E221XS3
Statistique pour les SHS en licence 1

C. Joutard, C. Lavergne, L. Piccinini & C. Trottier

Université Paul Valéry - Montpellier 3

Année universitaire 2012-2013

Organisation

- Informations et documents (à imprimer) sur le site
  `http://www.univ-montp3.fr/miap/ens/`
  lien "StatL1S2"

Introduction

La statistique est une discipline constituée d'un ensemble de méthodes visant d'une part à :

1. collecter
2. organiser
3. présenter
4. synthétiser

une **information**, et d'autre part à :

1. analyser cette information
2. modéliser le phénomène observé
3. tirer des conclusions
4. prendre des décisions

Ce cours en L1 va se focaliser sur la première partie qui constitue la statistique descriptive.

La statistique est une discipline qui permet de mettre en évidence des phénomènes tout en ne fournissant en aucun cas d'explication.
L'explication, l'interprétation ... sont l'affaire du praticien : psychologue, médecin, sociologue ...

Elle est un outil précieux d'aide à l'analyse, qu'il est nécessaire de connaître suffisamment pour s'en servir. Il s'agit de comprendre ses principales notions, la logique qui les sous-tend pour mettre en œuvre des techniques sans pour autant en connaître les détails des fondements mathématiques.

Le mot statistique(s) recouvre différentes réalités.

Il est utilisé dans le langage courant au **pluriel** pour désigner des chiffres, des tableaux. On parle des statistiques du chômage, des accidents de la route ... Elles désignent alors directement les observations faites que l'on appelle les données.

Dans le langage courant, elles peuvent aussi désigner toutes les quantités calculées à partir de ces données : moyennes, pourcentages ...

Dans ce cours, on utilisera le **singulier** pour désigner la discipline elle-même.

Exemple 1 : Accueil des jeunes handicapés

On s'intéresse à la répartition des établissements d'accueil d'enfants et jeunes handicapés en Languedoc-Roussillon au 1er janvier 2005 selon le type de handicap des personnes majoritairement accueillies.

Type de handicap	**Nombre d'établissements**
Déficients mentaux	43
Polyhandicapés	16
Troubles du comportement	17
Handicapés moteurs	3
Handicapés sensoriels	5

Exemple 2 : Évaluation de la difficulté d'un examen

À la sortie d'un examen de statistique en L2 de psychologie à l'université Paul Valéry - Montpellier 3, on a demandé à des étudiants sélectionnés au hasard d'évaluer la difficulté de l'épreuve selon 4 catégories : Très difficile (1), Difficile (2), Facile (3), Très facile (4). On a obtenu les réponses suivantes :

3	2	2	1	3	2	2	4	1	2
1	1	3	1	3	2	2	2	3	2
3	3	3	3	1	2	2	3	2	3
4	3	2	2	1	1	1	2	2	2

Exemple 3 : La démence sénile

La démence sénile est définie par un déclin significatif des capacités intellectuelles, comme la mémoire ou le raisonnement. Environ 10% de la population des 65 ans et plus montrent une telle détérioration. Certains cas de démence sénile sont connus sous le nom de maladie d'Alzheimer.

En aout 1994, le Dr David Masur a mené une étude visant à évaluer les capacités prédictives d'une batterie de tests psychologiques pour la survenue d'une démence sénile dans le futur proche.

Des personnes en bonne santé agées de plus de 60 ans, libres de suivre des traitements, ont été soumis à une série de tests. Les tests ont été gradués sur une échelle de scores : faible, modéré et élevé. Les sujets ont été ensuite suivis sur 4 années pour déterminer si des symptômes cliniques de la démence sénile sont apparus.

Le tableau suivant résume les résultats de l'étude :

Score aux tests	**État clinique**		
	Pas de démence	Démence	
Faible	2	11	13
Modéré	49	43	92
Élevé	202	10	212
	253	64	317

Exemple 4 : L'absentéisme salarié

L'absentéisme salarié est, pour certaines entreprises, un des problèmes majeurs. En 1984, un expert étudiant l'absentéisme des 200 salariés du service Expéditions d'une grande firme, a relevé sur les 50 premiers les chiffres suivants :

6	4	4	6	0	6	11	5	10	8	4	8	4	7	7	3	2	3	6	2
4	3	6	1	3	2	4	6	6	6	6	8	3	3	6	2	3	2	4	0
8	3	6	0	1	6	5	13	11	6										

Ce sont les nombres de jours d'absence au cours de l'année de chacun des employés (les congés de longue maladie étant exclus). Les 50 employés ont été rangés en ordre alphabétique.

Exemple 5 : Complexité d'une situation routière et vitesse de traitement de l'information

Une étude a été menée pour vérifier l'effet de la complexité d'une situation routière sur la vitesse de traitement de l'information chez les automobilistes. Pour cela, on mesure le temps de réaction (en ms) à un test d'identification d'une cible visuelle parmi des distracteurs. Le niveau de complexité de la situation routière est défini par le nombre d'éléments présents à l'écran du simulateur.
60 automobilistes ont été assignés aléatoirement à un niveau donné de complexité (10 par niveau).

On a mesuré les résultats suivants :

Niveau de complexité					
1	**2**	**3**	**4**	**5**	**6**
2010	1600	1535	1960	1915	2609
1708	1667	1472	1854	2089	2611
2145	1546	1378	1603	2387	2800
1844	1480	1550	1870	2126	2344
1825	1880	1695	1955	2274	2817
2000	1700	1685	2000	2105	2522
1912	1801	1766	1699	2469	2356
1952	1787	1762	1882	2381	2792
2071	1799	1574	1689	2043	2421
1762	1480	1465	1732	1978	2833

Exemple 6 : Âge des décès dûs à l'alcoolisme

Une étude de l'Inserm nous renseigne sur la répartition par tranche d'âge des décès dûs à l'alcoolisme et à la psychose alcoolique.

Âge	0-24	25-34	35-44	45-54	55-64	65-74	75-84	85-99
Nb décès	1	13	65	112	101	58	28	8

Chapitre 1 : Description d'une situation statistique

Il s'agit d'identifier les différents ingrédients d'une situation statistique. Ces ingrédients sont : les individus, la (ou les) variables et les données.

I - Les individus

Réponse à la question : "**sur qui** porte l'étude ?"

Une analyse statistique débute par l'identification précise du groupe d'**individus** soumis à l'étude. Il peut s'agir :

- des élèves d'une école,
- des membres donateurs d'une grande association,
- des électeurs d'une région ...

Pour désigner un individu, on parle aussi d'unité statistique. Les individus ne sont pas nécessairement des "personnes". Par exemple :

- les entreprises du bâtiment,
- les clubs sportifs,
- les stations de ski ...

La **totalité** des individus du groupe sur lequel porte l'étude constitue **la population**. La population est donc l'ensemble de **tous** les individus visés par l'étude.

Il est souvent **impossible** (ou au moins très peu pratique) d'étudier la population dans son ensemble. Dans ce cas, on se contente d'en extraire une **partie** (un sous-ensemble) que l'on appelle **échantillon**.

Les individus qui constituent l'échantillon sont donc extraits de la population étudiée. Choisir, dans la population, les individus qui seront réellement observés au cours de l'étude **est tout un art** ! Cela constitue une branche de la statistique que l'on appelle **la théorie de l'échantillonage**. Il s'agit en effet que l'échantillon soit **représentatif** de la population.

Quand on le pourra, on donnera des précisions sur la façon dont les individus ont été sélectionnés dans la population. À défaut, on se contentera de préciser leur nombre. Le **nombre** d'individus qui composent l'échantillon est appelé **la taille** de l'échantillon.

Un échantillon constitue une vue nécessairement **partielle**, **approximative** de la population ... mais on espère bien que l'information qu'il porte nous permette de tirer des conclusions pour la population entière. Il s'agit alors **à partir de l'échantillon** d'**inférer** des **propriétés sur la population** : ce domaine constitue la **statistique inférentielle** au programme en L2 et L3.

Dans les quelques cas où l'on a pu étudier l'ensemble de tous les individus de la population, on parle alors de **recensement**. L'échantillon correspond alors à la population entière.

Exemple 1 : Accueil des jeunes handicapés

Exemple 1

Les individus sont **les établissements d'accueil** des enfants et jeunes handicapés en Languedoc-Roussillon au 1er janvier 2005.
Tous les établissements ont été étudiés : il s'agit d'un **recensement**.
La taille de la population (qui correspond à l'échantillon ici) est **84**.

Exemple 2 : Évaluation de la difficulté d'un examen

Exemple 2

Les individus sont **les étudiants de l'UPV** inscrits en L2 de Psychologie et passant l'épreuve de statistique.
On a choisi des étudiants au hasard. Il s'agit d'**un échantillon**, tous les étudiants n'ont pas été interrogés. L'échantillonage a probablement été attentif aux horaires de sortie de l'épreuve.
La taille de l'échantillon est **40**.

Exemple 3 : La démence sénile ▸ Exemple 3

Les individus sont **des personnes en bonne santé agées de plus de 60 ans**.
Il s'agit nécessairement d'un **échantillonage** mais aucune précision n'a été donnée sur la façon dont il a été réalisé.
La taille de l'échantillon est **317**.

Exemple 4 : L'absentéisme salarié ▸ Exemple 4

Les individus sont **les salariés du service Expéditions de la grande firme**.
Il ne s'agit pas d'un échantillonage car tous les salariés de ce service ont été étudiés. C'est un **recensement**.
La taille de la population est **50**.

Exemple 5 : Complexité d'une situation routière et vitesse de traitement de l'information ▸ Exemple 5

Les individus sont **des automobilistes**. La sélection des individus n'a pas été précisée mais on sait qu'ils ont été assignés à différentes conditions expérimentales.
Il s'agit d'un **échantillon** de taille **60**.

Exemple 6 : Âge des décès dûs à l'alcoolisme ▸ Exemple 6

Les individus sont **les décès dûs à l'alcoolisme** en France (étude de l'Inserm). Aucune information n'est donnée pour savoir s'il s'agit d'un échantillonage ou non mais étant donné les effectifs affichés, il semble qu'il ne s'agisse que d'un échantillon et non de la population entière.
La taille de l'échantillon est **386**.

II - La ou les variable(s)

Réponse à la question : "**sur quoi** porte l'étude ?"

Une analyse statistique se poursuit en identifiant précisément la (ou les) caractéristique(s) retenue(s) sur les individus. On parle de **caractère** ou **variable** que l'on observe, que l'on mesure sur chaque individu. Il peut s'agir par exemple :

- du choix d'une activité scolaire
- du montant d'un don
- de la tendance politique
- du nombre de salariés
- de la taille du club (en fonction du nombre de licenciés)
- de l'enneigement

Une variable est désignée par une **lettre majuscule** : X, Y, U ...
L'observation qui en est faite **varie** d'un individu à l'autre.

On appelle **modalités** les réponses faites par les individus à une variable. Un individu n'a **qu'une seule** réponse possible. Sa réponse est désignée par une **lettre minuscule**, par exemple x_3 la réponse faite par l'individu numéro 3 de l'échantillon à la variable X.

On distingue l'ensemble des modalités **observées** de l'ensemble des modalités **observables**. Il est en effet possible qu'au travers des individus de l'échantillon toutes les réponses n'aient pas été rencontrées, soit parce que l'ensemble des modalités observables est infini, soit parce que l'échantillon n'a pas pu recouvrir l'ensemble des possibilités.

On désignera par $\mathcal{U}_X$ l'ensemble des modalités de la variable X. Cela peut-être par exemple :

- $\mathcal{U}_X = \{\text{lecture, sport, peinture, musique}\}$
- $\mathcal{U}_Y = [0; 1000]$
- $\mathcal{U}_Z = \{\text{gauche,droite}\}$
- $\mathcal{U}_T = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
- $\mathcal{U}_U = \{\text{petit, moyen, gros}\}$
- $\mathcal{U}_V = [0; 400]$

Dans le cas où l'ensemble des modalités est **fini**, on note C son **cardinal**. On a alors :

$$\mathcal{U} = \{m_1, m_2, ..., m_C\},$$

Lorsque $C = 2$, la variable est dite **dichotomique**.

Il est fondamental pour la suite de l'analyse statistique de savoir étudier la **structure** de cet ensemble de modalités.

1. Certaines modalités sont des **noms** : la variable est dite **qualitative**. Les modalités sont aussi appelées des **niveaux** et on regarde s'il existe un **ordre naturel** sur ces modalités :
   - si non, variable qualitative **nominale**
   - si oui, variable qualitative **ordinale**

2. D'autres modalités sont des **chiffres** (on a compté, mesuré ...) : la variable est dite **quantitative**. On parle alors de **valeurs** plutôt que de modalités :
   - si les valeurs sont isolées les unes des autres, variable quantitative **discrète**
   - si les valeurs sont prises dans des intervalles, variable quantitative **continue**

Exemple 1 : Accueil des jeunes handicapés

▸ Exemple 1

X : "Type de handicap des jeunes accueillis"

$\mathcal{U}_X =$ { Déficients mentaux, Polyhandicapés, Troubles du comportement, Handicapés moteurs, Handicapés sensoriels }

variable qualitative nominale

Exemple 2 : Évaluation de la difficulté d'un examen

▸ Exemple 2

Y : "Difficulté de l'examen de statistique"

$\mathcal{U}_Y = \{$Très difficile, Difficile, Facile, Très facile$\}$

ou après codage :

$\mathcal{U}_Y = \{1,2,3,4\}$

variable qualitative ordinale

Exemple 3 : La démence sénile

▸ Exemple 3

X : "Score au test"

Y : "État clinique"

$\mathcal{U}_X = \{$Faible, Modéré, Élevé$\}$

variable qualitative ordinale

$\mathcal{U}_Y = \{$Pas de démence, Démence$\}$

variable dichotomique et qualitative nominale

Exemple 4 : L'absentéisme salarié ▸ Exemple 4

Z : "Nombre de jours d'absence"

$\mathcal{U}_Z = \{0, 1, 2, .., 13\}$

variable quantitative discrète

Exemple 5 : Complexité d'une situation routière et vitesse de traitement de l'information ▸ Exemple 5

U : "Temps de réaction"

V : "Niveau de complexité"

$\mathcal{U}_U = [0, 3000]$

variable quantitative continue

$\mathcal{U}_V = \{1, 2, 3, 4, 5, 6\}$

variable qualitative ordinale

Exemple 6 : Âge des décès dûs à l'alcoolisme ▸ Exemple 6

V : "Âge du décédé"

$\mathcal{U}_V = [0, 100]$

variable quantitative continue

III - Les donnée(s)

Réponse à la question : "**quel relevé** des observations ?"

La description d'une situation statistique se termine par une identification précise de l'**information** dont on dispose. Il est **rare** que l'on vous fournisse le **relevé complet** des observations faites. Il faut pour autant réussir à l'**imaginer**. Souvent on présente des tableaux qui comportent déjà un certain **résumé** de l'observation.

On appellera **données brutes** le **tableau** ou la **liste** des observations réalisées : c'est le **relevé pratique de l'information**. Aucune opération n'a encore été réalisée.

↪ Le **tableau** se présente sous la forme :

- en **lignes** : les **individus**
- en **colonnes** : la (ou les) **variable(s)**

Il y a autant de lignes que d'individus et autant de colonnes que de variables.

Numéro de l'individu (son identifiant)	Variable (X)
1	m_3
2	m_5
3	m_5
4	m_1
$\vdots$	$\vdots$
n	m_2

Par exemple :

Numéro de l'enfant	Choix d'activité (X)
1	peinture
2	peinture
3	sport
4	lecture
$\vdots$	$\vdots$
n	sport

↪ La **liste** se présente sous la forme :

$$x_1 = m_3,\ x_2 = m_5,\ x_3 = m_5,\ x_4 = m_1,\ \ldots\ ,\ x_n = m_2$$

ou encore (lorsqu'il n'y a pas d'ambiguïté) :

$$m_3,\ m_5,\ m_5,\ m_1,\ \ldots\ ,\ m_2$$

Par exemple,

$$x_1 = \text{peinture},\ x_2 = \text{peinture},\ x_3 = \text{sport},\ x_4 = \text{lecture},\ \ldots\ ,\ x_n = \text{sport}$$

ou encore

$$\text{peinture, peinture, sport, lecture, } \ldots\ \text{, sport}$$

Il s'agit enfin de décrire l'opération faite, la **transformation réalisée** sur les **données brutes** pour obtenir le tableau fourni.

Souvent, on **regroupe** les individus dont la **réponse** à la variable est **identique** (dans le tableau des données brutes, ils apparaissent avec la même modalité) et on les compte. On associe alors à chaque **modalité** de la variable, l'**effectif** (le nombre) d'individus ayant eu cette réponse.

On résume cela dans le tableau suivant :

Nom de la variable	m_1	m_2	m_3	$\ldots$	m_C
Effectifs	n_1	n_2	n_3	$\ldots$	n_C

Il décrit la **répartition des individus** selon les différentes modalités de la variable. On appelle cela le tableau de la **distribution** : c'est l'objet du chapitre 2.

Exemple 1 : Accueil des jeunes handicapés

Exemple 1

Le tableau des données brutes devait se présenter sous la forme :

Numéro du centre d'accueil	Type de handicap accueilli
1	Handicapés moteurs
2	Handicapés sensoriels
3	Handicapés sensoriels
⋮	⋮
84	Déficients mentaux

Ces données brutes ont été regroupées selon le type de handicap et on a compté le nombre d'établissements (individus) pour chacun de ces types. Le **tableau** fourni est donc celui de la **répartition des individus** (centres d'accueil) **selon le type de handicap** : **tableau de la distribution en effectifs de la variable X : "Type de handicap"**.

Exemple 2 : Évaluation de la difficulté d'un examen

▸ Exemple 2

Le tableau des données brutes devait se présenter sous la forme :

Numéro de l'étudiant	Avis
1	Facile
2	Difficile
3	Difficile
⋮	⋮
40	Difficile

L'information fournie est la **liste des données brutes**.

Exemple 3 : La démence sénile

Exemple 3

Le tableau des données brutes devait se présenter sous la forme :

Numéro de l'individu	Score aux tests (X)	État clinique (Y)
1	Élevé	Pas de démence
2	Modéré	Pas de démence
3	Modéré	Démence
⋮	⋮	⋮
317	Faible	Démence

Ces données brutes ont été regroupées à la fois selon le score aux tests (X) et selon l'état clinique (Y). On a compté le nombre d'individus à chaque croisement d'une modalité de X et d'une modalité de Y. Le **tableau** fourni est donc celui de la **répartition des individus (en effectifs) selon les modalités au croisement des 2 variables X et Y : tableau de la distribution conjointe en effectifs des variables X et Y.**

Exemple 4 : L'absentéisme salarié ▸ Exemple 4

Le tableau des données brutes devait se présenter sous la forme :

Numéro du salarié	Nombre de jours d'absence
1	6
2	4
3	4
⋮	⋮
50	6

L'information fournie est la **liste des données brutes**.

Exemple 5 : Complexité d'une situation routière et vitesse de traitement de l'information

▸ Exemple 5

Le tableau des données brutes devait se présenter sous la forme :

No d'automobiliste	Tps réaction (U)	Niv. complexité (V)
1	2010	1
2	1708	1
3	2145	1
⋮	⋮	⋮
60	2833	6

• Pour la variable V, on a regroupé dans une même colonne les réponses concernant les individus ayant passé l'expérience avec le même niveau de complexité mais on ne les a pas comptés pour autant.

• Pour la variable U, aucun regroupement n'a été effectué. On fournit à chaque fois les données brutes.

Le tableau fourni est donc un tableau de données brutes dans lequel on a réarrangé dans une même colonne les données d'un même niveau de complexité.

Exemple 6 : Âge des décès dûs à l'alcoolisme ▸ Exemple 6

Le tableau des données brutes devait se présenter sous la forme :

Numéro du décès	Âge de la personne
1	18
2	89
3	46
⋮	⋮
386	39

Aucun regroupement direct n'est possible. On a formé des classes de valeurs (tranches d'âge) et on a regroupé les décés pour lesquels l'âge du décédé appartenait à une même classe. On les a alors comptés. Le tableau fourni est donc celui de la **répartition des individus (décès) selon les tranches d'âge** : c'est **le tableau de la distribution de la variable V regroupée en classes**.

À SAVOIR

La **description ou modélisation d'une situation statistique** consiste à identifier les 3 éléments suivants :

1. **individu** : entité sur laquelle l'observation est réalisée
   - **population** : ensemble des individus sur lesquels porte l'étude mais non nécessairement observés
   - **échantillon** : ensemble des individus de la population qui ont été observés. Le nombre de ces individus s'appelle la taille.
2. **variable** : objet de l'observation réalisée
   - **modalité** : réponse de l'individu à la variable
   - **nature** :

$$\begin{cases} \text{- qualitative} & \begin{cases} \text{- nominale} \\ \text{- ordinale} \end{cases} \\ \text{- quantitative} & \begin{cases} \text{- discrète} \\ \text{- continue} \end{cases} \end{cases}$$

3. **présentation des données** :
   - **données brutes** : liste des toutes les réponses des individus interrogés
   - **données transformées**

Chapitre 2 : Distribution et distribution cumulée

Les **réponses** des n individus de l'échantillon à la variable X varient d'un individu à l'autre :

$$x_1,\ x_2,\ \dots,\ x_n$$

Dans le cas où l'on peut ranger ces observations, on notera :

$$x_{(1)},\ x_{(2)},\ \dots,\ x_{(n)}$$

l'**échantillon ordonné**.
Construire le tableau de la **distribution** et de la **distribution cumulée** constitue un premier regard essentiel sur la **répartition de ces réponses**.

