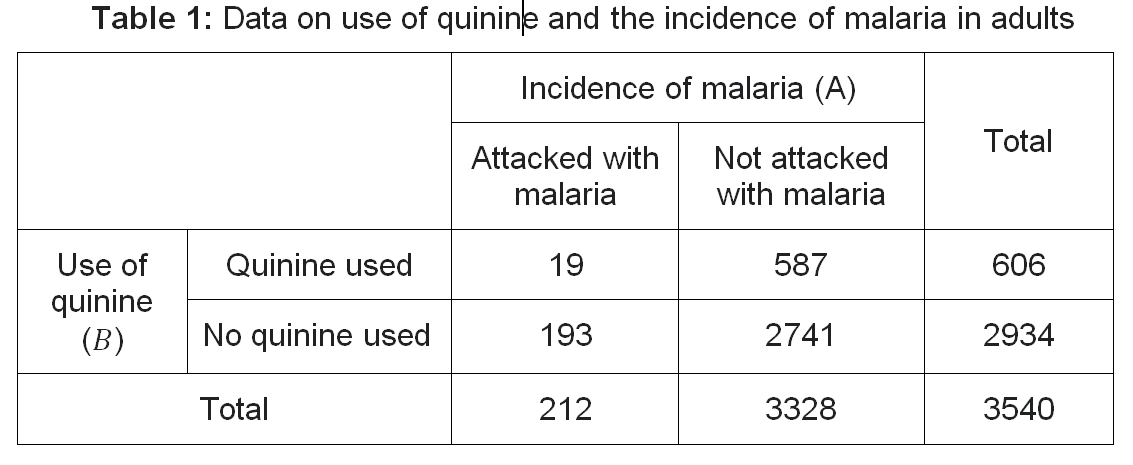
**Joint Distributions of Attributes**

In some investigations, it may be appropriate to collect data on ***two characters on each individual*** or item at the same time. For examples,

* the data may relate to proficiency in English (very good/good/mediocre/bad/very bad) and proficiency in mathematics (very good/good/mediocre/bad/very bad) of a group of students.
* data may relate to economic status (rich/middle-class/poor) and the level of education (illiterate/primary/high-school/university) of a group of adults.
* data may relate to use of quinine (used/not used) and incidence of malaria (attacked with malaria/not attacked with malaria).
* data may relate to height (in cm) and weight (in Kg) of a group of children.
* data may relate to length (in cm) of green jute and weight (in gm) of dry jute fibre.
* data may relate to yield (in kg) of crops and rainfall (in cm) in an area.

It can be noted that the characters may be attributes or variables. The object here would be to look for any relationship that may be obtained among the characters.

Let us consider first the cases where all the characters are attributes. Table 1 below shows the data on use of quinine and incidence of malaria of a state of India in a summary form.



In Table 1, the figure in each cell stands for the number of individuals (i.e. frequency) corresponding to a pair of forms (or levels) of the two attributes. For example, number of adults attacked with malaria among those administered quinine is 19 and number of adults attacked with malaria among those not administered quinine is 193, and so on. Such a table showing joint frequency distribution of the attribute is known as ***Contingency table***.

**Joint (frequency) distribution of the attributes:** The cell-frequencies, together with their grand total, show how the two attributes vary jointly in the given group of individuals. Therefore, the cell-frequencies, together with their grand total, are called as the joint (frequency) distribution of the attributes.

**Marginal frequency distribution:** The row-totals (marginal frequencies), together with the grand total, give marginal distribution of an attribute. Similarly, the column-totals, together with the grand totals, give the marginal distribution of the other attribute.

In the present example, the row-totals along with the grand total, give the marginal frequency distribution of “use of quinine, and the column-totals, along with grand total, give the marginal frequency distribution of “incidence of malaria”.

**Conditional frequency distribution:** The frequencies in a row along with the corresponding row-total, as well as the frequencies in a column along with the corresponding column-total, give the conditional frequency distribution.

In the present example, The frequencies in the first row together with the row-total 606, give the conditional frequency distribution of the attribute “incidence of malaria” for the adults used quinine. The frequencies in the first column together with the column-total 212, in the same way, give the conditional frequency distribution of the attribute “use of quinine” for adults “attacked with malaria”.

It may be noted that in Table 1, we might consider the relative frequencies, instead of the frequencies, which would also give the joint, marginal and conditional distributions of the attributes, although in a different form.

