

EXEMPLE D'UN TABLEAU DE CONTINGENCE

Données

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

Données triées par sexe, puis par couleur

	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

Tableau de contingence

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

POPULATION ACTIVE EMPLOYÉE DANS LA RÉGION MÉTROPOLITAINE DE MONTRÉAL ZONE DE RÉSIDENCE, SELON LE SEXE ET LA PROFESSION, 1991

Zone de résidence	Professions					TOTAL toutes professions
	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.	
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

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

Variable dépendante		Variables indépendantes	Méthode	
Aucune		2 variables catégoriques	Analyse de tableau de contingence	... à 2 dimensions
		Plus de 2 var. catégo.		... à plus de 2 dimensions
Continue		Continues ou discrètes	Régression multiple	
		Discrètes	Analyse de variance	
Catégorique	2 catégories	Continues ou discrètes	Logit ou probit	... binomial
	Plus de 2 cat.			... multinomial

SCHÉMA D'UN TABLEAU DE CONTINGENCE AVEC SUBDIVISION DES COLONNES

	Profession, sexe									
Zone de résidence	Directeurs, gérants, administrateurs et assimilés			Professionnels, enseignants et cols blancs spécialisés			...	TOTAL toutes professions		
	F	H	T	F	H	T		F	H	T
Montréal							...			
...							...			
Total							...			

TABLEAU 2
POPULATION ACTIVE EMPLOYÉE DANS LA RÉGION MÉTROPOLITAINE DE MONTRÉAL
ZONE DE RÉSIDENCE SELON LA PROFESSION, 1991

Zone de résidence	Professions						Répartition ($p_{i\bullet}$)
	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	
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	

FRÉQUENCES RELATIVES ET PROBABILITÉS DANS UN TABLEAU DE CONTINGENCE

NOTATION

Fréquences

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
$x_{\bullet\bullet} = \sum_i \sum_j x_{ij}$	nombre total d'observations dans le tableau

Exemples (Chap. 4-1, tableau 2)

$x_{23} = 107\ 822$, 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

Fréquences relatives (probabilités)

$p_{ij} = \frac{x_{ij}}{x_{\bullet\bullet}}$	fraction du nombre total d'observations qui appartient à la colonne j et qui se trouve dans la ligne i
$p_{i\bullet} = \frac{x_{i\bullet}}{x_{\bullet\bullet}}$	fraction du nombre total d'observations qui se trouve dans la ligne i
$p_{\bullet j} = \frac{x_{\bullet j}}{x_{\bullet\bullet}}$	fraction du nombre total d'observations qui se trouve dans la colonne j
$p_{j/i\bullet} = \frac{x_{ij}}{x_{i\bullet}}$	fraction des observations dans la ligne i qui appartient à la colonne j
$p_{i/\bullet j} = \frac{x_{ij}}{x_{\bullet j}}$	fraction des observations de la colonne j qui se trouve dans la ligne i

FRÉQUENCES RELATIVES ET PROBABILITÉS DANS UN TABLEAU DE CONTINGENCE DIFFÉRENTS CONCEPTS DE PROBABILITÉ

$p_{ij} = \frac{x_{ij}}{x_{..}}$ 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.

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

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

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

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

$p_{j/i.} = \frac{x_{ij}}{x_{i.}} = \frac{x_{ij}/x_{..}}{x_{i.}/x_{..}} = \frac{p_{ij}}{p_{i.}}$ est la probabilité **conditionnelle** d'appartenir à j , **étant**

donné que l'on appartient à i .

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

$p_{i/.j} = \frac{x_{ij}}{x_{.j}} = \frac{x_{ij}/x_{..}}{x_{.j}/x_{..}} = \frac{p_{ij}}{p_{.j}}$ est la probabilité **conditionnelle** d'appartenir à i , **étant**

donné que l'on appartient à j .

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

FRÉQUENCES RELATIVES ET PROBABILITÉS DANS UN TABLEAU DE CONTINGENCE IDENTITÉS

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

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

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

L'INDÉPENDANCE STATISTIQUE

TABLEAU 3

Population active employée dans la Région métropolitaine de Montréal

Répartition entre les zones de résidence selon la profession, 1991

Zone de résidence	Professions					TOTAL toutes professions
	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.	
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

TABLEAU 6

Population active employée dans la Région métropolitaine de Montréal

Composition professionnelle des zones de résidence, 1991

Zone de résidence	Professions					TOTAL toutes professions
	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.	
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

