République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mouloud Mammeri de Tizi Ouzou
Faculté des Sciences
Département de Mathématiques

Cours de Statistique descriptive

Préparé par Mr Yousfi smail

Table des matières

Introduction

La ” Statistique est apparue pour la premier fois au royaume Unis grâce aux travaux de Karl Pearson, le fondateur du premier département de statistique au monde, et aux travaux de Ronald Fisher, un pionnier dans le domaine des plans d'expérience. Cette nouvelle discipline englobe l'application de la statistique à la biologie, à la médecine, aux sciences de la nature et de l'histoire naturel des êtres vivants. Par ailleurs, les ” Statistiques ” sont la science des données, qui est à la fois une science, une méthode et un ensemble de techniques. Ceci implique la collection, la classification, le résumé, l'organisation, l'analyse et l'interprétation d'une information numérique. De ce fait, la science de la Biostatistique englobe à la fois : la conception des expériences biologiques ; la collecte des informations ; la compilation et analyse des données chiffrées de ces expériences et l'interprétation des résultats en vue d'avancer une conclusion.

La Biostatistique est l'une des matières fondamentales intégrées dans le programme de tous les étudiants de Biologie de l'Université de Tizi Ouzou. Le but recherché par ces programmes d'enseignement est l'initiation de l'étudiant aux traitements de données liées aux thématiques biologiques.

Le polycopié est equivalent à 6 heures de cours et consacré à la définition de la statistique dans sa composante descriptive, et cela en expliquant la signification des différents concepts et notions adoptées dans la démarche statistique, suivi d'un détail sur la nature des caractères étudiées en statistique et les différentes techniques d'organisation et de structuration d'une série de données. Ce polycopié expose également les différents paramètres descriptifs d'une série statistique simple à savoir les paramètres de position et les paramètres de dispersion.

Chapitre 1

Statistique desriptive univariée

1.1 Introduction

L'origine du mot "*Statistique*" est "*status*" qui signifie état qui, a l'origine, cette discipline concernait exclusivement les affaires de l'État, en général par des études méthodiques des faits sociaux définissant cet État, par des procédés numériques (dénombrements, inventaires, recensements,...).

La statistique est un ensemble de techniques d'interprétation mathématique appliquées à des phénomènes pour lesquels une étude exhaustive de tous les facteurs est impossible à cause de leur grand nombre ou de leur complexité. Pour certain elle est une branche des mathématiques(les anglo-saxons) pour d'autre une discipline à part entière hors des mathématiques.

1.2 La statistique descriptive

La statistique descriptive est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données. Cette démarche a pour but de :

- Résumer et synthétiser l'information contenue dans la série statistique ;
- Mettre en évidence ses propriétés ;
- Suggérer des hypothèses relatives à la population dont est issu l'échantillon.

Les Outils utilisés sont :

- Les Tableaux ;
- Les Graphiques ;
- Les indicateurs.

Le type d'outils utilisés dépend de :

- La nature de la série (uni ou multidimensionnelle) ;
- La nature des variables (quantitatives ou qualitatives).

Si les données ne sont relatives qu'à une seule variable, on parle de statistique descriptive *" univariée"*. Dans le cas où l'on s'intéresse à deux variables simultanément, on met en œuvre la statistique descriptive *" bivariée "*. Si l'ensemble de données provient de l'observation de plusieurs variables, on doit faire appel aux méthodes de la statistique descriptive *" multivariée "*. Ce document est consacré à la présentation des outils et méthodes de la statistique descriptive univariée.

1.3 Notions de base de la statistique descriptive

1.3.1 Élément ou unité statistique :

Qui peut être

- Un individu (êtres vivants) : humain, animal, végétal ... ;
- Un sujet : modules enseignés en biologie, les nationalités, les métiers ou professions... ;
- Un objet : Table, chaise, verrerie de laboratoire ;

- Une association (dans les études écologiques en général) : une parcelle d'herbe, une association d'arbustes...

1.3.2 Population :

C'est un ensemble d'éléments possédant au moins une caractéristique commune et exclusive permettant de l'identifier et la distinguer sans ambiguïté de toutes les autres.

Exemple 1.3.1. :

- Une population algérienne ;
- Une population estudiantine ;
- Une population de plantes médicinales ;
- Une population de poissons d'eau douce.

1.3.3 Échantillon :

Pour des raisons techniques ou économiques, il n'est généralement pas possible de collecter des données sur tous les éléments de la population. En outre, si cette opération est possible il est rarement utile de la faire, car l'analyse d'un groupe restreint d'éléments extraits de la population fournit généralement des résultats de précision satisfaisante. Cette petite partie de la population qu'on va examiner s'appelle " échantillon ".

Exemple 1.3.2. :

- Etude de 20 étudiants pris à partir d'une population de 57.
- Etude de 5 régions prises à partir d'une population de 25.
- Etude de 5 modules pris à partir d'une population de 13.
- Etude de 200 patients pris à partir d'une population de 660.

