Université Moulay Ismail
FST-Errachidia

Cours du module M136-S3-MIP

Statistique déscriptive et Probabilité

Pr : M.Haddaoui

Département : Mathématique

أعوذ بالله من الشيطان الرجيم

بسم الله الرحمن الرحيم

الحمد لله رب العالمين ۝ الرحمن الرحيم ۝ مالك يوم الدين ۝ إياك نعبد وإياك
نستعين ۝ اهدنا الصراط المستقيم ۝ صراط الذين أنعمت عليهم غير
المغضوب عليهم ولا الضالين ۝

إهـــــــداء

إلـــى أمـــي الغـاليـــة رحمـــــهـــا الله

إلـــى أبـــي العزيـــــز حفـــظـــه الله

إلـــى زوجتي الحبيـــبة أيـــدهـــا الله

إلى أولادي:إبتهال،عطاء أسيد،وفقهـم الله

إلـى أصهاري وأبناءهـم بـارك فيهـم الله

إلـــى إخوتي وإخوانـــي أعانهـــم الله

إلـــى أصدقائي وزملائي يسر لهـــم الله

إلى أساتذتي ومشايخي وكل من علمني تقبل منـهم الله

إلى كل من أحبنا أو أحببناه في الله وإلى كل المؤمنين بالله

والصلاة والسلام على رسـول الله وآله وصحبه وإخوانه وحزبه ومـــن والاه

Table des matières

Première partie

Statistique descriptive

Introduction

Aussi loin que l'on remonte dans le temps, les civilisations ont toujours senti le besoin de disposer d'informations sur leurs sujets ou sur les biens qu'ils possèdent et produisent. Mais les recensements de population et de ressources, les statistiques (du latin 'status' : état) sont restés purement descriptifs jusqu'au 17ème siècle. Puis s'est développé le calcul des probabilités et des méthodes statistiques sont apparues en Allemagne, en Angleterre et en France. Beaucoup de scientifiques de tout horizon ont apporté leur contribution au développement de cette science : PASCAL, HUYGENS, BERNOULLI, MOIVRE, LAPLACE, GAUSS, MENDEL, PEARSON, FISCHER... Actuellement, la statistique s'applique à la plupart des disciplines : agronomie, biologie, démographie, économie, sociologie, linguistique, psychologie...
Le terme "statistique" est issu du latin 'statisticum', c'est-à-dire qui a trait à l'état. Les statistiques descriptives regroupent les méthodes dont l'objectif principal est la description des données étudiées. Cette description des données se fait à travers leur représentation graphique, et le calcul de résumés numériques. L'objectif de la statistique descriptive est :
* décrire de façon synthétique et avec un minimum d'effort des données observées ;
* la collecte, l'organisation, la présentation de données ;
* la modélisation et la construction de résumés numériques permettant de décrire et d'analyser des phénomènes repérés ;
* l'extrapolation des résultats partiels en vue de déduire des précisions globales (Statistique Inférentielle).

Chapitre 1

Concepts généraux

1.1 Terminologie

1. **Population** : ensemble (souvent noté Ω) que l'on observe et qui sera soumis a une analyse statistique. Chaque élément de cet ensemble est un individu ou unité statistique (souvent noté ω).
2. **Echantillon** : c'est un sous ensemble E de la population Ω considérée. Le nombre d'individus dans l'échantillon est la taille de l'échantillon.
3. **Caractère** : c'est le sujet de l'étude statistique et porte aussi le nom de variable statistique. Plus généralement, on appelle caractère toute application X de la population Ω dans un ensemble E, dont les éléments x sont appelés modalités du caractère X (ou valeurs du caractère). Il existe différents types de variables statistiques :
   - **Caractère qualitatif** : lorsque la variable ne se prête pas a des valeurs numériques (exemple : opinions politiques, couleurs des yeux, sexe, profession, nationalité...). Elle peut être ordonnée ou non, dichotomique ou non.
   - **Caractère quantitatif** (**ou mesurable**) : lorsque la variable peut être exprimée numériquement. Elle est discontinue si elle ne prend que des valeurs isolées les unes des autres. Une variable discontinue qui ne prend que des valeurs entières est dite discrète (exemple : nombre d'enfants d'une famille). Elle est dite continue lorsqu'elle peut prendre toutes les valeurs d'un intervalle fini ou infini (exemple : diamètre de pièces, salaires...).
4. **Série ou distribution statistique** : l'ensemble $X(\Omega) \subset E$ des modalités est parfois appelé distribution statistique ou série statistique ou encore variable statistique. On dit alors que l'on a effectué un regroupement des données brutes. De plus, pour chaque valeur x_i de modalité (du caractère) constatée, on détermine le nombre d'individus n_i ayant présenté cette valeur du caractère, nombre appelé effectif associé à la modalité. L'ensemble des couples $(x_i, n_i)_{i \in J}$ (modalité,effectif) déterminé est parfois appelé distribution statistique. On va étudier dans ce cours deux types de distribution :
   - **Distribution statistique univariée** (**ou simple**) : c'est une série statistique à un caractère ou à une dimension. Par exemple, dans le cas d'une série statistique quantitatif obtenue lorsque nous nous intéressons

à un caractère élémentaire, dont l'ensemble des modalités $X(\Omega)$ est un sous-ensemble de $\mathbb{R}$.

— **Distribution statistique bivariée (ou double)** : c'est une série statistique à deux caractères obtenue lorsque à chaque individu sont associés deux caractères élémentaires, plus précisément un couple de caractères élémentaires, ou encore un caractère à valeurs dans le produit cartésien $E \times F$. Par exemple, dans le cas d'une série statistique quantitatif $X(\Omega) \subset \mathbb{R}^2$.

1.2 Remarques

Remarque 1.1. Soit X l'application définie de la population Ω dans l'ensemble E, on peut écrire

$$\begin{aligned} X : \Omega &\to E \\ \omega &\mapsto X(\omega) := x. \end{aligned}$$

Donc on a

$$\begin{aligned} X : &\text{ le caractère} \\ \Omega : &\text{ la population} \\ \omega : &\text{ l'individu} \\ E : &\text{ l'ensembe des modalités} \\ X(\Omega) : &\text{ l'ensembe des modalités possibles} \\ X(\omega) := x : &\text{ modalité.} \end{aligned}$$

Remarque 1.2. On peut transformer toute statistique qualitative à une statistique quantitative à l'aide d'un codage des valeurs possibles du caractère. Par exemple, pour le caractère 'sexe', on utilise le codage usuel suivant : 1 = masculin et 2 = féminin.

Remarque 1.3. Dans le cas d'une série double, nous nous intéressons pour chaque individu au couple (x,y) de réponses et nous effectuons le regroupement des données par rapport à ces couples, alors que dans le cas de l'étude des deux séries simples associées, nous effectuons le regroupement des données séparément sur chacun des deux caractères X et Y ; nous obtenons alors des résultats plus concis, mais au prix d'une perte d'information.

1.3 Exemples

Exemple 1.1. Étude des membres des familles d'un quartier donné.

$$\begin{aligned} \text{Population} : &\text{ Ensembles des familles du quartier,} \\ \text{Individu} : &\text{ Une famille,} \\ \text{Caractère} : &\text{ Membre de famille,} \\ \text{Type} : &\text{ Quantitatif discret.} \end{aligned}$$

Exemple 1.2. Étude des couleurs des voitures d'une ville donnée.

Population : Ensembles des voitures de la ville,
Individu : Voiture,
Caractère : Couleur,
Type : Qualitatif.

