Introduction

La théorie des probabilités permet de modéliser les phénomènes aléatoires (initialement développée à partir des jeux de hasard, puis étendue à l'ensemble des sciences expérimentales). Cette théorie permet aussi de construire des modèles de ces phénomènes et d'y effectuer des calculs théoriques (On peut prédire les fréquences d'apparition d'événements à partir d'un modèle probabiliste d'un jeu de hasard).

Le mot hasard vient de l'arabe : az-zahr (le dé). En français ce mot désigne tout d'abord un jeu de dés, puis plus généralement un évenement aléatoire.

Á partir du $17^{ième}$ siècle plusieurs principes de calcul de probabilités ont été avancés, parmi eux : Les travaux de Pierre de Fermat, Blaise Pascal et Christian Huygens, puis Pierre-Simon de Laplace, Abraham de Moivre, jacque Bernouli et Denis Siméon Poisson ($18^{ième}$ siècle), et Carl Friedrich Gauss et Henri Poincaré ($19^{ième}$ siècle).

C'est à partir de 1933 que Andrei Kolmogorov introduit rigoureusement la théorie des probabilités, cela nécessite le développement des théories de la mesure et de l'intégration.

La statistique est la discipline qui sert à étudier des phénomènes à travers la collecte de données, l'élaboration des procédures pour traiter, analyser et interpréter les resultats de ces données afin de les rendre compréhensibles par tous.

Le terme " statistique " remonte au latin classique " status " i.e : état qui, par une série d'évolution aboutit au français statistique

$$(status \rightarrow stato \rightarrow statista(1633) \rightarrow statistica(1672) \rightarrow statisticus(1771) \rightarrow statistique).$$

Vers la même époque statistik apparait en allemend, tandis que les anglais utilisent l'expression political arithmetic jusqu'en 1798, date à laquel le mot statistics entre dans le dictionnaire de cette langue.

Aux temps anciens, la statistique consiste qu'à la collection d'information sur les états, actuellement , la statistique désigne à la fois un ensemble de données d'observation et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation.

Ce document est structuré en trois chapitres :
Le premier chapitre concerne les concepts de base de la statistique descriptive, les tableaux statistiques : Cas de caractère qualitatif (Représentation circulaire par des secteurs, Représentation

en tuyaux d'orgue, Diagramme en bandes), cas de caractère quantitatif (Diagramme en batons, Histogramme, Polygone).

Le deuxième chapitre est consacré pour la représentation numérique des données : les caractéristiques de tendance centrale ou de position (La médiane, Les quartiles, Intervalles interquartile, Le mode, La moyenne arithmétique, La moyenne arithmétique pondérée, La moyenne arithmétique géométrique, La moyenne harmonique, La moyenne quadratique), et les caractéristiques de dispersion (L'étendu, L'écart type, L'écart absolue moyen, Le coefficient de variation).

Le dernier chapitre permet aux étudiants de comprendre les notions fondamentaux en probabilités : Analyse combinatoire, calcul des probabilités, les variables aléatoires.

Chapitre 1

Notions de base et vocabulaire statistique

1.1 Introduction

Ce chapitre permet de définir les notions de base de la statistique descriptive, et les tableaux statistiques dont des différentes représentations graphiques d'un caractère qualitatif, et quantitatif (discret et continu) seront traitées.

1.2 Concepts de base de la statistique

Population : c'est l'ensemble de tous les éléments concernés par l'étude, appelée aussi univers. Par exemple la section A de la $1^{ère}$ année MI.

Individus : c'est chaque élément de la population, applées aussi unités statistiques peuvent être des êtres humains, des objets ...

Caractère (Variable) : c'est la propriété caractéristique des éléments de la population, les différents caractères étudies habitiellement sont : l'âge, la taille, le poids, la nationalité, le groupe sanguin ...

