Master 1 MRHDS

Traitement statistique des données

Cyrille Joutard et Jean-Michel Kosianski

Université Paul Valéry-Montpellier 3

Année universitaire 2012-2013

Organisation

1) 1h30 Cours - 1h30 TD par semaine

2) Informations et documents sur le site

   http://www.univ-montp3.fr/miap/ens/

   cliquer sur le lien “M1MRHDS: Traitement statistique des données”

Introduction

Les statistiques: Ensemble de données, d'observations relatives à des groupes d'individus, présentées sous forme de tableaux numériques, de graphiques ou synthétisées par des résumés numériques.

La statistique: Ensemble de méthodes ayant pour objectif la collecte, la présentation, le traitement de données d'une part, l'analyse de ces données, leur modélisation et la prise de décision d'autre part.

Ce cours portera principalement sur la **Statistique descriptive** (on s'intéresse à décrire et à synthétiser les données étudiées) $\neq$ **Statistique inférentielle** (on cherche à tirer des conclusions générales à partir des données observées sur un échantillon) $\hookrightarrow$ utilise notamment les probabilités.

Exemple 1: Notes d'étudiants et filières d'étude

On a noté dans le tableau ci-dessous la répartition de 100 étudiants d'une université par filière d'étude et par classe de notes obtenues à une épreuve d'un concours commun.

Note	$[0, 6[$	$[6, 10[$	$[10, 14[$	$[14, 20]$
Filière A	26	6	4	1
Filière B	12	9	3	1
Filière C	1	4	5	6
Filière D	10	8	3	1

Exemple 2: Sondage sur une mesure gouvernementale

Un échantillon est prélevé au hasard dans la population française pour sonder l'opinion sur une mesure du gouvernement. On a obtenu les résultats suivants: 689 sont opposés à la mesure, 894 indifférents et 417 favorables.

Exemple 3: Habitants d'une résidence de Montpellier

Dans une résidence de Montpellier comprenant 11 appartements, on s'intéresse au nombre de personnes habitant dans chaque appartement. On a obtenu les réponses suivantes:

$$1, 3, 1, 0, 2, 2, 4, 1, 3, 1, 2$$

Exemple 4: Produits d'une grande surface

Dans le tableau ci-dessous, on a réparti 700 produits provenant d'une grande surface française selon leur classe de prix (en euros).

Prix	]0, 10[	[10, 15[	[15,30[	[30,45[	[45,150]	Total
Effectifs	110	205	235	70	80	700

Exemple 5: Couleurs des yeux et des cheveux

On souhaite savoir s'il existe un lien entre la couleur des cheveux et la couleur des yeux chez les 445 élèves d'une école primaire.
On a relevé les effectifs suivants:

Couleur yeux \ cheveux	Brun	Noir	Roux	Blond
Marron	113	72	7	39
Gris-vert	38	41	10	27
Bleu	20	17	8	53

Chapitre I. Description d'une situation statistique

1) Population et individus

L'ensemble des **individus** sur lequel porte l'étude s'appelle la **population**. Comme en général, on ne peut pas étudier la population en entier, on en extrait une partie appelée **échantillon**. Les individus constituant l'échantillon sont donc extraits de la population étudiée. L'échantillon est censé être représentatif de cette population (note: le processus de sélection des individus de l'échantillon fait parti d'une branche de la statistique appelée échantillonnage qui ne sera pas étudiée dans ce cours).
Lorsque l'échantillon correspond à la population dans son entier on parle de recensement.

Pour désigner un **individu**, on parle aussi d'unité statistique. Notons qu'un individu (ou unité statistique) n'est pas nécessairement une personne (exemple: une voiture, une ville française, un appartement,...)

On notera n**=taille de l'échantillon** = nombre d'individus de l'échantillon = effectif total

Exemple 1: Notes d'étudiants et filières d'étude

Les individus sont des étudiants d'une université.

Même si ce n'est pas précisé on peut penser qu'il s'agit d'un échantillon (tous les étudiants n'ont pas été interrogés).

La taille de l'échantillon est $n = 100$.

Exemple 2: Avis sur une mesure gouvernementale

Les individus sont des personnes habitant en France.

Il s'agit d'un échantillon (a priori représentatif de la population française).

La taille de l'échantillon est $n = 2000$.

Exemple 3: Habitants d'une résidence

Les individus sont les appartements d'une résidence de Montpellier.
Il s'agit d'un recensement car on a étudié tous les appartements de la résidence. La taille de le la population (=échantillon) est $n = 11$.

Exemple 4: Produits d'une grande surface

Les individus sont des produits d'une grande surface française.
Il pourrait s'agir d'un échantillon mais ce n'est pas clairement spécifié.
La taille de l'échantillon est $n = 700$.

Exemple 5: Couleurs des yeux et des cheveux

Les individus sont les enfants d'une école primaire.
C'est un recensement.
La taille de la population (=échantillon) est $n = 445$.

2) Variable(s)

Le (ou les) caractère(s) étudié(s) et mesuré(s) sur les individus est (sont) appelé(s) la (ou les) **variable(s)**. On la (les) notera par des lettres majuscules: X, Y, Z.... On observe et mesure la (ou les) variable(s) sur chaque individu de l'échantillon.

Exemple de variables: âge, taille, salaire, note à un examen, nombre d'enfants, mention au bac, etc....

On appelle **modalités** les réponses faites par les individus à une variable. Pour un individu donné, on n'a qu'une seule réponse possible (par variable). On notera ces réponses par des lettres minuscules. Par exemple, x_2 représente la réponse faite par l'individu numéro 2 de l'échantillon à la variable X.

Si on étudie 2 variables X et Y (au lieu d'une seule), (x_5, y_5) représente les réponses faites par l'individu numéro 5 de l'échantillon aux 2 variables X et Y.

On fait la distinction entre l'ensemble des modalités observées et l'ensemble des modalitées observables (les réponses "possibles"). Il se peut très bien que toutes les réponses "possibles" n'aient pas été observées parmi les n individus constituant l'échantillon étudié.

En effet soit l'ensemble des modalités observables est infini (ex: modalités de la variable "taille" en cm) soit certaines réponses n'ont tout simplement pas été proposées par les n individus de l'échantillon.

On notera $\mathcal{M}_X$ l'ensemble des modalités de la variable X. On pourrait avoir par exemple:

- $\mathcal{M}_X = \{$"Passable", "Assez bien", "Bien", "Très bien"$\}$.
- $\mathcal{M}_X = \{1, 2, 3, 4, 5, 6\}$.
- $\mathcal{M}_X = [140, 200]$.
- $\mathcal{M}_X = \{$"musique", "théatre", "danse", "autre"$\}$.

