LICENCE Scientifique

Cours – Henri IMMEDIATO

Statistiques

STATISTIQUE

Chapitre I - GENERALITES.

I. 1. OBJET DE LA STATISTIQUE

Le but de la statistique est de dégager les significations de données, numériques ou non, obtenues au cours de l'étude d'un phénomène.
Il faut distinguer les **données statistiques** qui sont les résultats d'observations recueillies lors de l'étude d'un phénomène, et la **méthode statistique** qui a pour objet l'étude rationnelle des données.
La méthode statistique comporte plusieurs étapes.

I. 1. 1. La statistique descriptive ou déductive.

C'est l'ensemble des méthodes à partir desquelles on recueille, ordonne, réduit, et condense les données.
A cette fin, la statistique descriptive utilise des paramètres, ou synthétiseurs, des graphiques et des méthodes dites d'analyse des données (l'ordinateur a facilité le développement de ces méthodes).

I. 1. 2. La statistique mathématique ou inductive

C'est l'ensemble des méthodes qui permettent de faire des **prévisions**, des interpolations sur une population à partir des résultats recueillis sur un échantillon.
Nous utilisons des raisonnements **inductifs** c'est-à-dire des raisonnements de passage du particulier au général.
Cette statistique utilise des repères de référence qui sont les **modèles théoriques** (lois de probabilités).
Cette statistique nécessite la recherche d'échantillons qui représentent le mieux possible la diversité de la population entière ; il est nécessaire qu'ils soient constitués au hasard ; on dit qu'ils résultent d'un **tirage non exhaustif**.
L'étude sur échantillon se justifie pour réduire le coût élevé et limiter la destruction d'individus pour obtenir la réponse statistique.

I. 2. VOCABULAIRE STATISTIQUE

I. 2. 1. Population

C'est l'ensemble des unités ou individus sur lequel on effectue une analyse statistique.

$\Omega = \{\omega_1, \ldots, \omega_N\}$ avec $\text{card}(\Omega) = N$ fini

Ce vocabulaire est hérité du 1er champ d'application de la statistique : la démographie (Vauban (1633-1707) effectua des recensements pour des études économiques et militaires).

Exemples de populations.

Les véhicules automobiles immatriculés en France
La population des P.M.E. d'un pays
Les salariés d'une entreprise
Les habitants d'un quartier

I. 2. 2. Echantillon

C'est un ensemble d'individus prélevés dans une population déterminée

Exemple d'échantillon.

L'échantillon des véhicules automobiles immatriculés dans un département.

I. 2. 3. Caractère

C'est un trait déterminé *C* présent chez tous les individus d'une population sur laquelle on effectue une étude statistique.

- Un caractère est dit **quantitatif** s'il est mesurable.

Exemples de caractères quantitatifs.

La puissance fiscale d'un véhicule automobile.
Le chiffre d'affaire d'une P.M.E.
L'âge, le salaire des salariés d'une entreprise.

- Un caractère est dit **qualitatif** s'il est repérable sans être mesurable.

Exemples de caractères qualitatifs.

La couleur de la carrosserie d'un véhicule automobile
Le lieu de travail des habitants d'un quartier
Le sexe et la situation matrimoniale des salariés d'une entreprise

I. 2. 4. Modalités

Ce sont les différentes situations M_i possibles du caractère.

Les modalités d'un caractère doivent être **incompatibles** et **exhaustives** ; tout individu doit présenter une et une seule modalité.
Les modalités d'un caractère qualitatif sont les différentes rubriques d'une nomenclature ; celles d'un caractère quantitatif sont les mesures de ce caractère.
L'ensemble des modalités est noté *E*.

Pour un caractère quantitatif, la mesure du caractère peut être un nombre entier pris parmi un ensemble limité ; nous dirons qu'il est **discret**.

Exemple de caractère quantitatif discret.

Le nombre d'enfants d'une famille (fratrie)

Dans certains cas la mesure du caractère peut être un nombre décimal pris parmi un ensemble de valeurs possibles très important (plusieurs dizaines ou plusieurs centaines).
Pour permettre une étude et notamment une représentation graphique plus simple, nous sommes conduits à effectuer un regroupement en classes (5 à 20 classes) ; nous dirons alors que le caractère est **continu**.

Dans ces deux situations, nous dirons que le caractère quantitatif est défini par ses modalités (valeurs discrètes ou classes).

Les modalités d'un caractère quantitatif peuvent être prises dans $\mathbb{R}$ ou $\mathbb{R}^n$.

Exemples d'ensembles de modalités.

Nombre d'enfants dans une fratrie : $\{M_i\} = \{x_i\}=\{0, 1, 2, 3, ...\}$, M_i ? $\mathbb{N}$.
L'âge, la taille et le poids d'un groupe d'individus représentent globalement une modalité définie dans $\mathbb{R}^3$ (à condition que chacune de ces variables soit discrète)

L'ensemble des modalités d'un caractère peut être établi à priori avant l'enquête (une liste, une nomenclature, un code) ou après enquête.
On constitue l'ensemble des valeurs prises par le caractère.

Les caractères étudiés sur une population peuvent être **mixtes** :

Exemple de caractère mixte.

L'ensemble des salariés d'une entreprise peut être représenté par un caractère mixte que nous pourrons exploiter globalement ou plus efficacement en extrayant une partie des données.
Le sexe, de modalités : *H* ou *F* (codé par 1 ou 2)
L'âge, de modalités : 18, 19, 20, ... ou [16, 20], [21, 25], ...
Le salaire mensuel, de modalités : 6000, 6500, 7000, ... ou [6000, 6500[, [6500, 7500[, ...
La situation matrimoniale, de modalités : marié, célibataire, veuf, divorcé, vivant maritalement.

I. 3. NOTION DE DISTRIBUTION STATISTIQUE

Considérons une population ? = {? $_1$, ... , ? $_N$}.
Dans cette population, considérons un caractère *C* et soit *E* l'ensemble des modalités du caractère *C*, *card* (*E*) = *p*.
On note A_i l'ensemble des individus de ? présentant la modalité M_i du caractère *C*, *i* = 1, ... , *p*.

Les A_i forment une partition de ? : $A_i \cap A_j = \emptyset$ pour *i* ? *j*, et $\bigcup_{i=1}^{i=p} A_i = ?$.

Nous définissons $n_i = card\,(A_i)$.
n_i est l'**effectif** de la modalité M_i.

On appelle **variable statistique** toute application *X* de ? dans *E* qui, à chaque individu ? de la population, associe une modalité M_i du caractère *C*.
L'effectif n_i d'une modalité M_i est le cardinal de l'image réciproque A_i de M_i par *X* :

$n_i = card\ (A_i) = Card\ (X^{-1}\ (M_i))$

Une variable statistique s'identifie à l'ensemble des triplets $\{(M_i, A_i, n_i)\}$, i ? [1, p].

En pratique, le statisticien se contente souvent de l'ensemble des doublets $\{(M_i, n_i)\}$, i ? [1, p], sans se préoccuper de savoir qui sont les n_i individus de la population présentant la modalité M_i du caractère C et constituant l'ensemble A_i.

On appelle aussi **distribution statistique** l'ensemble des doublets $\{(M_i, n_i)\}$, i ? [1, p].

Exemples de variables statistiques.

Le nombre d'enfants d'une fratrie : $x_1 = 0$, $n_1 = 50$; $x_2 = 1$, $n_2 = 70$; $x_3 = 2$, $n_3 = 20$.

La taille d'une population : $M_1 =$ [150, 160 [, $n_1 = 50$; $M_2 =$ [160, 175 [, $n_2 = 100$.

Les marques de véhicules automobiles : M_1 = "Renault", $n_1 = 15\ 000$; M_2 = "Citroën", $n_2 = 10\ 000$

La **fréquence** de la modalité M_i est, par définition : $f\ (A_i) = \frac{n_i}{N} = f_i$, $N = \sum_{i=1}^{i=p} n_i$.

La notion d'effectif d'une modalité est une notion absolue, elle ne permet pas directement les comparaisons.

La notion de fréquence est une notion relative, elle permet directement les comparaisons.

Remarque.

Si le caractère C ne présente qu'une modalité a dans la population, on parle de **variable, ou de distribution, statistique constante** $\{(a, ?, N)\}$.

Chapitre II - ANALYSE UNIVARIEE.

(Statistique descriptive à un caractère)

II. 1. REPRESENTATION GRAPHIQUE

La représentation graphique des données relatives à un caractère unique repose sur la proportionnalité des longueurs, ou des aires, des graphiques, aux effectifs, ou aux fréquences, des différentes modalités du caractère.

II. 1. 1. Caractère qualitatif.

Pour un caractère qualitatif, on utilise principalement trois types de représentation graphique : le **diagramme en bâtons**, la représentation par **tuyaux d'orgue** et la représentation par **secteurs**. Lorsque le caractère étudié est la répartition géographique d'une population, la représentation graphique est un **cartogramme**.

a) Diagramme en bâtons.

Nous portons en abscisse les modalités, de façon arbitraire.
Nous portons en ordonnée des segments dont la **longueur** est proportionnelle aux effectifs (ou aux fréquences) de chaque modalité.
Nous appelons **polygone statistique**, ou **diagramme polygonal**, la ligne obtenue en joignant les sommets des bâtons.

b) 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.

c) Secteurs.

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.
Ces diagrammes conviennent très bien pour des données politiques ou socio-économiques.

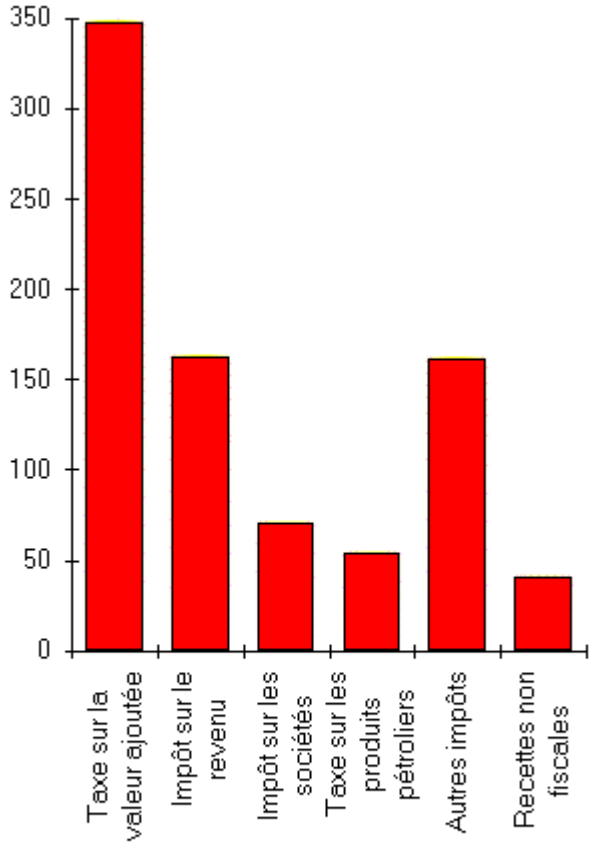
d) Exemple.

En 1982, les recettes du budget de l'Etat se présentaient de la façon suivante (en milliards de francs) :

Taxe sur la valeur ajoutée	348
Impôt sur le revenu	163
Impôt sur les sociétés	71
Taxe sur les produits pétroliers	54
Autres impôts	161
Recettes non fiscales	41
TOTAL	**838**

Le caractère étudié, la nature des recettes du budget de l'Etat, est un caractère qualitatif.

Dans la représentation en **tuyaux d'orgue**, les différentes modalités du caractère (les diverses sources de recettes du budget de l'Etat) sont représentées par des segments sur l'axe des ordonnées. Pour chaque abscisse on porte un rectangle dont la longueur est proportionnelle au montant correspondant de la recette (effectif).



Histogramme des effectifs

Dans la représentation par **diagramme en bâtons**, les différentes modalités du caractère (les diverses sources de recettes du budget de l'Etat) sont représentées par des points sur l'axe des ordonnées. Pour chaque abscisse, on porte un segment vertical dont la longueur est proportionnelle au montant correspondant de la recette (rectangle de largeur nulle).

Dans le **diagramme circulaire**, chaque secteur a une surface proportionnelle à l'importance de la recette dans le budget. L'angle au centre représentant une modalité est donc proportionnelle à l'importance de la recette dans le budget.

Source de recettes	Recette		Angle (degrés)	Angle cumulé	
	MF	%		$\widehat{xOy}$	$\widehat{xOy}_{-180°}$
Taxe sur la valeur ajoutée	348	41,5	149,5	149,5	
Impôt sur le revenu	163	19,5	70,0	219,5	39,5
Impôt sur les sociétés	71	8,5	30,5	250,0	70,0
Taxe sur les produits pétroliers	54	6,4	23,2	273,2	93,2
Autres impôts	161	19,2	69,2	342,4	162,4
Recettes non fiscales	41	4,9	17,6	360,0	180,0
TOTAL	**838**	100			

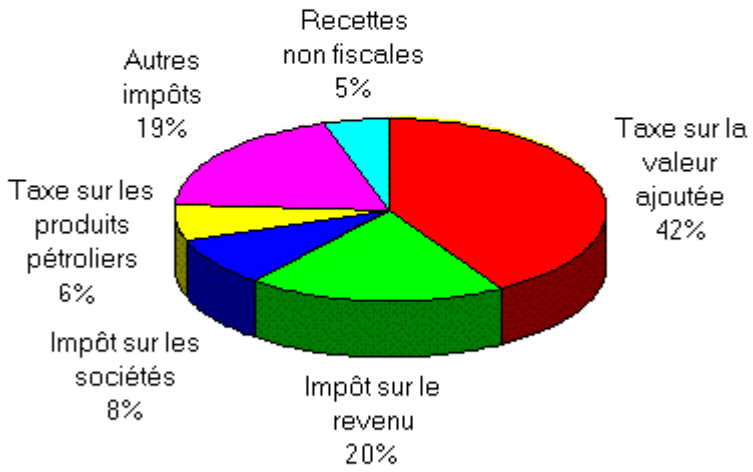


Diagramme circulaire des fréquences

e) Cartogrammes.

Un cartogramme est une carte géographique dont les secteurs géographiques sont coloriés avec une couleur différente suivant l'effectif ou suivant la fréquence du caractère étudié.

II. 1. 2. Caractère quantitatif.

La variable statistique est la mesure du caractère.
Celle-ci peut être discrète ou continue.
Il existe deux types de représentation graphique d'une distribution statistique à caractère quantitatif :

— Le **diagramme différentiel** correspond à une représentation des effectifs ou des fréquences.

— Le **diagramme intégral** correspond à une représentation des effectifs cumulés, ou des fréquences cumulées.

a) Variable statistique discrète.

— Diagramme différentiel : diagramme en bâtons, des effectifs ou des fréquences.

La différence avec le cas qualitatif consiste en ce que les abscisses ici sont les valeurs de la variable statistique.

— Diagramme intégral : courbe en escaliers des effectifs cumulés ou des fréquences cumulées.

Exemple.

En vue d'établir rationnellement le nombre de postes de travail nécessaires pour assurer à sa clientèle un service satisfaisant, une agence de voyage a fait relever, minute par minute, le nombre d'appels téléphoniques reçus au cours d'une période de 30 jours. Cette opération a fourni, pour la tranche horaire de pointe qui se situe entre onze heures et midi, les résultats suivants :

Nombre d'appels téléphoniques par minute	Nombre de minutes
0	93
1	261
2	416
3	393
4	308
5	174
6	93
7	42
8 et plus	20
TOTAL	1 800

La population étudiée est celle des 1 800 minutes composant la durée totale des appels dans la tranche horaire de onze heures à midi pendant 30 jours.
Le caractère observé est le nombre d'appels téléphoniques : c'est un caractère quantitatif et la variable statistique correspondante, qui ne peut prendre que des valeurs entières, est discrète.
La représentation des effectifs est identique à celle des fréquences : seule change l'échelle verticale.

La représentation graphique **différentielle** correcte est le **diagramme en bâtons**.
A chaque valeur x_i de la variable, portée en abscisse, on fait correspondre un segment vertical de longueur proportionnelle à la fréquence f_i de cette valeur.

Le regroupement des valeurs extrêmes de la variable en une seule classe (nombre d'appels supérieur ou égal à 8) interdit normalement la représentation graphique de ce dernier segment.
Mais, étant donnée la fréquence quasi négligeable de cette classe, l'inconvénient n'est pas bien grand et l'on pourra représenter par un segment à l'abscisse 8, la fréquence des appels de durée 8 ou plus.

Nombre d'appels téléphoniques par minute	Nombre de minutes (effectifs)	Fréquence (%)	Fréquence cumulée (%)
0	93	5,2	5,2
1	261	14,5	19,7
2	416	23,1	42,8
3	393	21,8	64,6
4	308	17,1	81,7
5	174	9,7	91,4
6	93	5,2	96,6
7	42	2,3	98,9
8 et plus	20	1,1	100,0
TOTAL	1 800	100,0	

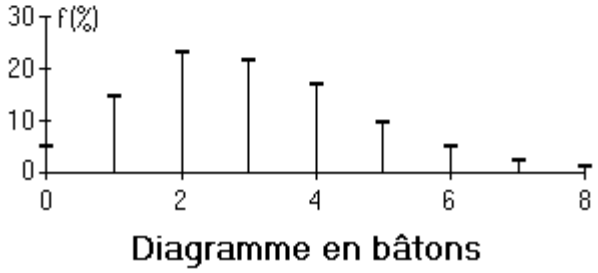
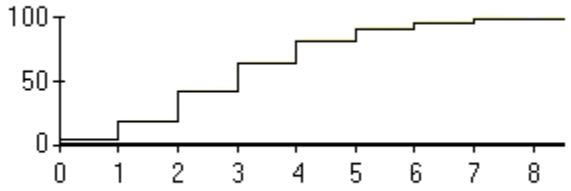


Diagramme en bâtons

La représentation graphique **intégrale** correcte est la **courbe en escalier** : les fréquences des diverses valeurs de la variable statistique correspondent aux hauteurs des marches de la courbe en escalier.



Courbe en escalier

b) Variable statistique continue.

Les observations sont regroupées en classes.
Chaque classe possède une certaine **amplitude**, qui est la longueur de l'intervalle définissant la classe.
Le rapport entre l'effectif d'une classe et son amplitude s'appelle la **densité d'effectif**.
Le rapport entre la fréquence d'une classe et son amplitude s'appelle la **densité de fréquence**.

— Diagramme différentiel : **histogramme des densités**.

Nous portons en abscisse les classes représentant les modalités et en ordonnées des rectangles dont la longueur est proportionnelle à la densité d'effectif ou à la densité de fréquence.
L'**aire** d'un rectangle de cet histogramme est alors proportionnelle à l'effectif ou à la fréquence de la classe.

— Diagramme intégral : **courbe cumulative** des effectifs ou des fréquences.
La courbe cumulative des fréquences doit représenter la fonction de répartition de la variable statistique.

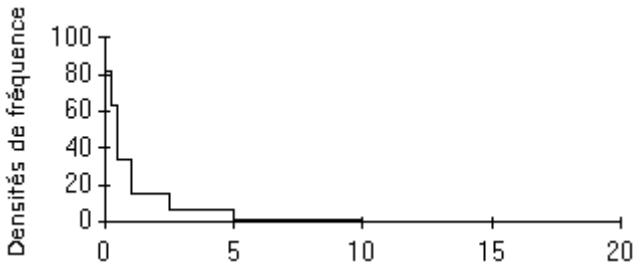
Exemple.

La Fédération nationale de la réparation et du commerce de l'automobile a effectué une enquête auprès de ses adhérents visant à mieux connaître la structure de ce secteur. Cette opération a fourni la répartition suivante des entreprises de la réparation de du commerce de l'automobile selon leur chiffre d'affaires annuel.
La masse de chiffres d'affaires correspondant aux entreprises de la première et de la dernière classes s'élève respectivement à 1 714 et 110 145 millions de francs.

Chiffre d'affaires (millions de F)	Nombre d'entreprises
Moins de 0,25	13 712
0,25 à moins de 0,50	10 674
0,50 à moins de 1,00	11 221
1,00 à moins de 2,50	15 496
2,50 à moins de 5,00	10 043
5,00 à moins de 10,00	3 347
10,00 et plus	3 147
TOTAL	67 640

La population étudiée est celle des entreprises de la réparation et du commerce de l'automobile.
Le caractère observé est le chiffre d'affaires.
C'est un caractère quantitatif et la variable statistique correspondante est continue.

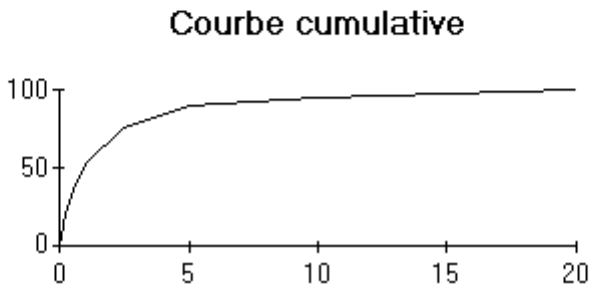
La représentation graphique différentielle correcte est l'**histogramme des densités de fréquences**.
Pour la première et la dernière classes, l'amplitude de la classe n'est pas connue.
On détermine alors la moyenne de la classe, qu'on considère comme la valeur centrale de la classe (quand on construit un histogramme, on fait l'hypothèse implicite que les effectifs sont répartis uniformément à l'intérieur de la classe, la moyenne de la classe est alors le centre de la classe).
Pour la première classe, la moyenne du chiffre d'affaires est $\frac{1714}{13712}$ = 0,125, de sorte que la première classe est la classe [0,00 , 0,25 [.
Pour la dernière classe, la moyenne du chiffre d'affaires est $\frac{110145}{3147}$ = 35, de sorte que la dernière classe est la classe [10,00 , 60,00 [.



La représentation graphique intégrale correcte est la **courbe cumulative des fréquences**.
Pour que chaque point expérimental représente la fonction de répartition, il faut prendre pour abscisses les **limites supérieures des classes** et, pour ordonnées, les **fréquences cumulées** correspondantes.

Comme la variable statistique est continue, on tracera une courbe cumulative continue, et non une courbe en escalier, de façon qu'à une valeur de fréquence cumulée corresponde une et une seule valeur de variable.
Entre deux points expérimentaux, on trace un segment de droite représentant l'interpolation linéaire, ou bien une courbe lissée, asymptotiquement tangente à l'horizontale d'ordonnée 100.

Chiffre d'affaires (millions de F)	Nombre d'entreprises (effectif)	Fréquence (%)	Amplitude de classe	Densité de fréquence	Fréquence cumulée
Moins de 0,25	13 712	20,3	0,25	81,2	20,3
0,25 à moins de 0,50	10 674	15,8	0,25	63,2	36,1
0,50 à moins de 1,00	11 221	16,6	0,50	33,2	52,7
1,00 à moins de 2,50	15 496	22,9	1,50	15,3	75,6
2,50 à moins de 5,00	10 043	14,8	2,50	5,9	90,4
5,00 à moins de 10,00	3 347	4,9	5,00	0,98	95,3
10,00 et plus	3 147	4,7	50,0	0,094	100,0
TOTAL	67 640				



II. 2. PARAMETRES CARACTERISTIQUES

Le but de l'étude statistique est aussi de résumer des données par des paramètres ou synthétiseurs.
Il existe 3 types de paramètres :
— paramètres de position (ou de tendance centrale)
— paramètres de dispersion
— paramètres de forme (asymétrie, aplatissement, concentration)

II. 2. 1. Paramètres de position

Les paramètres de position (mode, médiane, moyenne) permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.

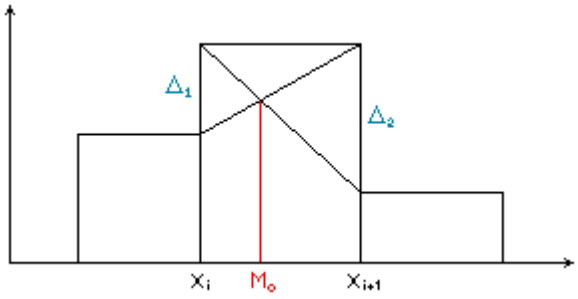
II. 2. 1. 1. Le mode

Le mode, noté M_o, est la modalité qui admet **la plus grande fréquence** :

$$f(M_o) = Max(f_i) \; ; \; i \in [\,1, p\,]$$

Il est parfaitement défini pour une variable qualitative ou une variable quantitative discrète.

Pour une variable quantitative continue nous parlons de **classe modale** : c'est la classe dont la densité de fréquence est maximum.
Si les classes ont même amplitude la densité est remplacée par l'effectif ou la fréquence et nous retrouvons la définition précédente.
Nous définissons le **mode**, pour une variable quantitative continue, en tenant compte des densités de fréquence des 2 classes adjacentes par la méthode suivante.



La classe modale $[\,x_i, x_{i+1}\,[$ étant déterminée, le mode M_o vérifie :

$$\frac{M_o - x_i}{\Delta_1} = \frac{x_{i+1} - M_o}{\Delta_2}$$

Dans une proportion, on ne change pas la valeur du rapport en additionnant les numérateurs et en additionnant les dénominateurs :

$$\frac{M_o - x_i}{\Delta_1} = \frac{x_{i+1} - M_o}{\Delta_2} = \frac{x_{i+1} - x_i}{\Delta_1 + \Delta_2}$$

$$\boxed{M_o = x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(x_{i+1} - x_i).}$$

Remarques.

Lorsque les classes adjacentes à la classe modale ont des densités de fréquences égales, le mode coïncide avec le centre de la classe modale.
Le mode dépend beaucoup de la répartition en classes.
Une variable statistique peut présenter plusieurs modes locaux : on dit alors qu'elle est **plurimodale**.
Cette situation est intéressante : elle met en évidence l'existence de plusieurs sous-populations, donc l'hétérogénéité de la population étudiée.

II. 2. 1. 2. La médiane

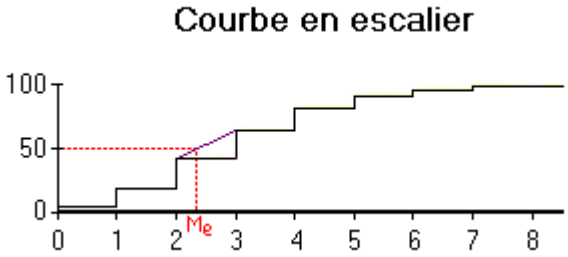
La médiane M_e est telle que l'effectif des observations dont les modalités sont inférieures à M_e est égal à l'effectif des observations dont les modalités sont supérieures à M_e.

Cette définition n'a de sens que si les modalités sont toutes ordonnées.
Dans le cas d'une variable qualitative il est parfois possible de choisir un ordre.
Exemple : niveau d'études scolaires : école primaire < 1er cycle < CAP < BEP < Bac < BTS < DEUG < ....
Une variable quantitative X doit être définie dans $\mathbb{R}$.

Détermination pratique de la médiane.

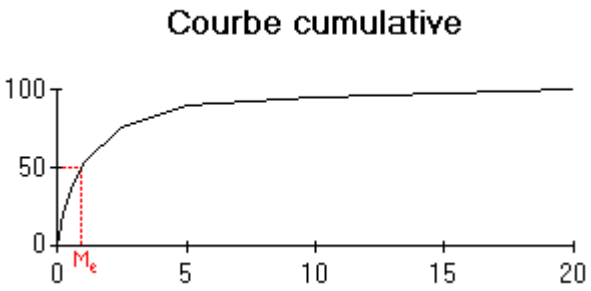
· Cas d'une variable discrète.

Reprenons l'exemple de II.1.2.a de variable discrète (appels téléphoniques).
La fréquence cumulée est 42,8 % pour $x = 2$, et 64,6 % pour $x = 3$.
L'intervalle [2, 3 [est appelé **intervalle médian**.
Dans l'intervalle médian, la médiane est calculée par **interpolation linéaire**.



· Cas d'une variable continue :

Reprenons l'exemple de II.1.2.b de variable continue (entreprises automobiles).
La fréquence cumulée est 36,1 % pour $x = 0{,}50$, et 52,7 % pour $x = 1{,}00$.
L'intervalle [0,50, 1,00 [est l'**intervalle médian**.
Dans l'intervalle médian, la médiane est calculée par **interpolation linéaire**.



Remarques

La médiane ne dépend que de l'ordre des modalités , elle n'est donc pas influencée par les observations aberrantes.
La médiane partage l'histogramme des fréquences en 2 parties d'aires égales.

II. 2. 1. 3. La moyenne

La moyenne $\overline{X}$ ne se définit que pour une variable statistique quantitative.

Pour une variable statistique discrète $\{(x_i, n_i)\}_{1 \le i \le p}$ à valeurs dans $\mathbb{R}$, la moyenne $\overline{X}$ est la moyenne arithmétique des modalités pondérées par les effectifs :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{i=p} n_i x_i = \frac{1}{N}\sum_{\omega \in \Omega} X(\omega), \text{ avec } N = \sum_{i=1}^{i=p} n_i.$$

Pour une variable statistique discrète $\{((x_{ij})_{1 \le j \le q}, n_i)\}_{1 \le i \le p}$ à valeurs dans $\mathbb{R}^q$, la moyenne $\overline{X}$ est encore la moyenne arithmétique des modalités dans $\mathbb{R}^q$, pondérées par les effectifs :

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{i=p} n_i \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iq} \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{i=p} n_i x_{i1} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{i=p} n_i x_{iq} \end{pmatrix} = \begin{pmatrix} \overline{X_1} \\ \vdots \\ \overline{X_q} \end{pmatrix}.$$

$\overline{X}$ est le "point moyen" qui résume le nuage de points de $\mathbb{R}^q$.
Il caractérise un individu moyen représentatif du nuage de données.

Exemple.

L'étude de 21 familles a conduit à la distribution suivante suivante le nombre d'enfants dans la famille :

Nombre d'enfants x_i	0	1	2	3	4	5
Nombre de familles n_i	5	3	6	1	3	3

Le nombre moyen d'enfants par famille est $\overline{X} = \frac{1}{N}\sum_{i=1}^{i=p} n_i x_i = \frac{1}{21}(0 \times 5 + 1 \times 3 + 2 \times 6 + 3 \times 1 + 4 \times$

$3 + 5 \times 3) = \frac{45}{21} = \frac{15}{7}$.

Naturellement, cette moyenne ne représente pas une "famille moyenne" mais donne une estimation du nombre d'enfants dans une famille dont est extrait l'échantillon : nous pourrons dire que, dans cette population, il faudra, en moyenne, 7 familles pour avoir 15 enfants, ou que 100 familles auront, en moyenne, 214 enfants.

a) Propriétés de la moyenne.

Somme.

La somme $X + Y$ de deux variables statistiques X et Y est définie par :

$$(X + Y)(\omega) = X(\omega) + Y(\omega), \text{ pour tout } \omega \in \Omega.$$

Nous avons alors écrire :

$$\overline{X+Y} = \frac{1}{N}\sum_{\omega\in\Omega}(X+Y)(\omega) = \frac{1}{N}\sum_{\omega\in\Omega}(X(\omega)+Y(\omega)) = \frac{1}{N}\sum_{\omega\in\Omega}X(\omega) + \frac{1}{N}\sum_{\omega\in\Omega}Y(\omega) = \overline{X} + \overline{Y}$$

$$\boxed{\overline{X+Y} = \overline{X} + \overline{Y}}$$

Produit par un scalaire

Le produit λX d'une variable statistique X par un nombre réel λ est défini par :

$$(\lambda X)(\omega) = \lambda X(\omega), \text{ pour tout } \omega \in \Omega.$$

Nous pouvons alors écrire :

$$\overline{\lambda X} = \frac{1}{N}\sum_{\omega\in\Omega}(\lambda X)(\omega) = \frac{\lambda}{N}\sum_{\omega\in\Omega}X(\omega) = \lambda\,\overline{X}.$$

$$\boxed{\overline{\lambda X} = \lambda\,\overline{X}.}$$

Ecart moyen à la moyenne.

$$\overline{X - \overline{X}} = \frac{1}{N}\sum_{\omega\in\Omega}(X - \overline{X})(\omega) = \frac{1}{N}\sum_{\omega\in\Omega}(X(\omega) - \overline{X}) = \frac{1}{N}\sum_{\omega\in\Omega}X(\omega) - \overline{X} = 0$$

$$\boxed{\overline{X - \overline{X}} = 0}$$

b) Moyenne conditionnée.

Soit Ω^* une sous-population de Ω (exemple : nombre d'enfants d'une fratrie d'origine étrangère dans une population donnée).

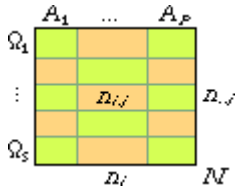
Soit X^* la restriction à Ω^*.d'une variable statistique $X = \{(x_i, A_i, n_i)\}$, $i \in [\,1, p\,]$, sur Ω.

On pose : $A_i^* = A_i \cap \Omega^*$, $n_i^* = Card\,(A_i^*) = Card\,(A_i \cap \Omega^*)$, $n^* = Card\,(\Omega^*)$.

$X^* = \{(x_i, A_i^*, n_i^*)\}$, $i \in [\,1, p\,]$.

X^* est une variable statistique sur Ω^*.

Sa moyenne est $\overline{X^*} = \frac{1}{n^*} \sum_{i=1}^{i=p} n_i^* x_i = \frac{1}{n^*} \sum_{\omega \in \Omega^*} X^*(\omega) = \frac{1}{n^*} \sum_{\omega \in \Omega^*} X(\omega)$.



Considérons maintenant une partition de Ω en s sous-populations $\Omega_1, \dots, \Omega_s$.

Soit $X = \{(x_i, A_i, n_i)\}$, $i \in [\,1, p\,]$, une variable statistique sur Ω.

Chaque sous-population Ω_j, $j \in [\,1, s\,]$, définit une variable statistique X_j sur Ω_j, qui est la restriction de X à Ω_j.

On pose $n_{ij} = Card(A_i \cap \Omega_j)$, $n_{.j} = Card(\Omega_j) = \sum_{i=1}^{i=p} n_{ij}$, $j \in [\,1, s\,]$.

On a $n_i = Card(A_i) = \sum_{j=1}^{j=s} n_{ij}$, $i \in [\,1, p\,]$.

La moyenne de X_j est $\overline{X_j} = \frac{1}{n_{.j}} \sum_{i=1}^{i=p} n_{ij} x_i$.

On peut alors définir une nouvelle variable statistique sur Ω, qu'on appelle la **moyenne conditionnée** de X pour la partition $\{\Omega_1, \dots, \Omega_s\}$:

$$M_C(X) = \{(\overline{X_j}, \Omega_j, n_{.j})\},\ j \in [\,1, s\,].$$

La moyenne de cette variable statistique est :

$$\overline{M_C(X)} = \frac{1}{N} \sum_{j=1}^{j=s} n_{.j} \overline{X_j} = \frac{1}{N} \sum_{j=1}^{j=s} \sum_{i=1}^{i=p} n_{ij} x_i = \frac{1}{N} \sum_{i=1}^{i=p} \left(\sum_{j=1}^{j=s} n_{ij} \right) x_i = \frac{1}{N} \sum_{i=1}^{i=p} n_i x_i = \overline{X}.$$

$$\boxed{\overline{M_C(X)} = \overline{X}.}$$

Cette relation constitue le **théorème de la moyenne conditionnée**.

Exemple.

Soit Ω une population de commerçants, partitionnée en trois catégories disjointes :
- A : les supermarchés,
- B : les moyennes surfaces,
- C : les petits détaillants.

Soit X le prix du litre d'huile.

Soit $\overline{X_A}$ le prix moyen du litre d'huile dans les supermarchés : c'est le quotient entre le prix de vente total de l'huile dans les supermarchés, et le nombre total de litres vendus dans les supermarchés.

De même, soit $\overline{X_B}$, le prix moyen du litre d'huile dans les moyennes surfaces.

De même, soit $\overline{X_C}$, le prix moyen du litre d'huile chez les petits détaillants.

La relation précédente (théorème de la moyenne conditionnée) permet de calculer le prix moyen du litre d'huile en prenant le barycentre des prix moyens $\overline{X_A}$, $\overline{X_B}$, $\overline{X_C}$, affectés des nombres de litres d'huile vendus par chaque catégorie de commerçants (moyenne pondérée par les fréquences).

c) Moyenne d'une variable continue.

La variable est connue par ses classes et la fréquence associée à chaque classe.

$$[\, e_i, e_{i+1} \,[, f_i = \frac{n_i}{N}.$$

Supposons que nous connaissions le point moyen $\overline{X_i}$ de chaque classe $[\, e_i, e_{i+1} \,[$.
Alors, d'après le théorème de la moyenne conditionnée, la moyenne de X est donnée par :

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{i=p} n_i \overline{X_i} = \sum_{i=1}^{i=p} f_i \overline{X_i}.$$

Nous allons faire le calcul dans deux hypothèses.

Première hypothèse.

Dans chaque classe, toutes les observations sont **concentrées au centre de la classe** : $x_i = \frac{1}{2} (e_i + e_{i+1})$.

$$\overline{X_i} = \frac{1}{n_i} n_i x_i = x_i$$
$$\overline{X} = \sum_{i=1}^{i=p} f_i \overline{X_i} = \sum_{i=1}^{i=p} f_i x_i$$

Deuxième hypothèse.

Dans chaque classe, la répartition des observations est **uniforme**.
Alors, par raison de symétrie, la moyenne $\overline{X_i}$ d'une classe est la valeur centrale $x_i = \frac{1}{2} (e_i + e_{i+1})$ de la classe.
On a encore :

$$\overline{X} = \sum_{i=1}^{i=p} f_i \overline{X_i} = \sum_{i=1}^{i=p} f_i x_i$$

Conclusion : dans le cas d'une variable statistique continue, pour effectuer le calcul du point moyen, l'hypothèse de répartition uniforme dans chaque classe est équivalente à l'hypothèse d'une concentration de toutes les modalités d'une classe au centre de la classe.

d) Généralisation de la notion de moyenne.

Soit $X = \{(x_i, n_i)\}$, $i \in [\, 1, p \,]$, une variable statistique quantitative discrète à valeurs dans R_+^*, $N = \sum_{i=1}^{i=p} n_i$.

Soit $\varphi : R_+^* \to R$ une application monotone (injection croissante ou décroissante) continue.
Alors $\varphi(X) = \{(\varphi(x_i), n_i)\}$, $i \in [\, 1, p \,]$, est une variable statistique quantitative discrète à valeurs dans R.