Exemple 1.3. Une enquête réalisée dans un petit village porte sur le nombre d'enfants à charge par famille, les résultats sont regroupés dans la série statistique suivante :

$$0-1-4-2-2-2-3-3-3-3-3-4-4-4-4-3-1$$
$$0-1-4-2-2-2-3-3-3-3-3-4-4-4-4-3-2$$
$$0-1-4-2-2-2-3-3-3-3-3-4-4-4-4-1.$$

On a

Population : Ensembles des familles du village,
Individu : Une famille,
Caractère : Nombre d'enfants à charge,
Type : Quantitatif discret.

De plus le nombre des familles étudiées est $N = 50$.

Chapitre 2

Distribution univariée

2.1 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 de cette section, nous considérons la situation suivante :

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

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

2.1.1 Effectif partiel - Effectif cumulé

Définition 2.1. Pour chaque valeur x_i, on pose par définition

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

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

Définition 2.2. Pour chaque valeur x_i, on pose par définition

$$N_i = \sum_{k=1}^{k=i} n_k = n_1 + ... + n_i.$$

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. De plus on a

$$Card(\Omega) := N = \sum_{k=1}^{k=q} n_k.$$

Exemple 2.1. On considère l'exemple 3.2. On note X le nombre d'enfants. Les résultats sont données par ce tableau :

Table 2.1

x_i	0	1	2	3	4
n_i	3	5	10	17	15
N_i	3	8	18	35	50

2.1.2 Fréquence partielle - Fréquence cumulée

Définition 2.3. Pour chaque valeur x_i, on pose par définition

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

f_i s'appelle la fréquence partiel de x_i.

Définition 2.4. Pour chaque valeur x_i, on pose par définition

$$F_i = \sum_{k=1}^{k=i} f_k = f_1 + ... + f_i.$$

La fréquence cumulé F_i d'une valeur est la somme de la fréquence de cette valeur et de toutes les valeurs des fréquences qui la précèdent. De plus, on a

$$F_q = \sum_{k=1}^{k=q} f_k = f_1 + ... + f_q = 1$$

Exemple 2.2. On considère l'exemple 3.2 :

Table 2.2

x_i	0	1	2	3	4
n_i	3	5	10	17	15
f_i	$0,06$	$0,1$	$0,2$	$0,34$	$0,3$
F_i	$0,06$	$0,16$	$0,36$	$0,7$	1

2.1.3 Pourcentage partiel - Pourcentage cumulé

Définition 2.5. Pour chaque valeur x_i, on pose par définition

$$p_i = \frac{n_i}{N} \times 100 = f_i \times 100\%.$$

p_i s'appelle le pourcentage partiel de x_i.

Définition 2.6. Pour chaque valeur x_i, on pose par définition

$$P_i = \sum_{k=1}^{k=i} p_k = p_1 + ... + p_i.$$

Le pourcentage cumulé P_i d'une valeur est la somme du pourcentage de cette valeur et de toutes les valeurs des pourcentages qui précèdent. De plus, on a

$$P_q = \sum_{k=1}^{k=p} p_k = p_1 + ... + p_q = 100\%$$

Exemple 2.3. On considère l'exemple 3.2 :

TABLE 2.3

x_i	0	1	2	3	4
n_i	3	5	10	17	15
f_i	0,06	0,1	0,2	0,34	0,3
$p_i\%$	6	10	20	34	30
$P_i\%$	6	16	36	70	100

2.1.4 Fonction de répartition

Définition 2.7. La fonction de répartition F est la fonction définie de $\mathbb{R} \to [0, 1]$ par

$$F(x) = \begin{cases} 0, & \text{si } x < x_1, \\ F_i, & \text{si } x_i \leq x < x_{i+1}, \\ 1, & \text{si } x \geq x_q, \end{cases}$$

2.1.5 Paramètres de position

Définition 2.8. Une série statistique discrète $X = (x_i, n_i)_{i\in\{1,2,...,q\}}$ est dite ordonnée si $i < j \Rightarrow x_i < x_j$.

Définition 2.9. Pour chaque $X = (x_i, n_i)_{i\in\{1,2,...,q\}}$ série statistique ordonnée discrète, on pose par définition

1. Le mode : toute valeur de la modalité dont l'effectif est maximum. Autrement dit,
$$Mode = Mo = \{x_i : n_i = \max_{j\in\{1,2,...,q\}} n_j\}.$$
2. La moyenne arithmétique : le nombre réel noté $\overline{X}$ défini par
$$\overline{X} = \frac{1}{N} \sum_{i=1}^{q} n_i x_i, \text{ où } N = \sum_{i=1}^{q} n_i$$

3. La médiane : toute modalité $\{x_i$ telle que :

$$\sum_{j/x_j<x_i} \leq \frac{N}{2} \text{ et } \sum_{j/x_j>x_i} \leq \frac{N}{2}.$$

4. Le quantile d'ordre $\alpha, 0 < \alpha < 1$: la modalité de cette série pour laquelle l'effectif cumulé représente le quotient α de l'effectif total. Par exemple, les quartiles $\alpha = \frac{1}{4}$, les déciles $\alpha = \frac{1}{10}$, les centiles $\alpha = \frac{1}{100}$.
   — Si $\alpha = \frac{1}{4}$, alors le premier quartile noté Q_1 est la valeur du caractère pour un effectif cumulé de 25% de l'effectif total.
   — Si $\alpha = \frac{1}{2}$, alors le deuxième quartile noté Q_2 est la valeur du caractère pour un effectif cumulé de 50% de l'effectif total, autrement dit c'est la médiane.
   — Si $\alpha = \frac{3}{4}$, alors le troisième quartile noté Q_3 est la valeur du caractère pour un effectif cumulé de 75% de l'effectif total. De plus, on appelle $[Q_1, Q_3]$ l'intervalle interquartile et le nombre $Q_3 - Q_1$ l'écart interquartile qui est la largeur de l'intervalle interquartile.
   — Si $\alpha = \frac{i}{10}N, i \in \{1, 2, ..., 9\}$, alors on parle du $i^{ème}$ décile.
   — Si $\alpha = \frac{i}{100}N, i \in \{1, 2, ..., 99\}$, alors on parle du $i^{ème}$ centile.

Remarque 2.1. Une série statistique peut avoir plus qu'une valeur modale et plus d'une médiane. De plus, on peut caractériser la médiane par son effectif cumulé, et on obtient que la médiane est la plus petite modalité dont l'effectif cumulé est supérieur ou égale à la moitié de l'effectif total.

