4

Statistique à une dimension

A LA FIN DE CE CHAPITRE, IL FAUT ÊTRE CAPABLE DE...

1. Connaitre le vocabulaire de la statistique ;
2. Distinguer les distributions observées et continues ;
3. Représenter les séries statistiques en utilisant des graphiques adaptés aux distributions ;
4. Calculer les paramètres de centrage et de dispersion des séries statistiques.

4.1 Introduction

La statistique [1] est la science qui procède à l'étude méthodique à partir de modélisations mathématiques, des modes d'utilisation et de traitement de données, c'est-à-dire de l'information, dans le but de conduire et d'étayer une réflexion ou de prendre une décision en situation concrète soumise aux aléas de l'incertain.

La statistique descriptive étudie ces modes d'utilisation et de traitement de données, à un premier niveau, dans la perspective de produire essentiellement des descriptions des informations.

La statistique inférentielle les étudie à un second niveau dans la perspective d'étendre ces informations décrites à un domaine de validité non exploré directement, avec, si possible, un contrôle des risques encourus dans ce raisonnement inductif [2].

La démarche statistique comprend :

- la collecte des données ;
- le traitement des données collectées, aussi appelé la statistique descriptive ;
- l'interprétation des données, aussi appelée l'inférence statistique, qui s'appuie sur la théorie des sondages et la statistique mathématique ;
- la présentation afin de rendre les données compréhensibles par tous

Cette distinction ne consiste pas à définir plusieurs domaines étanches. En effet, le traitement et l'interprétation des données ne peuvent se faire que lorsque celles-ci ont été collectées. La statistique a des règles et des méthodes sur la collecte des données, pour que celles-ci puissent être correctement interprétées.

4.1.1 La démarche statistique

Recueil des données

L'enquête statistique est toujours précédée d'une phase où sont déterminés les différents caractères à étudier.

L'étape suivante consiste à choisir la population à étudier ou le modèle à utiliser. Il se pose alors le problème de l'échantillonnage : choix de la méthode d'échantillonage (au sens large : cela peut être un sondage d'opinion en interrogeant des humains, ou bien le ramassage de roches pour déterminer la nature d'un sol en géologie), la taille de l'échantillon et ses propriétés.

Que ce soit pour un recueil total (recensement) ou partiel (sondage), des protocoles sont à mettre en place pour éviter les erreurs de mesures qu'elles soient accidentelles ou répétitives (biais).

Le pré-traitement des données est extrêmement important ; en effet, une transformation des données initiales (un passage au logarithme, par exemple), peuvent considérablement faciliter les traitements statistiques suivants.

Traitement des données

Le résultat de l'enquête statistique est une série de nombres (tailles, salaires) ou de données qualitatives (langues parlées, marques préférées). Pour pouvoir les exploiter,

1. d'après la référence bilbliographique **??**
2. http :\\jean-claude.regnier.pagesperso-orange.fr\joao_claudio\statisti\statistique_def.htm

il va être nécessaire d'en faire un classement et un résumé visuel ou numérique. Il sera parfois nécessaire d'opérer une compression de données. C'est le travail de la statistique descriptive. Il sera différent selon que l'étude porte sur une seule variable ou sur plusieurs variables.

Étude d'une seule variable Le regroupement des données, le calcul des effectifs, la construction de graphiques permettent un premier résumé visuel du caractère statistique étudié. Dans le cas d'un caractère quantitatif continu, l'histogramme en est la représentation graphique la plus courante.

Les valeurs numériques d'un caractère statistique se répartissent dans $\mathbb{R}$, il est nécessaire de définir leurs positions. En statistique, on est en général en présence d'un grand nombre de valeurs. Or, si l'intégralité de ces valeurs forme l'information, il n'est pas aisé de manipuler plusieurs centaines voire milliers de nombres, ni d'en tirer des conclusions. Il faut donc calculer quelques valeurs qui vont permettre d'analyser les données : c'est le rôle des réductions statistiques. Celles-ci peuvent être extrêmement concises, réduites à un nombre : c'est le cas des valeurs centrales et des valeurs de dispersion. Certaines d'entre elles (comme la variance) sont élaborées pour permettre une exploitation plus théorique des données

On peut aussi chercher à comparer deux populations. On s'intéressera alors plus particulièrement à leurs critères de position, de dispersion, à leur boîte à moustaches ou à l'analyse de la variance.

Étude de plusieurs variables Les moyens informatiques permettent aujourd'hui d'étudier plusieurs variables simultanément. Le cas de deux variables va donner lieu à la création d'un nuage de points, d'une étude de corrélation (statistique) éventuelle entre les deux phénomènes ou étude d'une régression linéaire.

Mais on peut rencontrer des études sur plus de deux variables : c'est l'analyse multidimensionnelle dans laquelle on va trouver l'analyse en composantes principales, l'analyse en composantes indépendantes, la régression linéaire multiple et le data mining. Aujourd'hui, le data mining (appelé aussi knowledge discovery) s'appuie sur la statistique pour découvrir des relations entre les variables de très vastes bases de données. Les avancées technologiques (augmentation de la fréquence des capteurs disponibles, des moyens de stockage, et de la puissance de calcul) donnent au data mining un vrai intérêt.

Interprétation et analyse des données

L'inférence statistique a pour but de faire émerger des propriétés d'un ensemble de variables connues uniquement à travers quelques unes de leurs réalisations (qui constituent un échantillon de données).

Elle s'appuie sur les résultats de la statistique mathématique, qui applique des calculs mathématiques rigoureux concernant la théorie des probabilités et la théorie de l'information aux situations où on n'observe que quelques réalisations (expérimentations) du phénomène à étudier.

Sans la statistique mathématique, un calcul sur des données (comme une moyenne), n'est qu'un indicateur. C'est la statistique mathématique qui lui donne le statut d'estimateur dont on souhaite maîtriser le biais, l'incertitude et autres caractéristiques sta-

tistiques. On cherche en général à ce que l'estimateur soit sans biais, convergeant et efficace.

On peut aussi émettre des hypothèses sur la loi générant le phénomène général, par exemple « la taille des enfants de 10 ans en France suit-elle une loi gaussienne ? ». L'étude de l'échantillon va alors valider ou non cette hypothèse : c'est ce qu'on appelle les tests d'hypothèses. Les tests d'hypothèses permettent de quantifier la probabilité avec laquelle des variables (connues seulement à partir d'un échantillon) vérifient une propriété donnée.

Enfin, on peut chercher à modéliser un phénomène "*a posteriori*". La modélisation statistique doit être différenciée de la modélisation physique. Dans le second cas, des physiciens (c'est aussi vrai pour des chimistes, biologistes, ou tout autre scientifique), cherchent à construire un modèle explicatif d'un phénomène, qui est soutenu par une théorie plus générale décrivant comment les phénomènes ont lieu en exploitant le principe de causalité. Dans le cas de la modélisation statistique, le modèle va être construit à partir des données disponibles, sans nécessairement un "*a priori*" sur les mécanismes entrant en jeux. Ce type de modélisation s'appelle aussi modélisation empirique. Compléter une modélisation statistique par des équations physiques (souvent intégrées dans les pré traitements des données) est toujours positif.

Un modèle est avant tout un moyen de relier des variables à expliquer Y à des variables explicatives X, par une relation fonctionnelle :

$$Y = F(X)$$

4.1.2 Domaines d'application

La statistique est utilisée dans des domaines très variés comme :

- en géophysique, pour les prévisions météorologiques, la climatologie, la pollution, les études des rivières et des océans ;
- en démographie : le recensement permet de faire une photographie à un instant donné d'une population et permettra par la suite des sondages au moyen d'échantillons convenablement choisis ;
- en sciences économiques et sociales, et en économétrie : l'étude du comportement d'un groupe de population ou d'un secteur économique s'appuie sur des statistiques. C'est dans cette direction que travaille Eurostat. Les questions environnementales s'appuient également sur des données statistiques ;
- en sociologie : les sources statistiques constituent des matériaux d'enquête, et les méthodes statistiques sont utilisées comme techniques de traitement des données ;
- en marketing : le sondage d'opinion devient un outil pour la décision ou l'investissement ;
- dans les jeux de hasard et les paris tels que le Lotto ou les paris équestres, pour "prévoir" les résultats ;
- en physique : l'étude de la mécanique statistique et de la thermodynamique statistique, permet de déduire du comportement de particules individuelles un comportement global (passage du microscopique au macroscopique) ;
- en métrologie, pour tout ce qui concerne les systèmes de mesure et les mesures elles-mêmes ;
- en médecine et en psychologie, tant pour le comportement des maladies que leur fréquence ou la validité d'un traitement ou d'un dépistage ;

- en archéologie, appliquée aux vestiges (céramologie...) ;
- en écologie, pour l'étude des communautés végétales et des écosystèmes ;
- en assurance et en finance (calcul des risques,...).

