UQÀM

Fiche d'exercices - Chapitre 2

STT 1000 - Automne 2025

Les prochaines questions portent sur un exemple classique popularisé par Ronald Fisher en 1936, soit le jeu de données Iris. Cinq caractéristiques ont été notées sur 150 fleurs de genre iris: la longueur et la largeur d'un sépale, ainsi que la longueur et la largeur d'un pétale, et finalement l'espère à laquelle chaque iris appartient, parmi trois espèces représentées: setosa (se), versicolor (ve), et virginica (vi). Les mesures numériques sont en centimètres. Toutes les données dont il est question sont des sous-ensembles des données Iris.

Exercice 2.1. Déterminez quels sont les individus, quelle est la population, et à quel type appartient chacune des cinq variables du jeu de données Iris.

Exercice 2.2. Considérons l'échantillon de 11 longueurs de pétales suivant:

$$1.5,\ 5.6,\ 4.2,\ 1.5,\ 5.6,\ 3.8,\ 1.7,\ 4.4,\ 5.6,\ 6.3,\ 4.7. \tag{1}$$

(a) Calculez la moyenne, la médiane, la moyenne tronquée au niveau $\alpha = 0.2$, le mode, la variance, l'écart-type et l'étendue de cet échantillon.

(b) Répétez la partie (a) après avoir converti les données en pouces (vous pouvez utiliser la formule approximative 1 pouce $\approx$ 2.5cm).

(c) Répétez la partie (a) après avoir centré-réduit l'échantillon.

(d) En se référant à nouveau à l'échantillon original (2), tracez un histogramme des données basé sur une partition en 6 sous-intervalles de longueur 1cm.

(e) Comment votre histogramme en (d) serait-il affecté si vous travailliez plutôt avec les données centrées-réduites?

(f) Tracez un diagramme en boîte à moustache des données.

(g) Comment votre diagramme en (f) serait-il affecté si vous travailliez plutôt avec les données centrées-réduites?

Exercice 2.3. Considérons l'échantillon de 20 espèces suivant:

$$\text{se, ve, vi, ve, ve, ve, se, ve, ve, se, ve, se, ve, ve, se, se, se, ve, vi, se} \tag{2}$$

(a) Pour chaque modalité M_j de la variable espèce, construisez les échantillons de variables indicatrices $\mathbb{1}_{M_j}(x_1), \ldots, \mathbb{1}_{M_j}(x_n)$.

(b) Déduisez de la partie (a) les fréquences et les proportions échantillonnales de chaque modalité.

(c) Utilisez un diagramme à barres et un diagramme circulaire pour représenter ces données.

Exercice 2.4. Supposons qu'un échantillon de 40 largeurs de pétales d'iris setosa et qu'un échantillon de 25 largeurs de pétales d'iris virginica ont été tirés.

(a) On a observé une moyenne de 0.25 pour les iris setosa et une moyenne de 1.98 pour les iris virginica. Avec cette information, est-il possible de calculer la moyenne de l'échantillon combiné de $40 + 25 = 65$ iris? Si oui, calculez-la.

(b) De plus, on a observé une variance de 0.013 pour les iris setosa et une moyenne de 0.063 pour les iris virginica. Avec cette information et celle de la partie (a), est-il possible de calculer la variance de l'échantillon combiné? Si oui, calculez-la.
Indice: rappelons que la variance échantillonnale satisfait

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

(c) Et si, en (b), vous connaissiez seulement les variances des deux échantillons, mais pas leurs moyennes. Serait-il possible de calculer la variance de l'échantillon combiné?

Exercice 2.5. Considérons le diagramme en boîte à moustache suivant (Figure 1), qui a été produite avec l'entièreté des observations, c'est-à-dire 50 iris de chacune des trois espèces.

Dites, avec justification, si chacun des énoncés suivants est vrai ou faux.

(a) Les distributions des largeurs de pétale des trois espèces d'iris sont symétriques.

(b) La largeur médiane chez les virginica est plus grande que la largeur médiane chez les versicolor, qui est elle-même plus grande que la largeur médiane chez les setosa.

(c) La largeur de pétale de n'importe quel iris setosa est plus petite que la largeur de pétale de n'importe quel iris versicolor.

