CHAPITRE 4-1
L'ANALYSE DES TABLEAUX DE CONTINGENCE

Plan

CHAPITRE 4-1
L'ANALYSE DES TABLEAUX DE CONTINGENCE

4-1.1 Introduction

4-1.1.1 QU'EST-CE QU'UN TABLEAU DE CONTINGENCE ?

Un tableau de contingence est une façon de présenter des données d'énumération (de comptage) d'individus classés en catégories. C'est donc dire qu'un tableau de contingence est déjà le résultat d'un traitement des données, puisque les individus (observations) ont dû être préalablement classés, puis comptés.

La forme du tableau est déterminée par le schème de classement utilisé. Le schème de classement doit être constitué de catégories mutuellement exclusives et il doit être exhaustif : dans le langage de la théorie des ensembles, les catégories doivent constituer une *partition* de l'univers, de sorte que chaque individu appartienne à une, et à une seule catégorie.

Les catégories sont définies au moyen d'une ou de plusieurs variables de classement (variables catégoriques), qui correspondent à autant d'attributs (dimensions) des individus. Chacun des individus observés est décrit aux fins du classement par les valeurs de ses attributs. Tous les individus ayant la même description (les mêmes valeurs d'attributs) sont comptés et leur nombre est inscrit dans la cellule correpondante du tableau de contingence qui en résulte. Le tableau de contingence a autant de dimensions qu'il y a de variables de classification et autant de cellules qu'il y a de combinaisons de catégories.

Voyons un petit exemple de construction d'un tableau de contingence à partir de données brutes. Soit le tableau d'observations suivant, où les observations ont déjà été classées par sexe, puis par couleur d'yeux :

	Prénom	**Sexe**	**Couleur d'yeux**
1	Bernadette	F	Bleus
6	Sophie	F	Bleus
4	Marie	F	Noirs
2	Jean-Pierre	M	Bleus
3	Marc	M	Noirs
5	Pierre	M	Noirs

On reconnaît dans le tableau qui précède la structure matricielle des données brutes. De ces données, on peut tirer un tableau de contingence de la couleur des yeux par sexe :

	Sexe		
Couleur d'yeux	**F**	**M**	**Total**
Bleux	2	1	3
Noirs	1	2	3
Total	3	3	6

Ce tableau de contingence a aussi une structure matricielle, mais il ne s'agit plus de données brutes : le tableau de contingence est le résultat d'un traitement. On remarque qu'un tableau de contingence à une seule variable catégorique est tout bonnement un tableau de fréquences.

Tableau 1 – Population active employée dans la Région métropolitaine de Montréal

Zone de résidence, selon le sexe et la profession, 1991

	Professions					
Zone de résidence	Directeurs, gérants, administra-teurs et assimilés	Profession-nels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**
Femmes						
Ville de Montréal	24 025	58 204	76 450	24 385	28 825	**211 889**
Reste de la CUM	22 575	42 207	70 003	14 065	17 435	**166 285**
Couronne Nord	16 785	31 699	63 491	11 975	18 630	**142 580**
Couronne Sud	18 365	35 674	65 290	10 485	19 380	**149 194**
Hors RMR	3 265	7 535	11 089	3 190	3 565	**28 644**
Total Femmes	**85 015**	**175 319**	**286 323**	**64 100**	**87 835**	**698 592**
Hommes						
Ville de Montréal	32 336	55 045	43 546	65 340	46 850	**243 117**
Reste de la CUM	39 146	39 920	37 819	46 173	28 749	**191 807**
Couronne Nord	33 287	27 560	31 170	62 852	29 329	**184 198**
Couronne Sud	36 006	32 464	30 600	58 778	29 721	**187 569**
Hors RMR	8 270	8 590	8 270	22 305	9 099	**56 534**
Total Hommes	**149 045**	**163 579**	**151 405**	**255 448**	**143 748**	**863 225**
Total hommes et femmes						
Ville de Montréal	56 361	113 249	119 996	89 725	75 675	**455 006**
Reste de la CUM	61 721	82 127	107 822	60 238	46 184	**358 092**
Couronne Nord	50 072	59 259	94 661	74 827	47 959	**326 778**
Couronne Sud	54 371	68 138	95 890	69 263	49 101	**336 763**
Hors RMR	11 535	16 125	19 359	25 495	12 664	**85 178**
Total H + F	**234 060**	**338 898**	**437 728**	**319 548**	**231 583**	**1 561 817**

Source : Statistique Canada, Recensement de 1991

Ce tableau de contingence a trois dimensions : la zone de résidence, la profession et le sexe. La zone de résidence comprend 5 catégories, la profession en comprend 5 et il y a 2 sexes. Le tableau contient donc $5 \times 5 \times 2 = 50$ cellules, auxquelles s'ajoutent les lignes et colonnes de totaux et sous-totaux.

4-1.1.2 L'ANALYSE DES TABLEAUX DE CONTINGENCE PARMI LES MÉTHODES D'ANALYSE MULTIVARIÉE

L'analyse multivariée, au sens large, désigne l'ensemble des méthodes d'analyse statistique qui traitent simultanément plus d'une variable. C'est à l'analyse multivariée que l'on recourt notamment pour

- mesurer le degré d'association entre deux ou plusieurs variables ;
- estimer les paramètres d'une relation entre deux ou plusieurs variables ;
- évaluer à quel point les différences entre deux ou plusieurs groupes d'observations sont significatives ;
- tenter de prédire à quel groupe appartient un individu, à partir de ses autres caractéristiques ;
- essayer de discerner une structure dans un ensemble de données.

Plusieurs techniques d'analyse multivariée distinguent les variables *dépendantes* et les variables *indépendantes*. Les variables dépendantes sont celles dont on veut prédire la valeur ; les autres variables sont appelées indépendantes [1]. On peut classer les méthodes d'analyse multivariée selon le nombre de variables dépendantes et indépendantes, et selon que les unes et les autres sont des variables discrètes ou continues [2].

Le tableau suivant présente un classement des méthodes qui sont abordées dans ce manuel.

---

1 Les termes « variable dépendante » et « variable indépendante » viennent des sciences expérimentales où le chercheur fixe de manière « indépendante » la valeur de certaines variables (comme, par exemple, le dosage d'un traitement), pour observer ensuite l'effet sur la variable « dépendante ». Les variables indépendantes sont parfois appelées « explicatives ». Cette expression doit cependant s'employer avec prudence, à cause de la connotation de causalité qu'elle véhicule. Dans un modèle à une seule équation, la variable dépendante s'appelle aussi « endogène », c'est-à-dire déterminée à l'intérieur du modèle, tandis que les variables indépendantes sont « exogènes », c'est-à-dire déterminées à l'extérieur du modèle. On appelle aussi les variables indépendantes « stimuli » ; les variables dépendantes sont alors des « réponses ». En anglais, on trouve les couples *predictor/criterion*, *stimulus/response*, *task/performance*, *input/output*.

2 Cela découle de l'échelle de mesure associée à chaque variable : les variables catégoriques sont discrètes, alors que les variables rationnelles et les variables d'intervalle sont traitées le plus souvent comme continues. Pour ce qui est des variables ordinales, il existe peu de méthodes qui leur soient spécifiquement adaptées ; en pratique, on les traite souvent comme continues, mais alors l'interprétation des résultats doit tenir compte de la nature ordinale des variables.

<table>
<tr><th colspan="2">Variable dépendante</th><th>Variables indépendantes</th><th colspan="2">Méthode</th></tr>
<tr><td colspan="2" rowspan="2">Aucune</td><td>2 variables catégoriques</td><td rowspan="2">Analyse de tableau de contingence</td><td>... à 2 dimensions</td></tr>
<tr><td>Plus de 2 var. catégo.</td><td>... à plus de 2 dimensions</td></tr>
<tr><td colspan="2" rowspan="2">Continue</td><td>Continues ou discrètes</td><td colspan="2">Régression multiple</td></tr>
<tr><td>Discrètes</td><td colspan="2">Analyse de variance</td></tr>
<tr><td rowspan="2">Catégorique</td><td>2 catégories</td><td rowspan="2">Continues ou discrètes</td><td rowspan="2">Logit ou probit</td><td>... binomial</td></tr>
<tr><td>Plus de 2 cat.</td><td>... multinomial</td></tr>
</table>

Ce chapitre aborde l'analyse des tableaux de contingence. Cette méthode permet d'examiner les relations entre plusieurs variables catégoriques. Dans l'analyse des tableaux de contingence, aucune des variables ne joue le rôle de variable dépendante.

4-1.1.3 RÈGLES DE PRÉSENTATION D'UN TABLEAU DE CONTINGENCE

Le principe général de présentation d'un tableau de contingence est le même que pour n'importe quel tableau : tout doit être mis en oeuvre pour que le lecteur sache parfaitement de quoi il s'agit. On pourrait dire que les données du tableau doivent être accompagnées des *métadonnées* indispensables à leur compréhension.

Les principales règles de présentation à respecter généralement sont les suivantes :

1. Le tableau est coiffé d'un titre qui...
   - identifie la population (au sens statistique) ou, le cas échéant, l'échantillon auquel se rapporte le tableau (ici, la population active employée dans la RMR de Montréal en 1991) ; noter que l'identification de la population comprend, lorsque c'est pertinent, une référence à la zone géographique et à la période de temps ;
   - indique quelles sont les unités de mesure utilisées (milliers de personnes, millions de dollars ou... ; cet élément peut être omis ici, puisqu'il s'agit du nombre de personnes) ;
   - identifie les dimensions du tableau (variables catégoriques de classement ; ici, la zone de résidence, le sexe et la profession).

