Faculté des sciences exactes et sciences de la nature et de la vie

Département : Sciences exactes et sciences de la nature et de la vie de

Module : Biostatistiques (2019-2020)

Chapitre I : Statistique descriptive bivariée

1- Introduction

Une population ou un échantillon d'individus peuvent être décrits par une ou plusieurs variables à la fois. Le but de la description dans le cas de plusieurs variables, est de rechercher une liaison ou association entre ces variables au sein de la population ou l'échantillon choisi. Dans ce chapitre, l'étude sera limitée au cas de deux variables qui peuvent être toutes les deux quantitatives, qualitatives, ou l'une quantitative et l'autre qualitative (cas mixte).

2 – Présentation des données

2-1- Cas de deux variables quantitatives X et Y

Après avoir déterminé le nombre k de classes pour chacune des deux variables, on affecte les n couples de valeurs observées (Xi,Yi) aux classes croisées correspondantes. La synthèse de cette opération de répartition se présente sous la forme d'un tableau à deux entrées (ou tableau de lignes pour les classes c_i de X et colonnes pour les classes b_j de Y) appelé **tableau de contingence.** Chaque case du tableau contient la **fréquence absolue croisée** (nombre) n_{ij} d'individus pour lesquels la valeur Xi $\in$ à la classe c_i (i=1,2,3....,p) de X et la valeur Yi $\in$ à la classe b_j (j=1,2,3....,q) de Y pour le couple (Xi,Yi), comme indiqué dans le tableau suivant :

Classe de X	**Classes de Y**						**Total**
	b_1	b_2	:	:	:	b_q	
c_1	n_{11}	n_{12}				n_{1p}	$n_{1\bullet}$
c_2	n_{21}	n_{22}				n_{2p}	$n_{2\bullet}$
:							
:							
:							
c_p	n_{p1}	n_{p2}				n_{pq}	$n_{p\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$				$n_{\bullet q}$	$n_{\bullet\bullet}$

Remarque : l'indice i (classes de X) est l'indice des lignes, j (classes de Y) est l'indice des colonnes. Les valeurs c_i et b_j sont les centres de classes.

2-2 - Caractéristiques du tableau :

1 –fréquences absolues croisées : n_{ij} avec : $\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}=n=n_{\bullet\bullet}$

2 – fréquences absolues marginales de X : pour chaque classe i : $\sum_{j=1}^{q} n_{ij}=n_{i\bullet}$ (sommation sur la ligne i, donnent la distribution de X seulement) et : $\sum_{i=1}^{p} n_{i\bullet}=n_{\bullet\bullet}$.

3 - fréquences absolues marginales de Y : pour chaque classe j : $\sum_{i=1}^{p} n_{ij}=n_{\bullet j}$ (sommation sur la colonne j, donnent la distribution de Y seulement) et : $\sum_{j=1}^{q} n_{\bullet j}=n_{\bullet\bullet}$

4- fréquences relatives croisées : $f_{ij}=n_{ij}/n_{\bullet\bullet}$, avec : $\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}=1$

5- fréquences relatives marginales de X : pour chaque classe i : $\sum_{j=1}^{q} f_{ij}=f_{i\bullet}$ (sommation sur la ligne i , donnent la distribution de X seulement) avec : $\sum_{i=1}^{p} f_{i\bullet}=1$.

6- fréquences relatives marginales de Y : pour chaque classe j : $\sum_{i=1}^{p} f_{ij}=f_{\bullet j}$ (sommation sur la colonne j, donnent la distribution de Y seulement) avec : $\sum_{j=1}^{q} f_{\bullet j}=1$.

Remarque :

Ces différentes caractéristiques peuvent être représentées graphiquement à l'aide d'histogrammes.

2-3 - Caractéristiques sur les variables :

- **Moyenne et écart-type globaux de X :**

$$\overline{X}=\frac{1}{n_{\bullet\bullet}}\left(\sum_{i=1}^{p} n_{i\bullet}\times c_i\right)=\sum_{i=1}^{p} f_{i\bullet}\times c_i$$

$$\sigma_X=\sqrt{\frac{1}{n_{\bullet\bullet}}\left(\sum_{i=1}^{p} n_{i\bullet}\times(c_i-\overline{X})^2\right)}$$

- **Moyenne et écart-types globaux de Y :**

$$\overline{Y}=\frac{1}{n_{\bullet\bullet}}\left(\sum_{j=1}^{q} n_{\bullet j}\times b_j\right)=\sum_{j=1}^{q} f_{\bullet j}\times b_j$$

$$\sigma_Y=\sqrt{\frac{1}{n_{\bullet\bullet}}\left(\sum_{j=1}^{q} n_{\bullet j}\times(b_j-\overline{Y})^2\right)}$$

2-4 - Mesures de liaison entre les variables X et Y

Ces mesures permettent d'indiquer le degré de liaison entre les variables X et Y chez les individus de la population ou l'échantillon considérés.

a)- le coefficient de corrélation :

Il est calculé par l'expression mathématique suivante :

$$r(X,Y)=\frac{Cov(X,Y)}{\sigma_X\times\sigma_Y}=\frac{\frac{1}{n_{\bullet\bullet}}\left(\sum_{i=1}^{p}\sum_{j=1}^{q}n_{ij}\times(c_i-\overline{X})\times(b_j-\overline{Y})\right)}{\sigma_X\times\sigma_Y}=\frac{\frac{1}{n_{\bullet\bullet}}\left(\sum_{i=1}^{p}\sum_{j=1}^{q}n_{ij}\times c_i\times b_j\right)-\overline{X}\times\overline{Y}}{\sigma_X\times\sigma_Y}$$

Le terme $Cov(X,Y)$ désigne covariance entre X et Y appelée **<u>variance mixte</u>**.

