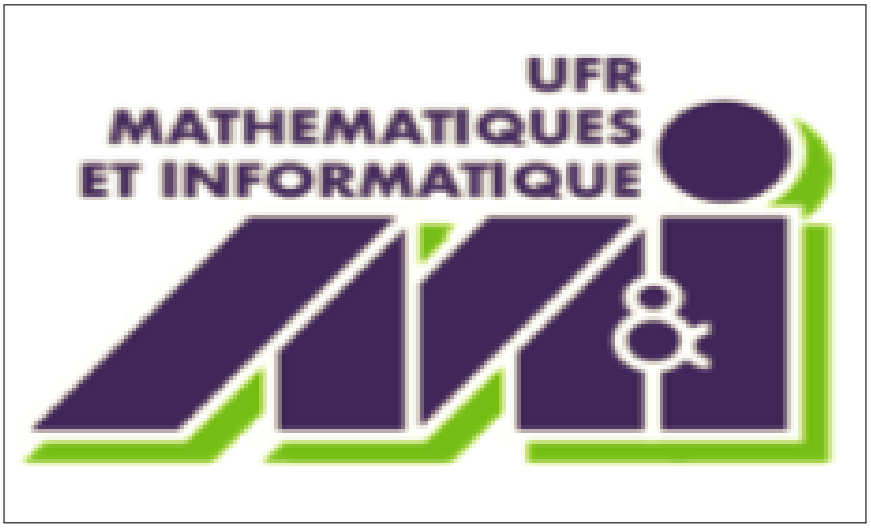
Probabilités-Statistique

Université de San-Pedro
UFR Agriculture, Ressources Halieulitique et Agro-Industrie
(ARHAI)

Licence 1

Prof. Monsan Vincent
Prof. Armel yodé



Université Félix Houphouët-Boigny

Table des matières

Chapitre 1

Introduction

La statistique est l'ensemble des méthodes et des techniques destinées à la collecte, l'exploration, l'analyse et l'interprétation des données. Elle a pour objectif de mettre en évidence des informations cachées dans ces données en vue généralement de prendre une décision concernant le phénomène ayant généré ces données. La statistique se divise généralement en deux grandes parties :

- la statistique descriptive qui a pour but d'obtenir un résumé des données ;
- la statistique inférentielle qui a pour but d'utiliser les données afin de rechercher des modèles ou de faire des prévisions.

1.1 Terminologie de base

Population. C'est l'ensemble sur lequel porte l'étude statistique. La population que l'on envisage en statistique dépend du domaine que l'on traite, et peut donc aussi bien être constituée d'êtres humains que d'animaux voire d'objets.

Individu ou unité statistique. C'est un élément de la population.

Echantillon. C'est un sous-ensemble de la population ; l'échantillon doit être représentatif de la population, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité ; en effet, les informations obtenues à partir de l'échantillon doivent pouvoir être étendues, sans erreur grave, à l'ensemble de la population.

Enquête statistique. C'est l'opération consistant à collecter des données sur l'ensemble des individus d'un échantillon ou éventuellement la population entière.

Recensement. C'est une enquête statistique effectuée sur toute la population.

Sondage. C'est une enquête statistique effectuée sur un échantillon de la population.

Caractère. C'est une grandeur ou un attribut observable sur un individu ; parfois, on emploie le terme de variable statistique au lieu de caractère.

Modalité. C'est un état du caractère ; les modalités d'un caractère sont exhaustives et incompatibles, c'est à dire que chaque individu présente une et une seule modalité du caractère.

Série statistique : C'est la suite des valeurs du caractère observée sur chaque individu de l'ensemble étudié (population ou échantillon).

1.2 Caractères

On distingue deux types de caractères : le caractère qualitatif et le caractère quantitatif.

1.2.1 Caractère qualitatif

Le caractère est dit qualitatif si ses modalités sont non mesurables. Le caractère qualitatif est dit ordinal s'il existe un ordre entre ses modalités. Dans le cas contraire, il est dit qualitatif nominal.

Exemple 1.2.1. *Caractère qualitatif ordinal.*

- *Population : la classe.*
- *Individu : un étudiant*
- *Caractère : décision du jury*
- *Modalités : ajourné, passable, assez-bien, bien, très bien.*

Exemple 1.2.2. *Caractère qualitatif nominal.*

- *Population : la classe.*
- *Individu : un étudiant*
- *Caractère : groupe sanguin.*
- *Modalités : A, B, AB et O.*

1.2.2 Caractère quantitatif

Lorsque les modalités d'un caractère sont mesurables, on dit que ce caractère est quantitatif.

1.2.2.1 Caractère quantitatif discret

Le caractère quantitatif est dit discret lorsqu'il ne peut prendre que des valeurs isolées notées par exemple $x_1, x_2, \ldots, x_k$ où k est le nombre de modalités.

Exemple 1.2.3.

- *Population : le personnel d'une entreprise*
- *Individu : un employé*
- *Caractère : nombre d'enfants*
- *Modalités :* 0, 1, 2, 3, 4, 5, 6 *et* 7.

1.2.2.2 Caractère quantitatif continu

Le caractère quantitatif est dit continu lorsqu'il peut prendre n'importe quelle valeur d'un intervalle de l'ensemble des nombres réels $\mathbb{R}$. Dans ce cas, l'intervalle des valeurs possibles est divisé en k classes

$$[a_0, a_1[,\ [a_1, a_2[, \ldots, [a_{k-1}, a_k[, \quad \text{où} \quad a_0 < a_1 < \ldots < a_{k-1} < a_k.$$

a_{j-1} et a_j sont les frontières de la j-ième classe, $c_j = \frac{a_{j-1}+a_j}{2}$ est le centre de celle-ci. L'amplitude de cette classe est $a_j - a_{j-1}$. On supposera que les observations d'une classe sont concentrées au centre.

Exemple 1.2.4. - *Population : l'ensemble des ouvriers d'une entreprise*

- *Individu : un ouvrier*
- *Caractère : salaire mensuel net (en milliers francs)*
- *Modalités :* $[80,100[$, $[100,110[$, $[110,120[$, $[120,130[$ *et* $[130,150[$.

Le centre de la classe $[80,100[$ *est :*

$$\frac{80+100}{2} = 90.$$

La répartition en classes des données nécessite de définir a priori le nombre de classes J et donc l'amplitude de chaque classe. Il existe des formules qui nous permettent d'établir le nombre de classes et l'amplitude pour une série statistique de n observations.

— La règle de Sturge : $J = 1 + 3.3 \times \log_{10}(n)$

— La règle de Yule : $J = 2.5 \times n^{1/4}$.

L'amplitude de classe est obtenue ensuite de la manière suivante :

$$\text{amplitude} = \frac{x_{max} - x_{min}}{J}$$

où x_{max} (resp. x_{min}) désigne la plus grande (resp. la plus petite) valeur observée.

1.3 Effectif, fréquences

On observe un caractère X présentant k modalités sur n individus. L'effectif n_i de la i-ème modalité du caractère est le nombre d'individus qui possède cette modalité. On a

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

On appelle fréquence de la i-ème modalité le rapport

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

La fréquence est la proportion par rapport au nombre d'observations des individus pour lesquels le caractère prend la valeur x_i ou appartient à la classe $[a_i, a_{i+1}[$. Elle est un nombre réel compris entre 0 et 1. Nous avons

$$\sum_{i=1}^{k} f_i = 1.$$

On exprime la fréquence souvent en pourcentage :

$$f_i = \frac{n_i}{n} \times 100\,\%$$

Dans ce cas, nous avons :

$$\sum_{i=1}^{k} f_i = 100.$$

1.3.1 Effectifs cumulés, Fréquences cumulées

On suppose que les modalités du caractère quantitatif étudié sont rangées par ordre croissant. L'effectif cumulé croissant de la i-ème modalité x_i est la somme des effectifs des modalités inférieures ou égales à cette modalité :

$$N_i = \sum_{j=1}^{i} n_j = n_1 + \cdots + n_i.$$

Le nombre n_i représente le nombre d'observations inférieures ou égale à x_i.

La fréquence cumulée croissante de la i-ème modalité x_i est la somme des fréquences des modalités inférieures ou égales à cette modalité :

$$F_i = \sum_{j=1}^{i} f_j = \frac{N_i}{n}.$$

Cette fréquence représente la proportion (ou le pourcentage) des observation inférieures ou égales à la i-ème modalité x_i du caractère quantitatif si il est discret ou bien inférieures à la borne supérieure du i-ème intervalle s'il est continu.

L'effectif cumulé décroissant de la i-ème modalité x_i est la somme des effectifs des modalités supérieures ou égales à cette modalité :

$$D_i = \sum_{j=i}^{k} n_j.$$

La fréquence cumulée décroissante de la i-ème modalité est la somme des fréquences des modalités supérieures ou égales à cette modalité :

$$G_i = \sum_{j=i}^{k} f_j = \frac{D_i}{n}.$$

1.4 Présentation générale des tableaux statistiques

On considère un échantillon de taille n issu d'une population. Pour chaque individu, on fait une observation concernant le caractère X comportant k modalités M_1, M_2,..., M_k. On obtient une série statistique $x_1,\ldots,x_n$. Les données recueillies, appelées données brutes, sont soumises à un premier traitement afin d'en faciliter à la fois la présentation et l'exploitation. Cela consiste à classer chacun des n individus dans les k sous-ensembles définis par les diverses modalités du caractère X. Pour chaque modalité M_i, on pourra inscrire dans le tableau statistique son effectif n_i, son effectif cumulé croissant ou décroissant, sa fréquence f_i et sa fréquence cumulée croissante ou décroissante. On prendra toujours soin de préciser dans la présentation du tableau :

- la population étudiée et le caractère ;
- l'origine du renseignement.

La présentation des données sous forme de tableaux est intéressante car elle propose un premier résumé. On dégage ainsi les tendances de la population. Ces tableaux vont nous permettre de faire des représentations graphiques. L'idée sera de rendre compte visuellement du résumé que nous avons commencé. Ensuite, pour les caractères quantitatifs, nous chercherons à résumer numériquement l'information.

Modalité	Effectif	Effectif Cumulé	Fréquence	Fréquence cumulée
M_1	n_1	n_1	$f_1 = \frac{n_1}{n}$	$F_1 = f_1$
M_2	n_2	$n_1 + n_2$	$f_2 = \frac{n_2}{n}$	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M_j	n_j	$n = \sum_{i=1}^{j} n_i$	$f_j = \frac{n_j}{n}$	$F_j = \sum_{i=1}^{j} f_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M_k	n_k	$n = \sum_{i=1}^{k} n_i$	$f_k = \frac{n_k}{n}$	$F_k = \sum_{i=1}^{k} f_i = 1$
Total	n		1	

Exemple 1.4.1. Caractère quantitatif discret : *Répartition des ménages selon le nombre de pièces du logement occupé*

Nombre de pièces	*Nombre de ménages*	*Fréquence*	*Fréquence cumulée croissante*	*Fréquence cumulée decroissante*
1	*20*	$\frac{20}{200} = 0.1$	*0.1*	*1*
2	*40*	0.2	*0.3*	*0.9*
3	*40*	0.2	*0.5*	*0.7*
4	*60*	*0.3*	*0.8*	*0.5*
5	*40*	*0.2*	*1*	*0.2*
Total	*200*	*1*		

Exemple 1.4.2. Caractère quantitatif continu : *Répartition des ouvriers selon leur salaire mensuel net (en milliers francs).*

Salaire	*Effectif*	*Fréquence (%)*	*Fréquence cumulées croissante (%)*	*Fréquence cumulées decroissantes (%)*
$[80, 100[$	*26*	*18.6*	*18.6*	*100*
$[100, 110[$	*33*	*23.5*	*42.1*	*81.4*
$[110, 120[$	*64*	*45.8*	*87.9*	*57.9*
$[120, 130[$	*7*	*5.0*	*92.9*	*12.1*
$[130, 150[$	*10*	*7.1*	*100*	*7.1*
Total	*140*	*100*		

Exemple 1.4.3. Caractère qualitatf ordinal : *Répartition de* 50 *personnes selon le dernier diplôme obtenu.*

Diplôme	*Nombre de personnes*	*Fréquence*	*Fréquence cumulée croissante*	*Fréquence cumulée decroissante*
Sans diplôme	*4*	$\frac{4}{50} =$		*1*
Primaire	*11*			
Secondaire	*14*			
Supérieur	*21*			
Total	*50*			

Exemple 1.4.4. Caractère qualitatif nominal : *Répartition des touristes et visiteurs arrivés à l'aéoroport Félix Houphouet-Boigny par continent de provenance en 1999.*

Continent	*Effectif*	*Fréquence (%)*
Afrique	*168238*	
Europe	*164542*	
Amérique	*27540*	
Asie	*15058*	
Océanie	*1014*	
Total		*100*

Source : Ministère de l'Economie et des Finances, 2007.

Remarque 1.4.1. *Il s'agit d'un caractère qualitatif nominal. Les fréquences cumulées sont sans intérêt car il n'existe pas de relation d'ordre entre les modalités.*

Chapitre

2 Représentations graphiques

2.1 Introduction

La représentation graphique a pour objectif de visualiser la distribution des données. Dans ce chapitre, nous passons en revue les principales représentations graphiques utilisées dans les analyses statistiques. Selon le type de variable statistique étudié, on a recours à des graphiques différents.

2.2 Diagrammes à secteurs

Les diagrammes à secteurs conviennent pour représenter les effectifs et les fréquences des caractères qualitatifs ou des caractères quantitatifs discrets. Un diagramme en secteurs est un graphique constitué d'un cercle divisé en secteurs dont les angles au centre sont proportionnels aux effectifs (ou aux fréquences). L'angle α_i d'une modalité d'effectif n_i ou de fréquence f_i est donné en dégrés par

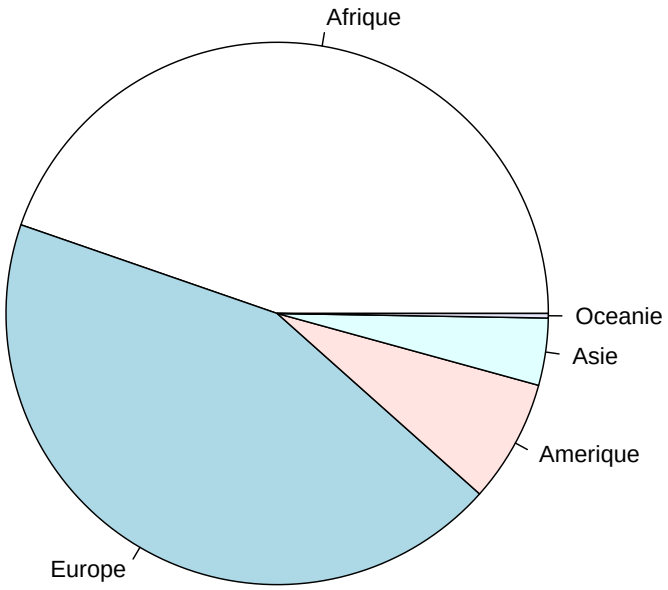
$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360.$$

Exemple 2.2.1. *Caractère qualitatif :*

Continent	Effectif	Fréquence (%)
Afrique	168238	44.7
Europe	164542	43.72
Amérique	27540	7.32
Asie	15058	4.00
Océanie	1014	0.27
Total	376392	100

TABLE 2.1 – Répartition des touristes et visiteurs arrivés à l'aéoroport Félix Houphouet-Boigny par continent de provenance en 1999

Source : Ministère de l'Economie et des Finances, 2007.



2.3 Diagramme en barres, diagramme en bâtons

Les diagrammes en barres et les diagrammes en bâtons conviennent pour représenter les fréquences des caractères qualitatifs ou quantitatifs discrets. Les modalités du caractère sont en abscisse et les fréquences sont en ordonné. Dans le cas d'un caractère qualitatif nominal, la position des modalités n'a pas de signification particulière. Si le caractère est qualitatif ordinal ou quantitatif discret, on placera les modalités dans leur ordre naturel.

— **Le digramme en barres :** à chaque modalité du caractère, on associe un rectangle de base constante dont la hauteur est proportionnelle à la fréquence.
— **Le diagramme en bâtons :** à chaque modalité du caractère, on fait correspondre un segment vertical de longueur proportionnelle à la fréquence de cette modalité.