1.3.4 Échantillonnage :

C'est l'opération ou la méthode qui consiste à prélever une partie de la population (échantillon). Cette méthode doit assurer la représentativité de cette population c'est-à-dire qu'elle doit refléter fidèlement sa composition et sa structure. Il existe plusieurs méthodes d'échantillonnage qui varient en fonction de la nature de l'étude envisagée. On peut citer l'échantillonnage : Stratifié, par degré, systématique. Le plus utilisé est l'échantillonnage aléatoire et simple qui est basé sur le principe que tous les éléments de la population ont une probabilité égale (non nulle) de faire partie de l'échantillon : c'est une méthode

1.3.5 Caractère (variable ou mesure) :

C'est une propriété possédée par les unités statistiques permettant de les décrire et de les distinguer les unes des autres. Toute unité statistique peut être étudiée selon un ou plusieurs caractères :

Exemple 1.3.3. :

- Couleur des yeux ;
- Poids des souris ;
- Superficie d'une pièce ;
- La température de l'air.

1.3.6 Modalité :

Ce sont les diverses situations (cas, état, valeur) susceptibles d'être prises par le caractère. Un caractère peut posséder une ou plusieurs modalités.

Exemple 1.3.4. :

- Couleur des yeux : vert, bleu, noir.
- Poids des souris (en grammes) : 15, 18, 20, 39.
- Superficie d'une pièce (en m^2) : 3, 5, 6.
- La température de l'air (en °C) : 8, 16, 27, 30, 38.

1.3.7 Effectifs :

On distinguent :

- L'effectif de la valeur x_i d'un caractère est le nombre d'individus de la population ayant cette valeur, elle est notée n_i.
- L'effectif total n est la somme de tous les effectifs. On écrit alors :

$$n = n_1 + n_2 + n_3 + \dots$$

— En rangeant les valeurs du caractère dans l'ordre croissant, on peut calculer l'effectif cumulé croissant en faisant la somme des effectifs de cette valeur et de tous ceux qui la précèdent.

$$N_i = n_1 + n_2 + \dots + n_i$$

Exemple 1.3.5. Dans une promotion de 20 étudiants en Biologie, voici les notes obtenues au dernier examen de Biostatistique : 10, 14, 12, 15, 7, 8, 10, 11, 12, 18, 2, 4, 12, 13, 14, 15, 19, 11, 9, 0. On va calculer les effectifs et les effectifs cumulés. Pour les effectifs on cherche le nombre d'étudiants ayant la note x_i. Pour les effectifs cumulés on fait la somme des effectifs de la note + la somme des effectifs de toutes les notes qui la précédent. On obtient alors le tableau suivant :

Notes x_i	0	2	4	7	8	9	10	11	12	13	14	15	18	19
Effectifs n_i	1	1	1	1	1	1	2	2	3	1	2	2	1	1
Effectifs cumulés N_i	1	2	3	4	5	6	8	10	13	14	16	18	19	20

Remarque 1.3.1. Afin de s'assurer que le calcul des effectifs cumulés est bien correcte, la dernière valeur de l'effectif cumulée doit correspondre aux nombre total d'individus, dans cet exemple égal à 20.

1.3.8 Fréquence d'une modalité ou d'une classe

La fréquence d'une valeur x_i du caractère notée f_i est le quotient de l'effectif de la valeur par l'effectif total c 'est-à-dire :

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

En rangeant les valeurs du caractère dans l'ordre croissant, on peut calculer les fréquences cumulées croissantes (notée F_i) en faisant la somme des fréquences de cette valeur et de tous ceux qui la précèdent,

$$\boxed{F_i = f_1 + f_2 \dots + f_i.}$$

Exemple 1.3.6. Dans l'exemple (1.3.5) on calcul les effectifs et les effectifs cumulés, les résultats sont résumés dans le tableau suivant.

Notes x_i	0	2	4	7	8	9	10	11	12	13	14	15	18	19
fréquences f_i	0,05	0,05	0,05	0,05	0,05	0,05	0,1	0,1	0,15	0,05	0,1	0,1	0,05	0 ,05
fréquences cumulés F_i	0,05	0,1	0,15	0,2	0,25	0,3	0,4	0,5	0,65	0,7	0,8	0,9	0,95	1

1.3.9 Nature des caractères :

On distingue deux grandes familles de caractères ; les caractères qualitatifs et les caractères quantitatifs.

Caractère qualitatif :

un caractère est dit qualitatif lorsque ses modalités ne sont pas mesurables. Le nombre de valeurs que peut prendre la variable est limité. Il existe au sein de ce type deux échelles : nominale et ordinale.

Échelle nominale chaque modalité est exprimée par un nom ou un cod e

Exemple 1.3.7. (Cas des noms)
- Nationalité : Algérienne, Tunisienne ;
- Les différentes séquences nucléotidiques (ADN ou ARN) ;
- Les hormones : œstradiol, progestérone
- État matrimoniale : marié, célibataire, veuf, divorcé ;
- Sexe : féminin, masculin ;
- Profession : enseignant, médecin ;

Exemple 1.3.8. (Cas des codes)
- État matrimoniale : marié (1), célibataire (2), veuf (3), divorcé (4) ;
- Sexe : féminin (1), masculin (2) ;
- Profession : enseignant (1), médecin (2) ;
- Nationalité : Algérienne (1), Tunisienne (2)

Échelle ordinale Chaque modalité est explicitement significative du rang pris par chaque individu pour le caractère considéré.

