L'analyse des données d'enquêtes
- Rappels –

Régis Schlagdenhauffen

Attraction et indépendance entre des variables

- Le *tableau croisé* est l'outil de base du sociologue dès qu'il est confronté à des données d'enquête.
- L'outil principal pour étudier les relations entre variables qualitatives est le tableau croisé (parfois tri croisé). La question étant souvent de savoir **dans quelle mesure une variable dépend-elle de l'autre ?** Dès lors une des variables dépendante est "à expliquer", l'autre est la variable indépendante ou "explicative".
- Les conventions veulent que l'ont mette en ligne la variable préalablement connue (variable dite aussi « explicative ») et en colonne la variable nouvelle, celle dont on veut rendre compte (variable dite « à expliquer »), et de toujours calculer les pourcentages *en ligne*.

Calculer l'effectif théorique

- Deux variables sont indépendantes s'il n'existe pas de lien entre les deux.
- Pour savoir cela, il faut commencer par calculer les **effectifs théoriques** qui devraient composer chacune des cases du tableau si les deux variables étaient parfaitement indépendantes.
- De manière générale l'effectif théorique peut être calculé de la manière suivante dans un tableau:
- Effectif d'indépendance

= Produit des marges divisé par le total

= Effectif total de la colonne *Effectif total de la ligne / Effectif total de l'échantillon.

Illustration du calcul de l'effectif théorique

- ***Les hommes, les femmes et le jardinage…***

	S'occupe d'un jardin		
	Oui	Non	Total
Hommes	965	1439	2404
%	40,1	59,9	100
Femmes	1052	1541	2593
%	40,6	59,4	100
Total	2017	2980	4997
%	40,4	59,6	100

Légende : En ligne, les variables explicatives donc connues, en colonne, les variables à expliquer en effectifs (N=4997) et en-dessous, en %.

Interprétation: En moyenne environ 40% des gens s'occupent d'un jardin et cette proportion moyenne est celle des deux sexes. Cependant, si l'on regarde les choses de près, on doit dire que les hommes s'occupent moins d'un jardin que les femmes puisque ils s'en occupent pour 40,1% d'entre eux contre 40,6% chez les femmes. **Cet écart hommes/femmes est-il notable ou négligeable ?**
Pour le savoir, nous devons calculer les effectifs théoriques…

Ex de calcul pour « homme » + « oui » : (2017*2404)/4997=970,94

- En procédant au calcul des effectifs théoriques nous obtenons le tableau suivant qui met en lumière les effectifs observés, théoriques et les écarts entre les deux types d'effectifs

	S'occupe d'un jardin		
	Oui	Non	Total
Homme : observé	965	1439	2404
théorique	970,4	1433,6	
écart	-5,4	+5,4	
Femme : observé	1052	1541	2593
théorique	1046,6	1546,4	
écart	+5,4	-5,4	
Total	2017	2980	4997

Cependant, les résultats ici restent vagues pour nous indiquer s'il y a une différence en fonction du sexe du point de vue du rapport des hommes et des femmes au jardinage, cela d'autant plus que nous n'avons que des chiffres et non des pourcentages. Nous pouvons affiner la chose en distinguant « **jardin d'agrément** » et « **jardin potager** »

Exercice à partir du tableau suivant : Calculer les effectifs théoriques, les pourcentages d'écarts entre effectifs théoriques et observés et conclure

S'occupe d'un jardin d'agrément			
	Oui	Non	Total
Hommes	745	1659	2404
%	31,0	69,0	100
Femmes	932	1661	2593
%	35,9	64,1	100
Total	1677	3320	4997
%	33,6	66,4	100

- Procédons en attendant une analyse concernant le rapport hommes/femmes au « **jardin potager** ». On observe les résultats suivants:

	S'occupe d'un jardin potager		
	Oui	Non	Total
Hommes	671	1733	2404
%	27,9	72,1	100
Femmes	520	2073	2593
%	20,1	79,9	100
Total	1191	3806	4997
%	23,8	76,2	100

On calcule ensuite les effectifs et pourcentages théoriques puis la distance entre les pourcentages observés et théoriques, c'est à dire l'écart à l'indépendance.
Les résultats et calculs effectués sont consignés dans le tableau suivant …

	Oui observé	Oui théorique	Non observé	Non théorique	Ecart Oui	Ecart Non	Total
Hommes	671	(1191*2404/4997) = **573**	1733	1831			2404
%	27,9	(573/2404)= 23,8	72,1	76,2	**+4,1%**	**-4,1**	100
Femmes	520	**618**	2073	1975			2593
%	20,1	(618/2593)= 23,8	79,9	76,2	**-3,7%**	**+3,7**	100
Total	1191	**1191**	3806	3806			4997
%	23,8		76,2				100

Conclusion : Le « jardin potager » est marqué masculin avec tous les stéréotypes associés : nécessité de la force physique, production utile, alors que les stéréotypes associés au « jardin d'agrément » sont "la grâce et la délicatesse féminine" associées à la cueillette des roses.