Exemple 2.3.1. *Le tableau suivant donne la répartition selon le groupe sanguin de 50 individus pris au hasard dans une population :*

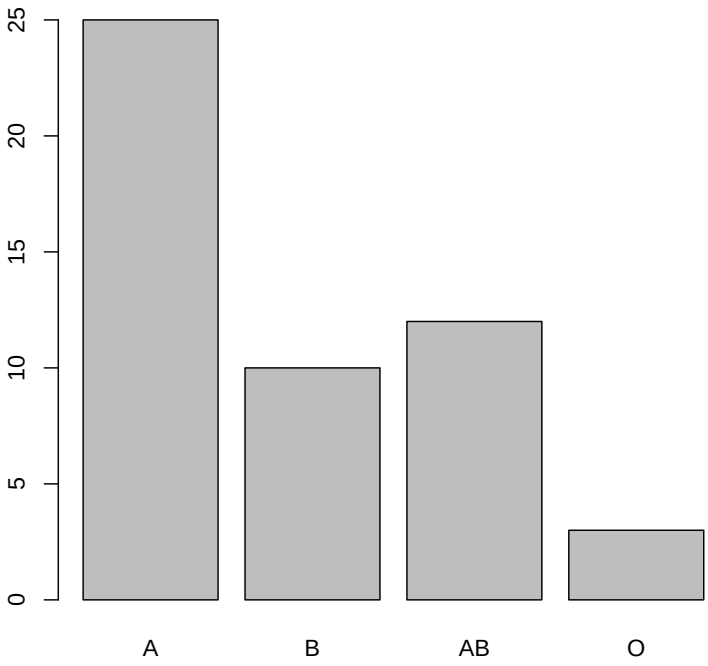
Groupe sanguin	*A*	*B*	*AB*	*O*
Effectif	25	*10*	12	3

1. *Déterminer la variable statistique et son type.*

   Variable statistique : groupe sanguin
   Nature : qualitative nominale.

2. *Donnez une représentation graphique qui fasse apparaître l'importance relative des différents groupes sanguins.*

   Nous pouvons faire un diagramme en barres ou un diagramme en secteurs



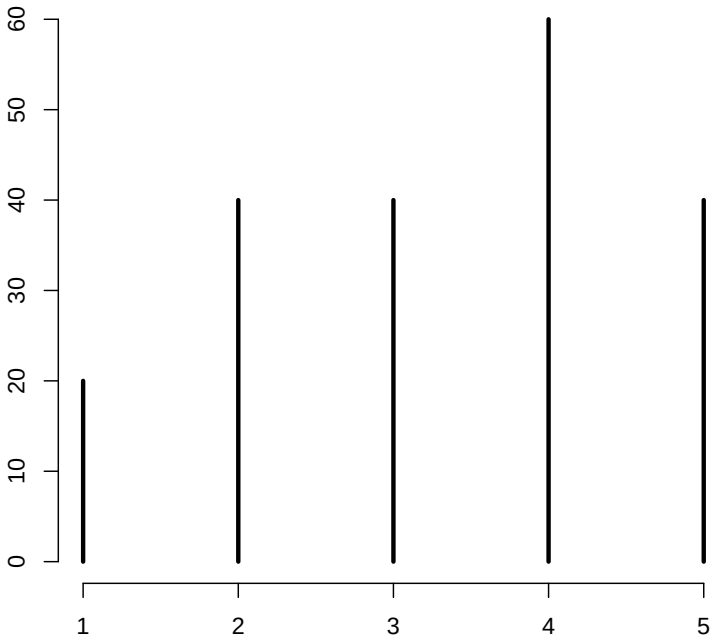
Exemple 2.3.2. *A Cauphygombokro, en vue d'instaurer la taxe d'habitation, une enquête portant sur le nombre de pièces du logement occupé a été réalisée auprès des ménages. Cette enquête a donné les résultats suivants :*

Nomrbe de pièces	*Nombre de ménages*
1	*20*
2	*40*
3	*40*
4	*60*
5	*40*

1. *Caractériser la distribution (population, individu, caractère, nature du caractère, modalités)*

   Population : l'ensemble des ménages de Cauphygombokro
   Individu : un ménage
   Caractère : nombre de pièces du logement occupé
   Modalités : 1,2,3,4,5.

2. *Tracer le diagramme en bâtons.*



2.4 Histogramme

L'histogramme est la représetation graphique de la distribution des effectifs ou des fréquences d'une variable statistique continue. Pour construire l'histogramme, on place en abscisse les différentes extrémités a_i des classes, puis on trace, pour chaque classe, un rectangle parallèle aux axes, de telle sorte que la partie parallèle à l'axe des abscisses ait une longueur correspondant à l'amplitude de la classe et que la surface du rectangle soit proportionnelle à l'effectif (ou à la fréquence) de la classe (ceci afin de bien visualiser l'importance de chaque classe). Deux classes de même amplitude sont directement comparables. Cette comparaison ne peut être étendue à des classes d'amplitude différente. Pour effectuer la comparaison correctement, nous allons construire les histogrammes en respectant le protocole ci-dessus :

- Choix de l'unité d'amplitude u : on retiendra par exemple le pgcd des diverses amplitudes.
- Expression des amplitudes dans cette nouvelle unité d'amplitude :

$$e_i = \frac{a_i - a_{i-1}}{u}$$

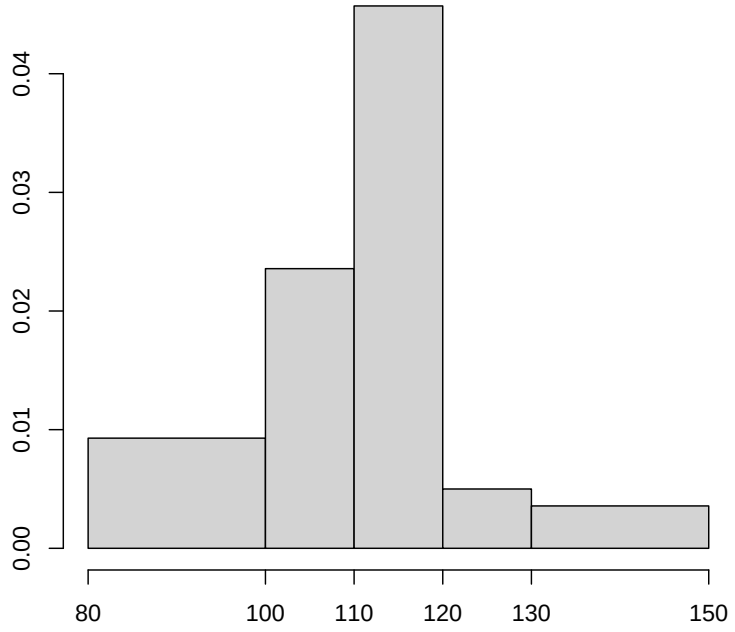
- La hauteur h_i de chaque rectangle est égale à

$$h_i = \frac{f_i}{a_i}$$

de telle sorte que la surface des rectangles représentatifs est égale à la fréquence de la classe correspondante ; h_i est la fréquence par unité d'amplitude de la classe i.

Exemple 2.4.1. *Répartition des ouvriers selon leur salaire mensuel net (en milliers francs.*

Salaire	*Effectif*	*Fréquence (%)*	*Fréquence cumulées (%)*
[80,100[	*26*	*18.6*	*18.6*
[100,110[	*33*	*23.5*	*42.1*
[110,120[	*64*	*45.8*	*87.9*
[120,130[	*7*	*5.0*	*92.9*
[130,150[	*10*	*7.1*	*100*
Total	*140*	*100*	



2.5 Diagramme de fréquences cumulées

2.5.1 Cas d'un caractère qualitatif ordinal

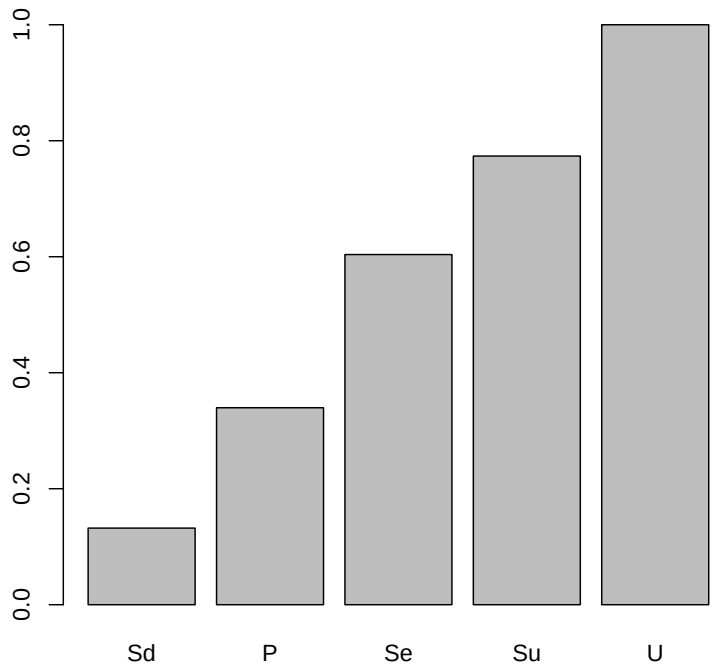
Exemple 2.5.1. *On interroge* 50 *personnes sur leur dernier diplôme obtenu (Sans diplôme, Primaire, Secondaire, Supérieur non universitaire, Universitaire). On a obtenu la série statistique suivante (Sd= Sans diplôme, P=Primaire, Se=Secondaire, Su=Supérieur, U=Universitqire)*

Sd Sd Sd Sd P P P P P P P P P P Se Se Se Se Se Se Se Se Se Se Se Se Su Su Su Su Su Su Su Su Su Su U U U U U U U U U U U U.

```
   Eff EffCum Freq FreqCum
Sd   7      7 0.13    0.13
P   11     18 0.21    0.34
```

```
Se  14     32 0.26    0.60
Su   9     41 0.17    0.77
U   12     53 0.23    1.00
```

Les fréquences cumulées d'une variable qualitative ordinale sont représentées au moyen d'un diagramme en barres.



2.5.2 Cas d'un caractère quantitatif discret

C'est la représentation graphique de la fonction F_X définie par

$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1 \\ F_i & \text{si } x_i \le x < x_{i+1} \ i = 1, \ldots, k-1, \\ 1 & \text{si } x \ge x_k \end{cases}$$

ou

$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1 \\ N_i & \text{si } x_i \le x < x_{i+1} \ i = 1, \ldots, k-1, \\ n & \text{si } x \ge x_k \end{cases}$$

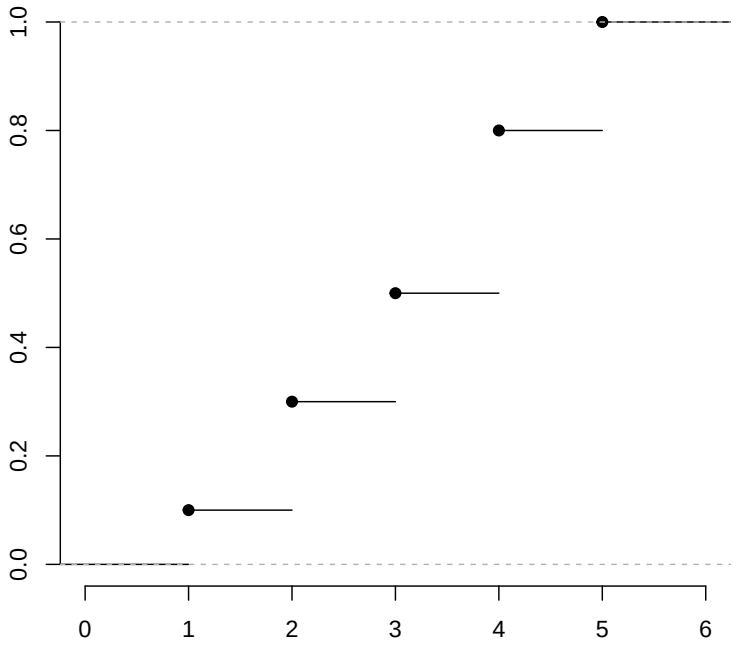
Exemple 2.5.2. *Répartition des ménages selon le nombre de pièces du logement occupé. Le tableau statistique est :*

	Effectif	Frequence	Frequence_Cumulees
1	20	10	10
2	40	20	30
3	40	20	50
4	60	30	80
5	40	20	100

A partir de ce tableau, nous déduisons :

$$F(x) = \begin{cases} 0 & si\ x < 1 \\ 10 & si\ 1 \leq x < 2 \\ 30 & si\ 2 \leq x < 3 \\ 50 & si\ 3 \leq x < 4 \\ 80 & si\ 4 \leq x < 5 \\ 100 & x \geq 5 \end{cases}$$

La courbe des fréquences cumulées est donc :



2.5.3 Cas d'un caractère quantitatif continu

La courbe cumulative est la représentation graphique de la fonction cumulative. Les observations étant groupées par classe, on ne connait de cette fonction que les valeurs qui correspondent aux extrémités supérieures de chaque classe et pour lesquelles elle est égale à la fréquence cumulée F_i :

$$F(a_i) = F_i$$

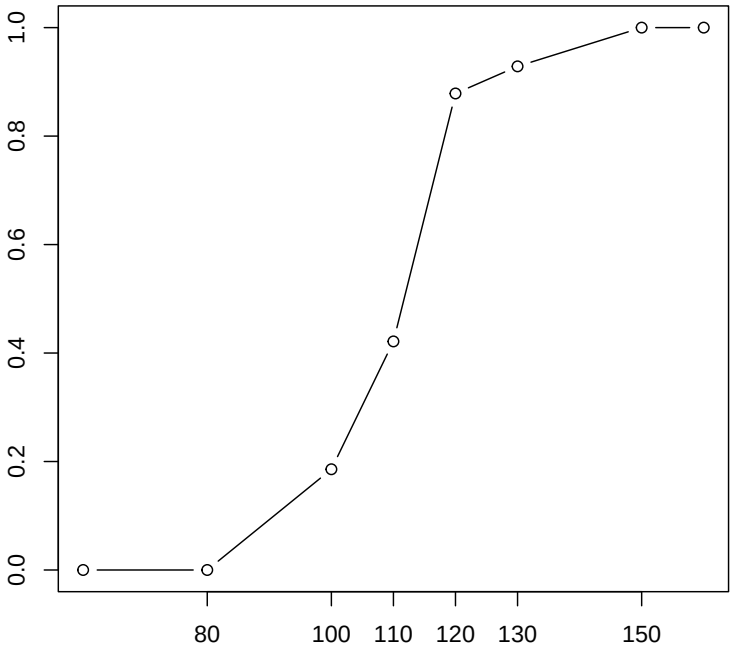
Exemple 2.5.3. *Répartition des ouvriers selon leur salaire mensuel net (en milliers francs.*

Salaire	*Effectif*	*Fréquence (%)*	*Fréquence cumulées (%)*
[80,100[	*26*	*18.6*	*18.6*
[100,110[	*33*	*23.5*	*42.1*
[110,120[	*64*	*45.8*	*87.9*
[120,130[	*7*	*5.0*	*92.9*
[130,150[	*10*	*7.1*	*100*
Total	*140*	*100*	

Dans notre exemple, nous avons :

$$\begin{aligned} F(80) &= 0 \\ F(100) &= 0.186 \\ F(110) &= 0.421 \\ F(120) &= 0.879 \\ F(130) &= 0.929 \\ F(150) &= 1. \end{aligned}$$

La courbe des fréquences cumulées est donc :



Chapitre 3

Paramètres numériques

On distingue les paramètres de tendance centrale (ou de position ou de localisation), les paramètres de dispersion, les paramètres de concentration et les paramètres de forme.

3.1 Paramètres de tendance centrale

Les paramètres de tendance centrale ont pour objet de résumer la série d'observations par une valeur considérée comme représentative. Selon les cas, certains sont plus appropriés que d'autres.

