Developments in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 19, Ljubljana: FDV, 2003

On the Normalization of χ^2 based Contingency Indices

Antonio Mango[1]

Abstract

In this paper a new way to construct a χ^2 based contingency coefficient other than Cramèr's V^2 and Tschuprov's T^2 is proposed.

The author explains his point of view on the opportunity to select a particular element of the Class of Fréchet, to which the table of contingency under analysis belongs, to get the maximum value of χ^2.

Both a manual way and a computational one are exposed to obtain the maximum contingency table.

1 Premise

Every statistical textbook and statistical package for social sciences gives the *Cramèr*'s and the *Tschuprov*'s solutions for the normalized χ^2 based contingency coefficients when the contingency tables are not square or have no couples of marginal row and column frequencies with equal coordinates.

Both solutions furnish distorted coefficients because they refer to tables with dimension or marginal frequencies different from those which characterize the tables under study, holding only in account the total frequency, N, the number of rows, r, and the number of columns, c.

We start from the following three tables to underline the type of distortions those solutions introduce:

$m\backslash M$	A	B	$\sum$
a	50	10	60
b	20	20	40
$\sum$	70	30	100

Table 1

$m\backslash M$	A	B	C	$\sum$
a	30	20	10	60
b	7	13	20	40
$\sum$	37	33	30	100

Table 2

$m\backslash M$	A	B	C	D	$\sum$
a	30	20	3	7	60
b	7	13	17	3	40
$\sum$	37	33	20	10	100

Table 3

their respective χ^2 values are:

$$\chi^2_1 = 12.70,\ \chi^2_2 = 15.75 \text{ and } \chi^2_3 = 18.70 \tag{1.1}$$

[1] Dipartimento di Matematica e Statistica, Università degli Studi di Napoli "Federico II", mango@unina.it

which, according to *Cramèr*'s solution expressed by

$$V^2 = \frac{\chi^2}{n\,(q-1)}$$

where $q = min\,[r, c]$, turn into the coefficients:

$$V_1^2 = .0898,\ V_2^2 = .111 \text{ and } V_3^2 = .207 \tag{1.2}$$

and, according to *Tschuprov*'s

$$T^2 = \frac{\chi^2}{n\sqrt{(r-1)\,(c-1)}}$$

become:

$$T_1^2 = .127,\ T_2^2 = .112 \text{ and } T_3^2 = .108. \tag{1.3}$$

The expressions which appear at the denominators of V^2 and T^2 represent the particular theoretical maximum values of χ^2 we criticize.

Since we will use the notion of Class of Fréchet, we shortly remember that it deals with the set of all contingency tables having same dimension and same marginal frequencies.

On the basis of a different idea we maintain that the maximum value of χ^2 has to refer to an element of the Class of Fréchet to which the table under analysis belongs and that this value must be calculated on the table presenting the *highest cell frequency concentration.*

We remember an interesting work of Diaconis and Efron (1985) in which the *Volume test* for independence is proposed.

The procedure for the construction of the test foresees the production, by computer simulation, of a sequence of contingency tables all belonging to the same Class of Fréchet. The greatest value of χ^2 chosen from those associated to these tables, can be used for normalization.

If we repeat this procedure several times we can see that the set of these χ^2 values extends to an *upper limit* which coincides with the value that we propose. The procedure of Diaconis and Efron is justified in the logic of the problems of inference dealt with, while the value of χ^2 we propose has a descriptive meaning and is a parameter itself, not an estimate and can be calculated very simply and quickly.

The following three tables, that we call *maximum contingency tables*, have the suitable properties and correspond to the previous ones,

$m\backslash M$	A	B	$\sum$
a	60	0	60
b	10	30	40
$\sum$	70	30	100

Table 1bis

$m\backslash M$	A	B	C	$\sum$
a	37	23	0	60
b	0	10	30	40
$\sum$	37	33	30	100

Table 2bis

$m\backslash M$	A	B	C	D	$\sum$
a	37	0	20	3	60
b	0	33	0	7	40
$\sum$	37	33	20	10	100

