Statistiques descriptives

Symboles et Notations

Symbole	Signification
[ ]	La partie entière.
Card(Ω)	Le cardinal : nombre d'éléments de l'ensemble Ω.
:=	Est défini comme étant (symbole d'affectation).
N	Ensemble des nombres entiers naturels.
Z	Ensemble des nombres entiers relatifs.
R	Ensemble des nombres réels.
R^2	Ensemble des couples de nombres réels.
$\sum_{i=1}^{n}$	La somme pour i variant de 1 à n.
$V.S$	La variable statistique
Me	La médiane.
Me^+	Me par valeur supérieure.
Me^-	Me par valeur inférieure.
M_0	Le mode.
$\overline{x}$	La moyenne d'une série statistique X.
σ_X	L'écart-type de X.
Var(X)	La variance de X.
Cov(X,Y)	La covariance entre les variables X et Y.
ρ_{XY}	Le coefficient de corrélation entre les variables X et Y.
F_x	La fonction s'appelle la fonction de répartition du caractère X

Chapitre 1

Généralités sur la statistique

La statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques.

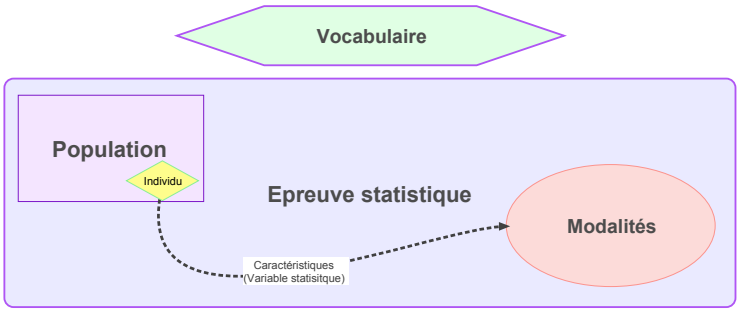
L'analyse des données est utilisée pour d'écrire les phénomènes étudiés, faire des prévisions et prendre des décisions à leur sujet. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes.

Les données étudiées peuvent être de toute nature, ce qui rend la statistique utile dans tous les champs disciplinaires et explique pourquoi elle est enseignée dans toutes les filières universitaires, de l'économie à la biologie en passant par la psychologie et bien sûr les sciences de l'ingénieur. La statistique consiste à :

- Recueillir des données.
- Présenter et résumer ces données.
- Tirer des conclusions sur la population étudiée et d'aider à la prise de décision.
- En présence de données dépendant du temps, nous essayons de faire de la prévision.

1.1 Vocabulaire

Les statistiques consistent en diverses méthodes de classement des données tels que les tableaux, les histogrammes et les graphiques, permettant d'organiser un grand nombre de données. Les statistiques se sont développées dans la deuxième moitié du XIXe siècle dans le domaine des sciences humaines (sociologie, économie, anthropologie, ...). Elles se sont dotées d'un vocabulaire particulier.



1.1.1 Épreuve statistique

Les statistiques descriptives visent à étudier les caractéristiques d'un ensemble d'observations comme les mesures obtenues lors d'une expérience. L'expérience est l'étape préliminaire à toute étude statistique. Il s'agit de prendre "contact" avec les observations. De manière générale, la méthode statistique est basée sur le concept suivant.

Définition 1
L'épreuve statistique est une expérience que l'on provoque.

Exemple 1 (La durée de vie des lampes)
Imaginons le cas suivant : un fabricant d'ampoules électriques ayant le choix entre 4 types de filaments se propose d'étudier l'influence de la nature du filament sur la durée de vie des ampoules fabriquées. Pour ce faire, il va faire fabriquer 4 échantillons d'ampoules identiques, sauf en ce qui concerne le filament, faire brûler les ampoules jusqu'à extinction, puis comparer les résultats obtenus.

1.1.2 Population

En statistique, on travaille sur des populations. Ce terme vient du fait que la démographie, étude des populations humaines, a occupé une place centrale aux débuts de la statistique, notamment au travers des recensements de population. Mais, en statistique, le terme de population s'applique à tout objet statistique étudié, qu'il s'agisse d'étudiants (d'une université ou d'un pays), de ménages ou de n'importe quel autre ensemble sur lequel on fait des observations statistiques. Nous définissons la notion de population.

Définition 2
On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté Ω.

Exemple 2
- *On considère l'ensemble des étudiants de la section A. On s'intéresse aux nombre de frères et sœurs de chaque étudiant. Dans ce cas*

$$\Omega = \text{ensemble des étudiants.}$$

- *Si l'on s'intéresse maintenant a la circulation automobile dans une ville, la population est alors constituée de l'ensemble des véhicules susceptibles de circuler dans cette ville à une date donnée. Dans ce cas*

$$\Omega = \text{ensemble des véhicules.}$$

1.1.3 Individu (unité statistique)

Une population est composée d'individus. Les individus qui composent une population statistique sont appelés unités statistiques.

Définition 3
On appelle individu tout élément de la population Ω, *il est noté* ω (ω *dans* Ω).

Remarque 1
L'ensemble Ω *peut être un ensemble de personnes, de choses ou d'animaux...*
L'unité statistique est un objet pour lequel nous sommes intéressés à recueillir de l'information.

Exemple 3
- *Dans l'exemple indiqué ci-dessus, un individu est tout étudiant de la section.*
- *Si on étudie la production annuelle d'une usine de boîtes de boisson en métal (canettes). La population est l'ensemble des boîtes produites durant l'année et*

une boîte constitue un individu.

1.1.4 Caractère (variable statistique)

La statistique « descriptive », comme son nom l'indique cherche à décrire une population donnée. Nous nous intéressons au caractéristique des unités qui peuvent prendre différentes valeurs.

Définition 4
On appelle caractère (ou variable statistique, dénotée V.S) toute application

$$X : \Omega \to C.$$

L'ensemble C est dit : ensemble des valeurs du caractère X (c'est ce qui est mesuré ou observé sur les individus)

Exemple 4
Taille, température, nationalité, couleur des yeux, catégorie socioprofessionnelle ...

Remarque 2
Soit Ω un ensemble. On appelle et on note $Card(\Omega)$, le nombre d'éléments de Ω.

$$Card(\Omega) := \text{nombre d'éléments de } \Omega = N.$$

1.1.5 Modalités

Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci.

Exemple 5
- *Variable est " situation familiale "*
  Modalités sont " célibataire, marié, divorcé "

- *Variable est" statut d'interrupteur "*
  Modalités sont " 0 et 1 ".
- *Variable est " catégories socio-professionnelles "*
  Modalités sont " Employés, ouvriers, retraités,... "

Les modalités sont les différentes situations dans lesquelles les individus peuvent se trouver à l'égard du caractère considéré.

1.2 Types des caractères

Nous distinguons deux catégories de caractères : les caractères qualitatifs et les caractères quantitatifs.

1.2.1 Caractère qualitatif

Les caractères qualitatifs sont ceux dont les modalités ne peuvent pas être ordonnées, c'est-à-dire que si l'on considère deux caractères pris au hasard, on ne peut pas dire de l'un des caractères qu'il est inférieur ou égal à l'autre. Plus précisément, nous avons la définition suivante.

Définition 5
Les éléments de C sont représentés par autre chose que des chiffres.

