Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 3 Issue 10, October - 2016

The role of statistical independence in contingency tables

Erjola CENAJ[1]
Departament of Mathematics, Mathematical and Physical Engineering Faculty
Polytechnic University of Tirana
Tirana, Albania
erjola_cenaj@yahoo.com

Luela PRIFTI[2]
Departament of Mathematics, Mathematical and Physical Engineering Faculty
Polytechnic University of Tirana
Tirana, Albania
luela_p@yahoo.com

***Abstract*— This paper discusses the statistical independence can be observed in a contingency table when the table is viewed as a matrix. Statistical independence in a contingency table is represented as a special form of linear dependence, where all the rows or columns are described by one row or column, respectively. The use chi-square test and the rank of matrices for statistical independence of an application.**

Keywords—Statistical Independence, Contingency Table, Matrix Contingency, chi-square test

I. Introduction

In this paper, a statistical independence in a contingency table is focused on from the viewpoint of granular computing. The first important observation is that a contingency table compares two attributes with respect to information granularity. It is shown from the definition that statistifcal independence in a contingency table is a special form of linear depedence of two attributes. Especially, when the table is viewed as a matrix, the above discussion shows that the rank of the matrix is equal to 1.0. The second important observation is that matrix algebra is a key point of analysis of this table. A contingency table can be viewed as a matrix and several operations and ideas of matrix theory are introduced into the analysis of the contingency table. The paper is organized as follows: Section II discusses the characteristics of contingency tables, shows the conditions on statistical independence for a 2 × 2 table and gives those for a 2×3 and 2 × n table.

Section III discusses statistical independence from matrix theory , chi square test and A survey was made showing statistical independence through the rank of matrix.

II. Contingency Table

A. Two-way Contingency Table

From the viewpoint of information systems, acontingency table summarizes the relation between two attributes with respect to frequencies. However, this study focuses on more statistical interpretation of this table.

Definition 1 Let X_1 and X_2 denote binary attributes in an attribute space A. A contingency table is a table of a set of the meaning of the following formulas:

$|[X_1=0]|$, $|[X_1=1]|$, $|[X_2=0]|$, $|[X_2=1]|$, $|[X_1=0\wedge X_2=0]|$,
$|[X_1=0\wedge X_2=1]|$, $|[X_1=1\wedge X_2=0]|$, $|[X_1=1\wedge X_2=1]|$,
$|[X_1=0\vee X_1=1]|$, $|[X_1=0\vee X_1=1]|$ (=N)

This table is arranged into the form shown in Table 1, where:

$|[X_1=0]|=x_{11}+x_{21}=x_{.1}$, $|[X_1=1]|=x_{12}+x_{22}=x_{.2}$,
$|[X_2=0]|=x_{11}+x_{12}=x_{1.}$, $|[X_2=1]|=x_{21}+x_{22}=x_{2.}$,
$|[X_1=0\wedge X_2=0]|=x_{11}$, $|[X_1=0\wedge X_2=1]|=x_{21}$,
$|[X_1=1\wedge X_2=0]|=x_{12}$, $|[X_1=1\wedge X_2=1]|=x_{22}$,
$|[X_1=0\vee X_1=1]|=x_{.1}+x_{.2}=x_{..}$,
$|[X_2=0\vee X_2=1]|=x_{1.}+x_{2.}=x_{..}$(=N)

Table 1

	$X_1=0$	$X_1=1$	Total
$X_2=0$	x_{11}	x_{12}	x_1
$X_2=1$	x_{21}	x_{22}	$x_{2.}$
Total	$x_{.1}$	$x_{.2}$	$x_{..}$

B. Contingency Table(m × n)

Two-way contingency table can be extended into a contingency table for multinominal attributes.

Definition 2 Let X_1 and X_2 denote multinominal attributes in an attribute space A which have m and n values. A contingency tables is a table of a set of the meaning of the following formulas:

$|[X_1=A_j]|$, $|[X_2=B_i]|$, $|[X_1=A_j\wedge X_2=B_i]|$,
$|[X_1=A_1\wedge X_1=A_2\ldots\wedge X_1=A_m]|$,
$|[X_2=B_1\wedge X_2=A_2\ldots\wedge X_2=A_n]|$=(N)
$(i=1,2,\ldots,n, m=1,2,\ldots,m), N=x_{..}$

This table is arranged into the form shown in Table 1, where:

$|[X_1=A_j]|=\sum_{i=1}^{m}x_{1i}=x_{.j}$, $|[X_2=B_i]|=\sum_{j=1}^{n}x_{j1}=x_{i.}$,
$|[X_1=A_j\wedge X_2=B_i]|=x_{ij}$, $x_{..}=N$, $(i=1,2,\ldots,n, m=1,2,\ldots,m)$.