3.1.1 Le mode

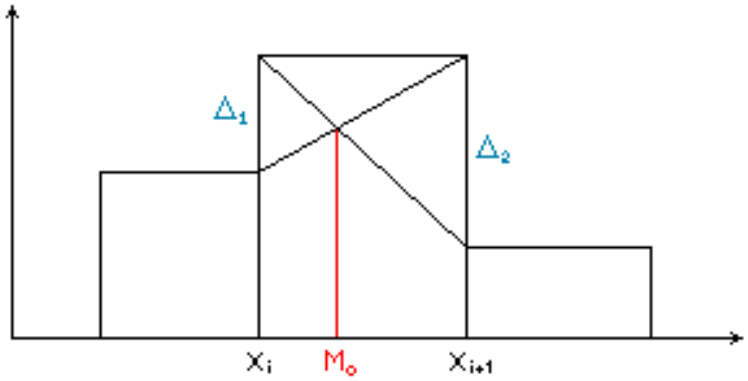
Le mode est la valeur la plus fréquente du caractère. Il peut être calculé pour tous les types de caractère (quantitatif ou qualitatif). Le mode n'est pas nécessairement unique.

3.1.1.1 Caractère quantitatif discret

Le mode d'un caractère quantitatif discret est la valeur pour laquelle la fréquence est la plus élevée. Graphiquement, le mode est la modalité qui correspond au sommet du diagramme en bâton.

3.1.1.2 Caractère quantitatif continu

Le mode est plus difficile à définir dans le cas d'un caractère quantitatif continu. Lorsque les données sont regroupées en classes, on définit la classe modale. La classe modale n'est pas la classe de plus grande fréquence mais la classe de plus grande densité c'est à dire de plus grande fréquence par amplitude. Il est néanmoins possible de déterminer une valeur unique comme mode.



La classe modale $[x_i, x_{i+1}[$ étant déterminée, le mode M_0 est égale est :

$$M_0 = x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(x_{i+1} - x_i).$$

Lorsque les classes adjacentes à la classe modale ont des densités de fréquences égales, le mode coincide avec le centre de la classe modale. Le mode dépend beaucoup de la répartition en classes.

3.1.2 La moyenne arithmétique

3.1.2.1 Données brutes

Pour une série statistique $x_1, x_2, \ldots, x_n$, on définit la moyenne par

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

C'est la somme de toutes les observations divisée par le nombre total des observations.

3.1.2.2 Données rangées : caractère quantitatif discret

Pour un caractère quantitatif discret dont les n observations sont rangées selon ses k modalités $x_1, \ldots, x_k$ d'effectifs respectifs $n_1, \ldots, n_k$, la moyenne est

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{k} n_i x_i.$$

3.1.2.3 Données rangées : caractère quantitatif continu

Pour un caractère quantitatif continu dont les n observations ont été réparties dans k intervalles $([a_{i-1}, a_i[)_{i=1,\ldots,k}$, la moyenne est

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$$

où $c_i = \dfrac{a_{i-1} + a_i}{2}$ est le centre de la classe $[a_{i-1}, a_i[$.

3.1.2.4 Remarques

- La moyenne n'est pas nécessairement une valeur observable du caractère.
- La moyenne est sensible aux valeurs extrêmes ou atypiques.

3.1.3 La moyenne géométrique

Si $x_i \geq 0$, alors la moyenne géométrique est définie par

$$G = \left(\prod_{i=1}^{n} x_i^{n_i}\right)^{1/n} = \left(x_1 \times x_2 \times ... \times x_n\right)^{1/n}.$$

La moyenne géométrique s'utilise, par exemple, quand on veut calculer la moyenne de taux d'intérêt.

Exemple 3.1.1. *Supposons que les taux d'intérêt pour* 4 *années consécutives soient respectivement de* 5*,* 10*,* 15*, et* 10%*. Que va-t-on obtenir après* 4 *ans si je place* 100 *francs ?*

Solution.
— *Après* 1 *an on a,* $100 \times 1.05 = 105$.
— *Après* 2 *ans on a,* $100 \times 1.05 \times 1.1 = 115.5$
— *Après* 3 *ans on a,* $100 \times 1.05 \times 1.1 \times 1.15 = 132.825$.
— *Après* 4 *ans on a,* $100 \times 1.05 \times 1.1 \times 1.15 \times 1.1 = 146.1075$.
Si on calcule la moyenne arithmétique des taux on obtient

$$\overline{x} = \frac{1.05 + 1.10 + 1.15 + 1.10}{4} = 1.10$$

Si on calcule la moyenne géeométrique des taux, on obtient

$$G = (1.05 \times 1.10 \times 1.15 \times 1.10)^{\frac{1}{4}} = 1.099431377.$$

Le bon taux moyen est bien G *et non* $\overline{x}$*, car si on applique* 4 *fois le taux moyen* G *aux* 100 *francs, on obtient*

$$100 \times G^4 = 100 \times 1.0994313774 = 146.1075.$$

3.1.4 La moyenne harmonique

Si $x_i \geq 0$ alors la moyenne harmonique, H, est l'inverse de la moyenne arithmétique des inverses des obervations :

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}.$$

On peut par exemple la moyenne harmonique pour les vitesses.

Exemple 3.1.2. *Un cycliste parcourt* 4 *étapes de* 100 *km. Les vitesses respectives pour ces étapes sont de* 10 *km/h,* 30 *km/h,* 40 *km/h,* 20 *km/h. Quelle a été sa vitesse moyenne ?*

Solution. *Il a parcouru la première étape en* $10h$*, la deuxième en* $3h20$ *la troisième en* $2h30$ *et la quatrième en* $5h$*. Il a donc parcouru le total des* $400km$ *en :*

$$10 + 3h20 + 2h30 + 5h = 20h50 = 20.8333h.$$

Sa vitesse moyenne est donc

$$Moy = \frac{400}{20.8333} = 19.2\,km$$

Si on calcule la moyenne arithmétique des vitesses, on obtient

$$\overline{x} = \frac{10+30+40+20}{4} = 25\,km/h.$$

Si on calcule la moyenne harmonique des vitesses, on obtient

$$H = \frac{4}{\frac{1}{10}+\frac{1}{30}+\frac{1}{40}+\frac{1}{20}} = 19.2\ km/h.$$

La moyenne harmonique est donc la manière appropriée de calculer la vitesse moyenne.

3.1.5 La médiane

La médiane M_e est la valeur du caractère pour laquelle la fréquence cumulée est égale à 0.5. Elle correspond donc au centre de la série statistique classée par ordre croissant ou à la valeur pour laquelle 50% des valeurs observées sont supérieures et 50% sont inférieures.

3.1.5.1 Caractère quantitatif discret

On procède ainsi après avoir rangé les n observations $x_1, x_2, \ldots, x_n$ par ordre croissant $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

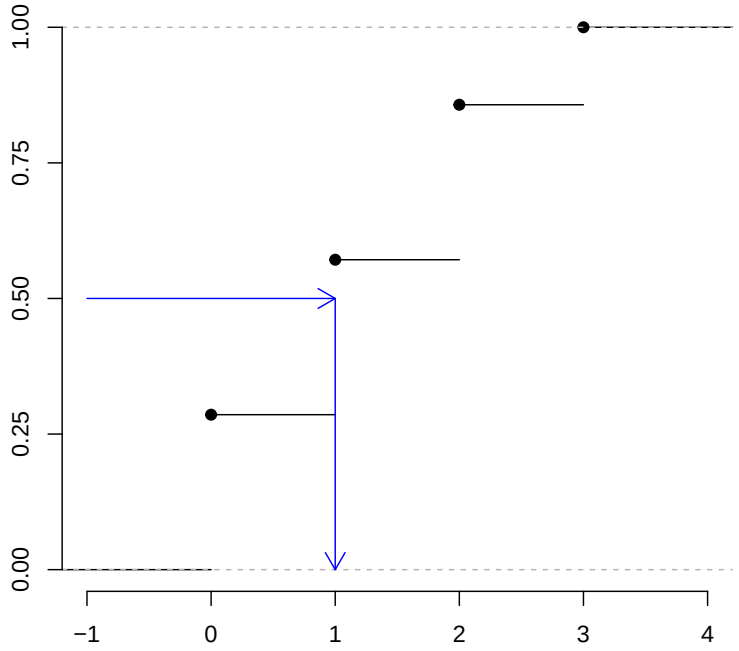
- si n est impair, alors $n = 2m+1$ et la médiane est la valeur $M_e = x_{(m+1)}$.

Exemple 3.1.3. *On considère la série statistique :*

$$3 \quad 2 \quad 1 \quad 0 \quad 0 \quad 1 \quad 2$$

On ordonne la série

$$0 \quad 0 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3$$



- si n est pair, alors $n = 2m$ et une médiane est une valeur quelconque entre $x_{(m)}$ et $x_{(m+1)}$; $(x_{(m)}, x_{(m+1)})$ est appelé intervalle médian. Dans ce cas, on prend souvent le milieu comme médiane, c'est à dire

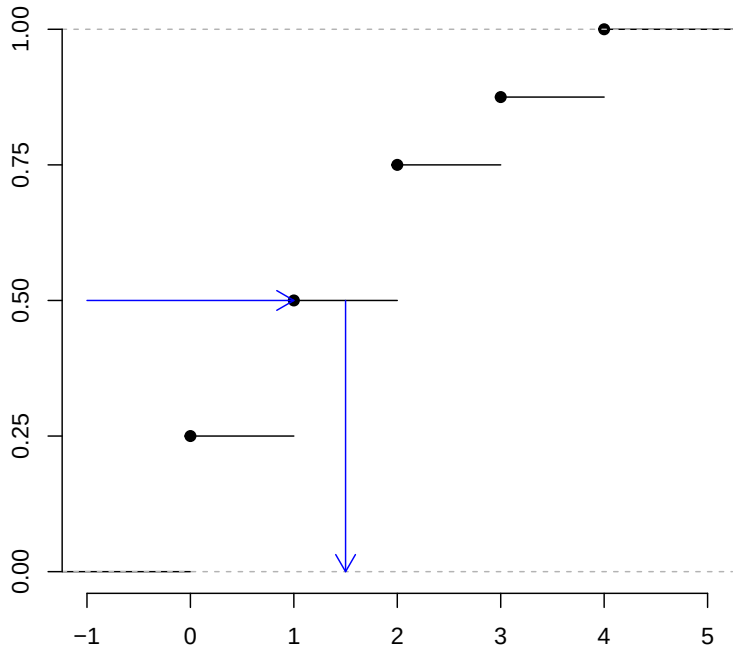
$$M_e = \frac{x_{(m)} + x_{(m+1)}}{2}.$$

Exemple 3.1.4. *On considère la série statistique :*

3 2 1 0 0 1 2 4

On ordonne la série

0 0 1 1 2 2 3 4



3.1.5.2 Caractère quantitatif continu

On utilisera la méthode de l'interpolation linéaire exposée ci-dessous.

3.1.5.3 Remarques

- La médiane peut être calculéee pour un caractère quantitatif et pour un caractère qualitatif ordinal.
- La médiane est plus robuste que la moyenne car elle n'est pas influencée par les valeurs extrêmes.
- La médiane est influencée par le nombre d'observations.

3.1.6 Les quantiles

Le quantile d'ordre α est la valeur x_α du caractère qui laisse une proportion α des observations en dessous et $1-\alpha$ des observations au dessus d'elle. Les fractiles sont les quantiles qui partitionnent les données triées en classes de taille égale. Les fractiles les plus utilisés sont les quartiles, les déciles et les centiles.

Les quartiles sont au nombre de trois.

- Le premier quartile Q_1 est le quantile d'ordre $\frac{1}{4}$; c'est la valeur du caractère telle qu'il ait **25%** des observations qui lui soient inférieures et **75%** supérieures.
- Le deuxième quartile Q_2 est le quantile d'ordre $\frac{1}{2}$, est la médiane.

- Le troisième quartile Q_3est la quantile d'ordre $\frac{3}{4}$; c'est la valeur du caractère telle que 75% des observations lui soient inférieures et 25% supérieures.

Les quartiles Q_1, Q_2 et Q_3 partagent la série ordonnée en quatre groupes de même effectif (25% chacun).

Remarque 3.1.1. *Un décile est l'une des neuf valeurs qui partagent la série ordonnée en* 10 *groupes de même effectif (*10% *chacun). Un centile est l'une des cent valeurs qui partagent la série ordonnée en* 100 *groupes de même effectif (*1% *chacun).*

Détermination pratique de la médiane
On utilise le tableau des effectifs cumulés ou des fréquences cumulées.

Caractère quantitatif discret : s'il existe une modalité x_j du caractère telle que $N_{j-1} < \alpha n \leq N_j$ ou $F_{j-1} < \alpha \leq F_j$ alors le quantile d'ordre α est x_j.
Caractère quantitatif continu : soit la première classe dont la fréquence empirique est supérieure ou égale à α. Notons là $C_i = [a_{i-1}, a_i[$ et appelons F_i sa fréquence cumulée. Si $F_i = \alpha$, le quantile est a_i. Dans le cas contraire, $F_i > \alpha$, considérons les points de coordonnées (a_{i-1}, F_{i-1}) et (a_i, F_i), F_{i-1} est la fréquence cumulée de la classe précédant C_i si elle existe, 0 sinon. La droite passant par ces deux points passe par un point d'ordonées α dont l'abscisse est x_α.

a_{i-1}	F_{i-1}
x_α	α
a_i	F_i

On tire x_α à partir de la formule suivante :

$$\frac{x_\alpha - a_{i-1}}{\alpha - F_{i-1}} = \frac{a_i - a_{i-1}}{F_i - F_{i-1}}.$$

Par suite

$$x_\alpha = a_{i-1} + (a_i - a_{i-1})\frac{\alpha - F_{i-1}}{F_i - F_{i-1}}.$$

Exemple 3.1.5. *Répartition des ouvriers selon leur salaire mensuel net (en milliers francs.*

Salaire	*Effectif*	*Fréquence (%)*	*Fréquence cumulées (%)*
[80,100[	*26*	*18.6*	*18.6*
[100,110[	*33*	*23.5*	*42.1*
[110,120[	*64*	*45.8*	*87.9*
[120,130[	*7*	*5.0*	*92.9*
[130,150[	*10*	*7.1*	*100*
Total	*140*	*100*	

100	18.6
Q_1	25
110	42.1

Par suite

$$Q_1 = 100 + (110 - 100)\frac{25 - 18.6}{42.1 - 18.6} = 102.72$$

110	42.1
Q_2	50
120	87.9

Par suite

$$Q_2 = 110 + (120-110)\frac{50-42.1}{87.9-42.1} = 111.72$$

110	42.1
Q_3	75
120	87.9

Par suite

$$Q_3 = 110 + (120-110)\frac{75-42.1}{87.9-42.1} = 117.18$$

3.1.7 Boîte à moustaches

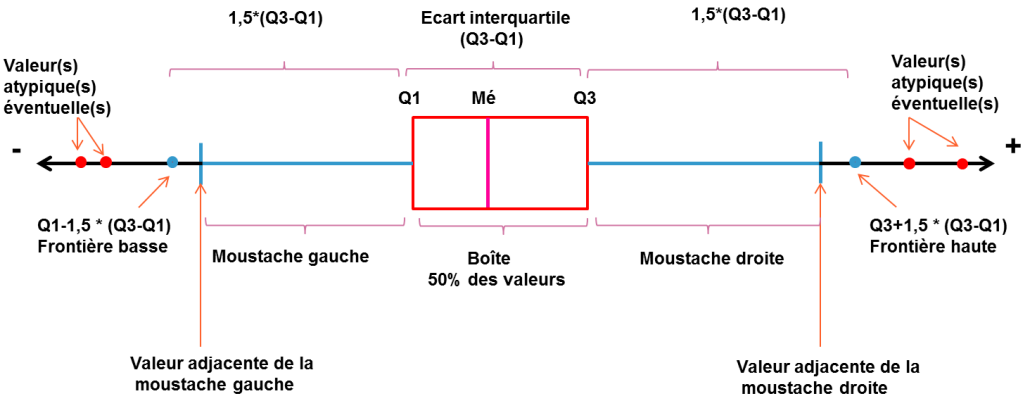
La boîte à moustaches ou boxplot est un diagramme qui permet de représenter la distribution d'un caractère. Ce diagramme est composé de :

- un rectangle qui s'étend du premier au troisième quartile ; le rectangle est divisé par une ligne correspondant à la médiane ;
- ce rectangle est complété par deux segments de droites ; pour les dessiner, on calcule d'abord les bornes

$$b^- = Q_1 - 1.5(Q_3 - Q_1)$$

$$b^+ = Q_3 + 1.5(Q_3 - Q_1).$$

Les valeurs au-délà des moustaches sont des valeurs hors norme éventuellement suspectes ou aberrantes mais pas nécessairement.



