UNIVERSITÉ DE GENÈVE

Introduction à la Statistique

Statistiques descriptives

Stéphane Guerrier, Mucyo Karemera, Samuel Orso & Lionel Voirol

Data Analytics Lab

Licence : CC BY NC SA 4.0

Statistiques descriptives et inférentielles

Statistiques descriptives : décrire les données que nous avons collectées qui composent l'échantillon

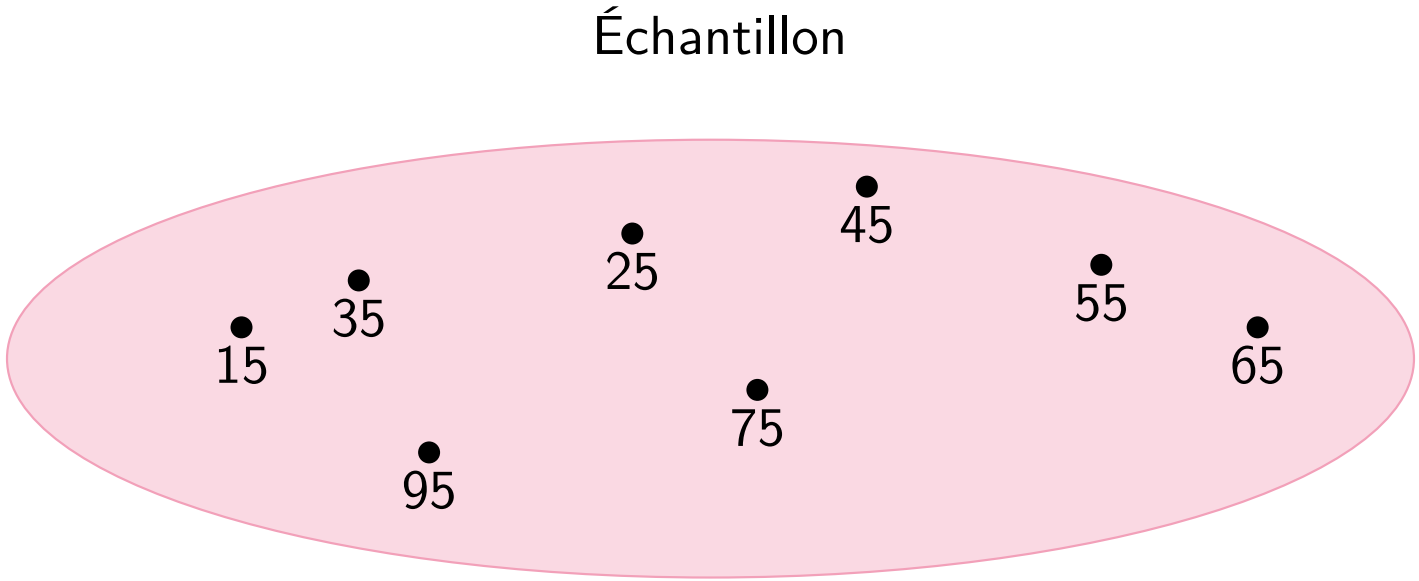
Inférence statistique : faire des généralisations sur un ensemble plus large, la population

Description des échantillons

Quelle est sa tendance centrale ?

Quelle est sa dispersion ou variabilité ? Combien de bruit contient les données ?

Quelle est la forme de la distribution ? Est-elle symétrique ?



Tendance centrale de la distribution

Mesures de la tendance centrale de la distribution

Moyenne : additionner les données et diviser par le nombre d'observations.

Médiane : un nombre égal d'observations plus grandes et plus petites que la médiane.

Moyenne

Additionner les données et diviser par le nombre d'observations

Exemples :

