R2.08 Statistiques descriptives
BUT Informatique

Claire Wolfersperger

2022

Chapitre 1

Généralités sur les statistiques

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 : de l'économie à la biologie en passant par la psychologie et l'informatique. 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.

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

Exemple 2 *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.

Definition 3 *On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté* Ω.

Exemple 4 *Si l'on s'intéresse à 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* $\Omega =$*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.

Definition 5 *On appelle individu tout élément de la population Ω, il est noté ω (ω dans Ω).*

Exemple 6 *Si on étudie la production annuelle d'une usine de puces électroniques. La population est l'ensemble des puces produites durant l'année et une puce électronique 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.

Definition 7 *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 8 *Taille, température, nationalité, couleur des yeux, catégorie socioprofessionnelle ...*

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

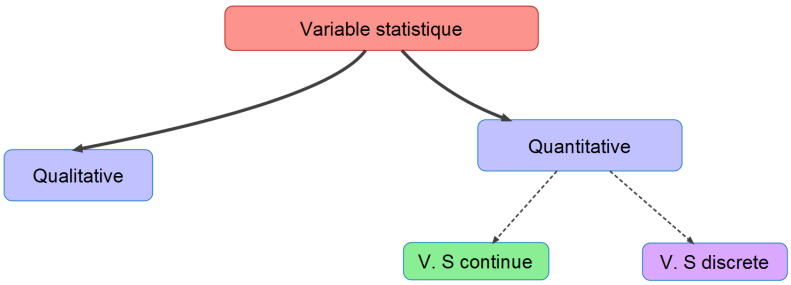
1.1.5 Modalités

Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci. Ce sont les différentes situations dans lesquelles les individus peuvent se trouver à l'égard du caractère considéré.

Exemple 9 *: Si la variable est " situation familiale ", les modalités sont " célibataire, marié, divorcé ".*
Si la Variable est" statut d'interrupteur ", les modalités sont " 0 et 1 ".

1.2 Types de 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.

Definition 10 *Les éléments de C sont représentés par autre chose que des chiffres.*

Exemple 11 *Le caractère : "état d'une maison" est un caractère qualitatif car 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 ou le salaire d'un individu sont des caractères quantitatifs. Donc, nous avons la définition suivante.

Definition 12 *L'ensemble des valeurs est représenté par des chiffres. De même, il est partagé en deux sortes de caractères, discret et 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. Par exemple, le temps de réalisation d'une tâche.

1.3 Exercices

Exercice 1

La variable statistique "couleur de maisons d'un quartier" est-elle :

- ◯ *qualitative* ◯ *quantitative*
- ◯ *discrète* ◯ *continue*

La variable statistique "revenu brut" est-elle :

- ◯ *qualitative* ◯ *quantitative*
- ◯ *discrète* ◯ *continue*

La variable statistique "nombre de maisons vendues par ville" est-elle :

- ◯ *qualitative* ◯ *quantitative*
- ◯ *discrète* ◯ *continue*

Exercice 2 *Parmi ces assertions, préciser celles qui sont vraies, celles qui sont fausses.*

1. *On appelle variable, une caractéristique que l'on étudie.*
2. *La tâche de la statistique descriptive est de recueillir des données.*
3. *La tâche de la statistique descriptive est de présenter les données sous forme de tableaux, de graphiques et d'indicateurs statistiques.*
4. *En Statistique, on classe les variables selon différents types.*
5. *Les valeurs des variables sont aussi appelées modalités.*
6. *Pour une variable qualitative, chaque individu statistique ne peut avoir qu'une seule modalité.*
7. *Pour faire des traitements statistiques, il arrive qu'on transforme une variable quantitative en variable qualitative.*
8. *La variable quantitative poids d'automobile peut être reclassée en variable qualitative en compacte, intermédiaire et grosse.*
9. *En pratique, lorsqu'une variable quantitative discrète prend un grand nombre de valeurs distinctes, on la traite comme continue.*

Exercice 3 *Proposer des exemples de variable quantitative transformée en variable qualitative. Préciser les modalités de cette dernière.*

Exercice 4 *Pour chacune des variables suivantes, préciser si elle est qualitative, quantitative discrète ou quantitative continue :*

1) Revenu annuel	*2) Citoyenneté*	*3) Distance*	*4) Taille*
5) Lieu de résidence	*6) Âge*	*7) Couleur des yeux*	*8) Nombre de langues parlées*

Exercice 5 *Pour les sujets d'étude qui suivent, spécifier : l'unité statistique, la variable statistique et son type :*

1. *Étude du temps de validité des lampes électriques.*
2. *Étude de l'absentéisme des ouvriers, en jours, dans une usine.*
3. *Répartition des étudiants d'une promotion selon la mention obtenue sue le diplôme du Bac.*
4. *On cherche à modéliser le nombre de collisions impliquant deux voitures sur un ensemble de 100 intersections routières choisies au hasard dans une ville. Les données sont collectées sur une période d'un an et le nombre d'accidents pour chaque intersection est ainsi mesuré.*

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 : $X : \Omega \to \{x_1; x_2; \cdots; x_n\}$ avec $Card(\Omega) = N$ est le nombre d'individus dans notre étude.

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

Exemple 13 *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

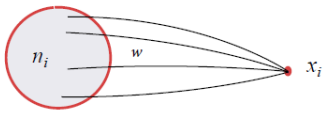
Nous avons Ω qui représente l'ensemble des familles, ω une famille, X le nombre d'enfants par famille ($X : \omega \to X(\omega)$).