I - La distribution

a - Distribution en effectifs

Lorsqu'il y a des **répétitions** dans les réponses :

- il s'agit dans un premier temps de regrouper tous les individus dont la réponse est identique
- puis de les compter

On associe ainsi à chaque modalité $\mathbf{m_k}$ de $\mathcal{U}_X$ son **effectif**, noté $\mathbf{n_k}$, c'est-à-dire le nombre d'individus ayant eu cette modalité pour réponse. On forme le tableau :

Variable X	m_1	m_2	$\ldots$	m_C	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n

Définition : Une distribution en effectif est donc la liste des modalités de la variable et de leur effectif associé.
Cette distribution se présente le plus souvent sous forme d'un tableau mais ce peut être aussi une liste de couples (modalité, effectif).

b - Distribution en fréquences

S'intéresser à la distribution en fréquences, c'est ramener **en proportion à la taille de l'échantillon** les différents effectifs obtenus. C'est ainsi plus facile à **interpréter** et à **comparer**.

Par exemple : annoncer un effectif de 18 pour la modalité m_1 ne dit pas du tout la même chose lorsque la taille de l'échantillon vaut 50 que lorsqu'elle vaut 500.

On associe ainsi à chaque modalité $\mathbf{m_k}$ de $\mathcal{U}_X$ sa **fréquence** :

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

On forme alors le tableau :

Variable X	m_1	m_2	$\ldots$	m_C	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1

Remarques :

- une fréquence est toujours comprise entre 0 et 1
- un pourcentage n'est qu'une écriture d'un chiffre à virgule : $0.42 = 42\%$
- la somme des effectifs vaut n (la taille de l'échantillon) et la somme des fréquences vaut 1 :

$$\sum_{k=1}^{C} n_k = n \qquad \sum_{k=1}^{C} f_k = 1$$

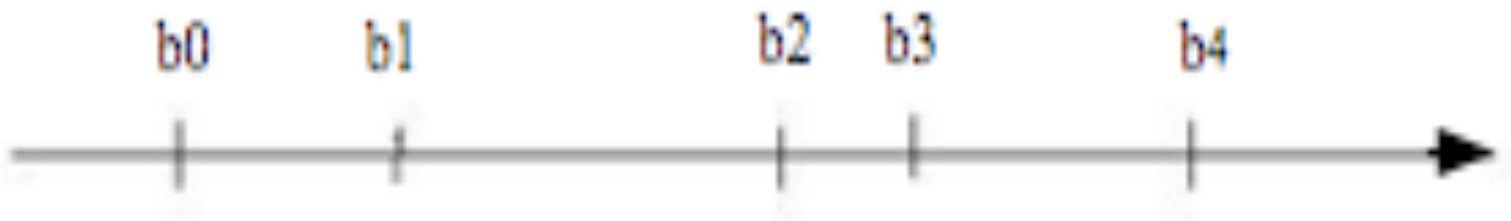
c - Cas particulier des variables quantitatives continues

Il n'y a **pas de répétitions** dans les réponses. Il n'y a donc pas de regroupement immédiat. On forme des **classes de valeurs**. C'est un découpage de $\mathcal{U}$ en **intervalles successifs**.

Définir ce découpage c'est :

- choisir le nombre de classes C
- définir la borne inférieure et la borne supérieure de chaque intervalle

Cette définition doit être propre : il ne doit pas y avoir de "trou", la fin d'une classe est le début de la suivante.



On dit que la variable a été **regroupée en classes**. Mais attention, ce regroupement en classes implique une **perte d'information**. On oublie la valeur exacte observée pour l'individu pour ne retenir que son appartenance à une classe.

On forme alors le tableau :

Variable X	$[b_0; b_1[$	$[b_1; b_2[$	$\ldots$	$[b_{C-1}; b_C[$	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1

Le **choix des classes** est **délicat** :

- leur nombre :
  - pas trop petit pour éviter une perte d'information trop importante
  - pas trop grand pour rendre l'information lisible
- leurs bornes :
  - les classes ne sont pas nécessairement de même **largeur**

II - La distribution cumulée

La définition de la **distribution cumulée** n'a du sens que lorsqu'il existe un **ordre naturel** sur les modalités. On ne parlera donc pas de distribution cumulée dans le cas d'une variable qualitative nominale.

Lorsque l'on sait ranger les modalités selon un ordre, il s'agit alors de **cumuler les effectifs (ou les fréquences) selon l'ordre croissant des modalités**.

Ceci permet ensuite de répondre à des questions telles que :

- quel est le pourcentage d'individus dont la réponse est **plus petite que** ... ?
- combien d'individus ont une valeur **plus grande que** ... ?
- quelle est la fréquence d'individus dont la valeur est **comprise entre** ... **et** ... ?

a - Cas des variables qualitatives ordinales

Les modalités sont rangées selon un ordre naturel :

$$m_1 < m_2 < ... < m_k < ... < m_C$$

On forme alors le tableau contenant distribution et distribution cumulée :

Variable X	m_1	m_2	...	m_C	Total
Effectifs n_k	n_1	n_2	...	n_C	n
Fréquences f_k	f_1	f_2	...	f_C	1
Eff. cum. N_k	$N_1 = n_1$	$N_2 = n_1 + n_2$	...	$N_C = n$	
Fréq. cum. F_k	$F_1 = f_1$	$F_2 = f_1 + f_2$	...	$F_C = 1$	

b - Cas des variables quantitatives discrètes

Les valeurs peuvent être bien sûr rangées :

$$v_1 < v_2 < ... < v_k < ... < v_C$$

On forme alors le tableau contenant distribution et distribution cumulée :

Variable X	v_1	v_2	...	v_C	Total
Effectifs n_k	n_1	n_2	...	n_C	n
Fréquences f_k	f_1	f_2	...	f_C	1
Eff. cum. N_k	$N_1 = n_1$	$N_2 = n_1 + n_2$	...	$N_C = n$	
Fréq. cum. F_k	$F_1 = f_1$	$F_2 = f_1 + f_2$	...	$F_C = 1$	

À l'aide des fréquences cumulées, on définit **la fonction F** telle que :

$$\forall k \in \{1, .., C\} \qquad F(v_k) = F_k = \sum_{\ell=1}^{k} f_\ell$$

On peut aussi cumuler les fréquences selon un ordre décroissant :

Variable X	v_1	$\dots$	v_{C-1}	v_C	Total
Effectifs n_k	n_1	$\dots$	n_{C-1}	n_C	n
Fréquences f_k	f_1	$\dots$	f_{C-1}	f_C	1
Eff. cum. N_k	$N_1 = n_1$	$\dots$	N_{C-1}	$N_C = n$	
Fréq. cum. F_k	$F_1 = f_1$	$\dots$	F_{C-1}	$F_C = 1$	
Fréq. cum. décr. G_k	$G_1 = 1$	$\dots$	$G_{C-1} = f_C + f_{C-1}$	$G_C = f_C$	

Ainsi à l'aide des fréquences cumulées décroissantes, on définit **la fonction G** telle que :

$$\forall k \in \{1, .., C\} \quad G(v_k) = G_k = \sum_{\ell=k}^{C} f_\ell$$

Attention :

$$F(v_k) + G(v_k) = 1 + f_k$$

donc

$$F(v_k) + G(v_k) > 1$$

c - Cas des variables quantitatives continues

Les classes sont mises dans l'ordre :

$$b_0 < b_1 < \cdots < b_{C-1} < b_C$$

On forme alors le tableau contenant distribution et distribution cumulée :

Variable X	$[b_0; b_1[$	$[b_1; b_2[$	...	$[b_{C-1}; b_C[$	Tot.
Effectifs n_k	n_1	n_2	...	n_C	n
Fréquences f_k	f_1	f_2	...	f_C	1

Eff. cum. N_k	0	N_1	N_2	...	N_{C-1}	n
Fréq. cum. F_k	0	F_1	F_2	...	F_{C-1}	1
Fr. cum. décr. G_k	1	G_1	G_2	...	G_{C-1}	0

Attention : les cumuls se font au niveau des bornes de classes.

Variable X	$[b_0; b_1[$	$[b_1; b_2[$	$\ldots$	$[b_{C-1}; b_C[$	Tot.
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1

Eff. cum. N_k	0	N_1	N_2	$\ldots$	N_{C-1}	n
Fréq. cum. F_k	0	F_1	F_2	$\ldots$	F_{C-1}	1
Fr. cum. décr. G_k	1	G_1	G_2	$\ldots$	G_{C-1}	0

Dans ce cas, les fonctions **F** et **G** vérifient :

$$F(b_0) = 0 \; ; \; F(b_1) = f_1 \; ; \; F(b_2) = f_1 + f_2 \; ; \; \ldots \; ; \; F(b_C) = 1$$

$$G(b_0) = 1 \; ; \; \ldots \; ; \; G(b_{C-2}) = f_C + f_{C-1} \; ; \; G(b_{C-1}) = f_C \; ; \; G(b_C) = 0$$

On peut montrer pour les variables quantitatives continues (*et uniquement dans ce cas !*) que **F** et **G** vérifient :

$$F + G = 1$$

Exemple 1 : Accueil des jeunes handicapés

Exemple 1

Tableau de la distribution (en effectifs et en fréquences) :

Type hand.	Déf. ment.	Polyhand.	Tr. comp.	Hand. mot.	Hand. sens.	Total
Effectifs n_k	43	16	17	3	5	84
Fréq. f_k	0.51	0.19	0.20	0.04	0.06	1

Exemple 2 : Évaluation de la difficulté d'un examen

▸ Exemple 2

Tableau de la distribution et de la distribution cumulée :

Difficulté exam.	Très difficile	Difficile	Facile	Très facile	Total
Effectifs n_k	9	17	12	2	40
Fréquences f_k	0.225	0.425	0.300	0.050	1
Eff. cum. N_k	9	26	38	40	
Fréq. cum. F_k	0.225	0.650	0.950	1	

Exemple 3 : La démence sénile

Exemple 3

Tableau de la distribution de la variable "État clinique" :

État clinique	Pas de démence	Démence	Total
Effectifs n_k	253	64	317
Fréq. f_k	0.80	0.20	1

Tableau de la distribution et de la distribution cumulée de la variable "Score aux tests" :

Score aux tests	Faible	Modéré	Élevé	Total
Effectifs n_k	13	92	212	317
Fréquences f_k	0.04	0.29	0.67	1
Eff. cum. N_k	13	105	317	
Fréq. cum. F_k	0.04	0.33	1	

Exemple 4 : L'absentéisme salarié ▸ Exemple 4

Tableau de la distribution et de la distribution cumulée :

Nbjours	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
n_k	12	8	20	32	28	8	52	8	16	0	4	8	0	4	200
f_k (en %)	6	4	10	16	14	4	26	4	8	0	2	4	0	2	100
N_k	12	20	40	72	100	108	160	168	184	184	188	196	196	200	
F_k (en%)	6	10	20	36	50	54	80	84	92	92	94	98	98	100	
G_k (en%)	100	94	90	80	64	50	46	20	16	8	8	6	2	2	

NB : $F(5) = 54\%$ et $G(5) = 50\%$ $(= 1 - F(4))$

Exemple 5 : Compléxité d'une situation routière et vitesse de traitement de l'information

▸ Exemple 5

Regroupement en 9 classes.

Tps	[1300; 1500[	[1500; 1700[	[1700; 1800[	
n_k	5	11	8	
f_k	0.083	0.183	0.133	
N_k	0	5	16	24
F_k	0	0.083	0.267	0.400
G_k	1	0.917	0.733	0.600

[1800; 1900[	[1900; 2000[	[2000; 2200[	[2200; 2400[	
7	6	9	5	
0.117	0.100	0.150	0.083	
31	37	46	51	
0.517	0.617	0.767	0.850	
0.483	0.383	0.233	0.150	

[2400; 2700[	[2700; 3000[	Tot.
5	4	60
0.083	0.067	1
56	60	
0.933	1	
0.067	0	

NB : $F(2000) = 0.617$ et $G(2000) = 0.383$

Exemple 6 : Âge des décès dûs à l'alcoolisme

▸ Exemple 6

Variable déjà regroupée en classes à redéfinir proprement.
Tableau de la distribution et de la distribution cumulée.

Âge		[0 ;25[		[25 ;35[		[35 ;45[		[45 ;55[	
n_k		1		13		65		112	
f_k		0.003		0.034		0.168		0.290	
N_k	0		1		14		79		191
F_k	0		0.003		0.036		0.205		0.495
G_k	1		0.997		0.964		0.795		0.505

[55 ;65[	[65 ;75[	[75 ;85[	[85 ;100[	Total
101	58	28	8	386
0.262	0.150	0.073	0.021	1
292	350	378	386	
0.756	0.907	0.979	1	
0.244	0.093	0.021	0	

NB : $F(65) = 75.6\%$ et $G(65) = 24.4\%$

III - Quelques exemples : prudence !

a - Exemple 1

On donne les effectifs (nombre de patients) de chaque catégorie dans un hôpital. Une catégorie est en fait un service dans lequel le patient est hospitalisé :

phlébologie	gériatrie	chirurgie	autre
50	45	37	78

On demande à un marchand de fruits et legumes le nombre de kilos qu'il a vendus sur différents produits. Il répond :

poireaux	carottes	raisin	pommes de terre
50	45	37	78

b - Exemple 2

Dans un dossier du journal Le Monde de février 2005, on a pu lire l'enquête suivante : *Pour chacune des catégories suivantes, dites-moi si elle constitue pour vous actuellement en France :*

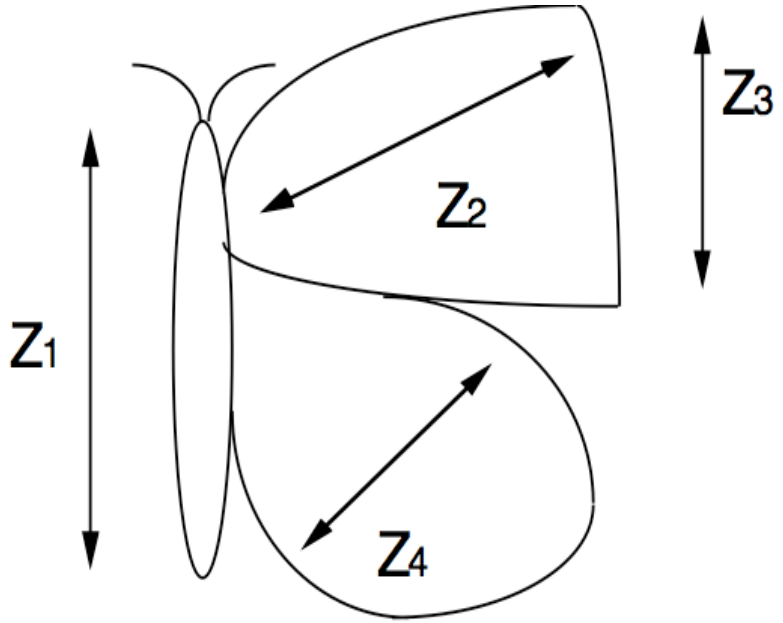
1. *un groupe à part dans la société*
2. *un groupe ouvert aux autres*
3. *des personnes ne formant pas spécialement un groupe*
4. *ne se prononcent pas*

La tableau de résultats fourni était :

en %	1	2	3	4
Les musulmans	57	18	19	6
Les maghrébins	48	21	24	7
Les juifs	36	26	31	7
Les homosexuels	32	31	32	5
Les noirs	19	37	39	5
Les catholiques	11	41	44	4

c - Exemple 3 : Mesures des papillons

	Z_1	Z_2	Z_3	Z_4
1	22	35	24	19
	24	31	21	22
	27	36	25	15
	27	36	24	23
	21	33	23	18
	26	35	23	32
	27	37	26	15
.	22	30	19	20
.	25	33	22	22
.	30	41	28	17
	24	39	27	21
	⋮	⋮	⋮	
	⋮	⋮	⋮	
23	24	38	26	21



d - Exemple 4 : Consommations

	PaO	PaA	ViO	ViA	PdT	Leg	Rai	Plat
AGRI	167	1	163	23	41	8	6	6
SAAG	162	2	141	12	40	12	4	15
PRIN	119	6	69	56	39	5	13	41
CSUP	87	11	63	111	27	3	18	39
CMOY	103	5	68	77	32	4	11	30
EMPL	111	4	72	66	34	6	10	28
OUVR	130	3	76	52	43	7	7	16
INAC	138	7	117	74	53	8	12	20

167 francs sont dépensés en moyenne annuellement par un exploitant agricole pour l'achat de pain ordinaire (étude INSEE 1972).

e - Exemple 5 : Élections régionales 1992

	PC	MajP	Ecol	UPF	FN	CPNT
alsa	0	6	9	20	11	1
aqui	6	20	9	32	8	10
auve	4	9	5	24	4	1
bour	3	14	7	24	8	1
bret	3	19	12	41	7	1
cent	8	18	8	32	11	0
cham	3	9	4	22	8	3
frco	1	10	5	21	6	0
ilef	17	33	37	85	37	0
lang	8	14	7	24	13	1
limo	7	13	3	19	1	1
lorr	3	16	11	32	10	1
mipy	5	27	7	43	6	3
nord	15	27	14	30	15	12
bnor	1	9	8	24	5	0
hnor	5	14	8	19	8	0
ploi	3	20	13	48	8	2
pica	6	9	9	22	8	3
poit	3	13	7	25	5	2
paca	10	30	6	43	34	0
rhal	12	29	21	65	29	1

À SAVOIR

1. **distribution** : liste des modalités et de leur effectif ou fréquence associé(e).
   - dans le but de construire la distribution d'une variable quantitative continue, une modalité correspondra alors à une classe de valeurs.

2. **distribution cumulée** : liste des modalités et de leur effectif ou fréquence cumulé(e) associé.

Chapitre 3 : Représentations graphiques

Pour **chaque type de variables**, nous présentons les **points importants** à respecter pour la **représentation graphique** de la **distribution** et, le cas échéant, de la **distribution cumulée**.

Il s'agit de visualiser graphiquement la **répartition des individus** sur les modalités de la variable.

Nous terminerons avec une critique de graphiques produits par un logiciel.

I - Cas d'une variable qualitative nominale

La distribution de la variable X est fournie par le tableau :

Variable X	m_1	m_2	$\ldots$	m_C	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1

$\hookrightarrow$ **diagramme en barres séparées**

1. Tracer un **axe horizontal** portant le **nom** de la variable et y positionner les **modalités** de la variable (ni ordre, ni distance n'ont de sens ici - l'axe n'est pas orienté)
2. Choisir une échelle sur l'**axe vertical** pour y positionner les **effectifs** ou les **fréquences** (l'axe est orienté)
3. Associer à chaque modalité un **trait** ou **rectangle** (l'épaisseur n'ayant pas de signification) **vertical** de **hauteur** correspondant à son effectif ou sa fréquence

II - Cas d'une variable qualitative ordinale

Distribution et distribution cumulée de la variable X sont fournies par :

Variable X	m_1	m_2	$\ldots$	m_C	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1
Eff. cum. N_k	$N_1 = n_1$	N_2	$\ldots$	$N_C = n$	
Fréq. cum. F_k	$F_1 = f_1$	F_2	$\ldots$	$F_C = 1$	

Nous envisageons ici l'*échelle continue sous-jacente* à la variable ordinale.

- **Distribution - diagramme en barres juxtaposées**
  1. Tracer un **axe horizontal** portant le **nom** de la variable (l'axe est orienté) et y positionner les **modalités** de la variable entre 2 délimiteurs (répartis régulièrement)
  2. Choisir une échelle sur l'**axe vertical** pour y positionner les **effectifs** ou les **fréquences** (l'axe est orienté)
  3. Associer à chaque modalité un **rectangle vertical** (la base correspond à la modalité entre 2 délimiteurs) de **hauteur** correspondant à son effectif ou sa fréquence

- **Distribution cumulée** -

La distribution cumulée existe. Cependant l'échelle n'étant pas numérique, nous n'en faisons **pas de représentation graphique**.

III - Cas d'une variable quantitative discrète

Distribution et distribution cumulée de la variable X sont fournies par :

Variable X	v_1	v_2	$\ldots$	v_C	Total
Effectifs n_k	n_1	n_2	$\ldots$	n_C	n
Fréquences f_k	f_1	f_2	$\ldots$	f_C	1
Eff. cum. N_k	$N_1 = n_1$	N_2	$\ldots$	$N_C = n$	
Fréq. cum. F_k	$F_1 = f_1$	F_2	$\ldots$	$F_C = 1$	
Fréq. cum. décr. G_k	$G_1 = 1$	G_2	$\ldots$	$G_C = f_C$	

- **Distribution** - **diagramme en bâtons**

1. Choisir une échelle sur l'**axe horizontal** (l'axe est orienté) pour y positionner les **valeurs** (en respectant l'échelle)
2. Choisir une échelle sur l'**axe vertical** pour y positionner les **effectifs** ou les **fréquences** (l'axe est orienté)
3. Associer à chaque valeur un **bâton vertical** de **hauteur** correspondant à son effectif ou sa fréquence

• **Distribution cumulée** - L'échelle étant numérique discrète, on complète la définition de la **fonction de répartition** par :

$$\begin{array}{ll} \forall k \in \{1, .., C\} & F(v_k) = F_k = \sum_{\ell=1}^{k} f_\ell \\ \forall x < v_1 & F(x) = 0 \\ \forall x \in [v_k; v_{k+1}[& F(x) = F_k \\ \forall x \geq v_C & F(x) = 1 \end{array}$$

Les cumuls se font par **sauts successifs** (la fonction est constante sur chaque intervalle).