Lorsque l'ensemble des modalités est fini (en particulier ce n'est pas un intervalle comme dans l'exemple $\mathcal{M}_X = [140, 200]$), on note K son cardinal (c'est-à-dire le nombre d'éléments de l'ensemble). Dans le cas général, on aura alors $\mathcal{M}_X = \{m_1, \ldots, m_K\}$.

Si $K = 2$ (2 modalités pour la variable), on dit que la variable est **dichotomique** (exemple: réponse (Oui/Non) à une question).

Nous nous intéressons à présent à la nature des ces variables (qui dépend de la structure de l'ensemble des modalités):

- **Variable qualitative**: Lorsque les modalités d'une variable sont des mots (ou des "codes" utilisés pour désigner des mots), on dit que la variable est qualitative. Dans ce cas les modalités sont aussi appelées niveaux (de la variable). Il existe 2 types de variables qualititatives :

  Variable qualitative nominale: Il n'existe pas d'ordre naturel entre les différentes modalités (ou niveaux) de la variable (ex: catégories socio-professionnelles, couleurs des cheveux,....)

  Variable qualitative ordinale: Il existe un ordre naturel entre les modalités et il est donc possible d'ordonner ces modalités les unes par rapport aux autres (ex: mentions au bac, niveaux de difficulté d'un test,...)

- **Variable quantitative**: Les modalités d'une variables quantitative sont des nombres (obtenus par comptage, mesure, etc..). On parle alors de valeurs (de la variable) plutôt que de modalités. Il existe 2 types de variables quantitatives :

  Variable quantitative discrète: les valeurs sont en général des nombres entiers obtenus par dénombrement (ex: Nombre d'enfants d'une famille, Nombre d'étages d'un immeuble, etc..)

  Variable quantitative continue: les valeurs sont en général des nombres décimaux lus par un instrument de mesure réel ou imaginaire. Elles s'expriment souvent dans une unité (ex: taille en cm, salaire en euros, durée en min,...). Notons qu'entre deux valeurs distinctes, il existe en théorie une infinité de valeurs possibles.

Exemple 1: Notes d'étudiants et filières d'étude

X: "Filière d'étude"

Y: "Note"

$\mathcal{M}_X = \{A, B, C, D\}$.

Variable qualitative nominale

$\mathcal{M}_Y = [0, 20]$.

Variable quantitative continue

Exemple 2: Avis sur une mesure gouvernementale

X: "Avis"

$\mathcal{M}_X = \{$"Opposé", "Indifférent", "Favorable"$\}$.

Variable qualitative ordinale

Exemple 3: Habitants d'une résidence

X: "Nombre de personnes (habitant dans un appartement)"

$\mathcal{M}_X = \{0, 1, 2, 3, 4\}$.

Variable quantitative discrète

Exemple 4: Produits d'une grande surface

X: "Prix"

$\mathcal{M}_X =]0, 150]$.

Variable quantitative continue

Exemple 5: Couleurs des yeux et des cheveux

X: "Couleur des yeux"

Y: "Couleur des cheveux"

$\mathcal{M}_X = \{$"Marron", "Gris-vert", "Bleu"$\}$

Variable qualitative nominale

$\mathcal{M}_Y = \{$"Blond", "Brun", "Noir", "Roux"$\}$

Variable qualitative nominale

3) Données

On s'intéresse maintenant à la manière dont les données (l'ensemble des réponses des individus) sont présentées. En général, les données dont on dispose ont déjà subi des transformations pour les rendre plus "lisibles". Les données dans leur forme "originelle" (aucune transformation n'a été effectuée) sont appelées **données brutes**. Elles peuvent se présenter sous la forme d'un tableau avec en lignes les individus et en colonnes les variables. Dans le cas d'une seule variable X, on aurait alors par exemple:

Numéro de l'individu (identifiant)	Variable X
1	m_3
2	m_2
3	m_4
4	m_1
$\vdots$	$\vdots$
n	m_1

Exemple 2: Avis sur une mesure gouvernementale: Le tableau des données brutes pouvait se présenter sous la forme:

Numéro de l'individu	Avis
1	Indifférent
2	Indifférent
3	Opposé
4	Favorable
⋮	⋮
n	Opposé

Les données brutes pourraient aussi être simplement présentées sous la forme d'une liste de n réponses (1 par individu) à la variable X. Par exemple, on pourrait avoir: x_1 = Indifférent, x_2 = Indifférent, x_3 = Opposé, x_4 = Favorable, . . ., x_n= Opposé,
ou encore plus simplement:
Indifférent, Indifférent, Opposé, Favorable, . . ., Opposé.

Exemple 1: Notes d'étudiants et filières d'étude

X = Filière d'étude (variable qualitative nominale).
Y = Note obtenue à l'épreuve (variable quantitative continue).

Les ensembles des modalités pour X et Y sont $\mathcal{M}_X = \{A, B, C, D\}$ et $\mathcal{M}_Y = [0, 20]$. Le tableau des données brutes pouvait se présenter sous la forme:

Numéro de l'étudiant	Filière (X)	Note (Y)
1	B	10.5
2	A	4.5
3	B	12
4	D	8.5
⋮	⋮	⋮
n	C	14

Souvent, on regroupe les individus qui ont donné la même réponse (c'est-à-dire la même modalité) et on compte leur nombre. Donc pour chaque modalité (dans le cas d'un ensemble de modalités de cardinal fini), on obtient l'effectif associé, c'est-à-dire le nombre d'individus (dans l'échantillon) ayant choisi cette modalité. Pour une variable X avec K modalités, on a alors le tableau suivant donnant la répartition des individus selon les K modalités:

Variable X	m_1	m_2	$\cdots$	m_K
Effectifs n_k	n_1	n_2	$\cdots$	n_K

Ce tableau donne la **distribution en effectifs** de la variable X, notion qui fait l'objet du chapitre suivant.

En résumé, lorsque l'on veut décrire une situation statistique, on doit clairement identifier:

- Les **individus** (qui est l'objet de l'étude?) et éventuellement la **population** et l'**échantillon** (en donnant la **taille de l'échantillon** et en précisant si possible s'il s'agit d'un recensement).
- La (ou les) **variable(s)** (sur quoi porte l'étude?) en précisant leur **nature** et leurs **modalités**.
- Les **données** (comment se présente le relevé des observations? données brutes? tableau de distribution en effectifs? etc...).

Chapitre II. Distribution et distribution cumulée

1) Distribution

Soit X une variable observée sur un échantillon de taille n. On note $x_1, x_2, \ldots, x_n$ les réponses (appartenant à l'ensemble des modalités $\mathcal{M}_X$) données par les n individus de l'échantillon à X. On appelle distribution de X la répartition des réponses suivant les modalités de X.

1.1) Distribution en effectifs

Pour chaque modalité, on compte le nombre d'individus ayant pour réponse cette même modalité (ceci revient à compter les répétitions dans les réponses données par les n individus). On notera m_k **la kème modalité** de X et n_k **l'effectif** associé à cette modalité (c'est-à-dire le nombre d'individus dont la réponse à X est m_k).

