Cours et Exercices
de
Statistique descriptive
Niveau Licence 1

Table des matières

Introduction

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.

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).

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.

Chapitre 1

Concepts de base

1. Eléments de vocabulaire

Pour faire une étude statistique, il faut d'abord préciser l'ensemble des **individus** ou **unités statistiques** sur lequel portera l'étude. Cet ensemble s'appelle la **population**.
Exemple : Animeaux, des champs agricoles, être humains etc.

En statistique descriptive c'est l'ensemble des individus effectivement étudiés, sans chercher à étendre les constatations faites à une population plus vaste, ce qui relève de la statistique inférentielle.

Ensuite identifier les critères selon les quels l'étude doit être faite. Ces critères sont appelés des **caractères** que l'on désigne généralemenr sous le nom de **variables statistiques**.

Exemple : Sexe, âge, revenu, catégorie socio-professionnelle, nombre d'enfants, rendement d'une region agricole etc.

Il faut aussi identifier les valeurs possible des caractères, une valeur possible du caractère est applée une **modalité**. Il faut faire la difference entre les modalités observées c'est à dire prise par au moins un individu de la population et celles non observées.

Les modalités d'un caractère doivent être incompatibles et exhaustives ; tout individu doit présenter une et une seule modalité.

Une série statistique est la suite des valeurs prises par une ou plusieurs variables pour chacun des individus d'une population donnée ou d'un échantillon de la population.

2. Echantillonnage statistique

Pour receuillir des informations sur une population statistique, l'on dispose de deux méthodes :

- La méthode exhaustive ou recensement où chaque individu de la population est étudié selon le ou les caractères étudiés.
- La méthode des sondages ou éccchantillonnage qui conduit à n'examiner qu'une fraction de la population, un échantillon.

L'écchantillonnage représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans la population donnée.
Pour que les résultats observés soient généralisables à la population statistique, l'échantillon doit être représentatif de cette dernière, i.e., il doit refléter fidèlement sa composition et sa complexité. Seul l'échantillonnage aléatoire asure la représentativité de l'échantillon.

Un échantillon est qualifié d'aléatoire lorsque chaque individu de la population a une probabilité connu et non nulle d'appartenir à l'échantillon.

Exemple 1. *Un questionnaire est distribué à 150 personnes dans la cour d'un établissement secondaire. Il comporte diverses questions. La population = l'ensemble des elèves de cet établissement. L'échantillon = les étudiants ayant répondu au questionnaire. Un individu est une personne interrogée. Les variables correspondent aux questions posées : l'âge, la taille, la couleur des yeux, etc.*

3. Les types de variables statistiques

Le type d'une variables dependent de la nature de ses modalités. On distingue plusieurs types de variables :

3.1. Variables qualitatives

Une variable est dite qualitative lorsque les réponses possibles à la question posée, ou les modalités, ne correspondent pas à une quantité mesurable par un nombre mais appartiennent à un groupe de catégories.

Exemple 2. *le sexe, la couleur des yeux, la mention au baccalauréat, la fréquence d'une activité (jamais, rarement, parfois, souvent, très souvent).*

on distingue :
- les variables **qualitatives nominales** : il n'y a pas d'hiérachie entre les differentes modalités ; exemple : sexe, couleur des yeux.
- les varibles **qualitatives ordinales** : les differentes modalités peuvent être ordonnées de manière naturelle ; exemple : la mention au baccalauréat, la frequence d'une activité, niveau d'études scolaires : école primaire $<$ 1er cycle $<$ CAP $<$ BEP $<$ Bac $<$ BTS $<$ DEUG $<$ .....

Remarque 1.1. *Certaines variables nominales peuvent être désignées par un code numerique, qui n'a pas de valeur quantité. Exemple : le code postal, le sexe (*$1 =$ *garçon,* $2 =$ *fille)*

3.2. Variables quantitatives

Les réponses correspondent à des quantités mesurables et sont données sous forme de nombre. On distingue :

- Les variables quantitatives discrètes : elles prennent leurs valeurs dans un ensemble discret, le plus souvent fini. Exemple : le nombre d'enfants, la pointure du pied.
- les variables quantitatives continues : elles peuvent prendre toutes les valeurs d'un intervalle réel. Exemple : la taille des individus, une note à un examen.

Remarque 1.2. *L'âge peut être vu et traité comme une variable quantitative discrète ou continue suivant la précision que l'on choisit et le nombre de valeurs qu'il prend au sein de la population. Il peut également exister des variables basées sur l'âge qui sont qualitatives. Si dans un sondage on pose la question "quelle est votre tranche d'âge parmi les possibilités suivantes : - de* 25 *ans, entre* 25 *et* 45*, entre* 40 *et* 60 *et* +60 *ans", on peut voir la variable "tranche d'âge" comme une variable qualitative ordinale*

4. Séries statistiques

Une série statistique correspond aux modalités des caractères étudiés sur une population donnée ou un échantion d'indivudus de la population.

Supposons qu'on a observé n individus selon p caractères ou variables statistiques ϵ_k $(k = 1, 2, ..., p)$. A chaque individu, on associe une modalité et une seule pour chaque caractère donné. Une série statistique est la donnée d'une suite $x_1, x_2, \cdots, x_n$ de valeurs prises par les variables ϵ_k $(k = 1, 2, ..., p)$ pour chacun des n individus observés, où $x_i = (x_{i1},\ x_{i2},\ ...,\ x_{ip})$, avec x_{ik} la valeur observée du caractère ϵ_k pour l'individu u_i.
Ces données peuvent se présenter sous forme d'un tableaux **individus** $\times$ **variables** comme suit :

		Variables				
		1	$\cdots$	j	$\cdots$	p
Individus	1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	x_{1i}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n	x_{n1}	$\cdots$	x_{nj}	$\cdots$	x_{np}

Si $p = 1$ (un seul caractère), on parle de séries statistiques univariées et bivariées si $p = 2$ (deux caractères).

Chapitre 2

Séries statistiques univarieés

1. Tableaux statistiques et représentations graphiques

1.1. Les variables qualitatives

Exemple 3. *On s'intéresse à la variable "couleur des yeux" sur un groupe de* 20 *personnes. On code chaque modalité de la manière suivante : M=marron, V=vert, N=noir, B=bleu. On obtient la série statistique suivante :*
M,V,M,M,M,M,M,N,M,N,M,M,B,M,M,M,B,M,M,M.

Tableaux de distribution de fréquence absolues, relatives et cumulées

Exemple 4. *Pour l'exemple précédent, on remplit le tableau suivant :*

Couleur des yeux	*M*	*V*	*N*	*B*	*Total*
Effectif					
Proportion					

Tableau-type : On choisit une notation pour la variable, par exemple : X. n désigne le nombre d'individus dans l'échantillon. on note $C_1, \ldots, C_k$ les k modalités de la variable. Pour $1 \leq j \leq k$, on note

- n_j l'effectif associé à la modalité C_j (le nombre d'individus pour lesquels la valeur prise par la variable est C_j),
- $f_j = n_j/n$ la fréquence relative ou proportion associée à cette modalité,
- et si la variable est qualitative **ordinale** : $N_j = n_1 + n_2 + \cdots + n_j$ resp. $F_j = f_1 + f_2 + \cdots + f_j$ la frequence absolue (effectif) cumulée croissante resp. la frequence relative cumulée croissante pour cette modalité
  (avec la convention : $F_0 = 0$). Elle n'a de sens que si la variable est qualitative ordinale et si les modalités $C_1, C_2, \ldots, C_k$ sont ordonnées suivant l'odre croissant naturel (ou hiérachique

ascendant) qui règne parmi ces modalités. Exemple : niveau d'études scolaires : école primaire $<$ 1er cycle $<$ CAP $<$ BEP $<$ Bac $<$ BTS $<$ DEUG $<$ ....

Le tableau suivant est un tableau-type qui permet de résumer les données.

