

**Analyse statistique :
Étude des réalités par région,
par types de risques, vulnérabilités et facteurs de risque par milieu réalisée
par la FQCEDI et le RIFVEH**

Marc Bourdeau¹

Professeur de statistique, Université de Montréal

Résumé

Nous indiquons en premier lieu quelques clés pour l'interprétation des tableaux croisés à partir des résidus qu'on y calcule. Le résidu d'une cellule est l'écart normalisé entre ce qui est attendu comme contingent de la cellule sous hypothèse d'indépendance comparativement à celui qui est observé sur les données. Ce sont des écarts à l'indépendance calculés pour chaque cellule.

Nous présentons ensuite deux tableaux croisés tirés d'une enquête auprès d'intervenants dans les de risques dans quatre régions du Québec, où la question se pose de savoir s'il y a des différences selon les régions pour les (1) lieux de risques, et (2) des facteurs de risque. Sur ces deux tableaux, les clés d'analyse sont mises en action.

Pour le premier de ces tableaux, les types d'abus ou risques se présentent selon la même structure d'une région à l'autre. Dans le second des ces tableaux qui concerne les lieux d'abus, les incidences des fréquences diffèrent d'une région à l'autre. Nous présentons ces dernières constatations pour le milieu résidentiel.

Enfin, en annexe, nous rapportons l'ensemble des tableaux de contingence que nous avons examinés dans cette étude à l'exception de ceux expliqués dans le corps du texte. Leurs interprétations sont laissées au lecteur intéressé. Elles se construisent comme celles qui sont présentées dans le corps du texte.

1. Clés pour l'interprétation des tableaux croisés

Un tableau qui croise les modalités de deux variables, ou mesures, contient dans chaque cellule le contingent ou nombre des sujets qui ont une certaine modalité pour la première variable et une certaine modalité pour l'autre mesure. On parle aussi de tableaux de contingences à deux entrées qualitatives (catégoriques).

Les écarts à l'indépendance d'un tableau croisé sont décrits de la façon la plus parlante par les écarts normalisés entre ce qui observé et ce qui est attendu sous l'hypothèse de l'indépendance entre les deux mesures. Ce sont les *contributions* (r), dont la somme des carrés donne la valeur du Chi² du tableau.

¹ <mailto:Marc.Bourdeau@polymtl.ca> , <http://www.mgi.polymtl.ca/marc.bourdeau/Consultations> .

Plus ce Chi2 est grand et plus les écarts à l'indépendance sont élevés. On mesure la probabilité de dépassement de ce Chi2, notée p (« p-value »). De petites valeurs de la probabilité de dépassement renvoient à de grandes valeurs des contributions, un grand Chi2 au total.

De grandes contributions montrent que les écarts à l'indépendance des deux caractères sont importants. Plus cette probabilité est petite, plus on est forcé d'admettre que les écarts à l'indépendance ne seraient pas dus à des hasards de l'échantillonnage, mais à des différences significatives.

Pour un Chi2 significatif au seuil conventionnel ($p < 0,05$), on peut déterminer les cellules qui sont surreprésentées ou sous représentées.

Les grandes valeurs positives des contributions indiquent des surreprésentations : i.e. qu'on attendrait des effectifs plus petits s'il y avait indépendance des modalités des deux mesures correspondant à la cellule surreprésentée.

Les grandes valeurs négatives des contributions indiquent une sous représentation de la cellule : si les modalités correspondant à la cellule étaient indépendantes, on y trouverait un effectif plus grand.

Nous n'indiquerons pas quelles sont les grandeurs des contributions suffisantes pour les rendre significatives (en un sens statistique à préciser). L'examen se fera intuitivement, les tests n'apportant le plus souvent que des confirmations difficiles à valider.²

2. Les incidences des abus ou risques selon les régions et le type d'abus³

Tableau 1. Tableau croisant les incidences des abus selon le type d'abus et les 4 régions : 1, la région Chaudière-Appalaches; 2, Outaouais; 3, Chicoutimi; 4, Estrie. Suivi du tableau de contributions (dites aussi « r ») dont la somme des carrés constitue la valeur du Chi2 du tableau. Enfin la valeur du Chi2 et sa probabilité de dépassement (« p-value »).