On peut calculer sa moyenne $\overline{\varphi(X)} = \frac{1}{N} \sum_{i=1}^{i=p} n_i \varphi(x_i)$.

$\overline{\varphi(X)}$ est un nombre réel, compris entre la valeur minimum et la valeur maximum de $\varphi(x_i)$, $i \in [\, 1, p \,]$.
Comme φ est une injection continue, il existe un unique $\overline{X_\varphi} \in R_+^*$ tel que $\varphi(\overline{X_\varphi}) = \overline{\varphi(X)}$

$\overline{X_\varphi}$ est appelé la φ-**moyenne** de *X*.

Exemples de φ-moyennes.

1. Si φ est l'application identique définie par φ (*x*) = *x*, la φ-moyenne de *X* est la **moyenne arithmétique** de *X*, c'est la moyenne au sens ordinaire.
2. Si φ est définie par φ (*x*) = x^2, nous obtenons la **moyenne quadratique** $\overline{X}_q$ de *X*, définie par $\overline{X}_q{}^2 = \frac{1}{N}\sum_{i=1}^{i=p} n_i x_i^2$.
3. Si φ est définie par φ (*x*) = $\frac{1}{x}$, nous obtenons la **moyenne harmonique** $\overline{X}_h$ de *X*, définie par $\frac{1}{\overline{X}_h} = \frac{1}{N}\sum_{i=1}^{i=p} n_i \frac{1}{x_i}$.
4. Si φ est définie par φ (*x*) = *ln* (*x*), nous obtenons la **moyenne géométrique** $\overline{X}_g$ de *X*, définie par

$ln\,(\overline{X}_g) = \frac{1}{N}\sum_{i=1}^{i=p} n_i\, ln\,(x_i)$, soit $\overline{X}_g = \left(\prod_{i=1}^{i=p} x_i^{n_i}\right)^{\frac{1}{N}}$

Propriétés des φ-moyennes.

Pour une variable statistique *X*, les différentes moyennes, harmonique, géométrique, arithmétique, quadratique, sont liées par la relation :

$$\boxed{\overline{X}_h \le \overline{X}_g \le \overline{X} \le \overline{X}_q.}$$

Il y a égalité si, et seulement si, toutes les valeurs de *X* sont égales.
La moyenne géométrique est bien adaptée à l'étude des phénomènes de croissance.
La moyenne harmonique est utilisée pour les calculs d'indices économiques.

II. 2. 2. Paramètres de dispersion

Les paramètres de dispersion (étendue, intervalle interquartile,) sont calculés pour les variables statistiques quantitatives.
Ils ne donnent pas une information complète sur une variable statistique X : en effet, deux variables qui ont la même moyenne peuvent se présenter avec des dispersions très différentes.
L'histogramme, ou le diagramme, des fréquences donnent déjà une idée qualitative de la dispersion.

II. 2. 2. 1. Etendue

Soit X une variable statistique réelle discrète.
L'étendue ω de X est la différence entre la plus grande valeur de X et la plus petite valeur de X.

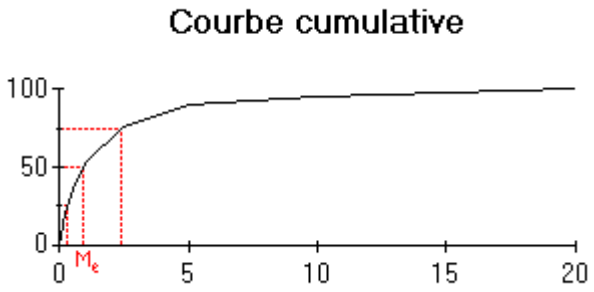
$$\omega = x_{max} - x_{min}$$

Ce paramètre est souvent utilisé dans les contrôles de fabrication, pour lesquels on donne, a priori, des marges de construction.
Son intérêt est limité par le fait qu'il dépend uniquement des valeurs extrêmes, qui peuvent être des valeurs aberrantes.

II.2.2.2. Quartiles et déciles.

a) Variable statistique continue.

Pour une variable statistique quantitative réelle continue X, on appelle **quartiles** les nombres réels Q_1, Q_2, Q_3, pour lesquels les fréquences cumulées de X sont respectivement 0,25, 0,50, 0,75.
Ce sont les valeurs pour lesquelles l'ordonnée de la **courbe cumulative des fréquences** est respectivement égale à 0,25, 0,50, 0,75.
Les quartiles partagent l'étendue en quatre intervalles qui ont le même effectif.

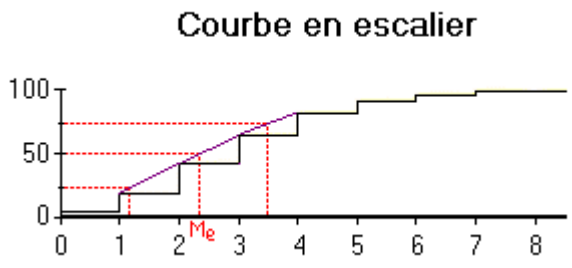


Le deuxième quartile, Q_2, est égal à la médiane.

L'**intervalle interquartile** est la différence entre les valeurs du troisième et du premier quartiles : $Q_3 - Q_1$.
L'intervalle $[Q_1, Q_3]$ contient 50 % des valeurs de X.

b) Variable statistique discrète.

Pour une variable statistique réelle discrète X, la courbe des fréquences cumulées est une **courbe en escalier**.
S'il existe une valeur de x pour laquelle la fréquence cumulée est 0,25 (resp. 0,50, 0,75), le quartile correspondant est cette valeur de X.
Sinon, les quartiles seront déterminés par **interpolation linéaire** entre deux valeurs.



c) Déciles et percentiles.

Les 9 déciles sont les nombres réels qui partagent l'étendue en dix intervalles de même effectif.
Utilisation : en matière de salaires, le rapport $\frac{D_9}{D_1}$ est un paramètre de dispersion fréquemment utilisé.

Les 99 percentiles sont les nombres réels qui partagent l'étendue en cent intervalles de même effectif.

II.2.2.3. Ecart absolu moyen.

a) Définition.

Soit $X = \{(x_i, n_i)\}_{1 \le i \le p}$ une variable statistique réelle.
On appelle **écart absolu moyen** de X la moyenne arithmétique des valeurs absolues des écarts de X à sa moyenne :

$$e = \frac{1}{N} \sum_{i=1}^{i=p} n_i \mid x_i - \overline{X} \mid$$

On pourrait aussi définir l'écart absolu moyen de X par rapport à sa médiane, ou par rapport à un nombre réel a quelconque.

$$e = \frac{1}{N} \sum_{i=1}^{i=p} n_i \mid x_i - a \mid$$

On peut démontrer que l'écart absolu moyen par rapport à un nombre réel a est minimum lorsque a est égal à la moyenne $\overline{X}$ de X.

b) Calcul pratique.

Lorsque les observations sont groupées par classe, on adopte généralement pour valeur de variable statistique le centre de chaque classe.
L'écart absolu moyen présente un inconvénient majeur : il ne se prête pas facilement aux calculs algébriques, à cause de la valeur absolue.

II.2.2.4. Variance et écart-type.

a) Définition.

Soit $X = \{(x_i, n_i)\}_{1 \le i \le p}$ une variable statistique réelle.
On appelle **variance** de X, la moyenne arithmétique des carrés des écarts de X à sa moyenne :

$$s^2(X) = \frac{1}{N} \sum_{\omega \in \Omega} (X(\omega) - \overline{X})^2 = \frac{1}{N} \sum_{i=1}^{i=p} n_i (x_i - \overline{X})^2$$

On appelle **écart-type** de X la racine carrée $s(X)$ de la variance de X.
$S = N s^2(X)$ est la somme des carrés des écarts : $S = \sum_{i=1}^{i=p} n_i (x_i - \overline{X})^2$

b) Formule de la variance.

En développant le carré $(x_i - \overline{X})^2$, la formule de définition de la variance peut être écrite :

$$s^2(X) = \frac{1}{N} \sum_{i=1}^{i=p} n_i x_i^2 - \overline{X}^2 = \overline{X^2} - \overline{X}^2$$

$$\boxed{s^2(X) = \overline{X^2} - \overline{X}^2}$$

Cette formule (la variance est égale à la moyenne du carré moins le carré de la moyenne) est appelée **formule de la variance**, ou formule de König.
Elle peut s'écrire sous la forme :

$$s^2(X) = \frac{1}{N} \left(\sum_{i=1}^{i=p} n_i x_i^2 - \frac{1}{N} \left(\sum_{i=1}^{i=p} n_i x_i \right)^2 \right)$$

c) Généralisation à $\mathbf{R}^q$.

Dans R, la distance euclidienne $d(X(\omega), \overline{X})$ entre $X(\omega)$ et $\overline{X}$, est l'écart absolu $| X(\omega) - \overline{X} |$, de sorte que la variance peut être écrite :

$$s^2(X) = \frac{1}{N} \sum_{\omega \in \Omega} (d(X(\omega), \overline{X}))^2.$$

Dans $\mathbf{R}^q$, on peut définir la **distance euclidienne** $d(X(\omega), \overline{X})$ entre $X(\omega) = \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_q(\omega) \end{pmatrix}$ et $\overline{X} = \begin{pmatrix} \overline{X_1} \\ \vdots \\ \overline{X_q} \end{pmatrix}$, par la formule

$$(d(X(\omega), \overline{X}))^2 = \sum_{j=1}^{j=q} (X_j(\omega) - \overline{X_j})^2 = \sum_{j=1}^{j=q} (d(X_j(\omega), \overline{X_j}))^2$$

La variance d'une variable statistique à valeurs dans $\mathbf{R}^q$, est alors définie par :

$$\begin{aligned} s^2(X) &= \frac{1}{N} \sum_{\omega \in \Omega} (d(X(\omega), \overline{X}))^2 \\ &= \frac{1}{N} \sum_{\omega \in \Omega} \sum_{j=1}^{j=q} (X_j(\omega) - \overline{X_j})^2 \\ &= \sum_{j=1}^{j=q} \left(\frac{1}{N} \sum_{\omega \in \Omega} (d(X_j(\omega), \overline{X_j}))^2 \right) \\ &= \sum_{j=1}^{j=q} s^2(X_j) \\ &= \sum_{j=1}^{j=q} (\overline{X_j^2} - (\overline{X_j})^2) \end{aligned}$$

Si X présente p modalités $x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iq} \end{pmatrix}$, $i \in [\, 1, p \,]$, il vient, en notant n_i l'effectif de la modalité x_i $\left(N = Card(\Omega) = \sum_{i=1}^{i=p} n_i \right)$:

$$s^2(X) = \frac{1}{N}\sum_{i=1}^{i=p} n_i \left(\sum_{j=1}^{j=q} (x_{ij} - \overline{X_{\cdot}})^2 \right)$$
$$= \frac{1}{N}\sum_{j=1}^{j=q} \left(\sum_{i=1}^{i=p} n_i (x_{ij} - \overline{X_{\cdot}})^2 \right)$$
$$= \sum_{j=1}^{j=q} \left(\frac{1}{N}\sum_{i=1}^{i=p} n_i (x_{ij} - \overline{X_{\cdot}})^2 \right)$$

$$\boxed{s^2(X) = \sum_{j=1}^{j=q} s^2(X_j) = \frac{1}{N}\sum_{j=1}^{j=q}\sum_{i=1}^{i=p} n_i (x_{ij} - \overline{X_{\cdot}})^2}$$

d) Propriétés de la variance.

1. La variance est toujours un nombre réel positif.
   En effet, c'est une somme de carrés.

2. La variance est nulle si, et seulement si, X possède une seule valeur.

   En effet, une somme de carrés $s^2(X) = \frac{1}{N}\sum_{\omega\in\Omega} (d(X(\omega), \overline{X}))^2$ est nulle si, et seulement si, chaque carré est nul.

3. $s^2(a + bX) = b^2 s^2(X)$, quels que soient les nombres réels a et b.
   En effet, si X est à valeurs réelles, on a :

$$\overline{(a+bX)^2} = \overline{a^2 + b^2 X^2 + 2abX} = a^2 + b^2\overline{X^2} + 2ab\overline{X}$$
$$\overline{a+bX} = a + b\overline{X}$$
$$(\overline{a+bX})^2 = a^2 + b^2(\overline{X})^2 + 2ab\overline{X}$$
$$s^2(a+bX) = \overline{(a+bX)^2} - (\overline{a+bX})^2 = b^2(\overline{X^2} - (\overline{X})^2) = b^2 s^2(X).$$

$$\boxed{s^2(a + bX) = b^2 s^2(X).}$$

Puis, si X est à valeurs dans R^q, on a :

$$s^2(a + bX) = \sum_{j=1}^{j=q} s^2(a + bX_j) = \sum_{j=1}^{j=q} b^2 s^2(X_j) = b^2 \sum_{j=1}^{j=q} s^2(X_j) = b^2 s^2(X).$$

e) Inertie par rapport à un point *a*.

On appelle **inertie** d'une variable statistique X par rapport à un point a, la moyenne du carré de la distance de X au point a :

$$\boxed{I_a(X) = \frac{1}{N}\sum_{\omega\in\Omega} (d(X(\omega), a))^2}$$

L'inertie de X par rapport au point moyen $\overline{X}$ est la variance de X.

Propriété.

L'inertie $I_a(X)$ est minimale lorsque a est égal à $\overline{X}$.

La valeur minimum de l'inertie est donc la variance de X.

En effet, soit $d = a - \overline{X}$.

Dans R^q, cette relation s'écrit : $\begin{pmatrix} d_1 \\ \vdots \\ d_q \end{pmatrix} = \begin{pmatrix} a_1 - \overline{X_1} \\ \vdots \\ a_q - \overline{X_q} \end{pmatrix}$.

$X(\omega)$ est une modalité $x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iq} \end{pmatrix}$ de X, d'effectif n_i, $i \in [\,1, p\,]$.

$$I_a(X) = \frac{1}{N} \sum_{\omega \in \Omega} (d(X(\omega), a))^2 = \frac{1}{N} \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_i (x_{ij} - a_j)^2$$

Ecrivons $x_{ij} - a_j$ sous la forme :

$$x_{ij} - a_j = x_{ij} - \overline{X_j} + \overline{X_j} - a_j$$

Il vient alors :

$$(x_{ij} - a_j)^2 = (x_{ij} - \overline{X_j})^2 + (\overline{X_j} - a_j)^2 + 2(x_{ij} - \overline{X_j})(\overline{X_j} - a_j)$$

$$I_a(X) = \frac{1}{N} \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_i (x_{ij} - \overline{X_j})^2 + \frac{1}{N} \sum_{i=1}^{i=p} n_i \sum_{j=1}^{j=q} (\overline{X_j} - a_j)^2 + 2 \frac{1}{N} \sum_{i=1}^{i=p} n_i \sum_{j=1}^{j=q} (x_{ij} - \overline{X_j})(\overline{X_j} - a_j)$$

$$= s^2(X) + \sum_{j=1}^{j=q} (\overline{X_j} - a_j)^2 + 2 \frac{1}{N} \sum_{j=1}^{j=q} (\overline{X_j} - a_j) \left(\sum_{i=1}^{i=p} n_i (x_{ij} - \overline{X_j}) \right)$$

Par définition de $\overline{X_j}$, on a $\sum_{i=1}^{i=p} n_i (x_{ij} - \overline{X_j}) = 0$.

Posons :

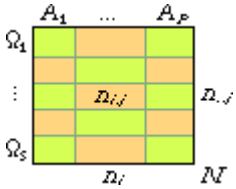
$$d^2 = \sum_{j=1}^{j=q} (\overline{X_j} - a_j)^2$$

Il reste :

$$\boxed{I_a(X) = s^2(X) + d^2.}$$

$s^2(X)$ est un nombre réel positif qui ne dépend pas de a.
d^2 est un nombre réel positif, sa valeur minimum est 0.
$I_a(X)$ est minimum lorsque d^2 est nul, c'est-à-dire lorsque $a_j = \overline{X_j}$ pour tout $j \in [\,1, q\,]$, soit $a = \overline{X}$.

f) Variance conditionnée.



Considérons maintenant une partition de Ω en s sous-populations $\Omega_1, \ldots, \Omega_s$.
Soit $X = \{(x_i, A_i, n_i)\}$, $i \in [\,1, p\,]$, une variable statistique quantitative discrète sur Ω, à valeurs dans R.
Chaque sous-population Ω_j, $j \in [\,1, s\,]$, définit une variable statistique X_j sur Ω_j, qui est la restriction de X à Ω_j.

On pose $n_{ij} = Card\,(A_i \cap \Omega_j)$, $n_{.j} = Card\,(\Omega_j) = \sum_{i=1}^{i=p} n_{ij}$, $j \in [\,1, s\,]$.

On a $n_i = Card\,(A_i) = \sum_{j=1}^{j=s} n_{ij}$, pour tout $i \in [\,1, p\,]$.

La moyenne de X_j est $\overline{X_j} = \frac{1}{n_{.j}} \sum_{i=1}^{i=p} n_{ij}\, x_i$.

La variance de X_j est $s^2\,(X_j) = \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij}\, {x_i}^2 - \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij}\, x_i \right)^2 \right)$

La **moyenne conditionnée** de X pour la partition $\{\Omega_1, \ldots, \Omega_s\}$ a été définie par la variable statistique :

$$M_C\,(X) = \{(\overline{X_j}, \Omega_j, n_{.j})\},\ j \in [\,1, s\,],\ \text{avec } N = \sum_{j=1}^{j=s} n_{.j}$$

La moyenne de cette variable statistique est : $\overline{M_C(X)} = \overline{X}$.

Sa variance est :

$$\begin{aligned} s^2\,(M_C\,(X)) &= \frac{1}{N} \left(\sum_{j=1}^{j=s} n_{.j}\, {\overline{X_j}}^2 - \frac{1}{N} \left(\sum_{j=1}^{j=s} n_{.j}\, \overline{X_j} \right)^2 \right) \\ &= \frac{1}{N} \left(\sum_{j=1}^{j=s} \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij}\, x_i \right)^2 - \frac{1}{N} \left(\sum_{j=1}^{j=s} \sum_{i=1}^{i=p} n_{ij}\, x_i \right)^2 \right) \\ &= \frac{1}{N} \left(\sum_{j=1}^{j=s} \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij}\, x_i \right)^2 - \frac{1}{N} \left(\sum_{i=1}^{i=p} n_i\, x_i \right)^2 \right) \end{aligned}$$

On peut définir une nouvelle variable statistique sur Ω, qu'on appelle la **variance conditionnée** de X pour la partition $\{\Omega_1, \ldots, \Omega_s\}$:

$$s_C{}^2\,(X) = \{(s^2\,(X_j), \Omega_j, n_{.j})\},\ j \in [\,1, s\,],\ \text{avec } N = \sum_{j=1}^{j=s} n_{.j}$$

La moyenne de cette variable statistique est : $\overline{s_C{}^2(X)} = \frac{1}{N} \sum_{j=1}^{j=s} n_{.j}\, s^2\,(X_j)$.

Sa variance est $s^2\,(s_C{}^2\,(X)) = \frac{1}{N} \left(\sum_{j=1}^{j=s} n_{.j}\, (s^2\,(X_j))^2 - \frac{1}{N} \left(\sum_{j=1}^{j=s} n_{.j}\, s^2\,(X_j) \right)^2 \right)$

On a alors :

$$\begin{aligned} N\, \overline{s_C{}^2(X)} &= \sum_{j=1}^{j=s} n_{.j}\, s^2\,(X_j) = \sum_{j=1}^{j=s} \left(\sum_{i=1}^{i=p} n_{ij}\, {x_i}^2 - \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij}\, x_i \right)^2 \right) \\ &= \sum_{i=1}^{i=p} \left(\sum_{j=1}^{j=s} n_{ij} \right) {x_i}^2 - \sum_{j=1}^{j=s} \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij}\, x_i \right)^2 \\ &= \sum_{i=1}^{i=p} n_i\, {x_i}^2 - \sum_{j=1}^{j=s} \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij}\, x_i \right)^2 \\ \overline{s_C{}^2(X)} &= \frac{1}{N} \left(\sum_{i=1}^{i=p} n_i\, {x_i}^2 \right) - \frac{1}{N} \left(\sum_{j=1}^{j=s} \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij}\, x_i \right)^2 \right) \end{aligned}$$

$$\overline{s_C^2(X)} + s^2(\mathrm{M}_C(X)) = \frac{1}{N}\left(\sum_{i=1}^{i=p} n_i x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{i=p} n_i x_i \right)^2 \right) = s^2(X)$$

La relation :

$$\boxed{s^2(X) = \overline{s_C^2(X)} + s^2(\mathrm{M}_C(X))}$$

constitue le **théorème de la variance conditionnée** : la variance de *X* est la somme de la moyenne de la variance conditionnée de *X* et de la variance de la moyenne conditionnée de *X*.

– Le terme $\overline{s_C^2(X)}$ s'appelle la **variance intraclasse**. Il traduit la variation de *X* autour de sa moyenne, dans la partition $\{\Omega_1, \ldots, \Omega_s\}$.

– Le terme $s^2(\mathrm{M}_C(X))$ s'appelle la **variance interclasse**. Il traduit la variation de la moyenne de *X* dans la partition $\{\Omega_1, \ldots, \Omega_s\}$.

Note : Ce résultat peut être étendu à une variable statistique discrète à valeurs dans R^q.

g) Variance d'une variable statistique réelle continue.

Les classes $[\, e_i, e_{i+1} \,[$, de fréquences $f_i = \frac{n_i}{N}$, $i \in [\, 1, p \,]$, forment une partition de $X(\Omega)$.
La variance de *X* s'obtient :

— en calculant la variance $s_i^2(X)$ de *X* dans chaque classe,

— en faisant la moyenne de ces variances (moyenne de la variance conditionnée) : $\sum_{i=1}^{i=p} f_i s_i^2(X)$

— en calculant la variance de la moyenne de *X* dans chaque classe (variance de la moyenne conditionnée) : $\sum_{i=1}^{i=p} f_i (\overline{X_i} - \overline{X})^2$

— en faisant la somme de la moyenne de la variance conditionnée et de la variance de la moyenne conditionnée :

$$s^2(X) = \sum_{i=1}^{i=p} f_i s_i^2(X) + \sum_{i=1}^{i=p} f_i (\overline{X_i} - \overline{X})^2$$

1°/ Dans l'hypothèse où toutes les observations sont concentrées au milieu de la classe $x_i = \frac{e_i + e_{i+1}}{2}$, la variance $s_i^2(X)$ de *X* dans chaque classe, est nulle, $s^2(X) = \sum_{i=1}^{i=p} f_i (x_i - \overline{X})^2$. On retrouve la formule du cas discret.

$$\boxed{s^2(X) = s^2(U)}$$

où $x_i = \frac{e_i + e_{i+1}}{2}$ est le centre de la classe d'indice *i* et *U* est la variable statistique $\{(x_i, n_i)\}$, $i \in \{1, \ldots, p\}$.

2°/ Dans l'hypothèse où la répartition des valeurs de *X* dans chaque classe est uniforme, au terme $\sum_{i=1}^{i=p}$

$f_i\,(\overline{X_i} - \overline{X})^2 = \sum_{i=1}^{i=p} f_i\,(x_i - \overline{X})^2$, s'ajoute un terme correctif $\sum_{i=1}^{i=p} f_i\,s_i^{\,2}\,(X)$ qui tient compte de la variation de X dans chaque classe.
Pour calculer ce terme complémentaire, il faut calculer la variance d'une variable répartie uniformément sur un intervalle.

Lemme.

La variance d'une variable statistique répartie uniformément sur un intervalle de longueur a est $\frac{a^2}{12}$.

Démonstration du lemme.

On peut utiliser la formule de la variance : la variance est égale à la moyenne du carré moins le carré de la moyenne.
La moyenne du carré est

$$\overline{X^2} = \frac{1}{a}\int_{e_i}^{e_i+a} x^2\,dx = \frac{1}{a}\left[\frac{x^3}{3}\right]_{x=e_i}^{x=e_i+a} = \frac{1}{3a}[\,(e_i+a)^3 - e_i^{\,3}\,] = \frac{1}{3a}(3\,e_i^{\,2}\,a + 3\,e_i\,a^2 + a^3)$$

$$= \frac{a^2}{3} + e_i^{\,2} + e_i\,a$$

Le carré de la moyenne est

$$\overline{X}^2 = \left[\frac{1}{2}[e_i + (e_i + a)]\right]^2 = \left[e_i + \frac{a}{2}\right]^2 = \frac{a^2}{4} + e_i^{\,2} + e_i\,a.$$

La variance de X dans l'intervalle $[e_i, e_i + a]$ est donc :

$$s_i^{\,2}\,(X) = \left[\frac{a^2}{3} + e_i^{\,2} + e_i\,a\right] - \left[\frac{a^2}{4} + e_i^{\,2} + e_i\,a\right] = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12}$$

Le terme correctif $\sum_{i=1}^{i=p} f_i\,s_i^{\,2}\,(X)$ est donc donné par :

$$\sum_{i=1}^{i=p} f_i\,s_i^{\,2}\,(X) = \frac{1}{12}\sum_{i=1}^{i=p} f_i\,(e_{i+1} - e_i)^2.$$

Dans le cas où toutes les classes ont la même amplitude $e_{i+1} - e_i = a$, le terme correctif est :

$$\sum_{i=1}^{i=p} f_i\,s_i^{\,2}\,(X) = \frac{a^2}{12}\sum_{i=1}^{i=p} f_i = \frac{a^2}{12}$$

et la variance de X est donnée par :

$$s^2\,(X) = \sum_{i=1}^{i=p} f_i\,(x_i - \overline{X})^2 + \frac{a^2}{12} = s^2\,(U) + \frac{a^2}{12}$$

$$\boxed{s^2\,(X) = s^2\,(U) + \frac{a^2}{12}}$$

où $x_i = \frac{e_i + e_{i+1}}{2}$ est le centre de la classe d'indice i et U est la variable statistique $\{(x_i, n_i)\}$, $i \in \{1, \ldots, p\}$.

II.2.2.5. Coefficient de variation.

Pour une variable statistique réelle X, on appelle **coefficient de variation** le rapport

$$c = \frac{s(X)}{\overline{X}}$$

Pour une variable statistique X à valeurs dans $\mathbb{R}^q$, le coefficient de variation est défini par :

$$c = \frac{s(X)}{d(0, \overline{X})}.$$

Le coefficient de variation est un nombre sans dimension qui permet de comparer deux variables statistiques de natures différentes.
On remarquera que, au signe près, c'est l'écart-type de la variable statistique $\frac{X}{\overline{X}}$ ou $\frac{X}{d(0, \overline{X})}$.

II.2.2.6. Moments.

Soit X une variable statistique quantitative réelle.
On appelle **moment d'ordre *r*** de X, la quantité :

$$m_r = \frac{1}{N} \sum_{\omega \in \Omega} [X(\omega)]^r = \frac{1}{N} \sum_{i=1}^{i=p} n_i x_i^r$$

Pour $r = 0$: $m_0 = 1$.
Pour $r = 1$: $m_1 = \overline{X}$. Le moment d'ordre 1 est la moyenne.
Pour $r = 2$: $m_2 = \overline{X^2}$.

On appelle **moment centré d'ordre *r*** de X, la quantité :

$$\mu_r = \frac{1}{N} \sum_{\omega \in \Omega} [X(\omega) - \overline{X}]^r = \frac{1}{N} \sum_{i=1}^{i=p} n_i (x_i - \overline{X})^r$$

Pour $r = 0$: $\mu_0 = 1$.
Pour $r = 1$: $\mu_1 = 0$.
Pour $r = 2$: $\mu_2 = s^2(X) = m_2 - m_1^2$. Le moment centré d'ordre 2 est la variance.

II.2.2.7. Conclusion.

Centrer et réduire une variable statistique quantitative X consiste la remplacer par $\frac{X - \overline{X}}{s(X)}$:

— $X - \overline{X}$ pour la centrer (moyenne 0)
— diviser par $s(X)$ pour la réduire (écart-type 1).

La variable $X' = \frac{X - \overline{X}}{s(X)}$ a pour moyenne 0 (elle est **centrée**) et pour écart-type 1 (elle est **réduite**).

Par exemple, si nous considérons la variable statistique continue 

théorique dont la densité de fréquence est

$$h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \text{ (loi de Gauss),}$$

sa moyenne est 0 et son écart-type est 1 : c'est une variable centrée réduite et la courbe de densité de fréquence associée est appelée la **courbe en cloche**, ou **courbe de Gauss**.
Un problème intéressant sera de comparer la courbe de densité de fréquence d'une variable statistique quantitative à cette courbe en cloche.

II. 2. 3. Paramètres de forme

Nous définissons les paramètres de forme pour une variable statistique quantitative, discrète ou continue, à valeurs réelles.

II. 2. 3. 1. Coefficient d'asymétrie.

a) Définition.

Il existe plusieurs coefficients d'asymétrie. Les principaux sont les suivants.

Le coefficient d'asymétrie de **Pearson** fait intervenir le mode M_o : quand il existe, il est définie par

$$P = \frac{\overline{X} - M_o}{s(X)}.$$

Le coefficient d'asymétrie de **Yule** fait intervenir la médiane et les quartiles, il est défini par

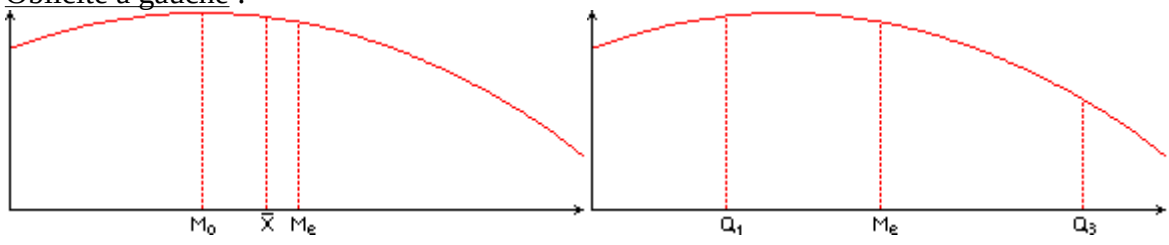
$$Y = \frac{Q_1 + Q_3 - 2M_e}{2(Q_3 - Q_1)}.$$

Le coefficient d'asymétrie de **Fisher** fait intervenir les moments centrés, il est défini par

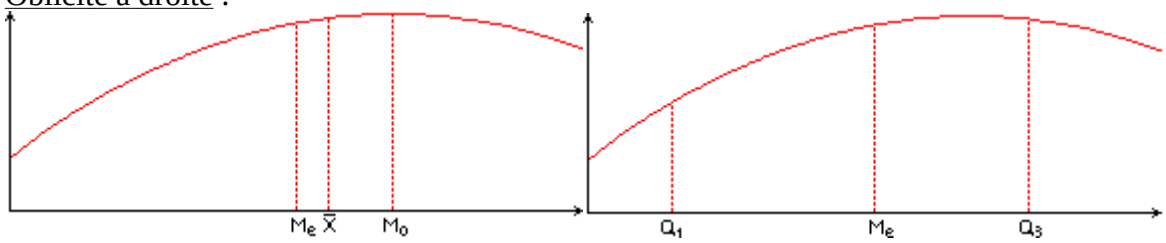
$$F = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\mu_3}{s^3(X)}.$$

Lorsque le coefficient d'asymétrie est positif, la distribution est plus étalée à droite : on dit qu'il y a **oblicité à gauche**.
Lorsque le coefficient d'asymétrie est négatif, la distribution est plus étalée à gauche : on dit qu'il y a **oblicité à droite**.

Oblicité à gauche :



Oblicité à droite :



On utilise souvent un coefficient d'asymétrie de **Pearson basé sur les moments centrés** : $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$.

Ce coefficient d'asymétrie est toujours positif.
Il est nul pour une distribution à densité de fréquence symétrique, telle la loi de Gauss.

b) Exemples.

1°/ Considérons la variable statistique X de distribution :

x_i	-1	4
n_i	4	1

$M_o = -1$; $\mu_3 = \frac{1}{5}(4 \times (-1)^3 + 1 \times 4^3) = 12$; $\mu_2 = \frac{1}{5}(4 \times (-1)^2 + 1 \times 4^2) = 4$.

$$P = \frac{\overline{X} - M_o}{s(X)} = \frac{1}{2} > 0 \text{ : oblicité à gauche.}$$

$$F = \frac{\mu_3}{s^3(X)} = \frac{3}{2} > 0 \text{ : oblicité à gauche.}$$

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} = \frac{9}{4}.$$

2°/ Considérons la variable statistique X de distribution :

x_i	-4	1
n_i	1	4

$M_o = 1$; $\mu_3 = \frac{1}{5}(1 \times (-4)^3 + 4 \times 1^3) = -12$; $\mu_2 = \frac{1}{5}(1 \times (-4)^2 + 4 \times 1^2) = 4$.

$$P = \frac{\overline{X} - M_o}{s(X)} = -\frac{1}{2} < 0 \text{ : oblicité à droite.}$$

$$F = \frac{\mu_3}{s^3(X)} = -\frac{3}{2} < 0 \text{ : oblicité à droite.}$$

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} = \frac{9}{4}.$$

II. 2. 3. 2. Coefficient d'aplatissement.

Là encore plusieurs définitions sont possibles.

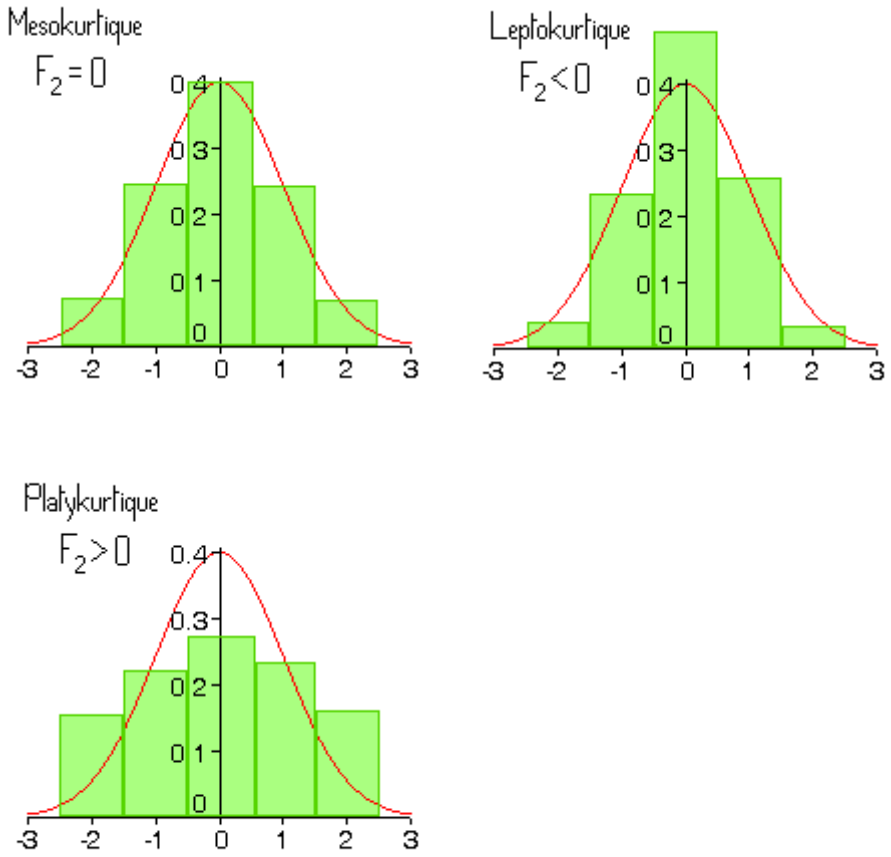
Le coefficient d'aplatissement de Pearson est $\beta_2 = \frac{\mu_4}{{\mu_2}^2}$.

Le coefficient d'aplatissement de Yule est $F_2 = \frac{\mu_4}{{\mu_2}^2} - 3$.

On peut se demander pourquoi – 3 ?
C'est parce que, en Probabilités, on peut démontrer que le coefficient d'aplatissement de Pearson pour une variable aléatoire réelle qui suit une loi de Gauss, est égal à 3.

Il est alors naturel, pour comparer l'applatissement d'une distribution statistique à l'aplatissement d'une variable de Gauss, d'introduire le coefficient $F_2 = \beta_2 - 3$.

Si F_2 est égal à 0, le polygone statistique de la variable réduite a le même aplatissement qu'une courbe en cloche, on dit que la variable est **mésokurtique**.
Si F_2 est > 0, le polygone statistique de la variable réduite est moins aplati qu'une courbe en cloche, on dit que la variable est **leptokurtique**.
Si F_2 est < 0, le polygone statistique de la variable réduite est plus aplati qu'une courbe en cloche, on dit que la variable est **platykurtique**.



II. 2. 3. 3. Indice de concentration de Gini.

a) Courbe de Lorenz.

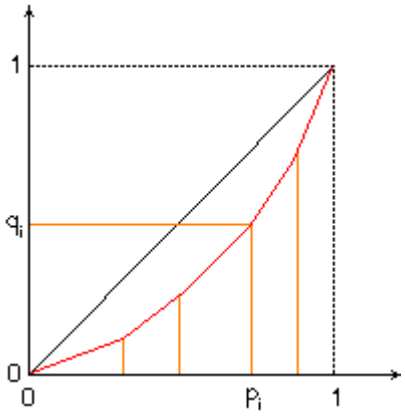
La notion de concentration ne s'applique qu'à des variables statistiques quantitatives à valeurs strictement positives.
Elle se comprendra facilement sur un exemple.

Considérons la distribution des salaires dans la populations des salariés d'une entreprise.
Les salaires sont divisés en n classes : la i^e classe, [e_i, e_{i+1} [a, pour centre, x_i et, pour effectif, n_i.
On note p_i la fréquence cumulée de e_{i+1} : c'est la proportion de salariés dont le salaire est strictement plus petit que e_{i+1}.
On note q_i la proportion de masse salariale représentée par les salariés dont le salaire est strictement

plus petit que e_{i+1}.

$$q_i = \frac{\sum_{k=1}^{k=i} n_k x_k}{\sum_{k=1}^{k=n} n_k x_k} = \frac{\sum_{k=1}^{k=i} n_k x_k}{n \overline{X}} = \frac{1}{\overline{X}} \sum_{k=1}^{k=i} f_k x_k = \sum_{k=1}^{k=i} f_k \frac{x_k}{\overline{X}}$$

On appelle courbe de concentration, ou courbe de Lorenz, la ligne polygonale joignant les points de corrdonnées (p_i, q_i).