2. Des « sous-titres » indiquent à quelle variable correspondent les différentes dimensions du tableau (ici, les lignes correspondent aux zones de résidence, les colonnes aux professions et le tableau est divisé en parties selon la troisième dimension, le sexe) ;
3. L'entête de chaque colonne, ligne ou partie du tableau indique à quelle catégorie de la variable correspond cette colonne, ligne ou partie du tableau ;
4. Le tableau contient des lignes et colonnes de totaux, ainsi que le grand total (1 561 817) ; les lignes et colonnes de totaux sont clairement identifiées et mises en relief (ici, par l'utilisation de caractères gras) ; de même, toute la troisième partie du tableau, intitulée « Total hommes et femmes » est constituée des totaux des cellules correspondantes des deux premières parties.
5. Enfin, la source des données est indiquée (ici, en termes généraux ; l'idéal est de donner une référence bibliographique complète).

Un tableau de contingence peut aussi contenir les éléments suivants :

- des proportions ou des pourcentages
- des sous-totaux
- des renvois et les notes correspondantes

Si le tableau contient des proportions ou des pourcentages, il faut qu'il soit évident s'il s'agit de proportions (fractions comprises entre zéro et un) ou de pourcentages (compris entre zéro et cent). De plus, il faut indiquer clairement par rapport à quel total ont été calculés les pourcentages ou proportions (pourcentage de quoi ?). Une façon de le faire est d'inscrire « 100 % » là où cela s'applique. Enfin, il faut éviter de surcharger un tableau au point d'en rendre la lecture difficile : mieux vaut parfois présenter deux tableaux, un pour les nombres et un second pour les pourcentages.

Il en est de même des sous-totaux. Par exemple, dans le tableau précédent, on pourrait juger utile de présenter le sous-total pour la CUM (somme des deux premières lignes de chaque partie). Les sous-totaux doivent être libellés de telle façon que le lecteur sache exactement ce qui a été additionné. Et il faut éviter de surcharger le tableau. À cet égard, il est parfois préférable de présenter les mêmes données en deux tableaux : un premier tableau détaillé (parfois relégué en annexe) et un second, plus agrégé (où les agrégats sont des sous-totaux).

Les notes permettent de préciser les titres, sous-titres ou entêtes sans les allonger indûment. Elles peuvent aussi être utilisées pour donner la définition de certains termes ou pour énoncer des formules employées dans le calcul des chiffres du tableau.

Enfin, ajoutons qu'il y a plusieurs façons de structurer un tableau à plus de deux dimensions. Dans l'exemple précédent, la troisième dimension, le sexe, correspond aux différentes parties du tableau. On peut aussi utiliser la technique de la subdivision des lignes ou des colonnes ; la subdivision des colonnes est illustrée schématiquement ci-après.

Schéma d'un tableau de contingence avec subdivision des colonnes

	Profession, sexe									
Zone de résidence	Directeurs, gérants, administra-teurs et assimilés			Profession-nels, enseignants et cols blancs spécialisés			...	**TOTAL toutes professions**		
	F	H	T	F	H	T		F	H	T
Montréal							...			
...										
Total										

Il est recommandé, lorsqu'on utilise cette méthode de présentation, de rapprocher vers l'intérieur les variables dont on veut examiner l'interaction. La structure représentée dans le schéma qui précède conviendrait bien à l'étude de l'interaction entre sexe et zone de résidence, la profession jouant le rôle d'une variable de contrôle (les variables de contrôle sont discutées ci-après, en 6) ; le format antérieur est mieux adapté à l'examen de la relation entre profession et zone de résidence, alors que le sexe est pris comme variable de contrôle.

4-1.2 Fréquences relatives et probabilités dans un tableau de contingence

Bien que les méthodes qui vont être présentées se généralisent aux tableaux de plus de deux dimensions, nous allons nous en tenir ici à l'analyse des tableaux à deux dimensions, plus simple. Nous allons donc réduire le tableau précédent à deux dimensions en supprimant la dimension sexe [3]. Pour cela, il suffit de faire la somme des femmes et des hommes, comme dans le tableau suivant, où les nombres sont identiques à ceux de la troisième partie du tableau 1.

---

[3] Il faut cependant être conscient du fait que cela détruit de l'information. Cette pratique est donc à éviter en recherche et elle n'est tolérable ici que pour des motifs pédagogiques.

Tableau 2 – Population active employée dans la Région métropolitaine de Montréal
Zone de résidence selon la profession, 1991

	Professions						
Zone de résidence	Directeurs, gérants, administrateurs et assimilés	Professionnels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**	Répartition ($p_{i\bullet}$)
Ville de Montréal	56 361	113 249	119 996	89 725	75 675	455 006	*0,29*
Reste de la CUM	61 721	82 127	107 822	60 238	46 184	358 092	*0,23*
Couronne Nord	50 072	59 259	94 661	74 827	47 959	326 778	*0,21*
Couronne Sud	54 371	68 138	95 890	69 263	49 101	336 763	*0,22*
Hors RMR	11 535	16 125	19 359	25 495	12 664	85 178	*0,05*
Total	234 060	338 898	437 728	319 548	231 583	1 561 817	*1,00*
Répartition ($p_{\bullet j}$)	*0,15*	*0,22*	*0,28*	*0,20*	*0,15*	*1,00*	

Nous allons utiliter la notation suivante :

x_{ij}	nombre d'observations de la colonne *j* dans la ligne *i*
$x_{\bullet j} = \sum_i x_{ij}$	nombre total d'observations de la colonne *j*
$x_{i\bullet} = \sum_j x_{ij}$	nombre total d'observations de la ligne *i*
$x_{\bullet\bullet} = \sum_i \sum_i x_{ij}$	nombre total d'observations dans le tableau

Cette notation applique la convention selon laquelle une sommation sur l'une ou l'autre des dimensions se représente en remplaçant l'indice correspondant par un point. Par exemple, dans le tableau de la population active par zone de résidence et par profession, nous avons :

$x_{23} = 107\,822$, le nombre d'employés de bureau habitant la CUM hors Montréal

$x_{\bullet 3} = 437\,728$, le nombre d'employés de bureau employés dans la RMR

$x_{2\bullet} = 358\,092$, le nombre de personnes employées dans la RMR habitant la CUM hors Montréal

$x_{\bullet\bullet} = 1\,561\,817$, le nombre total de personnes employées dans la RMR

L'analyse d'un tableau de contingence porte sur la structure des données davantage que sur l'ordre de grandeur des nombres. C'est pourquoi les analyses sont généralement formulées en termes des fréquences relatives. Celles-ci se calculent tout simplement en divisant les nombres par le total pertinent.

Les fréquences relatives s'interprètent comme des probabilités. Ainsi, $p_{34} = \frac{74827}{1561817} = 0{,}048$ est la probabilité qu'un individu tiré au hasard parmi les 1 561 817 personnes employées dans la RMR fasse partie de la profession *Ouvrier* **et** qu'il habite la Couronne Nord. Au dénominateur de la fréquence relative, on trouve donc le nombre d'individus parmi lesquels se fait le tirage (1 561 817) et au numérateur, on a le nombre d'individus parmi ceux-ci qui possèdent la ou les caractéristiques examinées (74 827).

Différents calculs de fréquences relatives correspondent aux différents concepts de probabilité. On peut ainsi calculer des probabilités *conjointes*, *marginales* ou *conditionnelles*.

$p_{ij} = x_{ij} / x_{\bullet\bullet}$ est la probabilité **conjointe** d'appartenir à la fois à *i* et à *j*.

> Ex. : $p_{34} = \frac{74827}{1561817} = 0{,}048$ est la probabilité de faire partie de la profession *Ouvrier* **et** d'habiter la Couronne Nord.

$p_{i\bullet} = x_{i\bullet} / x_{\bullet\bullet} = \sum_j x_{ij} / x_{\bullet\bullet} = \sum_j p_{ij}$ est la probabilité **marginale** d'appartenir à *i*.

> Ex. : $p_{3\bullet} = \frac{326778}{1561817} = 0{,}209$ est la probabilité d'habiter la Couronne Nord, quelle que soit la catégorie professionnelle.

$p_{\bullet j} = x_{\bullet j} / x_{\bullet\bullet} = \sum_i x_{ij} / x_{\bullet\bullet} = \sum_i p_{ij}$ est la probabilité **marginale** d'appartenir à *j*.

> Ex. : $p_{\bullet 4} = \frac{319548}{1561817} = 0{,}205$ est la probabilité d'appartenir à la profession *Ouvrier*, quelle que soit la zone de résidence.

$p_{j/i\bullet} = {x_{ij}}/{x_{i\bullet}} = \dfrac{x_{ij}/x_{\bullet\bullet}}{x_{i\bullet}/x_{\bullet\bullet}} = {p_{ij}}/{p_{i\bullet}}$ est la probabilité **conditionnelle** d'appartenir à *j*, **étant donné** que l'on appartient à *i*.

> Ex. : $p_{4/3\bullet} = \dfrac{74827}{326778} = 0{,}229$ est la probabilité d'appartenir à la profession *Ouvrier*, **étant donné** que la zone de résidence est la Couronne Nord.

$p_{i/\bullet j} = {x_{ij}}/{x_{\bullet j}} = \dfrac{x_{ij}/x_{\bullet\bullet}}{x_{\bullet j}/x_{\bullet\bullet}} = {p_{ij}}/{p_{\bullet j}}$ est la probabilité **conditionnelle** d'appartenir à *i*, **étant donné** que l'on appartient à *j*.

> Ex. : $p_{3/\bullet 4} = \dfrac{74827}{319548} = 0{,}234$ est la probabilité d'habiter la Couronne Nord, **étant donné** que l'on appartient à la profession *Ouvrier*.

