République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Centre Universitaire Nour Bachir - El-Bayadh

Institut des Sciences

Département des Sciences de la Nature et de la Vie

Polycopié de cours

Module de Biostatistique

Destiné aux étudiants de 3ème LMD/Licence 3

Biochimie, Microbiologie et Ecologie

Réalisé par :

1- **Dr. BOUKHOBZA Abdelyamine**
2- **Dr. ALAMI Mohamed**
3- **Dr. AZOUZ Noureddine**

Année Universitaire 2023/2024

AVANT-PROPOS

Ce polycopié de cours de biostatistique est destiné en premier lieu aux étudiants de 3ème année (Biochimie, Microbiologie et Ecologie) des sciences de la nature et de la vie.

Cette modeste contribution c'est une plateforme pour l'enseignant du premier cycle de l'enseignement supérieur dont la tache pédagogique est immense due à son caractère général. Le but de ce polycopie est de mettre à la disposition de l'enseignant un document lui servant de point de départ pour étendre et approfondir ses connaissances dans cette discipline scientifique.

Ce polycopié est une synthèse des cours de biostatistique qui assure à l'étudiant de département des sciences de la nature et de la vie de centre universitaire Nour Bachir El-Bayad de comprendre les notions de la statistique. Ce travail est le fruit d'un effort intrinsèque émanant de mon expérience propre en tant qu'enseignant de cours de biostatistique.

SOMMAIRE

La statistique descriptive

Etude de corrélation et Régression

Estimation ponctuelle et par intervalle de confiance

REFERENCES

La statistique descriptive

OBJECTIF

- Comprendre l'intérêt de l'utilisation biostatistique en santé, biologie et environnement ;
- Apprendre les principales techniques de statistique descriptive univariée et bivariée.
- Différencier entre statistique et statistiques
- Comprendre le lien entre probabilités et statistique(s)
- Appréhender la notion de variabilité et de ses sources
- Savoir définir les différents types de variables et les choisir en fonction du problème étudié
- Comprendre les différences entre démarche d'observation et démarche expérimentale

I. Historique

De tous temps, les chefs d'Etat ont souhaité déterminer la puissance des nations qu'ils dirigeaient à l'aide de recensements partiels ou complets (population, territoire, production,…) Dès 3000 av. J.-C., on trouve mention de collectes d'observations sur les biens et les personnes en Mésopotamie.

En 1200 av. J.-C., des évaluations de productions agricoles sont effectuées en Chine.Au début de notre Ere a lieu un dénombrement des richesses de l'Empire Romain, rendu célèbre par sa mention dans l'Evangile de Luc.

Au Moyen Age, des relevés sont exécutés sur l'ordre de Charlemagne puis de Guillaume le Conquérant. Dans les deux cas, le but est de se faire une idée plus précise des richesses du pays.

Au XVIIe Siècle, pour éviter le recensement lourd et onéreux, William Petty (1623-1687) met au point une méthode de comptage de la population de Londres sur base des proportions moyennes entre :

- Les maisons
- Les feux (ménages) par maison
- La composition des familles

Au XIXe Siècle, les recensements proprement dits reprennent de l'importance et, en 1853, a lieu à Bruxelles le 1er Congrès International de Statistique, sous l'impulsion d'Adolphe Quételet (1796-1874, astronome et mathématicien belge, un des fondateurs de la science statistique). L'objectif de ce congrès est d'uniformiser les techniques de compilation des statistiques nationales, en vue de faciliter les comparaisons.

Au début du XXe Siècle, un débat oppose les partisans des recensements (réalisés sur l'ensemble de la population) et des sondages (réalisés sur un échantillon représentatif de la population).Les recensements ne sont pas toujours possibles, ni souhaitables. Dans certains cas, ils peuvent être trop chers (comme, par exemple, des enquêtes sur toute la population d'un pays). Ils peuvent aussi contenir des erreurs. Parfois, ils sont carrément aberrants (mesurer la solidité moyenne d'un type de voiture en lançant toutes les voitures de ce type contre un mur serait commercialement inacceptable).

Pour pallier ces inconvénients, on a recours au sondage statistique, qui consiste à :

Déduire les propriétés de toute *une population* à partir de l'analyse *d'un échantillon*.Il est capital que l'échantillon soit choisi et analysé de manière adéquate. En particulier, il faut que l'échantillon soit représentatif de la population. Un échantillon non représentatif est dit biaisé.

Au début du XXe siècle, beaucoup de journaux américains réalisent des « votes de paille » en demandant leur avis par écrit à plusieurs millions de personnes quelques semaines avant les élections présidentielles.

En 1936, le Literary Digest prédit, à l'aide d'un échantillon de 2.400.000 électeurs, la victoire du candidat républicain N. Landon. George Gallup, grâce à un sondage sur 4000 personnes judicieusement choisies, prévoit, quand à lui, la victoire du démocrate Franklin D. Roosevelt. La victoire de ce dernier sonne le glas des votes de paille dont les échantillons sont souvent biaisés (les cartes du Literary Digest avaient été envoyées aux abonnés du téléphone et aux propriétaires de voitures, cet électorat aisé était plus favorable aux républicains).

Il y a statistiques et Statistique. "Les statistiques" - familièrement: les "stats" - ce sont les données chiffrées (moyennes, pourcentages, indices de toute sorte) des mass media et qu'on rencontre aujourd'hui dans tous les secteurs possibles et imaginables: statistiques officielles (INSEE), sondages, etc. "La statistique" - au singulier, l s'agit d'un ensemble de méthodes, de raisonnements, de règles de décision permettant de présenter, d'analyser des données et d'extrapoler à une population les résultats obtenus dans le cadre d'une étude ponctuelle en prenant en compte la variabilité des phénomènes observés. Objet de notre cours.

II. Biostatistique :

Biostatistique est un champ scientifique constitué par l'application de la ***science statistique*** à **la *biologie*** et à ***la médecine***. La biostatistique : ensemble des méthodes qui ont pour objet :
La collecte des données ; Afficher et tabuler les données ; Le traitement et L'interprétation des données ; puis affichés les résultats et prend la décision.

III. Terminologies

III. 1. Population :

La *population* est l'ensemble des individus (ou unités statistiques) auxquels on décide de s'intéresser. Habituellement désignée par N, est grande, ou même infinie.

Exemple :

La population peut être l'ensemble des étudiants de centre universitaire Elbayadh.

III. 2. Échantillon :

C'est un sous ensemble de la population considérée. Le nombre d'individus dans *l'échantillon* est la taille de l'échantillon.

Exemple :

Echantillon peut être sous ensemble des étudiants de la SNV de centre universitaire Elbayadh.

III.3. Individu : (ou unités statistiques)

Un élément de cette population ou de ce échantillon est appelé *individu*.

Exemple :

Individu peut être un étudiants de la SNV ou de centre universitaire Elbayadh.

III. 4. Taille :

La *taille* c'est le nombre d'observations constituant d'un échantillon (n) ou d'une population (N).

Exemple :

Le nombre d'observations dans l'échantillon de la SNV est 50 étudiants, ou 200 étudiants dans le centre universitaire Elbayadh.

III. 5. Caractère :

C'est la propriété ou l'aspect singulier que l'on se propose d'observer dans la population ou l'échantillon. *Un caractère* qui fait le sujet d'une étude porte aussi le nom de *variable statistique*. On se pose la question «Sur quoi porte l'étude ? Qu'est-ce qui est étudié ?»

Exemples :

Si je relève leur couleur favorite, le caractère c'est leur couleur préférée.

Si je mesure la hauteur des sommets de montagnes, le caractère c'est la hauteur.

Concernant un groupe de personnes, on peut s'intéresser à leur sexe, taille, poids,….

III. 6. Les modalités :

Les différentes manières d'être que peut présenter un caractère, *Les modalités* correspondent donc à l'ensemble des valeurs possibles. Une modalité est la valeur prise par une variable statistique qu'elle soit qualitative ou quantitative.

Exemple :

1- Le sexe est un caractère qui présente deux modalités : féminin ou masculin.
2- Quant au nombre d'enfant par famille, les modalités de ce caractère peuvent être 0,1,2,3,4,…….,20 .

IV. Différents types de variables statistiques (Caractère) :

IV.1. Caractère (variable) qualitatif :

Lorsque la variable ne se prête pas à des valeurs numériques, elle est dite *qualitative*.

Exemple :

Opinions politiques, couleurs des yeux, religion...) .Elle peut être ordonnée ou non, dichotomique ou non.

IV.2. Caractère (variable) qualitatif Nominal :

La variable est dite *qualitative nominale* quand les modalités ne peuvent pas être ordonnées.

Exemple :

Couleur des yeux : noir, vert,marron,…etc).

IV.3. Caractère (variable) qualitatif Ordinal :

La variable est dite *qualitative ordinale* quand les modalités peuvent être *ordonnées.* (Exmple niveau d'étude : primaire,secondaire, supérieur).Le fait de pouvoir ou non ordonner les modalités est parfois discutable. Par exemple : dans les catégories socio-professionnelles, on admet d'ordonner les modalités : 'ouvriers', 'employés', 'cadres'. Si on ajoute les modalités 'sans profession', 'enseignant', 'artisan', l'ordre devient beaucoup plus discutable.

IV.4. Caractère (variable) quantitatif :

Lorsque la variable peut être exprimée numériquement, elle est dite *quantitative* (ou mesurable).Dans ce cas, elle peut être *discontinue (discret)* ou *continue*.

IV.5. Caractère (variable) quantitatif Discontinu (discret) :

Il est *discontinu* s'il ne prend que des valeurs isolées. Une variable discontinue qui ne prend que des valeurs entières est dite discrète,(représentées par des entiers naturels (N) .

Exemple :

Le nombre d'enfants d'une famille, 0,2,3….etc.

IV.6. Caractère (variable) quantitatif Continu :

Il est dit *continu* lorsqu'il peut prendre toutes les valeurs d'un intervalle fini ou infini.

Exemple :

Taille, Poids, Age,...etc.

V. Série statistique :

On appelle *série statistique* la suite des valeurs prises par une variable X sur les unités d'observation. Le nombre d'unités d'observation est noté $\boldsymbol{n}$. Les valeurs de la variable X sont notées. $x_1,\ldots,x_i,\ldots,x_n$.

Exemple :

La série suivante résulte d'une courte enquête auprès de quelques personnes pour connaitre leur âge : 18 21 19 19 17 18 22 23 18 23.

V. 1. L'effectif :

L'*effectif* est le nombre d'individus pour lesquels une variable statistique a pris une valeur donnée on le note n_i (i est l'indice de la classe). Si, sur 150 familles, 50 ont 2 enfants, on dira que l'effectif n_i correspondant à la valeur $x_i = 2$ du variable "nombre d'enfants", est 50.

V.1.1. Effectifs cumulés :

Résultat de l'addition, de proche en proche, des effectifs d'une distribution observée, soit en commençant par le 1er :

V.1.2. Effectifs cumulés croissants :

$N_1 = n_1$, $N_2 = n_1 + n_2$, ... , $N_i = n_1 + n_2 + ... + n_i$

Soit en commençant par le dernier :

V.1.3. Effectifs cumulés décroissants :

$N'_K = n_K$, $N'_{K-1} = n_K + n_{K-1}$, ... , $N_i' = n_K + n_{K-1} + ... + n_i$

Exemple :

Le tableau suivent présenté l'effectif , Effectifs cumulés croissants et Effectifs cumulés décroissants

Nombre d'appels	Effectif n_i	Effectifs cumulés croissants	Effectifs cumulés décroissants
0	2	2	96
1	14	16	94
2	23	39	80
3	24	63	57
4	18	81	33
5	9	90	15
6	6	96	6
Total :	96		

V.1.4. Effectif total :

L'*Effectif total* le nombre d'observations, d'une série statistique brute, nombre d'individus de la population étudiée. Il est égal à la somme des effectifs associés aux différentes modalités, valeurs ou classes :

$$n = \sum_{i=1}^{K} n_i$$

Exemple :

L'*Effectif total* de l'exemple précédent (Le tableau) égal à la somme des effectifs :

N=2+14+23+24+18+9+6=96

V. 2. La fréquence :

La *fréquence* est une proportion d'individus de la population ou de l'échantillon appartenant à la classe : on la note f_i. f_i et n_i sont liés par :

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

Où ***N*** est le nombre total d'individus dans la population.

Remarque : On peut remplacer f_i par $f_i \times 100$ qui représente alors *un pourcentage.*

Donc *le pourcentage* est *:* $P = f_i \times 100$

On a toujours :

$$\sum_{i=1}^{k} n_i = N \qquad 0 \leq f_i \leq 1 \qquad \sum_{i=1}^{k} f_i = 1$$

Où *k* représente le nombre de classes.

Exemple :

Sur 150 familles, 50 ont 2 enfants, on dira que la fréquence f_i correspondant à la valeur $x_i = 2$ de la variable "nombre d'enfants", est : = 0.33 soit 50/150 = 1/3 ou 33.33 %.

V.1.1. Fréquences cumulées :

Résultat de l'addition, de proche en proche, des fréquences d'une distribution observée, soit en commençant par le 1er :

V.1.2. Fréquences cumulées croissantes

$F_1 = f_1$, $F_2 = f_1 + f_2$, ... , $F_i = f_1 + f_2 + ... + f_i$;

Soit en commençant par le dernier :

V.1.3. Fréquences cumulées décroissantes

$F'_K = f_K$, $F'_{K-1} = f_K + f_{K-1}$, ... , $F'_i = f_K + f_{K-1} + ... + f_i$.

Exemple :

Nombre d'appels	Fréquences en %	Fréquences cumulées croissantes	Fréquences cumulées décroissantes
0	2.08	2.08	100
1	14.58	16.66	97.92
2	23.96	40.62	83.34
3	25.00	65.62	59.38
4	18.75	84.37	34.38
5	9.38	93.75	15.63
6	6.25	100	6.25
Total :	100		

V.3. Les moyennes :

Il existe différentes moyenne tel que la moyenne arythmique, Moyenne quadratique, Moyenne harmonique et Moyenne géométrique.

V.3.1. Moyenne arythmique :

La *moyenne arythmique* correspond à une répartition « équitable » de la grandeur mesurée sur tous les individus, La moyenne est la somme des grandeurs mesurées divisée par le nombre d'individus.

1- Distribution simple :

Pour un échantillon de *n* individus, la moyenne est calculée par :

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{N} \quad ; \quad \overline{X} \cong \frac{1}{\mathrm{N}} \sum_{i=1}^{k} x_i$$

Exemple :

• Dans un échantillon de 9 personnes, le poids moyen vaut :

$$\overline{\mathrm{X}} = \frac{45+68+89+74+62+56+49+52+63}{9} = 62 \quad \mathrm{kg}$$

• Dans le second échantillon de 10 personnes, le poids moyen vaut :

$$\overline{\mathrm{X}} = \frac{45+49+52+55+56+62+63+68+74+89}{10} = 61{,}3 \quad \mathrm{kg}$$

2- Distribution fréquentée :

La moyenne est le nombre obtenu en additionnant les produits de ces valeurs par leurs coefficients et en divisant le résultat par la somme des coefficients.

$$\overline{X} \cong \frac{1}{\mathrm{N}} \sum_{i=1}^{k} n_i . x_i \quad \text{(La moyenne pondérée)}$$

$$\overline{X} \cong \sum_{i=1}^{k} x_i . f_i \quad \text{(Avec } f \text{ la fréquence)}$$

Pour des données groupées en classes, on peut calculer une valeur approximative de la moyenne en supposant que tous les individus d'une classe se situent au centre de celle-ci.