Remarquons que

$$\sum_{k=1}^{K} n_k = n_1 + n_2 + \ldots + n_K = n,$$

c'est-à-dire la somme des effectifs est égale à n.

On a le tableau suivant pour la **distribution en effectifs** de X:

Variable X	m_1	m_2	$\cdots$	m_K	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n

1.2) Distribution en fréquences

On s'intéresse à la proportion des différents effectifs par rapport à l'effectif total n (taille de l'échantillon). Il est ainsi plus facile de comparer et d'interpréter.

Notons f_k **la fréquence** associée à la kème modalité:

$$f_k = \frac{n_k}{n}.$$

Une fréquence est toujours comprise entre 0 et 1. On peut aussi l'écrire sous forme de pourcentage (ex: $0.35 = 35\%$).

On a $\sum_{k=1}^{K} f_k = 1$, c'est-à-dire la somme des fréquences est égale à 1 (ou 100% s'il s'agit de pourcentages).

Le tableau ci-dessous donne la **distribution en fréquences** de X:

Variable X	m_1	m_2	$\cdots$	m_K	Total
Fréquences f_k	f_1	f_2	$\cdots$	f_K	1

1.3) Cas particulier de la variable quantitative continue

Vu l'infinité des valeurs (ou modalités) observables pour une variable quantitative continue, il n'y a pas de répétitions et donc pas de regroupement immédiat par modalité. On va former des classes de valeurs, c'est-à-dire découper l'ensemble des modalités $\mathcal{M}$ en intervalles (classes) successifs. Pour ceci on doit :

- choisir le nombre de classes (ou intervalles) K.
- choisir la borne inférieure et la borne supérieure de chaque intervalle.

Il ne doit pas y avoir d'espace vide entre 2 intervalles (ou classes) successifs, la borne supérieure d'une classe est la borne inférieure de la classe qui suit. En même temps les deux intervalles doivent être disjoints (exemple: $[1, 3[$ et $[3, 6[$).

On parle de regroupement en classes de la variable. Notons que ce regroupement en classes implique forcément une perte d'information vu qu'on ne connaît plus la valeur exacte de chaque observation (ou réponse observée pour chaque individu) mais seulement son appartenance à une classe.

On a le tableau suivant pour la distribution (en effectifs et fréquences) de X:

Variable X	$[b_0; b_1[$	$[b_1; b_2[$	$\cdots$	$[b_{K-1}; b_K[$	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n
Fréquences f_k	f_1	f_2	$\cdots$	f_K	1

En ce qui concerne le nombre de classes à choisir, on doit faire attention à ce qu'il ne soit pas trop petit pour éviter une trop grande perte d'information mais pas trop grand non plus pour que l'information reste assez lisible.

2) Distribution cumulée

On parle de **distribution cumulée** uniquement lorsqu'il existe un ordre naturel sur les modalités (c'est le cas des variables quantitatives discrètes ou continues et des variables qualitatives ordinales mais pas des variables qualitatives nominales). Lorsqu'on peut donc ranger les modalités selon un ordre, il s'agit de cumuler les effectifs (ou les fréquences) selon l'ordre croissant des modalités.

Etudier la distribution cumulée d'une variable permet de répondre à des questions du type:

Quelle est la proportion d'individus dont la réponse est inférieure à...?

Quelle est la proportion d'individus dont la réponse est comprise entre... et ...?

Quelle est la proportion d'individus dont la réponse est supérieure à...?

Pour la kème modalité, les formules sont les suivantes:

Effectifs cumulés N_k: $N_k = n_1 + n_2 + \ldots + n_k$.

Fréquences cumulées F_k: $F_k = f_1 + f_2 + \ldots + f_k$.

2.1) Cas des variables qualitatives ordinales

Les modalités sont rangées selon un ordre naturel.

$$m_1 < m_2 < \ldots < m_k < \ldots < m_K$$

On a alors le tableau suivant pour la distribution et la distribution cumulée (en effectifs et en fréquences) de X:

Variable X	m_1	m_2	$\cdots$	m_K	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n
Fréquences f_k	f_1	f_2	$\cdots$	f_K	1
Effectifs cumulés N_k	$N_1 = n_1$	N_2	$\cdots$	$N_K = n$	
Fréquences cumulées F_k	$F_1 = f_1$	F_2	$\cdots$	$F_K = 1$	

2.2) Cas des variables quantitatives discrètes

Les valeurs peuvent bien entendu être rangées dans l'ordre.

$$v_1 < v_2 < \ldots < v_k < \ldots < v_K$$

Le tableau ci-dessous donne la distribution et la distribution cumulée (en effectifs et en fréquences) de X:

Variable X	v_1	v_2	$\cdots$	v_K	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n
Fréquences f_k	f_1	f_2	$\cdots$	f_K	1
Effectifs cumulés N_k	$N_1 = n_1$	N_2	$\cdots$	$N_K = n$	
Fréquences cumulées F_k	$F_1 = f_1$	F_2	$\cdots$	$F_K = 1$	

2.3) Cas des variables quantitatives continues

Les classes sont rangées dans l'ordre.

$$[b_0; b_1[< [b_1; b_2[< \cdots < [b_{K-1}; b_K[$$

Le tableau ci-dessous donne la distribution et la distribution cumulée (en effectifs et en fréquences) de X:

Variable X	|	$[b_0; b_1[$	|	$[b_1; b_2[$	|	$\cdots$	|	$[b_{K-1}; b_K[$	|	Total
Bornes	b_0		b_1		b_2	$\cdots$	b_{K-1}		b_K		
Effectifs n_k	|	n_1	|	n_2	|	$\cdots$	|	n_K	‖	n
Fréquences f_k	|	f_1	|	f_2	|	$\cdots$	|	f_K	‖	1
Eff. cum. N_k	0		N_1		N_2	$\cdots$	N_{K-1}		n		
Fréq. cum. F_k	0		F_1		F_2	$\cdots$	F_{K-1}		1		

Note: On écrit en général les effectifs (et les fréquences) cumulées au niveau des bornes des classes.

3) Exemples

Exemple 1: Notes d'étudiants et filières d'étude

Les deux tableaux ci-dessous donnent d'une part la distribution (en effectifs et fréquences) de la variable X = "Filière d'étude" (ou distribution marginale de X, voir chapitre III) et d'autre part la distribution et la distribution cumulée de la variable Y = "Note obtenue à l'épreuve"

Filière	A	B	C	D	Total
Effectifs n_k	37	25	16	22	100
Fréquences f_k (%)	37	25	16	22	100

Note	$[0,6[$	$[6,10[$	$[10,14[$	$[14,20]$	Total
Bornes	0 6	10	14	20	
n_k	49	27	15	9	100
f_k (%)	49	27	15	9	100
N_k	0 49	76	91	100	
F_k (%)	0 49	76	91	100	

Exemple 2: Avis sur une mesure gouvernementale

Le tableau ci-dessous donne la distribution (en effectifs et fréquences) et la distribution cumulée (en effectifs et fréquences) de X="Avis".