**Notions of Independence and Association**

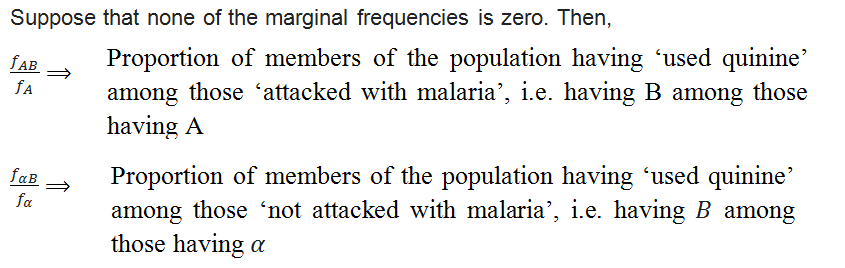
Suppose the individuals under consideration constitute the population itself and not just a sample from the population. Let the two forms of may be denoted by (the positive form, indicating the presence of the character ) and (the negative form, indicating the absence of the character , and similarly, the two forms of may be denoted by and .

The four cell frequencies may be denoted by , , and , and the total by . Also the marginal frequencies for the attribute may be denoted by and , and the marginal frequencies for the attribute by and . Then Quinine-Malaria data can be presented as

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Incidence of malaria (A) | | Total |
| Attacked with malaria () | Not attacked with malaria () |
| Use of quinine () | Quinine used  () |  |  |  |
| No quinine used  () |  |  |  |
| Total | |  |  |  |

Thus,

and



**Independence:** If proportions and be equal, we may say that the presence or absence of the character in an individual does not in any way determine whether will be present. and then be called statistically unrelated or independent.

**Association:** As opposed to the notion of independence, there is the notion of association. Thus and are said to be associated if they are not independent.

**Condition for independence**

For A and B to be independent, we must have

.......................................................................(1)

.....................................................................(2)

*It is the defining equation for the independence of and .*

Actually, equation (1) also implies

........................................................................(3)

........................................................................(4)

.......................................................................(5)

Suppose and are not independent, i.e. are associated. Then two distinguished cases may arise.

This implies that and occur together more frequently than they would have if they had been independent. In this case, the attributes are said to be positively associated.

This implies that and occur together less frequently than they would have if they had been independent. In this case, the attributes are said to be negatively associated.

**Measures of association for classification**

A measure of association should fulfil certain criteria. For example,

* It should be independent of the total frequency , and thus depends on the relative frequencies in the cells rather than on the frequencies.
* It should be zero in the case of independence, negative in case of negative association and positive in the case of positive association.
* It should increase from its lowest possible value through zero to its highest possible value as we proceed from perfect negative association through independence to perfect positive association.
* It should preferably vary between two definite limits, like -1 and +1.

Let, denotes the difference between the actual frequency for the cell (, ) and the value that it should assume if and are independent, i.e.

It may be noted that should be zero if and are truly independent, and so it may serve as the basis for a measure of association. Keeping all the criteria in mind, different measures of association are proposed in literature.

is called as the Yule's ***coefficient of association*** between and .

Any value between –1 to +1 tells us the degree of relationship between two attributes A and B. If = 1, A and B has perfect positive association. If = –1, A and B possess perfect negative association. If = 0, A and B are independent.

Conventionally, if > 0.5 the association between two attributes is considered to be of high order and the value of less than 0.5 shows low degree of association between two attributes.

is called as the ***coefficient of colligation*** and it has the same general properties as .

**Example 1:** Analysis of data on Table 1

Each of the measures indicates only a slight negative association between the two attributes. In other words, there is only slight evidence in support of the belief that use of quinine followed by exemption from attack of malaria.

**Manifold two-way () classification**

Two attributes may again be denoted by and . Let occurs in one of forms: , , ..., and occurs in one of forms: , , ...,

Suppose, of the individuals under study, have the form of together with the form of . Then, is the cell-frequency of the ()th cell or of the combination .

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | | | | | | **Total** |
|  |  |  | **.** | **.** |  |
|  |  |  |  |  | . | . |  |  |
|  |  |  |  | . | . |  |  |
|  |  |  |  | . | . |  |  |
| **.** | . | . | . |  |  |  |  |
| **.** | . | . | . |  |  |  |  |
|  |  |  |  |  |  |  |  |
| **Total** | |  |  |  | . | . |  |  |

Then,

* the marginal frequency of is ()
* the marginal frequency of is ()
* the frequencies (together with ) define the joint distribution of A and B
* the marginal distribution of A is defined by and
* the marginal distribution of B is defined by
* the conditional distributions of for given forms of are represented by the rows of the two-way frequency table,
* the conditional distributions of for given forms of are represented by the columns of the table.

**Independence of the attributes and**

Attributes and are considered to be statistically independent if

for each

or equivalently, if

for all ()

If on the other hand,

for any pair (), then A and B will be said to be associated.

