

Feuille TD 1

Exercice 1.

1. Donner des exemples de populations avec des échantillons.
2. Donner des exemples de variables
(a) qualitatives, (b) quantitatives, (c) discrètes, (d) continues.

Exercice 2. On considère l'échantillon $A C B A B C C C A A C A$.

1. Déterminer le tableau avec les fréquences et les fréquences relatives de l'échantillon.
2. Quels sont les modes de l'échantillon ?
3. Dessiner un camembert avec la distribution de l'échantillon.

Exercice 3. On considère l'échantillon 1 3 9 20 6 3.

1. Déterminer le tableau avec les fréquences, les fréquences relatives et les fréquences relatives cumulées de l'échantillon.
2. Quels sont les modes de l'échantillon ?
3. Dessiner un diagramme de colonnes avec les fréquences relatives de l'échantillon.
4. Trouver un moyen de représenter graphiquement les fréquences relatives cumulées de l'échantillon.

Exercice 4.

1. Déterminer la moyenne et la médiane de l'échantillon 1 3 9 20 6 3.
2. Donner des exemples de populations pour lesquelles la médiane est plus adaptée pour caractériser la population que la moyenne et vice versa.

Exercice 5. Déterminer la variance et l'écart type de l'échantillon 1 3 9 20 6 3.

Exercice 6. On considère un échantillon à valeurs réelles.

1. On ajoute à toutes les valeurs un nombre réel c . Comment est-ce que cela affecte la moyenne, l'écart type et la médiane ?
2. On multiplie toutes les valeurs par un nombre réel c . Comment est-ce que cela affecte la moyenne, l'écart type et la médiane ?

Exercice 7. Soit x la variable qui décrit les caractéristiques numériques d'un échantillon de taille L .

1. La déviation de la caractéristique x_i de l'individu i est

$$d_i := x_i - \bar{x}.$$

Montrer que $\bar{d} = 0$.

2. Montrer que la variance $V(x)$ de l'échantillon peut être obtenue par la formule

$$V(x) := \frac{1}{L-1} \sum_{i=1}^L d_i^2 = \frac{1}{L-1} \left(\sum_{i=1}^L x_i^2 - \frac{(\sum_{i=1}^L x_i)^2}{L} \right) = \frac{L}{L-1} (\overline{x^2} - (\bar{x})^2).$$

3. Que vaut $\frac{V(10x)}{V(x)}$?

Il arrive parfois que si on considère un jeu de données globalement, on observe une certaine tendance, alors que si on sépare les données en plusieurs catégories, on observe une tendance différente - apparemment contradictoire. Ce phénomène s'appelle **Paradoxe de Simpson**. Les deux exercices suivants illustrent comment cela peut se produire et pourquoi il faut être prudent avant de tirer des conclusions des données, et la nécessité de chercher des variables latentes qui puissent expliquer la contradiction apparente.

Exercice 8.

Le journal local a examiné les deux hôpitaux de la ville, et a constaté qu'à l'hôpital Cochin, 79% des patients des six derniers mois ont survécu, tandis qu'à l'hôpital Conté 90 % des patients ont survécu. Le tableau ci-dessous résume les résultats.

	survie	décès	Total	taux de survie
Cochin	790	210	1000	79.0
Conté	900	100	1000	90.0

Dans une étude plus approfondie, il a été observé que les patients étaient catégorisés lors de l'admission comme étant en condition raisonnable (ou meilleure) ou en condition médiocre (ou pire). Quand les taux de survie ont été examinés pour ces groupes, les tableaux suivants ont été obtenus :

Patients admis avec condition raisonnable ou meilleure :

	survie	décès	Total	taux de survie
Cochin	580	10	590	
Conté	860	30	890	

Patients admis avec une médiocre condition ou pire :

	survie	décès	Total	taux de survie
Cochin	210	200	410	
Conté	40	70	110	

1. Remplissez les quatre cases dans la dernière colonne dans les deux tableaux ci-dessus avec les pourcentages corrects.
2. Comparez les pourcentages dans le premier tableau avec ceux des deux tableaux suivants. Est-ce que vous observez quelque chose d'étrange ?
3. Quel hôpital choisiriez-vous, et pourquoi ? Quelle est la variable latente ?

Exercice 9. Au cours d'une période d'embauche récente, un grand magasin hypothétique, U-Prix, a embauché 62% des hommes qui ont postulé, et 14% des femmes. Un procès a été envisagé puisque ces chiffres semblaient indiquer qu'il y avait une discrimination fondée sur le genre. Lors d'un examen plus approfondi, il a été constaté que les offres d'embauche ne concernaient que deux des départements d'U-Prix : un département de matériel et un département de vêtements pour femmes. Le département de matériel a embauché 60 des 80 candidats hommes et 15 femmes sur 20. Le département de vêtements pour femmes a embauché 2 candidats hommes sur 20 et 30 femmes sur 300 candidates.

	Candidats hommes	Hommes admis	Candidates femmes	Femmes admises
Matériel	80	60	20	15
Vêtements pour femmes	20	2	300	30
Total	100	62	320	45

1. Quel pourcentage de candidats hommes pour le département de matériel a été embauché par U-prix ?
2. Quel pourcentage de femmes candidates pour le département de matériel a été embauché ?
3. Quel pourcentage de candidats hommes pour le département de vêtements pour femmes a été embauché ?
4. Quel pourcentage de femmes candidates pour le département des vêtements pour femmes a été embauché ?
5. Vous êtes avocat d'une femme plaignante. Comment justifieriez-vous qu'il y a une discrimination basée sur le genre ?
6. Vous êtes avocat d'U-prix. Comment justifieriez-vous qu'il n'y a pas de discrimination basée sur le genre ?
7. Comment voteriez-vous si vous étiez le juge ou si vous étiez membre d'un jury ?