

Statistiques descriptives

Vincent Arel-Bundock



Notre chemin

1 Centralité

2 Dispersion

3 Association



Centralité

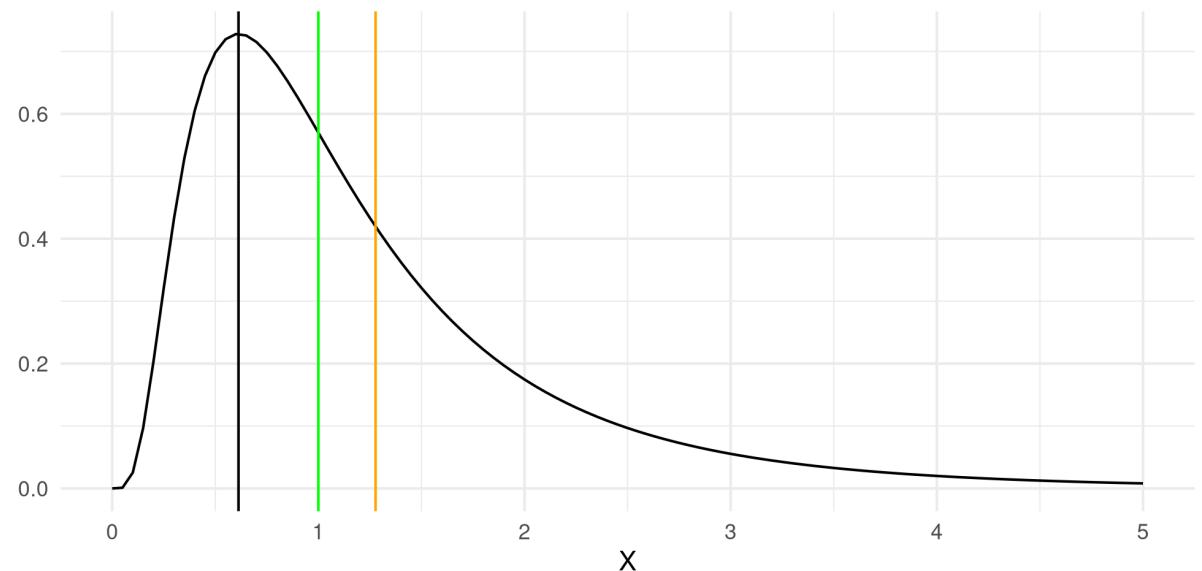


Centralité

4 mesures de centralité:

1. Mode
2. Médiane
3. Moyenne
4. Espérance

Distribution log-normale
Mode: Noir
Médiane: Vert
Moyenne: Orange



Mode

Définition:

- | Valeur la plus fréquente d'un ensemble.

Exemple:

$$\{1, 2, 2, 3, 4, 7, 8, 8, 8\}$$

Mode = 8

Moyenne

Définition:

Pour un ensemble X de taille n , la moyenne est représentée par \bar{X} et se calcule:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Exemple:

Si X contient 5 éléments $\{1, 3, 4, 8, 9\}$, la moyenne est:

$$\bar{X} = \frac{1}{5}(1 + 3 + 4 + 8 + 9) = 5$$

Médiane

Définition:

Valeur qui sépare une série ordonnée de nombres en deux parties contenant le même nombre d'éléments.

Exemple

$$\begin{array}{ccc} 1, 3, & \boxed{4} & 8, 9 \\ \underbrace{}_2 \text{ éléments} & \boxed{4} & \underbrace{}_2 \text{ éléments} \end{array}$$

Médiane = 4

Médiane: Avantage

La médiane est plus "robuste" que la moyenne, au sens où elle est moins sensible aux valeurs extrêmes ou aberrantes.

Considérez les deux ensembles suivants:

$$\{1, 2, 3, 4, 5\} \quad \text{Moyenne} = 3; \text{ Médiane} = 3$$

$$\{1, 2, 3, 4, 100\} \quad \text{Moyenne} = 22; \text{ Médiane} = 3$$

La moyenne est fortement affectée par la valeur extrême 100, mais la médiane ne change pas.

