Statistiques à deux variables (partie 2)

I) Tableau de contingence

Nous étudions deux caractères simultanément sur chaque individu d'une population de taille n.
Les deux listes de valeurs des caractères forment deux variables X et Y.

Deux mises en forme des résultats peuvent être employées, suivant l'étude menée :

1. Deux séries de valeurs données en listes (voir la partie 1 du cours).
2. Deux séries de valeurs données ainsi que des **effectifs** (ce qui nous intéresse dans ce chapitre).

On a dans ce deuxième cas un **tableau de contingence**, de la forme suivante par exemple :

Y \ X	x_1	x_2	x_3
y_1	n_{11}	n_{12}	n_{13}
y_2	n_{21}	n_{22}	n_{23}
y_3	n_{31}	n_{32}	n_{33}

-en colonnes (par exemple) les différentes valeurs de X : les x_i.
-en lignes les différents valeurs de Y : les y_j (pas forcément en même nombre que celles de X).
-en contenu (dans "chaque case") : l'effectif correspondant à x_i et y_j, que l'on notera alors n_{ji}.

Exemple : Pour l'élection présidentielle, deux candidats sont en présence, Monsieur A et Madame B.
Dans un village de Bretagne, il y a 500 électeurs, dont 50 sont retraités, 100 sont chômeurs, et 350 sont actifs. Les résultats des élections sont les suivants :

Y (Électeurs) \ X (Candidats)	Monsieur A	Madame B	Blancs/Abstentions
Retraités	24	16	10
Actifs	122	148	80
Chômeurs	36	27	37

On peut ensuite calculer les sous-totaux des colonnes et des lignes : cela s'appelle les **effectifs marginaux** , ainsi que l'effectif total (ici 500).
Les **effectifs marginaux** correspondent à chaque valeur de X : on trouve ici 182, 191 et 127.
Les **effectifs marginaux** correspondent à chaque valeur de Y : on trouve ici 50, 350 et 100.

Y (Électeurs) \ X (Candidats)	Monsieur A	Madame B	Blancs/Abstentions	Effectifs marginaux
Retraités	24	16	10	50
Actifs	122	148	80	350
Chômeurs	36	27	37	100
Effectifs marginaux	182	191	127	Total=500

Grâce à ces données, on peut ensuite calculer $E(X)$, $V(X)$, $E(Y)$ et $V(Y)$, etc (mais vu les contraintes cette année, nous ne développerons pas cela).

II) Test du χ^2

Le test du carré de contingence χ^2 (que l'on prononce Khi deux) va nous permettre de déterminer la dépendance ou l'indépendance de deux variables qualitatives ou quantitatives, à partir de la distribution d'effectifs obtenue auprès d'un échantillon de répondants (sous forme de tableau de contingence).

A) Exemple

On veut vérifier s'il y a un lien entre le sexe et les résultats scolaires. Autrement dit : les filles travaillent-elles mieux que les garçons (ou inversement) ? Y a-t-il vraiment un lien entre résultats et le fait d'être un garçon ou une fille ?
On travaille à partir du tableau suivant, avec les **effectifs observés** :

Sexe / Résultats scolaires	Filles	Garçons	**Total**
Très faibles	8	20	28
Plutôt faibles	14	45	59
Plutôt forts	32	31	62
Très forts	30	20	50
Total	84	116	200

Il s'agit ici d'effectifs observés. On note O_{ij} une valeur numérique du tableau de contingence où i représente l'indice de la modalité prise par la première variable (ici les Résultats scolaires) et j représente l'indice de la modalité prise par la deuxième variable (ici le Sexe).
Par exemple, $O_{11} = 8$, $O_{12} = 20$, $O_{21} = 14$, $O_{41} = 30$, $O_{32} = 31$ etc.

B) Les deux hypothèses

On pose deux hypothèses :
-L'hypothèse nulle (notée H_0) : le sexe et les résultats scolaires sont indépendants.
-La contre-hypothèse (notée H_1) : il y a un lien entre les résultats scolaires et le sexe.

On considère que l'hypothèse H_0 est vraie tout au long du test. C'est uniquement à l'issue du test que l'on gardera cette hypothèse ou qu'on la rejettera.

C) Calculs des effectifs théoriques (calculés)

Puisqu'il y a 84 filles sur 200, il y a donc 42 % de filles.
Puisqu'il y a 116 garçons sur 200, il y a donc 58 % de garçons.

On calcule ensuite le contenu "théorique" de chaque case, à partir de la fréquence des filles et des garçons.
Par exemple, il y a 28 élèves dans la catégorie "Très faibles", et 42% de filles. Or $28 \times 42 \div 100 = 11,76$, la valeur théorique des filles "Très faibles" est donc 11,76.
De même, il y a 28 élèves dans la catégorie "Très faibles", et 58% de garçons. Or $28 \times 58 \div 100 = 16,24$, la valeur théorique des garçons "Très faibles" est donc 16,24.

