Fiche méthodologique n° 14

Étudier des séries statistiques

Pour les exercices d'application, on considèrera l'exemple de série statistique suivant :

a	4	6	5	1	8	2	7	5	3	4	6	1
b	5	13	6	2	47	2	26	7	1	3	12	3
effectif	1	3	2	1	1	1	4	2	1	2	2	1
a	3	2	4	7	5	3	1	6	7	8	2	7
b	3	1	4	22	5	2	1	15	25	49	3	29
effectif	1	4	2	1	2	3	1	1	2	2	1	1

1 Étudier une série statistique univariée

On considère un caractère quantitatif x et y sur une population de taille $n \geqslant 1$, de modalités $(x_1, x_2, \dots, x_p)$ où $p \geqslant 1$. Pour chaque $i \in [\![1,p]\!]$, on note n_i l'effectif de la modalité x_i et $\mathrm{f}_i = n_i/n$ la fréquence associée. Ainsi :

$$\boxed{\sum_{i=1}^{p} n_i = n} \quad \text{et} \quad \boxed{\sum_{i=1}^{p} \mathrm{f}_i = 1}.$$

1.1 Représenter graphiquement une série statistique univariée

Le diagramme circulaire du caractère x est un disque découpé en p secteurs associés aux modalité x_i et d'aires proportionnelles aux effectifs n_i.

Le diagramme en bâtons du caractère x est la représentation graphique des segments verticaux d'abscisses x_i et de hauteurs proportionnelles aux effectifs n_i.

Si les modalités du caractère x sont regroupées par classes du type $[c_i - \frac{a_i}{2}, c_i + \frac{a_i}{2}[$ (où c_i désigne le centre de la classe de modalités et a_i l'amplitude), alors l'histogramme du caractère x est la représentation graphique des rectangles de base $[c_i - \frac{a_i}{2}, c_i + \frac{a_i}{2}[$ et d'aires proportionnelles aux effectifs n_i, donc de hauteurs proportionnelles à n_i/a_i.

Exercice d'application 1
Représenter le diagramme circulaire et le diagramme en bâtons du caractère a ainsi que le diagramme circulaire et l'histogramme du caractère b dont on regroupera les modalités par classes d'amplitude 5.

Si les modalités du caractère x sont ordonnées dans l'ordre croissant, c'est-à-dire si $x_1 < x_2 < \cdots < x_p$, alors la courbe des fréquences cumulées est la courbe représentative de la fonction :

$$\mathrm{F} : t \mapsto \begin{cases} 0 & \text{si } t < x_1 \\ \sum_{i=1}^{k} \mathrm{f}_i & \text{si } t \in [x_k, x_{k+1}[\quad \text{où } k \in [\![1, p-1]\!] \\ 1 & \text{si } t \geqslant x_p \end{cases}.$$

Remarque. Dans ce cas, la courbe des fréquences cumulées est une courbe constante par morceaux (donc discontinue), passant par les points $\mathrm{F}(x_k) = \sum_{i=1}^{k} \mathrm{f}_i$ où $k \in [\![1,p]\!]$.

Si les modalités du caractère x sont regroupées par classes du type $[c_i - \frac{a_i}{2}, c_i + \frac{a_i}{2}[$ avec $c_1 - \frac{a_1}{2} < c_1 + \frac{a_1}{2} = c_2 - \frac{a_2}{2} < c_2 + \frac{a_2}{2} = \cdots = c_p - \frac{a_p}{2} < c_p + \frac{a_p}{2}$, alors la courbe des fréquences cumulées est la courbe représentative de la fonction :

$$\mathrm{F} : t \mapsto \begin{cases} 0 & \text{si } t < c_1 - \frac{a_1}{2} \\ \sum_{i=1}^{k-1} \mathrm{f}_i + \frac{t-(c_k - \frac{a_k}{2})}{a_k} \mathrm{f}_k & \text{si } t \in [c_k - \frac{a_k}{2}, c_k + \frac{a_k}{2}[\quad \text{où } k \in [\![1, p]\!] \\ 1 & \text{si } t \geqslant c_p + \frac{a_p}{2} \end{cases}.$$

Remarque. Dans ce cas, la courbe des fréquences cumulées est une courbe continue affine par morceaux, passant par les points $\mathrm{F}(c_k - \frac{a_k}{2}) = \sum_{i=1}^{k-1} \mathrm{f}_i$ et $\mathrm{F}(c_k + \frac{a_k}{2}) = \sum_{i=1}^{k} \mathrm{f}_i$ où $k \in [\![1,p]\!]$.

Exercice d'application 2
Représenter la courbe des fréquences cumulées du caractère a ainsi que celle du caractère b dont on regroupera les modalités par classes d'amplitude 5.

1.2 Calculer des caractéristiques de position

Un mode est une modalité d'effectif maximal et une classe modale est une classe de modalités d'effectif maximal.

La moyenne du caractère x est définie par :

$$\boxed{\overline{x} = \frac{1}{n} \sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} \mathrm{f}_i x_i}.$$

Si les modalités sont regroupées par classes, alors la moyenne est définie par une formule similaire obtenue en remplaçant les modalités par les centres de classes.