En réalité, pour une variable statistique continue, on ne connaît la courbe de Lorenz que pour les extrémités des classes : l'interpolation linéaire suppose que la répartition des valeurs de la variable à l'intérieur de chaque classe est uniforme.

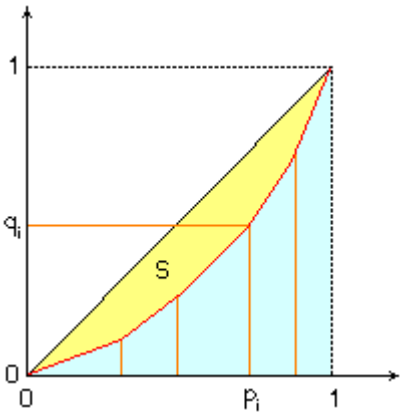
Dans le cas d'une variable discrète, on adopte aussi la représentation par une ligne polygonale.

La courbe de Lorenz est toujours inscrite dans le carré [0, 1] × [0, 1]. Cette courbe se caractérise par les traits suivants.

1°/ Les **points extrêmes** sont les points (0, 0) et (1, 1) puisque 0 % de la population reçoit 0 % de de la masse salariale et 100 % de la population reçoit 100 % de la masse salariale.
2°/ La courbe est nécessairement **convexe vers le bas**.
Cela résulte du fait que la pente du segment qui correspond, par exemple, aux points d'abscisses 0, 50 et 0,60, ne peut être inférieure à celle du segment correspondant aux abscisses 0,40 et 0,50 puisque, par définition, on considère des classes successives disposant chacune d'une part croissante de la masse salariale totale.
3°/ Enfin, et surtout, la courbure de la courbe de Lorenz peut être interprétée comme un **indice d'inégalité**.
En effet, dans une situation hypothétique d'égalité absolue, la courbe prendrait la forme d'un segment de droite (diagonale du carré) tendue entre les points (0, 0) et (1, 1).
De même, dans une situation d'inégalité extrême où la quasi-totalité de la masse salariale serait détenue par une infime minorité de la population, la courbe de Lorenz tendrait à longer l'axe des *p*, avant de remonter brutalement vers le point (1, 1).

b) Indice de Gini.

L'indice de Gini (du nom du statisticien italien Corrado Gini qui a proposé en 1912 cet indice pour les distributions de salaires et de revenus), quant à lui, est obtenu en déterminant la surface *S* comprise entre la courbe de Lorenz et la diagonale et en rapportant cette surface à la surface du demi-carré dans lequel s'inscrit cette courbe.
Comme la surface du carré est 1, l'indice de Gini est le double de l'aire *S* comprise entre la courbe de Lorenz et la diagonale du carré.
Très souvent, la surface *S* peut être déterminée avec suffisamment de précisions de manière graphique.
Numériquement, on peut calculer l'indice de Gini par la formule :



$$g = 2\,S = 1 - \sum_{i=1}^{i=n-1} (p_{i+1} - p_i)(q_{i+1} + q_i) = 1 - \sum_{i=1}^{i=n-1} f_{i+1}(q_{i+1} + q_i)$$

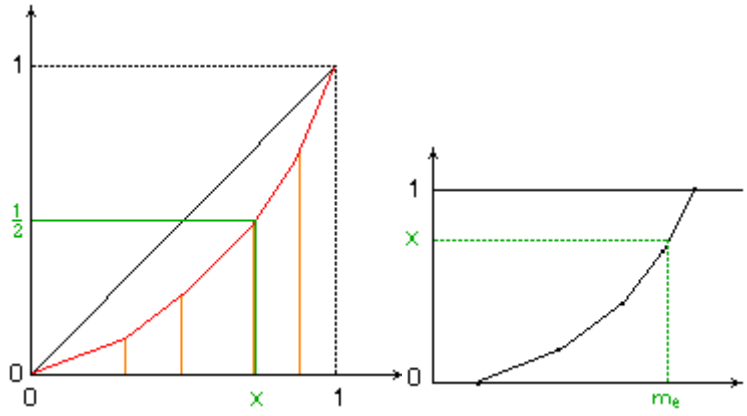
Dire que $g = 0$, c'est dire que la courbe de Lorenz coïncide avec la diagonale du carré (égalité absolue).
Dire que $g = 1$, c'est dire que la courbe de Lorenz longe d'abord l'axe des p, puis la droite $p = 1$ (inégalité maximale).
De façon générale, l'indice de Gini peut être interprété comme ayant une valeur d'autant plus grande que l'inégalité est grande : il constitue donc une bonne mesure de l'inégalité.

Applications.

L'indice de Gini permet de mesurer les inégalités scolaires, les inégalités de statut, les inégalités de salaires, etc.

c) Médiale.

La **médiale** d'une variable statistique X est la valeur de X qui partage la masse globale en deux parties égales.
Sur la courbe de Lorenz, la moitié de la masse globale correspond à l'ordonnée $\frac{1}{2}$.
Le point d'ordonnée $\frac{1}{2}$ a une abscisse x qui correspond à une fréquence cumulée x.
La valeur correspondante de X s'obtient en prenant l'abscisse du point d'ordonnée x sur le diagramme cumulatif des fréquences.



Si la variable statistique X est définie par $\{(x_i, n_i)\}$, $i \in [1, p]$, soit $\overline{X} = \frac{1}{N} \sum_{i=1}^{i=p} n_i x_i$, avec $N = \sum_{i=1}^{i=p} n_i$.

Pour une variable continue, x_i représente le centre de la i^{e} classe.

On pose $r_i = \frac{n_i x_i}{N \overline{X}}$. On a : $\sum_{i=1}^{i=p} r_i = 1$.
Dans notre exemple, r_i représente la fraction de la masse salariale globale gagnée par les personnes dont le salaire est x_i.
La médiale de X est la médiane de la variable statistique $\{(x_i, r_i)\}$, $i \in [1, p]$.
La médiale n'est pas le salaire gagné par l'employé qui est "au milieu de la file", mais le salaire gagné par le salarié qui permet d'atteindre la moitié de la masse salariale totale.

La comparaison des valeurs de la médiale et de la médiane constitue une mesure de la concentration.
Lorsque l'écart entre la médiale et la médiane est important par rapport à l'étendue de la distribution de la variable, la concentration est forte.
Si la distribution est égalitaire, la concentration est faible et l'écart entre la médiale et la médiane est faible.

La médiale est toujours **supérieure à la médiane**, puisque 50 % des effectifs cumulés croissants ne permettent jamais d'atteindre 50 % de la masse totale.

Chapitre III - ANALYSE BIVARIEE.

(Variables statistiques à deux dimensions)

III.1. DEFINITIONS.

III.1.1. Variable statistique à deux dimensions.

Considérons une population finie Ω (*Card* (Ω) = *N*) sur laquelle nous étudions deux caractères (qualitatifs ou quantitatifs réels) *A* et *B*.
Désignons par A_i, $i \in [1, p]$, les modalités observées du caractère *A*, par B_j, $j \in [1, q]$, les modalités observées du caractère *B*.
Appelons C_{ij} l'ensemble des $\omega \in \Omega$ présentant, à la fois, la modalité A_i du caractère *A* et la modalité B_j du caractère *B*.
Appelons n_{ij} le cardinal de C_{ij}.
$N = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij}$.

On appelle **variable statistique à deux dimensions** l'ensemble *Z* des triplets $((A_i, B_j), C_{ij}, n_{ij})$, pour $i \in [1, p]$ et $j \in [1, q]$, pour lesquels n_{ij} n'est pas nul.

Les C_{ij} forment une partition de Ω.

Le nombre $n_{i.} = \sum_{j=1}^{j=q} n_{ij}$ des individus $\omega \in \Omega$ présentant la modalité A_i du caractère *A*, permet de définir une variable statistique *X* à une dimension.
Le nombre $n_{.j} = \sum_{i=1}^{i=p} n_{ij}$ des individus $\omega \in \Omega$ présentant la modalité B_j du caractère *B*, permet de définir une variable statistique *Y* à une dimension.
Le couple (*X*, *Y*) est une **variable conjointe** : c'est une variable statistique à deux dimensions si l'on en élimine les modalités conjointes (A_i, B_j) dont l'effectif est nul.

En pratique, on admettra que, pour une variable statistique *Z* à deux dimensions :
— des modalités conjointes (A_i, B_j) peuvent avoir un effectif n_{ij} nul,
— pour tout $j \in [1, q]$, il existe au moins un $i \in [1, p]$ tel que n_{ij} ne soit pas nul,
— pour tout $i \in [1, p]$, il existe au moins un $j \in [1, q]$ tel que n_{ij} ne soit pas nul.
Dans ce cas, une variable statistique à deux dimensions est une variable conjointe, couple de deux variables statistiques à une dimension.

Une telle variable statistique à deux dimensions peut se représenter par un tableau à double entrée appelé **tableau de contingence**.

A \ B	$B1$	$\cdots$	B_j	$\cdots$	B_q	Total
A_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$	$\vdots$
A_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p.}$
Total	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.q}$	N

La fréquence de la modalité conjointe (A_i, B_j) est $f_{ij} = \frac{n_{ij}}{N}$.

La fréquence de la modalité A_i est $f_{i.} = \frac{n_{i.}}{N} = \sum_{j=1}^{j=q} f_{ij}$.

La fréquence de la modalité B_j est $f_{.j} = \frac{n_{.j}}{N} = \sum_{i=1}^{i=p} f_{ij}$.

Ces fréquences sont parfois appelées des "**pondérations**".

Elles vérifient les égalités : $\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} f_{ij} = \sum_{i=1}^{i=p} f_{i.} = \sum_{j=1}^{j=q} f_{.j} = 1$.

III.1.2. Variables marginales. Variables conditionnelles.

III.1.2.1. Variables marginales.

Soit $Z = \{((A_i, B_j), C_{ij}, n_{ij})\}$, $i \in [1, p]$, $j \in [1, q]$, une variable statistique à deux dimensions.

A \ B	$B1$	$\cdots$	B_j	$\cdots$	B_q	Total
A_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$	$\vdots$
A_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p.}$
Total	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.q}$	N

Considérons les variables statistiques
$X = \{(A_i, C_{i.}, n_{i.})\}$, $i \in [1, p]$,

définie par $C_{i.} = \bigcup_{j=1}^{j=q} C_{ij}$ et $n_{i.} = \sum_{j=1}^{j=q} n_{ij}$, et

$Y = \{(B_j, C_{.j}, n_{.j})\}$, $j \in [1, q]$,

définie par $C_{.j} = \bigcup_{i=1}^{i=p} C_{ij}$ et $n_{.j} = \sum_{i=1}^{i=p} n_{ij}$.

Les variables statistiques X et Y ainsi définies sont appelées les **variables marginales** de Z.
Leur distribution est représentée par les marges du tableau de contingence.

III.1.2.2. Variables conditionnelles.

Considérons la j^e colonne du tableau de contingence :

B / A	B_j
A_1	n_{1j}
⋮	⋮
A_i	n_{ij}
⋮	⋮
A_p	n_{pj}
Total	$n_{.j}$

Ce tableau représente une variable statistique dont les modalités sont les A_i, $i \in [1, p]$ pour lesquels les n_{ij} ne sont pas nuls.

A ces modalités, est associée une partition de $C_{.j} = \bigcup_{i=1}^{i=p} C_{ij}$ par les C_{ij} non vides, pour j fixé, avec, pour effectifs, les n_{ij} non nuls.

Cette variable statistique $\{(A_i, C_{ij}, n_{ij})\}$, $i \in [1, p]$, définie par une colonne du tableau de contingence, est appelée la **variable *X* conditionnée par B_j**, ou variable X **conditionnelle** pour B fixé.

Pour cette variable conditionnelle, nous pouvons définir la **fréquence conditionnelle** de la modalité A_i par $f_{i|j} = \frac{n_{ij}}{n_{.j}}$.

On peut définir ainsi q variables conditionnelles, correspondant aux q colonnes du tableau de contingence (autant qu'il existe de modalités du caractère B).

B / A	$B1$	…	B_j	…	B_q	Total
A_i	n_{i1}	…	n_{ij}	…	n_{iq}	$n_{i.}$

De la même façon, nous pouvons définir pour chaque ligne du tableau de contingence une variable Y conditionnée par A_i, avec une fréquence conditionnelle de la modalité B_j donnée par $f_{j|i} = \frac{n_{ij}}{n_{i.}}$.

Remarque.

Si les deux variables X et Y sont quantitatives et jouent des rôles symétriques, il est intéressant d'étudier les variables conditionnelles des deux types.

Exemple : taille et poids d'étudiants.

Si l'une des variables est qualitative et l'autre quantitative, alors seul le conditionnement par la variable qualitative présente un intérêt.

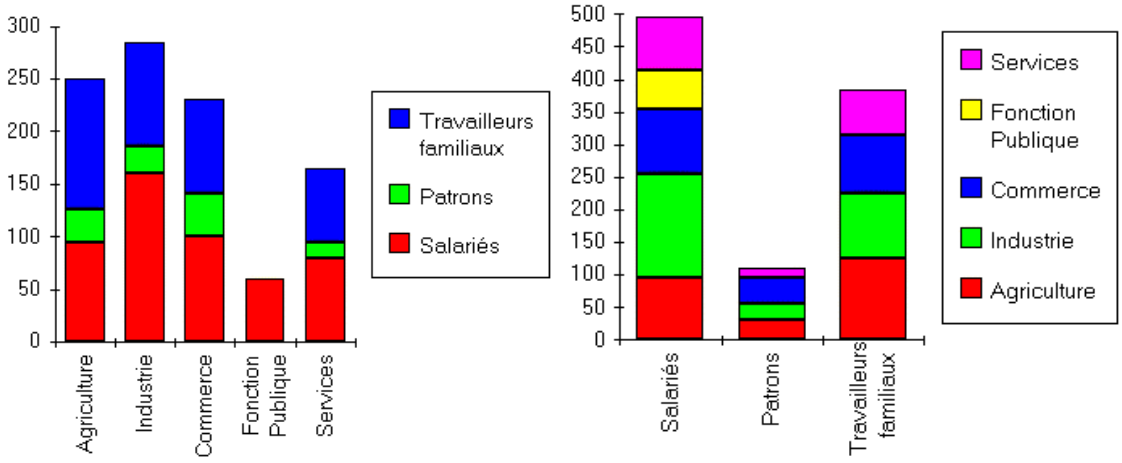
III.2. REPRESENTATION GRAPHIQUE.

III.2.1. Variable qualitative.

Pour une variable qualitative *Z* à deux dimensions, les données du tableau de contingence seront représentées par un **diagramme en tuyaux d'orgue**.

Exemple.

Statut / *Branche*	*Salariés*	*Patrons*	*Travailleurs familiaux*	Total
Agriculture	95	30	125	250
Industrie	160	25	100	285
Commerce	100	40	90	230
Fonction Publique	60	0	0	60
Services	80	15	70	165
Total	495	110	385	990



III.2.2. Variable quantitative.

III.2.2.1. Nuage de points.

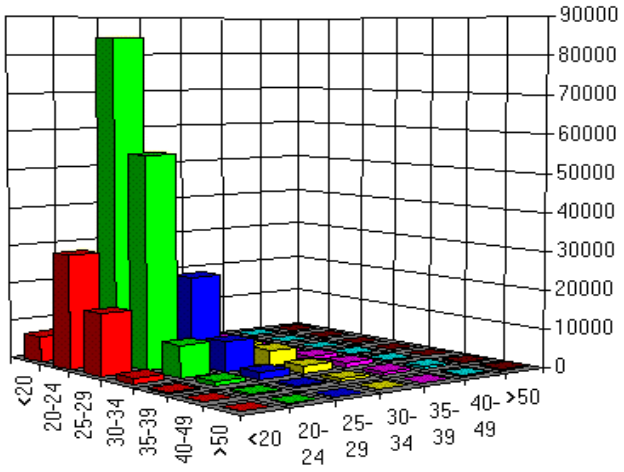
Pour une variable quantitative, discrète ou continue, on peut utiliser une représentation par un **nuage de points** dans un plan.
On peut remplacer chaque point par un **cercle** délimitant une aire proportionnelle à l'effectif ou à la fréquence.

III.2.2.2. Stéréogramme.

Dans certains cas, on peut faire une représentation dans R^3 :
- stéréogramme en bâtons pour une variable discrète.
- stéréogramme en histogramme pour une variable continue.

Exemple : Mariages célébrés en 1962, suivant l'âge des époux (1^e colonne : âge de l'époux, 1^e ligne : âge de l'épouse).

	<20	20-24	25-29	30-34	35-39	40-49	>50
<20	6756	3051	180	15	3	3	0
20-24	29416	84556	13430	1205	168	50	10
25-29	15893	54978	22774	3890	651	113	14
30-34	1789	8289	7809	4111	1021	244	15
35-39	255	1304	1996	2078	1232	362	20
40-49	66	283	447	733	852	697	120
>50	6	46	59	83	145	336	472



III.2.3. Variable mixte.

Dans le cas d'une variable mixte, ayant une composante qualitative et une composante quantitative, on utilise une représentation dans R^2 ou dans R^3 en plaçant de facon arbitraire les modalités de la variable qualitative sur l'un des axes.

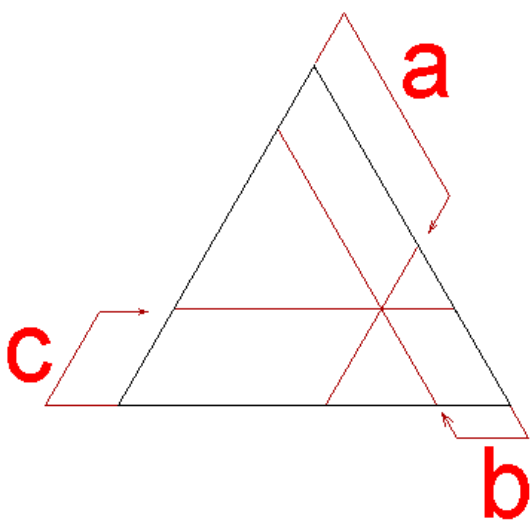
III.2.4. Autres représentations.

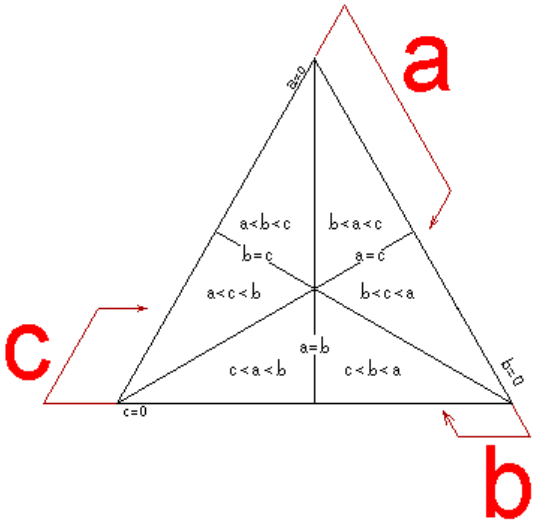
III.2.4.1. Représentation en étoile.

La représentation en étoile permet de représenter un phénomène périodique.
Par exemple, l'évolution d'un indice de prix peut se représenter par douze rayons équidistants représentant les mois avec, sur chaque rayon, les indices de prix pour le mois correspondant, d'année en année (spirale des prix).

III.2.4.2. Représentation triangulaire.

La représentation graphique triangulaire est utilisée pour représenter une quantité constante, fractionnée en trois parties variables (de somme constante).
Le principe de cette représentation repose sur le fait qu'étant donné un point à l'intérieur d'un triangle équilatéral, si l'on trace à partir de ce point des parallèles aux trois côtés, la somme des longueurs des segments déterminés par ces parallèles du point choisi aux côtés du triangle, est constante et égale à la longueur du côté du triangle équilatéral.
En particulier, on utilisera cette représentation triangulaire si la grandeur à représenter est somme de trois grandeurs représentées par des pourcentages.





Dans cette représentation, les **côtés** du triangle correspondent à la valeur 0 de l'une des trois composantes.
Les **sommets** du triangle correspondent à la valeur 0 de deux des trois composantes.
Les **milieux des côtés** correspondent à la valeur 0 de l'une des trois composantes et à la valeur 50 % des deux deux autres composantes.
Le **centre** du triangle correspond à l'égalité des trois grandeurs représentées.
Les **hauteurs** du triangle correspondent à l'égalité de deux des trois facteurs, ce qui permet de diviser l'aire du triangle en zones caractérisées par un critère précis.

Exemple.

A une date donnée, on répartit les différents secteurs d'activité selon le pourcentage d'entreprises escomptant une augmentation, une diminution, ou une stabilité, de leur activité pour la période à venir. La représentation du point dans un diagramme triangulaire, permet de suivre à travers le temps l'évolution des pronostics pour une même branche d'activité (analyse des réponses des chefs d'entreprise à l'enquête trimestrielle sur la conjoncture économique).

III.3. CARACTERISTIQUES MARGINALES ET CONDITIONNELLES.

III.3.1. Caractéristiques marginales.

Soit $Z = \{(x_i , y_j), C_{ij} , n_{ij})\}$, $i \in [1, p]$, $j \in [1, q]$, une variable statistique quantitative à deux dimensions, de variables marginales
$X = \{(x_i , C_{i.} , n_{i.})\}$, $i \in [1, p]$, et $Y = \{(y_j , C_{.j} , n_{.j})\}$, $j \in [1, q]$.

$$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} = N$$

X et Y sont des variables statistiques quantitatives, discrètes ou continues.

Pour une variable continue, les valeurs sont celles des moyennes des classes (centre de classes sous l'hypothèse de répartition uniforme des valeurs à l'intérieur d'une classe).

III.3.1.1. Moyennes marginales.

Les moyennes marginales de Z sont les moyennes des variables marginales X et Y :

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{i=p} n_{i.}\, x_i \; ; \; \overline{Y} = \frac{1}{N} \sum_{j=1}^{j=q} n_{.j}\, y_j \; .$$

III.3.1.2. Variances marginales.

Les variances marginales de Z sont les variances des variables marginales X et Y :

$$s^2(X) = \frac{1}{N} \sum_{i=1}^{i=p} n_{i.} (x_i - \overline{X})^2 = \frac{1}{N} \left(\sum_{i=1}^{i=p} n_{i.}\, x_i^{\,2} - \frac{1}{N} \left(\sum_{i=1}^{i=p} n_{i.}\, x_i \right)^2 \right)$$

$$s^2(Y) = \frac{1}{N} \sum_{j=1}^{j=q} n_{.j} (y_j - \overline{Y})^2 = \frac{1}{N} \left(\sum_{j=1}^{j=q} n_{.j}\, y_j^{\,2} - \frac{1}{N} \left(\sum_{j=1}^{j=q} n_{.j}\, y_j \right)^2 \right)$$

III.3.2. Caractéristiques conditionnelles.

Soit $Z = \{(x_i , y_j), C_{ij} , n_{ij})\}$, $i \in [1, p]$, $j \in [1, q]$, une variable statistique quantitative à deux dimensions, de variables conditionnelles

$$Z_{|Y=y_j} = \{(x_i , C_{ij} , n_{ij})\},\ i \in [1, p], \text{ et } Z_{|X=x_i} = \{(y_j , C_{ij} , n_{ij})\},\ j \in [1, q].$$

avec

$$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} = N$$

III.3.2.1. Moyennes conditionnelles.

Les moyennes conditionnelles de Z sont les moyennes de ses variables conditionnelles :

$\overline{Z_{|Y=y_j}} = \frac{1}{n_{.j}} \sum_{i=1}^{i=p} n_{ij} x_i$, notée aussi, de façon simplifiée, $\overline{X_j}$.

Cette notation simplifiée sera utilisée systématiquement : dans le cas d'une moyenne, l'indice représente toujours le conditionnement.

$$\overline{Z_{|X=x_i}} = \frac{1}{n_{i.}} \sum_{j=1}^{j=q} n_{ij} y_j = \overline{Y_i}$$

III.3.2.2. Variances conditionnelles.

Les variances conditionnelles de *Z* sont les variances de ses variables conditionnelles.

$$s^2(Z|Y=y_j) = \frac{1}{n_{.j}} \sum_{i=1}^{i=p} n_{ij} (x_i - \overline{Z_{|Y=y_j}})^2 = \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij} x_i^2 - \frac{1}{n_{.j}} \left(\sum_{i=1}^{i=p} n_{ij} x_i \right)^2 \right) = s_j^2(X)$$

$$s^2(Z|X=x_i) = \frac{1}{n_{i.}} \sum_{j=1}^{j=q} n_{ij} (y_j - \overline{Z_{|X=x_i}})^2 = \frac{1}{n_{i.}} \left(\sum_{j=1}^{j=q} n_{ij} y_j^2 - \frac{1}{n_{i.}} \left(\sum_{j=1}^{j=q} n_{ij} y_j \right)^2 \right) = s_i^2(Y)$$

Là encore, la notation simplifiée sera utilisée systématiquement : un indice pour la variance représente le conditionnement.

III.3.3. Covariance.

Pour une variable statistique quantitative *Z* à deux dimensions, de variables marginales *X* et *Y*, on définit la covariance de *X* et *Y* par l'expression :

$$Cov(X, Y) = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (x_i - \overline{X})(y_j - \overline{Y})$$

Nous remarquons que la variance a la même dimension qu'une variance.
D'ailleurs, nous avons $Cov(X, X) = s^2(X)$ et $Cov(Y, Y) = s^2(Y)$.
De plus, si l'on remarque que l'on a :

$$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} = N$$
$$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} x_i = \sum_{i=1}^{i=p} n_{i.} x_i = N \overline{X}$$
$$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} y_j = \sum_{j=1}^{j=q} n_{.j} y_j = N \overline{Y}$$

la formule de définition de la covariance peut s'écrire :

$$Cov(X, Y) = \frac{1}{N} \left(\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} x_i y_j - \frac{1}{N} \left(\sum_{i=1}^{i=p} n_{ij} x_i \right) \left(\sum_{j=1}^{j=q} n_{ij} y_j \right) \right) = \overline{XY} - \overline{X}\,\overline{Y}$$

La formule $Cov(X, Y) = \overline{XY} - \overline{X}\,\overline{Y}$ est appelée **formule de la covariance**.

Propriétés de la covariance.

$Cov(a X + b, c Y + d) = a c Cov(X, Y)$, pour a, b, c, d dans $\mathbb{R}$.

En effet :

$\overline{aX+b} = a\,\overline{X} + b,$
$\overline{cY+d} = c\,\overline{Y} + d,$
$\overline{(aX+b)(cY+d)} = a\,c\,\overline{XY} + a\,d\,\overline{X} + b\,c\,\overline{Y} + b\,d.$

$Cov\,(a\,X + b, c\,Y + d) = \overline{(aX+b)(cY+d)} - \overline{aX+b}\;\overline{cY+d}$

$= a\,c\,\overline{XY} + a\,d\,\overline{X} + b\,c\,\overline{Y} + b\,d - (a\,\overline{X} + b)(c\,\overline{Y} + d)$
$= a\,c\,\overline{XY} + a\,d\,\overline{X} + b\,c\,\overline{Y} + b\,d - a\,c\,\overline{X}\,\overline{Y} - b\,c\,\overline{Y} - a\,d\,\overline{X} - b\,d$
$= a\,c\,(\overline{XY} - \overline{X}\,\overline{Y})$
$= a\,c\,Cov\,(X, Y)$

III.3.4. Relations entre caractéristiques marginales et caractéristiques conditionnelles.

III.3.4.1. Moyenne.

La moyenne marginale est la moyenne pondérée des moyennes conditionnelles.

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{i=p} n_{i.}\,x_i = \frac{1}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,x_i = \frac{1}{N}\sum_{j=1}^{j=q}\sum_{i=1}^{i=p} n_{ij}\,x_i = \frac{1}{N}\sum_{j=1}^{j=q} n_{.j}\,\overline{X_j}$$

De même :

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{i=p} n_{i.}\,\overline{Y_i}$$

Nous retrouvons là un résultat déjà établi (Théorème de la moyenne conditionnée, II.2.1.3.b).

III.3.4.2. Variance.

La variance marginale est la somme de la moyenne pondérée des variances conditionnelles et de la variance pondérée des moyennes conditionnelles.

$$s^2(X) = \frac{1}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(x_i - \overline{X})^2 = \frac{1}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(x_i - \overline{X_j} + \overline{X_j} - \overline{X})^2$$
$$= \frac{1}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(x_i - \overline{X_j})^2 + \frac{1}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(\overline{X_j} - \overline{X})^2 + \frac{2}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(x_i - \overline{X_j})(\overline{X_j} - \overline{X})$$

et l'on a :

$$\sum_{i=1}^{i=p} n_{ij}\,(x_i - \overline{X_j})^2 = n_{.j}\,s_j^2(X)$$
$$\sum_{i=1}^{i=p} n_{ij} = n_{.j}$$
$$\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(\overline{X_j} - \overline{X})^2 = \sum_{j=1}^{j=q} n_{.j}\,(\overline{X_j} - \overline{X})^2 = N\,\overline{s^2(X_j)}$$

$$\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(x_i-\overline{X_j})(\overline{X_j}-\overline{X}) = \sum_{j=1}^{j=q}(\overline{X_j}-\overline{X})\sum_{i=1}^{i=p} n_{ij}\,(x_i-\overline{X_j})$$
$$= \sum_{j=1}^{j=q}(\overline{X_j}-\overline{X})\left(\sum_{i=1}^{i=p} n_{ij}\,x_i - \sum_{i=1}^{i=p} n_{ij}\,\overline{X_j}\right)$$
$$= \sum_{j=1}^{j=q}(\overline{X_j}-\overline{X})(n_{.j}\,\overline{X_j} - n_{.j}\,\overline{X_j}) = 0.$$

Il reste donc seulement :

$$s^2(X) = \frac{1}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(x_i-\overline{X_j})^2 + \frac{1}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}\,(\overline{X_j}-\overline{X})^2$$

$$s^2(X) = \frac{1}{N}\sum_{j=1}^{j=q} n_{.j}\,s_j^{\,2}(X) + \frac{1}{N}\sum_{j=1}^{j=q} n_{.j}\,(\overline{X_j}-\overline{X})^2$$

ce qui traduit le résultat annoncé, qui peut s'écrire aussi (Théorème de la variance conditionnée, II.2.2.4.f) :

$$s^2(X) = \overline{s_j^{\,2}(X)} + s^2(\overline{X_j})$$

De même, la variance marginale de Y est donnée par la formule :

$$s^2(Y) = \frac{1}{N}\sum_{i=1}^{i=p} n_{i.}\,s_i^{\,2}(Y) + \frac{1}{N}\sum_{i=1}^{i=p} n_{i.}\,(\overline{Y_i}-\overline{Y})^2$$
$$s^2(Y) = \overline{s_i^{\,2}(Y)} + s^2(\overline{Y_i})$$

Remarque.

La variance traduit la dispersion de la distribution.
La dispersion de la distribution marginale de X résulte de deux facteurs :

— La dispersion des distributions conditionnées autour de leurs moyennes : c'est le premier terme, $\frac{1}{N}\sum_{i=1}^{i=p} n_{i.}\,s_i^{\,2}(Y)$, qu'on appelle la **variance intra-population**, et qu'on note $s_w^{\,2}(Y)$ (*w* pour within).
— La dispersion des moyennes conditionnelles autour de la moyenne : c'est le deuxième terme, $\frac{1}{N}\sum_{i=1}^{i=p} n_{i.}\,(\overline{Y_i}-\overline{Y})^2$, qu'on appelle la **variance inter-population**, et qu'on note $s_b^{\,2}(Y)$ (*b* pour between).

$$s^2(Y) = s_w^{\,2}(Y) + s_b^{\,2}(Y)$$

III. 4. REGRESSION ET CORRELATION.

En présence d'une distribution statistique de deux variables (X, Y), il est possible d'étudier les distributions marginales, les distributions conditionnelles, mais cette étude ne fournit pas d'interprétation des résultats.

Dans certains cas, nous pouvons nous poser la question suivante.
La connaissance d'une modalité de la variable X apporte-t-elle une information supplémentaire sur les modalités de la variable Y ?
La réponse à cette question est du domaine de la **régression** : dans un tel cas, on dit que X est la variable explicative et Y la variable expliquée.

Dans d'autres cas, aucune des deux variables ne peut être privilégiée : la liaison stochastique entre X et Y s'apprécie alors de façon symétrique par la mesure de la **corrélation**.

Exemple : X est la température moyenne mensuelle, Y est le volume des émissions de gaz destiné au chauffage.
Dans cet exemple, X est la variable explicative et Y la variable expliquée.
Il est à noter qu'une variable explicative X peut être une variable qualitative.

III.4.1. Régression et corrélation.

Soient X et Y des variables réelles quantitatives et $Z = (X, Y)$.
Considérons la variable statistique $(X, \overline{Z_{|X}})$ à valeurs dans $\mathbb{R}^2$ définie par :
$\{((x_i , \overline{Y_i}), f_{i.})\}$, $i \in [1, p]$
où $f_{i.} = \frac{n_i}{N}$.
Nous appellerons cette variable la **variable statistique de régression de *Y* en *X***.

III.4.1.1. Courbe de régression.

On appelle **courbe de régression** de Y en X, le graphe, ou courbe représentative, de l'application $f : x \mapsto \overline{Y_i}$.

Si X est une variable discrète, la courbe de régression est une succession de points $(x_i , \overline{Y_i})$.
Si X est une variable continue, la courbe de régression sera formée de segments de droite joignant les points $(x_i , \overline{Y_i})$, où les x_i représentent les centres des classes.
On peut dire que la courbe de régression est la représentation graphique de la variable statistique définie précédemment.

III.4.1.2. Propriétés.

a) Le point moyen de la variable de régression de *Y* en *X* est le point moyen de Z.

En effet :
$\sum f_{i.} x_i = \overline{X}$ et $\sum f_{i.} \overline{Y_i} = \overline{Y} \Rightarrow \sum f_{i.} (x_i , \overline{Y_i}) = (\sum f_{i.} x_i , \sum f_{i.} \overline{Y_i}) = (\overline{X}, \overline{Y}) = \overline{(X,Y)} = \overline{Z}$

b) *Cov* (*X*, $\overline{Z_{|X}}$) = *Cov* (*X*, *Y*).

En effet :

$Cov\ (X, \overline{Z_{|X}}) = \sum f_{i.}\ (x_i - \overline{X})(\overline{Y_i} - \overline{Y})$

$= \sum f_{i.}\ x_i\ \overline{Y_i} - \sum f_{i.}\ x_i\ \overline{Y} - \sum f_{i.}\ \overline{X}\overline{Y_i} + \sum f_{i.}\ \overline{X}\ \overline{Y}$
$= \sum f_{i.}\ x_i\ \overline{Y_i} - \overline{X}\ \overline{Y} - \overline{X}\overline{\overline{Y_i}} + \overline{X}\ \overline{Y}$

$\overline{\overline{Y_i}} = \overline{Y}$

$Cov\ (X, \overline{Z_{|X}}) = \sum f_{i.}\ x_i\ \overline{Y_i} - \overline{X}\ \overline{Y}$

$= \sum_{i=1}^{i=p} f_{i.}\ x_i \sum_{j=1}^{j=q} \frac{f_{ij}}{f_{i.}}\ y_j - \overline{X}\ \overline{Y}$
$= \overline{XY} - \overline{X}\overline{Y}$
$= Cov\ (X, Y)$

c) $s^2\ (\overline{Z_{|X}}) = s_b^{\,2}\ (Y)$.

En effet, comme on a $\overline{\overline{Y_i}} = \overline{Y}$, il résulte de la définition :

$$s^2\ (\overline{Z_{|X}}) = \sum f_{i.}\ (\overline{Y_i} - \overline{Y})^2 = s_b^{\,2}\ (Y)$$

Notons que $s_b^{\,2}\ (Y)$, variance inter-population, n'est pas la variance marginale $s^2\ (Y)$ de Y.

III.4.1.3. Rapport de corrélation.

La variance marginale de Y est donnée par la formule :

$$s^2\ (Y) = s_w^{\,2}\ (Y) + s_b^{\,2}\ (Y)$$

où la variance intra-population $s_w^{\,2}\ (Y)$ est donnée par la formule $s_w^{\,2}\ (Y) = \sum f_{i.}\ s_i^{\,2}\ (Y)$ (moyenne des variances conditionnelles)
et la variance inter-population $s_b^{\,2}\ (Y)$ par la formule $s_b^{\,2}\ (Y) = \sum f_{i.}\ (\overline{Y_i} - \overline{Y})^2$ (variance de la moyenne conditionnelle).

Imaginons une variable $Z = (X, Y)$ pour laquelle $\overline{Z_{|X=x_i}} = \overline{Y_i}$ soit très proche de $\overline{Y}$, pour tout $i \in [1, p]$.
Alors la variance inter-population $s_b^{\,2}\ (Y)$ sera faible et la courbe de régression de Y en X variera peu autour de $\overline{Y}$.
Inversement, si les $\overline{Z_{|X=x_i}}$ sont très dispersés autour de $\overline{Y}$, la variance inter-population $s_b^{\,2}\ (Y)$ sera grande, ce qui veut dire que la courbe de régression de Y en X variera en grandes dents de scie autour de $\overline{Y}$.
Autrement dit, la valeur de la variance inter-population $s_b^{\,2}\ (Y)$ influence directement la courbe de régression.
Nous dirons que $s_b^{\,2}\ (Y)$ est la part de la variance marginale $s^2\ (Y)$ qui est **expliquée par la régression** de Y en X.
Nous parlerons simplement de **variance expliquée**.

Le terme $s_w^2(Y)$, quant à lui, est d'autant plus faible que les $s_i^2(Y)$ sont faibles, donc que les valeurs de Y varient peu, pour chaque x_i, autour de $\overline{Y_i}$.
Ce terme n'a pas d'influence sur la courbe de régression de Y en X (qui fait intervenir seulement les x_i et les $\overline{Y_i}$) : nous l'appelons la **variance résiduelle**.

a) Définition.

Le rapport entre la variance expliquée $s_b^2(Y)$ et la variance marginale totale $s^2(Y)$ est appelé **rapport de corrélation**.
On le note $\eta^2_{Y|X}$:

$$\eta^2_{Y|X} = \frac{s_b^2(Y)}{s^2(Y)}$$

Il peut aussi être calculé par la formule :

$$\eta^2_{Y|X} = 1 - \frac{\text{variance résiduelle}}{\text{variance totale}}.$$

b) Propriétés.