Avis	Opposé	Indifférent	Favorable	Total
Effectifs n_k	689	894	417	2000
Fréquences f_k	0.3445	0.447	0.2085	1
Eff. cum. N_k	689	1583	2000	
Fréq. cum. F_k	0.3445	0.7915	1	

Exemple 3: Habitants d'une résidence

On représente dans le tableau ci-dessous la distribution et la distribution cumulée (en effectifs et en fréquences) de X="Nbre de personnes (habitant dans un appartement)".

Nbre de personnes	0	1	2	3	4	Total
Effectifs n_k	1	4	3	2	1	11
Fréquences f_k	0.091	0.364	0.273	0.182	0.091	1
Eff. cum. N_k	1	5	8	10	11	
Fréq. cum. F_k	0.091	0.455	0.728	0.910	1	

Exemple 4: Produits d'une grande surface

On donne dans le tableau ci-dessous la distribution et la distribution cumulée (en effectifs et en fréquences) de X="Prix".

X	]0, 10[	[10, 15[	[15,30[	[30,45[	[45,150]	Tot.
b_k	0	10	15	30	45	150	
n_k	110	205	235	70	80	700
f_k (%)	15.7	29.3	33.6	10.0	11.4	100
N_k	0	110	315	550	620	700	
F_k (%)	0	15.7	45	78.6	88.6	100	

4) Représentations graphiques

Dans cette partie du cours, nous donnons une représentation graphique de la répartiton des individus selon leurs modalités.

Pour chaque type de variable, nous présentons les principaux éléments permettant de représenter graphiquement la distribution et lorsque cela est possible la distribution cumulée.

4.1) Variable qualitative nominale

La distribution de X est donnée dans le tableau suivant

Variable X	m_1	m_2	$\cdots$	m_K	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n
Fréquences f_k	f_1	f_2	$\cdots$	f_K	1

Distribution: diagramme en barres séparées

- On trace un axe horizontal portant le nom de la variable et on y positionne les modalités de la variable (ici l'ordre et la distance entre modalités n'ont pas de sens, l'axe n'est pas orienté).
- Sur l'axe vertical, on place les effectifs ou les fréquences après avoir choisi une échelle (l'axe est orienté).
- Au dessus de chaque modalité, on trace un trait ou un rectangle (la largeur n'a pas de signification) vertical de hauteur égale à l'effectif (ou à la fréquence) correspondant.

4.2) Variable qualitative ordinale

La distribution et la distribution cumulée de X sont données ci-dessous:

Variable X	m_1	m_2	$\cdots$	m_K	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n
Fréquences f_k	f_1	f_2	$\cdots$	f_K	1
Effectifs cumulés N_k	$N_1 = n_1$	N_2	$\cdots$	$N_K = n$	
Fréquences cumulées F_k	$F_1 = f_1$	F_2	$\cdots$	$F_K = 1$	

Distribution: diagramme en barres juxtaposées

- Sur un axe horizontal portant le nom de la variable, on positionne les modalités de la variable entre deux délimiteurs répartis régulièrement.
- Sur l'axe vertical, on place les effectifs ou les fréquences après avoir choisi une échelle (l'axe est orienté).
- Au dessus de chaque modalité, on trace un rectangle vertical (la base du rectangle correspond à la modalité entre 2 délimiteurs) de hauteur égale à l'effectif (ou à la fréquence) correspondant.

Distribution cumulée: graphe des fréquences cumulées

- On trace un axe horizontal portant le nom de la variable et on y positionne les modalités de la variable (dans l'ordre) entre deux délimiteurs répartis régulièrement.
- Sur l'axe vertical, on positionne les fréquences cumulées en respectant l'échelle choisie (l'axe est orienté de 0 à 1).
- Au niveau de chaque délimiteur, on place un point correspondant à sa fréquence cumulée.
- On relie les points par des morceaux de droite.

4.3) Variable quantitative discrète

La distribution et la distribution cumulée de X sont données ci-dessous:

Variable X	v_1	v_2	$\cdots$	v_K	Total
Effectifs n_k	n_1	n_2	$\cdots$	n_K	n
Fréquences f_k	f_1	f_2	$\cdots$	f_K	1
Effectifs cumulés N_k	$N_1 = n_1$	N_2	$\cdots$	$N_K = n$	
Fréquences cumulées F_k	$F_1 = f_1$	F_2	$\cdots$	$F_K = 1$	

Distribution: Diagramme en bâtons

- On trace un axe horizontal portant le nom de la variable et on y positionne les valeurs (modalités) de la variable après avoir choisi une échelle (l'axe est orienté).
- Sur l'axe vertical, on place les effectifs ou les fréquences en respectant l'échelle choisie (l'axe est orienté).
- Au dessus de chaque modalité on trace un bâton de longueur égale à l'effectif (ou à la fréquence) correspondant.

Distribution cumulée: graphe de la fonction de répartition empirique
$\hookrightarrow$ graphe en escalier

- On positionne les valeurs (modalités) de la variable sur l'axe horizontal en respectant l'échelle choisie (l'axe est orienté).
- Sur l'axe vertical, on place les fréquences cumulées en respectant l'échelle choisie (l'axe est orienté de 0 à 1).
- A chaque valeur, on associe un point correspondant à sa fréquence cumulée.
- On trace des morceaux de droite horizontale (le cumul se fait par des sauts pour chaque valeur).

4.4) Variable quantitative continue

Distribution: Histogramme

Dans le cas d'une variable quantitative continue, on doit prendre en compte l'amplitude (ou largeur) des classes.
Par exemple, "15 individus sont âgés entre 20 et 30 ans" n'a pas du tout la même signification que "15 individus sont âgés entre 20 et 22 ans". Les effectifs (et fréquences) sont les mêmes mais dans le 2ème cas la concentration des observations est beaucoup plus forte.

Notons a_k **l'amplitude** de la kème classe.