Exemple 2.4. On considère l'exemple 3.2 :

1. Le mode :
$$Mode = Mo = \{x_i : n_i = \max_{j\in\{1,2,...,5\}} n_j\} = 3.$$

2. La moyenne arithmétique :
$$\overline{X} = \frac{1}{50}\sum_{i=1}^{5} n_i x_i = \frac{3\times 0 + 5\times 1 + 10\times 2 + 17\times 3 + 15\times 4}{50} = 2,72$$

3. La médiane : c'est3 car pour la modalité 3 on a :
$$\sum_{j/x_j<x_i} = 3 + 5 + 10 = 18 \leq \frac{N}{2} = 25 \text{ et } \sum_{j/x_j>x_i} = 15 \leq \frac{N}{2} = 25$$

4. Les quartiles : $Q_1 = 2$, $Q_2 = 3$, $Q_3 = 4$.

2.1.6 Paramètres de dispersion

Définition 2.10. Pour chaque $X = (x_i, n_i)_{i\in\{1,2,...,q\}}$ série statistique ordonnée discrète, on pose par définition

1. L'étendue : la différence entre la plus grande et la plus petite valeur observée. Autrement dit,
$$Etendue = E = \max_{i\in\{1,2,...,q\}} x_i - \min_{i\in\{1,2,...,q\}} x_i.$$

2. L'écart moyen : le nombre réel positif noté $e(X)$ défini par

$$e(X) = \frac{1}{N}\sum_{i=1}^{q} n_i|x_i - \overline{X}|.$$

3. La variance : le nombre réel positif noté $V(X)$ défini par

$$V(X) = \frac{1}{N}\sum_{i=1}^{q} n_i(x_i - \overline{X})^2.$$

4. L'écart type : le nombre réel positif noté $\sigma(X)$ défini par

$$\sigma(X) = \sqrt{V(X)}.$$

Remarque 2.2.
1. L'écart moyen et la variance mesurent la dispersion des valeurs du caractère de la moyenne arithmétique.
2. La plupart des valeurs du caractère se trouvent dans l'intervalle $[\overline{X}-\sigma(X), \overline{X}+\sigma(X)]$. Ainsi, d'autant que $\sigma(X)$ est petit d'autant que les valeurs du caractère s'approchent de la moyenne arithmétique de la série statistique. Par conséquent, l'écart-type $\sigma(X)$ est le plus signifiant des paramètres des dispersion.

Exemple 2.5. On considère l'exemple 3.2 :

1. L'étendue :

$$Etendue = E = \max_{i\in\{1,2,...,5\}} x_i - \min_{i\in\{1,2,...,5\}} x_i = 4 - 0 = 4.$$

2. L'écart moyen :

$$\begin{aligned} e(X) &= \frac{1}{N}\sum_{i=1}^{q} n_i|x_i - \overline{X}| \\ &= \frac{3\times|0-2,72| + 5\times|1-2,72| + 10\times|2-2,72| + 17\times|3-2,72| + 15\times|4-2,72|}{50} \\ &= 0,9584. \end{aligned}$$

3. La variance :

$$\begin{aligned} V(X) &= \frac{1}{N}\sum_{i=1}^{q} n_i(x_i - \overline{X})^2 \\ &= \frac{3\times(-2,72)^2 + 5\times(-1,72)^2 + 10\times(-0,72)^2 + 17\times(0,28)^2 + 15\times(1,28)^2}{50} \\ &= 1,3616. \end{aligned}$$

4. L'écart type :

$$\sigma(X) = \sqrt{V(X)} = 1,1668.$$

Définition 2.11. Soient $a \in \mathbb{R}^*, b \in \mathbb{R}$ et $X = (I_i, n_i)_{i\in\{1,2,\dots,q\}}$ une série statistique ordonnée discrète, on définit la série statistique ordonnée discrète $aX+b$ par $aX+b = (x'_i, n'_i)_{i\in\{1,2,\dots,q\}}$ où $x'_i = ax'_i + b$ et $n'_i = n_i$.

Définition 2.12. Soient $a \in \mathbb{R}^*, b \in \mathbb{R}$ et $X = (x_i, n_i)_{i\in\{1,2,\dots,q\}}$ une série statistique ordonnée discrète, on définit $E(X^2)$ par

$$E(X^2) = \frac{1}{N}\sum_{i=1}^{q} n_i x_i^2.$$

Proposition 2.1. *Soient $a \in \mathbb{R}^*, b \in \mathbb{R}$ et $X = (I_i, n_i)_{i\in\{1,2,\dots,q\}}$. On a alors*

1. $V(X) = E(X^2) - E(X)^2$
2. *L'étendue de $aX + b = a\times$ l'étendue de X.*
3. *La moyenne arithmétique de $aX + b$:*

$$E(aX + b) = aE(X) + b.$$

4. *L'écart moyen de $aX + b$:*

$$e(aX + b) = ae(X).$$

5. *La variance de $aX + b$:*

$$V(aX + b) = a^2V(X).$$

6. *L'écart type de $aX + b$:*

$$\sigma(aX + b) = |a|\sigma(X).$$

2.2 Étude d'une variable statistique continue

On appelle variable statistique (V.S) continue (ou caractère continu) toute application de Ω dans $\mathbb{R}$ qui prend une infinité de valeurs dans un certain intervalle fini (par exemples : taille, poids...). Dans toute la suite de cette section, nous considérons le caractère continu suivant :

$$X : \Omega \to \mathbb{R} = \bigcup I_i$$

avec $Card(\Omega) := N$ est le nombre d'individus dans notre étude. Notons $I_i = [x_{i-1}, x_i[$ et posons $X(\Omega) = \bigcup_{i=1}^{q} I_i$ tel que pour tout $i, j \in \{1, 2, ..., q\}$ on a $I_i \cap I_j = \emptyset$. On note aussi $c_i = \frac{x_{i-1}+x_i}{2}$ le centre de la classe I_i.

2.2.1 Effectif partiel - effectif cumulé

Définition 2.13. Pour chaque classe I_i, on pose par définition

$$n_i = Card\{\omega \in \Omega : X(\omega) \in I_i\} = Card(X^{-1}(I_i)).$$

n_i s'appelle effectif partiel de la classe I_i.

Définition 2.14. Pour chaque valeur x_i, on pose par définition

$$N_i = \sum_{k=1}^{k=i} n_k = n_1 + ... + n_i = Card(X^{-1}(\bigcup_{k=1}^{i} I_k)).$$

L'effectif cumulé N_i d'une classe est la somme de l'effectif de cette classe et de tous les effectifs des classes qui précèdent. De plus, on a $Card(\Omega) := N = \sum_{k=1}^{k=q} n_k$.

Exemple 2.6. Une enquête réalisée dans un établissement, porte sur la taille(en m) des élèves du 1ère Bac-SM. Les résultats sont regroupés dans la série statistique suivante :

$$1,40 - 1,51 - 1,32 - 1,51 - 1,62 - 1,73 - 1,52 - 1,53 - 1,65 - 1,71 -$$
$$1,31 - 1,42 - 1,64 - 1,78 - 1,69 - 1,70 - 1,66 - 1,48 - 1,62 - 1,60.$$

Les résultats sont données par ce tableau :

Table 2.4

I_i	$[1,3;1,4[$	$[1,4;1,5[$	$[1,5;1,6[$	$[1,6;1,7[$	$[1,7;1,8[$
n_i	2	3	4	7	4
N_i	2	5	9	16	20

2.2.2 Fréquence partielle - Fréquence cumulée

Définition 2.15. Pour chaque valeur x_i, on pose par définition

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

f_i s'appelle la fréquence partiel de x_i.

Définition 2.16. Pour chaque classe I_i, on pose par définition

$$F_i = \sum_{k=1}^{k=i} f_k = f_1 + ... + f_i.$$

La fréquence cumulé F_i d'une classe est la somme de la fréquence de cette classe et de tous les fréquences des classes qui précèdent. De plus, on a

$$F_q = \sum_{k=1}^{k=q} f_k = f_1 + ... + f_q = 1$$

Exemple 2.7. On considère l'exemple 3.1 :

Table 2.5

I_i	$[1,3;1,4[$	$[1,4;1,5[$	$[1,5;1,6[$	$[1,6;1,7[$	$[1,7;1,8[$
n_i	2	3	4	7	4
f_i	$0,1$	$0,15$	$0,2$	$0,35$	$0,2$
F_i	$0,1$	$0,25$	$0,45$	$0,8$	1

2.2.3 Pourcentage partiel - Pourcentage cumulé

Définition 2.17. Pour chaque classe x_i, on pose par définition

$$p_i = \frac{n_i}{N} \times 100 = f_i \times 100\%.$$

p_i s'appelle le pourcentage partiel de x_i.