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)

Definition 14 *Pour chaque valeur x_i, on pose par définition $n_i = Card(\{\omega \in \Omega : X(\omega) = x_i\})$.*
ni est le nombre d'individus qui ont le même x_i, c'est l'effectif partiel de x_i.



Exemple 15 *Dans l'exemple 13 précédent, 66 est le nombre de familles qui ont 2 enfants.*

2.1.2 Effectif cumulé croissant

Definition 16 *Pour chaque valeur x_i, on pose par définition : $N_i = n_1 + n_2 + ... + n_i$. L'effectif cumulé croissant N_i d'une valeur est la somme de l'effectif de cette valeur et de tous les effectifs des valeurs qui précèdent.*

Exemple 17 *Pour l'exemple 13, compléter le tableau suivant :*

x_i	0	1	2	3	4	5	6
N_i							

Quel est le nombre de familles qui ont moins de deux enfants ?

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 $N = Card(\Omega) = \sum_{i=1}^{n} n_i$. Dans notre exemple précédent, nous avons $N = 200$.

2.2 Fréquence partielle - Fréquence cumulée

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

2.2.1 Fréquence partielle (fréquence relative)

Definition 18 *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.

Interprétation : Si on écrit f_i sous la forme d'un pourcentage, il représente le pourcentage des ω tel que $X(\omega) = x_i$.

Exemple 19 *Dans l'exemple 13, $f_3 = \frac{n_3}{N} = \frac{66}{200} = 0,33$.*
Il y a donc 33% de familles dont le nombre d'enfants égale à 2.

On en déduit la propriété suivante :

Proposition 20 *Soit f_i défini comme précédemment. Alors, $\sum_{i=1}^{n} f_i = 1$.*

Démonstration : Rappelons que $\sum_{i=1}^{n} n_i = N$. Donc $\sum_{i=1}^{n} f_i = \frac{1}{N} \sum_{i=1}^{n} n_i = \frac{N}{N} = 1$.

2.2.2 Fréquence cumulée croissante

Definition 21 *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 croissante 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 22 *Pour l'exemple 13, compléter le tableau suivant :*

x_i	0	1	2	3	4	5	6
F_i							

Quel est le pourcentage de familles qui ont moins de deux enfants ?

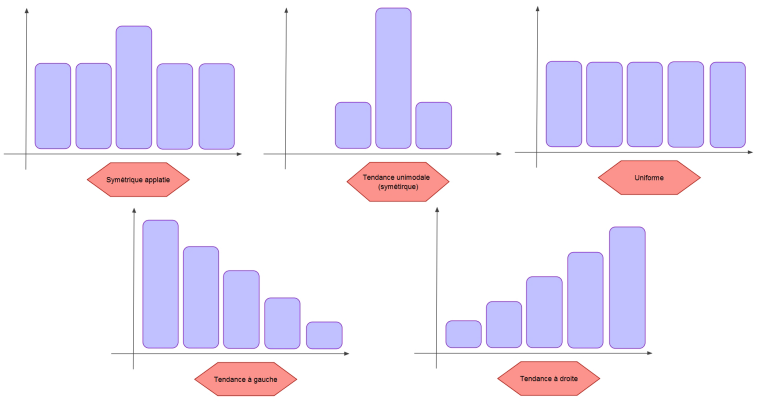
Nous avons vu que les tableaux sont un moyen souvent indispensable, en tous cas très utile, de classification et de présentation des unités d'une population statistique. Dans le paragraphe suivant, nous allons voir comment on traduit ses tableaux en graphique permettant aussi de résumer d'une manière 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'oeil les principales caractéristiques (mais on perd une quantité d'informations) comme dans l'exemple suivant :



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'orgue : 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.

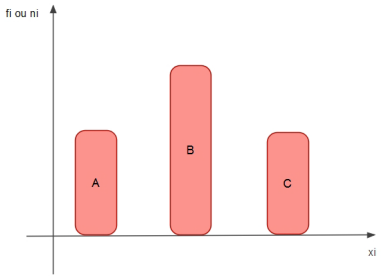
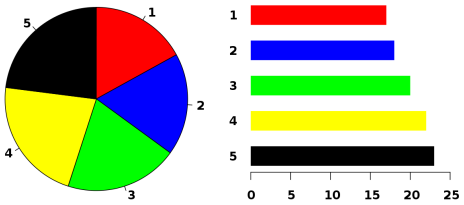


Diagramme circulaire : Les diagrammes circulaires, ou semi-circulaires, consistent à partager un disque ou un demi-disque, en tranches, ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence, de la modalité.

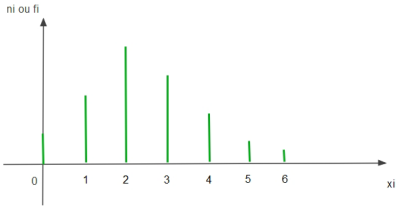


La mesure de l'angle en degré d'un secteur est proportionnelle à son effectif.
Si d_i est l'angle du secteur de la modalité i d'effectif n_i, alors $d_i = \dfrac{n_i \times 360}{N}$.

2.3.2 Distribution à caractère quantitatif discret

A partir de l'observation d'une variable quantitative discrète, deux diagrammes peuvent représenter cette variable : le diagramme en bâtons et le diagramme cumulatif (voir paragraphe suivant). Pour l'illustration, nous prenons l'exemple 13 précédent.

Diagramme en bâtons : À chaque x_i correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs représentés.