<u>Remarque importante</u> : $\boxed{-1\le r(X,Y)\le 1}$

- Si : $0\le r(X,Y)\le 1$ alors : X et Y sont positivement corrélées, elles varient dans le même sens chez les individus de la population.
- Si : $-1\le r(X,Y)\le 0$ alors : X et Y sont négativement corrélées, elles varient en sens opposé chez les individus de la population.

3- Cas de deux variables qualitatives A et B

Si on considère deux variables qualitatives, l'une notée A à p catégories (modalités) ou classes dénommées A1, A2,…, Ap ; et l'autre B à q catégories ou modalités dénomméesB1, B2 ; …, Bq respectivement sur un échantillon de n individus, les observations obtenues peuvent etre groupées en classes croisées selon le tableau suivant :

Classe de B **Classe de A**	B1	B2	..	Bj	..	Bq	**Effectifs marginaux**
A1	n_{11}	n_{12}	..	n_{1j}	..	n_{1p}	$n_{1\bullet}$
A2	n_{21}	n_{22}	..	n_{2j}	..	n_{2p}	$n_{2\bullet}$
..	..	..	..	..	..	..	..
Ai	n_{i1}	n_{i2}		nij		nip	
..	..	..	..	..	..	..	..
Ap	n_{p1}	n_{p2}	..	..	..	n_{pq}	$n_{p\bullet}$

Effectifs marginaux	$n_{\bullet 1}$	$n_{\bullet 2}$	..	..	..	$n_{\bullet q}$	**n**

La distribution de ces deux variables est synthétisée par le même type de table précédent et on peut calculer les mêmes caractéristiques sur ce tableau (fréquences absolues croisées, fréquences absolues marginales, fréquences relatives croisées, fréquences relatives marginales).

- **Mesures d'association entre deux variables qualitatives**

a) Le coefficient χ^2 (khi-deux) de vraisemblance du lien de Pearson : Il mesure le lien par la quantité d'information mutuelle échangée entre les deux variables. Il est calculé par :

$$L(A,B) = \chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\left[\frac{\left(n_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n_{\bullet\bullet}}}\right]$$

b) Le coefficient ϕ (phi) : $\phi = \sqrt{\chi^2 / n_{\bullet\bullet}}$ **(attention :** $0 \le \phi \le 1$**)**

3- Cas mixte (une variable quant. X et une variable qualit. A)

Si la variable qualitative possède p catégories A_1 , A_2 , …., A_p , les observations de X pour les différentes catégories de A peuvent être résumées par le tableau suivant :

Classes de A	Observations de X ou Echantillons de X (valeurs de X)	Total	moyenne
A_1	X_{11}, X_{12} , ……., X_{1n1}	$\sum_{j=1}^{n1} X_{1j}$	$\overline{X}_1$
A_2	X_{21}, X_{22}, ……., X_{2n2}	$\sum_{j=1}^{n2} X_{2j}$	$\overline{X}_2$
:			
:			
:			
A_p	X_{p1} , X_{p2} , ……., X_{pnp}	$\sum_{j=1}^{np} X_{pj}$	$\overline{X}_p$

Les effectifs n_1 , n_2 ,….., n_p représentent le nombre de valeurs de X (taille de l'échantillon) pour chaque catégorie A_i (i=1,2,….,p) de A.

Pour vérifier l'influence de la variable qualitative A (appelée **facteur**) sur la variation de X, on compare les écarts (variations) entre les moyennes catégorielles $\overline{X}_i$ de X et la moyenne globale $\overline{X}$ (ces variations sont appelées variations factorielles) et les écarts (variations résiduelles) résiduels entre les valeurs observées Xij de X et les moyennes catégorielles $\overline{X}_i$.

On effectue ce qu'on appelle une **analyse de variance** ou de covariance. Cela revient à calculer :

- **Variations factorielles** : $SCF=\sum_{i=1}^{p} n_i \times(\overline{X}_i-\overline{X})^2$, avec : $\overline{X}_i=(\sum_{j=1}^{ni} X_{ij})/n_i$, $\overline{X}=(\sum_{i=1}^{p}\sum_{j=1}^{ni} X_{ij})/n$, $\sum_{i=1}^{p} n_i=n$

- **Variations résiduelles** : $SCR=\sum_{i=1}^{p}\sum_{j=1}^{ni} (X_{ij}-\overline{X}_i)^2$

- **Le rapport de Fisher** : $F=\frac{SCF/(p-1)}{SCR/(n-p)}$

  L'effet réel ou significatif de la variable qualitative A sur la variable quantitative X se vérifie par ce rapport F qu'on compare avec des valeurs de référence (valeurs critiques) de la loi de Fisher qui sera étudiée ultérieurement.