

# Croisement d'une variable qualitative et d'une variable quantitative

A.B. Dufour & M. Royer

---

## Table des matières

<b>Introduction</b>	<b>1</b>
<b>Exercice 1</b>	<b>5</b>
<b>Exercice 2</b>	<b>6</b>
<b>Exercices supplémentaires</b>	<b>7</b>

## Introduction

Pour étudier la relation entre une variable qualitative et une variable quantitative, on décompose la variation totale en variation intergroupe et en variation intragroupe. Pour mesurer l'intensité de la relation (toujours d'un point de vue descriptif), on peut calculer un paramètre appelé rapport de corrélation.

### La notion de variation

La variance d'une variable quantitative peut être perçue selon le point de vue descriptif ou le point de vue inférentiel. Ce terme général peut désigner :

- la variance descriptive mesurée sur une groupe de  $n$  individus

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

- la variance estimée de la population à partir d'un échantillon de  $n$  individus

$$\widehat{\sigma^2} = \frac{n}{n-1} s^2$$

que l'on peut encore écrire

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

C'est pourquoi, on préférera travailler sur la variation totale c'est-à-dire la somme des carrés des écarts à la moyenne :

$$vartot = \sum_{i=1}^n (x_i - \bar{x})^2$$

Sous  $\mathbb{R}$ , nous écrirons  
soit :

```
vartot <- fonction(x) {
  res <- sum((x - mean(x))^2)
  return(res)
}
```

soit :

```
vartot2 <- fonction(x) {
  res <- var(x) * (length(x) - 1)
  return(res)
}
```

*Exercice.* Vérifier que les deux formules sont identiques. Considérons la variable quantitative  $X$ , note obtenue par 15 étudiants.

```
notes <- c(13, 11, 10, 11, 12, 5, 8, 7, 2, 4, 16, 17,
          13, 16, 15)
vartot(notes)
```

[1] 301.3333

```
vartot2(notes)
```

[1] 301.3333

## La notion de variation intergroupe

Reprenons la variable `note` précédente. Les 15 étudiants sont répartis dans  $p = 3$  groupes : (1) ceux qui ont suivi la moitié des cours, (2) ceux qui ne sont jamais venus, (3) ceux qui ont suivi tous les cours.

Maintenant, l'objectif est de savoir si la note est liée au choix des étudiants de participer ou non aux cours. Supposons que tous les étudiants aient la même note, qu'ils participent beaucoup, moyennement, ou pas du tout au cours. Alors cette valeur commune serait égale à la moyenne calculée sur l'échantillon global. Pour évaluer l'erreur réalisée si on considère que les étudiants ont la même note, on calcule le carré des écarts entre les valeurs mesurées et la moyenne globale, c'est-à-dire la variation totale vue ci-dessus.

Si on considère que la note dépend du choix des étudiants de participer ou non aux cours, alors la valeur est la moyenne du groupe d'appartenance. Pour évaluer l'erreur réalisée si on considère que la note des étudiants est liée au suivi des cours, on va calculer le carré des écarts entre la moyenne du groupe et la moyenne globale. Cette quantité est appelée *variation inter-groupes*.

$$\text{varinter} = \sum_{k=1}^p n_k (\bar{x}_k - \bar{x})^2$$

où  $\bar{x}_k$  désigne la note des étudiants du groupe  $k$  et  $n_k$ , le nombre d'étudiants appartenant à ce même groupe.

Sous  $\mathbb{R}$ , nous écrirons :

```
varinter <- fonction(x, gpe) {
  moyennes <- tapply(x, gpe, mean)
  effectifs <- tapply(x, gpe, length)
  res <- (sum(effectifs * (moyennes - mean(x))^2))
  return(res)
}
```

Notez que pour éviter de retaper une fonction plusieurs fois, vous pouvez utiliser dans le menu **Fichier**, l'item **Nouveau Script**.

*Exercice.* Considérons que les étudiants soient classés ici par groupe de 5. Calculer la variation inter-groupes.

```
suivi <- as.factor(rep(c("1", "2", "3"), rep(5, 3)))
varinter(notes, suivi)
```

[1] 264.1333

## Le rapport de corrélation

Pour étudier la relation entre une variable qualitative et une variable quantitative, on calcule le rapport de corrélation noté  $\eta^2$  :

$$\eta^2 = \frac{\sum_{k=1}^p n_k (\bar{x}_k - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Si le rapport est proche de 0, les deux variables ne sont pas liées.

Si le rapport est proche de 1, les variables sont liées.

