

# Statistique descriptive bivariée

Marie-Ange Remiche

Université de Namur

# Conditions d'analyse

Soit  $(\underline{X}, \underline{Y})$  la distribution statistique d'un couple de variables mesurées sur un échantillon dont l'effectif total est  $n$ . Nous avons plusieurs manières de décrire cette distribution, à savoir

- sous forme de données brutes, soit  $(X_i, Y_i)$  pour  $i = 1, \dots, n$ ,
- sous forme de valeurs distinctes. Soit  $(x_1, \dots, x_p)$  avec  $x_1 < x_2 \dots < x_p$  pour la variable  $X$  et  $(y_1, \dots, y_q)$  avec  $y_1 < y_2 \dots < y_q$  pour la variable  $Y$ , nous avons alors le **tableau de contingence**

	$y_1$	$\dots$	$y_j$	$\dots$	$y_q$
$x_1$	$n_{11}$	$\dots$	$n_{1j}$	$\dots$	$n_{1q}$
$\vdots$					
$x_i$	$n_{i1}$	$\dots$	$n_{ij}$	$\dots$	$n_{iq}$
$\vdots$					
$x_p$	$n_{p1}$	$\dots$	$n_{pj}$	$\dots$	$n_{pq}$

$n_{ij}$  représente l'**effectif** de l'échantillon pour lequel sont observées la mesure  $x_i$  pour la variable  $X$  et la mesure  $y_j$  pour la variable  $Y$ , avec  $i \in \{1, \dots, p\}$  et  $j \in \{1, \dots, q\}$ .

Le tableau de contingence fournit ce qu'on appelle la **distribution jointe** de la série statistique  $(\underline{X}, \underline{Y})$ .

On peut aisément y adjoindre les totaux des lignes et des colonnes. On obtient alors la représentation suivante

	$y_1$	...	$y_j$	...	$y_q$	Total
$x_1$	$n_{11}$	...	$n_{1j}$	...	$n_{1q}$	$n_{1\bullet}$
$\vdots$						
$x_i$	$n_{i1}$	...	$n_{ij}$	...	$n_{iq}$	$n_{i\bullet}$
$\vdots$						
$x_p$	$n_{p1}$	...	$n_{pj}$	...	$n_{pq}$	$n_{p\bullet}$
Total	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet q}$	$n$

## Définition

- L'effectif marginal de  $x_i$  est donné par

$$n_{i\bullet} \stackrel{\text{def}}{=} \sum_{j=1}^q n_{ij},$$

celui de  $y_j$  est défini comme


$$n_{\bullet j} \stackrel{\text{def}}{=} \sum_{i=1}^p n_{ij},$$

## Définition

- la fréquence marginale de  $x_i$  est donnée par

$$f_{i\bullet} \stackrel{\text{def}}{=} \sum_{j=1}^q \frac{n_{ij}}{n},$$

*(Handwritten blue circle around  $\frac{n_{ij}}{n}$  and blue text  $\{i\}$  above the sum)*



celle de  $y_j$  est définie comme

$$f_{\bullet j} \stackrel{\text{def}}{=} \sum_{i=1}^p \frac{n_{ij}}{n},$$

## Définition

- la **distribution marginale des effectifs** de  $X$  est  $(x_i, n_{i\bullet})_{i=1,\dots,p}$ , celle de  $Y$  est  $(y_i, n_{\bullet i})_{i=1,\dots,q}$ ,
- la **distribution marginale des fréquences** de  $X$  est  $(x_i, f_{i\bullet})_{i=1,\dots,p}$ , celle de  $Y$  est  $(y_i, f_{\bullet i})_{i=1,\dots,q}$ , avec

$$f_{i\bullet} \stackrel{\text{def}}{=} \frac{n_{i\bullet}}{n},$$

$$f_{\bullet i} \stackrel{\text{def}}{=} \frac{n_{\bullet i}}{n}.$$

## Exemple

Nous avons les codes suivants

```
table(BD$age,BD$genre) → donne table contingence
```

```
addmargins(table(BD$age,BD$genre))
```

```
addmargins(prop.table(table(BD$age,BD$genre)))
```

## Définition

La **fréquence conditionnelle** de la valeur  $y_j$  sachant que  $X = x_i$ , notée  $f_{j|i}^{Y|X}$  est donnée par

$$\begin{aligned} f_{j|i}^{Y|X} &\stackrel{\text{def}}{=} \frac{n_{ij}}{n_{i\bullet}} \\ &= \frac{f_{ij}}{f_{i\bullet}} \end{aligned}$$