Naturellement, ces probabilités ou fréquences relatives s'additionnent à 1 lorsque la sommation couvre l'ensemble des possibilités :

$$\sum_i \sum_j p_{ij} = \sum_i p_{i\bullet} = \sum_j p_{\bullet j} = 1$$

$$\sum_j p_{j/i\bullet} = {\sum_j x_{ij}}/{x_{i\bullet}} = 1$$

$$\sum_i p_{i/\bullet j} = {\sum_i x_{ij}}/{x_{\bullet j}} = 1$$

4-1.3 Test de l'hypothèse d'indépendance dans un tableau de contingence

4-1.3.1 PRÉSENTATION INTUITIVE

Le tableau 3 donne, pour chaque profession, la distribution des individus entre les zones de résidence. On constate sans trop de surprise que les individus appartenant à des professions différentes se répartissent différemment dans l'espace, entre les zones de résidence. Mais ces différences sont-elles significatives ?

On examine cette question en comparant les distributions observées avec une distribution qui serait hypothétiquement la même pour toutes les professions ; cette distribution est simplement donnée par la distribution de l'ensemble (dernière colonne du tableau).

Mais comment décider si les différences sont « significatives » ou non ? On procède pour cela à un *test d'hypothèse* (pour une discussion approfondie des tests d'hypothèse, voir le chapitre 2-3). L'hypothèse à tester est que les distributions sont identiques, et que les différences observées ne sont que des « accidents » dûs au hasard.

Tableau 3 – Population active employée dans la Région métropolitaine de Montréal
Répartition entre les zones de résidence selon la profession, 1991

	Professions					
Zone de résidence	Directeurs, gérants, administra-teurs et assimilés	Profession-nels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**
Ville de Montréal	0,241	0,334	0,274	0,281	0,327	0,291
Reste de la CUM	0,264	0,242	0,246	0,189	0,199	0,229
Couronne Nord	0,214	0,175	0,216	0,234	0,207	0,209
Couronne Sud	0,232	0,201	0,219	0,217	0,212	0,216
Hors RMR	0,049	0,048	0,044	0,080	0,055	0,055
Total	1,000	1,000	1,000	1,000	1,000	1,000

Ce test comporte trois étapes :

6. mesurer l'écart entre ce qui est observé et l'hypothèse ;
7. déterminer quelle est la probabilité qu'un écart aussi grand ou plus grand soit l'effet du hasard (plus l'écart est grand, moins il est probable qu'il soit uniquement dû au hasard) ;
8. prendre une décision.

Première étape : mesurer l'écart

Pour mesurer l'écart entre les observations et l'hypothèse, il faut d'abord se donner une représentation exacte de l'hypothèse. On calcule donc les fréquences que l'on *devrait théoriquement* avoir si les distributions étaient identiques (tableau 4).

Tableau 4 – Population active employée dans la Région métropolitaine de Montréal
Fréquences théoriques dans l'hypothèse de distributions identiques

	Professions					
Zone de résidence	Directeurs, gérants, administra-teurs et assimilés	Profession-nels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**
Ville de Montréal	68 189,0	98 731,6	127 523,8	93 094,3	67 467,4	455 006,0
Reste de la CUM	53 665,1	77 702,2	100 361,9	73 265,7	53 097,1	358 092,0
Couronne Nord	48 972,2	70 907,4	91 585,6	66 858,8	48 454,0	326 778,0
Couronne Sud	50 468,6	73 074,1	94 384,0	68 901,8	49 934,5	336 763,0
Hors RMR	12 765,1	18 482,7	23 872,7	17 427,4	12 630,0	85 178,0
Total	234 060,0	338 898,0	437 728,0	319 548,0	231 583,0	1 561 817,0

Les fréquences théoriques sont calculées simplement en appliquant au total de chaque colonne la distribution de l'ensemble (dernière colonne du tableau des répartitions) :

$$x^*_{ij} = x_{\bullet j} p_{i\bullet}$$

où l'astérisque sert à distinguer les fréquences théoriques des fréquences observées. Par exemple, [4]

$$x^*_{54} = x_{\bullet 4} \times p_{5\bullet} = 319548 \times 0{,}0545378 = 17427{,}4$$

[4] Dans la formule qui suit, la valeur de la probabilité est représentée à 7 décimales, de façon à obtenir des fréquences théoriques exactes étant donné que le multiplicateur est de l'ordre des centaines de mille. Cette précision est souhaitable ici pour rendre plus clair le développement qui suit, mais en pratique, elle n'est pas nécessaire.

On remarque que les totaux de lignes et de colonnes du tableau 4 sont égaux à ceux du tableau 2 des valeurs observées. Ce n'est pas un accident : cela découle de la formule de calcul.

$$\sum_i x_{ij}^* = \sum_i x_{\bullet j} p_{i\bullet} = \sum_i x_{\bullet j} \left(\frac{x_{i\bullet}}{x_{\bullet\bullet}} \right) = x_{\bullet j} \frac{\sum_i x_{i\bullet}}{x_{\bullet\bullet}} = x_{\bullet j} \frac{x_{\bullet\bullet}}{x_{\bullet\bullet}} = x_{\bullet j}$$

$$\sum_j x_{ij}^* = \sum_j x_{\bullet j} p_{i\bullet} = p_{i\bullet} \sum_j x_{\bullet j} = p_{i\bullet} x_{\bullet\bullet} = \frac{x_{i\bullet}}{x_{\bullet\bullet}} x_{\bullet\bullet} = x_{i\bullet}$$

Une fois calculées les fréquences théoriques, il faut mesurer l'écart entre l'ensemble des fréquences théoriques et l'ensemble des fréquences observées. Pour ce faire, on applique la formule

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij} - x_{ij}^* \right)^2}{x_{ij}^*}$$

Cette statistique s'appelle le Khi-deux de Pearson, aussi dénotée X^2 comme dans la formule.

Tableau 5 – Population active employée dans la Région métropolitaine de Montréal

Calcul du Khi-deux

	Professions					
Zone de résidence	Directeurs, gérants, administra-teurs et assimilés	Profession-nels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**
Ville de Montréal	2 051,7	2 134,6	444,4	121,9	998,5	5 751,1
Reste de la CUM	1 209,3	252,0	554,5	2 316,5	900,1	5 232,4
Couronne Nord	24,7	1 913,6	103,3	949,6	5,1	2 996,2
Couronne Sud	301,7	333,4	24,0	1,9	13,9	675,0
Hors RMR	118,5	300,8	853,4	3 734,7	0,1	5 007,5
Total	3 706,0	4 934,4	1 979,6	7 124,6	1 917,6	19 662,2

Les valeurs du tableau 5 sont les contributions des cellules individuelles au Khi-deux.

Ainsi, pour la cinquième cellule de la quatrième colonne,

$$\frac{(25\ 495 - 17\ 427{,}4)^2}{17\ 427{,}4} = 3\ 734{,}7$$

Le Khi-deux est simplement la somme de tous les éléments de ce tableau : 19 662,2

Deuxième étape : déterminer la probabilité

Pourquoi cette formule-là en particulier ? La réponse à cette question relève de la théorie de l'induction statistique. Cette formule est utilisée parce que, grâce à la statistique mathématique, on connaît la distribution de probabilité du Khi-deux ainsi calculé. Celui-ci a en effet une distribution asymptotique bien connue : c'est la Loi du χ^2 (on dit « Khi-deux », puisque le symbole χ est la lettre grecque « khi »). Ce résultat s'applique à la condition que l'on se situe dans le cadre d'un certain *modèle d'échantillonnage* ; un modèle d'échantillonnage est un modèle qui décrit le processus aléatoire par lequel on suppose que sont générés les écarts des fréquences observées par rapport aux fréquences théoriques (voir le chapitre 2-2). Nous n'examinerons pas ce modèle ici ; nous nous contenterons de dire qu'il est assez général pour que l'on puisse appliquer le test d'hypothèse du Khi-deux de Pearson à une grande variété de situations (voir ci-après, 4-1.4).

Il importe de préciser ici que, lorsque l'on utilise la Loi du χ^2, il faut tenir compte de ce que l'on appelle le nombre de *degrés de liberté*, parce que les probabilités données par la Loi du χ^2 en dépendent. Pour le test d'hypothèse du Khi-deux de Pearson, le nombre de degrés de liberté est égal à

$(C - 1) \times (L - 1)$, où C est le nombre de colonnes et L, le nombre de lignes du tableau.

Dans notre exemple (tableaux 2 à 5), C est le nombre de professions et L, le nombre de zones ; le nombre de degrés de liberté est donc égal à

$(5 - 1)\ (5 - 1) = 16$

Une parenthèse sur le nombre de degrés de liberté. La Loi du Khi-deux est représentée par une courbe dont la forme varie légèrement selon le nombre de valeurs que le hasard est, pour ainsi dire, « libre » de perturber. Dans le tableau de contingence, les totaux de lignes et de colonnes sont fixés. Ainsi, dans chacune des C colonnes, une fois que $(L - 1)$ valeurs ont été perturbées « librement » par ce diable de hasard, la dernière valeur de la colonne est déterminée par la différence entre le total et les $(L - 1)$ autres valeurs ; de même, dans chacune des L lignes, une

fois que ($C - 1$) valeurs ont été perturbées « librement », la dernière valeur de la colonne est déterminée par la différence entre le total et les ($C - 1$) autres valeurs. Donc, dans l'ensemble du tableau, une fois que ($C - 1$) ($L - 1$) « perturbations » ont été introduites, les autres valeurs sont déterminées par la nécessité de respecter les totaux marginaux.

