Chapitre 15

Statistique descriptive univariée

La statistique descriptive consiste à analyser une série de données, représentant des valeurs d'une grandeur mesurable ou des caractéristiques non mesurables, relatives à un échantillon d'une population donnée.

15.1 Un peu de vocabulaire statistique

• Un n-échantillon est une partie d'une population complète composée de n individus choisis au hasard.

• Un caractère est une propriété liée aux individus d'une population.
⋆ Un caractère peut être une grandeur mesurable (par exemple la taille). On parle alors de caractère *quantitatif*.
⋆ Un caractère peut être une caractéristique non mesurable (par exemple la couleur des yeux). On parle alors de caractère *qualitatif*.
Au sein d'un échantillon d'individus, les différentes valeurs ou caractéristiques non mesurables que peut prendre le caractère étudié s'appellent les modalités.

Dans la suite de ce chapitre, on s'intéresse à un caractère quantitatif.

On considère les modalités (donc les valeurs deux à deux distinctes) $x_1, \dots, x_p$ avec $1 \leq p \leq n$ que prend une série statistique relative à un n-échantillon.
• Pour tout $i \in [\![1, p]\!]$, l'effectif de la modalité x_i, noté n_i, est le nombre d'individus dont le caractère est égal à la valeur x_i.
• Pour tout $i \in [\![1, p]\!]$, la fréquence de la modalité x_i, notée f_i, est le rapport : $f_i = \dfrac{n_i}{\sum_{k=1}^{p} n_k}$.
• Supposons que la série $x_1, \dots, x_p$ est ordonnée par *ordre croissant*. Les sommes $\sum_{k=1}^{i} f_k$, $1 \leq i \leq p$, sont appelées les fréquences cumulées.

15.2 Caractéristiques de position et de dispersion

Dans les trois prochaines définitions, on considère les modalités (donc les valeurs deux à deux distinctes) $x_1, \dots, x_p$ avec $1 \leq p \leq n$ que prend une série statistique relative à un n-échantillon.

Définition 1 (Moyenne d'une série statistique)

La moyenne, notée $\overline{x}$, est définie par : $\overline{x} = \dfrac{1}{n} \sum_{k=1}^{p} n_k x_k$.

Remarque : La moyenne d'une série statistique est un indicateur de position. Pour faire simple, elle permet de résumer l'ensemble des valeurs de la série à un seul nombre. Encore faut-il préciser si cela a du sens...

Définition 2 (Variance empirique d'une série statistique)

La variance empirique, notée V, est définie par : $V = \frac{1}{n}\sum_{k=1}^{p} n_k(x_k - \overline{x})^2$.

Remarques :
- La variance empirique est la « moyenne du carré des écarts à la moyenne ».
- La variance empirique d'une série statistique est un indicateur de dispersion. Elle exprime à quel point les données fluctuent autour de la moyenne. La moyenne d'une série statistique n'a de sens que si sa variance empirique est très proche de 0.
- On parle de variance empirique car elle est obtenue à partir des données.
- On peut réécrire la formule de la variance empirique à l'aide de la formule de Kœnig : $V = \overline{x^2} - \overline{x}^2$.

Définition 3 (Ecart type d'une série statistique)

L'écart type, noté σ, est défini par : $\sigma = \sqrt{V}$.

Remarques :
- L'écart type d'une série statistique est un indicateur de dispersion. Il exprime à quel point les données fluctuent autour de la moyenne. La moyenne d'une série statistique n'a de sens que si son écart type est très proche de 0.
- L'avantage de l'écart type par rapport à la variance empirique est qu'il s'agit d'une grandeur sans unité.
- Le rapport entre l'écart type et la moyenne, σ/m, est appelé le *coefficient de variation* et mesure la dispersion relative des données par rapport à la moyenne.

On définit maintenant les notions de médiane et de quartiles d'une série statistique $x_1, \ldots, x_n$ (données non nécessairement distinctes) qui nous permettront de tracer la boîte à moustache de cette dernière.

Définition 4 (médiane d'une série statistique)

La médiane d'une série statistique, notée m, dont les modalités sont ordonnées par ordre croissant est une valeur qui partage la population en deux groupes de même effectif.

- Cas où n est pair : Il existe $q \in \mathbb{N}^*$ tel que $n = 2q$. Alors, $m = \frac{1}{2}(x_q + x_{q+1})$.
- Cas où n est impair : Il existe $q \in \mathbb{N}$ tel que $n = 2q+1$. Alors, $m = x_{q+1}$.

Définition 5 (Quartiles d'une série statistique)

Les quartiles, notés Q_1, Q_2 et Q_3, d'une série statistique sont trois valeurs de la série ordonnée qui la partagent en quatre séries de même effectif (25% de l'effectif total).