La **fréquence conditionnelle** de la valeur  $x_i$  sachant que  $Y = y_j$ , notée  $f_{i|j}^{X|Y}$  est donnée par

$$\begin{aligned} f_{i|j}^{X|Y} &\stackrel{\text{def}}{=} \frac{n_{ij}}{n_{\bullet j}} \\ &= \frac{f_{ij}}{f_{\bullet j}} \end{aligned}$$



## Exemple

Nous avons les codes suivants

```
prop.table(table(BD$age,BD$genre),2)  
addmargins(prop.table(table(BD$age,BD$genre),2),1)
```

```
AgeFnFemme<-  
  addmargins(prop.table(table(BD$age,BD$genre),2),1)  
AgeFnFemme[,1]
```

```
plot(x=(rownames(AgeFnFemme)[1:(nrow(AgeFnFemme)-1)]),  
      y=AgeFnFemme[1:(nrow(AgeFnFemme)-1),1],  
      type="l",xlab="Age",ylab="Fréquence conditionnelle")
```

## Exemple

Nous avons le diagramme en bâtons superposés pour les effectifs ou le diagramme en bâtons juxtaposés.

```
barplot ( table (BD$maux , BD$genre ) )
```

```
barplot ( table (BD$maux , BD$genre ) , beside=TRUE)
```

## Exemple

Nous avons le diagramme en bâtons superposés pour les fréquences conditionnelles ou le diagramme en bâtons juxtaposés.

```
Tf=prop.table(table(BD$maux,BD$genre),2)
```



```
barplot(Tf))
```

```
barplot(Tf, beside=TRUE)
```



# Une variable qualitative et une variable quantitative

## Exemple

```
boxplot (BD$age~BD$maux)
```

# Deux variables quantitatives

## Exemple

```
plot(BD$taille ,BD$poids)
```

## Définition

La **moyenne conditionnelle** de la variable  $X$  sachant que  $Y = y_j$  est définie par

$$\begin{aligned}\overline{X}_j &\stackrel{\text{def}}{=} \frac{1}{n_{\bullet j}} \sum_{i=1}^p n_{ij} x_i \\ &= \sum_{i=1}^p f_{i|j}^{X|Y} x_i.\end{aligned}$$

## Définition

La **moyenne conditionnelle** de la variable  $Y$  sachant que  $X = x_i$  est définie par

$$\begin{aligned}\overline{Y}_i &\stackrel{\text{def}}{=} \frac{1}{n_{i\bullet}} \sum_{j=1}^q n_{ij} y_j \\ &= \sum_{j=1}^q f_{j|i}^{Y|X} y_j.\end{aligned}$$

## Définition

La **variance conditionnelle** de la variable  $X$  sachant que  $Y = y_j$  est définie par

$$\begin{aligned} S_j^2(\underline{X}) &\stackrel{\text{def}}{=} \frac{1}{n_{\bullet j}} \sum_{i=1}^p n_{ij} (x_i - \bar{X}_j)^2 \\ &= \sum_{i=1}^p f_{i|j}^{X|Y} x_i^2 - \bar{X}_j^2. \end{aligned}$$



## Définition

La **variance conditionnelle** de la variable  $Y$  sachant que  $X = x_i$  est définie par

$$\begin{aligned} S_i^2(\underline{Y}) &\stackrel{\text{def}}{=} \frac{1}{n_{i\bullet}} \sum_{j=1}^q n_{ij} (y_j - \bar{Y}_i)^2 \\ &= \sum_{j=1}^q f_{j|i}^{Y|X} y_j^2 - \bar{Y}_i^2. \end{aligned}$$

## Définition

Les séries statistiques  $\underline{X}$  et  $\underline{Y}$  sont dites **indépendantes** si et seulement si pour tout  $i \in \{1, \dots, p\}$  et pour tout  $j \in \{1, \dots, q\}$ , nous avons

$$f_{i|j}^{X|Y} = f_{i\bullet}$$

$$f_{j|i}^{Y|X} = f_{\bullet j}.$$

## Propriété

Lorsque  $\underline{X}$  et  $\underline{Y}$  sont indépendantes, nous avons

$$f_{ij} = f_{i\bullet} f_{\bullet j},$$

pour tout  $i \in \{1, \dots, p\}$  et pour tout  $j \in \{1, \dots, q\}$ .

