Statistique à une variable

Exercice 1

Pour étudier le nombre d'enfants de moins de 18 ans par famille, on choisit un échantillon de familles et pour chacune d'elles, on note le nombre d'enfants. La répartition des familles de l'échantillon suivant le nombre d'enfants est donnée par le tableau :

nombre k d'enfants	0	1	2	3	4	5	6	7	8
nombre de familles ayant k enfants	91	146	104	63	47	33	10	4	2

1. Construire le diagramme en bâtons des fréquences de la série statistique.
2. Déterminer et représenter la fonction de répartition.
3. Calculer le nombre moyen d'enfants par famille dans l'échantillon.
4. Donner la médiane et les quartiles de cette série.
5. Dessinez le diagramme en boîte à moustache.

Exercice 2

Les salaires mensuels payés aux ouvriers d'une entreprise se répartissent comme suit :
102 ouvriers gagnent entre 2400 et 2999 francs,
104 ouvriers gagnent entre 3000 et 3299 francs,
163 ouvriers gagnent entre 3300 et 3599 francs,
121 ouvriers gagnent entre 3600 et 3899 francs,
57 ouvriers gagnent entre 3900 et 4199 francs,
48 ouvriers gagnent entre 4200 et 5000 francs.

1. Dessinez l'histogramme des fréquences et le polygone des fréquences cumulées.
2. Calculez le mode, la médiane.
3. Calculez le salaire mensuel moyen.

Exercice 3

On considère les statistiques suivantes sur les taux de réussites au baccalauréat de deux lycées :

	Lycée A	Lycée B	Total
Échecs	63	16	79
Réussites	2037	784	2821
Total	2100	800	2900
Taux d'échec	0,030	0,020	0,027

Quel lycée choisiriez-vous ? Une deuxième étude, plus fine, sépare les individus en deux groupes, ceux qui sont issus d'un milieu défavorisé et les autres :

	Favorisé			Défavorisé		
	Lycée A	Lycée B	Total	Lycée A	Lycée B	Total
Échecs	6	8	14	57	8	65
Réussites	594	592	1186	1443	192	1635
Total	600	600	1200	1500	200	1700
Taux d'échec	0,010	0,013	0,016	0,038	0,040	0,038

Quel lycée choisiriez-vous ? Expliquer le paradoxe (on observera que pour chaque lycée, le taux d'échec du premier tableau est une moyenne pondérée des deux taux du deuxième tableau par une formule que l'on détaillera).

Exercice 4

Les salaires annuels des 30 employés d'une entreprise sont les suivants (en centaines d'euros), présentés par ordre croissant :

100	100	100	110	120	140	150	150	150	160
160	160	180	180	190	190	200	200	200	210
220	230	230	250	260	290	340	410	420	530

1. Donner la médianes et les quartiles de cette série.
2. Calculer la moyenne.
3. Tracer l'histogramme en regroupant les données en classes de longueur 50.

4*. Les temps sont durs ; il faut faire des économies. Peut-on baisser la masse salariale de 25 000 euros en respectant les contraintes suivantes :
– les salaires inférieurs à la médiane ne sont pas modifiés,
– si X gagne plus que Y alors, après la modification, c'est toujours le cas et leur différence de salaire est divisée par au plus 2,
– un salaire ne baisse pas de plus de 10% ?
Si oui, proposer une répartition respectant les contraintes. Sinon dire pourquoi.

Exercice 5

En 2007, le taux brut de mortalité en Inde est inférieur à celui de la France : 8 pour 1000 contre 9 pour 1000. Pourtant à tout âge le taux de mortalité est inférieur en France à ce qu'il est en Inde. Expliquer.

Exercice 6

Soit $(x_1, ...x_n)$ une suite de données numériques. Notons $\bar{x}$ et s les moyennes et écarts type associés.

1. Soit a un réel, que valent les moyennes et écarts type des suites $(x_i - a)$ et (x_i/a) ?
2. Que valent la moyenne et l'écarts type de la suite $(x_i - \bar{x})/s$?

Exercice 7

Soit x un ensemble de données séparé en deux sous-ensembles y et z de taille n_y et n_z, montrer que

$$\bar{x} = p_y\bar{y} + p_z\bar{z}, \quad p_y = \frac{n_y}{n_y + n_z}, \quad p_z = \frac{n_z}{n_y + n_z}$$
$$s_x^2 = \{p_y s_y^2 + p_z s_z^2\} + \{p_y(\bar{y} - \bar{x})^2 + p_z(\bar{z} - \bar{x})^2\}.$$

Pour la seconde identité, on commencera par montrer que

$$\sum (y_i - \bar{x})^2 = \sum (y_i - \bar{y})^2 + n_y(\bar{y} - \bar{x})^2.$$

$\bar{x}$ est donc une moyenne pondérée des moyennes. s_x^2 est la somme de deux termes, le premier étant la moyenne pondérée des variances, appelée variance intra-classe ; montrer que le second, appelé variance inter-classe, peut s'interpréter comme la variance d'une certaine variable aléatoire.

Exercice 8 *

Soit $(x_1, ...x_n)$ une suite de données numériques. Montrer que la médiane est la valeur pour laquelle la somme des distances des données à cette valeur est minimale. On remarquera que la fonction $y \to \sum_i |x_i - y|$ est continue, affine par morceaux, avec une dérivée entière sur chaque morceau. On pourra traiter séparément les cas "n pair" et "n impair".