Exemple 1.3.9.
- Degré d'intelligence : pas intelligent (0), peu intelligent (1), moyennement intelligent(2), très intelligent (3) ;
- Forme des fruits : petite (1), moyenne (2), grosse (3) ;
- Abondance/Dominance : peu abondant (1), abondant (2), très abondant (3).

Caractère quantitatif :

un caractère est dit quantitatif si ses modalités s'expriment par des nombres dont les opérations de types sommes et produits sont possibles sur les valeurs des modalités. Le nombre de valeurs que peut prendre la variable est illimité. On distingues deux catégories de caractère quantitative :

Les caractères quantitatifs discrets sont des caractères dont les modalités sont des nombres isolés, pas nécessairement entiers.

Exemple 1.3.10.
- Nombre de pièces d'un immeuble ;
- Nombre d'enfants d'une famille ;
- Nombre de personnes atteintes par une maladie.

Les caractères quantitatifs continus sont des caractères dont les modalités sont définies sur un intervalle (continu) de valeur donné appelé domaine de variation et défini par les valeurs minimales et maximales. Ses valeurs sont regroupés dans des classes (petit intervalles) qu'on note $e_i =]\min(e_i),\ \max(e_i)]$

Exemple 1.3.11. Notes des étudiants, la taille, le poids, l'âge.

Nombre de classes : Dans le cas continu, les résultats sont regroupés en classes à cause de leur grande masse. Discrétiser une variable quantitative c'est, mathématiquement, transformer un vecteur de nombres réels en un vecteur de nombres entiers nommés " indices de classe ". En statistiques, discrétiser c'est à la fois réaliser cette transformation mathématique, nommer et justifier les classes. Un bon découpage correspond à des classes homogènes et séparées, ce qui correspond respectivement aux notions statistiques de faible variance intra-classe et de forte variance interclasse. Mais d'autres critères sont possibles, comme l'équirépartition, le respect d'un nombre minimal de données par classe, etc. Voici ci-après (tableau suivant) le nombre de classe en fonction du nombre d'élément.

Nombre d'éléments	Nombre de classes
10	2 à 3
20	3 à 4
50	4 à 6
100	5 à 8

TABLE 1.1: Nombre de classes en fonction du nombre d'éléments d'une série statistique donnée.

Règle de STURGE :

Soit un échantillon de N valeurs observées. Le mathématicien Herbert Sturges (1882-1958) a proposé une valeur approximative pour le nombre de classe k en fonction de la taille N de l'échantillon :

$$k = 1 + \log_2 N = 1 + \log_2 N$$

où $\log_2$ est le logarithme en base 2. Le résultat ne sera pas, en général, entier. Il donne une appréciation de ce qui ferait un bon découpage.

Amplitude d'une classes : l'amplitude d'une classe e_i est la longueur de l'intervalle définissant cette classe (noté a_i), c 'est-à-dire :

$$a_i = |e_i| = \max(e_i) - \min(e_i)$$

Centre d'une classes : le centre d'une classe e_i est égale à :

$$c(e_i) = \frac{\max(e_i) + \min(e_i)}{2}$$

Étendue d'une série statistique Dans le cas continu, on s'interesse en particulier à l'étendue des valeurs prise par la variable X qui vaut

$$E(X) = \max\{x_1, x_2, \ldots, x_n\} - \min\{x_1, x_2, \ldots, x_n\}$$

1.3.10 Tableaux statistiques

Un tableau statistique constitue un résumé ou une synthèse numérique des résultats d'une distribution statistique, on distingue trois formes de tableaux statistiques qui sont fonction de l'objectif envisagé et de la nature du caractère étudié.

Tableau brut

Après la collecte des données, celles-ci apparaissent de façon brute. Sous cette forme, elles sont peu informatives. Il nous faut donc des moyens pour en extraire un maximum d'informations.

Tableau de dénombrement

Il est de la forme suivante :

Element i	modalité x_i
1	x_1
2	x_2
.	.
.	.
.	.

TABLE 1.2: :
tableu brut

Modalités x_i	Effectifs n_i	Fréquences relatives f_i
x_1	n_1	f_1
x_2	n_2	f_2
x_3	n_3	f_3
.	.	.
.	.	.
.	.	.
x_k	n_k	f_k
Total	N	1

TABLE 1.3: Tableau de dénombrement

tableau de distribution des fréquences

il est de la forme

Classes	Centre c_i	Effectifs n_i	Fréquences relatives f_i
$e_1 = [a_0,\ a_1[$	c_1	n_1	f_1
$e_2 = [a_1,\ a_2[$	c_2	n_2	f_2
$e_3 = [a_2,\ a_3[$	c_3	n_3	f_3
.	.	.	.
.	.	.	.
.	.	.	.
$e_k = [a_{k-1},\ a_k[$	c_k	n_k	f_k
Total	–	N	1

TABLE 1.4: Tableau de destribution

Exemple 1.3.12. (Tableau statistique pour un caractère qualitatif) :
Les groupes sanguins de 100 étudiants est resumés dans le tableau suivant :