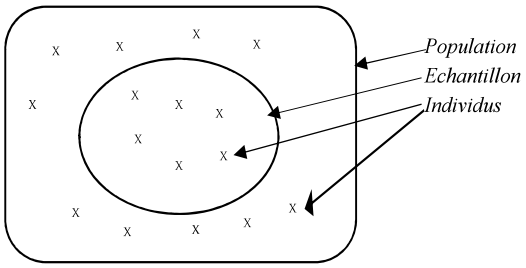
Modalités : ce sont les différentes situations possibles d'un caractère. Par exemple, le caractère sexe présente deux modalités (Féminin, Masculin).

Les modalités doivent être incompatibles (un individu ne peut appartenir à plus d'une modalité à la fois).

Effectif (Fréquence absolue) : est le nombre d'individu présentant chaque modalité.

L'échantillon : est un sous ensemble de la population souvent le nombre des individus d'une population est assez grand, alors le traitement des résultats sera très délicat, dans ce cas on doit prendre un sous ensemble de la population choisi aléatoirement pour avoir toutes les propriétés qui existent dans la population.

Taille : le nombre des éléments de l'échantillon.



Les caractères statistiques se décomposent en deux types :

Caractère qualitatif : Tout caractère qu'on ne peut pas mesurer. Par exemple : le sexe, la profession, la nationnalité, région d'habitation, ...

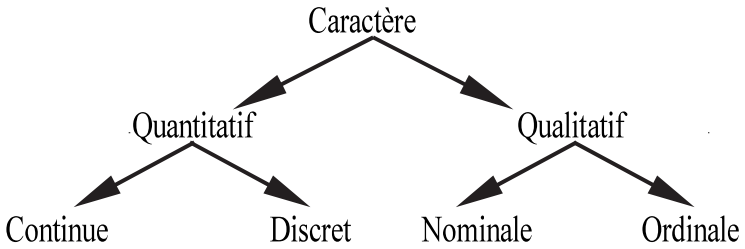
Il y a deux types de caractère qualitatif :

— Caractère qualitatif nominale : ses modalités ne sont pas naturellement ordonnés (statut matrimonial, sexe).

— Caractère qualitatif ordinale : ses modalités peut être doté d'une relation d'ordre (niveau d'instruction).

Caractère quantitatif : C'est un caractère mesurable autrement dit on peut assosier à chaque individu de la population une valeur réelle, par exemple : la taille, le poids, la moyenne des étudiants.

Remarque 1.1 *Par fois, pour raisons de traitement on essaye de traduire à l'aide d'un codage les caratères qualitatifs en nombres réeles pour quelle devienne une variable quantitative.*



1.3 Les tableaux statistiques

Au cours d'une étude statistique les données sont recueillies de façon désordonnée. Cette masse d'informations ne peut fournir des renseignements que si l'information en question est classée et mise en ordre, ce classement ne peut se faire que dans des tableaux statistiques qui servent de documentation statistique.

On traduit les tableaux statistiques par des graphiques de façon à visualiser le comportement du caractère statistique.

1.3.0.1 Fréquence absolue (effectif) et fréquence relative

Considérons une population composé de n individus décrits selon caractère X, constitué de la modalité $x_1, x_2, ..., x_k$.

La présentation de cette information dans un tableaux consiste à dénombrer le nombre d'individus possédant chaque modalité puis de les ordonner.

Le tableaux théorique est le suivant

Modalités X_i`du caractères`	x_1	x_2	...	x_i	...	x_k
Nombre d'individus n_i	n_1	n_2	...	n_i	...	n_k

Effectif (ou fréquence absolue) : on appelle effectif ou fréquence absolue le nombre n_i d'individu ayant pris la modalité x_i du caractère X.