Exemple 6
L'état d'une maison : on peut considérer les modalités suivantes
- *Ancienne.*
- *Dégradée.*
- *Nouvelle.*

– *Rénovée.*

1.2.2 Caractère quantitatif

Les caractères quantitatifs sont des caractères dont les modalités peuvent être ordonnées. Ainsi, l'âge, la taille de vie ou le salaire d'un individu sont des caractères quantitatifs. Donc, nous avons la définition suivante.

Définition 6
L'ensemble des valeurs est représenté par des chiffres. De même, il est partagé en deux sortes de caractères, discret et continu (voir l'exemple).

Exemple 7

– *Le salaire d'employés d'une usine.*
Modalités : $10000da$, $20000da$...
Type : Discret.

– *La rigidité des ressorts.*
Modalités : $[10, 20]$ N/m
Type : continu.

En général, la variable quantitative discrète est une variable ne prenant que des valeurs entières (plus rarement décimales). Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible. Citons, par exemple, le nombre de maisons par quartier d'une ville. Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises, mais des intervalles. C'est le cas lorsque nous avons un grand nombre d'observations distinctes.

La statistique descriptive a pour objectif de synthétiser l'information contenue dans les jeux de données au moyen de tableaux, figures ou résumés numériques. Les variables statistiques sont analysées différemment selon leur nature (quantitative, qualitative).

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.

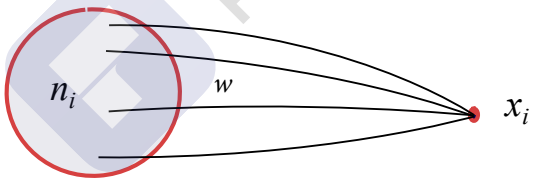


FIGURE 2.1: Le nombre d'individus qui prennent la valeur x_i.

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.

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).

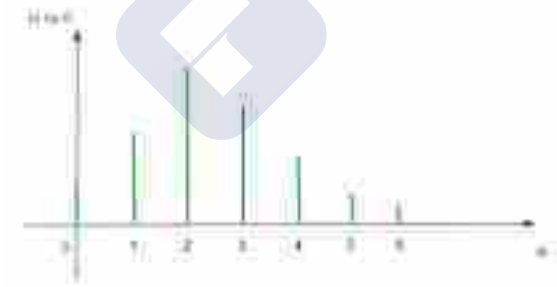


FIGURE 2.5: Diagramme à bâtons

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.

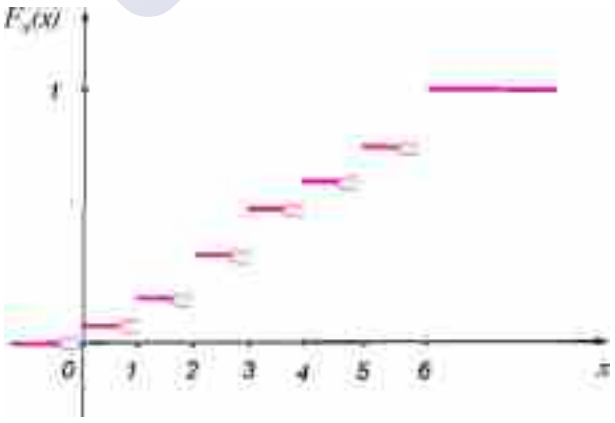


FIGURE 2.6: Représentation d'une variable quantitative discrète par la courbe cumulative.

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)

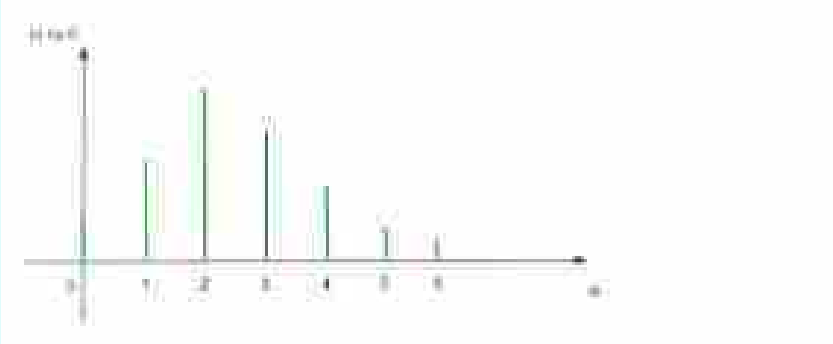
Les indicateurs statistiques de tendance centrale (dits aussi de position) considérés fréquemment sont la moyenne, la médiane et le mode.

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).*

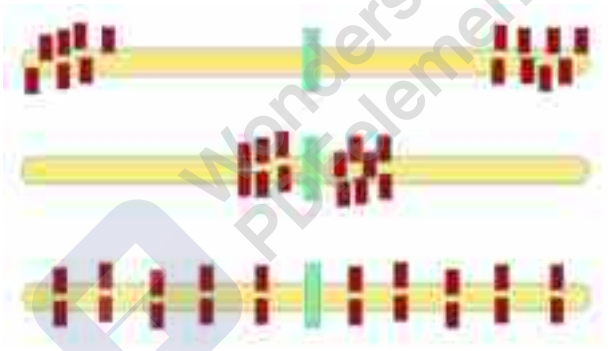


FIGURE 2.7: La dispersion d'une série statistique autour de sa moyenne

2.6 Exercices

Exercice 7

- Le tableau suivant donne la répartition selon le groupe sanguin de 40 individus pris au hasard dans une population,

Groupes sanguins	*A*	*B*	*AB*	*O*
L'effectif	*20*	*10*	n_3	*5*

1. *Déterminer la variable statistique et son type.*
2. *Déterminer l'effectif des personnes ayant un groupe sanguin AB.*

3. Donner toutes les représentations graphiques possibles de cette distribution.

Solution 1 - La population dans cette étude est les 40 personnes. Donc $N = 40$. La variable statistique est le groupe sanguin des individus et elle est qualitative.

2 - L'effectif total est égal à 40. Par conséquent,

$$N = 40 = \sum_{i=1}^{4} n_i.$$

Alors,

$$20 + 10 + n_3 + 5 = 40.$$

Ce qui implique que $n_3 = 5$.

3- Nous avons deux représentations possibles "Tyaux d'orgue" et "Diagramme en secteur".

Chapitre 3

Étude d'une variable statistique continue

Nous rappelons qu'une variable statistique (V.S) quantitative concerne une grandeur mesurable. Ses valeurs sont des nombres exprimant une quantité et sur lesquelles les opérations arithmétiques (addition, multiplication, etc,...) ont un sens. Nous allons dans ce chapitre se focaliser sur la V.S quantitative continue.

3.1 Caractère continu

Définition 11
On appelle V.S continue (ou caractère continu) toute application de Ω et à valeurs réelles et qui prend un nombre "important" de valeurs (Les caractères continus sont ceux qui ont une infinité de modalités).

Exemple 16
Soit Ω l'ensemble des nouveaux nés au C.H.U d'une ville pendant les 3 premiers mois de 2017. Nous désignons par X le poids des nouveaux nés. On suppose que

$$x_{min} = 2.701 \quad et \quad x_{max} = 5.001.$$

Remarque 9
Comment étudier ce caractère ?