Table 2. Contingency Table (m × n)

	A_1	A_2	...	A_n	Total
B_1	x_{11}	x_{12}	...	x_{1n}	$x_{1.}$
B_2	x_{21}	x_{22}		x_{2n}	$x_{2.}$
⋮	⋮	⋮	⋱	⋮	⋮
B_m	x_{m1}	x_{m2}	...		$x_{m.}$
Total	$x_{.1}$	$x_{.2}$	...	$x_{.n}$	$x_{..}$

C. *Statistical Independence in 2 × 2*

Contingency Table Let us consider a contingency table shown in Table 1. Statistical independence between X_1 and X_2 gives:

P([X_1=0, X_2=0)]=P([X_1=0]) P([X_2=0])

P([X_1=0, X_2=1)]=P([X_1=0]) P([X_2=1])

P([X_1=1, X_2=0)]=P([X_1=1]) P([X_2=0])

P([X_1=1, X_2=1)]=P([X_1=1]) P([X_2=1])

Since each probability is given as a ratio of each cell to N, the above equations are calculated as:

$$\frac{x_{11}}{N} = \frac{x_{11}+x_{12}}{N} \cdot \frac{x_{11}+x_{21}}{N},$$
$$\frac{x_{12}}{N} = \frac{x_{11}+x_{12}}{N} \cdot \frac{x_{12}+x_{22}}{N},$$
$$\frac{x_{21}}{N} = \frac{x_{21}+x_{22}}{N} \cdot \frac{x_{11}+x_{21}}{N},$$
$$\frac{x_{22}}{N} = \frac{x_{21}+x_{22}}{N} \cdot \frac{x_{12}+x_{22}}{N}$$

Since $N = \sum_{ij} x_{ij}$, the following formula will be obtained from these four formulae.

$x_{11}x_{22} = x_{12}x_{21}$ or $\frac{x_{11}x_{22}}{x_{12}x_{21}} = 1$.

Theorem 1

If two attributes in a contingency table shown in Table 1 are statistical indepedent, the following equation holds: $x_{11}x_{22} - x_{12}x_{21} = 0$. (1)

D. *Statistical Independence in 2 × 3 Contingency Table*

Let us consider a 2 × 3 contingency table shown in Table 3. Statistical independence between X_1 and X_2 gives:

Table 3. contingency table 2 × 3

	X_1=0	X_1=1	X_1=2	Total
X_2=0	x_{11}	x_{12}	x_{13}	$x_{1.}$
X_2=1	x_{21}	x_{22}	x_{23}	$x_{2.}$
Total	$x_{.1}$	$x_{.2}$	$x_{.3}$	$x_{3.}$

P([X_1=0, X_2=0)]=P([X_1=0]) P([X_2=0])

P([X_1=0, X_2=1)]=P([X_1=0]) P([X_2=1])

P([X_1=0, X_2=2)]=P([X_1=0]) P([X_2=2])

P([X_1=1, X_2=0)]=P([X_1=1]) P([X_2=0])

P([X_1=1, X_2=1)]=P([X_1=1]) P([X_2=1])

P([X_1=1, X_2=2)]=P([X_1=1]) P([X_2=2])

Since each probability is given as a ratio of each cell to N, the above equations are calculated as:

$$\frac{x_{11}}{N} = \frac{x_{11}+x_{12}+x_{13}}{N} \cdot \frac{x_{11}+x_{21}}{N} \quad (2)$$
$$\frac{x_{12}}{N} = \frac{x_{11}+x_{12}+x_{13}}{N} \cdot \frac{x_{12}+x_{22}}{N} \quad (3)$$
$$\frac{x_{13}}{N} = \frac{x_{11}+x_{12}+x_{13}}{N} \cdot \frac{x_{13}+x_{23}}{N} \quad (4)$$
$$\frac{x_{21}}{N} = \frac{x_{21}+x_{22}+x_{23}}{N} \cdot \frac{x_{11}+x_{21}}{N} \quad (5)$$
$$\frac{x_{22}}{N} = \frac{x_{21}+x_{22}+x_{23}}{N} \cdot \frac{x_{12}+x_{22}}{N} \quad (6)$$
$$\frac{x_{23}}{N} = \frac{x_{21}+x_{22}+x_{23}}{N} \cdot \frac{x_{13}+x_{23}}{N} \quad (7)$$

From equation (2) and (5),

$$\frac{x_{11}}{x_{21}} = \frac{x_{11}+x_{12}+x_{13}}{x_{21}+x_{22}+x_{23}}$$

In the same way, the following equation will be obtained:

$$\frac{x_{11}}{x_{21}} = \frac{x_{12}}{x_{22}} = \frac{x_{13}}{x_{23}} = \frac{x_{11}+x_{12}+x_{13}}{x_{21}+x_{22}+x_{23}}$$

Thus, we obtain the following theorem:

Theorem 2 If two attributes in a contingency table shown in Table 3 are statistical indepedent, the following equations hold: $x_{11}x_{22} - x_{12}x_{21} = x_{23}x_{23} - x_{13}x_{22} = x_{13}x_{21} - x_{11}x_{23} = 0$.