Données : 1, 2, 2, 3, 4

Moyenne $= (1 + 2 + 2 + 3 + 4)/5 = 2.4$

Données : 10, 12, 56, 78, 113, 1209

Moyenne $= (10 + 12 + 56 + 78 + 113 + 1209)/6 = 246.3$

Moyenne

Données

$$\{x_1,\ x_2,\ x_3,\ \ldots,x_n\}$$

Moyenne

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Médiane

Trier les données et considérer la valeur de l'observation centrale

Données : 1, 2, **2**, 3, 4

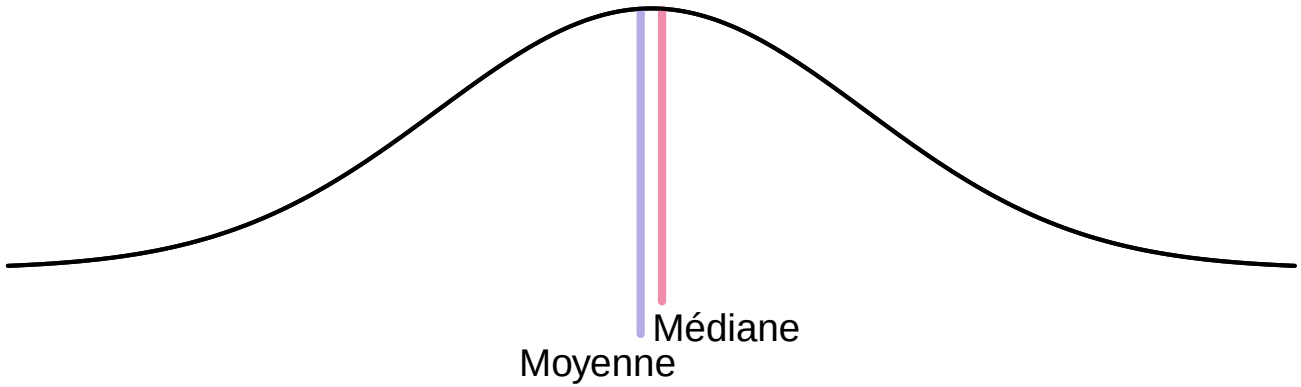
Médiane = 2

Données : 10, 12, **56, 78**, 113, 1209

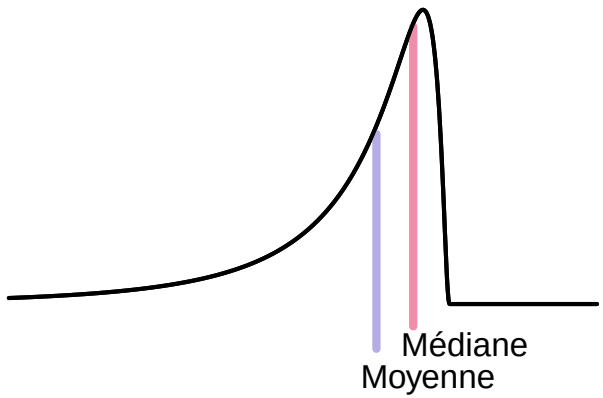
Médiane = (56+78)/2 = 67

Moyenne versus médiane

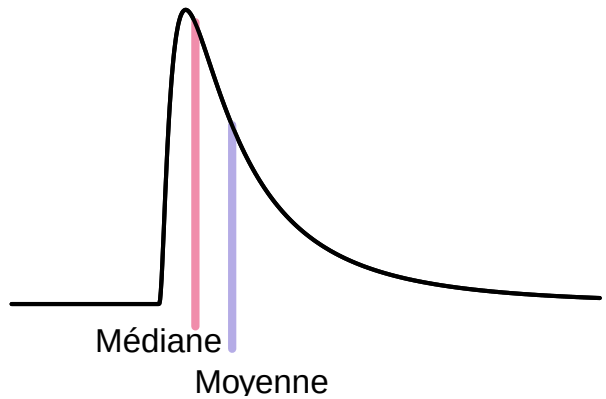
La moyenne et la médiane sont proches pour des distributions symétriques.



La moyenne se déplace dans la direction de l'asymétrie d'une distribution.



Asymétrie à gauche



Asymétrie à droite

Valeurs aberrantes

Valeur aberrante = une valeur qui ne correspond pas au reste

Données : 3, 6, 7, 10, **157**

Moyenne $= \frac{1}{5}(3+6+7+10+157) = 36.6$

Médiane $= 7$

La médiane est résistante aux valeurs aberrantes.

Résumé en 5 nombres

Médiane

Minimum, Maximum

Quartiles : observation centrale au-dessus et en dessous de la médiane

Trouver les quartiles

Données : $7, 23, 75, 82, 34, 91, 10$

Trier, on obtient 7, 10, 23, 34, 75, 82, 91

Trouver la médiane, on obtient 34

Observations en dessous de la médiane : 7, 10, 23

Premier quartile $Q1 = 10$

Observations au-dessus de la médiane : 75, 82, 91

Troisième quartile $Q3 = 82$

Autre exemple

Données : $7, 8, 22, 38, 48, 62$

Médiane = $(22+38)/2 = 30$

Premier Quartile : 7, 8, 22

$$Q_1 = 8$$

Troisième Quartile : 38, 48, 62

$$Q_3 = 48$$

Mesurer la dispersion

Quelle est la variabilité des données ?

Étendue = Maximum - Minimum

Étendue Interquartile (IQR) : $Q3 - Q1$

Écart-Type (s) : Racine carrée de la moyenne des distances quadratiques à la moyenne.

Écart-type de l'échantillon

Formule

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

Calcul simplifié lorsque la moyenne de l'échantillon est donnée

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

5 Étapes faciles

Calculer la moyenne $\bar{x}$

L'élever au carré

Calculer la somme des x^2

Trouver la différence (somme des x_i^2) $- n\bar{x}^2$

Diviser par $n - 1$

Prendre la racine carrée

Exemple : 7, 8, 3

Moyenne $= \bar{x} = 6$

Carré de la moyenne $= \bar{x}^2 = 36$

(Somme des x^2) $= 7^2 + 8^2 + 3^2 = 49 + 64 + 9 = 122$

(Somme des x^2) $- n\bar{x}^2 = 122 - 3(36) = 122 - 108 = 14$

$\frac{1}{n-1}\ 14 = \frac{1}{3-1}\ 14 = 7$

$s = \sqrt{7} = 2.645$

IQR versus s

L'IQR, comme la médiane, ne dépend pas des observations les plus grandes (ou les plus petites)

L'IQR est résistant aux valeurs aberrantes

s dépend de toutes les données et peut être sensible aux observations éloignées (valeurs aberrantes)

Résumé en 5 nombres

$$(\text{Minimum}, Q_1, \text{Médiane}, Q_3, \text{Maximum})$$

Exemple : 25, 78, 97, 133, 193, 212, 215, 274
Médiane : (133+193)/2 = 163 Partie inférieure : 25, 78, 97, 133

$$Q1 = (78 + 97)/2 = 87.5$$

Partie supérieure : 193, 212, 215, 274

$$Q3 = (212 + 215)/2 = 213.5$$

$$(25, 87.5, 163, 213.5, 274)$$

Variables

Variable : l'aspect qui diffère d'un sujet à un autre, d'un individu à un autre, par exemple l'orientation politique, l'âge, le sexe, le revenu, etc.

Données : la valeur des variables, par exemple : "Conservateur", "19", "Homme", "15 000$" , etc.

Deux types de variables

Variables quantitatives ou numériques

Nombres, mesures

Âge, taille, distance parcourue, heures de sommeil, revenu

Variables catégorielles

Classifier chaque observation

Nationalité, langue maternelle, satisfaction du cours, niveau d'études

Quatre sous-types de variables

Variables quantitatives ou numériques : (i) continues et (ii) discrètes

(i) Âge (21,25 ans), taille (1,67 m), distance parcourue (42,195 kms), heures de sommeil (8,4 h), revenu (24,48 CHF/h)