Table 3bis

The χ^2 values of these tables, we point out with $\chi^2_{\max}$, are:

$$\chi^2_{\max\,1} = 64.29,\ \chi^2_{\max\,2} = 70.96 \text{ and } \chi^2_{\max\,3} = 90.25 \tag{1.4}$$

will be used to normalize the χ^2 indices reported in 1.1 to obtain the corresponding coefficients of contingency A^2 that we propose, obviously expressed by:

$$A^2 = \frac{\chi^2}{\chi^2_{\max}} \tag{1.5}$$

which assumes the following values:

$$A_1^2 = .198 \ , \ A_2^2 = .222 \text{ and } A_3^2 = .207. \tag{1.6}$$

We pick up the results 1.2, 1.3 and 1.6 in the following table for a simpler comparison:

$Tab.\ n.$	V^2	T^2	A^2
1	.090	.127	.198
2	.111	.112	.222
3	.207	.108	.207

We see that *Cramer*'s and *Tschuprov*'s indices assume smaller values to those of the proposed index because they are related to values of χ^2 which are necessarily higher.

Now we will give the demonstration of the mathematical legitimacy of the proposed procedure for $two-by-two$ tables and introduce a compact expression that directly furnishes the coefficient of contingency for these tables.

2 The two-by-two case

2.1 The legitimacy of the procedure

We show that the proposed procedure is mathematically valid for $two-by-two$ contingency tables.

Starting from the following generic table:

$m\backslash M$	A	B	$total$
a	x	$s-x$	s
b	$t-x$	$n+x-s-t$	$n-s$
$total$	t	$n-t$	n

(2.1)

the related χ^2 expression is:

$$\chi^2 = \frac{n\left(nx-st\right)^2}{st\left(n-s\right)\left(n-t\right)} \tag{2.2}$$

which is a parabolic function in x and, therefore, has two relative maxima at the extremities of its interval of definition.

Let us consider the Class of Fréchet of 2.1, we hypothesize that the greatest of the two relative maxima is in correspondence of the element of the class which presents a value of x, that we indicate as x^* defined as

$$x^* = \min\left[\max\left(s, n-s\right), \max\left(t, n-t\right)\right] \tag{2.3}$$

in the cell corresponding to the two greater among the marginal frequencies of line and column, as shown in the following table, where, for simplicity of writing we put $\max[s, n-s] = [s, n-s]^*$ and $\max[t, n-t] = [t, n-t]^*$:

$m \backslash M$	A	B	$total$
a	x^*	$[s, n-s]^* - x^*$	$[s, n-s]^*$
b	$[t, n-t]^* - x^*$	$n + x^* - [s, n-s]^* - [t, n-t]^*$	$n - [s, n-s]^*$
$total$	$[t, n-t]^*$	$n - [t, n-t]^*$	n

(2.4)

whose χ^2 index, we indicate with $\chi^2(x^*)$, is obtained by:

$$\chi^2(x^*) = \frac{n\left(|n-2s|\,|n-2t| + n\left(||n-2s| - |n-2t|| - n\right)\right)^2}{16st(n-s)(n-t)}. \tag{2.5}$$

The minimum value that x may assume in 2.4, say x_*, is given by:

$$x_* = \max[t, n-t] + \max[s, n-s] - n = \frac{1}{2}\left(|2t-n| + |2s-n|\right) \ . \tag{2.6}$$

and the corresponding value of χ^2, by:

$$\chi^2(x_*) = \frac{n\left(n\,|2s-n| + n\,|2t-n| - 2st\right)^2}{4st(n-s)(n-t)}$$

we will show that

$$\chi^2(x^*) - \chi^2(x_*) \geqslant 0. \tag{2.7}$$

The 2.7 can assume the form:

$$\frac{n\left(n^2 - |n-2s|\,|n-2t|\right)^2 - 4\left(n\,||n-2s| - n\,|n-2t||\right)^2}{16st(n-s)(n-t)} \geqslant 0$$

to show the validity of our affirmation it is sufficient to verify the non-negativity of the quantity:

$$\left(n^2 - |n-2s|\,|n-2t|\right)^2 - 4\left(n\,||n-2s| - n\,|n-2t||\right)^2$$

Such quantity may be rewritten this way:

$$n^4 + 6n^2\,|n-2s|\,|n-2t| + |n-2s|^2\,|n-2t|^2 + 4n\left(|n-2s|^2 + |n-2t|^2\right)$$

which certainly gives non negative results for any n, s and t, as it was expected.

