Statistiques
Moyenne, Médiane

On considère une population de n individus, numérotés de 1 à n. Sur cette population on mesure un caractère x : taille, poids, salaire moyen, ... On note x_i la valeur du caractère pour l'individu numéro i.

On obtient ainsi $(x_i)_{i=1}^n$ une série statistique simple, à une variable. On numérote les individus de manière à ce que la série soit ordonnée. En pratique, on est en général obligé de trier les valeurs de la série.

$$x_1 \leq x_2 \leq \cdots \leq x_n.$$

Lorsque la série prend plusieurs fois les même valeurs, on utilise les séries statistiques pondérées :

$$\begin{aligned}
(y_j, n_j)_{j=1}^p, \qquad & p \leq n, \\
y_1 < y_2 < \cdots < y_n, & \\
\{y_1, y_2, \cdots, y_p\} &= \{x_1, x_2, \cdots, x_n\}, \\
n_j &= \text{ cardinal } \{i \in \{1, 2, \cdots, n\} \text{ tel que } x_i = y_j\}, \\
N_j &= \text{ cardinal } \{i \in \{1, 2, \cdots, n\} \text{ tel que } x_i \leq y_j\}, \\
f_j &= \frac{n_j}{n}, \\
F_j &= \frac{N_j}{n}.
\end{aligned}$$

C'est à dire que :

- p est le nombre de valeurs que prend la série $(x_i)_{i=1}^n$,
- $\{y_1, y_2, \cdots, y_p\}$ est l'ensemble des valeurs prise par la série $(x_i)_{i=1}^n$.,
- n_j est l'effectif des individus qui prennent la valeur y_j, i.e. le nombre de fois que la série $(x_i)_{i=1}^n$ prend la valeur y_j,
- N_j est l'effectif cumulé pour la valeur y_j, i.e. le nombre de fois que la série $(x_i)_{i=1}^n$ prend une valeur infrieure ou égale à y_j,
- f_j est la fréquence ou proportion d'individus de la population qui prennent la valeur y_j pour le caractère x,
- $F_j = \dfrac{N_j}{n}$ est la fréquence cumulée.

Lorsque l'on a beaucoup de valeurs, ou un caractére continu, on peut regrouper la série $(x_i)_{i=1}^n$ par classe :

$$([a_{j-1}, a_j[, n_j)_{j=1}^p,$$

où n_j est l'effectif des individus dont la valeur du caractère est dans l'intervalle $[a_{j-1}, a_j[$. Ainsi, on a toujours

$$n_1 + n_2 + \cdots + n_p = n.$$

1 Calcul de la moyenne

On note $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ la moyenne statistique de la série $(x_i)_{i=1}^n$.

Lorsque la série $(x_i)_{i=1}^n$ est regroupée par classe on note $\widetilde{x} = \dfrac{1}{n}\sum_{j=1}^{n} b_j n_j$ la moyenne aprrochée de la série $(x_i)_{i=1}^n$, où b_j désigne le milieu de l'intervalle $[a_{j-1}, a_j[$.

1. Montrez que $\overline{x} = \sum_{j=1}^{p} y_j f_j$.
2. Soit $h = \max_{j=1}^{p} |a_j - a_{j-1}|$, vérifiez que $|\overline{x} - \widetilde{x}| \leq \dfrac{h}{2}$.
   Quand a-t-on égalité : $|\overline{x} - \widetilde{x}| = \dfrac{h}{2}$?

2 Ecart-type

Soit $f(x) = \dfrac{1}{n}\sum_{i=1}^{n} |x_i - x|^2$.

On note $s = \sqrt{f(\overline{x})}$, l'écart-type et, s^2 la variance de la série $(x_i)_{i=1}^n$.

1. Montrez que $f > 0$ dès que $p > 1$.
2. Minimiser f sur $\mathbb{R}$.
3. En déduire que la valeur moyenne de la série est le nombre qui minimise l'écart quadratique moyen.
4. Calculer s en fonction $a = \overline{x} - x_1$ pour $n = 2$.

3 Stabilité de la moyenne

On suppose que l'on cherche à mesurer une quantité ξ précisément. On fait plusieurs mesures $(x_i)_{i=1}^n$ de cette quantité. En raison d'erreurs de mesures on a

$$x_i = \xi + \varepsilon_i,$$

où ε_i désigne l'erreur dûe à la mesure numéro i.

On va étudier la présicion obtenue sur ξ en prenant comme approximation de ξ la moyenne : $\overline{x}$. On remarquera dans tous les cas suivants que bien que chaque x_i est une mauvaise approximation de ξ, $\overline{x}$ donnera une bonne approximation de ξ.