Espérance

Distinction importante entre statistiques calculées à partir de:

- Population entière
- Échantillon, c.-à-d. sous-groupe de la population

Espérance:

Moyenne d'une population entière, plutôt que d'un échantillon.

ou

Une moyenne à "long terme," calculée après un grand nombre de répétitions de l'expérience qui produit la variable X .

Espérance

Exemple:

- Moyenne de la taille des hommes canadiens

Problème:

- Trop coûteux de mesurer tout le monde

Solution:

- Échantillon de 1000 personnes.

Deux quantités:

- \bar{X}
 - Taille moyenne dans l'échantillon
- $E[X]$
 - Taille moyenne qu'on obtiendrait si on pouvait mesurer la taille de tous les hommes canadiens

Espérance

L'espérance de X s'écrit $E[X]$ et se calcule ainsi:

$$E[X] = \sum_{\forall x \in X} x \cdot P(X = x)$$

Pour toutes les valeurs x possibles de la variable X , nous multiplions x par la probabilité d'observer x , et nous prenons la somme.

Espérance

L'espérance de X s'écrit $E[X]$ et se calcule ainsi:

$$E[X] = \sum_{\forall x \in X} x \cdot P(X = x)$$

Pour toutes les valeurs x possibles de la variable X , nous multiplions x par la probabilité d'observer x , et nous prenons la somme.

Si X représente le lancer d'un dé, l'ensemble X est composé des chiffres de 1 à 6. La probabilité d'obtenir n'importe lequel de ces chiffres est égale à 1/6.

L'espérance est:

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3,5$$

Espérance

Une variable "binaire" ou "dichotomique" peut prendre seulement deux valeurs: 0 et 1.

Si Z est une variable dichotomique avec cette probabilité :

$$P(Z) = \begin{cases} P(Z = 0) = \frac{3}{5} \\ P(Z = 1) = \frac{2}{5} \end{cases}$$

alors l'espérance est égale à :

$$E[Z] = 1 \cdot \frac{2}{5} + 0 \cdot \frac{3}{5} = 40\%$$

L'espérance d'une variable binaire est égale à la probabilité d'observer un 1.

Espérance conditionnelle

L'espérance conditionnelle s'écrit comme suit:

$$E[Y|X = x],$$

et se dit "valeur attendue de Y lorsque X est égale à x ."

Cette équation indique qu'il faut calculer l'espérance de la variable Y en considérant seulement les observations pour lesquelles la variable X est égale à x .

Espérance conditionnelle et indépendance

Rappel:

$$\text{Si } Y \perp X, \text{ alors } P(Y|X) = P(Y)$$

Similairement:

$$\text{Si } Y \perp X, \text{ alors } E[Y|X] = E[Y]$$

Espérance conditionnelle et indépendance

Rappel:

$$\text{Si } Y \perp X, \text{ alors } P(Y|X) = P(Y)$$

Similairement:

$$\text{Si } Y \perp X, \text{ alors } E[Y|X] = E[Y]$$

Important!

Ceci nous permettra de déterminer si on peut donner une interprétation causale à des analyses statistiques.

Dispersion



Dispersion

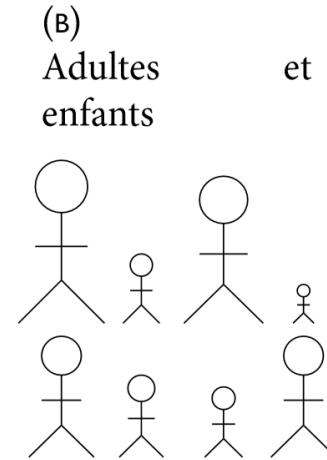
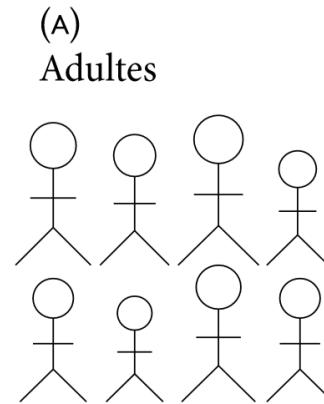
Est-ce que beaucoup d'observations s'éloignent du centre de la distribution?