Groupe sanguin de x_i	Effectifs n_i	Fréquences relatives f_i
A	40	0.40
B	43	0.43
AB	12	0.12
O	05	0.05
Total	100	1

TABLE 1.5: Répartition des étudiants en fonction de leurs groupes sanguins

Exemple 1.3.13. (Cas d'un caractère quantitatif discret) Un laboratoire dispose de 20 lots d'animaux, on s'interesse alors au nombre de souris dans chacun des lots. Les résultats obtenus sont résumés dans le tableau suivant :

x_i	0	1	5	10	12	15	18	20	Total
n_i	4	1	2	5	4	1	1	2	20
N_i	4	5	7	12	16	17	18	20	
f_i	0.20	0.05	0.10	0.25	0.20	0.05	0.05	0.1	1
F_i	0.20	0.25	0.35	0.60	0.80	0.85	0.90	1	

TABLE 1.6: repartition des souris sur les 20 lots

Exemple 1.3.14. (cas continu)
On s'intéresse à la taille (cm) de 20 étudiants, les résultats obtenus sont :

$$140, 144, 150, 156, 142, 146, 152, 157, 143, 147, 153, 158, 143, 148, 154, 159, 144, 150, 155, 163.$$

Dans ce cas, on doit regrouper cette série en classes. Par la règle de Sturges le nombre de classe est

$$k = 1 + \log_2 N = 1 + \frac{\ln 20}{\ln 2} = 5.32 \simeq = 5.$$

On calcul ensuite l'amplitude de chaque classe, qui est égal

$$a_i = \frac{E(X)}{k} = \frac{163 - 140}{5} = 4.6 \simeq 5.$$

On obtient alors le tableau suivant

Classes	x_i	n_i	f_i	N_i	F_i
$[140 - 145[$	142.5	6	0.30	6	0.30
$[145 - 150[$	147.5	3	0.15	9	0.45
$[150 - 155[$	152.5	5	0.25	14	0.70
$[155 - 160[$	157.5	5	0.25	19	0.95
$[160 - 165[$	162.5	1	0.05	20	1
Total	-	20	1		

1.3.11 Représentation graphique de données statistiques

Cas d'un caractère qualitatif

Diagramme en bâtonnet

Le diagramme en bâtonnets (ou tuyaux d'orgue) est une représentation graphique de la distribution de fréquences d'une variable qualitative. Les "bâtonnets" sont bien séparés pour indiquer les différentes catégories. La hauteur d'un bâtonnet est proportionnelle à la fréquence de la catégorie correspondante (Figure 1.1).

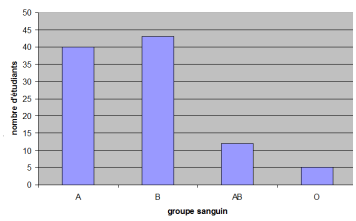


FIGURE 1.1: diagramme en bâtonnet des groupes sanguins de 100 étudiants

Le camembert (diagramme circulaire) dans le diagramme circulaire, chaque secteur a une surface proportionnelle à la fréquence de chaque modalité (Figure 1.2).

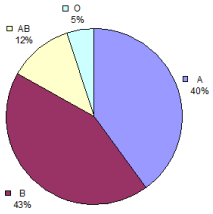


FIGURE 1.2: Diagramme en circulaire des groupes sanguins de 100 étudiants

Cas d'un caractère quantitatif

Il existe deux types de représentation graphique d'une distribution statistique à caractère quantitatif :

- Le diagramme différentiel correspond à une représentation des effectifs ou des fréquences.
- Le diagramme intégral correspond à une représentation des effectifs cumulés, ou des fréquences cumulées.

Variable statistique discrète :

Le diagramme différentiel en bâtons est réalisé en fonction des effectifs ou des fréquences, à la différence du cas qualitatif, les abscisses sont des valeurs de la variable (voir figure 1.3).

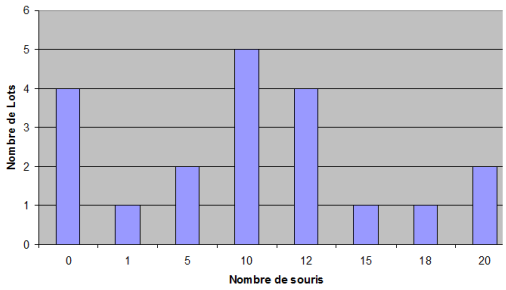


FIGURE 1.3: Repartition des nombre de souris sur les 20 lots

Diagramme (courbe) intégral : c 'est une courbe en escalier réalisée en fonction des effectifs cumulés ou des fréquences cumulées. Dans cette représentation les effectifs ou les fréquences des diverses valeurs de la variable statistique correspondent aux hauteurs des marches de la courbe (voir figure 1.4).