La preuve par le jardin d'agrément …

S'occupe d'un jardin d'agrément			
	Oui	Non	Total
Hommes	745	1659	2404
%	31,0	69,0	100
Femmes	932	1661	2593
%	35,9	64,1	100
Total	1677	3320	4997
%	33,6	66,4	100

Puis, on procède au calcul de la distance entre effectif observé et théorique (c'est-à-dire le **calcul de l'écart à l'indépendance**)

	Oui (%)	Non (%)
Hommes $_{Obs}$	31	69
Hommes $_{Theo}$	34	66
Distance$_{(Obs-Theo)}$	**-3**	**+3**
Femmes $_{Obs}$	35,9	64,1
Femmes $_{Théo}$	33,5	64,5
Distance$_{(Obs-Theo)}$	**+0,4**	**-0,4**
Total	33,6	66,4

Une lecture rapide consiste à repérer le signe des écarts à l'indépendance (par le biais des écarts au pourcentage moyen). Elle nous donne la structure suivante :

S'occupe d'un jardin d'agrément		
	Oui	Non
Hommes	-	+
Femmes	+	-

Conclusion: On voit que les femmes sont plus nombreuses que les hommes à s'occuper d'un jardin d'agrément. **Le jardin d'agrément est marqué féminin**.

Calculer les effets du hasard …

L'écart pondéré

- Imaginons une population de 1000 étudiants classés selon leur série du bac et selon leur destination l'année suivante : université, classes préparatoires aux grandes écoles, IUT et formations professionnalisantes.

Série	Université	Classes prep.	Prof.	Total
Littéraire	130	20	50	200
Eco.et soc.	200	20	80	300
Scientifique	100	50	50	200
Tech.et pro.	70	10	220	300
Total	500	100	400	1000

- Dans ce tableau, isolons deux cases : "littéraires allant à l'université" et "scientifiques allant dans une classe préparatoire", et calculons pour chaque cas l'écart à l'indépendance.

	Littéraires Université	Scientifiques Classes préparatoires
Observé	130	50
Théorique	500 x 200 / 1000 = 100	100 x 200 / 1000 = 20
Ecart	130 - 100 = 30	50 - 20 = 30

- L'écart pour les deux cases est bien le même, égal à trente (30) individus.
- Question: **ce même écart à l'indépendance a-t-il la même signification dans les deux cas ?**

- La réponse est bien entendu NON car entre les littéraires et les scientifiques, le poids de l'écart n'est pas le même. Afin d'équilibrer les résultats ont fait appel à un **Coefficient de pondération** : Théorique/Ecart. Il est de 0,3 dans le premier cas et 1,5 dans le deuxième car 30/100=0,3 et 30/20=1,5.

	Littéraires Université	Scientifiques Classes préparatoires
Observé	130	50
Théorique	500 x 200 / 1000 = 100	100 x 200 / 1000 = 20
Ecart	130 - 100 = 30	50 - 20 = 30

- Grâce au coefficient de pondération, il devient possible de calculer **l'Ecart pondéré** qui est le produit de l'écart par un coefficient de pondération. Dans le premier cas, il est de 9 individus ; dans le deuxième de 45.

Rapport écart/théorique	0,3	1,5
Ecart pondéré	30 x 0,3 = 9	30 x 1,5 = 45

- L'écart pondéré est appelé le **khi-deux** d'une case.
- Le khi-deux d'une case est la même chose que l'écart pondéré : on peut ainsi considérer le khi-deux de toutes les cases d'un tableau comme la somme des écarts pondérés.
- Un écart apporte toujours une certaine information, mais pour l'indice khi-deux, elle est pondérée par son rapport au théorique qui l'amplifie ou la réduit selon le cas.

- **Ex**: Les sorties au cinéma selon le sexe (par mois)

Obs / Ind	Jamais	Moins d'une fois/mois	Plus d'une fois/mois	Total
Femme	261 / 248	180 / 168	69 / 94	510 / 510
Homme	225 / 238	150 / 162	116 / 91	491 / 491
Total	486 / 486	330 / 330	185 / 185	1001 / 1001

- Cela nous permet ensuite de calculer le tableau des Ecarts absolus.