On obtient ainsi un tableau avec les **effectifs théoriques** :

Résultats scolaires \ Sexe	Filles	Garçons	**Total**
Très faibles	11,76	16,24	28
Plutôt faibles	24,78	34,22	59
Plutôt forts	26,46	36,54	62
Très forts	21,00	29,00	50
Total	84 (soit 42%)	116 (soit 58%)	200

Il s'agit ici d'effectifs calculés. On note C_{ij} une valeur numérique du tableau de contingence où i représente l'indice de la modalité prise par la première variable (ici les Résultats scolaires) et j représente l'indice de la modalité prise par la deuxième variable (ici le Sexe).
Par exemple, $C_{11} = 11,76$, $C_{12} = 16,24$, $C_{21} = 24,78$, $C_{41} = 21,00$, $C_{32} = 36,54$ etc.

On peut regrouper ces données (effectifs observés O_{ij} et effectifs calculés C_{ij}), ce qui nous facilitera le travail pour la suite :

Filles		Garçons	
8	11,76	20	16,24
14	24,78	45	34,22
32	26,46	31	36,54
30	21,00	20	29,00

D) Calcul du χ^2

Il s'agit maintenant de comparer les effectifs observés (notés O_{ij}) et les effectifs théoriques calculés (notés C_{ij}).

On calcule le carré le contingence χ^2 grâce à la formule :

$$\chi^2 = \sum_{i=1}^{\text{nbe lignes}} \sum_{j=1}^{\text{nbe colonnes}} \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$$

En pratique, ce n'est pas difficile à calculer (mais un peu long). Sur notre exemple :

$$\frac{(8-11,76)^2}{11,76} \approx 1,202 \qquad \frac{(20-16,24)^2}{16,24} \approx 0,871$$
$$\frac{(14-24,78)^2}{24,78} \approx 4,690 \qquad \frac{(45-34,22)^2}{34,22} \approx 3,396$$
$$\frac{(32-26,46)^2}{26,46} \approx 1,160 \qquad \frac{(31-36,54)^2}{36,54} \approx 0,840$$
$$\frac{(30-21,00)^2}{21,00} \approx 3,857 \qquad \frac{(20-29,00)^2}{29,00} \approx 2,793$$

On additionne ensuite tous ces résultats, et on obtient :

$\chi^2 \approx 1,202 + 4,689 + 1,160 + 3,857 + 0,871 + 3,396 + 0,840 + 2,793.$

On trouve $\chi^2 \approx 18,809$.

E) Seuil de signification

Le seuil de signification α est (pour simplifier) la marge d'erreur que l'on est près à accepter pour rejeter ou retenir l'hypothèse H_0.

Le **seuil de signification** α est souvent choisi à 5% (qui est un seuil assez classique), c'est à dire que le risque est de 5% de conclure à tort qu'il existe une dépendance entre les deux variables étudiées.

On utilise bien sûr d'autres seuils, comme nous le verrons en exercice.

F) Degré de liberté

Dans un tableau de contingence, le degré de liberté dd_l sera obtenu de la façon suivante :

$$dd_l = \text{(nombre de lignes -1)}\times\text{(nombre de colonnes -1)}.$$

Dans notre exemple, il y a 4 lignes (Très faibles, Plutôt faibles, Plutôt forts, Très forts) et deux colonnes (Filles, Garçons).
Donc $dd_l = (4-1)\times(2-1) = 3\times 1 = 3$.

G) Règle de décision et interprétation finale

Que faire de ce résultat du χ^2 ? Comment l'interpréter ?

Plus la valeur de χ^2 est élevé, plus le degré d'association entre les deux variables est grande (plus ils sont "dépendants" l'un de l'autre).

On va ensuite utiliser une table de valeurs du χ^2 critique (noté χ^2_c), que l'on trouve dans les livres (ou les logiciels de statistiques), dont voici un exemple :

Table des valeurs du χ^2_c				
α / ddl	0,01	0,02	0,05	0,10
1	6,64	5,41	3,84	2,71
2	9,21	7,82	5,99	4,61
3	11,3	9,84	7,82	6,25
4	13,3	11,7	9,49	7,78
5	15,1	13,4	11,1	9,24
6	16,8	15,0	12,6	10,6
7	18,5	16,6	14,1	12,0
8	20,1	18,2	15,5	13,4
9	21,7	19,7	16,9	14,7
10	23,2	21,2	18,3	16,0

On peut maintenant conclure notre test d'indépendance. On accepte ou on rejette l'hypothèse nulle H_0 selon la règle de décision suivante :

Règle de décision :

-Si $\chi^2 < \chi^2_c$, on accepte l'hypothèse de départ H_0 (les variables sont indépendantes).

-Sinon, on rejette l'hypothèse de départ H_0, et on accepte l'hypothèse H_1 (les variables sont dépendantes).

On a choisi $\alpha = 5\ \%$ (soit 0,05) dans cet exemple, et on a trouvé dans notre exemple $dd_l = 3$ et $\chi^2 = 18,809$. D'après la table, le χ^2_c théorique (ou critique) est $7,82$.

On a dans notre exemple $\chi^2 > \chi^2_c$. On rejette donc l'hypothèse de départ, les variables sont donc dépendantes !

Il y a donc une dépendance (avec une marge d'incertitude) entre les notes et le fait d'être un garçon ou une fille. Les notes (dans notre exemple) semblent liées au fait d'être un garçon ou une fille.

Je vous conseille fortement d'aller voir ces deux vidéos afin de compléter/revoir cette notion du test du χ^2 dont je me suis très fortement inspiré pour ce cours.

https ://urlz.fr/fpvQ

https ://urlz.fr/fpvU