4.2 Le signe de sommation

4.2.1 Définition

Supposons que nous ayons à manipuler algébriquement la somme des inverses des 100 premiers naturels non nuls. Écrire cette somme, terme après terme, demanderait un certain temps et couvrirait une partie non négligeable de la page. Ce n'est ni commode, ni efficace, ni recommandé.
Pour gagner du temps et réduire les écritures, cette somme peut être notée

$$1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{100}$$

les "petits points" remplaçant tous les termes allant de $\frac{1}{4}$ jusqu'à $\frac{1}{99}$. En pratique, cette manière de faire est souvent utilisée pour écrire des sommes comptant un grand nombre de termes. Mais, elle possède une faiblesse. Est-on sûr que les "petits points" seront compris de la même façon par tous ? Ainsi, l'écriture 2 + 4 + ... + 128 représente-t-elle la somme des naturels pairs de 2 à 128 ou la somme des puissances de 2 de $2 = 2^1$ à $128 = 2^7$?
Réduisant les écritures au maximum et évitant l'ambiguïté des "petits points", le signe de sommation $\sum$ offre le moyen idéal de noter ce type de somme.
Avec ce symbole, la somme des inverses des 100 premiers naturels non nuls s'écrit :

$$\sum_{k=1}^{100} \frac{1}{k}$$

se lit «somme des $\frac{1}{k}$ pour k allant de 1 jusqu'à 100 et signifie qu'il faut considérer la somme dont les termes s'obtiennent en remplaçant k successivement par 1, 2, 3, ... , 100 dans $\frac{1}{k}$. Dans une telle écriture, k est appelée variable de sommation. Quant à l'expression figurant "à l'intérieur" du signe de sommation, il est appelé terme général de la somme.
De la même manière, la somme des nombres pairs de 2 à 128 s'écrit

$$\sum_{k=1}^{64} 2k$$

et la somme des puissances de 2 allant de 2 à 128 se note

$$\sum_{k=1}^{7} 2^k$$

THÉORIE

4.2.2 Propriétés

Première propriété

La variable de sommation est muette. Cette propriété signifie que le nom de la variable de sommation peut être choisi arbitrairement.

$$\sum_{i=1}^{n} A(i) = \sum_{j=1}^{n} A(j) = \sum_{k=1}^{n} A(k)$$

En fait, la somme (résultat de l'addition des termes) est indépendante de la variable de sommation.

Deuxième propriété

La mise en évidence d'un facteur commun à tous les termes de la somme s'effectue en «sortant» ce facteur de la sommation, c'est-à-dire en le plaçant devant le signe de sommation.
Nous avons

$$\begin{aligned}\sum_{i=1}^{n} rA(i) &= rA(1) + rA(2) + \ldots + rA(n) \\ &= r\left(A(1) + A(2) + \ldots + A(n)\right) \\ &= r\sum_{i=1}^{n} A(i)\end{aligned}$$

Troisième propriété

Lorsque le terme général d'une somme est lui-même une somme, l'expression peut être scindée en deux comme le montre la formule suivante :

$$\sum_{i=1}^{n} \left(A(i) + B(i)\right) = \sum_{i=1}^{n} A(i) + \sum_{i=1}^{n} B(i)$$

Quatrième propriété

Une somme à terme général constant est égale au produit de ce terme par le nombre de termes :

$$\sum_{i=1}^{n} A = nA$$

4.3 Définitions et vocabulaire

Définition: *Une population U désigne un ensemble d'individus sur lequel porte une certaine analyse statistique.*
Un élément de cette population est un individu, généralement représenté par un indice $i = 1, ..., N$ où N est la taille de la population

Lorsque la population est restreinte, on peut étudier l'entièreté de la population.

Exemple:

Une classe dans une école, ...

Lorsque la population est trop étendue, on ne peut l'étudier dans son entièreté. On est invité à appliquer une technique de sondage.

Exemple:

La population d'un pays, les travailleurs d'une multinationale, les voitures produites annuellement par une usine, ...

Définition: *Une analyse statistique exhaustive étudie l'entièreté de la population U.*
Une échantillon s est un ensemble d'individus extrait d'une population importante et qui est "représentatif" de cette population dans le cadre de l'étude menée.

$$s \subset U$$

Remarquons que la réalisation des sondages est fort complexe, surtout dans le cadre de la définition d'un échantillon ***représentatif***.

Définition: *L'objet d'une étude statistique est un caractère X. Ce caractère peut prendre plusieurs valeurs appelées modalités. Elles sont en général notées :*

$$x_i \quad i \in U$$

Définition: *Un caractère peut être quantitatif lorsque les modalités sont mesurables et qualitatif dans le cas contraire.*
Un caractère quantitatif peut être discret s'il prend un ensemble fini de valeurs ou continu s'il peut prendre un nombre infini de valeurs.

THÉORIE

Exemples:

- Soit une classe d'une école où l'on s'intéresse aux notes sur 20 obtenues par les élèves à l'examen d'histoire.
  On pourra effectuer une étude exhaustive de cette population. Le caractère est quantitatif car ce sont des notes exprimées souvent en nombres naturels. En outre, ce caractère est discret (il prend des valeurs naturelles entre 0 et 20).
- Soit une école où l'on s'intéresse à la taille des élèves.
  Sauf si l'on est dans une école de village, on devra prélever une échantillon de cette population. Le caractère est quantitatif mais continu (il peut prendre toute une série de valeurs).
- Soit une école où l'on s'intéresse à la couleur des cheveux des élèves.
  Dans ce cas, le caractère est qualitatif car la désignation de la couleur ne s'exprime pas par un nombre.

4.4 Tableaux statistiques

4.4.1 Tableaux bruts

Définition: *Un tableau dans lequel les nombres ne sont pas disposés dans un ordre mathématique particulier, mais placés tels qu'ils ont été recueillis, est appelé tableau brut.*
Le nombre de données figurant dans le tableau brut (autrement dit, le nombre d'individus) est appelé effectif total et noté N.

Exemple:

Lors de la première épreuve d'un concours[a], les participants ont été soumis à un questionnaire à choix multiple (QCM) comportant 15 questions. Voici le nombre de bonnes réponses fournies par chacun des 59 candidats.

1	12	5	12	0	6	4	5	11	5	9	6
9	8	7	6	10	6	9	9	7	10	7	4
8	2	10	11	7	15	6	7	10	5	6	11
3	10	7	5	5	8	5	6	7	6	4	7
9	7	8	7	5	7	14	8	14	7	9	

Cette liste de nombres constitue un tableau statistique (ou une série statistique).

a. Tout ce qui suit est fortement inspirés de la référence **??**

4.4.2 Distributions observées

Le tableau brut est trop touffu ; il renferme trop d'informations. Tel quel, il ne donne pas une idée claire des résultats. Pour tirer parti des données, nous allons procéder à leur recensement. Nous noterons

$$x_1, x_2, x_3, ..., x_p$$

les nombres différents compris dans le tableau brut, classés par valeurs croissantes. Ces nombres constituent les ***modalités*** de la variable statistique étudiée, à savoir le résultat individuel au QCM.

Exemple:

Voici le distribution observée correspondant au QCM.

x_k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_k	1	1	1	1	3	8	8	12	5	6	5	3	2	0	2	1

Définition: *Le nombre d'apparition d'une modalité donnée d'un caractère est appelé un effectif*[a]*. Il est noté :*

$$n_i$$

si x_i est la valeur à laquelle il est associé. La somme de tous les effectifs est l'effectif global de la population. Il est noté :

$$N = \sum_{i=1}^{k} n_i$$

On appelle fréquence d'une modalité le rapport :

$$f_i = \frac{n_i}{N}$$

a. On supposera qu'il y a k modalités différentes dans la population étudiée.