Les observations ordonnées forment une série statistique (on distribution statistique) qui est constituée de l'ensemble des données et les effectifs correspondant.

$$\left\{(x_i, n_i)\,,\ i = \overline{1,k}\right\} \text{ est une série statistique.}$$

$$\sum_{i=1}^{k} n_i = n = n_1 + n_2 + ... + n_k.$$

Féquence (Fréquence relative) : on appelle fréquence ou fréquence relative de la modalité x_i le nombre $f_i = \dfrac{n_i}{n}$: c'est la proportion d'individus ayant pris la modalité x_i.

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + ... + f_k = 1.$$

1.3.1 Représentation graphique d'un caractère qualitatif

Lorsque le caractère est qualitatif, les modalités x_i sont ordonnées selon les effectifs n_i ($\nearrow$ ou $\searrow$) .

Exemple 1.1 *La répartition des travailleurs d'une entreprise selon la qualification à été comme suit : 10 Ingénieurs, 30 Employés, 140 Ouvriers et 20 Techniciens.*

Pour interpréter ces données, on doit les ranger dans un tableaux statistique

Qualification	*Nombre de travailleurs* n_i	*Fréquance relative* f_i
Ouvriers	*140*	*0.7*
Employés	*30*	*0.15*
Techniciens	*20*	*0.10*
Ingénieurs	*10*	*0.05*
Total	*200*	*1*

1.3.1.1 Représentation en tuyaux d'orgue

Ce type de représentation s'obtient en construisant autant de colonnes que de modalités du caractère. Ces colonnes sont des rectangles de base constante et de hauteur proportionnelle aux f_i (ou n_i) .

1.3.1.2 Représentation en diagramme circulaire

Le diagramme circulaire permet de visualiser la part relative de chaque modalité du caractère. Le support de cette représentation est un cercle divisé en autant de portions que de modalités du caractère.

L'angle θ_i indiquant la portion est donné par :

$$\theta_i = 360^\circ \times \frac{n_i}{n} = 360^\circ \times f_i$$

Exemple 1.2 *La répartition des travailleurs selon la qualification*

$\theta_1 = 360^\circ \times f_1 = 360^\circ \times 0.7 = 252^\circ$
$\theta_2 = 360^\circ \times f_2 = 360^\circ \times 0.15 = 54^\circ$
$\theta_3 = 360^\circ \times f_3 = 360^\circ \times 0.10 = 36^\circ$
$\theta_4 = 360^\circ \times f_4 = 360^\circ \times 0.05 = 18^\circ$

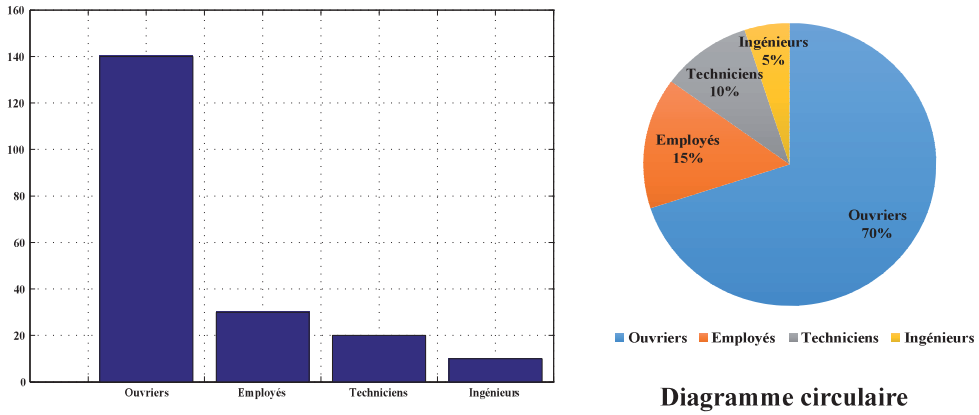


FIGURE 1.1: *Présentation graphique de la répartition des ouvriers : En tuyaux d'orgue (à gauche) et Diagramme circulaire (à droite).*

1.3.2 Représentation graphique d'un caractère quantitatif (cas discret)

Exemple 1.3 *Distribution statistique du nombre de pièces par logement*

Nombre de pièces x_i	*1*	*2*	*3*	*4*	*5*	*6*	*Total*
Nombre de logement n_i	*5*	*10*	*20*	*30*	*25*	*10*	*100*
f_i	*0.05*	*0.1*	*0.20*	*0.30*	*0.25*	*0.10*	*1*

Tabelaux statistique de la répartition des logements selon le nombre de pièces.