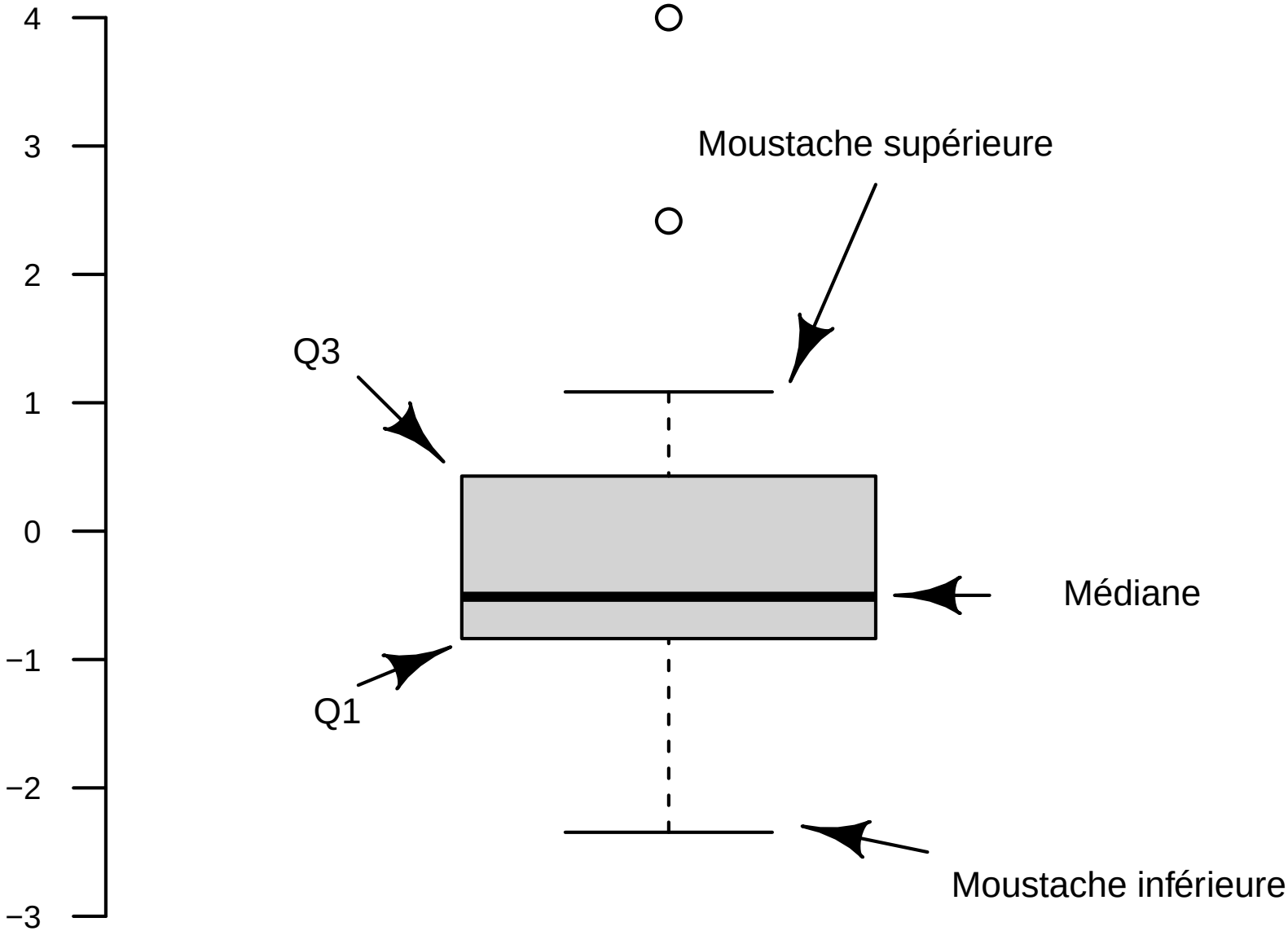
(ii) Âge (21 ans), taille (2 m), distance parcourue (42 kms), heures de sommeil (8 h), revenu (1'000 CHF mensuel brut)

Variables catégorielles : (iii) nominales et (iv) ordinales

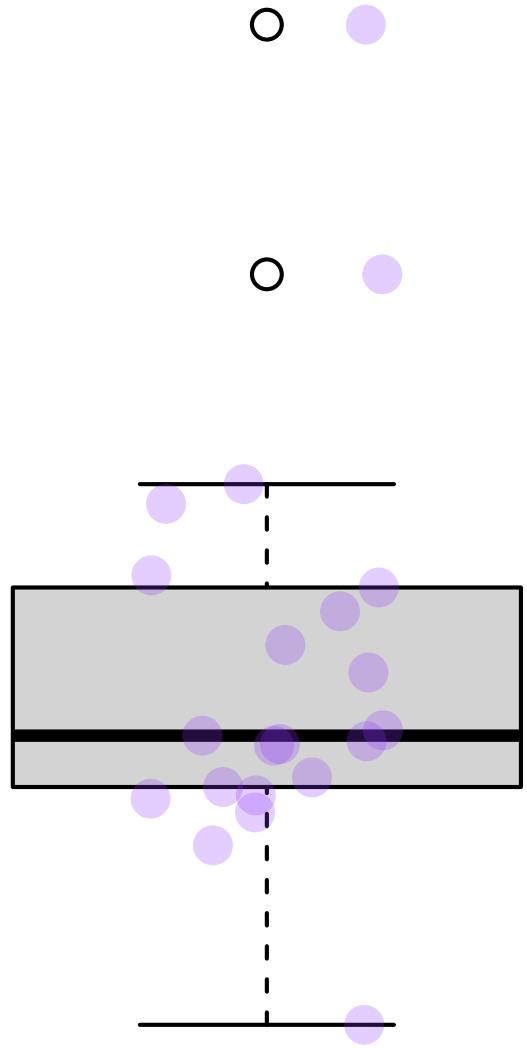
(iii) Nationalité (Suisse, Français, Allemand, Italien, ...), langue maternelle (Français, Anglais, Allemand, Italien, ...)

(iv) satisfaction du cours (neutre, satisfait, très satisfait, très très satisfait), niveau d'études (Bachelor, Master, Doctorat)

