ONGOING RESEARCH
RECHERCHE EN COURS

PERCENTAGE OF MAXIMUM DEVIATION FROM INDEPENDENCE (PEM): COMMENT ON LEFEVRE & CHAMPELY'S "ANALYSE D'UN TABLEAU DE CONTINGENCE" ARTICLE

par

Philippe Cibois
(University of Versailles-St-Quentin ; phcibois@wanadoo.fr)

The article by Lefèvre and Champely (2009) is of major interest because it proposes using a bootstrap procedure to construct confidence intervals for measuring the overall liaison between rows and columns in a contingency table, as well as at the level of cells. They use Cramer's V (1946) for the overall link and the PEM (Cibois 1993) for local liaison.

This use of the bootstrap procedure is quite convincing because it helps construct confidence intervals in cases where one does not know the probability distribution of an indicator (as with the case of the PEM). The contribution of Lefèvre and Champely is therefore very important since it will complement the current use of the PEM.

Before specifying that it is possible to improve the Cramer's V using the overall PEM, we will describe the logic of the PEM by considering the local situation for the example of Lefèvre and Champely which crosses the age of individuals with their physical activity or sport. The purpose of the local PEM is to give an indicator of the strength of the liaison of attraction (or repulsion).

Table 1:

N=	No Practice	Less than once a week	Once per week or more	Total row
Age 50-54	56	7	34	97
Age 55-59	35	7	40	82
Age 60-65	74	8	32	114
Total column	165	22	106	293

As an example, examine the cell at the intersection of third row and first column (not practicing a sport or physical activity and having 60 to 65 years) where the observed frequency is 74.

If there were independence, the theoretical number of subjects would be equal to the product of the margins, divided by the total : 114 x 165 / 293 = 64.20

The deviation from independence is equal to the observed difference between observed and theoretical subjects: 74 - 64.20 = 9.80

To find out if deviation is weak or strong, we must see what is the maximum that can be in this cell when taking into account the margins that serve as the reference universe. In this case, the 165 non-practitioners may not be in the age group 60 to 65 years because there are only 114 individuals in this age group. On the other hand, the 114 in this age group may all be non practicers. The lower margins of the two is therefore the maximum number of subjects.

In this maximum case, the deviation from independence would therefore be: 114 - 64.20 = 49.80, and since it is a maximum, this value can serve as a reference.

The observed deviation compared to the deviation in the maximum case is equal to : 9.80 / 49.80 = 0.197 or 19.7%

As a general rule, if n_{ij} is the observed frequency, n_i and n_j the margins and n the total, and the theoretical number $t_{ij} = n_i \times n_j / n$, then the local PEM, PEM_{ij}, can be defined as follows :
$PEM_{ij} = ((n_{ij} – t_{ij}) / (\min(n_i, n_j) – t_{ij})) \times 100$

Now take an example of repulsion in the same column (no practice), but with the age group 55-59 years.

The theoretical number under the assumption of independence is: 82 x 165 / 293 = 46.18

The deviation from independence is 35 - 46.18 = -11.18, and since this deviation is negative, there is a negative connection, a repulsion, a shortfall in relation to independence. The situation corresponding to the maximum of this deficit would be if there would be no one observed in this case. The deficit would be equal to 0 minus the theoretical number 46.18 or -46.18.

The observed deviation compared to the maximum represents: -11.18 / -46.18 = 0.242 or 24.2%. As this PEM reflects a repulsion, by convention one gives it a negative sign to distinguish it from an PEM measuring an attraction.

As a general rule, we have : $PEM_{ij} = ((n_{ij} - t_{ij}) / 0 - t_{ij})) \times 100$ (it will be seen that this case it is the ratio deviation / theoretical number).

FROM LOCAL PEM TO OVERALL PEM

Regardless of the bootstrap method proposed by Lefèvre and Champely, it is possible to know if a PEM associated with a deviation is significant by combining all other rows into one, and all other columns of the table into a single one too.