3 mesures:

1. Variance
2. Écart type
3. Écart interquartile

GRAPHIQUE 4.1. –

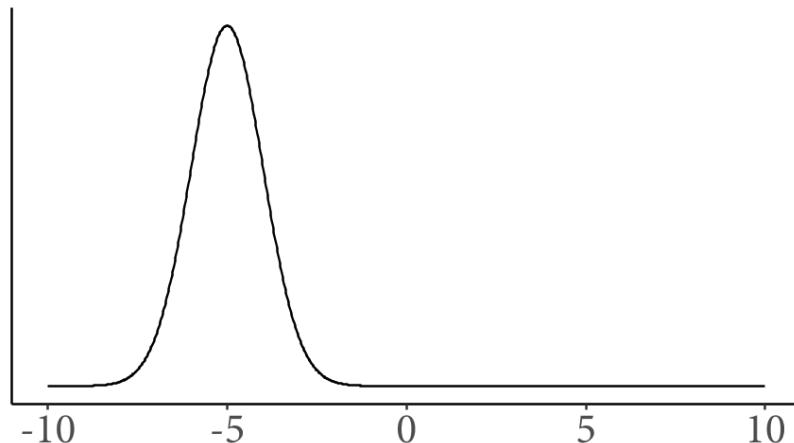
Les tailles risquent d'être plus dispersées si l'échantillon comprend des adultes et des enfants que si l'échantillon comprend seulement des adultes.



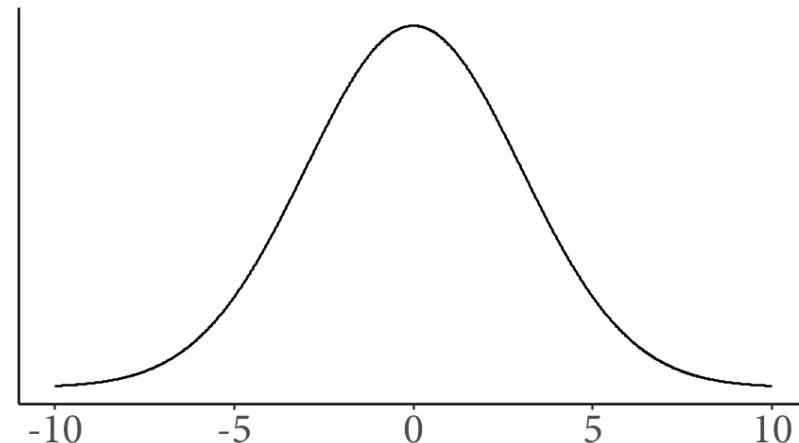
Dispersion

Deux distributions normales avec différents niveaux de dispersion:

$$\mu = -5, \sigma = 1$$



$$\mu = 0, \sigma = 3$$



Variance

La variance d'un ensemble X composé de n éléments s'écrit σ_X^2 ou $\text{Var}(X)$, et se calcule:

$$\sigma_X^2 = \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Plus la variance est grande, plus les valeurs ont tendance à s'éloigner du centre de la distribution.

Variance

$$\sigma_X^2 = \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Par exemple, l'ensemble $X = \{0, 1, 2\}$ a une moyenne de 1 et une variance de:

$$\sigma_X^2 = \frac{1}{3} [(0 - 1)^2 + (1 - 1)^2 + (2 - 1)^2] = \frac{2}{3}$$

L'ensemble $Z = \{-2, 1, 4\}$ a aussi une moyenne de 1, mais une plus grande variance:

$$\sigma_Z^2 = \frac{1}{3} [(-2 - 1)^2 + (1 - 1)^2 + (4 - 1)^2] = 6$$

Écart type

Problème:

- La formule de la variance prend le carré des déviations par rapport à la moyenne.
- $\sigma_X^2 = \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

Conséquence:

- La variance ne peut pas être interprétée sur la même échelle que la variable originale.