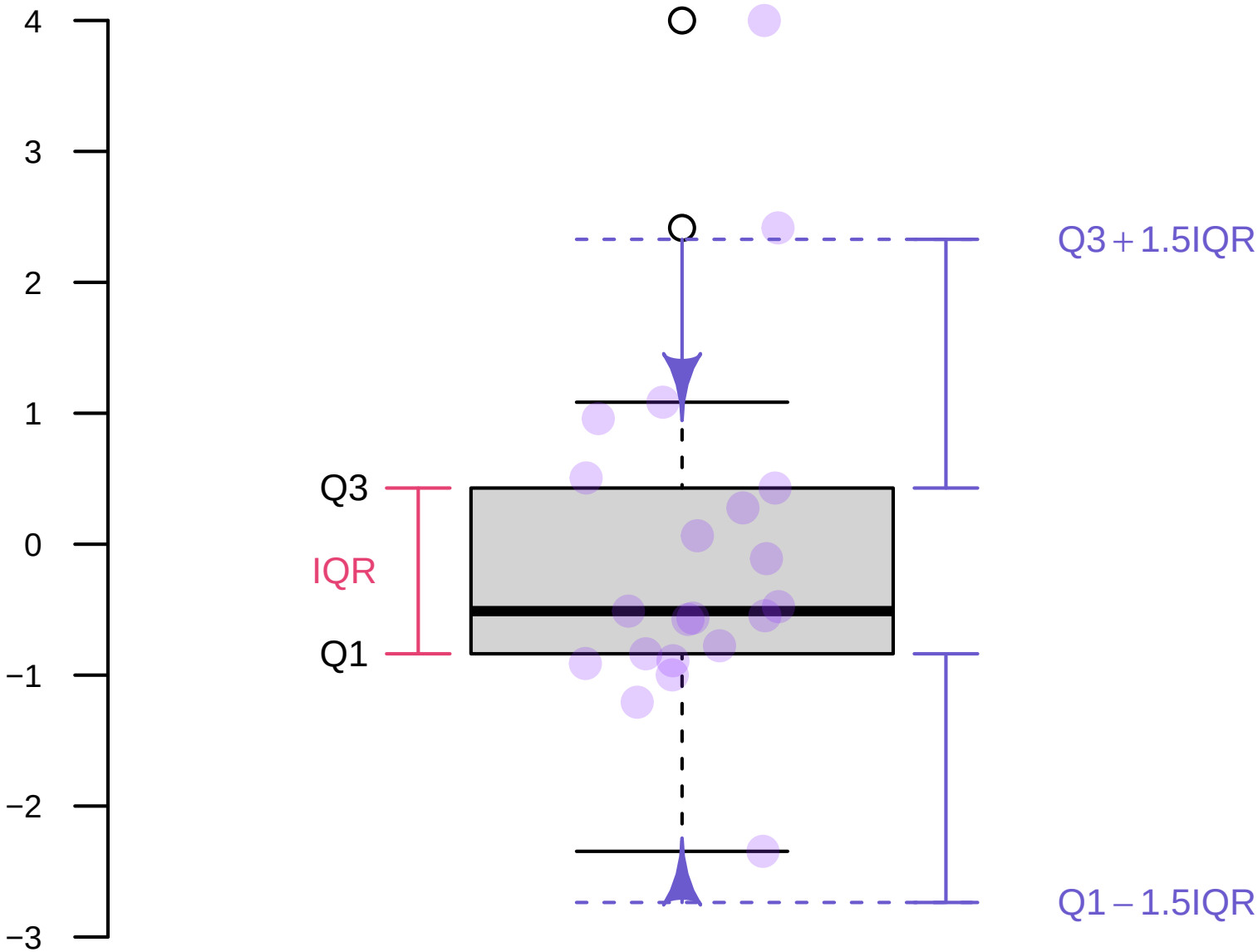
Boxplot


4
3
2
1
0
−1
−2
−3
Moustache supérieure
Q3
Médiane
Q1
Moustache inférieure

Boxplot



Boxplot



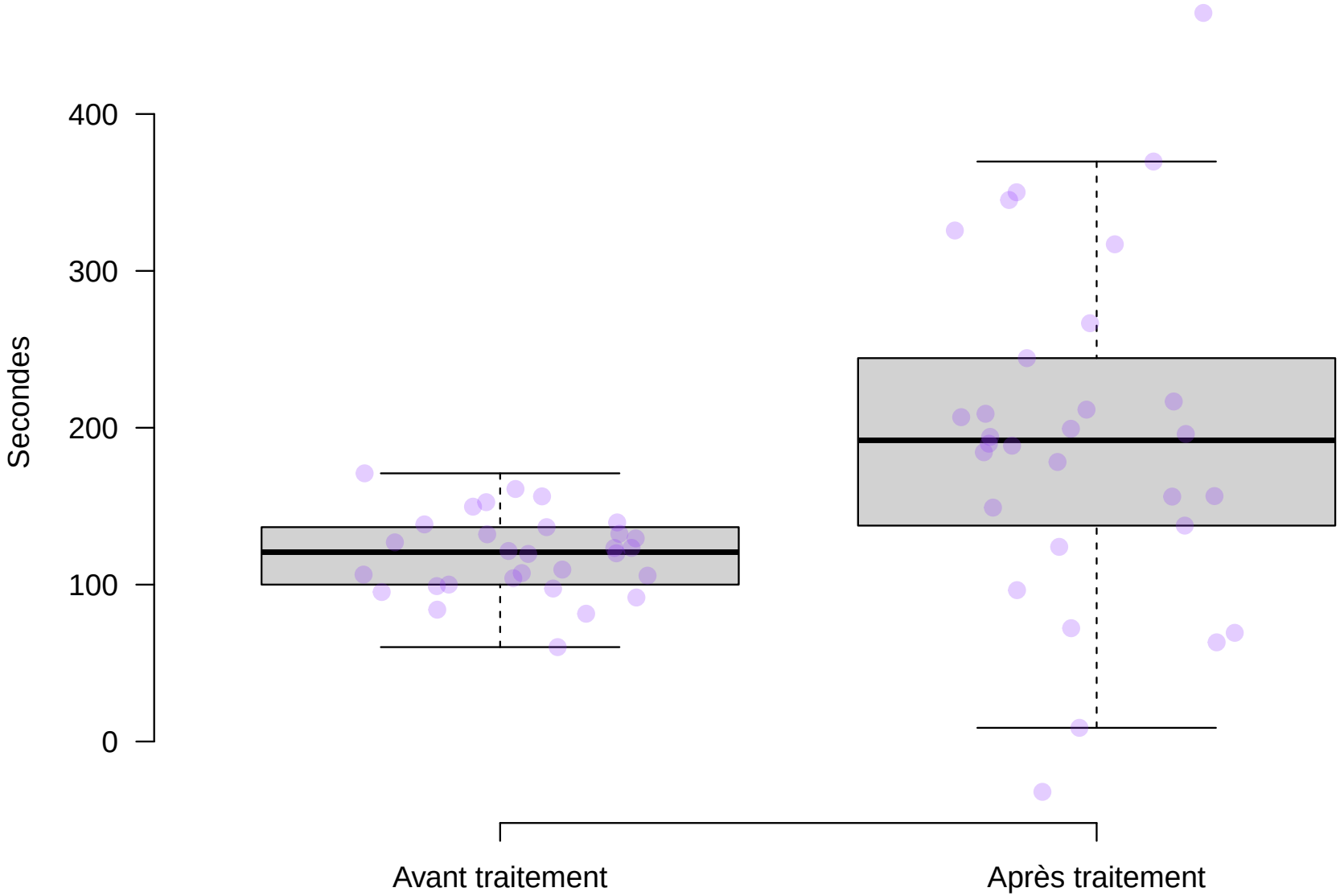
Comparaison entre groupes

Boxplot côte à côte pour comparer deux ensembles de données ou plus.

Ont-ils le même centre ? La même forme ? La même dispersion ?

La différence entre les médianes est-elle beaucoup plus grande que la variabilité des données ?