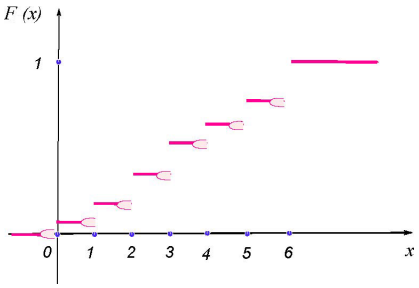
2.3.3 Représentation sous forme de courbe et fonction de répartition

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.

Definition 23 *Soit la fonction* $F : \mathbb{R} \to [0, 1]$ *définie par :*
$F(x)$ =*pourcentage des individus dont la valeur du caractère est* $\leqslant x$.
Cette fonction s'appelle la fonction de répartition du caractère X.

Remarque : Pour tout $i \in \{1, \cdots, n\}$, on a $F(x_i) = F_i$.
La courbe de F passe par les points (x_1, F_1), (x_2, F_2), ... et (x_n, F_n).

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



Proposition 24 *La fonction de répartition satisfait pour tout* $i \in \{1, \cdots, n\}$ *:*

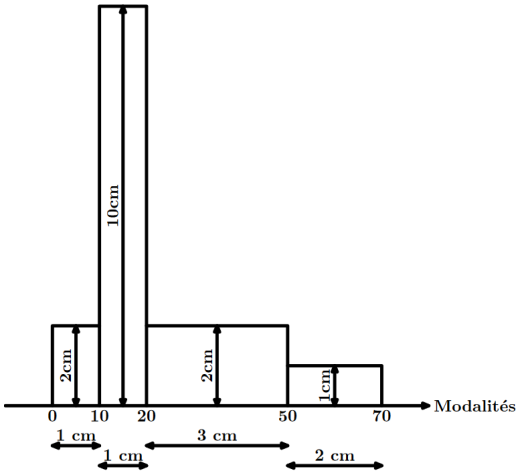
- *l'égalité :* $F(x_i) = F_i$
- *l'expression :* $F(x) = \begin{cases} 0, & si\ x < x_1, \\ F_1, & si\ x_1 \leqslant x < x_2, \\ F_i, & si\ x_i \leqslant x < x_{i+1}, \\ 1, & si\ x \geqslant x_n. \end{cases}$

2.3.4 Distribution à caractère quantitatif continu

Lorsque la variable est quantitative continue, on utilise un histogramme.
Les modalités sont des intervalles $m_1 = [a_1, a_2[$, $m_2 = [a_2, a_3[$, $\ldots$, $m_p = [a_p, a_{p+1}[$.
L'amplitude de m_1 est $a_2 - a_1$, l'amplitude de m_2 est $a_3 - a_2$ etc. À chaque modalité $[a_i, a_{i+1}[$ correspond un rectangle. Les largeurs des rectangles sont proportionnelles aux amplitudes des modalités correspondantes. Les surfaces des rectangles sont proportionnelles aux effectifs des modalités correspondantes.

Exemple : On s'intéresse à l'âge d'un échantillon de 1000 personnes.

Modalités	$[0, 10[$	$[10; 20[$	$[20; 50[$	$[50; 70[$
Effectifs	100	500	300	100



2.4 Paramètres de position

Les indicateurs statistiques de tendance centrale (dits aussi de position) considérés fréquemment sont la moyenne, la médiane et le mode. Il existe d'autres paramètres de position qui ne donnent pas une tendance centrale : les quartiles.

2.4.1 Le mode

Definition 25 *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 noté* M_0*.*

Exemple : Dans l'exemple 13, le mode est $M_0 =$...

Remarque : Il peut y avoir aucun mode, un mode ou plusieurs modes.

2.4.2 La médiane

Definition 26 *On appelle médiane, notée* Me*, la première valeur* x_i *de la V.S* X *telle que* $F(x_i) \geqslant 0,5$*.*
La médiane partage la série statistique en deux groupes de même effectif.

Exemple : Dans l'exemple 13, la médiane est $Me =$ ... car

2.4.3 La moyenne

Definition 27 *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 $N = Card(\Omega)$.*

On peut donc exprimer et calculer la moyenne dite "arithmétique" avec des effectifs ou avec des fréquences.

Exemple : Dans l'exemple 13, la moyenne est $\overline{x} = 2,46$ car

Cela signifie que dans ce village, les familles ont en moyenne 2,46 enfants à charge. Ce chiffre ne correspond pas à un fait concret. La valeur de la moyenne est abstraite.

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. 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.

Proposition 28 *La moyenne est linéaire : quand on ajoute (resp. multiplie) une même quantité à toutes les valeurs d'une série, la nouvelle moyenne s'obtient en ajoutant (resp. multipliant) cette quantité à l'ancienne moyenne.*

2.4.4 Les quartiles

Definition 29 *On appelle premier quartile, noté Q_1, la première valeur x_i de la V.S X telle que $F(x_i) \geqslant 0,25$.*
On appelle troisième quartile, noté Q_3, la première valeur x_i de la V.S X telle que $F(x_i) \geqslant 0,75$.

Exemple : Dans l'exemple 13, les quartiles sont :

2.5 Paramètres de dispersion (variabilité)

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

2.5.1 L'étendue

Definition 30 *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.

Exemple : Dans l'exemple 13, l'étendue est ...

2.5.2 La variance

Definition 31 *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.

Exemple : Dans l'exemple 13, la variance est ...

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

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

.