L'INDÉPENDANCE STATISTIQUE

TABLEAU 4

Population active employée dans la Région métropolitaine de Montréal

Fréquences théoriques dans l'hypothèse de distributions identiques

Zone de résidence	Professions					TOTAL toutes professions
	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.	
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

Exemple de calcul

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

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

Totaux de lignes et de colonnes égaux aux totaux observés

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

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

L'INDÉPENDANCE STATISTIQUE

TABLEAU 5

Population active employée dans la Région métropolitaine de Montréal

Calcul du Khi-deux

Zone de résidence	Professions					TOTAL toutes professions
	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.	
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

Exemple de calcul d'une cellule

Pour la cellule [5,4] (5e ligne, 4e colonne), $\frac{(x_{54} - x_{54}^*)^2}{x_{54}^*} = \frac{(25\,495 - 17\,427,4)^2}{17\,427,4} = 3\,734,7$

Calcul du Khi-deux

$$\chi^2 = \sum_i \sum_j \frac{(x_{ij} - x_{ij}^*)^2}{x_{ij}^*} = 19\,662,2$$

Détermination de la probabilité critique

SI les deux variables catégoriques sont indépendantes,

ALORS le χ^2 a une distribution asymptotique χ^2 avec $(C - 1)(L - 1)$ degrés de liberté
(C = nombre de colonnes et L = nombre de lignes du tableau)

χ^2 de Pearson =	19 662
Nombre de lignes =	5
Nombre de colonnes =	5
Degrés de liberté =	16
Probabilité critique =	inférieure à $2,4 \times 10^{-300}$

L'INDÉPENDANCE STATISTIQUE GÉNÉRALISATION

Définition

Un événement A est indépendant d'un autre événement B si la probabilité que l'événement A se réalise est la même, que B se réalise ou non.

Dans un tableau de contingence à 2 dimensions, les 2 variables catégoriques sont indépendantes si : $p_{i/\bullet j} = p_{i\bullet}$.

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

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

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

Fréquences théoriques sous l'hypothèse d'indépendance

Quelles fréquences s'attendrait-on à observer si les 2 variables catégoriques étaient indépendantes ?

Si on avait $p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$, on observerait x_{ij}^*

$$\frac{x_{ij}^*}{x_{\bullet\bullet}} = \left[\frac{x_{i\bullet}}{x_{\bullet\bullet}} \right] \left[\frac{x_{\bullet j}}{x_{\bullet\bullet}} \right], \text{ c'est-à-dire } x_{ij}^* = \frac{x_{i\bullet} \cdot x_{\bullet j}}{x_{\bullet\bullet}}$$

Le tableau des fréquences théoriques x_{ij}^* est un tableau *bi-proportionnel* :

- toutes les lignes i sont proportionnelles entre elles : $\frac{x_{ij}^*}{x_{i\bullet}} = \frac{x_{\bullet j}}{x_{\bullet\bullet}}$ pour toute ligne i
- toutes les colonnes j sont proportionnelles entre elles : $\frac{x_{ij}^*}{x_{\bullet j}} = \frac{x_{i\bullet}}{x_{\bullet\bullet}}$ pour toute colonne j

TABLEAU BI-PROPORTIONNEL DES FRÉQUENCES THÉORIQUES : NOMBRES ANTICIPÉS SOUS L'HYPOTHÈSE D'INDÉPENDANCE

Zone de résidence	Professions					TOTAL toutes professions
	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.	
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

Bi-proportionnel

- Chaque ligne est proportionnelle à la ligne « Total »
- Chaque colonne est proportionnelle à la colonne « Total »

Formule de calcul et exemple

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

$$x_{54}^* = \frac{x_{5\bullet} \cdot x_{\bullet 4}}{x_{\bullet\bullet}} = \frac{85178 \times 319548}{1561817} = 17427,4$$

Respect automatique des totaux marginaux

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

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

Équivalence des deux approches

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

LES MILLE ET UNE APPLICATIONS DU TEST DU KHI-DEUX DE PEARSON AUX TABLEAUX DE CONTINGENCE TEST SUR UNE SEULE CELLULE DU TABLEAU

Test du Khi-deux de Pearson avec 1 degré de liberté

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

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

X ² de Pearson =	0,11
Nombre de lignes =	2
Nombre de colonnes =	2
Degrés de liberté =	1
Probabilité critique	0,74

LES MILLE ET UNE APPLICATIONS... (SUITE)

TESTS D'HOMOGENÉITÉ

Test d'homogénéité entre deux ou plusieurs groupes ou échantillons

	Groupe A	Groupe B	Total A+B
Catégories			
Total			

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

	Groupe A	Reste de la population	Total
Catégories		Calculer par soustraction	
Total			