Graphe de la fonction de répartition

$\hookrightarrow$ graphe **en escalier**

1. Choisir une échelle sur l'**axe horizontal** (l'axe est orienté) pour y positionner les **valeurs** (en respectant l'échelle)
2. Choisir une échelle sur l'**axe vertical** pour y positionner les **fréquences cumulées** (l'axe est orienté de 0 à 1)
3. Associer à chaque valeur un **point** correspondant à sa fréquence cumulée et tracer des **morceaux de droite horizontale** entre 2 valeurs observées

III - Cas d'une variable quantitative continue

Distribution et distribution cumulée de la variable X sont fournies par :

Variable X	$[b_0; b_1[$	$[b_1; b_2[$	$\dots$	$[b_{C-1}; b_C[$	Total
Effectifs n_k	n_1	n_2	$\dots$	n_C	n
Fréquences f_k	f_1	f_2	$\dots$	f_C	1

Eff. cum. N_k	0	N_1	N_2	$\dots$	N_{C-1}	n
Fréq. cum. F_k	0	F_1	F_2	$\dots$	F_{C-1}	1

• Distribution - histogramme

Il est nécessaire de prendre en compte les **amplitudes** des classes (ou **largeurs** des classes). On note $\mathbf{a_k}$ l'amplitude de la classe k.

Par exemple, affirmer que 10 individus mesurent entre 155 cm et 165 cm ne signifie pas du tout la même chose que d'affirmer que 10 individus mesurent entre 160 et 162 cm. Les effectifs sont identiques (les fréquences aussi) mais dans le 2ème cas, il y a une **concentration** beaucoup plus forte d'observations.

Nous allons donc calculer pour chaque classe sa **densité de fréquence**. C'est le rapport entre sa fréquence et son amplitude :

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

Cela représente la fréquence d'individus par unité de largeur de la classe et mesure donc bien **la concentration**.

On complète le tableau :

Variable X	$[b_0; b_1[$	$[b_1; b_2[$	$\dots$	$[b_{C-1}; b_C[$	Total
Effectifs n_k	n_1	n_2	$\dots$	n_C	n
Fréquences f_k	f_1	f_2	$\dots$	f_C	1
Eff. cum. N_k	0 N_1	N_2	$\dots$	N_{C-1}	n
Fréq. cum. F_k	0 F_1	F_2	$\dots$	F_{C-1}	1
Amplitudes a_k	a_1	a_2	$\dots$	a_C	
Densités d_k	d_1	d_2	$\dots$	d_C	

1. Choisir une échelle sur l'**axe horizontal** (l'axe est orienté) pour y positionner les **bornes des classes** (en respectant l'échelle)
2. Choisir une échelle sur l'**axe vertical** pour y positionner les **densités de fréquence** (l'axe est orienté)
3. Associer à chaque classe un **rectangle vertical** de base l'intervalle correspondant à la classe et de **hauteur** correspondant à sa densité de fréquence

Par cette construction, la **surface** de chaque rectangle, égale à la **hauteur** multipliée par la **largeur** (i.e. la **densité** multipliée par l'**amplitude**), représentera donc la fréquence associée à la classe.
Dans cette représentation graphique, les surfaces ont donc un sens et la **surface totale** vaut :

$$\sum_{k=1}^{C} a_k \times d_k = \sum_{k=1}^{C} f_k = 1$$

• **Distribution cumulée** - L'échelle étant numérique continue, on complète la définition de la **fonction de répartition** par :

$$\begin{array}{ll} \forall k \in \{1, .., C\} & F(b_k) = F_k = \sum_{\ell=1}^{k} f_\ell \\ \forall x < b_0 & F(x) = 0 \\ \forall x \in [b_k; b_{k+1}[& F(x) : \text{droite reliant } F_k \text{ et } F_{k+1} \\ \forall x \geq b_C & F(x) = 1 \end{array}$$

Sur cette **échelle continue**, les cumuls ne se font pas en un lieu (une valeur de la variable) précis. Ils sont au contraire **progressifs** à l'intérieur de chaque classe. On approche alors cette progression par une **droite** : la progression est régulière. On fait donc une **approximation linéaire** à l'intérieur de chaque classe. Ceci est équivalent à faire une hypothèse de **distribution uniforme** des valeurs dans chaque classe.

Graphe de la fonction de répartition

$\hookrightarrow$ graphe **linéaire par morceaux**

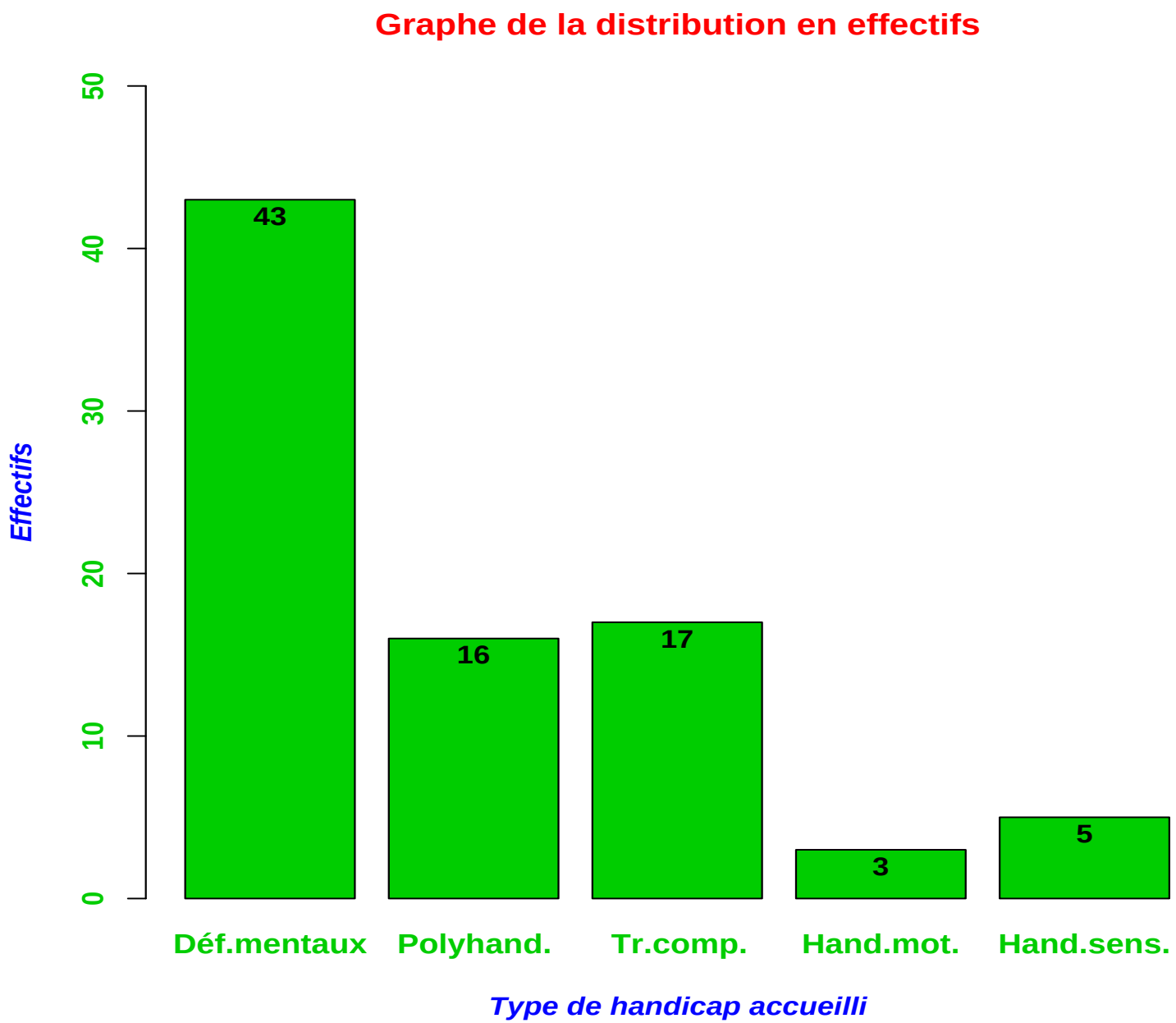
1. Choisir une échelle sur l'**axe horizontal** (l'axe est orienté) pour y positionner les **bornes des classes** (en respectant l'échelle)
2. Choisir une échelle sur l'**axe vertical** pour y positionner les **fréquences cumulées** (l'axe est orienté de 0 à 1)
3. Positionner, au niveau chaque borne, un **point** correspondant à sa fréquence cumulée
4. **Relier** les différents points par des **morceaux de droite**

V - Graphiques - logiciel R

Exemple 1 : Accueil des jeunes handicapés

Tableau de la distribution (en effectifs et en fréquences) :

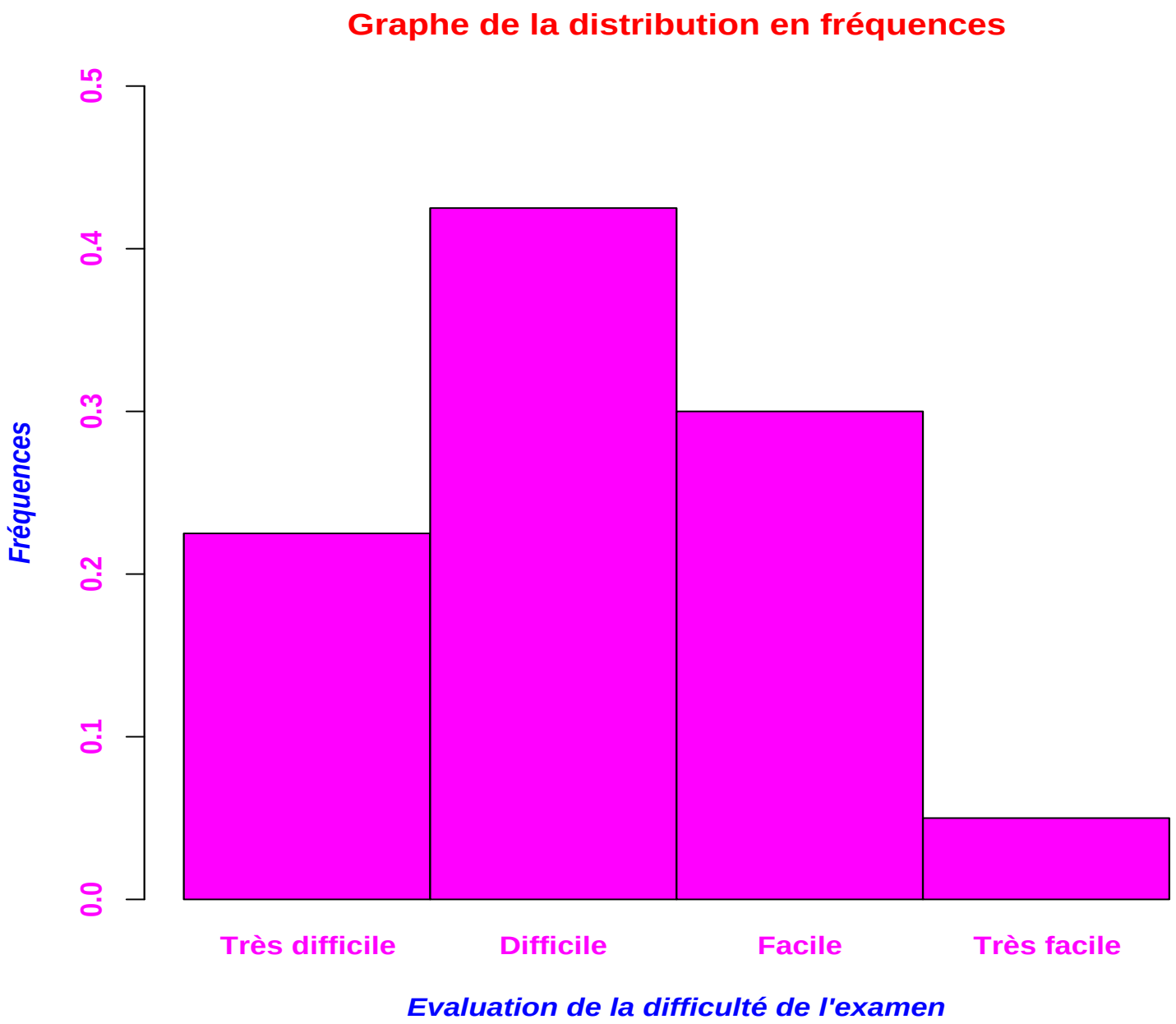
Type hand.	Déf. ment.	Polyhand.	Tr. comp.	Hand. mot.	Hand. sens.	Total
Effectifs n_k	43	16	17	3	5	84
Fréq. f_k	0.51	0.19	0.20	0.04	0.06	1


Graphe de la distribution en effectifs
Effectifs
0
10
20
30
40
50
43
16
17
3
5
Déf.mentaux
Polyhand.
Tr.comp.
Hand.mot.
Hand.sens.
Type de handicap accueilli

Exemple 2 : Évaluation de la difficulté d'un examen

Tableau de la distribution et de la distribution cumulée :

Difficulté exam.	Très difficile	Difficile	Facile	Très facile	Total
Effectifs n_k	9	17	12	2	40
Fréquences f_k	0.225	0.425	0.300	0.050	1
Eff. cum. N_k	9	26	38	40	
Fréq. cum. F_k	0.225	0.650	0.950	1	


Graphe de la distribution en fréquences
Fréquences
0.0
0.1
0.2
0.3
0.4
0.5
Très difficile
Difficile
Facile
Très facile
Evaluation de la difficulté de l'examen

Exemple 3 : La démence sénile

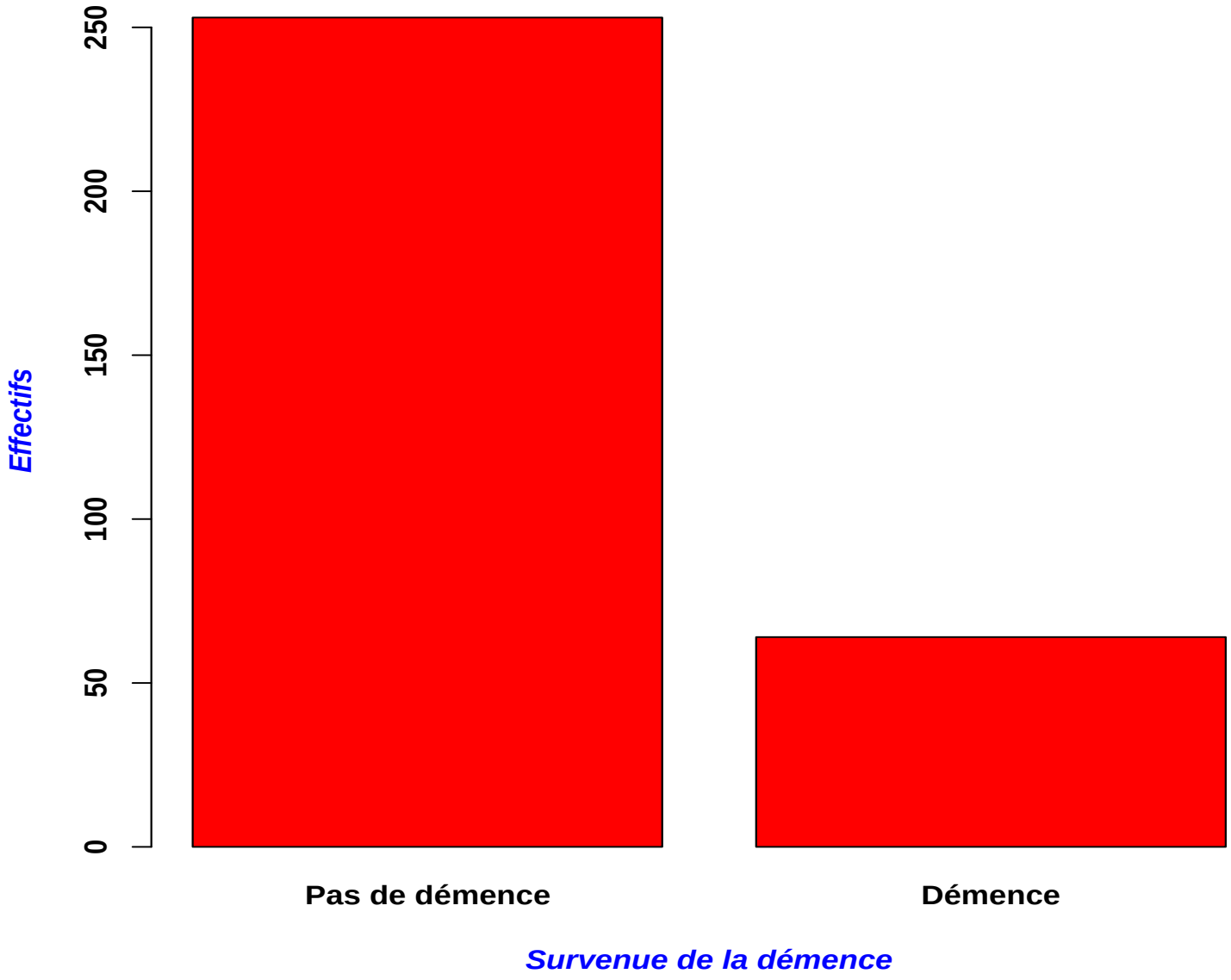
Tableau de la distribution de la variable "État clinique" :

État clinique	Pas de démence	Démence	Total
Effectifs n_k	253	64	317
Fréq. f_k	0.80	0.20	1

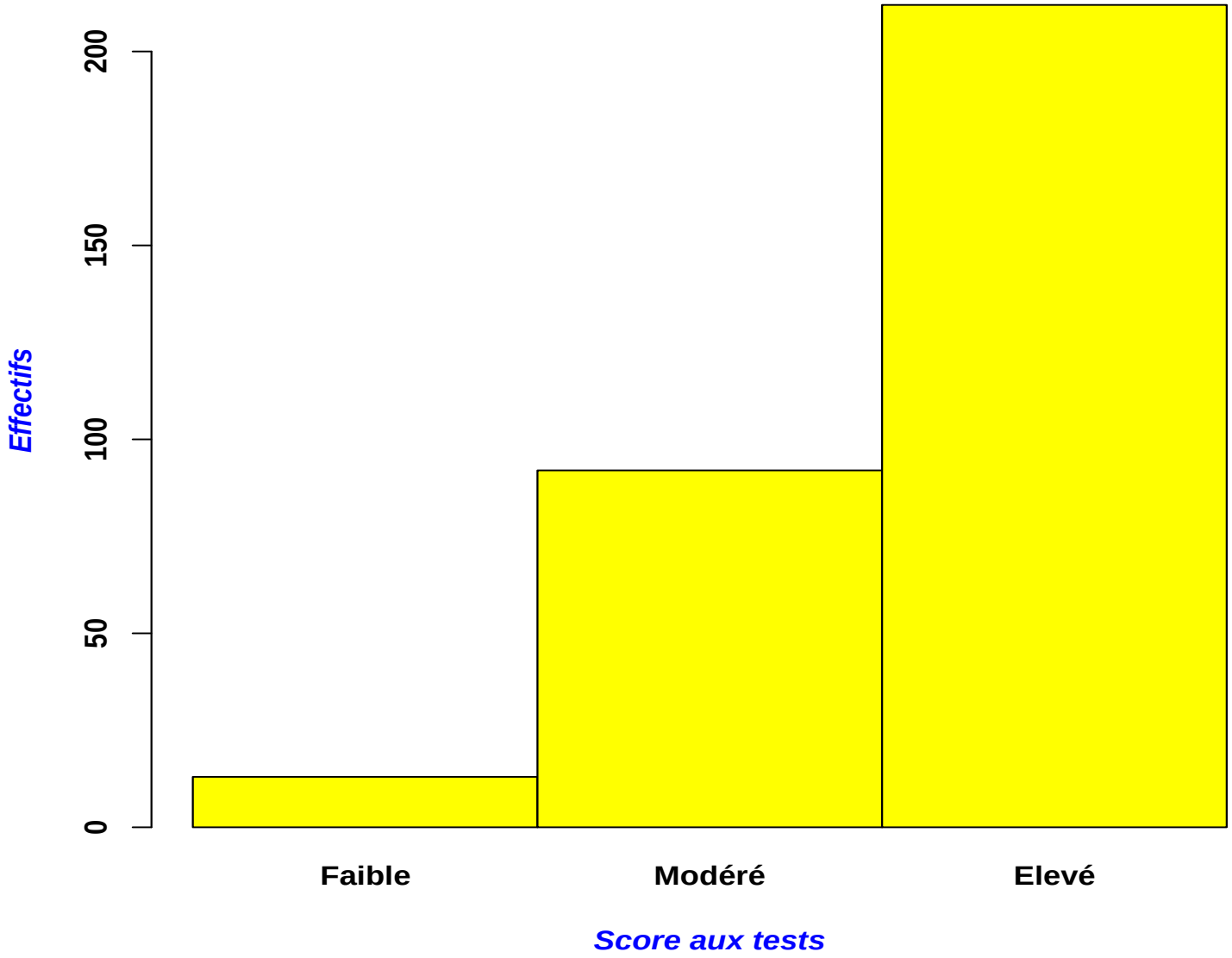
Tableau de la distribution et de la distribution cumulée de la variable "Score aux tests" :