Variable X	C_1	C_2	...	C_k	Totales
Fréquence absolue ou effectif	n_1	n_2	...	n_k	n
Fréquence relative ou proportion	$f_1 = n_1/n$	$f_2 = n_2/n$	...	$f_k = n_k/n$	1
Fréquence relative cumulée*	$F_1 = f_1$	$F_2 = f_1 + f_2$	...	$F_k = 1$	pas de sens

*Attention : uniquement dans le cas de variables qualitatives ordinales.

Représentations graphiques

Pour une variable ou 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.

- Diagramme en bâtons : en abscisse sont disposées les différentes modalités, de façon arbitraire aux quelles on associe des segments espacés entre eux dont les longeurs (en ordonnée) sont proportionnelles à l'effectif ou à la fréquence relative de chaque modalité. Preciser le nom des axes, le nom du graphique et la source des informations

Nous appelons polygone statistique, ou diagramme polygonal, la ligne obtenue en joignant les sommets des bâtons.

Exemple :

Caractère : catégorie socio-professionnelle.

Ouvriers = O, Cadre moyen = CM, Cadre supérieur = CS.

caractère	O	CM	CS
Effectifs	20	10	5

- Diagramme en tuyaux d'orgue : en abscisse sont disposées les différentes modalités, de façon arbitraire aux quelles on associe des rectangles espacés entre eux, de largeur constente, dont la hauteur (en ordonnée) sont proportionnelle à l'effectif ou à la fréquence relative de chaque modalité. Preciser le nom des axes, le nom du graphique et la source des informations. Dans le cas d'une variable qualitative ordinale, on peut également construire le diagramme en tuyau d'orgue des effectifs ou des proportions cumulés.

Exemple 5.

- Diagrammes en secteurs : chaque modalité est représentée par un secteur de disque dont l'angle est proportionnel à l'effectif ou à la fréquence de la modalité (ou pourcentage).

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

Dans un diagramme circulaire (cercle complet), l'effetif total ou la fréquence relative 1 (ou le pourcentage 100%) correspond à l'angle 360°.

Pour représenter les données sur un diagramme semi-circulaire (demi-cercle), il suffira de calculer les mesures des secteurs angulaires par rapport à 180°.

Exemple 6.

1.2. Les variables quantitatives discrètes

Exemple 7. *On s'intéresse à la variable "pointure" (que l'on notera P) sur un groupe de* 20 *personnes. On obtient la série statistique suivante :*
39, 43, 38, 39, 39, 42, 44, 44, 48, 40, 44, 43, 41, 37, 39, 38, 45, 41, 44, 44.

Tableaux de distribution de fréquences

Exemple 8. *Pour la variable P, on remplie le tableau suivant :*

P	*37*	*38*	*39*	*40*	*41*	*42*	*43*	*44*	*45*	*46*	*47*	*48*
Effectif												
Proportion												
Proportion cumulée												

On note $v_1, v_2, \ldots, v_k$ les k valeurs différentes que peut prendre la variable avec $v_i < v_j$ si $i < j$ (on n'en rencontrera pas pas d'exemple dans ce cour, mais une variable discrète peut prendre une infinité de valeurs). Pour $1 \leq j \leq n$, on note n_j l'effectif des individus pour lesquels la variable prend la valeur v_j. On note f_j la fréquence relative ou proportion pour la valeur v_j et $F_j = f_1 + \cdots + f_j$ la j-ème fréquence relative cumulée (avec la convention : $F_0 = 0$). On résume habituelement les données comme dans le tableau-type suivant :

Valeurs prises par la variable	v_1	v_2	$\ldots$	v_k	Total
Fréquence absolue	n_1	n_2	$\ldots$	n_k	n
Fréquence relative	$f_1 = n_1/n$	$f_2 = n_2/n$	$\ldots$	$f_k = n_k/n$	1
Fréquence relative cumulée $\nearrow$	$F_1 = f_1$	$F_2 = f_1 + f_2$	$\ldots$	$F_k = 1$	pas de sens

On définit de même pour la valeur v_j la fréquence cumulée décroissante :

$$F_j^* = \frac{1}{n}(n_j + ... + n_k) = f_j + ... + f_k.$$

La quantité $N_j^* = n_j + ... + n_k$ est appelée effectif cumulé décroissant.

Représentation graphique

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.

- Diagramme différentiel : diagramme en bâtons

On trace un graphique avec

- sur l'axe des abscisses les différentes valeurs prises par la variable, placées en **respectant une échelle**,
- en ordonné les fréquences relatives ou les fréquences absolues.
- Pour chaque valeur v_j on construit un bâton vertical à l'abscisse v_j, de hauteur proportionnel a la fréquence de la valeur v_j.

Exemple : pointure.

Nous appelons polygone statistique, ou diagramme polygonal, la ligne obtenue en joignant les sommets des bâtons.

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

Fonction de répartition empirique

La fonction de répartition empirique permet de décrire la série statistique de manière complète. Elle est définie sur $\mathbb{R}$ et prend ses valeurs dans $[0, 1]$. Pour x dans $\mathbb{R}$, elle est définie par

$$F(x) = \begin{cases} 0 & \text{si } x < v_1 \\ F_j & \text{si } \quad v_j \leq x < v_{j+1} \\ 1 & \text{si } \quad v_k \leq x \end{cases}$$

Exemple 9. *Pointure*

1.3. Les variables quantitatives continues

Exemple 10. *On s'intéresse à la taille, notée T et exprimée en mètre, de* 20 *individus. On a obtenu la série statistique suivante :*
1,72 ; 1,87 ; 1,66 ; 1,73 ; 1,64 ; 1,77 ; 1,80 ; 1,81 ; 1,60 ; 1,78 ; 1,83 ; 1,75 ; 1,70 ; 1,58 ; 1, 68 ; 1,66 ; 1,93 ; 1,75 ; 1,80 ; 1,85.

Tableaux de distribution de fréquences-fréquences cumulées

Les données brutes de la variable pour chaque individu sont notées $x_1, \ldots, x_n$. Elle peuvent prendre n'importe quelle valeur dans un interval de $\mathbb{R}$ et il est très rare d'avoir deux fois la même valeur pour deux individus différents. Il serait donc unitile de tracer un diagramme en bâton comme dans le cas d'une variable discrète : il consisterait en un amoncellemment illisible de bâton de hauteur $1/n$. On choisir donc de faire un **Regroupement en classe**.

- L'intervalle où la variable prend ses valeurs est divisé en k classes : $[b_0, b_1[, \quad [b_1, b_2[, \ldots, \quad [b_{k-1}, b_k[$ (il est possible d'avoir des bornes infinies).

- Pour $1 \leq j \leq n$, on note n_j l'effectif associé à la classe $[b_{j-1}, b_j[$, $f_j = n_j/n$ la fréquence relative associé à cette classe et $F_j = f_1 + \cdots + f_j$ la j-ème fréquence cumulée (avec la convention $F_0 = 0$)
- On note $a_j = b_j - b_{j-1}$ l'amplitude de la classe $[b_{j-1}, b_j[$.
- On note $d_j = f_j/a_j$ la densité de proportion pour la classe $[b_{j-1}, b_{j]}$.