It is notable that this discussion can be easily extended into a 2xn contingency table where n > 3.

Thus, Theorem 3 If two attributes in a contingency table (2×k(k = 2, ··· , n)) are statistical indepedent, the following equations hold: $x_{11}x_{22} - x_{12}x_{21} = x_{21}x_{23} - x_{13}x_{22} = \cdots = x_{1n}x_{21} - x_{11}x_{n3} = 0$.

III. CONTINGENCY TABLE

The meaning of the above discussions will become much clearer when we view a contingency table as a matrix.

Definition 3

A corresponding matrix $C_{T_{a,b}}$is defined as a matrix the element of which are equal to the value of the corresponding contingency table $T_{a,b}$ of two attributes a and b, except for marginal values.

Definition 4

The rank of a table is defined as the rank of its corresponding matrix. The maximum value of the rank is equal to the size of (square) matrix, denoted by r.

The contingency matrix of Table $2T(X_1, X_2)$is defined as $(C_{T_{X_1,X_2}})$as below:

$$N = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

a) *Independence of 2 × 2 Contingency Table*

Let us assume that a contingency table is given as Table 1. Then the corresponding matrix $(C_{T_{X_1,X_2}})$ is given as:

$$M = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$$

Proposition 1 The rank will be:

$$r(M) = \begin{cases} 2, & if \ |M| \neq 0 \\ 1, & if \ |M| = 0 \end{cases}$$

From Theorem 1,

Theorem 7 If the rank of the corresponding matrix of a 2times2 contingency table is 1, then two attributes in a given contingency table are statistically independent. Thus,

$$r(M) = \begin{cases} 2, \text{dependent} \\ 1, \text{statistical independent} \end{cases}$$

This discussion can be extended into 2 × 3 tables. According to Theorem 3, the following theorem is obtained.

Theorem 8

If the rank of the corresponding matrix of a 2times3 contingency table is 1, then two attributes in a given contingency table are statistically independent. Thus,

$$rank = \begin{cases} 2, \text{dependent} \\ 1, \text{statistical independent} \end{cases}$$

This discussion can be extended into 2 × n tables. According to Theorem 3, the following theorem is obtained.

Theorem 9 If the rank of the corresponding matrix of a 2 × n contigency table is 1, then two attributes in a given contingency table are statistically independent.

Thus, $r(N) = \begin{cases} 2, \text{dependent} \\ 1, \text{statistical independent} \end{cases}$

b) Chi-Square Test for Independence

The test is applied when we have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

c) Application

Respondents for their data device was made byphonequestion for iPhone and age.70 The respondents areclassified according to two criteria

a. Equipping with mobile phone iPhone

b. Age

Results are summarized in a contingency table with (2 × 3)

	18-30	*31-45*	*>45*	*Totale*
Yes	8	11	2	21
No	22	10	17	49
Totale	30	21	19	70

H_0: IPhone telephone device is independent of age

H_a: IPhone telephone device is not independent of age.

The test statistic is χ^2=8.11 >5.99 = $\chi^2_{0,05}$

then it is not true H_0

matrix corresponding contingency table is

$A = \begin{pmatrix} 8 & 11 \\ 22 & 10 \end{pmatrix}$ r(A) = 2 then mobile device with iPhone is dependent on age.

So we give the same result with both methods with chi-square test and the range of matrices.

References

[1] A. Agresti, Categorical Data Analysis, 2nd ed., University of Florida, 2003, pp.36-78.

[2] Sh. Tsumoto, Statistical Independence as Linear Independence Published by Science B. V, March 2003, Volume 82, Issue 4,vol. 82., pp.274-285.

[3] Sh. Tsumoto, S. Hirano, Linear independence in a contingency table. Published in 2005 IEEE International Conference on Granular Computing, vol.2.

[4] Sh. Tsumoto. Statistical independence and Determinants in a Contingency table.Interpretation of Pearson Reziduals based on linear Algebra. Volume 90, pp251-267.

[5] Sh. Tsumoto, S. Hirano, Role of Marginal Distribution in a Contingency table, Proceedings of NAFIPS 2006, IEEE press, 2006.