Exemple : résultats de tests pulmonaires avant et après traitement



L'histogramme

Exemple : *Heures de sommeil*

$$\Big\{12, 8.5, 7.2, 7.3, 7.7, 6, 6.5, 4.5, 3, 1.2,$$
$$1.3, 2, 2, 3.8, 6.6, 8.5, 5.9, 4.6, 5.6, 6.7\Big\}$$

Variable : Nombre d'heures de sommeil

Valeurs $= [0, 12]$

Exemple : Heures de sommeil

Trier les données :

$$\Big\{1.2, 1.3, 2, 2, 3, 3.8, 4.5, 4.6, 5.6, 5.9,$$
$$6, 6.5, 6.6, 6.7, 7.2, 7.3, 7.7, 8.5, 8.5, 12\Big\}$$

Choisir les *intervalles de classes* souhaités :

1-2 heures, 2-5 heures, 5-8 heures, 8-12 heures

4 intervalles de classe

4 blocs inégalement espacés

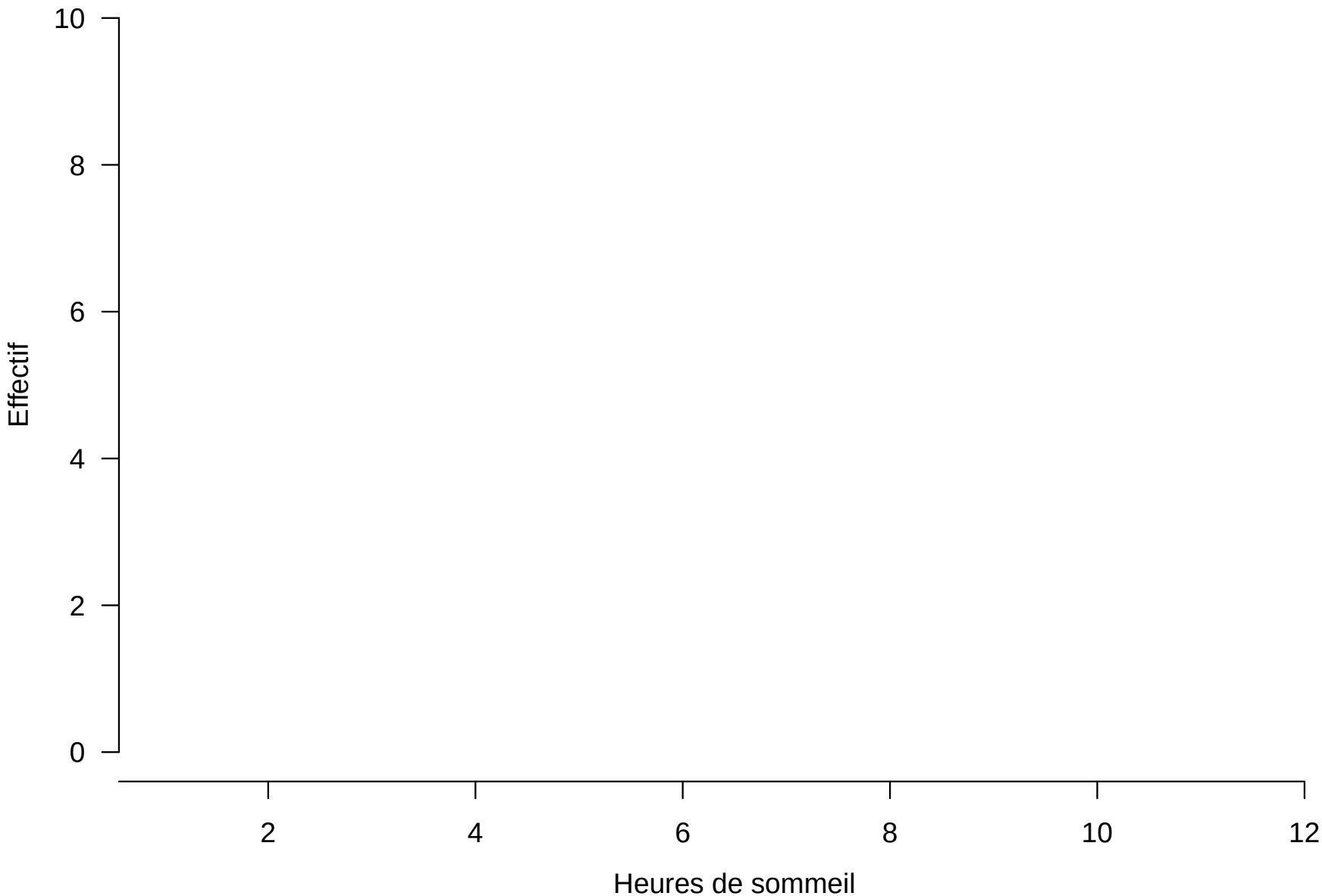
Exemple : Heures de sommeil

Comment dessiner les blocs ? Compter le nombre de points de données dans chaque classe :

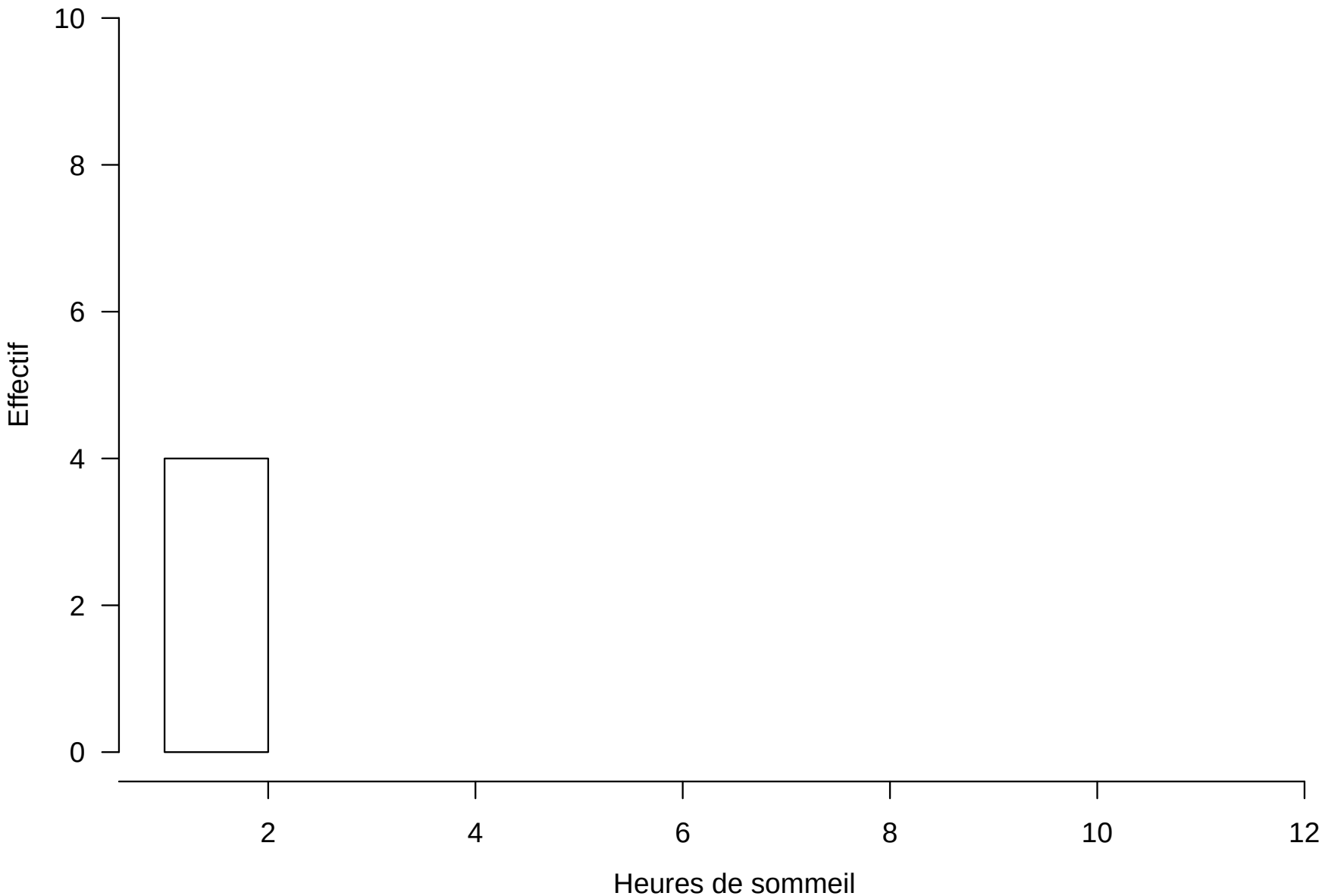
Heures de sommeil (X)	*Effectifs*	*Proportions*
$1 < X \leq 2$	4	4/20=0.2
$2 < X \leq 5$	4	4/20=0.2
$5 < X \leq 8$	9	9/20=0.45
$8 < X \leq 12$	3	3/20=0.15