La répresentation graphique adéquate d'un caractère statistique (variable) discret est le diagramme en bâtons : à chaque valeur x_i de la variable statistique on fait correspondre un bâton dont la hauteur est proportionnelle à n_i ou f_i.

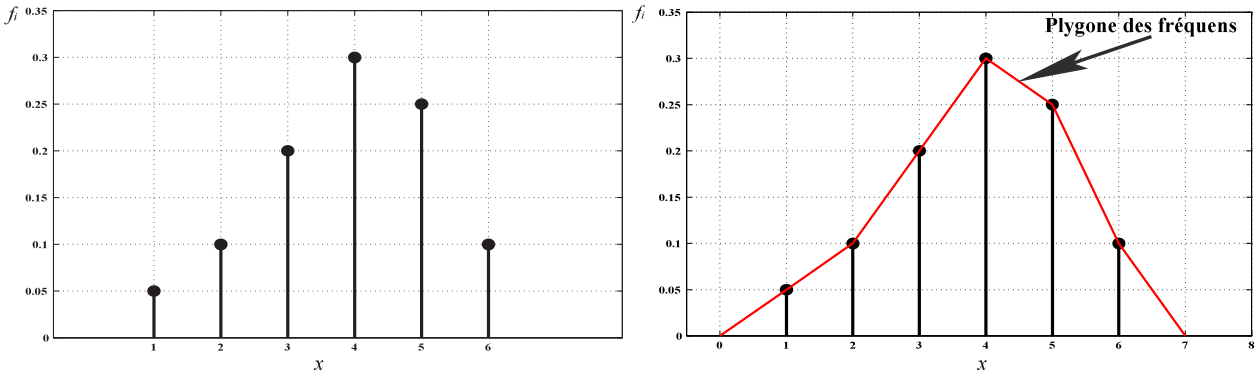


FIGURE 1.2: *Diagramme en Batons.*

Remarque 1.2 *Si l'on joint les sommets des bâtons on obtient le polygônes des effectifs (ou des fréquences).*

Effectif cumulée (resp. Fréquence cumulée) : On appelle effectif cumulée (resp. Fréquence cumulée) le nombre N_i (resp. F_i) telle que :

$$N_i = \sum_{j=1}^{i} n_j \quad (\text{resp. } F_i = \sum_{j=1}^{i} f_j),$$

la fréquence F_i répond à la question quelle est la proportion d'individus ayant le caractère inferieur à x_{i+1} ou superieur ou égale à x_i.

La courbe cumulative ou polygône des fréquences cumulées (absolues ou relatives) est la représentation graphique de ces fréquences cumulées.

Dans le cas d'une variable statistique discrète, la courbe cumulative est la représentation d'une fonction en escalier dont les paliers horizontaux ont pour coordonnées (x_i, F_i). Cette fonction appelée fonction de répartition emperique, définie par :

$$\begin{array}{l} F : \mathbb{R} \longrightarrow [0,1]\,. \\ x \longrightarrow F(x)\,, \end{array}$$

telle que

$$F(x) = \begin{cases} 0 & \text{si} \quad x < x_1 \\ f_1 & \text{si} \quad x_1 \leq x < x_2 \\ f_1 + f_2 & \text{si} \quad x_2 \leq x < x_3 \\ \vdots & \\ \displaystyle\sum_{j=1}^{i} f_j & \text{si} \quad x_i \leq x < x_{i+1} \\ \vdots & \\ 1 & \text{si} \quad x \geq x_k. \end{cases}$$