Score aux tests	Faible	Modéré	Élevé	Total
Effectifs n_k	13	92	212	317
Fréquences f_k	0.04	0.29	0.67	1
Eff. cum. N_k	13	105	317	
Fréq. cum. F_k	0.04	0.33	1	

Graphe de la distribution



Graphe de la distribution



Score aux tests	**État clinique**		
	Pas de démence	Démence	
Faible	2	11	13
Modéré	49	43	92
Élevé	202	10	212
	253	64	317

Graphe de la distribution conjointe

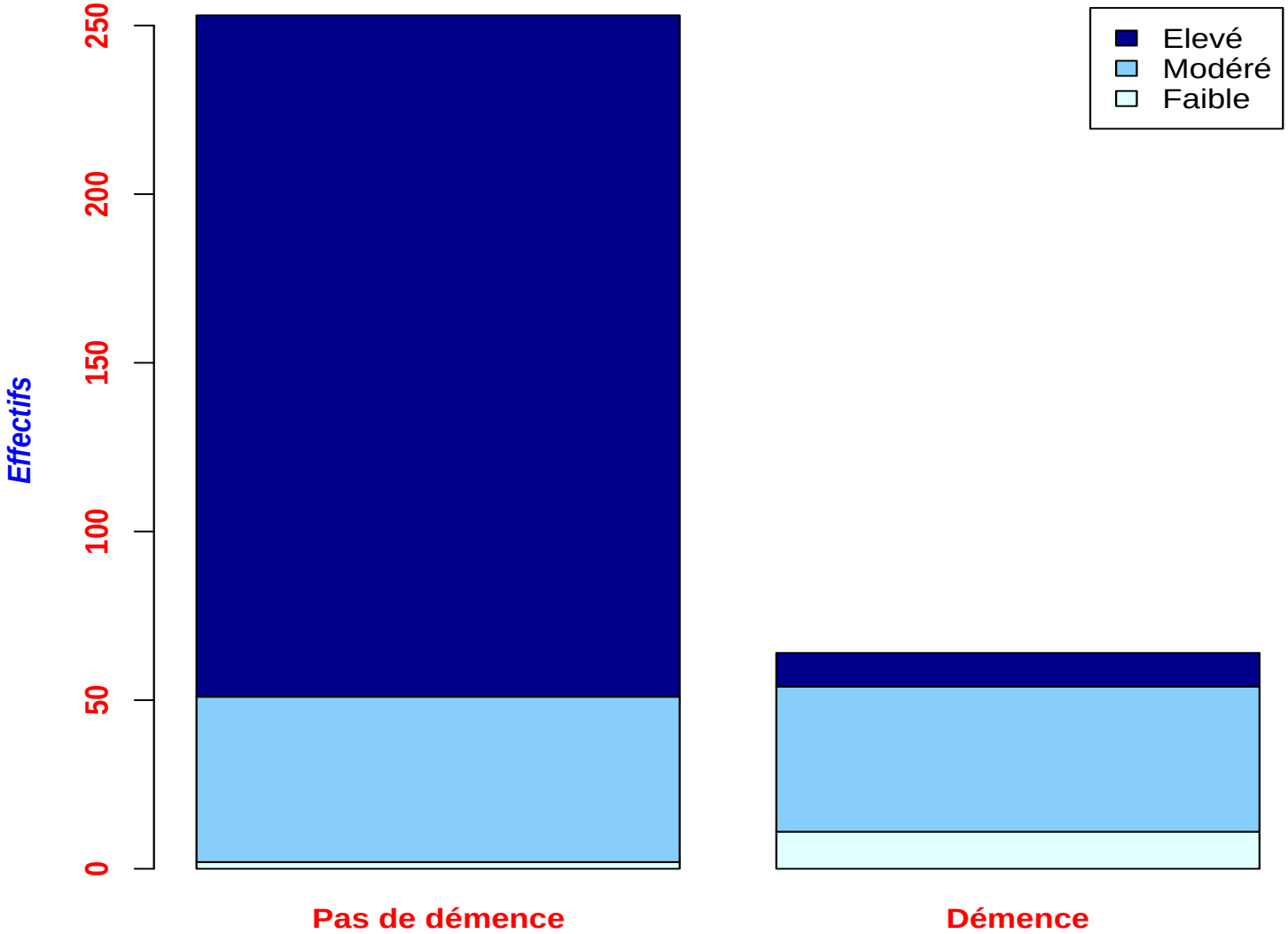


Tableau des distributions de la variable “État clinique” conditionnellement à chaque niveau de la variable “Score aux tests” :

Score	**État clinique**	Pas de démence	Démence	Total
Faible	Effectifs n_k	2	11	13
	Fréq. f_k	0.154	0.846	1
Modéré	Effectifs n_k	49	43	92
	Fréq. f_k	0.533	0.467	1
Elevé	Effectifs n_k	202	10	212
	Fréq. f_k	0.953	0.047	1

Graphe des distributions conditionnelles de la variable 'Survenue de la démence'

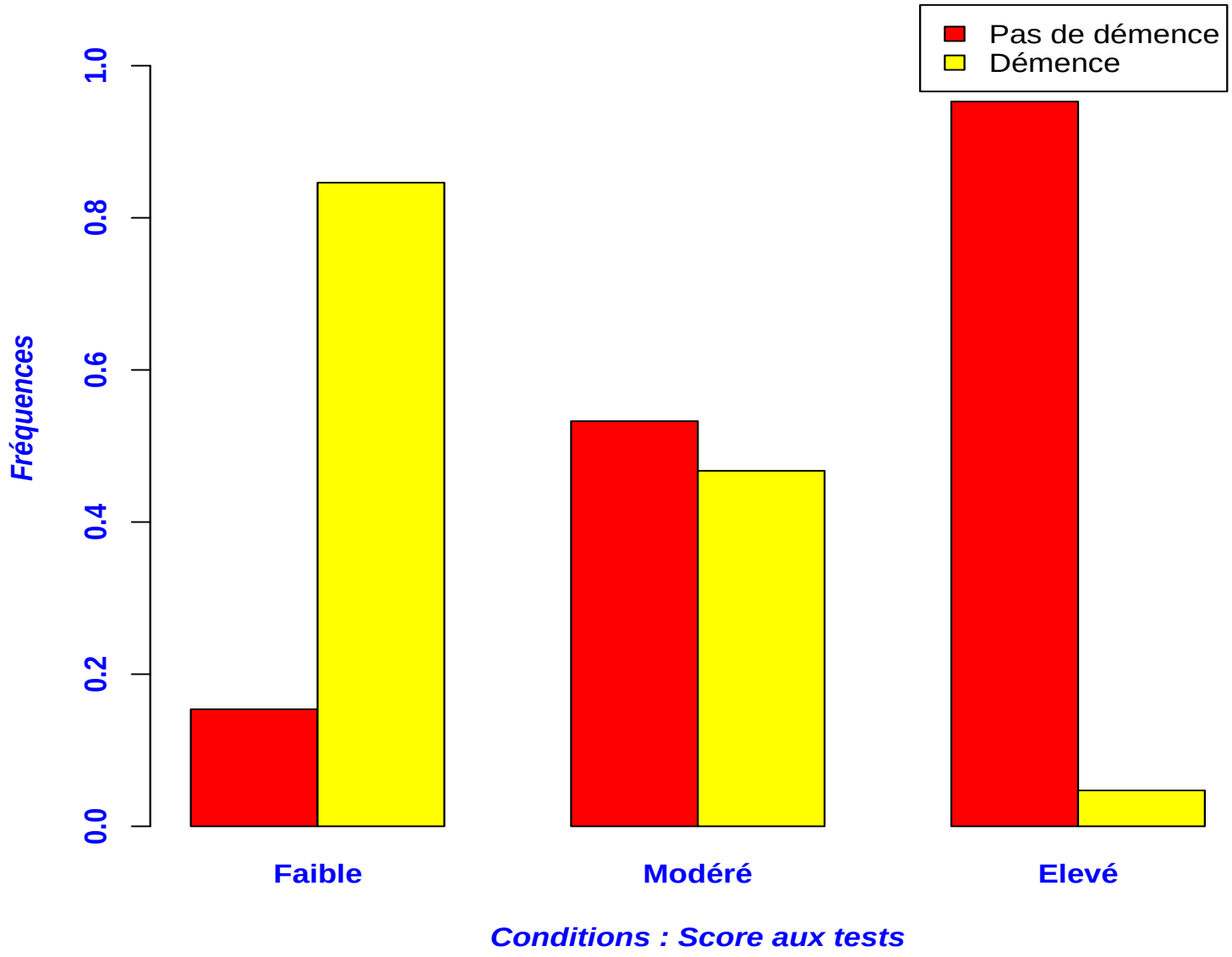
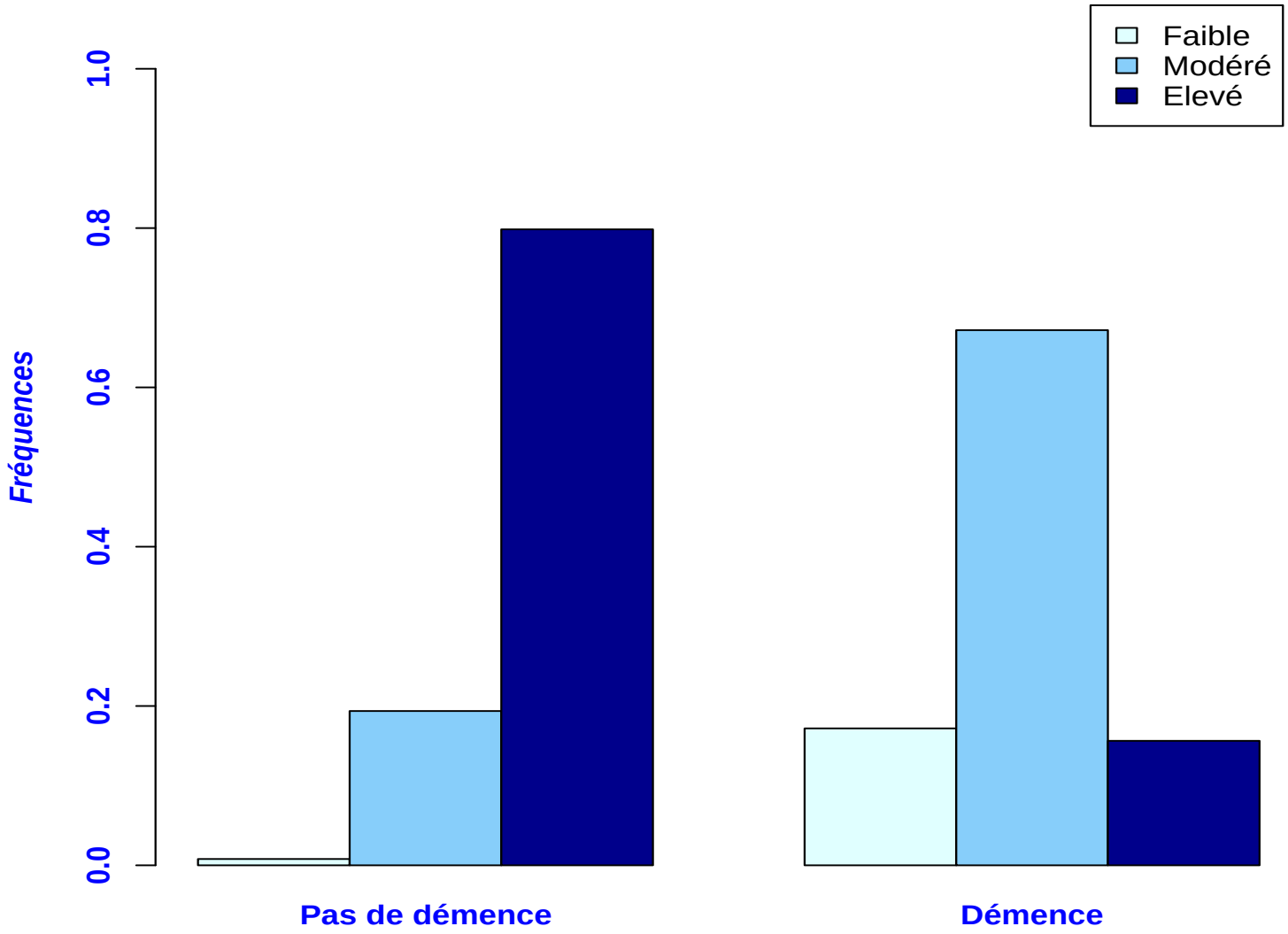


Tableau des distributions de la variable “Score aux tests” conditionnellement à chaque niveau de la variable “État clinique” :

Etat clinique	**Score aux tests**	Faible	Modéré	Élevé	Total
Pas de démence	Effectifs n_k	2	49	202	253
	Fréquences f_k	0.008	0.194	0.798	1
	Eff. cum. N_k	2	51	253	
	Fréq. cum. F_k	0.008	0.202	1	
Démence	Effectifs n_k	11	43	10	64
	Fréquences f_k	0.172	0.672	0.156	1
	Eff. cum. N_k	11	54	64	
	Fréq. cum. F_k	0.172	0.844	1	

Graphe des distributions conditionnelles de la variable 'Score aux tests'

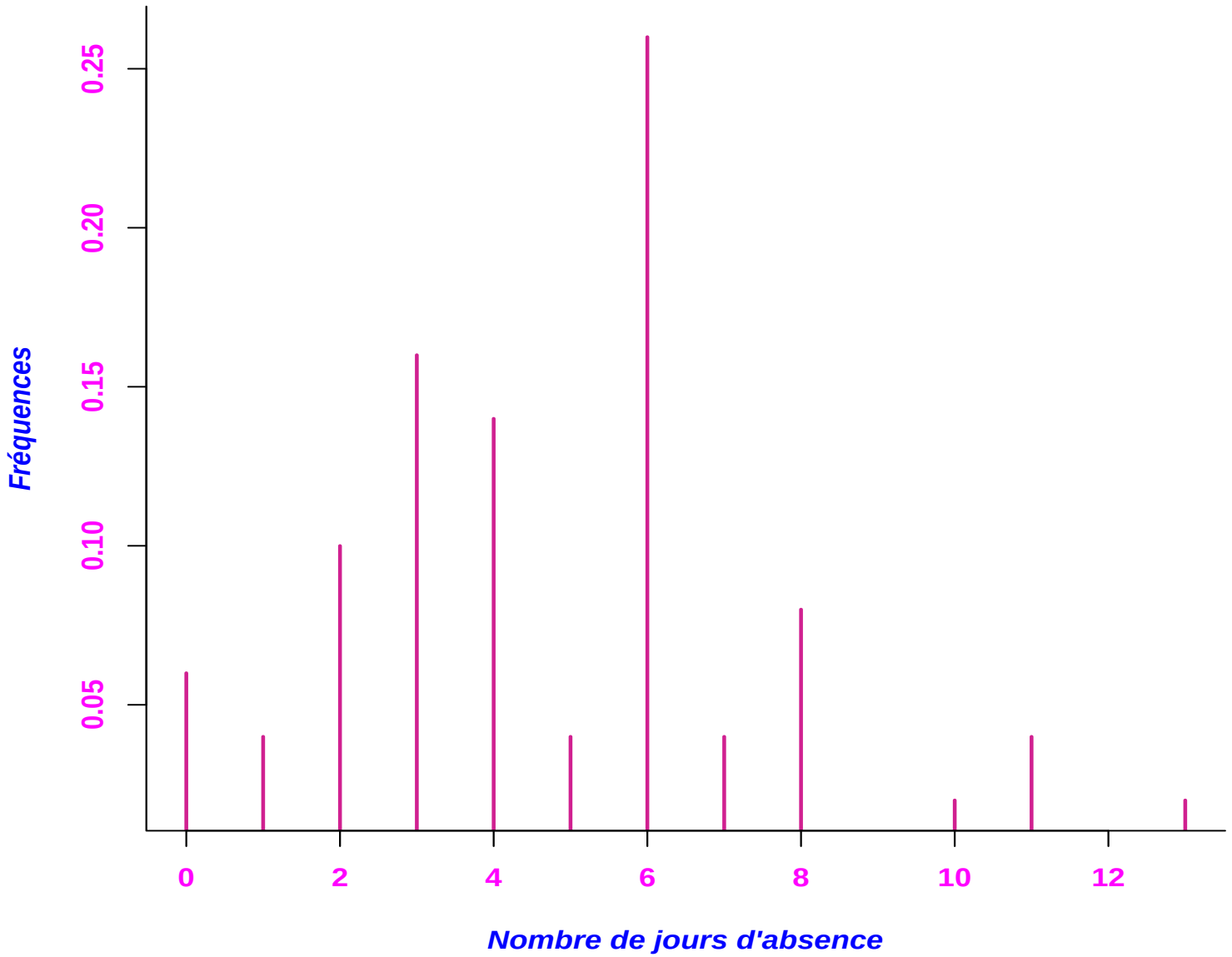


Exemple 4 : L'absentéisme salarié

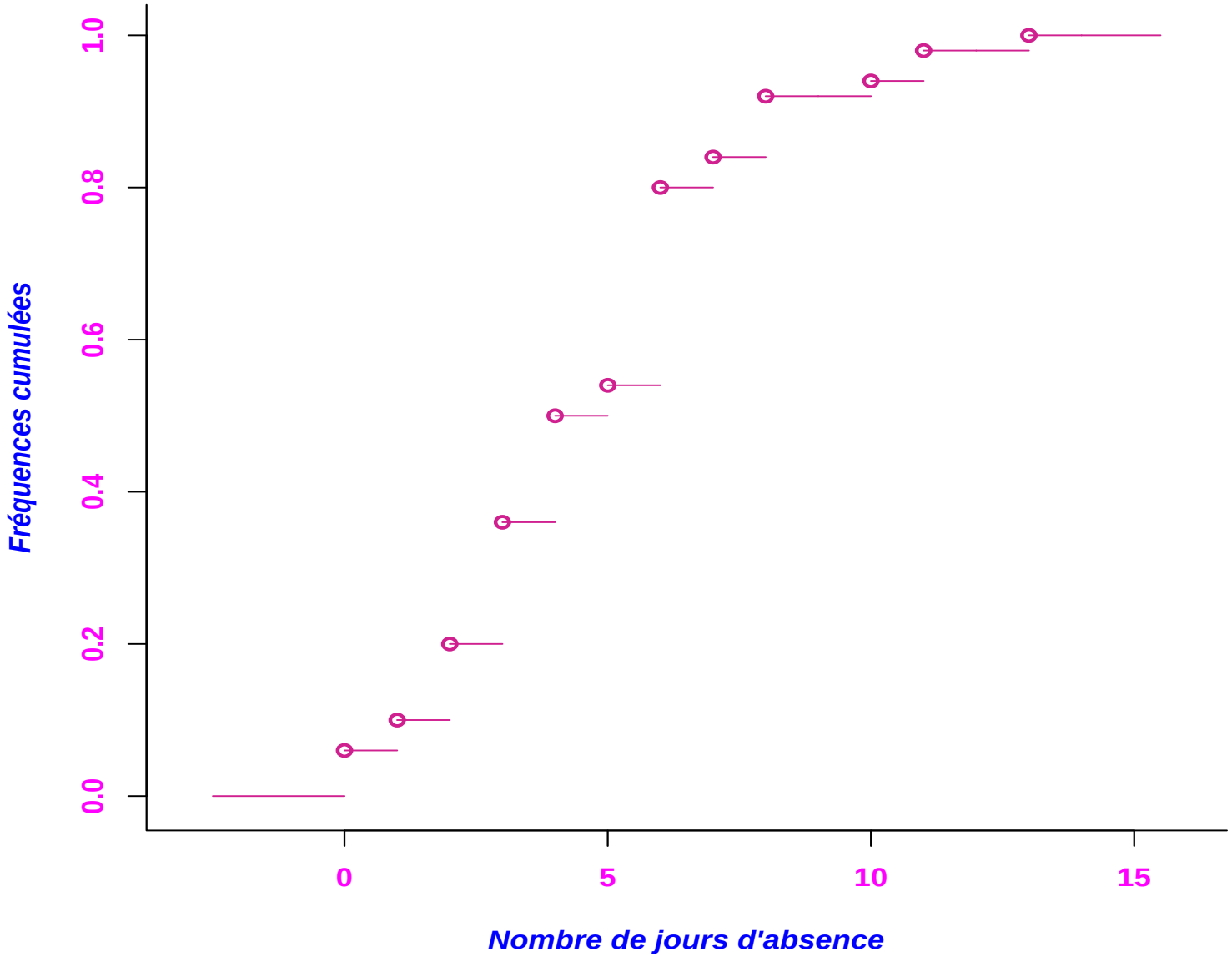
Tableau de la distribution et de la distribution cumulée :