2.5.3 L'écart-type

Definition 33 *La quantité $\sigma(X) = \sqrt{Var(X)}$ s'appelle l'écart-type de la V.S. X.*

Le paramètre $\sigma(X)$ mesure la distance moyenne entre $\overline{x}$ et les valeurs de X. 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).

2.6 Exercices

Exercice 6 *Le tableau suivant donne la répartition selon le groupe sanguin de 40 individus pris au hasard dans une population :*

Groupe sanguin	*A*	*B*	*AB*	*O*
Effectif	*20*	*10*	*x*	*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 et les réaliser.*

Exercice 7 *Le gérant d'un magasin vendant des articles de consommation courante a relevé pour un article particulier qui semble connaître une très forte popularité, le nombre d'articles vendus par jour. Son relevé a porté sur les ventes des mois de Mars et Avril, ce qui correspond à 52 jours de vente. Le relevé des observations se présente comme suit : 7 13 8 10 9 12 10 8 9 10 6 14 7 15 9 11 12 11 12 5 14 11 8 10 14 12 8 5 7 13 12 16 11 9 11 11 12 12 15 14 5 14 9 9 14 13 11 10 11 12 9 15.*

1. *Quel type est la variable statistique étudiée.*
2. *Déterminer le tableau statistique en fonction des effectifs, des fréquences, des effectifs cumulés et des fréquences cumulés.*
3. *Tracer le diagramme en bâtons associé à la variable X.*
4. *Déterminer F la fonction de répartition de X.*
5. *Calculer le mode M_0 et la moyenne arithmétique $\overline{x}$.*
6. *Déterminer à partir du tableau puis à partir de la fonction de répartition, la valeur de la médiane Me et des quartiles Q_1 et Q_3.*
7. *Calculer l'étendue, la variance, l'écart-type et l'écart interquartile.*

Exercice 8 *On considère deux groupes d'étudiants. Nous relevons leurs notes d'examens dans les deux tableaux suivants :*

Note (groupe A)	8	9	10	11
Effectif	2	2	1	1

Note (groupe B)	6	8	9	13	14
Effectif	2	2	2	1	1

Calculer les paramètres de position et de dispersion de ces deux séries puis les comparer.

Exercice 9 *Une entreprise, où le salaire mensuel moyen est de $2339,50\,€$ propose une augmentation généralisée du salaire de ses employés, selon deux modalités possibles :*

- *Modalité 1 : tous les salaires augmentent de 10%*
- *Modalité 2 : tous les salaires augmentent de $200\,€$.*

1. *Déterminer quel serait le salaire moyen si la modalité 1 est choisie. Même question avec la modalité 2.*
2. *L'entreprise réalise un vote auprès de ses employés pour savoir quelle modalité choisir. A votre avis, quelle modalité va être choisie par les employés ?*
3. *La répartition des salaires dans l'entreprise est la suivante :*

Salaire	*1 450*	*1 510*	*1 925*	*5 125*
Nombre d'employés	*15*	*10*	*15*	*10*

Calculer les paramètres de position et de dispersion de cette série.

4. *Le résultat du vote montre que les employés préfèrent la modalité 2. Expliquer pourquoi*

Exercice 10 *Voici le tableau des pourcentages obtenu pour la variable " Mode de logement" :*

x_i	"Cité U"	"Studio"	"Résidence"	"Maison"	"Autre"	TOTAL
%	4.8	16.5	38.6	28.6	11.6	100

Sachant que la taille de l'échantillon $N = 189$, retrouver les effectifs pour chaque modalité.

Exercice 11 *Au poste de péage, on compte le nombre de voitures se présentant sur une période de 5 min. Sur 100 observations de 5 min, on obtient les résultats suivants :*

Nombre de voitures	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Nombre d'observations	*2*	*8*	*14*	*20*	*19*	*15*	*9*	*6*	*2*	*3*	*1*	*1*

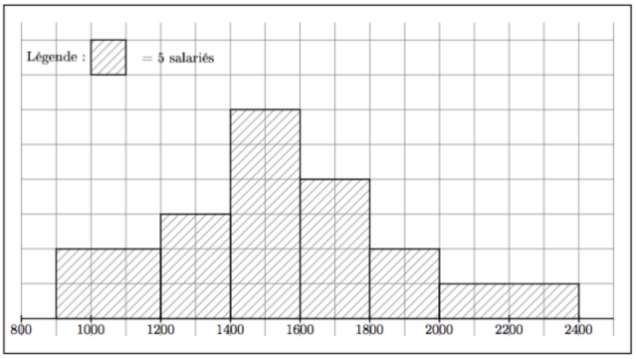
1. *Quel est le type de la variable statistique étudiée ?*
2. *Construire la table des fréquences et le diagramme en bâtons en fréquences de la série du nombre de voitures.*
3. *Déterminer F la fonction de répartition et tracer la courbe cumulative des fréquences.*
4. *Calculer le mode M_0 et la moyenne arithmétique $\overline{x}$.*
5. *Déterminer la valeur de la médiane Me et des quartiles Q_1 et Q_3.*
6. *Calculer l'étendue, la variance, l'écart-type et l'écart interquartile.*

Exercice 12 *On considère une variable statistique qui prend les valeurs suivantes :*
188, 169, 181, 180, 159, 180, 164, 184, 177, 159, 187, 174, 170, 173, 175, 150, 171, 166, 173, 182, 167, 173, 174, 175, 166, 173, 180, 188, 165, 173, 178, 182, 175, 186, 162, 193, 173, 184, 184, 184, 182, 187, 193, 161, 193, 169, 163, 179, 179, 184, 182, 182, 172, 175, 193, 170, 176, 166, 177, 163, 191, 171, 189, 183, 178, 197, 166, 187, 180, 172, 181, 167, 177, 177, 186, 174, 168, 182, 183, 182, 170, 186, 193, 167, 184, 159, 169, 192, 172, 167, 178, 176, 177, 175, 172, 175, 182, 176, 179, 188.
Même si les tailles ne prennent que des valeurs entières, il n'est pas opportun de la représenter comme une V.S discrète puisque dans ce cas on se retrouverait avec 34 modalités différentes ! C'est pour cette raison qu'on va considérer cette variable statistiques comme continue prenant ses valeurs dans l'intervalle $[150; 200]$.