Si C_i est *le centre de la classe* et *ni* le nombre d'individus dans celle-ci, la formule approchée s'écrit :

$$\overline{X} \cong \frac{1}{N}\sum_{i=1}^{k} n_i.C_i$$

Avec $C_i = \frac{a+b}{2}$ ou *a* et *b* sont les bornes inferieur et supérieur de la classe.

Exemple :

Les notes suivantes de 12 étudiants en biologie au cours de premier semestre.

Notes (x_i)	14	14.5	9	15	20	16
Effectif (n_i)	1	2	1	3	2	3
Fréquences (f)	1/12=0.08	2/12=0.16	1/12=0.08	3/12=0.25	2/12=0.16	3/12=0.25

$$\bar{X} = \frac{14 \times 1 + 14.5 \times 2 + 9 \times 1 + 15 \times 3 + 20 \times 2 + 16 \times 3}{1 + 2 + 1 + 3 + 2 + 3} = 15.4$$

$$\bar{X} = 14 \times 0.08 + 14.5 \times 0.16 + 9 \times 0.08 + 15 \times 0.25 + 20 \times 0.16 + 16 \times 0.25 = 15.4$$

Exemple :

Dans l'exemple précédent 9 personnes (page 7) avec une répartition comme suite :

Classe	**Centre C_i**	**Nombre n_i**
45-55	50	3
55-65	60	3
65-75	70	2
75-85	80	0
85-95	90	1

$$\overline{X} \cong \frac{3\times 50 + 3\times 60 + 2\times 70 + 0\times 80 + 1\times 90}{9} = 62{,}2 \quad \text{kg}$$

Dans l'exemple précédent, la formule approchée donne un poids moyen de 62,2 kg au lieu de 62 kg. La formule approchée donnera des résultats d'autant meilleurs que :

- Les classes seront étroites
- Le nombre d'individus par classe sera grand.

V.3.2. Moyenne quadratique :

Une moyenne qui trouve des applications lorsque l'on a affaire à des phénomènes présentant un caractère sinusoïdal avec alternance de valeurs positives et de valeurs négatives. Elle est, de ce fait, très utilisée en électricité. Elle permet notamment de calculer la grandeur d'un ensemble de nombre. Elle s'écrit :

Moyenne quadratique (x_1,.....,x_n) $\overline{Q} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$

Exemple :

Prenons un rapide exemple : considérons les nombre suivants {-2, 5, -8, 9, -4 } Nous pouvons en calculer la moyenne arithmétique avec l'inconvénient de voir se neutraliser les valeurs positives et négatives et d'aboutir à un résultat nul sans que cela ne nous apprenne quoi que ce soit.

En effet $\overline{X} = 0$

Le calcul de la moyenne quadratique pour la même série donne : $\bar{Q} = 6,16$.

V.3.3. Moyenne harmonique :

La moyenne harmonique se définit comme l'inverse de la moyenne arythmique des inverses.

$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{X_i}}$ (Cas discrète) $H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{C_i}}$ (Cas continue)

Xi	n_i	n_i/x_i
1	40	40
2	80	40
3	60	20
4	20	5
total	200	105

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

$$H = \frac{200}{105} = 1.904$$

Xi	n_i	C_i	n_i/C_i
0-5	70	2.5	28
5-10	110	7.5	14.66
10-15	80	12.5	6.4
15-20	40	17.5	2.28
total	300		51.34

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{c_i}}$$

$$H = \frac{300}{51.34} = 5.843$$

V.3.4. Moyenne géométrique :

La moyenne géométrique est un instrument permettant de calculer des taux moyens, notamment des taux moyens annuels. Son utilisation n'a un sens que si les valeurs ont un caractère multiplicatif. La moyenne géométrique de n valeurs positives x_i est la racine $ni^{ème}$ du produit de ces valeurs. Notée , elle s'écrit :

Les prix de l'immobilier ancien ont augmenté ces 10 dernières années de la façon suivante :

Année	Variation annuelle (%)
1	9,2
2	12,7
3	8,8
4	7,7
5	3,9
6	1,7
7	0,9
8	2,2
9	4,7
10	3,3

$$\hat{G} = \sqrt[n]{x_1 x\, x_2 x \dots \dots \dots \dots x x_n}$$

En utilisant la moyenne arithmétique simple, on obtiendrait une évolution moyenne de (9,2+12,7+8,8+7,7+3,9+1,7+0,9+2,2+4,7+3,3)/10 = 55,1 / 10 = 5,51 % mais ce résultat est faux compte tenu de la relation entretenue par les taux d'une année sur l'autre. L'utilisation de la moyenne géométrique permet de solutionner ce problème :

$$\hat{G} = \sqrt[10]{9{,}2x12{,}7x8{,}8x7{,}7x3{,}9x1{,}7x0{,}9x2{,}2x4{,}7x3{,}3}$$

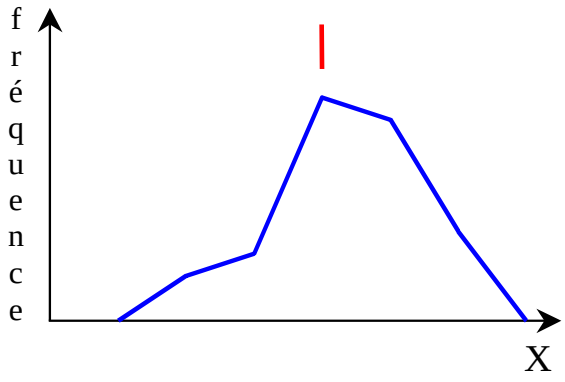
$$\hat{G} = \sqrt[10]{1\,611\,964{,}4588}$$

$$\hat{G} = 4{,}1757$$

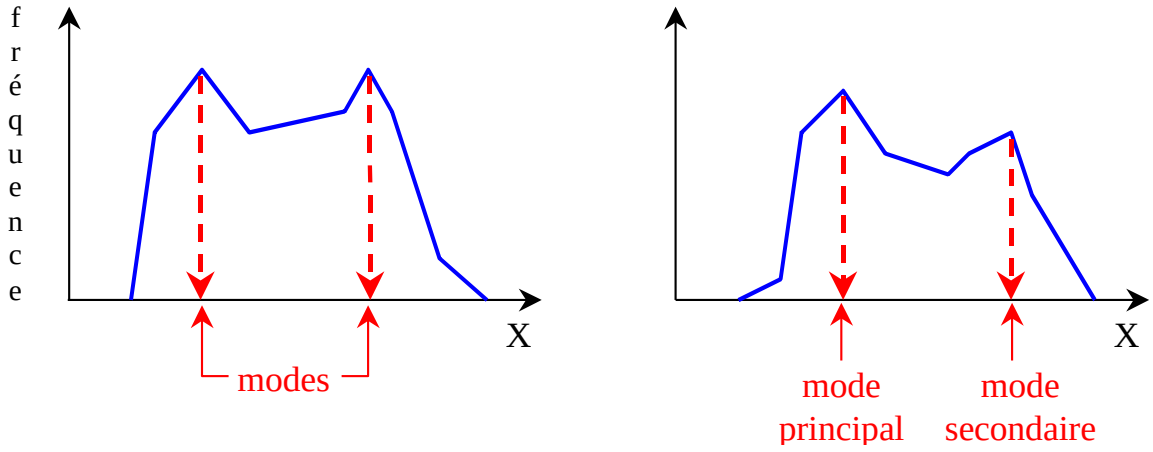
V.4. Le mode :

Le *mode* correspond au sommet de la distribution : le mode est la valeur la plus fréquente c'est la valeur la plus « à la mode ».

On appelle distribution *unimodale*, une distribution présentant un seul mode.



Une distribution *bimodale* est une distribution présentant deux modes.



Une distribution *multimodale* est une distribution présentant plusieurs modes (2,3,…). Elle est souvent le reflet d'une population composée de plusieurs sous-populations distinctes.

Par exemple, le polygone des fréquences ci- dessous, qui représente la distribution de la taille des individus dans une population adulte, présente deux modes. Ceux-ci sont le reflet de la présence de deux sous-populations : Les femmes et les hommes, ces derniers étant généralement plus grands.

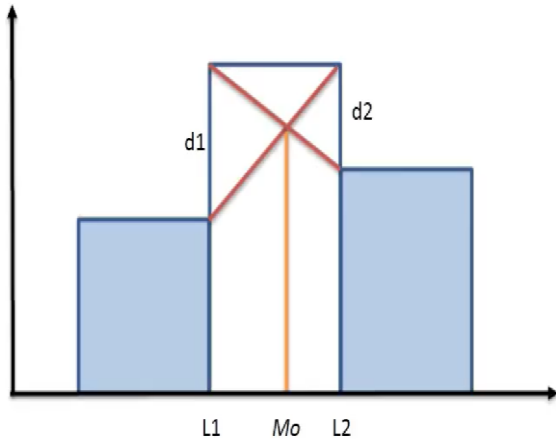
Exemple :

Les valeurs sont illustrées dans le tableau

x_i	N_i
1	5
3	20
8	9
10	11

L'effectif le plus grand est 20 donc le mode est 3.

1- Cas de même amplitude :



$$Mo = L1 + \frac{d1}{d1 + d2}\,(L_2 - L_1)$$

L_1 : limite inférieure de la classe modale.

L_2 : limite supérieure de la classe modale.

d_1 : différence des effectifs entre la classe modale et la classe précédente.

d_2 : différence des effectifs entre la classe modale

NB : *Lorsque les classes adjacentes à la classe modale ont des densité de fréquences égales ; le Mode coincide avec le centre dela classe modale ($d_1=d_2$) .*

Si les classes sont de même amplitude la classe modale est celle qui a le plus grand effectif (fréquence).

x_i	N_i
[0-5[	2
[5-10[	7
[10-15[	18
[15-20[	3

a_i = *borne sup - borne Inf , ou (Ei)*

$a_i = L_2-L_1$, $d_1=n_i-n_{i-1}$, $d_2=n_i-n_{i+1}$,

L'effectif le plus grand est 18 le classe modale [10-15[.

$$Mo = L1 + ai\frac{d1}{d1+d2} \quad ; \quad Mo = 10 + 5\frac{18-7}{(18-7)+(18-3)} = 12.11$$

2- Cas des différentes amplitudes

Si les *amplitudes* sont inégales pour calculer le mode, en remplaçant les effectif n_i par les amplitudes corrigés $Hi = \frac{ni}{ai}$

x_i	N_i	a_i	H_i
[0-5[	2	5	0.4
[5-10[	7	5	1.4
[10-20[	11	10	1.1
[20-40[	3	20	0.15

$$Mo = L1 + ai\frac{d1}{d1+d2} \qquad Mo = 5 + 5\frac{\mathbf{1.4}-0.4}{(1.4-0.4)+(1.4-1.1)} = 8.84$$

V.5. La médiane :

La *médiane* correspond au milieu de la distribution, La médiane est la valeur pour laquelle il y a autant d'individus à gauche qu'à droite dans l'échantillon. Pour déterminer la médiane d'un échantillon ou d'une population :

1- On classe les individus par ordre croissant ;

2-On prend celui du milieu.

Il existe quelques cas particuliers pour lesquels une formule est de mise.

Si une série compte *n* données, alors le rang de la donnée médiane sera :

Rang de la médiane ($=\frac{n+1}{2}$)^{e} donnée. Deux cas peuvent survenir :

Si ***n*** est un nombre *impair*, alors $\frac{n+1}{2}$ sera un nombre entier et on pourra aller chercher directement la médiane.

Si ***n*** est un nombre *pair*, alors $\frac{n+1}{2}$ sera un nombre décimal. Dans ce cas, la valeur de la médiane entre les deux rangs adjacents ($a=\frac{n}{2}$, $b=\frac{n}{2}+1$) on détermine la médiane en faisant la moyenne des données centrales de la distribution .

$$\text{médiane} = \frac{a + b}{2}$$

Exemple :

Soit un échantillon de 9 personnes dont le poids est :

45 – 68 – 89 – 74 – 62 – 56 – 49 – 52 – 63 kg

Classés par ordre croissant (pour des séries impair la valeur de la médiane se trouve au rang $Me=\frac{n+1}{2}$

$$\underbrace{45 - 49 - 52 - 56}_{4} - \underbrace{62}_{\text{médiane}} - \underbrace{63 - 68 - 74 - 89}_{4} \quad \text{kg}$$

Si le nombre d'individus est pair, on prend la moyenne entre les deux valeurs centrales :

$$\underbrace{45 - 49 - 52 - 55 - 56}_{5} - \underbrace{62 - 63 - 68 - 74 - 89}_{5}$$

$$\text{médiane} = \frac{56 + 62}{2} = 59 \quad \text{kg}$$

Lorsqu'on obtient un numéro demi entier (ex :(10+1)/2= 5,5), on calcule la moyenne des deux valeurs adjacentes En règle générale, la valeur de la médiane entre les deux rangs adjacentes.

Calcul de la médiane pour les grands échantillons répartis en classes

- Déterminez le numéro d'ordre de la médiane.
- Déterminez dans quelle classe elle se situe à l'aide du tableau des *nombres cumulés* (total des individus de cette classe et des précédentes).
- Rangez par ordre croissant les éléments (individus) de cette classe.
- Sélectionnez l'élément (individu) correspondant au numéro choisi.

Exemple :

Soient les pourcentages obtenus par 49 élèves à un examen, rangés par classes de 10 pourcents de large :

Classe	nombre	nombre cumulé
1-10	2	2
11-20	4	6
21-30	5	11
31-40	8	19
41-50	7	26
51-60	9	35
61-70	6	41
71-80	6	47
81-90	2	49

49 individus la médiane porte le n°25 $Me = \frac{n+1}{2}$

$\left[\frac{49+1}{2} = 25\right] \Longrightarrow$ dans la classe 41-50

Car, d'après le tableau des nombres cumulés, cette classe contient les individus portant les numéros d'ordre 20 à 26.

Examinons le contenu de cette classe :

46 – 42 – 45 – 44 – 50 – 43 – 49
Rangeons-les par ordre croissant :

42 – 43 – 44 – 45 – 46 – 49 – 50

Il y a 19 individus dans les classes précédentes (nombre cumulé).

Le premier de cette classe porte le n°20 et nous devons choisir le 25e.

Numéro :	20	21	22	23	24	25	26
Valeur :	42	43	44	45	46	49	50

La médiane vaut donc 49.

Pour obtenir une valeur plus précise de la médiane, on procède à *une interpolation linéaire*.la valeur de la médiane peut être lue sur le graphique ou calculée analytiquement.