À l'aide d'une table du Khi-deux ou de la fonction `Loi.Khideux` du tableur Excel (`Chidist` en anglais), on peut maintenant déterminer la probabilité que l'écart mesuré entre les fréquences observées et les fréquences théoriques soit si grand : la valeur de `Loi.Khideux(19662 ;16)` est inférieure à $2{,}4 \times 10^{-300}$.

Troisième étape : prendre une décision

Une probabilité de $2{,}4 \times 10^{-300}$, c'est une probabilité bien mince ! Il est extrêmement improbable que les déviations des fréquences observées par rapport aux fréquences théoriques soient uniquement dues au hasard. En fait, cela est tellement improbable que, à moins de circonstances exceptionnelles, la bonne décision à prendre est de rejeter cette hypothèse et de conclure au contraire qu'il y a décidément un rapport entre la profession et la zone de résidence.

4-1.3.2 MÊMES DONNÉES, NOUVELLE QUESTION… MÊME RÉPONSE ?!

Nous venons d'examiner la question de savoir si les individus appartenant à des professions différentes se répartissent entre les zones de résidence de façon significativement différente. Nous allons maintenant nous demander si la composition professionnelle de la population employée est significativement différente d'une zone de résidence à l'autre. Les données pertinentes se trouvent dans le tableau 6 ci-après, qui donne, pour chaque zone de résidence, la distribution des individus entre les professions.

Plus précisément, nous voulons tester l'hypothèse qu'il n'y a pas de différence significative entre les zones du point de vue de la composition professionnelle des personnes employées qui y demeurent. On compare donc, dans le tableau 6, les différentes distributions à celle que l'on *devrait théoriquement* avoir si les distributions étaient identiques, qui est la distribution de l'ensemble (dernière ligne).

Tableau 6 – Population active employée dans la Région métropolitaine de Montréal
Composition professionnelle des zones de résidence, 1991

	Professions					
Zone de résidence	Directeurs, gérants, administra-teurs et assimilés	Profession-nels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**
Ville de Montréal	0,124	0,249	0,264	0,197	0,166	1,000
Reste de la CUM	0,172	0,229	0,301	0,168	0,129	1,000
Couronne Nord	0,153	0,181	0,290	0,229	0,147	1,000
Couronne Sud	0,161	0,202	0,285	0,206	0,146	1,000
Hors RMR	0,135	0,189	0,227	0,299	0,149	1,000
Total	0,150	0,217	0,280	0,205	0,148	1,000

On constate qu'il y a en effet des différences entre les zones de résidence quant à la composition professionnelle. Pour voir si ces différences sont significatives, on procède de la même manière qu'auparavant, en commençant par calculer des fréquences théoriques. Mais, oh, surprise ! les fréquences théoriques auxquelles on arrive sont identiques à celles qui avaient été calculées pour examiner la question de la distribution entre les zones pour les différentes professions (le lecteur peut le vérifier par lui-même). Inutile de continuer : la conclusion sera forcément la même.

Évidemment rien de cela n'est un accident. Dans le premier cas, nous avions

$$x^*_{ij} = x_{\bullet j} \times p_{i\bullet}$$

Par exemple, $x^*_{54} = x_{\bullet 4} \times p_{5\bullet} = 319548 \times 0{,}0545378 = 17427{,}4$

Dans le cas présent,

$$x^*_{ij} = x_{i\bullet} \times p_{\bullet j}$$

Par exemple, $x^*_{54} = x_{5\bullet} \times p_{\bullet 4} = 85178 \times 0{,}2046002 = 17427{,}4$

Les deux formules donnent le même résultat numérique parce qu'elles sont strictement équivalentes :

$$x^*_{ij} = x_{\bullet j}\ p_{i\bullet} = x_{\bullet j}\ \frac{x_{i\bullet}}{x_{\bullet\bullet}} = x_{i\bullet}\ \frac{x_{\bullet j}}{x_{\bullet\bullet}} = x_{i\bullet}\ p_{\bullet j}$$

D'ailleurs, dans les calculs pratiques, on passe généralement directement du tableau des fréquences observées à celui des fréquences théoriques, au moyen de la formule

$$x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}}$$

Le tableau des fréquences théoriques est donc *bi-proportionnel* : les colonnes sont proportionnelles entre elles, et les lignes aussi.

4-1.3.3 GÉNÉRALISATION : L'INDÉPENDANCE STATISTIQUE DANS UN TABLEAU DE CONTINGENCE

L'analyse d'un tableau de contingence repose sur le postulat que le nombre d'individus observés dans les cellules du tableau dépend d'une structure sous-jacente. L'analyse a pour but de découvrir cette structure. Nous évoquerons plus loin, brièvement, le modèle loglinéraire qui sert à représenter cette structure. Nous nous penchons pour le moment sur un aspect particulier de cette structure, l'indépendance statistique.

Qu'est-ce que l'indépendance statistique ? En théorie des probabilités, un événement aléatoire *A* est indépendant d'un autre événement *B* si la probabilité que l'événement *A* se produise demeure la même, que l'événement *B* se produise ou non. Par exemple, dans le tableau de la population active par zone de résidence et par profession, il y a indépendance entre les variables zone de résidence et profession si, pour un individu choisi au hasard, la probabilité d'habiter dans une zone donnée est la même, quelle que soit la profession de cet individu ; symétriquement, il y a indépendance si la probabilité d'appartenir à un groupe professionnel donné est la même, quelle que soit la zone de résidence de l'individu.

Par exemple, disons que l'événement *A* est « l'individu habite la Couronne Nord » et l'événement *B* est « l'individu est un employé de bureau » ; s'il y avait indépendance, la probabilité qu'un individu choisi au hasard habite la Couronne Nord (probabilité de l'événement *A*) est la même, que cet individu soit employé de bureau (événement *B*) ou non.

Examinons de plus près comment l'indépendance se manifeste dans un tableau de contingence. Pour le voir, il faut d'abord interpréter les fréquences relatives du tableau comme des probabilités observées, que l'on confrontera aux probabilités théoriques du modèle ou de l'hypothèse. Ainsi, pour un individu choisi au hasard parmi les 1 561 817 travailleurs recensés de la RMR, la probabilité qu'il soit un ouvrier *et* qu'il habite la Couronne Sud est donnée par

$$p_{44} = \frac{x_{44}}{x_{\bullet\bullet}} = \frac{69263}{1561817} = 0{,}044$$

On calcule suivant le même principe les probabilités *marginales* observées. Ainsi, la probabilité qu'un individu choisi au hasard parmi les 1 561 817 travailleurs recensés de la RMR soit un ouvrier est donnée par

$$p_{\bullet 4} = \sum_i p_{i4} = \sum_i \left(\frac{x_{i4}}{x_{\bullet\bullet}} \right) = \frac{x_{\bullet 4}}{x_{\bullet\bullet}} = \frac{319548}{1561817} = 0{,}205$$

Et la probabilité qu'un individu choisi au hasard parmi les 1 561 817 travailleurs recensés de la RMR habite la Couronne Sud est donnée par

$$p_{4\bullet} = \sum_j p_{4j} = \sum_j \left(\frac{x_{4j}}{x_{\bullet\bullet}} \right) = \frac{x_{4\bullet}}{x_{\bullet\bullet}} = \frac{336763}{1561817} = 0{,}216$$

Et l'indépendance ? *Si* la probabilité d'être ouvrier est indépendante de la zone de résidence, la fraction d'ouvriers dans chaque zone devrait être égale à $p_{\bullet 4}$, c'est-à-dire à 20,5 %. Sachant que la fraction des travailleurs qui habitent la Couronne Sud est égale à $p_{4\bullet}$, c'est-à-dire à 21,6 %, alors, parmi les 1 561 817 travailleurs recensés de la RMR, ceux qui sont ouvriers et qui habitent la Couronne Sud devraient représenter 20,5 % de 21,6 % du total, c'est-à-dire

$$p_{\bullet 4} \times p_{4\bullet} = 0{,}205 \times 0{,}216 = 0{,}044$$

Cela, bien sûr, *sous condition que* les deux événements (être ouvrier et habiter la Couronne Sud) soient indépendants. En l'occurrence, il se trouve que le résultat est très proche de la valeur de p_{44}, ce qui laisse croire qu'en effet, les deux événements pourraient être indépendants.

Mais cette analyse est incomplète, parce que chacune des deux variables, zone de résidence et profession, comprend plus de deux catégories. De façon plus générale, donc, si deux variables sont indépendantes, on s'attend à ce que la probabilité observée d'appartenir à la fois à la

catégorie *i* de la première variable et à la catégorie *j* de la seconde soit égale au produit des probabilités marginales :

$p_{ij} = p_{i\bullet} \times p_{\bullet j}$, pour toutes les paires *i*, *j*

Un autre cheminement

On peut arriver à la même conclusion par un autre chemin, à partir de la même définition, à savoir : « Un événement *A* est indépendant d'un autre événement *B* si la probabilité que l'événement *A* se produise demeure la même, que *B* se produise ou non ». Dans le langage de la théorie des probabilités, cet énoncé équivaut à dire que la probabilité conditionnelle de *A* est égale à sa probabilité marginale, c'est-à-dire, dans un tableau de contingence à 2 dimensions : $p_{i/\bullet j} = p_{i\bullet}$.

Puisque $p_{i/\bullet j} = p_{ij} \Big/ p_{\bullet j}$, cela implique $p_{i\bullet} = p_{ij} \Big/ p_{\bullet j}$, c'est-à-dire $p_{ij} = p_{i\bullet}\ p_{\bullet j}$.

De façon équivalente, les 2 variables catégoriques sont indépendantes si : $p_{j/i\bullet} = p_{\bullet j}$.

Puisque $p_{j/i\bullet} = p_{ij} \Big/ p_{i\bullet}$, cela implique $p_{\bullet j} = p_{ij} \Big/ p_{i\bullet}$, c'est-à-dire $p_{ij} = p_{i\bullet}\ p_{\bullet j}$.

