CHAPTER-7

Categorical Data Analysis

⇨ **Categorical Variable and Categorical Data:** ∽ A variable which takes values on a scale consisting of some categories is called a categorical variable. For example, "Political Idealogy" is a categorical variable which takes values on a scale consisting of several categories, say, 'liberal', 'moderate' & 'conservative'. Any data collected on a categorical variable is called Categorical Data.

Response & Explanatory Variables: Most statistical techniques distinguish between response or dependent variables and explanatory or independent variables. For instance, regression models describe how the mean of response variable such as the selling prices of houses changes according as the values of explanatory variable such as square footage and location.

Nominal & Ordinal Variables: Categorical variables have two primary types of scales. Variables having categories without a natural ordering are called nominal variable. e.g. religious affiliation with the categories catholic, protestant, jewish, Muslims and others. For nominal variables, the orders of listing of the categories is inselevant (not necessary).

Many categorical variables do have ordered categories, such variables are called ordinal variables. Examples are social class: upper, middle, lower; & patient conditions: good, fair, serious.

Attribute: ∽ When we record the sex of each newborn baby during a month or the language of each book in a library, the data are not numbers initially. We get numbers if subsequently, we note the number of male babies and that of female babies, or the numbers of books written in English, the number written in Hindi, the number written in Bengali and so forth. For this type of data, the characters observed is not expressible in numerical terms. Such a character is, therefore, called a qualitative characters or an attribute.

Dichotomy :~ A classification of the simple kind considered, in which each class is divided into two sub-classes and no more, has been termed by logicians classification, or to use the more strictly applicable term, division by dichotomy. The classification of most statistics are not dichotomous, for most usually a class is divided into more than two subclasses, but dichotomy is the fundamental case.

Notations :~ For theoretical purposes, it is necessary to have some simple notations for the classes formed and for the numbers of observation assigned to each. Let us denote the several attributes by A, B, C, ..., etc. An object or an individual possessing the attribute A will be termed simply by A. The class, all the members of which possess the attribute A will be termed as the class-A. The absence of the attributes A, B, C, ...; we shall employ the notations $\alpha, \beta, \gamma, \ldots$ accordingly. Thus, if A represents the attribute blindness; α represents sight. If B stands for deafness, β stands for hearing.

Generally, 'α' is equivalent to 'not A'. The class α is equivalent to the class none of the members of which possesses the attribute A.

In case of combination of attributes, such as, 'AB' represents the combination of blindness and deafness. If a third attribute be noted, e.g. — insanity, denoted by C & the absence of it by γ; then the class 'ABC' includes those who are at once deaf, blind and insane; 'ABγ' includes those who are deaf and blind but not insane.

Class-frequencies :~ The no. of observation assigned to any class is termed as the 'class-frequency'. Class-frequencies will be denoted by putting the class-symbols within parenthesis.

Orders of classes and class-frequency :~

For 2 attributes, the class frequencies of different orders are

Order 0 : N

1st order : (A), (B), (α), (β) [(A) denotes the class frequency of the class A]

2nd order : (AB), (Aβ), (αB), ($\alpha\beta$)

Any class frequency can always be expressed in terms of frequencies of higher orders classes, e.g.

$(A) = (AB) + (A\beta)$

$(AB) = (ABC) + (AB\gamma)$

Ultimate Class frequencies :- A class of highest order is known as an ultimate class and its frequency an ultimate class frequency.

Every class frequency can be written as a sum of certain ultimate class frequencies.

For any frequency can be analysed into higher frequencies and the process needs stop only when we have reached the frequency of the highest order, e.g. with 3 attributes,

$$(A) = (AB) + (A\beta)$$
$$= (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$$

The total no. of class-frequencies :—

Order zero : N

Order one : (A) (B) (C)
(α) (β) (γ)

Order two : (AB) (AC) (BC)
$(A\beta)$ $(A\gamma)$ $(B\gamma)$
(αB) (αC) (βC)
$(\alpha\beta)$ $(\alpha\gamma)$ $(\beta\gamma)$

Order three : (ABC) (αBC)
$(AB\gamma)$ $(\alpha B\gamma)$
$(A\beta C)$ $(\alpha\beta C)$
$(A\beta\gamma)$ $(\alpha\beta\gamma)$

Thus, here we have 3^3 distinct class-frequencies in all. In general, for n attributes there are 3^n distinct class-frequencies, provided we count N as a frequency of order zero.

Of order zero, there is a single class N; of order one, there are $2n$ classes; of order two, there are $\binom{n}{2}2^2$ classes; hence of order r, there are $\binom{n}{r}2^r$ classes.

$\therefore$ No of class frequencies $= \sum_{r=0}^{n}\binom{n}{r}2^r = (1+2)^n = 3^n$

Consistency :→ Any class frequencies which have been or might have been observed within one, and the same population may be said to be consistent with one another. They conform with one another, and don't, in any way, conflict.

Symbols :> We define, $A.N = (A)$, for any attribute.

similarly, $\alpha.N = (\alpha)$

Now, $(A+\alpha)N = (A)+(\alpha) = N$

$\Rightarrow (A+\alpha) = 1$

$\Rightarrow \alpha = 1-A$

Therefore, in any symbolic expression we can replace the operators A by $1-\alpha$ or α by $1-A$.

Again, $(AB) = A.(B) = B.(A)$

A little reflection will show that the operative symbols therefore obey the laws of algebra.

$$\therefore (\alpha\beta) = \alpha\beta.N$$
$$= (1-A).(1-B).N$$
$$= (1-B-A+AB).N$$
$$= N-(A)-(B)+(AB)$$

similarly, $(\alpha\beta\gamma) = \alpha\beta\gamma.N$

$$= (1-A)(1-B)(1-C).N$$
$$= (1-A-B-C+AB+BC+AC-ABC).N$$
$$= N-(A)-(B)-(C)+(AB)+(BC)+(AC)-(ABC)$$

Condition for Consistency :— The attributes denoted by capitals may be termed positive attributes, on the other hand, attributes denoted by Greek letters, may be termed as negative attributes.

The necessary and sufficient condition for the consistency of a set of independent class-frequencies is that no ultimate class-frequency be negative.

If two attributes are noted, there are 4 ultimate frequencies — (AB), $(A\beta)$, (αB), $(\alpha\beta)$. Expressing them in terms of positive classes, we find the following conditions —

i) $(AB) \geq 0$

ii) $(AB) \geq (A)+(B)-N$ [i.e. $(\alpha\beta) \geq 0$]

iii) $(AB) \leq (A) \Leftrightarrow (A\beta) \geq 0$

iv) $(AB) \leq (B) \Leftrightarrow (\alpha B) \geq 0$

For 3 attributes the conditions that the 8 ultimate frequencies are not negative will be found to lead to the following —

i) $(ABC) \geq 0$
ii) $(ABC) \geq (AB) + (AC) - (A)$
iii) $(ABC) \geq (AB) + (AC) - (B)$
iv) $(ABC) \geq (AB) + (AC) - (C)$
v) $(ABC) \leq (AB)$
vi) $(ABC) \leq (AC)$
vii) $(ABC) \leq (BC)$
viii) $(ABC) \leq N - (A) - (B) - (C) + (AB) + (BC) + (AC)$

Example:→ Show that neither x nor y can exceed $1/4$ when the followings are given:

$\frac{(A)}{N} = x$, $\frac{(B)}{N} = 2x$, $\frac{(C)}{N} = 3x$ & $\frac{(AB)}{N} = \frac{(BC)}{N} = \frac{(CA)}{N} = y$.

Soln.→ From condition of consistency,

$(AB) \leq (A)$

$\Rightarrow y \leq x$ ——— (*)

$(BC) \geq (B) + (C) - N \Rightarrow \frac{(BC)}{N} \geq \frac{(B)}{N} + \frac{(C)}{N} - 1$

$\Rightarrow y \geq 2x + 3x - 1$

$\Rightarrow y \geq 5x - 1$ [Using (*)]

$\Rightarrow x \geq 5x - 1$

$\Rightarrow x \leq 1/4$

$\Rightarrow y \leq x \leq 1/4$.

INDEPENDENCE:

Attribute	B	β	TOTAL
A	(AB)	$(A\beta)$	(A)
α	(αB)	$(\alpha\beta)$	(α)
TOTAL	(B)	(β)	N

Two related attributes A & B are said to be independent if

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$$

$$\Leftrightarrow \frac{(AB)}{(B)} = \frac{(AB) + (A\beta)}{(B) + (\beta)} = \frac{(A)}{N}$$

$$\Leftrightarrow \frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N} \quad \text{——— (**)}$$

"If the attributes A and B are independent, the proportion of 'AB's in the population is equal to the proportion of A's multiplied by the proportion of B's." → fundamental rule.