Les intervalles n'ont pas nécessairement la même longueur.

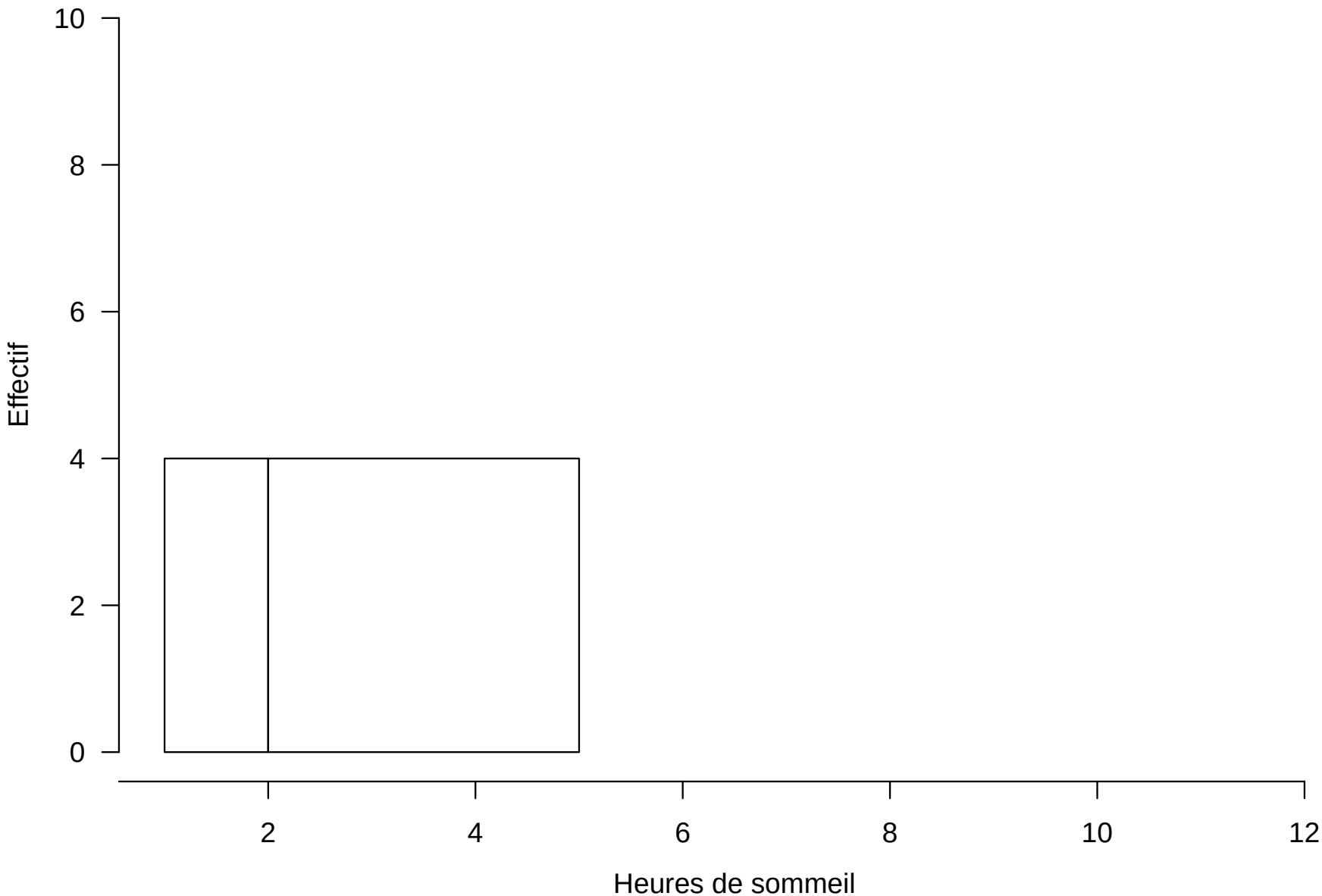
Exemple : Heures de sommeil



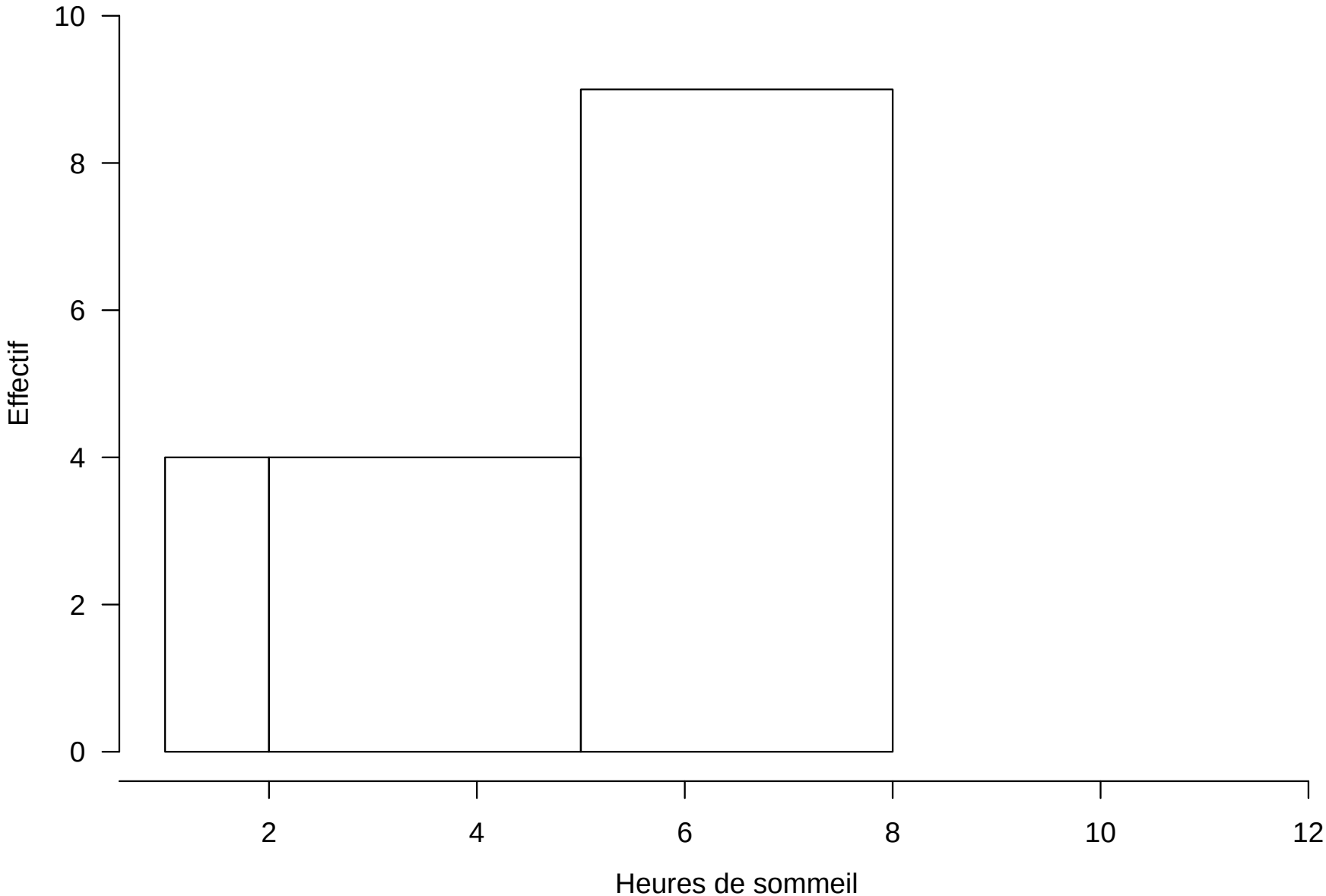
Exemple : Heures de sommeil



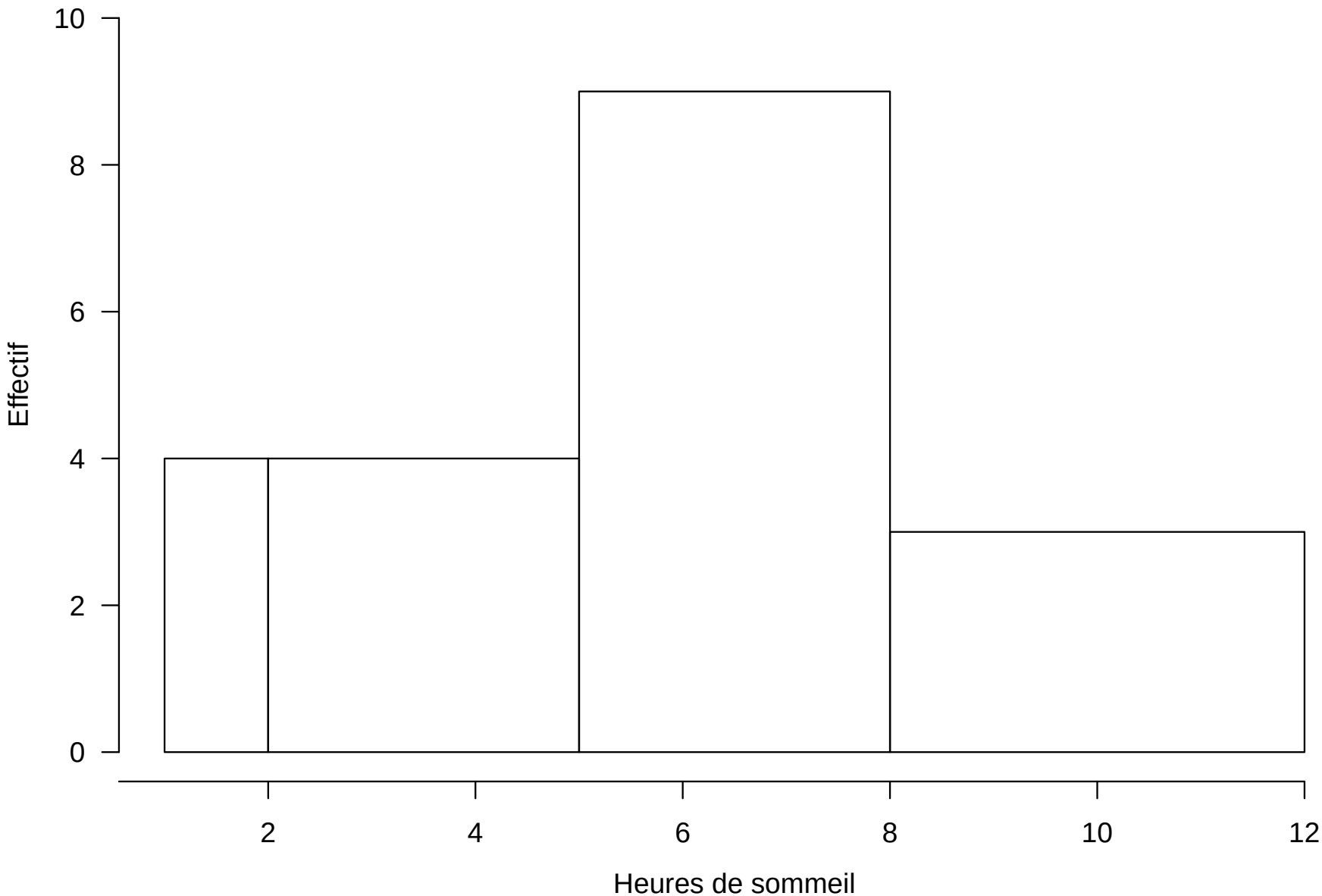
Exemple : Heures de sommeil



Exemple : Heures de sommeil



Exemple : Heures de sommeil



Exemple : Orientation politique

HARVARD Kennedy School
INSTITUTE OF POLITICS

La 13ème enquête bisannuelle des jeunes sur la politique et le service public (novembre 2007, N=2 526, jeunes de 18-24 ans)