LES MILLE ET UNE APPLICATIONS... (SUITE)

TEST DE L'HYPOTHÈSE D'UNE DISTRIBUTION PARTICULIÈRE

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>

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

Calcul des fréquences théoriques

	Population Costa Rica 1984	Reste
Hommes	$\frac{(630\,995 - 640\,307)^2}{640\,307}$	$\frac{\{(0,5 Y - 630\,995) - [0,5 (Y - 1\,280\,614)]\}^2}{0,5 (Y - 1\,280\,614)} = \frac{(640\,307 - 630\,995)^2}{0,5 (Y - 1\,280\,614)}$
Femmes	$\frac{(649\,619 - 640\,307)^2}{640\,307}$	$\frac{\{(0,5 Y - 649\,619) - [0,5 (Y - 1\,280\,614)]\}^2}{0,5 (Y - 1\,280\,614)} = \frac{(640\,307 - 649\,619)^2}{0,5 (Y - 1\,280\,614)}$

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

Probabilité critique = $7,4 \times 10^{-61}$

CONDITIONS DE VALIDITÉ DU KHI-DEUX DE PEARSON

La distribution du χ^2 , comme distribution asymptotique de la statistique du Khi-deux de Pearson, est généralement considérée comme une approximation satisfaisante si :

$$x_{..} > 10 \times L \times C \text{ (Legendre et Legendre, 1998, p. 218)}$$

(C est le nombre de colonnes et L, le nombre de lignes du tableau)

Cette condition est suffisante, mais elle n'est pas nécessaire !

Comme règle empirique, on peut appliquer la règle suivante :

SI il y a une ou plusieurs cellules avec une **fréquence théorique < 5**

ALORS le test du Khi-deux risque de ne pas être valide

Que faire si le test n'est pas valide ?

- **Regrouper des catégories** : fusionner des lignes ou colonnes ne contenant qu'un petit nombre d'observations pour obtenir des fréquences théoriques plus élevées.

ATTENTION ! Le regroupement doit se faire de façon scientifiquement pertinente.

- **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)
- En dernier recours, écarter de l'analyse des **catégories qui ne peuvent pas être regroupées de façon scientifiquement pertinente.**

PROPRIÉTÉS NUMÉRIQUES DU KHI-DEUX DE PEARSON

Propriétés :

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

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

PROPRIÉTÉS NUMÉRIQUES DU χ^2 DE PEARSON

DÉMONSTRATION DE

$$\chi^2 \leq x_{..} \text{Min}(L - 1, C - 1)$$

La démonstration fait appel à une autre formule de calcul du Khi-deux

$$\begin{aligned} \chi^2 &= \sum_i \sum_j \frac{(x_{ij}^* - x_{ij})^2}{x_{ij}^*} = \sum_i \sum_j \frac{\left[(x_{ij}^*)^2 - 2x_{ij}^* x_{ij} + x_{ij}^2 \right]}{x_{ij}^*} \\ \chi^2 &= \sum_i \sum_j x_{ij}^* - 2 \sum_i \sum_j x_{ij} + \sum_i \sum_j \frac{x_{ij}^2}{x_{ij}^*} = x_{..} - 2x_{..} + \sum_i \sum_j \frac{x_{ij}^2}{\left(\frac{x_{i.} x_{.j}}{x_{..}} \right)} \end{aligned}$$

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

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

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

et que, d'autre part,

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

Il s'ensuit que

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

et

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

Q.E.D.

SENSIBILITÉ DU χ^2 AU NOMBRE D'OBSERVATIONS – ILLUSTRATION NUMÉRIQUE NO 1

Tableaux de contingence											
$p_{i\bullet}$				$p_{i\bullet}$				$p_{i\bullet}$			
15	10	25	0,5	30	20	50	0,5	60	40	100	0,5
10	15	25	0,5	20	30	50	0,5	40	60	100	0,5
25	25	50		50	50	100		100	100	200	
0,5	0,5		$\leftarrow p_{\bullet j}$	0,5	0,5		$\leftarrow p_{\bullet j}$	0,5	0,5		$\leftarrow p_{\bullet j}$

Fréquences théoriques												
12,5			12,5	25	25			25	50	50		
12,5			12,5	25	25			25	50	100		
25			25	50	50			50	100	200		
Calcul du Khi-deux												
0,5		0,5	1		1	2		2				
0,5		0,5	1		1	2		2				
Khi-deux =	2		Khi-deux =	4		Khi-deux =	8					
n.lignes=	2		n.lignes=	2		n.lignes=	2					
n. col.=	2		n. col.=	2		n. col.=	2					
d.l.=	1		d.l.=	1		d.l.=	1					
Prob.crit.=	0,157		Prob.crit.=	0,046		Prob.crit.=	0,005					