1. $0 \leqslant \eta^2_{Y|X} \leqslant 1$.

Cette propriété résulte directement de la formule de définition $\eta^2_{Y|X} = \frac{s_b^2(Y)}{s^2(Y)}$ et de la formule $s^2(Y) = s_w^2(Y) + s_b^2(Y)$, dans laquelle tous les termes sont positifs.

2. $\eta^2_{Y|X} = 0 \Leftrightarrow s_b^2(Y) = 0 \Leftrightarrow \overline{Y_i} = \overline{Y}$, $\forall\, i \in [1, p]$.

Dans un tel cas, la courbe de régression est parallèle à l'axe des x.
Nous dirons que Y est **non corrélée** avec X : en clair, cela veut dire que la connaissance de X ne donne aucune information sur Y.

Naturellement et de façon symétrique, si l'on a $\eta^2_{X|Y} = 0$, X est non corrélée avec Y et la courbe de régression de X en Y est parallèle à l'axe des y.
Si l'on a, à la fois, $\eta^2_{Y|X} = 0$ et $\eta^2_{X|Y} = 0$, on dit qu'il y a **absence réciproque de corrélation**.

3. $\eta^2_{Y|X} = 1 \Leftrightarrow s_w^2(Y) = 0 \Leftrightarrow y_j = \overline{Y_i}$, $\forall\, i \in [1, p]$, $\forall\, j \in [1, q]$.

Dans un tel cas, à chaque valeur x_i de X correspond une valeur et une seule de Y : il y a une liaison fonctionnelle $Y = f(X)$ entre X et Y.
Si, de plus, on a aussi $\eta^2_{X|Y} = 1$, la liaison fonctionnelle entre X et Y est **biunivoque**.

4. En pratique, nous aurons toujours $0 < \eta^2_{Y|X} < 1$.

Dans ce cas, plus $\eta^2_{Y|X}$ est voisin de 1, plus la dépendance de Y par rapport à X est forte et,

inversement, plus $\eta^2_{Y|X}$ est voisin de 0, moins la dépendance de *Y* par rapport à *X* est forte.

Le rapport de corrélation $\eta^2_{Y|X}$ ne caractérise que l'intensité de la corrélation de *Y* par rapport à *X* et non le sens de la liaison entre les deux.
Il reste invariant si l'on effectue sur *Y* un changement d'origine ou d'échelle.

En effet : $s_b^2(aY+b) = a^2 s_b^2(Y)$ et $s^2(aY+b) = a^2 s^2(Y)$, de sorte que le rapport $\frac{s_b^2(Y)}{s^2(Y)}$ ne change pas.
Comme ce rapport ne tient pas compte de la nature de la courbe de régression, son emploi reste valable quelle que soit la nature de cette courbe de régression.

III.4.1.4. Indépendance et corrélation.

Etant donnée une variable statistique quantitative réelle à deux dimensions $Z = (X, Y)$, nous dirons que la variable statistique ***X* est indépendante de *Y*** si les variables statistiques *Y* et $Z_{|X=x_i}$ ont la même distribution pour tout $i \in [1, p]$, c'est-à-dire si, et seulement si, l'on a :

$$\frac{n_{i1}}{n_1} = \ldots = \frac{n_{ij}}{n_{.j}} = \ldots = \frac{n_{iq}}{n_{.q}}, \forall\, i \in [1, p]$$

Dans ce cas, la valeur commune de ces rapports est :

$$\frac{n_{i1}}{n_1} = \ldots = \frac{n_{ij}}{n_{.j}} = \ldots = \frac{n_{iq}}{n_{.q}} = \frac{n_{i1}+\cdots+n_{iq}}{n_1+\cdots+n_{.q}} = \frac{n_{i.}}{N}$$

et les lignes du tableau de contingence sont proportionnelles.

De façon symétrique, *Y* est indépendante de *X* si, et seulement si, l'on a :

$$\frac{n_{1j}}{n_{1.}} = \ldots = \frac{n_{ij}}{n_{i.}} = \ldots = \frac{n_{pj}}{n_{p.}} = \frac{n_{.j}}{N}, \forall\, j \in [1, q]$$

et, dans ce cas, les colonnes du tableau de contingence sont proportionnelles.

Remarque : *X* est indépendante de *Y* $\Leftrightarrow$ *Y* est indépendante de *X*.

En effet :

X est indépendante de *Y* $\Leftrightarrow \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N}, \forall\, i \in [1, p], \forall\, j \in [1, q]$

$\Leftrightarrow \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N}, \forall\, i \in [1, p], \forall\, j \in [1, q]$

$\Leftrightarrow$ *Y* est indépendante de *X*.

Au lieu de dire "*X* est indépendante de *Y*", on peut donc dire "*X* et *Y* sont indépendantes", la relation est symétrique.

Propriétés.

a) Courbes de régression de variables indépendantes.

Si *X* et *Y* sont indépendantes, les variables statistiques *Y* et $Z_{|X=x_i}$ ont la même distribution pour tout $i \in [1, p]$, elles ont donc la même moyenne, $\overline{Y} = \overline{Y_i}$ pour tout $i \in [1, p]$.
Il en résulte :

$$s_b^{\,2}(Y) = \sum f_{i.} \left(\overline{Y_i} - \overline{Y}\right)^2 = 0$$

$$\eta^2_{Y|X} = \frac{s_b^{\,2}(Y)}{s^2(Y)} = 0$$

De façon symétrique, si X et Y sont indépendantes, Y et X sont indépendantes, les variables statistiques X et $Z_{|Y=y_j}$ ont la même distribution pour tout $j \in [1, q]$, de sorte que l'on a aussi :

$$s_b^{\,2}(X) = \sum f_{.j} \left(\overline{X_j} - \overline{X}\right)^2 = 0$$

$$\eta^2_{X|Y} = \frac{s_b^{\,2}(X)}{s^2(X)} = 0$$

Ainsi, dans le cas où X et Y sont indépendantes, la courbe de régression de Y en X est une parallèle à l'axe des x et la courbe de régression de X en Y est une parallèle à l'axe des y.

On notera que si l'indépendance a pour conséquence le parallélisme des courbes de régression aux axes de coordonnées, en revanche, les courbes de régression peuvent être parallèles aux axes de coordonnées sans que, pour autant, les variables soient indépendantes.
Il ne suffit pas que les moyennes conditionnelles soient identiques pour assurer l'indépendance, il faut encore que les distributions conditionnelles soient identiques. Or plusieurs distributions peuvent avoir la même moyenne sans nécessairement être identiques.

L'absence réciproque de corrélation n'entraîne pas l'indépendance.

Les propriétés du rapport de corrélation peuvent être résumées dans le tableau suivant, qui est un tableau d'équivalence (il se lit dans les deux sens).

$\eta^2_{Y\|X}$ \ $\eta^2_{X\|Y}$	$\eta^2_{X\|Y} = 0$	$0 < \eta^2_{X\|Y} < 1$	$\eta^2_{X\|Y} = 1$
$\eta^2_{Y\|X} = 0$	Absence réciproque de corrélation	Cas général d'absence de corrélation de Y par rapport à X	Liaison fonctionnelle $y \mapsto x$ Absence de corrélation de Y par rapport à X
$0 < \eta^2_{Y\|X} < 1$	Cas général d'absence de corrélation de X par rapport à Y	Cas général	Cas général de liaison fonctionnelle non réciproque $y \mapsto x$
$\eta^2_{Y\|X} = 1$	Liaison fonctionnelle $x \mapsto y$ Absence de corrélation de X par rapport à Y	Cas général de liaison fonctionnelle non réciproque $x \mapsto y$	Liaison fonctionnelle réciproque

b) Critères d'indépendance.

1- Pour que X et Y soient indépendantes, il faut et il suffit que l'on ait :

$$n_{ij} = \frac{n_{i.}\, n_{.j}}{N}, \text{ pour tout } (i, j) \in [1, p] \times [1, q].$$

En effet, la relation précédente peut s'écrire :

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N}, \forall\, i \in [1, p], \forall\, j \in [1, q],$$

ce qui signifie que *X* est indépendante de *Y*.

2- Pour que *X* et *Y* soient indépendantes, il faut et il suffit que l'on ait :

$$f_{ij} = f_{i.}\, f_{.j}, \text{ pour tout } (i, j) \in [1, p] \times [1, q].$$

C'est simplement une autre façon d'écrire le critère précédent, avec

$$f_{ij} = \frac{n_{ij}}{N}, f_{i.} = \frac{n_{i.}}{N}, f_{.j} = \frac{n_{.j}}{N}.$$

c) Si *X* et *Y* sont indépendantes, leur covariance est nulle.

En effet, la covariance de *X* et *Y* est donnée par la formule de la covariance :
$Cov\,(X, Y) = \overline{XY} - \overline{X}\,\overline{Y}$
Lorsque *X* et *Y* sont indépendantes, nous avons :
$\overline{XY} = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} f_{ij}\, x_i\, y_j = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} f_{i.}\, f_{.j}\, x_i\, y_j = \sum_{i=1}^{i=p} f_{i.}\, x_i \sum_{j=1}^{j=q} f_{.j}\, y_j = \overline{X}\,\overline{Y}$
de sorte que la covariance est nulle.

La réciproque est fausse : la covariance peut être nulle sans que les variables soient indépendantes.

III. 4. 2. Méthode des moindres carrés.

III.4.2.1. Propriété de la courbe de régression.

Soit $Z = \{(x_i , y_j), C_{ij} , n_{ij})\}$, $i \in [1, p]$, $j \in [1, q]$, une variable statistique quantitative à deux dimensions, de variables marginales
$X = \{(x_i , C_{i.} , n_{i.})\}$, $i \in [1, p]$, et $Y = \{(y_j , C_{.j} , n_{.j})\}$, $j \in [1, q]$.
$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} = N.$

Pour chaque valeur x_i de X, on sait calculer la moyenne conditionnelle de Y pour X fixé :
$\overline{Y_i} = \frac{1}{n_{i.}} \sum_{j=1}^{j=q} n_{ij} y_j$, avec $n_{i.} = \sum_{j=1}^{j=q} n_{ij}$, pour tout $i \in [1, p]$.
La courbe de régression de Y en X joint les points R_i de coordonnées $(x_i , \overline{Y_i})$, $i \in [1, p]$.

Pour tout $i \in [1, p]$, considérons un point $A_i = (x_i, y'_i)$.

On appelle **somme des carrés des écarts**, en abrégé SCE, l'expression :
$S = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (y'_i - y_j)^2$
et **carré moyen**, en abrégé CM, l'expression :
$CM = \frac{S}{N} = \frac{1}{N} \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (y'_i - y_j)^2 = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} f_{ij} (y'_i - y_j)^2$

La somme des carrés des écarts s'écrit :
$S = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (y'_i - \overline{Y_i} + \overline{Y_i} - y_j)^2$
$= \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (y'_i - \overline{Y_i})^2 + \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (\overline{Y_i} - y_j)^2 + 2 \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (y'_i - \overline{Y_i})(\overline{Y_i} - y_j)$
$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (y'_i - \overline{Y_i})(\overline{Y_i} - y_j) = \sum_{i=1}^{i=p} (y'_i - \overline{Y_i}) \sum_{j=1}^{j=q} n_{ij} (\overline{Y_i} - y_j) = \sum_{i=1}^{i=p} (y'_i - \overline{Y_i}) (n_{i.} \overline{Y_i} - n_{i.} \overline{Y_i}) = 0$
$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (\overline{Y_i} - y_j)^2 = \sum_{i=1}^{i=p} n_{i.} s_i^2 (Y)$
$\sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij} (y'_i - \overline{Y_i})^2 = \sum_{i=1}^{i=p} n_{i.} (y'_i - \overline{Y_i})^2$
$S = \sum_{i=1}^{i=p} n_{i.} (y'_i - \overline{Y_i})^2 + \sum_{i=1}^{i=p} n_{i.} s_i^2 (Y)$
Le terme $\sum_{i=1}^{i=p} n_{i.} s_i^2 (Y)$ ne dépend pas du choix des y'_i.

S prendra donc une valeur minimum, lorsque $\sum_{i=1}^{i=p} n_{i.} (y'_i - \overline{Y_i})^2$ est nul, c'est-à-dire lorsque $y'_i = \overline{Y_i}$ pour tout $i \in [1, p]$.
Autrement dit :

La courbe de régression est la ligne qui rend minimum la somme des carrés des écarts.

C'est donc celle qui ajuste au mieux une courbe au nuage de points (x_i, y_j).
Pour cette courbe, le carré moyen (CM, en abrégé), prend aussi sa valeur minimum, qui est donnée par :

$$CM = \frac{1}{N} \sum_{i=1}^{i=p} n_{i.}\, s_i^{\,2}(Y) = s_w^{\,2}(Y)$$

Le carré moyen correspondant à la ligne de régression est la **variance résiduelle**.

III.4.2.2. Ajustement linéaire.

Si la ligne de régression de Y en X tracée sur le nuage de points (x_i, y_j) se rapproche globalement d'une droite, nous pouvons chercher directement, par la **méthode des moindres carrés ordinaires**, en abrégé MCO, la droite qui s'ajuste le mieux au nuage de points.
Soit $y = a + b\,x$ l'équation d'une droite.
Pour tout $i \in [1, p]$, considérons le point $A_i = (x_i, y'_i = a + b\,x_i)$ de la droite.
On peut associer à la droite la somme des carrés des écarts :

$$S = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij}\,(y'_i - y_j)^2 = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} n_{ij}\,(a + b\,x_i - y_j)^2$$

Le carré moyen associé est :

$$CM = \frac{S}{N} = \sum_{i=1}^{i=p} \sum_{j=1}^{j=q} f_{ij}\,(a + b\,x_i - y_j)^2$$

C'est la moyenne du carré de $(a + b\,X - Y)$.
Or la variance de $(a + b\,X - Y)$ est égale à la moyenne du carré, moins le carré de la moyenne,

$$s^2(a + b\,X - Y) = CM - \overline{a + b\,X - Y}^{\,2}$$

On obtient donc :

$$CM = \overline{a + b\,X - Y}^{\,2} + s^2(a + b\,X - Y) = (a + b\,\overline{X} - \overline{Y})^2 + s^2(a + b\,X - Y)$$

On sait, par ailleurs, que la variance de $a + b\,X - Y$ est donnée par :

$$s^2(a + b\,X - Y) = s^2(b\,X - Y) = b^2\,s^2(X) - 2\,b\,Cov(X, Y) + s^2(Y)$$

On peut écrire aussi :

$$b^2\,s^2(X) - 2\,b\,Cov(X, Y) + s^2(Y) = s^2(X)\left(b^2 - 2\,b\,\frac{Cov(X,Y)}{s^2(X)} \right) + s^2(Y)$$

$$= s^2(X)\left(b - \frac{Cov(X,Y)}{s^2(X)} \right)^2 + s^2(Y) - \frac{Cov^2(X,Y)}{s^2(X)}$$

$$= s^2(X)\left(b - \frac{Cov(X,Y)}{s^2(X)} \right)^2 + s^2(Y)\left(1 - \left(\frac{Cov(X,Y)}{s(X)\,s(Y)} \right)^2 \right)$$

Or, la variance $b^2\,s^2(X) - 2\,b\,Cov(X, Y) + s^2(Y)$ de $b\,X - Y$ est positive pour tout $b \in \mathbb{R}$, puisque toute variance est positive.
Donc le discriminant réduit de ce polynôme de degré 2 en b est négatif : $Cov^2(X, Y) \leqslant s^2(X)\,s^2(Y)$, et, dans l'expression

$$s^2(X)\left(b - \frac{Cov(X,Y)}{s^2(X)} \right)^2 + s^2(Y)\left(1 - \left(\frac{Cov(X,Y)}{s(X)\,s(Y)} \right)^2 \right)$$

le terme $s^2(Y)\left(1 - \left(\frac{Cov(X,Y)}{s(X)\,s(Y)} \right)^2 \right)$, qui ne dépend pas du choix de a et b, est toujours positif.

La conclusion est que le carré moyen s'écrit finalement comme somme de trois termes positifs dont le troisième ne dépend ni de a ni de b :

$$CM = (a + b\,\overline{X} - \overline{Y})^2 + s^2(X)\left(b - \frac{Cov(X,Y)}{s^2(X)} \right)^2 + s^2(Y)\left(1 - \left(\frac{Cov(X,Y)}{s(X)\,s(Y)} \right)^2 \right)$$

Cette somme prend sa valeur minimum lorsque les deux premiers termes sont nuls :

$$a + b\,\overline{X} - \overline{Y} = 0$$

$b = \dfrac{Cov(X,Y)}{s^2(X)}$

L'équation de la **droite ajustée par la méthode des moindres carrés** est donc :

$$\boxed{(y - \overline{Y}) = (x - \overline{X})\,\frac{Cov(X,Y)}{s^2(X)}}$$

La valeur de b obtenue est aussi celle qui rend minimum la variance

$s^2(a + b\,X - Y) = s^2(X)\left(b - \dfrac{Cov(X,Y)}{s^2(X)} \right)^2 + s^2(Y)\left(1 - \left(\dfrac{Cov(X,Y)}{s(X)\,s(Y)} \right)^2 \right)$.

Nous noterons $(X, \Delta_{Y|X})$ la variable statistique $\{((x_i, a + b\,x_i), f_{i.})\}$, $i \in [1, p]$.

Cette variable statistique est appelée la **variable statistique de régression linéaire de *Y* en *X***.

La représentation graphique de cette variable est donnée par la **droite ajustée par la méthode des moindres carrés ordinaires**.

Cette droite est parfois appelée la *droite de régression* de Y en X.

Le coefficient b est alors appelé le *coefficient de régression* de Y en X.

Il vaut mieux réserver ces dénominations à la droite de régression du modèle théorique probabiliste associé à la population et parler, ici, seulement de droite ajustée par la méthode des moindres carrés ordinaires.

Propriétés de la variable statistique $(X, \Delta_{Y|X})$.

1. Le point moyen est celui de *Z*.

En effet, on a : $\sum f_{i.}\,x_i = \overline{X}$ et $\sum f_{i.}\,(a + b\,x_i) = a + b\,\overline{X} = \overline{Y}$.

La relation $a + b\,\overline{X} = \overline{Y}$ montre que la droite ajustée par la méthode des moindres carrés ordinaires passe par ce point moyen $(\overline{X}, \overline{Y})$.

2. $Cov\,(X, \Delta_{Y|X}) = Cov\,(X, Y)$.

En effet :

$Cov\,(X, \Delta_{Y|X}) = \sum_{i=1}^{i=p} f_{i.}\,(x_i - \overline{X})(a + b\,x_i - (a + b\,\overline{X}))$

$= b \sum_{i=1}^{i=p} f_{i.}\,(x_i - \overline{X})^2$

$= b\,s^2(X)$

$= Cov\,(X, Y)$

puisque $b = \dfrac{Cov(X,Y)}{s^2(X)}$.

3. $s^2(\Delta_{Y|X}) = b^2\,s^2(X) \neq s^2(Y)$.

En effet, par définition : $s^2(\Delta_{Y|X}) = s^2(a + b\,X)$

et comme on a toujours $s^2(a + b\,X) = b^2\,s^2(X)$, il vient $s^2(\Delta_{Y|X}) = b^2\,s^2(X) = \left(\dfrac{Cov(X,Y)}{s(X)} \right)^2$

En général, $b^2 s^2(X)$ est différent de $s^2(Y)$, sinon on aurait $s^2(Y) = b^2 s^2(X) = \left(\frac{Cov(X,Y)}{s(X)}\right)^2$, donc :

$Cov(X, Y) = s(X)\, s(Y)$ ou $Cov(X, Y) = -s(X)\, s(Y)$
Dans le premier cas, la variance de $a + b X - Y$ est nulle :

$$s^2(a + b X - Y) = s^2(Y)\left(1 - \left(\frac{Cov(X,Y)}{s(X)\,s(Y)}\right)^2\right) = 0$$

et $Y = a + b X$, avec $b = \frac{Cov(X,Y)}{s^2(X)} = \frac{s(Y)}{s(X)} > 0$.

Dans le deuxième cas, la variance de $a + b X - Y$ est nulle aussi et $Y = a + b X$, avec $b = \frac{Cov(X,Y)}{s^2(X)} = -\frac{s(Y)}{s(X)} < 0$.

Variable statistique $(Y, \Delta_{X|Y})$.

C'est la variable statistique associée à la régression de X en Y.
L'équation de la droite ajustée par la méthode des moindres carrés ordinaires aux couples (y_j, x_i) a pour équation :

$$(x - \overline{X}) = (y - \overline{Y})\,\frac{Cov(X,Y)}{s^2(Y)}$$

Nous avons les propriétés suivantes, analogues aux précédentes :
$Cov(Y, \Delta_{X|Y}) = Cov(Y, X) = Cov(X, Y)$

$$s^2(\Delta_{X|Y}) = \frac{Cov^2(X,Y)}{s^2(Y)} \neq s^2(X)$$

III.4.2.3. Coefficient de corrélation linéaire.

Les variables $(X, \Delta_{Y|X})$ et $(Y, \Delta_{X|Y})$ représentent un résumé de la variable $Z = (X, Y)$.
Il est nécessaire de définir un nouveau paramètre pour mesurer la validité de ce résumé.
On appelle coefficient de corrélation linéaire le rapport :

$$r = \frac{Cov(X,Y)}{s(X)\,s(Y)}$$

Propriétés du coefficient de corrélation linéaire.

1. Coefficient de corrélation linéaire et rapport de corrélation.

Le carré du coefficient de corrélation linéaire, qu'on appelle aussi le **coefficient de détermination**, est donné par la formule :

$$r^2 = \left(\frac{Cov(X,Y)}{s(X)\,s(Y)}\right)^2 = \frac{s^2(\Delta_{Y|X})}{s^2(Y)} = \frac{s^2(\Delta_{X|Y})}{s^2(X)}$$

Il détermine la part de variance de Y qui est expliquée par la régression linéaire de Y en X (ou, respectivement, la part de variance de X expliquée par la régression linéaire de X en Y).
Le coefficient de détermination joue donc, pour la régression linéaire de Y en X, le même rôle que le

rapport de corrélation pour la régression de Y en X.
En particulier, pour la ligne de régression de Y en X, nous avions trouvé, pour carré moyen minimum, la variance résiduelle

$$s_w^2(Y) = (1 - \eta^2_{Y|X})\, s^2(Y).$$

Pour la régression linéaire de Y en X, la valeur minimum du carré moyen est $(1 - r^2)\, s^2(Y)$.
Cette valeur minimum est nécessairement plus grande que la variance résiduelle, qui est un minimum absolu :

$$0 \leqslant (1 - \eta^2_{Y|X})\, s^2(Y) \leqslant (1 - r^2)\, s^2(Y) \leqslant s^2(Y)$$
$$0 \leqslant (1 - \eta^2_{Y|X}) \leqslant (1 - r^2) \leqslant 1$$
$$0 \leqslant r^2 \leqslant \eta^2_{Y|X} \leqslant 1$$

En particulier, le coefficient de corrélation linéaire r est compris entre –1 et 1 :

$$-1 \leqslant r \leqslant 1.$$

L'égalité de r^2 et de $\eta^2_{Y|X}$ traduit la propriété que la ligne de régression de Y en X est une droite ; on dit alors que Y présente une *corrélation linéaire* avec X.

2. Cas où $r = 0$.

S'il n'y a pas de corrélation entre Y et X, $\eta^2_{Y|X}$ est nul donc aussi $r = 0$.
Dans ce cas, les droites de régression sont parallèles aux axes.
Nous ne pouvons pas en conclure l'indépendance de X et de Y.

3. Cas où $r^2 = 1$.

Si $r^2 = 1$, alors $\eta^2_{Y|X} = 1$, il y a une relation fonctionnelle liant X et Y.
Et cette relation fonctionnelle est linéaire.
En effet, dire que $r^2 = 1$, c'est dire que $Cov^2(X, Y) = s^2(X)\, s^2(Y)$.
Dans ce cas :

$$s^2(a + b\,X - Y) = s^2(X)\left(b - \frac{Cov(X,Y)}{s^2(X)} \right)^2 + s^2(Y)\left(1 - \left(\frac{Cov(X,Y)}{s(X)\,s(Y)} \right)^2 \right).$$

se réduit, avec $b = \frac{Cov(X,Y)}{s^2(X)}$, à $s^2(a + b\,X - Y) = 0$, ce qui veut dire que tous les points sont sur la droite ajustée par la méthode des moindres carrés : il existe une relation fonctionnelle linéaire entre X et Y, $Y = a + b\,X$, avec $b > 0$ si $r = 1$, et $b < 0$ si $r = -1$.
Plus r est proche de 1 ou de –1, plus la corrélation linéaire est forte.

III.4.2.4. Prédicteur et estimation.

En l'absence d'information, l'estimation la meilleure que nous puissions donner d'une valeur inconnue prise par Y est sa moyenne .
Si Y est en corrélation avec X, la connaissance de la valeur x_i de X, permet d'améliorer l'estimation de Y.
Nous dirons que $\overline{z_{|X}}$ et $\Delta_{Y|X}$ sont des **prédicteurs** de Y.

Nous avons :
$m\,(\overline{Z_{|X}}) = \overline{Y}$ et $m\,(\Delta_{Y|X}) = \overline{Y}$
$s^2\,(\overline{Z_{|X}}) = s_b{}^2\,(Y) = \eta^2{}_{Y|X}\,s^2\,(Y)$ et $s^2\,(\Delta_{Y|X}) = r^2\,s^2\,(Y)$

La mesure de la validité d'un prédicteur de *Y* se mesure par le rapport de sa variance à la variance de *Y* :

$$\frac{s^2\left(\Delta_{Y|X}\right)}{s^2(Y)} = r^2 \text{ et } \frac{s^2\left(\overline{Z_{|X}}\right)}{s^2(Y)} = \eta^2{}_{Y|X}$$

Plus le rapport est proche de 1, plus la variance du prédicteur est proche de la variance de *Y*, donc plus la variance résiduelle est faible et moins le nuage de points est dispersé autour du prédicteur, donc meilleur est le prédicteur.
$\eta^2{}_{Y|X}$ ou r^2 mesure donc la précision du prédicteur et nous pouvons dire que $\overline{Z_{|X}}$ est un prédicteur meilleur que $\Delta_{Y|X}$, puisque $\eta^2{}_{Y|X}$ est plus grand que r^2.

III.4.2.5. Généralisation du modèle.

L'ajustement linéaire peut, par des changements de variables, permettre l'ajustement d'autres modèles non linéaires.

1. Modèle exponentiel.

Si l'étude de la corrélation entre *Y* et *X* met en évidence que le taux de variation instantané de *Y* par rapport à *X* est constant (*X* pouvant être la variable "temps", dans le cas d'une chronique, ou série chronologique), alors nous avons, théoriquement :
$\frac{dy}{y} = k\,dx$, soit $y = y_0\,c^x$.
En posant $z = ln\,y$, $a = ln\,y_0$, $b = ln\,c$, il vient $z = a + b\,x$.
On est ramené à un modèle linéaire.
Dans la pratique, on vérifie si le taux de variation expérimental est sensiblement constant en calculant, pour chaque intervalle Δx le rapport $\frac{1}{\Delta x}\frac{\Delta y}{y}$.
La mise en évidence de ce modèle est obtenue en utilisant un **papier semi-logarithmique**, avec une échelle logarithmique en ordonnée et une échelle arithmétique en abscisse.
Un tel modèle est très utilisé en matière économique : étude des fonctions de production, de consommation, étude du chiffre d'affaire, etc.

2. Modèle à élasticité constante.

Si l'étude de la corrélation entre *Y* et *X* met en évidence que l'élasticité est constante, nous avons théoriquement (l'élasticité est le rapport entre la variation relative de *y* et la variation relative de *x*) :
$\frac{dy}{y} = k\,\frac{dx}{x}$, soit $y = y_0\,x^b$.
Si nous posons $z = ln\,y$, $t = ln\,x$, $a = ln\,y_0$, nous avons $z = a + b\,t$.
On est ramené à un modèle linéaire.
Dans la pratique, on vérifie que l'élasticité est constante en calculant, pour chaque intervalle Δx, le rapport $\frac{x}{\Delta x}\frac{\Delta y}{y}$.

La mise en évidence de ce modèle est obtenue en utilisant un **papier log-log**, avec une échelle logarithmique en abscisses et une échelle logarithmique en ordonnées.
Un tel modèle est, lui aussi, très utilisé en matière économique : étude des dépenses pour un poste particulier relativement aux dépenses totales du ménage.

Chapitre 4 - REGRESSION ORTHOGONALE DANS R².

4. 1. NOTION D'ESPACE VECTORIEL EUCLIDIEN.

4.1.1. Espace vectoriel $\mathbb{R}^n$.

Soit n un entier strictement positif et $\mathbb{R}$ le corps des nombres réels.
L'ensemble $\mathbb{R}^n$ des n-uples $(x_1, \dots, x_n)$ de nombres réels est muni de sa structure usuelle d'**espace vectoriel réel**, définie par les opérations :

$$(x_1, \dots, x_n) + (x'_1, \dots, x'_n) = (x_1 + x'_1, \dots, x_n + x'_n)$$
$$\lambda\,(x_1, \dots, x_n) = (\lambda\, x_1, \dots, \lambda\, x_n),\ \forall\ \lambda \in \mathbb{R}.$$

Notations.

On identifiera un élément $X = (x_1, \dots, x_n)$ de $\mathbb{R}^n$ avec la matrice $X = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix}$ à n lignes et 1 colonne.

La transposée de cette matrice est la matrice ${}^tX = \begin{pmatrix} x_1 \dots x_n \end{pmatrix}$ à 1 ligne et n colonnes.

Les opérations dans $\mathbb{R}^n$ sont alors définies par des opérations sur les matrices :

<u>Addition :</u>

$$\begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} x'_1 \\ \vdots \\ x'_i \\ \vdots \\ x'_n \end{pmatrix} = \begin{pmatrix} x_1 + x'_1 \\ \vdots \\ x_i + x'_i \\ \vdots \\ x_n + x'_n \end{pmatrix}$$
$$\begin{pmatrix} x_1 \dots x_n \end{pmatrix} + \begin{pmatrix} x'_1 \dots x'_n \end{pmatrix} = \begin{pmatrix} x_1 + x'_1 \dots x_n + x'_n \end{pmatrix}$$

<u>Multiplication par un scalaire :</u>

$$\lambda \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \vdots \\ \lambda x_i \\ \vdots \\ \lambda x_n \end{pmatrix}.$$
$$\lambda \begin{pmatrix} x_1 \dots x_n \end{pmatrix} = \begin{pmatrix} \lambda\, x_1 \dots \lambda\, x_n \end{pmatrix}$$

Dans $\mathbb{R}^n$, les n éléments e_i, $i \in \{1, \dots, n\}$, dont toutes les coordonnées sont nulles, sauf la i^e qui vaut

1, forment une base, appelée la **base canonique** de $\mathbb{R}^n$.

Tout élément $X = (x_1, \ldots, x_n)$ de $\mathbb{R}^n$ s'écrit de manière unique sous la forme

$$X = \sum_{i=1}^{i=n} x_i \, e_i$$

4.1.2. Produit scalaire dans $\mathbb{R}^n$.

Soit Φ une application de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}$.
On notera aussi $< X \mid \Phi \mid Y >$ ou $< X \mid Y >_\Phi$, le nombre réel $\Phi (X, Y)$.

4.1.2.1. Définition.

On appelle **produit scalaire** dans $\mathbb{R}^n$ toute application Φ de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}$ qui possède les propriétés suivantes :

a) Bilinéarité.

— Linéarité par rapport à la première variable :
$\Phi (X + X', Y) = \Phi (X, Y) + \Phi (X', Y)$ et $\Phi (\lambda X, Y) = \lambda \Phi (X, Y)$, quels que soient λ dans $\mathbb{R}$, X, X' et Y dans $\mathbb{R}^n$;
cette propriété s'écrit aussi

$$< X + X' \mid \Phi \mid Y > = < X \mid \Phi \mid Y > + < X' \mid \Phi \mid Y >$$

— Linéarité par rapport à la deuxième variable :
$\Phi (X, Y + Y') = \Phi (X, Y) + \Phi (X, Y')$ et $\Phi (X, \lambda Y) = \lambda \Phi (X, Y)$, quels que soient λ dans $\mathbb{R}$, X, Y et Y' dans $\mathbb{R}^n$;
cette propriété s'écrit aussi

$$< X \mid \Phi \mid Y + Y' > = < X \mid \Phi \mid Y > + < X \mid \Phi \mid Y' >$$

b) Symétrie.

$\Phi (X, Y) = \Phi (Y, X)$, quels que soient X et Y dans $\mathbb{R}^n$:

$$< X \mid \Phi \mid Y > = < Y \mid \Phi \mid X >$$

c) Positivité.

$\Phi (X, X)$ est un nombre réel supérieur ou égal à 0, quel que soit X dans $\mathbb{R}^n$:

$$< X \mid \Phi \mid X > \geq 0$$

d) Non dégénérescence.

$\Phi (X, X) = 0$ entraîne $X = 0$:

$$< X \mid \Phi \mid X > = 0 \Rightarrow X = 0.$$

Autrement dit, le vecteur 0 = (0, ... , 0, ... , 0) de $\mathbb{R}^n$ est l'unique solution de l'équation $\Phi(X, X) = 0$.

On dit aussi qu'un produit scalaire sur $\mathbb{R}^n$ est une **forme bilinéaire symétrique positive non dégénérée**.
Le mot "forme" fait simplement référence au fait que les valeurs sont des scalaires.
Lorsqu'il est muni d'un produit scalaire, $\mathbb{R}^n$ est appelé un **espace vectoriel euclidien**.

4.1.2.2. Exemples.

a) Produit scalaire canonique.

L'application de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}$ définie par :

$$((x_1, \dots, x_n), (y_1, \dots, y_n)) \mapsto < X \mid Y > = {}^t X\, Y = \begin{pmatrix} x_1 & \dots & x_j & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{i=n} x_i\, y_i$$

est un produit scalaire sur $\mathbb{R}^n$ qu'on appelle le **produit scalaire canonique** de $\mathbb{R}^n$.
En effet, les propriétés de bilinéarité, de symétrie, de positivité et de non dégénérescence sont pratiquement évidentes à vérifier.

b) Produit scalaire défini par une matrice diagonale à éléments positifs.

Considérons une matrice réelle *M* à *n* lignes et *n* colonnes dont tous les éléments en dehors de la diagonale principale sont nuls ($m_{ij} = 0$, quels que soient les entiers *i* et *j* dans $\{1, \dots, n\}$ avec $i \neq j$) (on dit alors que *M* est une **matrice diagonale**) et dont les éléments de la diagonale principale sont des **nombres réels strictement positifs** ($m_{ii} > 0$ quel que soit l'entier *i* dans $\{1, \dots, n\}$).

Alors l'application :

$$(X, Y) \mapsto < X \mid M \mid Y > = {}^t X\, M\, Y = \begin{pmatrix} x_1 & \dots & x_j & \dots & x_n \end{pmatrix} M \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \sum_{ij} m_{ij}\, x_j\, y_i = \sum_i m_{ii}\, x_i\, y_i$$

est un produit scalaire sur $\mathbb{R}^n$. La matrice *M* est appelée la **matrice des poids** (les "poids" sont les éléments de la diagonale).

En effet, les propriétés de bilinéarité, de symétrie, de positivité et de non dégénérescence sont pratiquement évidentes à vérifier.

- Le produit scalaire canonique correspond au cas où la matrice *M* est la matrice unité $\boldsymbol{I}_n$ (tous les éléments de la diagonale sont égaux à 1 et les éléments en dehors de la diagonale sont 0) : tous les poids sont égaux à 1.
- Autre exemple : $M = D_{\frac{1}{n}} = \frac{1}{n}\boldsymbol{I}_n$. Tous les poids sont égaux à $\frac{1}{n}$ et la somme des poids vaut 1.

4.1.2.3. Propriétés.

a) Matrice d'un produit scalaire.

Pour tout produit scalaire Φ sur $\mathbb{R}^n$, on peut écrire :

$$\Phi(X, Y) = \Phi\left(\sum_i x_i e_i, \sum_j y_j e_j\right) = \sum_{ij} \Phi(e_i, e_j)\, x_i y_j = \begin{pmatrix} x_1 \dots x_i \dots x_n \end{pmatrix} M_\Phi \begin{pmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_n \end{pmatrix}$$

La matrice $M_\Phi = [\Phi(e_i, e_j)]$ s'appelle la **matrice du produit scalaire** Φ dans la base canonique.
Cette matrice est une matrice symétrique : $\Phi(e_i, e_j) = \Phi(e_j, e_i)$.
Les éléments de sa diagonale sont des nombres réels strictement positifs : $\Phi(e_i, e_i) > 0$.
Remarquons ces propriétés ne sont pas suffisantes : une matrice symétrique dont les éléments de la diagonale sont des nombres réels strictement positifs ne définit pas forcément un produit scalaire.

Par exemple, la matrice $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ a un déterminant qui vaut $-3 < 0$, donc elle possède deux valeurs propres réelles de signe opposé (3 et – 1) et la forme bilinéaire $((x_1, x_2),(y_1, y_2)) \mapsto (x_1, x_2)\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ qu'elle définit n'est pas un produit scalaire car le "produit scalaire" du vecteur propre (1, – 1) pour la valeur propre négative, par lui-même, est un nombre réel strictement négatif $((1\ -1)\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = -2)$.
La matrice $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ n'est donc pas la matrice d'un produit scalaire sur $\mathbb{R}^2$, bien qu'elle soit symétrique et que les éléments de sa diagonale soient strictement positifs.

En réalité, pour qu'une matrice carrée symétrique réelle soit la matrice d'un produit scalaire, il faut et il suffit que toutes ses valeurs propres, qui sont toujours des nombres réels, soient strictement positives. Ce résultat sera démontré, dans sa généralité, en analyse.

b) Norme d'un vecteur.

Si Φ est un produit scalaire sur $\mathbb{R}^n$, le nombre réel positif $\| X \|_\Phi = \sqrt{\Phi(X, X)}$ s'appelle la **Φ-norme** de X, ou **Φ-longueur** de X.
Quand il n'y a pas de confusion à craindre, on parlera simplement de norme ou de longueur, qu'on notera $\| X \|$ au lieu de $\| X \|_\Phi$.

On dit qu'un vecteur est **normé pour Φ** si sa Φ-longueur est 1.