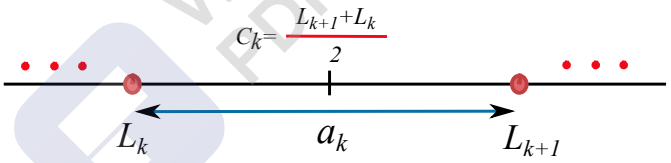
Réponse *: Partager les valeurs prises par X en classes de valeurs.*

3.1.1 Classe de valeurs

Définition 12
On appelle classe de valeurs de X un intervalle de type $[a, b[$ tel que $X \in [a, b[$ si et seulement si $a \leq X(w) < b$, c'est à dire, que les valeurs du caractère sont dans la classe $[a, b[$.

Dès qu'un caractère est identifié en tant que continu, ces modalités $C_k = [L_k, L_{k+1}[$ sont des intervalles avec

- L_k : borne inférieure.
- L_{k+1} : borne supérieure.
- $a_k = L_{k+1} - L_k$: son amplitude, son pas ou sa longueur.
- $C_k = x_k = (L_{k+1} + L_k)/2$: son centre.



Remarque 10
On supposera dans tous les cas étudiés que la distribution à l'intérieur des classes est uniforme (voir Figure 3.1). Cette hypothèse permet de justifier le fait qu'on choisisse le centre des classes comme représentant.

3.1.2 Nombre de classes

En combien de classes partageons-nous les valeurs ? la réponse n'est pas unique. Soit N l'effectif total. Nous pouvons considérer dans ce cours trois réponses à titre d'exemple.

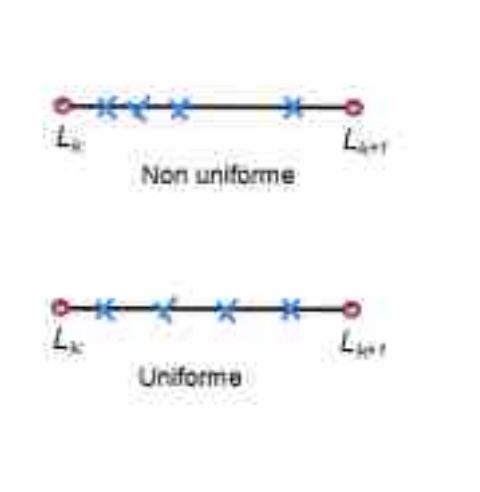


FIGURE 3.1: Une représentation de la distribution des valeurs à l'intérieur d'une classe.

1. Une réponse : $\sqrt{N}$, $[\sqrt{N}]$ (partie entière) ou $[\sqrt{N}]+1$. Donc, le nombre de classes

$$k \backsimeq \sqrt{N}.$$

Exemple 17
Considérons 30 valeurs entre 56.5 cm et 97.8 cm. Dans ce cas, $k = \sqrt{30}$ et on prend $k = 6$.

2. Une réponse : la formule de Sturge

$$k = 1 + 3.3 \log_{10}(N).$$

3. Une réponse : la formule de Yule

$$k = 2.5 \sqrt[4]{N}.$$

Remarque 11
De ce fait, on peut avoir plusieurs tableaux statistiques selon le nombre de classes.

Exemple 18
Si on prend $N = 30$, alors le nombre de classes est donné, par exemple, par

- *soit la formule de Sturge* $k = 1 + 3.3\log_{10}(30) \backsimeq 6$,

- *soit la formule de Yule* $k = 2.5\sqrt[4]{30} \backsimeq 6$.

Nous mentionnons que les deux formules sont presque pareils si $N \ll 200$.

Nous rappelons maintenant la définition de l'étendu. De plus, dans le cas continue nous parlons aussi du pas ou de la longueur de la classe.

Définition 13

Le nombre

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

s'appelle étendu de X. *Dans ce cas, on peut définir le pas par*

$$a_i := \frac{étendu}{nombre\ de\ classes} = \frac{x_{max} - x_{min}}{k}.$$

3.1.3 Effectif et fréquence d'une classe

Définition 14

La quantité

$$n_i := Card\{w \in \Omega : \quad X(w) \in C_i\}$$

s'appelle effectif partiel de C_i.

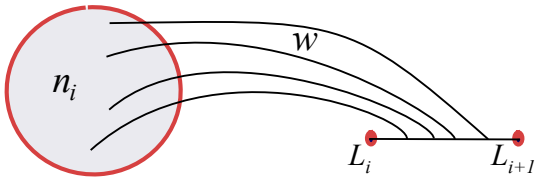


FIGURE 3.2: Le nombre d'individus qui prennent des valeurs x_i dans C_i.

Définition 15

Le nombre

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

est appelé la fréquence partielle de C_i.

Définition 16

On appelle l'effectif cumulé de C_i la quantité

$$N_i := \sum_{j=1}^{i} n_j.$$

Définition 17

On appelle la fréquence cumulée de C_i la quantité

$$F_i := \sum_{j=1}^{i} f_j.$$

Remarque 12

Nous avons, comme dans le chapitre précédent, les interprétations suivantes :

- n_i : *est le nombre d'individus dont les valeurs des caractères sont dans la classe* C_i,
- f_i : *est le pourcentage des* w *tel que* $X(w) \in C_i$,
- N_i : *est égale au* $Card\{w : \quad X(w) \in C_1 \cup C_2 \cup ... \cup C_i\}$,
- F_i : *est le pourcentage des* w *tel que*

$$X(w) \in C_1 \cup ... \cup C_i.$$

3.2 Représentation graphique d'un caractère continu

3.2.1 Histogramme des fréquences (ou effectifs)

Nous pouvons représenter le tableau statistique par un histogramme. Nous reportons les classes sur l'axe des abscisses et, au-dessus de chacune d'elles, nous traçons un rectangle dont l'aire est proportionnelle à la fréquence f_i (ou l'effectif n_i) associée. Ce graphique est appelé l'histogramme des fréquences (voir Figure 3.3).

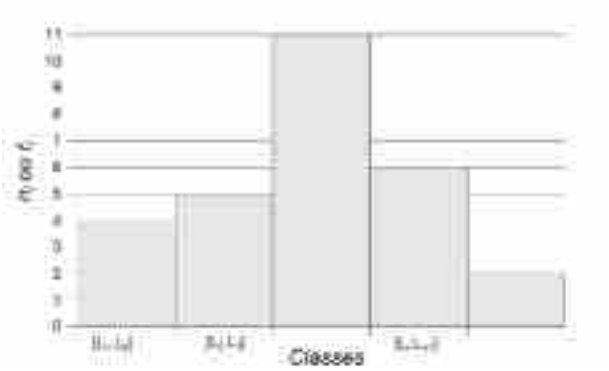


FIGURE 3.3: Histogramme des fréquences ou des éffctifs.

3.2.2 Fonction de répartition

Notation : Nous allons noter par

$$C_i = [x_{\min} = a_0, x_{\min+1} = a_1[.$$

Définition 18

La fonction $F_x : \mathbb{R} \to [0,1]$ définie par $F_x(x)$ représente le pourcentage des individus tel que la valeur de leur caractère est inférieure ou égale à x. Elle est donnée par

$$F_x(x) = \begin{cases} 0, & si \quad x < a_0, \\ \dfrac{f_1}{h}(x - a_0), & si \quad a_0 \leq x < a_1, \\ F_i + \dfrac{f_{i+1}}{h}(x - a_i), & si \quad a_i \leq x < a_{i+1}, \\ 1, & si \quad x \geq a_n, \end{cases}$$

et elle s'appelle la fonction de répartition de X.