(d) La largeur de pétale de n'importe quel iris versicolor est plus petite que la largeur de pétale de n'importe quel iris virginica.

(e) La médiane globale du jeu de données combiné (c'est-à-dire les 150 observations sans différencier par espèce) est d'environ 1.3.

(f) Il n'y a pas de valeur extrême parmi les iris setosa.

(g) Le troisième quartile du jeu de données combiné est d'environ 2.3.

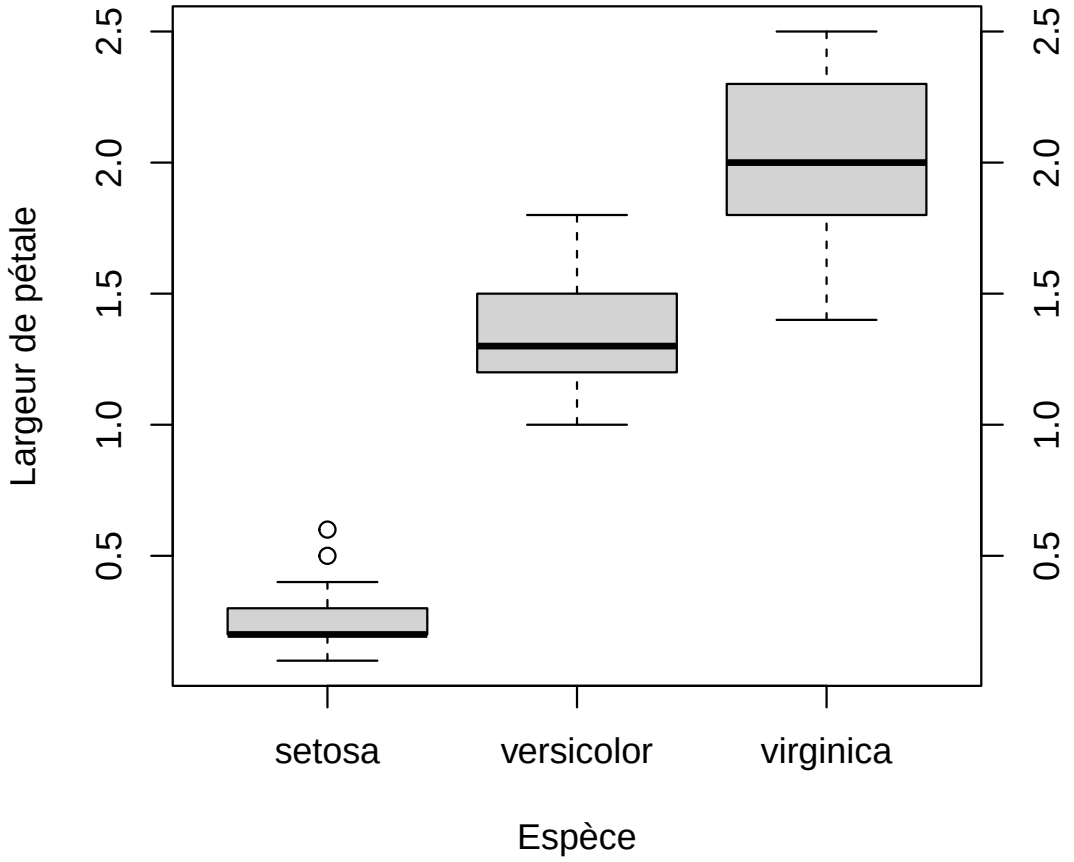

Figure 1: Largeur de pétale en fonction de l'espèce d'iris.

Exercice 2.6. Considérons quatre jeux de données bivariés, représentés par les quatre nuages de points suivants (Figure 2).

(a) Pour chaque jeu de données, y a-t-il un lien visible entre les deux variables? Si oui, est-il linéaire? Est-il positif ou négatif?

(b) Lequel des quatre nuages de points illustre le lien le plus fort?

(c) Lequel des quatre nuages de points illustre les données les plus dispersées? N'oubliez-pas de considérer l'échelle sur les axes de chaque graphique.

Exercice 2.7. Chacun des tableaux croisés à la page suivante contient les proportions observées de deux variables qualitatives. Complétez-les.