Sous  $\mathbb{R}$ , nous écrirons :

```
eta2 <- fonction(x, gpe) {
  res <- varinter(x, gpe)/vartot(x)
  return(res)
}
```

*Exercice.* Que peut-on dire finalement des notes et du suivi des cours par les étudiants ?

```
eta2(notes, suivi)
```

[1] 0.8765487

Oui, nous le reconnaissons, l'exemple est un peu démagogique.

**Remarque.** D'une manière générale, la variation totale est la somme de la variation inter-groupes et de la variation intragroupe. Cette dernière est la somme pondérée des variances calculées à l'intérieur de chaque groupe.

$$\text{varintra} = \sum_{k=1}^p n_k s_k^2$$

où  $s_k^2$  est la variance descriptive au sein du groupe  $k$ . Elle s'obtient facilement par différence entre la variation totale et la variation inter-groupes. Nous avons alors la relation suivante, fondamentale en statistiques :

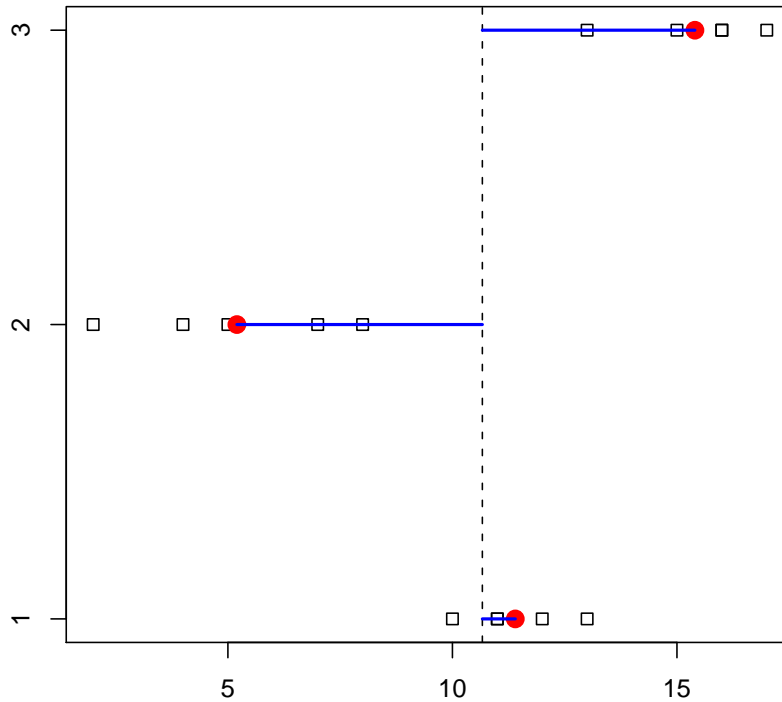
Variation Totale = Variation Inter-groupes + Variation intragroupe
--

### Représentation Graphique

Afin de bien visualiser la relation entre une variable quantitative et une variable qualitative, nous avons construit la représentation suivante.

- Les groupes sont représentés en vertical, la variable quantitative en horizontal
- Un carré blanc représente un individu
- Les points rouges représentent les moyennes dans chaque groupe
- La ligne en pointillé représente la moyenne de l'ensemble des individus
- Les traits bleus représentent les écarts entre les moyennes des groupes et la moyenne de l'ensemble soit une visualisation de la variation intergroupe.

```
graphnf <- function(x, gpe) {
  stripchart(x ~ gpe)
  points(tapply(x, gpe, mean), 1:length(levels(gpe)),
        col = "red", pch = 19, cex = 1.5)
  abline(v = mean(x), lty = 2)
  moyennes <- tapply(x, gpe, mean)
  traitnf <- function(n) segments(moyennes[n], n, mean(x),
    n, col = "blue", lwd = 2)
  sapply(1:length(levels(gpe)), traitnf)
}
graphnf(notes, suivi)
```



Remarquez que plus les traits bleus sont grands, plus les moyennes des groupes s'éloignent de la moyenne de l'ensemble et révèlent ainsi une différence c'est-à-dire un lien entre la variable qualitative et la variable quantitative.

*Exercice.* Modifier le vecteur `notes`, refaire les calculs et le graphique.

## Exercice 1

Afin de détecter de jeunes talents, il est intéressant de relever les particularités morphologiques des handballeurs de haut niveau.

Dufour et al. ont effectué en 1987 (1) différentes mesures chez des handballeurs de Nationales 1A, 1B et Nationale 2 ainsi que sur les appelés du contingent, sportifs de haut niveau rassemblés au bataillon de Joinville (BJ). Nous nous limiterons ici à la mesure de l'empan (distance du pouce à l'auriculaire, main et poignet posés bien à plat, doigts écartés au maximum) pour la main porteuse du ballon, en fonction du poste de jeu.