Nous allons donc calculer la **densité de fréquence** associée à chaque classe:

$$d_k = \frac{f_k}{a_k}$$

Nous pouvons alors compléter le tableau de distribution en effectifs et en fréquences comme ceci :

Variable X	$\vert$	$[b_0; b_1[$	$\vert$	$[b_1; b_2[$	$\vert$	$\cdots$	$\vert$	$[b_{K-1}; b_K[$	$\vert$	Total
Bornes	b_0		b_1		b_2	$\cdots$	b_{K-1}		b_K	
Effectifs n_k	$\vert$	n_1	$\vert$	n_2	$\vert$	$\cdots$	$\vert$	n_K	$\Vert$	n
Fréquences f_k	$\vert$	f_1	$\vert$	f_2	$\vert$	$\cdots$	$\vert$	f_K	$\Vert$	1
Eff. cum. N_k	0		N_1		N_2	$\cdots$	N_{K-1}		n	
Fréq. cum. F_k	0		F_1		F_2	$\cdots$	F_{K-1}		1	
Amplitudes a_k	$\vert$	a_1	$\vert$	a_2	$\vert$	$\cdots$	$\vert$	a_K	$\Vert$	
Densités d_k	$\vert$	d_1	$\vert$	d_2	$\vert$	$\cdots$	$\vert$	d_K	$\Vert$	

- On trace un axe horizontal portant le nom de la variable et on y positionne les bornes des classes après avoir choisi une échelle (l'axe est orienté).
- Sur l'axe vertical, on place les densités après avoir choisi une échelle (l'axe est orienté).
- Pour chaque modalité, on trace un rectangle de hauteur correspondant à sa densité de fréquence et de largeur correspondant à son amplitude.

Notons que la surface de chaque rectangle, égale à la largeur multipliée par la hauteur (donc ici l'amplitude multipliée par la densité), représente la fréquence associée à chaque classe.

Distribution cumulée: graphe de la fonction de répartition empirique
$\hookrightarrow$graphe linéaire par morceaux

- On trace un axe horizontal portant le nom de la variable et on y place les bornes des classes après avoir choisi une échelle (l'axe est orienté).
- Sur l'axe vertical, on place les fréquences cumulées en respectant l'échelle choisie (l'axe est orienté de 0 à 1).
- Au niveau de chaque borne on place un point correspondant à sa fréquence cumulée.
- On relie les points par des morceaux de droite.

4.5) Exemples

Exemple 1: Notes d'étudiants et filières d'étude

Tableau de la distribution (en effectifs et fréquences) de la variable X = "Filière d'étude" (ou distribution marginale de X, voir chapitre III)

Filière	Filière A	Filière B	Filière C	Filière D	Total
Effectifs n_k	37	25	16	22	100
Fréq. f_k (%)	37	25	16	22	100

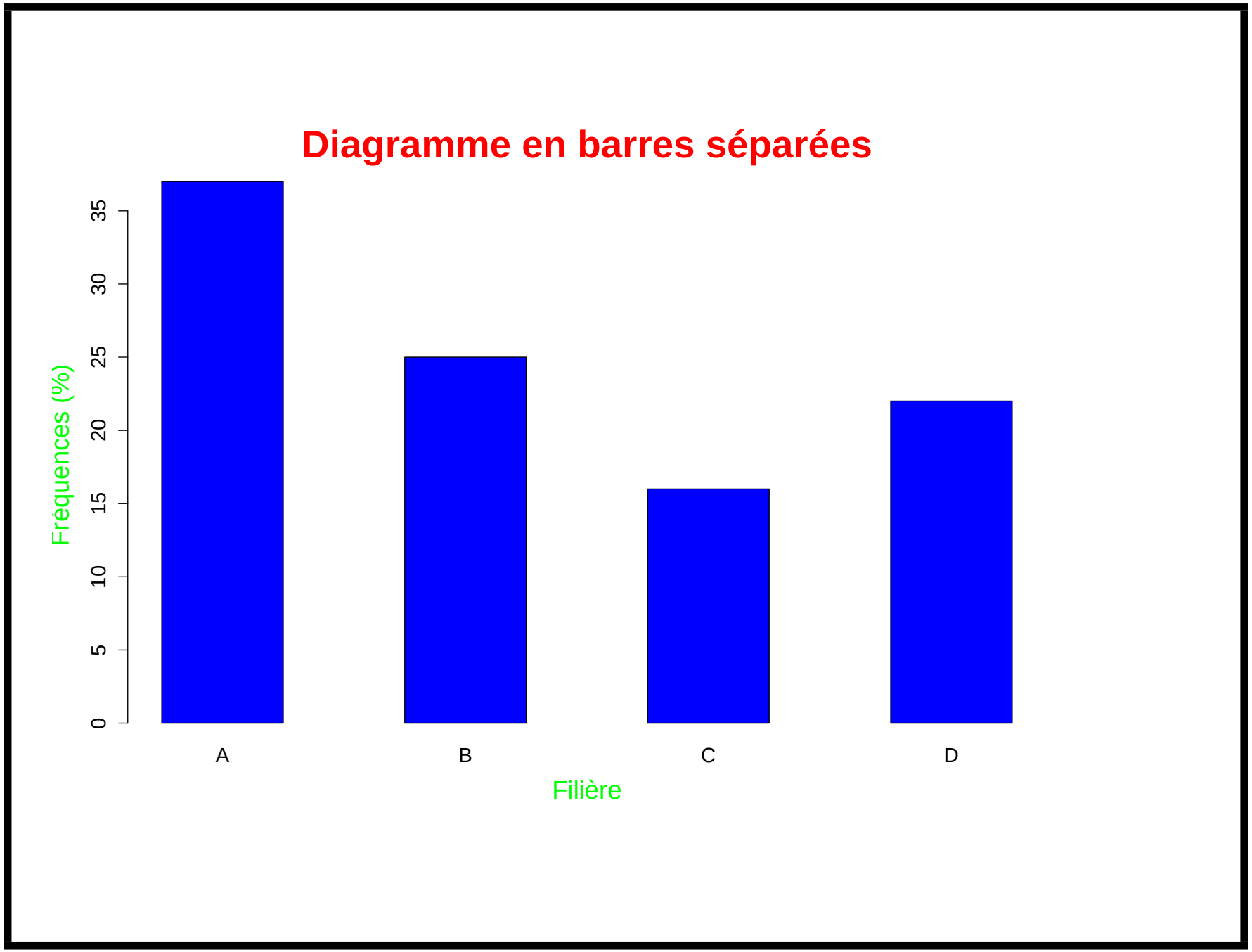


Diagramme en barres séparées
Fréquences (%)
0
5
10
15
20
25
30
35
A
B
C
D
Filière

Exemple 2: Avis sur une mesure gouvernementale

Tableau de la distribution et de la distribution cumulée de X="Avis".

