

Statistiques descriptives et Introduction à R

Partie 2 : Statistiques bivariées

Edmond Sanou

Université d'Evry Val d'Essonne

Avril 2023



Sommaire

① Statistique bivariée

Sommaire

① Statistique bivariée

Liaison entre deux variables quantitatives

Sommaire

① Statistique bivariable

Liaison entre deux variables quantitatives

Nuage de points

- Le nuage de points est un outil graphique qui permet de mettre en évidence la relation entre deux variables. Généralement la variable de cause est représentée en abscisse et la variable de réponse en ordonnée.

Remarques :

- Quand vous étudiez simultanément deux variables quantitatives X et Y , il faut **toujours** tracer le nuage de points $(x_i, y_i)_{i=1, \dots, n}$ avant de postuler une éventuelle relation linéaire entre les deux variables.
- Certains cas non linéaires peuvent se ramener à des cas linéaires via un changement de variable adéquat

Nuage de points : Exemple

Exemple : Révision et note d'examen

- Tableau

Etudiants	Maxime	Stan	Mariem	Célia	David
Nombre d'heures de révision avant l'examen	2	3	4	1	6
Note de l'examen sur 20	11	13	14	8	18

- Nuage de points

```
heures <- c(2,3,4,1,6)
```

```
notes <- c(11,13,14,8,18)
```

```
plot(heures, notes,
```

```
  main = "Nuage de points (révision (X), note d'examen (Y))",
```

```
  xlab = "Nombre d'heures de révision avant l'examen (X)",
```

```
  ylab = "Note de l'examen (Y)")
```

- Donner la nature de la relation entre les deux variables

Covariance

La formule de la covariance est donnée comme suit :

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Sous R :

- Fonction cov

```
heures <- c(2,3,4,1,6)
```

```
notes <- c(11,13,14,8,18)
```

```
cov(heures, notes)
```

```
## [1] 7.05
```

Une covariance positive indique que lorsque les valeurs de X sont supérieures à leur moyenne, les valeurs de Y le sont aussi.

Coefficient de corrélation linéaire

On appelle **coefficient de corrélation linéaire** (ou *coefficient de Bravais-Pearson*) de X et Y la statistique

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Avec $s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ et s_x et s_y les écarts-types de X et Y .

Sous R :

- La fonction est cor

```
heures <- c(2,3,4,1,6)
```

```
notes <- c(11,13,14,8,18)
```

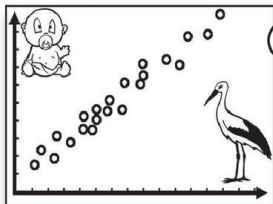
```
cor(heures, notes)
```

```
## [1] 0.9902115
```

Coefficient de corrélation linéaire : Interprétation

- On a $-1 \leq r_{xy} \leq 1$
- Si $r_{xy} = -1$ (respectivement si $r_{xy} = 1$), on a une corrélation négative (respectivement positive). Si $r_{xy} = 0$, il n'y a pas de corrélation.
- La présence de corrélation linéaire n'induit pas un lien de causalité

**Corrélation
ne veut pas dire
causalité**



En Alsace, les villes
qui ont le plus de cigognes
ont aussi le plus de bébés.

C'est la preuve que
ce sont bien les cigognes
qui apportent les bébés.

Ou alors
tout simplement
il y a plus de bébé
et plus de cigognes
dans les villes
avec le plus de
population...

**Il arrive que les deux
valeurs dépendent toutes
deux d'un même
troisième paramètre**



© Belin Éducation/Humensis, 2019 Sciences économiques et sociales 2de

© Christophe Michel

Coefficient de corrélation : Interprétation

- L'absence de corrélation n'implique pas l'absence de relation entre les deux variables.

Exemple : Quelle est la corrélation entre $(x_1, \dots, x_{40}) = (-20, -19, \dots, 20)$ et $y_i = |x_i|$?

- Un coefficient proche de 1 ne signifie pas pour autant que la relation est linéaire.

Exemple : Quelle est la corrélation entre $(x_1, \dots, x_{20}) = (0, 1, \dots, 20)$ et $y_i = \sqrt{x_i}$.

Droite d'ajustement linéaire (1)

- La droite d'ajustement linéaire ou droite de tendance permet d'ajuster linéairement un nuage de points deux variables quantitatives.
- En partant du nuage de points précédent dans lequel la relation semble linéaire, on cherche une droite qui résume au mieux le nuage de points (i.e. une ordonnée à l'origine et un coefficient directeur).