Variable statistique continue :
Histogramme et polygone des effectifs ou des fréquences :

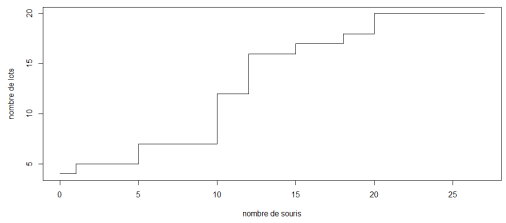


FIGURE 1.4: Courbe intégral des nombre de souris sur les 20 lots

L'histogramme est une représentation graphique (en tuyaux d'orgue) de la distribution des effectifs ou des fréquences d'une variable quantitative. Souvent, les "tuyaux" sont accolés pour montrer la continuité de la variable. La hauteur du tuyau est proportionnelle à l'effectif ou la fréquence de la classe correspondante (Figure 5).

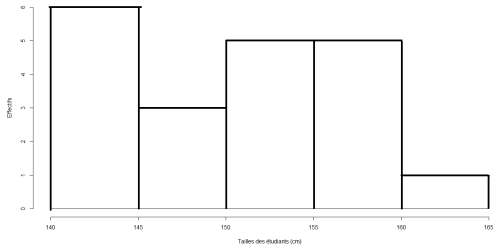


FIGURE 1.5: La distribution de la taille des étudiants, histogramme des effectifs

Le polygone des effectifs ou des fréquences : est une autre représentation graphique (en ligne brisée) de la distribution des effectifs ou des fréquences d'une variable quantitative. Pour tracer le polygone, on joint les points milieu du sommet des rectangles adjacents par un segment de droite. Le polygone est fermé aux deux bouts en le prolongeant sur l'axe horizontal (voir figure 1.7).

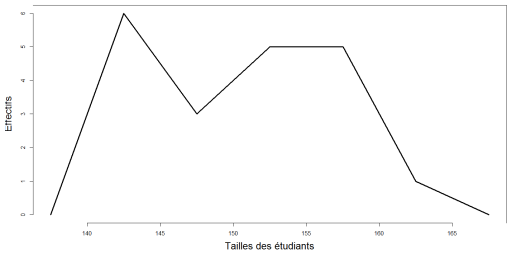


FIGURE 1.6: La distribution de la taille des étudiants(polygones des effectifs)

Diagramme (courbe intégral) la courbe intégrale en fonction des effectifs cumulés ou des fréquences cumulées appelée parfois ogive (Figure 6). Une telle représentation est utile par exemple pour un calcul graphique de la valeur médiane (voir paragraphe **??**

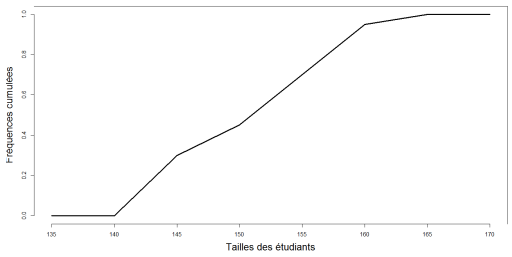


FIGURE 1.7: Polygones des fréquences cumulées.

1.3.12 Description numérique

Dans une description statistique des données ils est intéressant d'avoir la possibilité de résumé un partie de l'information par des paramètres numériques appelés valeurs atypique. Ils existent deux grandes familles de ces valeurs : paramètres de tendance centrale et et paramètres de dispersion.

Tendance centrale ou Paramètres de position

Le mode Le mode est la valeur la plus fréquente d'une distribution. Il se calcule toujours à partir d'un dénombrement des modalités du caractère. Comme pour le tableau de dénombrement, il faut distinguer le cas des caractères discrets et des caractères continus.
Caractère discret : pour un caractère quantitatif discret le mode noté Mod est la modalité qui a la fréquence la plus élevée (ou l'effectif le plus élevé).

Exemple 1.3.15. :

- 10, 11, 12, 10, 10, 10, 9, 14 ? $Mod = 10$ (4 fois), la distribution unimodale.
- 10, 11, 12, 10, 10, 12, 12, 9, 14 ? Deux modes : $Mod_1 = 10$ (3 fois) et $Mod_2 = 12$ (3 fois), la distribution est bimodale.

Caractère quantitatif continu : Les modalités étant en nombre infini, il est peu probable que deux éléments aient la même valeur. Dans ce cas, le mode ne peut pas être défini directement, il faut au préalable établir une partition en classes. Le mode est avant tous situer dans une classe appelée "***classe modale***", la deuxième étape consiste à déterminer une valeur approchée du mode en utilisant la formule suivante : Si e_i est la classe modale (la classe contenant le plus grands effectifs ou la plus grande fréquence, alors :

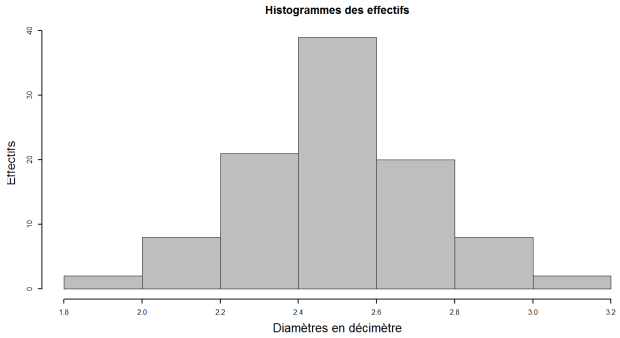
$$Mod = \min(e_i) + a_i \frac{\triangle_1}{\triangle_1 + \triangle_2} \tag{1.1}$$

- a_i l'amplitude de la classe modale.
- $\triangle_1$ c'est l'excès de la classe modale par rapport à la classe précédente.
- $\triangle_2$ c'est l'excès de la classe modale par rapport à la classe suivante.