Voilà la définition exacte de l'indépendance statistique entre deux variables catégoriques. Il saute aux yeux que cette définition est parfaitement symétrique par rapport aux deux variables. On peut également constater que les fréquences théoriques dont il a été question précédemment sont celles que l'on s'attendrait à voir dans l'hypothèse de l'indépendance statistique. En effet,

$$x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}} = \left(\frac{x_{i\bullet}}{x_{\bullet\bullet}}\right)\left(\frac{x_{\bullet j}}{x_{\bullet\bullet}}\right) x_{\bullet\bullet} = p_{i\bullet}\ p_{\bullet j}\ x_{\bullet\bullet}$$

Le tableau des fréquences théoriques (tableau 4) est donc une représentation exacte de l'hypothèse d'indépendance.

Mais pourquoi s'intéresser à l'hypothèse d'indépendance ? Parce que, si deux variables sont indépendantes, aucune des deux ne peut être considérée comme ayant une influence sur l'autre (noter que, dans cette formulation, aucune des deux variables n'est identifiée comme devant jouer le rôle de variable indépendante ou « explicative »).

Il peut arriver que des données se conforment parfaitement à l'hypothèse d'indépendance. Le plus souvent cependant, les données seront différentes de ce que prédit le modèle

d'indépendance. Si les données ne s'en écartent « pas trop », le modèle pourra être jugé acceptable, pourvu que l'on admette qu'il n'est qu'une approximation et qu'entre la réalité et le modèle, il intervient un élément aléatoire, que nous avons appelé « perturbation due au hasard ». Les hypothèses que l'on fait quant à cet élément aléatoire permettent de baliser l'incertitude quant au « vrai » modèle.

Le test d'hypothèse que nous avons décrit précédemment est donc une procédure d'induction statistique qui encadre la décision de rejeter ou non le modèle de l'indépendance statistique.

4-1.3.4 UN AUTRE TEST : LE TEST DU RAPPORT DE VRAISEMBLANCE

On utilise aussi la statistique du rapport de vraisemblance (plus exactement, moins deux fois le logarithme du rapport des fonctions de vraisemblance). Cette statistique est définie pour un tableau rectangulaire par [5]

$$G^2 = -2\sum_i \sum_j x_{ij} \ln\left(\frac{x^*_{ij}}{x_{ij}}\right)$$

$$G^2 = 2\sum_i \sum_j x_{ij} \ln\left(\frac{x_{ij}}{x^*_{ij}}\right)$$

Comme le Khi-deux de Pearson, G^2 a une distribution asymptotique χ^2 avec, sous contrainte de l'hypothèse d'indépendance, $(L-1)(C-1)$ degrés de liberté [6].

Le calcul de cette variable-test est illustré ci-après pour le tableau de la population active selon la profession et la zone de résidence.

---

5 La formule donnée ci-après est bien la bonne. Elle diffère de la définition informelle qu'en donne Upton (1981, p. 36, définition du Y^2).

6 Que X^2 et G^2 aient toutes deux la même distribution asymptotique n'implique pas qu'elles aient la même valeur.

Tableau 7 – Population active employée dans la Région métropolitaine de Montréal

Calcul de la statistique du rapport de vraisemblance G^2

	Professions					
Zone de résidence	Directeurs, gérants, administra-teurs et assimilés	Profession-nels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**
Ville de Montréal	-10737,1	15536,0	-7301,1	-3307,6	8687,8	2 878,0
Reste de la CUM	8632,4	4548,4	7730,8	-11793,9	-6442,2	2 675,5
Couronne Nord	1112,0	-10634,5	3126,5	8425,2	-492,4	1 536,8
Couronne Sud	4049,5	-4765,5	1517,9	362,2	-826,5	337,6
Hors RMR	-1168,8	-2200,5	-4057,2	9699,2	34,0	2 306,7
Total	1 888,0	2 484,0	1 016,8	3 385,1	960,7	9 734,6

Les valeurs du tableau 7 sont les contributions des cellules individuelles au G^2. Ainsi, pour la cinquième cellule de la quatrième colonne,

$$25\,495 \times \ln\left(\frac{25\,495}{17\,427{,}4}\right) = 9\,699{,}2$$

Le G^2 est simplement égal au double de la somme de tous les éléments de ce tableau : 19 469,2. La probabilité critique correspondante est inférieure à $2{,}4 \times 10^{-300}$

4-1.4 Un regard approfondi sur le Khi-deux de Pearson

4-1.4.1 Les mille et une applications du test du Khi-deux de Pearson aux tableaux de contingence

Test sur une seule cellule du tableau

Chacun des termes de la double sommation qui forme le Khi-deux peut être interprété comme la « contribution » de la cellule correspondante au Khi-deux. Cela permet de repérer les cellules les plus « déviantes » par rapport à l'hypothèse.

On peut même tester formellement l'hypothèse qu'une cellule en particulier du tableau est significativement « déviante ». Il suffit pour cela de construire un tableau où sont agrégées toutes les autres lignes et colonnes. Par exemple, si l'on veut faire le test pour la cellule [*h*, *k*], on construit un tableau agrégé 2 × 2 sur le modèle suivant

x_{hk}	$\sum_{j} x_{hj}$
$\sum_{i} x_{ik}$	$\sum_{i \neq h} \sum_{j \neq k} x_{ij}$

On applique ensuite à ce tableau le test du Khi-deux de Pearson avec 1 degré de liberté : $(L - 1) \times (C - 1) = (2 - 1) \times (2 - 1) = 1$

Considérons par exemple, la fraction des employés qui habitent à l'extérieur de la RMR et qui appartiennent aux professions *Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc*. Dans le tableau 5, on voit que cette cellule du tableau ne contribue que pour 0,1 à la valeur totale du Khi-deux. On peut tester l'hypothèse que cette déviation n'est pas significative par rapport à l'hypothèse d'indépendance. À partir du tableau de contingence, on construit le tableau agrégé qui suit.

Tableau 8 – Tableau agrégé
Population active employée, Région métropolitaine de Montréal
Zone de résidence, selon le sexe et la profession, 1991

	Toutes professions sauf →	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	Total
RMR	1 257 720	218 919	85 178
Hors RMR	72 514	12 664	1 476 639
Total	231 583	1 330 234	1 561 817

Source : Statistique Canada, Recensement de 1991

La valeur du Khi-deux calculé à partir de ce tableau est de 0,11, ce qui est bien loin du 19 662,2 obtenu avec le tableau détaillé. La probabilité critique associée à 0,11 est de 74 %, ce qui ne permet décidément pas de rejeter l'hypothèse d'indépendance dans le tableau agrégé.

Test d'homogénéité entre deux ou plusieurs groupes ou échantillons

Il arrive souvent que l'on ait à analyser des tableaux qui comparent deux groupes d'individus répartis entre plusieurs catégories. Pour deux groupes, le tableau de contingence prend la forme suivante :

	Groupe *A*	**Groupe *B***	**Total *A*+B**
Catégories			
Total			

Un *test d'homogénéité* entre deux ou plusieurs groupes vise à déterminer si, du point de vue de leur répartition entre les catégories d'une variable de classification donnée, les deux groupes sont significativement différents ou non. La clé pour appliquer le test du Khi-deux à une telle situation est de voir que le groupe d'appartenance est une seconde variable catégorique.

On pourrait vouloir comparer, par exemple, la répartition des hommes et des femmes entre les professions. On reconnaîtra facilement que la question de savoir si la répartition des femmes entre les professions est significativement différente de celle des hommes n'est rien d'autre que la question de l'indépendance entre la variable *Profession* et la variable *Sexe*.

Test d'homogénéité entre une sous-population et le reste de la population

Le test d'homogénéité sert notamment à comparer un groupe particulier avec le reste de la population. On l'utilise en particulier pour comparer un échantillon à la population dont il est tiré, pour voir si l'échantillon est représentatif de certaines caractéristiques connues de la population.

Supposons par exemple que l'on fasse un sondage en interviewant des résidants de Montréal choisis au hasard à l'intersection des rues Sainte-Catherine et Jeanne-Mance. Si la variable linguistique est importante aux fins de l'étude, on voudra vérifier à la fin si la proportion de francophones et d'anglophones interrogés est représentative de la composition linguistique de Montréal. Pour ce faire, on construira un tableau selon le schéma suivant :

	Groupe *A*	**Reste de la population**	**Total**

Catégories		Calculer par soustraction	
Total			

Pour notre exemple, les catégories pertinentes sont naturellement *Francophone*, *Anglophone* et *Autre* [7]. Les données relatives au *Groupe A* sont celles de l'échantillon ou autre groupe particulier à l'étude ; celles de la colonne *Total* peuvent être obtenues de sources officielles, comme le Recensement. On calcule les chiffres du *Reste de la population* par soustraction [8].

Test de l'hypothèse d'une distribution particulière

De façon plus générale, le test du Khi-deux peut servir à évaluer n'importe quelle hypothèse sur la distribution d'un ensemble d'individus entre des catégories [9]. Pour ce faire, on traite l'ensemble d'individus étudié comme un échantillon tiré d'une population infinie qui est distribuée selon l'hypothèse à évaluer.

Par exemple, selon le recensement de la population de 1984 au Costa Rica, ce pays comptait alors 630 995 hommes et 649 619 femmes (tableau 9). Confrontons ces chiffres à l'hypothèse d'une distribution 50-50 entre les sexes.