De manière générale, si *a* et *b* sont les bornes de la classe contenant la médiane, F(a) et F(b) les valeurs de la fréquence cumulée croissante en *a* et *b*, alors

$$Me = a + (b - a)\frac{0.5 - F(a)}{F(b) - F(a)}$$

Si L_1 et L_2 sont les bornes de la classe contenant la médiane, F_1 et F_2 les valeurs de l'effectif croissante en F_1 , F_2 $P_2=N/2$ alors

$$Me = L_1 + (L_2 - L_1)\frac{P_2 - F_1}{F_2 - F_1}$$

Exemple :

Dans une classe les tailles des étudiants en cm sont les suivants :150-153-158-161-164-165-166-168-169-163-170-173-176-178-177-180-182-185-187-188.Calculez la médiane de cette série.

Solution :

Modalité (X) en cm	Effectifs (Ni)	ECC (E cumulé)	Fréquences (Fi)	FCC (F cumulé)
[150-160[	3	3	3/20=0.15	0.15
[160-170[	8	11	8/20=0.4	0.55
[170-180[	5	16	5/20=0.25	0.80
[180-190[	4	20	4/20=0.20	1
Total	20		1	

$Me = a + (b - a)\frac{0.5-F(a)}{F(b)-F(a)}$,

$Me = 160 + (170 - 160)\frac{0.5-0.15}{0.55-0.15}$, $Me = 168.75$

$Me = L_1 + (L_2 - L_1)\frac{P_2-F_1}{F_2-F_1}$ $P_2=N/2 =20/2=10$

$Me = 160 + (170 - 160)\frac{10-3}{11-3}$, $Me = 168.75$

V.6. Quartiles :

Les *quartiles* permettent de séparer une série statistique en quatre groupes de même effectif (à une unité près).

Un quart des valeurs sont inférieures au premier quartile Q1.

Un quart des valeurs sont supérieures au troisième quartile Q3.

Comment déterminer les quartiles Q1 et Q3 d'une série de N valeurs ?

On calcule la quantité ¼ de N = 1/4×N =N:4.

On calcule la quantité ¾ de N = 3/4×N = 3×N : 4.

Deux cas sont possibles : Soit le résultat est entier (la division tombe juste), soit non.

Cas n°1: le résultat est entier (la division tombe juste)

- On vérifie que les valeurs sont rangées par ordre croissant.

- Q1 est la $n^{ème}$ valeur où n = N:4

- Q3 est le $n'^{ème}$ valeur où l'entier n' = ¾ de N = 3/4×N = 3×N : 4

Exemple :

Prenons les valeurs rangées dans l'ordre croissant :

1-3-3-3-5-5-**6**-7-7-8-8-8-9-9-10-10-10-10-11-11-12-**13**-13-13-14-15-16-19

Il y a N = 28 valeurs, qui est divisible par 4 car 28:4=7 qui est entier

n=N:4 = 7 donc Q1 = la $7^{ème}$ valeur de la série rangée dans l'ordre croissant= **6**.

et n' = 3N:4 = 21 donc Q3 = la $21^{ème}$ valeur de la série rangée dans l'ordre croissant= **13**.

Cas n°2: le résultat n'est pas entier

- On vérifie que les valeurs sont rangées par ordre croissant ;

- On arrondit le décimal N:4 à l'entier supérieur : l'entier n ; Q_1 est la $n^{ème}$ valeur ;

- On arrondit le décimal ¾ de N=3/4×N=3N:4 à l'entier supérieur: l'entier n';Q_3 est la $n'^{ème}$ valeur.

Exemple :

Prenons les valeurs rangées dans l'ordre croissant :

3-5-5-6-7-**8**-8-9-9-10-10-10-10-11-11-12-13-**13**-13-14-15-16-19

Il y a N = 23 valeurs ;

N:4 = 5,75 donc Q1 est la 6ème valeur de la série rangée dans l'ordre croissant donc **Q1= 8**

3N:4 = 17,25 donc Q3 est la 18 ème valeur de la série rangée dans l'ordre croissant donc **Q3= 13**

Comment interpréter des quartiles donnés ?

Si on connait les quartiles Q1 et Q3 d'une série, que peut-on en déduire ?

Au moins un quart (25%) des valeurs sont inférieures ou égales à Q1.

Au moins trois quarts (75%) des valeurs sont inférieurs ou égales à Q3.

Exemple :

Dans une classe, les notes présentent un premier quartile Q1 égal à 10 et un troisième quartile égal à 14. On peut dire que :

Au moins un quart des élèves à une note inférieure ou égale à 10.

Environ un quart des élèves a moins de 10, (et environ trois quarts des élèves ont plus de 10)

Au moins trois quarts des élèves a une note inférieure ou égale à 14.

Environ trois quarts des élèves a moins de 14, (et environ un quart des élèves ont plus 14)

L'intervalle interquartile est l'intervalle] 10 ; 14[.

Environ la moitié des élèves a une note entre 10 et 14Environ la moitié des valeurs se trouvent dans l'intervalle interquartile [Q1 ; Q3].

La méthode graphique est en particulier la plus employée

Pour calculer la valeur des quartiles on fait une *intrpolation linéaire* suivante : pour k=1,2,3**.**

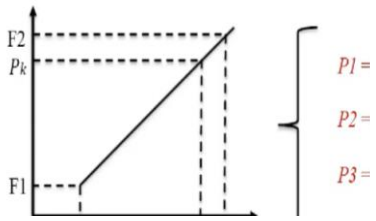
$$Q_k = L_1 + (L_2 - L_1)\frac{P_k - F_1}{F_2 - F_1}$$

L_1 : limite inférieure de la classe quartile.

L_2 : limite supérieure de la classe quartile.

F_1 : effectifs cumulé croissant de la classe précédente.

d_2 : effectifs cumulé croissant de la classe quartile.



$P_k = \frac{n}{4}$ $pour\ k = 1$; $P_k = \frac{n}{2}$ $pour\ k = 2$, $P_k = \frac{3n}{4}$ $pour\ k = 3$

Exemple :

Considérons la série statistique suivante :

classe		Effectif	ECC
]0	5]	15	15
]5	10]	62	77
]10	15]	94	171
]15	20]	90	261
]20	25]	86	347
]25	30]	58	405
]30	35]	28	433
]35	40]	11	444
]40	45]	4	448
]45	50]	2	450

Pour Q_1,on détermine la 1er classe quartile :

La classe [10,15[est la première classe dont l'effectif cumulé croissant est supérieur ou égal à 25% de l'effectif total, soit 112,5.

(P_1=N/4=450/4=112,5).

$$Q_k = L1 + \frac{P_k - F_1}{F_2 - F_1}(L_2 - L_1)$$

$$D'ou : \ Q_1 = 10 + (15 - 10)\frac{112{,}5 - 77}{171 - 77} = 11{,}88$$

Q1 = 11,88

Pour Q_2,on détermine la 2ème classe quartile :

La classe [15,20[est la première classe dont l'effectif cumulé croissant est supérieur ou égal à 50% de l'effectif total, soit 225.

(P_2=N/2=450/2=225).

$$D'ou: Q_2 = 15 + (20 - 15)\frac{225 - 7171}{261 - 171} = 18$$

$$Q_2 = Me = 18$$

Pour Q_3,on détermine la 3ème classe quartile :

La classe [20,25[est la première classe dont l'effectif cumulé croissant est supérieur ou égal à 75% de l'effectif total, soit 337,5.

(P_3=3N/4=3*450/4=337,5)

$$D'ou: Q_3 = 20 + (25 - 20)\frac{337,5 - 261}{347 - 261} = 24,44$$

$$Q_3 = 24,44$$

Remarque :

Si en utilise les fréquences la même loi est existé avec un petit changement pour P_1, P_2 et P_3, on utilise P_1=*0.25* , P_2=*0.5* et P_3=*0.75* ,puis on replace F_1 et F_2 par les fréquences cumulé au lieu d'effectif cumule.

Exemple :

Considérons la série statistique de l'exemple precedent :

Modalité (X) en cm	Effectifs (Ni)	Fréquences (Fi)	FCC (F cumulé)
[150-160[	3	3/20=0.15	0.15
[160-170[	8	8/20=0.4	0.55
[170-180[	5	5/20=0.25	0.80
[180-190[	4	4/20=0.20	1
Total	20	1	

Pour Q1,on détermine la 1er classe quartile :

La classe [160,170[est la première classe dont la fréquence cumulée croissant est supérieur ou égal à 25% de la fréquence total, soit 0,25.

$$Q_1 = a + (b - a)\frac{0.25 - F(a)}{F(b) - F(a)}$$

$$Q_1 = 160 + (170 - 160)\frac{0.25 - 0,15}{0,55 - 0,15}$$

$$Q_1 = 162,5$$

Pour Q2,on détermine la 2ème classe quartile :

La classe [160,170[est la première classe dont la fréquence cumulée croissant est supérieur ou égal à 50% de la fréquence total, soit 0,5.

$$Q_2 = a + (b - a)\frac{0.5 - F(a)}{F(b) - F(a)}$$

$$Q_2 = 160 + (170 - 160)\frac{0.5 - 0{,}15}{0{,}55 - 0{,}15}$$

$$Q_2 = 168{,}75$$

Pour Q_3,on détermine la 3ème classe quartile :

La classe [170,180[est la première classe dont la fréquence cumulée croissant est supérieur ou égal à 75% de la fréquence total, soit 0,75.

$$Q_3 = a + (b - a)\frac{0.5 - F(a)}{F(b) - F(a)}$$

$$Q_3 = 170 + (170 - 180)\frac{0.75 - 0{,}55}{0{,}80 - 0{,}55}$$

$$Q_3 = 178$$

V.7. Déciles :

Les *déciles* permettent de séparer une série statistique en dix groupes de même effectif (à une unité près).

Un dixième des valeurs sont inférieures au premier décile D1.

Un dixième des valeurs sont supérieures au neuvième décile D9.

Comment déterminer les déciles D1 et D9 d'une série de N valeurs ?

On calcule la quantité 1/10 de N = 1/10×N =N:10.

Deux cas sont possibles : soit le résultat est entier (la division tombe juste), soit non.

Cas n°1: le résultat est entier (la division tombe juste)

- On vérifie que les valeurs sont rangées par ordre croissant ;

- D1 est la n ème valeur où n = N:10 ;

- D9 est le n' ème valeur où l'entier n' = 9/10 de N = N ×9/10 = 9×N :10.

Exemple :

Prenons les valeurs rangées dans l'ordre croissant :

1-3-**3**-3-5-5-6-7-7-8-8-8-9-9-10-10-10-10-11-11-12-12-13-13-13-13-**14**-15-16-19

Il y a N = 30 valeurs, qui est divisible par 10 car 30:10=3 qui est entier

n=N:10 = 3 donc D1 est la 3ème valeur de la série rangée dans l'ordre croissant donc **D1=3**

n' = 9N:10 = 27 donc D9 est la 27ème valeur de la série rangée dans l'ordre croissant donc **D9= 14**

Cas n°2 : le résultat n'est pas entier

- On vérifie que les valeurs sont rangées par ordre croissant ;

- On arrondit le décimal N:10 à l'entier supérieur : l'entier n ; D1 est la nème valeur ;

- On arrondit le décimal 9/10 de N = 9/10×N = 9×N :10 à l'entier supérieur : l'entier n' ;D9 est la n' ème valeur.

Exemple :

Prenons les valeurs rangées dans l'ordre croissant :

3-5-**5**-6-7-8-8-9-9-10-10-10-10-11-11-12-13-13-13-14-**15**-16-19

Il y a N = 23 valeurs;

N:10 = 2,3 donc D1 est la 3ème valeur de la série rangée dans l'ordre croissant **donc D1= 5**

9N:10 = 20,7 donc D9 est la 21 ème valeur de la série rangée dans l'ordre croissant donc **D9= 15**

Comment interpréter des déciles donnés ?

Si on connait les déciles D1 et D9 d'une série, que peut-on en déduire ?

Au moins un dixième (10%) des valeurs sont inférieures ou égales à D1.

En pratique : environ 10% des valeurs sont inférieurs à D1.

Au moins neuf dixièmes (90%) des valeurs sont inférieures ou égales à D9.

En pratique : environ 10% des valeurs sont supérieurs à D9.

VI. Paramètres de dispersion

Voici les deux séries statistiques qui représentent les notes des étudiants dans deux spécialités différentes illustrés dans les tableaux suivants :

Tableau (A)

Xi	9	10	11
Effectif (ni)	3	4	3

Tableau (B)

Xi	1	10	19
Effectif (ni)	3	4	3

On Calcule la moyenne, le mode et la médiane de ces deux séries statistiques.

(A) $\overline{X} = 10, Mo = 10, Me = 10$.

(B) $\overline{X} = 10, Mo = 10, Me = 10$.

On remarque que les deux séries statistiques port la même moyenne, le même mode et la même médiane mais les valeurs sont différents ces paramètre qui sont les paramètres de position ne pas éclaires les valeurs de ces séries, pour cela on a besoin des autres paramètres pour identifies notre séries qui sont les paramètres de dispersion tel que l'Etendue et la variance.

VI.1. L'étendue :

L'*Etendue* est simplement la différence entre la plus grande et la plus petite valeur observée.

$$e = X_{max} - X_{min},$$

Le calcul de l'*étendue* est très simple. Il donne une première idée de la dispersion des observations. C'est un indicateur très rudimentaire et il existe des indicateurs de dispersion plus élaborés.

Exemple :

Quelle est l'étendue de la série statistique suivante : 10-390-395-405-410-1000

L'étendue de cette séries est :

$e = X_{max} - X_{min}$, e=1000 – 10 = 990

Si les valeur sont classes : [150-160[,[160-170[,[170-180[,[180-190[

$e = X_{max} - X_{min}$, e=190 – 150 = 40.

VI.2. L'étendue interquartile :

Le premier quartile Q_1 est l'individu ayant 25 % de l'échantillon au-dessus, le deuxième quartile Q_2 est l'individu ayant 50 % de l'échantillon : *c'est donc la médiane* au-dessus, le troisième quartile Q_3 est l'individu ayant 75 % de l'échantillon. *L'étendue interquartile* est la différence entre le troisième et les premiers quartiles .On appellent aussi l'écart interquartile et La distance interquartile.

$$(EIQ) = Q_3 - Q_1 .$$

Exemple :

Si on a dans une série statistique le 1er quartile = 65 kg et le 3me quartile = 76 kg.
Etendue interquartile (EIQ) = Q_3-Q_1=76 – 65 = 11 kg.

VI.3. La variance :

La *variance* est la somme des carrés des ´écarts `a la moyenne divisée par le nombre d'observations :

1er cas : série non classée

$$V_x = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2$$

2ème cas : série classée

$$V_x = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i(x_i - \bar{x})^2$$

Dans le cas d'une variable statistique continue, x_i représente le centre de la i ème classe. La variance est donc toujours positive ou nulle. Les formules ci-dessus imposent de calc̄uler les différences $(x_i-x)^2$ ce qui est assez fastidieux. On peut éviter cet inconvénient en utilisant le théorème de Koenig.