Définition 2.18. Pour chaque classe I_i, on pose par définition

$$P_i = \sum_{k=1}^{k=i} p_k = p_1 + ... + p_i.$$

Le pourcentage cumulé P_i d'une classe est la somme du pourcentage de cette classe et de tous les pourcentages des classes qui précèdent. De plus, on a

$$P_q = \sum_{k=1}^{k=p} p_k = p_1 + ... + p_q = 100\%$$

Exemple 2.8. On considère l'exemple 3.1 :

TABLE 2.6

I_i	$[1,3;1,4[$	$[1,4;1,5[$	$[1,5;1,6[$	$[1,6;1,7[$	$[1,7;1,8[$
n_i	2	3	4	7	4
f_i	$0,1$	$0,15$	$0,2$	$0,35$	$0,2$
$p_i\%$	10	15	20	35	20
$P_i\%$	10	25	45	80	100

2.2.4 Fonction de répartition

Définition 2.19. La fonction de répartition F est la fonction définie de $\mathbb{R} \to [0,1]$ par

$$F(x) = \begin{cases} 0, & \text{si } x < x_1, \\ F_i + \frac{f_i}{x_{i+1}-x_i}(x - x_i), & \text{si } x_i \leq x < x_{i+1}, \\ 1, & \text{si } x \geq x_q, \end{cases}$$

et on a $F(x_i) = F_i$

2.2.5 Paramètres de tendance centrale

Définition 2.20. Soit $X = (I_i, n_i)_{i\in\{1,2,\ldots,q\}}$ une série statistique regroupée en classes. On pose par définition :

1. La classe modale : toute classe dont le rapport $\frac{\text{l'effectif de la classe}}{\text{largeur de la classe}}$ est maximum. De plus, si $I_j = [L_j, L_{j+1}[$ est la classe modale, alors le mode est défini par la quantité
$$Mo =: L_j + (L_{j+1} - L_j)\frac{n_j - n_{j+1}}{n_j - n_{j-1} + n_j - n_{j+1}}.$$

2. Les quartiles Q_i d'ordre $\alpha = \frac{i}{4}$: la modalité de cette série pour laquelle l'effectif cumulé représente le quotient α. On a :
$$\text{si } N_{j-1} < \frac{i \times N}{4} \leq N_j \text{ alors } Q_i \in I_j = [L_j, L_{j+1}[\text{ et } Q_i = L_j + \frac{\alpha N - N_{j-1}}{N_j - N_{j-1}}(L_{j+1} - L_j)$$
. La médiane $Me = Q_2$ est la valeur qui sépare les données statistiques en deux parties égales, $[Q_1, Q_3]$ l'intervalle interquartile et le nombre $Q_3 - Q_1$ l'écart interquartile.

3. La moyenne arithmétique : c'est le nombre réel noté $E(X)$ défini par
$$E(X) = \frac{1}{N}\sum_{i=1}^{q} n_i c_i, \text{ où } N = \sum_{i=1}^{q} n_i c_i \text{le centre de la classe} I_i$$

Exemple 2.9. On considère l'exemple 3.1 :

1. La classe modale : $I_4 = [L_j = 1,6; L_{j+1} = 1,7[$
2. Le mode :
$$\begin{aligned} Mo &=: L_j + (L_{j+1} - L_j)\frac{n_j - n_{j+1}}{n_j - n_{j-1} + n_j - n_{j+1}} \\ &=: 1,6 + (1,7 - 1,6)\frac{7-4}{7-4+7-4} = 1,65. \end{aligned}$$
3. La moyenne arithmétique :
$$\begin{aligned} E(X) &= \frac{1}{N}\sum_{i=1}^{5} n_i c_i \\ &= \frac{1}{N}(2 \times 1,35 + 3 \times 1,45 + 4 \times 1,55 + 7 \times 1,65 + 4 \times 1,75 \\ &= 1,60. \end{aligned}$$
$$\overline{X} = \frac{1}{50}\sum_{i=1}^{5} n_i x_i = \frac{3 \times 0 + 5 \times 1 + 10 \times 2 + 17 \times 3 + 15 \times 4}{50} = 2,72$$
4. La médiane :
si $N_{j-1} = 2 < \dfrac{i \times N}{4} = 5 \leq N_j = 5$ alors $Q_i \in I_j = [L_j = 1,4, L_{j+1} = 1,5[$ et
$$\begin{aligned} Me = Q_2 &= L_j + \frac{\alpha N - N_{j-1}}{N_j - N_{j-1}}(L_{j+1} - L_j) \\ &= 1,4 + \frac{10-2}{5-2}(1,5 - 1,4) = 1,46. \end{aligned}$$

Remarque 2.3.

1. Une série statistique regroupée en classes peut avoir plus d'une classe modale et plus d'une médiane. De plus, si les largeurs des classes sont toutes égales, une classe modale est une classe dont l'effectif est maximum.
2. On peut caractériser la médiane par son effectif cumulé : la médiane est la plus petite modalité dont l'effectif cumulé est supérieur ou égale à la moitié de l'effectif total.
3. On peut avoir plusieurs tableaux statistiques (selon le nombre de classes) pour regrouper une série statistique en classes.
4. Il existe plusieurs relations pour déterminer une largeur ($k \in \mathbb{N}$) uniforme pour toutes les classes, on site par exemple :
   — la formule de racine :
   $$k = [\sqrt{N}] + \varepsilon \text{ où } \varepsilon = 0 \text{ ou } 1$$
   — la formule de Sturge :
   $$k = [1 + 3,3 \times \log_{10}(N)] + \varepsilon = [1 + 1,43 \times \ln(N)] + \varepsilon \text{ où } \varepsilon = 0 \text{ ou } 1$$
   — la formule de Yule
   $$k = [2,5 \times \sqrt[4]{N}] + \varepsilon \text{ où } \varepsilon = 0 \text{ ou } 1$$

2.2.6 Paramètres de dispersion

Définition 2.21. Pour chaque $X = (I_i, n_i)_{i\in\{1,2,...,q\}}$ série statistique ordonnée continue, on pose par définition

1. L'étendue : la différence entre la plus grande et la plus petite valeur observée.
2. L'écart moyen : le nombre réel positif noté $e(X)$ défini par

$$e(X) = \frac{1}{N}\sum_{i=1}^{q} n_i|c_i - E(X)|.$$

3. La variance : le nombre réel positif noté $V(X)$ défini par

$$V(X) = \frac{1}{N}\sum_{i=1}^{q} n_i(c_i - E(X))^2.$$

4. L'écart type : le nombre réel positif noté $\sigma(X)$ défini par

$$\sigma(X) = \sqrt{V(X)}.$$

Exemple 2.10. On considère l'exemple 3.1 :

1. L'étendue :

$$Etendue = E = \max_{i\in\{1,2,...,5\}} x_i - \min_{i\in\{1,2,...,5\}} x_i = 1,78 - 1,31 = 0,47.$$

2. L'écart moyen :

$$\begin{aligned} e(X) &= \frac{1}{N}\sum_{i=1}^{q} n_i|c_i - E(X)| \\ &= \frac{2\times 0,24 + 3\times 0,14 + 4\times 0,04 + 7\times 0,06 + 4\times 0,16}{20} \\ &= 0,11. \end{aligned}$$

3. La variance :

$$\begin{aligned} V(X) &= \frac{1}{N}\sum_{i=1}^{q} n_i(c_i - E(X)^2 \\ &= \frac{2\times (0,24)^2 + 3\times (0,14)^2 + 4\times (0,04)^2 + 7\times (0,06)^2 + 4\times (0,16)^2}{20} \\ &= 0,02. \end{aligned}$$

4. L'écart type :

$$\sigma(X) = \sqrt{V(X)} = 0,12.$$

Définition 2.22. Soient $a \in \mathbb{R}^*, b \in \mathbb{R}$ et $X = (I_i, n_i)_{i\in\{1,2,...,q\}}$ une série statistique ordonnée continue. On définit la série statistique ordonnée continue $aX + b$ par $aX + b = (I'_i, n'_i)_{i\in\{1,2,...,q\}}$ où $I'_i = \{ax + b/x \in I_i\}$ et $n'_i = n_i$.