Par exemple, dans $\mathbb{R}^2$ muni du produit scalaire canonique, la longueur de $X = (x_1, x_2)$ est $\| X \| = \sqrt{x_1^2 + x_2^2}$ et le vecteur (1, 0) est normé.

c) Angle de deux vecteurs.

Etant donnés deux vecteurs X et Y de $\mathbb{R}^n$ et un produit scalaire Φ sur $\mathbb{R}^n$, pour tout nombre réel λ, on a :

$$\Phi (X + \lambda Y, X + \lambda Y) = \| X + \lambda Y \|_\Phi ^2 \geq 0$$
$$\lambda ^2 \Phi (Y, Y) + \lambda (\Phi (Y, X) + \Phi (X, Y)) + \Phi (X, X) \geq 0$$
$$\lambda ^2 \Phi (Y, Y) + 2 \lambda \Phi (X, Y) + \Phi (X, X) \geq 0$$
$$\| Y \|_\Phi ^2 \lambda ^2 + 2 < X \mid Y >_\Phi \lambda + \| X \|_\Phi ^2 \geq 0$$

Comme cette relation est vraie pour tout nombre réel λ, c'est que le discriminant de ce trinôme du deuxième degré est négatif :

$$(< X \mid Y >_\Phi) ^2 - \| X \|_\Phi ^2 \| Y \|_\Phi ^2 \leq 0$$
$$| < X \mid Y >_\Phi | \leq \| X \|_\Phi \| Y \|_\Phi$$

Cette inégalité, valable pour tous vecteurs X et Y de $\mathbb{R}^n$ constitue l'**inégalité de Schwarz**.
Si les deux vecteurs X et Y sont différents de 0, leur longueur n'est pas nulle, le produit de leurs longueurs n'est pas nul, le rapport $\frac{\langle X|Y\rangle_\Phi}{\|X\|_\Phi \|Y\|_\Phi}$ est compris entre –1 et 1, et il existe donc un angle compris entre 0 et π radians dont le cosinus est égal au rapport $\frac{\langle X|Y\rangle_\Phi}{\|X\|_\Phi \|Y\|_\Phi}$.

Par définition, cet angle unique α compris entre 0 et π, vérifiant :

$$\cos \alpha = \frac{\langle X|Y\rangle_\Phi}{\|X\|_\Phi \|Y\|_\Phi} = \frac{\langle X|\Phi|Y\rangle}{\|X\|_\Phi \|Y\|_\Phi}$$

est appelé l'**angle des deux vecteurs non nuls** X et Y.

d) Orthogonalité.

Etant donnés deux vecteurs X et Y de $\mathbb{R}^n$ et un produit scalaire Φ sur $\mathbb{R}^n$, on dit que X et Y sont **Φ-orthogonaux** (ou simplement "orthogonaux" s'il n'y a pas de confusion à craindre) si, et seulement si, leur produit scalaire est nul :

$$\Phi (X, Y) = < X \mid Y >_\Phi = 0$$

Exemples :
— 0 est Φ-orthogonal à tout vecteur de $\mathbb{R}^n$.
— L'angle de deux vecteurs non nuls Φ-orthogonaux est $\frac{\pi}{2}$.

— La base canonique de $\mathbb{R}^n$ muni du produit scalaire canonique est formée de vecteurs normés orthogonaux deux à deux : on parle alors de **base orthonormée**.

e) Projeté orthogonal.

Soient X et Y deux vecteurs non nuls de $\mathbb{R}^n$ et Φ un produit scalaire sur $\mathbb{R}^n$.
Il existe un unique vecteur Z de $\mathbb{R}^n$, proportionnel à Y et tel que $X – Z$ soit orthogonal à Y.

Démonstration.

Pour tout vecteur Z on peut écrire :

$$< X - Z \mid Y >_\Phi = < X \mid Y >_\Phi - < Z \mid Y >_\Phi$$

Si l'on prend un Z proportionnel à Y, on a $Z = a\,Y$, donc :

$$< X - Z \mid Y >_\Phi = < X \mid Y >_\Phi - a < Y \mid Y >_\Phi = < X \mid Y >_\Phi - a \,\| Y \|_\Phi^2.$$

Pour que $X - Z$ soit orthogonal à Y., soit $< X - Z \mid Y >_\Phi = 0$, il faut et il suffit que l'on prenne $a = \frac{\langle X|Y\rangle_\Phi}{\|Y\|_\Phi^2}$.

L'unique vecteur $Z = \frac{\langle X|Y\rangle_\Phi}{\|Y\|_\Phi^2}\ Y$, proportionnel à Y et tel que $X - Z$ soit orthogonal à Y, s'appelle le **projeté orthogonal** de X sur Y.

Propriété du projeté orthogonal.

Le projeté orthogonal Z_0 de X sur Y est le vecteur Z de $\mathbb{R}^n$ proportionnel à Y, qui minimise $\| X - Z \|_\Phi^2$.

Démonstration.

Soit Z un vecteur proportionnel à Y.
Soit $Z_0 = \frac{\langle X|Y\rangle_\Phi}{\|Y\|_\Phi^2}\ Y$ le projeté orthogonal de X sur Y.

$$\| X - Z \|_\Phi^2 = \| X - Z_0 + Z_0 - Z \|_\Phi^2 .$$

Comme Z est proportionnel à Y et que Z_0 est proportionnel à Y, la différence $Z_0 - Z$ est proportionnelle à Y.
Or $X - Z_0$ est orthogonal à Y, donc $X - Z_0$ est orthogonal à $Z_0 - Z$ qui est proportionnel à Y.
Il est résulte que l'on a :

$$\| X - Z \|_\Phi^2 = \| X - Z_0 + Z_0 - Z \|_\Phi^2 = \| X - Z_0 \|_\Phi^2 + \| Z_0 - Z \|_\Phi^2 \geq \| X - Z_0 \|_\Phi^2.$$

Et cette inégalité montre que $\| X - Z \|_\Phi^2$ atteint son minimum lorsque $Z = Z_0$.

4.2. APPROCHE EUCLIDIENNE DE LA REGRESSION.

Considérons une variable statistique quantitative bidimensionnelle (X, Y) à valeurs dans $\mathbb{R}^2$, définie dans une population Ω de taille n.
Elle est définie par l'ensemble des couples $\{ (X(\omega), Y(\omega)) \}_{\omega \in \Omega}$.
$\mathbb{R}^2$ est l'**espace des individus**.
La variable statistique est représentée par un nuage de points dans $\mathbb{R}^2$ et chaque point du nuage statistique représente un individu de la population Ω.

4.2.1. Espace des variables.

Les n valeurs $X(\omega)$ de X pour les n individus de la population peuvent être considérées comme les

coordonnées d'un vecteur de $\mathbb{R}^n$.

Ce vecteur est noté encore $X = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix}$.

Les n valeurs $Y(\omega)$ de Y pour les n individus de la population peuvent être considérées comme les coordonnées d'un vecteur de $\mathbb{R}^n$.

Ce vecteur est noté encore $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$.

L'espace $E = \mathbb{R}^n$ apparaît alors comme l'**espace des variables**.
Chaque élément de E peut être considéré comme les valeurs d'une variable statistique quantitative réelle définie sur Ω.

4.2.2. Produit scalaire.

Dans cet espace des variables, la matrice $D_{\frac{1}{n}} = \frac{1}{n} \boldsymbol{I}_n$, où $\boldsymbol{I}_n$ est la matrice unité à n lignes et n colonnes, définit un **produit scalaire** :

$$< X \mid Y >_{D_{\frac{1}{n}}} = < X \mid D_{\frac{1}{n}} \mid Y > = \sum_i \frac{1}{n} x_i y_i = \frac{1}{n} \sum_i x_i y_i = \frac{1}{n} < X \mid Y >$$

en notant $< X \mid Y >$ le produit scalaire canonique de $\mathbb{R}^n$.

On note $\mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ le vecteur dont toutes les coordonnées sont égales à 1.

On l'appelle le **vecteur unité** de $\mathbb{R}^n$.

On remarquera que ce vecteur unité est normé, sa longueur est $\| \mathbb{1}_n \|_{D_{\frac{1}{n}}} = \frac{1}{n} \sum_i 1 \times 1 = \frac{1}{n} \times n = 1$.

4.2.3. Moyenne d'une variable statistique.

La moyenne $\overline{X}$ de la variable statistique X est donnée par :

$$\overline{X} = \frac{1}{n} \sum_\omega X(\omega) = \frac{1}{n} \sum_i x_i = \frac{1}{n} \sum_i x_i \times 1 = < X \mid D_{\frac{1}{n}} \mid \mathbb{1}_n >. = < X \mid \mathbb{1}_n >_{D_{\frac{1}{n}}}$$

La moyenne de X est le produit scalaire de X par le vecteur unité $\mathbb{1}_n$.

Notons X_0 la variable centrée correspondant à X : pour chaque individu ω de la population, sa valeur est $X(\omega) - \overline{X}$:

$$X_0 = \begin{pmatrix} x_1 - \overline{X} \\ \vdots \\ x_i - \overline{X} \\ \vdots \\ x_n - \overline{X} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} - \overline{X} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} = X - \overline{X}\, 1_n .$$

$$\boxed{X = X_0 + \overline{X}\, 1_n = X_0 + < X \mid 1_n >_{D_{\frac{1}{n}}} 1_n}$$

4.2.4. Variance d'une variable statistique.

$$s^2(X) = \overline{X_0^2} = \frac{1}{n} \sum_i (x_i - \overline{X})^2 = < X_0 \mid D_{\frac{1}{n}} \mid X_0 > = \| X_0 \|^2$$

$$\boxed{s^2(X) = \| X_0 \|^2}$$

La variance de X est le carré de la norme de la variable centrée.

4.2.5. Covariance.

La covariance de deux variables quantitatives réelles X et Y définies sur Ω est la moyenne du produit des variables centrées :

$$Cov(X, Y) = \frac{1}{n} \sum_i (x_i - \overline{X})(y_i - \overline{Y}) = < X_0 \mid D_{\frac{1}{n}} \mid Y_0 > = < X_0 \mid Y_0 >_{D_{\frac{1}{n}}}$$

$$\boxed{Cov(X, Y) = < X_0 \mid D_{\frac{1}{n}} \mid Y_0 > = < X_0 \mid Y_0 >_{D_{\frac{1}{n}}}}$$

La covariance est le produit scalaire des variables centrées.

4.2.6. Coefficient de corrélation linéaire.

$$r_{XY} = \frac{Cov[X,Y]}{s[X]\, s[Y]} = \frac{\left\langle X_0 \middle| D_{\frac{1}{n}} \middle| Y_0 \right\rangle}{\|X_0\|_\Phi \|Y_0\|_\Phi} = cos\,(X_0, Y_0)$$

$$\boxed{r_{XY} = cos\,(X_0, Y_0)}$$

Le coefficient de corrélation linéaire est le cosinus de l'angle des variables centrées.

4.2.7. Prédicteur linéaire.

Soient Y la variable à expliquer, X la variable explicative, X_0 et Y_0 les variables centrées.
Le prédicteur linéaire $\Delta_{Y \mid X}$ est $y^* = a + b\,x$ ou $y^* - \overline{Y} = b\,(x - \overline{X})$, soit $y_0^* = b\,x_0$.
Il est représenté par la **droite de régression** de Y en X dans l'espace des individus.

Le coefficient b s'obtient par $b = \frac{Cov[X,Y)}{s^2[X)} = \frac{Cov[X,Y)}{s^2[X_0)} = \frac{\langle X_0 | Y_0 \rangle_{D_{\frac{1}{n}}}}{\|X_0\|^2_{D_{\frac{1}{n}}}}$.

D'après ce qui précède (4.1.2.3.e), $b\,X_0 = \frac{\langle X_0 | Y_0 \rangle_{D_{\frac{1}{n}}}}{\|X_0\|^2_{D_{\frac{1}{n}}}} X_0$ est le projeté orthogonal de Y_0 sur X_0, $Y_0 - b\,X_0$ est orthogonal à X_0 et b est la valeur qui minimise l'expression

$$S^2 = \frac{1}{n} \sum_i (Y_{0i} - b\,X_{0i})^2 = \| Y_0 - b\,X_0 \|^2_{D_{\frac{1}{n}}} = s^2(Y - b\,X) = s^2(Y - a - b\,X) = s^2(Y - Y^*) = s^2(Y_0 - Y_0^*)$$

Le prédicteur linéaire de la variable centrée Y_0 est le projeté orthogonal de Y_0 sur X_0 dans $\mathbb{R}^n$. C'est la variable Y_0^* qui minimise la variance de $Y_0 - Y_0^*$.

Nous avons alors :

$$s^2(Y) = \| Y_0 \|^2_{D_{\frac{1}{n}}} = \| Y_0 - b\,X_0 + b\,X_0 \|^2_{D_{\frac{1}{n}}} = \| Y_0 - b\,X_0 \|^2_{D_{\frac{1}{n}}} + \| b\,X_0 \|^2_{D_{\frac{1}{n}}}$$

$$s^2(Y) = S^2_{min} + b^2 \| X_0 \|^2_{D_{\frac{1}{n}}} = S^2_{min} + \left(\frac{Cov[X,Y)}{s^2[X)} \right)^2 s^2(X) = S^2_{min} + \left(\frac{Cov[X,Y)}{s[X)\,s[Y)} \right)^2 s^2(Y)$$

$$s^2(Y) = S^2_{min} + r_{XY}^2\, s^2(Y).$$

Nous retrouvons la variance résiduelle S^2_{min} et la variance expliquée par la régression $r_{XY}^2\, s^2(Y)$.

De façon symétrique, si X est la variable explicative et Y la variable explicative, nous aurons une expression :

$$s^2(X) = S'^2_{min} + r_{XY}^2\, s^2(X).$$

avec la variance résiduelle S'^2_{min} et la variance expliquée par la régression $r_{XY}^2\, s^2(X)$.

4. 3. REGRESSION ORTHOGONALE. AXE PRINCIPAL.

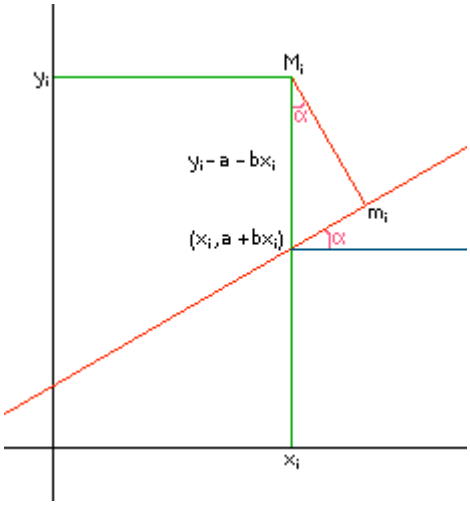
Soit $\mathbb{R}^2$ l'espace des individus, muni du produit scalaire canonique et de la base canonique $\{e_1, e_2\}$ qui, on l'a vu, est orthonormée pour ce produit scalaire.
Si aucune des variables statistiques, *X* ou *Y* ne peut s'interpréter par rapport à l'autre, il n'y a pas de raison de privilégier la régression linéaire de *Y* par rapport à *X* ou la régression linéaire de *X* par rapport à *Y*.
Nous sommes alors conduits à un autre point de vue, celui de la **réduction des données**.

4.3.1. Introduction.

Nous cherchons alors dans $\mathbb{R}^2$ une droite (*D*) qui minimise la somme $S_\perp{}^2$ des carrés des distances des points du nuage de points à la droite.
La solution est donnée par la **droite de régression orthogonale**.

a) Calcul du terme constant *a*.

L'équation de la droite de régression orthogonale est de la forme $y = a + b\,x$.



b est la tangente de l'angle de la droite avec l'axe des abscisses : $b = tan\ \alpha$.

$$\| M_i m_i \|^2 = cos^2 \alpha\, (y_i - a - b\,x_i)^2 = \frac{1}{1+b^2}\,(y_i - a - b\,x_i)^2$$

En introduisant le point moyen $(\overline{X}, \overline{Y})$, on peut écrire :

$$\frac{1}{n}\sum_{i=1}^{i=n} \| M_i m_i \|^2 = \frac{1}{1+b^2} \times \frac{1}{n}\sum_{i=1}^{i=n} (y_i - \overline{Y} - b\,(x_i - \overline{X}) + (\overline{Y} - a - b\,\overline{X}))^2$$

$$= \frac{1}{1+b^2} \times \frac{1}{n}\sum_{i=1}^{i=n} (y_i - \overline{Y} - b\,(x_i - \overline{X}))^2 + \frac{1}{1+b^2}\,(\overline{Y} - a - b\,\overline{X})^2$$

$$+ 2\,\frac{1}{1+b^2} \times (\overline{Y} - a - b\,\overline{X})\,\frac{1}{n}\sum_{i=1}^{i=n} (y_i - \overline{Y} - b\,(x_i - \overline{X})$$

Les relations $\overline{Y} = \frac{1}{n}\sum_{i=1}^{i=n} y_i$ et $\overline{X} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$ entraînent que le dernier terme de la somme est nul.

Il reste :

$$\frac{1}{n}\sum_{i=1}^{i=n} \| M_i m_i \|^2 = \frac{1}{1+b^2} \times \frac{1}{n}\sum_{i=1}^{i=n} (y_i - \overline{Y} - b\,(x_i - \overline{X}))^2 + \frac{1}{1+b^2}\,(\overline{Y} - a - b\,\overline{X})^2$$

Quel que soit la valeur de *b*, cette somme sera la plus petite possible lorsque le deuxième terme est nul : $\overline{Y} = a + b\,\overline{X}$.
Ce résultat signifie que **le point moyen est sur la droite de régression orthogonale** et que, lorsque *b* est connu, le terme constant *a* est donné par :

$$\boxed{a = \overline{Y} - b\,\overline{X}}$$

Puisque le point moyen $G = (\overline{X}, \overline{Y})$ est sur la droite de régression orthogonale, nous le prendrons comme **origine** dans $\mathbb{R}^2$.
La droite de régression orthogonale a une équation de la forme

$$y_0 = b\, x_0,$$

avec $y_0 = y - \overline{y}$ et $x_0 = x - \overline{x}$.

b) Analyse en composantes principales (ACP).

En fait, la forme de la relation précédente fait disparaître la symétrie initiale entre les rôles de X et Y : ce n'est pas sous cette forme que nous exprimerons l'équation de la droite (D) de régression orthogonale.

Etant donnée une droite (D) passant par l'origine G, on considère plutôt le vecteur unitaire de $\mathbb{R}^2$ orthogonal à la droite (D) :

$$u_1 = \begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \text{ avec } \alpha^2 + \beta^2 = 1.$$

Le vecteur unitaire u porté par la droite (D) est $u = \begin{pmatrix}\beta\\ -\alpha\end{pmatrix}$.

La droite (D) est l'ensemble des points $M = (x, y)$ vérifiant $< u_1 \mid \overrightarrow{GM} > = 0$, soit $\alpha\, x_0 + \beta\, y_0 = 0$.

Etant donné un point M_i du nuage de points et sa projection orthogonale m_i sur la droite D, le vecteur $\overrightarrow{Gm_i}$ est le projeté orthogonal de $\overrightarrow{GM_i}$ sur le vecteur u : $\overrightarrow{Gm_i} = < \overrightarrow{GM_i} \mid u > u = (\beta\, x_{i0} - \alpha\, y_{i0}) \begin{pmatrix}\beta\\ -\alpha\end{pmatrix}$

$$\overrightarrow{m_i M_i} = \overrightarrow{GM_i} - \overrightarrow{Gm_i} = \begin{pmatrix}x_{i0}\\ y_{i0}\end{pmatrix} - (\beta\, x_{i0} - \alpha\, y_{i0}) \begin{pmatrix}\beta\\ -\alpha\end{pmatrix} = \begin{pmatrix}(1-\beta^2)\, x_{i0} + \alpha\beta\, y_{i0}\\ \alpha\beta\, x_{i0} + (1-\alpha^2)\, y_{i0}\end{pmatrix} = \begin{pmatrix}\alpha^2\, x_{i0} + \alpha\beta\, y_{i0}\\ \alpha\beta\, x_{i0} + \beta^2\, y_{i0}\end{pmatrix} = (\alpha\, x_{i0} + \beta\, y_{i0}) \begin{pmatrix}\alpha\\ \beta\end{pmatrix}$$

$$\| \overrightarrow{m_i M_i} \|^2 = (\alpha\, x_{i0} + \beta\, y_{i0})^2\, (\alpha\ \beta) \begin{pmatrix}\alpha\\ \beta\end{pmatrix} = (\alpha\, x_{i0} + \beta\, y_{i0})^2\, (\alpha^2 + \beta^2) = (\alpha\, x_{i0} + \beta\, y_{i0})^2$$

$$\frac{1}{n} \sum_{i=1}^{i=n} \| M_i\, m_i \|^2 = \frac{1}{n} \sum_{i=1}^{i=n} (\alpha\, x_{i0} + \beta\, y_{i0})^2 = < \alpha\, X_0 + \beta\, Y_0 \mid D_{\frac{1}{n}} \mid \alpha\, X_0 + \beta\, Y_0 > = \| \alpha\, X_0 + \beta\, Y_0 \|^2_{D_{\frac{1}{n}}}.$$

La recherche de la droite de régression orthogonale se ramène donc à une question que l'on peut envisager d'un double point de vue :

— soit rechercher, dans l'espace des individus $\mathbb{R}^2$, un vecteur unitaire $u_1 = \begin{pmatrix}\alpha\\ \beta\end{pmatrix}$, avec $\alpha^2 + \beta^2 = 1$, qui minimise la somme

$$S_\perp^2 = \frac{1}{n} \sum_{i=1}^{i=n} \| M_i\, m_i \|^2 = \frac{1}{n} \sum_{i=1}^{i=n} (\alpha\, x_{i0} + \beta\, y_{i0})^2,$$

— soit rechercher, dans l'espace des variables $\mathbb{R}^n$, un vecteur $\alpha\, X_0 + \beta\, Y_0$, combinaison linéaire fictive des deux variables centrées X_0 et Y_0, avec $\alpha^2 + \beta^2 = 1$, qui minimise $\| \alpha\, X_0 + \beta\, Y_0 \|^2_{D_{\frac{1}{n}}}$, c'est-à-dire un vecteur de l'hyperplan défini par X_0 et Y_0, de norme minimum pour le produit scalaire défini par la matrice diagonale $D_{\frac{1}{n}}$, sous la contrainte $\alpha^2 + \beta^2 = 1$.

Sous la deuxième forme, la résolution du problème est appelée l'**analyse en composantes principales**.

4.3.2. Définitions.

Appellons Z la matrice $\begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix}$ des variables centrées,

a) Inertie totale.

On appelle inertie totale du nuage de points de $\mathbb{R}^2$ par rapport à l'origine G des axes, la quantité :

$$I_T = \frac{1}{n} \sum_{i=1}^{i=n} \| \overrightarrow{GM_i} \|^2 = \frac{1}{n} \sum_{i=1}^{i=n} (x_{i0}^2 + y_{i0}^2) = s^2(X) + s^2(Y).$$

b) Inertie statistique.

On appelle inertie statistique du nuage de points de $\mathbb{R}^2$ par rapport à une direction Δ de $\mathbb{R}^2$ définie par un vecteur unitaire u, la quantité :

$$I_S(u) = \frac{1}{n} \sum_{i=1}^{i=n} \| \overrightarrow{Gm_i} \|^2$$

où $\overrightarrow{Gm_i}$ est le projeté orthogonal de $\overrightarrow{GM_i}$ sur u.

Le rapport $\frac{I_S(u)}{I_T}$ est le **taux d'inertie totale expliquée par la direction u**.

Par exemple, l'inertie statistique du nuage de points par rapport à l'axe des x est la variance de X et l'inertie statistique du nuage de points par rapport à l'axe des y est la variance de Y.

c) Inertie mécanique.

On appelle inertie mécanique du nuage de points de $\mathbb{R}^2$ par rapport à une direction Δ définie par un vecteur unitaire u, la quantité :

$$I_M(u) = \frac{1}{n} \sum_{i=1}^{i=n} \| \overrightarrow{m_i M_i} \|^2$$

où $\overrightarrow{Gm_i}$ est le projeté orthogonal de $\overrightarrow{GM_i}$ sur u.

Par exemple, l'inertie mécanique du nuage de points par rapport à l'axe des x est la variance de Y et l'inertie mécanique du nuage de points par rapport à l'axe des y est la variance de X.

Le théorème de Pythagore $\| \overrightarrow{GM_i} \|^2 = \| \overrightarrow{Gm_i} \|^2 + \| \overrightarrow{m_i M_i} \|^2$ entraîne :

$$I_M(u) = I_T - I_S(u).$$

d) Axes principaux, ou factoriels.

On appelle **premier axe factoriel** du nuage de points de $\mathbb{R}^2$, l'axe dont la direction définie par un vecteur unitaire u maximise l'inertie statistique $I_S(u)$.

La direction définie par le vecteur *u* est appelée la **direction principale**, ou **direction factorielle**.

On remarquera que, comme le premier axe factoriel maximise $I_S(u)$, il minimise $I_M(u)$: il donne donc la solution de notre problème, c'est-à-dire la droite de régression orthogonale.

e) Matrice des variances-covariances.

Pour $u = \begin{pmatrix} \beta \\ -\alpha \end{pmatrix}$, l'inertie statistique $I_S(u) = \frac{1}{n} \sum_{i=1}^{i=n} \| \overrightarrow{Gm_i} \|^2$ s'écrit, avec $\overrightarrow{Gm_i} = < \overrightarrow{GM_i} \mid u > u = (\beta\, x_{i0} - \alpha\, y_{i0}) \begin{pmatrix} \beta \\ -\alpha \end{pmatrix}$, sous la forme :

$$I_S(u) = \frac{1}{n} \sum_{i=1}^{i=n} (\beta\, x_{i0} - \alpha\, y_{i0})^2 = \beta^2 \times \frac{1}{n} \sum_{i=1}^{i=n} x_{i0}^2 + \alpha^2 \times \frac{1}{n} \sum_{i=1}^{i=n} y_{i0}^2 - 2\, \alpha\, \beta \times \frac{1}{n} \sum_{i=1}^{i=n} x_{i0}\, y_{i0}$$

Et comme on sait que :

$$\frac{1}{n} \sum_{i=1}^{i=n} x_{i0}^2 = s^2(X),\ \frac{1}{n} \sum_{i=1}^{i=n} y_{i0}^2 = s^2(Y),\ \frac{1}{n} \sum_{i=1}^{i=n} x_{i0}\, y_{i0} = Cov(X, Y),$$

l'inertie statistique devient :

$$I_S(u) = \beta^2 s^2(X) + \alpha^2 s^2(Y) - 2\, \alpha\, \beta\, Cov(X, Y) = (\beta\ -\alpha) \begin{pmatrix} s^2[X] & Cov[X,Y] \\ Cov[X,Y] & s^2[Y] \end{pmatrix} \begin{pmatrix} \beta \\ -\alpha \end{pmatrix} = {}^t u\, A\, u$$

La matrice

$$A = \begin{pmatrix} s^2[X] & Cov[X,Y] \\ Cov[X,Y] & s^2[Y] \end{pmatrix} = \frac{1}{n} \begin{pmatrix} x_{10} & \cdots & x_{n0} \\ y_{10} & \cdots & y_{n0} \end{pmatrix} \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix}$$

s'appelle la **matrice des variances-covariances**.

En introduisant la matrice $Z = \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix}$ des variables centrées, la matrice des variances-covariances s'écrit sous les formes :

$$\boxed{A = \begin{pmatrix} s^2[X] & Cov[X,Y] \\ Cov[X,Y] & s^2[Y] \end{pmatrix} = \frac{1}{n} \begin{pmatrix} x_{10} & \cdots & x_{n0} \\ y_{10} & \cdots & y_{n0} \end{pmatrix} \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix} = \frac{1}{n} {}^t Z\, Z = {}^t Z\, D_{\frac{1}{n}}\, Z}$$

et l'inertie totale est la trace de cette matrice, somme des éléments diagonaux $s^2(X)$ et $s^2(Y)$:

$$I_T = Tr(A)$$

1 e remarque : valeurs propres.

La matrice des variances-covariances *A* est, comme on le voit, symétrique réelle.
Une valeur propre de *A* est un nombre réel λ tel qu'il existe un vecteur *v* non nul vérifiant $A\, v = \lambda\, v$.
Les valeurs propres de *A* sont donc les nombres réels λ tels que le noyau de l'endomorphisme

(application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^2$) défini par la matrice $A - \lambda \boldsymbol{I}_2$ ne soit pas réduit à 0.
Dire que le noyau n'est pas réduit à 0, c'est dire que l'application linéaire n'est pas injective, donc qu'elle n'est pas bijective (puisque, dans $\mathbb{R}^2$, injective = bijective) : pour cela, il faut et il suffit que son déterminant soit nul.
Les valeurs propres sont donc les solutions de l'équation :

$$Dét\,(A - \lambda \boldsymbol{I}_2) = 0$$
$$\lambda^2 - (s^2(X) + s^2(Y))\,\lambda + s^2(X)\,s^2(Y) - (Cov(X, Y))^2 = 0$$

Le discriminant de cette équation du deuxième degré est :

$$(s^2(X) + s^2(Y))^2 - 4\,(s^2(X)\,s^2(Y) - (Cov(X, Y))^2) = (s^2(X) - s^2(Y))^2 + 4\,(Cov(X, Y))^2 \geq 0$$

La matrice A possède donc, ainsi qu'on l'avait déjà dit pour toute matrice symétrique réelle, deux valeurs propres réelles λ_1 et λ_2 :
— la somme de ces valeurs propres est la **trace** de la matrice, somme des éléments de la première diagonale :
$\lambda_1 + \lambda_2 = s^2(X) + s^2(Y) \geq 0$.
— le produit de ces valeurs propres est le **déterminant** de la matrice :
$\lambda_1\,\lambda_2 = s^2(X)\,s^2(Y) - (Cov(X, Y))^2 \geq 0$ (d'après l'inégalité de Schwarz).

Les deux valeurs propres de la matrice des variances-covariances sont donc des nombres réels positifs : il est très improbable que l'une soit nulle (il faudrait, pour cela, que le coefficient de corrélation linéaire soit rigoureusement égal à 1, en valeur absolue, ce qui ne saurait se produire que si X et Y sont déduits l'un de l'autre par une relation linéaire, ou si X et Y sont constantes. Il est très improbable aussi que les deux valeurs propres soient égales : il faudrait pour cela que la covariance de X et Y soit strictement égale à 0 et que les variances de X et Y soient strictement égales, ce qui ne se produit jamais en pratique.

Dans le cas général, on peut donc appeler λ_1 et λ_2 les **valeurs propres de la matrice des variances-covariances**, rangées par ordre décroissant :

$$\lambda_1 > \lambda_2 > 0.$$

$$\lambda_1 = \frac{1}{2}\left(s^2(X) + s^2(Y) + \sqrt{\left[s^2(X) - s^2(Y)\right]^2 + 4\left[Cov(X,Y)\right]^2} \right)$$
$$\lambda_2 = \frac{1}{2}\left(s^2(X) + s^2(Y) - \sqrt{\left[s^2(X) - s^2(Y)\right]^2 + 4\left[Cov(X,Y)\right]^2} \right)$$

2e remarque : vecteurs propres.

On démontre aussi, en algèbre, que $\mathbb{R}^2$ possède une **base propre orthonormée**, c'est-à-dire une base $\{ u_1, u_2 \}$, orthonormée pour le produit scalaire canonique, formée de vecteurs propres de la matrice A :

$$A\,u_1 = \lambda_1\,u_1 \text{ et } A\,u_2 = \lambda_2\,u_2,$$

avec

$$\| u_1 \|^2 = 1, \| u_2 \|^2 = 1, < u_1 \mid u_2 > = 0.$$

Ces vecteurs propres peuvent se calculer.
Soit λ une valeur propre. On a :

$$\begin{pmatrix} s^2[X]-\lambda & Cov[X,Y] \\ Cov[X,Y] & s^2[Y]-\lambda \end{pmatrix}\begin{pmatrix} s^2[Y]-\lambda \\ -Cov[X,Y] \end{pmatrix} = \begin{pmatrix} (s^2[X]-\lambda)(s^2[Y]-\lambda)-(Cov[X,Y])^2 \\ (s^2[Y]-\lambda)Cov[X,Y]-(s^2[Y]-\lambda)Cov[X,Y] \end{pmatrix} = \begin{pmatrix} Det(A-\lambda I_2) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0$$

donc le vecteur $\begin{pmatrix} s^2[Y]-\lambda \\ -Cov[X,Y] \end{pmatrix}$ est un **vecteur propre pour la valeur propre λ**.
Le carré de la norme de ce vecteur pour le produit scalaire canonique est donné par :

$$(s^2(Y)-\lambda \quad -Cov(X,Y))\begin{pmatrix} s^2[Y]-\lambda \\ -Cov[X,Y] \end{pmatrix} = (s^2(Y)-\lambda)^2 + (Cov(X,Y))^2$$

On peut donc prendre pour vecteur normé relatif à la valeur propre λ, le vecteur

$$u = \frac{1}{\sqrt{(s^2[Y]-\lambda)^2+(Cov[X,Y])^2}}\begin{pmatrix} s^2[Y]-\lambda \\ -Cov[X,Y] \end{pmatrix}$$

Le produit scalaire des deux vecteurs propres ainsi obtenu est nul, parce que la relation $\lambda_1 + \lambda_2 = s^2(X) + s^2(Y)$ entraîne :

$$(s^2(Y)-\lambda_1 \quad -Cov(X,Y))\begin{pmatrix} s^2[Y]-\lambda_2 \\ -Cov[X,Y] \end{pmatrix} = (\lambda_2 - s^2(X) \quad -Cov(X,Y))\begin{pmatrix} s^2[Y]-\lambda_2 \\ -Cov[X,Y] \end{pmatrix} = -Dét(A-\lambda_2 \boldsymbol{I}_2) = 0$$

Les deux vecteurs $\begin{pmatrix} s^2[Y]-\lambda_1 \\ -Cov[X,Y] \end{pmatrix}$ et $\begin{pmatrix} s^2[Y]-\lambda_2 \\ -Cov[X,Y] \end{pmatrix}$ forment une base de $\mathbb{R}^2$ parce que le déterminant de leurs coordonnées n'est pas nul :

$$-Cov(X,Y)\times(s^2(Y)-\lambda_1) + Cov(X,Y)\times(s^2(Y)-\lambda_2) = Cov(X,Y)\times(\lambda_1-\lambda_2) \neq 0$$

de sorte que les deux vecteurs ne sont pas proportionnels.

Les deux vecteurs :

$$u_1 = \frac{1}{\sqrt{(s^2[Y]-\lambda_1)^2+(Cov[X,Y])^2}}\begin{pmatrix} s^2[Y]-\lambda_1 \\ -Cov[X,Y] \end{pmatrix}$$
$$u_2 = \frac{1}{\sqrt{(s^2[Y]-\lambda_2)^2+(Cov[X,Y])^2}}\begin{pmatrix} s^2[Y]-\lambda_2 \\ -Cov[X,Y] \end{pmatrix}$$

forment donc une **base propre orthonormée** de $\mathbb{R}^2$.

Remarquons que, au lieu de prendre pour vecteur propre pour la valeur propre λ, le vecteur $\begin{pmatrix} s^2[Y]-\lambda \\ -Cov[X,Y] \end{pmatrix}$, on aurait pu prendre aussi le vecteur $\begin{pmatrix} -Cov[X,Y] \\ s^2[X]-\lambda \end{pmatrix}$ qui lui est proportionnel (le déterminant de la matrice de ces vecteurs est le déterminant de la matrice $A - \lambda \boldsymbol{I}_2$).

4.3.3. Diagonalisation de la matrice des variances-covariances.

Soit $V = \begin{pmatrix} \frac{s^2[Y]-\lambda_1}{\sqrt{(s^2[Y]-\lambda_1)^2+(Cov[X,Y])^2}} & \frac{s^2[Y]-\lambda_2}{\sqrt{(s^2[Y]-\lambda_2)^2+(Cov[X,Y])^2}} \\ \frac{-Cov[X,Y]}{\sqrt{(s^2[Y]-\lambda_1)^2+(Cov[X,Y])^2}} & \frac{-Cov[X,Y]}{\sqrt{(s^2[Y]-\lambda_2)^2+(Cov[X,Y])^2}} \end{pmatrix}$ la matrice des coordonnées des vecteurs propres u_1 et u_2.

$$V\,e_1 = u_1,\ V\,e_2 = u_2.$$

V donne, par produits, pour image d'une base orthonormée, une base orthonormée : c'est ce qu'on appelle une matrice "orthogonale", ce qui veut dire que son inverse est égale à sa transposée :

$$\boxed{V^{-1} = {}^tV}$$

Pour le vérifier, remarquons que, puisque les bases $\{ e_1, e_2 \}$ et $\{ u_1, u_2 \}$ sont orthonormées, les coordonnées des vecteurs s'obtiennent par produits scalaires :

$$u_1 = < u_1 \mid e_1 > e_1 + < u_1 \mid e_2 > e_2$$
$$u_2 = < u_2 \mid e_1 > e_1 + < u_2 \mid e_2 > e_2$$

de sorte que la matrice V, qui a, pour colonnes, les vecteurs u_1 et u_2 dans la base $\{ e_1, e_2 \}$, est :

$$V = \begin{pmatrix} \langle u_1 | e_1 \rangle & \langle u_2 | e_1 \rangle \\ \langle u_1 | e_2 \rangle & \langle u_2 | e_2 \rangle \end{pmatrix}$$

et les relations inverses :

$$e_1 = < e_1 \mid u_1 > u_1 + < e_1 \mid u_2 > u_2$$
$$e_2 = < e_2 \mid u_1 > u_1 + < e_2 \mid u_2 > u_2$$

montrent que la matrice inverse de V est la matrice :

$$V^{-1} = \begin{pmatrix} \langle e_1 | u_1 \rangle & \langle e_2 | u_1 \rangle \\ \langle e_1 | u_2 \rangle & \langle e_2 | u_2 \rangle \end{pmatrix}$$

qui, compte tenu de la symétrie du produit scalaire, est la transposée de V.

$$V^{-1} = \begin{pmatrix} \langle u_1 | e_1 \rangle & \langle u_1 | e_2 \rangle \\ \langle u_2 | e_1 \rangle & \langle u_2 | e_2 \rangle \end{pmatrix} = {}^t V$$

Il résulte alors des relations $V\,e_1 = u_1$ et $V\,e_2 = u_2$, que l'on a :

$${}^tV\,u_1 = V^{-1}\,u_1 = e_1\ ;\ {}^t V\,u_2 = V^{-1}\,u_2 = e_2$$

Considérons maintenant la matrice $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, matrice diagonale des valeurs propres de A.