If there is no sort of relationship of any kind between two attributes A and B, we expect to find the same proportion of A's amongst the B's as amongst the not B's.

We have earlier the criterion of independence for A & B –

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \qquad \text{——— (*)}$$

If this relation holds good, then the following equations must also hold ——

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$$

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$$

$$\frac{(A\beta)}{(A)} = \frac{(\alpha B)}{(\alpha)}$$

(*) ⟶ gives ⟶ $\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$

$$\Leftrightarrow \frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$$

i.e. $\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$, etc.

NOTE: If 2 attributes A & B are independent, then i) α, β are also independent, ii) A, β are also independent.

i) For, $(\alpha\beta) = N - (A) - (B) + (AB)$

$$= N - (A) - (B) + \frac{(A)(B)}{N}$$

$$= N - (A) - \frac{(B)}{N}(N - (A))$$

$$= \frac{(N-(A))(N-(B))}{N}$$

$$= \frac{(\alpha)(\beta)}{N}$$

ii) $\frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$

$$\Rightarrow (A\beta) = \frac{(A)(\beta)}{N}$$

Interpretation: ⇒ The advantages of (**) over (*) is that it gives expressions for the second order frequency in terms of the frequencies of the 1st order and the total no. of observations alone.

The criteria of independence may be expressed in yet a 3rd form, i.e. in terms of 2nd order frequencies alone.

If A and B are independent then

$$(AB)\cdot(\alpha\beta) = \frac{(A)(B)}{N}\cdot\frac{(\alpha)(\beta)}{N} = \frac{(A)(B)(\alpha)(\beta)}{N^2}$$

$$\Rightarrow (AB)\cdot(\alpha\beta) = (\alpha B)(A\beta)$$

$$\Leftrightarrow \frac{(AB)}{(\alpha B)} = \frac{(A\beta)}{(\alpha\beta)} \quad \text{——— (1)}$$

$$\text{or,} \quad \frac{(AB)}{(A\ \beta)} = \frac{(\alpha B)}{(\alpha\beta)} \quad \text{——— (2)}$$

(1) may be read : "The ratio of A's to α's amongst the B's is equal to the ratio of A's to α's amongst the β's.

(2) can be interpreted in a similar way.

Example :→ If the second-order frequencies have the following values, are A and B independent or not ?

$(AB) = 110,\ (\alpha B) = 90,\ (A\beta) = 290,\ (\alpha\beta) = 510$

Soln.→ Clearly, $(AB)(\alpha\beta) = 110 \times 510$

$(A\beta)(\alpha B) = 290 \times 90$

So, $(AB)(\alpha\beta) > (\alpha B)(A\beta)$

So, A and B are not independent.

ASSOCIATION :→ A and B are said to be positively associated (or, simply associated) if

$$(AB) > \frac{(A)\cdot(B)}{N}$$

A and B are said to be negatively associated (or, simply disassociated) if

$$(AB) < \frac{(A)\cdot(B)}{N}$$

In common language, one speaks of A and B as being 'associated' if they appear together in a number of cases. But here, by positive association between A and B, we mean that they appear together in a greater no. of cases than is to be expected if they are independent.

Complete Association or Disassociation :—

The circumstances in which the association between two attributes is said to be complete, can be explained into two cases ——

i) We may say that for complete association, all A's must be B's and all B's must be A's, i.e. $[(A\beta)=0]$ & $[(\alpha B)=0]$ otherwise. Therefore, we see that A's and B's occur in the population in equal numbers.

ii) We may adopt a rather wider meaning and say that all A's are B's or all B's are A's, according to whether the A's or the B's are in the minority.

Similarly, complete disassociation may be taken either as the case when no A's are B's and no α's are β's, or more widely as the case when either of these statements is true.

Thus two attributes are completely associated if one of them can't occur without the other, though the other may occur without the one.

The symbols $(AB)_0$ and δ := We define $(AB)_0 = \frac{(A)(B)}{N}$, the value of (AB) under the assumption that the attributes are independent.

We shall use the other symbols

$$(\alpha\beta)_0 = \frac{(\alpha)(\beta)}{N}, \quad (\alpha B)_0 = \frac{(\alpha)(B)}{N}, \quad (A\beta)_0 = \frac{(A)(\beta)}{N}$$

If δ denote the excess of (AB) over $(AB)_0$, then keeping the marginal totals fixed, the table reduces to the form—

Attribute	B	β	TOTAL
A	$(AB)_0 + \delta$	$(A\beta)_0 - \delta$	(A)
α	$(\alpha B)_0 - \delta$	$(\alpha\beta) + \delta$	(α)
TOTAL	(B)	(β)	N

Define, $\delta = (AB) - (AB)_0$ and also, quite generally we have—

$$\delta = (AB) - (AB)_0 = (\alpha\beta) - (\alpha\beta)_0 = (A\beta)_0 - (A\beta) = (\alpha B)_0 - (\alpha B)$$

Note that, $\delta = (AB) - (AB)_0 = (AB) - \frac{(A)(B)}{N}$

$$\delta = \frac{N(AB) - (A)(B)}{N}$$

$$\Rightarrow \delta = \frac{1}{N}\left[N(AB) - \{(AB)+(A\beta)\}\{(AB)+(\alpha B)\}\right]$$

$$= \frac{1}{N}\left[\{(AB)+(A\beta)+(\alpha B)+(\alpha\beta)\}(AB) - \{(AB)+(A\beta)\}\{(AB)+(\alpha B)\}\right]$$

$$= \frac{1}{N}\left[(AB)(\alpha\beta) - (A\beta)(\alpha B)\right]$$

Clearly, $\delta = 0$, $\delta > 0 \Leftrightarrow$ (+ve) association, $\delta < 0 \Leftrightarrow$ (-ve) association

$\Leftrightarrow$ A and B are independent.

'δ' determines uniquely the departure from independence. But it may be of interest to measure the intensity of association. It is called 'δ measure' for association between 2 attributes.

Coefficient of association :

$$\delta = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{N}$$

1) Yule's coefficient of association : As a measure of intensity of association between 2 attributes A and B, Yule gave the following coefficient of association:

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

If A and B are independent, $\delta = 0 \Leftrightarrow Q = 0$.

* 2)

	A	α	
B	$k(AB)$	αB	
β	$k(A\beta)$	$\alpha\beta$	
	$k(A)$	(α)	

PROPERTIES :

1) $Q = \frac{N\delta}{(AB)(\alpha\beta) + (\alpha B + A\beta)}$ increases as δ increases.

* 2) If all the term containing A are multiplied by a constant, the value of Q remains unaltered. Similar things hold for α, β and B.

Hence, Q is independent of A's and α's in the data.

∴ Relative proportion of A's and α's are $\frac{k(A)}{k(A)+\alpha}$, $\frac{(\alpha)}{k(A)+(\alpha)}$

Now, $Q^* = \frac{k(AB)(\alpha\beta) - k(A\beta)(\alpha B)}{k(AB)(\alpha\beta) + k(A\beta)(\alpha B)} = Q$

Hence, Q is independent of the relative proportion of A's.

N.P. Q is a symmetric measure (coefficient) and it increases from -1 to 1 as the extent of association increases from perfect -ve to perfect +ve. Shown in property 3).

3) $-1 \le Q \le 1$

Proof:→ $Q = \dfrac{(AB)(\alpha\beta) - (\alpha B)(A\beta)}{(AB)(\alpha\beta) + (\alpha B)(A\beta)} = \dfrac{a-b}{a+b}$,

where $a = (AB)(\alpha\beta) \ge 0$ & $b = (\alpha B)(A\beta) \ge 0$

Note that, $Q = 1 - \dfrac{2b}{a+b} \le 1$, —①

and, $Q = -1 + \dfrac{2a}{a+b} \ge -1$ —②

Combining ① & ② ⇒ $-1 \le Q \le 1$.

Marginal cases :→ $Q = 0 \Leftrightarrow b = 0$ iff A & B are independent.

$Q = 1$ iff $b = 0$

$\Leftrightarrow (\alpha B)(A\beta) = 0$

$\Leftrightarrow (\alpha B) = 0$, or, $(A\beta) = 0$

$\Leftrightarrow (AB) = (B)$, or, $(AB) = (A)$

$\Leftrightarrow$ A & B are in complete association.

$Q = -1$ iff A & B are in complete disassociation.