1. *Si on découpe l'intervalle* $[150; 200]$ *en 5 intervalles d'amplitude 10, déterminer le tableau statistique en fonction des effectifs et des fréquences.*
2. *Tracer l'histogramme correspondant.*
3. *Reprenons les données en les étudiant suivant d'autres modalités. Compléter le tableau suivant et construire l'histogramme correspondant :*

X	$[150; 165[$	$[165; 170[$	$[170; 175[$	$[175; 180[$	$[180; 190[$	$[190; 200[$
Effectif						
Amplitude						
Largeur rectangle						

Exercice 13 *Voici un histogramme représentant la répartition des salaires dans une entreprise :*



1. *Retrouver le tableau des effectifs et des fréquences de cette série statistique.*
2. *Calculer les indicateurs de position et de dispersion de cette variable statistique.*

Exercice 14 *Un groupe de 8 élèves a obtenu les moyennes suivantes en mathématiques au cours du premier trimestre de l'année scolaire : 8, 8, 9, 10, 10, 11, 12, 12.*

1. *Déterminer la moyenne et l'écart type de la série, ainsi que la médiane et l'écart interquartile.*
2. *Au deuxième trimestre, deux nouveaux élèves rejoignent le groupe, avec des moyennes respectives de 4 et 14.*
   (a) *Déterminer les mêmes paramètres statistiques de la nouvelle série.*
   (b) *Quelle est à 1% près la variation en pourcentage de chacun des paramètres ?*
   Lequel des deux couples d'indicateurs paraît le moins sensible aux valeurs extrêmes, entre moyenne/écart type et médiane/écart interquartile ?

Chapitre 3

Étude d'une variable statistique à deux dimensions

Dans le chapitre précédent, nous avons présenté les méthodes qui permettent de résumer et représenter les informations relatives à une variable. Un même individu peut être étudié à l'aide de plusieurs caractères (ou variables). Par exemple, les salaries en regardant leur ancienneté et leur niveau d'étude, la croissance d'un enfant en regardant son poids et sa taille. Dans la suite, nous introduisons l'étude globale des relations entre deux variables (en nous limitant au cas de deux variables). Donc, soit Ω une population et

$$\begin{array}{rccc} Z & : & \Omega & \to & \mathbb{R}^2 \\ & & \omega & \mapsto & Z(\omega) = (X(\omega), Y(\omega)) \end{array}$$

ou directement

$$\begin{array}{rccc} (X,Y) & : & \Omega & \to & \mathbb{R}^2 \\ & & \omega & \mapsto & Z(\omega) = (X(\omega), Y(\omega)) \end{array}$$

Dans ce cas, Z est dite variable statistique à deux dimensions avec $Card(\Omega) = N$, avec N un entier. Le couple (X, Y) est appelé le couple de la variable statistique.

Exemples :

- On observe simultanément sur un échantillon de 200 foyers, le nombre d'enfants X et le nombre de chambres 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.

3.1 Représentation des séries statistiques à deux variables

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

Présentation 1 : A chaque ω_i, on associe (x_i, y_i). On rassemblera les données comme dans le tableau suivant :

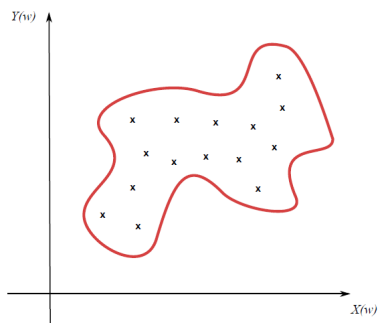
ω_i	ω_1	ω_2	...	ω_N
Variable X	$X(\omega_1)$	$X(\omega_2)$	...	$X(\omega_N)$
Variable Y	$Y(\omega_1)$	$Y(\omega_2)$	...	$Y(\omega_N)$

Cette représentation sera notée "présentation 1". Nous allons utiliser les notations suivantes : $x_i = X(\omega_i)$ et $y_i = Y(\omega_i)$.

Exemple : Soit Ω l'ensemble de 8 étudiants. 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.

ω_i	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
$X(\omega)$	8	2	6	6	11	10	7	2
$Y(\omega)$	9	10	11	7	14	16	12	5

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



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 sera notée "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, $n_{ij} = Card(\{\omega \in \Omega : Z(\omega) = z_{ij}\})$. Nous notons par f_{ij} la fréquence du couple (x_i, y_j). Cette fréquence est donnée par $f_{ij} = \dfrac{n_{ij}}{N}$ avec $N = Card(\Omega) = \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = n_{11} + n_{12} + ... + n_{1l} + n_{21} + n_{22} + ... + n_{2l} + ... + n_{k1} + n_{k2} + ... + n_{kl}$.