2.2 A compact expression for the index A

A compact expression for the proposed index for a $two - by - two$ table simply follows from 1.5, 2.2 and 2.5:

$$A^2 = \frac{16\left(Nx - fg\right)^2}{\left(|2f-N|\,|2g-N| + N\,||2f-N| - |2g-N| - N^2|\right)^2} \tag{2.8}$$

or more simply:

$$A = \frac{4\left(Nx - fg\right)}{\left(|2f-N|\,|2g-N| + N\,||2f-N| - |2g-N| - N^2|\right)} \tag{2.9}$$

3 Remarks for higher dimension tables

3.1 Remark 1

The variability of the function:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{f_{ij}^2}{f_{i.}f_{.j}} - 1\right) \tag{3.1}$$

defined on the Class of Fréchet with values in $R^+ \cup \{0\}$, depends on the ratios $\frac{f_{ij}^2}{f_{i.}f_{.j}}$.

We observe that

$$\frac{f_{ij}^2}{f_{i.}f_{.j}} \leq 1$$

because f_{ij} can overcome neither $f_{i.}$ nor $f_{.j}$, obviously:

$$\frac{f_{ij}^2}{f_{i.}f_{.j}} = 1 \text{ if } f_{ij} = f_{i.} = f_{.j} \tag{3.2}$$

and

$$\frac{f_{ij}^2}{f_{i.}f_{.j}} < 1 \text{ otherwise}$$

as it is generally the case.

We can obtain the maximum value for expression 3.1 when condition 3.2 is verified in $r = c$ case, since we encounter r non null ratios, it is:

$$\chi^2_{\max,r=c} = n\,(r-1)$$

otherwise:

$$\chi^2_{\max,r\neq c} < n\,(r-1)\,.$$

In general, ratio $\frac{f_{ij}^2}{f_{i.}f_{.j}}$ draws near the unity when f_{ij} extends to $\min\,[f_{i.}, f_{.j}]$.

The proposed procedure gives the generic element of a Class of Fréchet as the highest number of ratios with the greatest numerator which is compatible with its marginal frequencies, we may then argue that it furnishes the maximum value of χ^2 for that class.

3.2 Remark 2

If we develop algebraically the expression of χ^2 with an arbitrary number of degrees of freedom we can obtain a ratio having at the denominator the product of all raw and marginal frequencies and at the numerator a quadratic form in so many variables as the number of degrees of freedom of the table.

If we, for instance, consider the following 2×3 table:

$m\backslash M$	A	B	C	$total$
a	x	y	$s-x-y$	s
b	$f-x$	$g-y$	$n-f-g-s+x+y$	$n-s$
$total$	f	g	$n-f-g$	n

with the assumptions:

$$\begin{array}{l} s \geqslant n - s \\ f \geqslant g \geqslant n - f - g \\ s - x \geqslant n - s \end{array}$$

we note that the varying part of χ^2 is:

$$x^2\left(gn^2 - g^2n\right) - 2fgnsx - 2fgnsy + 2fgnxy + y^2\left(fn^2 - f^2n\right)$$

It represents a parabolic surface with four relative maxima among which the absolute maximum, with certainty, is obtained for:

$$x = \min\left[s, t\right]$$

and

$$y = \min\left[f - x, g\right]$$

which represent the highest frequencies for the chosen cells, this is very easy to verify even if tedious.

This way of reasoning cannot be spent as a procedure of mathematical induction, but we believe that it can be extended to the generic $h*k$ dimensional case. Besides, we believe that there is also a statistical way of reasoning: If we accept the idea that the maximum contingency table ought to belong to the Class of Fréchet of the table under analysis, any procedure that takes into account the real dimension and marginal frequencies of the table has to be preferred to the Cramèr's and to the Tschuprov's solutions.