Définition: *On appelle effectif et fréquence cumulée respectivement la somme des effectifs et des fréquences inférieurs ou égaux à la modalité concernée.*

$$N_k = \sum_{i=1}^{k} n_i$$

$$F_k = \sum_{i=1}^{n} f_i$$

Exemple:

Dans le cas de l'exemple QCM :

x_k	n_k	N_k	f_k	F_k
0	1	1	0,017	0,017
1	1	2	0,017	0,034
2	1	3	0,017	0,051
3	1	4	0,017	0,068
4	3	7	0,051	0,119
5	8	15	0,136	0,254
6	8	23	0,136	0,390
7	12	35	0,203	0,593

x_k	n_k	N_k	f_k	F_k
8	5	40	0,085	0,678
9	6	46	0,102	0,780
10	5	51	0,085	0,864
11	3	54	0,051	0,915
12	2	56	0,034	0,949
13	0	56	0	0,949
14	2	58	0,034	0,983
15	1	59	0,017	1,000

4.4.3 Tableaux groupés

Les modalités d'une variable statistique continue sont trop nombreuses pour être traitées séparément comme pour les variables discrètes étudiées aux paragraphes précédents. Pour une telle variable, il est préférable de procéder à un groupement des données, ce qui permet d'éviter des distributions observées très longues dans lesquelles les fréquences, trop petites, n'auraient plus vraiment de signification.

Exemple:

Pour donner un aperçu du procédé, considérons le nombre de kilomètres parcourus par jour de travail par un chauffeur de taxi pendant une année .

220	244	230	323	303	155	242	288	182	250	239	285
216	264	234	189	241	237	245	296	181	315	326	245
254	223	271	240	263	172	344	276	340	157	223	269
286	232	275	199	251	245	308	233	215	274	223	289
215	276	252	340	276	327	148	325	241	260	304	289
179	185	289	214	178	242	240	202	205	275	164	326
293	246	259	215	208	279	305	324	254	291	358	373
241	372	252	225	288	257	318	293	182	114	287	210
305	339	211	310	268	231	339	234	321	339	156	407
328	313	222	247	191	197	215	289	352	327	235	151
272	223	171	282	286	303	247	239	250	251	309	344
266	293	364	243	297	275	218	133	141	243	229	240
189	271	270	224	298	178	207	247	195	322	335	272
190	293	132	233	427	282	203	304	246	312	242	327
310	271	245	273	250	256	295	220	224	260	185	304
247	182	240	334	253	143	192	169	315	253	316	288
284	226	274	193	224	227	270	170	292	270	333	278
238	258	264	358	256	307	220	275	313	177	256	205
250	283	214	228	291	193	312	303	228	293	325	276
303	266	261									

Exemple:

Un recensement partiel des données produit le tableau suivant.

x_i	i	x_i	i	x_i	i	x_i	i	x_i	i	x_i	i	x_i	i	x_i	i
114	1	181	1	214	2	235	1	257	1	279	1	305	2	333	1
132	1	182	3	215	4	237	1	258	1	282	2	307	1	334	1
133	1	185	2	216	1	238	1	259	1	283	1	308	1	335	1
141	1	189	2	218	1	239	2	260	2	284	1	309	1	339	3
143	1	190	1	220	3	240	4	261	1	285	1	310	2	340	2
148	1	191	1	222	1	241	3	263	1	286	2	312	2	344	2
151	1	192	1	223	4	242	3	264	2	287	1	313	2	352	1
155	1	193	2	224	3	243	2	266	2	288	3	315	2	358	2
156	1	195	1	225	1	244	1	268	1	289	4	316	1	364	1
157	1	197	1	226	1	245	4	269	1	291	2	318	1	372	1
164	1	199	1	227	1	246	2	270	3	292	1	321	1	373	1
169	1	202	1	228	2	247	4	271	3	293	5	322	1	407	1
170	1	203	1	229	1	250	4	272	2	295	1	323	1	427	1
171	1	205	2	230	1	251	2	273	1	296	1	324	1		
172	1	207	1	231	1	252	2	274	2	297	1	325	2		
177	1	208	1	232	1	253	2	275	4	298	1	326	2		
178	2	210	1	233	2	254	2	276	4	303	4	327	3		
179	1	211	1	234	2	256	3	278	1	304	3	328	1		

Définition: *L'amplitude (ou étendue) d'un tableau statistique est la différence entre les valeurs extrêmes qu'il contient, donc entre la dernière et la première modalité, lorsqu'elles sont rangées dans l'ordre croissant (série ordonnée).*

Exemple:

Pour notre exemple, cette amplitude vaut 427 - 114 = 313.

Définitions: *Établir un groupement dans une série statistique, c'est répartir les données dans un certain nombre d'intervalles contigus, appelés classes.*
Le choix des classes s'effectue selon les critères suivants.

- *Les extrémités des intervalles doivent être des limites effectives de mesure.*
- *La réunion de toutes les classes doit recouvrir l'ensemble des données de la série statistique.*
- *Le nombre de classes est généralement être compris entre 5 et 20.*

Nous désignons par p le nombre de classes.

THÉORIE

Définitions: *Les extrémités des intervalles sont appelés limites de classes ; dans l'ordre croissant, nous les désignons par* $L_1, L_2, ...L_{p+1}$.
Avec cette notation, les classes sont $[L_1, L_2[, ..., [L_k, L_{k+1}[, ..., [L_p, L_{p+1}[$.
Chaque classe est caractérisée par son centre et sa largeur. Le centre de la $k^{ème}$ *classe est*

$$x_k = \frac{L_{k+1} + L_k}{2}$$

La largeur de la $k^{ème}$ *classe* $[L_k, L_{k+1}[$ *est :*

$$L_{k+1} - L_k$$

Tous les nombres appartenant à une même classe sont assimilés au centre de la classe. Dans les calculs, toutes les données appartenant à une même classe sont remplacées par le centre de cette classe. Ceci réduit fortement le volume de calculs et constitue un des principaux avantages du groupement des données par classes. L'effectif n_k *de la* $k^{ème}$ *classe est le nombre de données du tableau brut qui lui appartiennent. Sa fréquence* f_k *est égale au rapport de son effectif* n_k *à l'effectif total* N.
L'effectif cumulé N_k *et la fréquence cumulée* F_k *de la* $k^{ème}$ *classe se définissent de la même manière que pour les modalités d'un tableau de données.*

Remarque: Les grandeurs physiques (longueur, temps, poids, ...) varient de manière continue de sorte que les mesurer exactement est impossible. Toute mesure est entachée d'erreur, due notamment à l'imprécision inévitable de l'appareil de mesure. Ainsi, lorsque nous déclarons qu'un objet que nous venons de mesurer avec une latte a 52 mm de long, il est peu vraisemblable que sa longueur soit exactement égale à 52 mm. Le simple fait que la latte est graduée en millimètres entraîne une imprécision de l'ordre du millimètre. Il faut donc considérer que 52 mm n'est qu'une valeur approchée de la véritable longueur de l'objet et admettre que sa longueur réelle est comprise entre 51,5 et 52,5 mm. Dans le cas présent, ces nombres 51,5 et 52,5 sont des limites effectives de mesure.
Une limite effective de mesure se trouve généralement à mi-chemin entre deux graduations successives de l'appareil de mesure : en effet, entre deux graduations, notre oeil choisira toujours celle qui est la plus proche.

Exemple:

Dans notre exemple, si nous optons pour des classes de largeur constante 25 dont les centres sont eux-mêmes des multiples de 25, nous obtenons le tableau groupé suivant.