Tableau des écarts absolus

	Jamais	Mois d'une fois/mois	Plus d'une fois/mois
Femme	+13 (=261-248)	+12	-25
Homme	-13	-&2	+25

- Toutefois il est difficile de juger de l'importance, c'est à dire de la force de l'attraction ou de la répulsion entre les modalités à la seule vue des écarts, d'où **l'écart relatif.**
- Ecart relatif = (Effectif observé - Effectif théorique) / Effectif théorique

Tableau des écarts relatifs

	Jamais	-1x/mois	+1x/mois
Feme	+0,052 =(261-248)/248	+0,071	-0,266
Homme	-0,055	-0,074	+0,275

- Nous venons de voir comment l'écart entre deux cases peut être mesuré. Soit par l'écart absolu qui exprime une distance "absolue" ou brute et qui présente un nombre d'individus (donc un poids), soit l'écart relatif qui exprime un lien de répulsion ou d'attraction et qui présente l'intensité de ce lien.
- Une bonne manière de faire la synthèse entre ces deux aspects est de tenir compte à la fois de l'écart absolu et de l'écart relatif en les multipliant.
- Cette multiplication des indicateurs permet de supprimer les signes négatifs. Elle a en plus le mérite de tenir compte de l'intensité d'un lien d'attraction ou de répulsion tout en pondérant l'intensité de ce lien par les effectifs sur lesquels elle porte afin de ne pas accorder trop d'importance à des liens portant sur des effectifs trop faibles et, inversement, d'accorder toute son importance aux intensités de liens portant sur de nombreux individus.

- ***Distance entre deux cases = écart absolu x écart relatif***

= (Effectif observé - effectif théorique) x (effectif observé - effectif théorique/effectif théorique)

= (effectif observé-effectif théorique)^2 / effectif théorique

- Cette formule vérifie les propriétés attendues d'une distance: elle est toujours positive et elle est d'autant plus élevée que les écarts (absolu et relatif) entre cellules sont grands.
- La distance du Khi^2 peut être définie comme la somme des distances entre cases, c'est-à-dire comme la somme des contributions de chaque cellule. Ces distances sont appelées « Contributions au Khi^2 ».

Tableau des contributions pour chaque cellule

	Jamais	-1x/mois	+1x/mois
Femme	0,68 $=(261-248)^2/248$ $=13^2/248$	0,86	6,65
Homme	0,71	0,89	6,87

Formellement, cette distance s'écrit

$$\text{Distance du Khi2} = \sum_{\text{ensemble des cellules}} (\textit{Effectif observé} - \textit{Effectif théorique})2/\ \textit{Effectif théorique}$$

- Aussi, plus le tableau est grand, plus le nombre de contributions sera grand et donc, mécaniquement, plus la distance du Khi^2 sera grande. **C'est ce qu'on appelle le « degré de liberté »** (*ddl*).
- ddl = (nombre de colonnes -1)x(nombre de lignes -1). Plus le tableau est de grande taille (nombreuses lignes, nombreuses colonnes), plus le *ddl* est élevé.
- Dans le cas de notre exemple, la distance du Khi2 vaut 0,68+0,86+6,65+0,71+0,89+6,87 = **16,66**

- Conclusion: La valeur de la distance se situe clairement du côté des valeurs qu'il est peu probable d'obtenir si les variables sont indépendantes. On peut donc dire que ce n'est pas le hasard qui explique la distribution des valeurs dans le tableau: une autre explication s'impose donc.
- Les variables ne sont donc pas indépendantes. L'hypothèse d'indépendance est rejetée. Donc, **il existe un lien entre le sexe et la fréquentation des cinémas** !

La procédure du χ^2 en 6 étapes

- ✓ 1) Formuler les hypothèses statistiques H_0 et H_1
- ✓ 2) Construire le tableau des fréquences théoriques à partir du tableau de contingence.
- ✓ 3) Comparer les écarts entre les fréquences observées et les fréquences théoriques, entre la situation réelle et la situation d'indépendance.
- ✓ 4) Mesurer l'intensité des écarts entre les fréquences observées et les fréquences théoriques : calcul du χ^2
- ✓ 5) Définir le seuil de signification, calculer le nombre de degré de liberté, déterminer la valeur critique du khi carré avec la table de distribution et définir la règle de décision
- ✓ 6) Appliquer la règle de décision : conclusion du test (Rejet ou acceptation de H0)

Mesurer l'intensité de la corrélation

- **Le Φ « Phi »**

Il est utile dans les **tableaux 2x2** avec des variables nominales. Sa formule est simple :

$$\Phi = \sqrt{\frac{\chi^2}{N}}$$

Plus il est proche 1, plus les variables sont dépendantes. Donc, plus est proche de 0, plus il y a indépendance!

- **Le V de Cramer**

Pour les tableaux supérieurs à 2x2

V de Cramer = √ (Khi-deux / Khi-deux max) = $V = \sqrt{\chi^2 / N(k-1)}$

ou k = est le plus petit du nombre de rangées
ou de de colonnes et N = l'effectif total.

Plus V est proche de zéro, plus il y a indépendance entre les deux variables. Il vaut 1 en cas de complète dépendance puisque le χ^2 est alors égal au χ^2 max.

- FIN -