Autre expression de la variance : Théorème de KOENIG

1er cas : série non classée

$$V_x = (\frac{1}{n}\sum_{i=1}^{k} x^2{}_i) - \bar{x}^2$$

2ème cas : série classée

$$V_x = (\frac{1}{n}\sum_{i=1}^{k} n_i x^2{}_i) - \bar{x}^2 \quad = (\sum_{i=1}^{k} f_i x^2{}_i) - \bar{x}^2$$

Exemple :

Soit la série statistique :

2, 3, 4, 4, 5, 6, 7, 9 de taille 8, N=8

On a $\overline{X} = \frac{1}{N}\sum_{i=1}^{k} x_i$

$= (2 + 3 + 4 + 4 + 5 + 6 + 7 + 9)/ 8 = 5$

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

$= 1/ 8 [(2-5)^2 + (3-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2 + (9-5)^2]$

$= 1/ 8 [9 + 4 + 1 + 1 + 0 + 1 + 4 + 16] = 36/ 8 = 4.$

VI.4. L'écart type :

En mathématiques, l'écart type (aussi orthographié écart-type) est une mesure de la dispersion des valeurs d'un échantillon statistique ou d'une distribution de probabilité. Il est défini comme la racine carrée de la variance.

L'*écart type* sert à mesurer la dispersion d'un ensemble de données. Plus il est faible, plus les valeurs sont regroupées autour de la moyenne. Par exemple pour la répartition des notes d'une classe, plus l'écart type est faible, plus la classe est homogène. À l'inverse, s'il est plus important, les notes sont moins resserrées. Dans le cas d'une notation de 0 à 20, l'écart type minimal est 0 (notes toutes identiques), et peut valoir jusqu'à 10 si la moitié de la classe a 0/20 et l'autre moitié 20/20.

$$\sigma = \sqrt{V_x}\ ,\ \sigma = \sqrt{\frac{1}{n}\sum\left(X - \overline{X}\right)^2}\ ,\ \sigma = \sqrt{\sigma^2}\ \ ,\ S = \sqrt{S^2}$$

Exemple :

La taille en cm de 20 nouveau-nés dans une maternité un jour donné.

Valeurs (x_i)	48.5	48.7	49	49.2	49.5	50	50.5	51.3	52
Effectifs (n_i)	2	4	3	1	2	2	3	1	2
Fréquences (f_i)	2/20	4/20	3/20	1/20	2/20	2/20	3/20	1/20	2/20

Calculer la variance et écart-type

$$\overline{X} = \frac{1}{\mathrm{N}} \sum_{i=1}^{k} n_i x_i = \frac{993.8}{20} = 49.69cm$$

$$V_x = \frac{1}{\mathrm{n}} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 \Rightarrow V = 1.1959$$

$$\sigma = \sqrt{V_x} \Rightarrow \sqrt{1.1959}$$

$$\sigma = 1.09357$$

Autre expression

$$V = \frac{1}{\mathrm{N}} \sum_{i=1}^{k} n_i x^2{}_i - (\bar{x})^2 \Rightarrow V = \frac{1}{20}.49405.84 - 49.69^2$$

$$\Rightarrow V = 1.1959$$

$$\sigma = \sqrt{V_x} \Rightarrow \sqrt{1.1959}$$

$$\sigma = 1.09357$$

VI.5. Le coefficient de variation

Le coefficient de variation est défini comme le rapport entre l'écart-type et la moyenne :

$$c_v = \frac{\sigma}{\mu}$$

Comparaison avec l'écart type :

Avantages

L'écart-type seul ne permet le plus souvent pas de juger de la dispersion des valeurs autour de la moyenne. Si par exemple une distribution a une moyenne de 10 et un écart-type de 1 (CV de 10 %), elle sera beaucoup plus dispersée qu'une distribution de moyenne 1000 et d'écart-type 10 (CV de 1 %).

Ce nombre est sans unité, c'est une des raisons pour lesquelles il est parfois préféré à l'écart type. En effet, pour comparer deux séries de données d'unités différentes, l'utilisation du coefficient de variation est plus judicieuse.

Inconvénients

Quand la moyenne est proche de zéro, le coefficient de variation va tendre vers l'infini et sera par conséquent très sensible aux légères variations de la moyenne.

Contrairement à l'écart type, le coefficient de variation ne peut être utilisé directement pour construire un intervalle de confiance autour de la moyenne.

Exemple :

Même exemple précédente :

$\overline{X} = 49.69cm$

$\sigma = 1.09357$

Cv=1.09357/ 49.69=022 donc 22%.

VII. Présentation graphique des données

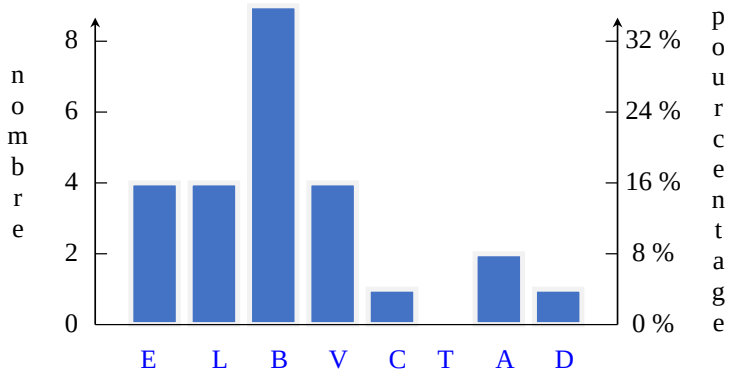
On distingue les méthodes de représentation d'une variable statistique en fonction de la nature de cette variable (qualitative ou quantitative). Les représentations recommandées et les plus fréquentes sont les tableaux et les diagrammes (graphe). Le graphique est un support visuel qui permet : La synthèse : visualiser d'un seul coup d'œil les principales caractéristiques (mais on perd une quantité d'informations).

VII.1. Distribution à caractère qualitatif

A partir de l'observation d'une variable qualitative, deux diagrammes permettent de représenter cette variable : le diagramme en bandes (dit tuyaux d'orgue) et le diagramme à secteurs angulaires (dit camembert).

VII.1.1. Diagramme en barres (ou en tuyaux d'orgue) :

C'est le *diagramme à barres*. Nous portons en abscisses les modalités, de façon arbitraire. Nous portons en ordonnées des rectangles dont la longueur est proportionnelle aux effectifs, ou aux fréquences, de chaque modalité.



Ce diagramme à barres peut aussi donner le pourcentage d'individus dans chaque catégorie.

Exemple :

Diagramme représentant la distribution d'une variable qualitative : les modalités sont placées en abscisse, formant des bases de rectangles égales et équidistantes, et les effectifs

(ou fréquences) en ordonnée, suivant une échelle arithmétique. Les surfaces des rectangles obtenus sont proportionnelles aux effectifs (ou aux fréquences).

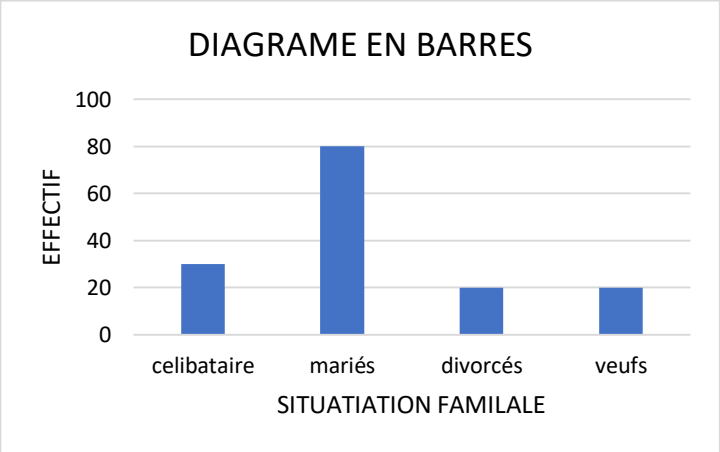
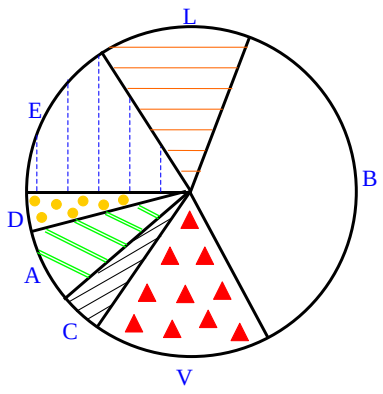


Diagramme en barres (ou en tuyaux d'orgue

VII.1.2. Diagramme circulaire (ou à secteurs circulaires, ou en camembert) :

Le diagramme *sectoriel* ou « *camembert* » ou semi-circulaires se prête très bien à la représentation des pourcentages. On dessine un disque découpé en secteurs ou « morceaux de tarte ». L'angle au centre de chaque secteur est proportionnel au pourcentage d'individus dans la catégorie correspondante.



Le diagramme en secteur

Le degré d'un secteur est déterminé à l'aide de la règle de trois de la manière suivante :

N $\longrightarrow$ 360◦

$n_i \longrightarrow$ d_i (degré de la modalité i).

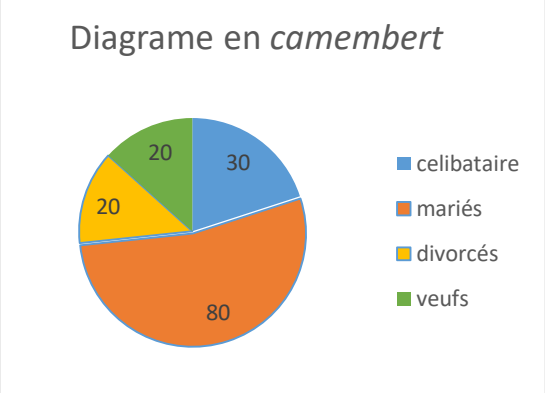
Donc, $d_i = \frac{n_i x360}{N}$ *Pour effectif*

$d_i = f_i x360$ *Pour fréquence*

$d_i = \frac{P_i x360}{100}$ *Pour pourcentage*

Exemple :

Diagramme permettant de représenter la distribution d'une variable qualitative : les modalités sont représentées par des portions de disque proportionnelles à leur effectif, ou à leur fréquence.



L'angle d_i est proportionnel à l'effectif, où à la fréquence par exemple pour représenter 15 % :

$$d_i = 0.15x360 = 54° \textit{ Pour fréquence}$$

$$d_i = \frac{15x360}{100} = 54° \quad \textit{Pour pourcentage}$$

NB : Il existe dans littératures que le nombre de coupe ne doit pas dépassé six tranches.

VII.1.3. Distribution à caractère quantitatif discret

A partir de l'observation d'une variable quantitative discrète, deux diagrammes permettent de représenter cette variable : le diagramme en bâtons et le diagramme cumulatif (voir ci-dessous). Les valeurs sont placées en abscisse, les effectifs (ou fréquences) en ordonnée, au moyen de segments verticaux.

VII.1.4. Diagramme en bâtons

Pour l'illustration, nous prenons l'exemple de nombre d'enfants par famille).

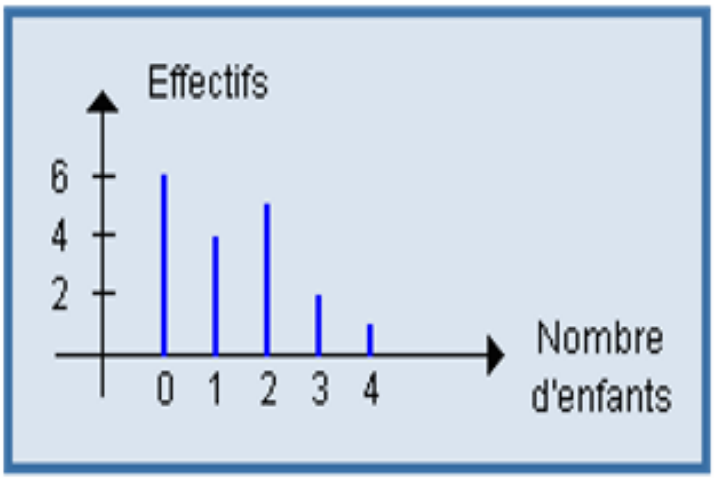


Diagramme en bâtons

Représentation sous forme de courbe et fonction de répartition

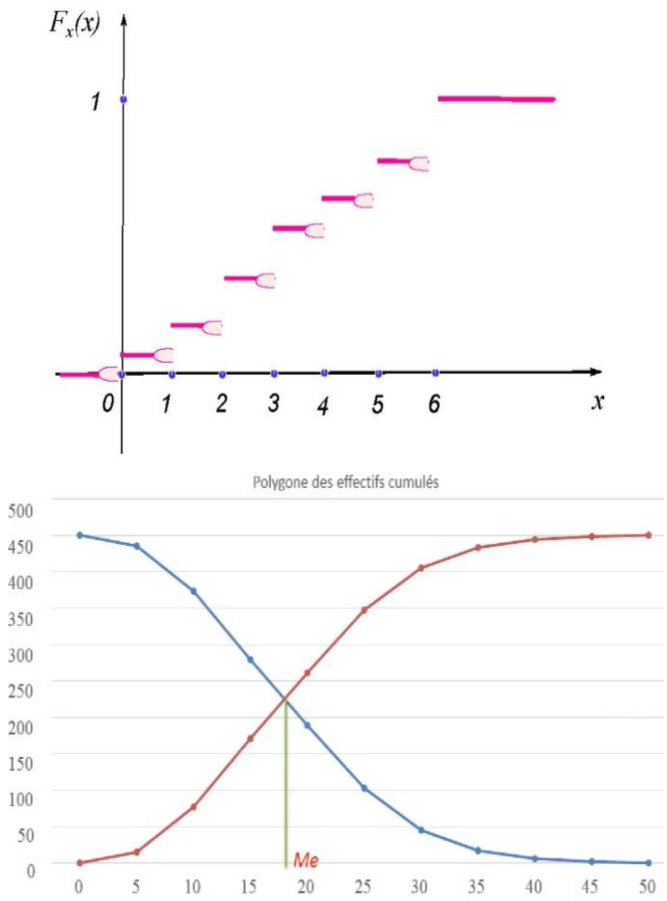
Nous avons déjà abordé les *distributions cumulées* d'une variable statistique. Nous allons dans cette partie exploiter ses *valeurs cumulées* pour introduire la notion de *la fonction de répartition*. Cette notion ne concerne que les variables quantitatives.

– *l'expression,* $F_x(x) = \begin{cases} 0, & si \quad x < x_1, \\ F_1, & si \quad x_1 \leq x < x_2, \\ F_i, & si \quad x_i \leq x < x_{i+1}, \\ 1, & si \quad x \geq x_n. \end{cases}$.

Et elle s'appelle la fonction de répartition de X.

Exemple :

Les valeurs de la fonction de répartition empirique sont les fréquences cumulées.