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

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

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

3.2 Exercice

Exercice 15 *Nous considérons 10 salariés qui sont observés à l'aide de deux variables "âge" et "salaire". Les informations brutes sont données dans le tableau suivant :*

Salaire	*1200*	*1480*	*1500*	*1640*	*1642*	*1780*	*1820*	*1980*	*1990*	*2150*
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 200 €.*
2. *Calculer f_{21}, f_{12}, f_{45} et f_{33}.*
3. *Déterminer les effectifs marginaux de X et de Y . Tracer le nuage de points.*
4. *Déterminer le tableau statistique des deux séries marginales X et Y .*

3.3 Description numérique

3.3.1 Caractéristique des séries marginales

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

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

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

Remarque : Dans le cas continu, x_i et y_j représentent respectivement le centre des classes de X et Y, c'est-à-dire, $x_i = \dfrac{L_i + L_{i+1}}{2}$ et $y_j = \dfrac{C_j + C_{j+1}}{2}$

Exemple : Les moyennes $\overline{x}$ et $\overline{y}$ dans l'exercice 15 sont :

Nous définissons maintenant la variance de X et la variance de Y comme suit :

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

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

Les écart-types de X et Y sont donnés par $\sigma_X = \sqrt{Var(X)}$ et $\sigma_Y = \sqrt{Var(Y)}$.

3.3.2 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, on a $\sigma_{X_j} = \sqrt{Var(X_j)}$ avec

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

Série conditionnelle par rapport à Y

Elle est notée par Y/x_i (ou Y_i) et on dit que c'est la série conditionnelle de Y sachant que $X = x_i$. Nous calculons aussi 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

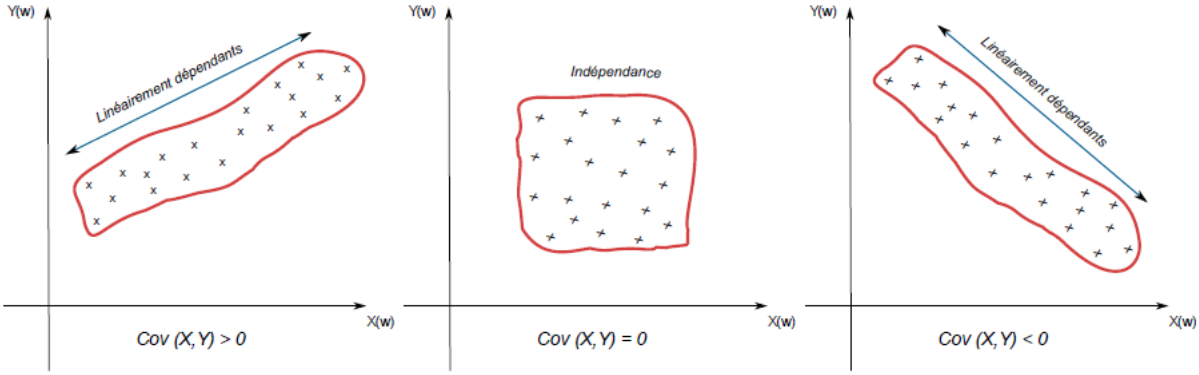
$$\overline{y_i} = \sum_{j=1}^{l} f_{j/i} y_j = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} n_{ij} y_j$$

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

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

3.3.3 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 suivante).



La covariance se calcule par l'expression suivante

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

ou

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

Remarques :

- Dans le cas où nous avons un tableau des données brutes "représentation 1" (nous néavons pas d'effectifs), on a $\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i$; $\overline{y} = \frac{1}{N}\sum_{i=1}^{n} y_i$ et $\overline{xy} = \frac{1}{N}\sum_{i=1}^{n} x_i y_i$.
- La covariance est une notion qui généralise la variance. En effet, $Cov(X,X) = Var(X)$ et $Cov(Y,Y) = Var(Y)$ car $Cov(X,X) = \overline{xx} - \overline{x} \times \overline{x} = \overline{x^2} - \overline{x}^2 = Var(X)$

Definition 34 *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 $Cov(X,Y) = 0$(réciproque est fausse).

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.

3.4 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.

3.4.1 Coefficient de corrélation

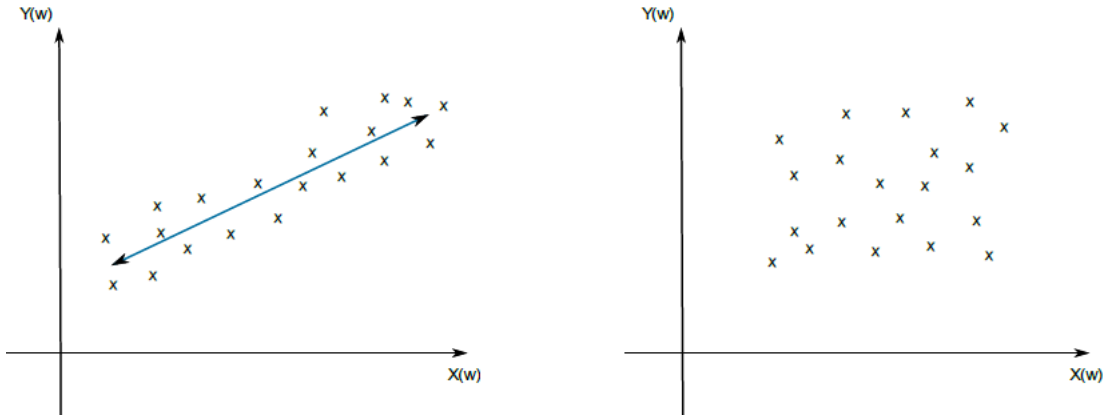
Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone.
Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires (voir ci-dessous). Il existe d'autres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