Exemple 1.4 *Lors d'une expérience biologique sur la fréquence d'un champignon bien précis dans un certain milieu, l'expérimentateur a constaté que la distribution des fréquences de ces champignons sur n sites, de ce milieu, peut-être résumer comme suit :*

						$\sum$
X_i	*5*	*6*	*7*	*8*	*9*	
f_i	*0.05*	*0.10*	*0.40*	*0.30*	*0.15*	*1*

Pour l'exemple **1.3.3** : répartition des champignons dans le milieu étudié on aura :

						$\sum$
X_i	5	6	7	8	9	
f_i	0.05	0.10	0.40	0.30	0.15	1
$F_i \nearrow$	0.05	0.15	0.55	0.85	1	

Remarque 1.3 *La fonction F est discontinue en chaque point de la variable statistique.*

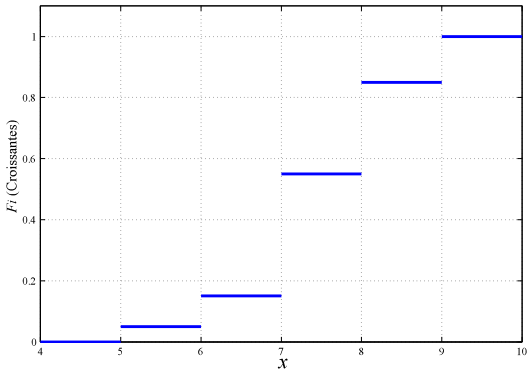


FIGURE 1.3: *Diagramme des fréquences cumulatives croissantes.*

1.3.3 Représentation graphique d'un caractère quantitatif (cas continu)

Dans le cas d'une variable statistique continue, pour établir le tableaux statistique, il faut effectuer au préalable une répartition en classe des données. Une classe est définie par la donnée de ses extremitée inferieur et superieur : par convention $[e_{i-1}, e_i]$.

Cela nécessite de définir le nombre de classe et l'amplitude associée à chaque classe.

La i^{ème} classe par exemple est donnée par $[e_{i-1}, e_i]$ où

e_{i-1} : l'extrémité inferieur.

e_i : l'extrémité superieur.

a_i : (e_{i-1}, e_i) amplitude de la i^{ème} classe.

Le centre de cette classe est $x_i = \dfrac{e_i + e_{i-1}}{2}$.

Remarque 1.4

1. *En génerale, on choisit des classes de même amplitude.*
2. *Le choix du nombre de classes et de leur amplitude se fait en fonction de l'effectif total (n) de la population.*
3. *Toute diminution du nombre de classes et toute augmentation de l'amplitude, conduit à une pert d'information.*
4. *Règle de Sturge : nombre de classe* $\simeq 1 + (3,3 \log N)$, *et l'amlplitude est* $a = \dfrac{\max x_i - \min x_i}{\text{nombre de classes}} = \dfrac{\text{étudue de la}}{\text{nombre de classes}}$.
   Le tableaux statistique d'une variable statistique continue se présente sous la forme suivante :

Classes	*Centre* x_i	n_i	f_i
$[e_0, e_1]$	x_1	n_1	f_1
$[e_1, e_2]$	x_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[e_{i-1}, e_i]$	x_k	n_k	f_k

5. *Les modalités d'une variable statistique continues sont les centres des classes*

$$x_i = \frac{e_i + e_{i-1}}{2}.$$

6. *Pour représenter graphiquement une variable statistique continues on utilise l'histogramme : qui consiste une généralisation du diagramme en bâtons à la notion de classe.*