Nous expliquons cette formulation de la fonction de répartition dans cette remarque.

Remarque 13

Nous calculons $F_x(x)$ par extrapolation (voir Figure 3.4). Nous avons déjà $F(L_i) = F_i$. De plus,

$$\tan(\alpha) = \frac{F(L_{i+1}) - F(L_i)}{L_{i+1} - L_i} = \frac{F(x) - F(L_i)}{x - L_i}.$$

Ce qui implique la formule de la fonction de répartition

$$F(x) = \frac{f_{i+1}}{h}(x - L_i) + F_i.$$

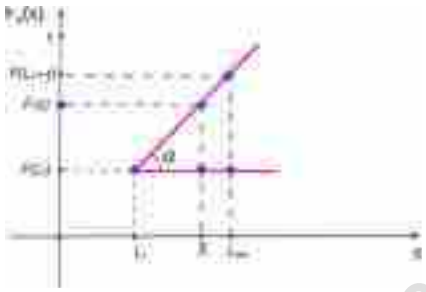


FIGURE 3.4: Le calcul de $F_x(x)$ par extrapolation.

La courbe de F_x est nulle avant a_0, constante égale à 1 après a_n et joint les points $(a_0, 0)$, (a_1, F_1),...,$(a_n, 1)$ par des segments de droites (voir Figure 3.5).

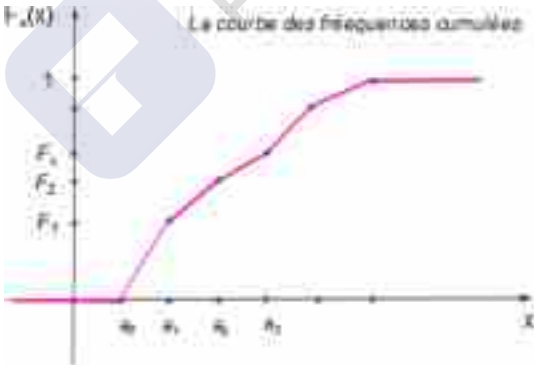


FIGURE 3.5: La courbe des fréquences cumulées.

3.3 Paramètres de tendance central

On note par C_i le centre de la classe C_i et nous considérons f_i la fréquence partielle de C_i.

FIGURE 3.6: Le centre de la classe.

La moyenne

Définition 19
La quantité

$$\overline{x} = \sum_{i=1}^{n} f_i C_i$$

s'appelle la moyenne de X.

Le mode

La définition suivante permet de comprendre la démarche à suivre pour calculer le mode d'une manière exacte et qui se trouve dans une des classes appelée "classe modale".

Définition 20
Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif partiel (ou la plus grande fréquence partielle). La quantité

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} a_i$$

s'appelle le mode avec (voir Figure 3.7)

- *L_i : la borne inférieure de la classe modale.*
- *a_i : le pas de la classe modale.*
- *$\Delta_1 = n_0 - n_1$, $\Delta_2 = n_0 - n_2$ ou bien $\Delta_1 = f_0 - f_1$, $\Delta_2 = f_0 - f_2$.*
- *n_0 et f_0 sont l'effectif et la fréquence associés à la classe modale.*
- *n_1 et f_1 sont l'effectif et la fréquence de la classe qui précède la classe modale.*
- *n_2 et f_2 sont l'effectif et la fréquence de la classe qui suit la classe modale.*

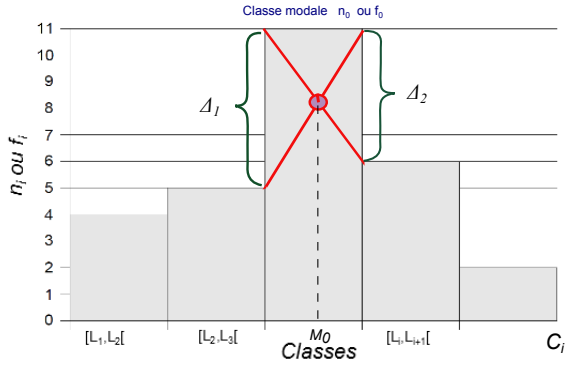


FIGURE 3.7: Représentation ou détermination graphique du mode (cas continu).

Remarque 14
L'expression du mode donnée ci-dessus est déterminée à partir de l'intersection des deux segments représentés dans la Figure 3.7. Cette notion n'est pas unique.

La médiane

Définition 21
C'est la valeur Me telle que $F(Me) = 0.5$. Cette valeur est unique.

Noue pouvons la déterminer graphiquement ou par calcule.

1. **Première méthode** : Graphiquement à partir de la formule

$$\tan(\alpha) = \frac{F(L_{i+1}) - F(L_i)}{L_{i+1} - L_i} = \frac{0.5 - F(L_i)}{Me - L_i}.$$

   Plus précisément, dans la figure 3.8, nous mettons $F(x) = 0.5$ et $x = Me$.

2. **Deuxième méthode** : En utilisant directement la fonction de répartition donnée par

$$F(x) = \frac{f_{i+1}}{h}(x - L_i) + F_i.$$

   Nous retrouvons donc

$$0.5 = \frac{f_{i+1}}{h}(Me - L_i) + F_i.$$

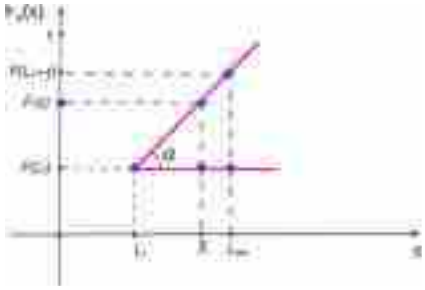


FIGURE 3.8: Le calcul de la médiane par extrapolation.

3.4 Paramètres de dispersion

Définition 22

La variance est la quantité

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

Remarque 15

Pour le calcul, on utilise (voir Chapitre 2, Théorème 1)

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

Définition 23

La quantité

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

s'appelle l'écart type de la V.S X.

Nous généralisons la notion de la médiane dans la définition suivante.

Définition 24

Pour $i \in \{1, 2, 3\}$, la quantité Q_i tel que $F(Q_i) = \frac{i}{4}$ s'appelle le i^{em} quartile.

observe les notes de math3 X et de statistique Y.

– *Une entreprise mène une étude sur la liaison entre les dépenses mensuelles en publicité X et le volume des ventes Y qu'elle réalise.*

4.1 Représentation des séries statistiques à deux variables

Les séries statistiques à deux variables peuvent être présentées de deux façons.

Présentation 1

A chaque w_i, on associé (x_i, y_i), c'est à dire,

$$w_i \longrightarrow (x_i, y_i).$$

On rassemblera les données comme dans le tableau suivant

w_i	w_1	w_2	...	w_N
Variable X	$X(w_1)$	$X(w_2)$	...	$X(w_N)$
Variable Y	$Y(w_1)$	$Y(w_2)$	...	$Y(w_N)$

Cette représentation on la notera "présentation 1". Nous allons utiliser toujours les notations suivantes :

$$x_i := X(w_i)$$

et $y_i := Y(w_i)$.