Exemple 11. *de la taille*

T	$[1,50; 1,65[$	$[1,65; 1,70[$	$[1,70; 1,75[$	$[1,75; 1,80[$	$[1,80; 1,85[$	$[1,85; 2,00[$
Effectif						
Proportion						
Proportion cumulée						
Amplitude						
Densité de proportion						

Remarque 2.1.
- *la densité de la proportion permet de comparer les effectifs dans chaque classe en tenant compte de la taille de ces classes (cf. la notion de densité de la population en géographie).*
- *Dans le cas de classes qui ont toutes les même longueur, il n'est pas nécessaire de calculer la densité de proportion, il est suffisant d'étudier les fréquences relatives ou absolues (qui sont directement proportionnelle a la densité de proportion).*

Tableau-type

Variable X	$[b_0, b_1[$	$[b_1, b_2[$	$\ldots$	$[b_{k-1}, b_k[$	Total
Fréq. relative	$f_1 = n_1/n$	$f_2 = n_2/n$	$\ldots$	$f_k = n_k/n$	1
Fréq. relative cumulée	$F_1 = f_1$	$F_2 = f_1 + f_2$	$\ldots$	$F_k = 1$	
Amplitude	$a_1 = b_1 - b_0$	$a_2 = b_2 - b_1$	$\ldots$	$a_k = b_{k-1} - b_k$	
Densité de proportion	$d_1 = f_1/a_1$	$d_2 = f_2/a_2$	$\ldots$	$d_k = f_k/a_k$	

Remarque 2.2. *Contrairement au cas d'une variable qualitative ou discrète, ce tableau représente une perte d'information par rapport aux données brutes*

Représentation graphique

- Diagramme différentiel : histogramme des densités.

Sur l'axe des abscisses sont placées les bornes des classes représentant les modalités en respectant une échelle. Pour chaque classe, on élève un rectangle de hauteur (ordonnée) proportionnelle à la densité de proportion ou d'effectif.

Exemple de taille T :

Remarque 2.3. *On représente la* ***densité de proportion ou d'effectif*** *et non pas les fréquences relatives ou absolues.*

Consequence 1. *L'aire d'un rectangle est proportionnelle à la fréquence (absolues ou relatives) de la classe correspondante. En effet, pour le rectangle conrespondant à la classe* $[b_j, b_{j-1}[$ *l'aire est*

$$(b_j - b_{j-1}) \times d_j = f_j.$$

Dans la pratique, on utilise la règle de construction suivante :

Vérifier si les amplitudes des différentes classes sont identiques.

- Si les amplitudes sont identiques, on représente sur l'axe des abscisse les classes par des segments de même longueur. On associe à chaque classe un rectangle dont la hauteur est proportionnelle à l'effectif ou à la fréquence.

- Si les amplitudes sont non identiques, on choisit une unité d'amplitude U et on construit l'histogramme de telle sorte que la hauteur du rectangle de la classe $[b_{j-1}, b_j[$ soit proportionnelle à l'effectif par unité d'amplitude $\frac{n_i}{a_i}U$ associé.

- Diagramme intégral : courbe cumulative des effectifs ou des fréquences.

La courbe cumulative des fréquences doit représenter la fonction de répartition de la variable statistique.

Fonction de répartition empirique

Pour x une valeur dans l'intervalle $[b_{j-1}, b_j[$, on approche la proportion d'individus pour lesquels la variable est inférieure ou égale à x par l'aire de l'histogramme entre les abscisses b_{j-1} et x notée $F(x)$:

$$F(x) = f_1 + f_2 + \cdots + f_{j-1} + (x - b_{j-1}) \times d_j = F_{j-1} + (x - b_{j-1}) \times d_j$$

On a ainsi définie une fonction F qui vaut 0 sur $]-\infty, b_0[$ et 1 sur $[b_k, +\infty[$. Elle vaut F_j en b_j et $F_{j-1} + (x - b_{j-1}) \times d_j$ sur $[b_{j-1}, b_j[$. cette fonction, affine par morceaux, est appelée **fonction de répartition empirique** de la variable X.

Exemple 12. *Fonction de répartition empirique de la variable* T.

2. Résumés numériques d'une variable statistique

2.1. Paramètre de tendance centrale

Le mode

Le mode rend compte de l'endroit où les données sont le plus concentrées.

Le mode, noté Mo, est la modalite la plus frequente ou dominante dans la population i.e. celle qui admet la plus grande frequence : $f(Mo) = \max_{i \in [1,k]}(f_i)$.

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

Pour une variable quantitative continue regroupée en classe, 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 :

$$Mo = x_m + a \times \frac{\Delta i}{\Delta i + \Delta s}$$

avec
x_m : limite inferieure de la classe d'effectif (par unité d'amplitude) maximal
a : l'amplitude de la classe modale
Δi : Ecart d'effectif (par unité d'amplitude) entre la classe modale et la classe inferieure la plus proche
Δs : Ecart d'effectif (par unité d'amplitude) entre la classe modale et la classe superieure la plus proche

Exemple 13. *Pointure, taille.*

Remarque :

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.

La moyenne

On note $\{x_1, x_2, \ldots, x_n\}$ la série statistique. La moyenne est définie par :

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Exemple 14. *pointure, taille*

Cas d'une variale discrète : si $v_1, v_2, \ldots, v_k$ sont les k valeurs prises par la variable X, n_j l'effectif et f_j la fréquence relative correspondant à la valeur v_j, on peut réécrire :

$$\bar{x} = \frac{n_1 v_1 + n_2 v_2 + \cdots + n_k v_k}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i v_i = \sum_{i=1}^{k} f_i v_i$$

Exemple 15. *Pointure.*

Cas d'une variable continue regroupée en classes : la variable X est regroupée das les classes $[b_{j-1}, b_j[$ $(1 \leq j \leq n)$, les fréquences relatives associées à ces classe sont notées f_j, $1 \leq j \leq n$. Lorsque les données brutes ne sont plus accessibles et qu'on ne dispose que des données regroupées en classe, on calcule une **moyenne approchée** grâce à des représentant des classes(leur centre) : $c_j = (b_j + b_{j-1})/2$, par la formule :

$$\bar{x}_{app} = f_1 c_1 + f_2 c_2 + \cdots + f_k c_k = \sum_{i=1}^{k} f_j c_j$$

Exemple : calcul d'une moyenne approchée de la variable "taille" à partir du groupement en classes.
Propriétés de la moyenne : si on fait le changement de variable $Y = aX + b$ (traduction sur la série statistiques : $y_i = ax_i + b$,$1 \leq i \leq n$), alors

$$\bar{y} = a\bar{x} + b$$

Exemple 16. *calcul de la taille moyenne en mètres.*

La médiane

La médiane Me correspond au centre des valeurs observées classées par ordre croissant

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)},$$

ou à la valeur pour laquelle 50% des valeurs observées sont supérieures et 50% sont inférieures.

a) Cas d'une variable discrète :

– si n est impair, la médiane est la $\frac{n+1}{2}$-ième valeur observée :

$$Me = x_{(\frac{n+1}{2})}.$$

– si n est pair, une médiane est une valeur quelconque entre la $\frac{n}{2}$-ième valeur observée : $x_{(\frac{n}{2})}$ et la $\frac{n}{2} + 1$-ième valeur observée : $x_{(\frac{n}{2}+1)}$. On parle donc d'intervalle médian. On peut prendre comme médiane $x_{(\frac{n}{2})}$ ou $x_{(\frac{n}{2}+1)}$. Mais il peut être commode de prendre le milieu :

$$Me = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

Lorsqu'on dispose d'une distribution observé, la médiane est définie grâce aux distributions cumulées à savoir :

- la distribution cumulée croissante $N(x)$ représentant le nombre d'observation inférieures ou égales à x.

- la distribution cumulée décroissante $N^*(x)$ correspondant au nombre d'observation supérieures ou égales à x.

ou grâce aux proportions cumulées ou à la fonction de répartition empirique (graphiquement).

La médiane est la valeur Me qui vérifie l'équation

$$N(Me) = N^*(Me).$$

La solution de cette équation est soit unique, soit indéterminée (intervalle médian). Dans ce dernier cas, on prend pour médiane la moyenne des valeurs qui définissent cet intervalle.

Dans la pratique, on procède comme suit :

Soit $v_1 < v_2 < \cdots < v_k$ les k valeurs différentes de la variable.