$[L_k, L_{k+1}[$	x_k	n_k	N_k	f_k	F_k
$[112,5; 137,5[$	125	3	3	0,013	0,013
$[137,5; 162,5[$	150	7	10	0,Q30	0,043
$[162,5; 187,5[$	175	15	25	0,065	0,108
$[187,5; 212,5[$	200	18	43	0,078	0,186
$[212,5; 237,5[$	225	34	77	0,147	0,333
$[237,5; 262,5[$	250	47	124	0,203	0,537
$[262,5; 287,5[$	275	36	160	0,156	0,693
$[287,5; 312,5[$	300	35	195	0,152	0,844
$[312,5; 337,5[$	325	21	216	0,091	0,935
$[337,5; 362,5[$	350	10	226	0,043	0,978
$[362,5; 387,5[$	375	3	229	0,013	0,991
$[387,5; 412,5[$	400	1	230	0,004	0,996
$[412,5; 437,5[$	425	1	231	0,004	1,000

4.5 Représentation graphique des séries statistiques

4.5.1 Diagrammes relatifs à une distribution observée

Diagramme en bâtonnets

Le plan étant muni d'un système d'axes perpendiculaires, portons-y les points de coordonnées (x_k, n_k) et relions chacun de ces points à l'axe des abscisses par un trait vertical. Nous obtenons le diagramme en bâtonnets des effectifs. Dans un tel diagramme, la longueur d'un "***bâtonnet***" représente donc l'effectif de la modalité correspondante.

Exemple:

Dans le cas des QCM :

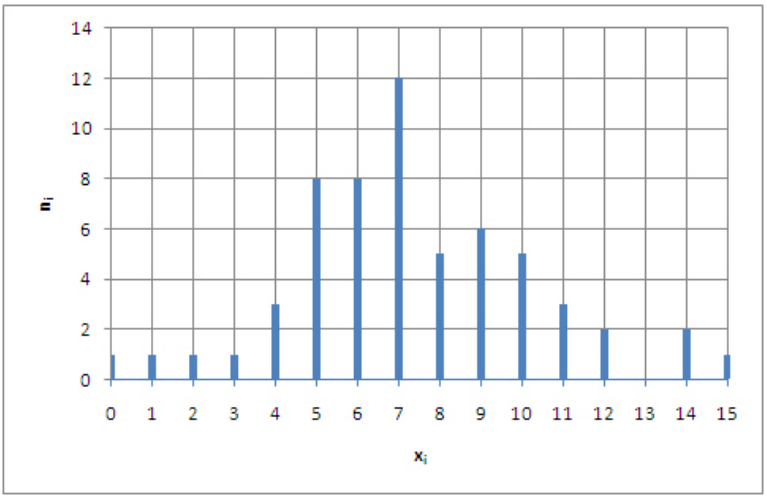
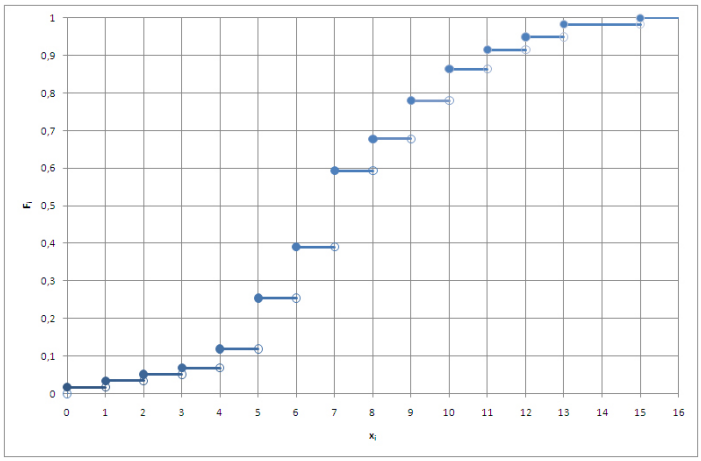


Diagramme des fréquences cumulées

Le diagramme cumulatif des fréquences est le graphique de la fonction en "escalier"
.

$$F : \mathbb{R} \to [0,1] : x \to \begin{cases} 0 \text{ si } x < x_1 \\ F_k \text{ si } x_k \leq x < x_{k+1} \\ 1 \text{ si } x \geq x_p \end{cases} \quad (\text{avec } 1 \leq k < p)$$



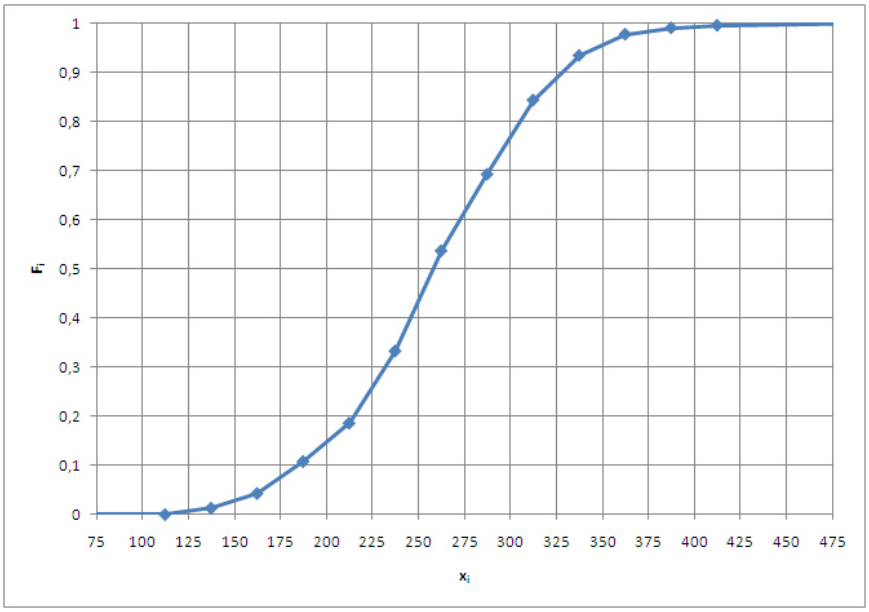
4.5.2 Diagrammes relatifs à un tableau groupé

Polygone des fréquences cumulées

Le diagramme des fréquences cumulées n'est pas adapté à la représentation des données d'un tableau groupé. En effet, dans ce diagramme, la fréquence cumulée reste évidemment constante entre deux modalités consécutives. Dans le cas d'un tableau groupé, il est raisonnable d'admettre que la fréquence cumulée croît de manière continue entre deux centres consécutifs de classes. Lors du groupement des données du tableau brut, toutes celles qui étaient comprises entre deux centres de classe consécutifs ont été réparties entre les classes correspondantes. En supposant que, dans une même classe, les données sont réparties uniformément, nous en déduisons que la fréquence cumulée croît linéairement à l'intérieur d'une classe.
Pour représenter les données d'un tableau groupé, nous construirons le polygone des fréquences cumlulées en plaçant les points de coordonnées

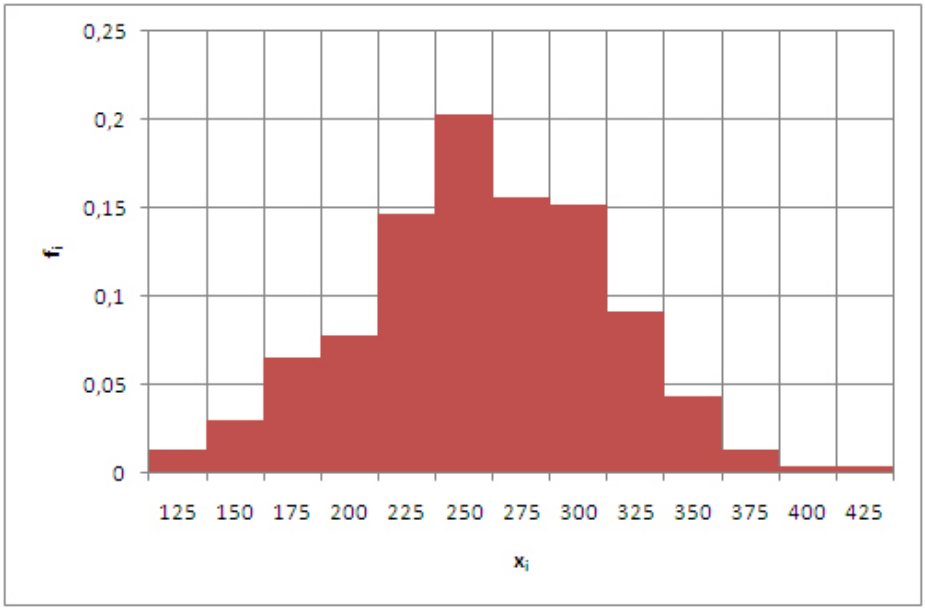
$$(L_1, 0), (L_2, F_1), ..., (L_{k+1}, F_k), ..., (L_{p+1}, 1)$$

dans le plan muni d'un système d'axes perpendiculaires et en reliant ces points de proche en proche par des segments de droite. Ces segments de droite traduisent ce que nous avons admis ci-dessus, à savoir que la fréquence cumulée croît linéairement à l'intérieur de chaque classe.