2) Yule's coefficient of colligation :→

Define, $Y = \dfrac{1 - \sqrt{(A\beta)(\alpha B)/(AB)(\alpha\beta)}}{1 + \sqrt{(A\beta)(\alpha B)/(AB)(\alpha\beta)}}$

$= \dfrac{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)} + \sqrt{(A\beta)(\alpha B)}}$

Show that → $Q = \dfrac{2Y}{1+Y^2}$

Proof:→ $Y = \dfrac{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)} + \sqrt{(A\beta)(\alpha B)}}$

$\Rightarrow \dfrac{1+Y}{1-Y} = \dfrac{\sqrt{(AB)(\alpha\beta)}}{\sqrt{(A\beta)(\alpha B)}}$

$\Rightarrow \dfrac{(1+Y)^2}{(1-Y)^2} = \dfrac{(AB)(\alpha\beta)}{(A\beta)(\alpha B)}$

$\Rightarrow \dfrac{(1+Y)^2 - (1-Y)^2}{(1+Y)^2 + (1-Y)^2} = \dfrac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$

$\Rightarrow \dfrac{4Y}{2(1+Y^2)} = Q$

$\Rightarrow Q = \dfrac{2Y}{1+Y^2}$.

N.P. Y has the same properties as Q, i.e. $-1 \le Y \le 1$.
=, holds, i.e. (+1) for complete association & (-1) for complete disassociation. If A & B are independent, then Y vanishes.

Manifold Classification :→

In stead of dividing the population under consideration into two parts by a simple dichotomy, we may also divide it into a number of parts by a simple process. For instance, we can extend the dichotomy of the population of men into "those with blue eyes" and "those not with blue eyes" to a threefold division: "those with blue eyes", "those with brown eyes" and "those with neither blue nor brown eyes"; or into a fourfold division by adding a fresh category, "those with grey eyes";

Generally, our population may be divided first according to s heads, $A_1, A_2, \ldots, A_s$; each of the classes so obtained into t heads, $B_1, B_2, \ldots, B_t$; each of these into u heads, $C_1, C_2, \ldots, C_u$; and so on.

This is called "manifold classification."

The theory of manifold classification involving n attributes is rather complicated, but its fundamental principles are very similar to dichotomy; A straight forward extension of the methods already discussed will give the following results:

(a) There are $s \times t \times u \times \ldots$ ultimate classes.
(b) The total no. of classes, including N and the ultimate classes, is $(s+1)(t+1)(u+1)\ldots$
(c) The data are consistent if, and only if, every ultimate class-frequency is not negative.
(d) The data are completely specified by $s \times t \times u \times \ldots$ algebrically independent class-frequencies. Even if all these are not given, it may be possible to set limits to the other class-frequencies.

Contingency Table :—

Let us consider a classification in which the attribute A is s-fold and B is t-fold. Clearly, here we have 'st' classes of the type $A_i B_j$; $i = 1(1)s$, $j = 1(1)t$.

Then we can arrange our data in a table as follows —

B \ A	A_1	A_2	A_m	A_s	Totals
B_1	(A_1B_1)	(A_2B_1)	(A_mB_1)	(A_sB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	(A_mB_2)	(A_sB_2)	(B_2)
⋮	⋮	⋮	⋮	⋮	⋮
B_n	(A_1B_n)	(A_2B_n)	(A_mB_n)	(A_sB_n)	(B_n)
⋮	⋮	⋮	⋮	⋮	⋮
B_t	(A_1B_t)	(A_2B_t)	(A_mB_t)	(A_sB_t)	(B_t)
Totals	(A_1)	(A_2)	(A_m)	(A_s)	N

Such a table is called a Contingency Table.

Alternative Notations:

$$n_{ij} = (A_iB_j)$$
$$n_{i0} = (A_i)$$
$$n_{0j} = (B_j)$$

Coefficients of contingency :~ The characteristics A & B are said to be completely independent in the population at large, if for all values of m and n, we must have —

$$(A_mB_n) = \frac{(A_m)(B_n)}{N} = (A_mB_n)_0 \text{ , i.e. } n_{ij} = \frac{n_{i0}\cdot n_{0j}}{N}$$

If, however, A and B are not completely independent, (A_mB_n) and $(A_mB_n)_0$ will not be identical for all the values of m and n. The deviation from independence in that particular cell will be measured by,

$$\delta_{mn} = (A_mB_n) - (A_mB_n)_0$$
$$= (A_mB_n) - \frac{(A_m)(B_n)}{N}$$
$$\text{i.e. } \delta_{ij} = n_{ij} - \frac{n_{i0}\cdot n_{0j}}{N}$$

Properties :→

(1) In General $\delta_{mn} \neq \delta_{nm}$.

(2) <u>The δ's are not algebraically independent:</u>

We have, in fact, for any particular m,

$$\delta_{m1} + \delta_{m2} + \cdots + \delta_{mn} + \cdots + \delta_{mt}$$

$$= \sum_{j=1}^{t} \delta_{mj}$$

$$= (AmB_1) - \frac{(Am)(B_1)}{N} + (AmB_2) - \frac{(Am)(B_2)}{N} + \cdots + (AmB_t) - \frac{(Am)(B_t)}{N}$$

$$= (Am) - \frac{(Am)}{N}\{(B_1) + (B_2) + \cdots + (B_t)\}$$

$$= (Am) - \frac{(Am)}{N} \times N$$

$$= 0.$$

OR

$$\left[\sum_{j=1}^{t} \delta_{mj} = \sum_{j=1}^{t} \left\{ (AmB_j) - \frac{(Am)(B_j)}{N} \right\} \right.$$

$$= \sum_{j=1}^{t} (AmB_j) - \frac{(Am)}{N} \sum_{j=1}^{t} (B_j)$$

$$= (Am) \div \frac{(Am)}{N} \cdot N$$

$$\left. = 0 \right]$$

$\therefore$ For a particular n, $\sum_{i=1}^{s} \delta_{in} = 0$.

<u>Note</u>, there are st δ-quantities. In virtue of the relationship we have just proved, for any particular m only $(t-1)$ of the t quantities δ_{mn} are independent. Similarly, for any n only $(s-1)$ are independent. Hence, the total number of independent δ's is $(s-1)(t-1)$.

SOME MEASURES OF ASSOCIATION :→ (C.U.)

(1) **Mean Square Contingency** :— We may define a measure of association in terms of so called 'square contingency'

$$\chi^2 = \sum_{m=1}^{s}\sum_{n=1}^{t}\left(\frac{\delta^2_{mn}}{(AmBn)_0}\right)$$

and call χ^2 the "square contingency".

We then define,

$$\phi^2 = \frac{\chi^2}{N}$$

and call ϕ^2 the "mean-square contingency".

Clearly, χ^2 and ϕ^2, being the sums of squares, can't be negative. Then vanish if, and only if, every δ-number vanishes, in which case A and B are independent. & ϕ^2 is not suitable for Comparative study.

$$\chi^2 = \sum_{m=1}^{s}\sum_{n=1}^{t}\frac{\delta^2_{mn}}{(AmBn)_0}$$

$$= N\left\{\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{(AmBn)^2}{(Am)(Bn)} - 1\right\}$$

$$= N\left\{\sum_{m=1}^{t}\frac{(AmBm)^2}{(Am)(Bm)} - 1\right\} \quad [\text{where } s=t]$$

$$= N(t-1)$$

$\Leftrightarrow \phi^2 = t-1$ .. Hence the limits of ϕ^2 vary in different systems.

~~Karl Pearson's coefficient of mean-square contingency~~

Properties :— (a) ϕ^2 is non-negative since it is the sum of squares and ϕ^2 can take any non-negative any non-negative value, i.e. $0 \le \phi^2 \le \infty$.

(b) $\phi^2 = 0$ iff $\delta_{mn} = 0 \ \forall\ m,n$.
iff A and B are independent.

Limitations :— (a) Since χ^2 can take any non-negative value then the limits of ϕ^2 vary in different situations. Hence ϕ^2 is not suitable to form a coefficient of association.

(b) In $t \times t$ table, in case of complete association so that $(Am) = (Bm) = (AmBm)\ \forall\ m$ & $(AmBn) = 0\ \forall\ m \neq n$, i.e. when only the leading diagonal frequencies are non-zero.