	Incidence des abus							
Région	Phy	Sexuel	Psychol	Nég.	Droits	Fin.	Pouvoirs	Rangée
1	109	95	294	291	226	103	164	1282
2	119	110	283	323	245	96	183	1359
3	130	121	331	358	260	95	188	1483
4	113	99	297	315	218	102	161	1305
Colonne	471	425	1205	1287	949	396	696	5429

	Contributions (r)						
Région	Phy	Sexuel	Psychol	Nég.	Droits	Fin.	Pouvoirs
1	-0,2	-0,5	0,6	-0,7	0,1	1,0	0,0
2	0,1	0,4	-1,1	0,0	0,5	-0,3	0,7
3	0,1	0,5	0,1	0,3	0,0	-1,3	-0,2
4	0,0	-0,3	0,4	0,3	-0,7	0,7	-0,5

Chi2	p-value
,77	0,95

² Le lecteur intéressé pourra se rapporter aux grands traités sur les tableaux de contingence, ou encore à la référence de base suivante : Shelby J. Haberman (1973), « The analysis of residuals in cross-classified tables », *Biometrics* **20**, 205-221.

³ On admet dans cette section et la suivante que les règles d'interprétation du Chi2, de sa 'p-value' et des contributions, expliquées dans la section précédente, sont assimilées et mises en action.

Dans ce tableau, on admet l'indépendance des deux caractères (questions) car la probabilité de dépassement, le 'p-value', est bien supérieure au seuil critique de $p=0,05$. On admet l'hypothèse suivante : quelle que soit la région, les proportions des incidences des types d'abus semblent bien les mêmes.

Les contributions rapportées en deuxième partie du tableau, les écarts à l'indépendance donc, sont donc trop petites pour justifier le rejet d'une différence des incidences des types d'abus selon les régions.

En fait, dans 95% des cas où l'hypothèse d'indépendance est vraie, on obtient des écarts au moins aussi grands que ceux observés dans ce tableau. Rejeter l'hypothèse de l'indépendance des deux caractères avec d'aussi petites contributions entraîne un risque de se tromper dans 95% des cas.

Dans le cas de l'acceptation de l'hypothèse de l'indépendance des deux caractères, on peut justifier, sans craindre de gommer des différences qui donnent un sens aux variations selon les régions, d'agrèger les lignes du tableau et observer les proportions des incidences des types d'abus⁴.

Tableau 2. Le tableau des incidences des types d'abus sans tenir compte du croisement avec la variable 'région'. On note que trois des types d'abus sont plus fréquents que les autres, de deux à trois fois plus fréquents.

	Incidence des abus toutes régions confondues							Total rangée
	Phy	Sexuel	Psychol	Nég.	Droits	Fin.	Pouvoirs	
Nb obs.	471	425	1205	1287	949	396	696	5429
Freq. Rel. %	8,7	7,8	22,2	23,7	17,5	7,3	12,8	100

3. Les lieux des abus, incidences selon les régions

Tableau 3. Tableau croisant les incidences des fréquences des facteurs de risques d'abus et les régions pour le milieu résidentiel : 'P' pour peu ; 'S' pour souvent ; 'TS' pour très souvent. Suivi du tableau de contributions (dites aussi « r ») dont la somme des carrés constitue la valeur du Chi2 du tableau; enfin la valeur du Chi2 et sa probabilité de dépassement (« p-value »).

⁴ On pourrait aussi agréger les colonnes, mais cela n'aurait pas grand sens ici.

Tableau 3.

	Incidences des fréquences d'abus dans le milieu résidentiel			
Région	P	S	TS	Rg
1	306	204	31	541
2	293	139	67	499
3	286	229	95	610
4	280	181	67	528
Colonne	1165	753	260	2178

	Contributions (r)		
Région	P	S	T
1	1,0	1,2	-4,2
2	1,6	2,6	1,0
3	-2,2	1,2	2,6
4	-0,1	0,1	0,5

Chi2	P-value
43,5	0

Tout au contraire de la situation décrite à la section précédente, on rejette ici l'hypothèse de l'indépendance des deux caractères » (Tab. 3) : les incidences des fréquences des abus dans le milieu résidentiel est différent de région à région⁵.

C'est à l'examen des résidus qu'on trouve que dans la région 1 (Chaudière-Appalaches) on a une grande sous représentation de la fréquence TS, au contraire de la région 3 (Chicoutimi) qui, elle, présente une fréquence trop grande de TS, mais aussi une sous représentation marquée de la fréquence P.

Ce qu'il en est exactement demanderait de retourner aux données. C'est une des utilités de l'examen (des contributions) des tableaux croisés : pourquoi cette situation dans la région 1? Etc. Le lecteur impliqué dans ces études pourra formuler d'autres questions et hypothèses.