Nbjours	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
n_k	12	8	20	32	28	8	52	8	16	0	4	8	0	4	200
f_k (en %)	6	4	10	16	14	4	26	4	8	0	2	4	0	2	100
N_k	12	20	40	72	100	108	160	168	184	184	188	196	196	200	
F_k (en%)	6	10	20	36	50	54	80	84	92	92	94	98	98	100	

Diagramme en bâtons de la distribution en fréquences



Graphe de la fonction de répartition



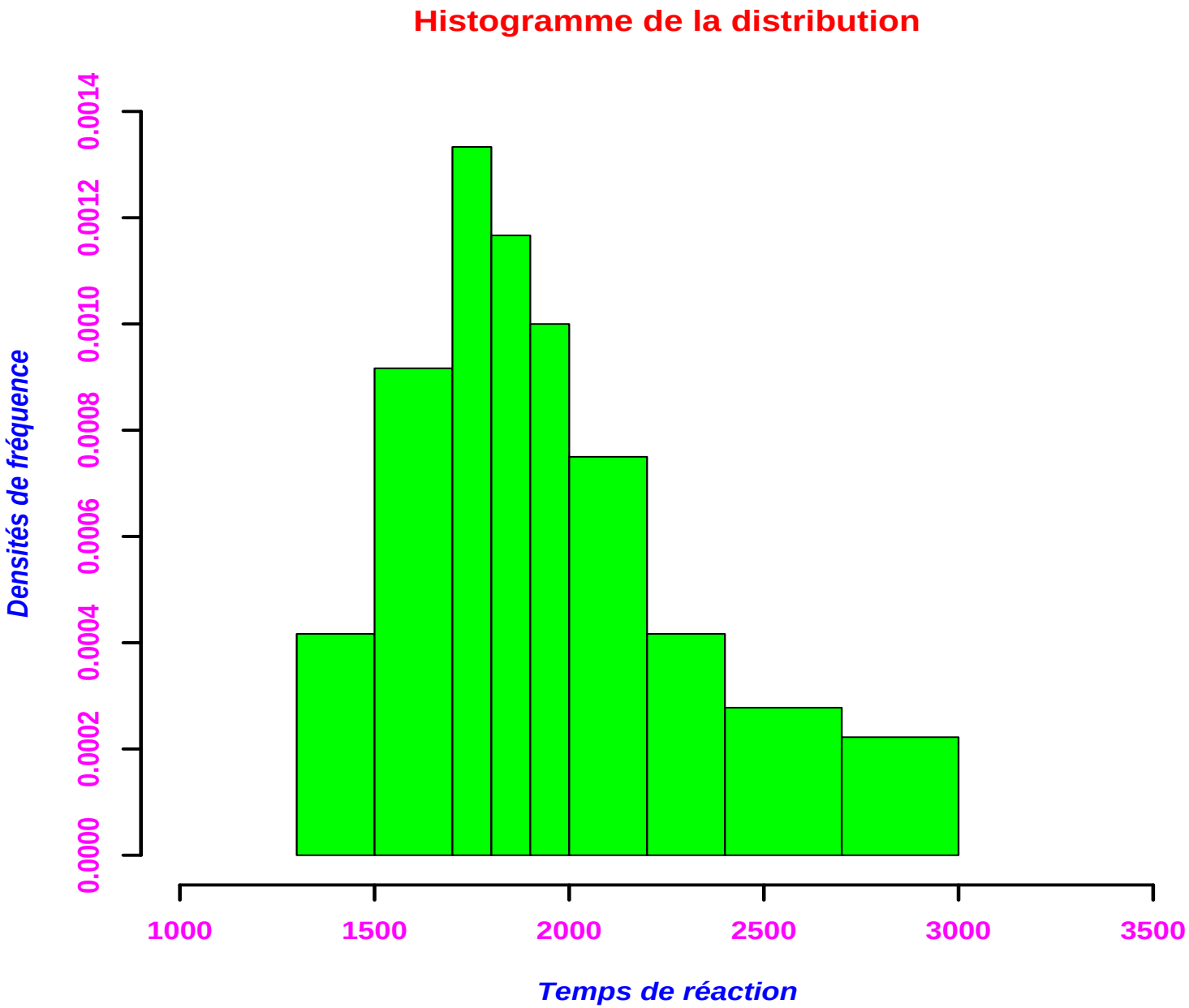
Exemple 5 : Compléxité d'une situation routière et vitesse de traitement de l'information

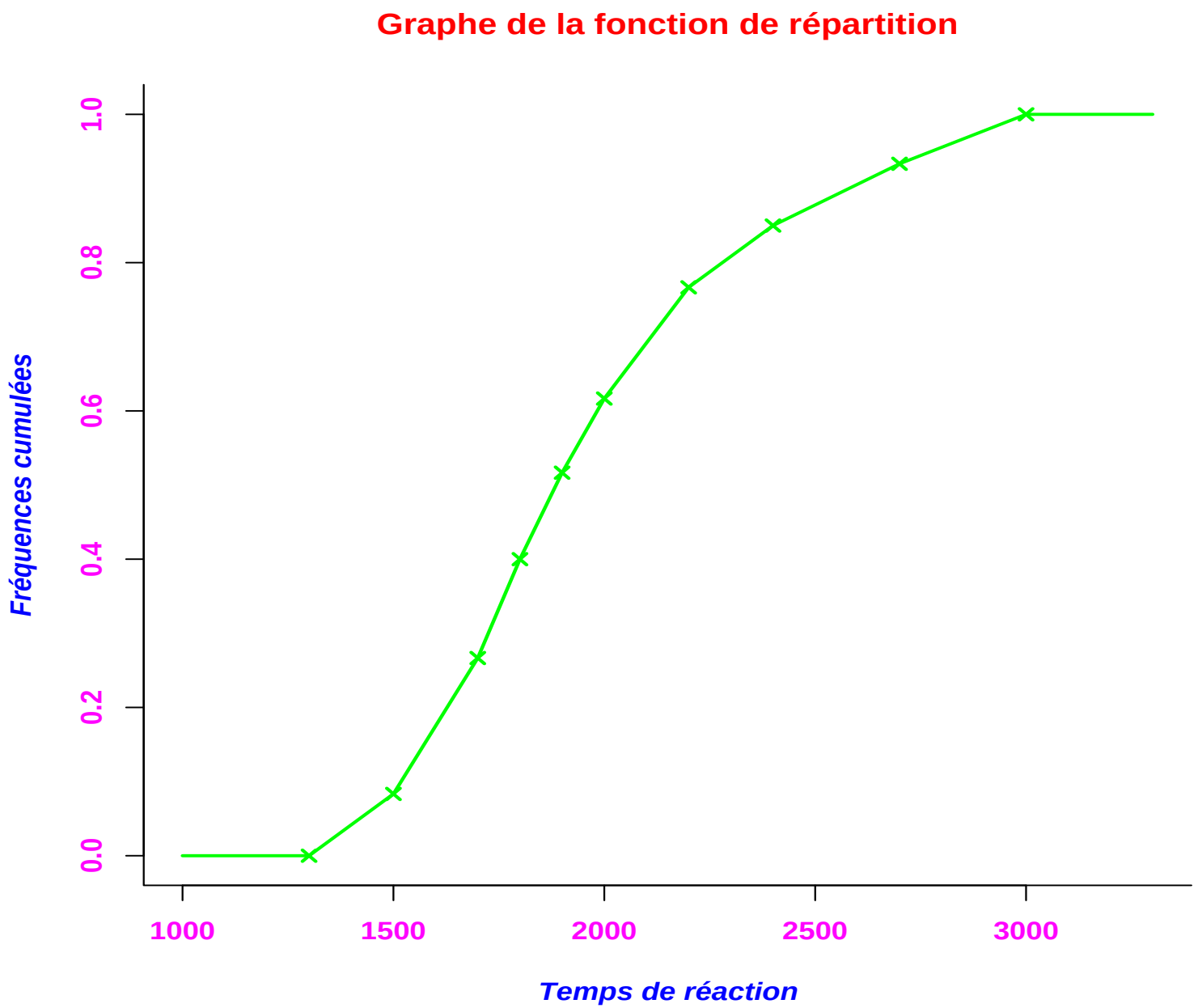
Regroupement en 9 classes.

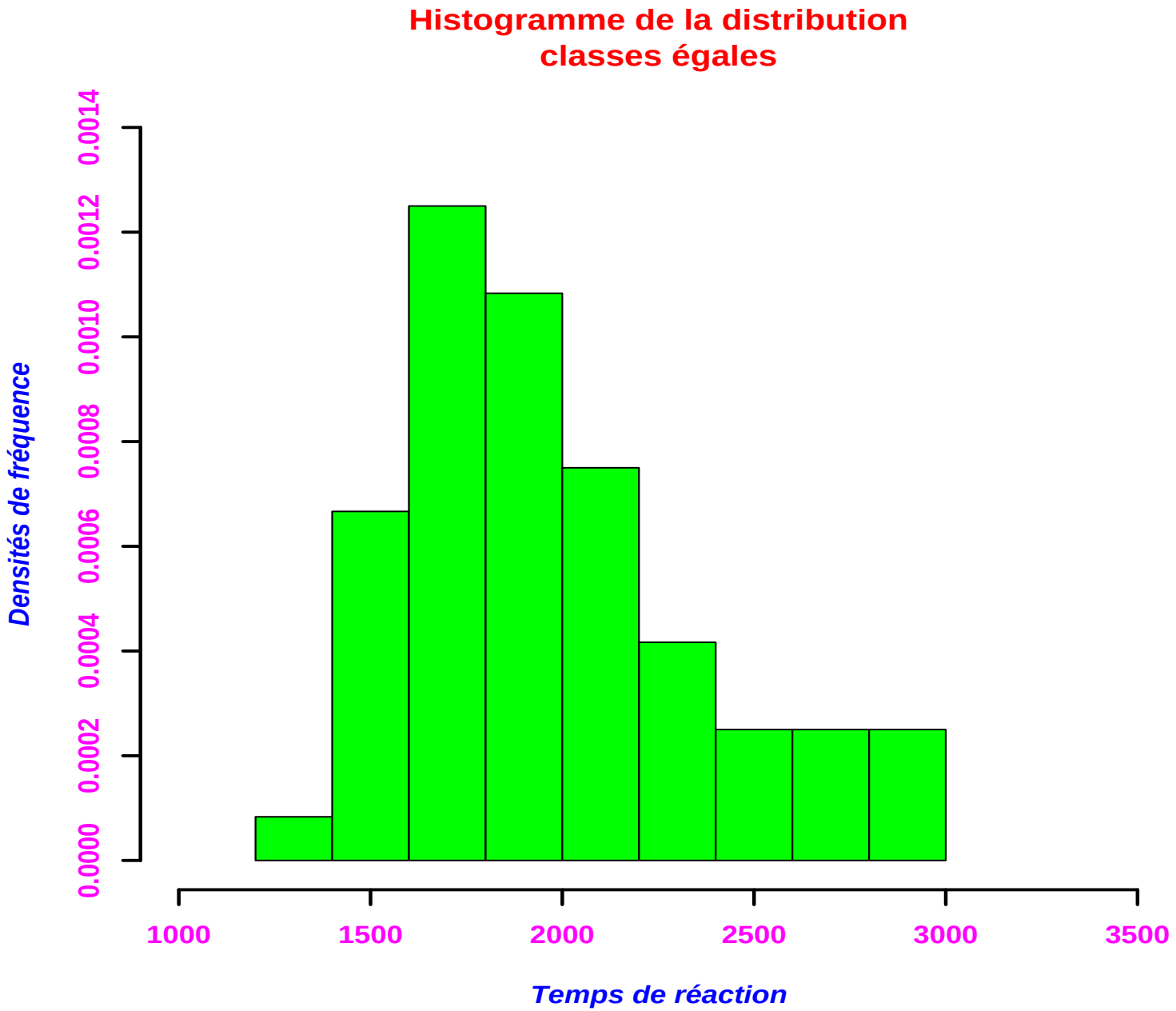
Tps	[1300; 1500[	[1500; 1700[	[1700; 1800[	
f_k	0.083	0.183	0.133	
F_k	0	0.083	0.267	0.400
a_k	200	200	100	
$d_k \times 10000$	4.17	9.17	13.33	

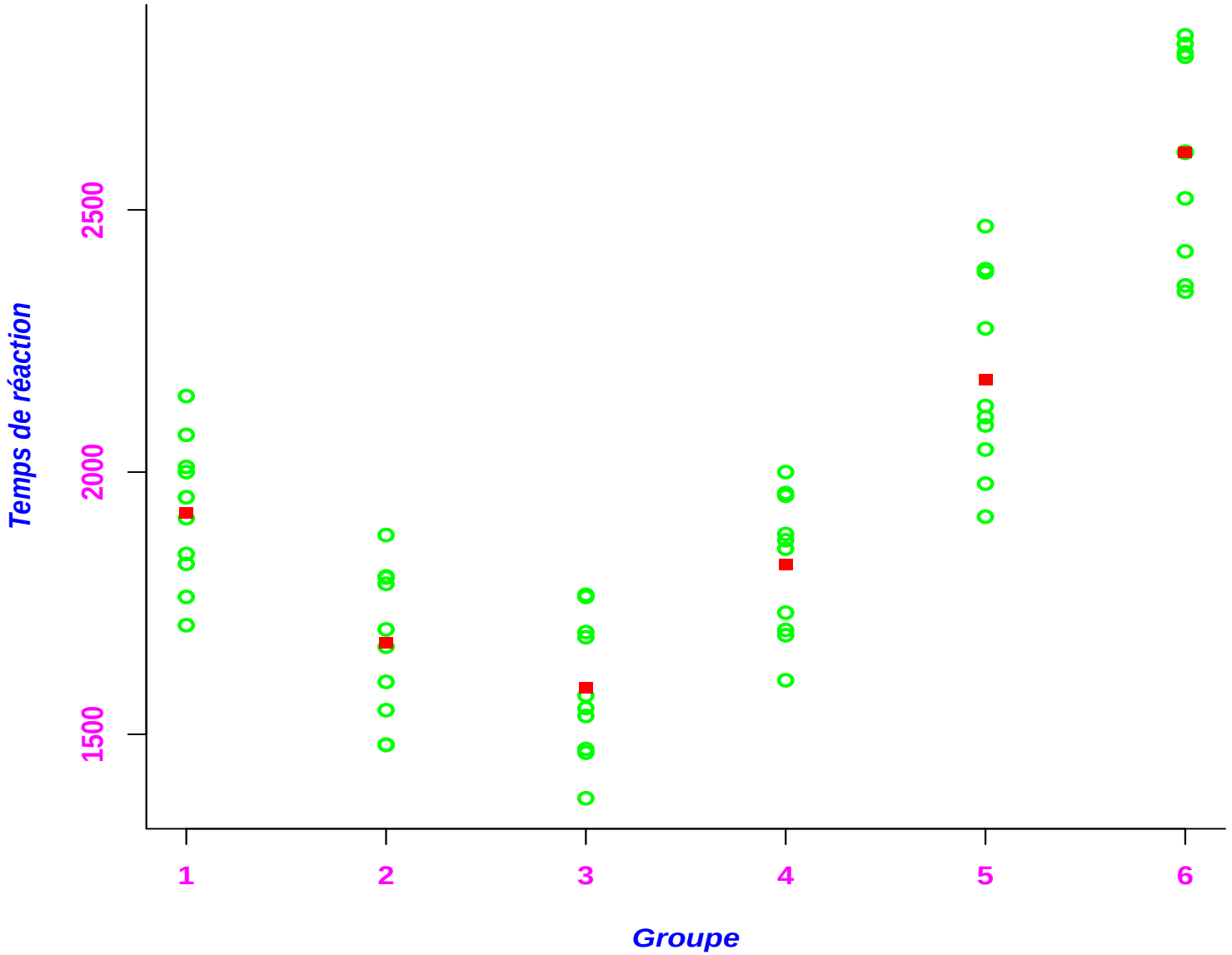
[1800; 1900[	[1900; 2000[	[2000; 2200[	[2200; 2400[	
0.117	0.100	0.150	0.083	
0.517	0.617	0.767	0.850	
100	100	200	200	
11.67	10	7.5	4.17	

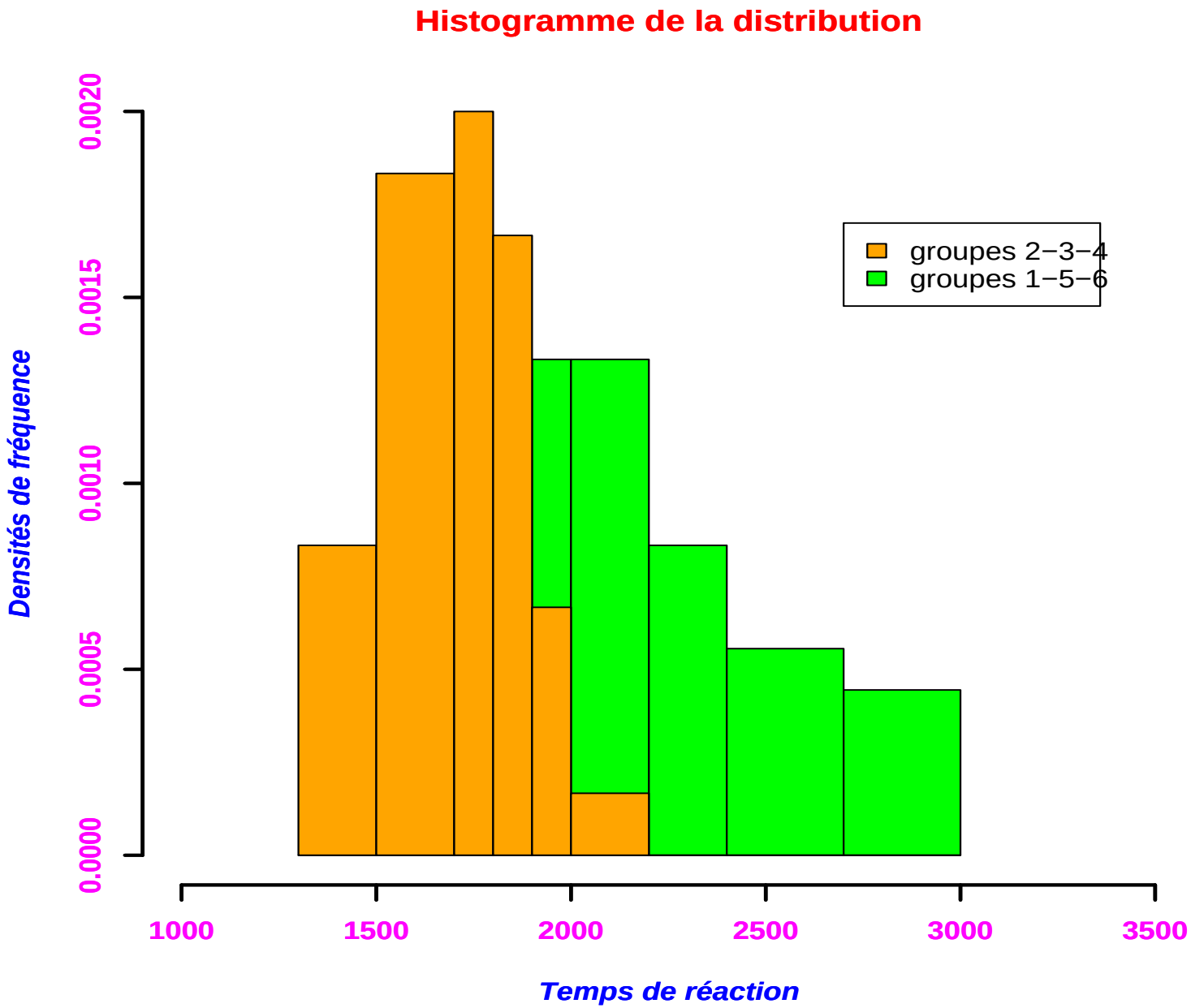
[2400; 2700[	[2700; 3000[	Tot.
0.083	0.067	1
0.933	1	
300	300	
2.78	2.22	