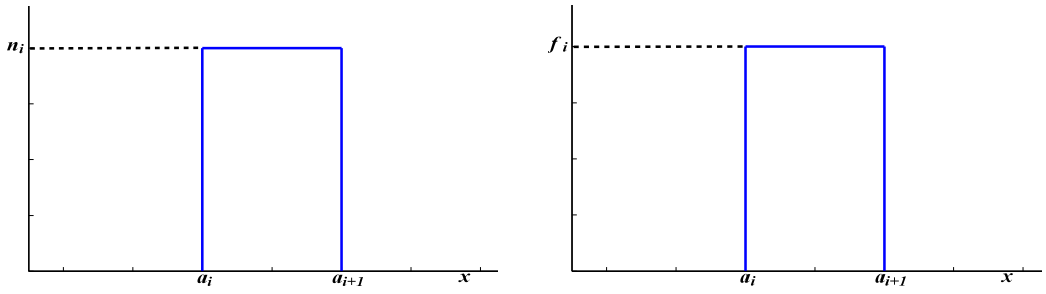


FIGURE 1.4: *Illustration de représentation graphique d'une classe*

Exemple 1.5 *Les données ci-dessous sont issues d'une expérience dans laquelle la concentration de calcium dans le plasma a été mesurée chez 40 personnes ayant subi l'administration d'un traitement hormonal.*

X	*[10,16[*	*[16,22[*	*[22,28[*	*[28,34[*	*[34,40[*	*[40,46[*
n_i	*4*	*6*	*17*	*8*	*4*	*1*
f_i	*0.100*	*0.150*	*0.425*	*0.200*	*0.100*	*0.025*

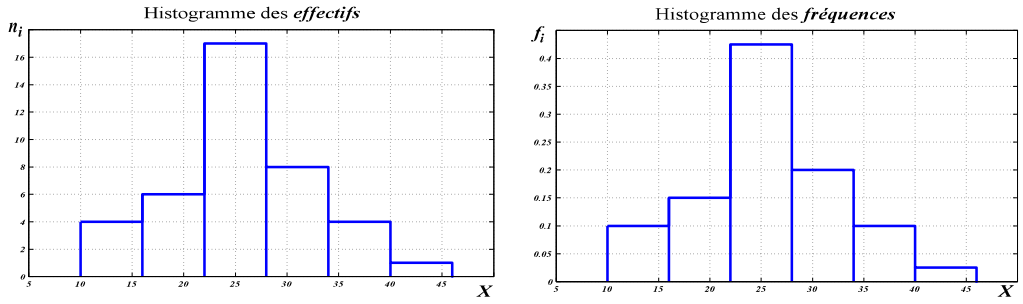


FIGURE 1.5: *Histogramme des effectifs et Histogramme des fréquences*

1.3.3.1 Séries statistiques à classes égales (même amplitude)

Chaque classe est représentée par un réctangle dont la base est *l'amplitude* et la hauteur au proportionnel à la fréquence (absolue ou relative).

Exemple 1.6 *On considère la répartition des ouvriers d'une entreprise selon le salaire mensuel (en DA). Les résultats sont donnés dans le tableaux suivant :*

Classes	n_i	f_i	$F_i^{\nearrow}$	$F_i^{\searrow}$
$[3,4[$	16	0.16	$0.16 = F(4)$	$1 = F(3)$
$[4,5[$	22	0.22	$0.38 = F(5)$	$0.84 = F(4)$
$[5,6[$	44	0.44	$0.88 = F(6)$	$0.62 = F(5)$
$[6,7[$	08	0.08	$0.90 = F(7)$	$0.18 = F(6)$
$[7,8[$	10	0.10	$1 = F(8)$	$0.10 = F(7)$
Total	100	1		

Le polygone des effectifs ou des fréquences est la ligne joignant les milieux de cotés superieur des réctangles.