La médiane est définie par l'abscisse d'intersection de la courbe des fréquences cumulées avec la droite d'équation $y = 1/2$.
Remarque. Dans le cas où la courbe des fréquences cumulées est constante par morceaux, l'intersection peut-être vide ou se faire selon un segment. La médiane est alors définie par l'abscisse du point de discontinuité ou par le milieu du segment d'intersection.

Exercice d'application 3
Calculer le mode, la moyenne et la médiane du caractère a ainsi que la classe modale, la moyenne et la médiane du caractère b dont on regroupera les modalités par classes d'amplitude 5.

1.3 Calculer des caractéristiques de dispersion

L'étendue est différence entre la modalité la plus haute et la modalité la plus basse.

L'écart type du caractère x est définie par :

$$\boxed{s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2} = \sqrt{\sum_{i=1}^{p} \mathrm{f}_i(x_i - \overline{x})^2}}.$$

Si les modalités sont regroupées par classes, alors l'écart type est définie par une formule similaire obtenue en remplaçant les modalités par les centres de classes.

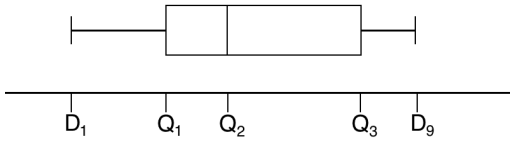
On peut également calculer l'écart type à l'aide de la formule de König-Huygens pour la variance :

$$\boxed{s_x^2 = \overline{(x^2)} - (\overline{x})^2}.$$

Exercice d'application 4
Calculer l'étendue et l'écart type du caractère a ainsi que ceux du caractère b dont on regroupera les modalités par classes d'amplitude 5.

Pour $k \in [\![1, 3]\!]$, le k-ième quartile Q_k est définie par l'abscisse d'intersection de la courbe des fréquences cumulées avec la droite d'équation $y = k/4$. Pour $k \in [\![1, 9]\!]$, le k-ième décile D_k est définie de manière similaire avec la droite d'équation $y = k/10$.

La boîte à moustaches est une représentation de D_1, Q_1, Q_2, Q_3 et D_9 du type :



Exercice d'application 5
Représenter les boîte à moustaches des caractères a et b.

2 Étudier une série statistique bivariée

On considère deux caractères quantitatifs x et y sur une population de taille $n \geqslant 1$, de modalités conjointes $((x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p))$ où $p \geqslant 1$. Pour chaque $i \in [\![1, p]\!]$, on note n_i l'effectif de la modalité conjointe (x_i, y_i) et $\mathrm{f}_i = n_i/n$ la fréquence associée.

2.1 Représenter un nuage de points ou de disques

Le nuage de points du couple (x, y) est la représentation graphique de l'ensemble des points de coordonnées (x_i, y_i).

Le nuage de disques du couple (x, y) est la représentation graphique de l'ensemble des disques centrés aux points de coordonnées (x_i, y_i) et d'aires proportionnelles aux effectifs conjoints n_i, donc de rayons proportionnels à $\sqrt{n_i}$. Le coefficient de proportionnalité est choisi en fonction de l'échelle du graphique.

Exercice d'application 6
Représenter les nuages de disques des couples (a, b) et $(a, \ln(b))$ sur deux graphiques.

2.2 Calculer le coefficient de corrélation affine

La covariance du couple (x, y) est définie par :

$$\boxed{s_{x,y} = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{p} \mathrm{f}_i(x_i - \overline{x})(y_i - \overline{y})}.$$

On peut également calculer la covariance à l'aide de la formule de König-Huygens pour la covariance :

$$\boxed{s_{x,y} = \overline{(xy)} - (\overline{x})(\overline{y})}.$$

Le coefficient de corrélation affine du couple (x, y) est définie par :

$$\boxed{\rho_{x,y} = \frac{s_{x,y}}{s_x s_y}}.$$

Le coefficient de corrélation affine vérifie $|\rho_{x,y}| \leqslant 1$ avec égalité si et seulement si x et y sont en relation affine. Un coefficient de corrélation affine proche de -1 ou $+1$ justifie un ajustement affine.

Exercice d'application 7
Calculer les coefficients de corrélation affine des couples (a, b) et $(a, \ln(b))$.

3 Réaliser un ajustement affine

La droite de régression affine du couple (x, y) obtenue par la méthode des moindres carrés est la droite passant par le point moyen de coordonnées $(\overline{x}, \overline{y})$ et de coefficient directeur $s_{x,y}/s_x^2$, c'est-à-dire la droite d'équation cartésienne :

$$\boxed{y = \frac{s_{x,y}}{s_x^2}(x - \overline{x}) + \overline{y}}.$$

Exercice d'application 8
Représenter sur deux graphiques les nuages de points des couples (a, b) et $(a, \ln(b))$ ainsi que les droites respectives de régression affine obtenues par la méthode des moindres carrés. Déterminer une relation de régression entre a et b du type $b = \lambda q^a$ avec $(\lambda, q) \in \mathbb{R}^2$ et représenter la courbe de cette régression sur le même graphique que celui comportant le nuage de points du couple (a, b). Quel modèle d'ajustement du couple (a, b) est le plus approprié ?