Ce diagramme est utilisé notamment pour comparer un même caractère dans deux ou plusieurs échantillons de tailles différentes.

3.2 Paramètres de dispersion

Exemple 3.2.1. *Deux groupes d'étudiants ont été observés selon la note obtenue en statistique descriptive :*

Groupe 1	*2*	*5*	*10*	*10*	*10*	*15*	*18*
Groupe 2	*8*	*9*	*10*	*10*	*10*	*11*	*12*

Pour le groupe 1 : $M_{01} = M_{e1} = \overline{X}_1 = 10$
Pour le groupe 2 : $M_{02} = M_{e2} = \overline{X}_2 = 10$.
On remarque que les deux séries présentent un même mode, une même médiane et une même moyenne. Cependant, leur distribution se fait d'une manière nettement différente. En effet, contrairement au groupe 1, les notes du groupe 2 ne s'écartent pas trop des valeurs centrales $(Me = \bar{X} = 10)$. *Ainsi, les indicateurs de tendance centrale peuvent s'avèrer insuffisant pour permettre à eux seuls de résumer et de comparer deux ou plusieurs séries statistiques, d'où la nécessité de calculer d'autres indicateurs dits de dispersion.*

Les paramètres de dispersion servent à préciser la variabilité de la série statistique, c'est à dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

3.2.1 L'étendue

On appelle étendue l'écart entre la plus grande valeur et la plus petite valeur. Posons

$$x_{min} = \min(x_1, \ldots, x_n) \quad x_{max} = \max(x_1, \ldots, x_n).$$

L'etendue est définie par

$$E = x_{max} - x_{min}.$$

Plus l'étendue est faible, plus la série est moins dispersée. L'inconvénient majeur de l'étendue est qu'il ne dépend que des valeurs extrêmes qui sont souvent exceptionnelles et aberrantes.

3.2.2 L'écart moyen absolu

Pour un caractère quantitatif discret dont les n observations sont rangées selon ses k modalités $x_1, \ldots, x_k$ d'effectifs respectifs $n_1, \ldots, n_k$, l'écart absolu moyen est le nombre

$$EMA = \frac{1}{n} \sum_{i=1}^{k} n_i \mid x_i - \bar{x}_n \mid .$$

Pour un caractère quantitatif continu dont les n observations ont été réparties dans k intervalles $([a_i, a_{i+1}[)_{i=1,\ldots,k}$, l'écart absolu moyen est le nombre

$$EMA = \frac{1}{n} \sum_{i=1}^{k} n_i \mid c_i - \bar{x}_n \mid,$$

où $c_i = \dfrac{a_i + a_{i+1}}{2}$ est le centre de la classe $[a_i, a_{i+1}[$.

Remarque 3.2.1. *On appelle écart absolu par rapport à la médiane* M_e *:*

$$EMA_1 = \frac{1}{n} \sum_{i=1}^{k} n_i \mid x_i - M_e \mid .$$

Cet indicateur de dispersion tient compte de tous les écarts entre les valeurs observées et la moyenne arithmétique. Son inconvénient est qu'il n'est pas commode pour le calcul algébrique vu la présence de l'expression de la valeur absolue. Une solution alternative consiste à considérer la moyenne des carrés des écarts et de calculer ensuite la racine carrée.

3.2.3 Variance, écart-type

Pour un caractère quantitatif discret dont les n observations sont rangées selon ses k modalités $x_1,\dots,x_k$ d'effectifs respectifs $n_1,\dots,n_k$,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2.$$

Pour un caractère quantitatif continu dont les n observations ont été réparties dans k intervalles $([a_i, a_{i+1}[)_{i=1,\dots,k}$, la variance est

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \bar{x})^2.$$

où $c_i = \dfrac{a_i + a_{i+1}}{2}$ est le centre de la classe $[a_i, a_{i+1}[$.

L'écart-type σ est la racine carrée de la variance.

La variance mesure la dispersion des valeurs autour de la moyenne. La variance est exprimée dans le carré de l'unité de mesure de la variable. C'est la raison pour laquelle on ne doit pas interpréter la variance mais plutôt sa racine carrée : l'écart-type. L'écart-type est utilisé comme un indicateur de la dispersion de la série statistique. Plus il est grand, plus la dispersion des observations autour de la moyenne de la variable est forte, plus la population est hétérogène.

3.2.4 L'écart inter-quartile

L'intervalle interquartile est l'intervalle $[Q_1, Q_3]$. L'écart interquartile est défini par

$$IQ = Q_3 - Q_1.$$

Nous avons 50% des observations qui se trouvent entre Q_1 et Q_3. Ainsi, 50% des observations s'étalent sur un intervalle de longueur égale à $Q_3 - Q_1$. Plus l'intervalle interquartiles est petit, plus la dispersion est faible et plus la population est homogène.
Cette quantité mesure la dispersion autour de la médiane. Plus IQ est grand, plus il existe des valeurs éloignées de la médiane.

3.2.5 Le Coefficient de variation

Le coefficient de variation CV est défini comme le rapport de l'écart-type à la moyenne :

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

C'est un nombre sans dimension qui mesure la proportion de la moyenne expliquée par l'écart-type. Le coefficient de variation permet de comparer deux ou plusieurs distributions exprimées dans des unités différentes et qui n'ont pas le même ordre de grandeur (les moyennes sont différentes). Le coefficient de variation est souvent exprimé en pourcentage. Plus le coefficient de variation est faible, plus la dispersion est faible et plus la population est homogène.

3.3 Les paramètres de concentration

La notion de concentration tient une place importante dans les études économiques ; on parle de concentration des entreprises, de concentration du pouvoir ou de la richesse, etc. L'étude de concentration ne s'applique qu'à des variables statistiques continues à valeurs positives et cumulables. Il est clair qu'elle ne peut s'appliquer à des ensembles d'individus classés selon l'âge, la taille ou le poids, parce que la somme des âges par exemple d'une population n'a pas de signification. Elle a pour but de mesurer les inégalités de répartition d'une masse totale.

3.3.1 La médiale

La médiale est la valeur du caractère qui partage la valeur totale ou la masse totale en deux parties égales. La médiale se détermine par interpolation linéaire sur les valeurs globales relatives cumulées croissantes.
Soit X un caractère continu dont les observations sont rangées dans les classes $[a_{i-1}, a_i[$, $k = 1, \ldots, k$. Soit n_i l'effectif de la classe $[a_{i-1}, a_i[$ et $c_i = \dfrac{a_{i-1} + a_i}{2}$ son centre.

- On appelle $n_i c_i$ la valeur globale (v.g.) associée à la classe $[a_{i-1}, a_i[$.
- $\sum_{i=1}^{n} n_i c_i$ est appelée valeur totale ou masse totale du caractère étudié.
- $q_i = \dfrac{n_i c_i}{\sum_{i=1}^{n} n_i c_i}$ est la valeur globale relative (v.g.r.) associée à la classe $[a_{i-1}, a_i[$. q_i désigne la part, dans la valeur totale, detenur par les individus ayant une valeur du caractère appartenant à la classe $[a_{i-1}, a_i[$.
- $V(a_i) = V_i = \sum_{j=1}^{i} q_j$ est appelée valeur globale relative cumulée croissante (v.g.r.c.c). Elle indique la part, dans la valeur totale, detenue par les individus ayant une valeur du caractère appartenant à la classe $[a_{i-1}, a_i[$.

La médiale M vérifie $V(M) = 0.5$. La détermination de la médiale se fait en deux étapes :

1. Soit la première classe $[a_{i-1}, a_i[$ dont la valeur globale relative cumulée croissante V_i est supérieure ou égale à 0.5. Si $V_i = 0.5$ alors la médiale est $M = F_i$. Sinon, nous avons $V_{i-1} < 0.5 < V_i$.
2. Par interpolation linéaire, on calcule la valeur de la médiale :

a_{i-1}	V_{i-1}
M	0.5
a_i	V_i

$$\frac{a_i - a_{i-1}}{V_i - V_{i-1}} = \frac{M - a_{i-1}}{0.5 - V_{i-1}} \Leftrightarrow M = a_{i-1} + \frac{0.5 - V_{i-1}}{V_i - V_{i-1}}(a_i - a_{i-1}).$$

Exemple 3.3.1. *La médiale est le niveau de salaire qui divise en deux la masse salariale : les salaires inférieurs à la médiale représentent la moitié de la masse salariale et ceux supérieurs à la médiale représentent aussi la moitié de la masse salariale.*

Modalié	Effectif	Centre	Valeur globale	Valeur gloable relative	Valeur gloable relative cumulée
$[a_0,a_1[$	n_1	$c_1=\frac{a_0+a_1}{2}$	n_1c_1	$q_1=\frac{n_1c_1}{\sum_{i=1}^{k}n_1c_1}$	$V_1=q_1$
$[a_1,a_2[$	n_2	$c_2=\frac{a_1+a_2}{2}$	n_2c_2	$q_2=\frac{n_2c_2}{\sum_{i=1}^{k}n_ic_i}$	$V_2=q_1+q_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_{i-1},a_i[$	n_i	$c_i=\frac{a_{i-1}+a_i}{2}$	n_ic_i	$q_i=\frac{n_ic_i}{\sum_{i=1}^{k}n_ic_i}$	$V_i=\sum_{j=1}^{i}q_j$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_{k-1},a_k[$	n_k	$c_k=\frac{a_0+a_1}{2}$	n_kc_k	$q_k=\frac{n_kc_k}{\sum_{i=1}^{k}n_ic_i}$	$V_k=\sum_{j=1}^{k}q_j=1$
Total	n				

TABLE 3.1 – Tableau de calcul de la médiale

Classe de salaire (en milliers francs)	*Effectif*	*Centre de classe*	*Masse salariale*	*Valeur globale relative*	*Valeur glaobale relative cumulée*
[80,100[	*26*	*90*	*2340*	*15.15*	*15.15*
[100,110[	*33*	*105*	*3465*	*22.44*	*37.59*
[110,120[	*64*	*115*	*7360*	*47.67*	*85.26*
[120,130[	*7*	*125*	*875*	*5.67*	*90.93*
[130,150[	*10*	*140*	*1400*	*9.07*	*100*
Total	*140*		*15440*		

110	37.59
M	50
120	85.26

$$M=110+(120-110)\times\frac{50-37.59}{65.26-37.59}.$$

3.3.2 L'écart entre médiane et médiale

On appelle écart médiale-médiane d'une série statistique, le nombre défini par :

$$\Delta M=M-M_e.$$

Cet écart nous fournit un premier renseignement sur la concentration d'une distribution statistique.

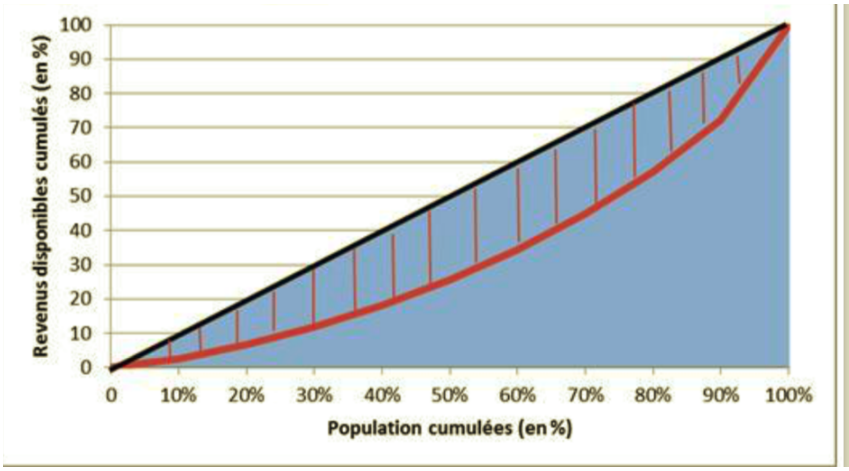
- Si $\Delta M=0\Leftrightarrow M=M_e$ alors alors la concentration est nulle et la répartition de la valeur totale est parfaitement égalitaire.
- Si $\Delta M\neq 0$ alors la répartition de la valeur totale n'est pas égalitaire. Cependant, aucune information sur l'intensité de cette inégalité ne peut être avancée.
- Pour comparer la concentration de deux ou plusieurs séries statistiques, on peut utiliser le rapport $\frac{\Delta M}{E}$. La concentration d'une série est d'autant plus forte que le rapport est élevé. (E représente l'etendue de la série).

3.3.3 La courbe de Lorenz

La courbe de Lorenz est obtenue en reliant, par des segments de droites, les points de coordonnées (F_i, V_i), $i = 0, \ldots, k$ avec $(F_0, V_0) = (0, 0)$. Plus la courbe de Lorenz s'éloigne de la première bissectrice, plus la concentration est forte et plus la répartition est inégalitaire.

3.3.4 L'indice de Gini

L'indice de Gini ou coefficient de Gini est le double de l'aire comprise entrelacourbe de concentration et la première bissectrice (zone hachuré en rouge). Il mesure le niveau d'inégalité de la répartition d'une variable dans la population. L'indice de Gini est compris entre 0 (égalité parfaite) et 1 (inégalité parfaite). Une baisse de l'indice de Gini indique une diminution globale des inégalités. A l'inverse, une élévation de l'indice reflète une augmentation globale des inégalités.



L'indice de Gini I est alors

$$I = 2S = 1 - \sum_{i=1}^{k} \frac{n_i}{n}(V_i + V_{i-1}).$$

3.4 Paramètres de forme

Les paramètres de forme permettent d'avoir une idée staisfaisante et plus précise sur la forme de la distribution. On distingue les coefficients d'asymétrie et les coefficients d'aplatissement.
Une distribution est dite symétrique si les observations également dispersées de part et d'autre de la valeur centrale. Dans le cas contraire, la distribution est dite asymétrique ou dissymétrique.

3.4.1 Moments

Pour un caractère quantitatif discret dont les n observations sont rangées selon ses k modalités $x_1, \ldots, x_k$ d'effectifs respectifs $n_1, \ldots, n_k$, le moment centré d'ordre r est défini par

$$\mu_r = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^r.$$

Pour un caractère quantitatif continu dont les n observations ont été réparties dans k intervalles $([a_i, a_{i+1}[)_{i=1,\dots,k}$, le moment centré d'ordre r est défini par

$$\mu_r = \frac{1}{n}\sum_{i=1}^{k} n_i (c_i - \bar{x})^r,$$

où $c_i = \dfrac{a_i + a_{i+1}}{2}$ est le centre de la classe $[a_i, a_{i+1}[$.

Remarque 3.4.1. $\mu_0 = 1$, $\mu_1 = 0$ *et* μ_2 *est la variance.*

3.4.2 Asymétrie

Le coefficient d'asymétrie de Pearson

Dans une distribution faiblement asymétrique, c'est la position du mode par rapport à la moyenne (ou à la médiane) qui caractérise l'asymétrie. Le coefficient d'asymétrie de Pearson est défini par :

$$s = \frac{\bar{X} - M_0}{\sigma}.$$

Le coefficient d'asymétrie de Fisher

Le Coefficient d'asymétrie de Fisher permet de quantifier le degré de déviation de la forme de la distribution par rapport à une distribution symétrique. Il est défini par

$$s = \frac{\mu_3}{\mu_2^{3/2}}.$$

Le coefficient d'asymétrie de Yule

On compare ici létalement de la courbe de distribution à gauche de la médiane et l'étalement à droite et à rapporter leur diiférence à leur somme. Le coefficient d'asymétrie de Yule est défini par :

$$s = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}.$$

Interprétation

Quelque soit la formule adoptée, nous avons l'interprétation suivante. Ces coefficients n'ont d'intérât que dans la mesure où ils permettent de comparer les formes de deux ou plusieurs distributions ; bien entendu, les comparaison ne sont valables que si la même formule est retenue pour les diverses distributions.

1. $s = 0$ indique une distribution parfaitement symétrique. Dans ce cas $M_e = M_0 = \bar{x}$.

2. $s > 0$ indique une distribution unimodale étalée vers la droite.
   Dans ce cas $M_0 < M_e < \bar{x}$

3. $s < 0$ indique une distribution unimodale étalée vers la gauche.
   Dans ce cas $\bar{x} < M_e < M_0$

FIGURE 3.1 – $s = 0$: la distribution symétrique.

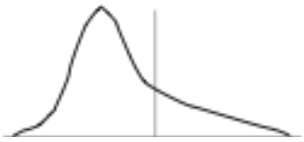


FIGURE 3.2 – $s > 0$: la distribution étalée à droite.

