Chapitre 2

Étude d'une variable statistique discrète

Le caractère statistique peut prendre un nombre fini raisonnable de valeurs (note, nombre d'enfants, nombre de pièces, ...). Dans ce cas, le caractère statistique étudié est alors appelé un caractère discret.

Dans toute la suite du chapitre, nous considérons la situation suivante :

$$X : \Omega \to \{x_1, x_2, ..., x_n\},$$

avec $\text{Card}(\Omega) := N$ est le nombre d'individus dans notre étude.

Nous allons utiliser souvent l'exemple ci-dessous pour illustrer les énoncés de ce chapitre.

Exemple 8

Une enquête réalisée dans un village porte sur le nombre d'enfants à charge par famille. On note X le nombre d'enfants, les résultats sont données par ce tableau :

x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
n_i *(Effectif)*	*18*	*32*	*66*	*41*	*32*	*9*	*2*

Nous avons

- Ω *ensemble des familles.*
- ω *une famille.*
- X *nombre d'enfants par famille*

$$X : \omega \to X(\omega).$$

On lit, à la famille ω, on associe $X(\omega)$ = le nombre d'enfants de cette famille.

2.1 Effectif partiel - effectif cumulé

On étudie ici un caractère statistique numérique représenté par une suite x_i décrivant la valeur du caractère avec i varie de 1 à k.

2.1.1 Effectif partiel (fréquence absolue)

Définition 7

Pour chaque valeur x_i, on pose par définition

$$n_i = Card\{\omega \in \Omega : \quad X(\omega) = x_i\}.$$

n_i : le nombre d'individus qui ont le même x_i, ça s'appelle effectif partiel de x_i.

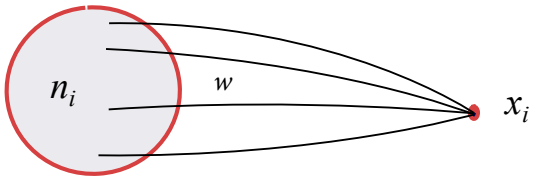


FIGURE 2.1: Le nombre d'individus qui prennent la valeur x_i.

Exemple 9

Dans l'exemple précédent, 66 *est le nombre de familles qui ont* 2 *enfants.*

x_i	$\cdots$	2	$\cdots$
n_i *(Effectif)*	$\cdots$	66	$\cdots$

2.1.2 Effectif cumulé

Définition 8
Pour chaque valeur x_i, on pose par définition

$$N_i = n_1 + n_2 + ... + n_i.$$

L'effectif cumulé N_i d'une valeur est la somme de l'effectif de cette valeur et de tous les effectifs des valeurs qui précèdent.

Exemple 10
Dans l'exemple précédent : 50 *est le nombre de familles qui ont un nombre d'enfant inférieur à* 1. *Nous le regardons dans le tableau suivant :*

x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
N_i	*18*	*50*	*116*	*157*	*189*	*198*	*200*

Interprétation : N_i est le nombre d'individus dont la valeur du caractère est inférieur ou égale à x_i. De ce fait, l'effectif total est donné par

$$N = \text{card}\{\Omega) = \sum_{i=1}^{n} n_i.$$

Dans notre exemple précédent, nous avons $N = 200$.

2.2 Fréquence partielle - Fréquence cumulée

Typiquement les effectifs n_i sont grands et il est intéressant de calculer des grandeurs permettant de résumer la série.

2.2.1 Fréquence partielle (fréquence relative)

Définition 9
Pour chaque valeur x_i, on pose par définition

$$f_i := \frac{n_i}{N}.$$

f_i s'appelle la fréquence partielle de x_i. La fréquence d'une valeur est le rapport de l'effectif de cette valeur par l'effectif total.

Remarque 3
On peut remplacer f_i par $f_i \times 100$ qui représente alors un pourcentage.

Interprétation : $f_i =$ est le pourcentage des ω tel que $X(\omega) = x_i$.

Exemple 11
Dans l'exemple précédent, $0,33 :=$ il y a 33% de familles dont le nombre d'enfants égale à 2. Ce pourcentage est calculé de la façon suivante ($N = 200$) :

x_i	$\cdots$	2	$\cdots$
n_i (Effectif)	$\cdots$	66	$\cdots$
N_i (Effectif)	$\cdots$	$\frac{66}{200} = 0.33$	$\cdots$

Nous pouvons conclure la propriété suivante.