1. s'il existe une valeur v_j telle que $N_{j-1} < \frac{n}{2} < N_j$, alors $Me = v_j$ (en posant $N_0 = 0$, si $j = 1$) ;
2. s'il existe une valeur v_j telle que $N_j = \frac{n}{2}$, alors

$$Me = \frac{v_j + x_{j+1}}{2}.$$

Exemple 17. *pointure*

b) Cas d'une variable continue.

La médiane est définie comme la solution de l'équation :

$$F(Me) = 0,5$$

où F est la fonction de répartion empirique de la variable. On sait que cette solution existe parce que F est continue, et $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$. Si de plus F est strictement croissante, la solution Me est unique. La méthode pratique est la suivante :

1. S'il existe une borne de classe b_j telle que la proportion cumulée sur la classe $[b_{j-1}, b_j[$ est exatement $0,5$, alors la **médiane** est ce b_j.
2. Sinon, alors il existe une classe $[b_{j-1}, b_j[$ telle que

$$F(b_{j-1}) < 0,5 < F(b_j) \quad \text{ou} \quad N(b_{j-1}) < \frac{n}{2} < N(b_j).$$

Cette classe est la première sur laquelle la frequence cumulée dépasse $0,5$. Pour $x \in [b_{j-1}, b_j[$, $F(x) = F_{j-1} + (x - b_{j-1}) \times d_j$. Mais en particulier :

$$F(Me) = F_{j-1} + (Me - b_{j-1}) \times d_j = 0,5$$

d'où

$$Me = b_{j-1} + \frac{0,5 - F_{j-1}}{d_j}$$

Ou encore, en terme de b_j et de F ou N :

$$Me = b_{j-1} + (b_j - b_{j-1}) \times \frac{0,5 - F(b_{j-1})}{F(b_j) - F(b_{j-1})}$$

ou

$$Me = b_{j-1} + (b_j - b_{j-1}) \times \frac{\frac{n}{2} - N(b_{j-1})}{N(b_j) - N(b_{j-1})}$$

Cette méthode peut se traduire graphiquement ent utilisant le graphe de la fonction de répartition empirique et le théorème de Thalès.

Exemple 18. *médiane de la variable "taille", regroupée en classes.*

Méthode graphique avec la fonction de répartition empirique

Quantiles

a) cas d'une variable continue

Soit X une variable quantitative continue, de fonction de répartition empirique F. On suppose qu'on dispose de la répartition en classe des observations.
Le Quantile d'ordre p de X est la solution notée q_p de :

$$F(q_p) = p.$$

Cela signifie qu'une proportion d'environ p des observations est inferieur à q_p et qu'une proportion d'environ $1-p$ des données est supérieure à q_p.

Quantiles particuliers

- Quartiles : quantiles correspondant aux proportions multiples de $0,25$ (un quart). On note Q_1 le premier quartile, qui correspond à $q_{0,25}$, Q_3 le troisième quartile, qui correspond à $q_{0,75}$. La médiane est le deuxième quartile $Q_2 = q_{0,5}$.
- Déciles : quantiles correspondant aux proportions multiples de $0,1$: $q_{0,1}$ (premier décile), $q_{0,2}$ (deuxième décile), etc.
- Percentiles ou centiles : quantiles correspondant aux proportions multiples de $0,01$. Par exemple, le 65ème percentile est le quantile $q_{0,65}$

Calcul du quantile q_p : même méthode que pour le calcul de la médiane.

1. S'il existe une borne de class b_j telle que la proportion cumulée sur la classe $[b_{j-1}, b_j[$ est exatement p, autrement dit : $F(b_j) = p$, alors $q_p = b_j$.
2. Sinon, alors il existe une classe $[b_{j-1}, b_j[$ telle que

$$F(b_{j-1}) < p < F(b_j) \quad \text{ou} \quad N(b_{j-1}) < np < N(b_j).$$

Cette classe est la première sur laquelle la fréquence cumulée dépasse p. Pour $x \in [b_{j-1}, b_j[$, $F(x) = F_{j-1} + (x - b_{j-1}) \times d_j$. Mais en particulier :

$$F(q_p) = F_{j-1} + (q_p - b_{j-1}) \times d_j = p$$

D'où

$$q_p = b_{j-1} + \frac{p - F_{j-1}}{d_j}$$

Ou encore, en terme des b_j et de F ou N :

$$q_p = b_{j-1} + (b_j - b_{j-1}) \times \frac{p - F(b_{j-1})}{F(b_j) - F(b_{j-1})}$$

ou

$$q_p = b_{j-1} + (b_j - b_{j-1}) \times \frac{np - N(b_{j-1})}{N(b_j) - N(b_{j-1})}$$

Exemple 19. *troisième quartile de la variable "taille"*

b) cas d'une variable discrète

On procède de la même façon comme pour la médiane : à partir de la série ordonnée, le quartile q_p est défini par :

- si np est un nombre entier, alors

$$q_p = \frac{x_{(np)} + x_{(np+1)}}{2}$$

- si np n'est pas un nombre entier, alors

$$q_p = x_{(\lceil np \rceil)},$$

où $\lceil np \rceil$ représente le plus petit entier supérieur ou égale à np.

Lorsqu'on dispose d'une distribution observé, le quartile q_p est défini comme suit :

Soit $v_1 < v_2 < \cdots < v_k$ les k valeurs différentes de la variable.

1. s'il existe une valeur v_j telle que $N_{j-1} < np < N_j$, alors $q_p = v_j$ (en posant $N_0 = 0$, si $j = 1$) ;

2. s'il existe une valeur v_j telle que $N_j = np$, alors

$$q_p = \frac{v_j + v_{j+1}}{2}.$$

Exemple 20. *troisième quartile de la variable "pointure".*

Utilisation des paramètres de tendance centrale

Robustesse

La médiane est plus **robuste** que la moyenne : une ou plusieurs données erronnées ne font pratiquement, voire pas du tout, changer la médiane, alors qu'elles peuvent affecter considérablement la moyenne.

Assymétrie

La comparaison de la médiane et de la moyenne permet de détecter des assymétries de données :

Si la distribution des valeurs est symétrique, la valeur de la médiane est proche de la valeur de la moyenne arithmétique. $Me \simeq \bar{x}$.

De façon générale on a :

- $Mo = Me = \bar{x} \Longrightarrow$ distribution symétrique,
- $Mo < Me < \bar{x} \Longrightarrow$ distribution dissymétrique à gauche,
- $Mo > Me > \bar{x} \Longrightarrow$ distribution dissymétrique à droite

2.2. Paramètres de dispersion

Il est possible que deux variables statistiques aient la même valeur centrale mais complètement différentes du point de vue de la concentration ou dispersion des valeurs observées autour de cette valeur centrale. Il est donc nécessaire de trouver des mesures permetant d'apprécier la dispersion d'une série statistique ou d'une distribution observée.

L'étendue

Soit x_{min} la plus petite observation et x_{max} la plus grande. On définie **l'étendue** $e = x_{max} - x_{min}$. Elle a la même unité que l'unité de la variable. Elle n'est pas très informative car elle ne tient pas du tout compte de la répartition des données à l'intérieur de l'intervalle $[x_{min}, x_{max}]$.

Exemple 21. *étendu de la variable "taille"*

L'intervalle inter-quartile

On appelle **intervalle inter-quartile** l'intervalle $[Q_1, Q_3]$, qui contient environ 50% des observations. La **distance inter-quartile** $Q_3 - Q_1$ est une mésure de dispersion.