Remarque 1.5

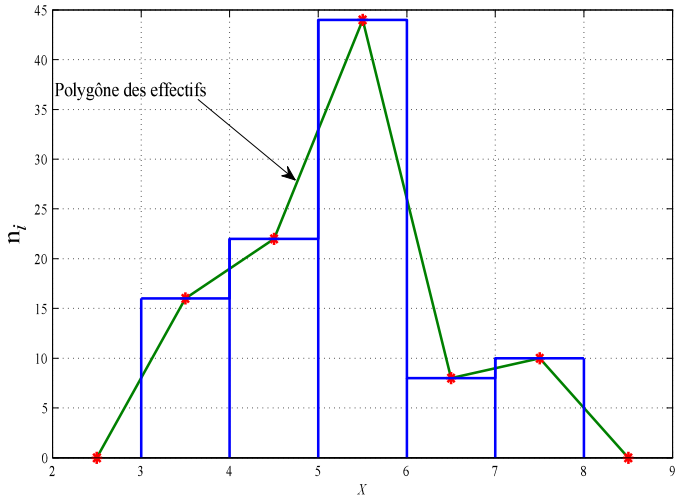


FIGURE 1.6: *Histogramme et polygone des effectifs*

— *L'aire de chaque réctangle, dans le cas des effectifs, est égale à* $n_i \times a$ *et l'aire de tout l'histogramme est* $\sum n_i \times a = a\sum n_i = a \times n$.
— *L'aire de chaque réctangle, dans le cas des fréquence, est égale à* $f_i \times a$ *et l'aire de tout l'histogramme est* $\sum f_i \times a = a\sum f_i = a$.

1.3.3.2 Séries statistiques à classes inégales (amplitudes différentes)

Chaque classe est représentée par un réctangle dont la base est égale à *l'amplitude* et *la hauteur* égale à la fréquence divisée par l'amplitude (fréquences corrigées).

Remarque 1.6 *Dans ce cas, l'aire de chaque réctangle est*

$$a \times \frac{n_i}{a_i} = n_i(corrigee),$$

est proportionnelle à la fréquence ou l'éffectif de la classe.
En pratique, pour simplifier les calculs, on choisit l'amplitude unité $a = pgcd\,(a_i)$ *ou bien "a" l'amplitude la plus répetée.*

Exemple 1.7 *La répartition d'un groupes d'individus par taille (cm) est donnée par le tableau suivant :*

Taille(cm)	n_i	f_i	a_i	$f_i = \frac{f_i}{a_i} \times a$ *(f_i corrigées)*	$F_i \searrow$
$[150, 160[$	33	0.33	10	0.165	$1 = F\,(x \leq 150)$
$[160, 170[$	14	0.14	10	0.070	$0.84 = F\,(160)$
$[170, 175[$	21	0.21	05	0.210	$0.62 = F\,(170)$
$[175, 180[$	16	0.16	05	0.160	$0.18 = F\,(175)$
$[180, 185[$	11	0.11	05	0.110	$0.10 = F\,(180)$
$[185, 190[$	05	0.05	05	0.050	*0.05=F(185)*
Total	100	1			*0*$= F(X \geq 190)$

$a = 5cm = pgcd(10, 5)$, *a est aussi l'amplitude la plus répétée.*

Soit $x \in [e_{i-1}, e_i[$, déterminons la valeur de la fonction de répartition empérique au point x i.e $F(x)$.