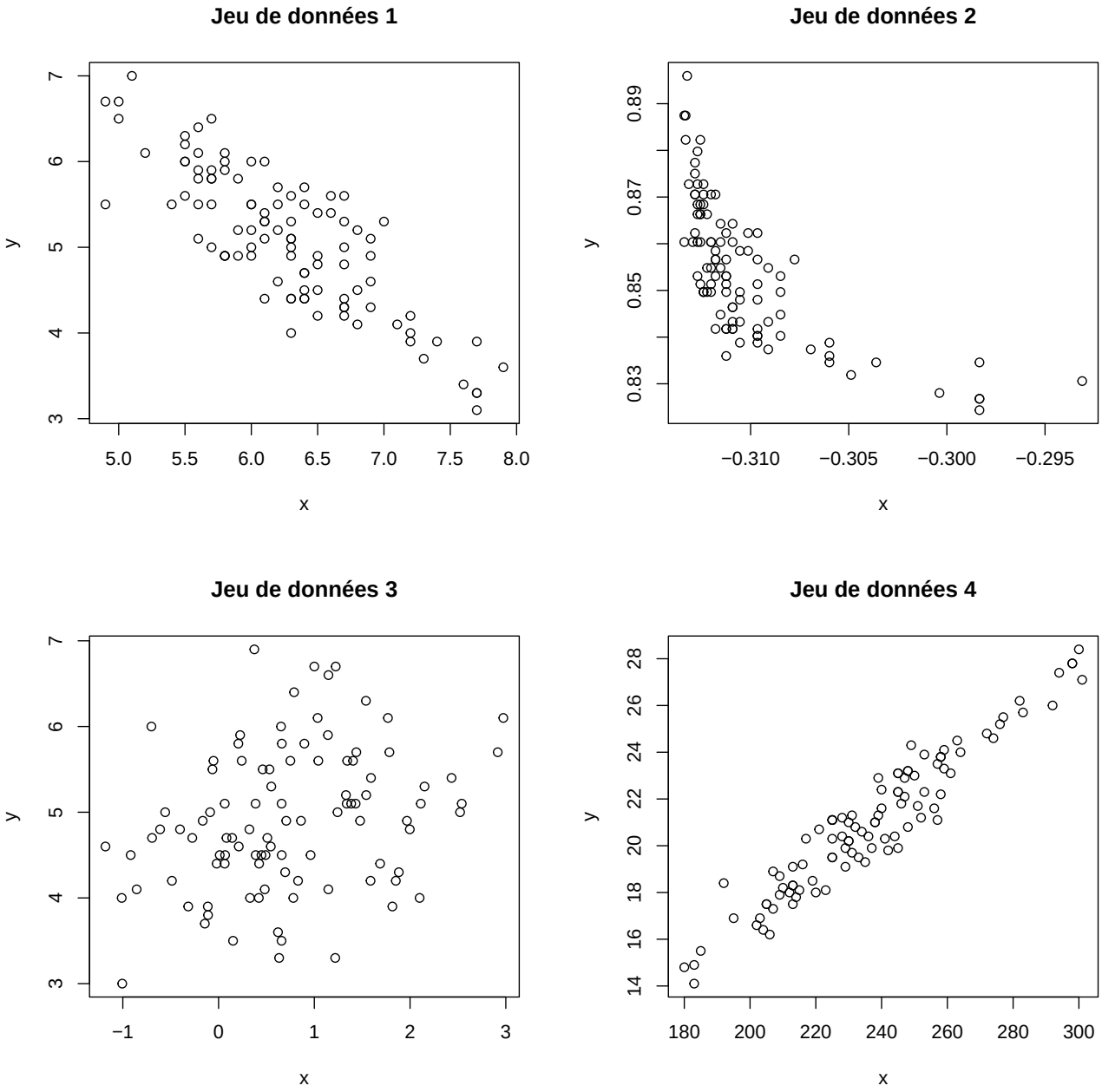

Figure 2: Nuages de points.

	A	B	Total
A	0.05		
B	0.17	0.05	
C	0.14	0.06	
D	0.28	0.01	
Total			1

	A	B	C	Total
A		0.02	0.15	0.2
B		0.09		0.51
C	0.2			0.29
Total	0.55	0.13	0.32	