- 1^{ier} quartile : $Q_1 = \min\left\{x_i, \sum_{x_k \leq x_i} f_k \geq 0,25\right\}$.
- $2^{\text{ième}}$ quartile : $Q_2 = m$.
- $3^{\text{ième}}$ quartile : $Q_3 = \min\left\{x_i, \sum_{x_k \leq x_i} f_k \geq 0,75\right\}$.

Remarques :
- L'écart inter-quartile est le nombre $Q_3 - Q_1$.

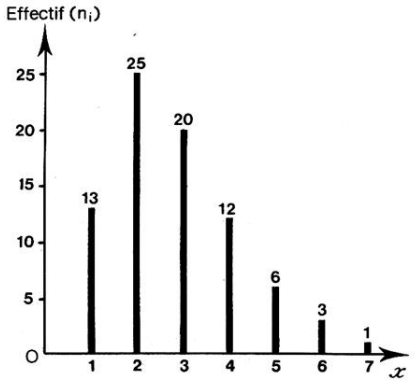
• On définit de manière analogue les déciles $D_1, \ldots, D_9$ d'une série qui sont neuf valeurs de la série ordonnée qui la partagent en dix séries de même effectif (10% de l'effectif total). L'écart inter-décile est le nombre $D_9 - D_1$.

Il existe d'autres indicateurs de position et de dispersion.
⋆ Le mode d'une série statistique est la valeur du caractère la plus fréquente (indicateur de position).
⋆ L'étendue d'une série ordonnée par ordre croissant est l'écart entre la valeur maximale et la valeur minimale du caractère (indicateur de dispersion).

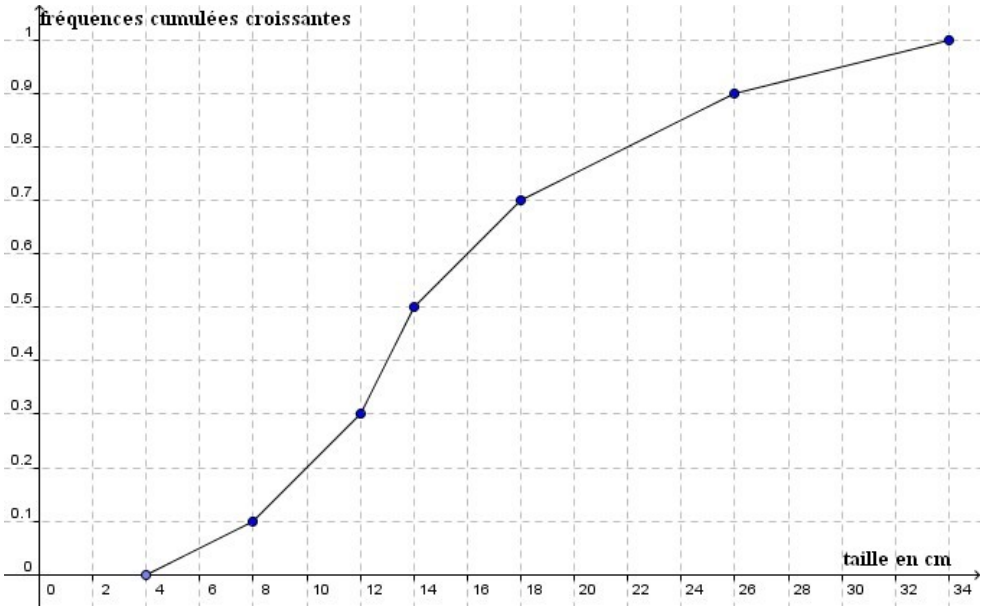
15.3 Représentations graphiques d'une série statistique

On considère les modalités (donc les valeurs deux à deux distinctes) classées par *ordre croissant* $x_1, \ldots, x_p$ avec $1 \leq p \leq n$ que prend une série statistique S relative à un n-échantillon.

• La série statistique S peut être représentée sous la forme d'un diagramme en bâtons. Pour ce faire, on trace un repère orthogonal (O, Ox, Oy) puis on place en abscisse les modalités x_i, $1 \leq i \leq p$, et en ordonnées les effectifs n_i, $1 \leq i \leq p$, correspondants (ou bien les fréquences). On termine en reliant pour tout $i \in [\![1, p]\!]$ le point de coordonnées $(x_i, 0)$ au point de coordonnées (x_i, n_i).



• La série statistique S peut aussi être représentée sous la forme d'une courbe des fréquences cumulées. Pour ce faire, on trace un repère orthogonal (O, Ox, Oy) puis on place en abscisse les modalités x_i, $1 \leq i \leq p$, et en ordonnées les fréquences cumulées $\sum_{k=1}^{i} f_k$, $1 \leq i \leq p$, correspondantes. Le diagramme des fréquences cumulées est la courbe joignant les points de coordonnées $\left(x_i, \sum_{k=1}^{i} f_k\right)$, $1 \leq i \leq p$.



• La boîte à moustache est un diagramme simple permettant de représenter les caractéristiques suivantes d'une série statistique : son étendue, ses quartiles et sa médiane.