"Lorsqu'il s'agit de la plupart des questions politiques, vous considérez-vous comme libéral, modéré ou conservateur ? (Si modéré, demander : en tant que modéré, vers quel côté penchez-vous)"

Orientation politique	Total	Étudiants	Non-étudiants
Libéral	32%	34%	31%
Modéré penchant libéral	14%	18%	13%
Modéré	21%	17%	23%
Modéré penchant conservateur	12%	12%	12%
Conservateur	21%	19%	22%

Résumé des catégories

Proportion d'échantillon
Comptages (Chaque catégorie a un certain nombre d'occurrences)
On obtient des proportions/pourcentages

Orientation politique	Total
Libéral	32%
Modéré penchant libéral	14%
Modéré	21%
Modéré penchant conservateur	12%
Conservateur	21%

Proportion d'échantillon

$$\hat{p}_{\text{Libéral}} = \frac{\#\text{ de personnes libérales}}{\text{nombre total de répondants}} = 0.32$$

$$\hat{p}_{\text{Modéré- libéral}} = \frac{\#\text{ de personnes modérées penchant libéral}}{\text{nombre total de répondants}} = 0.14$$

$$\hat{p}_{\text{Modéré}} = \frac{\#\text{ de personnes modérées}}{\text{nombre total de répondants}} = 0.21$$

$$\hat{p}_{\text{Modéré-conservateur}} = \frac{\#\text{ de personnes modérées penchant conservateur}}{\text{nombre total de répondants}} = 0.12$$

$$\hat{p}_{\text{Conservateur}} = \frac{\#\text{ de personnes conservatrices}}{\text{nombre total de répondants}} = 0.21$$

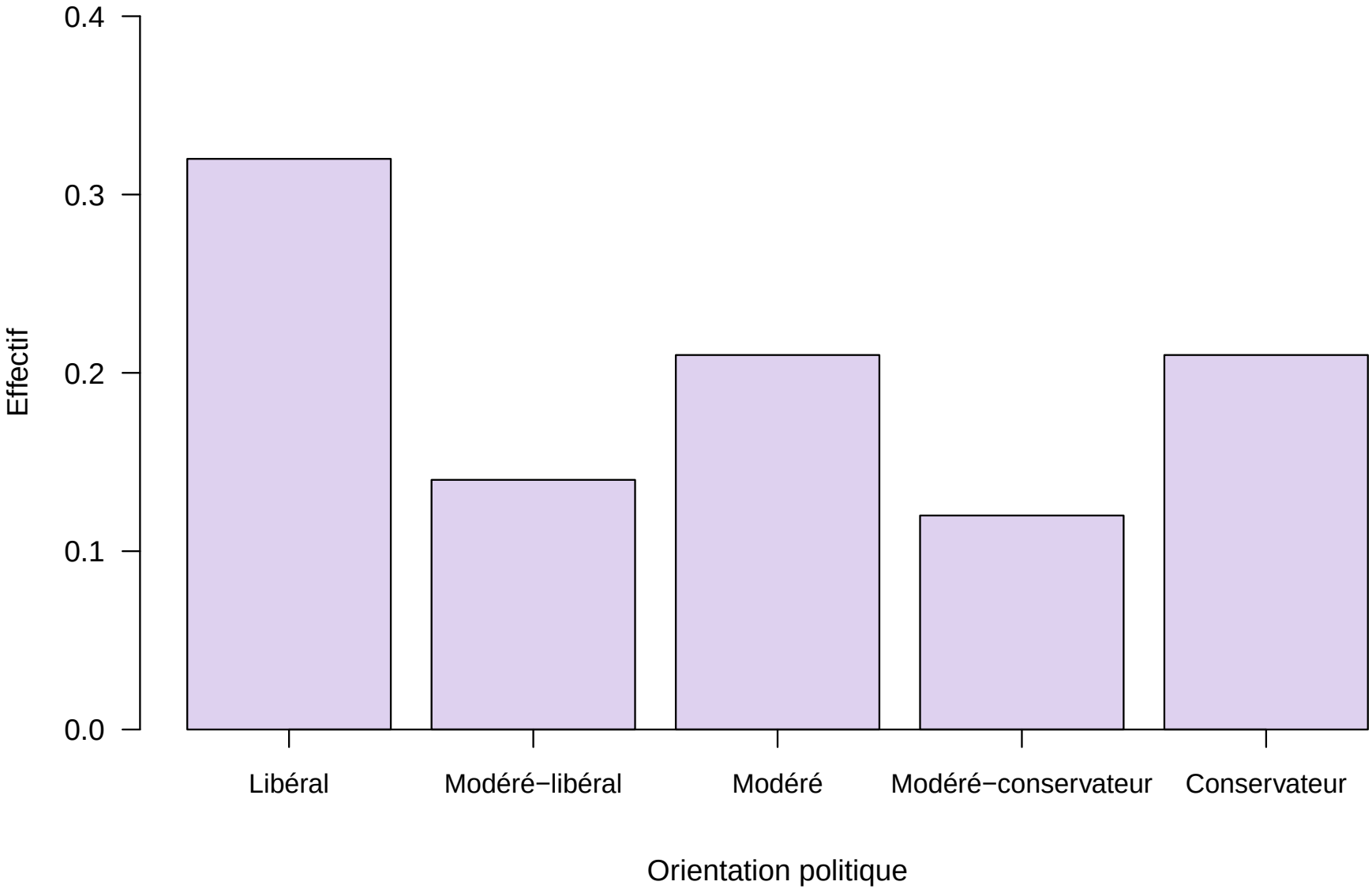
Visualiser les données catégorielles

Donner une image claire de ce que contiennent les données

Souligner les différences/similitudes

Les barplot sont généralement les meilleurs

Barplot


0.4
0.3
0.2
0.1
0.0
Effectif
Libéral
Modéré−libéral
Modéré
Modéré−conservateur
Conservateur
Orientation politique

Statistiques

Type de variable	Statistiques	Graphiques
Catégorielle	Proportions	Barplot
Quantitative Continue	Moyenne, Écart type Médiane, IQR Résumé à 5 nombres	Histogramme Boxplot