Ecriture matricielle

On est dans le cadre suivant

$$(S) \begin{cases} y_1 &= \beta_1 + \beta_2 x_1 \\ \vdots &= \vdots \\ y_n &= \beta_1 + \beta_2 x_n \end{cases} \Leftrightarrow y = \mathbf{X}\beta$$

Avec $\mathbf{X} = (\mathbf{1}, x) \in \mathbb{R}^{n \times 2}$, $\mathbf{1} = (1, \dots, 1)' \in \mathbb{R}^n$, $y = (y_1, \dots, y_n)'$ et $\beta = (\beta_1, \beta_2)'$.

- On a donc n équations et 2 inconnues, le système est surdimensionné
- Pour “résumer au mieux” le nuage de points, il nous faut un critère. Ce sera le critère des moindres carrés ordinaires.
- On cherche deux paramètres $\hat{\beta}_1, \hat{\beta}_2$ tels que la somme des carrés des distances des points à la droite soit minimale.

MCO(2)

Avec ce critère la solution de ce problème de minimisation est

$$\begin{cases} \hat{\beta}_2 &= \frac{s_{xy}}{s_x^2} \\ \hat{\beta}_1 &= \bar{y} - \hat{\beta}_2 \bar{x} \end{cases}$$

où $s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ est la covariance entre X et Y et s_x^2 la variance de X . Pour le démontrer il suffit de résoudre le programme de minimisation suivant

$$(\hat{\beta}_1, \hat{\beta}_2) = \arg \min_{(\beta_1, \beta_2) \in \mathbb{R}^2} \sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2$$

Sous R

- Ajouter le centre du nuage de point

```
plot(heures, notes,  
     main = "Nuage de points (révision (X), note d'examen (Y))",  
     xlab = "Nombre d'heures de révision avant l'examen (X)",  
     ylab = "Note del'exament (Y)")  
  
points(mean(heures), mean(notes), col = "red", pch=15)
```

Sous R : modélisation

- Modèle linéaire à l'aide de la fonction `lm`

```
modele <- lm(notes ~ heures) #Y ~ X  
coefficients(modele)
```

```
## (Intercept)      heures  
##      6.702703      1.905405
```


Sous R : Visualisation droite ajustement

- Ajouter la droite d'ajustement linéaire au nuage de point

```
modele <- lm(notes ~ heures)
```

```
plot(heures, notes,  
     main = "Nuage de points (révision (X), note d'examen (Y))",  
     xlab = "Nombre d'heures de révision avant l'examen (X)",  
     ylab = "Note del'exament (Y)")
```

```
points(mean(heures), mean(notes), col = "red", pch=15)
```

```
abline(modele)
```

Analyse de la variance

Il est possible d'écrire la décomposition suivante

$$\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

avec $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

- On appelle $\hat{u} = y - \hat{y}$ le vecteur des résidus
- La qualité d'ajustement peut mesurer aussi en faisant le rapport de la variance expliquée sur la variance totale de y (i.e. le coefficient de détermination R^2)

Coefficient de détermination

- C'est la part de variance de Y expliquée par la relation $y = \hat{\beta}_1 + \hat{\beta}_2 x$
- On a $0 \leq R^2 \leq 1$ qui représente la part de variabilité de Y expliquée par le modèle.
- Plus R^2 est proche de 1 meilleur est l'ajustement.
- Si $R^2 = 0$ alors la variance expliquée est nulle donc la variable X n'a aucune influence linéaire sur Y
- Si $R^2 = 1$ alors la variance résiduelle est nulle donc on a une relation linéaire exacte

Sur R

- Afficher les variances totale, résiduelle et expliquée

```
var(notes) ## variance totale Y
```

```
## [1] 13.7
```

```
var(residuals(modele)) ## variance résiduelle
```

```
## [1] 0.2668919
```

```
var(fitted(modele)) ## variance expliquée
```

```
## [1] 13.43311
```

Remarques

- Dans la corrélation linéaire, il n'y a aucune causalité entre les variables
- On voit bien au travers de ces droites que l'ajustement linéaire est un cadre de travail adapté pour expliquer une variable Y comme la conséquence d'une variable X . On rentre là dans le domaine de la modélisation et logiquement par la suite de la prévision
- De plus, les méthodes utilisées dans ce cours sont totalement déterministes et ne renseignent pas sur la significativité des paramètres obtenus.