Définition 2.23. Soient $a \in \mathbb{R}^*, b \in \mathbb{R}$ et $X = (I_i, n_i)_{i\in\{1,2,\ldots,q\}}$ une série statistique ordonnée continue, on définit $E(X^2)$ par

$$E(X^2) = \frac{1}{N}\sum_{i=1}^{q} n_i c_i^2, \text{ où } c_i \text{ le centre de l'intervalle } I_i$$

Proposition 2.2. *Soient $a \in \mathbb{R}^*, b \in \mathbb{R}$ et $X = (I_i, n_i)_{i\in\{1,2,\ldots,q\}}$. On a alors*

1. $V(X) = E(X^2) - E(X)^2$
2. *L'étendue de $aX + b = a\times$ l'étendue de X.*
3. *La moyenne arithmétique de $aX + b$:*

$$E(aX + b) = aE(X) + b.$$

4. *L'écart moyen de $aX + b$:*

$$e(aX + b) = ae(X).$$

5. *La variance de $aX + b$:*

$$V(aX + b) = a^2V(X).$$

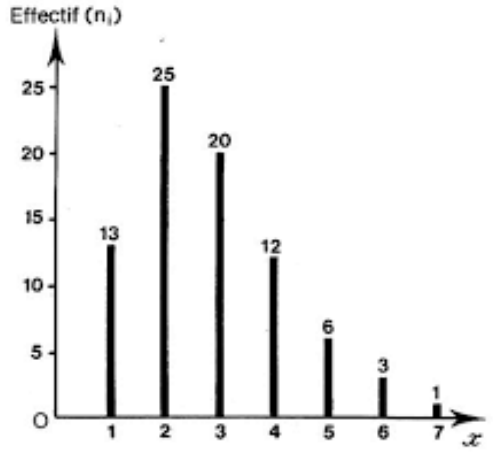
6. *L'écart type de $aX + b$:*

$$\sigma(aX + b) = |a|\sigma(X).$$

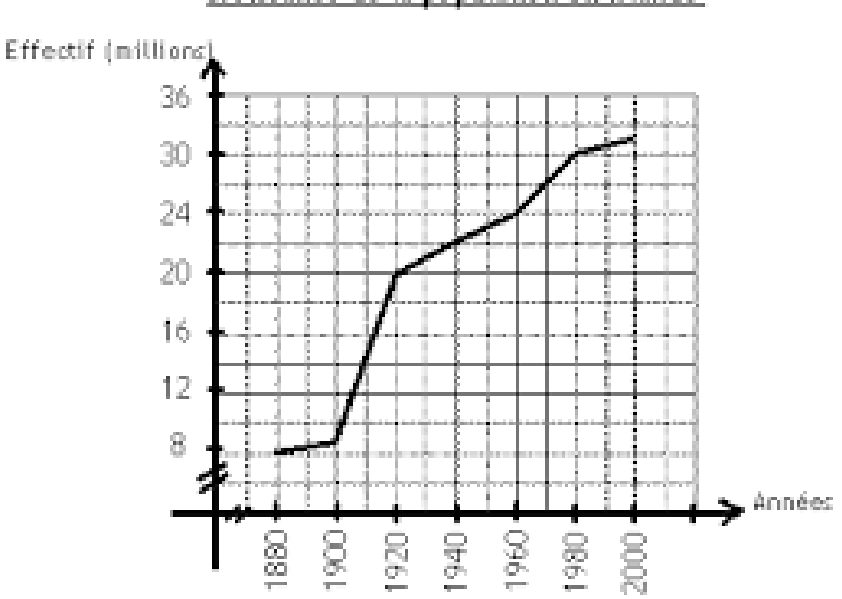
2.3 Représentation graphiques des résultats

Il est souvent souhaitable de présenter les résultats observés d'une série statistique sous forme graphique. Dans le cas des statistiques, on parlera souvent de diagramme au lieu de représentation graphique. Il existe plusieurs type de diagrammes, on site par exemple :

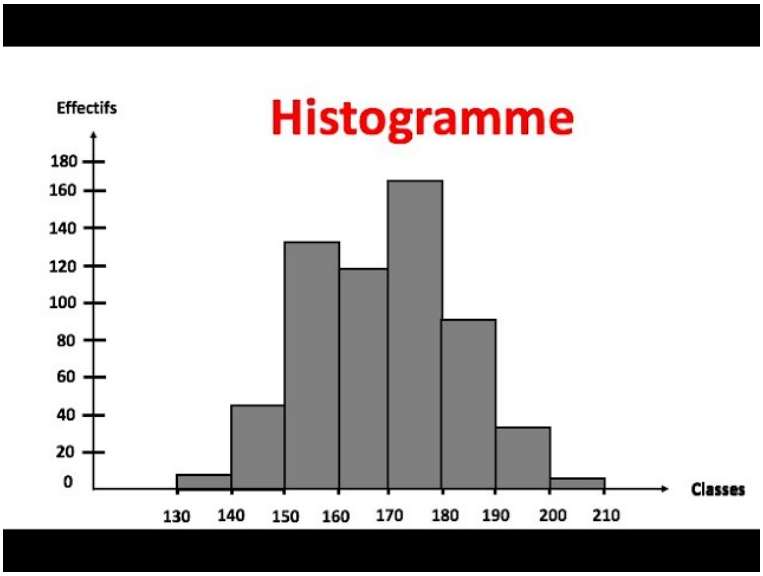
1. Diagramme en bâtons :
   Ce type est utilisé pour les séries statistiques discrètes.



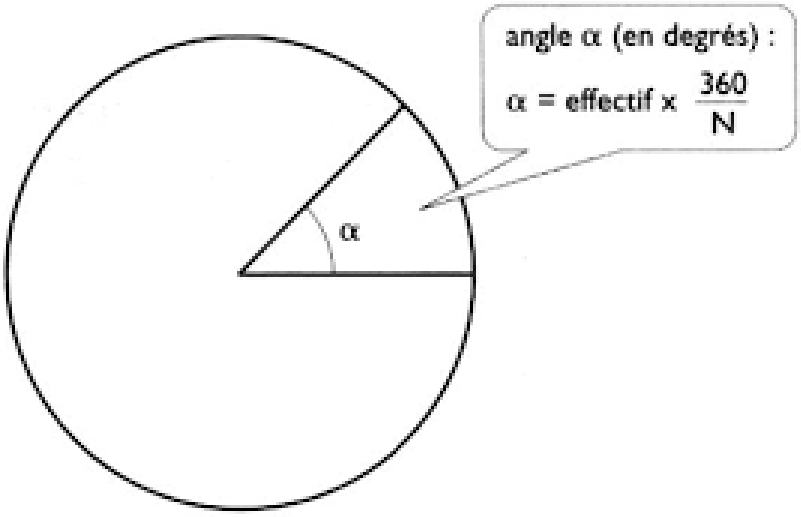
2. diagramme à lignes brisées :
   Il est souvent utilisé pour les séries statistiques discrète dépendant du temps.