Why? $\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{\delta^2_{mn}}{(A_mB_n)_0}$

$$= \sum_{m=1}^{s}\sum_{n=1}^{t}\frac{N}{(A_m)(B_n)}\left[(A_mB_n) - \frac{(A_m)(B_n)}{N}\right]^2$$

$$= N\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{\left\{(A_mB_n)^2 - \frac{2(A_mB_n)(A_m)(B_n)}{N} + \frac{(A_m)^2(B_n)^2}{N^2}\right\}}{(A_m)(B_n)}$$

$$= N\sum_{m=1}^{s}\sum_{n=1}^{t}\left\{(A_mB_n)^2 + \frac{(A_m)(B_n)}{N^2}\left((A_m)(B_n) - 2(A_mB_n)N\right)\right\}\Big/(A_m)(B_n)$$

$$= N\left[\sum_{m=1}^{s}\sum_{n=1}^{t}\left\{\frac{(A_mB_n)^2}{(A_m)(B_n)}\right\} + \frac{1}{N^2}\sum_{m=1}^{s}\sum_{n=1}^{t}\left\{(A_m)(B_n) - 2(A_mB_n)N\right\}\right]$$

$$= N\left[\sum_{m=1}^{s}\sum_{n=1}^{t}\frac{(A_mB_n)^2}{(A_m)(B_n)} - 1\right]$$

P.C.U.

(2) Karl Pearson's coefficient of mean-square contingency:

The quantity ϕ^2 is not quite suitable in itself to form a coefficient, because its limits vary in different cases. The upper-limit is infinite as N increases. Pearson, therefore, proposed the coefficient C, defined by,

$$C = \sqrt{\frac{\chi^2}{N+\chi^2}} = \sqrt{\frac{\phi^2}{1+\phi^2}}.$$

This is called the "coefficient of mean-square contingency". In general, no sign should be attached to the root, for the coefficient merely shows whether two characters are or are not independent; but in certain cases a conventional sign may be used. As for example — slight pigmentation of eyes and hair appear to go together, and the contingency may be regarded as positive. If slight pigmentation of eyes had been associated with marked pigmentation of hair, the contingency might have been regarded as negative.

(2.) Properties :—

(1) C vanishes iff $\delta_{mn} = 0 \ \forall \ (m, n)$
i.e. $C = 0$ iff A and B are independent of each other.

(2) The coefficient has one serious disadvantages.

(c.v.) $\therefore \ 0 \leq C < 1$ since $0 \leq \chi^2 < N + \chi^2$

Limitations :—

(1) The coefficient C has one serious drawback that it never reaches the value 1 even in the case of perfect association.

(2) In t x t table, in case of complete association,

$$\chi^2 = N(t-1) \text{ and}$$

$$C = \sqrt{\frac{N(t-1)}{N + N(t-1)}} = \sqrt{\frac{t-1}{t}}$$

Note that, no greater association than the case

$$(A_m B_m) = (A_m) = (B_m) \ \forall \ m, \text{ and}$$

$$(A_m B_n) = 0 \ \forall \ m \neq n$$

can be imagined in t x t table. Hence, $\chi^2 \leq N(t-1)$ and $C \leq \sqrt{\frac{t-1}{t}}$. The upper limits of C vary for different systems and hence it is not suitable for comparative study.

(c.v.)

(3) Tschuprow's Coefficient :— Even in case of complete association the value of C is not unity and its maximum depends on the number of rows and columns of the table. To remedy the defect to which we have reffered above, Tschuprow proposed an alternative coefficient T; defined by,

$$T^2 = \left\{ \frac{\chi^2}{N\sqrt{(s-1)(t-1)}} \right\}$$

$$= \frac{\phi^2}{\sqrt{(s-1)(t-1)}}$$

$T = 1$, in case of complete association in a t x t table but it will not be true in case of an s x t classification where $s \neq t$.

The coefficient varies between 0 and 1.

(4) [redacted] Cramer's V^2 :

In an $s \times t$ table, if $s > t$ and $(A_m B_m) = (A_m)(B_m)$, $m = 1(1)t$; then $\chi^2 = N(t-1)$ but if $s < t$ & $(A_m B_m) = (A_m)(B_m)$, $m = 1(1)t$; then $\chi^2 = N(s-1)$.

Hence for any $s \times t$ table, $\chi^2 \leq N \min\{s-1, t-1\}$.

As Cramer pointed out, we may avoid these difficulties by defining a new measure:

$$V^2 = \frac{\chi^2}{N \min\{s-1, t-1\}}$$

Clearly V^2 lies between 0 and 1.

When, $s = t = 2$, $V^2 = \frac{\chi^2}{N} = \phi^2$

Evidently, $V^2 = T^2$ when the table is square, otherwise V^2 exceeds T^2. Although the difference won't be very large unless s and t varies widely.

We also see that —

$$\frac{C^2}{T^2} = \frac{\sqrt{(s-1)(t-1)}}{1+\phi^2}$$

Remark :→ These measures of association, discussed above, are applicable to both nominal and ordinal data.

ORDINAL TRENDS :— Let us consider two jointly distributed attributes A and B. Further assume that the level of each can be arranged according to their degree of possession. Hence it may happen that — as the responses on the attribute A increase towards its higher level, responses on B also increase towards its higher level or the responses on B decrease towards its lower level. Such type of association is often reffered to as a monotone trend association.

Concordance & Discordance :—

A pair is said to be concordant pair if the subject ranked higher on A also ranked higher on B. A pair is said to be discordant pair if the subject ranking higher on A, ranks lower on B. A pair is said to be tied if the subjects have same classification on A and/or on B.

Ordinal measures of association :—

Here we shall consider some measures based on the no. of concordance and the no. of discordance pairs.

(1) **Goodman-Kruskal Gamma :—** Let C & D be respectively the no. of concordance pairs and the no. of discordance pairs. If the pairs are united on both variables, then $\frac{C}{C+D}$ is the proportion of concordance pairs and $\frac{D}{C+D}$ is the proportion of discordance pairs. The difference $\left(\frac{C}{C+D} - \frac{D}{C+D}\right)$ $= \left(\frac{C-D}{C+D}\right)$ is termed as Goodman-Kruskal Gamma (γ).

i.e. $\gamma = \frac{C-D}{C+D}$.

If $\gamma \gtrless 0$, i.e. $\frac{C}{C+D} \gtrless \frac{D}{C+D}$, then the majority of the pairs are in concordance (or discordance), i.e. characters A and B have the +ve (or, -ve) association.

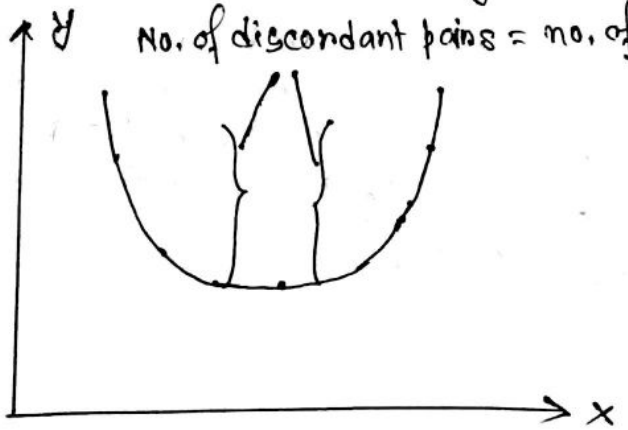
Properties :—

(a) The Goodman-Kruskal Gamma (γ) treats the characters symmetrically — any one of them can be taken as response variable.

(b) Clearly, $|\gamma| \leq 1$, i.e. $-1 \leq \gamma \leq 1$, since $C \geq 0$ and $D \geq 0$.

(c) Under independence γ vanishes but the converse is not true. i.e. Independence $\Rightarrow \gamma = 0$ but the converse is not necessarily true. For a U-shaped joint-distn., one can have $C = D$, i.e. $\gamma = 0$ and the characters are not independent (pic:(c))

(d) $\gamma = \pm 1$ iff the characters are monotonically related.
$\gamma = 1 \Leftrightarrow D = 0$ iff the characters have monotonic increasing relation.
$\gamma = -1 \Leftrightarrow C = 0$ iff the characters have monotonic decreasing relation.



pic : (c)

REMARK: → Note that,

$$C = 2\sum_i\sum_j n_{ij}\left(\sum_{r>i}\sum_{k>j} n_{rk}\right)$$

$$D = 2\sum_i\sum_j n_{ij}\left(\sum_{r>i}\sum_{k<j} n_{rk}\right)$$

Let, T_A = No of tied pairs on the characters A

$$= \binom{n_{10}}{2} + \binom{n_{20}}{2} + \cdots + \binom{n_{s0}}{2}$$

$$= \sum_{i=1}^{s} \frac{n_{i0}(n_{i0}-1)}{2}$$, since n_{i0} members have the same character A.

