Le PEM, pourcentage de l'écart maximum
A propos de l'article de : Brice Lefèvre et Stéphane Champely "Méthodes statistiques globales et locales d'analyse d'un tableau de contingence par les tailles d'effet et leurs intervalles de confiance"

Philippe Cibois

Université de Versailles-St-Quentin

phcibois@wanadoo.fr

L'article de Lefèvre et Champely[1] a le grand intérêt de proposer, en utilisant une procédure bootstrap, des intervalles de confiance pour la mesure de la liaison globale entre lignes et colonnes dans un tableau de contingence ainsi qu'au niveau de la case. Ils utilisent le V de Cramér (1946) pour la liaison globale et le PEM (Cibois 1993) pour la liaison locale.

Cette utilisation de la procédure Bootstrap est tout à fait convaincante car elle permet dans les cas où l'on ne connait pas la distribution théorique d'un indicateur (et c'est le cas du PEM) de se doter d'un intervalle de confiance. L'apport de Lefèvre et Champely est donc ici très important car il va permettre de compléter la pratique actuelle du PEM.

Avant de préciser qu'il est tout à fait possible d'améliorer le V de Cramér en utilisant le PEM global, nous indiquerons la logique du PEM en considérant le cas local à partir de l'exemple de Lefèvre et Champely qui croise l'âge d'individus avec leur pratique d'une activité physique ou sportive. Le but du PEM local est de se donner un indicateur de la force de la liaison d'attraction (ou de répulsion).

N=	Non Pratiquant	Moins d'une fois par sem	Une fois par sem. & +	Total en ligne
50-54 ans	56	7	34	97
55-59 ans	35	7	40	82
60-65 ans	74	8	32	114
Total en colonne	165	22	106	293

Tableau 1

Prenons comme exemple d'attraction la case à l'intersection de la 3e ligne et de la première colonne (non pratiquant d'une activité physique et sportive et ayant de 60 à 65 ans) où l'effectif observé est de 74.

S'il y avait indépendance, l'effectif théorique serait égal au produit des marges divisé par le total : 114 x 165 / 293 = 64,20

L'écart à l'indépendance est donc égal à la différence entre effectif observé et effectif théorique : 74 – 64,20 = 9,80

[1] Brice lefèvre et Stéphane Champely, "Méthodes statistiques globales et locales d'analyse d'un tableau de contingence par les tailles d'effet leurs intervalles de confiance", *Bulletin de Méthodologie Sociologique*, n°103, July 2009.

Pour savoir si cet écart est faible ou fort, il faut voir quel est l'effectif maximum qui pourrait être dans cette case en tenant compte des marges qui servent d'univers de référence. Dans le cas présent, les 165 non pratiquants ne peuvent être dans la tranche d'âge de 60 à 65 ans puisqu'ils il n'y a que 114 individus dans cette tranche d'âge. Par contre les 114 de cette tranche d'âge peuvent tous être non pratiquant. La plus faible des deux marges est donc l'effectif maximum.

Dans le cas du maximum, l'écart à l'indépendance serait donc de : 114 – 64,20 = 49,80 et, puisque c'est un maximum, cette valeur peut servir de référence.

L'écart observé représente par rapport à l'écart dans le cas du maximum une proportion de : 9,80 / 49,80 = 0,197 soit 19,7%

D'une manière générale si n_{ij} est l'effectif observé, n_i et n_j les marges et n l'effectif total, t_{ij} l'effectif théorique = n_i x n_j /n, le PEM local PEM_{ij} se définit comme suit : $PEM_{ij} = ((n_{ij} – t_{ij}) / (\min(n_i, n_j) – t_{ij})) \times 100$

Soit maintenant un exemple de répulsion dans la même colonne (non pratiquant) mais avec la tranche d'âge 55-59 ans.

L'effectif théorique sous l'hypothèse d'indépendance est de :
82 x 165 / 293 = 46,18

L'écart à l'indépendance est de 35 – 46,18 = -11,18 et, comme cet écart est négatif, il y a donc une liaison négative, une répulsion, un déficit par rapport à l'indépendance. La situation qui correspondrait au maximum de ce déficit serait le cas où il n'y aurait personne d'observé dans cette case. Le déficit serait alors égal à 0 moins l'effectif théorique de 46,18 soit -46,18.

