Feuille 1 : Statistiques descriptives

Exercice 1 (Démographie familiale)
Pour étudier le nombre d'enfants par foyer, on réalise un sondage auprès de 500 foyers. Le tableau suivant résume les données recueillies :

nombre d'enfants par foyer	0	1	2	3	4	5	6	7	8
nombre de foyers concernés	91	146	104	63	47	33	10	4	2

1. *Construire le diagramme en bâtons de la série statistique.*
2. *Déterminer le diagramme circulaire associé.*
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 (Répartition salariale)
Les salaires mensuels en euros payés aux 595 employé·e·s d'une entreprise se répartissent comme suit :

intervalle de salaire	$[500, 1000[$	$[1000, 1250[$	$[1250, 1500[$	$[1500, 2000[$	$[2000, 2500[$	$[2500, 3500[$
nombre d'employé·e·s	102	104	163	121	57	48

1. *Dessinez l'histogramme et le polygone des fréquences cumulées.*
2. *Déterminer la classe modale, l'intervalle de salaire médian.*
3. *Au vu des données, peut-on calculer le salaire mensuel moyen ? Déterminer la moyenne approchée en considérant le point milieu de chaque intervalle.*

Exercice 3 (Réussite au bac)
On considère les statistiques suivantes sur les taux de réussite 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

Au vu des nouvelles données, quel lycée choisiriez-vous ? Expliquer le paradoxe apparent.

Exercice 4 (Répartition salariale, le retour)
La série suivante représente les salaires annuels des 30 employé·e·s d'une entreprise, exprimés en milliers d'euros et par ordre croissant :

10	*10*	*10*	*11*	*12*	*14*	*15*	*15*	*15*	*16*
16	*16*	*18*	*18*	*19*	*19*	*20*	*20*	*20*	*21*
22	*23*	*23*	*25*	*26*	*29*	*34*	*41*	*42*	*53*

1. *Donner la médiane et les quartiles de cette série.*
2. *Calculer la moyenne.*
3. *Dessinez la boîte à moustache associée à la série.*
4. *Tracer l'histogramme en regroupant les données en classes de longueur 5, i.e. en agrégeant les salaires par tranche de 5000 euros. Calculer la moyenne approchée dans ce nouveau cas en considérant le point milieu de chaque intervalle.*

Exercice 5 (Mortalité)
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 (Centrer et réduire)
Soit $\mathrm{x} = (x_1, ...x_n)$ *une suite de données numériques. Notons* $\bar{\mathrm{x}}$ *et* s_{x} *la moyenne et écart type associés.*

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

Exercice 7 (Agrégation de données)
Soit x *un ensemble de données séparé en deux sous-ensembles* y *et* z *de taille* n_{y} *et* n_{z} *respectivement. On note* $\bar{\mathrm{x}}$, $\bar{\mathrm{y}}$, $\bar{\mathrm{z}}$ *les moyennes empiriques associées et* $s_{\mathrm{x}}, s_{\mathrm{y}}, s_{\mathrm{z}}$ *les écart types correspondants. Montrer que l'on a la relation*

$$\bar{\mathrm{x}} = p_{\mathrm{y}}\bar{\mathrm{y}} + p_{\mathrm{z}}\bar{\mathrm{z}}, \quad \text{où} \quad p_{\mathrm{y}} := \frac{n_{\mathrm{y}}}{n_{\mathrm{y}} + n_{\mathrm{z}}}, \quad p_z := \frac{n_{\mathrm{z}}}{n_{\mathrm{y}} + n_{\mathrm{z}}}.$$

Montrer que l'on a également la relation

$$s_{\mathrm{x}}^2 = \left(p_{\mathrm{y}} s_{\mathrm{y}}^2 + p_{\mathrm{z}} s_{\mathrm{z}}^2\right) + \left(p_{\mathrm{y}}(\bar{\mathrm{y}} - \bar{\mathrm{x}})^2 + p_{\mathrm{z}}(\bar{\mathrm{z}} - \bar{\mathrm{x}})^2\right).$$

La variance totale s_{x}^2 *est ainsi la somme de deux termes, le premier étant la moyenne pondérée des variances* s_{y}^2 *et* s_{z}^2*, appelée variance intra-classe ; le second terme est lui appelé variance inter-classe (pourquoi ?).*

Exercice 8 (Médiane comme minimiseur)
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".*