Similarly, T_B = No. of tied pairs on the characters B.

$$= \binom{n_{01}}{2} + \binom{n_{02}}{2} + \binom{n_{03}}{2} + \cdots + \binom{n_{0t}}{2}$$

$$= \sum_{j=1}^{t} \frac{n_{0j}(n_{0j}-1)}{2}$$

Now, T_{AB} = No. of tied pairs on both A and B

$$= \sum_{i=1}^{s}\sum_{j=1}^{t} \binom{n_{ij}}{2}$$

$$= \sum_i\sum_j \frac{n_{ij}(n_{ij}-1)}{2}$$

Note that → $\binom{N}{2} = C + D + T_A + T_B - T_{AB}$

(2) Kendall's τ : → Let (x_i, y_i) and (x_j, y_j) be two pairs of observation on two characters for a pair (i,j), $i<j$ of individuals.

Define, $a_{ij} = \text{sign}(x_i - x_j)$

$$= \begin{cases} 1, & \text{if } x_i > x_j \\ 0, & \text{if } x_i = x_j \\ -1, & \text{if } x_i < x_j \end{cases}$$

and, $b_{ij} = \text{sign}(y_i - y_j)$

Then, our measure of rank correlation will be based on the sum, $S = \sum_{i<j} a_{ij} b_{ij}$

$= C - D$, which is the total score.

Here, we have two choices to make a measure of association, if we wish to standardize S to lie in the range $[-1, 1]$ and attain ± 1 in the extreme cases of complete disassociation and complete association.

(a) Kendall's τ_a :⟶ If there are no ties, no a_{ij} or b_{ij} are zero. In that case, S would vary in between $-\binom{N}{2}$ & $\binom{N}{2}$.

i.e. $-\binom{N}{2} \le S \le \binom{N}{2}$. Hence, we define,

Kendall's $\tau_a = \dfrac{S}{\binom{N}{2}}$, as a coefficient of association.

Clearly, whether τ_a attains the end points or not, completely depends on the no. of zero scores.

If some scores a_{ij} or b_{ij} are zero, the τ_a can no longer attain ± 1.

[$S = \sum\sum_{i<j} a_{ij} b_{ij} = C - D$

If there are no ties, then $S = \pm\binom{N}{2}$ if all pairs are concordant or discordant.

If there are some ties, then —

$-\binom{N}{2} < S < \binom{N}{2}$ &

consequently τ_a can't attain ± 1.]

Note that, $\boxed{\tau_a = \dfrac{S}{\binom{N}{2}} = \dfrac{C-D}{\binom{N}{2}}}$, which is same as the kendall's τ_a.

(b) Kendall's τ_b :⟶ The correlation coefficient between two set of scores (a_{ij}) and (b_{ij}) is defined as —

$$\tau_b = \frac{\sum\sum_{i<j} a_{ij} b_{ij}}{\sqrt{\sum\sum_{i<j} a_{ij}^2}\sqrt{\sum\sum_{i<j} b_{ij}^2}}$$

In the presence of tied pairs, the measure τ_b will be of the form, $\tau_b = \dfrac{C-D}{\sqrt{\binom{N}{2} - T_A}\sqrt{\binom{N}{2} - T_B}}$

[Now, $\sum\sum_{i<j} a_{ij}^2$ = No. of pairs (i,j), $i<j$ for which $\{x_i > x_j\}$ or $\{x_i < x_j\}$

$= \binom{N}{2}$ − No. of tied pairs on A

$= \binom{N}{2} - T_A$, where $T_A = \sum_{i=1}^{s} \dfrac{n_{i0}(n_{i0}-1)}{2}$

& Similarly, $\sum\sum_{i<j} b_{ij}^2 = \binom{N}{2} - T_B$.]

If there are no ties, then $\tau_a = \tau_b$, but if there are some ties then $\tau_b > \tau_a$.

Remark:→ The Goodman-Kruskal γ can also be expressed as —

$$\gamma = \frac{C-D}{C+D} = \frac{C-D}{\binom{N}{2} - T_A - T_B + T_{AB}}$$

to note

It is important that γ, τ_a, τ_b all attain ± 1 if all the observations are in the cells along the longest diagonal of the table.

(3) **Somer's d :⟶** Somer proposed a measure of association for ordinal data, define,

$$d = \frac{C-D}{\binom{N}{2} - T_A} \text{ or } \frac{C-D}{\binom{N}{2} - T_B} \rightarrow \text{as a coefficient of association.}$$

In case of complete association, $(A_m B_m) = (A_m) = (B_m)$, $m = 1(1)\min(s,t)$, and all other cell frequencies are zero, i.e. all the observations lie in a longest diagonal of a table. Then In that case,

$$S = \sum_{i<j}\sum a_{ij}b_{ij} = \left\{\text{No. of pairs } (i,j), i<j \text{ for which } \begin{matrix} x_i < x_j \\ \text{or} > \end{matrix} (\text{or } \begin{matrix} y_i \leq y_j \\ \text{or} > \end{matrix})\right\}$$

$$= \{\text{No of pairs united on A (or B)}\}.$$

$$= \left\{\binom{N}{2} - T_A\right\} \text{ or } \left\{\binom{N}{2} - T_B\right\}.$$

Hence, $d = 1$.

But if all the observations lie on the longest diagonal of a table such that $(A_m B_{\overline{s-m+1}}) = (A_m) = (B_{\overline{s-m+1}})$ $\forall\, m = 1(1)\min(s,t)$ and all other frequencies are zero. (assuming that $s<t$),

Then, $S = \sum_{i<j}\sum a_{ij}b_{ij} = -\{\text{No. of pairs } (i,j), i<j \text{ for which } x_i > x_j (\text{or } y_i > y_j)\}$

$$= -\left\{\binom{N}{2} - T_B\right\} \text{ or } -\left\{\binom{N}{2} - T_A\right\}$$

and then $d = -1$.

Remark:→ In the above situations, kendall's τ_b behaves similarly.

Logistic Regression :—

Consider a clinical trial coherse patients are given a treatment and response (=1 if success ; = 0 if failure) is observed. In addition their age is also reported. Let y be the response variable and x be the age. Then we assumed

$$P(y=1) = \pi(x)$$

$$\pi(x) = \frac{e^{a+bx}}{1+e^{a+bx}} \quad ; a, b \text{ are unknown.}$$

Consider n points of data (x_i, y_i) ; $i = 1(1)n$. then the corresponding regression is termed as "Logistic regression".

$$a+bx = \log \frac{\pi(x)}{1-\pi(x)} = \log \text{ odds ratio representation.}$$

Here, coe can't use linear regression model i.e. $y = a+bx$, because for the unknown a and b may not take the value 0 or 1.

Ref:- Agresti :-

Let Y denote a response variable that can assume only two values, say 0 and 1. Denote the expected value of Y by $E(Y) = P(Y=1) = \pi$

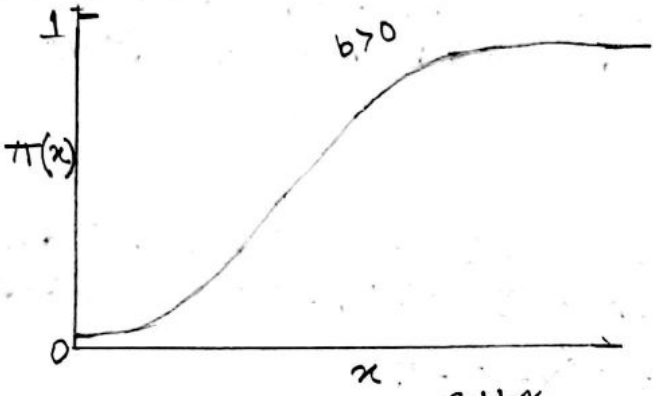
and suppose that we want to model $\pi(x)$ on the values of explanatory variables $X = (x_1, x_2, \ldots, x_k)$. The standard regression model has the form

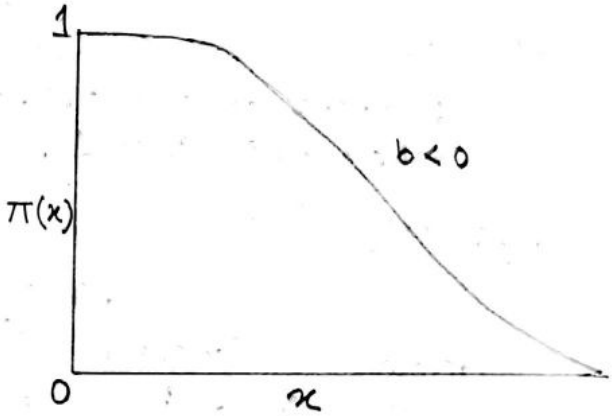
$$\pi(x) = a + b_1 x_1 + \cdots + b_k x_k .$$

There are several difficulties with using ordinary least square principle to fit a model of standard regression conditions that make least squares estimates optimal are not there. For instance, the variance of Y is $\pi(x)(1-\pi(x))$, which is not constant over the range of values of the explanatory variables. standard distributional statements for estimators do not apply, since Y is dichotomous rather than normally distributed.