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

SENSIBILITÉ DU χ^2 AU NOMBRE D'OBSERVATIONS – ILLUSTRATION NUMÉRIQUE NO 2

Tableau 12 – Échantillon fictif
Population active employée dans la Région métropolitaine de Montréal
Zone de résidence selon la profession, 1991

Zone de résidence	Professions					TOTAL toutes professions
	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.	
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

χ^2 de Pearson =	19,79
Nombre de lignes =	5
Nombre de colonnes =	5
Degrés de liberté =	16
Probabilité critique	0,23

UN NOUVEAU REGARD SUR LE QUOTIENT DE LOCALISATION

Formule de calcul des quotients de localisation

$$QL_{ij} = \frac{\left(\frac{x_{ij}}{x_{\bullet j}} \right)}{\left(\frac{x_{i\bullet}}{x_{\bullet\bullet}} \right)} \text{ ou, de manière équivalente, } QL_{ij} = \frac{\left(\frac{x_{ij}}{x_{i\bullet}} \right)}{\left(\frac{x_{\bullet j}}{x_{\bullet\bullet}} \right)}$$

Fréquences théoriques dans un tableau de contingence de l'emploi par branche et par zone

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

Réinterprétation des quotients de localisation

$$QL_{ij} = \frac{\frac{x_{ij}}{\left(\frac{x_{i\bullet} \cdot x_{\bullet j}}{x_{\bullet\bullet}} \right)}}{\frac{x_{ij}}{x_{ij}^*}} \text{ ou } x_{ij}^* \cdot QL_{ij} = x_{ij}$$

Le Khi-deux de Pearson et les quotients de localisation

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

Contribution d'une cellule au Khi-deux

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

UN NOUVEAU REGARD SUR LE QUOTIENT DE LOCALISATION

DÉMONSTRATION DE

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

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

$$\chi^2 = \sum_i \sum_j \frac{(x_{ij}^* QL_{ij} - x_{ij}^*)^2}{x_{ij}^*}$$

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

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

Q.E.D.

MESURES DE L'INTENSITÉ DE LA RELATION ENTRE DEUX VARIABLES CATÉGORIQUES

MESURES DÉRIVÉES DU KHI-DEUX DE PEARSON

Propriétés numériques non désirables :

- X^2 augmente avec le nombre d'observations $x_{..}$
- $X^2 \leq x_{..} \text{Min}(L - 1, C - 1)$

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
- serait comprise entre 0 et 1 plutôt qu'entre 0 et $x_{..} \text{Min}(L - 1, C - 1)$.

Quelques mesures dérivées du X^2 de Pearson

1. $\phi^2 = \frac{X^2}{x_{..}}$

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.

2. Le T^2 de Tschuprov :

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

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

3. Le V^2 de Cramer :

$$V^2 = \frac{\phi^2}{\text{Min}(L-1, C-1)} = \frac{X^2}{x_{..} \text{Min}(L-1, C-1)}$$

- 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 $L \neq C$.
- Le V^2 de Cramer est équivalent au ϕ^2 et au T^2 de Tschuprow lorsque $L = C = 2$.
- **Le V^2 de Cramer est la mesure qui a les propriétés les plus désirables.**

MESURES DE L'INTENSITÉ DE LA RELATION ENTRE DEUX VARIABLES CATÉGORIQUES : AUTRES MESURES (*TAU* ET *LAMBDA*)

Principe général

Mesure de l'intensité de la relation = mesure de la capacité de prédiction

Quelle prédiction ?

Prédiction de la variable dépendante au moyen de la variable indépendante

(On fait donc une distinction entre la variable dépendante et la variable indépendante)

Exemple :

Sachant dans quelle zone de résidence demeure un individu, prédire à quelle catégorie professionnelle il appartient

Variable dépendante correspond aux colonnes j (professions)

Variable indépendante correspond aux lignes i (zones de résidence)

Comment mesurer la capacité de prédiction ?

réduction relative moyenne des erreurs d'assignation

Formule générale

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

Les deux mesures diffèrent quant à la règle d'assignation, règle suivie pour prédire à quelle catégorie j appartient un individu.

MESURES DE L'INTENSITÉ DE LA RELATION ENTRE DEUX VARIABLES CATÉGORIQUES : 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

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

Valeurs limites

Le tau est égal à zéro lorsque les deux variables sont parfaitement indépendantes, c'est-à-dire lorsque $p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$. 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 .