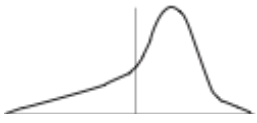


FIGURE 3.3 – $s < 0$: la distribution étalée à gauche.

FIGURE 3.4 – $\gamma = 0$: la distribution est normale.

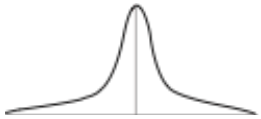


FIGURE 3.5 – $\gamma > 0$: la distribution est aigue.

3.4.3 L'aplatissement

Le coefficient d'aplatissement (kurtosis) permet de mesurer le relief ou la platitude d'une courbe issue d'une distribution de fréquences. On compare la courbe de fréquence de la distribution à la courbe de fréquence de la distribution normale considérée comme la distribution idéale. On fait apparaître ainsi l'aplatissement ou l'allongement au voisinage du mode. Quand la courbe est plus aplatie que la courbe normale, on dit qu'elle est platicurtique ; quand elle est plus aigue, on dit qu'elle est leptocurtique ; une courbe normale est dite mésocurtique. Le coefficient d'aplatissement de Fisher est :

$$\gamma = \frac{\mu_4}{\mu_2^2} - 3 \quad \mu_2 \neq 0.$$

1. $\gamma = 0$: la distribution est normale
2. $\gamma > 0$: la distribution est aigue.
3. $\gamma < 0$: la distribution est aplatie.

Remarque 3.4.2. *Les coefficients de kurtosis et de skewness peuvent être utilisés pour s'assurer que les variables suivent une distribution normale. On estime que le coefficient de symétrie ou skewness doit être inférieur à* 1 *et le coefficient d'aplatissement ou kurtosis doit être inférieur à* 1.5 *pour considérer que la variable suit bien une loi normale.*

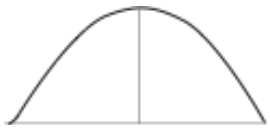


FIGURE 3.6 – $\gamma < 0$: la distribution est aplatie.

Chapitre 4

Statistiques à deux variables

4.1 Introduction

La question centrale de ce chapitre est relative aux statistiques bivariées (deux variables). Comment juger de l'intensité de la dépendance statistique esntre deux variables ?

Répondre statistiquement à cette question dépend de la nature des deux variables étudiées. Trois combinaisons sont possibles :

— Deux variables qualitatives :
— Une variable qualitative et une variable quantitative :
— Deux variables quantitatives

L'analyse dans ce cas n'est plus univariée mais bien bivariée. On analyse de manière simultanée les caractéristiques des individus suivant deux variables.

4.2 Généralités

4.2.1 Distribution conjointe

Soit une population comprenant n individus pour chacun desquels on a fait une observation concernant simultanément les caractères X et Y. Le caractère X comporte les k modalités $X_1, \cdots, X_k$ et le caractère Y, les l modalités $Y_1, \cdots, Y_l$. L'opération préliminaire de mise en ordre des observations va consister à classer chacun des n individus dans les $k \times l$ sous-ensembles définis par le croisement des caractères X et Y. A chacun des sous-ensembles correspond une case du tableau statistique à double entrée où figurent en ligne les modalités de X et en colonne les modalités de Y (tableau à k lignes et l colonnes). Ce tableau est appelé tableau de contingence.
On note n_{ij} l'effectif des individus présentant à la fois la modalité X_i et la modalité Y_j. La fréquence des individus présentant à la fois la modalité X_i et la modalité Y_j est

$$f_{ij} = \frac{n_{ij}}{n}.$$

La distribution conjointe des caractères X et Y est donnée par le tableau de contingence :

X \ Y	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_l
X_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1l}
X_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2l}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{il}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kl}

Exemple 4.2.1. *Deux variables qualitatives : répartition de* 22 *personnes selon le genre et le statut d'activité :*

Genre \ *Statut*	*Actifs occupés*	*Chômeurs*	*Inactifs*	*Total*
Masculin	*5*	*5*	*1*	*11*
Féminin	*4*	*3*	*4*	*11*
Total	*9*	*8*	*5*	*22*

Exemple 4.2.2. *Deux variables quantitatives continues : répartition de* 19 *adolescents selon la taille et le poids.*

Poids \ *Taille*	[20,40[	[40,60[	[60,80]
[120,140[	*1*	*0*	*0*
[140,160[	*6*	*4*	*0*
[140,160[	*0*	*6*	*2*

Exemple 4.2.3. *Une variable quantitative continue et une variable qualitative : Répartition d'un groupe de* 50 *personnes réparties par âge et par genre, tous âgés de moins de* 45 *ans.*

Age \ *Genre*	*Homme*	*Femme*
[0,18[	*10*	*20*
[18,45[	*5*	*15*

4.2.2 Distributions marginales

Le nombre d'individus présentant la modalité X_i du caractère X $n_{i\bullet}$ est

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

La fréquence de la modalité X_i est donnée par

$$f_{i\bullet} = \frac{n_{i\bullet}}{n}.$$

Le nombre d'individus présentant la modalité Y_j du caractère Y est

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

La fréquence de la modalité Y_j est donnée par

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}.$$

Nous avons

$$n = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j}$$

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

X \ Y	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_l	Total
X_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet l}$	n

X \ Y	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_l	Total
X_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1l}	$f_{1\bullet}$
X_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2l}	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{il}	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	f_{k1}	f_{k2}	$\cdots$	f_{kj}	$\cdots$	f_{kl}	$f_{k\bullet}$
Total	$f_{\bullet 1}$	$f_{\bullet 2}$	$\cdots$	$f_{\bullet j}$	$\cdots$	$f_{\bullet l}$	1

La distribution marginale de X est donnée par le tableau ci-dessous :

Modalités de X	Effectif	Fréquence
X_1	$n_{1\bullet}$	$f_{1\bullet}$
X_2	$n_{2\bullet}$	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$
X_i	$n_{i\bullet}$	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$
X_k	$n_{k\bullet}$	$f_{k\bullet}$
total	n	1

La distribution marginale de Y est donnée par le tableau ci-dessous :

Modalités de Y	Effectif	Fréquence
Y_1	$n_{\bullet 1}$	$f_{\bullet 1}$
Y_2	$n_{\bullet 2}$	$f_{\bullet 2}$
$\vdots$	$\vdots$	$\vdots$
Y_j	$n_{\bullet j}$	$f_{\bullet j}$
$\vdots$	$\vdots$	$\vdots$
Y_l	$n_{\bullet l}$	$f_{\bullet l}$
total	n	1

4.2.3 Distributions conditionnelles

Les distributions conditionnelles s'obtiennent en fixant la valeur d'une des deux variables. La distribution conditionnelle de Y sachant $X = X_i$ est donnée par :

Y$\mid X = X_i$	Y_1	$\cdots$	Y_j	$\cdots$	Y_l	Total
Effectif	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	$n_{i\bullet}$

Remarque 4.2.1. *Nous pouvons ainsi définir k distributions conditionnelles de Y.*

La distribution conditionnelle de X sachant $Y = Y_j$ est donnée par

$X \mid Y = Y_j$	X_1	$\cdots$	X_i	$\cdots$	X_k	Total
Fréquence	n_{1j}	$\cdots$	n_{ij}	$\cdots$	n_{kj}	$n_{\bullet j}$

Remarque 4.2.2. *Nous pouvons aussi définir l distributions conditionnelles de X.*

Exemple 4.2.4. *Deux variables qualitatives : répartition de 22 personnes selon le genre et le statut d'activité :*

Statut / *Genre*	*Actifs occupés*	*Chômeurs*	*Inactifs*	*Total*
Masculin	*5*	*5*	*1*	*11*
Féminin	*4*	*3*	*4*	*11*
Total	*9*	*8*	*5*	*22*

Statut | *Genre=Masculin*	*Actifs occupé*	*Chomeurs*	*Inactifs*	*Total*
Effectif	*5*	*5*	1	$n_{1\bullet} = 11$
frequence	$f_{1/1} = \frac{n_{11}}{n_{1\bullet}} = 0.4545$	$f_{2/1} = \frac{n_{12}}{n_{1\bullet}} = 0.4545$	$f_{3/1} = \frac{n_{13}}{n_{1\bullet}} = 0.091$	*1*

4.2.4 Indépendance

On dit que les caractères X et Y sont satistiquement indépendants dans l'ensemble des n individus considérés si toutes les distributions conditionnelles de X sont identiques à la distribution marginale en X.

$$\text{Indépendance entre } X \text{ et } Y \Longleftrightarrow \text{Pour tous } (i,j), \quad f_{i/j} = f_{i\bullet}$$

Puisque

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

alors

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

Ainsi, nous obtenons

$$\begin{aligned}\text{Indépendance entre } X \text{ et } Y &\iff \text{Pour tous } (i,j), \quad f_{ij} = f_{i\bullet} f_{\bullet j} \\ &\iff \text{Pour tous } (i,j), \quad n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n} \\ &\iff \text{Pour tous } (i,j), \quad n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n} = 0.\end{aligned}$$

Par symétrie :

$$\text{Indépendance entre } X \text{ et } Y \iff \text{Pour tous } (i,j), \quad f_{j/i} = f_{\bullet j}$$

Lorsque deux variables dépendent statistiquement l'une de l'autre, on cherche à évaluer l'intensité de leur liaison et, dans le cas de deux variables quantitatives, on examine si on peut les considérer liées par une relation linéaire.

4.3 Liaison entre deux caractères qualitatifs

4.3.1 Mesure de l'intensité de la liaison

L'intensité de la liaison entre deux caractères qualitatifs est mesurée par

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}.$$

Le χ^2 est toujours positif ou nul.

Exemple 4.3.1. *Prenons $k = 2$ et $l = 3$*

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}} \\ &= \sum_{i=1}^{2} \left(\frac{\left(n_{i1} - \frac{n_{i\bullet} n_{\bullet 1}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet 1}}{n}} + \frac{\left(n_{i2} - \frac{n_{i\bullet} n_{\bullet 2}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet 2}}{n}} + \frac{\left(n_{i3} - \frac{n_{i\bullet} n_{\bullet 3}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet 3}}{n}} \right) \\ &= \frac{\left(n_{11} - \frac{n_{1\bullet} n_{\bullet 1}}{n}\right)^2}{\frac{n_{1\bullet} n_{\bullet 1}}{n}} + \frac{\left(n_{12} - \frac{n_{1\bullet} n_{\bullet 2}}{n}\right)^2}{\frac{n_{1\bullet} n_{\bullet 2}}{n}} + \frac{\left(n_{13} - \frac{n_{1\bullet} n_{\bullet 3}}{n}\right)^2}{\frac{n_{1\bullet} n_{\bullet 3}}{n}} + \frac{\left(n_{21} - \frac{n_{2\bullet} n_{\bullet 1}}{n}\right)^2}{\frac{n_{2\bullet} n_{\bullet 1}}{n}} + \frac{\left(n_{22} - \frac{n_{2\bullet} n_{\bullet 2}}{n}\right)^2}{\frac{n_{2\bullet} n_{\bullet 2}}{n}} + \frac{\left(n_{23} - \frac{n_{2\bullet} n_{\bullet 3}}{n}\right)^2}{\frac{n_{2\bullet} n_{\bullet 3}}{n}}\end{aligned}$$

On sait que :

$$X \text{ et } Y \text{ sont indépendants} \iff f_{ij} = f_{i\bullet} \times f_{\bullet j} \;\; i = 1, \ldots, k, \; j = 1, \ldots, l.$$

Ainsi, nous avons

$$\begin{aligned}f_{ij} = f_{i\bullet} \times f_{\bullet j} &\iff \frac{n_{ij}}{n} = \frac{n_{\bullet j}}{n} \times \frac{n_{i\bullet}}{n} \\ &\iff n_{ij} = \frac{n_{\bullet j} \times n_{i\bullet}}{n}\end{aligned}$$

De ce fait on a $\chi^2 = 0$ si et seulement si $n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$. La quantité χ^2 mesure l'écart entre les effectifs observés n_{ij} et ceux attendus $\frac{n_{i\bullet} \times n_{\bullet j}}{n}$ sous l'hypothèse d'indépendance. On dira que X et Y ne sont pas indépendants si χ^2 est trop grand.

4.3.2 Coefficient de Cramer

Le coefficient de Cramer est défini par

$$C = \sqrt{\frac{\chi^2}{n \times \min(k-1, l-1)}}.$$

Nous avons $0 \leq C \leq 1$. Si $C \approx 0$, les deux caractères sont indépendants. Si $C = 1$, on parle de dépendance entre X et Y.

Exemple 4.3.2. *Prenons $k = 2$ et $l = 3$ Le coefficient de Cramer est défini par*

$$C = \sqrt{\frac{\chi^2}{n \times \min(2-1, 3-1)}} = \sqrt{\frac{\chi^2}{n \times \min(1,2)}} = \sqrt{\frac{\chi^2}{n}}.$$

4.3.3 Exercices

4.3.3.1 Exercice 1

Nous voulons étudier la liaison entre le type de musique X et l'âge Y. X a trois modalités (chansons, jazz, classique) et Y a quatre modalités (jeunes, adulte femme, adulte homme, vieux). Voici le tableau de contingence :

X \ Y	Jeunes	Adulte femme	Adulte homme	Vieux	Total
Chansons	69	172	133	27	401
Jazz	41	84	118	11	254
Classique	18	127	157	43	345
Total	128	383	408	81	1000

Etudions la liaison entre X et Y.

Nous avons

$$\chi^2 = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}} = 52.9138.$$

Le coefficient de Cramer est

$$C = \sqrt{\frac{\chi^2}{1000 \times \min(2,3)}} = \sqrt{\frac{\chi^2}{2000}} \approx 0.16.$$

La dépendance entre X et Y est très faible.

4.3.3.2 Exercice 2

Un site internet reçoit 113 457 visiteurs durant un mois. On désigne par X le navigateur internet utilisé et Y le système d'exploitation utilisé.

X \ Y	Windows	Mac	Linux
Chrome	14103	1186	427
Firefox	30853	4392	3234
Internet explorer	47389	23	0
Safari	668	6416	0
Autres	2974	40	1752

1. Identifier la population, sa taille ainsi que les variables étudiées en précisant leur type.
2. Quelle est la proportion de visiteurs sous Windows ?
3. Quelle proportion de visiteurs utilisent le navigateur Safari ?
4. Parmi les utilisateurs de Mac, quelle proportion utilise Chrome ?
5. Parmi les utilisateurs de Safari, quelle proportion est sous Windows ?
6. Représenter graphiquement la distribution des proportions par Navigateur pour chaque système d'exploitation. Les variables X et Y sont-elles indépendantes ?

4.4 Liaison entre deux caractères quantitatifs

4.4.1 Représentation graphique : nuage de points.

On suppose que les deux caractères X et Y sont quantitatifs. Pour chaque individu i, on connaît le couple de valeurs (X_i, Y_i) qui lui est attaché. Sur un graphique à axes de coordonnées rectangulaires, nous pouvons représenter chaque élément, par un point d'abscisse X_i et d'ordonnée Y_i. Ce graphique est appelé graphique de corrélation ou nuage de points. Schématiquement, le nuage peut revêtir trois aspects :

1. Les points représentatifs sont distribués sur toute la surface du graphique, à peu près comme s'ils avaient été placés au hasard. C'est le signe qu'il n'y a aucun lien entre les deux variables X et Y : on dit qu'elles sont indépendantes ;
2. Les points représentatifs sont , au contraire rangés le long d'une courbe (droite, arc de cercle,...). Une loi rigoureuse préside alors aux relations entre les deux variables. A chaque valeur de X correspond une seule valeur de Y. On dit qu'il y a liaison fonctionnelle entre Y et X
3. La plupart des phénomènes identifiés à des distributions à deux variables se trouvent entre ces deux extrèmes. Les points représentatifs se distribuent dans une région privilégiée du dessin. Moins le nuage de points a d'épaisseur et plus on se trouve proche de la liaison fonctionnelle : on dit qu'il y a une forte corrélation entre les deux variables. Inversement, plus le nuage de points s'étale, moins ses limites sont précises, plus on est proche de l'indépendance : la corrélation est faible.