## Exemple

```
independance<-chisq.test( table(BD$maux,BD$bronchite))  
independance
```

```
independance$expected  
independance$observed
```

## Définition

La **covariance** entre  $\underline{X}$  et  $\underline{Y}$ , notée  $\text{Cov}(\underline{X}, \underline{Y})$  est définie de la manière suivante

$$\text{Cov}(\underline{X}, \underline{Y}) \stackrel{\text{def}}{=} \sum_{1 \leq i \leq p, 1 \leq j \leq q} (x_i - \bar{X})(y_j - \bar{Y}) f_{ij}$$

## Remarque

Une autre équation existe pour définir la notion de covariance, soit

$$\text{Cov}(\underline{X}, \underline{Y}) \stackrel{\text{def}}{=} \sum_{1 \leq i \leq p, 1 \leq j \leq q} x_i y_j f_{ij} - \bar{X} \bar{Y}.$$

## Définition

Le **coefficient de corrélation** entre  $\underline{X}$  et  $\underline{Y}$ , notée  $R(\underline{X}, \underline{Y})$  est défini de la manière suivante

$$R(\underline{X}, \underline{Y}) \stackrel{\text{def}}{=} \frac{\text{Cov}(\underline{X}, \underline{Y})}{S(\underline{X})S(\underline{Y})}.$$

## Propriété

Le coefficient de corrélation entre  $X$  et  $Y$  est tel que

$$|R(\underline{X}, \underline{Y})| \leq 1.$$

## Propriété

Soit  $\underline{X}$  et  $\underline{Y}$  deux séries statistiques. Leur covariance respecte les identités suivantes,

$$\text{Cov}(\underline{X}, \underline{Y}) = \text{Cov}(\underline{Y}, \underline{X})$$

$$\text{Cov}(\underline{X}, \underline{X}) = S^2(\underline{X})$$

$$|\text{Cov}(\underline{X}, \underline{Y})| \leq \sqrt{S^2(\underline{X}) S^2(\underline{Y})}$$

## Exemple

```
femme<-subset (BD,BD$genre==0)
cov(femme$taille ,femme$poids)
plot(femme$taille ,femme$poids)
cor(femme$taille ,femme$poids)
```

# Association entre deux variables nominales

## Remarque

Si il y a indépendance entre les deux variables

$$n_{jk}^* = \frac{n_{j\bullet} n_{\bullet k}}{n}.$$

On peut mesurer "l'écart" à l'indépendance comme

$$e_{jk} = n_{jk} - n_{jk}^*$$

## Définition

La **mesure du Khi-deux**, notée  $D^2$ , donne la mesure de l'association qui existe entre deux variables nominales. Elle est donnée par

$$D^2 \stackrel{\text{def}}{=} \sum_{j=1}^p \sum_{k=1}^q \frac{e_{jk}^2}{n_{jk}^*}$$



## Définition

Soit  $R(X)$  le rang attribué à la valeur  $X$ .

Le **coefficient de corrélation de rang de Spearman**, noté  $r_S$  est défini comme

$$r_S \stackrel{\text{def}}{=} \frac{1/n \sum_{i=1}^n (R(X_i) - \bar{R}_X)(R(Y_i) - \bar{R}_Y)}{\sqrt{(1/n \sum_{i=1}^n [R(X_i) - \bar{R}_X]^2)(1/n \sum_{i=1}^n [R(Y_i) - \bar{R}_Y]^2)}}$$

où

$$\begin{aligned}\bar{R}_X &\stackrel{\text{def}}{=} 1/n \sum_{i=1}^n R(X_i) \\ &= \frac{n+1}{2}\end{aligned}$$

## Propriété

Nous avons

$$r_S = 1 - \frac{6 \sum_{i=1}^n (R(X_i) - R(Y_i))^2}{n(n^2 - 1)}$$

et

$$-1 \leq r_S \leq 1.$$

## Théorème

La **droite de régression** de  $Y$  par rapport à  $X$  est la droite

$$Y = aX + b$$

Elle est définie de telle manière qu'elle rend la quantité  $d(a, b)$ , définie comme

$$d(a, b) \stackrel{\text{def}}{=} \sum_{i=1}^n (Y_i - aX_i - b)^2.$$

la plus petite possible.

Dans ce cas, on peut établir que

$$a = \frac{\text{Cov}(\underline{X}, \underline{Y})}{S^2(\underline{X})}$$

$$b = \bar{Y} - a\bar{X}.$$

# La pente de la droite de régression

Rappelons la définition du coefficient de corrélation entre  $X$  et  $Y$ ,

$$R(\underline{X}, \underline{Y}) = \frac{\text{Cov}(\underline{X}, \underline{Y})}{S(\underline{X})S(\underline{Y})}.$$

Ainsi, nous avons

$$a = R(\underline{X}, \underline{Y}) \frac{S(\underline{Y})}{S(\underline{X})}.$$

Le rapport  $S(\underline{Y})/S(\underline{X})$  étant positif, seul  $R(\underline{X}, \underline{Y})$  indique la pente de la droite de la régression linéaire selon la méthode des moindres carrés. Positif, la droite est croissante ; négatif, elle est décroissante.