Avis	Opposé	Indifférent	Favorable	Total
Effectifs n_k	689	894	417	2000
Fréquences f_k	0.3445	0.447	0.2085	1
Eff. cum. N_k	689	1583	2000	
Fréq. cum. F_k	0.3445	0.7915	1	

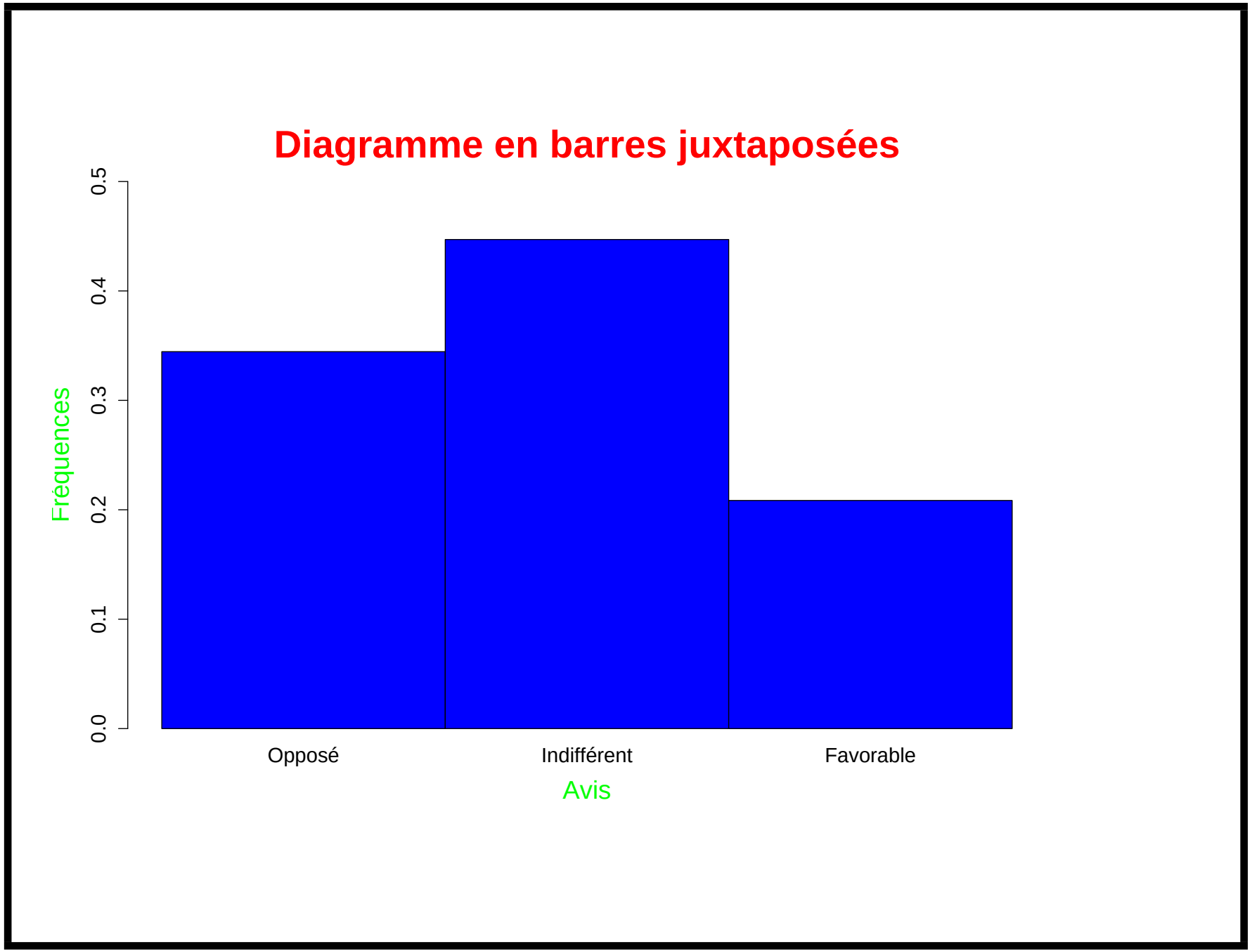
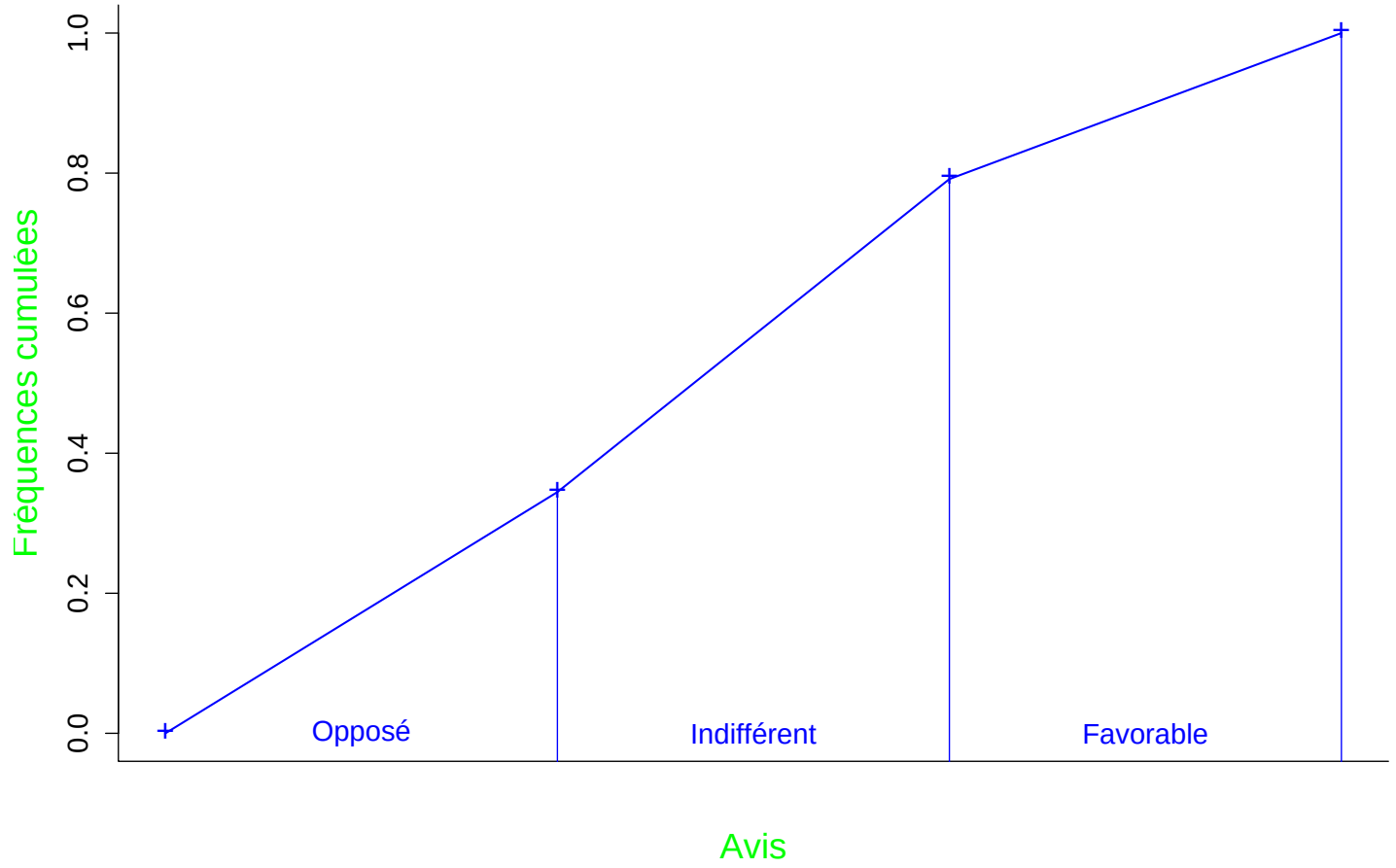