Tableau 9 – Population masculine et féminine, Costa Rica, 1984

	Données du recensement	Fréquences théoriques
Hommes	630 995	640 307
Femmes	649 619	640 307
Total	1 280 614	1 280 614

Source : http://populi.eest.ucr.ac.cr/observa/estima/cuadro1.htm

On calculera la valeur du Khi-deux comme

7 Nous passons ici sous silence la difficulté qu'il y a à définir l'appartenance linguistique de façon opérationnelle et la difficulté, plus grande encore, de trouver dans les données du Recensement de Statistique Canada l'information pertinente (la formulation des questions du Recensement qui portent sur l'appartenance linguistique est vivement critiquée).

8 Si le *Groupe A* ne représente qu'une fraction minime de l'ensemble, on peut en pratique calculer le Khi-deux entre le *Groupe A* et l'ensemble, même si ce n'est pas théoriquement exact.

9 Blalock (1972, p.312), Exercice No 3.

$$X^2 = \frac{(630\,995 - 640\,307)^2}{640\,307} + \frac{(649\,619 - 640\,307)^2}{640\,307} = 270{,}85$$

La probabilité critique, avec 1 degré de liberté, est égale à $7{,}4 \times 10^{-61}$, ce qui conduit à rejeter l'hypothèse que la distribution de la population entre hommes et femmes n'est pas significativement différente de la distribution 50-50.

En apparence, cette procédure diffère de celle qui a été utilisée jusqu'ici. Mais il n'en est rien. C'est que ce test repose sur la comparaison implicite entre la population étudiée et une population hypothétique de taille infinie, qui respecte la distribution hypothétique à tester. De façon explicite, voici le tableau de contingence sous-jacent à ce test.

Tableau 10 – Population masculine et féminine, Costa Rica, 1984

	Population Costa Rica 1984	Reste	Population hypothétique infinie
Hommes	630 995	$0{,}5 \times Y - 630\,995$	$0{,}5 \times Y$
Femmes	649 619	$0{,}5 \times Y - 649\,619$	$0{,}5 \times Y$
Total	1 280 614	$Y - 1\,280\,614$	Y

Source : http://populi.eest.ucr.ac.cr/observa/estima/cuadro1.htm

Calcul des fréquences théoriques

	Population Costa Rica 1984	Reste
Hommes	$\frac{(630\,995 - 640\,307)^2}{640\,307}$	$\frac{\{(0{,}5\,Y - 630\,995) - [0{,}5\,(Y - 1280\,614)]\}^2}{0{,}5\,(Y - 1280\,614)} = \frac{(640\,307 - 630\,995)^2}{0{,}5\,(Y - 1280\,614)}$
Femmes	$\frac{(649\,619 - 640\,307)^2}{640\,307}$	$\frac{\{(0{,}5\,Y - 649\,307) - [0{,}5\,(Y - 1280\,614)]\}^2}{0{,}5\,(Y - 1280\,614)} = \frac{(640\,307 - 649\,307)^2}{0{,}5\,(Y - 1280\,614)}$

Si Y est infiniment grand, la contribution de la colonne *Reste* à la valeur du Khi-deux est négligeable (infiniment petite), puisque le diviseur $0{,}5\,(Y - 1\,280\,614)$ est infiniment grand, de sorte que le calcul se ramène à ce qui avait été énoncé précédemment. Par ailleurs, le nombre de degrés de liberté est bien égal à $(C - 1)\,(L - 1)$ [10].

[10] Il y a des situations où le calcul des fréquences théoriques est soumis à plus d'une contrainte. Le nombre de degrés de liberté se calcule alors différemment. Voir Blalock (1972, Exercice No 3, p.312).

4-1.4.2 CONDITIONS DE VALIDITÉ DU TEST DU KHI-DEUX DE PEARSON

Le test du Khi-deux de Pearson est basé sur une approximation : la distribution du χ^2 est la ditribution *asymptotique* de la statistique du Khi-deux de Pearson. Pour que le test soit valide, il faut que l'approximation soit suffisamment bonne. En général, on considère que l'approximation est suffisamment bonne et que le test est valide, si le nombre total d'observations respecte la condition

$$x_{\bullet\bullet} > 10 \times L \times C ,$$

où C est le nombre de colonnes et L, le nombre de lignes du tableau (Legendre et Legendre, 1998, p. 218).

En pratique, la plupart des auteurs affirment que le test du Khi-deux de Pearson risque de ne pas être valide s'il a une ou plusieurs cellules dont la fréquence théorique est inférieure à 5 (Freund et Williams, 1973, p. 379).

Pour Legendre et Legendre (1998, p. 218), le test pourrait ne pas être valide si $x_{\bullet\bullet} < 5 \times L \times C$.

Cette condition est étroitement apparentée à la précédente : lorsque cette condition se réalise, il y a nécessairement au moins une cellule de fréquence théorique inférieure à 5, tel qu'il est démontré dans l'encart ci-après.

On a :

$$\underset{i}{MIN}\left[x_{i\bullet}\right] \le \frac{x_{\bullet\bullet}}{L} \text{ et } \underset{j}{MIN}\left[x_{\bullet j}\right] \le \frac{x_{\bullet\bullet}}{C} \text{, de sorte que}$$

$$\underset{i,j}{MIN}\left[x_{i\bullet} x_{\bullet j}\right] \le \frac{\left(x_{\bullet\bullet}\right)^2}{L \times C} \text{, c'est-à-dire } \underset{i,j}{MIN}\left[\frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}\right] \le \frac{x_{\bullet\bullet}}{L \times C}$$

Il en découle que, si $x_{\bullet\bullet} < 5 \times L \times C$, c'est-à-dire si $\frac{x_{\bullet\bullet}}{L \times C} < 5$,

alors la plus petite fréquence théorique $x^*_{ij} = \frac{x_{i\bullet}\, x_{\bullet j}}{x_{\bullet\bullet}}$ sera inférieure à 5.

Par contre, Cochran (1954) et Siegel (1956), cités par Legendre et Legendre (1998, p. 218) énoncent les conditions suivantes, moins restrictives, sous lesquelles le test du Khi-deux ne serait pas valide :

- Il y a une ou plusieurs cellules *ij* dont la fréquence théorique x_{ij}^* est inférieure à 1;

> Puisque $x_{ij}^* = \frac{x_{i\bullet}\, x_{\bullet j}}{x_{\bullet\bullet}}$, cette condition équivaut à dire qu'il y a au moins une ligne *i* et une colonne *j* telles que $x_{i\bullet}\, x_{\bullet j} < x_{\bullet\bullet}$.

ou bien

- Il y a 20 % des cellules *ij* dont la fréquence théorique x_{ij}^* est inférieure à 5.

Selon d'autres auteurs, même la dernière restriction énoncée (20 % des cellules...) serait encore trop sévère. Legendre et Legendre (1998, p. 218) citent Fienberg (1980), selon qui le test est valide à un seuil de signification de 5 %, pourvu que *toutes* les fréquences théoriques x_{ij}^* soient supérieures à 1.

En résumé, on retiendra que, lorsque l'on applique le test du Khi-deux à un tableau de contingence, il faut se méfier des résultats lorsque certaines des fréquences *théoriques* sont trop petites.

Que faire si l'on a des raisons de croire que le test du Khi-deux n'est pas valide ? Une première solution consiste à regrouper des catégories pour que soient fusionnées des lignes ou des colonnes ne contenant qu'un petit nombre d'observations. On aura ainsi des fréquences théoriques plus élevées dans les cellules fusionnées. Il ne faut cependant pas fusionner les catégories n'importe comment ! Le regroupement des catégories revient à modifier la manière d'opérationnaliser l'hypothèse (voir le début du chapitre 2-2). Cela doit donc être justifié en fonction du modèle conceptuel sous-jacent à l'étude.

Par ailleurs, il est souvent préférable de carrément écarter de l'analyse des catégories qui se prêtent mal à l'interprétation (par exemple, les réponses « Ne sait pas » dans les données d'enquête). On songera même parfois à écarter des catégories peu nombreuses qui ne sont pas dépourvues de contenu analytique, mais qui ne peuvent pas être regroupées avec d'autres pour former de nouvelles catégories qui sont pertinentes par rapport au modèle conceptuel.

4-1.4.3 QUELQUES PROPRIÉTÉS NUMÉRIQUES DU KHI-DEUX DE PEARSON

Le Khi-deux de Pearson possède notamment les propriétés suivantes :

9. Khi-deux est non négatif
10. Khi-deux est nul quand $x_{ij} = x_{ij}^*$ pour toutes les cellules *i*,*j* du tableau.
11. Khi-deux augmente avec le nombre d'observations $x_{\bullet\bullet}$
12. $X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$

    où *C* est le nombre de colonnes et *L*, le nombre de lignes du tableau et

    où l'expression $Min(L-1, C-1)$ représente la plus petite valeur parmi (*L* – 1) et (*C* – 1).

Les deux premières propriétés sont relativement évidentes lorsqu'on examine la formule de calcul

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*}$$

La troisième propriété est illustrée par l'exemple suivant.