Definition 35 *Le coefficient de corrélation est* $\rho_{XY} = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$

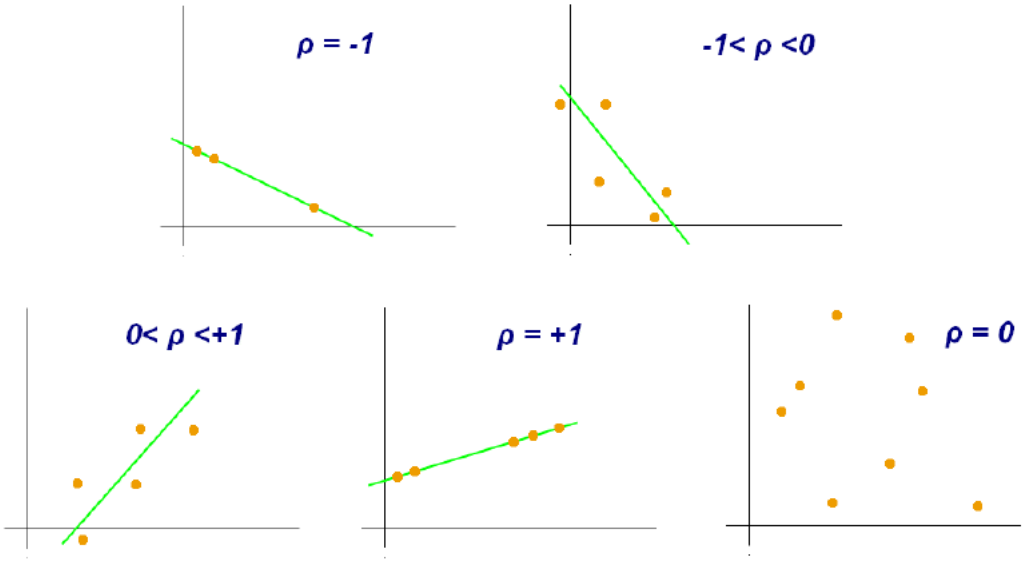
Proposition 36 *Le coefficient ρ_{XY} est un nombre compris entre -1 et 1.*

Le coefficient ρ_{XY} mesure le degré de liaison linéaire entre X et Y (voir figures suivantes). On a :

- Plus la valeur approchée de ρ_{XY} est proche de 1, plus X et Y sont liées linéairement.
- Plus la valeur approchée de ρ_{XY} est proche de 0, moins il y a de liaison linéaire entre X et Y .



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

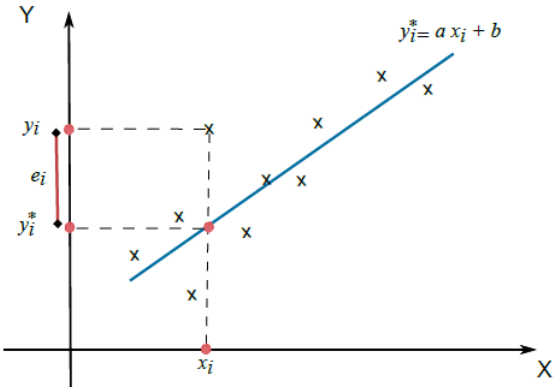


Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation.

Remarque : Si $\rho_{XY} = 0$ alors $Cov(X, Y) = 0$.

3.4.2 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.



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 la droite soit minimale. Cette distance matérialise 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 :

$$|e_i| = |y_i - y_i^*| = |y_i - ax_i - b|$$

. La méthode des moindres carrés consiste donc à minimiser la fonction U (la somme des erreur commise). 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, on obtient

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

c'est-à-dire

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

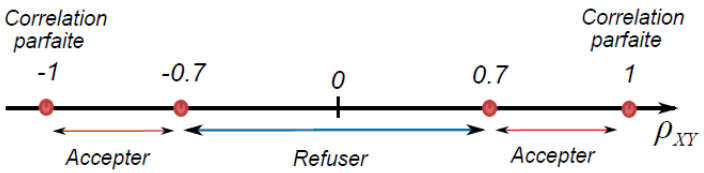
De même, après calcul, $\dfrac{\partial U}{\partial a} = 0$ implique :

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

Donc, la droite de régression, qui rend la distance entre elle et les points minimale, est donnée par :
$D(Y/X) : Y = aX + b$ avec $a = \dfrac{Cov(X,Y)}{Var(X)}$ et $b = \overline{y} - a\overline{x}$ ou
$D(X/Y) : X = a'Y + b'$ avec $a' = \dfrac{Cov(X,Y)}{Var(Y)}$ et $b' = \overline{x} - a'\overline{y}$.

Remarque : Le coefficient de corrélation ρ_{XY} permet de justifier l'ajustement linéaire. On adopte les critères suivants :

- Si $|\rho_{XY}| < 0,7$ alors l'ajustement linéaire est refusé (droite refusée).
- Si $|\rho_{XY}| \geqslant 0,7$ alors l'ajustement linéaire est accepté (droite acceptée).



