Statistiques descriptives et Introduction à R

Partie 2 : Statistiques bivariées

Edmond Sanou

Université d'Evry Val d'Essonne

Mars 2023

Sommaire

Sommaire

Sommaire

Tableau de contingence

Soient deux caractères qualitatifs X et Y à respectivement p et q modalités observés sur une même population de taille n. On appelle **tableau de contingence** des caractères X et Y, le tableau représentant la répartition des individus statistiques en fonction des couples de modalités des deux caractères. Il s'agit donc de la matrice des effectifs (n_{ij}) vérifiant les modalités i du caractère X et j du caractère Y.

- Un tableau de contingence peut aussi être défini à partir de la matrice des fréquences $f_{ij} = \frac{n_{ij}}{n}$.
- Il est aussi appelé tableau à double entrée ou tableau croisé

Notations (1)

Pour les effectifs

- n_{ij} est l'effectif des individus présentant les modalités i et j des variables X et Y.
- $n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$ est l'effectif total des individus présentant la modalité i pour le caractère X
- $n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$ est l'effectif total des individus présentant la modalité j pour le caractère Y
- $n = n_{\bullet\bullet} = \sum_{i=1}^{p} n_{i\bullet} = \sum_{i=1}^{p} n_{\bullet j}$ est l'effectif total
- $n_{i\bullet}$ et $n_{\bullet j}$ sont appelés **effectifs marginaux**

Notations (2)

Pour les fréquences

- $f_{ij} = \frac{n_{ij}}{n}$ est la proportion d'individus présentant les modalités i et j des variables X et Y.
- $f_{i\bullet} = \frac{n_{i\bullet}}{n}$ est la proportion d'individus présentant la modalité i pour le caractère X.
- $f_{\bullet j} = \frac{n_{\bullet j}}{n}$ est la proportion d'individus présentant la modalité j pour le caractère Y.
- $f_{i\bullet}$ et $f_{\bullet j}$ sont appelées **fréquences marginales**

Représentation

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

Table 1: Tableau de contingence.

- On peut substituer les fréquences en divisant les effectifs par n.

Sous R (1)

Exemple : Région d'habitation et sport préféré

- Le tableau de contingence des effectifs est donné par

```
table_contingence <- table(Sport, Region)
```

- Affichage du tableau

```
table_contingence
```

```
##       Foot Rugby
## Nord   100    80
## Sud     60   120
```

- Ajout des effectifs marginaux

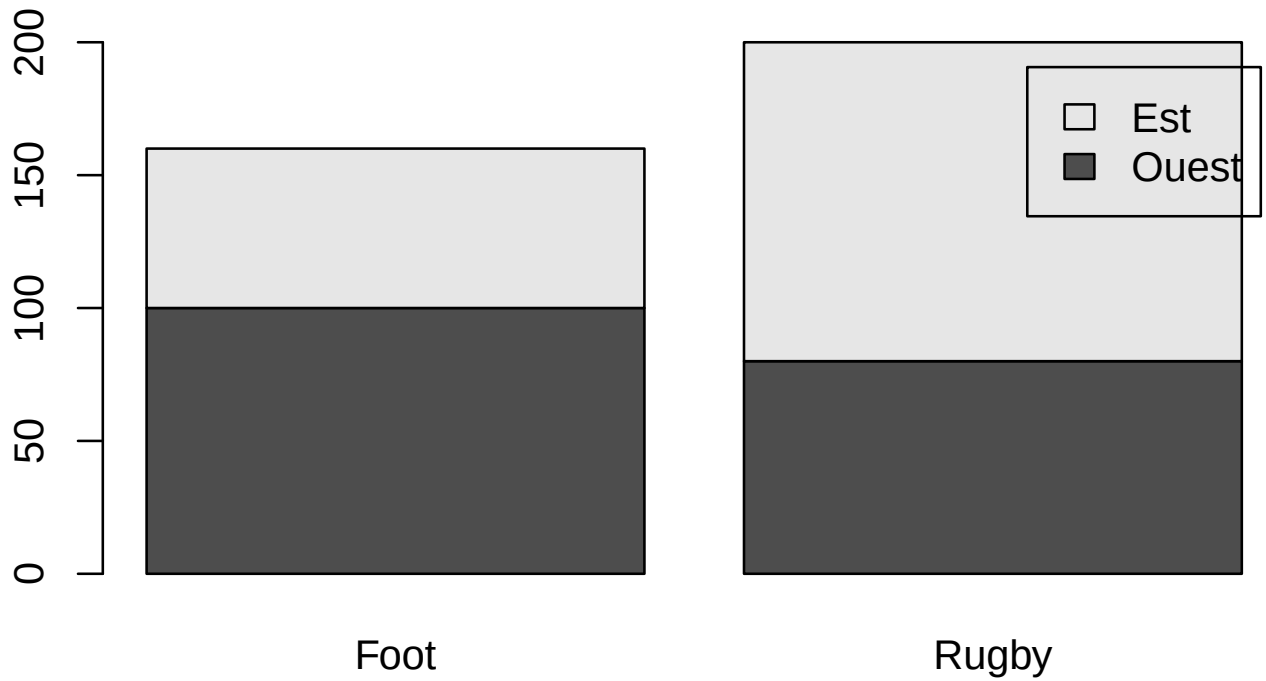
```
addmargins(table_contingence)
```

```
##       Foot Rugby Sum
## Nord   100    80 180
## Sud     60   120 180
## Sum    160   200 360
```