Exemple 21

Soit Ω l'ensemble de 8 étudiants. Nous avons le tableau suivant

w_i	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
$X(w)$	8	2	6	6	11	10	7	2
$Y(w)$	9	10	11	7	14	16	12	5

avec X représente le nombre d'heures passées à préparer l'examen de statistique par étudiant et Y représente la note sur 20 obtenue à l'examen par l'étudiant.

Lors de cette représentation, nous pouvons traduire le tableau associe dans une figure appelée "le nuage de points" ou "diagramme de dispersion" (voir Figure 4.1). Cette représentation est obtenue en mettant dans un repère cartésien chaque couple d'observation (x_i, y_j) par un point.

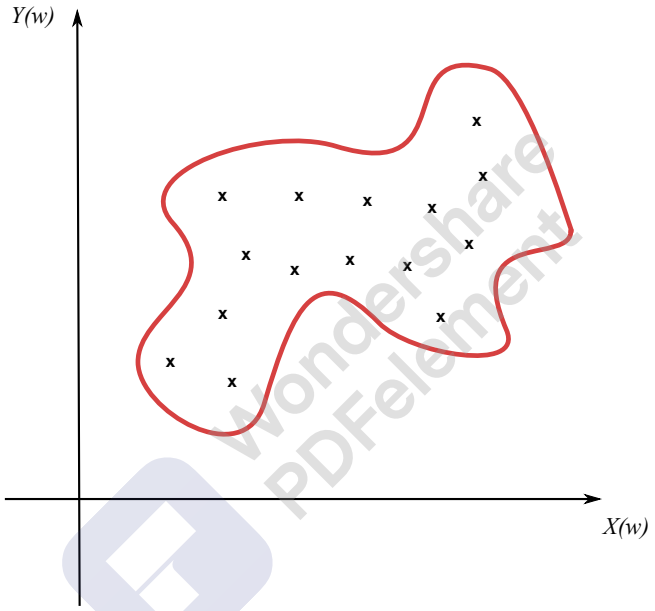


FIGURE 4.1: Représentation sous forme de nuage de points.

Présentation 2

Soit la variable statistique Z donnée par le couple (X, Y). Soient x_1, ..., x_k et y_1,...,y_l les valeurs prises respectivement par X et Y. Dans ce cas, nous définissons les valeurs de Z comme suite, pour i allant de 1 à k et pour j allant de 1 à l,

$$z_{ij} := (x_i, y_j).$$

La variable statistique Z prend $k \times l$ valeurs. Lors de cette étude, nous avons le tableau à double entrée (ou tableau de contingence) suivant (discrète ou continue)

X \ **Y**	$C'_1 = [L'_1, L'_2[$ **ou** y_1	...	$C'_l = [L'_l, L'_{l+1}[$ **ou** y_l	**Marginale** % à **X**
$C_1 = [L_1, L_2[$ **ou** x_1	n_{11} **ou** f_{11}	...	n_{1l} **ou** f_{1l}	$n_{1\bullet}$ **ou** $f_{1\bullet}$
$C_2 = [L_2, L_3[$ **ou** x_2	n_{21} **ou** f_{21}	...	n_{2l} **ou** f_{2l}	$n_{2\bullet}$ **ou** $f_{2\bullet}$
$C_3 = [L_3, L_4[$ **ou** x_3	n_{31} **ou** f_{31}	...	n_{3l} **ou** f_{3l}	$n_{3\bullet}$ **ou** $f_{3\bullet}$
$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$
$C_k = [L_k, L_{k+1}[$ **ou** x_k	n_{k1} **ou** f_{k1}	...	n_{kl} **ou** f_{kl}	$n_{k\bullet}$ **ou** $f_{k\bullet}$
Marginale % à **Y**	$n_{\bullet 1}$ ou $f_{\bullet 1}$	...	$n_{\bullet l}$ ou $f_{\bullet l}$	N

Cette représentation on l'a notera "présentation 2". A chaque couple (x_i, y_i), on a n_{ij} est l'effectif qui représente le nombre d'individus qui prennent en même temps la valeur x_i et y_i, c'est à dire,

$$n_{ij} := \text{Card}\{w \in \Omega \ : Z(w) = z_{ij}\}.$$

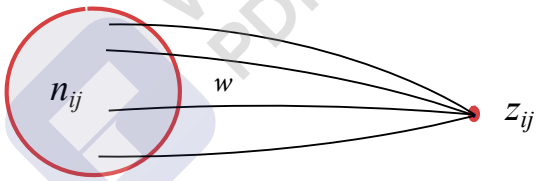


FIGURE 4.2: Le nombre d'individus qui prennent en même temps la valeur x_i et y_i.

Nous notons par f_{ij} la fréquence du coulpe (x_i, y_i). Cette fréquence est donnée par

$$f_{ij} := \frac{n_{ij}}{N},$$

avec

$$\begin{aligned} N &= \text{Card}(\Omega), \\ &= \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij}, \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}. \end{aligned}$$

Le calcul ou le développement de cette double série est donné par

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} &= n_{11}+n_{12}+n_{13}+...+n_{1l}\\ &\quad +n_{21}+n_{22}+n_{23}+...+n_{2l}\\ &\quad \ddots \quad \ddots \quad \ddots \quad \ddots\\ &\quad +n_{k1}+n_{k2}+n_{k3}+...+n_{kl}.\end{aligned}$$

Remarque 16

Nous avons la propriété suivante,

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1.$$

Lois marginales

Sur la marge du tableau de contingence, on peut extraire les données seulement par rapport à X et seulement par rapport à Y (voir le tableau de contingence établi auparavant).

1. Effectifs et fréquences marginale par rapport à Y : nous avons, pour $j = 1...l$,

$$n_{\bullet j} := \sum_{i=1}^{k} n_{ij},$$

et

$$f_{\bullet j} := \frac{n_{\bullet j}}{N} = \sum_{i=1}^{k} f_{ij}.$$

2. Effectifs et fréquences marginale par rapport à X : nous avons, pour $i = 1...k$,

$$n_{i\bullet} := \sum_{j=1}^{l} n_{ij},$$

et

$$f_{i\bullet} := \frac{n_{i\bullet}}{N} = \sum_{j=1}^{l} f_{ij}.$$

Remarque 17

Nous avons les propriétés suivantes

$$\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = N \qquad et \qquad \sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{l} f_{\bullet j} = 1.$$

Exercice 23

Nous considérons 10 *salariés qui sont observés à l'aide de deux variables "âge" et "salaire". Les informations brutes (pas encore traitées ou façonnées) sont données dans le tableau suivant,*

Salaire	6000	7400	7500	8200	8207	8900	9100	9900	9950	10750
Age	*15*	*26*	*20*	*43*	*47*	*37*	*52*	*34*	*50*	*44*

1. *Déterminer le tableau de contingence (X : âge, Y : salaire). Pour l'âge et pour le salaire, former respectivement des classes de pas de* 10 *ans et de* 1000 *Da.*
2. *Calculer* f_{21}, f_{12}, f_{45} *et* f_{33}.
3. *Déterminer les effectifs marginaux de* X *et de* Y. *Tracer le nuages de points.*
4. *Déterminer le tableau statistique des deux séries marginales* X *et* Y.