Proposition 1
Soit f_i défini comme précédemment. Alors,

$$\sum_{i=1}^{n} f_i = 1.$$

Démonstration. Rappelons que

$$\sum_{i=1}^{n} n_i = N.$$

Ce qui implique que

$$\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^{n} n_i = 1.$$

2.2.2 Fréquence cumulée

Définition 10
Pour chaque valeur x_i, on pose par définition

$$F_i = f_1 + f_2 + ... + f_i.$$

La quantité F_i s'appelle la fréquence cumulée de x_i.

Interprétation : $F_i =$ est le pourcentage des ω tel que la valeur $X(\omega)$ est inférieure ou égale à x_i.

Exemple 12
- Dans l'exemple précédent, 0.785 *représente* 78.5% *de familles dont le nombre d'enfants est inférieur ou égale à* 3.
- Dans un deuxième exemple, nous nous intéressons aux nombres d'erreurs d'assemblage sur un ensemble d'appareils,

Nombre d'erreurs	*Nombre d'appareils*	*Fréquences cumulées*
0	*101*	*0.26*
1	*140*	*0.61*
2	*92*	*0.84*
3	*42*	*0.94*
4	*18*	*0.99*
5	*3*	*1*

Nous avons 94% *des appareils qui ont un nombre d'erreurs d'assemblage inférieur ou égale à* 3.

Nous avons vu que les tableaux sont un moyen souvent indispensable, en tous cas très utile, de classification et de présentation des unités d'une population statistique. Dans le paragraphe suivant, nous allons voir comment on traduit ses tableaux en graphique permettant aussi de résumer d'une manière visuelle les données.

2.3 Représentation graphique des séries statistiques

On distingue les méthodes de représentation d'une variable statistique en fonction de la nature de cette variable (qualitative ou quantitative). Les représentations recommandées et les plus fréquentes sont les tableaux et les diagrammes (graphe).

Le graphique est un support visuel qui permet :

La synthèse : visualiser d'un seul coup d'œil les principales caractéristiques (mais on perd une quantité d'informations), voir Figure 2.2.

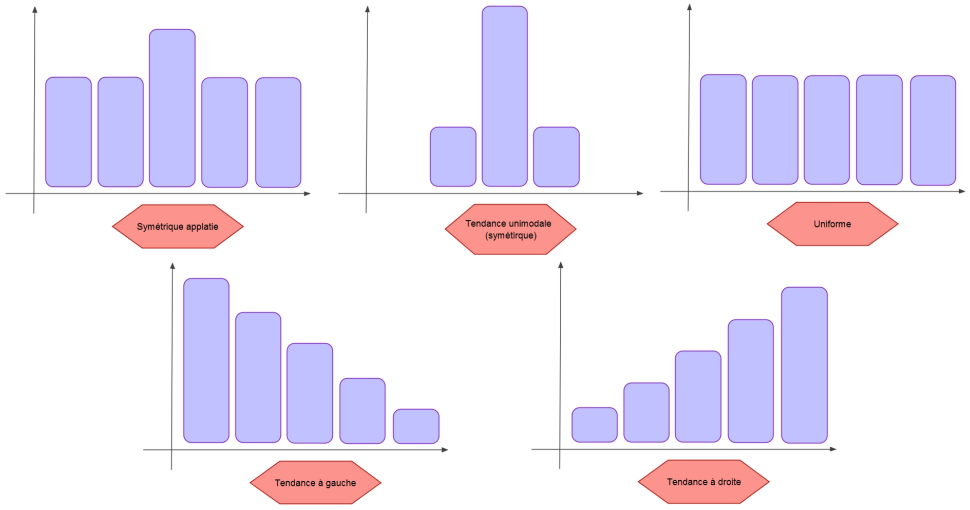


FIGURE 2.2: Quelques caractéristiques du graphique

La découverte : met en évidence les tendances.
Le contrôle : on aperçoit mieux les anomalies sur un graphique que dans un tableau.
La recherche des régularités : régularité dans le mouvement, répétition du phénomène.

2.3.1 Distribution à caractère qualitatif

A partir de l'observation d'une variable qualitative, deux diagrammes permettent de représenter cette variable : le diagramme en bandes (dit tuyaux d'orgue) et le diagramme à secteurs angulaires (dit camembert).