4.4.2 Covariance, coefficient de correlation linéaire

La covariance entre les caractères X et Y est défini par

$$\begin{aligned} Cov(X,Y) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}_n \overline{Y}_n. \end{aligned}$$

La covariance est un indice symétrique, c'est à dire, $Cov(X,Y) = Cov(Y,X)$ et peut prendre toute valeur (négative, nulle ou positive).
Le coefficient de correlation linéaire entre les caractères X et Y est défini par

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

où σ_X et σ_Y les écart-types respectifs de X et σ_Y, sont définis

$$\sigma_X = \left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \right)^{1/2} \qquad \sigma_Y = \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2 \right)^{1/2}.$$

Nous avons :

1. $-1 \leq r_{XY} \leq 1$.
2. Si $r_{XY} > 0$ alors les deux variables évoluent dans le même sens.
3. Si $r_{XY} < 0$ alors les deux variables n'évoluent pas dans le même sens.
4. $|r_{XY}| = 1 \Longleftrightarrow$ les n points (X_i, Y_i) sont alignés.
5. $r_{XY} = 0 \Longleftrightarrow$ Pas de liaison linéaire, mais possibilité d'une liaison d'un autre type.
6. X et Y indépendantes$\Longrightarrow r_{XY} = 0$.

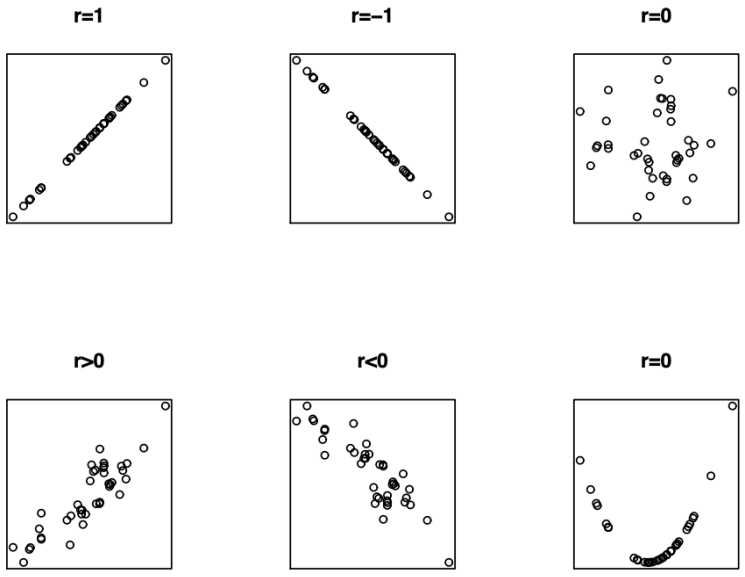


FIGURE 4.1 – Exemples de nuage de points et coefficients de corrélation

Remarque 4.4.1.

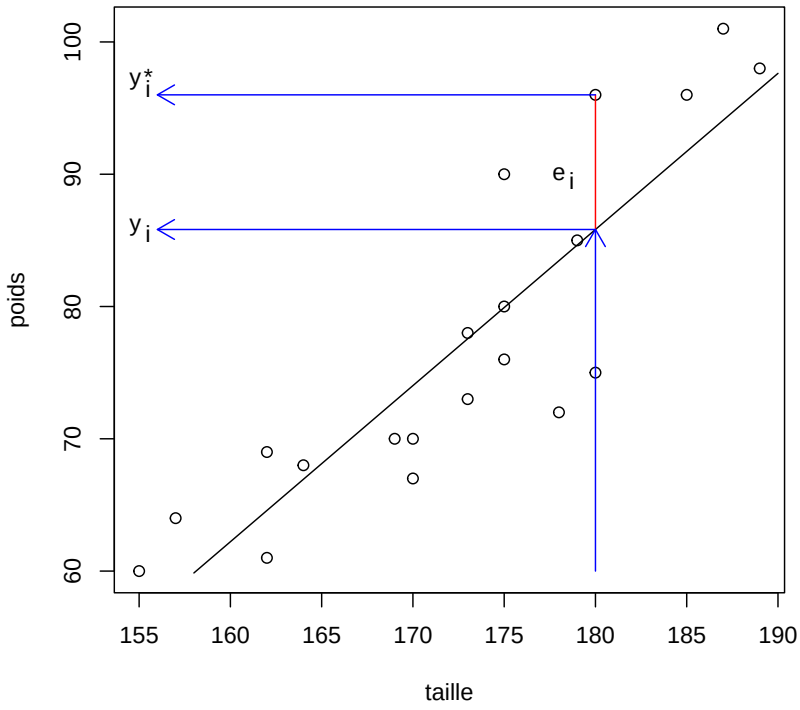
- *La covariance dépend des unités de mesure dans lesquelles sont exprimées X et Y. Le coefficient de corrélation est un indice de liaison sans unité.*
- *La covariance et le coefficient de corrélation ne permettent de mettre en évidence qu'une relation linéaire entre X et Y.*
- *Si deux variables sont statistiquement indépendantes (aucun lien), la corrélation est nulle, mais l'inverse est faux : il peut exister un lien autre que linéaire entre elles.*

4.4.3 Regression linéaire

Si $|r_{XY}| \simeq 1$, on peut supposer que X est cause de Y. Il est naturel de chercher, dans un ensemble donné de fonctions, la fonction de X approchant Y "le mieux possible" au sens d'un certain critère. On dit que l'on fait la regression de Y sur X. Si l'on choisit pour ensemble de fonctions celui des fonctions affines du type $(aX+b)$, on parle de regression linéaire. C'est le choix que l'on fait le plus fréquemment dans la pratique, le critère le plus usuel étant celui des moindres carrés.

Le critère des moindres carrés. Il consiste à minimiser la quantité

$$S(a,b) = \sum_{i=1}^{n} [Y_i - (aX_i + b)]^2.$$



Solution. La minimisation de S en a et b fournit la solution suivante :

$$a = \frac{Cov(X,Y)}{\sigma_X^2} \qquad b = \bar{y} - a\bar{x}.$$

La droite d'équation $y = ax + b$ est appelée droite de régression de Y sur X. Elle passe par le point $(\overline{X}_n, \overline{Y}_n)$.

4.4.4 Exemple Taux de cholestérol en fonction de l'âge

Sur un échantillon de 10 sujets d'âges différents, on a recueilli les données expérimentales suivant :

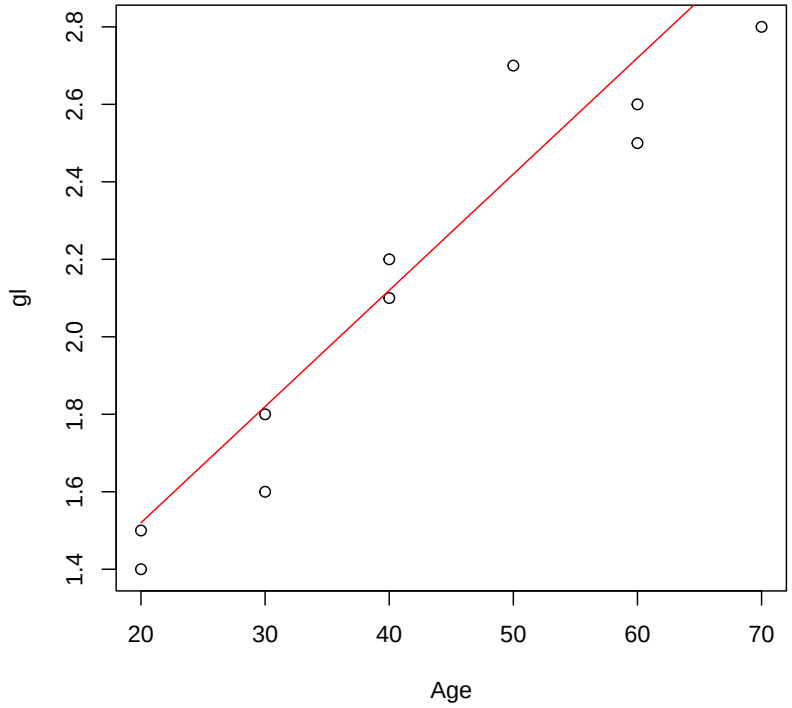
- âge en année

- la concentration sanguine du cholestérol (en g/L).

Age (X_i)	30	60	40	20	50	30	40	20	70	60
gl (Y_i)	1.6	2.5	2.2	1.4	2.7	1.8	2.1	1.5	2.8	2.6

Le taux de cholestérol est-il lié à l'âge ? La relation fonctionnelle est-elle linéaire ? Peut-on prévoir le taux de cholestérol attendu à 35 ans, 75 ans ?

1. Représentation du nuage de points.



Les points sont rangés le long d'une droite. On peut donc supposer l'existence d'une relation linéaire entre l'age et le taux de cholesterol.

2. Le coefficient de corrélation est donné par :

$$r_{XY} = \frac{\sum_{i=1}^{10} x_i y_i - 12\overline{x}_{12}\overline{y}_{12}}{\sqrt{\sum_{i=1}^{10} x_i^2 - 12(\overline{x}_{12})^2}\sqrt{\sum_{i=1}^{10} y_i^2 - 12(\overline{y}_{12})^2}} \approx 0.95$$

Le coefficient de corrélation est positif. Ce qui signifie que l'age et le taux de cholesterol évolue dans le même sens. De plus ils sont fortement corrélés ; ce qui confirme la relation linéaire entre l'age et le taux de cholesterol.

3. Estimation des paramètres

$$a = \frac{\sum_{i=1}^{10} x_i y_i - 12\overline{x}_{12}\overline{y}_{12}}{\sum_{i=1}^{10} x_i^2 - 12(\overline{x}_{12})^2} = 0.03$$

$$b = \overline{y}_{12} - \hat{a}\overline{x}_{12} = 0.92$$

4. La droite de regression est
$$gl = 0.03 * age + 0.92$$
. La droite de régression est en rouge.
5. Prévisions A 35 ans le taux de cholestérol prédit est $gl = 0.03 * 35 + 0.92 = 1.97$
A 75 ans le taux de cholestérol prédit est $gl = 0.03 * 75 + 0.92 = 3.17$

4.5 Caractère quantitatif et caractère qualitatif

4.5.1 Rapport de correlation

Soient n observations portant simultanément sur un caractère qualitatif X à k modalités et sur un caractère quantitatif Y. Les observations du caractère quantitatif Y se répartissent dans les k modalités de X. Nous notons $n_{i\bullet}$ le nombre d'observations de Y relatifs à la i-ème modalité de X, Y_{ij} la j-ème mesure de Y pour la i-ème modalité de X et $\overline{Y}_i$ la moyenne des observations dans la i-ème modalité

$$\overline{Y}_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{n_{i\bullet}} Y_{ij}.$$

La moyenne des observations de Y dans la popultion entière est

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} n_{i\bullet} \overline{Y}_i.$$

On définit :

- la variance intra-groupe
$$V_{intra} = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \sigma_i^2$$
avec
$$\sigma_i^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{n_{i\bullet}} (Y_{ij} - \overline{Y}_i)^2$$
- la variance inter-groupe
$$V_{inter} = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} (\overline{Y}_i - \overline{Y}_n)^2$$

Formule de décomposistion de la variance totale σ^2 :

$$\sigma^2 = V_{intra} + V_{inter}.$$

Le rapport de corrélation est défini par

$$\eta^2_{Y|X} = \frac{V_{inter}}{\sigma^2}$$

$\eta^2_{X|Y}$ est un nombre compris entre 0 et 1.

- $\eta^2_{X|Y} = 0 \Rightarrow V_{inter} = 0 \Rightarrow \overline{Y}_i = \overline{Y}_n$. Ce qui signifie que les moyennes de Y sont les mêmes dans toutes les modalités de X. En moyenne, les données ne diffèrent pas selon qu'elles se trouvent dans telle ou telle modalité de X.

- $\eta^2_{X|Y} = 1 \Rightarrow V_{intra} = 0 \Rightarrow Y_{ij} = \overline{Y}_i$. Les données diffèrent d'un groupe à l'autre mais à l'interieur même de chaque groupe, il n'y a aucune variabilité.

Remarque 4.5.1. *Si $\eta^2_{X|Y}$ est proche de 1, c'est que le caractère X explique une grande partie de la variabilité des données alors que si sa valeur est proche de 0, elle n'en explique que très peu.*

4.5.2 Exemple

Liaison entre le sexe (caractère X) et le salaire (caractère Y).
Le caractère X admet deux modalités : femme et homme.

Salaire des femmes

1955	1764	1668	1441	1970	1795	1716	1911	1660	2001
1744	1676	1695	1652	1626	1698	1656	1739	1789	1716
1684	1445	1646	1617	1630	1440	1850	1252	1493	1537

Salaire des hommes

2283	2010	1970	2019	1941	2024	2046	1962	1948	2071
2108	1880	2008	2119	2030	2014	1919	1837	2094	2169

Soient n_F, l'effectif des femmes ; $\overline{Y}_F$ la moyenne des salaires des femmes ; σ^2_F la variance des salaires des femmes ; n_H, l'effectif des hommes ; $\overline{Y}_H$ la moyenne des salaires des hommes ; σ^2_H la variance des salaires des hommes ; $\overline{Y}$ la moyenne générale des salaires (hommes et femmes).
Nous avons

$$n_F = 30 \quad \overline{Y}_F = 1682.2 \quad \sigma^2_F = 26959.56$$

$$n_H = 20 \quad \overline{Y}_H = 2022.6 \quad \sigma^2_F = 9925.44$$

$$\overline{Y} = \frac{30\overline{Y}_F + 20\overline{Y}_H}{30+20} = 1818.36$$

La variance inter-groupe est :

$$V_{inter} = \frac{1}{50}\left\{n_F\left(\overline{Y}_F - \overline{Y}\right)^2 + n_H\left(\overline{Y}_H - \overline{Y}\right)^2\right\} = 27809.32$$

La variance totale est $\sigma^2 = 47955.23$. Le rapport de correlation est

$$\eta_{Y|X} = \frac{V_{inter}}{\sigma^2} \approx 0.58.$$

On peut considérer que le caractère sexe explique environ 58% de la variabilité des salaires observés.

Chapitre 5

Analyse combinatoire

L'analyse combinatoire est un important outil dans de nombreuses branches des mathématiques, notamment dans la théorie des probabilités et en statistique.

5.1 Principe multiplicatif

Si une expérience peut se décomposer en k phases successives, ces dernières pouvant s'effectuer respectivement de $n_1, n_2, \ldots, n_k$ manières, alors l'expérience peut se ráliser de $n_1 \times n_2 \ldots n_k$ manières. différentes.

Exemple 5.1.1. *Combien de possibilités avons-nous de constituer un code de* 4 *chiffres ?*

5.2 Arrangements

Définition 5.2.1. *Un arrangement de* p *éléments choisis parmi* n *éléments est une disposition ordonnée de* p *de ces* n *éléments.*

On distingue les arrangements avec répétitions et les arrangements sans répétitions.

5.2.1 Arrangements sans répétitions

C'est le nombre d'arrangements que l'on peut faire avec p éléments choisis parmi n éléments, chacun d'eux ne peut figurer qu'une seule fois dans le même arrangement.

Définition 5.2.2. *Le nombre d'arrangements sans répétitions de* p *éléments choisis parmi* n *est*

$$A_n^p = \frac{n!}{(n-p)!}$$

où $n! = n \times (n-1) \times \ldots \times 2 \times 1$.

Exemple 5.2.1. *Le nombre d'arrangements sans répétitions que l'on peut faire avec deux éléments choisis parmi trois éléments a, b, c est* $A_3^2 = 6$. *Ces 6 arrangements sont : (a,b), (b,a), (a,c), (c,a), (b,c), et (c,b).*

Remarque 5.2.1. *Un arrangement sans répétitions est une permutation si* $p = n$. *Le nombre de permutations de* n *éléments est :*

$$A_n^n = n!$$

Exemple 5.2.2. *Le nombre de permutations de* 3 *éléments a, b, c est* $P_3 = 3! = 6$. *Ces 6 permutations sont : (a,b,c), (a,c,b), (b,a,c), (b,c,a), (c,a,b), et (c,b,a).*

Exemple 5.2.3. *Tirage sans remise : Une urne* U *contient* n *boules numérotés de* 1 *à* n. *On tire successivement* p *boules de* U *sans les remettre dans l'urne. Il y a* A_n^p *tirages différents possibles.*

5.2.2 Arrangements avec répétitions

C'est le nombre d'arrangements que l'on peut faire avec p éléments choisis parmi n éléments, chacun d'eux peut figurer plusieurs fois dans le même arrangement.