Tableau 11 – Sensibilité du Khi-deux au nombre d'observations

Illustration numérique

Tableaux de contingence

			$p_{i\bullet}$
15	10	25	0,5
10	15	25	0,5
25	25	50	
0,5	0,5		$\leftarrow p_{\bullet j}$

			$p_{i\bullet}$
30	20	50	0,5
20	30	50	0,5
50	50	100	
0,5	0,5		$\leftarrow p_{\bullet j}$

			$p_{i\bullet}$
60	40	100	0,5
40	60	100	0,5
100	100	200	
0,5	0,5		$\leftarrow p_{\bullet j}$

Fréquences théoriques

12,5	12,5	25
12,5	12,5	25
25	25	50

25	25	50
25	25	50
50	50	100

50	50	100
50	50	100
100	100	200

Calcul du Khi-deux

0,5	0,5
0,5	0,5

Khi-deux = 2
n.lignes= 2
n. col.= 2
d.l.= 1
Prob.crit.= 0,157

1	1
1	1

Khi-deux = 4
n.lignes= 2
n. col.= 2
d.l.= 1
Prob.crit.= 0,046

2	2
2	2

Khi-deux = 8
n.lignes= 2
n. col.= 2
d.l.= 1
Prob.crit.= 0,005

Les trois tableaux de contingence ci-haut ont des structures identiques. La seule chose qui les distingue est le nombre total d'observations : 25, 50 et 100. Le test du Khi-deux conduit à rejeter l'hypothèse d'indépendance dans le troisième cas et, de façon moins catégorique dans le second ; dans le premier cas cependant, on déciderait généralement que l'hypothèse ne peut pas être rejetée, du moins selon les critères habituellement utilisés en sciences sociales.

De façon plus générale, lorsque, pour une structure donnée, le nombre d'observations augmente proportionnellement dans toutes les cellules, la valeur du Khi-deux augmente dans la même proportion. Formellement, lorsque le nombre d'observations est multiplié par α, on a :

$$\sum_i \sum_j \frac{\left(\alpha x_{ij}^* - \alpha x_{ij}\right)^2}{\alpha x_{ij}^*} = \sum_i \sum_j \frac{\alpha^2 \left(x_{ij}^* - x_{ij}\right)^2}{\alpha x_{ij}^*} = \alpha \left[\sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*} \right]$$

Pourquoi s'intéresser à cette propriété ? Parce que si le nombre total d'observations $x_{\bullet\bullet}$ est grand, on peut être conduit à rejeter l'hypothèse d'indépendance (et donc, à considérer que les différences entre les distributions sont statistiquement significatives), alors qu'elles ne sont pas nécessairement *scientifiquement* significatives. Inversement, il peut arriver que des différences réelles apparaissent comme non significatives statistiquement, si le nombre d'observations est petit.

Supposons, par exemple, qu'au lieu d'utiliser des données du Recensement sur la profession et la zone de résidence, on prenait un échantillon de 1 sur 1000. Supposons aussi que, par le plus heureux des hasards, l'échantillon reflète au plus près la population, de sorte que les fréquences observées de l'échantillon soient égales à un millième des fréquences observées du Recensement, à une erreur d'arrondissement près (puisque l'on ne peut pas avoir des fractions de personnes dans l'échantillon). On obtiendrait alors le tableau suivant.

Tableau 12 – Échantillon fictif

Population active employée dans la Région métropolitaine de Montréal

Zone de résidence selon la profession, 1991

Zone de résidence	**Professions**					
	Directeurs, gérants, administra-teurs et assimilés	Profession-nels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**
Ville de Montréal	56	113	120	90	76	455
Reste de la CUM	62	82	108	60	46	358
Couronne Nord	50	59	95	75	48	327
Couronne Sud	54	68	96	69	49	337
Hors RMR	12	16	19	25	13	85
Total	234	339	438	320	232	1 562

Eh bien, avec les données de cet échantillon fictif, pourtant éminemment représentatif, la valeur du Khi-deux de Pearson n'est plus que de 19,79 et la probabilité correspondante est de 0,23 : on ne peut pas rejeter l'hypothèse d'indépendance !

Nous reviendrons sur ce point à propos des mesures de l'intensité de la relation entre deux variables catégoriques. D'ailleurs, la quatrième propriété du Khi-deux de Pearson,

$$X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$$

intervient dans la définition de certaines de ces mesures.

Démonstration de $X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$

La démonstration de cette dernière propriété fait appel à une formule de calcul du Khi-deux autre que celle que nous avons donnée précédemment. Cette nouvelle formule est dérivée de la première :

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*} = \sum_i \sum_j \frac{\left[\left(x_{ij}^*\right)^2 - 2\,x_{ij}^*\,x_{ij} + x_{ij}^2\right]}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j x_{ij}^* - 2\sum_i \sum_j x_{ij} + \sum_i \sum_j \frac{x_{ij}^2}{x_{ij}^*} = x_{\bullet\bullet} - 2x_{\bullet\bullet} + \sum_i \sum_j \frac{x_{ij}^2}{\left(\frac{x_{i\bullet}x_{\bullet j}}{x_{\bullet\bullet}}\right)}$$

$$X^2 = x_{\bullet\bullet}\left[\sum_i \sum_j \frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} - 1\right]$$

Pour démontrer la quatrième propriété, il suffit de constater que, d'une part,

$\frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} \leq \frac{x_{ij}}{x_{i\bullet}}$, de sorte que $\sum_{i=1}^{L}\sum_{j=1}^{C}\frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} \leq \sum_{i=1}^{L}\sum_{j=1}^{C}\frac{x_{ij}}{x_{i\bullet}} = \sum_{i=1}^{L}\frac{x_{i\bullet}}{x_{i\bullet}} = L$

et que, d'autre part,

$\frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} \leq \frac{x_{ij}}{x_{\bullet j}}$, de sorte que $\sum_{i=1}^{L}\sum_{j=1}^{C}\frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} \leq \sum_{i=1}^{L}\sum_{j=1}^{C}\frac{x_{ij}}{x_{\bullet j}} = \sum_{i=1}^{C}\frac{x_{\bullet j}}{x_{\bullet j}} = C$

Il s'ensuit que

$$X^2 = x_{\bullet\bullet}\left[\sum_i \sum_j \frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} - 1\right] \leq x_{\bullet\bullet}[L-1]$$

et

$$X^2 = x_{\bullet\bullet}\left[\sum_i \sum_j \frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} - 1\right] \leq x_{\bullet\bullet}[C-1]$$

Q.E.D.

4-1.4.4 *POST SCRIPTUM* : UN NOUVEAU REGARD SUR LE QUOTIENT DE LOCALISATION

Dans le chapitre 1-2, nous avons présenté le quotient de localisation comme un instrument pour analyser un tableau de l'emploi par branche et par ville ou région. Nous avions par ailleurs mentionné qu'un tableau de ce type est un tableau de contingence (à deux dimensions). On peut donc transposer le même calcul à n'importe quel tableau de contigence à deux dimensions (bien que le mot « localisation » soit quelque peu incongru dans certains contextes).

Plus intéressant, on peut réexaminer le quotient de localisation à la lumière des tests d'hypothèse appliqués aux tableaux de contingence. Plus précisément, il y a une relation très simple entre les quotients de localisation et les nombres attendus sous l'hypothèse de l'indépendance. Ces derniers sont donnés par

$$x_{ij}^* = \frac{x_{i\bullet}\, x_{\bullet j}}{x_{\bullet\bullet}}$$

Quant aux quotients de localisation, on les calcule à l'aide de la formule $QL_{ij} = \dfrac{\left(x_{ij} \middle/ x_{\bullet j} \right)}{\left(x_{i\bullet} \middle/ x_{\bullet\bullet} \right)}$ ou, de manière équivalente, $QL_{ij} = \dfrac{\left(x_{ij} \middle/ x_{i\bullet} \right)}{\left(x_{\bullet j} \middle/ x_{\bullet\bullet} \right)}$, ce qui revient à $QL_{ij} = \dfrac{x_{ij}}{\left(\dfrac{x_{i\bullet}\, x_{\bullet j}}{x_{\bullet\bullet}} \right)} = \dfrac{x_{ij}}{x_{ij}^*}$

Le quotient de localisation est donc le rapport de la fréquence observée sur la fréquence théorique sous l'hypothèse d'indépendance ; cette hypothèse se traduit, comme nous l'avons montré, par un tableau bi-proportionnel. Cette relation permet aussi d'exprimer le Khi-deux de Pearson en termes des quotients de localisation. En effet, puisque

$$x_{ij}^*\, QL_{ij} = x_{ij}$$

on a

$$X^2 = \sum_i \sum_j x_{ij}^* \left(QL_{ij} - 1 \right)^2$$

Démonstration de $X^2 = \sum_i \sum_j x_{ij}^* \left(QL_{ij} - 1 \right)^2$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij} - x_{ij}^* \right)^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^*\, QL_{ij} - x_{ij}^* \right)^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left[x_{ij}^* \left(QL_{ij} - 1\right)\right]^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^*\right)^2 \left(QL_{ij} - 1\right)^2}{x_{ij}^*}$$

Q.E.D.

Le Khi-deux de Pearson est une somme pondérée des carrés des déviations des quotients de localisation par rapport à la valeur repère 1 ; le poids de chaque cellule est la fréquence théorique sous l'hypothèse d'indépendance.

S'agissant de l'étude d'un tableau de l'emploi par branche et par ville ou région, il est bien évident que, dans l'immense majorité des cas, le test du Khi-deux conduira à rejeter catégoriquement l'hypothèse d'indépendance. L'intérêt du rapprochement que nous venons de faire réside plutôt dans l'interprétation que l'on peut donner à chacun des termes de la double sommation : c'est la contribution de la cellule correspondante au Khi-deux. En termes relatifs, le rapport

$$\frac{x_{ij}^* \left(QL_{ij} - 1\right)^2}{X^2}$$

est la part de la déviance totale (par rapport à la bi-proportionnalité) qui est attribuable à la cellule *i*,*j*.

4-1.5 Mesures de l'intensité de la relation entre deux variables catégoriques

Lorsqu'on rejette l'hypothèse d'indépendance, cela signifie que l'on décide qu'il existe entre les deux variables une relation statistiquement significative. Mais une relation statistiquement significative n'est pas nécessairement scientifiquement ou pratiquement pertinente ou importante. La nécessité de cette distinction ressort clairement à la lumière notamment de la troisième des propriétés discutées en 4.3 (le Khi-deux augmente avec le nombre d'observations). D'où, l'utilité de mesurer l'intensité de la relation entre deux variables catégoriques.

