Rappels de statistique

1 Vocabulaire

Individu: élément sur lequel porte l'étude statistique.

Population: ensemble de tous les individus. Il est très important de toujours bien choisir la population pour faire une étude statistique puisque la fiabilité des résultats de l'étude en dépendent.

Échantillon: partie de la population. On a très rarement accès à la population dans son ensemble, on étudie donc seulement un échantillon.

Caractère: propriété observée sur l'individu auquel on s'intéresse.
Un caractère est qualitatif s'il est lié à une observation ne faisant pas l'objet d'une mesure (par exemple la région de résidence de chaque français observée lors du dernier recensement).
Un caractère est quantitatif s'il est mesurable. Il peut être discret si les valeurs observées sont isolées (par exemple le nombre d'enfants par famille observé aussi lors du dernier recensement). Il peut être continu s'il peut prendre, au moins théoriquement, n'importe quelle valeur d'un intervalle de nombre réel (par exemple la taille de chaque individu).

Classe: sous-ensemble de la population rassemblant les individus qui ont la même valeur pour un caractère ou qui ont des valeurs contenues dans un intervalle (par exemple, les habitants de la région PACA, ou les familles de deux enfants, ou encore les habitants dont la taille en centimètre appartient à l'intervalle [164; 170[).

2 Effectifs et fréquences

L'effectif d'une classe est son nombre d'éléments. On dispose d'une série statistique quand on connait l'effectif n_i pour chaque valeur x_i du caractère. On peut la représenter sous la forme d'un tableau :

Valeurs prises par le caractère (ou classe)	x_1	x_2	$\cdots$	x_i	$\cdots$	x_p
Effectifs correspondants	n_1	n_2	$\cdots$	n_i	$\cdots$	n_p

L'effectif total est le nombre total d'individus observés :

$$n = n_1 + \ldots + n_p = \sum_{i=1}^{p} n_i.$$

La fréquence d'une classe est la proportion d'individus appartenant à cette classe. Ainsi la i-ème classe a pour fréquence :

$$f_i = \frac{n_i}{n}.$$

On remarque que :

$$\sum_{i=1}^{p} f_i = \frac{n_1 + \cdots + n_p}{n} = 1.$$

Si les valeurs des caractères $x_1, \ldots, x_p$ peuvent être ordonnées, on peut considérer les fréquences cumulées croissantes qui cumulent les fréquences associées aux valeurs du caractère inférieures ou égales à x_i. On note ces fréquences cumulées F_i et

$$F_i = \sum_{j=1}^{i} f_j \text{ pour } i = 1, \ldots, p.$$

On peut bien entendu définir de façon analogue les fréquences cumulées décroissantes, les effectifs cumulés croissants et décroissants.

3 Caractéristiques de position

Mode: Valeur correspondant à l'effectif maximal.

Valeurs maximale et minimale de la série: On note ces valeurs x_{max} et x_{min}.

Moyenne:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{p} n_i x_i}{n} = \sum_{i=1}^{p} f_i x_i.$$

Médiane: La médiane est la valeur telle que 50 % des observations de l'échantillon lui sont inférieures (et 50% lui sont supérieures).
Si le nombre d'observations est pair : la médiane est la moyenne entre les observations $\frac{n}{2}$ et $\frac{n+2}{2}$.
Si le nombre d'observations est impair : la médiane est l'observation $\frac{n+1}{2}$.

Premier et troisième quartiles: Le premier Q_1 et le troisième quartile Q_3 sont les valeurs qui partagent la série avec un quart de l'effectif avant Q_1 et trois quarts de l'effectif avant Q_3.

4 Mesures de dispersion

Variance: La variance est la quantité suivante

$$V(x) = \frac{\sum_{i=1}^{p} n_i (x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{p} n_i x_i^2}{n} - \overline{x}^2.$$

Écart type: L'écart type est simplement la racine carrée de la variance: $\sigma = \sqrt{V(x)}$.

Étendue: L'étendue est la différence entre la plus grande valeur du caractère et la plus petite: $x_{max} - x_{min}$.

Écart interquartiles: L'écart interquartile est le nombre $Q_3 - Q_1$. Il permet de mesurer la dispersion du caractère autour de la médiane.