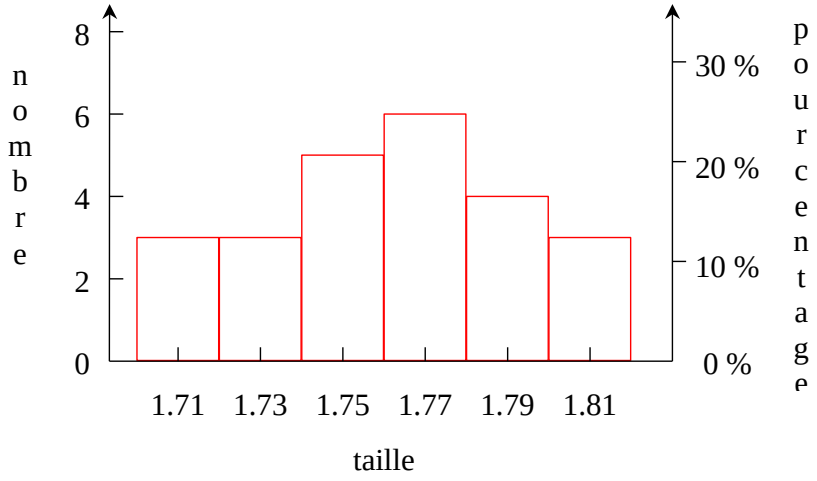
Représentation d'une variable quantitative discret par la courbe cumulative

VII.2. Représentation graphique d'un caractère continu

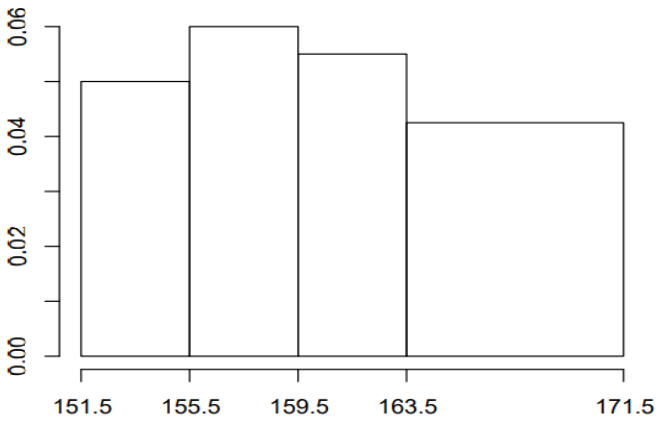
On rencontre aussi des *variables continues*. Dans ce cas, les résultats (numériques) peuvent prendre n'importe quelle valeur (éventuellement entre des limites inférieure et supérieure).

VII.2.1. Histogramme des fréquences (ou effectifs)

Nous pouvons représenter le tableau statistique par un histogramme. Nous reportons les classes sur l'axe des abscisses et, au-dessus de chacune d'elles, nous traçons un rectangle dont l'aire est proportionnelle à la fréquence fi (ou l'effectif ni) associée. Ce graphique est appelé l'histogramme des fréquences.



Les classes sont généralement repérées par leur centre, mais elles doivent être définies par leurs extrémités.

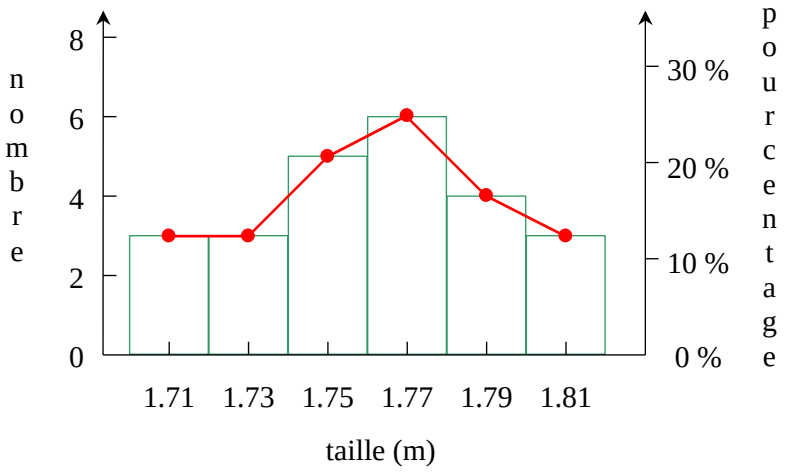


Si les deux dernières classes sont agrégées, comme dans la Figure, la surface du dernier rectangle est égale à la surface des deux derniers rectangles de l'histogramme. Dans le cas de classes de même amplitude la largeur des rectangles sont les mêmes mais ces les

classes n'ai pas le même amplitude la largeur des rectangles sont proportionnelle à largeur des classes.

VII.2.2. Polygone des fréquences ou des effectifs :

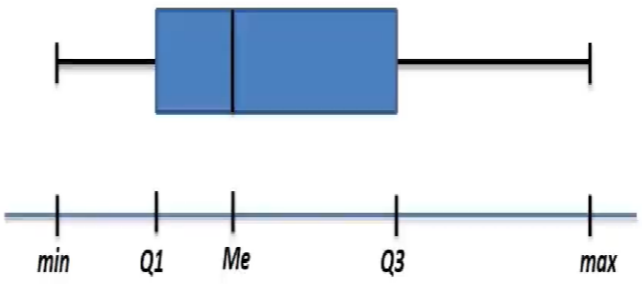
Pour obtenir ce polygone, on raccorde les sommets des barres, au centre de chaque classe, par des segments de droite.



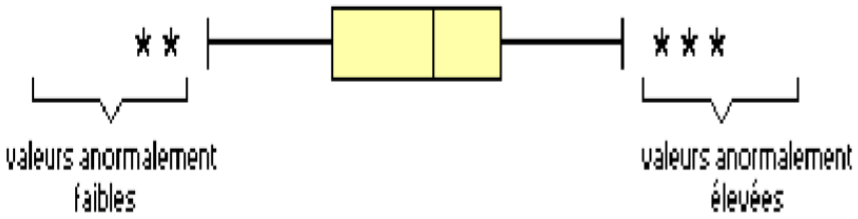
On obtient donc une série de points reliés par des segments de droite. *L'abscisse* de chaque point correspond au centre de la classe. La hauteur de chaque point (son *ordonnée*) correspond au *nombre* d'individus dans la classe (*polygone des effectifs*) ou au *pourcentage* d'individus dans la classe *(polygone des fréquences).*

VII.2.3. Diagramme en boîte (boîte à moustaches,Box-plot) :

Il s'agit d'un diagramme permettant de positionner les quartiles Q_1 , Q_2 , Q_3 , au moyen de rectangles de largeur arbitraire, prolongés par des "moustaches" de part et d'autre, de longueur au plus égale à une fois et demie $Q_3 - Q_1$.



Si la plus petite ou la plus grande valeur observée se trouvent à l'intérieur, on raccourcit les moustaches correspondantes ; si elles se trouvent à l'extérieur, on positionne à part les valeurs "aberrantes" qui dépassent des moustaches :



Ces diagrammes sont surtout utiles pour comparer rapidement l'allure générale de plusieurs distributions.

Exemple :

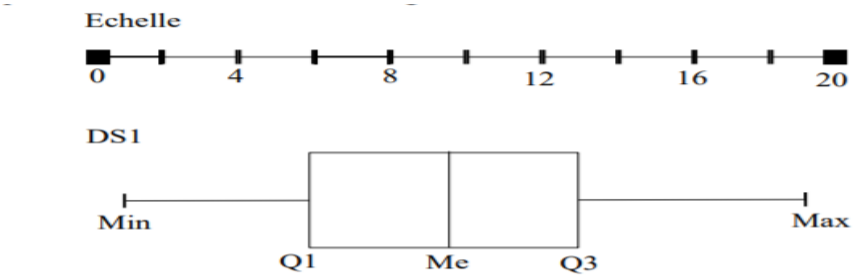
La médiane comme paramètre de position et l'intervalle interquartile comme paramètre de dispersion fournissent une bonne description d'une série statistique.

On utilise ces deux données pour construire un diagramme en boîte de la série

Soit une série de valeurs qui se résume en:

- le minimum Min = 1
- le 1er quartile Q1 = 6
- la médiane Me = 9,5
- le 3ème quartile Q3 = 13
- le maximum Max = 19

Ces 5 données permettent de construire un diagramme en boites :



VII.3. Choix de la largeur des classes :

La largeur choisie pour les classes dépendra :

- de la finesse de la représentation désirée (si on veut faire la distinction entre des individus dont la taille diffère de 5 cm, on ne va pas choisir des classes plus larges, par exemple 10 cm !)
- de la taille de l'échantillon étudié.

Pour que la représentation ait suffisamment de précision, il faut que chaque classe contienne, en général, un nombre suffisant d'individus.

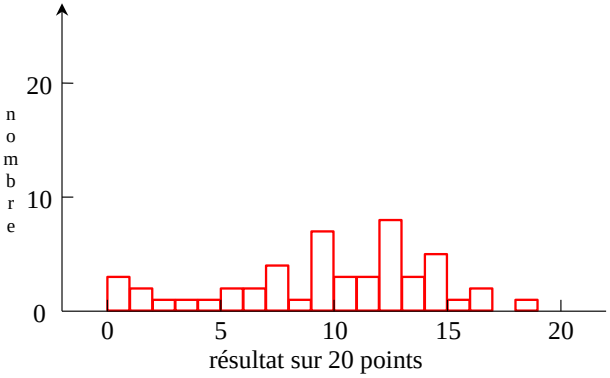
Exemple :

Les notes obtenues à un examen par 50 élèves sont données dans le tableau suivant :

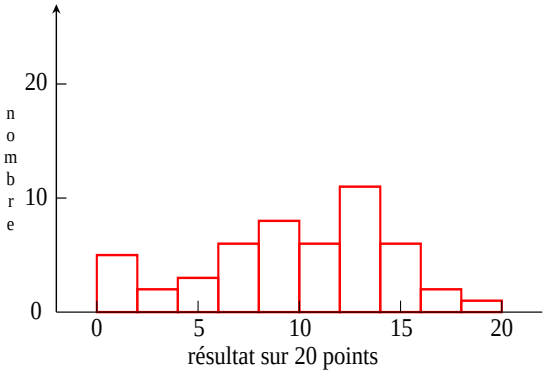
0.0	2.1	6.1	7.8	9.5	10.4	12.1	12.8	13.9	14.8
0.0	3.2	6.2	8.2	9.6	10.5	12.4	12.8	14.2	15.5
0.5	4.5	7.2	9.1	9.9	11.1	12.5	12.9	14.6	16.1
1.2	5.3	7.2	9.1	9.9	11.8	12.6	13.0	14.7	16.8
1.7	5.3	7.4	9.5	10.1	11.9	12.6	13.7	14.7	18.2

L'allure de l'histogramme change en fonction de la largeur choisie pour les classes :

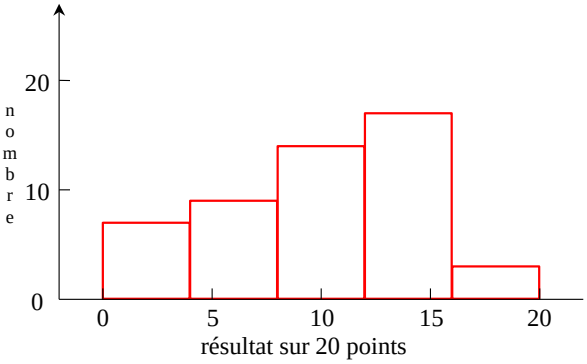
Classes de 1 cm:



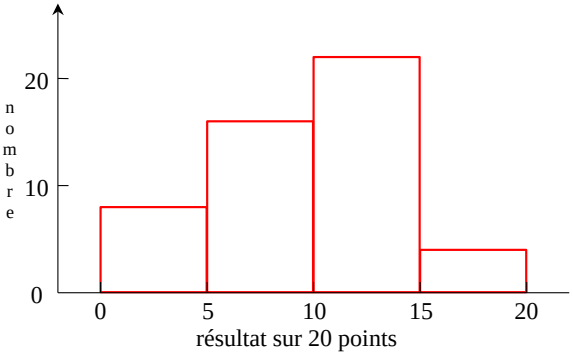
Classes de 2 cm:



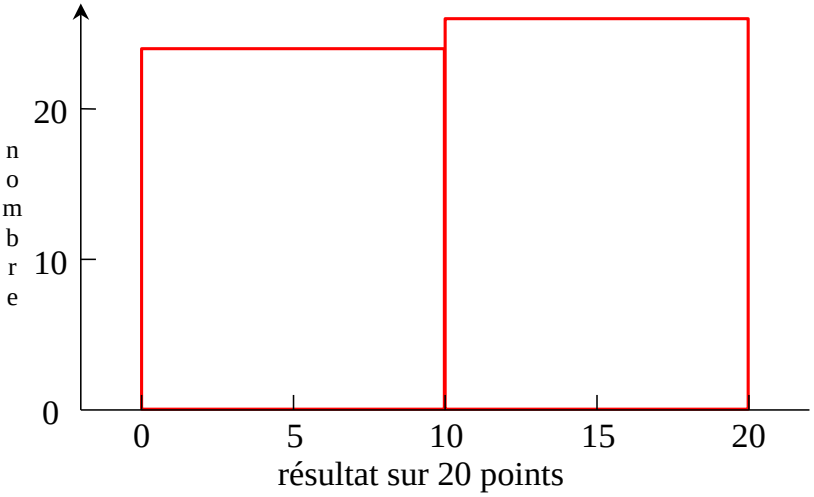
Classes de 4 cm:



Classes de 5 cm:



Classes de 10 cm:



VII.4. Nombre de classes :

En combien de classes partageons-nous les valeurs ? La réponse n'est pas unique. Soit N l'électif total. Nous pouvons considérer dans ce cours trois réponses à titre d'exemple

1-Une réponse : $\sqrt{N}$, $[\sqrt{N}]$ (partie entière) ou $[\sqrt{N}]+1$. Donc, le nombre de classes : $k = \sim\sqrt{N}$.

2-Une réponse : la formule de Sturge : $k = 1 + 3.3 \log_{10}(N)$.

3-Une réponse : la formule de Yule $k = 2.5 \sqrt[4]{N}$.

Nous mentionnons que les deux formules sont presque pareils si $N << 200$.

Exemple :

Considérons 30 valeurs entre 56.5 cm et 97.8 cm. Dans ce cas,

-Soit la formule de $[\sqrt{N}]$ (partie entière) $k = \sqrt{30}$ et on prend $k = \sim 6$.

- Soit la formule de Sturge $k = 1 + 3.3 \log_{10}(30) = \sim 6$,

- Soit la formule de Yule $k = 2.5 \sqrt[4]{30} = \sim 6$.

Conclusion de Représentations géométriques des distributions :

- Pour une variable qualitative : diagramme circulaire ("camembert").
- Pour une variable ordinale ou quantitative discrète : diagramme ou graphique en bâtons.
- Pour une variable ordinale ou quantitative continue : histogramme et courbe des fréquences cumulées.
- Pour deux variables quantitatives ou ordinales : *nuage de points*.

Etude de corrélation et Régression

I. Étude d'une variable statistique à deux dimensions

Dans la partie précédente, nous avons présenté les méthodes qui permettent de résumer et représenter les informations relatives à une variable. Un même individu peut être étudié à l'aide de plusieurs caractères (ou variables). Par exemple, la croissance d'un enfant en regardant son poids et sa taille. Dans la suite, nous introduisons l'étude globale des relations entre deux variables.

