

## Statistiques à une variable

### Exercice 1

On considère les données suivantes sur la distribution de la population active britannique (en milliers d'habitants)

|            | Juin 1959 | Juin 1966 |
|------------|-----------|-----------|
| Armée      | 565       | 417       |
| Employés   | 23242     | 24974     |
| Employeurs | 1677      | 1673      |
| Chômeurs   | 389       | 253       |

1. De combien (en %) a augmenté l'effectif des armées en Grande Bretagne ?
2. Même question pour l'effectif normalisé par le total de la population active
3. De combien diriez-vous que le chômage a augmenté en Grande Bretagne ?

### Exercice 2

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 3

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.

#### Exercice 4

On a relevé les salaires annuels d'un échantillon de 30 personnes. Les valeurs sont les suivantes (en centaines d'euro), présentées par ordre croissant :

|     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 186 | 201 | 211 | 238 | 260 | 277 | 277 | 281 | 292 | 295 |
| 318 | 325 | 325 | 328 | 336 | 431 | 432 | 496 | 543 | 567 |
| 571 | 585 | 623 | 625 | 696 | 747 | 788 | 871 | 944 | 959 |

- Donner la médiane de cette série.
- Calculer la moyenne et l'écart-type.
- Tracer l'histogramme en regroupant les données en classe de longueur 100.
- Calculer la médiane, la moyenne et l'écart-type sur les données regroupées en classes, comparer les résultats à ceux obtenus sur les données initiales.

#### Exercice 5

Soit  $(x_1, \dots, x_n)$  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'écart type de la suite  $(x_i - \bar{x})/s$  ?

#### Exercice 6

Soit  $(x_1, \dots, x_n)$  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  $y \rightarrow \sum_i |x_i - y|$  est continue, affine par morceaux, avec une dérivée entière sur chaque morceau. On pourra traiter séparément les cas « $n$  pair» et « $n$  impair».

#### 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.