A est la matrice, dans la base canonique $\{ e_1, e_2 \}$, d'un endomorphisme f.
Cet endomorphisme f se réduit à deux homothéties, de rapport λ_1 selon le vecteur u_1, et de rapport λ_2 selon le vecteur u_2.
Λ est donc la matrice, dans la base propre $\{ u_1, u_2 \}$, de l'endomorphisme f.
La matrice de l'application identique de $\mathbb{R}^2$ muni de la base $\{ u_1, u_2 \}$ dans $\mathbb{R}^2$ muni de la base $\{ e_1, e_2 \}$ donne, par produits, pour image du vecteur $u_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ le vecteur $u_1 = \frac{1}{\sqrt{(s^2[Y]-\lambda_1)^2+(Cov[X,Y])^2}} \begin{pmatrix} s^2[Y]-\lambda_1 \\ -Cov[X,Y] \end{pmatrix}$ et, pour image du vecteur $u_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ le vecteur $u_2 = \frac{1}{\sqrt{(s^2[Y]-\lambda_2)^2+(Cov[X,Y])^2}} \begin{pmatrix} s^2[Y]-\lambda_2 \\ -Cov[X,Y] \end{pmatrix}$. C'est donc la matrice V des vecteurs propres.

$$V = [\, Id_{\mathbb{R}^2}, \{ u_1, u_2 \}, \{ e_1, e_2 \}].$$

Réciproquement, la matrice de l'application identique de $\mathbb{R}^2$ muni de la base $\{ e_1, e_2 \}$ dans $\mathbb{R}^2$ muni de la base $\{ u_1, u_2 \}$ donne, par produits, pour image du vecteur $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ le vecteur $e_1 = \begin{pmatrix} \frac{s^2[Y]-\lambda_1}{\sqrt{(s^2[Y]-\lambda_1)^2+(Cov[X,Y])^2}} \\ \frac{s^2[Y]-\lambda_2}{\sqrt{(s^2[Y]-\lambda_2)^2+(Cov[X,Y])^2}} \end{pmatrix}$ et, pour image du vecteur $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ le vecteur $e_2 = \begin{pmatrix} \frac{-Cov[X,Y]}{\sqrt{(s^2[Y]-\lambda_1)^2+(Cov[X,Y])^2}} \\ \frac{-Cov[X,Y]}{\sqrt{(s^2[Y]-\lambda_2)^2+(Cov[X,Y])^2}} \end{pmatrix}$. C'est donc la matrice ${}^tV$ transposée et inverse de la matrice V des vecteurs propres.

$$ {}^tV = [\, Id_{\mathbb{R}^2}, \{ e_1, e_2 \}, \{ u_1, u_2 \}].$$

Le diagramme commutatif suivant :

$$\begin{array}{ccc} \mathbb{R}^2, \{ e_1, e_2 \} & \xrightarrow[A]{f} & \mathbb{R}^2, \{ e_1, e_2 \} \\ Id \uparrow {}^tV & & Id \downarrow V \\ \mathbb{R}^2, \{ u_1, u_2 \} & \xrightarrow[\Lambda]{f} & \mathbb{R}^2, \{ u_1, u_2 \} \end{array}$$

met en évidence la relation $f = Id \circ f \circ Id$.
En termes de produit de matrices, cette relation s'écrit :

$$\Lambda = V A {}^tV,$$

d'où l'on déduit aussitôt

$$A = {}^tV \Lambda V.$$

On dit qu'on a **diagonalisé la matrice** *A*.

4.3.4. Recherche des axes principaux.

Pour un vecteur normé *u*, posons $v = V u$.
On a ${}^tv = {}^tu\ {}^tV$.

$$\| v \|^2 = {}^tv\, v = {}^tu\ {}^tV\, V u = {}^tu\, u = \| u \|^2 = 1.$$

Le vecteur *v* est normé lui aussi.

L'inertie statistique par rapport à *u* s'écrit :

$$I_S(u) = {}^tu\, A\, u = {}^tu\ {}^tV \Lambda\, V u = {}^tv \Lambda\, v.$$

Dans $\mathbb{R}^2$ rapporté à la base $\{ u_1, u_2 \}$, notons $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$.

$$I_S(u) = {}^tv \Lambda\, v = (v_1 \quad v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda_1 v_1^2 + \lambda_2 v_2^2,$$
$$\text{avec } v_1^2 + v_2^2 = 1$$

Le problème de la recherche de la droite de régression orthogonale se ramène maintenant à la résolution du problème suivant :

Maximiser $\lambda_1 v_1^2 + \lambda_2 v_2^2$, sous la contrainte $v_1^2 + v_2^2 = 1$, avec $\lambda_1 > \lambda_2 > 0$.

C'est maintenant un problème facile à résoudre :

$$I_S(u) = \lambda_1 v_1^2 + \lambda_2 v_2^2 = \lambda_1 (1 - v_2^2) + \lambda_2 v_2^2 = \lambda_1 - (\lambda_1 - \lambda_2) v_2^2$$

La quantité $\lambda_1 - (\lambda_1 - \lambda_2) v_2^2$ avec $\lambda_1 > \lambda_2$ atteint sa valeur maximum λ_1 lorsqu'on prend $v_2 = 0$, donc $| v_1 | = 1$.
La direction du premier axe factoriel est donc définie par le vecteur *v* de coordonnées $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ dans la base $\{ u_1, u_2 \}$: $v = u_1$.

$$I_S(u_1) = \lambda_1$$

D'où le résultat, qu'on peut énoncer sous forme de **théorème** :

La direction du premier axe factoriel est définie par le vecteur propre associé à la plus grande valeur propre de la matrice des variances-covariances.

Le premier axe factoriel est la **droite de régression orthogonale**.

Comme **corollaire**, la direction perpendiculaire au premier axe factoriel définit le **deuxième axe**

factoriel : elle est définie par le vecteur propre associé à la plus petite valeur propre de la matrice des variances-covariances.
Le deuxième axe factoriel minimise l'inertie statistique $I_S(u)$: $I_S(u) = \lambda_2$ lorsque $|v_2| = 1$, donc $v_1 = 0$ et $v = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = u_2$ par exemple (on pourrait prendre aussi, bien sûr, $v = -u_2$, la direction définie serait la même).

$$I_S(u_2) = \lambda_2$$

Le taux d'inertie totale expliquée par le premier axe factoriel est le rapport $\frac{I_S(u_1)}{I_T} = \frac{\lambda_1}{s^2(X)+s^2(Y)} = \frac{\lambda_1}{\lambda_1+\lambda_2}$.
Le taux d'inertie totale expliquée par le deuxième axe factoriel est le rapport $\frac{I_S(u_2)}{I_T} = \frac{\lambda_2}{s^2(X)+s^2(Y)} = \frac{\lambda_2}{\lambda_1+\lambda_2}$.

La relation $\lambda_1 + \lambda_2 = s^2(X) + s^2(Y)$ (la somme des valeurs propres est la trace de la matrice des variances-covariances) s'écrit :

$$I_S(u_1) + I_S(u_2) = I_T.$$

La somme des inerties statistiques par rapport aux deux axes factoriels est l'inertie totale du nuage de points.
Chaque valeur propre de la matrice des variances-covariances correspond à l'inertie expliquée par l'axe factoriel correspondant.

4.3.5. Coordonnées factorielles et composantes principales.

Dans $\mathbb{R}^2$ rapporté à la base propre orthonormée $\{ u_1, u_2 \}$, les coordonnées des vecteurs $\overrightarrow{GM_i}$ s'appellent les **coordonnées factorielles**.
Comme la base $\{ u_1, u_2 \}$ est orthonormée, les coordonnées factorielles s'obtiennent par produit scalaire :

$$\overrightarrow{GM_i} = \langle \overrightarrow{GM_i} \mid u_1 \rangle u_1 + \langle \overrightarrow{GM_i} \mid u_2 \rangle u_2$$

Or la base canonique $\{ e_1, e_2 \}$ est, elle-même, orthonormée et l'on a, par conséquent :

$$\overrightarrow{GM_i} = \langle \overrightarrow{GM_i} \mid e_1 \rangle e_1 + \langle \overrightarrow{GM_i} \mid e_2 \rangle e_2 = x_{i0}\, e_1 + y_{i0}\, e_2$$

d'où :

$$\langle \overrightarrow{GM_i} \mid u_1 \rangle = x_{i0} \langle e_1 \mid u_1 \rangle + y_{i0} \langle e_2 \mid u_1 \rangle$$
$$\langle \overrightarrow{GM_i} \mid u_2 \rangle = x_{i0} \langle e_1 \mid u_2 \rangle + y_{i0} \langle e_2 \mid u_2 \rangle$$

Les coordonnées factorielles s'obtiennent donc par la formule matricielle :

$$\begin{pmatrix} \langle \overrightarrow{GM_i} | u_1 \rangle \\ \langle \overrightarrow{GM_i} | u_2 \rangle \end{pmatrix} = \begin{pmatrix} \langle e_1 | u_1 \rangle & \langle e_2 | u_1 \rangle \\ \langle e_1 | u_2 \rangle & \langle e_2 | u_2 \rangle \end{pmatrix} \begin{pmatrix} x_{i0} \\ y_{i0} \end{pmatrix} = {}^tV \begin{pmatrix} \langle \overrightarrow{GM_i} | e_1 \rangle \\ \langle \overrightarrow{GM_i} | e_2 \rangle \end{pmatrix}$$

$$\boxed{\begin{pmatrix} \langle \overrightarrow{GM_i} | u_1 \rangle \\ \langle \overrightarrow{GM_i} | u_2 \rangle \end{pmatrix} = {}^tV \begin{pmatrix} \langle \overrightarrow{GM_i} | e_1 \rangle \\ \langle \overrightarrow{GM_i} | e_2 \rangle \end{pmatrix} = {}^tV \begin{pmatrix} x_{i0} \\ y_{i0} \end{pmatrix}}$$

La matrice ${}^tV$ est ce qu'on appelle la **matrice du changement de base**.
Elle donne les nouvelles coordonnées (sur la base $\{ u_1, u_2 \}$) en fonction des anciennes (sur la base $\{ e_1, e_2 \}$).

Nous avons vu plus haut que cette matrice est la matrice de l'application identique, de $\mathbb{R}^2$ muni de la base $\{ u_1, u_2 \}$ dans $\mathbb{R}^2$ muni de la base $\{ e_1, e_2 \}$.

Les relations :

$$(< \overrightarrow{GM_i} \mid u_1 > \quad < \overrightarrow{GM_i} \mid u_2 >) = {}^t\begin{pmatrix} \langle \overrightarrow{GM_i} | u_1 \rangle \\ \langle \overrightarrow{GM_i} | u_2 \rangle \end{pmatrix} = {}^t\left({}^tV \begin{pmatrix} x_{i0} \\ y_{i0} \end{pmatrix} \right) = (x_{i0} \quad y_{i0})\, V, \text{ pour } i \in \{ 1, \dots, n \},$$

peuvent se condenser en une seule formule matricielle :

$$\boxed{L = Z\, V}$$

formule dans laquelle :

$$L = \begin{pmatrix} \langle \overrightarrow{GM_1} | u_1 \rangle & \langle \overrightarrow{GM_1} | u_2 \rangle \\ \vdots & \vdots \\ \langle \overrightarrow{GM_n} | u_1 \rangle & \langle \overrightarrow{GM_n} | u_2 \rangle \end{pmatrix}$$

est la matrice, à *n* lignes et 2 colonnes, dont les lignes sont les coordonnées factorielles du nuage de points dans $\mathbb{R}^2$ muni de la base $\{ u_1, u_2 \}$,

$$Z = \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix}$$

est la matrice, à *n* lignes et 2 colonnes, dont les colonnes sont les variables centrées $X - \overline{X}$ et $Y - \overline{Y}$,

$$V = \begin{pmatrix} \dfrac{s^2[Y] - \lambda_1}{\sqrt{(s^2[Y] - \lambda_1)^2 + (Cov[X,Y])^2}} & \dfrac{s^2[Y] - \lambda_2}{\sqrt{(s^2[Y] - \lambda_2)^2 + (Cov[X,Y])^2}} \\ \dfrac{-Cov[X,Y]}{\sqrt{(s^2[Y] - \lambda_1)^2 + (Cov[X,Y])^2}} & \dfrac{-Cov[X,Y]}{\sqrt{(s^2[Y] - \lambda_2)^2 + (Cov[X,Y])^2}} \end{pmatrix}$$

est la matrice des coordonnées des vecteurs propres orthonormés $\{ u_1, u_2 \}$ de la matrice des

variances-covariances, dans la base canonique $\{ e_1, e_2 \}$.

Les deux colonnes de la matrice L sont des éléments de l'espace des variables $\mathbb{R}^n$: on les appelle les **composantes principales** de la variable statistique (X, Y).

La première colonne de la matrice V est le vecteur propre u_1.
La première colonne de la matrice $L = Z\,V$ est donc le vecteur $L_1 = Z\,u_1$.
De même, la deuxième colonne de la matrice L est le vecteur $L_2 = Z\,u_2$.
Les deux composantes principales L_1 et L_2 de la variable statistique (X, Y) s'obtiennent ainsi par les formules :

$$L_1 = \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix} u_1 = \frac{1}{\sqrt{\left(s^2[Y]-\lambda_1\right)^2 + \left(Cov[X,Y]\right)^2}} \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix} \begin{pmatrix} s^2[Y]-\lambda_1 \\ -Cov[X,Y] \end{pmatrix}$$

$$L_2 = \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix} u_2 = \frac{1}{\sqrt{\left(s^2[Y]-\lambda_2\right)^2 + \left(Cov[X,Y]\right)^2}} \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix} \begin{pmatrix} s^2[Y]-\lambda_2 \\ -Cov[X,Y] \end{pmatrix}$$

avec les valeurs propres λ_1 et λ_2 de la matrice

$$A = \frac{1}{n} \begin{pmatrix} x_{10} & \cdots & x_{n0} \\ y_{10} & \cdots & y_{n0} \end{pmatrix} \begin{pmatrix} x_{10} & y_{10} \\ \vdots & \vdots \\ x_{n0} & y_{n0} \end{pmatrix} = \frac{1}{n}\,{}^t Z\,Z = {}^t Z\,D_{\frac{1}{n}}\,Z = \begin{pmatrix} s^2[X] & Cov[X,Y] \\ Cov[X,Y] & s^2[Y] \end{pmatrix}$$

des variances-covariances :

$$\lambda_1 = \frac{1}{2}\left(s^2(X) + s^2(Y) + \sqrt{\left(s^2[X]-s^2[Y]\right)^2 + 4\left(Cov[X,Y]\right)^2} \right)$$

$$\lambda_2 = \frac{1}{2}\left(s^2(X) + s^2(Y) - \sqrt{\left(s^2[X]-s^2[Y]\right)^2 + 4\left(Cov[X,Y]\right)^2} \right)$$

4.3.6. Propriétés des composantes principales.

a) Les composantes principales sont centrées.

$$\overline{L_1} = < L_1 \mid D_{\frac{1}{n}} \mid 1_n > = \frac{1}{n} < Z\,u_1 \mid 1_n > = \frac{1}{n}\,{}^t(Z\,u_1)\,1_n = \frac{1}{n}\,{}^t u_1\,{}^t Z\,1_n$$

$${}^t Z\,1_n = \begin{pmatrix} x_{10} & \cdots & x_{n0} \\ y_{10} & \cdots & y_{n0} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{i=n} x_{i0} \\ \sum_{i=1}^{i=n} y_{i0} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

puisque les variable X_0 et Y_0 sont centrées.
Il reste donc :

$$\overline{L_1} = \frac{1}{n}\,{}^t u_1 \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0$$

De même :

$$\overline{L_2} = < L_2 \mid D_{\frac{1}{n}} \mid 1_n > = \frac{1}{n} < Z\, u_2 \mid 1_n > = \frac{1}{n}\, {}^t(Z\, u_2)\, 1_n = \frac{1}{n}\, {}^t u_2\, {}^t Z\, 1_n = \frac{1}{n}\, {}^t u_2 \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0.$$

b) La variance d'une composante principale est la valeur propre correspondante.

Comme les composantes principales sont centrées, leur variance est le carré de leur norme pour le produit scalaire défini par $D_{\frac{1}{n}}$:

$$s^2(L_1) = \| L_1 \|^2_{D_{\frac{1}{n}}} = < L_1 \mid D_{\frac{1}{n}} \mid L_1 > = \frac{1}{n}\, {}^t L_1\, L_1 = \frac{1}{n}\, {}^t u_1\, {}^t Z\, Z\, u_1$$

$$\frac{1}{n}\, {}^t Z\, Z = \mathrm{A}$$

$$s^2(L_1) = {}^t u_1\, A\, u_1 = {}^t u_1\, \lambda_1\, u_1 = \lambda_1 \| u_1 \|^2 = \lambda_1$$

De même :

$$s^2(L_2) = < L_2 \mid D_{\frac{1}{n}} \mid L_2 > = \frac{1}{n}\, {}^t L_2\, L_2 = \frac{1}{n}\, {}^t u_2\, {}^t Z\, Z\, u_2$$
$$= {}^t u_2\, A\, u_2 = {}^t u_2\, \lambda_2\, u_2 = \lambda_2 \| u_2 \|^2 = \lambda_2$$

c) Les composantes principales sont non corrélées.

$$Cov(L_1, L_2) = < L_1 \mid D_{\frac{1}{n}} \mid L_2 > = \frac{1}{n}\, {}^t L_1\, L_2 = \frac{1}{n}\, {}^t u_1\, {}^t Z\, Z\, u_2$$
$$= \frac{1}{n}\, {}^t u_1\, A\, u_2 = \frac{\lambda_2}{n} < u_1 \mid u_2 > = 0$$

puisque les vecteurs u_1 et u_2 sont orthogonaux pour le produit scalaire canonique.

d) Reconstruction des données.

Les points du nuage centré sont définis par les vecteurs

$$\overrightarrow{GM_i} = x_{i0}\, e_1 + y_{i0}\, e_2 = < \overrightarrow{GM_i} \mid u_1 > u_1 + < \overrightarrow{GM_i} \mid u_2 > u_2.$$

Les projetés orthogonaux de ces vecteurs sur l'axe principal défini par u_1 sont les vecteurs :

$$\overrightarrow{Gm_i} = < \overrightarrow{GM_i} \mid u_1 > u_1 = < \overrightarrow{GM_i} \mid u_1 > (< u_1 \mid e_1 > e_1 + < u_1 \mid e_2 > e_2)$$

Les vecteurs $\overrightarrow{Om_i} = \overrightarrow{OG} + \overrightarrow{Gm_i}$ forment ce qu'on appelle l'**approximation de rang 1 du nuage de points** dans $\mathbb{R}^2$.
Les points m_i sont les projections orthogonales des points M_i sur la droite de régression orthogonale.

L'équation de la **droite de régression orthogonale**, sur laquelle se situe l'approximation de rang 1 du nuage de points, peut prendre l'une des formes équivalentes :

$$< \overrightarrow{GM} \mid u_2 > = 0$$
$$(x - \overline{X})\, (s^2(Y) - \lambda_2) = (y - \overline{Y})\, Cov(X, Y)$$
$$(x - \overline{X})\, (\lambda_1 - s^2(X)) = (y - \overline{Y})\, Cov(X, Y)$$

$$(x - \overline{x})\ Cov\ (X, Y) = (y - \overline{y})\ (s^2\ (Y) - \lambda_1)$$
$$(x - \overline{x})\ Cov\ (X, Y) = (y - \overline{y})\ (\lambda_2 - s^2\ (X))$$

Chapitre 5 - REGRESSION MULTIPLE.

5. 1. POSITION ET RESOLUTION DU PROBLEME.

5.1.1. Position du problème.

Considérons trois variables statistiques réelles centrées X_0, Y_0, Z_0, définies par n triplets (x_{0i}, y_{0i}, z_{0i}), $i \in [1, n]$.
Nous considérons Z_0 comme la variable à expliquer et X_0 et Y_0 comme les variables explicatives.
Nous supposons que les observations laissent à penser que le nuage de points dans $\mathbb{R}^3$ pourrait être modélisé par un plan.
Le problème de la régression linéaire multiple de Z_0 en X_0 et Y_0 consiste à trouver un prédicteur

$$\hat{z}_0 = a\, X_0 + b\, Y_0$$

de Z_0, tel que le nuage de points $(x_{0i}, y_{0i}, \hat{z}_{0i} = a\, x_{0i} + b\, y_{0i})$, $i \in [1, n]$, soit aussi proche possible du nuage de points (x_{0i}, y_{0i}, z_{0i}), $i \in [1, n]$, au sens des moindres carrés.

L'approche euclidienne de ce problème dans $\mathbb{R}^n$ consiste à trouver un $\hat{z}_0 = a\, X_0 + b\, Y_0 \in \mathbb{R}^n$ tel que $S^2 = \| Z_0 - \hat{z}_0 \|^2_{D_{\frac{1}{n}}}$ soit minimum.

Le problème est donc de trouver, dans $\mathbb{R}^n$, un vecteur $\hat{z}_0$ du plan (= sous-espace vectoriel de dimension 2) Π défini par X_0 et Y_0, tel que le vecteur $Z_0 - \hat{z}_0$ ait une longueur minimum (au sens du produit scalaire défini par la matrice des poids $D_{\frac{1}{n}}$).

La solution sera fournie par le projeté orthogonal $\hat{z}_0$ de Z_0 sur Π.

5.1.2. Projeté orthogonal sur un plan.

a) Définition.

Si nous connaissons une base orthonormée $\{ u_1, u_2 \}$ d'un sous-espace vectoriel Π de dimension 2, défini dans $\mathbb{R}^n$ par les deux vecteurs X_0 et Y_0, nous savons calculer le projeté orthogonal de Z_0 sur u_1, c'est le vecteur $\frac{\langle Z_0 | u_1 \rangle_{D_{\frac{1}{n}}}}{\| u_1 \|^2_{D_{\frac{1}{n}}}} u_1 = < Z_0 \mid u_1 >_{D_{\frac{1}{n}}} u_1$ et nous savons calculer aussi le projeté orthogonal $< Z_0 \mid u_2 >_{D_{\frac{1}{n}}} u_2$ de Z_0 sur u_2.

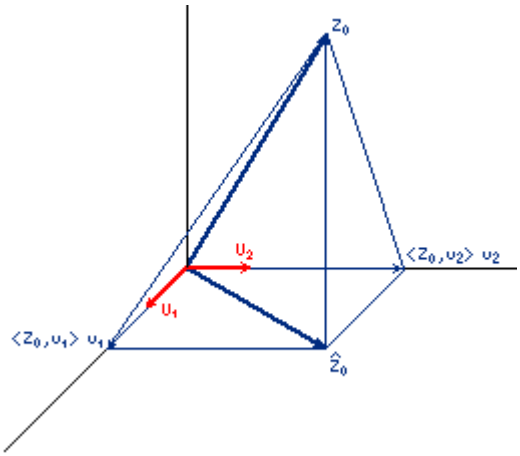
On appelle **projeté orthogonal** de Z_0 sur Π. l'unique vecteur $\hat{z}_0$ de Π tel que $Z_0 - \hat{z}_0$ soit orthogonal à Π.

Un tel vecteur existe et est unique.

Démonstration.

Notons $\hat{Z}_0$ le vecteur $< Z_0 \mid u_1 >_{D_1} u_1 + < Z_0 \mid u_2 >_{D_1} u_2$, somme des projetés orthogonaux de Z_0 sur les vecteurs u_1 et u_2.

$$< Z_0 - \hat{Z}_0 \mid u_1 >_{D_1} = < Z_0 \mid u_1 >_{D_1} - < \hat{Z}_0 \mid u_1 >_{D_1}$$
$$= < Z_0 \mid u_1 >_{D_1} - << Z_0 \mid u_1 >_{D_1} u_1 + < Z_0 \mid u_2 >_{D_1} u_2 \mid u_1 >_{D_1}$$
$$= < Z_0 \mid u_1 >_{D_1} - < Z_0 \mid u_1 >_{D_1} < u_1 \mid u_1 >_{D_1} + < Z_0 \mid u_2 >_{D_1} < u_2 \mid u_1 >_{D_1}$$
$$= < Z_0 \mid u_1 >_{D_1} - < Z_0 \mid u_1 >_{D_1}$$
$$= 0$$



$$< Z_0 - \hat{Z}_0 \mid u_2 >_{D_1} = < Z_0 \mid u_2 >_{D_1} - < \hat{Z}_0 \mid u_2 >_{D_1}$$
$$= < Z_0 \mid u_2 >_{D_1} - << Z_0 \mid u_1 >_{D_1} u_1 + < Z_0 \mid u_2 >_{D_1} u_2 \mid u_2 >_{D_1}$$
$$= < Z_0 \mid u_2 >_{D_1} - < Z_0 \mid u_1 >_{D_1} < u_1 \mid u_2 >_{D_1} + < Z_0 \mid u_2 >_{D_1} < u_2 \mid u_2 >_{D_1}$$
$$= < Z_0 \mid u_2 >_{D_1} - < Z_0 \mid u_2 >_{D_1}$$
$$= 0$$

Ainsi, $Z_0 - \hat{Z}_0$ est orthogonal à u_1 et à u_2, il est donc orthogonal à toute combinaison linéaire de u_1 et u_2, c'est-à-dire à tout élément de Π : on dit qu'il est orthogonal à Π.

Le projeté orthogonal de $\hat{Z}_0$ sur u_1 est

$$< \hat{Z}_0 \mid u_1 >_{D_1} u_1 = < Z_0 \mid u_1 >_{D_1} u_1.$$

Le projeté orthogonal de $\hat{Z}_0$ sur u_2 est

$$< \hat{Z}_0 \mid u_2 >_{D_1} u_2 = < Z_0 \mid u_2 >_{D_1} u_2.$$

Nous pouvons donc écrire :

$$\hat{Z}_0 = < Z_0 \mid u_1 >_{D_1} u_1 + < Z_0 \mid u_2 >_{D_1} u_2 = < \hat{Z}_0 \mid u_1 >_{D_1} u_1 + < \hat{Z}_0 \mid u_2 >_{D_1} u_2.$$

Réciproquement, si Z est un vecteur de Π tel que $Z_0 - Z$ soit orthogonal à Π, nous avons :

$$Z = < Z \mid u_1 >_{D_1} u_1 + < Z \mid u_2 >_{D_1} u_2 = < Z_0 \mid u_1 >_{D_1} u_1 + < Z_0 \mid u_2 >_{D_1} u_2 = \hat{Z}_0.$$

Le vecteur :

$$\hat{Z}_0 = < Z_0 \mid u_1 >_{D_{\frac{1}{n}}} u_1 + < Z_0 \mid u_2 >_{D_{\frac{1}{n}}} u_2$$

est donc l'unique vecteur de Π tel que $Z_0 - \hat{Z}_0$ soit orthogonal à Π : c'est, par définition, le projeté orthogonal de Z_0 sur Π.

La relation :

$$\hat{\hat{Z}}_0 = < \hat{Z}_0 \mid u_1 >_{D_{\frac{1}{n}}} u_1 + < \hat{Z}_0 \mid u_2 >_{D_{\frac{1}{n}}} u_2$$

signifie que le projeté orthogonal de $\hat{Z}_0$ sur le plan Π est $\hat{Z}_0$.

b) Propriété du projeté orthogonal.

Le projeté orthogonal de Z_0 sur Π est le vecteur Z de Π, qui minimise la quantité $\| Z_0 - Z \|^2_{D_{\frac{1}{n}}}$.

Démonstration.

Soit Z un vecteur appartenant au sous-espace Π.
Soit $\hat{Z}_0 = < Z_0 \mid u_1 >_{D_{\frac{1}{n}}} u_1 + < Z_0 \mid u_2 >_{D_{\frac{1}{n}}} u_2$ le projeté orthogonal de Z_0 sur Π.

$$\| Z_0 - Z \|^2_{D_{\frac{1}{n}}} = \| Z_0 - \hat{Z}_0 + \hat{Z}_0 - Z \|^2_{D_{\frac{1}{n}}}$$

Or $Z_0 - \hat{Z}_0$ est orthogonal à Π, donc orthogonal à tout élément de Π, donc $Z_0 - \hat{Z}_0$ est orthogonal à $\hat{Z}_0$ et à Z, donc aussi à $\hat{Z}_0 - Z$.
Le théorème de Pythagore s'applique :

$$\| Z_0 - \hat{Z}_0 + \hat{Z}_0 - Z \|^2_{D_{\frac{1}{n}}} = \| Z_0 - \hat{Z}_0 \|^2_{D_{\frac{1}{n}}} + \| \hat{Z}_0 - Z \|^2_{D_{\frac{1}{n}}}$$

$$\| Z_0 - Z \|^2_{D_{\frac{1}{n}}} = \| Z_0 - \hat{Z}_0 \|^2_{D_{\frac{1}{n}}} + \| \hat{Z}_0 - Z \|^2_{D_{\frac{1}{n}}}$$

Cette relation montre que $\| Z_0 - Z \|^2_{D_{\frac{1}{n}}}$ atteint sa valeur minimum $\| Z_0 - \hat{Z}_0 \|^2_{D_{\frac{1}{n}}}$ lorsque $Z = \hat{Z}_0$.

Notre problème initial se trouve résolu :

Le prédicteur $\hat{Z}_0 = a\, X_0 + b\, Y_0$ de Z_0 qui rend minimum la quantité $S^2 = \| Z_0 - \hat{Z}_0 \|^2_{D_{\frac{1}{n}}}$ est le projeté orthogonal de Z_0 dans le plan Π défini par X_0 et Y_0.

La seule chose qu'il nous reste à faire dans la suite, est d'expliciter ce projeté orthogonal en fonction des données (x_{0i}, y_{0i}, z_{0i}), $i \in [1, n]$.

5.1.3. Choix d'une base orthonormée { u_1, u_2 }.

Dans le plan Π défini par X_0 et Y_0, nous pouvons définir un premier vecteur normé u_1 par :

$$u_1 = \frac{X_0}{\|X_0\|_{D_{\frac{1}{n}}}} = \frac{X_0}{s(X)}.$$

On a, en effet : $s^2(X) = \| X_0 \|^2_{D_{\frac{1}{n}}}$.

Le projeté orthogonal de Y_0 sur X_0 est $\frac{\langle Y_0 | X_0 \rangle_{D_{\frac{1}{n}}}}{\|X_0\|^2_{D_{\frac{1}{n}}}} X_0$ et $Y_0 - \frac{\langle Y_0 | X_0 \rangle_{D_{\frac{1}{n}}}}{\|X_0\|^2_{D_{\frac{1}{n}}}} X_0$ est orthogonal à X_0.

Le carré de sa norme est donné par :

$$\left\| Y_0 - \frac{\langle Y_0 | X_0 \rangle_{D_{\frac{1}{n}}}}{\|X_0\|^2_{D_{\frac{1}{n}}}} X_0 \right\|^2_{D_{\frac{1}{n}}} = \| Y_0 \|^2_{D_{\frac{1}{n}}} + \left(\frac{\langle Y_0 | X_0 \rangle_{D_{\frac{1}{n}}}}{\|X_0\|^2_{D_{\frac{1}{n}}}} \right)^2 \| X_0 \|^2_{D_{\frac{1}{n}}} . - 2 \frac{\langle Y_0 | X_0 \rangle_{D_{\frac{1}{n}}}}{\|X_0\|^2_{D_{\frac{1}{n}}}} < Y_0 | X_0 >_{D_{\frac{1}{n}}}$$

$$= s^2(Y) - s^2(Y) \left(\frac{Cov[X,Y]}{s(X)\, s(Y)} \right)^2 = s^2(Y)\,(1 - r_{XY}^2) = \frac{s^2[X]\, s^2[Y] - (Cov[X,Y])^2}{s^2[X]}$$

On peut donc prendre dans le plan Π, pour vecteur normé u_2 orthogonal à u_1, le vecteur :

$$u_2 = \frac{1}{s(X)\sqrt{1 - r_{XY}^2}} \left(Y_0 - \frac{Cov[X,Y]}{s^2[X]} X_0 \right) = \frac{s(X)}{\sqrt{s^2[X]\, s^2[Y] - (Cov[X,Y])^2}} \left(Y_0 - \frac{Cov[X,Y]}{s^2[X]} X_0 \right)$$

Les vecteurs :

$$u_1 = \frac{X_0}{s(X)}$$

$$u_2 = \frac{s(X)}{\sqrt{s^2[X]\, s^2[Y] - (Cov[X,Y])^2}} \left(Y_0 - \frac{Cov[X,Y]}{s^2[X]} X_0 \right)$$

forment une base orthonormée du plan Π défini par X_0 et Y_0.

5.1.4. Calcul du projeté orthogonal de Z_0.

Soit

$$\hat{Z}_0 = < Z_0 | u_1 >_{D_{\frac{1}{n}}} u_1 + < Z_0 | u_2 >_{D_{\frac{1}{n}}} u_2$$

le projeté orthogonal de Z_0 sur Π.

La première composante est le projeté orthogonal de Z_0 sur u_1 :

$$< Z_0 | u_1 >_{D_{\frac{1}{n}}} u_1 = < Z_0 | \frac{X_0}{s(X)} >_{D_{\frac{1}{n}}} \frac{X_0}{s(X)} = \frac{Cov[X,Z]}{s^2[X]} X_0$$

C'est aussi le projeté orthogonal de Z_0 sur X_0.

La deuxième composante est le projeté orthogonal de Z_0 sur u_2 :

$$< Z_0 \mid u_2 >_{D_{\frac{1}{n}}} u_2 = < Z_0 \mid \frac{s[X]}{\sqrt{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}}\left(Y_0 - \frac{Cov[X,Y]}{s^2[X]}\,X_0 \right) >_{D_{\frac{1}{n}}} \frac{s[X]}{\sqrt{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}}$$
$$\left(Y_0 - \frac{Cov[X,Y]}{s^2[X]}\,X_0 \right)$$
$$= \frac{s^2[X]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\left(< Z_0 \mid Y_0 >_{D_{\frac{1}{n}}} - \frac{Cov[X,Y]}{s^2[X]} < Z_0 \mid X_0 >_{D_{\frac{1}{n}}} \right)\left(Y_0 - \frac{Cov[X,Y]}{s^2[X]}\,X_0 \right)$$
$$= \frac{s^2[X]\,Cov[Z,Y]-Cov[Z,X]\,Cov[X,Y]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\left(Y_0 - \frac{Cov[X,Y]}{s^2[X]}\,X_0 \right)$$

Au total, nous obtenons :

$$\hat{Z}_0 = \frac{Cov[X,Z]}{s^2[X]}\,X_0 + \frac{s^2[X]\,Cov[Z,Y]-Cov[Z,X]\,Cov[X,Y]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\left(Y_0 - \frac{Cov[X,Y]}{s^2[X]}\,X_0 \right)$$
$$= \frac{1}{s^2[X]}\left(Cov\,(X, Z) - \frac{s^2[X]\,Cov[Z,Y]-Cov[Z,X]\,Cov[X,Y]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2} Cov\,(X, Y) \right) X_0 +$$
$$\frac{s^2[X]\,Cov[Z,Y]-Cov[Z,X]\,Cov[X,Y]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\,Y_0$$
$$= \frac{s^2[Y]\,Cov[X,Z]-Cov[X,Y]\,Cov[Y,Z]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\,X_0 + \frac{s^2[X]\,Cov[Y,Z]-Cov[X,Y]\,Cov[X,Z]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\,Y_0$$

$$\boxed{\hat{Z}_0 = \frac{s^2[Y]\,Cov[X,Z]-Cov[X,Y]\,Cov[Y,Z]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\,X_0 + \frac{s^2[X]\,Cov[Y,Z]-Cov[X,Y]\,Cov[X,Z]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\,Y_0}$$

Cette expression est symétrique en X et Y.
On sait calculer les quantités qui interviennent dans cette expression en fonction des données (x_{0i}, y_{0i}, z_{0i}), $i \in [1, n]$.
On commence par calculer la matrice des variances-covariances :

$$A = \frac{1}{n}\begin{pmatrix} x_{01} & \cdots & x_{0n} \\ y_{01} & \cdots & y_{0n} \\ z_{01} & \cdots & z_{0n} \end{pmatrix}\begin{pmatrix} x_{01} & y_{01} & z_{01} \\ \vdots & \vdots & \vdots \\ x_{0n} & y_{0n} & z_{0n} \end{pmatrix} = \begin{pmatrix} s^2[X] & Cov[X,Y] & Cov[X,Z] \\ Cov[X,Y] & s^2[Y] & Cov[Y,Z] \\ Cov[X,Z] & Cov[Y,Z] & s^2[Z] \end{pmatrix}$$

Formellement, la relation $\hat{Z}_0 = \frac{s^2[Y]\,Cov[X,Z]-Cov[X,Y]\,Cov[Y,Z]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\,X_0 + \frac{s^2[X]\,Cov[Y,Z]-Cov[X,Y]\,Cov[X,Z]}{s^2[X]\,s^2[Y]-(Cov[X,Y])^2}\,Y_0$ peut se mémoriser comme un "déterminant" :

$$\begin{vmatrix} s^2[X] & Cov[X,Y] & X_0 \\ Cov[X,Y] & s^2[Y] & Y_0 \\ Cov[X,Z] & Cov[Y,Z] & \hat{Z}_0 \end{vmatrix} = 0$$

On a remplacé la dernière colonne de la matrice des variances-covariances par $\begin{pmatrix} X_0 \\ Y_0 \\ \hat{Z}_0 \end{pmatrix}$.