3. Histogramme :
   Ce type est souvent utilisé pour les séries statistiques continues.



4. Diagramme en secteur :
   Ce type est utilisé pour les deux types des séries statistiques et les donnés sont représentés par des secteurs d'angles proportionnels aux effectifs.



Chapitre 3

Étude d'une variable statistique bivarié

Dans les chapitres précédents, 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 finie, E et F deux ensembles

3.1 Définitions

Définition 3.1. On appelle série statistique bivarié (ou double) sur $\omega.$, toute application C définie par

$$\begin{aligned} C : \Omega &\to E \times F \\ \omega &\mapsto (X(\omega), Y(\omega)). \end{aligned}$$

Exemple 3.1.

1)A chaque individu, on associe son poids (X) et sa taille (Y).
2) A chaque voiture, on associe sa marque (X) et sa couleur (Y).

Définition 3.2. Soit $C = (X, Y)$ une série statistique double, on suppose que

$$X(\omega) = \{x_1, x_2, ...x_p\} \text{ et } Y(\omega) = \{y_1, y_2, ...y_q\}$$

on pose par définition

1. L'effectif du couple (x_i, y_j) : On appele l'effectif du couple (x_i, y_j), le nombre

$$n_{ij} = Card\{\omega \in \Omega : C(\omega) = (x_i, y_j)\}.$$

De plus, on a l'effectif total

$$N = Card(\Omega) = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$$

2. L'effectif associé à la modalité : On appelle effectif associé à la modalité x_i (resp y_j) le nombre

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \text{ resp}(n_{\bullet j} = \sum_{i=1}^{p} n_{ij}).$$

la famille $(x_i, n_{i\bullet})_{i\in\{1,2,...,p\}}$ s'appelle la première série marginale et la famille $(n_{\bullet j}, y_j)_{i\in\{1,2,...,p\}}$ s'appelle la deuxième série marginale.

3. Les série statistiques :
   — On note aussi la série statistique double prècédente par

$$C = (X, Y) = ((x_i, y_j), n_{ij})_{(i,j)\in\{1,2,...,p\}\times\{1,2,...,q\}}$$

   — la famille $(x_i, n_{i\bullet})_{i\in\{1,2,...,p\}}$ s'appelle la première série marginale.
   — la famille $(n_{\bullet j}, y_j)_{i\in\{1,2,...,p\}}$ s'appelle la deuxième série marginale.
   — Pour j fixé, la famille $(x_i, n_{ij})_{i\in\{1,2,...,p\}}$ s'appelle la série statistique conditionnelle de X sachant que le second caractère Y vaut y_j , on la note $X_{/_{Y=y_j}}$
   — Pour i fixé, la famille $(y_j, n_{ij})_{j\in\{1,2,...,q\}}$ s'appelle la série statistique conditionnelle de Y sachant que le premier caractère X vaut X_i , on la note $Y_{/_{X=x_i}}$.
4. Les fréquences :
   — Le nombre $f_{ij} = \frac{n_{ij}}{N}$ est appelé la fréquence du couple (x_i, y_j) ou fréquence conjointe, de plus on a $\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = 1$
   — Les nombres $f_{i\bullet} = \frac{n_{i\bullet}}{N}$ et $f_{\bullet j} = \frac{n_{\bullet j}}{N}$ sont appellés les fréquence marginales.

Remarque 3.1. Pour déterminer les séries statistiques marginales d'une série double, il suffit donc d'ajouter une colonne à droite du tableau statistique pour placer les sommes des $n_{i\bullet}$, de même $n_{\bullet j}$ s'obtient en faisant la somme des éléments de la j^{eme} colonne, d'ou le nom de séries "marginales".

Exemple 3.2. Nous effectuons un sondage auprès de nos étudiants en leur demandant leur note de mathématique au baccalauréat et le nombre de redoublements au cours de leur scolarité primaire et secondaire. Les résultats bruts obtenus sont les suivants :

$(0; 14) - (1; 14) - (0; 14) - (2; 11) - (1; 12) - (2; 11) - (1; 13) - (2; 13) - (1; 12)$
$(0; 12) - (1; 14) - (0; 14) - (0; 11) - (1; 12) - (3; 10) - (1; 13) - (3; 13) - (0; 12)$
$(3; 14) - (0; 14) - (0; 13) - (1; 14) - (0; 13) - (0; 11) - (3; 10).$

Soit X le caractère "nombre de redoublements" et Y le caractère "note au bac". Les tableaux statistiques des série simple X , Y et de la serie statistique double regroupant les données sont :

Table 3.1

x_i	0	1	2	3
n_i	10	8	3	4

Table 3.2

y_i	10	11	12	13	14
n_i	2	4	5	6	8

Table 3.3

$X \setminus Y$	10	11	12	13	14	$n_{i\bullet}$
0	0	2	2	2	4	10
1	0	0	3	2	3	8
2	0	2	0	1	0	3
3	2	0	0	1	1	4
$n_{\bullet j}$	2	4	5	6	8	25

3.2 Paramètres de distribution marginales

On considère la série statistique double $C = ((x_i, y_j), n_{ij})_{(i,j)\in\{1,2,\ldots,p\}\times\{1,2,\ldots,q\}}$. On définit les caractéristique marginales par

1. La moyenne marginale en X :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} n_{i\bullet}x_i.$$

2. La variance marginale en X : le nombre réel positif définit par

$$V(X) = \frac{1}{N}\sum_{i=1}^{q} n_{i\bullet}(x_i - \overline{X})^2.$$

3. L'écart type marginale en en X : le nombre réel positif définit par

$$\sigma(X) = \sigma_X = \sqrt{V(X)}.$$

de meme

1. La moyenne marginale en Y :
$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{p} n_{\bullet j} y_j.$$
2. La variance marginale en Y : le nombre réel positif définit par
$$V(Y) = \frac{1}{N}\sum_{i=1}^{q} n_{\bullet j}(y_j - \overline{Y})^2.$$
3. L'écart type marginale en en X : le nombre réel positif définit par
$$\sigma(Y) = \sigma_Y = \sqrt{V(Y)}.$$

Exemple 3.3. On considère le tableau 3.3, on a :

1. La moyenne marginale en X :
$$\begin{aligned}\overline{X} &= \frac{1}{N}\sum_{i=1}^{p} n_{i\bullet} x_i \\ &= \frac{10\times 0 + 8\times 1 + 3\times 2 + 4\times 3}{25} = 1,04.\end{aligned}$$
2. La variance marginale en X :
$$\begin{aligned}V(X) &= \frac{1}{N}\sum_{i=1}^{q} n_{i\bullet}(x_i - \overline{X})^2 \\ &= \frac{10\times(0-1,04)^2 + 8\times(1-1,04)^2 + 3\times(2-1,04)^2 + 4\times(3-1,04)^2}{25} \\ &= 1,15.\end{aligned}$$
3. L'écart type marginale en en X :
$$\sigma(X) = \sigma_X = \sqrt{V(X)} = 1,07.$$

de meme

1. La moyenne marginale en Y :
$$\begin{aligned}\overline{Y} &= \frac{1}{N}\sum_{j=1}^{q} n_{\bullet j} y_j \\ &= \frac{10\times 2 + 4\times 11 + 5\times 12 + 6\times 13 + 8\times 14}{25} = 12,56.\end{aligned}$$
2. La variance marginale en X :
$$\begin{aligned}V(Y) &= \frac{1}{N}\sum_{j=1}^{q} n_{\bullet j}(y_i - \overline{Y})^2 \\ &= \frac{2\times(-2,56)^2 + 4\times(-1,56)^2 + 5\times(-0,56)^2 + 6\times(0.44)^2 + 8\times(1,44)^2}{25} \\ &= 3,65.\end{aligned}$$
3. L'écart type marginale en en X :
$$\sigma(Y) = \sigma_Y = \sqrt{V(Y)} = 1,89.$$