Exemple 22. *intervalle inter-quartile de la variable "taille".*

La variance et l'écart-type

La **variance** est définie par :

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

L'expression suivante est la plus pratique pour le calcul de la variance :

$$Var(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - (\bar{x})^2$$

Preuve : en développant le carré dans la définition de la variance.
Pour une variable **quantitative discrète** en prenant la valeur v_j un nombre n_j de fois ou (ou avec la fréquence f_j), pour $1 \leq j \leq k$:

$$\begin{aligned} Var(X) &= \frac{1}{n}\sum_{j=1}^{k} n_j(v_j - \bar{x})^2 = \sum_{j=1}^{k} f_j(v_j - \bar{x})^2 \\ &= \left(\frac{1}{n}\sum_{j=1=}^{k} n_j v_j^2\right) - (\bar{x})^2 = \left(\sum_{j=1}^{k} f_j v_j^2\right) - (\bar{x})^2 \end{aligned}$$

Dans le cas le cas d'une variable continue pour laquelle on dispose seulement des **données regroupées en classes**, on peut faire un calcule approché similaire à celui de la moyenne approchée $\bar{x}_{app}$.

On calcule une valeur approchée de la variance, notée $Var_{app}(X)$. Toutes les expressions qui suivent sont équivalentes.

$$Var_{app}(X) = \frac{1}{n}\sum_{j=1}^{k} n_j(c_j - \bar{x}_{app})^2 = \sum_{j=1}^{k} f_j(c_j - \bar{x}_{app})^2$$
$$= \left(\frac{1}{n}\sum_{j=1}^{k} n_j c_j^2\right) - (\bar{x}_{app})^2 = \left(\sum_{j=1}^{k} f_j c_j^2\right) - (\bar{x}_{app})^2$$

où c_j est le centre de la j-ème classe, dotée de l'effectif n_j (ou de la fréquence relative f_j).

Propriétés de la variance

- La variance est toujour positive ou nullle. Elle est nulle si et seulement si toutes les observation sonts identiques :
$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \Leftrightarrow \forall i, x_i - \bar{x} = 0$$
- L'untité de la variance est l'untié de X au carré.

L'ecart-type σ_X est défini par :

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

Propriété : l'unité de σ_X est l'unité de X.
Plus σ_X est grand plus les modalités sont dispersés.

Exemple 23. *variance et ecart-type de la variable "pointure", de la variable "taille".*

Le coefficient de variation

La comparaison des dispersions de deux séries statistiques peut se faire grâce aux écart-types lorsque ses séries ont des moyennes du même ordre de grandeur et ne contient pas de valeurs aberrantes. Dans le cas contraire, on peut utiliser le coefficient de variation défini par

$$CV = \frac{\sigma_X}{\bar{x}}.$$

Ce paramètre est une mesure relative de dispersion et permet une interprétation plus appropriée. On l'exprime en général en pourcentage.

Rémarque importante : le coefficient de variation est défini pour les variables quantitatives dont les valeurs sont positives non toutes nulles.

2.3. Boîtes à moustaches

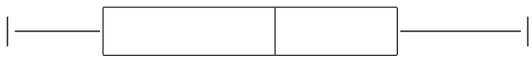
La boîte à moustaches est une représentation graphique qui permet de visualiser les quartiles ainsi que la dispersion des données et de repérer les données extrêmes ou *outliers*. Elle se fait couramment pour les variables quantitatives continues ou pour les variables quantitatives discrètes prenant un grand nombre de valeurs différentes. En revanche, elle n'a pas beaucoup d'intérêt pour une variable discrète prenant peu de valeurs différentes.
Elle est constituée :

- d'une **boîte** dont les bornes sont les premier et troisième quartile Q_1 et Q_3. A l'intérieur de la boîte figure la médiane Q_2.
- de **moustaches**. On définit tout d'abord deux bornes : $m_- = Q_1 - 1,5(Q_3 - Q_1)$ et $m_+ = Q_3 + 1,5(Q_3 - Q_1)$. On note m_{inf} la plus petite observation supérieure à m_-, et m_{sup} la plus grande observation inférieure à m_+. Soit :

$$m_{inf} = \min\{x_i : x_i \geq m_-\}$$

$$m_{sup} = \max\{x_i : x_i \leq m_+\}$$

  La moustache inférieure est le segment $[m_{inf}, Q_1]$. La moustache supérieure, de la même manière, est le segment $[Q_3, m_{sup}]$
- des **données extrêmes** éventuelles : les observations qui sont en dehors de la boîte et des moustaches, c'est à dire : supérieures à m_+ ou inférieures à m_-. On place ces données une à une quand on en dispose.

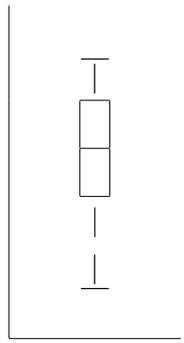


Remarque :

− Une boite et des moustaches courtes indiquent que la série est assez concentrée autour de sa médiane.
Au contraire une boite et des moustaches longues indiquent que la série est assez dispersée.

L'examen de la boite à moustaches permet d'avoir une idée de la symétrie de la distribution selon que la boîte et les moustaches sont symétriques ou, au contraire, de plus petite amplitude à gauche (asymétrie à gauche) ou à droite (asymétrie à droite).

− La représentation peut aussi se faire vertcalement, d'où l'appelation de "boite à pattes".



Exemple 24. *Boîte à moustache de la variable "taille" à partir de la série statistique de* 20 *observations.*

Dans le cas où on ne dispose pas des données brutes mais seulement des données regroupées en classes, on utilise les extrémités b_0 et b_k de la première et de la k-ème classe.

- la limite inférieure m_{inf} de la moustache inférieure est $\max\{m_-, b_0\}$ et la limite supérieure m_{sup} de la moustache supérieure est $\min\{m_+, b_k\}$.
- On ne peut pas placer les données extrêmes, sauf si elles sont fournies en plus.

Exemple 25. *Boîte à moustaches de la variable "taille" à partir des données regroupées.*

Chapitre 3

Séries statistiques bivariées

1. Généralités

On observe sur population P de n individus deux caractères X et Y. On obtient ainsi une série statistique composée de n couples d'observations du couple de variables (X, Y) : $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, appélée série statistiques bivariée, où (x_i, y_i) est le couple de valeurs observées du couple de variables (X, Y) pour l'individu i.
On suppose que X a I modalités notées $C_1, \ldots, C_I$ et Y a J modalitées notées $D_1, \ldots, D_J$.
Pour $1 \leq i \leq I$ et $1 \leq j \leq J$, on note n_{ij} le nombre d'individus qui vérifient à la fois les modalités C_i de X et D_j de Y.

1.1. Tableau de contingence

Dans le tableau de contingence, on regroupe les effectifs n_{ij}. On peut compléter le tableau de contingence en ajoutant les totaux en lignes et en colones.
On note $n_{i.} = n_{i1} + \cdots + n_{IJ} = \sum_{j=1}^{J} n_{ij}$ le total sur la ligne i de la table de contingence, $n_{.j} = n_{1j} + \cdots + n_{IJ} = \sum_{i=1}^{I} n_{ij}$ le total sur la colonne j de la table de contigence.

Y / X	D_1	D_2	$\ldots$	D_J	Total
C_1	n_{11}	n_{12}	$\ldots$	n_{1J}	$n_{1.}$
C_2	n_{21}	n_{22}	$\ldots$	n_{2J}	$n_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
C_I	n_{I1}	n_{I2}	$\ldots$	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.J}$	n

Exemple 26. *L'INSEE fournit les données suivantes relatives à la situation professionnelle des personnes habitant en France en* 2006*, immigrées ou non immigrées.*

situation quant à l'immigration / *Situation professionnelle*	*Immigrés*	*Non immigrés*	*Ensemble*
Actif ayant un emploi	2223906	23895180	26119096
Chômeur	559201	2845339	3404540
Retraité ou préretraités	963333	11901857	12865190
Elèves, étudiants, stagiaire	321533	4999097	5320630
Femme ou homme au foyer	486427	1926779	2413206
Autres inactifs	583016	12480429	13063445
Ensemble	5137416	58048681	63186098

Remarque 3.1. *La définition d'un immigré selon le Haut conseil à l'immigration, utilisée pour cette étude, est une personne née étrangère à l'étranger et résidant en France.*

1.2. Distribution marginale

La distribution marginale de la variable X est la donnée des **effectifs marginaux** $n_{1.}, \ldots, n_{I.}$. C'est la distribution de la variable X. On peut la présenter dans un tableau et calculer les fréquences $(f_{i.} = n_{i.}/n)$, qui sont les proportions associée à chaque modalité de la variable X. On peut calculer de même la distribution marginale de la variable Y.