Diagramme en barres juxtaposées
Fréquences
0.0
0.1
0.2
0.3
0.4
0.5
Opposé
Indifférent
Favorable
Avis

Graphe des fréquences cumulées

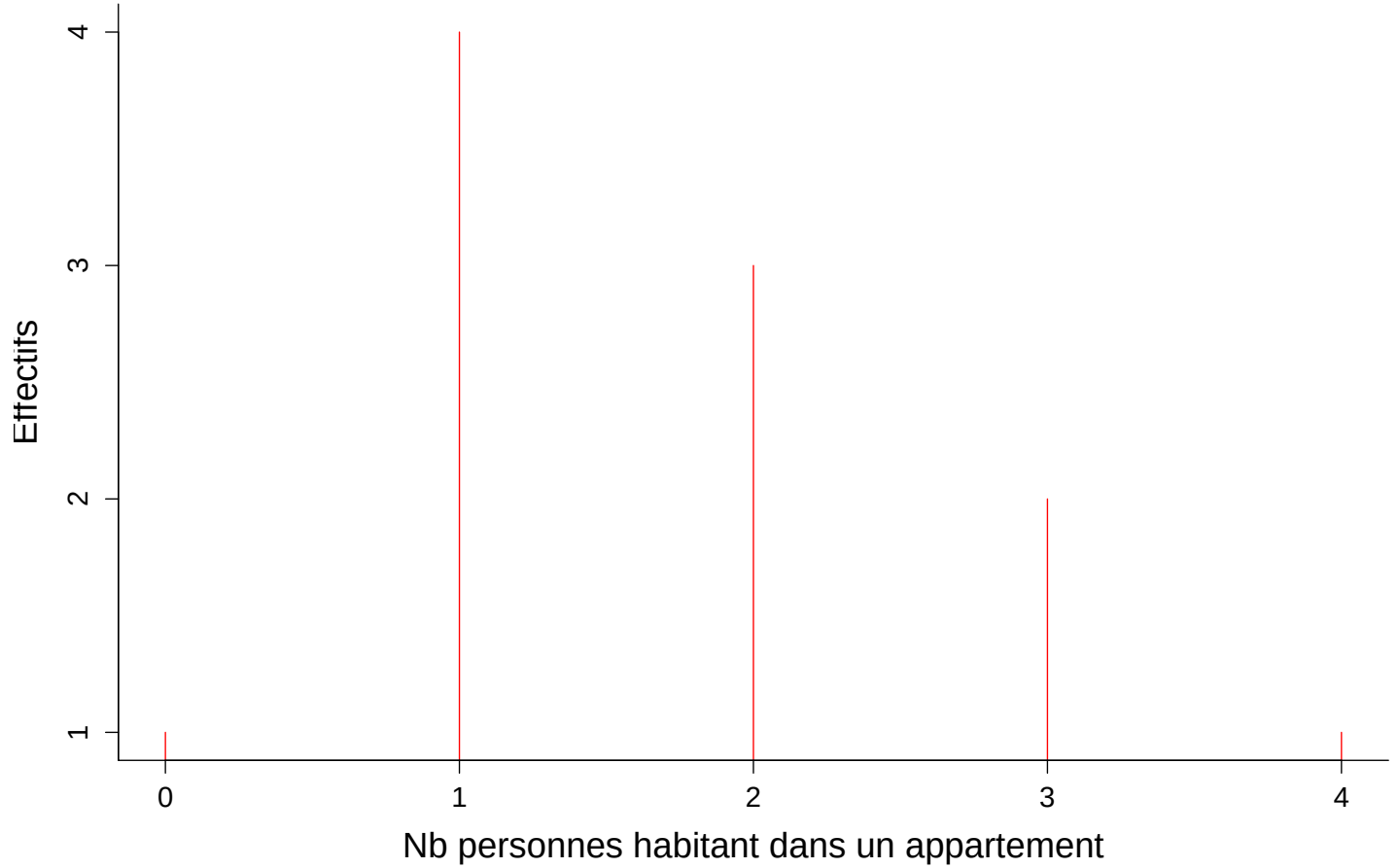


Exemple 3: Habitants d'une résidence

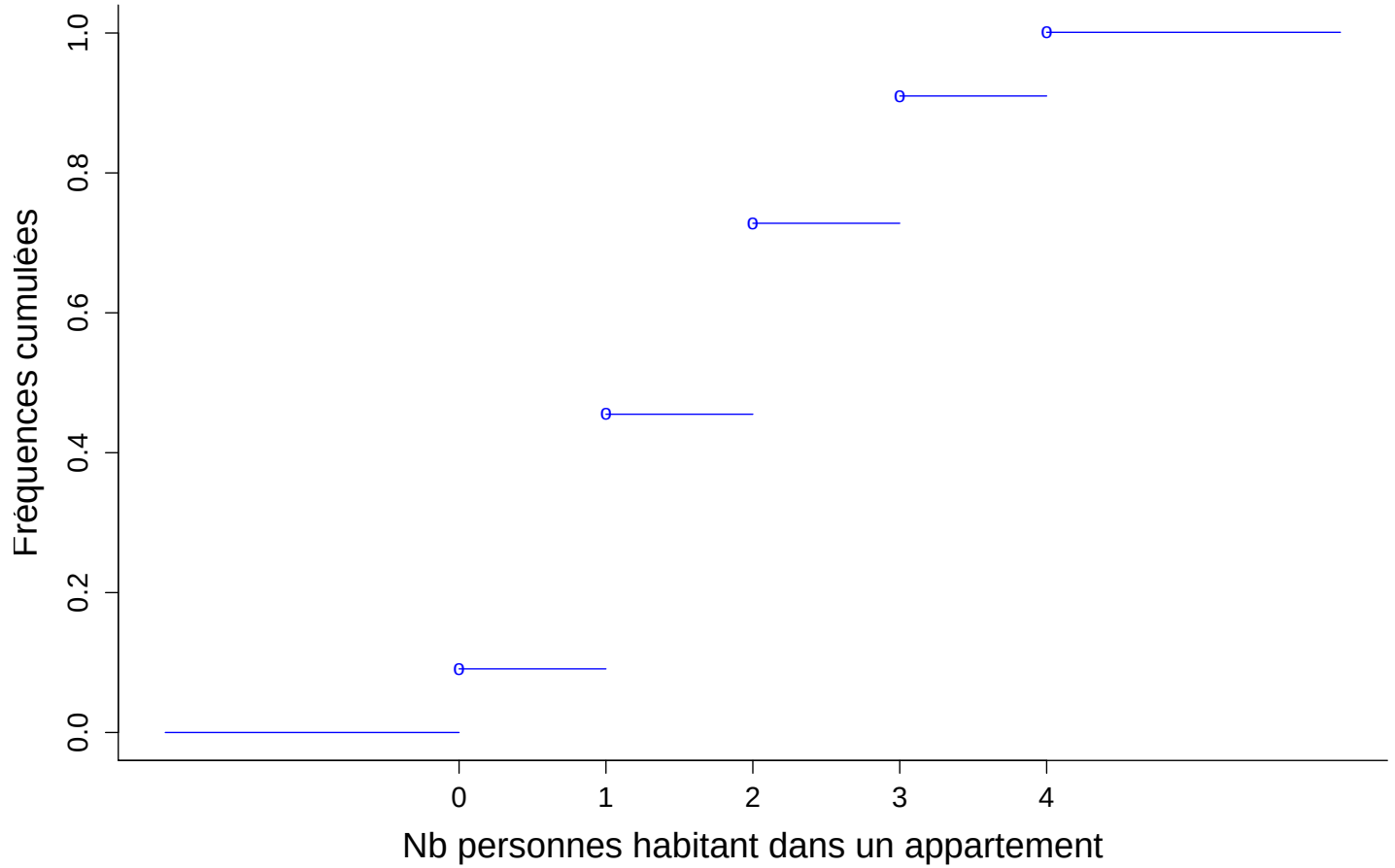
Tableau de la distribution et de la distribution cumulée de X="Nbre de personnes (habitant dans un appartement)".