We have the following 2 rows and 2 columns table with one degree of freedom for the first PEM calculated:

Table 2:

N=	No Practice	Other Practice	Total row
Age 60-65	74	40	114
Age else	91	88	179
Total column	165	128	293

If such a table has a significant chi-square, as is the case here (chi-square = 5.6, 1 degree of freedom, p = 0.017), its PEM (obviously the same as in Table 1 since the size of the reference cell, margins and total are the same) is deemed significant because it is derived from a significant table.

One can check that in the original table, the same four significant cases reported by the bootstrap procedure are significant (at the 5% level).

We will now develop a generalization of the PEM for the entire table by studying first the PEM of Table 2 with to 2 rows and 2 columns

Table 3:

PEM=	No Practice	Other Practice
Age 60-65	19.7	-19.7
Age else	-19.7	19.7

The diagonal symmetry reflects the diagonal symmetry of deviations from independence:

Table 4:

Deviation	No Practice	Other Practice
Age 60-65	9.8	-9.8
Age else	-9.8	9.8

The principle of generalization is to take into account the sum of positive deviations from independence, and then refere it to a situation where we have maximized the liaison on the diagonal of the table where are the attractions, by loading the diagonal as much as possible. The reasoning is very close to the case of the PEM for a single cell because in the reference case (60-65 years without sports), we can only put the smaller of the two margins. Thus the following table maximizes the attraction in the direction of observation.

Table 5:

N=	No Practice	Other Practice	Total row
Age 60-65	114	0	114
Age else	51	128	179
Total column	165	128	293

The two diagonal cells with frequecies 114 and 128 have the same maximum positive deviation from independence which is equal to 49.80 (114 - 128 = 64.20 -

78.20). The sum of deviations from independence is twice the deviation of each local PEM.

The same is true in the case of observed data and the overall PEM is the same as the local PEM in this case of a 2 x 2 table.

When the table not has only one degree of freedom, this is no longer the case, but you can use the same principle of maximizing the diagonal where the attractions are. In the original Table 1, we see that the SW-NE diagonal includes the attractions. The numerator of the overall PEM will be the sum of all positive differences to independence as shown below.

Table 6:

Observed data (table 1)

	NPRA	-1/S	1/S+	TOT.
5054	56	7	34	97
5559	35	7	40	82
6065	74	8	32	114
TOT.	165	22	106	293

Table 7:

Theoretical data

	NPRA	-1/S	1/S+	TOT.
5054	54.6	7.3	35.1	97.0
5559	46.2	6.2	29.7	82.0
6065	64.2	8.6	41.2	114.0
TOT.	165.0	22.0	106.0	293.0

Table 8:

Deviations from independance

	NPRA	-1/S	1/S+
5054	1.4	-0.3	-1.1
5559	-11.2	0.8	10.3
6065	9.8	-0.6	-9.2

Sum of positive deviations = 22.35

To form the denominator of the PEM, it is necessary to maximize the diagonal of Table 6: The algorithm is as follows: one starts with the reference cell at the bottom left, and there you put the smaller of the two margins.

Table 9:

	NPRA-	1/S	1/S+	TOT.
5054				97
5559				82

6065	**114**	**0**	**0**	**114**
TOT.	**165**	**22**	**106**	**293**

As this number is the margin row, the other two cells in the row can only be zero. However, there are still 165 - 114 = 51 individuals that must be put in the same column in the nearest row because it is consistent with the margin row of 82. The individuals in this column are entirely allocated in Table 10.

Table 10:

	NPRA-	**1/S**	**1/S+**	**TOT.**
5054	**0**			**97**
5559	**51**			**82**
6065	**114**	**0**	**0**	**114**
TOT.	**165**	**22**	**106**	**293**

As on the second row there are still 82 - 51 = 31 missing individual to be put on the second row, we can put only 22 in the second column (which corresponds to the margin in the second column) and the remainder in the third (Table 11).

Table 11:

	NPRA-	**1/S**	**1/S+**	**TOT.**
5054	**0**			**97**
5559	**51**	**22**	**9**	**82**
6065	**114**	**0**	**0**	**114**
TOT.	**165**	**22**	**106**	**293**

Table 11

The first row remains to becompleted where the 97 remaining individuals must be in the third column (Table 12).