Jusqu'à présent, nous nous sommes intéressés à des questions du type :

- Quelle est la taille moyenne des garçons âgés d'une vingtaine d'années ?
- Quelle est la probabilité pour qu'un médicament soit efficace ?
- Quelle fraction des barres métalliques produites par une usine sera-t-elle rejetée par le client ?
- Le poids moyen des fruits produits dans une firme est-il supérieur à 200 grammes ?

Dans toutes ces questions, nous étudions le comportement statistique d'une seule variable : taille, efficacité du médicament, longueur des barres, poids des fruits.

Il existe cependant toute une gamme de problèmes statistiques où l'on s'intéresse à la *relation entre plusieurs variables*.

Exemples :

- Les individus les plus grands sont-ils les plus lourds ?
- Le revenu d'une famille a-t-il une influence sur les résultats scolaires des enfants ?
- Y a-t-il une relation entre le tabagisme et les cancers du poumon ?
- Le rendement en céréales dépend-il de la quantité d'engrais utilisée ?
- La productivité d'une entreprise est-elle liée au salaire des ouvriers ou employés ?

Dans ces questions, nous désirons savoir si le comportement d'une variable est influencé par la valeur d'une autre variable :

Taille ⟷ Poids Revenu ⟷ Résultats

Tabagisme ⟷ Cancer Rendement ⟷ Engrais

La relation peut être *causale* ou non

Pour étudier les *relations* ou *corrélations* entre deux variables statistiques, on peut les porter sur un graphique.

Exemple :

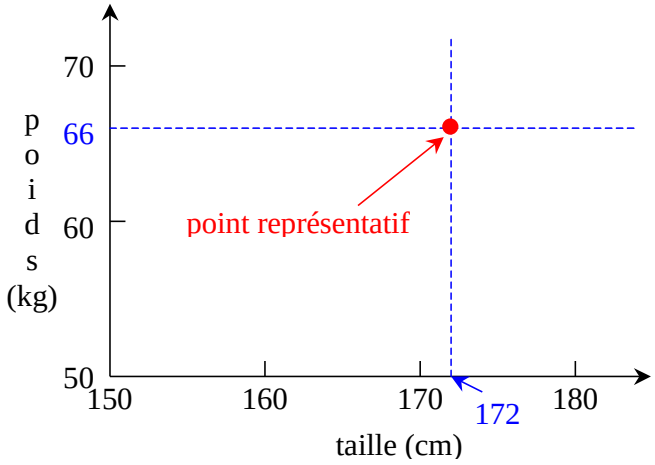
Relation entre la taille et le poids des individus pour chaque individu de l'échantillon, on porte sur un graphique :

• Sa taille en abscisse (*l'abscisse d'un point correspond à sa projection sur l'axe horizontal*)

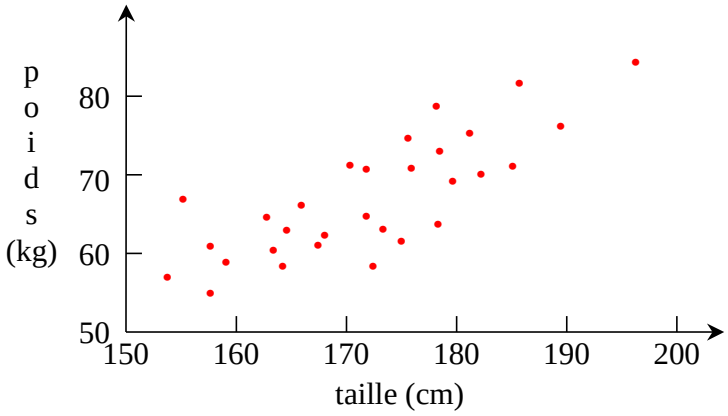
• Son poids en ordonnée (*l'ordonnée d'un point correspond à sa projection sur l'axe vertical*)

Chaque individu est donc, dans ce graphique, *représenté* par un point *(point représentatif)*

soit un individu mesurant 172 cm et pesant 66 kg.



Dans le graphe, il y aura donc autant de points qu'il y a d'individus dans l'échantillon.



Relation entre le poids et la taille dans un échantillon de 30 individus.

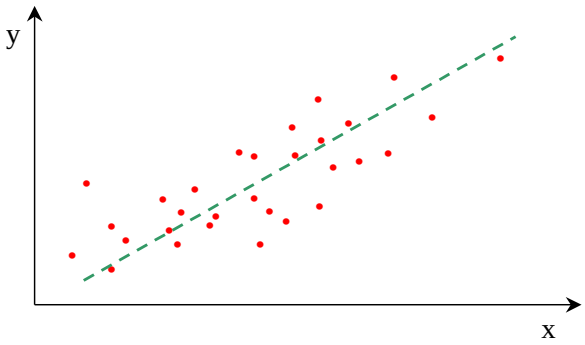
On peut (par la pensée ou réellement) tracer une droite qui passe au mieux par ces points (au milieu du "nuage" de points).

Si cette droite "monte", on dira qu'il y a *corrélation positive* entre les deux variables.

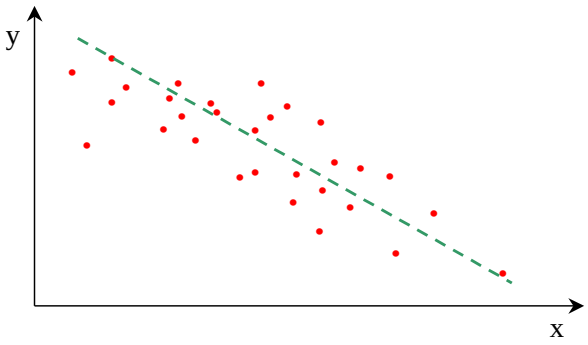
Si elle "descend", c'est une *corrélation négative*.

Si elle est "horizontale", ou si on ne peut pas décider, c'est qu'il y a *absence de corrélation*.

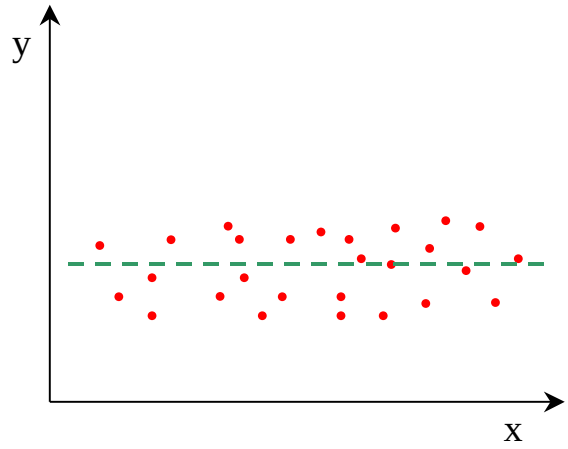
Corrélation positive:

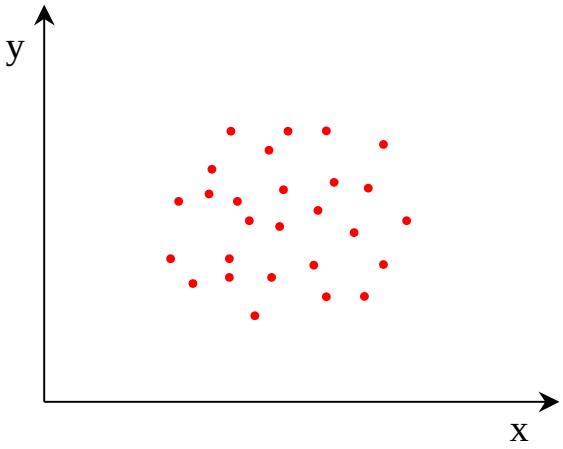


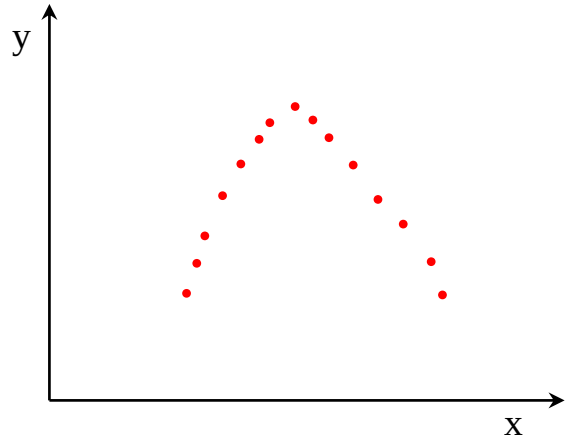
Corrélation négative:

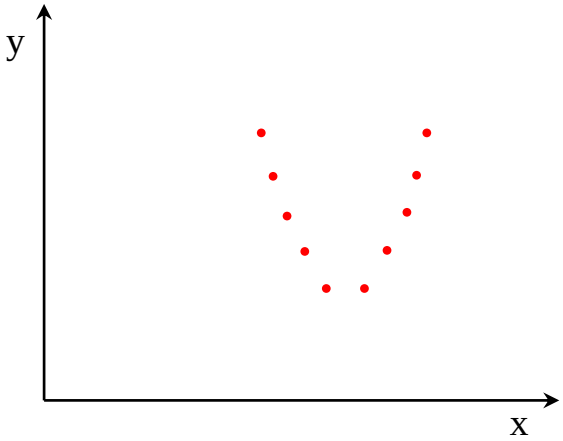


Absence de corrélation :



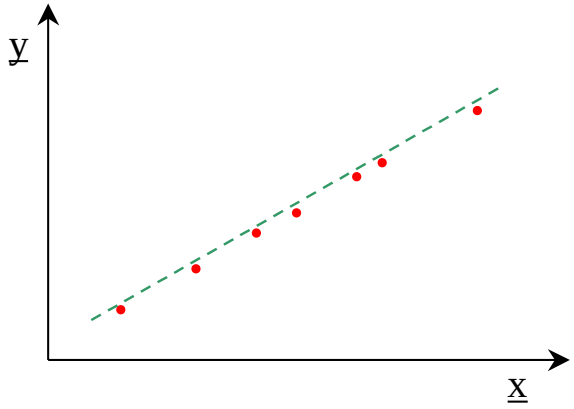




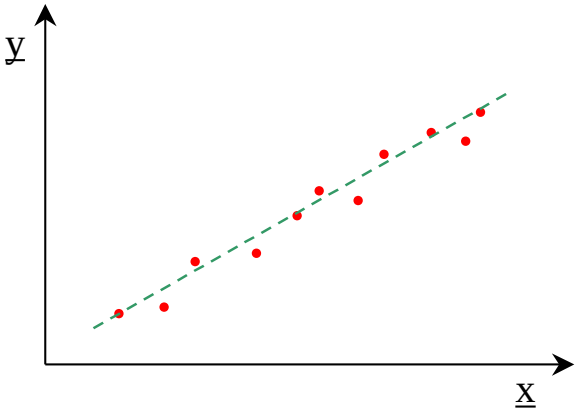


La *qualité* de la corrélation entre deux variables peut se mesure par la *dispersion* des points autour de la relation moyenne.

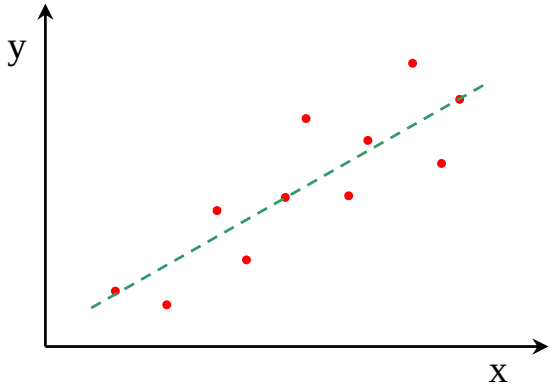
Corrélation parfaite :



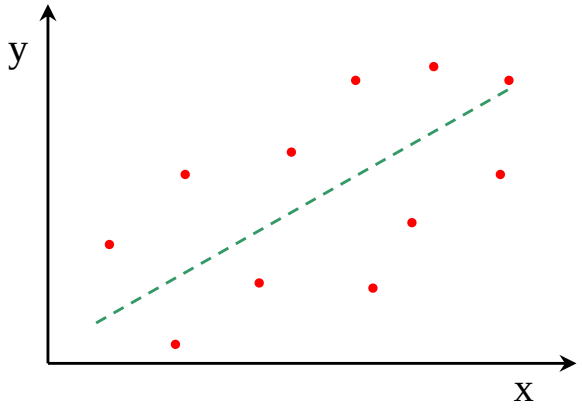
Corrélation Bonne (forte):



Corrélation (moyenne) :



Corrélation Mauvaise (faible):

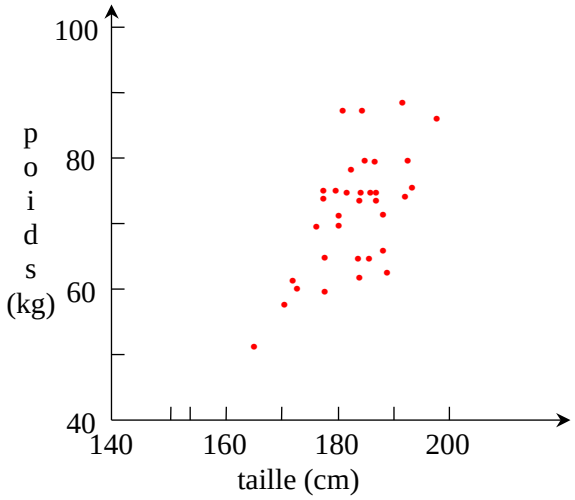


Exemple :

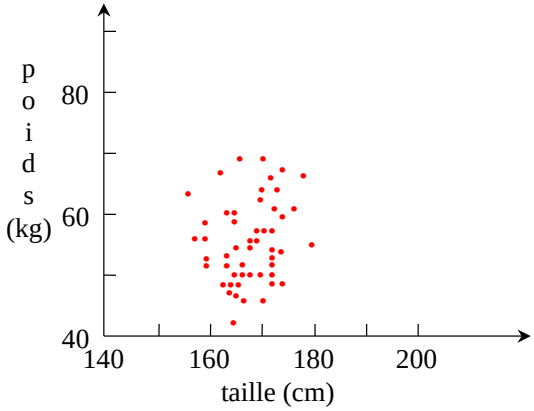
Corrélation entre le poids et la taille pour les garçons.

On constate une augmentation du poids avec la taille (corrélation positive) : les garçons les plus grands sont généralement les plus lourds.

Mais la dispersion des points est assez grande : la corrélation est assez faible.



Corrélation entre le poids et la taille pour les filles.



On ne constate pas de relation entre le poids et la taille (*absence de corrélation*): le poids des filles est indépendant de leur taille.(Les filles les plus grandes sont donc les plus minces)

II. La Covariance :

Si deux caractères quantitatifs ***x*** et ***y*** sont mesurés sur ***n*** individus, on peut considérer l'échantillon bidimensionnel comme un nuage de ***n*** points dans R^2 . Différentes caractéristiques statistiques permettent de résumer l'information contenue dans sa forme. Si $\overline{X}$ et $\overline{Y}$ désignent les moyennes empiriques des deux caractères, le point ($\overline{X}$, $\overline{Y}$) est *le centre de gravité du nuage*. Les variances empiriques S_x^2 et S_y^2 (σ_x^2 et σ_y^2) traduisent la dispersion des abscisses et des ordonnées. Pour aller plus loin dans la description, il faut calculer la covariance.