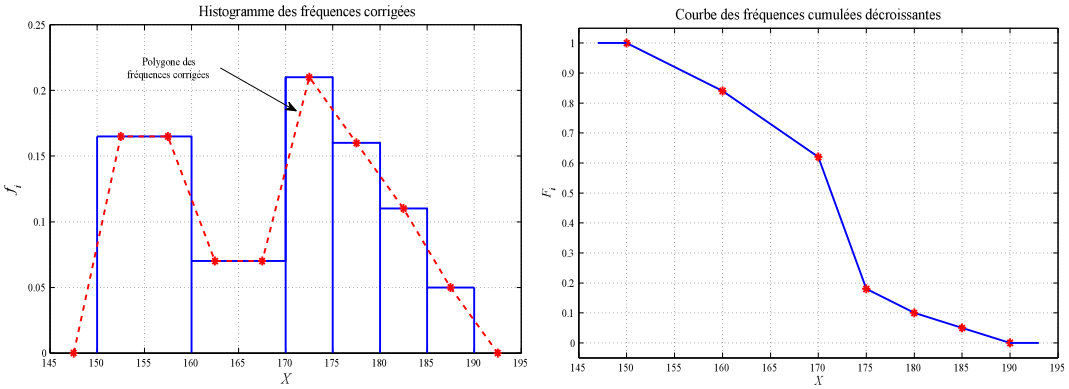


FIGURE 1.7: *Présentation graphique cas de classes inégales : Histogramme des fréquence corrigées (à gauche) et courbe des fréquences cumulées décroissante (à droite).*

où

$$\begin{aligned} tg\alpha &= \frac{F(e_i) - F(e_{i-1})}{e_{i-1} - e_i} = \frac{F(x) - F(e_{i-1})}{x - e_{i-1}} \\ &= \frac{f_i}{a_i} = \frac{F(x) - F(e_{i-1})}{x - e_{i-1}} \\ &\Rightarrow (x - e_{i-1}) \frac{f_i}{a_i} = F(x) - F(e_{i-1}). \end{aligned}$$

On aura :

$$F(x) = F(e_{i-1}) + \frac{x - e_{i-1}}{a_i} \times f_i.$$

D'où la fonction de répartition empérique d'une variable statistique continue de classes $[e_0, e_1[, ... [e_{k-1}, e_k[$ est

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0,1]. \\ x &\longrightarrow F(x), \end{aligned}$$

telle que

$$F(x) = \begin{cases} 0 & \text{si } x \leq e_0 \\ F(e_{i-1}) + \dfrac{x - e_{i-1}}{a_i} \times f_i & \text{si } x \in [e_{i-1}, e_i[\\ 1 & \text{si } x \geq e_k \end{cases}$$

Exemple 1.8 *Reprenons l'exemple de répartition des ouvries selon le salaire mensuel (en DA).*

Classes	f_i	$F_i^{\nearrow}$	$F_i^{\searrow}$
$[3,4[$	0.16	$0.16 = F(4)$	$1 = F(3)$
$[4,5[$	0.22	$0.38 = F(5)$	$0.84 = F(4)$
$[5,6[$	0.44	$0.88 = F(6)$	$0.62 = F(5)$
$[6,7[$	0.08	$0.90 = F(7)$	$0.18 = F(6)$
$[7,8[$	0.10	$1 = F(8)$	$0.10 = F(7)$
Total	1		

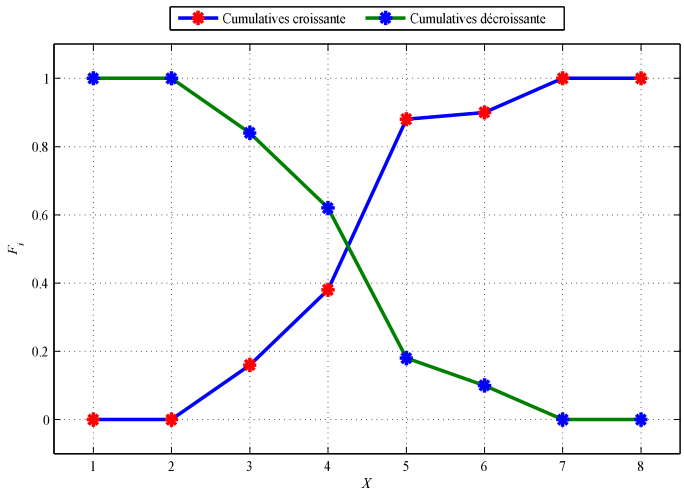


FIGURE 1.8: *Courbe des fréquences cumulatives croissantes et décroissantes*