Distribution marginale de X :

X	C_1	$\ldots$	C_I	Total
Effectif	$n_{1.}$	$\ldots$	$n_{I.}$	n
Proportion	$f_{1.} = n_{1.}/n$	$\ldots$	$f_{I.} = n_{I.}/n$	1

Distribution marginale de Y :

Y	D_1	$\ldots$	D_I	Total
Effectif	$n_{.1}$	$\ldots$	$n_{.J}$	n
Proportion	$f_{.1} = n_{.1}/n$	$\ldots$	$f_{.J} = n_{.J}/n$	1

Exemple 27. *Situation professionnelle de la population en France en* 2006

1.3. Distribution conditionlle

a) Profils-lignes

La distribution conditionelle de Y sachant la modalité de C_i de X est la distribution dont les proportions sont données dans le tableaux suivant :

$Y_{\|X=C_i}$	D_1	$\ldots$	D_I	Total
Proportion	$f_{1\|i} = n_{i1}/n_{i.}$	$\ldots$	$f_{J\|i} = n_{iJ}/n_{i.}$	1

où $f_{j|i} = \frac{n_{ij}}{n_{i.}}$

Une telle distribution est appelée profil-ligne. L'ensemble des profils-lignes peut être présenté dans un tableau :

Y_X / X	D_1	D_2	...	D_J	Total
C_1	$n_{11}/n_{1.}$	$n_{12}/n_{1.}$	...	$n_{1J}/n_{1.}$	1
C_1	$n_{21}/n_{2.}$	$n_{22}/n_{2.}$	...	$n_{2J}/n_{2.}$	1
...	...	...	...	...	
C_I	$n_{I1}/n_{I.}$	$n_{I2}/n_{I.}$	...	$n_{IJ}/n_{I.}$	1

Exemple 28. *Distribution conditionnelle de la variable " Situation quant à l'immigration" sachant la modalité " Actifs ayant un emploi" en France en 2006, ou : situation quant à l'immigration des actifs ayant un emploi en France en 2006.*

b) Profils-colones

De même, l'ensemble des distributions conditionnelles de X sachant les modalités de Y est l'ensemble des profils-colonnes, que l'on peut présenter dans le tableau suivant :

Y / $X_{\|Y}$	D_1	D_2	...	D_J
C_1	$n_{11}/n_{.1}$	$n_{12}/n_{.2}$	...	$n_{1J}/n_{.J}$
C_1	$n_{21}/n_{.1}$	$n_{22}/n_{.2}$	...	$n_{2J}/n_{.J}$
...	...	...	...	...
C_I	$n_{I1}/n_{.1}$	$n_{I2}/n_{.2}$	...	$n_{IJ}/n_{.J}$
Total	1	1	1	1

Exemple 29. *Ensemble des profils-colonnes du couple de variables "Situation professionnelle" et "Situation vis-à-vis de l'immigration".*

2. Laison linéaire entre deux variables quantitatives

On observe une série statistique $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ composée de n couples d'observations d'un couple de variables (X, Y). La méthodologie que nous présentons est valable aussi bien dans le cas où les données sont regroupées en modalités que dans le cas où elles sont regroupées en classes.
Dans le deuxième cas, on raisonne avec les centres des classes.
Nous supposons ici que les variables X et Y ont respectivement un nombre fini de modalités notées respectivement $X_1, ..., X_I$ et $Y_1, ..., Y_J$.
On désigne par n_{ij} le nombre d'individus qui vérifient à la fois les modalités X_i de X et Y_j de Y et par f_{ij} la fréquence associée.

Y / X	Y_1	Y_2	...	Y_J	Total
X_1	n_{11}	n_{12}	...	n_{1J}	$n_{1.}$
X_2	n_{21}	n_{22}	...	n_{2J}	$n_{2.}$
...	...	...	...	...	...
X_I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.J}$	n

Avec

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}; \quad n_{\cdot j} = \sum_{i=1}^{I} n_{ij}; \quad \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i\cdot} = \sum_{j=1}^{J} n_{\cdot j} = n$$

2.1. Caractéristiques marginales et conditionnelles

Moyennes et variances marginales

$$X : \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{I} n_{i\cdot} X_i, \qquad s^2(X) = \frac{1}{n}\sum_{i=1}^{I} n_{i\cdot}(X_i - \bar{X})^2$$

$$Y : \qquad \bar{Y} = \frac{1}{n}\sum_{j=1}^{J} n_{\cdot j} Y_j, \qquad s^2(Y) = \frac{1}{n}\sum_{j=1}^{J} n_{\cdot j}(Y_j - \bar{Y})^2$$

Moyennes et variances conditionnelles

$$X|Y = Y_j : \qquad \bar{X}_j = \frac{1}{n_{\cdot j}}\sum_{i=1}^{I} n_{ij} X_i, \quad s_j^2(X) = \frac{1}{n_{\cdot j}}\sum_{i=1}^{I} n_{ij}(X_i - \bar{X}_j)^2.$$

$$Y|X = X_i : \qquad \bar{Y}_i = \frac{1}{n_{i\cdot}}\sum_{j=1}^{J} n_{ij} Y_j, \quad s_i^2(Y) = \frac{1}{n_{i\cdot}}\sum_{j=1}^{J} n_{ij}(Y_j - \bar{Y}_i)^2.$$

2.2. Covariance

Définition 3.1. *On définit la* **covariance** *de X et deY par :*

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}[(X_i - \bar{X})(Y_j - \bar{Y})].$$

L'unité dans est exprimée la covariance est le produit des unités de X et de Y.

Remarque 3.2. *Lien avec la variance : $Cov(X,X) = Var(X)$*

Remarque 3.3. *Formule pratique :*

$$Cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}X_iY_j\right) - \bar{X}\bar{Y}.$$

Propriété 3.1. *Changement d'échelle : soient a, b, c, d des constantes réelles. On a*

$$Cov(aX+b, cY+d) = acCov(X,Y).$$

Proposition 3.1. *Expression de la variance d'une somme de variables :*

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

Proposition 3.2. *Inégalité de Cauchy-Schwarz :*

$$\|Cov(X,Y)\| \leq \sigma_X\sigma_Y.$$

Preuve : Pour tout réelle a, on peut développer grâce à la proposition 1 la quantité $Var(X + aY) \geq 0$:

$$\begin{aligned} Var(X+aY) &= Var(X) + Var(aY) + 2Cov(X,aY) \\ &= Var(X) + a^2Var(Y) + 2aCov(X,Y) \quad \text{par la propriéte 1} \\ &\geq 0 \end{aligned} \tag{3.1}$$

Le polynôme du second degré en a étant de signe constant, son discriminant est négatif ou nul :

$$4(Cov(X,Y))^2 - 4Var(X)Var(Y) \leq 0,$$

d'où l'égalité recherchée.
Remarquons au passage que le cas d'égalité se produit lorsque le discriminant de l'équation 3.1 est nul. Dans ce cas, l'équation admet une racine double :

$$\begin{aligned} a &= -\frac{2Cov(X,Y))}{2Var(Y)} = -\frac{Cov(X,Y)}{Var(Y)} \\ &= \begin{cases} -\frac{\sigma_X}{\sigma_Y} & \text{si} \quad Cov(X,Y) = +\sigma_X\sigma_Y \\ \frac{\sigma_X}{\sigma_Y} & \text{si} \quad Cov(X,Y) = -\sigma_X\sigma_Y \end{cases} \end{aligned}$$

Dans le premier cas, cela signifie que $X - \frac{\sigma_X}{\sigma_Y}Y$ a une variance nulle, donc est une constante, d'où

$$X = \frac{\sigma_X}{\sigma_Y}Y + constante.$$

Dans le second cas,

$$X = -\frac{\sigma_X}{\sigma_Y}Y + constante.$$

Ces deux cas sont les seuls cas d'égalité dans la proposition 2. Ils correspondent au fait que les variables Y et Y s'obtiennent l'une à partir de l'autre par une application affine.