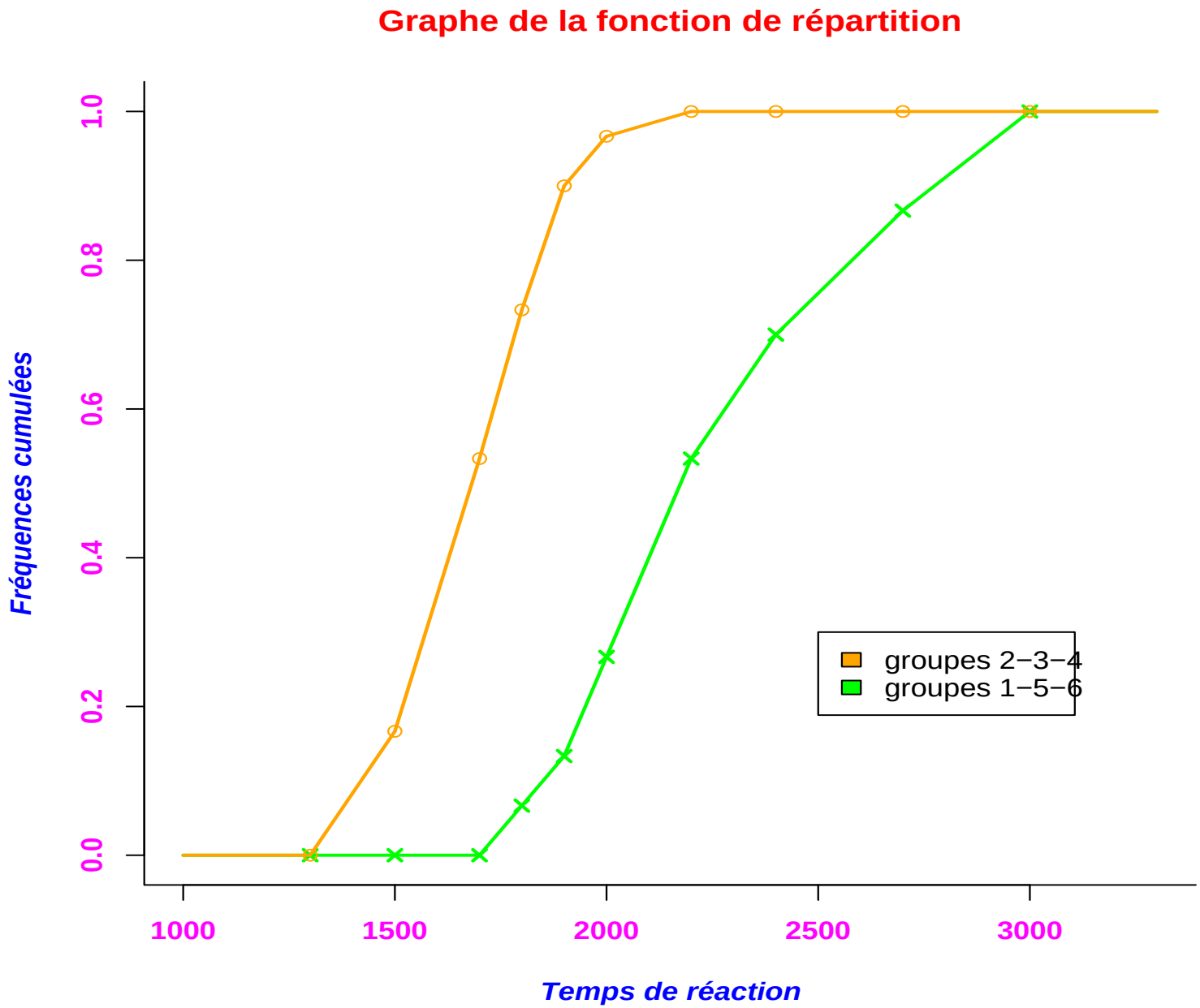

Histogramme de la distribution
Densités de fréquence
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0.0014
1000
1500
2000
2500
3000
3500
Temps de réaction


Graphe de la fonction de répartition
Fréquences cumulées
0.0
0.2
0.4
0.6
0.8
1.0
1000
1500
2000
2500
3000
Temps de réaction


Histogramme de la distribution
classes égales
Densités de fréquence
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0.0014
1000
1500
2000
2500
3000
3500
Temps de réaction


Temps de réaction
1500
2000
2500
1
2
3
4
5
6
Groupe

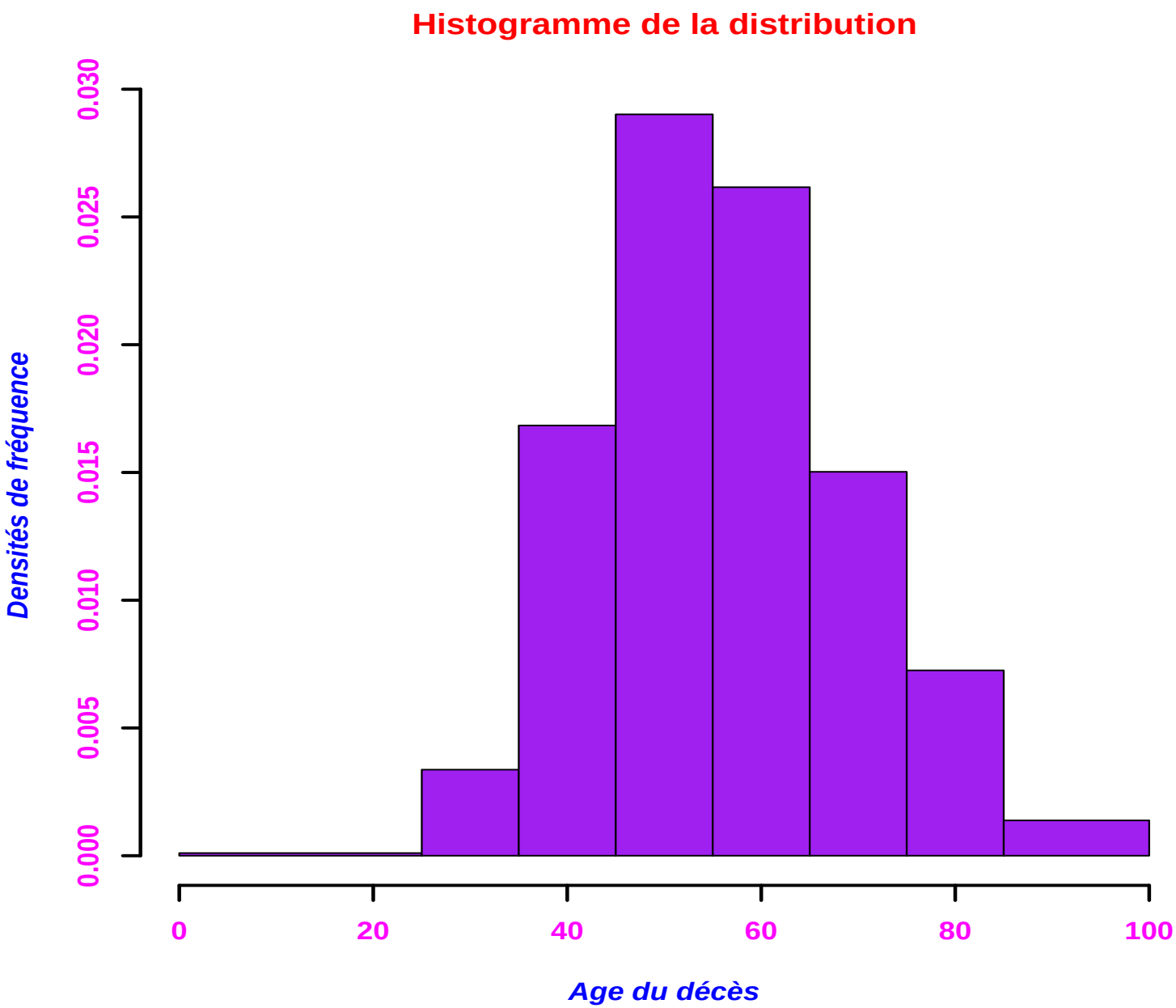

Histogramme de la distribution
Densités de fréquence
0.0000
0.0005
0.0010
0.0015
0.0020
1000
1500
2000
2500
3000
3500
Temps de réaction
groupes 2−3−4
groupes 1−5−6


Graphe de la fonction de répartition
Fréquences cumulées
0.0
0.2
0.4
0.6
0.8
1.0
1000
1500
2000
2500
3000
Temps de réaction
groupes 2−3−4
groupes 1−5−6

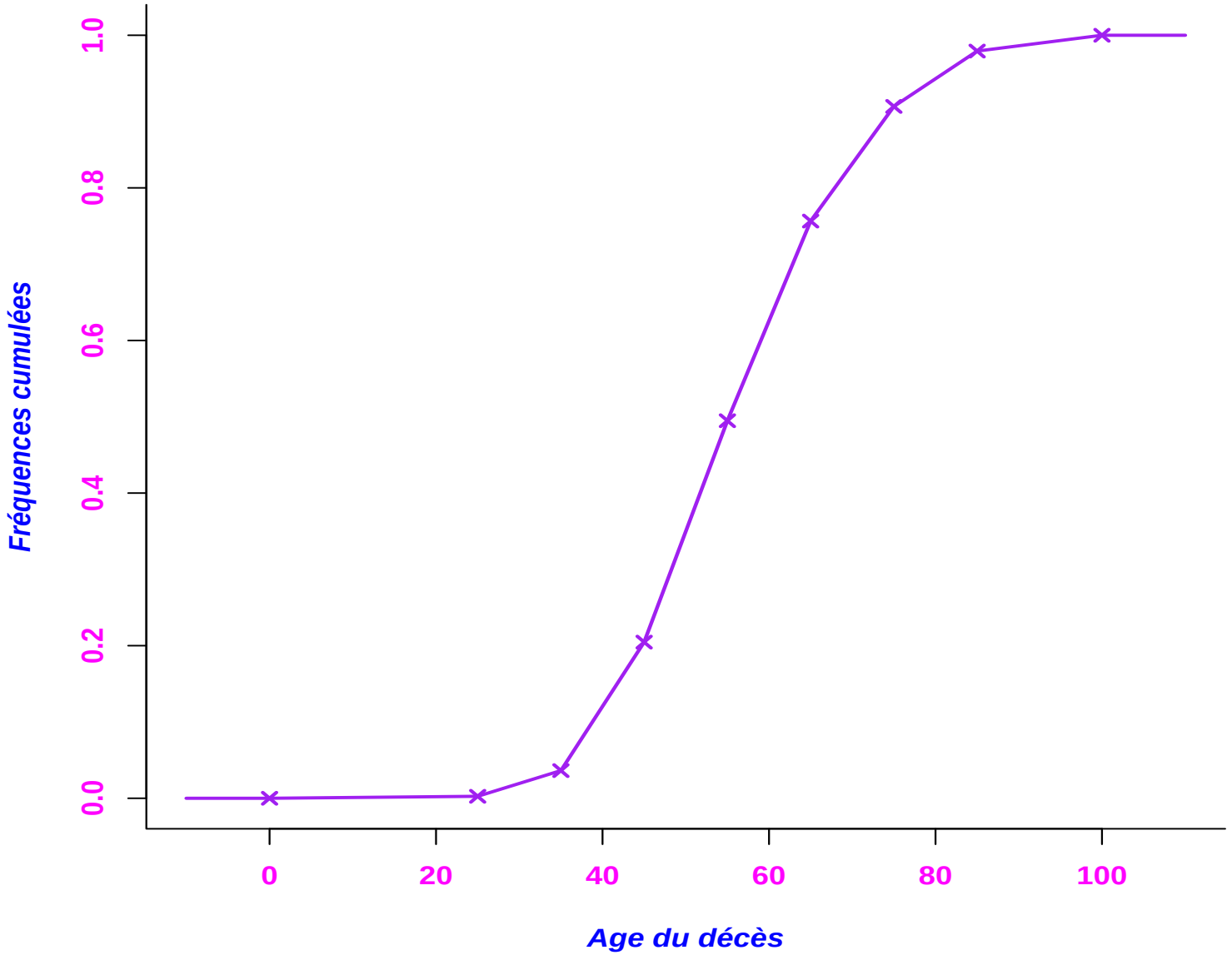
Exemple 6 : Âge des décès dûs à l'alcoolisme

Variable déjà regroupée en classes à redéfinir proprement.
Tableau de la distribution et de la distribution cumulée.

Âge	[0 ;25[	[25 ;35[	[35 ;45[	[45 ;55[	[55 ;65[	[65 ;75[	[75 ;85[	[85 ;100[	Total
f_k	0.003	0.034	0.168	0.290	0.262	0.150	0.073	0.021	1
F_k	0	0.003	0.036	0.205	0.495	0.756	0.907	0.979	1
a_k	25	10	10	10	10	10	10	15	
$d_k \times 10^3$	0.1	3.37	16.84	29.02	26.17	15.02	7.25	1.38	


Histogramme de la distribution
Densités de fréquence
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
20
40
60
80
100
Age du décès

Graphe de la fonction de répartition



À SAVOIR

1. **variable qualitative nominale** :
   - **distribution** : diagramme en barres séparées
2. **variable qualitative ordinale** :
   - **distribution** : diagramme en barres juxtaposées
3. **variable quantitative discrète** :
   - **distribution** : diagramme en bâtons
   - **distribution cumulée** : graphe de la fonction de répartition (en escalier)
4. **variable quantitative continue** :
   - **distribution** : histogramme
   - **distribution cumulée** : graphe de la fonction de répartition (linéaire par morceaux)

Chapitre 4 : Définitions et premiers indices simples

On se concentre, dans cette 2e partie du cours, sur les **variables quantitatives**. Les valeurs observées d'une variable se positionnent donc sur une **échelle** de valeur.

Tous les indices présentés visent à **résumer** la **distribution** de la variable, i.e. la **répartition** des **valeurs observées**.

I - Définitions

À la vue des observations d'une variable quantitative, on peut s'intéresser à résumer l'information de la distribution par des indices de :

• localisation : c'est une valeur qui reflète un endroit spécifique de l'échelle où se situent les valeurs observées.

$\longrightarrow$ une position sur l'échelle

• dispersion : c'est une valeur qui renseigne sur l'éloignement des valeurs observées les unes par rapport aux autres.

$\longrightarrow$ un "écartement" sur l'échelle

Attention : ces 2 types d'indices renseignent sur des notions très différentes.

II - Des indices simples

2 indices naturels de **localisation** :

- le minimum (min) : la valeur minimale observée.

$\longrightarrow$ à partir d'où se situent les valeurs sur l'échelle

- le maximum (max) : la valeur maximale observée.

$\longrightarrow$ jusqu'où se situent les valeurs sur l'échelle

1 indice naturel de **dispersion** (qui en découle) :

- l'étendue (et) : la valeur et = max - min.

$\longrightarrow$ éloignement entre la valeur **min** et la valeur **max**

III - Le mode d'une distribution

- Cas d'une variable quantitative discrète :

le mode est la valeur observable la plus fréquemment observée.

$\longrightarrow$ valeur la plus répétée dans l'échantillon, i.e. associée à l'effectif (ou la fréquence) le (la) plus élevé(e) dans le tableau de la distribution

$\longrightarrow$ valeur du bâton le plus haut dans la représentation graphique de la distribution et "marche" la plus haute dans la représentation graphique de la distribution cumulée.

- Cas d'une variable quantitative continue :

On parle d'abord de **classe modale** : la classe la plus représentée.

Attention : les classes ne sont pas toutes de même amplitude, pour les rendre comparable, on compare les densités.

$\longrightarrow$ classe de **densité maximale**

$\longrightarrow$ classe du rectangle le plus haut dans l'histogramme, i.e. de morceau de droite de pente maximale dans le graphe de la fonction de répartition.

On définit alors **le mode** comme le centre de la **classe modale**.

Le **mode** est un indice de localisation

À SAVOIR

1. **localisation** : c'est une position des valeurs sur l'échelle.
   - **minimum** : la valeur observée minimale
   - **maximum** : la valeur observée maximale
   - **mode** : le mode est une valeur observable qui concentre la plus grande quantité d'observations

2. **dispersion** : c'est un éloignement des valeurs les unes par rapport aux autres sur l'échelle.
   - **étendue** : l'écart entre le minimum et le maximum.

Chapitre 5 : Indices basés sur les rangs

On commence ici par **ranger** les individus par **ordre croissant** des valeurs de la variable.

La série des valeurs est dite **triée** et **ordonnée**.
On rappelle qu'on note alors l'**échantillon ordonné** :

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

$\longrightarrow$ le min est la première valeur dans ce classement : min $= x_{(1)}$
$\longrightarrow$ le max est la dernière valeur dans ce classement : max $= x_{(n)}$

On peut alors proposer comme **indice de localisation** une valeur qui **sépare** l'échantillon avec une certaine **proportion** d'individus ayant des **valeurs plus petites** que cette valeur là, et les autres individus ayant des **valeurs plus grandes**.

I - La médiane

Intuition : **La médiane ("Med")** est une valeur observable qui partage en **deux** effectifs égaux l'échantillon rangé par ordre croissant de la variable.

Exemples :

Quelle est la médiane des séries statistiques suivantes ?

- 3, 5, 6, 8, 10, 12, 14
- 4, 18, 12, 9, 7, 22, 10, 3, 6, 17, 14
- 1, 3, 2, 2, 3, 1, 4, 0, 2, 1, 3, 0, 2, 0, 1, 1, 3, 1, 3, 2, 2, 5, 1, 3, 5
- 1, 1, 2, 1, 2, 3, 3, 2, 1, 1, 2, 1, 3, 2, 0, 3, 3, 0, 2
- 5, 9, 19, 21, 24, 18, 43, 25, 26, 19

Définition de la médiane :

1. Lorsque le nombre d'observations est **impair**, n = 2 k + 1, alors **Med** est la $(k+1)^e$ observation de la série ordonnée : $Med = x_{(k+1)}$.
2. Lorsque le nombre d'observations est **pair**, n = 2 k, alors toutes les valeurs observables situées entre la k^e et la $(k+1)^e$ valeur sont candidates. Par convention, on choisira pour **Med** la valeur centrale (le milieu) entre la k^e et la $(k+1)^e$:

$$Med = \frac{x_{(k)} + x_{(k+1)}}{2}$$

si elle est observable, sinon la valeur observable immédiatement inférieure.

Caractérisation de la médiane :

La médiane (Med) doit vérifier simultanément les 2 propriétés P_1 et P_2 suivantes :

1. au moins 1 individu sur 2 dans l'échantillon a une valeur **inférieure ou égale** à Med :

$$P_1 \quad : \quad \begin{array}{c} freq(observations \leq Med) \geq \frac{1}{2} \\ F(Med) \geq \frac{1}{2} \end{array}$$

2. au moins 1 individu sur 2 dans l'échantillon a une valeur **supérieure ou égale** à Med :

$$P_2 \quad : \quad \begin{array}{c} freq(observations \geq Med) \geq \frac{1}{2} \\ G(Med) \geq \frac{1}{2} \end{array}$$

Aspects graphiques : Pour obtenir la médiane,

1. dans le cas d'une **variable quantitative discrète**
   - on cherche l'abscisse du ou des points d'ordonnée 0.5 du graphe de la fonction de répartition.
   - si plusieurs valeurs sont possibles (cas du plateau à hauteur 0.5), par la même convention, on prendra la valeur centrale si elle est observable ou la valeur observable immédiatement inférieure sinon.
2. dans le cas d'une **variable quantitative continue**,
   - à partir de la **fonction de répartition**, on cherche l'abscisse du point d'ordonnée 0.5 : soit par une simple lecture graphique, soit en calculant sa valeur par **interpolation linéaire** sur l'intervalle $[b_k; b_{k+1}]$ contenant cette médiane :

$$Med = b_k + (b_{k+1} - b_k) \times \frac{0.5 - F_k}{F_{k+1} - F_k}$$

   - à partir de l'**histogramme**, on cherche l'abscisse pour laquelle la surface cumulée à gauche est 0.5 : sur l'intervalle $[b_k; b_{k+1}]$ contenant cette médiane (F_k étant la surface à gauche de b_k) :

$$Med = b_k + (b_{k+1} - b_k) \times \frac{0.5 - F_k}{f_{k+1}}$$

Exercices

Un tableau de distribution d'une variable quantitative discrète :
Nbre d'enfants dans une famille d'étudiants

N^b d'enfants	N^b d'étudiants
1	7
2	99
3	47
4	12
5	6
≥ 6	9
Total	180

Donner le mode et la médiane de cette distribution ?

Un tableau de distribution d'une variable quantitative continue :
l'information étant non exhaustive (on ne dispose pas des données brutes) on ne peut calculer qu'une approximation de la médiane.

Longueur	Eff.	Fréq. (en %)	Fréq. Cumul. (en %)
de 30 à 34	6	4.00	4.00
de 34 à 36	6	4.00	8.00
de 36 à 38	20	13.33	21.33
de 38 à 40	30	20.00	41.33
de 40 à 42	37	24.67	66.00
de 42 à 44	23	15.33	81.33
de 44 à 46	20	13.33	94.67
de 46 à 50	8	5.33	100.00
Total	**150**	**100**	

Donner la classe modale et la classe médiane de cette distribution ?
Donner une valeur du mode, et une valeur approchée de la médiane ?

II - Les quartiles

Il y en a 3. Ils sont notés : Q_1, Med et Q_3.

Intuition : les 3 valeurs Q_1, Med, Q_3 sont des valeurs observables qui partagent en **quatre** effectifs égaux l'échantillon ordonné.

Q_1, Med, Q_3 sont des indices de localisation. Ils sont fondamentaux.

Exemples :

Quel est le 1er quartile (Q_1) des séries statistiques suivantes ?

- ($n = 12$) 3, 7, 8, 10, 11, 12, 14, 18, 20, 23, 24, 27
- ($n = 13$) 3, 7, 8, 10, 11, 12, 14, 18, 20, 23, 24, 27, 29
- ($n = 14$) 3, 7, 8, 10, 11, 12, 14, 18, 20, 23, 24, 27, 29, 30
- ($n = 15$) 3, 7, 8, 10, 11, 12, 14, 18, 20, 23, 24, 27, 29, 30, 35

Convention : quand plusieurs valeurs sont possibles, on choisit la valeur centrale (milieu) si elle est observable, sinon la valeur observable immédiatement inférieure.