4-1.5.1 MESURES DÉRIVÉES DU KHI-DEUX DE PEARSON

Le Khi-deux de Pearson possède, comme nous l'avons vu, certaines propriétés numériques non désirables comme mesure de l'intensité de la relation entre deux variables catégoriques :

13. X^2 augmente avec le nombre d'observations $x_{\bullet\bullet}$

14. $X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$

    où C est le nombre de colonnes et L, le nombre de lignes du tableau.

On cherche une mesure qui reflète la structure, et non le nombre d'observations, et qui, idéalement, serait comprise entre 0 et 1 plutôt qu'entre 0 et $x_{\bullet\bullet}\, Min(L-1, C-1)$.

Voici quelques-unes des mesures dérivées du Khi-deux de Pearson.

15. $\varphi^2 = \dfrac{X^2}{x_{\bullet\bullet}}$

    Ce « Phi-deux » (le symbole φ est la lettre grecque « phi ») est compris entre 0 et 1 pour les tableaux 2×2, mais en général, sa valeur maximum est considérablement plus élevée.

16. Le T^2 de Tschuprow :

$$T^2 = \frac{\varphi^2}{\sqrt{(L-1)(C-1)}} = \frac{X^2}{x_{\bullet\bullet}\sqrt{(L-1)(C-1)}}$$

    Sa valeur maximum est égale à 1 si $L = C$; autrement, elle est strictement inférieure à 1.

17. Le V^2 de Cramer :

$$V^2 = \frac{\varphi^2}{Min(L-1, C-1)} = \frac{X^2}{x_{\bullet\bullet} Min(L-1, C-1)}$$

    Le V^2 de Cramer est équivalent au T^2 de Tschuprow lorsque $L = C$, mais contrairement à ce dernier, il peut atteindre la valeur maximum de 1 lorsque $L \neq C$.

4-1.5.2 AUTRES MESURES (*TAU* ET *LAMBDA*)

Principe général

Le *tau* de Goodman et Kruskal (du nom de la lettre grecque τ, qui a la valeur phonétique de « T ») et le *lambda* (du nom de la lettre grecque λ, qui a la valeur phonétique de « L ») ne sont pas symétriques : ces mesures reposent sur une distinction entre la variable dépendante et la variable indépendante. Dans le cas où la variable dépendante correspond aux catégories *j*

(colonnes), le *tau* et le *lambda* mesurent l'intensité de la relation par la réduction relative moyenne des erreurs d'assignation que l'on fait lorsque l'on doit prédire à quelle catégorie *j* appartient un individu lorsqu'on sait à quelle catégorie *i* il appartient. Leur forme générale est donc

$$1 - \frac{\text{Nombre moyen d' erreurs d' assignation lorsqu' on connaît } i}{\text{Nombre moyen d' erreurs d' assignation lorsqu' on ne connaît pas } i}$$

Les deux mesures diffèrent quant à la règle suivie pour prédire à quelle catégorie *j* appartient un individu.

Le tau *de Goodman et Kruskal*

Règle d'assignation

Les individus sont distribués entre les catégories *j* proportionnellement aux $p_{\bullet j}$ si l'on ne connaît pas *i* et proportionnellement aux p_{ij} lorsque l'on connaît *i*.

Formule

$$\tau_J = 1 - \frac{\sum_i \sum_j p_{ij}\left(1 - \frac{p_{ij}}{p_{i\bullet}}\right)}{\sum_j p_{\bullet j}(1 - p_{\bullet j})} = 1 - \frac{\sum_i \sum_j p_{ij}\left(1 - p_{j\,/\,i\bullet}\right)}{\sum_j p_{\bullet j}(1 - p_{\bullet j})}$$

Valeurs limites

Le *tau* est égal à zéro lorsque les deux variables sont parfaitement indépendantes, c'est-à-dire lorsque $p_{ij} = p_{i\bullet}\, p_{\bullet j}$. Il est égal à 1 lorsque, dans chaque ligne *i* du tableau, il n'y a qu'une seule cellule non nulle, ce qui permet de prédire *j* avec certitude quand on connaît *i*.

Le lambda

Règle d'assignation

Lorsque l'on ne sait pas à quelle catégorie *i* appartiennent les individus, ils sont tous assignés à la catégorie qui contient le plus grand nombre d'observations, c'est-à-dire la catégorie pour laquelle la probabilité marginale $p_{\bullet j}$ est la plus grande ; lorsque l'on connaît *i*, les individus sont assignés à la catégorie *j* qui contient le plus grand nombre d'observations à l'intérieur de la

catégorie *i*, c'est-à-dire la catégorie pour laquelle la probabilité conditionnelle $p_{j/i\bullet}$ est la plus grande.

Formule

$$\lambda_J = 1 - \frac{\sum_i \left(1 - \frac{p_{i,Max}}{p_{i\bullet}}\right) p_{i\bullet}}{(1 - p_{\bullet Max})} = 1 - \frac{\sum_i (1 - p_{Max/i\bullet}) p_{i\bullet}}{(1 - p_{\bullet k})},$$

où $p_{\bullet Max} = \underset{j}{Max}\, p_{\bullet j}$, $p_{i,Max} = \underset{j}{Max}\, p_{ij}$ et $p_{Max/i\bullet} = \underset{j}{Max}\, p_{j/i\bullet}$.

Valeurs limites

Le *lambda* est égal à 1 lorsque, dans chaque ligne *i* du tableau, il n'y a qu'une seule cellule non nulle et que $p_{i,Max} = p_{i\bullet}$, ce qui permet de prédire *j* avec certitude quand on connaît *i*. Il est égal à zéro lorsque $p_{i,Max} = p_{i\bullet}\, p_{\bullet Max}$ pour tout *i*, même si les deux variables ne sont pas indépendantes, c'est-à-dire même si, pour les colonnes *j* autres que celle où se trouve $p_{i,Max} = \underset{j}{Max}\, p_{ij}$, on a $p_{ij} \neq p_{i\bullet}\, p_{\bullet j}$. Cette dernière propriété amène à préférer *tau*.

4-1.6 Les variables de contrôle dans les tableaux à plus de deux dimensions

Le tableau 1 comportait trois dimensions : le sexe, la zone de résidence et la profession. Jusqu'à maintenant, notre analyse n'a porté que sur les deux dernières et nous avons ignoré la possibilité qu'il existe des différences entre les femmes et les hommes. Mais il se pourrait bien que la structure du tableau de contingence *zone de résidence-profession* soit fort différente pour les femmes et les hommes. La rigueur scientifique exigerait que nous tenions compte de cette possibilité.

Plus généralement, quand on examine la relation entre deux variables catégoriques, il faut se demander si l'intensité ou la forme de cette relation ne pourrait pas être influencée par d'autres variables. Et si tel est le cas, il faut tenir compte de ces autres variables, qu'on appelle *variables de contrôle*. Cette expression vient du langage des sciences expérimentales, où le contexte du laboratoire permet de « contrôler » le niveau des variables qui pourraient influencer la relation à l'étude. Par exemple, si l'on fait l'essai d'un médicament sur des rats et que l'on croit que l'efficacité du médicament peut être influencée par l'alimentation, on fera des essais sur différents groupes de rats auxquels on appliquera différents régimes alimentaires « contrôlés ».

Une façon simple de tenir compte des variables de contrôle est d'examiner la relation à l'étude (tests d'hypothèse et mesure de l'intensité de la relation) séparément pour chaque groupe homogène d'individus. Dans notre exemple, cela voudrait dire examiner deux tableaux de contingence, un pour les femmes et un autre pour les hommes. Mais ce procédé comporte ses limites. En particulier, lorsqu'il y a plusieurs variables de contrôle, chacune ayant plusieurs catégories, le nombre de tableaux de contingence à analyser augmente rapidement. Par exemple, pour tenir compte du sexe et de l'âge, avec 5 tranches d'âge, il faut analyser 10 tableaux. De plus, quand le nombre d'observations est limité, le nombre d'observations dans plusieurs cellules peut être trop petit pour assurer la validité des tests (par exemple, avec 1000 observations, avec 5 professions, 5 zones de résidence, 5 groupes d'âge et 2 sexes, les fréquences théoriques seront de 4 *en moyenne* et il est fort probable que certaines cellules auront une fréquence théorique inférieure à 1).

Vu sous un autre angle, le problème est celui de la multiplicité des interactions possibles, qui augmente rapidement avec le nombre de variables (dimensions du tableau). Ainsi, dans un tableau à deux dimensions, il n'y a qu'une interaction possible et donc, qu'une hypothèse d'indépendance à tester. Dans un tableau à trois dimensions, il y a quatre interactions possibles : trois entre des paires de variables, et une entre les trois variables à la fois. Avec un tableau à quatre dimensions, il y a 17 interactions possibles (quatre pour chacun des quatre trios que l'on peut former en choisissant trois variables parmi quatre, plus une quadruple interaction impliquant toutes les variables)...

Le modèle *log-linéaire* constitue un cadre permettant d'examiner les différentes interactions possibles. Le modèle « saturé », qui inclut toutes les interactions possibles, reproduit parfaitement les données observées. Une version généralisée du test de l'hypothèse d'indépendance permet de sélectionner, parmi la multitude d'interactions possibles, lesquelles on doit retenir pour représenter la structure sous-jacente.

Pour en savoir plus long...

On peut consulter Upton (1981), où on trouvera une présentation informelle et pragmatique du modèle log-linéaire et un exemple de son utilisation dans le contexte des sciences régionales. Button *et al.* (1995) offrent un exemple plus récent d'utilisation du modèle log-linéaire.