A weighted least squares approach can be used to obtain more efficient estimates of the regression parameters in this model. The model itself is likely to be inaccurate in certain regions, however, if some x_i's are quantitative. They follows because the model predicts the impossible value $\pi < 0$ and $\pi > 1$ for sufficiently large or sufficiently small values of x_i. For a dichotomous response, $E(Y)$ can't be linearly related to x_i over an unbounded range of x_i values.

Logistic function :⟶ Because of the factors just discussed, it is often more appropriate to use a model that allows a curvilinear relationship between $E(Y)$ and each quantitative x_i. If we expect monotonic relationship then the regression curve will be





Here $\pi(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$ called the logistic function. This function is monotonic with $\pi(x) \downarrow 0$ or $\pi(x) \uparrow 1$ as $x \uparrow \infty$ depending on whether $b<0$ or $b>0$. Takes the value $\pi(x) = \frac{1}{2}$ at $x = -\frac{a}{b}$ and the curve has a steeper rate of increase around that value as $|b|$ increases. When $b>0$, it is the distribution function of the logistic random variable having mean $-\frac{a}{b}$ and s.d. $\frac{\pi}{\sqrt{3}b}$.

Here the odds of making response 1 instead of response 0 is

$$\frac{\pi(x)}{(1-\pi(x))} = e^{a+bx}$$

The odds increases multiplicatively by e^b for every unit of x. The log odds has the simple linear relationship

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = a+bx$$

The model of the log odds is called logistic regression model. In this relationship the logit transformation yields a linear relationship for the logit model. When there are several explanatory variables, the logit model generalisis to

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = a + b_1x_1 + \cdots + b_kx_k.$$

Logit model:→ In applied mathematics & statistics, the logit of a number p between 0 and 1 is

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log p - \log(1-p)$$

The logit function is the inverse of the sigmoid or logistic function. If p is a probability then $p/(1-p)$ is the corresponding odds and the logit of the probability is the logarithm of the odds.

Q) Why do we use logistic regression rather than ordinary linear regression?

Ans:—

1) If we use linear regression, the predicted values will become greater than one and less than zero if you move far enough on the x-axis. Such values are theoritically inadmissible.

2) One of the assumption of regression is that variance of y is constant accross values of x. this cannot be the case with a binary variable because the variance is $p(1-p)$. So, if p approaches to 1 or 0; variance approaches to zero.

3) The significance of testing of the weights b rest upon the assumption that errors of friction $(y-y')$ are normally distributed. Because y only takes values 0 and 1, this assumption is hard to justify. therefore, we can't use linear regression.

Relative Risk: — A value $\pi_1 - \pi_2$ of fixed size may have greater importance when both π_i are close to zero or 1 than when they are not. For a study comparing two treatments on the proportion of subjects who die, the difference between 0.010 and 0.001 and 0.410 and 0.401 are both 0.009. In such cases, the ratio of proportions is also informative. The relative risk is defined to be the ratio $\frac{\pi_1}{\pi_2}$. It can be any non negative real number. Relative risk of one corresponds to independence, e.g. for the proportions just given the relative risks are $\frac{0.010}{0.001} = 10$ and $\frac{0.410}{0.401} = 1.02$.

Comparing the rows on the 2nd response (failure) category gives a different relative risk, $\frac{1-\pi_1}{1-\pi_2}$.

2.0.

Odds Ratio: — For a probability π of a success, the odds are define to be $\Omega = \frac{\pi}{1-\pi}$. the odds are non-negative with $\Omega > 1$ when a success is more likely than a failure, e.g.

$\pi = 0.75 \Rightarrow \Omega = 3$, a success is thrice is as likely as a failure.

Inversely, $\pi = \frac{\Omega}{\Omega+1}$.

Refer again to 2×2 table, within rows i, the odds of success instead of failure are $\Omega_i = \frac{\pi_i}{1-\pi_i}$, the ratio of the odds Ω_1 and Ω_2 in the two rows,

$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$; is called the odds ratio.

For joint distn with cell probabilities $\{\pi_{ij}\}$, the equivalent defn of Ω_i could be

$\Omega_i = \frac{\pi_{i1}}{\pi_{i2}}$. Hence,

$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$

Hence, odds ratio is also called <u>cross-product ratio</u>.

Properties of Odds ratio: — The odds ratio can be equal to any non-negative number. The condition $\Omega_1 = \Omega_2$ (when all cell probabilities are +ve), an hence $\theta = 1$ corresponds to independence of X and Y. When $1 < \theta < \infty$ subjects in row 1 are more likely to have a success than subjects in row 2; i.e. $\pi_1 > \pi_2$. For instance, when $\theta = 4$, the odds of success in row 1 are 4 times the odds in row 2, this does not mean that $\pi_1 = 4\pi_2$, i.e. relative risk of 4. When $0 < \theta < 1$, $\pi_1 < \pi_2$, one of the cell probabilities vanishes, θ becomes 0 or ∞.

Values of θ further from 1 in a given direction represents stronger association. Two values represent the same association, but in opposite directions, when one is the inverse of the other.

C.J.

Relationship between Odds ratio & Relative risk: —

$$\text{Odds ratio} = \text{relative risk} \times \frac{1-\pi_2}{1-\pi_1}.$$

Their magnitude are similar whenever the probability π_i of the outcome of interest is close to zero for both groups thus when each π_i is small, the odds ratio provides a rough indication of the relative risk when it is not directly estimable, such as in case of controlled study.
We can compute the odds ratio, however, since it is determined by the conditional distributions in either direction. When the probability of the outcome of interest is very small, the population odds ratio and relative risk take similar values.

Logistic Regression:

Let Y denote a binary response variable, for instance Y might indicate the choice of car (new or used) or the diagonasis of cancer (present or absent). Each observation has one of 2 outcomes denoted by 0 and 1. Now for a response variable Y and an explanatory variable X, consider the data –

Setting of x	Values of Y	Totals
x_1	$y_{11}\ y_{12} \cdots y_{1N_1}$	$\sum_{j=1}^{N_1} y_{1j} = n_1$
x_2	$y_{21}\ y_{22} \cdots y_{2N_2}$	$\sum_{j=1}^{N_2} y_{2j} = n_2$
⋮		⋮
x_K	$y_{K1}\ y_{K2} \cdots y_{KN_K}$	$\sum_{j=1}^{N_K} y_{Kj} = n_K$

Here the true regression of y on x is given by the array mean:

$$\Pi(x_i) = \bar{y}_i = \sum_{j=1}^{N_i} y_{ij} \Big/ N_i = \frac{n_i}{N_i}$$

Since n_i is the number of y_{ij} which takes the value 1 when $x = x_i$ and $0 \le n_i \le N_i$. Hence, $0 \le \bar{y}_i \le 1$ and

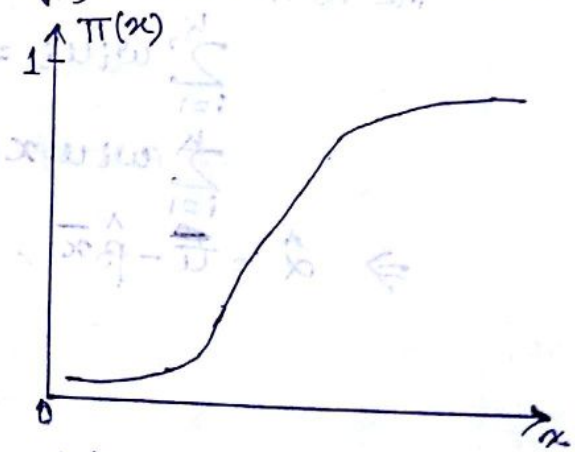
$$0 < \frac{\bar{y}_i}{1-\bar{y}_i} < \infty \text{ . Then, } -\infty < \log_e\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right) < \infty.$$

i.e. $\log_e\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right)$ can be any real number, the real nos are also the range of any linear predictor such as $(\alpha+\beta x)$. We can assume the regression model as $\log_e\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right) = \alpha+\beta x_i$

$$\Leftrightarrow \bar{y}_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

[Usually binary data, result from a non-linear relationship between $\Pi(x_i) = \bar{y}_i$ and x_i. A fixed change in x has less impact on $\Pi(x)$ when $\Pi(x)$ is near 0 or 1 than when $\Pi(x)$ is near 0.5.]

In practice, non linear relationship between $\Pi(x)$ and x is monotonic. Most important curve with the above shape is the logistic curve.