3.5 Exercices

Exercice 16 *Nous considérons 10 joueurs et soient :*

- *Y la variable qui représente le nombre de jeux auquel un joueur joue.*
- *X la variable qui représente le gain ou perte ($+1$ s'il gagne 10 € et -1 s'il perd 10 € et 0 sinon).*

Nous avons le tableau de contingence suivant :

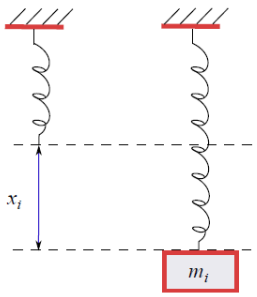
$X \setminus Y$	1	2	3	4	$n_{i\bullet}$
-1	0	1	2	2	
0	1	1	0	1	
1	0	1	1	0	
$n_{\bullet j}$					

1. *Compléter le tableau ci-dessus.*
2. *Calculer $Cov(X, Y)$.*

Exercice 17 *Dans un TP de physique, on a les données suivantes :*

x_i	0	0.5	1.1	1.5	1.9
m_i	0	10	20	30	40

La variable m_i représente les différentes masses appliquées comme dans le schéma ci-dessous et la variable x_i les hauteurs induits depuis l'état initial.



1. *Déterminer $D(m/x)$*
2. *Déterminer $D(x/m)$*
3. *Tracer le nuage de points et les deux droites. Représenter le point de coordonnées $(\overline{x}, \overline{y})$.*
4. *Peut-on déterminer x si $m = 51,75kg$?*

Exercice 18 *Le tableau de contingence suivant est entre le salaire mensuel X et l'ancienneté Y des cadres d'une entreprise.*

$X(\times 100) \setminus Y$	$[0, 8[$	$[8, 16[$	$[16, 24[$	$[24, 32[$	*Loi marginale*
$[20, 30[$	5	6	1	0	
$[30, 40[$	2	4	3	3	
$[40, 50[$	0	2	4	10	
Loi marginale					

1. *Etudier les séries marginales.*
2. *Déterminer si les variables X et Y sont indépendantes.*
3. *Etudier les séries conditionnelles X/y_3 et Y/x_2 et présenter les résultats pour chaque groupe de séries conditionnelles.*

Exercice 19 *Pour les données suivantes :*

X	1	2	7	4	6
Y	5	4	1	3	2

1. *Tracer le nuage de points.*
2. *Sans calcul, deviner le signe et la valeur du coefficient de corrélation.*
3. *Calculer le coefficient de corrélation, la pente et l'ordonnée à l'origine de la droite de régression.*

Exercice 20 *Une usine produit des pièces d'une machine. Pour chaque pièce (individu), on dispose du coût de sa production (€) et du temps nécessaire pour sa réalisation (en heures). Le tableau ci-après (série statistique) donne cette répartition :*

Individu	*1*	*2*	*3*	*4*	*5*
Temps X (en h)	*2*	*3*	*52*	*2*	*4*
Coût Y (en €)	*10*	*16*	*23*	*12*	*18*

1. *Calculer la moyenne de la variable statistique X.*
2. *Calculer la moyenne de la variable statistique Y .*
3. *Calculer l'écart-type de la variable statistique X.*
4. *Calculer l'écart-type de la variable statistique Y .*
5. *Calculer la covariance des variable statistiques X et Y .*
6. *En supposant qu'il existe une corrélation linéaire entre X et Y , déterminer cette droite de corrélation.*
7. *Calculer le coefficient de corrélation. Conclusion ?*
8. *Une nouvelle pièce est réalisée en 6 heures. Estimer le coût de production de cette pièce en utilisant la droite de corrélation établie.*

Exercice 21 *Soit X et Y deux variables statistiques mesurées sur un même individu. Par exemple, pour l'individu $n°3$, $X = 2$ et $Y = 8$.*

Individu	*1*	*2*	*3*	*4*	*5*
X	*3*	*4*	*2*	*5*	*3*
Y	*12*	*14*	*8*	*19*	*11*

1. *Calculer la moyenne de la variable statistique X.*
2. *Calculer la moyenne de la variable statistique Y .*
3. *Calculer l'écart-type de la variable statistique X.*
4. *Calculer l'écart-type de la variable statistique Y .*
5. *Calculer la covariance des variable statistiques X et Y .*
6. *En supposant qu'il existe une corrélation linéaire entre X et Y , déterminer cette droite de corrélation.*
7. *Calculer le coefficient de corrélation. Conclusion ?*

Exercice 22 *Une étude sur le chômage a été faite et qui s'intéresse à l'ancienneté du chômage (X) moins de 24 mois, et l'âge (Y) entre 20 et 35 ans. Les résultats sont donnés par le tableau de contingence suivant :*

$\boldsymbol{X} \setminus \boldsymbol{Y}$	$[20, 25[$	$[25, 30[$	$[30, 35[$
$[0, 6[$	*10*	*8*	*5*
$[6, 12[$	*8*	*9*	*4*
$[12, 18[$	*15*	*11*	*9*
$[18, 24[$	*3*	*6*	*2*

1. *Quel est le nombre d'individus qui ont une ancienneté de chômage moins d'un an ?*
2. *Déterminer les deux distributions marginales.*
3. *Déterminer la distribution de X conditionnelle à $Y = [25; 30]$, c'est à dire, $X/Y = [25; 30]$.*
4. *Les variables X et Y sont elles indépendantes ? Justifier.*
5. *Donner la moyenne arithmétique.*
6. *Calculer le coefficient de corrélation linéaire. Commenter.*
7. *Donner l'équation de la droite de régression de Y en fonction de X.*
8. *Quel sera l'âge d'une personne ayant une ancienneté de chômage de 15 mois.*