Les données sont dans le fichier "hand89.txt" que vous trouvez sur le site habituel : <http://pbil.univ-lyon1.fr/R/enseignement.fr>.

- 1) Calculer la moyenne et la variance de la longueur de l'empan pour l'ensemble des sportifs.
- 2) Calculer la moyenne de la longueur de l'empan en fonction de poste occupé par le sportif (utiliser la fonction `tapply`).
- 3) Faire la représentation graphique présentée dans l'introduction. Peut-on observer une différence ?
- 4) Vérifier que les écart-types sont du même ordre de grandeur quel que soit le poste.
- 5) On cherche à savoir s'il y a une relation entre la longueur de l'empan et le poste occupé par le joueur, ce qui revient à savoir si les longueurs d'empan sont différentes selon le poste occupé par le joueur.
  - a) Supposons que tous les joueurs ont la même longueur d'empan, quel que soit le poste qu'ils occupent dans l'équipe. Alors cette valeur commune serait égale à la moyenne calculée sur l'échantillon global.  
Pour évaluer l'erreur réalisée si on considère que les joueurs ont la même longueur d'empan, on calcule la somme des carrés des écarts entre les valeurs mesurées et la moyenne globale, c'est-à-dire la variation totale.
  - b) Si on considère que la longueur de l'empan des joueurs dépend du poste occupé, alors cette valeur est la moyenne du groupe d'appartenance. Pour évaluer l'erreur réalisée si on considère que la longueur de l'empan des joueurs dépend du poste occupé, on va calculer la somme des carrés des écarts entre la moyenne du groupe et la moyenne globale, pondérée par l'effectif du groupe c'est-à-dire la variation inter-groupes.
  - c) On remplace la longueur d'empan d'un sportif par la moyenne des longueurs d'empan de son groupe d'appartenance (poste occupé).
    - i) Créer un vecteur qui donne les nouvelles mesures de longueur d'empan, en utilisant les commandes `rep`, `mean` et `length`.
    - ii) Quelle est alors la moyenne globale ? La moyenne de chaque poste ?
    - iii) Calculer la variation de ce nouveau vecteur de données à l'aide de la fonction `var`.
    - iv) Vérifier qu'on retrouve la *variation inter-groupes* calculée à la question 4b.
  - d) On étudie maintenant la mesure de l'intensité de la relation entre l'empan et le poste de jeu.
    - i) Calculer le rapport de corrélation  $\eta^2$  en tant que rapport de la variation inter sur la variation totale.
    - ii) Utiliser la fonction `eta2` pour faire le calcul.
    - iii) Commenter la relation.

## Exercice 2

Le fichier "horizontal.txt" contient les détentes horizontales (variable "dente") pour 56 enfants pratiquant trois sports différents (variable "sport"), toujours accessible dans les fichiers de données du site pédagogique.

- 1) Calculer la moyenne et la variance de la longueur de la détente chez l'ensemble des enfants.
- 2) Calculer la moyenne de la longueur de la détente pour chaque groupe d'enfants (c'est-à-dire en fonction du sport pratiqué).
- 3) Faire une représentation graphique afin de répondre à la question : peut-on observer une différence en fonction des groupes ?
- 4) Vérifier que les écart-types sont du même ordre de grandeur quel que soit le type de sport pratiqué.
- 5) On cherche à savoir s'il y a une relation entre la longueur de la détente et le type de sport pratiqué.
  - a) Evaluer l'erreur réalisée si on considère que les enfants ont tous la même détente.
  - b) Evaluer l'erreur réalisée si on considère que la longueur de la détente des enfants dépend du type de sport pratiqué.
  - c) Conclure.

## Exercices supplémentaires

- Le fichier "Réaction.txt" contient les résultats des finales du 100m, 200m et 400m aux JO d'Atlanta. Ce sont les temps de réaction (en ms.) des sprinters finalistes. Existe-t-il une différence de temps de réaction (variable "temps") en fonction des trois types de courses (variable "course") ? Si oui, expliquer la relation ?
- Considérons les réponses au questionnaire passé au près des étudiants de L3, (fichier L3APA06.txt).
  - 1) Le rythme cardiaque des étudiants dépend-il du fait de pratiquer ou non un même sport plus de deux fois par semaine ?
  - 2) L'indice de Masse Corporelle *IMC* des étudiants dépend-il du sexe ?

## Références

- [1] A.B. Dufour, A.H. Rouard, J. Pontier, and L. Maurin, *Profil morphologique des handballeurs français de haut niveau*, Science et Motricité **2** (1987), 3-9.