An appropriate regression model is

$$\Pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}},$$

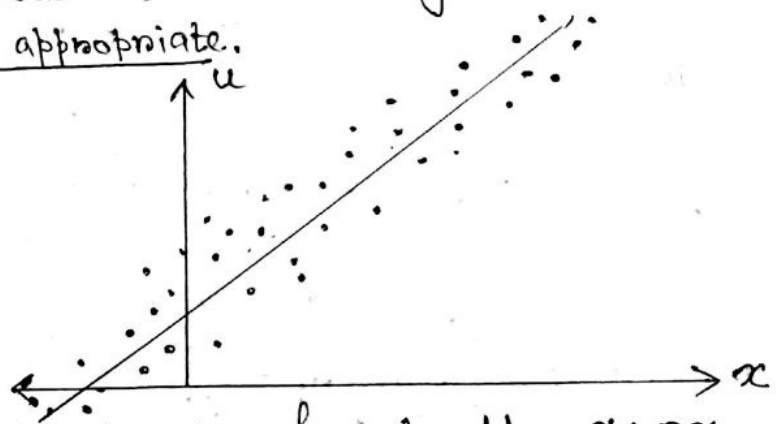
which is called logistic regression model.

Looking at the data & fitting of logistic regression equation

(C.U.)

Before fitting logistic model, look at the data to check whether the logistic regression is appropriate or not. Since Y takes any of the values 0 and 1, it is difficult to check by plotting Y-values for different x.

If the quantities $u_i = \log\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right) = \log\left(\frac{n_i}{N_i - n_i}\right)$ are plotted against x_i, $i = 1(1)k$, and the points (x_i, u_i) are near about a line, then, by definition, the logistic regression is appropriate.



(C.U.)

- **Fitting :** Consider the predicting formula $U_x = \alpha + \beta x$

When $x = x_i$, then the predicting value of $u_i = \log\left(\frac{\bar{y}_i}{1-\bar{y}_i}\right)$ is $U_{x_i} = \alpha + \beta x_i$ and the corresponding error $e_i = u_i - U_{x_i} = (u_i - \alpha - \beta x_i)$.

To determine α and β, we shall use method of weighted least squares and we shall minimise

$$S^2 = \sum_{i=1}^{k} w_i e_i^2 = \sum_{i=1}^{k} w_i (u_i - \alpha - \beta x_i)^2$$, where the weights $w_i = N_i \bar{y}_i (1-\bar{y}_i) = N_i \pi(x_i)(1-\pi(x_i))$, with respect to α and β.

The normal equations are :

$$\sum_{i=1}^{k} w_i u_i = \alpha \sum_{i=1}^{k} w_i + \beta \sum_{i=1}^{k} w_i x_i$$

$$\sum_{i=1}^{k} w_i u_i x_i = \alpha \sum_{i=1}^{k} w_i x_i + \beta \sum w_i x_i^2$$

$$\Rightarrow \hat{\alpha} = \bar{u} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{\sum_{i=1}^{k} w_i u_i x_i - N\bar{u}\bar{x}}{\sum w_i x_i^2 - N\bar{x}^2}$$, where

$$\bar{u} = \frac{\sum w_i u_i}{\sum w_i} \text{ and } \bar{x} = \frac{\sum w_i x_i}{\sum w_i}.$$

Question: — What is the rationale behind the Yule's coefficient of association?

Soln. → δ is a measure of independence of A and B. To get a normed measure, we divide $N\delta$, the difference between $(AB)(\alpha\beta)$ & $(A\beta)(\alpha B)$ by the sum $(AB)(\alpha\beta)+(A\beta+\alpha B)$ and the resulting coefficient is $Q = \dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$, which is suitable for comparison of two or more data sets.

Question: — What is the rationale behind Yule's coefficient of colligation?

Soln. → Empirical study reveals that the coefficient Q overestimate the extent of association. Hence, Yule provides a measure by considering lower orders of the cell frequencies as Y, a coefficient of colligation. Note that $Q = \dfrac{2Y}{1+Y^2} \geq Y$.

Question: — Let $x_i = \begin{cases} 1, & \text{ith individual possess A and similarly} \\ 0, & \text{ith individual possess } \alpha. \end{cases}$

$y_i = \begin{cases} 1, & \text{ith individual possess B} \\ 0, & \text{ith individual possess } \beta. \end{cases}$

Find the correlation coefficient between x and y based on the data $\{(x_i, y_i) : i = 1(1)n\}$.

Soln. →

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^{1/2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)^{1/2}}$$

$$= \frac{f_{11} - n\cdot\frac{f_{10}}{n}\cdot\frac{f_{01}}{n}}{\left\{f_{10} - \frac{f_{10}^2}{n}\right\}^{1/2}\left\{f_{01} - \frac{f_{01}^2}{n}\right\}^{1/2}}$$

$$= \frac{(f_{11}f_{22} - f_{12}f_{21})/n}{\left(\frac{f_{10}f_{20}}{n}\right)^{1/2}\left(\frac{f_{01}f_{02}}{n}\right)^{1/2}}$$

$$= \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{f_{10}f_{20}f_{01}f_{02}}}$$

$$= \frac{n\delta}{\sqrt{f_{10}f_{20}f_{01}f_{02}}}$$

y \ x	1	0	
1	f_{11}	f_{21}	f_{01}
0	f_{12}	f_{22}	f_{02}
	f_{10}	f_{20}	n

Hence, $r_{xy} = 0$ iff $\delta = 0$ iff A & B are independent.

Problem: For a 2×2 table, Yule introduces

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

i) S.T. in a 2×2 table, Goodman-Kruskal's $\gamma = Q$.

ii) S.T. the Yule coefficient can be written as $Q = \frac{\hat{\theta} - 1}{\hat{\theta} + 1}$, a monotonic transformation of $\hat{\theta}$ from $[0, \infty)$ to $[-1, 1]$.

Soln. →

i) $$\gamma = \frac{C - D}{C + D}$$

X \ Y	Y_1	Y_2
X_1	n_{11}	n_{12}
X_2	n_{21}	n_{22}

Here, C = No. of pairs (i, j), $i < j$ for which $\{x_i > x_j, y_i > y_i\}$ or $\{x_i < x_j, y_i < y_j\}$

$= n_{11} \times n_{22}$

and D = No. of pairs (i, j), $i < j$ for which $\{x_i > x_j, y_i < y_j\}$ or $\{x_i < x_j, y_i > y_j\}$

$= n_{21} \times n_{12}$

$\therefore \gamma = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = Q$

ii) $\hat{\theta}$ = the sample odds ratio or α-measure

$$= \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

$$\therefore Q = \frac{\hat{\theta} - 1}{\hat{\theta} + 1},$$

$$\frac{dQ}{d\hat{\theta}} = \frac{d}{d\hat{\theta}}\left(1 - \frac{2}{\hat{\theta}+1}\right) = \frac{2}{(\hat{\theta}+1)^2} > 0$$

$\Rightarrow Q$ is monotonic increasing in $\hat{\theta}$.

As $0 \le \hat{\theta} < \infty$, $1 \le \hat{\theta} + 1 < \infty$

$$\Rightarrow 0 < \frac{2}{\hat{\theta}+1} \le 2$$

$$\Rightarrow -1 \le 1 - \frac{2}{\hat{\theta}+1} \le 1$$

$$\Rightarrow -1 \le Q \le 1.$$

Categorical Data Analysis

Categorical Table :— Let us consider two categorical variable A and B, where A takes values on p categories, say, $A_1, A_2, \ldots, A_p$ and B takes values $B_1, B_2, \ldots, B_q$, on q categories.

Let us define the probability,

$$\pi_{ij} = P(A = A_i \cap B = B_j) \quad \text{for } i = 1(1)p, j = 1(1)q.$$

If we construct a table consisting of p rows for the p-categories of A and q columns for the q categories of B, then we get a $p \times q$ table as follows:

A \ B	B_1	B_2	$\ldots$	B_j	$\ldots$	B_q	
A_1	π_{11}	π_{12}	$\ldots$	π_{1j}	$\ldots$	π_{1q}	π_{10}
A_2	π_{21}	π_{22}	$\ldots$	π_{2j}	$\ldots$	π_{2q}	π_{20}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	π_{i1}	π_{i2}	$\ldots$	π_{ij}	$\ldots$	π_{iq}	π_{i0}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_p	π_{p1}	π_{p2}	$\ldots$	π_{pj}	$\ldots$	π_{pq}	π_{p0}
	π_{01}	π_{02}	$\ldots$	π_{0j}	$\ldots$	π_{0q}	

This table is called a $p \times q$ contingency table. Here —

$$\pi_{i0} = \sum_{j=1}^{q} \pi_{ij} \quad \& \quad \pi_{0j} = \sum_{i=1}^{p} \pi_{ij}$$

are called Marginal Probabilities.