Solution:

- Ramener la mesure de dispersion à la même échelle que la variable originale en prenant sa racine carrée

Écart type

Variance:

$$\sigma_X^2 = \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Écart type:

$$\sigma_X = \sqrt{\text{Var}(X)}$$

Écart interquartile

Problème:

- La moyenne, la variance et l'écart type peuvent être affectés par les valeurs extrêmes ou aberrantes.

Solution analogue à la médiane:

- L'écart interquartile est une mesure de dispersion plus robuste.

Écart interquartile

Centile

Un centile est l'une des 99 valeurs qui divisent une variable X en 100 groupes de taille identique.

Si on range toutes les observations d'une variable X en ordre croissant et on les divise en 100 groupes de taille identique :

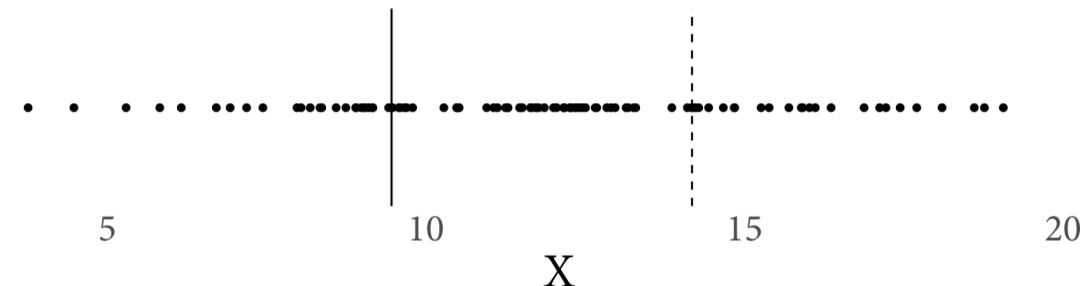
- Le 1er centile est la valeur qui sépare le 1% des plus petites valeurs de X du reste.
- Le 25e centile est la valeur qui sépare le 25% des plus petites valeurs de X du reste.
- Le 75e centile est la valeur qui sépare le 75% des plus petites valeurs de X du reste.

Écart interquartile

L'écart interquartile mesure la distance entre le 25e centile et le 75e centile.

GRAPHIQUE 4.2. –

25% des points se trouvent à gauche de la ligne pleine. 75% des points se trouvent à gauche de la ligne pointillée. L'écart interquartile est la distance entre les deux lignes verticales.



L'écart interquartile mesure la dispersion du 50% central des données.

Plus l'écart interquartile est grand, plus les observations ont tendance à s'éloigner du centre de la distribution.

Association



Mesures d'association

Jusqu'ici: Statistiques univariées

Maintenant: Statistiques bivariées

Objectif: Mesurer la force de l'association entre deux variables

- Covariance
- Corrélation
- Variables catégoriques

Covariance

La covariance mesure l'association linéaire entre deux variables. Elle se calcule :

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}),$$

où :

- n est le nombre d'observations
- \bar{X} est la moyenne de X
- \bar{Y} est la moyenne de Y

Covariance

La covariance mesure l'association linéaire entre deux variables. Elle se calcule :

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}),$$

où :

- n est le nombre d'observations
- \bar{X} est la moyenne de X
- \bar{Y} est la moyenne de Y

Positive:

- Valeurs élevées de X sont associées à des valeurs élevées de Y .

Négative:

- Valeurs élevées de X sont associées à des valeurs faibles de Y .

Covariance

Si $Cov(\text{Taille}, \text{Revenu}) > 0$, alors on peut conclure que les grands ont tendance à avoir un revenu élevé.