Tuyaux d'orgues

Nous portons en abscisses les modalités, de façon arbitraire. Nous portons en ordonnées des rectangles dont la longueur est proportionnelle aux effectifs, ou aux fréquences, de chaque modalité (voir Figure 2.3).

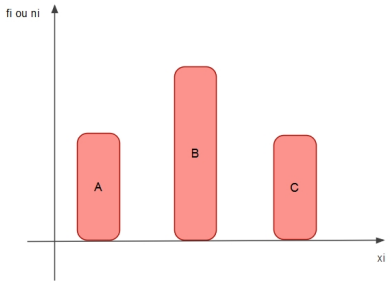


FIGURE 2.3: Tuyaux d'orgues

Diagramme par secteur (diagramme circulaire)

Les diagrammes circulaires, ou semi-circulaires, consistent à partager un disque ou un demi-disque, en tranches, ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence, de la modalité (voir Figure 2.4).

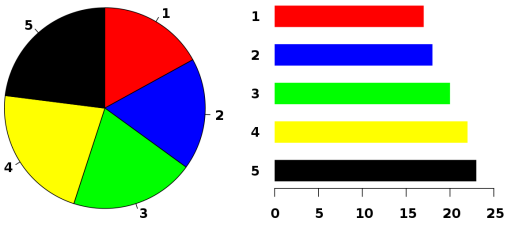


FIGURE 2.4: Diagramme par secteur

Le degré d'un secteur est déterminé à l'aide de la règle de trois de la manière suivante :

$$N \longrightarrow 360°$$

$$n_i \longrightarrow d_i \quad (\textit{degré de la modalité } i).$$

Donc,

$$d_i = \frac{n_i \times 360}{N}.$$

2.3.2 Distribution à caractère quantitatif discret

A partir de l'observation d'une variable quantitative discrète, deux diagrammes permettent de représenter cette variable : le diagramme en bâtons et le diagramme cumulatif (voir ci-dessous).

Pour l'illustration, nous prenons l'exemple précédent de départ (nombre d'enfants par famille). Nous rappelons le tableau statistique associe.

x_i	0	1	2	3	4	5	6
n_i	18	32	66	41	32	9	2

Diagramme à bâtons

On veut représenter cette répartition sous la forme d'un diagramme en bâtons. À chaque marque correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs représentés (voir Figure 2.5).

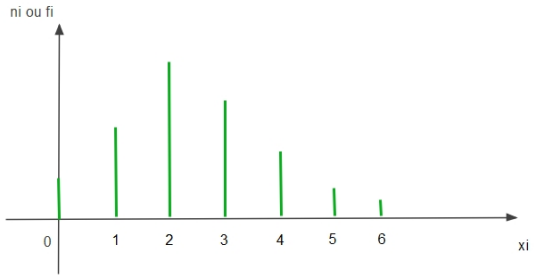


FIGURE 2.5: Diagramme à bâtons

2.3.3 Représentation sous forme de courbe et fonction de répartition

Nous avons déjà abordé les distributions cumulées d'une variable statistique. Nous allons dans cette partie exploiter ses valeurs cumulées pour introduire la notion de la fonction de répartition. Cette notion ne concerne que les variables quantitatives.

Soit la fonction $F_x : \mathbb{R} \to [0,1]$ définie par

$$F_x(x) := \text{pourcentage des individus dont la valeur du caractère est} \leq x.$$

Cette fonction s'appelle la fonction de répartition du caractère X.

Remarque 4
Pour tout $i \in \{1, \ldots, n\}$, *on a*

$$F_x(x_i) = F_i.$$

La courbe de F_x *passe par les points* (x_1, F_1), (x_2, F_2), ... *et* (x_n, F_n).

En se basant sur notre exemple, la courbe de F_x est représentée ci-dessous (Figure 2.6) sur

$$\mathbb{R} =]-\infty, 0[\cup [0,1[\cup .... \cup [6, +\infty[.$$

Dans ce cas, nous avons

- Si $x < 0$, alors $F_x(x) = 0$.
- Si $x \in [0,1[$, alors $F_x(x) = 0.09$.

  ...
- Si $x \geq 6$, alors $F_x(x) = 1$.