Définition 5.2.3. *Le nombre d'arrangements avec répétitions de* p *éléments choisis parmi* n *est* n^p.

Exemple 5.2.4. *Le nombre d'arrangements avec répétitions que l'on peut faire avec deux éléments choisis parmi trois éléments a, b, c est* $3^2 = 9$. *Ces 9 arrangements sont :* (a,a), (a,b), (b,a), (a,c), (c,a), (b,b), (b,c), (c,b) *et* (c,c).

Exemple 5.2.5. *Tirage avec remise : Une urne* U *contient* n *boules numérotés de* 1 *à* n. *On tire successivement* p *boules de* U *en remettant chaque fois dans l'urne la boule qu'on vient de tirer. Le nombre de tirages possibles est donc* n^p.

5.3 Combinaisons

Définition 5.3.1. *Une combinaison de* p *éléments choisis parmi* n *éléments est une disposition non ordonnée de* p *de ces* n *éléments.*

On distingue les combinaisons avec répétitions et les combinaisons sans répétitions.

5.3.1 Combinaisons sans répétitions

C'est le nombre de combinaisons que l'on peut faire avec p éléments choisis parmi n éléments, chacun d'eux ne peut figurer qu'une seule fois dans la même combinaison.

Définition 5.3.2. *Le nombre de combinaisons sans répétitions de* p *éléments choisis parmi* n *est :*

$$C_n^p = \frac{n!}{p!(n-p)!}.$$

Exemple 5.3.1. *Le nombre de combinaisons sans répétitions que l'on peut faire avec deux éléments choisis parmi trois éléments a, b, c est* $C_3^2 = 3$. *Ces 3 combinaisons sans répétitions sont :* (a,b), (a,c), *et* (b,c).

Exemple 5.3.2. *Une urne* U *contient* n *boules numérotée de* 1 *à* n. *On tire simultanément* p *boules de* U. *Le nombre de tirages possibles vaut le nombre de combinaisons de* p *éléments parmi* n.

5.3.2 Combinaisons avec répétitions

C'est le nombre de combinaisons que l'on peut faire avec p éléments choisis parmi n éléments, chacun d'eux peut figurer plusieurs fois dans la même combinaison.

Définition 5.3.3. *Le nombre de combinaisons avec répétitions de p éléments choisis parmi n est :*

$$K_n^p = C_{n+p-1}^p.$$

C'est le nombre de façons de choisir p éléments parmi n, avec répétition et sans tenir compte de l'ordre des éléments.

Exemple 5.3.3. *Le nombre de combinaisons avec répétitions que l'on peut faire avec deux éléments choisis parmi trois éléments a, b, c est $K_3^2 = C_4^2 = 6$. Ces 6 combinaisons sont :* (a,a), (a,b), (a,c), (b,b), (b,c) *et* (c,c)

Exemple 5.3.4. *Soit $E = \{R,V,B\}$. Alors (B,B,R,V,V) est une combinaison avec répétition de 5 éléments de E.*

Exemple 5.3.5. *On veut ranger 3 pantalons dans 2 tiroirs. Quel est le nombre de possibilités ?*

Exemple 5.3.6. *On souhaite répartir p chiffons dans n tiroirs. On note les tiroirs $t_1,\ldots,t_n$. A une répartition, on associe le mot $t_1,\ldots,t_1,t_2,\ldots,t_2,\ldots,t_n,\ldots,t_n$, où chaque t_i est répété autant de fois que le nombre de chiffons rangés dans le tiroir. On obtient une combinaison avec répétitions.*

Chapitre

6 Espace probabilisé

L'objet des probabilités est de modéliser des phénomènes aléatoires et de prédire avec certitude leur évolution ou les conséquences qu'ils peuvent engendrer.

6.1 Univers des possibles

Définition 6.1.1. *Une expérience $\mathscr{E}$ est qualifiée d'aléatoire si on ne peut pas prévoir par avance son résultat et si, répétée dans des conditions identiques, elle peut donner lieu à des résultats différents.*

Définition 6.1.2. *L'univers des possibles (ou univers), noté Ω est défini par l'ensemble de tous les résultats possibles qui peuvent être obtenus au cours d'une expérience aléatoire.*

La description explicite de l'ensemble Ω est la première étape dans la modélisation d'un phénomène aléatoire. On distingue les univers comprenant un nombre fini de résultats de ceux comprenant un nombre infini de résultats. Parmi les univers infinis, on distingue les univers infinis non dénombrables des univers infinis dénombrables. Par exemple, l'univers $\Omega = \{\omega_1, \ldots, \omega_i, \ldots\}$ est un univers infini dénombrable puisque l'on peut identifier chacun des éléments de Ω, même s'il en existe une infinité. En revanche, $\Omega = \mathbb{R}$ est un exemple d'univers infinis non dénombrables. Dans le cas d'un univers fini ou infini dénombrable, la taille de l'univers est appelée cardinal de Ω et est noté $\mathit{card}(\Omega)$.

Exemple 6.1.1. *Voici quelques expériences aléatoires et les univers des possibles correspondants :*

1. *On lance une pièce. On a $\Omega = \{pile, face\}$.*
2. *On jette un dé. On a $\Omega = \{1,2,3,4,5,6\}$.*
3. *On jette deux dés. On a*

$$\Omega = \{(i,j) : 1 \leq i,j \leq 6\} = \{(1,1),(1,2),(1,3),\ldots\}.$$

4. *Un bus est censé passer toutes les 30 minutes à l'école de police pour se rendre à Faya. Un passager arrive à l'arrêt de bus. On cherche à modéliser son temps d'attente. A priori, on peut supposer que ce temps d'attente est dans l'intervalle $\Omega = [0,30]$.*

6.2 Evénements, Tribu

Définition 6.2.1. *Un événement (ou une partie) A est un sous-ensemble de l'univers des possibles Ω vérifiant $A \subset \Omega$.*

Définition 6.2.2. *Un événement constitué d'un seul élément est un événement élémentaire (ou singleton).*

Définition 6.2.3. *Un événement certain correspond à l'univers des possibles Ω.*

Définition 6.2.4. *Un événement impossible est un événement qui ne se réalise jamais. Il correspond à l'ensemble vide, noté $\emptyset$*

Exemple 6.2.1. *On considère une expérience aléatoire correspondant au lancer d'un dé à 6 faces. L'univers est alors $\Omega = \{1,2,3,4,5,6\}$. L'événement " nombre pair ", noté A, correspond au sous-ensemble de l' univers Ω défini par $A = \{2,4,6\}$.*

Définition 6.2.5. *Soient deux événements A et B. La réalisation de l'événement C, défini par $C = A \cup B$ implique la réalisation de l'événement A ou de l'événement B, ou des deux événements A et B simultanément.*

Définition 6.2.6. *Soient deux événements A et B. La réalisation de l'événement D, défini par $D = A \cap B$ entraîne la réalisation de l'événement A et de l'événement B.*

Définition 6.2.7. *Deux événements A et B sont disjoints s'ils n'ont pas d'élément en commun, c'est à dire, $A \cap B = \emptyset$. Ces deux événements sont donc incompatibles : la réalisation simultanée de ces événements est impossible.*

Définition 6.2.8. *Deux événements A et $\overline{A}$ inclus dans un ensemble B sont complémentaires si leur union correspond à B, c'est à dire, $A \cup \overline{A} = B$ et leur intersection est vide.*

On note $\mathscr{P}(\Omega)$, l'ensemble de toutes les parties de Ω.

Définition 6.2.9. *Soit $\mathscr{A} \subset \mathscr{P}(\Omega)$. On dit que $\mathscr{A}$ est une tribu sur Ω si les trois conditions suivantes sont vérifiées :*

- $\Omega \in \mathscr{A}$
- *si $A \in \mathscr{A}$ alors $\bar{A} \in \mathscr{A}$ (stabilité par passage au complémentaire)*
- *si $(A_i)_{i \in I}$ est une famille dénombrable d'éléments de $\mathscr{A}$ alors $\bigcup_{i \in I} A_i \in \mathscr{A}$. (stabilité par réunion dénombrable)*

Remarque 6.2.1. *La tribu $\mathscr{A}$ sur Ω représente l'ensemble de tous les évènements suceptibles de se produire au cours de l'expérience aléatoire $\mathscr{E}$. Lorsque l'ensemble Ω est fini ou dénombrable, on choisira pour $\mathscr{A}$ l'ensemble de toutes les parties de Ω, c'est-à-dire, $\mathscr{A} = \mathscr{P}(\Omega)$.*

Le couple $(\Omega, \mathscr{A})$ est appelé espace probabilisable. Pour compléter la description d'un phénomène aléatoire, il nous reste à introduire la notion de mesure de probabilité.

6.3 Probabilité

Pour une expérience aléatoire donnée, une fois déterminé le couple $(\Omega, \mathscr{A})$ qui représente l'univers Ω associé à cette expérience et la tribu des évènements $\mathscr{A}$, on définit une application de $\mathscr{A}$ à valeurs dans $[0,1]$ qui à chaque évènement associe sa probabilité, c'est à dire la chance de réalisation de cet évènement.

Définition 6.3.1. *On appelle* **probabilité** *sur* $(\Omega, \mathscr{A})$ *une application* $\mathbb{P} : \mathscr{A} \to [0,1]$ *telle que :*

(i) $\mathbb{P}(\Omega) = 1$

(ii) si $(A_i)_{i\in I}$ *est une famille dénombrable d'éléments de* $\mathscr{A}$ *deux à deux disjoints ou incompatibles (i.e.* $\forall i \neq j,\ A_i \cap A_j = \emptyset$*) alors*

$$\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i).$$

On appelle **espace probabilisé** *le triplet* $(\Omega, \mathscr{A}, \mathbb{P})$.

Propriété 6.3.1. *1.* $\mathbb{P}(\emptyset) = 0$

2. L'évènement A *tel que* $\mathbb{P}(A) = 0$ *est dit presque impossible.*

3. L'évènement A *tel que* $\mathbb{P}(A) = 1$ *est dit presque certain.*

4. $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.

5. $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$.

6. Si $A_1 \subseteq A_2$ *alors* $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$.

La probabilité $\mathbb{P}$ est dite discrète dès que l'espace Ω est fini ou infini dénombrable. La tribu associée est alors généralement $\mathscr{P}(\Omega)$. Une probabilité sur un ensemble dénombrable est complètement déterminée par $\mathbb{P}(\{\omega\})$ pour tout $\omega \in \Omega$. En effet, pour tout $A \subset \Omega$:

$$\mathbb{P}(A) = \sum_{\omega\in A} \mathbb{P}(\{\omega\}).$$

Définition 6.3.2. *On suppose que* Ω *est un ensemble fini. On dit que l'on se trouve dans un cas d'équiprobabilité si tous les évènements élémentaires ont la même probabilité.*

Dans ce cas, nous avons $\forall \omega \in \Omega$

$$\mathbb{P}(\{\omega\}) = \frac{1}{Card(\Omega)}$$

et pour tout $B \in \mathscr{P}(\Omega)$

$$\mathbb{P}(B) = \frac{Card(B)}{Card(\Omega)}.$$

6.4 Conditionnement et indépendance

Soit $(\Omega, \mathscr{A}, \mathbb{P})$ un espace probabilisé. Dans ce chapitre, nous allons étudier deux notions importantes : le conditionnement et l'indépendance. Le conditionnement permet de prendre en compte une information supplémentaire dans le calcul d'une probabilité. L'indépendance rend compte du fait que deux évènements n'ont aucune incidence l'un sur l'autre.

6.4.1 Probabilité conditionnelle

Définition 6.4.1. *Soient* A *et* B *deux évènements tels que* $\mathbb{P}(B) > 0$. *On appelle probabilité conditionnelle de* A *sachant que* B*, le réel défini par*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

L'application $A \longmapsto \mathbb{P}(A|B)$ *définit une probabilité sur* $(\Omega, \mathscr{A})$.

Proposition 6.4.1. ***Formule des probabilités composées.*** *oit $A_0,\dots,A_n$ une suite d'évènements telle que $\bigcap_{i=0}^{n} A_i \neq \emptyset$. Alors, on a*

$$\mathbb{P}(\bigcap_{i=0}^{n} A_i) = \mathbb{P}(A_0) \times \mathbb{P}(A_1|A_0) \times \mathbb{P}(A_2|A_0 \cap A_1) \times \dots \times \mathbb{P}(A_n|A_0 \cap A_1 \cap \dots \cap A_{n-1}).$$

Exemple 6.4.1. *Pour $n = 1$, on a*

$$\mathbb{P}(A_0 \cap A_1) = \mathbb{P}(A_0) \times \mathbb{P}(A_1|A_0).$$

Pour $n = 2$, on a

$$\mathbb{P}(A_0 \cap A_1 \cap A_2) = \mathbb{P}(A_0) \times \mathbb{P}(A_1|A_0) \times \mathbb{P}(A_2|A_0 \cap A_1).$$

Définition 6.4.2. *Une famille finie d'évènements $(A_i)_{1\leq i\leq n}$ deux à deux incompatibles tels que $\bigcup_{i=1}^{n} A_i = \Omega$ est appelée* ***système complet d'évènements****.*

Théorème 6.4.1. ***Formule des probabilités totales.***
Soit $\{B_1,\dots,B_n\}$ un système complet d'évènements. Alors, nous avons

$$\forall A \in \mathscr{A} \quad \mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

Exemple 6.4.2. *Une urne contient des boules blanches et nores, marquées ou non. On suppose que parmi les boules marquées, il y a 30% de boules blanches et parmi les non marquées 60%. Par ailleurs, on sait que 80% des boules sont marquées. Quelle est la probabilité de tirer une boule blanche ?*

Solution. *On note*

B = *"la boule est blanche"*

M = *"la boule est marquée"*

On a

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \cap M) + \mathbb{P}(B \cap M^c) \\ &= \mathbb{P}(M) \times \mathbb{P}(B|M) + \mathbb{P}(M^c) \times \mathbb{P}(B|M^c) \\ &= \frac{80}{100} \times \frac{30}{100} + \frac{20}{100} \times \frac{60}{100} = \frac{36}{100}.\end{aligned}$$

Théorème 6.4.2. *(**Formule de Bayes**). Soit $\{B_1,\dots,B_n\}$ un système complet d'évènements et A un évènement tel que $\mathbb{P}(A) > 0$. Alors, nous avons*

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(B_i)\mathbb{P}(A|B_i)}{\sum_{k=1}^{n} P(B_k)\mathbb{P}(A|B_k)}.$$

Exemple 6.4.3. *Le quart d'une population est vacciné contre le choléra. Au cours d'une épidémie, on constate qu'il y a parmi les malades un vacciné pour 4 nonvaccinés, et qu'il y a un malade sur 12 parmi les vaccinés. Quelle est la probabilité qu'un non-vacciné tombe malade ?*

6.4.2 Indépendance

Définition 6.4.3. *Soient A et B deux évènements. On dit que A et B sont indépendants si* $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Si A est tel que $\mathbb{P}(A) > 0$, l'indépendance de A et B s'écrit encore $\mathbb{P}(B|A) = \mathbb{P}(B)$ et on retrouve la notion intuitive d'indépendance : le fait que A se soit réalisé ne change rien quant à la probabilité que B se réalise.

Proposition 6.4.2. *Si A et B sont indépendants, alors il en va de même pour :*

- *les évènements $\bar{A}$ et B ;*
- *les évènements A et $\bar{B}$;*
- *les évènements $\bar{A}$ et $\bar{B}$*

Définition 6.4.4. *Les évènements $A_1, \ldots, A_n$ sont dits mutuellement indépendants si*

$$\forall I \subset \{1, \ldots, n\}, \quad \mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Le fait que les événements $A_1, \ldots, A_n$ sont indépendants deux à deux indépendants entraîne qu'ils sont indépendants deux à deux. La ré ciproque est fausse.

Exercice 6.4.1. *On lance une pièce deux fois de suite. Soit A l'événement "obtenir face au premier jet", B l'événement "obtenir face au deuxième jet" et C l'evénement "obtenir deux résultats différents. Momtrer que les évènements A, B, C sont deux à deux indépendants, mais la famille $\{A, B, C\}$ n'est pas indépendamte.*

Chapitre 7

Variables aléatoires discrètes

7.1 Généralités

Soit $(\Omega, \mathscr{A}, \mathbb{P})$ un espace probabilisé. La variable aléatoire X traduit une situation liée à l'expérience aléatoire modélisée par l'espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$.