Nbre de personnes	0	1	2	3	4	Total
Effectifs n_k	1	4	3	2	1	11
Fréquences f_k	0.091	0.364	0.273	0.182	0.091	1
Eff. cum. N_k	1	5	8	10	11	
Fréq. cum. F_k	0.091	0.455	0.728	0.910	1	

Diagramme en batons de la distribution en effectifs



Graphe de la fonction de répartition

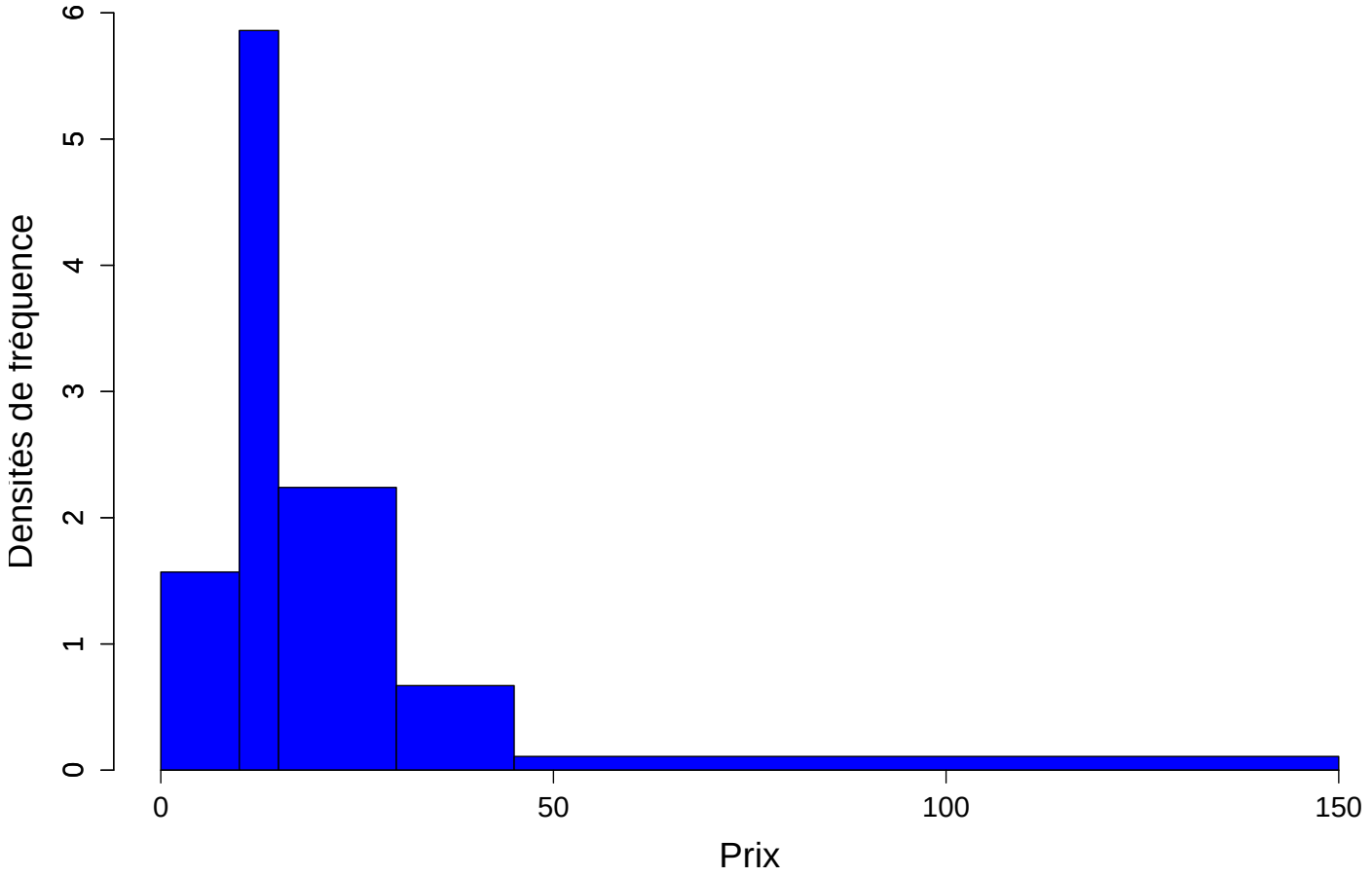


Exemple 4: Produits d'une grande surface

Tableau de la distribution, de la distribution cumulée de X="Prix" et des densités de fréquences (calculées grâce à la formule $d_k = f_k/a_k$).

Prix	]0, 10[	[10, 15[	[15,30[	[30,45[	[45,150]	Total
Bornes	0	10	15	30	45	150	
n_k	110	205	235	70	80	700
f_k (%)	15.7	29.3	33.6	10.0	11.4	100
N_k	0	110	315	550	620	700	
F_k (%)	0	15.7	45	78.6	88.6	100	
a_k	10	5	15	15	105	/
d_k	1.57	5.86	2.24	0.67	0.109	/

Histogramme de la distribution



Graphe de la fonction de répartition