Cette courbe s'appelle "la courbe cumulative des fréquences". La courbe cumulative est une courbe en escalier représentant les fréquences cumulées relatives.

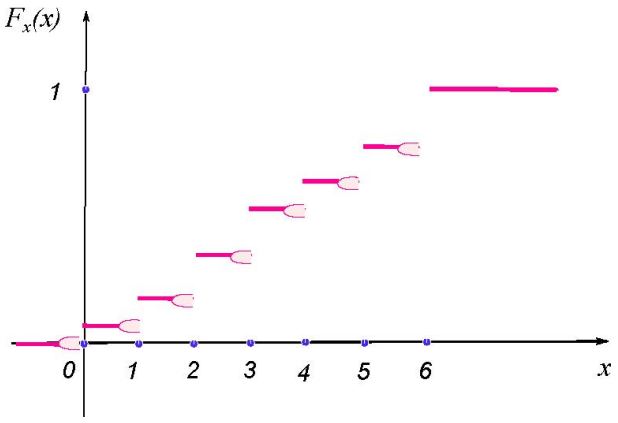


FIGURE 2.6: Représentation d'une variable quantitative discrète par la courbe cumulative.

Proposition 2

La fonction de répartition satisfait, pour $i \in \{1, \ldots, n\}$,

- *l'égalité,* $F_x(x_i) = F_i$,
- *l'expression,*

$$F_x(x) = \begin{cases} 0, & si \quad x < x_1, \\ F_1, & si \quad x_1 \leq x < x_2, \\ F_i, & si \quad x_i \leq x < x_{i+1}, \\ 1, & si \quad x \geq x_n. \end{cases}$$

2.4 Paramètres de position (caractéristique de tendance centrale)

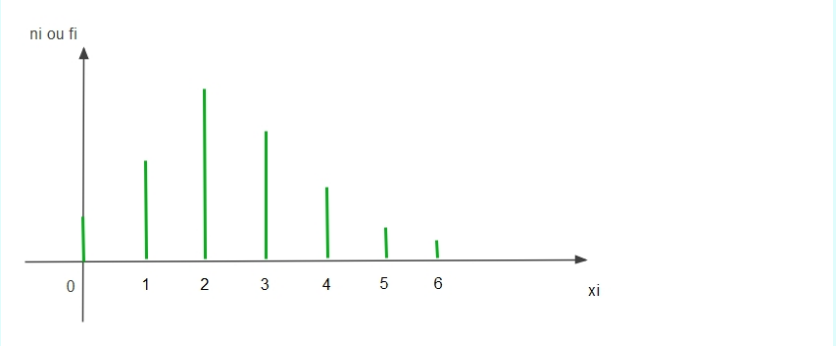
Les indicateurs statistiques de tendance centrale (dits aussi de position) considérés fréquemment sont la moyenne, la médiane et le mode.

Le mode

Le mode d'une V.S est la valeur qui a le plus grand effectif partiel (ou la plus grande fréquence partielle) et il est dénoté par M_0.

Exemple 13

Dans l'exemple précédent, le mode est égal à 2 qui correspondant au plus grand effectif.



Remarque 5
On peut avoir plus d'un mode ou rien.

La médiane

On appelle médiane la valeur Me de la V.S X qui vérifie la relation suivante :

$$F_x(Me^-) < 0.5 \leq F_x(Me^+) = F_x(Me).$$

La médiane partage la série statistique en deux groupes de même effectif.

Exemple 14
Dans l'exemple précédent, la relation

$$F_x(0) = 0 < 0.5 \leq F_x(0^+) = 0.09$$

n'est pas satisfaite. Donc, la médiane est différente de 0. *Par contre, nous avons*

$$F_x(2^-) = 0.25 < 0.5 \leq F_x(2^+) = F(2) = 0.58.$$

Donc, $Me = 2$.

La moyenne

On appelle moyenne de X, la quantité

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

avec $N = \text{Card}(\Omega)$. On peut donc exprimer et calculer la moyenne dite "arithmétique" avec des effectifs ou avec des fréquences.

Exemple 15
Si $\overline{x} = 2.46$, *alors nous avons au moyenne une famille de quartier a* 2.46 *d'enfants.*

La valeur de la moyenne est abstraite. Comme dans l'exemple précédent, $\overline{x} = 2.46$ est un chiffre qui ne correspond pas à un fait concret.