4 A manual procedure toward $\chi^2_{\max}$

From Table 3bis we can reach the maximum contingency table in three steps equal to the number of *degrees of freedom* of the table:

$m\backslash M$	A	B	C	D	$total$
a	30	20	3	7	60
b	7	13	17	3	40
$total$	37	33	20	10	100

(4.1)

the couple of maximum row and column frequencies is $(60, 37)$, we put the minimum between the coordinates of the couple in their crossing cell and get:

$m\backslash M$	A	B	C	D	$total$	$rest$
a	37				60	23
b					40	40
$total$	37	33	20	10	100	
$rest$	0	33	20	10		

(4.2)

We choose the second couple of maximum row and column rest frequencies, it is $(40, 33)$, and put the minimum coordinate in the crossing cell:

$m\backslash M$	A	B	C	D	$total$	$rest$
a	37				60	23
b		33			40	7
$total$	37	33	20	10	100	
$rest$	0	0	20	10		

(4.3)

For the third time we repeat the assignment of the highest frequency possible in a cell, it is the minimum of the coordinates of the couple $(23, 20)$, and have:

$m\backslash M$	A	B	C	D	$total$	$rest$
a	37		20		60	3
b		33			40	7
$total$	37	33	20	10	100	
$rest$	0	0	0	0		

(4.4)

We can complete the assignment of frequencies to the cells *by subtraction* and finally get the maximum contingency table:

$m\backslash M$	A	B	C	D	$total$	$rest$
a	37	0	20	3	60	0
b	0	33	0	7	40	0
$total$	37	33	20	10	100	
$rest$	0	0	0	0		

(4.5)

5 Some final comments

The aim of this work is to give social researchers a coefficient of contingency based on the χ^2 index more coherent with the structure of the population under study than Cramer's and Tschuprov's.

Some other hypotheses may be argued on the idea of maximum contingency table, to get a maximum value of χ^2 for normalization, based on a particular function chosen from those that can be established between the variables that is concretized in the individualization of a particular element of the Class of Fréchet.

The proposed procedure gives the highest value of χ^2 associated to the Class of Fréchet in univocal, simple and fast way, the procedure of Diaconis and Efron has this value as upper limit which can be reached only by chance and necessarily though the use of a computer.

6 A computer program for $\chi^2_{\max}$

This program[2], written in the language of MATLAB, gives the value of $\chi^2_{\max}$ introducing parameters R, number of rows, C, number of columns:

[2]This program has been developed by Dr. Luigi Arpaia.

```
function matr=contingency(row,col)
% Construction maximum contingency table
row=row';
col=col';
R=length(row);
C=length(col);
matr=zeros(R,C);
range=(R-1)*(C-1);
for i=1:range
[Srow,Irow]=sort(row);
[Scol,Icol]=sort(col);
if Srow(R)¿Scol(C)
matr(Irow(R),Icol(C))=Scol(C);
row(Irow(R))=abs(Srow(R)-Scol(C));
col(Icol(C))=0;
else
matr(Irow(R),Icol(C))=Srow(R);
col(Icol(C))=abs(Srow(R)-Scol(C));
row(Irow(R))=0;
end
end
```

References

[1] Bonferroni, C.E. and Brambilla, F. (1941): Studi sulla correlazione e sulla connessione, Istituto di Statistica, Università Bocconi di Milano, Milano.

[2] Castellano, V. (1956): Contributi alla teoria della correlazione e della connessione tra due variabili, *Metron*, **18**.

[3] Diaconis, P. and Efron, B. (1985): Testing for indipendence in a two-way table: new interpretation of the chi-square statistic. *The Annals of Statistics*, **13**, 845-874.

[4] Diaconis, P. and Gangolli, A. (1995): Rectangular arrays with fixed margins. In D. Aldous, P. Diaconis, J. Spencer, and J.M. Steele (Eds.): *Discrete Probability and Algorithms*, New York: Springer.

[5] Herzel, A. (1968): Alcuni problemi di massimi e minimi condizionati di interesse statistic, M*etron*, **27**, 1-2.

[6] Holmes, R.B. and Jones, L.K. (1996): On uniform generation of two-way tables with fixed margins and the conditional volume test of Diaconis and Efron. *The Annals of Statistics*, **24**, 64-68.

[7] Landenna, G. (1956): La dissomiglianza. *Statistica*, **1**.

[8] Landenna, G. (1957): Osservazioni sulla connessione, *Statistica*, **28**.

[9] Leti, G. (1967): La distribuzione delle tabelle della Classe di Fréchet. *Metron*, **26**.

[10] Leti, G. (1970): La distribuzione delle tabelle della Classe di Fréchet. *Metron*, **28**.

[11] Leti, G. (1976): Sulle distribuzioni massimanti e minimanti degli indici di omogeneità e eterogeneità delle distribuzioni doppie, Studi in onore di Paolo Fortunati. Università "La Sapienza" di Roma, Roma.