L'écart observé représente par rapport à l'écart dans le cas du maximum une proportion de : -11,18 / -46,18 = 0,242 soit 24,2% : comme ce PEM traduit une répulsion, par convention on lui donne un signe négatif pour le différencier d'un PEM mesurant une attraction.

D'une manière générale on écrira : $PEM_{ij} = ((n_{ij} – t_{ij}) / 0 – t_{ij})) \times 100$ (on remarquera qu'il s'agit dans ce cas du rapport écart / théorique.

Du PEM local au PEM global

Indépendamment de la méthode bootstrap proposée par Lefèvre et Champely, il est possible de savoir si un PEM associé à un écart est significatif en regroupant toutes les autres lignes du tableau dans une seule et toutes les autres colonnes du tableau dans une seule également

On a alors le tableau suivant à 2 lignes et 2 colonnes et à 1 degré de liberté pour le premier PEM calculé :

N=	Non Pratiquant	Autre Pratique	Total en ligne
60-65 ans	74	40	114
Autre âge	91	88	179
Total en colonne	165	128	293

Tableau 2

Si un tel tableau a un khideux significatif, et c'est le cas ici (khideux = 5,6, 1 degré de liberté, p= 0,017) son PEM, évidemment le même que dans le tableau 1 puisque l'effectif de la case de référence, les marges et le total sont les mêmes, est réputé significatif puisqu'il est dérivé d'un tableau significatif.

On peut vérifier que dans le tableau d'origine, sont significatifs (au seuil de 5%), les mêmes quatre cases déclarées significatives par la procédure bootstrap.

Nous allons maintenant mettre au point une procédure de généralisation du PEM à l'ensemble d'un tableau en étudiant d'abord les PEM du tableau 2 à 2 lignes et 2 colonnes

PEM=	Non Pratiquant	Autre Pratique
60-65 ans	19.7	-19.7
Autre âge	-19.7	19.7

Tableau 3

Cette symétrie diagonale n'est que le reflet de la symétrie diagonale des écarts à l'indépendance :

Ecarts à l'indépendance	Non Pratiquant	Autre Pratique
60-65 ans	9.8	-9.8
Autre âge	-9.8	9.8

Tableau 4

Le principe de la généralisation est de prendre en compte la somme des écarts observés positifs à l'indépendance et de la rapporter à une situation où l'on a maximisé la liaison en chargeant le plus qu'il est possible la diagonale du tableau où se font les attractions. Le raisonnement est très proche du cas du PEM pour une case car dans la case de référence (60-65 ans sans pratique sportive), on ne peut mettre au maximum que la plus petite des deux marges. On a alors le tableau suivant qui maximise l'attraction dans le sens de l'observation.

N=	Non Pratiquant	Autre Pratique	Total en ligne
60-65 ans	114	0	114
Autre âge	51	128	179
Total en colonne	165	128	293

Tableau 5

Les deux cases diagonales d'effectif 114 et 128 ont le même écart maximum positif à l'indépendance qui est égal à 49,80 (114 – 64,20 = 128 – 78,20). La somme des écarts à l'indépendance est le double de l'écart de chaque PEM local.

Il en est de même dans le cas des données observées et le PEM global est donc le même que le PEM local dans ce cas du tableau 2 x 2.

Quand le tableau n'est plus à un degré de liberté, il n'en est plus de même mais on peut garder le même principe qui consiste à maximiser la diagonale où se situe les attractions. Soit le tableau d'origine : on voit que c'est la diagonale SW-NE qui porte les attractions. Le numérateur du PEM global sera constitué par la somme de tous les écarts positifs à l'indépendance comme montré ci-dessous.