1. On se place dans le cas où les mesures sont indépendantes de même variance, i.e. $(\varepsilon_i)_{i=1}^n$ est une suite de variables aléatoires indépendantes, d'espérance nulle et de variance s^2.
   Montrez que $|\overline{x} - \xi| \leq 2\dfrac{s}{\sqrt{n}}$ avec une probabilité d'environ 95% pour un échantillon assez grand.
2. On suppose que $\varepsilon_i = s(-1)^i$ avec $s > 0$, montrez que $|\overline{x} - \xi| \leq \dfrac{s}{n}$.
   Si n est pair, vérifier que l'on a même $\overline{x} = \xi$.

3. On suppose que $(\varepsilon_i)_i$ est une suite k-périodique de moyenne nulle :

$$\varepsilon_{i+k} = \varepsilon_i \text{ pour tout } i, \quad \frac{1}{k}\sum_{i=1}^{k} \varepsilon_i = 0.$$

Montrez que $|\overline{x} - \xi| \leq \dfrac{A}{n}$, avec $A = \displaystyle\sum_{i=1}^{k} |\varepsilon_i|$.

Vérifiez que la suite $\varepsilon_i = \sin(i\alpha\pi)$ satisfont les hypothèses pour α rationnel.

4. On suppose que les sommes des perturbations sont toujours bornées. C'est à dire qu'il existe un nombre $C > 0$ tel que pour tout j :

$$\left| \sum_{i=1}^{j} \varepsilon_i \right| \leq C.$$

Montrez que $|\overline{x} - \xi| \leq \dfrac{C}{n}$.

Vérifiez que la suite $\varepsilon_i = \sin(i\alpha\pi)$ satisfont les hypothèses pour $\alpha \in I\!R$.

4 Quelques propriétés de la moyenne

1. Associativité : Si une population P de n individus de moyenne du caractère $\overline{x}$, est partagée entre deux populations P_a et P_b, de n_a, respectivement n_b individus et de moyenne du caractère $\overline{x_a}$, respectivement $\overline{x_b}$.

Montrez que $\overline{x} = \dfrac{n_a\overline{x_a} + n_b\overline{x_b}}{n}$.

2. Valeur aberrante : Soit ξ la "vraie" valeur de la moyenne de la série statisiques $(x_i)_{i=1}^{n}$. On suppose que $x_n = \xi$. On est dans le cas où une seule valeur de la série statistique est aberrante. Pour cela on remplace la vraie valeur de x_n par $\xi + A$, $|A| >> 1$.

   (a) Quelle erreur fait-on maintenant sur ξ en l'évaluant avec $\overline{x}$?

   (b) Justifier le procédé qui consiste à enlever les valeurs aberrantes avant de calculer les moyennes, les variances.

5 Calcul de la médiane

Lorsque la série $(x_i)_{i=1}^n$ est ordonnée, on définit la médiane suivant la parité de n :

$$Med \;=\; \begin{cases} x_k & \text{si n = 2k-1,} \\ \dfrac{x_k + x_{k+1}}{2} & \text{si n = 2k} \end{cases}$$

1. Calculer la médiane des séries suivantes $(1,3)$, $(1,1,1)$, $(2,1,1)$.
2. Vérifier que au moins 50% des individus ont une valeur du caractère inférieure ou égale à la médiane. De même, vérifier que au moins 50% des individus ont une valeur du caractère supérieure ou égale à la médiane.

   Ces deux propriétés caractérisent la médiane.
   On dit que la valeur médiane partage en "deux moitiés" la population.
3. Si la série est pondérée, montrez que

$$Med \;=\; \begin{cases} y_k & \text{si } F_{k-1} < 0,5 < F_k, \\ \dfrac{y_{k-1} + y_k}{2} & \text{si } F_k = 0,5 \end{cases}$$

4. Si la série est regroupée par classe. On obtient la classe médiane en appliquant l'algorithme du calcul de la médiane d'une série pondérée utilisant les fréquences cumulées.

   Vérifiez que la médiane appartient bien à la classe médiane obtenue.

6 L'écart absolue moyen

Soit $g(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} |x_i - x|$. On note $\mathcal{M}$ l'ensemble des points qui minimise g sur $\mathbb{R}$.

1. Représenter le graphe de g et expliciter $\mathcal{M}$ pour les séries suivantes : $(-1,1)$, $(-1,0,1)$.
2. Montrer que g est continue et affine sur les intervalles $]-\infty, y_1]$, $[y_1, y_2]$, $\ldots$, $[y_{p-1}, y_p]$, $[y_p, +\infty[$.

   Sur chaque intervalle, on exprimera la pente de g en fonction des fréquences cumulées.
3. En déduire que la médiane apartient toujours à $\mathcal{M}$.
   Plus précisément, montrez que $\mathcal{M}$ est un point ou un intervalle, et que, dans ce dernier cas, la médiane est le milieu de cet intervalle.
4. Ainsi la médiane est un nombre qui minimise l'écart absolu moyen.
   Montrez que $g(Med) \leq s$ où s désigne l'écart-type de la série.