On appelle covariance de ***x*** et ***y*** , et on note σ_{xy} la quantité :

$$Cov(x,y) = \frac{1}{N}\sum (x-\bar{x}).(y-\bar{y})$$

La covariance est symétrique ($\sigma_{xy} = \sigma_{yx}$) *[Cov(x,y)= Cov(y,x)]* et bilinéaire :

Si C***ov(x,y)>0*** ***x*** et ***y*** varient dans le même sens .

Si C***ov(x,y)<0*** ***x*** et ***y*** varient dans le sens contraire.

La covariance C***ov(x,x)=*** V***(x)***

On peut utiliser les formules suivantes :

$$\sigma_{xy} = \frac{\sum xiy_i}{n_i} - \bar{x}\bar{y} \qquad Cov(x,y) = \frac{\sum n_i (x-\bar{x}).(y-\bar{y})}{N} \qquad \sigma_{xy} = \frac{\sum n_i xiy_i}{N} - \bar{x}\bar{y}$$

III. Coefficient de corrélation linéaire

Le coefficient de corrélation linéaire, ou de Pearson, noté ***r*** et établi sur ***n*** observations, est égal à *la* *covariance* entre une variable explicative ***x*** et une variable à expliquer ***y***, rapportée au produit de leurs *écarts-types*. Compris entre -1 et 1, il mesure l'intensité de la relation entre les deux variables statistiques**.**

$$r = \frac{Cov(x,y)}{\sigma(x).\sigma(y)}$$

Ce coefficient nous montrent si il y a une corrélation linéaire de type (Y=aX+b) ou pas.

Le signe de la pente *a* donné le *sens* de corrélation, mais pas sa *qualité*.

a > 0 corrélation positive

a < 0 corrélation négative

a = 0 pas de corrélation

Le coefficient de corrélation est compris entre –1 et +1.

On peut utiliser les formules suivantes :

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad r = \frac{\sum (X-\bar{X}).(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2} \times \sqrt{\sum (Y-\bar{Y})^2}},$$

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2} \times \sqrt{n\sum Y^2 - (\sum Y)^2}} \text{ et } r = \frac{1}{n}\sum \left(\frac{X-\bar{X}}{S_X}\right)\left(\frac{Y-\bar{Y}}{Sy}\right)$$

Plus il s'éloigne de zéro, la corrélation est meilleure.

$r = +1$ corrélation positive parfaite

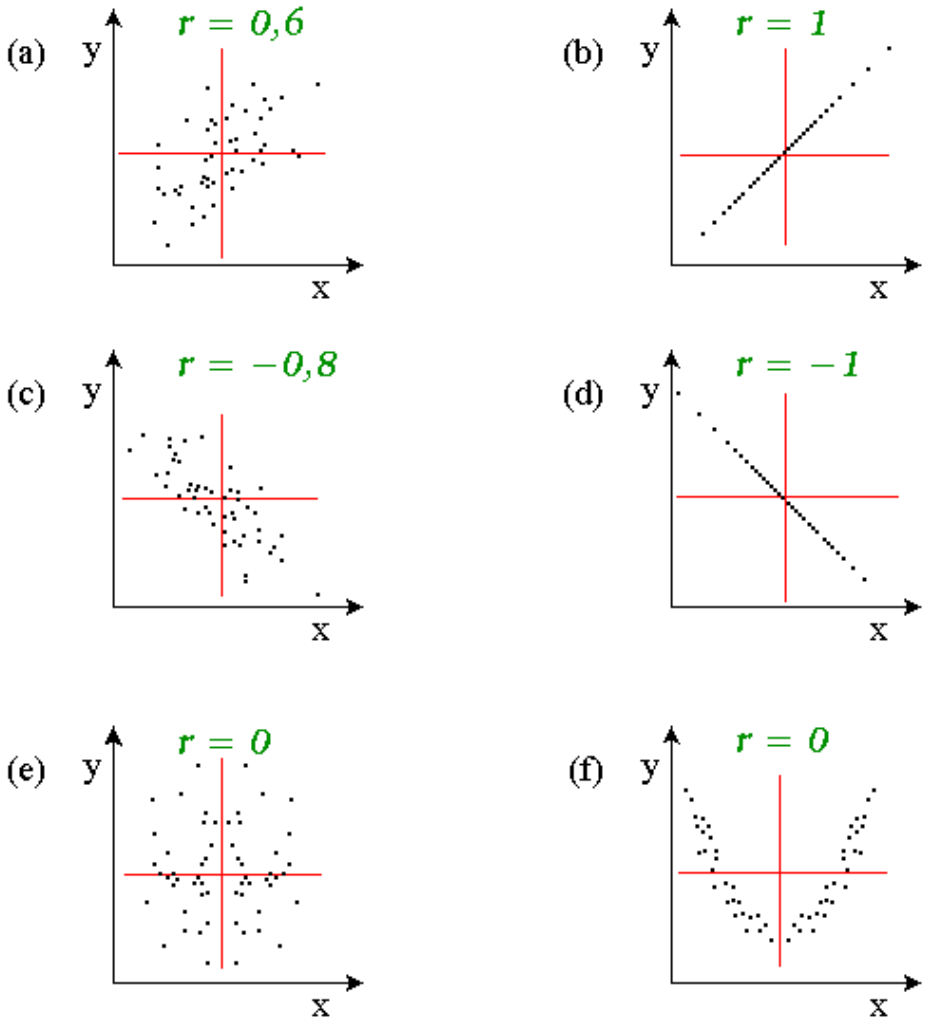
$r = -1$ corrélation négative parfaite

$r = 0$ absence totale de corrélation

$G(\overline{X}, \overline{Y})$ Centre de nuage.

Quelques exemples de corrélation

(Le coefficient de corrélation *r* est indiqué dans chaque cas)



Exemple :

Les tailles en cm et poids en kg de 10 étudiants.

Enfant	H	A	E	C	D	I	J	B	F	G
Tailles (X)	161	164	164	170	174	174	176	178	182	186
Poids (Y)	61	62	71	72	73	86	89	97	104	112

Calculer le coefficient de corrélation et commenter vos résultats.

Solution

Calcul de corrélation entre taille et poids de 10 étudiants

Enfant	Tailles (X)	Poids (Y)	X^2	Y^2	X*Y
H	161	61	25921	3721	9821
A	164	62	26896	3844	10168
E	164	71	26896	5041	11644
C	170	72	28900	5184	12240
D	174	73	30276	5329	12702
I	174	86	30276	7396	14964
J	176	89	30976	7921	15664
B	178	97	31684	9409	17266
F	182	104	33124	10816	18928
G	186	112	34596	12544	20832
Total	1729	827	299545	71205	144229

$$\bar{x} = 172.9 \ , \ \bar{y} = 82,7$$

$$V_x = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2$$

$$\sigma_x = 7,75cm \ , \ \sigma_y = 16,77kg$$

$$\sigma_{xy} = \frac{\sum xiy_i}{n_i} - \bar{x}\bar{y} \qquad \sigma_{xy} = \frac{144229}{10} - 172.9x82.7$$

$$\sigma_{xy} = 124.07$$

$$r = \frac{Cov(x,y)}{\sigma(x).\sigma(y)} = \frac{124.07}{7,75x16,77}$$

$$r = 0,95$$

Nous sommes en présence d'une corrélation positive forte, qui semble indiquer qu'il existe une relation linéaire entre la taille et le poids des étudiants.

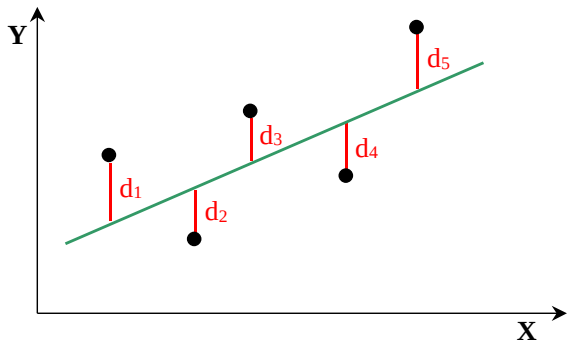
Nb : Taille en Cm ou m $\boldsymbol{Cov_{Cm} > Cov_m}$

IV. Méthode des moindres carrés :

Si on se contente de tracer à main levée la droite qui "passe au mieux" par les points représentatifs, différentes personnes vont obtenir des résultats différents.

Il existe une méthode mathématique pour déterminer la "meilleure" droite: c'est la *méthode des moindres carrés*.

Elle consiste, dans sa version la plus simple, à trouver la droite qui minimise les carrés des écarts des points représentatifs à cette droite.



Trouver la droite telle que la somme des carrés des écarts d_1, d_2,… soit minimale:

$$\sum d^2 = \text{minimum}$$

Soit : $Y = aX + b$

l'équation de la droite cherchée (*droite de régression*)

Les coefficients *a* et *b* peuvent être calculés à partir des formules suivantes:

Pente :

$$a = \frac{(X_1 - \overline{X})(Y_1 - \overline{Y}) + (X_2 - \overline{X})(Y_2 - \overline{Y}) + \cdots + (X_n - \overline{X})(Y_n - \overline{Y})}{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}$$

ou:

$$a = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

$$a = \frac{Cov(x, y)}{V(x)} \quad , \quad a' = \frac{Cov(x, y)}{V(y)}$$

Ordonnée à l'origine :

$$b = \overline{Y} - a.\overline{X}$$

Exemples :

1. Supposons un échantillon aléatoire de 4 firmes pharmaceutiques présentant les dépenses de recherche X et les profits Y suivants (en milliers de dollars):

X	**Y**
40	50
40	60
30	40
50	50

Trouvez la droite de régression et le coefficient de corrélation.

Calculons tout d'abord X et Y:

$$\overline{X} = \frac{1}{n}\sum X = \frac{1}{4}(40+40+30+50) = \frac{160}{4} = 40$$

$$\overline{Y} = \frac{1}{n}\sum Y = \frac{1}{4}(50+60+40+50) = \frac{200}{4} = 50$$

Complétons le tableau suivant:

X	Y	$X-\overline{X}$	$Y-\overline{Y}$	$(X-\overline{X})^2$	$(Y-\overline{Y})^2$	$(X-\overline{X})\,.(Y-\overline{Y})$
40	**50**	0	0	0	0	0
40	**60**	0	+10	0	+100	0
30	**40**	–10	–10	+100	+100	+100
50	**50**	+10	0	+100	0	0

On a donc:

$$\sum(X-\overline{X})^2 = 200$$
$$\sum(Y-\overline{Y})^2 = 200$$
$$\sum(X-\overline{X})(Y-\overline{Y}) = 100$$

Les coefficients de la droite de régression sont :

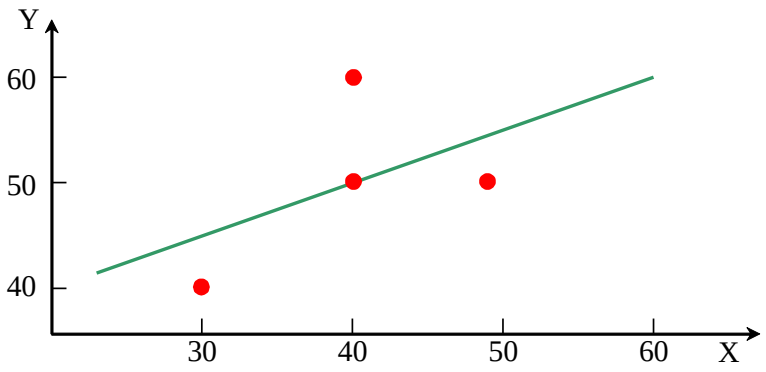
$$a = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} = \frac{100}{200} = 0{,}5$$
$$b = \overline{Y} - a.\overline{X} = 50 - 0{,}5\times 40 = 50 - 20 = 30$$

Y=aX+b ***y=0.5x+30***

Et le coefficient de corrélation :

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\times\sqrt{\sum(Y-\overline{Y})^2}} = \frac{100}{\sqrt{200}\times\sqrt{200}} = \frac{100}{200} = 0{,}5$$

La corrélation est *positive* et de *qualité moyenne*



2. La corrélation entre la taille (X) et le poids (Y) pour les garçons donne les résultats suivants :

   (a) droite de régression **Y = *a*X + *b***

   a = 0,816 b = -77,0

   (b) coefficient de corrélation

   r = 0,61

   La corrélation est donc positive, de qualité moyenne

3. De la même manière, pour les filles, on obtient :

   (a) droite de régression

   a = 0,239 b = 16,6

   (b) coefficient de corrélation

   r = 0,20

La corrélation est positive (les filles les plus grandes tendent à être les plus lourdes), mais de très mauvaise qualité (r proche de zéro).

Estimation ponctuelle et par intervalle de confiance

I .Introduction

"Estimer ne coûte presque rien Estimer incorrectement coûte cher"

Vieux proverbe chinois

Dans le domaine de la SNV comme dans de nombreux autres domaines, on a besoin de connaître certaines caractéristiques de la population ciblée par une étude. En règle générale, il est difficile voire impossible d'évaluer ces caractéristiques sur la population (effectif trop important, coût trop élevé, durée de réalisation trop longue). On est alors conduit à estimer ces caractéristiques à partir des observations sur un échantillon de plus petite taille issu de la population.

La qualité de l'estimation d'un paramètre tel que la moyenne, la variance ou la proportion d'un caractère et donc celle de l'information obtenue repose sur la prise en compte de plusieurs éléments.

- **La représentativité de l'échantillon**

  L'échantillon doit être représentatif de la population ciblée et une façon d'obtenir un échantillon représentatif est de réaliser un tirage au sort aléatoire, simple ou stratifié, parmi les N individus constituant la population.

- **La fluctuation d'échantillonnage**

  Un échantillon, même tiré au sort dans la population ciblée, n'est pas l'image exacte de la population en raison de ce que l'on appelle la fluctuation d'échantillonnage.

  Par exemple, si l'on tire au sort des échantillons dans une population contenant 20% de personnes avec des yeux bleu, on obtient des échantillons où la proportion de sujets aux yeux bleus fluctue autour de 20%. Ces fluctuations sont imprévisibles et le hasard peut conduire à des écarts (par rapport à 20%) plus ou moins importants. Cependant les grands écarts sont très peu probables et l'on pourra calculer des intervalles autour de la valeur observée dans l'échantillon tel qu'une grande proportion (ou probabilité) d'entre eux contiennent la valeur de 20%. Il s'agit des intervalles de confiance.

- **L'erreur commise sur l'estimation, ou impression**

  Elle est calculée à partir de l'intervalle de confiance et qui dépend en particulier de la taille de l'échantillon.

Dans ce chapitre nous allons apprendre à estimer, à l'aide des observations faites sur un échantillon, les paramètres suivants :

- - moyenne (μ) et variance $(\sigma^2)[ou\,écart-type(\sigma)]$ dans le cas d'une variable quantitative
- proportion (π) dans le cas d'une variable qualitative.