π_{ij}'s are called the Joint probabilities and with their probabilities we can define the conditional probabilities in this table as follows :

$$\pi_{j|i} = P(B = B_j \mid A = A_i)$$
$$= \frac{P(A = A_i \cap B = B_j)}{P(A = A_i)}$$
$$= \frac{\pi_{ij}}{\pi_{i0}}$$

Independence of two categorical variables: —

Two categorical variables A and B taking values on two categorical scales consisting of the categories $A_1, A_2, \ldots, A_p$ and $B_1, B_2, \ldots, B_q$, respectively, are independent if the conditional probability of assuming a value by one variable, say B, remains the same for any value, assumed by the other variable.

For every j, $\pi_{j|i} = P(B = B_j \mid A = A_i) = \pi_{0j} \quad \forall i$

i.e. for every j, $\frac{\pi_{ij}}{\pi_{i0}} = \pi_{0j} \quad \forall i$

$\Rightarrow \pi_{ij} = \pi_{i0} \cdot \pi_{0j} \quad \forall i, j$

Difference measures of independence in a 2×2 contingency table: —

Let there be two groups of subjects and let each of them be allowed to response on a binary variable. Let the categorical variable corresponding to the group of subject denoted by X and that corresponding to the response be denoted by Y. Then both X and Y are binary variables.

Response / Groups	Yes	No
Grp-I	π_1	$1-\pi_1$
Grp-II	π_2	$1-\pi_2$

In the adjacent table, let π_i $(i=1,2)$ be the proportion of 'Yes' response in the i-th group. Note that, here the probabilities π_1 and π_2 are actually conditional probability given for the group.

Now if the categorical variable X and Y are independent, Then $\pi_1 - \pi_2 = 0$. Now considering another table —

Response / Groups	Yes	No	No. of subjects
Grp-I	x_1	$n_1 - x_1$	n_1
Grp-II	x_2	$n_2 - x_2$	n_2

Now, we can judge independence from the difference of the observed proportion

$p_1 = \frac{x_1}{n_1}$ and $p_2 = \frac{x_2}{n_2}$,

If $p_1 - p_2$ is close to zero then we can predict independence, otherwise, association is believed to exist between the categorical variable X and Y.

Limitations: → Limitation of this measures of independence is that when π is close to 0 or 1, the difference may lead to a wrong decision about independence; for instance, let $\pi_1 = 0.010$ and $\pi_2 = 0.001$, then $\pi_1 - \pi_2 = 0.009$, again let $\pi_1^* = 0.410$ and $\pi_2^* = 0.401$, then $\pi_1^* - \pi_2^* = 0.009$, i.e. difference of probabilities are same in both the cases, though $\frac{\pi_1}{\pi_2} = 10$ and $\frac{\pi_1^*}{\pi_2^*} = 1.2$. Limits of this measure are -1 and $+1$, $\pi_1 - \pi_2 = +1 \Rightarrow$ extreme +ve association, $\pi_1 - \pi_2 = -1 \Rightarrow$ extreme -ve association.

Example:-

Groups	AGES	YES	NO	TOTAL
Grp-I	Age ≥ 70	60	0	$60 = n_1$
Grp-II	Age ≤ 20	0	40	$40 = n_2$

$\pi_1 = 1; \pi_2 = 0$

$\therefore \boxed{\pi_1 - \pi_2 = 1}$

$\pi_1 = 0; \pi_2 = 1$

$\therefore \boxed{\pi_1 - \pi_2 = -1}$

where, $\pi_1 = \frac{x_1}{n_1}$, $\pi_2 = \frac{x_2}{n_2}$.

ODDS RATIO:- Let us compare two groups of subjects on a binary response variable Y. Let the categorical variable Y be allowed to take values 1 and 2. Then we denote by X another categorical variable, also taking values 1 and 2, to denote the two groups of subjects:

Groups (X)	Response (Y) 1	2	
1	π_1	$1-\pi_1$	1
2	π_2	$1-\pi_2$	1

If, we consider the response '1' as "success" with probability π, then the odds of success is defined by $\frac{\pi}{1-\pi}$. In the present case the odds of success for the 1st group is

$\text{Odd } S_1 = \frac{\pi_1}{1-\pi_1}$ and that for the 2nd group is

$\text{Odd } S_2 = \frac{\pi_2}{1-\pi_2}$.

Then, we define the Odds Ratio of the two groups by $\theta = \frac{\text{Odds}_1}{\text{Odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$.

When two categorical variables X and Y are independent, $\pi_1 = \pi_2$, then the odds ratio $(\theta) = 1$.

When $\theta > 1$, we say that X and Y are +vely associated.

& $\theta < 1$, we say that X and Y are -vely associated.

PROPERTIES OF ODDS RATIO :—

1) Odds Ratio can take any non-negative value.

2) $\theta = 1 \Rightarrow$ Independence.
   $\theta > 1 \Rightarrow$ positive association.
   $\theta < 1 \Rightarrow$ negative association.

3) If θ and θ^* are two odds ratio s such that $\theta = \frac{1}{\theta^*}$; then we can calculate that degree of association is the same in both the cases, though the direction of association is opposite.

4) For a table with joint probabilities, the odds ratio is given by $\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$.

Proof:- In the table, we can write the odds ratio in terms of conditional probabilities

$$\pi_{1|1} = \frac{\pi_{11}}{\pi_{11}+\pi_{12}} \text{ and } \pi_{1|2} = \frac{\pi_{21}}{\pi_{21}+\pi_{22}}$$

Then, $\theta = \frac{\pi_{1|1}/(1-\pi_{1|1})}{\pi_{1|2}/(1-\pi_{1|2})} = \frac{\pi_{1|1}(1-\pi_{1|2})}{\pi_{1|2}(1-\pi_{1|1})}$

$= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$.

5) If any rows or any column of the table is interchanged, then the odds ratio is reversed.

Proof:-

	1	2
1	π_1	$1-\pi_1$
2	π_2	$1-\pi_2$

Odds ratio is $\theta = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$

Now, if the 1st and 2nd rows of the table are interchanged to get the new table.

	1	2
1	π_2	$1-\pi_2$
2	π_1	$1-\pi_1$

Then the odds ratio is

$$\theta^* = \frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)} = \frac{1}{\theta}.$$

6. If the orientation of the table is reversed then the odds ratio does not change.

Proof:→

X \ Y	1	2
1	π_1	$1-\pi_1$
2	π_2	$1-\pi_2$

Here the odds ratio is $\theta = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$

Now, if the orientation of this table is reversed to get the new table.

X \ Y	1	2
1	π_1	π_2
2	$1-\pi_1$	$1-\pi_2$

Here, the odds ratio is

$$\theta^* = \frac{\pi_1/\pi_2}{(1-\pi_1)/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} = \theta.$$

(e.g.) **PROBLEM :→** Show that sample odds ratio does not change when both cell counts within any row are multiplied by a non-zero constant or when both cell counts within any column are multiplied by a non-zero constant. Discuss the implication of the above result using a real life example.

ANS:→

	1	2
1	π_1	$1-\pi_1$
2	π_2	$1-\pi_2$

Odds ratio is

$$\theta = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Now, if the both cell counts with in any row are multiplied by a constant (non-zero), say, k, we get the new table.

	1	2
1	$k\pi_1$	$k(1-\pi_1)$
2	π_2	$1-\pi_2$

Then the odds ratio is

$$\theta^* = \frac{k\pi_1(1-\pi_2)}{\pi_2(k(1-\pi_1)}$$

$$= \theta$$

So, odds ratio does not change.

An implication of the multiplicative invariance property is that the sample odds ratio estimates the same characteristic (θ) even when we select disproportionately large or small samples from marginal categories of a variable. For instance, suppose a study investigates the association between vaccination and catching a certain strain of flu. For a retrospective design, the sample odds ratio estimates the same characteristic whether we randomly sample (1) 100 people who got the flu and 100 people who didn't, (2) 150 people who got the flu and 50 people who didn't, in each case classifying subjects on whether they took the vaccine. In fact, the odds ratio is equally valid for retrospective, prospective, or cross-sectional sampling designs. We would estimate the same characteristic if (3) we randomly sample 100 people who took the vaccine and 100 people who didn't, and then classify them on whether they got the flu, or (4) we randomly sample 200 people and classify them on whether they took the vaccine and whether they got the flu.

—— × ——