La moyenne arithmétique dont on vient d'indiquer la formule est dite moyenne pondérée ; cela signifie que chaque valeur de la variable est multipliée (pondérée) par un coefficient, ici par l'effectif n_i qui lui correspond. Dans ce cas, chaque valeur x_i de la variable intervient dans le calcul de la moyenne autant de fois qu'elle a été observée. On parle de moyenne arithmétique simple quand on n'effectue pas de pondération. Par exemple, si 5 étudiants ont pour âge respectif 18, 19, 20, 21 et 22 ans, leur âge moyen est donné par $(18 + 19 + 20 + 21 + 22)/5 = 20$ ans.

Remarque 6
Nous mentionnons qu'il existe d'autres moyennes que la moyenne arithmétique

2.5 Paramètres de dispersion (variabilité)

Les indicateurs statistiques de dispersion usuels sont l'étendue, la variance et l'écart-type.

L'étendue

La différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité

$$e = x_{\max} - x_{\min},$$

s'appelle l'étendue de la V.S X. Le calcul de l'étendue est très simple. Il donne une première idée de la dispersion des observations. C'est un indicateur très rudimentaire et il existe des indicateurs de dispersion plus élaborés (voir ci-dessous).

La variance

On appelle variance de cette série statistique X, le nombre

$$Var(X) = \sum_{i=1}^{n} f_i(\overline{x} - x_i)^2$$

On dit que la variance est la moyenne des carrés des écarts à la moyenne $\overline{x}$. Les « écarts à la moyenne » sont les $(\overline{x} - x_i)$, les « carrés des écarts à la moyenne » sont donc les $(\overline{x} - x_i)^2$. En faisant la moyenne de ces écarts, on trouve la variance.

Le théorème suivant (Théorème de König-Huygens) donne une identité remarquable reliant la variance et la moyenne, parfois plus pratique dans le calcule de la variance.

Théorème 1
Soit (x_i, n_i) une série statistique de moyenne $\overline{x}$ et de variance $Var(X)$. Alors,

$$Var(X) = \sum_{i=1}^{n} f_i x_i^2 - \overline{x}^2.$$

Démonstration. Par définition, nous avons

$$Var(X) = \sum_{i=1}^{n} f_i(\overline{x} - x_i)^2 = \frac{1}{N}\sum_{i=1}^{n} n_i(\overline{x} - x_i)^2 = \frac{\sum_{i=1}^{n} n_i(\overline{x} - x_i)^2}{\sum_{i=1}^{n} n_i}.$$

Donc,

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

Par égalité, nous avons

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

Ce qui implique que

$$Var(X) = \overline{x}^2 + \frac{\sum_{i=1}^{n} n_i x_i^2}{\sum_{i=1}^{n} n_i} - 2\overline{x}\overline{x} = -\overline{x}^2 + \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2.$$

Remarque 7
Dans l'utilisation de la formule du théorème précédent, il faut veiller à remplacer $\overline{x}$ par sa valeur approchée la plus précise possible.

L'écart type

La quantité

$$\sigma_X = \sqrt{Var(x)}$$

s'appelle l'écart type de la V.S X.

Remarque 8

Le paramètre σ_x mesure la distance moyenne entre $\overline{x}$ et les valeurs de X (voir Figure 2.7). Il sert à mesurer la dispersion d'une série statistique autour de sa moyenne.

- *Plus il est petit, plus les caractères sont concentrés autour de la moyenne (on dit que la série est homogène).*
- *Plus il est grand, plus les caractères sont dispersés autour de la moyenne (on dit que la série est hétérogène).*

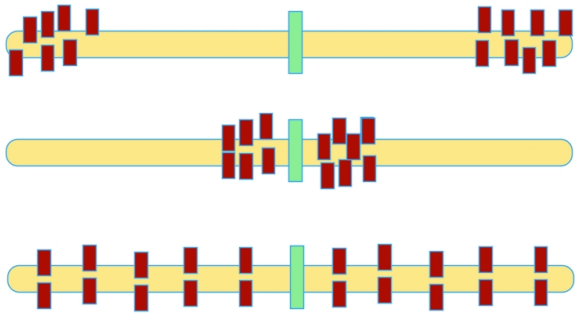


FIGURE 2.7: La dispersion d'une série statistique autour de sa moyenne