Sous R (2)

- Diagramme en barres des effectifs

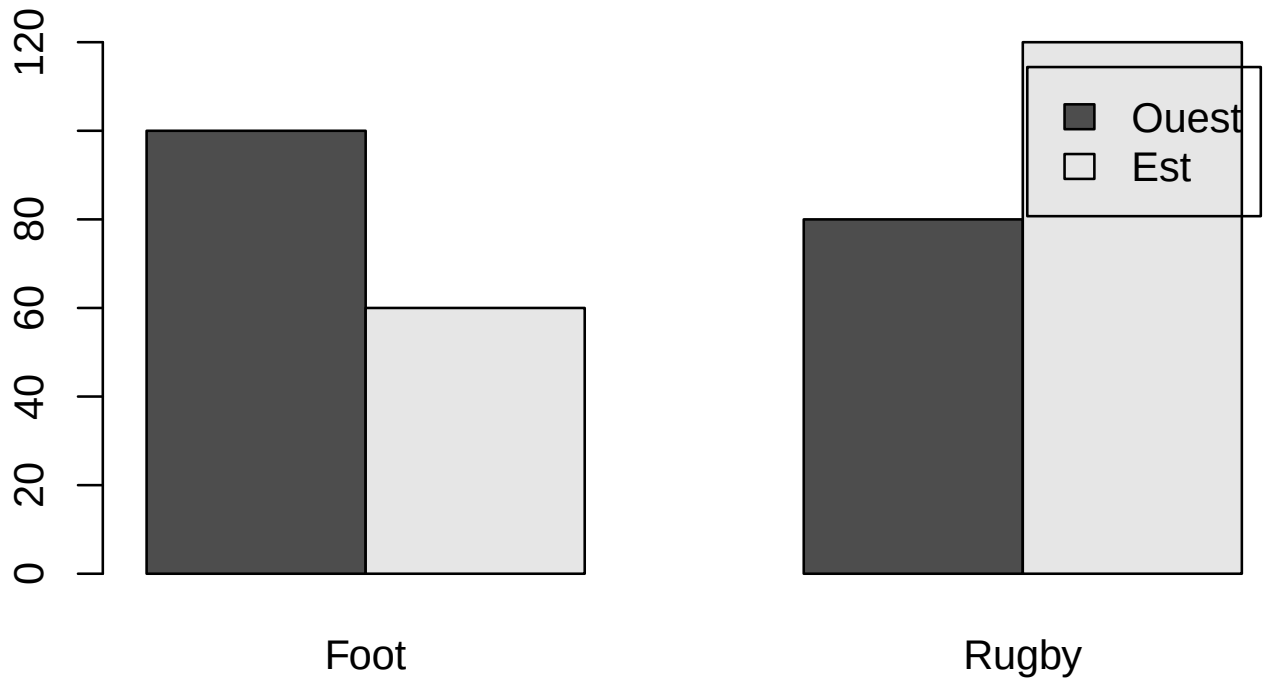
```
barplot(table_contingence,
        legend=rownames(table_contingence))
```