2.3. Coefficient de corrélation

Définition 3.2. *Le coefficient de corrélation $r(X,Y)$ est défini par :*

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

C'est un coefficient sans unité. Sa valeur absolue est invariante par translation et changement d'échelle des variables : pour toutes constantes réelles $a \neq 0$, b, $c \neq 0$, d,

$$r(aX+b, cY+d) = \frac{ac}{|ac|} r(X,Y).$$

Propriété 3.2. *il découle de la proposition 2 que*

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

De plus, les cas de l'égalité sont les suivantes :

- *$r(X,Y) = 1$ si et seulement si les deux variables satisfont une relation affine du type $Y = aX + b$ avec $a > 0$.*
- *$r(X,Y) = -1$ si et seulement si les deux variables satisfont une relation affine du type $Y = aX + b$ avec $a < 0$.*

Lorque le nuage des points (x_i, y_i) est exactement situé sur une droite (cas idéal), on est dans la situation où $r(X,Y) = \pm 1$. Lorsque $r(X,Y)$ est proche de ± 1 (pour fixer les idées : $r^2(X,Y) \geq \frac{3}{4}$, alors il y'a une laison linéaire importante entre X et Y. Lorsqu'au contraire $r(X,Y)$ est proche de 0, alors il n'existe pas de relation linéaire entre X et Y. Attention, il peut y avoir quand même un autre type de laison entre X et Y.

2.4. Régression linéaire

On suppose à présent que les observations du couple de variable (X,Y) satisfont une relation de la forme suivante,

$$y_i = ax_i + b + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.2}$$

où a et b sont des coefficients réels. Le terme ϵ_i désigne un *bruit*, c'est à dire une pertubation supposée *petite*. Dans ce cours, on ne cherchera pas à donner un sens précis à la mesure de ce bruit.
Disposant des observations $(x_i, y_i)_{i=1}^n$ du couple (X,Y), on cherche à trouver les coefficients a et b qui permettent le mieux d'ajuster les données à une relation du type (3.2), au sens du critère des moindres carrés. On cherche

$$\min_{a,b} \sum_{i=1}^{n} (y_i - b - ax_i)^2. \tag{3.3}$$

La solution, qui s'obtient en annulant les dérivées partielles de la fonction de (a, b) qui est minimisée en (3.3), est

$$\hat{a} = \frac{Cov(X,Y)}{Var(X)},$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x},$$

où $\bar{x}$ et $\bar{y}$ désigne les moyennes respective de X et Y. La droite des moindres carrés est la droite d'équation : $y = \hat{a}x + \hat{b}$. On peut remarquer qu'elle passe toujours par le barycentre $(\bar{x}, \bar{y})$ du nuage de points. Sa pente peut aussi s'écrire à l'aide du coefficient de corrélation : $\hat{a} = r(X,Y)\frac{\sigma_X}{\sigma_Y}$.

Prediction

Pour une valeur x_0 de la variable X qui ne fait pas partie des observations, on peut faire une prédiction de la valeur correspondante de Y en calculant l'ordonnée du point d'abscisse x_0 sur la droite des moindres carrés :

$$y_0 = \hat{a}x_0 + \hat{b}$$

2.5. Régression linéaire aprés transformation d'une variable

On suppose que les observations $(x_i, y_i)_{i=1}^n$ satisfont une rélation de type

$$y_i = af(x_i) + b + \epsilon_i,$$

Pour une certaine fonction f donnée et de bruit ϵ_i. On peut estimer les coefficients de la droite de régression de Y sur $f(X)$ par la méthode décrite auparavant.

3. Liaison entre deux variables qualitatives

3.1. Indépendance

Il y a indépendance stricte entre X et Y lorsque tous les profils-lignes sont identiques. Il sont dans ce cas tous identiques à la distribution marginale de Y.

De la même manière, l'indépendance a lieu lorsque tous les profils-colonnes sont égaux à la distribution marginale de X.

Ceci implique : pour tous i, j,

$$n_{ij} = \frac{n_{i.} \times n_{.j}}{n}. \tag{3.4}$$

Réciproquement, si (3.4) a lieu, alors il y a indépendance entre X et Y.

preuve :

On définit donc pour tout tableau conjoint : $n^*_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ effectif théorique. De même, on définit la fréquence théorique : $f^*_{ij} = \frac{n^*_{ij}}{n} = f_{i.} \times f_{.j}$.
Deux caractères X et Y sont indépendants si pour tout i, j l'effectif observé n_{ij} est égal à l'effectif théorique n^*_{ij}. Sinon, il y a une liaison entre X et Y.
Remarque : La notion d'indépendance est une notion symétrique. Si X est indépendant de Y alors Y est indépendant de X.

3.2. Liaison fonctionnelle

Les caractères X et Y étant considerés, X est lié fonctionnellement à Y si à chaque modalité Y_j de Y correspond une seule modalité possible de X i.e., $n_{ij} = 0$ pour tout j, sauf pour $i = \varphi(j)$, ou encore $n_{ij} = n_{.j}$ pour $i = \varphi(j)$. (Il faut que $I \leq J$).

Ainsi, dans chaque colonne du tableau de contingence, un terme et un seul est différent de 0. en revanche, une même ligne peut comporter plusieurs termes non nuls.
Remarque : lorsque X est lié fonctionnellement à Y et Y est lié fonctionnellement à X, on dit qu'il y a liaison fonctionnelle réciproque entre X et Y. Dans ce cas, les caractères X et Y doivent avoir le même nombre de modalités ($I = J$).

3.3. Mesure de la liaison

Les deux cas extrêmes que nous venons de présentés (indépendance et liaison fonctionnelle) se réalisent très rarement en pratique. Ainsi, en pratique, l'analyse de la liaison ou de l'indépendance de deux caractères revient à analyser si le tableau conjoint des deux caractères se rapproche du tableau des effectifs théoriques (cas d'indépendance) ou si en est le plus éloigné possible (cas de la liaison fonctionnelle).
La distance du χ^2 d'écart à l'indépendance permet de mesurer le degré de dépendance entre X et Y. Elle se base sur la comparaison entre n_{ij} et n^*_{ij}.

Définition 3.3. *La **distance du** χ^2 observée sur la série statistique $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ est définie par*

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}} \right) = n \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{\left(f_{ij} - f^*_{ij}\right)^2}{f^*_{ij}} \right)$$

Exemple 30. *Distance du χ^2 pour mesurer l'écart à l'indépendance entre les variables "situation quant à l'immigration" et "situation professionnelle" en France 2006.*

Propriété 3.3.
- *la grandeur $\chi^2 = 0$ si il y a indépendance stricte entre X et Y.*
- *la grandeur χ^2 est d'autant plus élevée que la laison est forte : il existe alors des cellules (i, j) avec une écart important $n_{ij} - n^*_{ij}$.*
- *l'inégalité suivante est toujours vérifiée : $\chi^2 \leq n \times \min\{I - 1, J - 1\} = \chi^2_{max}$.*
- *si $\chi^2 = \chi^2_{max}$, alors la liaison entre X et Y est parfaite (liaison fonctionnelle).*

Si $0 < \chi^2 < \chi^2_{max}$, on calcule le coefficient de Cramer :

Définition 3.4. *Le coefficient C de Cramer est défini par :*

$$C = \sqrt{\frac{\chi^2}{\chi^2_{max}}}.$$

Propriété 3.4.
- $0 \leq C \leq 1$
- *$C = 0$ lorsqu'il y a indépendance.*
- *De petites valeurs de C signifient que la liaison entre X et Y est trés faible.*
- *Des valeurs proches de 1 signifient qu'il y a une liaison très forte entre X et Y.*
- *Ce coefficient, qui varie entre 0 et 1, permet de comparer la laison entre plusieurs couples de variables.*

On apprécie la liaison à partir du tableau suivant :

C	0	]0 ;0,2]	]0,2 ;0,4]	]0,4 ;0,7]	]0,7 ;1[	1
liaison	nulle	faible	moyenne	forte	très forte	parfaite

Généralement, on compare la valeur du χ^2 observée : χ^2_{obs} à sa valeur tabulée : χ^2_{tab} lue dans une table de loi de $\chi^2(k)$ à k dégré de liberté, où $k = (I-1)(J-1)$.