Histogramme

Dans la représentation des données du tableau groupé sous forme d'histogramme, chaque classe est symbolisée par un rectangle. La base de ce rectangle repose sur l'axe des abscisses ; sa largeur est égale à la largeur de la classe et la mesure de son aire est proportionnelle à l'effectif de la classe.

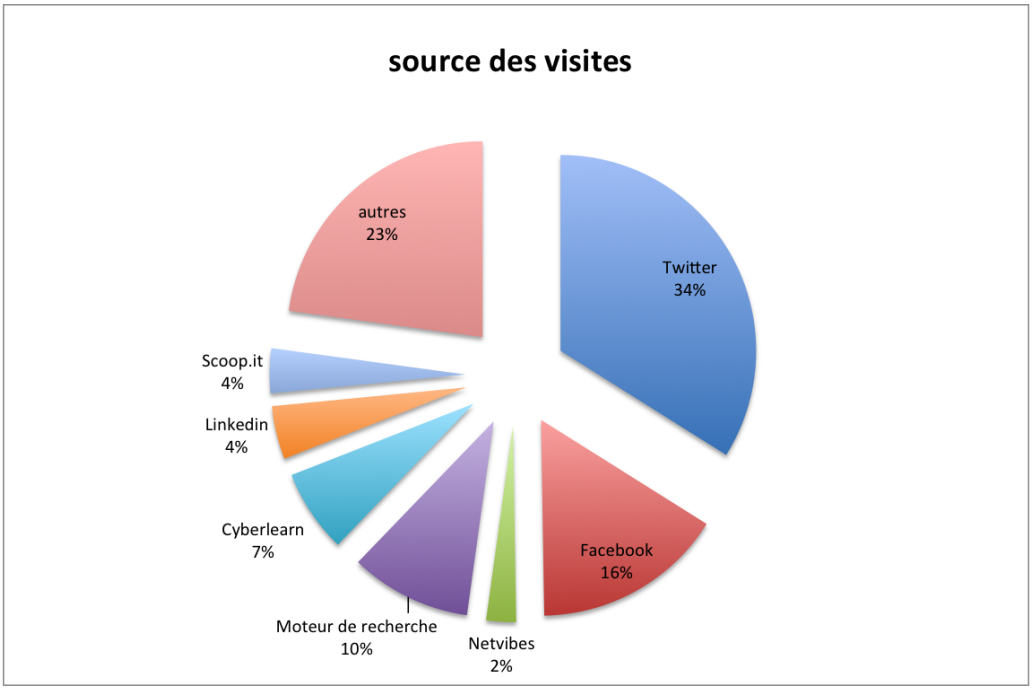


Dans un histogramme, la somme des mesures des aires des rectangles est égale à l'effectif total de la série statistique.

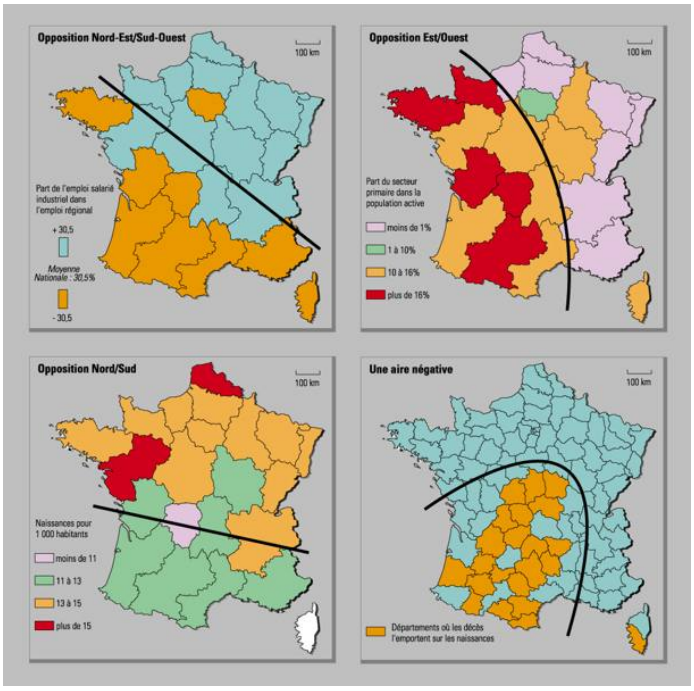
Autres représentations

Souvent pour illustrer les résultats d'une étude statistique, les médias, les magazines, les revues de vulgarisation scientifique, d'industries ou d'organes officiels, utilisent diverses formes de diagrammes : pictogrammes, tartes découpées en quartiers, histogrammes en colonnes ou en bandes, ... Leur caractère figuratif les rend très agréables

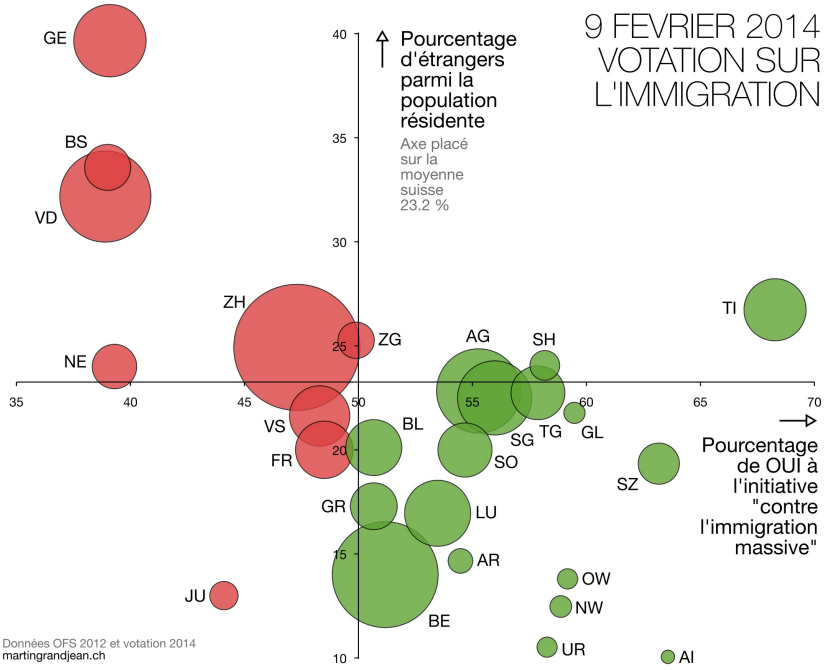
à consulter et leur confère une grande lisibilité. Cependant, en règle générale, ces diagrammes proviennent de l'analyse statistique d'une population divisée en classes selon des critères qualitatifs.
La qualité de telles représentations est, malheureusement, souvent de mauvaise qualité et difficilement interprétable...
Ci-dessous, quelques exemples de diagrammes trouvés sur des sites internet.



D'après https ://recherchemid.wordpress.com/2014/12/09/statistiques-du-blog-recherche-did



D'après http ://www.apex-cartographie.com/creation-livre-scolaire.asp



D'après http ://www.martingrandjean.ch/suisse-la-votation-sur-limmigration-en-un-graphique

4.6 Caractéristiques d'une série statistique :valeurs de centrage

4.6.1 Mode

Définition: *Dans une distribution observée, nous appelons mode, toute modalité dont l'effectif est maximal.*
Dans un tableau groupé, nous appelons classe modale, toute classe dont l'effectif est maximal ; le terme mode est alors attribué au centre d'une telle classe.

Une série statistique peut posséder plusieurs modes ; lorsqu'elle n'en possède qu'un seul, elle est dite unimodale.
Dans un diagramme en bâtonnets des effectifs ou des fréquences, le(s) mode(s) correspond(ent) au(x) point(s) d'ordonnée maximale du graphique.

Exemple:

Pour la variable statistique QCM (tableau 2), le mode est 7.
Pour la variable statistique «kilométrage-taxi», la classe modale est [237,5 ; 262,5 [et le mode 250.