Si $Cov(\text{Âge}, \text{Cheveux}) < 0$, alors on peut conclure que les vieux ont tendance à avoir moins de cheveux que les jeunes.

Corrélation

Problème:

- La covariance dépend de l'échelle des variables qui entrent dans son calcul.

Exemple:

- Si on mesure la covariance entre les tailles de parents et de leurs enfants en centimètres plutôt qu'en mètres, la covariance estimée sera 10,000 fois plus grande.

Solution:

- Normaliser la covariance en la divisant par le produit des écarts types des deux variables

Corrélation

Covariance:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}),$$

Correlation de Pearson:

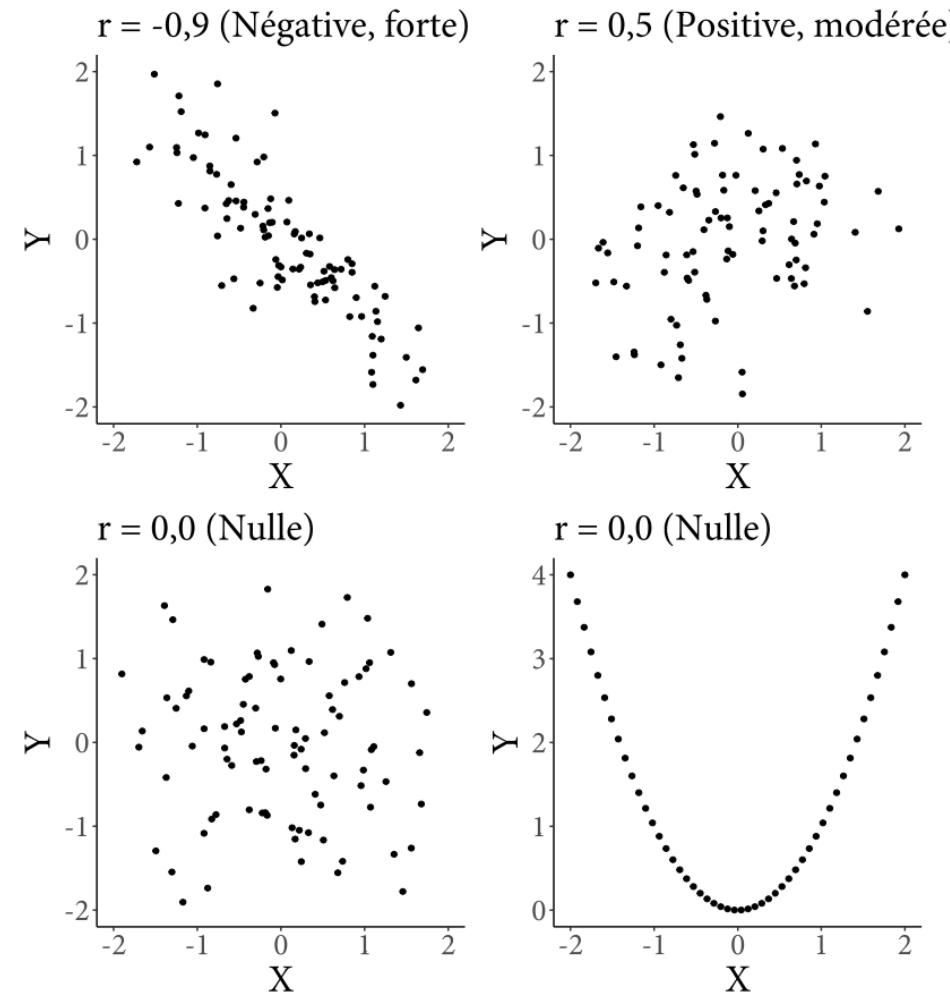
$$r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Corrélation

Propriétés utiles:

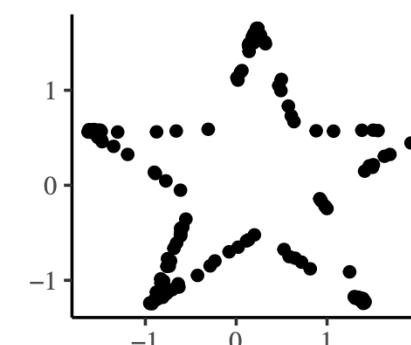
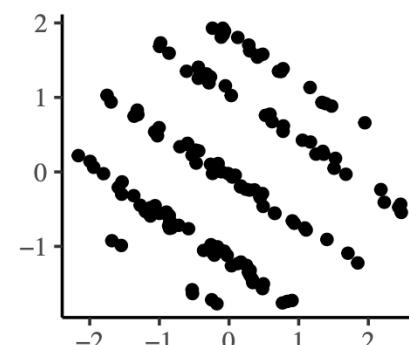
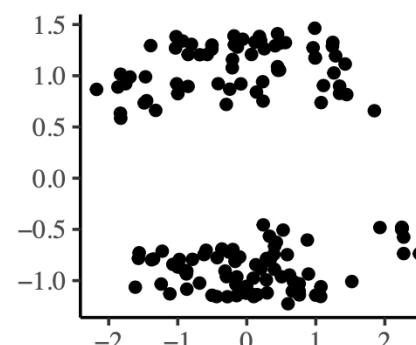
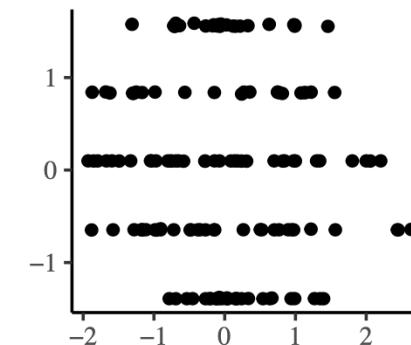
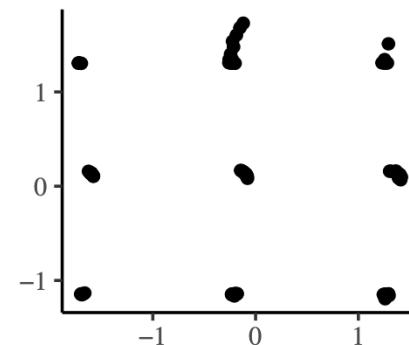
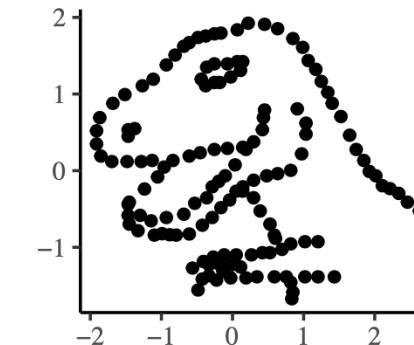
- $r_{XY} \in [-1, 1]$
- Signe:
 - $r_{XY} > 0$: relation linéaire entre X et Y est positive.
 - $r_{XY} < 0$: relation linéaire entre X et Y est négative.
 - $r_{XY} = 0$: pas d'association linéaire entre X et Y .
- Force:
 - Plus la corrélation approche les extrêmes (-1 ou 1), plus l'association entre les deux variables est forte.

Corrélation



Corrélation: Limites

Variables continues et relations *linéaires*!



Association entre variables catégoriques

Pour étudier l'association entre deux variables catégoriques (binaires, ordinaires, ou nominales), on peut employer un tableau de contingence comme celui-ci:

TABLEAU 4.1. –

Ce tableau de contingence suggère que, suivant l'accident du Titanic, le taux de survie des femmes était plus élevé que celui des hommes.

	Mort	Vivant
Femme	154	308
Homme	709	142

Association entre variables catégoriques

Il existe beaucoup de statistiques pour résumer l'association entre variables catégoriques. Par exemple:

- Variables nominales: V de Cramer
- Variables ordinaires: τ de Kendall

Merci!