Solution : En utilisant les hypothèses, nous considérons les classes suivantes,

$$[15, 25[,\ [25, 35[,\ [35, 45[,\ [45, 55[,$$

pour l'âge et

$$[6, 7[,\ [7, 8[,\ [8, 9[,\ [9, 10[,\ [10, 11[,$$

pour le salaire ($\times 1000$). De plus, nous avons

$$Nombre\ de\ classe = \frac{e}{a_{âge}} = \frac{x_{\max} - x_{\min}}{a_{âge}} = \frac{52 - 15}{10} = 3.7 \simeq 4\ classes,$$

pour l'âge et

$$Nombre\ de\ classe = \frac{e}{a_{sal}} = \frac{y_{\max} - y_{\min}}{a_{sal}} = \frac{10750 - 6000}{1000} = 4.75 \simeq 5\ classes,$$

pour le salaire. Cette série statistique est représentée par le tableau suivant,

Age \ Salaire	$[6,7[$	$[7,8[$	$[8,9[$	$[9,10[$	$[10,11[$	$n_{i\bullet}$	$f_{i\bullet}$
$[15,25[$	1	1	0	0	0	0	0.2
$[25,35[$	0	1	0	1	0	2	0.2
$[35,45[$	0	0	2	0	1	3	0.3
$[45,55[$	0	0	1	2	0	3	0.3
$n_{\bullet j}$	1	2	3	3	1	10	1
$f_{\bullet j}$	0.1	0.2	0.3	0.3	0.1	1	$\varnothing$

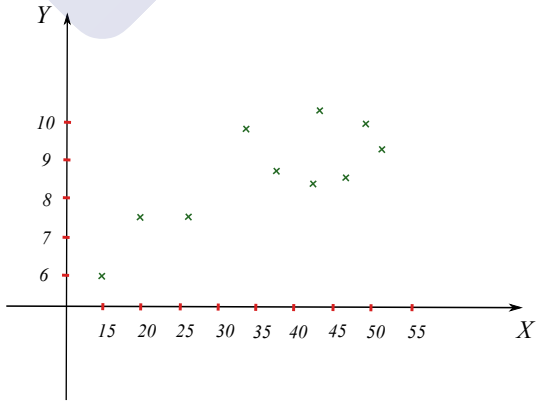
De ce fait, nous avons

$$f_{12} = \frac{n_{12}}{N} = \frac{1}{10} = 0.1, \qquad f_{21} = \frac{n_{21}}{N} = \frac{0}{10} = 0,$$

et

$$f_{45} = \frac{n_{45}}{N} = \frac{0}{10} = 0, \qquad f_{33} = \frac{n_{33}}{N} = \frac{2}{10} = 0.2.$$

Le nuage de points est tracé, à partir des données brutes, dans la figure suivante.



Enfin, les deux tableaux statistiques de X et de Y sont donnés, respectivement, par

X	$n_{i\bullet}$	$f_{i\bullet}$	x_i le centre
$[15, 25[$	2	0.2	20
$[25, 35[$	2	0.2	30
$[35, 45[$	3	0.3	40
$[45, 55[$	3	0.3	50

Y	$n_{i\bullet}$	$f_{i\bullet}$	y_i le centre
$[6, 7[$	1	0.1	6.5
$[7, 8[$	2	0.2	7.5
$[8, 9[$	3	0.3	8.5
$[9, 10[$	3	0.3	9.5
$[10, 11[$	1	0.1	10.5

4.2 Description numérique

4.2.1 Caractéristique des séries marginales

Dans le cas d'une variable statistique à deux dimensions X et Y, les moyennes sont données respectivement par

$$\overline{x} := \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i = \sum_{i=1}^{k} f_{i\bullet}x_i \quad \text{(moyenne } de\ X\text{)},$$

et

$$\overline{y} := \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j = \sum_{j=1}^{l} f_{\bullet j}y_j \quad \text{(moyenne } de\ Y\text{)}.$$

Remarque 18

Dans le cas continu, x_i et y_j représentent respectivement le centre des classes de X et Y, c'est à dire,

$$x_i = \frac{L_{i+1} + L_i}{2} \qquad et \qquad y_j = \frac{L_{j+1} + L_j}{2}.$$

Exemple 22

Nous calculons $\overline{x}$ et $\overline{y}$ pour l'exercice traité précédemment. Nous avons la moyenne d'âge

$$\overline{x} = \frac{1}{10}(40 + 60 + 120 + 150) = 37 \ ans.$$

et la moyenne du salaire

$$\overline{y} = \frac{1}{10}(6.5 + 15 + 25.5 + 28.5 + 10.5) \times 100 = 8600 \ Da.$$

Nous définissions maintenant la variance de X et la variance de Y comme suit,

$$Var(X) := \overline{x^2} - (\overline{x})^2, \qquad \text{avec} \quad \overline{x^2} := \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i^2 = \sum_{i=1}^{k} f_{i\bullet}x_i^2,$$

et

$$Var(Y) := \overline{y^2} - (\overline{y})^2, \qquad \text{avec} \quad \overline{y^2} := \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j^2 = \sum_{j=1}^{l} f_{\bullet j}y_j^2.$$

Les écarts-type de X et de Y sont donnés, respectivement, par

$$\sigma_X := \sqrt{Var(X)} \qquad \text{et} \qquad \sigma_Y := \sqrt{Var(Y)}.$$

4.2.2 Série conditionnelle

La notion de série conditionnelle est essentielle pour comprendre l'analyse de la régression. Un tableau de contingence se compose en autant de séries conditionnelles suivant chaque ligne et chaque colonnes.

Série conditionnelle par rapport à X

Elle est notée par X/y_j (ou X_j) et on dit que c'est la série conditionnelle de X sachant que $Y = y_j$. Nous calculons dans ce cas la fréquence conditionnelle $f_{i/j}$ (f_i sachant j), pour $i = 1, ..., k$, par

$$f_{i/j} := \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}.$$

Nous avons aussi la moyenne conditionnelle $\overline{x}_j$, c'est à dire la moyenne des valeurs de X sous la condition y_j, elle est définie par

$$\overline{x}_j := \sum_{i=1}^{k} f_{i/j}x_i = \frac{1}{n_{\bullet j}}\sum_{i=1}^{k} n_{ij}x_i.$$

Pour l'écart-type conditionnel, nous avons $\sigma_{X_j} := \sqrt{Var(X_j)}$ avec

$$Var(X_j) := \sum_{i=1}^{k} f_{i/j}(x_i - \overline{x}_j)^2 = \overline{x^2}_j - (\overline{x}_j)^2.$$

Série conditionnelle par rapport à Y

Elle est notée par Y/x_j (ou Y_j) et on dit que c'est la série conditionnelle de Y sachant que $X = x_i$. Nous calculons aussi dans ce cas la fréquence conditionnelle $f_{j/i}$ (f_j sachant i), pour $j = 1, ..., l$, par

$$f_{j/i} := \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}.$$

Nous avons aussi la moyenne conditionnelle $\overline{y}_i$, c'est à dire la moyenne des valeurs de Y sous la condition x_i, elle est définie par

$$\overline{y}_i := \sum_{j=1}^{l} f_{j/i} y_j = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} n_{ij} y_j \ .$$

Pour l'écart-type conditionnel, nous avons $\sigma_{Y_i} := \sqrt{Var(Y_i)}$ avec

$$Var(Y_i) := \sum_{j=1}^{l} f_{j/i} (y_j - \overline{y}_i)^2 = \overline{y^2}_i - (\overline{y}_i)^2.$$

4.2.3 Notion de covariance

Nous notons par $Cov(X, Y)$ la covariance entre les variables X et Y. La covariance est un paramètre qui donne la variabilité de X par rapport à Y (voir Figure 4.3).