Sous R (3)

- Diagramme en barres des effectifs

```
barplot(table_contingence,
        beside = TRUE,
        legend=rownames(table_contingence))
```



Notion de dépendance

Imaginons le tableau de contingence incomplet suivant

	Eco	STID	Math/Eco	Maths	Ensemble
Hommes					58.3%
Femmes					41.7%
Ensemble	25%	3.3%	11.7%	60.0%	100%

Table 2: Répartition des étudiants selon leur recrutement

Que peut-on s'attendre à trouver dans les cases vides si le sexe et la filière de recrutement sont totalement indépendants l'un de l'autre ? A priori, on devrait trouver des valeurs telles que

$$f_{ij} = f_{i\bullet} f_{\bullet j} \text{ ou } n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}, \quad \forall (i,j) \in \{1,\dots,p\} \times \{1,\dots,q\}.$$

Statistique du χ^2 (1)

On appelle **statistique du** χ^2, la statistique valant

$$D^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - n_{ij}^{\star})^2}{n_{ij}^{\star}}$$

avec $n_{ij}^{\star} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ l'**effectif théorique** des individus vérifiant les modalités i et j des deux caractères considérés et n_{ij} l'**effectif observé ou empirique.**

Statistique du χ^2 (2)

Il est possible d'écrire D^2 comme fonction des fréquences et des fréquences marginales

$$D^2 = n \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}$$

- Cette écriture permet de voir que pour deux strucures identiques, la taille va avoir une influence sur la valeur de D^2.
- On peut montrer que $D^2 \leq n \min(p-1, q-1)$.

Statistiques dérivées du χ^2

On appelle **coefficient de Cramer** ou **V de Cramer,** la statistique

$$V = \sqrt{\frac{D^2}{n \min(p-1; q-1)}}$$

On appelle **coefficient de contingence de Pearson,** la statistique

$$C = \sqrt{\frac{D^2}{D^2 + n}}$$

On appele **coefficient de Tschuprow,** la statistique

$$T = \sqrt{\frac{D^2}{n\sqrt{(p-1)(q-1)}}}$$

Quelques remarques

- On V, C et $T \in [0, 1]$
- Avant de mesurer le degré de dépendance, il faut faire un test d'indépendance du χ^2 (voir Cours de Statistiques inférentielles).
- En pratique, aucun effectif théorique ne doit être inférieur à 5
- Une fois que l'indépendance est rejetée, V et C permettent de mesurer le degré de dépendance.
- Qui dit dépendance **ne dit pas** relation de causalité.

Contribution au χ^2

On appelle contribution au χ^2 du couple (i, j), la statistique

$$C_{\chi^2}(i,j) = \frac{1}{D^2} \frac{(n_{ij} - n_{ij}^\star)^2}{n_{ij}^\star}$$

- Cette statistique permet d'identifier les couples (i, j) qui contribuent le plus à la dépendance entre X et Y
- Il est aussi utile de déterminer le signe de cette dépendance
  - Si $n_{ij} - n_{ij}^\star > 0$ on parlera d'association positive
  - Si $n_{ij} - n_{ij}^\star < 0$ on parlera d'association négative

Distribution conditionnelle

On appelle $i-$ième **profil ligne** la distribution conditionnelle, notée $(f_{\bullet j|i\bullet})_{j=1,\ldots,q}$ avec

$$f_{\bullet j|i\bullet} = \frac{f_{ij}}{f_{i\bullet}} = \frac{n_{ij}}{n_{i\bullet}} \quad \forall j = 1, \ldots, q.$$

On appelle $j-$ième **profil colonne** la distribution conditionnelle, notée $(f_{i\bullet|\bullet})_{i=1,\ldots,p}$ avec

$$f_{i\bullet|\bullet j} = \frac{f_{ij}}{f_{\bullet j}} = \frac{n_{ij}}{n_{\bullet j}} \quad \forall i = 1, \ldots, p$$

Notion de représentativité (1)

Soit le groupe d'individus vérifiant la modalité du caractère X. On souhaite caractériser cette catégorie d'unités statistiques en fonction de la variable Y. Parmi cette sous-population, on dit que les individus de modalité j du caractère Y sont **sur-représentés** (respectivement sous-représentés) si

$$f_{\bullet j|i\bullet} > f_{\bullet j} \qquad \text{respectivement} \qquad f_{\bullet j|i\bullet} < f_{\bullet j}$$

$$\frac{n_{ij}}{n_{i\bullet}} > \frac{n_{\bullet j}}{n} \qquad \text{respectivement} \qquad \frac{n_{ij}}{n_{i\bullet}} < \frac{n_{\bullet j}}{n}$$

Remarque :

Il ne faut pas comparer les $(f_{\bullet j|i\bullet})$, repérer la modalité j pour laquelle $(f_{\bullet j|i\bullet})$ est maximale et conclure à la représentativité.