Caractérisation :

$\longrightarrow$ Q_1 vérifie simultanément les 2 propriétés P_1 et P_2 suivantes :

1. au moins 1 individu sur 4 dans l'échantillon a une valeur **inférieure ou égale** à Q_1 : P_1 : $$\begin{aligned} freq(observations \leq Q_1) &\geq \frac{1}{4} \\ F(Q_1) &\geq \frac{1}{4} \end{aligned}$$
2. au moins 3 individus sur 4 dans l'échantillon ont une valeur **supérieure ou égale** à Q_1 : P_2 : $$\begin{aligned} freq(observations \geq Q_1) &\geq \frac{3}{4} \\ G(Q_1) &\geq \frac{3}{4} \end{aligned}$$

$\longrightarrow$ Q_3 vérifie simultanément les 2 propriétés P_1 et P_2 suivantes :

1. au moins 3 individus sur 4 dans l'échantillon ont une valeur **inférieure ou égale** à Q_3 : P_1 : $$\begin{aligned} freq(observations \leq Q_3) &\geq \frac{3}{4} \\ F(Q_3) &\geq \frac{3}{4} \end{aligned}$$
2. au moins 1 individu sur 4 dans l'échantillon a une valeur **supérieure ou égale** à Q_3 : P_2 : $$\begin{aligned} freq(observations \geq Q_3) &\geq \frac{1}{4} \\ G(Q_3) &\geq \frac{1}{4} \end{aligned}$$

Aspect graphique : Q_1, Med et Q_3 sont les valeurs où le graphe de la fonction de répartition F franchit respectivement les ordonnées 0.25, 0.5 et 0.75. On utilisera les mêmes techniques de calcul et convention que pour la médiane.

Comme pour la médiane, dans le cas d'une **variable quantitative continue**, on n'obtient donc que des **valeurs approchées** de ces indices en supposant une répartition uniforme des valeurs dans chaque classe.

Exercices

Donner la valeur des quartiles Q_1 et Q_3 dans les 2 exemples ci-dessus :

▸ Exercices

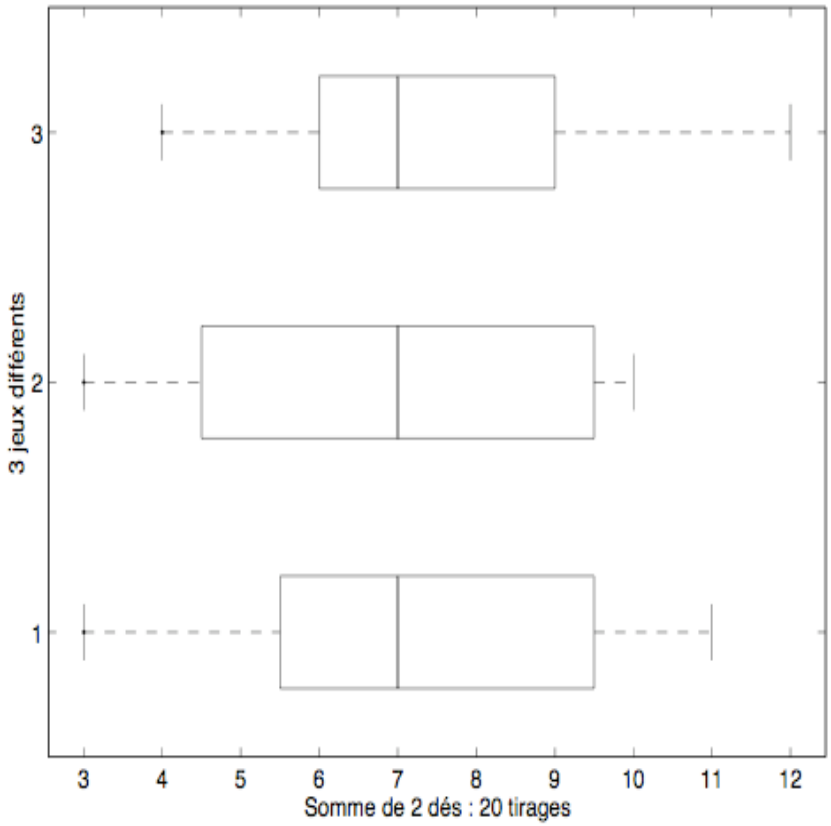
Définition :
À l'aide des quartiles, on peut définir **l'écart inter-quartiles (EIQ)** :

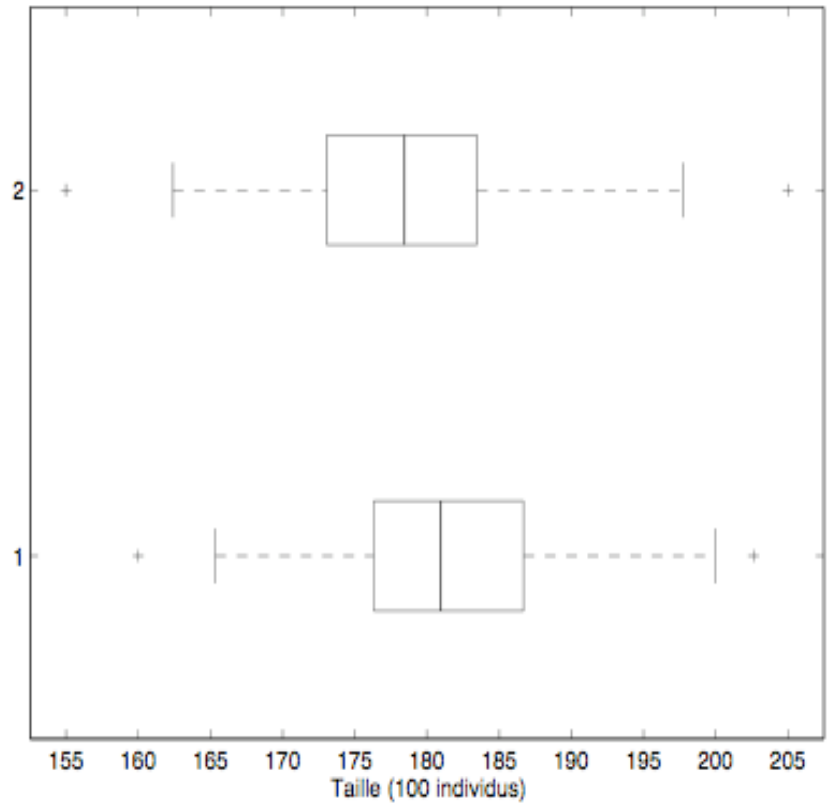
$$EIQ = Q_3 - Q_1$$

C'est l'écart entre le premier et le troisième quartile. C'est donc un indice de **dispersion**.

Représentation graphique : le box-plot

Il existe une représentation graphique fondée sur la médiane et les quartiles ("boîte à pattes", "boîte à moustaches", box and whiskers) :





Autres indices basés sur les rangs

- Les déciles (D_1 à D_9) :
  les 9 valeurs D_1, D_2, D_3, D_4, Med, D_6, D_7, D_8, D_9, partagent en **dix** effectifs égaux l'échantillon rangé par ordre croissant de la variable.
- Les centiles (C_1 à C_{99}) : ...

- De façon générale, un quantile (q_α) peut être associé à toute proportion α d'invididus dans l'échantillon ayant des valeurs plus petites que q_α et $1-\alpha$ ayant des valeurs plus grandes que q_α.

  Caractérisation : q_α vérifie simultanément les 2 propriétés :

$$\begin{array}{lrcl} P_1 : & freq(observations \leq q_\alpha) & \geq & \alpha \\ & F(q_\alpha) & \geq & \alpha \\ P_2 : & freq(observations \geq q_\alpha) & \geq & 1-\alpha \\ & G(q_\alpha) & \geq & 1-\alpha \end{array}$$

  q_α est une valeur observable de la variable.

Exemples : $q_{0.5} = q_{50\%}$=Med ou $q_{0.03} = q_{3\%} = C_3$

Tous ces indices sont des indices de localisation.

Intervalles basés sur les rangs

À l'aide des indices précédents, on peut construire de nombreux intervalles centrés autour de la médiane (intervalles centrés au sens des quantiles). Ce sont autant d'indicateurs des **positions centrales** des valeurs de l'échantillon.

- l'intervalle à 50% centré autour de la médiane : $[Q_1; Q_3]$
- l'intervalle à 80% centré autour de la médiane : $[D_1; D_9]$
- l'intervalle à 60% centré autour de la médiane : $[D_2; D_8]$
- l'intervalle à 95% centré autour de la médiane : $[q_{2.5\%}; q_{97.5\%}]$

A SAVOIR

Lorsque les observations ont été ordonnées, on peut alors définir :

1. comme paramètres de **localisation** :
   - **la médiane (Med)** : une valeur observable vérifiant qu'au moins 50% des valeurs de l'échantillon sont plus petites qu'elle et au moins 50% plus grandes qu'elle.
   - **les quartiles, déciles, centiles**
   - **le quantile d'ordre α (q_α)** : une valeur observable vérifiant qu'une proportion au moins α des valeurs de l'échantillon sont plus petites qu'elle et au moins 1-α sont plus grandes qu'elle.

2. comme paramètre de **dispersion** :
   - **l'écart inter-quartile** : l'écart entre le 1er et 3ème quartile.

Chapitre 6 : Distance, dispersion, indice de localisation centrale

I - Définition intuitive

Un indice de **localisation centrale** est :
une valeur "résumé" qui se situe **le mieux possible au milieu** des données.
Elle sera **proche de tous** les individus de l'échantillon à la fois.

On a donc besoin des 2 notions :

1. pour le mot "**proche**" on a besoin de définir une notion de **distance** entre deux valeurs.
2. pour l'expression "**proche de tous**" on a besoin de définir une notion de **dispersion** des valeurs de l'échantillon autour d'un point.

II - Distance et dispersion

Soit une variable X observée sur n individus. On continue de désigner par $x_1, x_2, \ldots, x_i, \ldots x_n$, les n valeurs observées.

Soit a une valeur réelle quelconque (une position quelconque sur l'échelle). On définit :

- $d(x_i, a)$ **: la distance** entre la valeur x_i et la valeur a

  On choisit généralement comme distance **naturelle** :
  1. la distance au sens des **valeurs absolues** : $d(x_i, a) = |x_i - a|$
  2. la distance au sens des **carrés** : $d(x_i, a) = (x_i - a)^2$

- $\mathbf{Disp}(a) = \sum_{i=1}^{n} d(x_i, a)$ **: la dispersion** totale des valeurs observées de la variable X (les valeurs de l'échantillon) autour de a et associée à la distance d.
  C'est un résumé (la somme) des distances de toutes les valeurs de l'échantillon à la valeur a.

Exemple numérique

pour la variable X , les x_i	5	8	9	15	18
pour $a = 2$, les $\lvert x_i - a \rvert$	3	6	7	13	16
pour $a = 2$, les $(x_i - a)^2$	9	36	49	169	256
pour $a = 8$, les $\lvert x_i - a \rvert$	3	0	1	7	10
pour $a = 8$, les $(x_i - a)^2$	9	0	1	49	100
pour $a = 12$, les $\lvert x_i - a \rvert$	7	4	3	3	6
pour $a = 12$, les $(x_i - a)^2$	49	16	9	9	36

	$a = 2$	$a = 8$	$a = 12$
$\text{Disp}^{(\lvert\rvert)}(a) = \sum_{i=1}^{5} \lvert x_i - a \rvert$	45	21	23
$\text{Disp}^{(2)}(a) = \sum_{i=1}^{5} (x_i - a)^2$	519	159	119

La valeur $a = 8$ est plus proche des données que $a = 2$ et $a = 12$ pour la dispersion $\text{Disp}^{(\lvert\rvert)}$

La valeur $a = 12$ est plus proche des données que $a = 2$ et $a = 8$ pour la dispersion $\text{Disp}^{(2)}$

Cas d'une variable discrète présentée sous la forme du tableau de distribution.

n_1 individus prennent la valeur v_1
n_2 individus prennent la valeur v_2
$\vdots$
n_C individus prennent la valeur v_C

La dispersion $\mathbf{Disp}(a) = \sum_{i=1}^{n} d(x_i, a)$ peut alors aussi s'écrire :

$$\mathrm{Disp}(a) = \sum_{k=1}^{C} n_k d(v_k, a)$$

Ce qui pour les 2 distances envisagées donne :

$$\mathrm{Disp}^{(|)}(a) = \sum_{k=1}^{C} n_k |v_k - a| \qquad \mathrm{Disp}^{(2)}(a) = \sum_{k=1}^{C} n_k (v_k - a)^2$$

Exemple

▸ Nbre d'enfants dans une famille d'étudiants : cas de la $\text{Disp}^{(\|)}$

							$\text{Disp}^{(\|)}$
	v_k	1	2	3	4	5	6	
	n_k	7	99	47	12	9	6	
$a=1$:	$\lvert v_k - a \rvert$	0	1	2	3	4	5	
$a=1$:	$n_k \times \lvert v_k - a \rvert$	0	99	94	36	36	30	295
$a=2$:	$\lvert v_k - a \rvert$	1	0	1	2	3	4	
$a=2$:	$n_k \times \lvert v_k - a \rvert$	7	0	47	24	27	24	**129**
$a=3$:	$\lvert v_k - a \rvert$	2	1	0	1	2	3	
$a=3$:	$n_k \times \lvert v_k - a \rvert$	14	99	0	12	18	18	161

La valeur $a=2$ est plus proche des données que $a=1$ et $a=3$ pour la dispersion $\text{Disp}^{(\|)}$

▸ Nbre d'enfants dans une famille d'étudiants : cas de la $\text{Disp}^{(2)}$

							$\text{Disp}^{(2)}$
	v_k	1	2	3	4	5	6	
	n_k	7	99	47	12	9	6	
$a=1$:	$(v_k-a)^2$	0	1	4	9	16	25	
$a=1$:	$n_k \times (v_k-a)^2$	0	99	188	108	144	150	689
$a=2$:	$(v_k-a)^2$	1	0	1	4	9	16	
$a=2$:	$n_k \times (v_k-a)^2$	7	0	47	48	81	96	279
$a=3$:	$(v_k-a)^2$	4	1	0	1	4	9	
$a=3$:	$n_k \times (v_k-a)^2$	28	99	0	12	36	54	**229**

La valeur $a=3$ est plus proche des données que $a=2$ et $a=1$ pour la dispersion $\text{Disp}^{(2)}$

III - Indice de localisation centrale

Un **indice de localisation centrale** est la valeur a :

- qui **minimise** la dispersion
- telle que $\text{Disp}(a) = \sum_{i=1}^{n} d(x_i, a)$ est **minimale**

$\longrightarrow$ Un tel indice **dépend** beaucoup de la **distance** d **choisie**.

a - Pour la dispersion $\text{Disp}^{(|)}$

pour la variable X , les x_i	5	8	9	15	18
pour $a = 2$, les $\lvert x_i - a \rvert$	3	6	7	13	16
pour $a = 8$, les $\lvert x_i - a \rvert$	3	0	1	7	10
pour $a = 9$, les $\lvert x_i - a \rvert$	4	1	0	6	9
pour $a = 10$, les $\lvert x_i - a \rvert$	5	2	1	5	8
pour $a = 12$, les $\lvert x_i - a \rvert$	7	4	3	3	6

	$a = 2$	$a = 8$	$a = 9$	$a = 10$	$a = 12$
$\text{Disp}^{(|)}(a)$	45	21	20	21	23

On peut montrer que $a = 9$ est **la valeur la plus proche** de ces données pour la dispersion $\text{Disp}^{(|)}$

Propriété

- Avec la distance $d(x_i, a) = |x_i - a|$, le minimum de la dispersion est obtenu pour a = Med (la médiane des observations).

▸ Nbre d'enfants dans une famille d'étudiants

- La valeur de la dispersion des observations autour de la médiane est :

$$\text{Disp}^{(||)}(\text{Med}) = \sum_{i=1}^{n} |x_i - \text{Med}|$$

C'est **l'écart absolu à la médiane**.

- En divisant l'expression précédente par le nombre d'observations :

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - \text{Med}|$$

on parle d'**écart absolu moyen à la médiane**.

b - Pour la dispersion $\text{Disp}^{(2)}$

pour la variable X , les x_i	5	8	9	15	18
pour $a = 2$, les $(x_i - a)^2$	9	36	49	169	256
pour $a = 8$, les $(x_i - a)^2$	9	0	1	49	100
pour $a = 10$, les $(x_i - a)^2$	25	4	1	25	64
pour $a = 11$, les $(x_i - a)^2$	36	9	4	16	49
pour $a = 12$, les $(x_i - a)^2$	49	16	9	9	36

	$a = 2$	$a = 8$	$a = 10$	$a = 11$	$a = 12$
$\text{Disp}^{(2)}(a)$	519	159	119	114	119

On peut montrer que :

$$\text{Disp}^{(2)}(11 + x) = 114 + 5 \times x^2$$

$a = 11$ est **la valeur la plus proche** des données pour la dispersion $\text{Disp}^{(2)}$

Propriété

- Avec la distance $d(x_i, a) = (x_i - a)^2$, le minimum de la dispersion est obtenu pour $a = \frac{1}{n}\sum_{i=1}^{n} x_i$ (la moyenne des observations notée $\overline{x}$).
- La valeur de la dispersion des observations autour de la moyenne est :

$$\text{Disp}^{(2)}(\overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

C'est **l'écart carré à la moyennne**.

- La dispersion **moyenne** des observations autour de la moyenne s'obtient alors en divisant la dispersion par le nombre d'observations :

$$\text{Var}(x) = \frac{1}{n}\text{Disp}^{(2)}(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

on parle d'**écart carré moyen à la moyenne**. Cette valeur s'appelle la **variance**.

"la moyenne des carrés des écarts à la moyenne".

IV - Moyenne, variance, écart-type

a - Définition

1. La **moyenne** a été définie comme la valeur de a qui minimise la dispersion des données de l'échantillon autour de a.
2. La **variance** a été définie comme la valeur de cette dispersion minimale moyenne (dispersion moyenne autour de la moyenne).
3. L'**écart-type** est défini comme la racine carrée de la variance

$$\sigma(x) = \sqrt{\mathrm{Var}(x)}$$

$\longrightarrow$ la moyenne est un **indice de localisation** ;
$\longrightarrow$ la variance et l'écart-type sont des **indices de dispersion** ;
mais contrairement à la variance, l'**écart-type** est un indice "compréhensible" puisque de **même unité** que les données.

Attention : il est incohérent d'associer à la moyenne une mesure de dispersion qu'elle ne minimise pas. Autrement dit, il est incohérent d'associer à la moyenne une distance autre que le carré.

b - Calcul de la moyenne et de la variance

Cas d'une variable quantitative discrète présentée sous la forme du tableau de distribution

n_1 individus prennent la valeur v_1
n_2 individus prennent la valeur v_2
$\vdots$
n_C individus prennent la valeur v_C

la **moyenne** peut alors aussi s'écrire :

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{C} n_k v_k$$

et la **variance** :

$$\mathrm{Var}(x) = \frac{1}{n} \sum_{k=1}^{C} n_k (v_k - \bar{x})^2$$

Exemple

▸ Nbre d'enfants dans une famille d'étudiants

$$\bar{x} = \frac{1}{180}(7v_1 + 99v_2 + 47v_3 + 12v_4 + 9v_5 + 6v_6) = 2.64.$$

							$\text{Disp}^{(2)}$
v_k	1	2	3	4	5	6	
n_k	7	99	47	12	9	6	
$(v_k-2)^2$	1	0	1	4	9	16	
$n_k(v_k-2)^2$	7	0	47	48	81	96	279
$(v_k-3)^2$	4	1	0	1	4	9	
$n_k(v_k-3)^2$	28	99	0	12	36	54	229
$(v_k-\bar{x})^2$	2.7	0.41	0.13	1.9	5.6	11.3	
$n_k(v_k-\bar{x})^2$	18.8	40.6	6.1	22.2	50.1	67.7	**205.5**

La moyenne $a = 2.64$ est la valeur la plus proche des données pour la dispersion $\text{Disp}^{(2)}$

La dispersion vaut 205.5, la variance vaut 1.142, l'écart-type vaut 1.07.

Cas d'une variable continue présentée sous la forme du tableau de distribution

On dispose de l'**information non exhaustive** suivante :

n_k individus prennent une valeur dans l'intervalle $[b_{k-1}; b_k[$

On choisit alors chaque **milieu de classe** comme "représentant" des valeurs dans la classe.

c_1 est le milieu de l'intervalle $[b_0; b_1[$

$\vdots$

c_k est le milieu de la classe $[b_{k-1}; b_k]$

On calcule alors une **approximation** de la moyenne par l'expression

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{C} n_k c_k$$

Remarque : cette expression est très similaire à celle pour une variable quantitative discrète avec ici C classes et C centres de classes à la place des valeurs v_k.

Exemple

Longueur	**Eff.** n_k	c_k	$n_k \times c_k$	**Fréq.** f_k	**Cumul** F_k
de 30 à 34	6	32	192	4.00	4.00
de 34 à 36	6	35	210	4.00	8.00
de 36 à 38	20	37	740	13.33	21.33
de 38 à 40	30	39	1170	20.00	41.33
de 40 à 42	37	41	1517	24.67	66.00
de 42 à 44	23	43	989	15.33	81.33
de 44 à 46	20	45	900	13.33	94.67
de 46 à 50	8	48	384	5.33	100.00
Total	**150**		**6102**	**100**	

$$\overline{x} \approx \mathbf{40.68}$$

Remarque : Calcul de la moyenne et la variance en utilisant les fréquences

Lorsque l'on dispose de l'information sous forme de **tableau de distribution**, on peut calculer à l'aide des fréquences ces 2 indices.