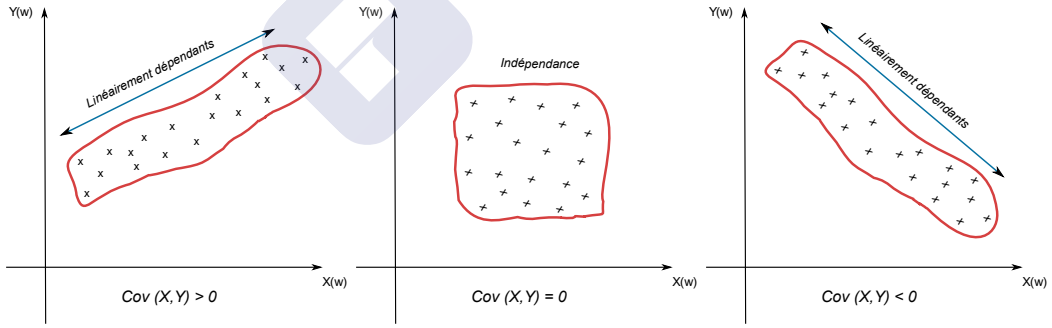


FIGURE 4.3: La covariance et la variabilité.

La covariance se calcule par l'expression suivante

$$Cov(X, Y) = \overline{xy} - \overline{x}\ \overline{y} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} x_i y_j - \overline{x}\ \overline{y}.$$

Nous avons aussi cette formule

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(x_i - \overline{x})(y_j - \overline{y}).$$

Remarque 19

Dans le cas où nous avons un tableau des données brutes "representation 1" (nous n'avons pas d'effectifs), nous avons les formules suivantes

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \qquad et \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{n} y_i.$$

De plus, nous avons

$$\overline{xy} = \frac{1}{N}\sum_{i=1}^{n} x_i y_i.$$

Remarque 20

La covariance est une notion qui généralise la variance, En effet,

$$Cov(X,X) = Var(X) \qquad et \qquad Cov(Y,Y) = Var(Y).$$

Cela provient de la définition, c'est à dire,

$$Cov(X,X) = \overline{xx} - \overline{x}\ \overline{x} = \overline{x^2} - \overline{x}^2 = Var(X).$$

Définition 25

On dit que deux variables statistiques X et Y sont indépendantes si et seulement si, pour tout i et j,

$$f_{ij} = f_{i\bullet} \times f_{\bullet j}.$$

Il suffit que cette égalité ne soit pas vérifiée dans une seule cellule pour que les deux variables ne soient pas indépendantes.. De manière équivalente, pour tout i et j,

$$N \times n_{ij} = n_{i\bullet} \times n_{\bullet j}.$$

Dans ce cas, si X et Y sont indépendantes alors (réciproque est fausse) $Cov(X,Y) = 0$.

Cette définition donne une interprétation intéressante de d'indépendance ; elle signifie que dans ce cas, les effectifs des modalités conjointes peuvent se calculer uniquement à partir des distributions marginales, supposées « identiques » aux distributions de X et Y dans la population ; en d'autres termes, si X et Y sont indépendantes, les observations séparées de X et de Y donnent la même information qu'une observation conjointe.

4.3 Ajustement linéaire

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y (la silhouette du nuage de points est étirée dans une direction), on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une équation de droite qui résume cette relation. Nous appelons cette démarche l'ajustement linéaire.

4.3.1 Coefficient de corrélation

Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires (voir ci-dessous). Il existe d'autres coefficients pour les relations non-linéaires et non-monotones, mais ils ne seront pas étudiés dans le cadre de ce cours.

Définition 26
La quantité

$$\rho_{XY} := \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

s'appelle le coefficient de corrélation.

Proposition 3
Le coefficient ρ_{XY} est compris entre $[-1,1]$, ou encore

$$|\,\rho_{XY}\,| \leq 1.$$

Le coefficient ρ_{XY} mesure le degré de liaison linéaire entre X et Y (voir Figure 4.4 et). Nous avons les deux caractéristiques suivantes (voir Figures 4.5 et 4.6) [1] :

1. Source : https ://en.wikipedia.org/wiki/Pearson_correlation_coefficient

– Plus le module de ρ_{XY} est proche de 1 plus X et Y sont liées linéairement.

– Plus le module de ρ_{XY} est proche de 0 plus il y a l'absence de liaison linéaire entre X et Y.

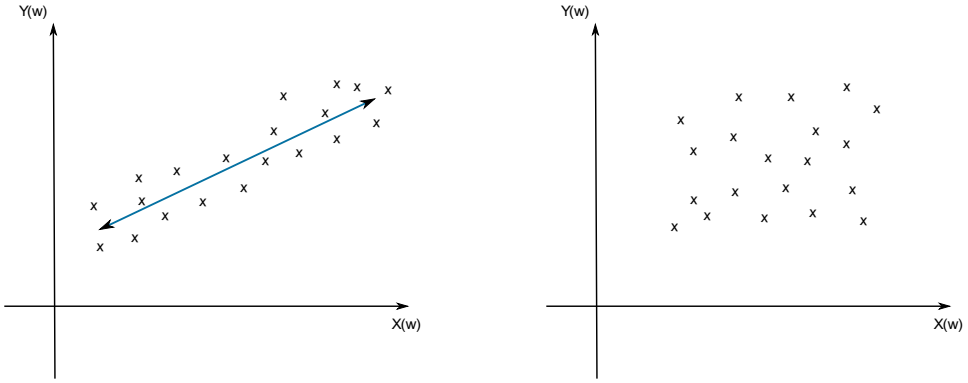


FIGURE 4.4: A gauche, le coefficient de corrélation est proche de 1. A droite, le coefficient de corrélation est proche de 0.

FIGURE 4.5: Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation .

Remarque 21
Par définition, si $\rho_{XY} = 0$, *alors* $Cov(X, Y) = 0$.

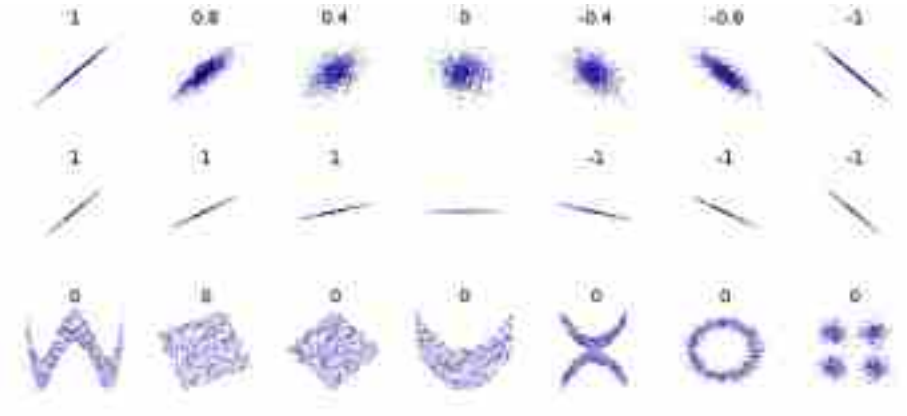


FIGURE 4.6: La corrélation reflète la non-linéarité et la direction d'une relation linéaire mais pas la pente de cette relation ni de nombreux aspects des relations non linéaires (en bas). La figure au centre a une pente de 0, mais dans ce cas, le coefficient de corrélation est indéfini car la variance de Y est nulle. .