Définition 7.1.1. *Une variable aléatoire X réelle est une application définie sur Ω à valeurs dans $\mathbb{R}$ telle que pour tout $x \in \mathbb{R}$,*

$$X^{-1}(]-\infty, x]) = \left\{\omega \in \Omega : X(\omega) \leq x\right\} \in \mathscr{A}.$$

C'est à dire que pour tout $x \in \mathbb{R}$, $X^{-1}(]-\infty, x])$ est un événement.

Étant donnés un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$ et une variable aléatoire réelle X, on peut construire de façon naturelle une probabilité sur $X(\Omega)$, l'ensemble des valeurs prises par la fonction X. Cette probabilité est appelée loi de la variable aléatoire X et est notée $\mathbb{P}_X$.

7.2 Variables aléatoires discrètes

Définition 7.2.1. *La variable aléatoire réelle X est dite discrète si $X(\Omega)$ est fini ou infini dénombrable.*

La loi de probabilité d'une variable aléatoire réelle discrète X est déterminée par :

1. $X(\Omega)$
2. $\mathbb{P}_X(\{x\}) = \mathbb{P}(X = x)$, pour tout $x \in X(\Omega)$

La probabilité d'un évènement A est donnée par

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x).$$

De plus, nous avons

$$\sum_{x \in X(\Omega)} \mathbb{P}(X = x) = 1.$$

7.3 Fonction de répartition

Définition 7.3.1. *Soit X une variable aléatoire réelle. On appelle fonction de répartition de X, la fonction F définie sur $\mathbb{R}$ à valeurs dans $[0,1]$ par :*

$$F(x) = \mathbb{P}(X \leq x).$$

Proposition 7.3.1. *On a :*

1. *F est une fonction en escaliers.*
2. *F est croissante ;*
3. *F est continue à droite ;*
4. *$\lim_{x\to+\infty} F(x) = 1$ et $\lim_{x\to-\infty} F(x) = 0$;*
5. *Pour tous réels a et b avec $a < b$,*

$$\mathbb{P}(a < X \leq b) = F(b) - F(a).$$

Proposition 7.3.2. *F est continue à droite en tout $x \in \mathbb{R}$ et*

$$\mathbb{P}(X = x) = F(x^+) - F(x^-) = F(x) - F(x^-)$$

où

$$F(x^+) = \lim_{t\to x, t>x} F(t).$$

$$F(x^-) = \lim_{t\to x, t<x} F(t).$$

Pour une variable aléatoire discrète :

$$F(x) = \sum_{t\leq x} \mathbb{P}(X = t) \quad \forall x \in \mathbb{R}..$$

Exemple 7.3.1. *On lance deux dés non pipés. L'univers associé à cette expérience est*

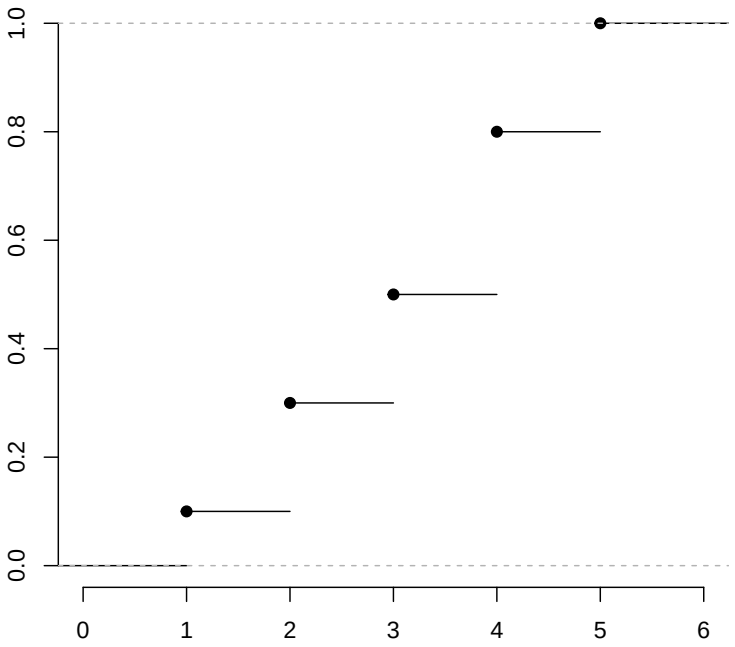
$$\Omega = \{(i,j) : 1 \leq i,j \leq 6\}.$$

Nous avons $card(\Omega) = 36$ et la probabilité sur Ω est définie par

$$\mathbb{P}(\{\omega\}) = \frac{1}{36}.$$

On s'interesse à la variable aléatoire discrète suivante : pour tout $\omega = (i,j) \in \Omega$, $X(\omega) = i + j$.

x	2	3	4	5	6	7	8	9	10	11	12
p_x	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



La fonction de répartition est

$$F(x) = \sum_{t \leq x} \mathbb{P}(X = t)$$
$$= \begin{cases} 0 & si\ x < 2 \\ 1/36 & si\ 2 \leq x < 3 \\ 3/36 & si\ 3 \leq x < 4 \\ 6/36 & si\ 4 \leq x < 5 \\ 10/36 & si\ 5 \leq x < 6 \\ 15/36 & si\ 6 \leq x < 7 \\ 21/36 & si\ 7 \leq x < 8 \\ 26/36 & si\ 8 \leq x < 9 \\ 30/36 & si\ 9 \leq x < 10 \\ 33/36 & si\ 10 \leq x < 11 \\ 35/36 & si\ 11 \leq x < 12 \\ 1 & si\ x \geq 12 \end{cases}$$

7.4 Caractéristiques des variables aléatoires discrètes

7.4.1 Espérance

Soit X et Y deux variables aléatoires réelle discrètes.

Définition 7.4.1. *On appelle* ***espérance*** *de* X*, le nombre réel*

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x).$$

Lorsque $\mathbb{E}(X)$ *est fini, on dit que* X *admet un moment d'ordre* 1*.*

Définition 7.4.2. *La variable aléatoire* X *est dite centrée si* $\mathbb{E}(X) = 0$*.*

Théorème 7.4.1. Théorème de transfert.
Soit g *une application définie sur* $\mathbb{R}$ *à valeurs dans* $\mathbb{R}$*. Nous avons*

$$\mathbb{E}(g(X)) = \sum_{x \in X(\Omega)} g(x)\mathbb{P}(X = x).$$

Proposition 7.4.1. *(Linéarité de l'espérance). Soit* $c \in \mathbb{R}$ *une constante. Alors on a*

$$\mathbb{E}[cX + Y] = c\mathbb{E}[X] + \mathbb{E}[Y].$$

Proposition 7.4.2. *On suppose que que* $X \leq Y$*. Alors, nous avons*

$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$

Proposition 7.4.3. *Soit* X *une variable aléatoire discrète positive. Alors, pour tout* $\varepsilon > 0$*, nous avons*

$$\mathbb{P}(X > \varepsilon) \leq \frac{\mathbb{E}(X)}{\varepsilon}.$$

L'espérance de X est la moyenne pondérée des valeurs que X peut prendre, les poids étant les probabilités que ces valeurs soient prises. C'est un indicateur de localisation. Néanmoins, la connaissance de l'espérance seule donne peu de renseignements sur X. Ainsi, elle s'accompagne de la variance qui caractérise la dispersion de X autour de sa moyenne $\mathbb{E}(X)$.

7.5 Variance, écart-type

Définition 7.5.1. *Soit* X *une variable aléatoire discrète. On appelle moment d'ordre* $k \geq 1$*, la quantité*

$$\mathbb{E}[X^k] = \sum_{x \in X(\Omega)} x^k\mathbb{P}(X = x).$$

Définition 7.5.2. *Soit* X *une variable aléatoire qui admet des moments d'ordre deux i.e.* $\mathbb{E}[X^2] < +\infty$*. On appelle variance de* X *la quantité*

$$var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[(X - \mathbb{E}(X))^2].$$

Définition 7.5.3. *On appelle écart-type* σ_X *la racine carrée de la variance :* $\sigma_X = \sqrt{var(X)}$*.*

Proposition 7.5.1. *Soient* X *et* Y *deux variables aléatoires réelles, et* a *et* b *deux constantes réelles. Alors on a*

- $Var(X) \geq 0$
- $var(aX + b) = a^2 var(X)$

Proposition 7.5.2. *(Inégalité de Bienaymé-Chebychev). Pour tout* $\varepsilon > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| > \varepsilon) \leq \frac{var(X)}{\varepsilon^2}.$$

7.6 Variables aléatoires discrètes indépendantes

Définition 7.6.1. *On dit que les variables aléatoires discrètes X et Y sont indépendantes si pour tout $x \in X(\Omega)$ et tout $y \in Y(\Omega)$, les événements $[X = x]$ et $[Y = y]$ sont indépendants, c'est à dire,*

$$\mathbb{P}\Big([X = x] \cap [Y = y]\Big) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

Proposition 7.6.1. *Soient X et Y deux variable aléatoires discrètes admettant des moments d'ordre 1. Alors la variable aléatoire XY admet un moment d'ordre 1. De plus, si X et Y sont indépendantes alors*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Définition 7.6.2. *Soient X et Y deux variable aléatoires discrètes admettant des moments d'ordre 1. On appelle covariance entre X et X, la quqntité*

$$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Si X et Y sont indépendantes alors $Cov(X,Y) = 0$.

Proposition 7.6.2. *Si X et Y admettent un moment d'ordre deux et sont indépendantes alors $X+Y$ a un moment d'ordre deux et*

$$Var(X+Y) = Var(X) + Var(Y).$$

Définition 7.6.3. *On appelle coefficient de corrélation linéaire entre X et Y, le nombre*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Le coefficient de corrélation permet de mesure l'intensité de la liaison linéaire entre les variables X et Y.

Proposition 7.6.3. *On a $-1 \leq \rho(X,Y) \leq 1$.*

Remarque 7.6.1.

- *Si $|\rho(X,Y)| = 1$ alors il existe une liaison linéaire entre X et Y :*

$$Y = aX + b \quad a,b \in \mathbb{R}.$$

- *Si $\rho(X,Y) = 0$ alors X et Y sont dits linéairement indépendantes.*
- *Si $\rho(X,Y) > 0$ alors X et Y évoluent dans le même sens.*
- *Si $\rho(X,Y) < 0$ alors X et Y évoluent en sens contraire.*
- *Si X et Y sont indépendantes alors $\rho(X,Y) = 0$*

Nous terminons ce chapitre par un résultat très important appelé **loi des grands** qui permet détudier le comportement de la moyenne empirique vue en statistique descriptive et la moyenne théorique vue dans ce cours.

Théorème 7.6.1. *Soit $X_1,\ldots,X_n,\ldots$ une suite de varibales aléatoires indépendantes et de même loi, d'espérance $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \ldots$ et de variance $Var(X_1) = Var(X_2) = \ldots$ Alors quel que soit $\varepsilon > 0$, on a*

$$\lim_{n\to+\infty} \mathbb{P}\Big(\Big|\frac{X_1 + \ldots + X_n}{n} - \mathbb{E}(X_1)\Big|\Big) = 0.$$

Cela signifie que pour tout réel $\varepsilon > 0$ fixé, on peut trouver un entier $N(\varepsilon)$ tel que pour $n \geq N(\varepsilon)$, la probabilité que la moyenne empirique s'écarte de la moyenne théoriaue d'au moins ε est très petite. En d'autres termes, pour n assez grand la probabilité que la moyenne empirique $\overline{X}_n$ soit proche de la moyenne théorique dans les coditions du théorème est proche de 1.

Chapitre 8

Quelques lois de probabilités discrètes

8.1 Loi uniforme discrète

$$X \hookrightarrow \mathcal{U}_N \Longleftrightarrow \begin{cases} X(\Omega) = \{1, \ldots, N\} \\ P(X = k) = \frac{1}{N}, \quad \forall k \in X(\Omega) \end{cases}$$

$$E(X) = \frac{N+1}{2}$$

et

$$var(X) = \frac{N^2 - 1}{12}.$$

Exemple 8.1.1. *Soit X le résultat d'un lancer de dé non truqué : alors $\forall i \in X(\Omega) = \{1,2,3,4,5,6\}$, $P(X = i) = \frac{1}{6}$; X suit la loi uniforme $\mathcal{U}_6$.*

8.2 Loi de Bernouilli

$$X \hookrightarrow \mathcal{B}(1,p) \Longleftrightarrow \begin{cases} X(\Omega) = \{0,1\} \\ P(X = 1) = p, \ P(X = 0) = 1 - p \end{cases}$$

$$E(X) = p$$

$$var(X) = p(1-p).$$

Cette variable modélise l'issue d'une expérience où l'on ne s'intéresse qu'au "succès" ou à l'"echec" de l'expérience.

Exemple 8.2.1. *Lancer d'une pièce de monnaie (pile ou face), qualité d'un produit (bon ou defectueux), sondage elctoral (pour ou contre).*

8.3 Loi binomiale

On réalise n fois successivement et d'une manière indépendante une expérience aléatoire de Bernouilli. La variable aléatoire égale au nombre de succès obtenus au cours des n épreuves

suit la loi binomiale $\mathscr{B}(n,p)$.

$$X \hookrightarrow \mathscr{B}(n,p) \Longleftrightarrow \begin{cases} X(\Omega) = \{0,\ldots,n\} \\ P(X=k) = C_n^k p^k (1-p)^{n-k}, \ \forall k \in X(\Omega) \end{cases}$$

$$E(X) = np$$

$$var(X) = np(1-p).$$

Cette loi modélise une succession de "succès" et d'"échecs", p étant la probabilité du succès.

8.4 Loi hypergéométrique

Soit une population de N individus parmi lesquels une proportion p (donc Np individus) possède un caractère. Il s'agit par exemple de la proportion des individus qui souffrent d'une maladie, ou de la proportion des pièces défectueuses dans un grand lot de fabrication. On prélève un échantillon de n individus parmi cette population (le tirage pouvant s'effectuer d'un seul coup ou au fur et à mesure mais sans remise). On note X la variable aléatoire égale au nombre d'individus de l'échantillon possédant le caractère envisagé. La loi de X est appelée loi hypergéométrique de paramètre N, n, p et notée $\mathscr{H}(N,n,p)$:

$$X \hookrightarrow \mathscr{H}(N,n,p) \Longleftrightarrow \begin{cases} X(\Omega) = \{max(0, n-(1-p)N), min(Np,n)\} \\ P(X=k) = \frac{C_{Np}^k C_{(1-p)N}^{n-k}}{C_N^n}, \ \forall k \in X(\Omega) \end{cases}.$$

$$E(X) = np.$$

8.5 Loi géométrique

C'est la loi du nombre d'essais (ou épreuves) nécessaires pour faire apparaître un évènement de probabilité p. C'est le cas de nombre d'examens necessaires pour réussir une épreuve en supposant que la probabilité de réussir à chaque passage de l'examen est de type p et que les résultats sont indépendants d'un examen vers un autre. Soit la variable X égale le nombre d'essais avant d'obtenir le premier succès :

$$X \hookrightarrow \mathscr{G}(p) \Longleftrightarrow \begin{cases} X(\Omega) = \mathbb{N}^* \\ P(X=k) = p(1-p)^{k-1}, \ \forall k \in X(\Omega) \end{cases}.$$

$$E(X) = \frac{1}{p}$$

$$var(X) = \frac{1-p}{p^2}.$$

Exemple 8.5.1. *On effectue des lancers indépendants d'une pièce, dont la probabilité d'obtenir face est p, jusqu'à l'obtention d'un "face". On note X la v.a.r égale au nombre de lancers nécessaires. On dit également que X est le temps d'attente du premier "face".*

8.6 Loi de Poisson

Pour modéliser des phénomènes rares (nombre d'accidents d'avion, nombre d'appels téléphoniques pendant un certain temps, nombre de pièces défectueuses dans une commande importante, nombre de suicides par an dans un pays donné...), on utilise la loi de Poisson (de paramètre $\lambda > 0$) :

$$X \hookrightarrow \mathscr{P}(\lambda) \Longleftrightarrow \begin{cases} X(\Omega) = \mathbb{N} \\ P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \ \forall k \in X(\Omega) \end{cases}$$

$$E(X) = var(X) = \lambda.$$