3.3 Paramètres de distribution conditionnelles

On considère la série statistique double $C = ((x_i, y_j), n_{ij})_{(i,j)\in\{1,2,...,p\}\times\{1,2,...,q\}}$. On définit les caractéristique marginales par

1. La fréquence conditionnelle de x_i sachant y_j :

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

2. La fréquence conditionnelle de y_j sachant x_i :

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

3. La distribution $(x_i, f_{i/j})_{i\in\{1,2,...,p\}}$ est appelée la distribution conditionnelle des fréquences de X sachant que $Y = y_j$.
4. La distribution $(y_j, f_{j/i})_{i\in\{1,2,...,q\}}$ est appelée la distribution conditionnelle des fréquences de Y sachant que $X = x_i$.

Proposition 3.1. *On considère les deux distributions* $(x_i, f_{i/j})_{i\in\{1,2,...,p\}}$ *et* $(y_j, f_{j/i})_{i\in\{1,2,...,q\}}$ *on a*

$$\forall(i,j) \in \{1,2,...,p\} \times \{1,2,...,q\}, f_{ij} = f_{i/j} \times f_{\bullet j} = f_{j/i} \times f_{i\bullet}$$

Définition 3.3. On considère les deux distributions $(x_i, f_{i/j})_{i\in\{1,2,...,p\}}$ et $(y_j, f_{j/i})_{i\in\{1,2,...,q\}}$, on dit que les caractères observés X et Y sont statistiquement indépendants si et seulement si

$$\forall(i,j) \in \{1,2,...,p\} \times \{1,2,...,q\}, f_{ij} = f_{\bullet j} \times f_{i\bullet}$$

Autrement dit

$$f_{i/j} = f_{\bullet j} = f_{i\bullet}.$$

Chapitre 4

Ajustement linéaire

Nous appelons cette démarche l'ajustement linéaire.

4.1 Covariance

Définition 4.1. On considère la série statistique double $C = ((x_i, y_j), n_{ij})_{(i,j)\in\{1,2,...,p\}\times\{1,2,...,q\}}$. On définit la covariance du couple (X, Y) et on note $cov(X, Y)$ (ou σ_{xy}) la moyenne :

$$cov(X, Y) = \sigma_{xy} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - \overline{X})(y_j - \overline{Y})$$

Proposition 4.1. *Soit (X, Y) une statistique double et soit $a, b(a \neq 0)$ deux réels. Alors on a :*

1. $cov(X, Y) = (\frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_iy_j) - (\overline{XY})$.
2. $cov(X, X) = V(X)$.
3. *La symétrie :* $cov(X, Y) = cov(Y, X)$
4. *La linéarité par rapport à X :* $cov(a.X + b, Y) = acov(Y, X)$
5. *La linéarité par rapport à Y :* $cov(X, a.Y + b) = acov(Y, X)$

Remarque 4.1. La covariance peut prendre des valeurs positives, négatives ou nulles.

4.2 Coéfficient de corrélation

Définition 4.2. On considère la série statistique double $C = ((x_i, y_j), n_{ij})_{(i,j)\in\{1,2,...,p\}\times\{1,2,...,q\}}$. On suppose que $\sigma_x \neq 0$ et $\sigma_y \neq 0$. On définit le coefficient de corrélation linéaire, le nombre réel :

$$\rho(X, Y) = \rho_{xy} = \frac{cov(X, Y)}{\sigma_x \times \sigma_x}$$

Proposition 4.2. *Soit (X, Y) une statistique double tel que $\sigma_x \neq 0$ et $\sigma_y \neq 0$. . Alors on a :*

$$-1 \leq \rho(X, Y) \leq 1.$$

Proposition 4.3. *Soit* (X, Y) *une statistique double tel que* $\sigma_x \neq 0$ *et* $\sigma_y \neq 0$. .

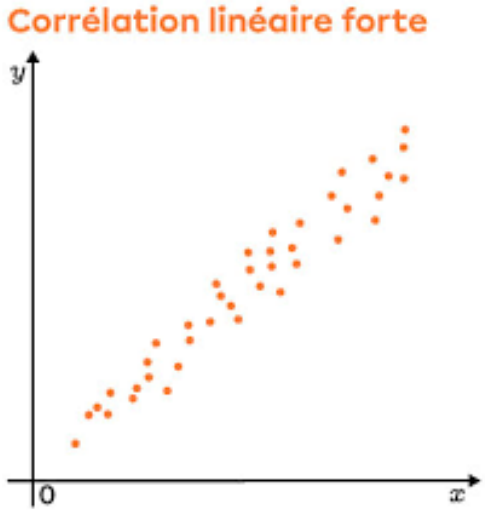
$$Si\ Y = aX + b(a \neq 0) \neq \quad Alors\ on\ a : \rho(X, Y) = \pm 1.$$

Remarque 4.2.

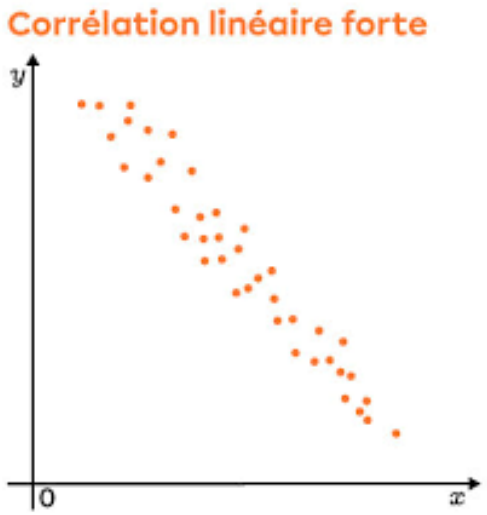
1. Le coefficient de corrélation dit de " Pearson" n'est pas sensible aux unités de chacune des variables.
2. Si $\rho(X, Y) = 1$ alors l'une des variables est fonction affine croissante de l'autre, de plus la corrélation est dite positive parfaite.
3. Si $\rho(X, Y) = -1$ alors l'une des variables est fonction affine décroissante de l'autre, de plus la corrélation est dite négative parfaite. .
4. Plus le coefficient est proche des valeurs extremes $-1et1$, plus la correlation linéaire entre les variables est forte.
5. Une correlation égale à 0 signifie que les variables sont linéairement independant.
6. Il ne faut pas croire qu'un coefficient de corrélation élevé induit une relation de causalité entre les deux caractères mesurés.

Exemple 4.1.