Si $\chi^2_{obs} > \chi^2_{tab}$, alors il y a assez d'évidence contre l'hypothèse d'indépendance : X et Y sont liés.
Question : Quelle est la conttribution des modalités à cette liaison ?

Exemple 31. *Calcul du C de Cramer pour mesurer l'écart à l'indépendance entre les variables "Situation quant à l'immigration" et " Situation professionnelle" en France en 2006.*

Il ne suffit pas d'établir ou réfuter une relation entre deux caractères qualitatifs. Il convient de donner la source de la relation ie les modalités qui ont contribué significativement à la construction de cette relation. Pour ce faire, on peut calculer la contribution relative de chaque case du tableau à la liaison.

3.4. Explication de la liaison : Contribution des modalités

Définition 3.5. *On appelle **contribution au** χ^2 du couple de modalités (C_i, D_j) et (X, Y) la quantité $\chi^2_{ij} = \frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}}$.*

Plus la contribution est forte, plus la laison entre les modalités C_i et D_j est importante.

Définition 3.6. *L'association entre les modalités C_i et D_j est dite **positive** si $n_{ij} - n^*_{ij} > 0$. Elle est **négative** si $n_{ij} - n^*_{ij} < 0$.*

La contribution relative associée au couple de modalités (C_i, D_j) est le rapport : $\frac{\chi^2_{ij}}{\chi^2}$. On l'exprime souvent en pourcentage et on note :

$$CR_{ij} = \frac{\chi^2_{ij}}{\chi^2} \times 100.$$

On en fait un tableau :

Y / X	D_1	...	D_j	...	D_J
C_i	CR_{ij}	...	CR_{ij}	...	CR_{ij}

On interprète le tableau en tenant compte de la valeur de $n_{ij} - n^*_{ij}$ pour savoir si la contribution est positive ou négative i.e. si les modalités sont simultanées ou antithétiques.

Exemple 32. *Liaison entre la modalité "Elèves, étudiants, stagiaires" de la variable "Situation professionnelle" et la modalité "Immigrés" de la variable "Situation quant à l'immigration".*

4. Liaison entre une variable qualitative et une variable quantitative

Soit X une variable quantitative et Y le caractère qualitatif ayant J modalités : $D_1, ..., D_J$. La méthodologie que nous présentons est valable aussi bien dans le cas où X est discret que dans le cas où il est continu. En effet, dans ce deuxième cas, si nous ne disposons que d'un tableau conjoint présentant X en classe, on peut raisonner avec les centres de classes. Nous supposons ici que X a un nombre fini de modalités notées $X_1, ..., X_I$.

4.1. Notations

Les J modalités du caractère Y divisent la population en J sous-populations.
On rappelle à cet effet que n_{ij} désigne le nombre d'individus présentant simultanément les modalités X_i de X et Y_j de Y. On notera pour la suite :

$n_{.j}$: le nombre d'individus de la sous-population j.
$\bar{X}$: Moyenne de X sur la population totale,
$\bar{X}_j$: Moyenne de X sur la sous-population j (moyenne conditionnelle)
$V(X)$: Variance de X sur la population totale,
$V_j(X)$: Variance de X sur la sous-population j (variance conditionnelle).

4.2. Caractéristiques conditionnelles

A partir de ces notations, on a :

La moyenne du groupe j : $\bar{X}_j = \frac{1}{n_{.j}} \sum_{i=1}^{I} n_{ij} X_i$

La variance du groupe j : $V_j(X) = \frac{1}{n_{.j}} \sum_{i=1}^{I} n_{ij} (X_i - \bar{X}_j)^2$.

On en déduit la moyenne et la variance de X sur la population totale en fonction des caractéristiques conditionnelles :

La moyenne sur toute la population s'écrit donc : $\bar{X} = \frac{1}{n}\sum_{j=1}^{J} n_j \bar{X}_j$

La Variance sur toute la population s'écrit : $V(X) = \frac{1}{n}\sum_{j=1}^{J} n_j V_j(X) + \frac{1}{n}\sum_{j=1}^{J} n_j(\bar{X}_j - \bar{X})^2$.

Cette relation est appelée équation d'analyse de la variance.

Soit :

`Variance totale = Moyenne des Variances + Variance des Moyennes`

`Variance Totale = Variance intra classe + Variance interclasse`

où :

$\texttt{VarIntra} = \texttt{Moyenne des Variances} = \frac{1}{n}\sum_{j=1}^{J} n_j V_j(X)$.

$\texttt{VarInter} = \texttt{Variance des Moyennes} = \frac{1}{n}\sum_{j=1}^{J} n_j(\bar{X}_j - \bar{X})^2$.

4.3. Indépendance

La distribution de X est independante de celle de Y si et seulement si on observe la même distribution de X à l'interieur de chacune des sous-populations definies par les modalités de Y. Soit les moyennes conditionnelles, moyennes calculées à l'interieur des sous-populations sont les mêmes. Il en est de même pour les variances conditionnelles.

En d'autres termes, pour tout $j = 1, .., J$, on a : $\bar{X}_j = \bar{X}$ et $V_j(X) = V(X)$.

Ainsi, pour montrer qu'une variable quantitative est indépendante d'un caractère qualitatif, on calcule les moyennes conditionnelles que l'on compare entre elles, ou à la moyenne marginale. Si elles sont les mêmes, on conclut à l'independance des deux caractères, sinon à une liaison.

4.4. Liaison fonctionnelle

X sera dit fonctionnellement lié à Y si et seulement si les individus d'une même sous population (définie par les modalités de Y) prennent des valeurs identiques de la variable X et que ces différentes valeurs sont les plus éloignées possibles. En d'autres termes, la donnée de la catégorie dans laquelle se trouve un individu permet de connaître parfaitement la valeur de X qui lui est affectée.

Ainsi, les variances conditionnelles sont égales à 0 et les moyennes conditionnelles sont le plus différenciées possible.

4.5. Mesure de la liaison

On mesure la liaison ici, à l'aide du **rapport de corrélation** $\eta_{X|Y}$ définie par :

$$\eta_{X|Y} = \sqrt{\frac{\texttt{VarInter}}{\texttt{VarTotale}}} \text{ soit } \eta^2_{X|Y} = \frac{\texttt{VarInter}}{\texttt{VarTotale}}.$$

$\eta^2_{X|Y}$ est la part de variation de X expliquée par Y dans la variation totale de X.
A partir de l'équation d'analyse de la variance, on déduit que : $0 \leq \eta_{X|Y} \leq 1$.

$\eta_{X|Y} = 0$ Independance (absence de corrélation) entre X et Y ;

$\eta_{X|Y} = 1$ Liaison parfaite (liaison fonctionnelle) entre X et Y ;

$\eta_{X|Y} \to 1$ Liaison d'autant plus forte ;

$\eta_{X|Y} \to 0$ Liaison d'autant plus faible.

Remarque : Dans le cas où la variable quantitative est continue, on peut bien le recodé en classe pour soit analyser la relation entre deux caractères qualitatifs soit travailler avec les centres de classes et donc refaire le travail en raisonnant avec les centres des classes.