Notion de représentativité (2)

En 1997, les afro-américains représentaient 12.5% de la population américaine et 47% de la population carcérale. Que penser de l'affirmation suivante : *La majorité des prisonniers sont Afro-américains ?*

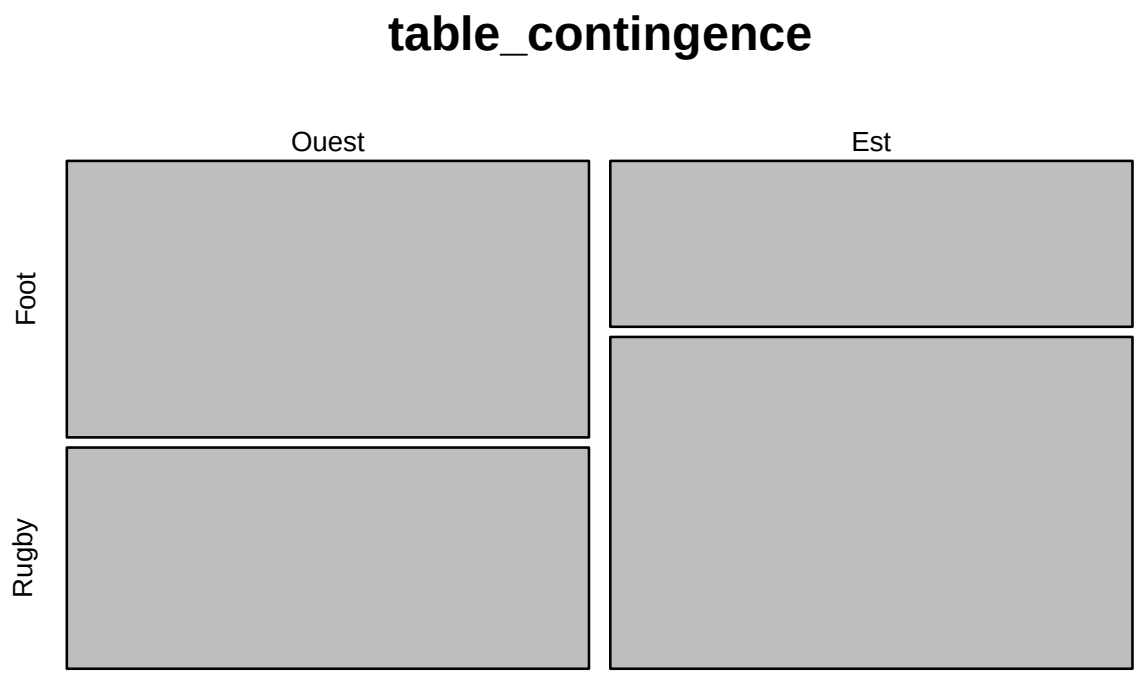
Représentation d'un couple de variables qualitatives : Diagramme mosaique

Ce diagramme vise à représenter un tableau de contingence avec des informations sur ses marges :

- Chaque colonne j possède une largeur proportionnelle à sa marge $n_{\bullet j}$
- Chaque case ij dans une colonne j possède une hauteur proportionnelle à $\frac{n_{ij}}{n_{\bullet j}}$.
- la surface de chaque case est donc proportionnelle à son effectif n_{ij}.

Diagramme mosaique (2)

```
plot(table_contingence)
```



Interprétation

- Si deux variables sont indépendantes les hauteurs des cases $i\bullet$ sont toujours les mêmes (proportionnelles à $n_{i\bullet}$).
- Plus le diagramme mosaïque semble être traversé de lignes horizontales, plus l'hypothèse d'indépendance semble vraisemblable.