Exemple 1.3.16. Chez un fabricant de tubes de plastiques, on a prélevé un échantillon de 100 tubes dont on a mesuré le décimètre en centimètre. Les résultats sont regroupés dans le tableaux suivant.

1.94	2.20	2.33	2.39	2.45	2.50	2.54	2.61	2.66	2.85
1.96	2.21	2.33	2.40	2.46	2.51	2.54	2.62	2.68	2.87
2.07	2.26	2.34	2.40	2.47	2.52	2.55	2.62	2.68	2.90
2.09	2.26	2.34	2.40	2.47	2.52	2.55	2.62	2.68	2.91
2.09	2.28	2.35	2.40	2.48	2.52	2.56	2.62	2.71	2.94
2.12	2.29	2.36	2.41	2.49	2.52	2.56	2.63	2.73	2.95
2.13	2.30	2.37	2.42	2.49	2.53	2.57	2.63	2.75	2.99
2.14	2.31	2.38	2.42	2.49	2.53	2.57	2.65	2.76	2.99
2.19	2.31	2.38	2.42	2.49	2.53	2.59	2.66	2.77	3.09
2.19	2.31	2.38	2.42	2.50	2.54	2.59	2.66	2.78	3.12

La règle de Sturge permet d'obtenir 7 classes d'amplitude $a_i = 0.2$ décimètre, ce qui permet de construire l'histogramme suivant :



La classe modale est donc la quatrième classe $e_4 = [2.4 - 2.6[$, le mode est égal alors :

$$\boxed{Mod = 2.4 + 0.2\frac{39 - 21}{(39 - 21) + (39 - 20)} = 2.50}$$

La médiane Les valeurs étant classées par ordre croissant, la médiane est la valeur du caractère qui partage la population (ou l'échantillon) en deux sous-ensembles d'effectifs égaux : 50% des valeurs lui sont supérieures et 50% lui sont inférieures.

• **Cas discret :** Soit $x_{(1)}, x_{(1)}, \ldots, x_{(n)}$ les valeurs de la variables statistiques classées par ordres croissant, on distingue alors deux cas :
n est **impair** : si $n = 2p + 1$ alors

$$\boxed{Med = x_{(p+1)}}$$

n est **pair** : si $n = 2p$ alors

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

Exemple 1.3.17. Dans l'exemple (1.3.13), on réarrange dans l'ordre croissant les valeurs de la variable sur les 20 lots, on obtient le tableau suivant :

$$0, 0, 0, 0, 1, 5, 5, 10, 10, 10, 10, 10, 12, 12, 12, 12, 15, 18, 20, 20$$

Comme $n = 20 = 2 \times 10$ (pair) donc

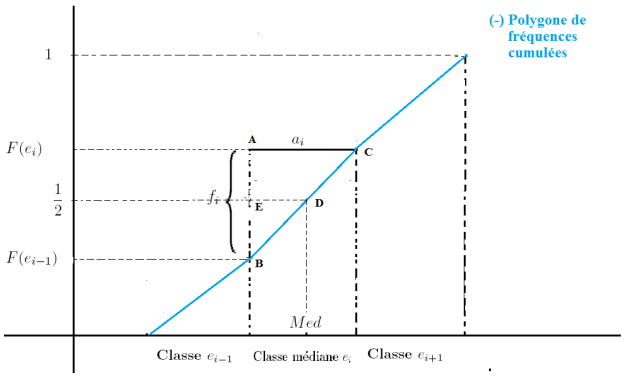
$$Med = \frac{x_{(10)} + x_{(11)}}{2} = \frac{10 + 10}{2} = 10$$

• **Cas continu :**Les modalités étant en nombre infini et généralement réarrangées en classe, la valeur de la médiane dans ce cas est calculée d'une manière approchée en utilisant le polygone des fréquences cumulées et la proposition suivante :

Propriété 1.1. Si F est la fonction correspondante au polygone des fréquences cumulées alors :

$$F(Med) = \frac{1}{2}. \tag{1.2}$$

D'autre part en appliquant



En appliquant le théorème de Tales sur les deux triangles ABC et EBD, on obtient :

$$\frac{|AB|}{|EB|} = \frac{|AC|}{|ED|} \tag{1.3}$$

$$= \frac{a_i}{|ED|} \tag{1.4}$$

$$\iff |ED| = a_i \times \frac{|EB|}{|AB|} \tag{1.5}$$

Graphiquement on a aussi :

$$Med = min(e_i) + |ED| = min(e_i) + a_i \times \frac{|EB|}{|AB|} \tag{1.6}$$

D'autre part

$$|EB| = \frac{1}{2} - F(e_{i-1}), \tag{1.7}$$

$$|AB| = F(e_i) - F(e_{i-1}). \tag{1.8}$$

En remplace ces deux dernière formule dans (1.6) on obtient la formule d'approximation graphique de la médiane suivante :

$$\boxed{Med = min(e_i) + a_i \times \frac{\frac{1}{2} - F(e_{i-1})}{F(e_i) - F(e_{i-1}).}} \tag{1.9}$$