Notons enfin, que rien n'empêche de constituer des données regroupées (agrégées) par lignes ou colonnes. Ici c'est l'agrégation des lignes des divers milieux par région qui présente un intérêt. On obtient le Tab. 4 croisant les régions et les milieux d'abus comme facteurs de risque.

Un seul élément d'interprétation de ce tableau ou l'hypothèse de l'indépendance des deux caractères est rejetés de façon très significative (*p'* presque nul): relativement aux autres régions, les milieux de garde semblent avoir une bien moins grande représentation des facteurs de risque pour la région 1 (Chaudière-Appalaches) qu'on attendrait sous l'hypothèse d'indépendance des deux caractères, à l'inverse de la région 4 (Estrie).

⁵ On trouvera en annexe l'ensemble des tableaux de cette étude qui viennent compléter les tableaux du corps du texte : les analogues du Tab. 3, pour les autres milieux.

Tableau 4. Tableau croisant le regroupement (collapsus) des fréquences par milieu, et les régions. Suivi du tableau de contributions, enfin la valeur du Chi2 et sa probabilité de dépassement (« p-value »).

Région	Milieux					Rangée
	Familial	Scolaire	De garde	Résidentiel	De jour, etc.	
1	368	122	81	541	328	1440
2	392	159	197	499	252	1499
3	475	146	232	610	252	1715
4	403	175	306	528	280	1692
Colonne	1638	602	816	2178	1112	6346

Région	Contributions (R ou résidus)				
	Familial	Scolaire	De garde	Résidentiel	De jour, etc.
1	-0,2	-1,2	-7,7	2,1	4,8
2	0,3	1,4	0,3	-0,7	-0,7
3	1,5	-1,3	0,8	0,9	-2,8
4	-1,6	1,1	6,0	-2,2	-1,0

Chi2-12	P-value
149,20	0

Terminons en rapportant l'agrégation du tableau précédent selon les régions (Tab. 5). Donc les colonnes du tableau précédent ainsi que leurs pourcentages.

On y constate que les deux milieux les plus à risque sont le milieu résidentiel (plus du tiers), suivi du milieu familial (le quart) et des centres de jour (le cinquième environ). Les milieux de garde et scolaire sont les moins à risque.

Tableau 5. Les milieux en tant que facteurs de risque d'abus, indépendamment des régions.

	Incidence des milieux					Total rangée
	Familial	Scolaire	De garde	Résidentiel	De jour, etc.	
Nb. Observés	1638	602	816	2178	1112	6346
Fréq. Rel. %	25,8	9,5	12,9	34,3	17,5	100

Rien au fond n'empêche les agrégations, mais il faut se méfier de gommer par là des différences de comportements qui seraient dignes d'être explorés^o, à des fins d'ailleurs pas toujours transparentes...

Un exemple fictif mais très significatif des abus d'agrégation possibles est rapporté par l'économiste Ivar Ekeland, ancien président de l'Université Paris-Dauphine et rédacteur de la revue *Pour la Science*. Il écrivait en août 2005 : «Un nombre égal de femmes et d'hommes se présentent à l'embauche dans une industrie, mais il y a plus d'hommes embauchés que de femmes : l'industrie fait-elle de la discrimination ? Les chiffres sont là, apparemment indubitables. Et pourtant ! Supposons, pour simplifier, que toute l'industrie se réduise à deux usines. Pour la première, il y a 120 candidats, 100 hommes et 20

femmes ; l'usine embauche 74 hommes, soit 74 %, et 15 femmes, soit 75 %. Pour la seconde, il y a 280 candidats, 100 hommes et 180 femmes ; l'usine embauche 49 hommes, soit 49 %, et 95 femmes, soit 53 %. Si discrimination il y a, elle est en faveur des femmes, puisque, dans chacune des usines, une candidate a plus de chances d'être recrutée qu'un candidat. Et pourtant, à l'échelle de l'industrie, on a recruté davantage d'hommes que de femmes, 123 contre 110 (pour 200 candidats de chaque sexe). Ce qui semblait une discrimination, quand nous ne retenons que les données agrégées, est un problème de distribution quand nous examinons des données plus détaillées : les hommes répartissent également leurs candidatures sur les deux usines, alors que les femmes se concentrent sur la seconde [...]. » Ces résultats sont un des exemples du paradoxe dit de Simpson.