- Pour une variable quantitative **discrète** :

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{k=1}^{C} n_k v_k \\ &= \sum_{k=1}^{C} \frac{n_k}{n} v_k \\ &= \sum_{k=1}^{C} f_k v_k \end{aligned}$$

et la **variance** :

$$\mathrm{Var}(x) = \sum_{k=1}^{C} f_k (v_k - \bar{x})^2$$

- Pour une variable quantitative **continue** :

On reprend les mêmes expressions en remplaçant les v_k par c_k.

Quelques remarques :

- On peut vérifier que :

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

"la moyenne des carrés moins le carré de la moyenne".

- La moyenne a toujours une précision plus "fine" que les observations. *Dans le cas d'une variable quantitative discrète elle ne sera donc jamais (ou presque) une valeur entière ; ce n'est pas nécessairement une valeur observable.*
- La médiane est de la même précision que les observations. *Dans le cas d'une variable quantitative discrète elle sera toujours une valeur entière ; ce sera toujours une valeur observable.*
- La proximité des 3 indices mode, médiane, moyenne traduit la symétrie de la distribution.
  Exemple : moyenne $\approx$ médiane $\approx$ **40.7**
  le mode est aussi dans l'intervalle [40 ; 42]

V - Intervalle basé sur la moyenne et l'écart-type

Pour une distribution symétrique et unimodale, on appelle **intervalle de dispersion centré autour de la moyenne**, l'intervalle :

$$\mathrm{ID}(X) = [\overline{x} - 2 \times \sigma(x), \overline{x} + 2 \times \sigma(x)]$$

Il contient approximativement 95% des valeurs **de la population**.

À SAVOIR

1. **Distance** : on définit la distance $d(x_i, a)$ entre une valeur x_i de l'échantillon et une position quelconque a sur l'échelle par :
   - **en valeur absolu** : $d(x_i, a) = |x_i - a|$
   - **au carré** : $d(x_i, a) = (x_i - a)^2$
2. **Dispersion** : on définit la dispersion $Disp(a)$ des valeurs de l'échantillon autour de la position a par : $Disp(a) = \sum_{i=1}^{n} d(x_i, a)$
3. La **médiane** minimise la dispersion en valeur absolue des observations de l'échantillon autour de a.
4. La **moyenne** minimise la dispersion au carré des observations de l'échantillon autour de a.
5. La **variance** est la dispersion au carré moyenne des observations de l'échantillon autour de la moyenne.
6. L' **écart-type** est la racine carré de la variance.
7. L' **intervalle de dispersion** est l'intervalle $[\overline{x} - 2\sigma(x), \overline{x} + 2\sigma(x)]$. Il correspond, pour une variable dont la distribution est symétrique, à un intervalle contenant environ 95% de ses valeurs dans la population.

Chapitre 7 : Données centrés réduites, Autres indices

I - Données centrées réduites

Il existe 2 **transformations** qui sont classiquement appliquées aux observations d'une même variable. Ce sont les opérations de :

1. **centrage** : on **retranche** la moyenne à toutes les valeurs

$$y_i = x_i - \overline{x}$$

les y_i sont appelées **données (valeurs) centrées**.

2. **centrage et réduction** : on **divise** toutes les valeurs centrées par **l'écart-type**

$$z_i = \frac{x_i - \overline{x}}{\sigma_x}$$

les z_i sont appelées **données (valeurs) centrées réduites**.

Les valeurs ainsi obtenues (les y_i et z_i) sont dites :

centrées = de moyenne nulle
réduites = de variance égale à 1

On a donc :

- pour la **variable centrée**

$$\overline{y} = 0 \quad \sigma_y = \sigma_x$$

- pour la **variable centrée, réduite**

$$\overline{z} = 0 \quad \sigma_z = 1$$

Exemple 1

						$\sum$
les x_i	5	8	9	15	18	**55**

$$\bar{x} = 11$$

						$\sum$
les $y_i = x_i - 11$	-6	-3	-2	4	7	**0**
les $y_i^2 = (x_i - 11)^2$	36	9	4	16	49	**114**

$$\sigma_x^2 = 22.8 \qquad \text{et} \qquad \sigma_x = 4.775$$

						$\sum$
les $z_i = \dfrac{(x_i - 11)}{\sigma}$	-1.26	-0.63	-0.42	0.84	1.47	**0**
les $z_i^2 = \dfrac{(x_i - 11)^2}{\sigma^2}$	1.58	0.39	0.18	0.70	2.15	**5**

Exemple 2

▸ Nbre d'enfants dans une famille d'étudiants :

$$\bar{x} = 2.64 \ , \ \sigma_x^2 = 1.142 \text{ et } \sigma_x = 1.07$$

							$\sum$
v_k	1	2	3	4	5	6	
n_k	7	99	47	12	9	6	**180**
$(v_k - \bar{x})$	-1.64	-0.64	0.36	1.36	2.36	3.36	
$n_k(v_k - \bar{x})$	-11.47	-63.25	16.97	16.33	21.25	20.17	**0**
$\frac{(v_k - \bar{x})}{\sigma_x}$	-1.532	-0.597	0.337	1.272	2.207	3.141	
$n_k \frac{(v_k - \bar{x})^2}{\sigma_x^2}$	16.46	35.39	5.37	19.47	43.94	59.35	**180**

Interprétation

1. La valeur centrée réduite (ou score centré réduit) permet de **situer** le score d'un individu **comparativement à un groupe** constitué de l'échantillon et ceci **indépendamment de l'échelle de mesure**.

   Pour l'individu numéro i,
   - un score centré réduit $z_i = 1.5$ signifie que son score brut x_i s'est situé au dessus de la moyenne à une distance de 1.5 écart-type.
   - un score centré réduit $z_i = -0.8$ signifie que son score brut x_i s'est situé en dessous de la moyenne à une distance de 0.8 écart-type.

2. Cela permet aussi de **comparer des mesures** qui n'ont pas été réalisées sur la même échelle de mesure.

   Par exemple :
   - le salaire de monsieur X en France est de 2500 euros bruts mensuel
   - le salaire de monsieur Y au Japon est de 296000 yens bruts mensuel

   Qui de monsieur X et monsieur Y a le meilleur salaire par rapport à sa nationalité ?

• En France, le salaire moyen est de 2750 euros et l'écart-type de 900 euros.

Le salaire centré réduit de monsieur X est donc : $\frac{2500 - 2750}{900} = -0.278$.

• Au Japon, le salaire moyen est de 297000 yens et l'écart-type de 150000 yens.

Le salaire centré réduit de monsieur Y est donc : $\frac{296000 - 297000}{150000} = -0.007$.

$\Rightarrow$ Monsieur Y est donc mieux payé dans son pays que monsieur X dans le sien mais cela reste dans les 2 cas des salaires inférieurs à leur moyenne (la moyenne dans leur pays).

II - 2 nouveaux indices

À l'aide des données centrées réduites, on définit 2 nouveaux indices :

- **le skewness**

$$sk_x = \frac{1}{n}\sum_i^n z_i^3 = \frac{1}{n}\sum_i^n (\frac{x_i - \overline{x}}{\sigma_x})^3$$

  le skewness "mesure" la **symétrie** d'une distribution
  le skewness d'une distribution symétrique est proche de 0.

- **le kurtosis**

$$k_x = \frac{1}{n}\sum_i^n z_i^4 = \frac{1}{n}\sum_i^n (\frac{x_i - \overline{x}}{\sigma_x})^4$$

  le kurtosis "mesure" les **excès** d'une distribution
  le kurtosis d'une distribution symétrique *"Classique"* est proche de 3.

Exemple 1

▸ Nbre d'enfants dans une famille d'étudiants :

$$\bar{x} = 2.64 \ , \ \sigma_x^2 = 1.142 \text{ et } \sigma_x = 1.07$$

							$\sum$
v_k	1	2	3	4	5	6	
n_k	7	99	47	12	9	6	**180**
$(v_k - \bar{x})$	-1.64	-0.64	0.36	1.36	2.36	3.36	
$\frac{(v_k - \bar{x})}{\sigma_x}$	-1.53	-0.60	0.34	1.27	2.21	3.14	
$n_k(\frac{(v_k - \bar{x})}{\sigma_x})^3$	-25.2	-21.1	1.8	24.7	96.7	186	**263**
$n_k(\frac{(v_k - \bar{x})}{\sigma_x})^4$	38.5	12.6	0.6	31.4	213.4	584.2	**881**

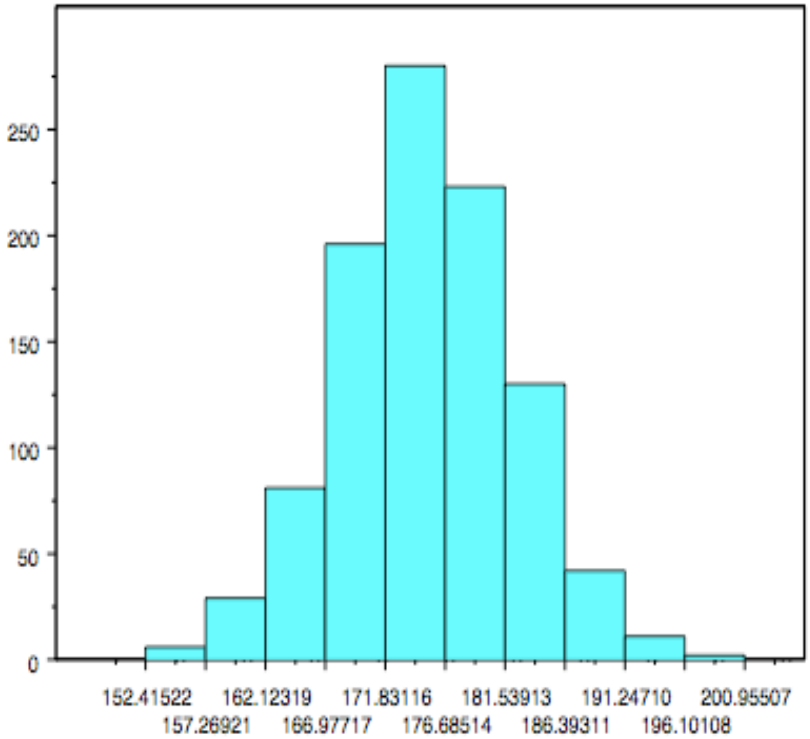
Ainsi :

$$\text{skewness} = \frac{263}{180} = 1.46 \qquad \text{kurtosis} = \frac{881}{180} = 4.89$$

Exemple 2

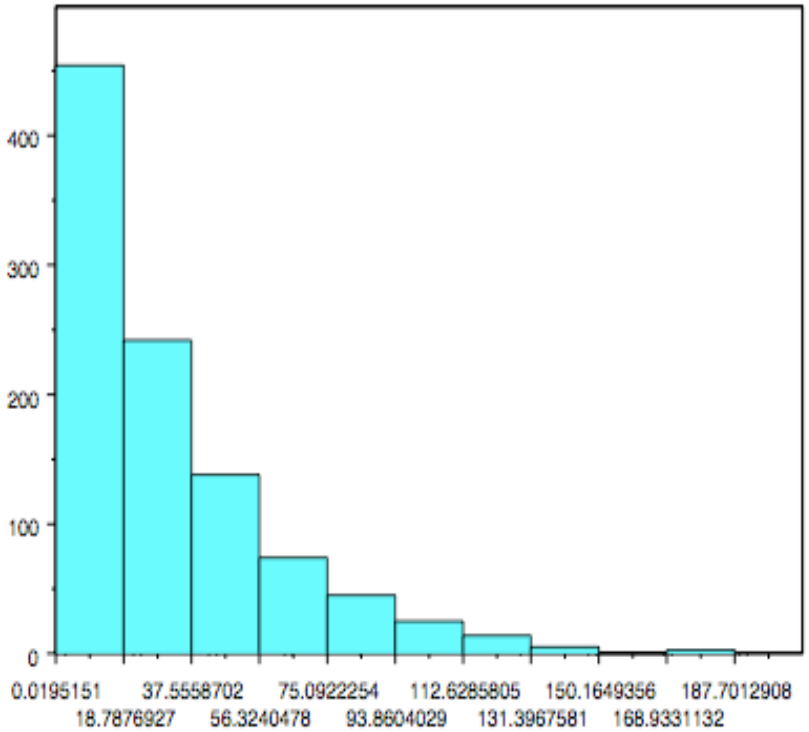
Deux variables continues (1000 observations)

la taille d'une population masculine



moyenne :	174.9
variance :	51.2
écart-type :	7.16
skewness :	0.047
kurtosis :	2.95

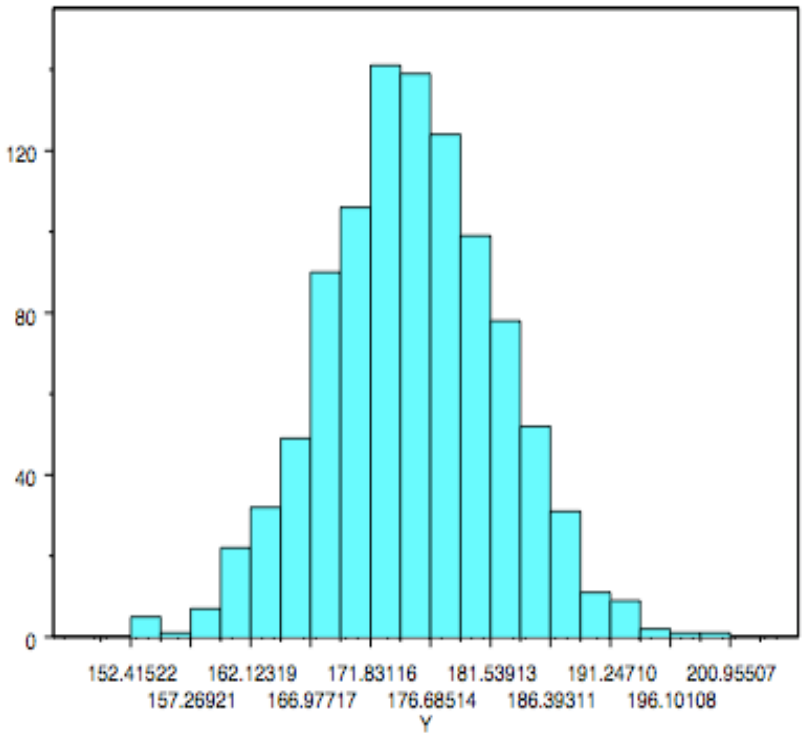
les précipitations pluvieuses

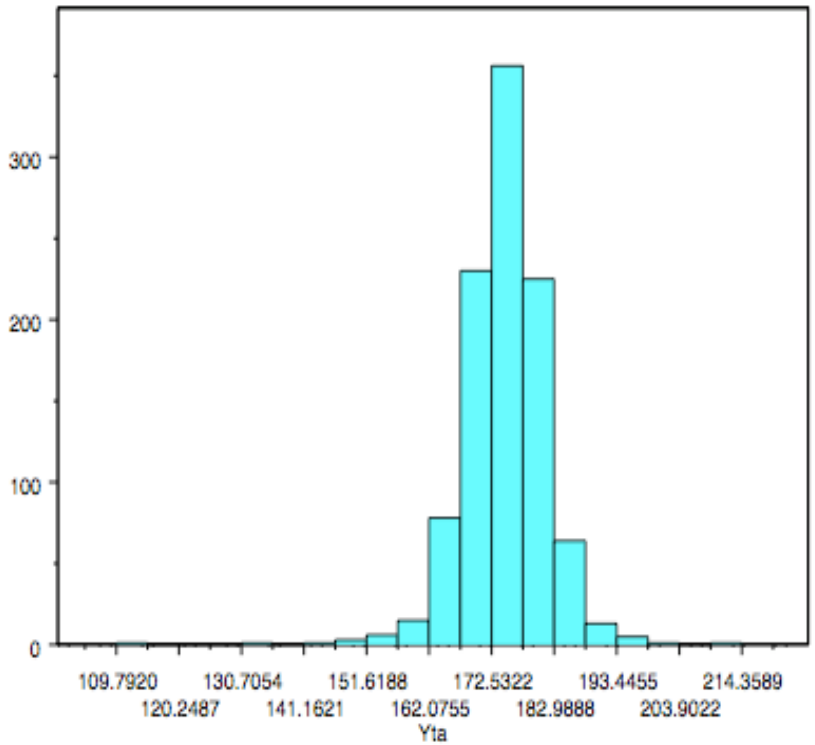


30.3
856.6
29.3
1.85
7.88

Exemple 3

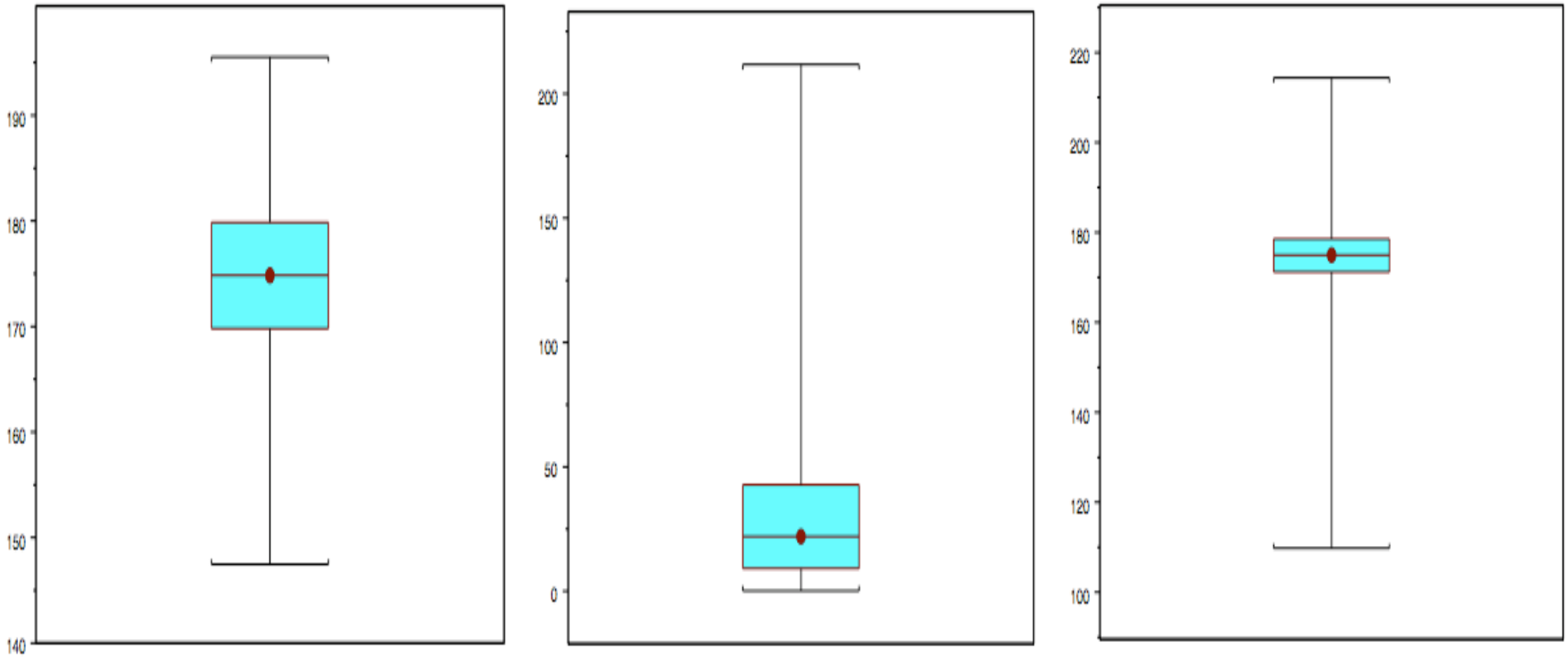
Histogrammes de 2 variables symétriques avec 20 classes (1000 observations)





moyenne :	174.9	174.7
variance :	51.2	47.5
écart-type :	7.16	6.89
skewness :	0.047	-0.94
kurtosis :	2.95	13.5

Box plot



Min :	147.46	0.091	109.79
1st Qu. :	169.82	9.109	171.20
Median :	174.85	21.97	174.98
3rd Qu. :	179.85	42.74	178.56
Max :	195.54	211.71	214.36

À SAVOIR

Données centrées réduites : à partir des valeurs x_i de l'échantillon, on construit l'échantillon des données centrées réduites z_i en calculant :

$$z_i = \frac{x_i - \overline{x}}{\sigma_x}$$

Ce nouvel échantillon des données centrées réduites vérifie :

$$\overline{z} = 0 \qquad \sigma_z = 1$$

On définit alors :

1. **skewness** :

$$sk_x = \frac{1}{n}\sum_i^n z_i^3 = \frac{1}{n}\sum_i^n (\frac{x_i - \overline{x}}{\sigma_x})^3$$

2. **kurtosis** :

$$k_x = \frac{1}{n}\sum_i^n z_i^4 = \frac{1}{n}\sum_i^n (\frac{x_i - \overline{x}}{\sigma_x})^4$$