THÉORIE

4.6.2 Médiane

Définition: *La médiane d'une série statistique ordonnée, notée $x_{1/2}$, est la «donnée» qui partage la série en deux parties de même effectif : les éléments de la première partie sont inférieurs à cette médiane et ceux de la seconde partie lui sont supérieurs.*

Méthode dans le cas d'une distribution observée

Si N est impair, alors $x_{1/2}$ est l'élément central de la série. Si N est pair, alors M est égal à la moyenne des deux éléments centraux de la série.

Exemples:

Pour la série 2,7,7,9,10,11,11,11,13, nous avons $x_{1/2}$= 10.
Pour la série 3, 3, 4, 8, 8, 9, 10, 10, 12, 13, 14, 16, nous obtenons $x_{1/2} = \frac{9+10}{2} = 9,5$
Pour la variable statistique QCM, la médiane est égale au 30ème nombre de la série ordonnée. Les effectifs cumulés du tableau 2 montre que celui-ci se trouve parmi les données de modalité 7. Nous avons donc $x_{1/2}$ = 7 (voir tableau du paragraphe 4.4.2).

Méthode dans le cas d'un tableau groupé

Si une des fréquences cumulées est égale à 0,5, alors $x_{1/2}$ est égale à l'extrémité de la classe correspondante (la limite supérieure de cette classe).
Si aucune des fréquences cumulées n'est égale à 0,5, alors il existe $k \in 1, 2, ..., p$ tel que

$$F_{k-1} < 0,5 < F_k$$

(en convenant ici que $F_0 = 0$). Dans ce cas, nous utiliserons une interpolation linéaire pour déterminer $x_{1/2}$.

Pour le calcul de la médiane par interpolation linéaire, considérons le polygone des fréquences cumulées (voir figure suivante) et supposons que

$$F_{k-1} < 0,5 < F_k$$

ce qui entraîne que nous devons rechercher la médiane dans la $k^{\text{ème}}$ classe du tableau groupé.

Les points $A(L_k, F_{k-1})$, $B(L_{k+1}, F_k)$ et $C(x_{1/2}; 0,5)$ étant alignés, les droites AB et AC ont le même coefficient angulaire. Nous avons donc

$$\frac{F_k - Fk - 1}{L_{k+1} - L_k} = \frac{F_k - 0,5}{L_{k+1} - x_{1/2}}$$

D'où, nous tirons

$$L_{k+1} - x_{1/2} = \frac{L_{k+1} - L_k}{F_k - Fk - 1}.(F_k - 0,5)$$

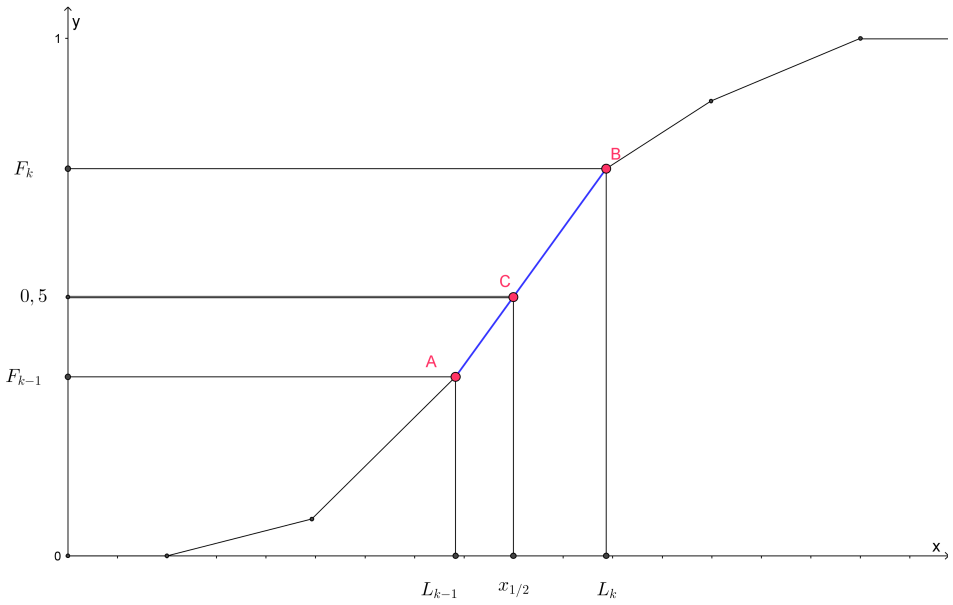


FIGURE 4.1 – Interpolation linéaire

ou

$$x_{1/2} = L_{k+1} - \frac{L_{k+1} - L_k}{F_k - Fk - 1}.(F_k - 0,5)$$

Cette formule ne doit pas être mémorisée ; pour calculer $x_{1/2}$, il suffit en pratique de reprendre le procédé suivi ci-dessus.

Exemples:

Pour la variable statistique "kilométrage-taxi", la fréquence cumulée 0,5 est comprise entre $F_5 = 0{,}333$ et $F_6 = 0{,}537$. Dès lors, $x_{1/2}$ est compris entre les limites de la classe [237,5 ; 262,5 [; nous avons donc (voir le tableau du paragraphe 4.4.3) :

$$x_{1/2} = 262,5 - \frac{262,5 - 237,5}{0,537 - 0,333}.(0,537 - 0,5) \approx 257,9$$

4.6.3 Moyenne arithmétique

Définition: *La moyenne d'une série statistique, notée $\bar{x}$, est la moyenne arithmétique de tous ses éléments.*
Dans le cas d'une distribution observée :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$$

Dans le cas d'un tableau groupé, les éléments d'une même classe sont remplacés par le centre de cette classe pour le calcul de la moyenne.

Exemple:

Pour la variable statistique QCM, la moyenne est $\bar{x} = 7,36$.
Pour la variable statistique "kilométrage-taxi", la moyenne est $\bar{x} = 258,55$

4.6.4 Généralisation : quartiles, déciles, centiles

La médiane divise une série statistique ordonnée en deux parties de même effectif. De manière analogue, les quartiles divisent la série statistique en quatre parties de même effectif, les déciles la divisent en dix parties de même effectif, les centiles en cent parties de même effectif.
Diviser une série en quatre, en dix ou en cent parties de même effectif n'est pas toujours possible, stricto sensu. Aussi, il va être nécessaire de définir de manière plus précise les quartiles, déciles et centiles ou, au moins, donner une méthode pour les calculer.

Cas d'une distribution observée

Pour une série statistique d'effectif N dont les données ne sont pas groupées, nous procédons selon le schéma suivant :

- Toutes les données, rangées par ordre croissant, sont numérotées de 0 jusqu'à $N-1$; le numéro ainsi attribué à une donnée est son rang.
- Le rang de chaque donnée est divisé par $N-1$. Ainsi. l'emplacement de chaque donnée est repéré par un nombre décimal allant de 0 à 1 par pas de $\frac{1}{N-1}$; ce nombre est dénommé rang décimal.
- Soit un réel α tel que $0 < \alpha < 1$.
  Si une donnée admet α pour rang décimal, alors cette donnée est appelée α-quantile et notée Q_α.
  Si α est compris entre deux données consécutives, Q_α est alors déterminé par une interpolation linéaire entre ces deux données. Ceci revient à utiliser la formule établie pour un tableau groupé en y remplaçant les limites supérieures de classes par les données qui encadrent α et les fréquences cumulées par les rangs décimaux de ces données.

Exemple:

Pour la variable statistique QCM comprenant 59 nombres, le premier quartile tombe entre les 15ème et 16ème données qui ont pour rangs décimaux $\frac{14}{58} \approx 0,2414$ et $\frac{15}{58} \approx 0,2586$. Ces données valant respectivement 5 et 6, nous obtenons $Q_{0,25} \approx 5,5$.
Quant au troisième quartile, il se situe entre les 44ème et 45ème données. On obtient $Q_{0,75} \approx 9$.