MESURES DE L'INTENSITÉ DE LA RELATION ENTRE DEUX VARIABLES CATÉGORIQUES : *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 $p_{\bullet Max} = \max_j p_{\bullet j}$, $p_{i,Max} = \max_j p_{ij}$ et $p_{Max|i\bullet} = \max_j p_{j|i\bullet}$.

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 $p_{i,Max} = \max_j p_{ij}$, on a $p_{ij} \neq p_{i\bullet} p_{\bullet j}$. Cette dernière propriété amène à préférer *tau*.

SEXE ET MOTIF DE MAGASINAGE ¹

Données		Sexe			
	Motif	H	F	PR	Tot.
0	Pas de réponse	1	1	0	2
1	L'unique motif est « Pour faire des achats »	17	10	0	27
2	Il y a plus d'un motif ET l'un d'eux est « Pour faire des achats »	6	11	0	17
3	L'unique motif est « Pour voir la marchandise... »	14	9	0	23
4	Il y a plus d'un motif ET l'un d'eux est « Pour voir la marchandise... » ET l'un d'eux n'est PAS « Pour faire des achats »	8	14	1	23
5	Dans tous les autres cas	34	37	0	71
	Total	80	82	1	163

Source : enquête du cours EUR-8112 à Place Versailles, automne 2000.

Fréquences théoriques		Sexe			
	Motif	H	F	PR	Tot.
0	Pas de réponse	1,0	1,0	0,0	2,0
1	L'unique motif est « Pour faire des achats »	13,3	13,6	0,2	27,0
2	Il y a plus d'un motif ET l'un d'eux est « Pour faire des achats »	8,3	8,6	0,1	17,0
3	L'unique motif est « Pour voir la marchandise... »	11,3	11,6	0,1	23,0
4	Il y a plus d'un motif ET l'un d'eux est « Pour voir la marchandise... » ET l'un d'eux n'est PAS « Pour faire des achats »	11,3	11,6	0,1	23,0
5	Dans tous les autres cas	34,8	35,7	0,4	71,0
	Total	80,0	82,0	1,0	163

Problème de validité du test :

- Il y a 3 cellules avec des fréquences observées < 5.
- Il y a plusieurs fréquences théoriques < 1 ; il y a 2 fréquences théoriques = 1.

Que faire ? Agréger ? Éliminer ?

¹ Plus exactement, motif de la présence à Place Versailles, selon la classification utilisée ici.

SEXE ET MOTIF DE PRÉSENCE À PLACE VERSAILLES

Données « nettoyées »		Sexe		
	Motif	H	F	Tot.
1	L'unique motif est « Pour faire des achats »	17	10	27
2	Il y a plus d'un motif ET l'un d'eux est « Pour faire des achats »	6	11	17
3	L'unique motif est « Pour voir la marchandise... »	14	9	23
4	Il y a plus d'un motif ET l'un d'eux est « Pour voir la marchandise... » ET l'un d'eux n'est PAS « Pour faire des achats »	8	14	22
5	Dans tous les autres cas	34	37	71
	Total	79	81	160

Source : enquête du cours EUR-8112 à Place Versailles, automne 2000.

Fréquences théoriques

Les fréquences théoriques ne sont pas exactement égales à celles qu'on obtient avant le nettoyage des données, mais les différences sont toutes inférieures à 0,01.

		Sexe		
	Motif	H	F	Tot.
1	L'unique motif est « Pour faire des achats »	13,3	13,7	27,0
2	Il y a plus d'un motif ET l'un d'eux est « Pour faire des achats »	8,4	8,6	17,0
3	L'unique motif est « Pour voir la marchandise... »	11,4	11,6	23,0
4	Il y a plus d'un motif ET l'un d'eux est « Pour voir la marchandise... » ET l'un d'eux n'est PAS « Pour faire des achats »	10,9	11,1	22,0
5	Dans tous les autres cas	35,1	35,9	71,0
	Total	79,0	81,0	160,0

$$\text{Khi-2} = 6,111436$$

nombre de col. = 2 ; nombre de lignes = 5

nombre de degrés de liberté = $4 = (2-1) \times (5-1)$

Probabilité critique = 0,190979 (19 %)

Conclusion ?

- On ne peut pas rejeter l'hypothèse d'indépendance (la probabilité, 19 %, est trop grande)
- Les données ne nous permettent pas d'affirmer qu'il y a une différence Hommes/Femmes.