Table 12:

	NPRA-	**1/S**	**1/S+**	**TOT.**
5054	**0**	**0**	**97**	**97**
5559	**51**	**22**	**9**	**82**
6065	**114**	**0**	**0**	**114**
TOT.	**165**	**22**	**106**	**293**

One can verify that we would come to the same result if we started by using the cell in the upper right. A presentation of the algorithm figures will be find in appendix.

Since this table is the maximum, the sum of positive deviation from independence may be used as the denominator for the overall PEM.

Table 13:

```
Deviancy from independance in the case of maximum
       NPRA  -1/S  1/S+
5054 -54.6  -7.3  61.9
5559   4.8  15.8 -20.7
6065  49.8  -8.6 -41.2
Sum of positive deviations =  132.38
```

The overall PEM is: 22,35 / 132,38 x 100 = 16,9%

Cramer's V for the same data, if we follow to the letter Cramer's publication (1946), where *q* is the smallest number of *r* rows or *s* columns, where *n* is the total, $V = \chi^2 / n (q-1)$, where chi-square is here equal to 10.11, n = 293 and q = 3.

$$V = 10.11 / 293 \times 2 = 0.017.$$

This expression varies between 0 and 1. The numerator is the observed chi-square, the denominator the chi-square that there will be at maximum : "the upper limit 1 is attained when and only when each row (when r>=s) or each column (when r <=s) contains one single element different from zero" (Cramer 1946: 443).

This is exactly the proportion of chi-square observed to chi-square maximum in the case of the maximum link and we can therefore compare it in percentage with the overall PEM

Overall PEM represents 16.9% of the maximum.

Cramer's V represents 1.7% of the maximum.

Cramer's index is very pessimistic because its maximum is very particular: it assumes that all data are grouped in a few cells; it does not take into account the value of margins.

We can also be consistent with Cramer's logic and take as the maximum chi-square the maximized Table 12. In this case, Chi-square / Chi-square max = 11.1 / 315.23 = 0.035 or 3.5% as a percentage.

The overall PEM is more realistic than Cramer's V because it takes into account the observed margins and because it takes into account the deviation from independence and not the contribution of the cell to Chi-square whose presence in Cramer's V is simply justified by the fact that it follows a Chi-square law, a requirement which can be overcome by a bootstrap procedure, as in the case of local PEM.

However, we notice that the algorithm requires an order on the rows and columns to have a single result. This is the case in the present example where rows are ordered by age and columns by the intensity of practice. In the general case, it is always possible to find an order for the rows and columns using the first factor of a correspondence analysis which suggested such an order (Benzécri 1976: 193).

REFERENCES

Benzécri, Jean-Paul, 1976, *L'analyse des données, 2, L'analyse des correspondances*, Paris, Dunod.

Lefèvre, Brice et Champely, Stéphane, 2009, Méthodes statistiques globales et locales d'analyse d'un tableau de contingence par les tailles d'effet et leurs intervalles de confiance, *Bulletin de Méthodologie Sociologique*, 104.

Cibois, Philippe, 1993, Le PEM, pourcentage de l'écart maximum: un indice de liaison entre modalités d'un tableau de contingence, *Bulletin de Méthodologie Sociologique*, 40, 43-63.

Cramer, Harald, *Mathematical Methods of Statistics*, Princeton, PUP, 1946.

APPENDIX

Algorithm to maximize the diagonal of the table (in this example the version for the first diagonal where i and j are initialized to 1), i is for rows (1 to ImaxRow), j for columns (1 to JmaxCol), MarginRow is initialized with the contents of the row margins, MarginCol with the column margins, TabMax is the terminal matrix.

```
Strt:
         If i > ImaxRow Or j > JmaxCol Then GoTo EndTab
         If MarginRow (i) > MarginCol (j) Then
                  R = MarginCol (j)
                  TabMax (i, j) = R
          Else
                  R = MarginRow (i)
                  TabMax (i, j) = R
          End If
          MarginRow (i) = MarginRow (i) - R
          MarginCol (j) = MarginCol (j) - R
          If MarginRow (i) = 0 Then i = i + 1: GoTo Strt
          If MarginCol (j) = 0 Then j = j + 1: GoTo Strt
EndTab:

         ----------------------------------------
```