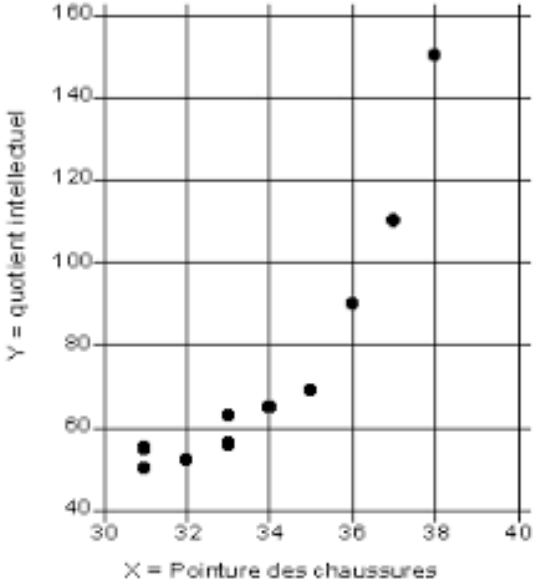
1. Forte corrélation linéaire positive :



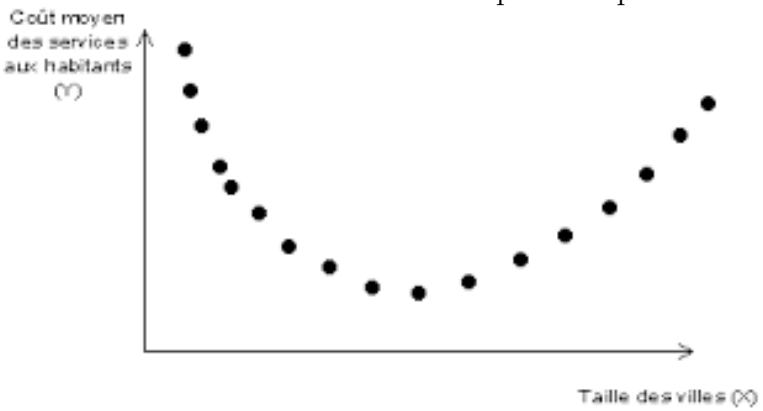
2. Forte corrélation linéaire négative :



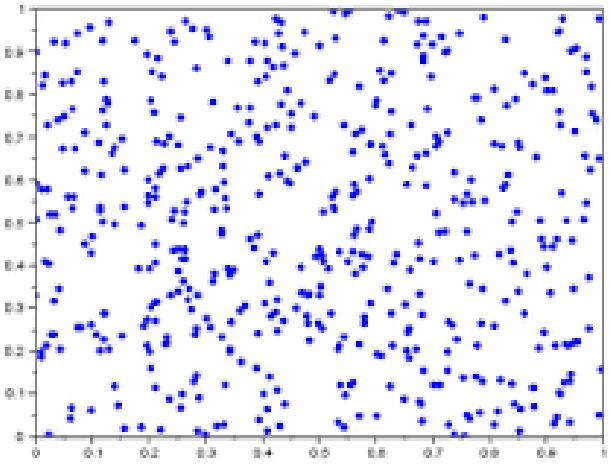
3. Faible corrélation négative :liaison exponentielle.



4. Pas de corrélation linéaire : Liaison parabolique.

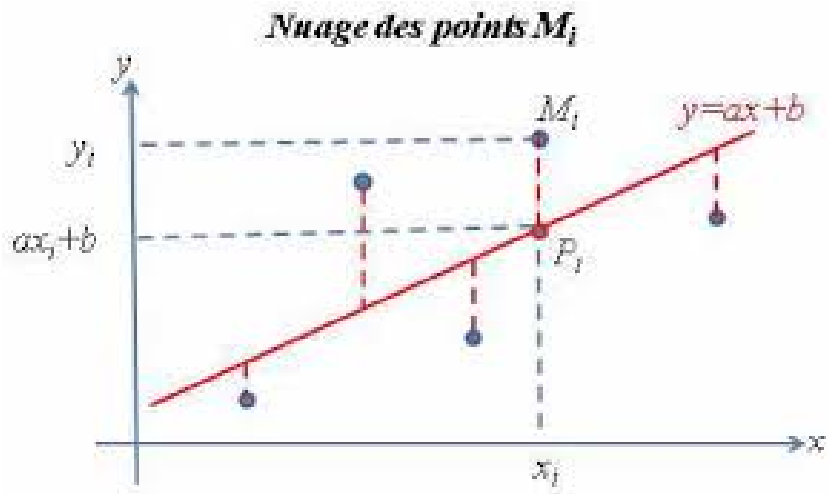


5. Pas de corrélation linéaire :variables indépendantes.



4.3 La méthode de moindres carrés

L'idée la méthode de moindres carrés 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. voir la fiqure suivante :



Pour cela, on cherche une droite $Y = aX + b$ telle que la distance entre le nuage de points et droite soit minimale,c'est à dire la différence entre le point réellement observé et le point prédit par la droite. 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. Soit $M_i(x_i, y_i)$ un point de N point du nuage, on pose

$$U(a,b) = \sum_i^n e_i^2 = \sum_i^n (y_i - ax_i - b)^2.$$

La méthode des moindres carrées consiste donc à minimiser la fonction U (la somme

des erreurs commises). Nous avons la condition de minimisation suivante,

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

On a donc,

$$\begin{cases} \frac{\partial U}{\partial a} = \sum_i^N (-2x_i)(y_i - ax_i - b), & ; \\ \frac{\partial U}{\partial b} = \sum_i^N (-2)(y_i - ax_i - b), & . \end{cases}$$

Aprés les calcules, on trouve

$$\begin{cases} \sum_i^N x_i y_i - a\sum_i^N x_i^2 - bN\overline{X} = 0, & ; \\ \sum_i^N y_i - a\sum_i^N x_i - bN = 0, & . \end{cases}$$

En multipliant la $1^{`ere}$ équation par $\frac{1}{N}$ et la $2^{`eme}$ équation par $-\frac{1}{N}\overline{X}$, nous obtenons le système suivant :

$$\begin{cases} \frac{1}{N}\sum_i^N x_i y_i - a\frac{1}{N}\sum_i^N x_i^2 - b\overline{X} = 0 & ; \\ -\overline{X}\ \overline{Y} + a\overline{X}^2 - b\overline{X} = 0 & . \end{cases}$$

En faisant la somme terme à terme des deux équations (1) et (2), on obtient

$$Cov(X,Y) - aV(X) = 0$$

ce qui donne

$$a = \frac{Cov(X,Y)}{V(X)}, \text{ et } b = \overline{Y} - a\overline{X}.$$

de plus on a

$$\min_{(a,b)\in\mathbb{R}^2} U(a,b) = N\frac{V(X)V(Y) - Cov(X,Y)^2}{V(X)} \geq 0.$$

4.4 La droite de régression

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y, on peut chercher à formaliser la relation moyenne entre ces deux variables à l'aide d'une équation de droite qui résume cette relation, Les calculs et les résultats précédents se généralisent aisément au cas où la statistique double $C = (X,Y) = ((x_i, y_j), n_{ij})_{(i,j)\in\{1,2,\ldots,p\}\times\{1,2,\ldots,q\}}$ n'est pas injective. on définit cette droite appellée droite de regression par

Définition 4.3.

1. La droite d'équation $(D_{Y/X})$: $y = ax+b$ avec $a = \frac{Cov(X,Y)}{V(X)}$, et $b = \overline{Y} - a\overline{X}$, s'appelle droite de régression de Y dans X.
2. La droite d'équation $(D_{X/Y})$: $x = \alpha x + \beta$ avec $\alpha = \frac{Cov(X,Y)}{V(Y)}$, et $\beta = \overline{X} - \alpha\overline{Y}$, s'appelle droite de régression de X dans Y.

Proposition 4.4. *Les deux droites de régression $(D_{Y/X})$ et $(D_{X/Y})$ passent par le même point de coordonnées $(\overline{X}, \overline{Y})$*

Remarque 4.3. On admet generalement qu'un coefficient de correlation $|\rho(X,Y)| \geq 0.75$ justifie la recherche d'un alignement statistique, en l'absence de renseignements complementaires.