5.2. COEFFICIENT DE CORRELATION MULTIPLE.

5.2.1. Définition.

Nous connaissons déjà les formules donnant les coefficients de corrélation linéaire entre deux variables :

$$r_{XY} = \frac{Cov[X,Y]}{s[X]\,s[Y]} = \frac{\langle X_0 | Y_0 \rangle_{D_{\frac{1}{n}}}}{s[X]\,s[Y]} \; ; \; r_{XZ} = \frac{Cov[X,Z]}{s[X]\,s[Z]} \; ; \; r_{YZ} = \frac{Cov[Y,Z]}{s[Y]\,s[Z]}.$$

Les coefficients de X_0 et Y_0 dans l'expression de $\hat{Z}_0$ deviennent :

$$\frac{s^2[Y]\,Cov[X,Z] - Cov[X,Y]\,Cov[Y,Z]}{s^2[X]\,s^2[Y] - (Cov[X,Y])^2} = \frac{s^2[Y]\,r_{XZ}\,s[X]\,s[Z] - r_{XY}\,s[X]\,s[Y]\,r_{YZ}\,s[Y]\,s[Z]}{s^2[X]\,s^2[Y] - (r_{XY}\,s[X]\,s[Y])^2} = \frac{s[X]\,s^2[Y]\,s[Z]}{s^2[X]\,s^2[Y]} \times \frac{r_{XZ} - r_{XY}\,r_{YZ}}{1 - r_{XY}^2} = \frac{s[Z]}{s[X]} \frac{r_{XZ} - r_{XY}\,r_{YZ}}{1 - r_{XY}^2}$$

et, en échangeant X et Y :

$$\frac{s^2[X]\,Cov[Y,Z] - Cov[X,Y]\,Cov[X,Z]}{s^2[X]\,s^2[Y] - (Cov[X,Y])^2} = \frac{s[Z]}{s[Y]} \frac{r_{YZ} - r_{XY}\,r_{XZ}}{1 - r_{XY}^2}$$

En reportant, dans l'expression de $\hat{Z}_0$, les expressions obtenues pour les coefficients, on obtient :

$$\hat{Z}_0 = \frac{s[Z]}{s[X]} \frac{r_{XZ} - r_{XY}\,r_{YZ}}{1 - r_{XY}^2} X_0 + \frac{s[Z]}{s[Y]} \frac{r_{YZ} - r_{XY}\,r_{XZ}}{1 - r_{XY}^2} Y_0$$

$$\frac{\hat{z}_0}{s[Z]} = \frac{r_{XZ} - r_{XY}\,r_{YZ}}{1 - r_{XY}^2} \frac{X_0}{s[X]} + \frac{r_{YZ} - r_{XY}\,r_{XZ}}{1 - r_{XY}^2} \frac{Y_0}{s[Y]}$$

Les vecteurs $\frac{X_0}{s[X]}$ et $\frac{Y_0}{s[Y]}$ sont normés pour le produit scalaire de $\mathbb{R}^n$: $\| X_0 \|_{D_{\frac{1}{n}}}^2 = s^2(X)$ et $\| Y_0 \|_{D_{\frac{1}{n}}}^2 = s^2(Y)$.

$$\left\| \frac{\hat{z}_0}{s[Z]} \right\|_{D_{\frac{1}{n}}}^2 = \frac{\| \hat{z}_0 \|_{D_{\frac{1}{n}}}^2}{s^2[Z]} = \left(\frac{r_{XZ} - r_{XY}\,r_{YZ}}{1 - r_{XY}^2} \right)^2 + \left(\frac{r_{YZ} - r_{XY}\,r_{XZ}}{1 - r_{XY}^2} \right)^2 + 2 \frac{r_{XZ} - r_{XY}\,r_{YZ}}{1 - r_{XY}^2} \frac{r_{YZ} - r_{XY}\,r_{XZ}}{1 - r_{XY}^2} \frac{\langle X_0 | Y_0 \rangle_{D_{\frac{1}{n}}}}{s[X]\,s[Y]}$$

$$= \frac{1}{(1 - r_{XY}^2)^2} \left(r_{XZ}^2 + r_{XY}^2\,r_{YZ}^2 - 2\,r_{XY}\,r_{XZ}\,r_{YZ} + r_{YZ}^2 + r_{XY}^2\,r_{XZ}^2 - 2\,r_{XY}\,r_{XZ}\,r_{YZ} + 2\,r_{XY}(r_{XZ}\,r_{YZ} - r_{XY}\,r_{XZ}^2 - r_{XY}\,r_{YZ}^2 + r_{XY}^2\,r_{XZ}\,r_{YZ}) \right)$$

$$= \frac{1}{(1 - r_{XY}^2)^2} \left(r_{XZ}^2 + r_{XY}^2\,r_{YZ}^2 - 2\,r_{XY}\,r_{XZ}\,r_{YZ} + r_{YZ}^2 + r_{XY}^2\,r_{XZ}^2 - 2\,r_{XY}\,r_{XZ}\,r_{YZ} + 2\,r_{XY}\,r_{XZ}\,r_{YZ} - 2\,r_{XY}^2\,r_{XZ}^2 - 2\,r_{XY}^2\,r_{YZ}^2 + 2\,r_{XY}^3\,r_{XZ}\,r_{YZ}) \right)$$

$$= \frac{1}{(1-r_{XY}^2)^2}\left(r_{XZ}^2 + r_{XY}^2 r_{XZ}^2 - 2 r_{XY}^2 r_{XZ}^2 + r_{YZ}^2 + r_{XY}^2 r_{YZ}^2 - 2 r_{XY}^2 r_{YZ}^2 - 2 r_{XY} r_{XZ} r_{YZ} - 2 r_{XY} r_{XZ} r_{YZ} + 2 r_{XY} r_{XZ} r_{YZ} + 2 r_{XY}^3 r_{XZ} r_{YZ}) \right)$$

$$= \frac{1}{(1-r_{XY}^2)^2}\left(r_{XZ}^2 - r_{XY}^2 r_{XZ}^2 + r_{YZ}^2 - r_{XY}^2 r_{YZ}^2 - 2 r_{XY} r_{XZ} r_{YZ} + 2 r_{XY}^3 r_{XZ} r_{YZ}) \right)$$

$$= \frac{1}{(1-r_{XY}^2)^2}\left(r_{XZ}^2 (1 - r_{XY}^2) + r_{YZ}^2 (1 - r_{XY}^2) - 2 r_{XY} r_{XZ} r_{YZ} (1 - r_{XY}^2) \right)$$

$$= \frac{1}{1-r_{XY}^2}\left(r_{XZ}^2 + r_{YZ}^2 - 2 r_{XY} r_{XZ} r_{YZ} \right)$$

Le coefficient :

$$R_{Z|XY} = \sqrt{\frac{1}{1-r_{XY}^2}\left(r_{XZ}^2 + r_{YZ}^2 - 2 r_{XY} r_{XZ} r_{YZ}\right)}$$

s'appelle le **coefficient de corrélation linéaire multiple** de Z en X, Y.

La variance du prédicteur de Z est donnée par :

$$s^2(\hat{Z}) = \| \hat{Z}_0 \|^2_{D_{\frac{1}{n}}} = R_{Z|XY}^2 \, s^2(Z)$$

5.2.2. Propriétés.

a) Validité du prédicteur de *Z*.

La variance de Z s'écrit :

$$s^2(Z) = s^2(Z_0) = \| Z_0 \|^2_{D_{\frac{1}{n}}} = \| Z_0 - \hat{Z}_0 + \hat{Z}_0 \|^2_{D_{\frac{1}{n}}} = \| Z_0 - \hat{Z}_0 \|^2_{D_{\frac{1}{n}}} + \| \hat{Z}_0 \|^2_{D_{\frac{1}{n}}}$$

Or $\| Z_0 - \hat{Z}_0 \|^2_{D_{\frac{1}{n}}}$ est la valeur minimum de la quantité $S^2 = \| Z_0 - \hat{Z} \|^2_{D_{\frac{1}{n}}}$ pour les $\hat{Z} \in \Pi$: $\| Z_0 - \hat{Z}_0 \|^2_{D_{\frac{1}{n}}} = S^2_{min}$, c'est la variance "**résiduelle**", donc

$$s^2(Z) = S^2_{min} + R_{Z|XY}^2 \, s^2(Z)$$

On retrouve la même formule de décomposition de la variance que pour la régression linéaire : la variance de Z est la somme de la variance expliquée $R_{Z|XY}^2 \, s^2(Z)$ par la régression linéaire multiple, et de la variance résiduelle $S^2_{min} = (1 - R_{Z|XY}^2) \, s^2(Z)$.

Plus le coefficient $R_{Z|XY}^2$ est proche de 1, plus la part de variance de Z expliquée par la régression linéaire multiple en X et Y est grande, donc meilleur est le prédicteur linéaire $\hat{Z}_0$.
La validité du prédicteur $\hat{Z}_0$ est mesurée par le coefficient $R_{Z|XY}^2$.

b) Calcul pratique du coefficient de corrélation linéaire multiple.

En pratique, le calcul du coefficient de corrélation linéaire multiple $R_{Z|XY}$ s'effectue de la façon

suivante :

— On calcule la **matrice des corrélations** de X et Y à partir de la matrice $V_{XY} = \begin{pmatrix} \frac{x_1-\overline{X}}{s(X)} & \frac{y_1-\overline{Y}}{s(Y)} \\ \vdots & \vdots \\ \frac{x_n-\overline{X}}{s(X)} & \frac{y_n-\overline{Y}}{s(Y)} \end{pmatrix}$ des données (X, Y) réduites :

$$C_{XY} = \begin{pmatrix} 1 & r_{XY} \\ r_{XY} & 1 \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \frac{x_1-\overline{X}}{s(X)} & \cdots & \frac{x_n-\overline{X}}{s(X)} \\ \frac{y_1-\overline{Y}}{s(Y)} & \cdots & \frac{y_n-\overline{Y}}{s(Y)} \end{pmatrix} \begin{pmatrix} \frac{x_1-\overline{X}}{s(X)} & \frac{y_1-\overline{Y}}{s(Y)} \\ \vdots & \vdots \\ \frac{x_n-\overline{X}}{s(X)} & \frac{y_n-\overline{Y}}{s(Y)} \end{pmatrix} = {}^t V_{XY} D_{\frac{1}{n}} V_{XY}.$$

— On calcule l'**inverse** de cette matrice des corrélations :

$$C_{XY}^{-1} = \frac{1}{1-r_{XY}^2} \begin{pmatrix} 1 & -r_{XY} \\ -r_{XY} & 1 \end{pmatrix}$$

— La matrice des **coefficients de corrélation linéaire** de X et Y avec Z, peut se calculer à partir de la matrice V_{XY} et de la variable centrée réduite $V_Z = \begin{pmatrix} \frac{z_1-\overline{Z}}{s(Z)} \\ \vdots \\ \frac{z_n-\overline{Z}}{s(Z)} \end{pmatrix}$ par la formule :

$$\begin{pmatrix} r_{XZ} \\ r_{YZ} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \frac{x_1-\overline{X}}{s(X)} & \cdots & \frac{x_n-\overline{X}}{s(X)} \\ \frac{y_1-\overline{Y}}{s(Y)} & \cdots & \frac{y_n-\overline{Y}}{s(Y)} \end{pmatrix} \begin{pmatrix} \frac{z_1-\overline{Z}}{s(Z)} \\ \vdots \\ \frac{z_n-\overline{Z}}{s(Z)} \end{pmatrix} = {}^t V_{XY} D_{\frac{1}{n}} V_Z.$$

— Le **coefficient de corrélation linéaire multiple** $R_{Z|XY}$ est donné par la formule :

$$R_{Z|XY}^2 = \frac{1}{1-r_{XY}^2} \left(r_{XZ}^2 + r_{YZ}^2 - 2\, r_{XY}\, r_{XZ}\, r_{YZ} \right) = (r_{XZ} \quad r_{YZ})\, C_{XY}^{-1} \begin{pmatrix} r_{XZ} \\ r_{YZ} \end{pmatrix}$$

formule que l'on peut écrire directement en fonction des **données centrées réduites** :

$$\boxed{R_{Z|XY}^2 = {}^t\left({}^t V_{XY} D_{\frac{1}{n}} V_Z \right) \left({}^t V_{XY} D_{\frac{1}{n}} V_{XY} \right)^{-1} \left({}^t V_{XY} D_{\frac{1}{n}} V_Z \right).}$$

Remarquons, à l'usage des débutants, qu'il ne faudrait pas écrire :

$$\left({}^t V_{XY} D_{\frac{1}{n}} V_{XY} \right)^{-1} = V_{XY}^{-1} D_{\frac{1}{n}}^{-1} {}^t V_{XY}^{-1}$$

puisque la matrice V_{XY}, à n lignes et 2 colonnes, n'est pas inversible, alors que la matrice produit $C = {}^t V_{XY} D_{\frac{1}{n}} V_{XY}$, à 2 lignes et 2 colonnes, est inversible.

5.2.3. Application : technique de la régression pas à pas.

Pour connaître le rôle de chacune des variables explicatives, on calcule les coefficients de

détermination r_{XZ}^2 et r_{YZ}^2 et le coefficient $R_{Z|XY}^2$.
Chacun de ces coefficients représente le pourcentage de variance de Z restitué par le prédicteur correspondant.
On conservera, pour prédicteur de Z le modèle qui restituera significativement le meilleur résultat :

$\hat{z}_0 = c\, X_0$
$\hat{z}_0 = d\, Y_0$
$\hat{z}_0 = a\, X_0 + b\, Y_0$.

La théorie de la régression multiple que nous venons d'exposer dans le cas de deux variables explicatives peut se généraliser au cas de p variables explicatives, avec $p > 2$.

Chapitre 6 - INITIATION A LA THEORIE DES SONDAGES.

6. 1. GENERALITES.

6.1.1. Introduction.

L'étude exhaustive d'un caractère donné dans une population est un **recensement**.
Elle se heurte souvent à une impossibilité matérielle : coût trop élevé, ou destruction des individus étudiés.

Les méthodes d'analise quantitative ont alors recours à la théorie des sondages, qui consiste à étudier un sous-ensemble de la population qu'on appelle un **échantillon**.

La théorie des sondages pose deux types de problèmes :

- L'échantillon doit être représentatif de la population : c'est la **théorie de l'échantillonnage**.
- Les techniques numériques utilisées sur les observations expérimentales doivent conduire à des résultats fiables, c'est-à-dire donnant une bonne représentation des paramètres inconnus de la population : c'est la **théorie de l'estimation** et des **tests**.

Les deux problèmes sont liés : la méthode d'échantillonnage utilisée a une influence sur les estimations obtenues.

En résumé, nous pouvons dire que la théorie des sondages est un outil mathématique permettant, à partir d'observations expérimentales partielles, de tenter d'atteindre une réalité inaccessible.

6.1.2. Avantages de la méthode d'enquêtes par sondages.

La méthode d'enquêtes par sondages présente sur le recensement (lorsqu'il est possible) les avantages suivants :

1. Coût plus réduit.
2. Plus grande vitesse d'exécution (notamment pour les sondages d'opinions).
3. Plus grande fiabilité des résultats : le personnel étant plus réduit, il peut être plus qualifié.
4. Moins de risque d'erreur : le volume des données à traiter est plus faible.
5. Plus grand champ d'application, notamment dans le cas de destruction des unités testées.

6.1.3. Etapes d'une enquête par sondage.

Pour effectuer une enquête par sondage, il est indispensable de respecter les instructions suivantes.

- Dresser une liste claire des objectifs de l'enquête.
- Etablir avec précision la population à échantillonner.
- Etablir une liste précise et courte des données à collecter.
- Définir le choix des méthodes de mesure : téléphone, convocations, visites à domicile, ...
- Etablir, lorsque c'est possible, le degré de précision désiré afin d'analyser le rapport des coûts et des avantages.
- Déterminer l'unité de l'échantillonnage : personne physique, collectivité, ...
- Etablir le plan de l'échantillonnage ou la méthode de sélection.

- Faire parfois une pré-enquête courte.
- Organiser le travail sur le terrain.
- Récolter les données, les présenter, les synthétiser par traitement statistique.
- Conserver les données pour pouvoir les réutiliser.

6.2. DIVERS TYPES DE SONDAGES.

Pour effectuer un sondage dans une population, c'est-à-dire pour en extraire un échantillon, deux types de méthodes sont employées : méthodes empiriques et méthodes aléatoires. Seules les méthodes aléatoires permettent d'utiliser la théorie de l'estimation.

6.2.1. Méthodes empiriques : sondages raisonnés.

Ce sont les plus connues du grand public et les plus utilisées par les instituts de sondage d'opinion. La précision de ces méthodes ne peut être calculée et leur réussite n'est que le résultat d'une longue pratique et de l'habileté professionnelle.
Les éléments sondés sont choisis dans la population suivant des critères fixés a pirori.

6.2.1.1. Méthode des unités types.

Elle repose sur l'idée suivante : les différentes variables attachées à un individu de la population n'étant pas indépendantes, un individu qui se trouve dans la moyenne de la population pour un certain nombre de caractères impportants, sera également peu différent de la moyenne pour les autres caractères.
La méthode consiste donc à diviser la population en un certain nombre de sous-ensembles relativement homogènes et à représenter chacun d'eux par une unité-type.
On choisit donc des unités d'individus que l'on considère comme fortement représentatives de certaines catégories de population : cantons-types, bureau de vote pilotes, dont les résultats observés sur de longues périodes figurent les résultats définitifs d'une région ou d'une ville, etc.

Exemple.

L'INSEE décomposa en 1942 la France en 600 régions agricoles et, dans chaque région, désigna un canton-ype.
Comme il y a en France environ 3000 cantons, la désignation de 600 cantons-types permettait de réduire d'un facteur 5 l'ampleur d'une étude des cantons.

6.2.1.2. Méthode des quotas.

L'enquêteur prélève librement son échantillon, à condition de respecter une composition donnée à l'avance (pourcentage fixé d'agriculteurs, d'ouvriers, de cadres, etc., par exemple).
Cette méthode est facile, mais aucun intervalle de confiance ne peut être donné.
Elle suppose implicitement que les catégories retenues pour la détermination des quotas sont pertinentes quant à l'objet de l'étude, ce qui est bien difficile à établir.
Pour diminuer l'arbitraire du choix, on impose à l'enquêteur des normes de déplacement géographique : c'est la **méthode de Politz**.

On utilise souvent des "**panels**", qui sont des échantillons permanents dont on étudie l'évolution.

Exemples.

- Panel d'audience à la télévision (médiamétrie, centres d'études d'opinion, ...).

- Panel de consommateurs (SECODIF : 4 500 ménages).
- Panel de détaillants (SOFRES).

Ces panels sont utilisés en marketing (lancement d'un produit, transfert de marques, etc.).

6.2.2. Méthodes aléatoires.

Les éléments sondés sont extraits **au hasard** d'une liste connue a priori de la population, appelée **base de sondage**.

Exemples.

1. Liste d'immatriculation des véhicules automobiles en France.
   C'est une très bonne base car elle est mise à jour régulièrement (cartes grises neuves, cartes grises à détruire).
2. Répertoire des entreprises (SIREN).
   Chaque entreprise possède un numéro d'immatriculation à neuf chiffres, un nom ou raison sociale, une adresse exacte.
3. L'annuaire téléphonique est une mauvaise base de sondage car d'une part, tout individu ne possède pas obligatoirement un téléphone et, d'autre part, un individu peut posséder un téléphone et ne pas figurer sur l'annuaire (la liste rouge représente environ 8 % des abonnés et l'annuaire ne recense pas les téléphones portables, soit environ 40 % des téléphones).

Les bases de sondages sont en général établies à partir des résultats d'un recensement et elles sont corrigées périodiquement entre deux recensements.
Le tirage de l'échantillon est effectué dans la base de sondage selon des critères spécifiques à chaque méthode (plan de sondage).
Cette méthode de travail ne laisse aucune initiative aux enquêteurs : il est très simple de contrôler leur travail.

6.2.2.1. Sondage élémentaire : échantillon aléatoire simple.

Dans un **échantillon aléatoire simple**, les éléments constituant l'échantillon sont extraits au hasard (à l'aide d'une table de nombres au hasard, par exemple) d'une liste de la population.
On extrait ainsi n individus d'une population de taille N.
Le tirage peut s'effectuer avec ou sans remise, renvoyant ainsi généralement à un modèle de loi binomiale (avec remise), ou hypergéométrique (sans remise).
Si le tirage s'effectue avec remise, l'échantillon aléatoire simple est dit indépendant (EASI = **Echantillon Aléatoire Simple et Indépendant**).

La méthode permet de calculer des intervalles de confiance, comme nous le verrons plus loin.
Le rapport $f = \frac{n}{N}$ s'appelle le **taux de sondage**.
Par exemple, l'INSEE utilise des taux de sondage de l'ordre de $\frac{1}{1\,500}$ pour les enquêtes sur les conditions de vie des ménages.

Exemple.

Nous voulons extraire un échantillon de 8 individus dans une population formée de 437 individus.
Nous numérotons les individus de la population de 1 à 437.
Nous considérons trois colonnes consécutives d'une page de nombres au hasard : ils forment des nombres au hasard à trois chiffres.

Nous lisons ces nombres de trois chiffres en ne retenant que ceux qui sont compris entre 001 et 437.
Lorsque nous avons retenus 8 nombres, notre échantillon est constitué des 8 individus désignés dans la population par ces huit nombres.
Selon que nous effectuons un tirage avec ou sans remise, nous garderons ou écarterons un individu déjà tiré.

L'inconvénient majeur de la méthode élémentaire est son coût : les individus tirés peuvent être très éloignés géographiquement.

6.2.2.2. Sondage stratifié.

La population étudiée Ω est partitionnée en q sous-populations Ω_1, Ω_2, ... , Ω_q, appelées "**strates**".
L'échantillon est constitué de la réunion de q échantillons choisis au hasard, un par strate : nous effectuons dans chaque strate un échantillonnage simple.

Exemple.

$\Omega = \{1, 2, 3, 4, 5\}$, $\Omega_1 = \{1, 2\}$, $\Omega_2 = \{3, 4, 5\}$.
Nous sélectionnons trois individus, dont un dans Ω_1 et deux dans Ω_2.
Nous obtenons l'un des six échantillons possibles.

Cette méthode se justifie par deux raisons essentielles :

1. — L'existence d'une stratification de fait, soit pour des raisons géographiques, soit pour des raisons administratives.

Exemple 1 : enquête sur les conditions de vie pénitentiaire en France.

La population est celle des détenus en France
Les strates sont les populations de détenus dans les divers établissements pénitentiaires.

Exemple 2 : enquête sur la consommation par un organisme disposant de bureaux départementaux.

La population est celle des consommateurs français.
Les strates sont les consommateurs de chaque département.

2. — Un caractère étudié dans la population peut varier sous l'influence d'un certain nombre de facteurs.

Pour éliminer au mieux les risques de biais, nous créons des strates homogènes et, dans chacune d'elles, nous extrayons un échantillon aléatoire simple.

Exemple.

Pour étudier la consommation de tabac, si nous estimons que l'âge et le sexe sont des facteurs très influents, nous partageons la population en strates du type :
– Hommes de moins de 20 ans,
– Hommes de 20 à 30 ans,

– etc.
– Femmes de moins de 20 ans,
– Femmes de 20 à 30 ans,
– etc.

De chaque strate, nous extrayons un échantillon aléatoire simple.

6.2.2.3. Echantillonnage systématique.

Les individus de la population Ω sont numérotés de 1 à N.
Pour sélectionner n individus, nous partageons la population en $k = \frac{N}{n}$ groupes : $\{1, \dots, k\}$, $\{1 + k, \dots, 2k\}$, ... , $\{1 + (n-1)k, \dots, N\}$.
Nous choisissons au hasard l'individu i par les individus numérotés de 1 à k.
Nous constituons notre échantillon des individus $\{i, i + k, i + 2k, \dots, i + (n-1)k\}$.
Le choix de l'individu i détermine entièrement la constitution de l'échantillon.

Exemple.

$\Omega = \{1, \dots, 20\}$, $k = 4$.
Les échantillons possibles sont : {1, 5, 9, 13, 17}, {2, 6, 10, 14, 18}, {3, 7, 11, 15, 19}, {4, 8, 12, 16, 20}.

Cette méthode est bien adaptée à la sélection de cartes dans un fichier, ou au prélèvement de pièces dans une fabrication pour un contrôle de qualité.
Elle présente une certaine analogie avec la méthode précédente d'échantillonnage stratifié.

6.2.2.4. Echantillonnage à plusieurs degrés.

La population Ω est divisée en sous-populations appelées unités primaires.
Chaque unité primaire est divisée en unités secondaires, etc.
Nous effectuons des tirages au hasard en cascade : nous tirons des unités primaires ; dans chaque unité primaire, nous tirons une unité secondaire, etc.

Exemple.

L'INSEE effectue des échantillonnages à quatre niveaux : départements, cantons, communes, ménages.

Cette méthode permet une exécution rapide.
Elle est économique, car elle focalise les tirages.

La méthode de tirage au hasard à chaque niveau peut varier suivant le cas, par exemple tirage proportionnel aux unités qu'il contient, ou tirage équiprobable.
Nous disons alors que nous pouvons avoir des tirages avec **probabilités inégales**.

Cas particulier : tirage par grappes.

Nous choisissons des grappes pour lesquelles nous gardons tous les "grains", ou individus.
Une "grappe" est un groupe d'individus de même nature.

Exemple : ménages d'un même immeuble.

6.2.2.5. Conclusion.

En pratique, les diverses méthodes aléatoires peuvent être mêlées pour améliorer le rendement.
Pour chacune d'elle, nous pourrons varier les critères de tirage au hasard de chaque individu : avec remise, sans remise, avec des probabilités égales ou inégales.

6.3. ESTIMATION DES PARAMETRES.

6.3.1. Notion de paramètre.

Nous considérons une population Ω de taille finie N.
Dans cette population, nous étudions un caractère quantitatif réel prenant les valeurs réelles x_i, $i \in \{1, ... N\}$.
La fonction de répartition empirique $F_N(x)$ est une fonction en escalier.
La variable statistique représentant le caractère étudié peut être une variable quantitative discrète ou continue.

Le problème est de modéliser la fonction de répartition empirique $F_N(x)$, par la fonction de répartition $F(x)$ d'une variable aléatoire X, discrète ou continue suivant le cas, vérifiant $F(x_i) = F_N(x_i)$, $i \in \{1, ... N\}$.

Nous dirons que $F(x)$ définit la **loi de référence** associée à une population hypothétique infinie, dite **population de référence**.

La population Ω est appelée la **population-mère**.

La connaissance de la loi de référence du caractère étudié est d'un grand intérêt pour la déduction statistique.
Elle constitue un modèle mathématique du phénomène étudié.
Cette distribution théorique peut dépendre d'un certain nombre de paramètres inconnus.

Les sondages permettent d'estimer deux types de paramètres :

- Les paramètres propres à la population-mère : moyenne, variance, etc.
- Les paramètres propres à la loi de référence : paramètre d'une loi de Poisson, paramètres d'une loi normale, etc.

6.3.2. Notion d'estimateur d'un paramètre de Ω.

6.3.2.1. Estimateur et estimation ponctuelle.

Soit X un caractère quantitatif de la population Ω.
Ce caractère prend les valeurs inconnues x_i, $i \in \{1, ... N\}$.
Un résumé de l'ensemble des valeurs $\{x_1, ... , x_N\}$ peut être défini par un ou plusieurs paramètres de Ω (moyenne, variance, proportion, etc.).

Soit y un tel paramètre de la population Ω.

Lorsque nous extrayons de la population un échantillon aléatoire simple E de taille n, nous pouvons calculer, avec les valeurs $\{x_1, ... , x_n\}$ prises par X dans l'échantillon, une estimation ponctuelle de y,

qui sera notée y^*.

Exemple.

Si y est la moyenne $\mu = \overline{X}$ de X, nous obtiendrons une estimation ponctuelle μ^* de la moyenne μ en prenant la moyenne arithmétique de l'échantillon :

$$\mu^* = \frac{1}{n} \sum_{i=1}^{i=n} x_i.$$

La valeur observée y^* n'est que l'une des valeurs possibles que l'on peut obtenir avec les divers échantillons possibles de taille n.
En réalité, avec une population de N individus, il y a un certain nombre, mettons k, d'échantillons possibles E_j de taille n, $j \in \{1, ..., k\}$ (k dépend de la méthode d'échantillonnage).
Chaque échantillon possible E_j de taille n possède une certaine probabilité p_j d'être tiré.
A chaque échantillon possible E_j de taille n est associée une estimation ponctuelle y_j^* de y.
A chaque estimation ponctuelle y_j^* de y est donc associée la probabilité p_j d'être observée.

Nous pouvons alors définir une variable aléatoire $\hat{y}$ prenant, pour chaque échantillon possible E_j de taille n, la valeur y_j^* avec la probabilité p_j.

Cette variable aléatoire $\hat{y}$ est appelée un **estimateur** du paramètre y.

Les valeurs de $\hat{y}$ sont les **estimations ponctuelles** de y.

La loi de probabilité de $\hat{y}$ s'appelle la **distribution d'échantillonnage** de $\hat{y}$.

On appelle **fluctuation d'échantillonnage**, la variation des estimations ponctuelles de y et **aléas d'échantillonnage** les causes de ces variations.

6.3.2.2. Caractéristiques d'un estimateur.

Il est logique de souhaiter que l'estimateur $\hat{y}$ prenne des valeurs aussi voisines que possible de la valeur inconnue y que nous voulons estimer.
Nous sommes conduits à définir un certain nombre de qualités que doit présenter un "bon" estimateur.

a) Estimateur sans biais.

Nous dirons que $\hat{y}$ est un estimateur sans biais du paramètre y, si, et seulement si, son espérance mathématique est y.

$$\boxed{\text{sans biais} \Leftrightarrow E(\hat{y}) = y}$$

Cette propriété traduit le fait qu'en moyenne, sur tous les échantillons possibles, nous retrouvons la valeur du paramètre que nous voulons estimer.

b) Estimateur robuste.

L'estimateur $\hat{y}$ d'un paramètre *y* possède une variance $\sigma_{\hat{y}}^2$ qui traduit la dispersion des valeurs de $\hat{y}$ autour de son espérance mathématique.
Cette variance dépend de la taille *n* de l'échantillon.
Nous dirons que $\hat{y}$ est un estimateur robuste, ou **convergent**, de *y* si la limite, lorsque *n* tend vers *N* de $\sigma_{\hat{y}}^2$ est nulle.

$$\text{robuste} \Leftrightarrow \lim_{n \to N} \sigma_{\hat{y}}^2 = 0$$

Cette propriété traduit le fait suivant : si nous connaissons la valeur prise par le caractère pour tous les individus de la population, la valeur de $\hat{y}$ est la valeur exacte *y* du paramètre.

Un **estimateur correct** est un estimateur sans biais et robuste.

c) Estimateur asymptotiquement gaussien.

Nous dirons qu'un estimateur $\hat{y}$ d'un paramètre *y* est **asymptotiquement gaussien** si, et seulement si, il vérifie la propriété suivante :

Lorsque *n* augmente indéfiniment, la fonction de répartition de $\frac{\hat{y} - E(\hat{y})}{\sigma_{\hat{y}}}$ tend uniformément vers la **fonction de répartition d'une variable normale centrée réduite**.

En pratique, dès que *n* est supérieur ou égal à 30, nous admettrons que la fonction de répartition de $\frac{\hat{y} - E(\hat{y})}{\sigma_{\hat{y}}}$ peut être remplacée par la fonction de répartition de la variable normale centrée réduite.

Lorsque *n* est suffisamment grand (en pratique $n \geq 30$), pour tout $\alpha \in [0, 1]$, le nombre réel positif u_α donné par :

$\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$, où Φ est la fonction de répartition de la variable normale centrée réduite,

vérifie :

$$P\left(\left| \frac{\hat{y} - E(\hat{y})}{\sigma_{\hat{y}}} \right| \leqslant u_\alpha \right) = 1 - \alpha.$$

En effet, comme la fonction de répartition de $\frac{\hat{y} - E(\hat{y})}{\sigma_{\hat{y}}}$ peut être remplacée par la fonction de répartition de la variable normale centrée réduite, dès que *n* est supérieur ou égal à 30, la symétrie de la loi normale donne :

$$P\left(\left|\frac{\hat{y}-E(\hat{y})}{\sigma_{\hat{y}}}\right| \leqslant u_\alpha\right) = \Phi(u_\alpha) - \Phi(-u_\alpha) = \Phi(u_\alpha) - (1-\Phi(u_\alpha)) = 2\,\Phi(u_\alpha) - 1 = 1-\alpha.$$

Les valeurs de la fonction de répartition Φ sont données par des tables.

Un **estimateur CAG** est un estimateur correct et asymptotiquement gaussien.

d) Amélioration d'un estimateur.

Etant donnés deux estimateurs $\hat{y}_1$ et $\hat{y}_2$ du même paramètre *y*, on dit que l'estimateur $\hat{y}_1$ est meilleur que l'estimateur $\hat{y}_2$ si l'espérance de $(\hat{y}_1 - y)^2$ est plus petite que l'espérance de $(\hat{y}_2 - y)^2$.

Ceci signifie simplement que l'on considère comme meilleur un estimateur dont les valeurs sont moins dispersées autour de la valeur de *y*.
Dans l'absolu, le meilleur estimateur d'un paramètre est celui dont pour lequel l'espérance de $(\hat{y} - y)^2$ est la plus petite possible.
Un estimateur sans biais dont la variance est minimale s'appelle un **estimateur précis**.
Pour un estimateur précis, l'espérance $E(\hat{y})$ est égale à *y* et la variance $\sigma_{\hat{y}}^2$ est minimale.

6.3.3. Notion d'intervalle de confiance.

6.3.3.1. Introduction.

Considérons un échantillon aléatoire simple *E*, de taille *n*, extrait de la population Ω (tirages au sort équiprobables, sans remise).
Dans cet échantillon, le caractère étudié prend les valeurs $\{x_1, \ldots, x_n\}$.
Nous pouvons considérer la valeur prise par le caractère étudié pour l'individu *i* de l'échantillon comme la valeur prise par une variable aléatoire *X*.
L'ensemble des valeurs $\{x_1, \ldots, x_n\}$ apparaît alors comme le résultat de *n* épreuves indépendantes sur la même variable aléatoire.
L'estimateur $\hat{y}$ d'un paramètre *y* apparaît alors comme une fonction de *n* variables aléatoires indépendantes X_i, $i \in \{1, \ldots, n\}$, de même loi de probabilité, qui est la loi de probabilité de *X*.
X s'appelle la **variable parente**.

La connaissance de la loi de probabilité de *X* permet de calculer la loi de probabilité de $\hat{y}$.

La variable aléatoire centrée réduite $\frac{\hat{y}-E(\hat{y})}{\sigma_{\hat{y}}}$ correspondant à $\hat{y}$, possède une espérance mathématique nulle et une variance égale à 1.

Exemple 1.

Nous étudions la taille des individus d'une population d'effectif *N*.
Pour cela nous extrayons un échantillon aléatoire simple et indépendant d'effectif *n*.
Soit μ la moyenne de la taille des individus de la population.
Soit *X* la variable aléatoire "taille d'un individu" : à chaque individu de l'échantillon est associé une

variable aléatoire indépendante "taille" X_i qui a la même loi de probabilité que la variable parente X.
L'estimateur

$$\hat{\mu} = \overline{X} = \frac{1}{n} \sum_{i=1}^{i=n} X_i$$

de la taille moyenne μ dans la population, a, pour valeur dans l'échantillon, la moyenne arithmétique des tailles des individus de l'échantillon.
Cet estimateur possède une loi de probabilité qui peut être calculée en fonction de la loi de probabilité de X.

Exemple 2.

Soit σ^2 la variance de la taille des individus de la population.
Soit X la variable aléatoire "taille d'un individu" : à chaque individu de l'échantillon est associé une variable aléatoire indépendante "taille" X_i qui a la même loi de probabilité que la variable parente X.
L'estimateur

$$\hat{\sigma}^2 = \frac{1}{n-1} \left(\sum_{i=1}^{i=n} X_i^2 - \frac{1}{n} \left(\sum_{i=1}^{i=n} X_i \right)^2 \right)$$

de la variance σ^2 de la taille dans la population, a, pour valeur dans l'échantillon, $\frac{n}{n-1} S^2(X)$ où $S^2(X)$ est la variance des tailles des individus de l'échantillon (variance d'échantillonnage).
Cet estimateur possède une loi de probabilité qui peut être calculée en fonction de la loi de probabilité de X.

6.3.3.2. Intervalle de confiance pour les grands échantillons.

Si $\hat{y}$ est un estimateur correct et asymptotiquement gaussien (estimateur CAG) d'un paramètre y, avec $E(\hat{y}) = y$, la relation

$$P\left(\left| \frac{\hat{y} - E(\hat{y})}{\sigma_{\hat{y}}} \right| \leqslant u_\alpha \right) = 1 - \alpha$$

s'écrit :

$$P(\hat{y} - u_\alpha \sigma_{\hat{y}} \leqslant \mu \leqslant \hat{y} + u_\alpha \sigma_{\hat{y}}) = 1 - \alpha.$$

L'événement $\hat{y} - u_\alpha \sigma_{\hat{y}} \leqslant \mu \leqslant \hat{y} + u_\alpha \sigma_{\hat{y}}$ a donc une probabilité $1 - \alpha$ de se réaliser lorsqu'on choisit au hasard un échantillon de taille $n \geq 30$.
Autrement dit, dans la population, la proportion des échantillons de taille $n \geq 30$ pour lesquels l'événement $\hat{y} - u_\alpha \sigma_{\hat{y}} \leqslant \mu \leqslant \hat{y} + u_\alpha \sigma_{\hat{y}}$ est réalisé est $1 - \alpha$.
Autrement dit encore, étant donné un échantillon de taille $n \geq 30$, choisi au hasard, la probabilité de réalisation de l'événement $\hat{y} - u_\alpha \sigma_{\hat{y}} \leqslant \mu \leqslant \hat{y} + u_\alpha \sigma_{\hat{y}}$ est $1 - \alpha$.

Or, pour un échantillon de taille n choisi au hasard, $\hat{y}$ prend la valeur y^* et $\sigma_{\hat{y}}$ une valeur $s_{\hat{y}}$, de sorte que $\hat{y} - u_\alpha \sigma_{\hat{y}}$ prend une valeur

$$y_1 = y^* - u_\alpha s_{\hat{y}}$$

et $\hat{y} + u_\alpha \sigma_{\hat{y}}$ prend la valeur

$$y_2 = y^* + u_\alpha s_{\hat{y}}$$

L'intervalle

$$\boxed{[y_1 ; y_2] = [\, y^* - u_\alpha s_{\hat{y}} ; y^* + u_\alpha s_{\hat{y}} \,]}$$

dans lequel la taille n de l'échantillon est supérieure ou égale à 30 et $\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$, s'appelle l'intervalle de confiance de y au risque α, ou **intervalle de confiance de y au niveau de confiance $1 - \alpha$**.

C'est un intervalle dans lequel la probabilité de trouver la vraie valeur de y est $1 - \alpha$.

