Université Claude Bernard - Lyon 1 **Semestre automne 2023-2024**

L2 - Informatique UE : Statistiques pour l'Informatique

Feuille de TD 1

Exercice 1.1

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

Exercice 1.2 On considère l'échantillon $A\,C\,B\,A\,B\,C\,C\,C\,A\,A\,C\,A$. (Il a une unique variable statistique).

1. Quelle est la nature de cette variable ? Déterminer le tableau avec les effectifs et les fréquences (relatives) de l'échantillon. Peut-on calculer des fréquences cumulées ?
2. Quels sont les modes de l'échantillon ?
3. Dessiner un diagramme circulaire avec la distribution de l'échantillon.

Exercice 1.3 On considère l'échantillon 1 3 9 20 6 3. (Il a une unique variable statistique).

1. Quelle est la nature de cette variable ? Déterminer le tableau avec les effectifs, les fréquences et les fréquences cumulées de l'échantillon.
2. Quels sont les modes de l'échantillon ?
3. Dessiner un diagramme en tuyau d'orgue avec les fréquences de l'échantillon.
4. Trouver un moyen de répresenter graphiquement les fréquences cumulées de l'échantillon.

Exercice 1.4

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 le milieu de la population que la moyenne et vice versa.
3. Déterminer la variance empirique, la variance non-biaisée et l'écart type (non-biaisé) de l'échantillon 1 3 9 20 6 3.

Exercice 1.5 On considére 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 1.6 Un atelier réalise le séchage de boues d'origine industrielle. Il obtient à la fin du processus des déchets. On a observé les poids suivants de déchets après le traitement de 100 kg de boues :

$$4,7 \quad 4,3 \quad 4,5 \quad 4,9 \quad 4,2 \quad 4,7 \quad 4,0 \quad 4,2 \quad 5,0 \quad 3,9 \quad 4,6 \quad 4,6$$

$$4,8 \quad 4,4 \quad 4,2 \quad 4,6 \quad 4,3 \quad 4,9 \quad 4,0 \quad 4,5 \quad 4,1 \quad 4,4 \quad 4,3 \quad 4,3$$

Notons x cette série statistique.

a) Etablir le tableau des effectifs, fréquences, fréquences cumulées.

b) Tracer un histogramme. On choisira la subdivision $a_0 = 3,9 \leq a_1 = 4.3$‘$a_2 = 4.5 \leq a_3 = 4.7 \leq a_4 = 5$.

c) Déterminer la médiane et les quartiles. Tracer le diagramme à moustache (boxplot).

Exercice 1.7 On prélève $n = 30$ échantillons de pluies provenant du sud de la Pologne. On mesure le pH de ces échantillons et on note cette série statistique $(y_i)_{1\leq i\leq 30}$:

4.60,	4.79,	4.81,	4.82,	4.86,	4.89,	5.03,	5.06,	5.10,	5.14
5.17,	5.18,	5.28,	5.28,	5.32,	5.44,	5.45,	5.55,	5.62,	5.63
5.64,	5.70,	5.77,	5.79,	5.81,	5.82,	5.83,	5.85,	5.97,	6.92

a) Faire un histogramme de ces données. On choisira un nombre de classes égal à l'entier supérieur à $1 + \log(n)/\log(2)$. (ici on prendra des intervalles égaux).

b) Calculer la médiane et les quartiles.

c) Tracer le diagramme à moustache (boxplot).

Exercice 1.8 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 non-biaisée $var(x)$ de l'échantillon peut être obtenue par les formules suivantes

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

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

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** ». L'exercice suivant illustre 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 1.9

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 (en %)
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 ?