Données observées (tableau 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

Tableau 6

Effectif théoriques

	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

Tableau 7

Ecarts à l'indépendance

	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

Somme des écarts positifs = 22.35

Tableau 8

Pour constituer le dénominateur du PEM, il faut maximiser la diagonale du tableau 6 : l'algorithme est le suivant : on part de la case de référence en bas à gauche et l'on y met la plus petite des deux marges.

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

Tableau 9

Comme cet effectif correspond à la marge ligne, les deux autres cases de la ligne ne peuvent être que nulles, par contre, il reste 165 – 114 = 51 individus que l'on doit mettre dans la même colonne, dans la ligne la plus proche car c'est compatible avec la marge ligne de 82, l'effectif de cette colonne est ainsi entièrement réparti dans le tableau 10.

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

Tableau 10

Comme sur la deuxième ligne il manque encore 82 – 51 = 31 à répartir, sur la deuxième ligne, nous ne pouvons en mettre que 22 en 2e colonne (ce qui correspond à la marge de la deuxième colonne) et le reste sur la troisième (tableau 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

Tableau 11

Il reste à compléter la première ligne où les 97 ne peuvent être qu'en 3e colonne (tableau 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

Tableau 12

On vérifiera que l'on serait arrivé au même résultat si on était parti de la case en haut à droite. On trouvera en annexe une présentation de l'algorithme.

Puisque ce tableau correspond au maximum, la somme des écarts positifs à l'indépendance pourra servir de dénominateur du PEM global.

Ecarts à l'indépendance dans le cas du 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

Somme des écarts positifs = 132.38

Tableau 13

Le PEM global est de 22,35 / 132,38 x 100 = 16,9%

Le V de Cramér pour les mêmes données est, si nous prenons à la lettre le texte de Cramér dans la publication de référence (1946), *q* étant le plus petit nombre de *r* lignes ou de *s* colonnes, *n* étant l'effectif total, $V = \chi^2 / n(q-1)$ où le khideux est égal ici à 10,11, n=293 et q=3. V = 10,11 / 293 x 2 = 0,017.

Cette expression varie entre 0 et 1. Le numérateur est le khideux observé, le dénominateur le khideux qu'il y aurait au 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" (Cramér 1946 : 443).

C'est bien le rapport du khideux observé au khideux maximum dans le cas de la liaison maximum et on peut donc le comparer en pourcentage avec le PEM global

Le PEM global représente 16,9% du maximum

Le V de Cramér représente 1,7% du maximum

L'indice de Cramér est très pessimiste car son maximum est très particulier, il suppose que toutes les données soient regroupées sur quelques cases : il ne prend pas en compte la valeur des marges.

On peut aussi rester dans la logique de Cramér et prendre comme Khideux maximum, celui du tableau maximisé que nous avons utilisé : dans ce cas Khideux / Khideux max = 11,1 / 315,23 = 0,035 soit 3,5% en pourcentage.

Le PEM global est plus réaliste que le V de Cramér car il tient compte des marges observées et qu'il prend en compte l'écart à l'indépendance et non la contribution de la case au khideux dont la présence dans le V de Cramér est simplement justifiée par le fait qu'on suit ainsi une loi du khideux, exigence dont on peut s'affranchir par une procédure bootstrap comme dans le cas du PEM local.

On notera cependant que l'algorithme utilisé suppose un ordre sur les lignes et sur les colonnes pour avoir un résultat unique. C'est le cas dans l'exemple traité où les lignes sont ordonnées par l'âge et les colonnes par l'intensité de la pratique. Dans le cas général, il est toujours possible de trouver un ordre pour les lignes et les colonnes en utilisant le premier facteur d'une analyse des correspondances qui propose un tel ordre (Benzécri 1976 : 193).

Bibliographie

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*, 103, July 2009.

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.

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

Annexe

Algorithme de maximisation de la diagonale du tableau (ici à titre d'exemple la version pour la première diagonale ou i et j sont initialisé à 1), i indice les lignes (de 1 à ImaxRow), j les colonnes (de 1 à JmaxCol), MarginRow est chargé au départ avec le contenu de la marge des lignes, MarginCol avec celui des colonnes, TabMax est la matrice terminale

```
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:
```