Plus α est grand, plus l'amplitude de l'intervalle de confiance est petite, puisque Φ est une fonction croissante.
Dans la pratique, en l'absence de précision contraire, nous conviendrons de prendre $\alpha = 5$ %.

Plus n est grand, plus la valeur de $\sigma_{\hat{y}}^2$ a des chances d'être proche de 0, donc plus la valeur de $\hat{y}$ a des chances d'être proche de y.
Nous pourrons ainsi calculer la valeur de n qui permet d'avoir un intervalle de confiance d'amplitude donnée.

Les **valeurs à retenir** de la fonction de répartition de la variable aléatoire normale centrée réduite sont, pour $\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$:

— $\Phi(1{,}645) = 0{,}950$, soit $u_{0,10} = 1{,}645$.

— $\Phi(1{,}960) = 0{,}975$, soit $u_{0,05} = 1{,}960$.

— $\Phi(2{,}575) = 0{,}995$, soit $u_{0,01} = 2{,}575$.

Ces valeurs donnent les intervalles de confiance aux niveaux de confiance 90 %, 95 %, 99 %.
La valeur utilisée par défaut est $u_{0,05} = 1{,}960$.

6. 4. ETUDE DU SONDAGE ELEMENTAIRE.

Soit Ω une population d'effectif *N* dont on étudie un caractère *X*.

Si *X* est un caractère quantitatif, les paramètres qui caractérisent ce caractère sont :

— la moyenne $\overline{X} = \mu = \frac{1}{N} \sum_{i=1}^{i=N} x_i$

— la variance $\sigma^2 = \frac{1}{N} \left(\sum_{i=1}^{i=N} x_i^2 - \frac{1}{N} \left(\sum_{i=1}^{i=N} x_i \right)^2 \right)$.

Si *X* est un caractère qualitatif à deux modalités *A* et *B*, le paramètre qui caractérise *X* est la proportion *p* d'individus présentant la modalité *A*.
Les paramètres sont inconnus.
La théorie de l'échantillonnage a pour but de les estimer au mieux.

6.4.1. Echantillon non exhaustif, tirage à probabilités égales.

Un tirage au hasard avec remise induit que chaque individu a une probabilité $\frac{1}{N}$ d'être tiré.

6.4.1.1. Caractère quantitatif.

a) Loi de probabilité induite par le tirage de l'échantillon.

Le tirage avec remise, d'un individu de W, peut être représenté par une variable aléatoire parente, notée encore *X*, dont la loi de probabilité est définie par :

$$P(X = x_i) = \frac{1}{N},\ i \in [1, N].$$

L'espérance mathématique de *X* est $E(X) = \sum_{i=1}^{i=N} \frac{1}{N} x_i = \frac{1}{N} \sum_{i=1}^{i=N} x_i = \mu$.

La variance de *X* est $Var(X) = E((X - \mu)^2) = \sigma^2$.

b) Estimateur de la moyenne de la population.

Constituer un échantillon de taille *n* par des tirages non exhaustifs équiprobables dans Ω, revient à définir *n* variables aléatoires indépendantes $X_1, \ldots, X_n$, qui suivent toutes la même loi que *X*.
Soit $\{x_1, \ldots, x_n\}$ la réalisation de l'échantillon *E*.

La moyenne arithmétique $\overline{x} = \frac{1}{n} \sum_{i=1}^{i=n} x_i$ est la réalisation par échantillonnage de la variable aléatoire

$$\boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{i=n} X_i.}$$

L'espérance mathématique de l'estimateur $\hat{\mu}$ est $E(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^{i=n} E(X_i) = \frac{1}{n} \times n\, E(X) = \mu$.

La variance de l'estimateur $\hat{\mu}$ est $\sigma_{\hat{\mu}}^2 = Var(\hat{\mu}) = \frac{1}{n^2} \sum_{i=1}^{i=n} Var(X_i) = \frac{1}{n^2} \times n\, Var(X) = \frac{\sigma^2}{n}$.

Par conséquent, $\hat{\mu}$ est un estimateur **sans biais** de μ ($E(\hat{\mu}) = \mu$) mais il n'est **pas robuste** ($\lim_{n \to N} \sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{N} \neq 0$).

c) Estimateur de la variance de la population.

La variance expérimentale de l'échantillon est $s^2 = \frac{1}{n} \sum_{i=1}^{i=n} (x_i - \overline{x})^2$.

C'est la réalisation par échantillonnage de la variable aléaoire "**variance d'échantillonnage**" :

$$S^2 = \frac{1}{n} \left(\sum_{i=1}^{i=n} X_i^2 - \frac{1}{n} \left(\sum_{i=1}^{i=n} X_i \right)^2 \right) = \frac{1}{n} \sum_{i=1}^{i=n} (X_i - \hat{\mu})^2$$

L'espérance mathématique de S^2 est

$$E(S^2) = E\left(\frac{1}{n} \sum_{i=1}^{i=n} (X_i - \hat{\mu})^2 \right) = \frac{1}{n} \sum_{i=1}^{i=n} E\left((X_i - \hat{\mu})^2 \right)$$

$$E(S^2) = \frac{1}{n} \sum_{i=1}^{i=n} E\left((X_i - \mu + \mu - \hat{\mu})^2 \right)$$

$$E(S^2) = \frac{1}{n} \sum_{i=1}^{i=n} E(X_i - \mu)^2 + \frac{1}{n} \sum_{i=1}^{i=n} E(\mu - \hat{\mu})^2 + \frac{2}{n} \sum_{i=1}^{i=n} E\left((X_i - \mu)(\mu - \hat{\mu}) \right)$$

Mais on a :

$$\frac{1}{n} \sum_{i=1}^{i=n} E(X_i - \mu)^2 = \frac{1}{n} \sum_{i=1}^{i=n} E\left(X_i - E(X_i) \right)^2 = \frac{1}{n} n \, Var(X) = \sigma^2.$$

$$\frac{1}{n} \sum_{i=1}^{i=n} E(\mu - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{i=n} E\left((\hat{\mu} - E(\hat{\mu}))^2 \right) = Var(\hat{\mu}) = \frac{\sigma^2}{n}.$$

$$\frac{2}{n} \sum_{i=1}^{i=n} E\left((X_i - \mu)(\mu - \hat{\mu}) \right) = \frac{2}{n} E\left((\mu - \hat{\mu}) \sum_{i=1}^{i=n} (X_i - \mu) \right) = \frac{2}{n} E\left((\mu - \hat{\mu})(n\hat{\mu} - n\mu) \right) = -2 E((\hat{\mu} - \mu)^2) = -2 Var(\hat{\mu}) = -2 \frac{\sigma^2}{n}.$$

Au total :

$$E(S^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2.$$

La variance d'échantillonnage n'est pas un estimateur sans biais de la variance σ^2 de la population : c'est un **estimateur biaisé**.

La linéarité de l'espérance mathématique montre que :

$$E\left(\frac{n}{n-1} S^2 \right) = \frac{n}{n-1} E(S^2) = \sigma^2,$$

de sorte que l'estimateur :

$$\hat{\sigma}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{i=n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{i=n} X_i\right)^2\right) = \frac{n}{n-1} S^2$$

est un estimateur **sans biais** de la variance σ^2 de la population : $E(\hat{\sigma}^2) = \sigma^2$.

6.4.1.2. Caractère qualitatif.

Le paramètre étudié inconnu est la proportion *p* d'individus de la population présentant la modalité *A* du caractère qualitatif.
Pour chaque individu de la population, nous pouvons définir une variable aléatoire de Bernoulli, prenant la valeur 1, avec la probabilité *p*, si l'individu est porteur de la modalité *A*, 0 sinon, avec la probabilité $q = 1 - p$.

Choisir un échantillon de taille *n*, c'est choisir un *n*-uple de variables aléatoires $(X_1, \ldots, X_n)$ de Bernoulli, indépendantes, de même paramètre *p*.
Soit $(x_1, \ldots, x_n)$ une réalisation de l'échantillon *E*.

La moyenne expérimentale $p^* = \frac{1}{n}\sum_{i=1}^{i=n} x_i$ est la réalisation par échantillonnage de la variable aléatoire $\hat{p} = \frac{1}{n}\sum_{i=1}^{i=n} X_i$, qui représente la fréquence de la modalité *A* dans l'échantillon.

Son espérance mathématique est $E(\hat{p}) = \frac{1}{n}\sum_{i=1}^{i=n} E(X_i) = \frac{1}{n} \times n\,p = p$.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{i=n} X_i$$

est un **estimateur sans biais** de la proportion *p* des individus de la population présentant la modalité *A* du caractère étudié.

Sa variance est $Var(\hat{p}) = \frac{1}{n^2}\sum_{i=1}^{i=n} Var(X_i) = \frac{1}{n^2} \times n\,p\,(1-p) = \frac{p(1-p)}{n}$.
Lorsque *n* tend vers *N*, cette variance ne tend pas vers 0, mais vers $\frac{p(1-p)}{N}$: l'estimateur $\hat{p}$ de *p* n'est **pas un estimateur robuste**.

Pour les échantillons de grande taille ($n \geq 30$), on peut définir l'intervalle de confiance de *p* correspondant au risque α, par :

$$[p_1, p_2] = \left[\, p^* - u_\alpha \sqrt{\frac{p^*(1-p^*)}{n}} \;;\; p^* + u_\alpha \sqrt{\frac{p^*(1-p^*)}{n}} \,\right]$$

avec $\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$.

6.4.2. Echantillon exhaustif, tirage à probabilités égales.

Un tirage au hasard sans remise induit que chaque échantillon de taille n a une probabilité $\frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$ d'être tiré.

6.4.2.1. Caractère quantitatif.

a) Estimation de la moyenne.

Soit x_{ij} la réalisation du caractère X pour le j^e individu de l'échantillon $E_i = (X_{i1}, \dots, X_{in})$.
La réalisation du i^e échantillon est un n-uple $(x_{i1}, \dots, x_{in})$.
La moyenne d'échantillonnage $\overline{x}_i = \frac{1}{n}\sum_{j=1}^{j=n} x_{ij}$ est la réalisation d'une variable aléatoire $\overset{\triangle}{\overline{X}}$ que nous allons définir.

Nous pouvons définir $\binom{N}{n}$ échantillons différents E_i, $i \in \left[\ 1\ ;\ \binom{N}{n}\ \right]$, de taille n, chacun ayant une probabilité $p_i = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$ d'être tiré au hasard.

Considérons la variable aléatoire $\hat{p}$ dont la loi de probabilité, uniforme, est définie par :

$$P\left(\overset{\triangle}{\overline{X}} = \overline{x}_i\right) = p_i,\ i \in \left[\ 1\ ;\ \binom{N}{n}\ \right].$$

Son espérance mathématique est :

$$E\left(\overset{\triangle}{\overline{X}}\right) = \sum_{i=1}^{i=\binom{N}{n}} p_i\, \overline{x}_i = \frac{1}{\binom{N}{n}} \sum_{i=1}^{i=\binom{N}{n}} \left(\frac{1}{n} \sum_{k=1}^{k=n} x_{ik} \right) = \frac{1}{\binom{N}{n}} \frac{1}{n} \sum_{i=1}^{i=\binom{N}{n}} \left(\sum_{k=1}^{k=n} x_{ik} \right).$$

La somme $\sum_{i=1}^{i=\binom{N}{n}}$ est une somme étendue à tous les échantillons de taille n.

Pour un k pris entre 1 et n, notons que x_{ik} est la valeur x_j du caractère X pour le k^e individu de l'échantillon, qui est le j^e individu de la population.
Cette valeur apparaît une fois dans tous les échantillons de taille n contenant cet individu de la population, mais pas forcément à la même place, c'est-à-dire pas forcément avec le même indice k.
Or il y a $\binom{N-1}{n-1}$ échantillons de taille n contenant cet individu, de sorte que la valeur x_j de X pour le j^e individu de la population, apparaît $\binom{N-1}{n-1}$ fois dans la somme $\sum_{i=1}^{i=\binom{N}{n}} \left(\sum_{k=1}^{k=n} x_{ik} \right)$.

Ce raisonnement est valable, bien sûr, pour tous les indices j de 1 à N.
Lorsque nous faisons la somme pour tous les échantillons de taille n, nous obtenons :

$$\sum_{i=1}^{i=\binom{N}{n}} \left(\sum_{k=1}^{k=n} x_{ik} \right) = \sum_{j=1}^{j=N} \binom{N-1}{n-1} x_j = \binom{N-1}{n-1} (x_1 + ... + x_N)$$

$$E\left(\hat{\bar{X}}\right) = \frac{1}{\binom{N}{n}} \frac{1}{n} \binom{N-1}{n-1} (x_1 + ... + x_N) = \frac{1}{\binom{N}{n}} \frac{1}{n} \binom{N-1}{n-1} N\,\mu = \frac{n![N-n)!}{N!} \frac{N}{n} \frac{[N-1)!}{[n-1)![N-n)!}\,\mu = \mu$$

Moralité : la moyenne d'échantillonnage $\hat{\bar{X}} = \frac{1}{n} \sum_{j=1}^{j=n} X_{ij}$ est un estimateur **sans biais** de la moyenne μ du caractère *X*.

b) Variance de la moyenne d'échantillonnage.

La variance de $\hat{\bar{X}}$ est donnée par $Var\left(\hat{\bar{X}}\right) = E\left(\hat{\bar{X}}^2\right) - \left(E\left(\hat{\bar{X}}\right)\right)^2 = E\left(\hat{\bar{X}}^2\right) - \mu^2$.

Calculons le terme :

$$E\left(\hat{\bar{X}}^2\right) = \sum_{i=1}^{i=\binom{N}{n}} p_i \, \bar{x}_i^{\,2}$$

$$E\left(\hat{\bar{X}}^2\right) = \frac{1}{\binom{N}{n}} \sum_{i=1}^{i=\binom{N}{n}} \bar{x}_i^{\,2} = \frac{1}{\binom{N}{n}} \sum_{i=1}^{i=\binom{N}{n}} \frac{1}{n^2} \left(\sum_{k=1}^{k=n} (x_{ik})^2 \right) = \frac{1}{\binom{N}{n}} \frac{1}{n^2} \sum_{i=1}^{i=\binom{N}{n}} \left(\sum_{k=1}^{k=n} (x_{ik})^2 \right)$$

$$= \frac{1}{\binom{N}{n}} \frac{1}{n^2} \left(\sum_{i=1}^{i=\binom{N}{n}} \left(x_{i1}^{\,2} + ... + x_{in}^{\,2} \right) + \sum_{i=1}^{i=\binom{N}{n}} \left(\sum_{\substack{j,k=1 \\ j \neq k}}^{n} x_{ij}\, x_{ik} \right) \right)$$

Pour tout individu de numéro *j* de Ω, il y a $\binom{N-1}{n-1}$ échantillons de taille *n* contenant cet individu, de sorte que $x_j^{\,2}$ apparaît $\binom{N-1}{n-1}$ fois dans la somme $\sum_{i=1}^{i=\binom{N}{n}} \left(x_{i1}^{\,2} + ... + x_{in}^{\,2} \right)$.

Et ceci est vrai pour les *N* individus de la population.
De sorte que l'on obtient :

$$\sum_{i=1}^{i=\binom{N}{n}} \left(x_{i1}^{\,2} + ... + x_{in}^{\,2} \right) = \binom{N-1}{n-1} \left(x_1^{\,2} + ... + x_N^{\,2} \right) = \binom{N-1}{n-1} N \left(\sigma^2 + \mu^2 \right) = \frac{N!}{[n-1)![N-n)!} (\sigma^2 + \mu^2)$$

Reste à calculer la somme $\sum_{i=1}^{i=\binom{N}{n}} \left(\sum_{\substack{j,k=1 \\ j \neq k}}^{n} x_{ij}\, x_{ik} \right)$

Dans chacun des $\binom{N}{n}$ échantillons de taille *n*, on forme $\frac{n[n-1)}{2}$ produits de la forme $x_{ij}\, x_{ik}$, avec $j \neq k$.
Dans l'ensemble des échantillons de taille *n*, on forme donc $\binom{N}{n} \frac{n[n-1)}{2}$ produits de deux valeurs de *X*

différentes.
Comme il existe $\frac{N(N-1)}{2}$ produits de deux valeurs de *X* différentes, chacun intervient $\binom{N}{n}\frac{n(n-1)}{N(N-1)}$ fois dans la somme étendue à l'ensemble des échantillons de taille *n*.
On obtient donc :

$$\sum_{i=1}^{i=\binom{N}{n}}\left(\sum_{\substack{j,k=1\\ j\neq k}}^{n} x_{ij}\,x_{ik}\right)=\binom{N}{n}\frac{n(n-1)}{N(N-1)}\sum_{\substack{j,k=1\\ j\neq k}}^{N} x_j\,x_k$$

Or on peut écrire aussi :

$$\sum_{\substack{j,k=1\\ j\neq k}}^{N} x_j\,x_k=\sum_{j=1}^{j=N} x_j\left(\sum_{k=1}^{k=N} x_k-x_j\right)=\left(\sum_{j=1}^{j=N} x_j\right)\left(\sum_{k=1}^{k=N} x_k\right)-\sum_{j=1}^{j=N} x_j^2$$

$$=\left(\sum_{j=1}^{j=N} x_j\right)^2-\sum_{j=1}^{j=N} x_j^2=(N\,\mu)^2-N\,(\sigma^2+\mu^2)=N\,((N-1)\,\mu^2-\sigma^2)$$

On obtient alors :

$$\sum_{i=1}^{i=\binom{N}{n}}\left(\sum_{\substack{j,k=1\\ j\neq k}}^{n} x_{ij}\,x_{ik}\right)=\binom{N}{n}\frac{n(n-1)}{N(N-1)}N\,((N-1)\,\mu^2-\sigma^2)=\frac{N!}{n!(N-n)!}\,n\,\frac{n-1}{N-1}\,((N-1)\,\mu^2-\sigma^2)=N\binom{N-2}{n-2}$$

$$((N-1)\,\mu^2-\sigma^2)$$

$$E\left(\overset{\triangle}{\bar{X}}^2\right)=\frac{1}{\binom{N}{n}}\frac{1}{n^2}\left(N\binom{N-1}{n-1}(\sigma^2+\mu^2)+N\binom{N-2}{n-2}((N-1)\,\mu^2-\sigma^2)\right)$$

$$E\left(\overset{\triangle}{\bar{X}}^2\right)=\frac{1}{\binom{N}{n}}\frac{N}{n^2}\left(\binom{N-1}{n-1}-\binom{N-2}{n-2}\right)\sigma^2+\frac{1}{\binom{N}{n}}\frac{N}{n^2}\left(\binom{N-1}{n-1}+(N-1)\binom{N-2}{n-2}\right)\mu^2$$

$$\frac{1}{\binom{N}{n}}\frac{N}{n^2}\left(\binom{N-1}{n-1}-\binom{N-2}{n-2}\right)=\frac{n!(N-n)!}{N!}\frac{N}{n^2}\left(\frac{(N-1)!}{(n-1)!(N-n)!}-\frac{(N-2)!}{(n-2)!(N-n)!}\right)$$

$$=\frac{(n-1)!(N-n)!}{(N-1)!}\frac{1}{n}\frac{(N-2)!}{(n-1)!(N-n)!}\left((N-1)-(n-1)\right)=\frac{1}{n}\frac{N-n}{N-1}$$

$$\frac{1}{\binom{N}{n}}\frac{N}{n^2}\left(\binom{N-1}{n-1}+(N-1)\binom{N-2}{n-2}\right)=\frac{n!(N-n)!}{N!}\frac{N}{n^2}\left(\frac{(N-1)!}{(n-1)!(N-n)!}+(N-1)\frac{(N-2)!}{(n-2)!(N-n)!}\right)$$

$$=\frac{(n-1)!(N-n)!}{(N-1)!}\frac{1}{n}\frac{(N-1)!}{(n-1)!(N-n)!}(1+(n-1))=1$$

$$E\left(\overset{\triangle}{\bar{X}}^2\right)=\frac{1}{n}\frac{N-n}{N-1}\,\sigma^2+\mu^2$$

$$Var\left(\overset{\triangle}{\bar{X}}\right)=E\left(\overset{\triangle}{\bar{X}}^2\right)-\mu^2=\frac{1}{n}\frac{N-n}{N-1}\,\sigma^2$$

$$\boxed{Var\left(\overset{\triangle}{\overline{X}}\right) = \frac{1}{n}\frac{N-n}{N-1}\,\sigma^2}$$

Moralité : lorsque *n* tend vers *N*, la variance de $\overset{\triangle}{\overline{X}}$ tend vers 0, l'estimateur $\overset{\triangle}{\overline{X}}$ de μ est **robuste**.

La moyenne d'échantillonnage $\overset{\triangle}{\overline{X}} = \frac{1}{n}\sum_{j=1}^{j=n} X_{ij}$ est un estimateur sans biais et robuste, donc **correct**, de μ.

On remarquera aussi que la présence du rapport d'exhaustivité $\frac{N-n}{N-1}$, inférieur à 1, fait que la variance de $\overset{\triangle}{\overline{X}}$ est plus faible lorsque l'échantillon est exhaustif que lorsqu'il est non exhaustif : les valeurs de $\overset{\triangle}{\overline{X}}$ sont moins dispersées autour de la moyenne μ lorsque l'échantillon est exhaustif.

c) Estimation de la variance.

La variance expérimentale de l'échantillon $s^2 = \frac{1}{n}\sum_{j=1}^{j=n}(x_{ij} - \overline{x}_i)^2$ est une réalisation de la variable aléatoire :

$$S^2 = \frac{1}{n}\sum_{j=1}^{j=n}(X_{ij} - \overset{\triangle}{\overline{X}})^2 = \frac{1}{n}\left(\sum_{j=1}^{j=n} X_{ij}^{\,2} - \frac{1}{n}\left(\sum_{j=1}^{j=n} X_{ij}\right)^2\right)$$

L'espérance mathématique de cette variable aléatoire est ;

$$E(S^2) = \frac{1}{n}\sum_{j=1}^{j=n} E\left((X_{ij} - \overset{\triangle}{\overline{X}})^2\right) = \frac{1}{n}\sum_{j=1}^{j=n} E\left((X_{ij} - \mu + \mu - \overset{\triangle}{\overline{X}})^2\right)$$
$$= \frac{1}{n}\sum_{j=1}^{j=n} E\left((X_{ij} - \mu)^2\right) + \frac{1}{n}\sum_{j=1}^{j=n} E\left((\mu - \overset{\triangle}{\overline{X}})^2\right) - \frac{2}{n}\sum_{j=1}^{j=n} E\left((X_{ij} - \mu)(\overset{\triangle}{\overline{X}} - \mu)\right)$$

Mais :

$E\left((X_{ij} - \mu)^2\right) = E\left((X_{ij} - E(X_{ij})^2\right) = Var(X_{ij}) = \sigma^2$.
$\frac{1}{n}\sum_{j=1}^{j=n} E\left((X_{ij} - \mu)^2\right) = \frac{1}{n} n\,\sigma^2 = \sigma^2$.

$$\frac{1}{n}\sum_{j=1}^{j=n} E\left((\mu - \overset{\triangle}{\overline{X}})^2\right) = \frac{1}{n}\sum_{j=1}^{j=n} Var\left(\overset{\triangle}{\overline{X}}\right) = \frac{1}{n} n\, Var\left(\overset{\triangle}{\overline{X}}\right) = Var\left(\overset{\triangle}{\overline{X}}\right) = \frac{1}{n}\frac{N-n}{N-1}\,\sigma^2$$

$$\sum_{j=1}^{j=n} E\left((X_{ij} - \mu)(\overset{\triangle}{\overline{X}} - \mu)\right) = E\left((\overset{\triangle}{\overline{X}} - \mu)\sum_{j=1}^{j=n}(X_{ij} - \mu)\right) = E\left((\overset{\triangle}{\overline{X}} - \mu)\, n\,(\overset{\triangle}{\overline{X}} - \mu)\right) = n\,E\left((\overset{\triangle}{\overline{X}} - \mu)^2\right) = n\, Var\left(\overset{\triangle}{\overline{X}}\right)$$

Il reste alors :

$$E(S^2) = \sigma^2 + \frac{1}{n}\frac{N-n}{N-1}\,\sigma^2 - \frac{2}{n} n\, Var\left(\overset{\triangle}{\overline{X}}\right) = \sigma^2 - \frac{1}{n}\frac{N-n}{N-1}\,\sigma^2 = \frac{n[N-1)-[N-n)}{n[N-1)}\,\sigma^2 = \frac{N[n-1)}{n[N-1)}\sigma^2$$

On voit donc que S^2 est un estimateur biaisé de σ^2, mais que, par linéarité de l'espérance mathématique :

$$\hat{\sigma}^2 = \frac{1-\frac{1}{N}}{1-\frac{1}{n}} S^2 = \frac{N-1}{N}\frac{1}{n-1}\left(\sum_{j=1}^{j=n} X_{ij}^2 - \frac{1}{n}\left(\sum_{j=1}^{j=n} X_{ij} \right)^2 \right)$$

est un **estimateur sans biais de la variance** σ^2.

6.4.2.2. Caractère qualitatif.

La fréquence d'échantillonnage $p^* = \frac{1}{n}\sum_{i=1}^{i=n} x_i$ de la modalité A du caractère qualitatif étudié est la valeur prise après échantillonnage par la variable aléatoire

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{i=n} X_i.$$

Mais nous avons vu, précédemment, que l'espérance mathématique et la variance de X_i, étaient données par :

$$E(X_i) = p$$
$$Var(X_i) = p(1-p).$$

L'étude précédente montre que nous pouvons écrire :

$$E(\hat{p}) = p$$
$$Var(\hat{p}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{i=n} X_i \right) = \frac{1}{n^2} Var\left(n\hat{\bar{X}} \right) = Var(\hat{\bar{X}}) = \frac{1}{n}\frac{N-n}{N-1} p(1-p).$$

Ainsi, $\hat{p}$ est un estimateur sans biais et robuste de p.

Sa réalisation $p^* = \frac{1}{n}\sum_{i=1}^{i=n} x_i$ dans un échantillon est une estimation ponctuelle sans biais de p.

Pour les grands échantillons, au niveau de confiance $1-\alpha$, la réalisation de l'intervalle de confiance de p sera donné par $[\, p_1 ; p_2 \,]$, avec

$$p_1 = p^* - u_\alpha \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p^*(1-p^*)}{n}}$$
$$p_2 = p^* + u_\alpha \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p^*(1-p^*)}{n}}$$

où u_α est défini par la relation $\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$, Φ étant la fonction de répartition de la variable normale centrée réduite.

6.4.3. Echantillon non exhaustif, tirage à probabilités inégales.

Soit $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ la population.
Nous étudions dans cette population un caractère quantitatif X de valeur x_j pour l'individu ω_j.
Notons p_j la probabilité de tirage de l'individu ω_j lors de la constitution de l'échantillon $\left(\sum_{j=1}^{j=N} p_j = 1 \right)$.

Tout tirage avec remise peut être schématisé par une variable aléatoire $\hat{\hat{X}}$ dont la loi de probabilité est définie par :

$$P(\hat{\hat{X}} = x_j) = p_j, \forall\, j \in [1 ; N].$$

Notons :

— $\mu = \frac{1}{N} \sum_{j=1}^{j=N} x_j$, la moyenne du caractère X dans la population.
— $\sigma^2 = \frac{1}{N}\left(\sum_{j=1}^{j=N} x_j^2 \right) - \frac{1}{N}\left(\sum_{j=1}^{j=N} x_j \right)^2$, la variance de X dans la population.

Ces paramètres sont inconnus, nous cherchons à les estimer.
Nous supposons connues la taille N de la population et les probabilités p_j associées aux valeurs x_j.

Notons, pour simplifier, $(x_1, \ldots, x_n)$ la réalisation d'un échantillon.

6.4.3.1. Estimation de la moyenne.

Considérons la variable aléatoire $\hat{\hat{X}}'$ définie par la loi de probabilité :

$$P\left(\hat{\hat{X}}' = \frac{x_j}{p_j} \right) = p_j, \forall\, j \in [1 ; N].$$

et soit :

$$\boxed{\hat{\hat{m}}' = \frac{1}{Nn} \sum_{i=1}^{i=n} \hat{\hat{X}}'_i}$$

la variable aléatoire de réalisation $m'^* = \frac{1}{Nn} \sum_{i=1}^{i=n} \frac{x_i}{p_i}$ dans l'échantillon.

Nous avons :

$$E(\hat{\hat{m}}') = \frac{1}{Nn} \sum_{i=1}^{i=n} E(\hat{\hat{X}}'_i) = \frac{1}{Nn} \sum_{i=1}^{i=n} \left(\sum_{j=1}^{j=N} p_j \frac{x_j}{p_j} \right) = \frac{1}{Nn} \sum_{i=1}^{i=n} N\mu = \frac{1}{n} \sum_{i=1}^{i=n} \mu = \frac{1}{n} \times n\,\mu = \mu$$

La relation $E(\hat{\hat{m}}') = \mu$ montre que la variable aléatoire $\hat{\hat{m}}'$ est un **estimateur sans biais de μ**.
Sa réalisation $m'^* = \frac{1}{Nn} \sum_{i=1}^{i=n} \frac{x_i}{p_i}$ dans l'échantillon est une estimation ponctuelle sans biais de μ.

6.4.3.2. Variance de l'estimateur de la moyenne.

Nous avons :

$$E(\hat{\hat{X}}') = \sum_{j=1}^{j=N} p_j \frac{x_j}{p_j} = N\,\mu$$

$$E(\hat{\hat{X}}'^{\,2}) = \sum_{j=1}^{j=N} p_j \frac{x_j^{\,2}}{p_j^{\,2}} = \sum_{j=1}^{j=N} \frac{x_j^{\,2}}{p_j}$$

$$Var(\hat{\hat{X}}') = \sum_{j=1}^{j=N} \frac{x_j^{\,2}}{p_j} - N^2\,\mu^2$$

Comme le tirage de l'échantillon est fait avec remise, les variables $\hat{\hat{X}}_i'$ sont indépendantes, et, par conséquent :

$$Var(\hat{\bar{m}}') = \left(\frac{1}{Nn}\right)^2 Var\left(\sum_{i=1}^{i=n} \hat{\hat{X}}_i'\right) = \frac{1}{N^2\,n^2} \sum_{i=1}^{i=n} Var(\hat{\hat{X}}_i')$$

$$= \frac{n}{N^2\,n^2} Var(\hat{\hat{X}}') = \frac{1}{N^2\,n} Var(\hat{\hat{X}}') = \frac{1}{N^2\,n}\left(\sum_{j=1}^{j=N} \frac{x_j^{\,2}}{p_j} - N^2\,\mu^2\right)$$

$$\boxed{Var(\hat{\bar{m}}') = \frac{1}{N^2\,n} \sum_{j=1}^{j=N} \frac{x_j^{\,2}}{p_j} - \frac{\mu^2}{n}}$$

Cette variance s'exprime à l'aide de l'ensemble des valeurs x_j, inconnues, prises par le caractère X dans la population Ω.
Il serait intéressant d'en avoir une estimation à partir de la réalisation $\{x_1, \ldots, x_n\}$ d'un échantillon.

6.4.3.3. Estimation de la variance de l'estimateur de la moyenne.

Soit $\hat{\hat{X}}'$ la variable aléatoire définie, comme dans IV.4.2.1. par la loi de probabilité :

$$P\left(\hat{\hat{X}}' = \frac{x_j}{p_j}\right) = p_j,\ \forall\, j \in [1\,;\,N].$$

Nous avons vu que l'espérance mathématique de cette variable aléatoire était égale à $N\,\mu$, qu'on peut estimer par $N\,\hat{\bar{m}}'$.
Considérons la variance d'échantillonnage de la variable aléatoire $\hat{\hat{X}}'$, c'est la variable aléatoire :

$$\hat{S}_1^{\,2} = \frac{1}{n} \sum_{i=1}^{i=n} (\hat{\hat{X}}_i' - N\,\hat{\bar{m}}')^2$$

L'espérance mathématique de $\hat{S}_1^{\,2}$ est :

$$E(\hat{S}_1^{\,2}) = E\left(\frac{1}{n} \sum_{i=1}^{i=n} (\hat{\hat{X}}_i' - N\,\hat{\bar{m}}')^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{i=n} E\left((\hat{X}_i' - N\,\hat{m}')^2 \right)$$
$$= \frac{1}{n}\sum_{i=1}^{i=n} E\left((\hat{X}_i' - N\,\mu + N\,\mu - N\,\hat{m}')^2 \right)$$
$$= \frac{1}{n}\sum_{i=1}^{i=n} E\left((\hat{X}_i' - N\,\mu)^2 \right) + \frac{1}{n}\sum_{i=1}^{i=n} E\left((N\,\mu - N\,\hat{m}')^2 \right) + \frac{2}{n}\sum_{i=1}^{i=n} E\left((\hat{X}_i' - N\,\mu)(N\,\mu - N\,\hat{m}') \right)$$
$$= \frac{1}{n}\sum_{i=1}^{i=n} Var\,(\hat{X}_i') + \frac{1}{n}\sum_{i=1}^{i=n} Var\,(N\,\hat{m}') + \frac{2}{n} E\left((N\,\mu - N\,\hat{m}') \sum_{i=1}^{i=n} (\hat{X}_i' - N\,\mu) \right)$$
$$= \frac{1}{n} \times n\,Var\,(\hat{X}') + \frac{1}{n} \times n\,N^2\,Var\,(\hat{m}') + \frac{2}{n} E\left((N\,\mu - N\,\hat{m}')(N\,n\,\hat{m}' - N\,n\,\mu) \right)$$
$$= Var\,(\hat{X}') + N^2\,Var\,(\hat{m}') - \frac{2}{n} \times n\,N^2\,Var\,(\hat{m}')$$
$$= Var\,(\hat{X}') - N^2\,Var\,(\hat{m}')$$
$$= n\,N^2\,Var\,(\hat{m}') - N^2\,Var\,(\hat{m}')$$
$$= (n-1)\,N^2\,Var\,(\hat{m}')$$

La relation $E\,(\hat{s}_1^{\,2}) = (n-1)\,N^2\,Var\,(\hat{m}')$, qui s'écrit aussi :

$$E\left(\frac{\hat{s}_1^{\,2}}{(n-1)\,N^2} \right) = Var\,(\hat{m}')$$

montre que

> La variable aléatoire $\frac{\hat{s}_1^{\,2}}{(n-1)\,N^2}$ est un estimateur sans biais de la variance $Var\,(\hat{m}')$

et sa réalisation dans l'échantillon :

$$\frac{1}{n(n-1)\,N^2}\sum_{i=1}^{i=n}\left(\frac{x_i}{p_i} - N\,m'^* \right)^2 = \frac{1}{n(n-1)\,N^2}\left(\sum_{i=1}^{i=n}\left(\frac{x_i}{p_i} \right)^2 - \frac{1}{n}\left(\sum_{i=1}^{i=n} \frac{x_i}{p_i} \right)^2 \right)$$

compte tenu de la relation $N\,m'^* = m'^* = \frac{1}{n}\sum_{i=1}^{i=n} \frac{x_i}{p_i}$, est une **estimation ponctuelle sans biais de la variance de** $\hat{m}'$.

$$\sigma_{\hat{m}'}^{\,2}{}^* = \frac{1}{n(n-1)\,N^2}\left(\sum_{i=1}^{i=n}\left(\frac{x_i}{p_i} \right)^2 - \frac{1}{n}\left(\sum_{i=1}^{i=n} \frac{x_i}{p_i} \right)^2 \right)$$

Cette estimation de la variance de $\hat{m}'$ permet de construire, pour les grands échantillons, un **intervalle de confiance de la moyenne** μ :

$$m'^* \pm u_\alpha\,\sigma_{\hat{m}'}^*.$$

EXTRAITS D'UNE TABLE DE NOMBRES AU HASARD

(Kendall et Babington Smith, table tirée de Christian Labrousse, Statistique, Tome2, Dunod, Paris, 1962)

02 22 85 19 48 74 55 24 89 69 15 53 00 20 88 48 95 08
85 76 34 51 40 44 62 93 65 99 72 64 09 34 01 13 09 74
00 88 96 79 38 24 77 00 70 91 47 43 43 82 71 67 49 90
64 29 81 85 50 47 36 50 91 19 09 15 98 75 60 58 33 15
94 03 80 04 21 49 54 91 77 85 00 45 68 23 12 94 23 44
42 28 52 73 06 41 37 47 47 31 52 99 89 82 22 81 86 55
09 27 52 72 49 11 30 93 33 29 54 17 54 48 47 42 04 79
54 68 64 07 85 32 05 96 54 79 57 43 96 97 30 72 12 19
25 04 92 29 71 11 64 10 42 23 23 67 01 19 20 58 35 93
28 58 32 91 95 28 42 36 98 59 66 32 15 51 46 63 57 10
64 35 04 62 24 87 44 85 45 68 41 66 19 17 13 09 63 37
61 05 55 88 25 01 15 77 12 90 69 34 36 93 52 39 36 23
98 93 18 93 86 98 99 04 75 28 30 05 12 09 57 35 90 15
61 89 35 47 16 32 20 16 78 52 82 37 26 33 67 42 11 93
94 40 82 18 06 61 54 67 03 66 76 82 90 31 71 90 39 27
54 38 58 65 27 70 93 57 59 00 63 56 18 79 85 52 21 03
63 70 89 23 76 46 97 70 00 62 15 35 97 42 47 54 60 60
61 58 65 62 81 29 69 71 95 53 53 69 20 95 66 60 50 70
51 68 98 15 05 64 43 32 74 07 44 63 52 38 67 59 56 69
59 25 41 48 64 79 62 26 87 86 94 30 43 54 26 98 61 38
85 00 02 24 67 85 88 10 34 01 54 53 23 77 33 11 19 68
01 46 87 56 19 19 19 43 70 25 24 29 48 22 44 81 35 40
42 41 25 10 87 27 77 28 05 90 73 03 95 46 88 82 25 02
03 57 14 03 17 80 47 85 94 49 89 55 10 37 19 50 20 37
18 95 93 40 45 43 04 56 17 03 34 54 83 91 69 02 90 72

Table de la fonction de répartition Φ de la variable normale centrée réduite

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6143
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9270	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de u.

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
$\Phi(u)$	0,998 65	0,999 04	0,999 31	0,999 52	0,999 66	0,999 76	0,999 841	0,999 928	0,999 968	0,999 997

La table donne les valeurs de $\Phi(u)$ pour u positif. Lorsque u est négatif,
il faut prendre le complément à 1 de la valeur lue dans la table : $\Phi(-u) = 1 - \Phi(u)$