X_k	n_k	$n_k x_k$	$n_k x_k^2$
0	1	0	0
1	1	1	1
2	1	2	4
3	1	3	9
4	3	12	48
5	8	40	200
6	8	48	288
7	12	84	588
8	5	40	320
9	6	54	486
10	5	50	500
11	3	33	363
12	2	24	288
14	2	28	392
15	1	15	225
Σ	59	434	3712

Cas d'un tableau groupé

Définition: *Soit α un réel tel que $0 < \alpha < 1$. On appelle α-quantile d'une série statistique l'abscisse Q_α du point du polygone cumulatif des fréquences dont l'ordonnée est égale à α.*
En particulier, $Q_{0,5}$ est la médiane et le second quartile, $Q_{0,25}$ est le premier quartile et $Q_{0,75}$ le troisième quartile. Enfin, Q_α est le $k^{ème}$ décile lorsque $\alpha = \frac{k}{10}$ et le $k^{ème}$ centile lorsque $\alpha = \frac{k}{100}$.

Sauf lorsque $\alpha = 0,5$, les α-quantiles sont des paramètres de position.
Les calculs effectués pour déterminer la médiane par interpolation peuvent être recommencés pour trouver Q_α. Ils conduisent au résultat suivant (voir paragraphe 4.6.2)

$$Q_\alpha = L_{k+1} - \frac{L_{k+1} - L_k}{F_k - Fk - 1}.(F_k - \alpha)$$

où F_{k-1} et F_k sont les fréquences cumulées qui encadrent et L_k et L_{k+1} les limites supérieures des classes correspondantes.

THÉORIE

Exemple:

Pour la variable statistique "kilométrage-taxi", nous obtenons en appliquant la formule ci-dessus, $Q_{0,25} \approx 223,38$ et $Q_{0,75} \approx 296,94$.

$[L_k, L_{k+1}[$	x_k	n_k	N_k	f_k	F_k
$[112,5; 137,5[$	125	3	3	0,013	0,013
$[137,5; 162,5[$	150	7	10	0,Q30	0,043
$[162,5; 187,5[$	175	15	25	0,065	0,108
$[187,5; 212,5[$	200	18	43	0,078	0,186
$[212,5; 237,5[$	225	34	77	0,147	0,333
$[237,5; 262,5[$	250	47	124	0,203	0,537
$[262,5; 287,5[$	275	36	160	0,156	0,693
$[287,5; 312,5[$	300	35	195	0,152	0,844
$[312,5; 337,5[$	325	21	216	0,091	0,935
$[337,5; 362,5[$	350	10	226	0,043	0,978
$[362,5; 387,5[$	375	3	229	0,013	0,991
$[387,5; 412,5[$	400	1	230	0,004	0,996
$[412,5; 437,5[$	425	1	231	0,004	1,000

4.6.5 Écart interquartile

Définition: *L'écart interquartile d'une série statistique est la différence $Q_{0,75} - Q_{0,25}$ entre les premier et troisième quartiles. C'est donc la largeur de l'intervalle interquartile $[Q_{0,25}, Q_{0,75}]$ comprenant la moitié de l'effectif total de la série.*

Exemple:

L'écart interquartile de la variable statistique QCM vaut 9 - 5,5 = 3,5.
Celui de la variable statistique "kilométrage-taxi" est égal à 296,94 - 223,38 = 73,66.

4.6.6 Écart interdécile

Définition: *L'écart interdécile d'une série statistique est la différence $Q_{0,9} - Q_{0,1}$ entre les premier et neuvième déciles. C'est donc la largeur de l'intervalle interdécile $[Q_{0,1}, Q_{0,9}]$ comprenant 80% de l'effectif total de la série.*

Exemple:

Pour la variable statistique QCM, les premier et neuvième déciles sont respectivement $Q_{0,1} = 4$ et $Q_{0,9} = 1$ et l'écart interdécile est donc égal à 11 - 4 = 7.
Pour la variable statistique "kilométrage-taxi", nous avons $Q_{0,1}$= 184,42 et $Q_{0,9}$ = 327,88 ; l'écart interdécile vaut donc 327,88 - 184,42 = 143,46.

4.6.7 Étendue

Comme paramètre de dispersion d'une série statistique, il y a lieu de considérer aussi son amplitude (ou étendue) (voir paragraphe 4.4.3).

4.7 Caractéristiques d'une série statistique : mesure de dispersion

Médiane et moyenne, très utiles, permettent de préciser autour de quelle valeur sont dispersées les données d'une série statistique. Cependant, ils ne permettent pas de déterminer la manière dont les données sont réparties (dispersées) autour de ces valeurs.

Exemple:

À partir du tableau de la variable QCM, nous tirons le tableau suivant dans lequel nous indiquons les écarts de chaque modalité par rapport à la moyenne, ainsi que les valeurs absolues de ces écarts.

X_k	n_k	$e_k = x_k - \mu$	$\lvert e_k \rvert$
0	1	- 7,36	7,36
1	1	- 6,36	6,36
2	1	- 5,36	5,36
3	1	- 4,36	4,36
4	3	- 3,36	3,36
5	8	- 2,36	2,36
6	8	-1,36	1,36
7	12	- 0,36	0,36
8	5	0,64	0,64
9	6	1,64	1,64
10	5	2,64	2,64
11	3	3,64	3,64
12	2	4,64	4,64
14	2	6,64	6,64
15	1	7,64	7,64

La moyenne des écarts e_k n'est d'aucune utilité. En effet, il est facile de vérifier qu'elle est toujours nulle. Par conséquent, ce ne sont pas les écarts eux-mêmes qui doivent être pris en considération, mais leurs valeurs absolues. De toute évidence, pour

mesurer la dispersion, le signe des écarts importe moins que leur ordre de grandeur. La moyenne des valeurs absolues des écarts, appelé écart absolu moyen, vaut 2,29 dans le cas de la variable statistique QCM, montrant par là qu'en moyenne, les données s'écartent relativement peu de la moyenne.
Cependant, l'usage de valeurs absolues rendent difficiles les développements théoriques. De ce fait, l'écart absolu moyen n'est pas très utilisé dans la pratique. Pour gommer les valeurs absolues tout en gardant l'idée que le signe de l'écart ne joue aucun rôle dans la dispersion des données, nous allons introduire les carrés des écarts dans les définitions.

4.7.1 Variance

Définition: *La variance d'une série statistique, notée V, est la moyenne des carrés des écarts entre les données et la moyenne de cette série.*

$$V = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - x_{1/2})^2 = \sum_{i=1}^{k} f_i(x_i - x_{1/2})^2$$

Pour la facilité du calcul, transformons cette formule.

$$V = \sum_{i=1}^{k} f_i(x_i - x_{1/2})^2 = \sum_{i=1}^{k} f_i(x_i^2 - 2x_{1/2}x_i + x_{1/2}^2)$$

$$= \sum_{i=1}^{k} f_i x_i^2 - 2x_{1/2}\sum_{i=1}^{k} f_i x_i + x_{1/2}^2\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} f_i x_i^2 - 2x_{1/2}^2 + x_{1/2}^2$$

Nous obtenons donc une nouvelle expression de V :

$$V = \sum_{i=1}^{k} f_i x_i^2 - x_{1/2}^2$$

Exemple:

Pour un calcul manuel ou pour présenter les calculs dans un tableur, une disposition des calculs pratiques est présentée dans les tableaux suivants (correspondant aux variables QCM et "kilométrage-taxi").

X_k	n_k	$n_k x_k$	$n_k x_k^2$
0	1	0	0
1	1	1	1
2	1	2	4
3	1	3	9
4	3	12	48
5	8	40	200
6	8	48	288
7	12	84	588
8	5	40	320
9	6	54	486
10	5	50	500
11	3	33	363
12	2	24	288
14	2	28	392
15	1	15	225
Σ	59	434	3712

Exemple:

$[L_k, L_{k+1}[$	x_k	n_k	$n_k x_k$	$n_k x_k^2$
$[112,5;137,5[$	125	3	375	46875
$[137,5;162,5[$	150	7	1050	157500
$[162,5;187,5[$	175	15	2625	459375
$[187,5;212,5[$	200	18	3600	720000
$[212,5;237,5[$	225	34	7650	1721250
$[237,5;262,5[$	250	47	11750	2937500
$[262,5;287,5[$	275	36	9900	2722500
$[287,5;312,5[$	300	35	10500	3150000
$[312,5;337,5[$	325	21	6825	2218125
$[337,5;362,5[$	350	10	3500	1225000
$[362,5;387,5[$	375	3	1125	421875
$[387,5;412,5[$	400	1	400	160000
$[412,5;437,5[$	425	1	425	180625
Σ	231		59725	16120625

Le calcul de la variance de la variable statistique QCM à partir du tableau donne

$$V = \frac{3712}{59} - \left(\frac{434}{59}\right)^2 \approx 8,8055$$

Celui de la variance de la variable statistique "kilométrage-taxi" à l'aide du tableau donne

$$V = \frac{16120625}{231} - \left(\frac{59725}{231}\right)^2 \approx 2938,2648$$

4.7.2 Ecart-type

La variance n'a pas la même "unité" que les écarts par rapport à la moyenne puisque sa définition fait intervenir les carrés de ces écarts.