Exemple 1.3.18. Reprenons les données de l'exemple (1.3.16), on obtient alors le tableau résumé suivant :

Classes	Centres des classes	n_i	f_i	N_i	F_i
$[1.8 - 2.0[$	1.9	2	0.02	02	0.02
$[2.0 - 2.2[$	2.1	8	0.08	10	0.10
$[2.2 - 2.4[$	2.3	21	0.21	31	0.31
$[2.4 - 2.6[$	2.5	39	0.39	70	0.70
$[2.6 - 2.8[$	2.7	20	0.2	90	0.9
$[2.8 - 3[$	2.9	8	0.08	98	0.98
$[3 - 3.2[$	3.1	2	0.02	100	1

Le polygone des fréquence cumulée est (voir figure 1.8)

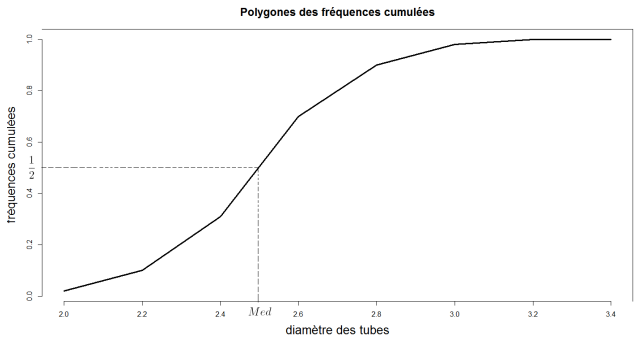


FIGURE 1.8: Calcul de la médiane graphiquement.

La médiane appartient donc à la classe $[2.4, 2.6[$ (classe médiane), en remplaçant dans la formule (1.9) les valeurs numériques correspondantes on obtient :

$$Med = 2.6 + 0.2\ \frac{0.5 - 0.31}{0.7 - 0.31} \simeq 2.7.$$

La moyenne Il s'agit d'une paramètre statistique caractérisant les éléments d'un ensemble de mesures et elle exprime la somme de toutes ces mesures divisée par l'effectif total de l'échantillon. Lorsque ces mesures représentent une quantité partagée entre des individus, la moyenne exprime la valeur qu'aurait chacun si le partage était équitable.

Cas discret :

Soit x une variable statistique discrète à k modalités $x_1,\ x_2, \ldots, x_k$ auxquelles on correspondent les effectifs $n_1,\ n_2, \ldots, n_k$, avec

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

On appelle moyenne arithmétique pondérée (ou moyenne) qu'on note $\bar{x}$, la quantité suivante :

$$\boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{i=k} n_i x_i} \tag{1.10}$$

Dans l'exemple (1.3.13), la moyenne vaut :

$$\bar{x} = \frac{4 \times 0 + 1 \times 1 + 2 \times 5 + 5 \times 10 + 4 \times 12 + 1 \times 15 + 1 \times 18 + 2 \times 20}{20} = 9.1$$

Cas continu : Dans le cas continu le nombre de mesures peut être très grands, on remplace alors dans (1.10) les x_i par les centres des des classes, on obtient alors :

$$\boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{i=k} n_i c_i} \tag{1.11}$$

Dans l'exemple (1.3.16), la moyenne vaut :

$$\bar{x} = \frac{2 \times 1.9 + 8 \times 2.1 + 21 \times 2.3 + 39 \times 2.5 + 20 \times 2.7 + 8 \times 2.9 + 2 \times 3.1}{100} = 2.498$$

Paramètres de dispersion

Les paramètres de positions ne permettent pas à eux seuls de décrire efficacement des données statistiques. les deux exemples suivants illustrent bien la problématique.

Exemple 1.3.19. Les deux séries de notes suivantes ont la même moyenne

- La moyenne des notes 9,10,11 est 10.
- La moyenne des notes 1, 10, 19 est 10.

Pour la premiere série de note, la moyenne est un bon indicateur de niveau des étudiants, ce qui n 'est pas le cas pour la deuxième séries de notes. D'où l'utilité d'introduire d'autres mesures afin de représenter aux mieux la variabilité contenue dans les données. C 'est l'objet du paragraphe suivant.

L'étendue : Pour une séries de valeurs l'étendue noté E est la différence entre la plus grande et la plus petite valeur.
Pour les deux séries de l'exemple précédent l'étendu est $E = 1$ et $E = 18$ respectivement.
Variance et écart type une mesure qui exprime la moyenne des carrés des écarts entre chaque observation (mesure sur chaque individus) et la moyenne des observations est appelé la variance de l'échantillon, qu'on note pas s^2 égale dans le cas discret à :

$$s^2 = \frac{1}{n}\sum_{i=1}^{i=k} n_i(x_i - \bar{x})^2, \tag{1.12}$$

et dans le cas continu à

$$s^2 = \frac{1}{n}\sum_{i=1}^{i=k} n_i(c_i - \bar{x})^2. \tag{1.13}$$

Proposition 1.1.