4.3.2 Droite de régression

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite.

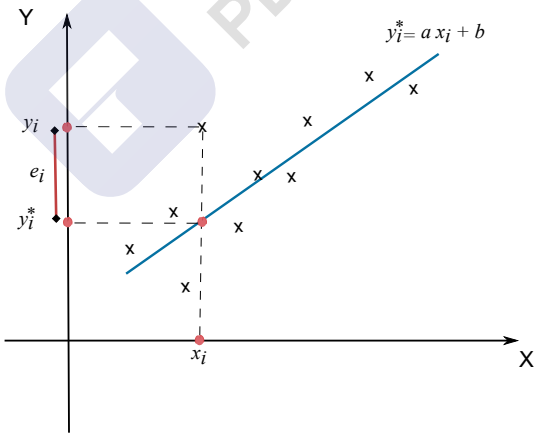


FIGURE 4.7: La droite la plus proche possible de chacun des points.

Pour cela, on utilise la méthode des moindres carrées. Cette méthode vise à expliquer un nuage de points par une droite qui lie Y à X, c'est à dire,

$$Y = aX + b,$$

telle que la distance entre le nuage de points et droite soit minimale. Cette distance matéria-

lise l'erreur, c'est à dire la différence entre le point réellement observé et le point prédit par la droite. Si la droite passe au milieu des points, cette erreur sera alternativement positive et négative, la somme des erreurs étant par définition nulle. Ainsi, la méthode des moindres carrés consiste à chercher la valeur des paramètres a et b qui minimise la somme des erreurs élevées au carré.

On pose

$$\sum_{i=1}^{n} e_i^2 = U(a,b),$$

avec e_i est l'erreur commise sur chaque observation, c'est à dire,

$$\mid e_i \mid = \mid y_i - y_i^\star \mid = \mid y_i - ax_i - b \mid .$$

La méthode des moindres carrées consiste donc à minimiser la fonction U (la somme des erreurs commises). Nous avons la condition de minimisation suivante,

$$\frac{\partial U}{\partial a} = \frac{\partial U}{\partial b} = 0,$$

avec

$$U(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

L'équation $\dfrac{\partial U}{\partial b} = 0$ donne

$$\sum_{i=1}^{n} -2(y_i - ax_i - b) = 0.$$

Ce qui implique que

$$\left(\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0 \right) \times \frac{1}{N}.$$

Par conséquent, nous obtenons

$$\overline{y} - a\overline{x} - b = 0,$$

c'est à dire,

$$\boxed{b = \overline{y} - a\overline{x}.}$$

De même, après calcule, $\dfrac{\partial U}{\partial a} = 0$ implique que

$$\boxed{a = \frac{Cov(X,Y)}{Var(X)}.}$$

Donc, la droite de régression, qui rend la distance entre elle et les points minimale, est

$$D(Y/X) : \ Y = aX + b,$$

$$a = \frac{Cov(X,Y)}{Var(X)} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}.$$

$$D(X/Y) : \ X = a'Y + b',$$

avec

$$a' = \frac{Cov(X,Y)}{Var(Y)} \qquad \text{et} \qquad b' = \overline{x} - a'\overline{y}.$$

Remarque 22

Le coefficient de corrélation ρ_{XY} permet de justifier le fait de l'ajustement linéaire. On adopte les critères numériques suivants (voir Figure 4.8),

- *Si $\mid \rho_{XY} \mid < 0.7$, alors l'ajustement linéaire est refusé (droite refusée).*
- *Si $\mid \rho_{XY} \mid \geq 0.7$, alors l'ajustement linéaire est accepté (droite acceptée).*

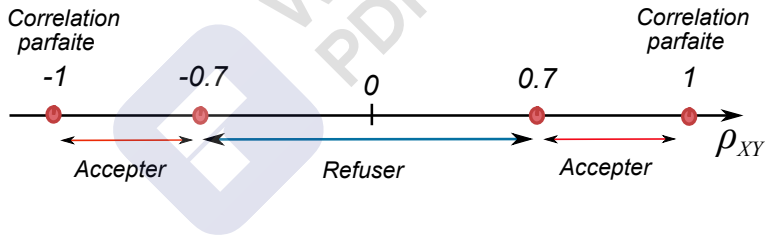


FIGURE 4.8: La zone d'acceptation ou de refus de l'ajustement linéaire.

Chapitre 5

Annexe historique

[1]*L'histoire de la "statistique" remonte à une époque très ancienne. Les activités statistiques (dénombrements) ont commencé bien avant la création du mot, l'application de la méthode et de l'analyse statistique.*

Depuis l'antiquité, les Empereurs, les Rois et les Hommes d'Eglise réalisaient des dénombrements de populations humaines et de terres pour les besoins de la guerre et de l'impôt.

Il y a plus de 4 ou 5000 ans, il existait déjà en Chine des descriptions chiffrées de la population et de l'agriculture.

Les Égyptiens de l'époque des Pharaons procédaient au dénombrement de la population ou du bétail.

A Rome, l'empereur Auguste fit procéder à une vaste enquête en dénombrant les soldats, les navires et les revenus publics.

Jusqu'au moyen âge, les seules "statistiques" existante étaient les dénombrements faits dans des buts divers : assiettes de l'impôt, répartition des terres, recrutement dans l'armée est effectués avec des méthodes diverses (recensements des personnes, enregistrements de certains actes d'état civil ...).

C'est à partir du XVIII siècle, qu'apparait le mot "statistique" crée par ACHENWAL en 1749 à partir du mot "STATISTA" (politique). Du simple dénombrement de populations humaines et de terres, la statistique est devenue une science qui a retenu et continue de retenir l'attention, non seulement des empereurs et de rois, mais surtout des personnes de sciences.

L'extension et l'utilisation du calcul des probabilités développé par J. BERNOULLI au 18ème siècle et l'application des études démographiques et sociales ont permis à cette science

1. Source : B. Oukacha and M. Benmessaoud, Statistique descriptive et calcul des probabilités, 2013.

de connaitre un essor considérable. Ainsi au 19 siècle, de la simple statistique descriptive, elle passe au stade de "Statistique Mathématique".

Depuis le 20ème siècle, les travaux de KARL PEARSON (1857-1936), de STUDENT (WILLIAM SEALY GOSSET, 1876-1937) et de RONAL FISHER (1890-1963) ont permis à cette science de connaitre un développement considérable et une application vaste et variée. La statistique utilise les techniques et des méthodes de collecte, de présentation, d'étude et d'analyse des données quantitatives.

La statistique n'est pas uniquement utilisée pour décrire, pour mieux connaitre un événement survenu dans le passé mais elle intervient de plus en plus dans les travaux de planification, dans le choix de prises de décisions et d'actions.