**Measure of association**

For constructing a measure of association the differences between the actual cell-frequencies and the values they should assume if the characters and are independent (i.e. ) may be taken into consideration, where

Then the quantity may serve as a measure of association, where

* The value of is zero if and only if and are independent (i.e. if and only if for all (),
* the higher the strength of association, the higher is the value of .
* depends too much on the total frequency , and theoretically it can be infinite.

A measure that does not suffer from this defect is proposed by Karl Pearson as follows:

is called Pearson’s **coefficient of contingency**.

* equals to zero if and only if for each (i.e. if and only if ).
* The larger the value of , the greater is the degree of association.
* If the number of rows and columns of a contingency table is equal to , the maximum value of is given by .
* The number of rows and columns determines the maximum value of , and it is never greater than one. *It does not attain the value unity even if and are perfectly associated.*

**Proof:**   *does not attain the value unity even if and are perfectly associated.*

Let us consider a table with classes for each of the two attributes, where every diagonal and for every non-diagonal cell. Surely, no greater degree of association than this can be imagined in the case. Yet since for and ,

and so,

To remove the above limitation, Tschuprow suggests an alternative coefficient:

Like , vanishes if and only if and are independent. But unlike , if the attributes and are perfectly associated in a table, then

**Example 2:** Classifications of 830 professional workers according to occupation group and activity status are shown in the following table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Activity status | | | Total |
| Emplo-yees | Emplo-yers | Own-account workers |
| Occupation group | Scientists & technicians | 169 | 21 | 140 | 330 |
| Medical & health services | 83 | 25 | 68 | 176 |
| Teachers | 286 | 10 | 28 | 324 |
|  | Total | 538 | 56 | 236 | 830 |

Find out the degree of association between occupation group and activity status.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 169 | 177540 | 0.16087 |
| 83 | 94688 | 0.07275 |
| 286 | 174312 | 0.46925 |
| 21 | 18480 | 0.02386 |
| 25 | 9856 | 0.06341 |
| 10 | 18144 | 0.00551 |
| 140 | 77880 | 9.25167 |
| 68 | 41536 | 0.11133 |
| 28 | 76464 | 0.01024 |
| Total | - | 1.16890 |

From the above table, we get

and coefficient of contingency,

and

Both and values indicate that there is moderate degree of association between the attributes.

**Exercise 1:** Compute a measure of association for the following data and comment.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Males | Females | Total |
| Tuberculosis of respiratory system | 3534 | 1319 | 4853 |
| Other forms of tuberculosis | 270 | 252 | 522 |
| Total | 3804 | 1571 | 5375 |

[Ans. Q=0.429]

**Exercise 2:** For the following classification, compute a measure of association. Would it be proper to attach a sign to this measure?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Left-eyed | Ambiocular | Right-eyed | Total |
| Left-handed | 48 | 25 | 52 | 125 |
| Ambidextrous | 32 | 13 | 25 | 70 |
| Right-handed | 94 | 33 | 91 | 218 |
| Total | 174 | 71 | 168 | 413 |

[Partial Ans. C=0.076]

**Association and Causal relationship**

There may exist an association between two attributes and because of the following reasons:

1. being a cause of
2. being a cause of
3. Both ad being caused by other characters or group of characters

It may be noted that there is causal relationship (i.e. cause-effect type) between and only in first two cases. Therefore, **an association between two attributes need not imply a causal relationship**.

For instance, consider the case of example 1, where data on use of quinine and the incidence of malaria in adults was analyzed. We observed that taking of quinine as a precautionary measure (B) is negatively associated with attack of malaria (A). However, it may be that economic condition (C) of the people examined has something to do with the observed association.

* It is well known that rich people are more health-conscious than the poor and are more likely to afford the use of quinine as a precautionary measure. Hence B and C are likely to be positively associated.
* Again, the rich live in more hygienic conditions than the poor and hence are less likely to be attacked with malaria. Thus A and C are likely to be negatively associated.
* Thus, the non-zero association between B and A may actually be due to the non-zero association of each of them with C.
* Similarly, the apparent independence of A and B may also be spurious, being due to the effect of a third attribute C on them.

**Spurious Association:** In statistics, a spurious association (or spuriousness) refers to a connection between two attributes that appears to be causal but is not. With spurious association, any observed dependencies between attributes are merely due to chance or are both related to some unseen confounder.

To make sure that the association is not spurious, we should study and together with other characters, say , etc., which are likely to have an influence on the former. That is to say, we should separately measure the association between and for each combination of forms of the others. The reason is that ***an apparent association between and may actually be due to the effect of those other characters on them.***

Therefore, while using a measure of association in looking for a causal relationship between two characters, A and B, we should consider them in conjunction with other characters, C, D etc. that are likely to have an effect on them. Only when A and B are found to be associated for fixed combinations of forms of those characters, it will be proper to say that one of them is a cause of the others.