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{i=k} n_i x_i^2 - \bar{x}^2 \\ &\text{ou} \\ s^2 &= \frac{1}{n}\sum_{i=1}^{i=k} n_i c_i^2 - \bar{x}^2. \end{aligned}$$

L'écart type par définition est la racine carrée de la variance c'est-à-dire égale à s.

Exemple 1.3.20. Dans l'exemple 1.3.13 on ajoute aux tableau (1.6) une ligne contenant les carrées des écarts entre chaque mesure et la moyenne de tous le mesures, on obtient le tableau suivant :

x_i	0	1	5	10	12	15	18	20	Total
n_i	4	1	2	5	4	1	1	2	20
N_i	4	5	7	12	16	17	18	20	
f_i	0.20	0.05	0.10	0.25	0.20	0.05	0.05	0.1	1
F_i	0.20	0.25	0.35	0.60	0.80	0.85	0.90	1	
$(x_i - \bar{x})^2$	82.81	65.61	16.81	0.81	8.41	34.81	79.21	118.81	407.28

TABLE 1.7: repartition des souris sur les 20 lots

La variance est donc égale à

$$s^2 = \frac{407.28}{20} = 20.364$$

L'étendue interquartile : Les quartiles sont les trois valeurs qui permettent de découper la distribution en quatre classes d'effectifs égaux, on les note $Q1$, $Q2$ et $Q3$.

Le premier quartile, noté $Q1$, est le point qui sépare la portion de 25% des valeurs les plus petites de la portion de 75% des valeurs les plus grandes.

Le troisième quartile, noté $Q3$, est le point qui sépare la portion de 25% des valeurs les plus grandes de la portion de 75% des valeurs les plus petites.

Le deuxième quartile Q2 correspond tout simplement à la médiane.

L'intervalle interquartile $(Q3 - Q1)$ est un paramètre de dispersion absolue qui correspond à l'étendue de la distribution une fois que l'on a retiré les 25% des valeurs les plus faibles et les 25% des valeurs les plus fortes. 50% des observations sont donc concentrées entre $Q1$ et $Q3$.

Cas discret : soit à calculer le quartile Qi :
Soit j la partie entière de $i.(n+1)/4$ et k la partie fractionnaire de $i.(n+1)/4$.
Soit $x_{(j)}$ et $x_{(j+1)}$ les valeurs des observations classées respectivement en j^{ieme} et $(j+1)^{ieme}$ position (lorsque les observations sont classées par ordre croisant).
Alors :

$$Qi = x_{(j)} + k \times (x_{(j+1)} - x_{(j)}) \tag{1.14}$$

Dans l'exemple (1.3.13), on réarrange dans l'ordre croissant les valeurs de la variable sur les 20 lots, on obtient le tableau suivant :

$$0, 0, 0, 0, 1, 5, 5, 10, 10, 10, 10, 10, 12, 12, 12, 12, 15, 18, 20, 20$$

Calculons Q1 :

$$j = [\frac{20+1}{4}] = [\frac{21}{4}] = 5$$

$$k = \frac{20+1}{4} - [\frac{21}{4}] = 5.25 - 5 = 0.25$$

Donc :

$$\boxed{Q1 = x_{(5)} + 0.25 \times (x_{(6)} - x_{(5)}) = 1 + 0.25 \times (5-1) = 2}$$

Calculons Q3 :

$$j = [\frac{3 \times (20+1)}{4}] = [\frac{63}{4}] = 15$$

$$k = \frac{3 \times (20+1)}{4} - [\frac{3 \times (20+1)}{4}] = 15.75 - 15 = 0.75$$

Donc :

$$\boxed{Q3 = x_{(15)} + 0.75 \times (x_{(16)} - x_{(15)}) = 12 + 0.75 \times (12-12) = 15.}$$

Cas continu : le calcul des deux quartiles $Q1$ et $Q3$ se fait en généralisant cette de la médiane, c 'est à dire le deuxième quartile $(Q2)$.
En général un quartile Qi vérifie la relation :

$$F(Qi) = \frac{i}{4}$$

ou F est la fonction correspondante au polygone des fréquences cumulées.
Donc si e_{Qi} est la classe contenant le quartile Qi alors

$$Qi = \min(e_{Qi}) + a_i \frac{\frac{i}{4} - F(e_{Qi-1})}{F(e_{Qi}) - F(e_{Qi-1})} \tag{1.15}$$

où $F(e_{Qi-1})$ (resp $F(e_{Qi})$) est la fréquence cumulées de la classe e_{Qi-1} (resp e_{Qi}).

Dans l'exemple (1.3.16), la classe qui contient le premier quartile Q_1 est $[2.2 \ - \ 2.4]$, on déduit alors :

$$\boxed{Q1 = 2.2 + 0.2 \times \frac{0.25 - 0.1}{0.31 - 0.1} \simeq 2.34}$$

La classe qui contient le premier quartile Q_3 est $[2.6 \ - \ 2.8]$, on déduit alors :

$$\boxed{Q3 = 2.6 + 0.2 \times \frac{0.75 - 0.7}{0.9 - 0.7} = 2.65.}$$