Définition: *L'écart-type d'une série statistique, notée σ, est la racine carrée de la variance de cette série.*

$$\sigma = \sqrt{V}$$

Exemple:

Pour la variable statistique QCM, nous avons $\sigma = 2,97$.
Pour la variable statistique "kilométrage-taxi", nous trouvons $\sigma = 54,21$.

4.7.3 Diagramme en boîte ou boîte à moustaches

Représenter la dispersion des données exploitant l'amplitude, les quartiles et déciles peut être réalisé au moyen d'un diagramme en boîte, aussi appelé de façon imagée «boîte à moustaches". Ce diagramme consiste en une boîte rectangulaire matérialisant l'intervalle interquartile. Celle-ci est prolongée par des segments (les «moustaches») dont les extrémités symbolisent les premier et neuvième déciles. La médiane est marquée par un trait dans la boîte et les valeurs extrêmes de la variable statistique par des points.

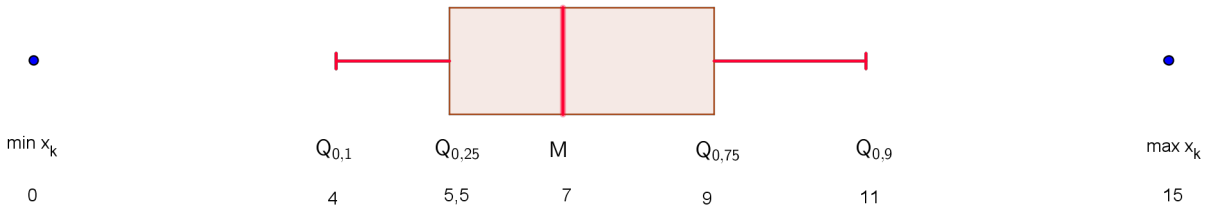


FIGURE 4.2 – Boite à moustache de la variable QCM

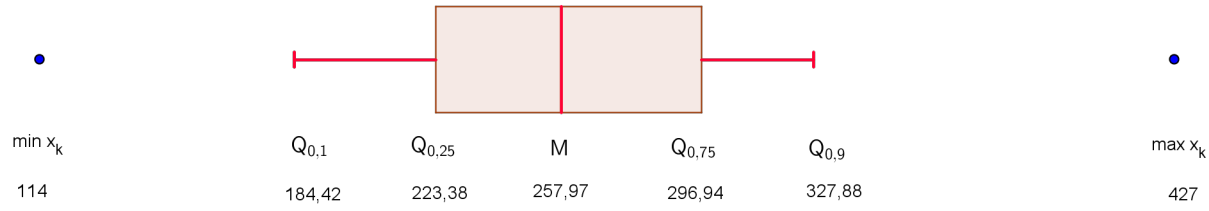


FIGURE 4.3 – Boite à moustache de la variable "kilométrage-taxi"

4.8 Exercices

Pour chacune des séries statistiques suivantes :

1. Analyser les données de manière graphiques (au moins histogrammes des effectifs, effectifs cumulés, fréquences et fréquences cumulées). D'autres graphes peuvent être présentés. Les graphiques seront commentés.
2. Caractériser entièrement les séries par des chiffres représentatifs qui seront analysés et commentés.
3. Tous les résultats devront être justifiés soit par un calcul soit par une phrase. Les statistiques ne tombent pas du ciel...

1. On a relevé les puissances en CV fiscaux des véhicules d'une société de location de voitures. Elles sont reprises dans le tableau suivant :

9	7	8	5	4	5	7	7	6	6	5	5
8	5	6	6	5	3	5	6	8	6	6	4
4	5	6	4	6	7	4	4	7	5	6	6
8	2	2	3	5	4	5	5	3	7	5	4
8	5	6	5	4	7	5	6	7	5	6	4
7	5	5	5	5	6	5	5	6	6	5	6

2. On a relevé les cotes d'une interrogation surprise (sur 10) d'un cours. Elles sont reprises dans le tableau suivant :

3	7	4	8	5	7	7	5	8	5	7	3	9
6	5	6	5	7	3	6	4	8	4	10	5	9
5	4	6	3	6	6	5	3	4	6	7	4	7
7	5	5	7	4	7	5	7	6	6	5	6	6

3. On a relevé le nombre d'enfants de familles d'un village. Ils sont repris dans le tableau suivant :

1	3	0	0	1	2	1	1	1	4
2	0	1	8	4	1	3	3	7	3
3	2	5	1	3	3	4	6	1	2
2	3	5	0	4	3	1	5	0	7

4. Une société immobilière dispose de 600 appartements dont les surfaces sont données par le tableau suivant :

Surface (en m^2)	% des appartements
$[20, 50[$	2
$[50, 60[$	15
$[60, 80[$	13
$[80, 100[$	22
$[100, 120[$	28
$[120, 145[$	20

5. Voici le nombre de minutes de connexion Internet d'un échantillon d'abonnés d'une compagnie spécialisée dans ce type de service :

Nombre de minutes de connexion	Nombre d'abonnés
$[0, 60[$	8
$[60, 90[$	20
$[90, 120[$	0
$[120, 150[$	60
$[150, 180[$	100
$[180, 210[$	12

6. Des analystes en finances viennent de composer un nouveau portefeuille REER pour les clients de la banque BANKO. Afin de vérifier son impact sur les clients, on prélève au hasard, 50 dépôts dans les comptes REER. Les montants déposés sont (en centaines d'€)

59	101	76	86	99	87	99	101	77	58
87	66	82	59	77	81	89	114	97	87
99	77	77	99	79	88	107	86	89	76
71	80	83	79	100	98	77	84	81	89
85	86	85	83	85	67	83	88	75	100

7. Des investisseurs analysent un nouveau type d'investissement : fabrication d'un nouveau type d'automobile électrique. Voici le nombre de ventes chez un concessionnaire spécialisé durant 25 jours choisis au hasard le printemps dernier :

186	183	196	194	193	193	189	199	200	192
190	186	194	191	187	188	197	195	196	190
180	188	186	198	199					

8. On a mesuré la taille des joueurs d'un club de basket. Elles sont reprises dans le tableau suivant :

175	201	195	185	203	185	188	185	192
176	185	180	181	190	182	185	172	197
193	189	177	186	190	182	185	172	197
181	191	175	170	204	187	180	182	188
191	198	190	175	192	186	185	178	169

9. On a relevé les montants des chèques émis pendant une journée dans un magasin (en €). Ils sont repris dans le tableau suivant :

127.00	390.00	410.00	540.50	190.00	280.00	210.00	425.00
350.00	742.00	176.00	138.50	120.00	355.20	472.00	140.00
170.20	150.30	170.00	595.00	792.50	688.20	100.50	672.00
240.00	200.10	185.00	205.00	175.00	182.50	150.10	180.10

10. On a relevé l'âge des employés d'une société. Ils sont repris dans le tableau suivant :

31	25	40	19	35	24	60	24	25	48	35	40	47	29	60
38	29	61	20	40	44	35	25	40	35	41	26	24	21	42
30	33	35	18	36	25	31	34	33	55	23	31	33	34	41
55	37	25	57	32	30	41	20	57	31	49	37	33	53	60
33	22	40	49	21	38	58	30	40	50	20	40	18	34	39
31	60	31	32	49	36	26	50	33	56	27	58	50	38	50

11. On a mesuré la distance parcourue chaque semaine entre le domicile et l'école par des élèves. Elles sont reprises dans le tableau suivant :

2	5	1	8	12	25	7
14	7	5	4	3	5	11
20	23	21	15	10	7	4
2	5	7	6	14	17	15
13	11	8	10	11	20	19