Nous envisagerons l'estimation ponctuelle et l'estimation par intervalle pour chacun des paramètres.

II. Procédure générale d'estimation d'un paramètre

- Soit une variable X (quantitative ou qualitative) et sa loi de distribution (ou de probabilité) pour la population ciblée telles que la loi de Gauss ou loi normale N(μ , σ), la loi binomiale B(n,p).

  On dispose des n observations de la variable X sur un échantillon (noté n-échantillon) issu de la population : x_1, x_2,.....x_i, ...x_n

NB : x_1, x_2,.....x_i, ...x_n sont les réalisations des n variables X_1, X_2,....X_i, ...X_n qui suivent la même loi de distribution que la variable X.

- On cherche à obtenir la valeur d'un paramètre θ (μ , σ ou π) de la loi de distribution de X.

  La procédure d'estimation a pour objectif de trouver la meilleure valeur approchée du paramètre θ à partir des n observations de l'échantillon.

  Cela nécessite de disposer du meilleur estimateur de ce paramètre.

  Un estimateur est une fonction mathématique reliant le n variables X_i, c'est donc une variable aléatoire.

Exemple :

L'expression de l'estimation de la moyenne est :

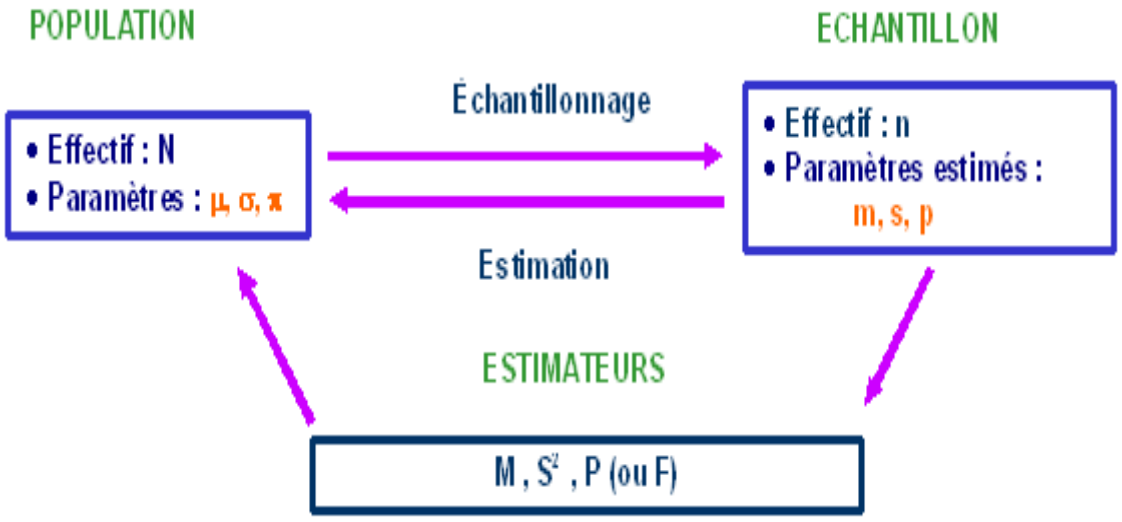
$$M = \frac{1}{n}\sum_{i=1}^{n} X_i$$

A partir des n observations xi de l'échantillon on obtiendra, en utilisant l'estimateur, une valeur du paramètre θ qui constitue l'estimation ponctuelle de ce paramètre.

L'estimation ponctuelle de la moyenne (μ) sera égale à :

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- L'estimation ponctuelle du paramètre θ étant influencée par les fluctuations d'échantillonnage, il faudra lui associée une estimation par intervalle : on calculera un intervalle de confiance (IC) qui nécessite de connaître la loi de distribution du paramètre estimé.
- La figure ci-dessous représente les étapes et les notations associées à la procédure d'estimation.



III. Estimation ponctuelle d'un paramètre

III.1.Introduction

Comme il a été décrit dans la procédure générale, l'estimation ponctuelle se fait à l'aide d'un estimateur qui est une variable aléatoire. L'estimation ponctuelle est la valeur que prend cette variable aléatoire dans l'échantillon observé.

III.2.Estimation ponctuelle d'une moyenne

Soit une population et une variable quantitative X de moyenne μ.

Si l'on disposait des valeurs x_i de X pour chacun des N individus de la population on pourrait **calculer** la valeur de la moyenne :

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

On dispose en général des observations x_i d'un échantillon issu de la population comprenant n individus. On pourra seulement **estimer** la valeur de la moyenne en utilisant l'**estimateur M** associé.

- **Estimateur** :

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Estimation ponctuelle** :

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

III.3.Estimation ponctuelle d'une proportion

Soit un échantillon de taille ***n*** et une variable aléatoire qualitative binaire décrite par une loi binomiale B(n,p).

La proportion observée ***p*** sur l'échantillon est une estimation de la probabilité π de la réalisation de l'événement étudié, nommé "succès".

L'estimateur *F* de la proportion est égal au nombre de réalisations de l'événement (a) rapporté à la taille de l'échantillon ***(n)*** soit ***F= a / n***.

III.4.Estimation ponctuelle d'une variance et d'un écart-type

Cas d'une variable quantitative

Soit une population et une variable quantitative X de moyenne μ et de variance σ^2

Si l'on disposait des valeurs x_i de X pour chacun des N individus de la population on pourrait **calculer** la valeur de la variance après avoir calculé la valeur de la moyenne et en déduire celle de l'écart-type :

$$\sigma^2=\frac{(x_1-\mu)^2+(x_2-\mu)^2+.....+(x_N-\mu)^2}{N}$$

On dispose en général des observations ***xi*** d'un échantillon issu de la population comprenant ***n*** individus. On pourra seulement **estimer** la valeur de la variance en utilisant **l'estimateur S^2** associé (couregie).

Estimateur :

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-m)^2$$

Estimation ponctuelle :

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-m)^2=\frac{(x_1-m)^2+(x_2-m)^2+.....+(x_n-m)^2}{n-1}$$

On en déduit l'estimation ponctuelle de l'écart-type σ :

$$s=\sqrt{s^2}$$

Cas d'une variable qualitative

On a vu que la variance d'une variable X décrite par une loi binomiale B(n,p) est égale à np(1- p) si l'on considère le nombre de "succès" sur l'échantillon de taille n.

Si l'on considère la proportion (ou pourcentage) de "succès", la variance de cette proportion est égale à p(1 – p) / n.

Exemple 1 :

On a constitué, à partir d'une population de sujets présentant une allergie particulière, un échantillon représentatif de 120 sujets. Le dosage d'un paramètre biologique B, spécifique de cette allergie dont la moyenne μ_B et la variance σ_{B^2} au sein de la population sont inconnues, a été effectué chez chacun des sujets de l'échantillon. Les résultats des dosages (xi), exprimés en UI, sont résumés par :

.

$$\sum x_i = 4\,200 \text{ et } \sum (x_i - m_B)^2 = 27\,045$$

où m_B est la moyenne calculée sur l'échantillon

- Estimation de la moyenne et de l'écart-type du paramètre biologique B au sein de la population.

L'estimation ponctuelle de la moyenne μ_B est égale à :

$$m_B = [\sum x_i] / n = 4\,200/120 = 35{,}0 \ UI$$

L'estimation ponctuelle de la variance σ_{B^2} est égale à :

$$s^2_B = [\sum (x_i - m_B)^2] / (n-1) = 27\,045 \ / \ 119 = 227{,}27 \ UI^2$$

L'estimation ponctuelle de l'écart-type σ_B est égale à :

$$s_B = \sqrt{s^2_B} = \sqrt{227{,}3} = 15{,}08 \ UI$$

Exemple 2

- Afin d'estimer le pourcentage d'enfants obèses ou en surcharge pondérale chez les enfants de 10 à 16 ans, on a constitué un échantillon de 300 enfants scolarisés dans cette tranche d'âge. On en dénombré 45 enfants obèses ou en surcharge pondérale.

- Estimation de la proportion π d'enfants obèses ou en surcharge pondérale et de sa variance (en supposant l'échantillon représentatif).

L'estimation ponctuelle de la proportion est égale à :

p = 45 / 300 = 0,15 soit 15%

L'estimation ponctuelle de la variance est égale à :

p (1 – p) / n = 0,15 (1 – 0,15) / 300 = 0,000425

IV. Estimation par intervalle d'un paramètre

IV.1. Définition d'un intervalle de confiance

- L'estimation ponctuelle d'un paramètre θ est influencée par les fluctuations d'échantillonnage. Il est donc nécessaire de définir une fourchette de valeurs de θ compatibles avec les observations qui constitue l'estimation par intervalle.

  Le calcul des bornes (θ_{inf}, θ_{sup}) de la fourchette de valeurs, appelé intervalle de confiance (IC), repose sur la loi de distribution du paramètre θ à estimer (moyenne μ, proportion π, variance σ^2). On fixe une probabilité, notée α et correspondant au seuil ou risque d'erreur, d'avoir un écart δ entre la valeur estimée (ou observée) et la valeur vraie de θ. Pour la moyenne on aura :

  $$|\delta| = |m-\mu| \quad tel\ que \quad Pr(|m-\mu|) = \alpha$$

- L'intervalle est symétrique en probabilité autour de la valeur estimée du paramètre.

  L'IC peut être défini de la manière suivante : $Pr(\theta \leqslant \theta_{inf}) = Pr(\theta \geqslant \theta_{sup}) = 1-\alpha/2$

  On l'exprime sous la forme : $IC\ (1-\alpha) = [\theta_{inf}, \theta_{sup}]$

  Si le risque d'erreur α est égal à 5% , on aura un IC à 95% noté IC(95).

- L'interprétation de l'IC est la suivante : si l'on constitue un grand nombre d'échantillons (théoriquement un nombre infini) de même taille (n) par tirage au sort, on obtiendra autant de valeurs estimées du paramètre et d'intervalles de confiance associés. Pour un risque d'erreur α de 5%, 95% des intervalles de confiance contiennent la vraie valeur (inconnue) du paramètre.

IV.2.Intervalle de confiance d'une moyenne

Distribution d'échantillonnage d'une moyenne

- Soit une variable X qui suit une loi normale N (μ , σ). L'estimateur M de la moyenne est aussi une variable aléatoire (que l'on notera $\overline{X}$ de moyenne μ et de variance σ^2 /n (pour un échantillon de taille n) qui suit une loi $N(\mu, \frac{\sigma}{\sqrt{n}})$

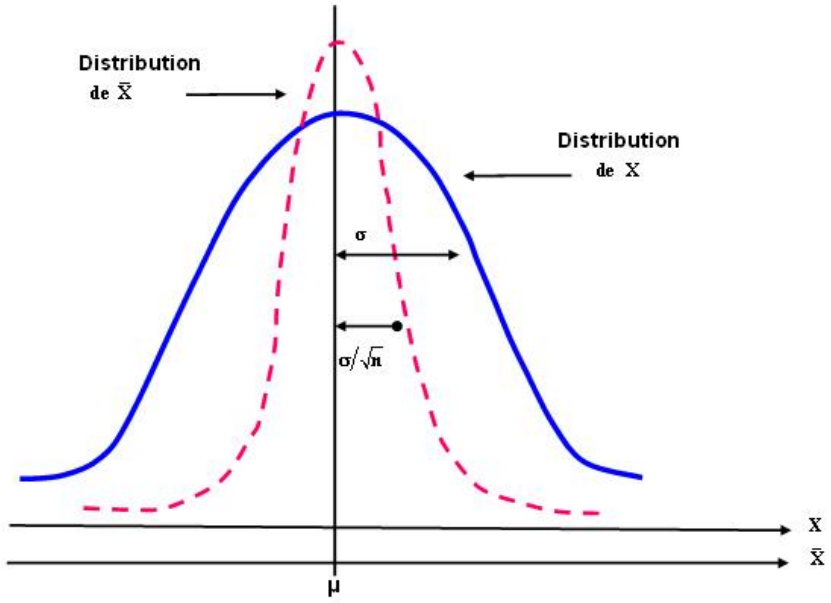
En effet, sachant que le n variables X_i suivent la même loi normale que X , on peut en déduire :

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Donc

$$E(M) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu$$

$$Var(M) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$



Si la variable X est normale (ou gaussienne) et σ^2 connue (situation rare), la distribution de la moyenne est une loi de Gauss $N(\mu, \frac{\sigma}{\sqrt{n}})$, quel que soit la taille n de l'échantillon.

Si la variable X est normale et σ^2 inconnue (estimée par s^2), ce qui est la situation la plus fréquente, la distribution de la moyenne dépend de la taille de l'échantillon :

- Si on a un grand échantillon (n ≥ 30) : la moyenne suit la loi $N(\mu, \frac{s}{\sqrt{n}})$
- Si on a un petit échantillon (n < 30) : la moyenne suit une loi de Student à n-1 ddl

Si la variable X n'est pas normale mais si l'échantillon est de grande taille (n > 30), la distribution de la moyenne est approximativement décrite par la loi $N(\mu, \frac{\sigma}{\sqrt{n}})$ si la variance σ^2 est connue

et $N(\mu, \frac{s}{\sqrt{n}})$ si la variance σ^2 est inconnue (estimée par s^2)

Expression de l'intervalle de confiance d'une moyenne

On dispose des valeurs estimées de la moyenne et de la variance (m et s^2) sur un n-échantillon aléatoire. Le calcul de l'intervalle de confiance de μ (correspondant à l'intervalle de fluctuation) au risque d'erreur α dépend de la loi de distribution de la moyenne donc de la taille de l'échantillon.

• Si on a un grand échantillon (n ≥ 30)

$\frac{m-\mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$ On aura donc :

$$Prob\left[-Z(\alpha) \leqslant \frac{m-\mu}{\sigma/\sqrt{n}} \leqslant +Z(\alpha)\right] = 1-\alpha$$

soit

$$Prob\left[m - Z(\alpha)\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant m + Z(\alpha)\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

A partir de la valeur de la moyenne m et de la variance s^2 , estimées sur l'échantillon, on peut calculer l'intervalle de confiance de μ au risque d'erreur α .

L'intervalle de confiance $IC(1-\alpha)$ de μ sera : $m \pm Z_\alpha \sqrt{\frac{s^2}{n}}$

Si σ^2 est connue on aura : $m \pm Z_\alpha \sqrt{\frac{\sigma^2}{n}}$

Il est exprimé de la manière suivante :

$$IC(1-\alpha) = \left[m - Z_\alpha \sqrt{\frac{s^2}{n}} \; ; m + Z_\alpha \sqrt{\frac{s^2}{n}}\right]$$

Pour α = 5 % $\Rightarrow$ $$IC(95) = \left[m - 1,96 \sqrt{\frac{s^2}{n}} \; ; m + 1,96 \sqrt{\frac{s^2}{n}}\right]$$

• Si on a un petit échantillon (n < 30)

A partir de la valeur de la moyenne m et de la variance s^2 , estimées sur l'échantillon, on peut calculer l'intervalle de confiance de μ au risque d'erreur α .

L'intervalle de confiance $IC(1-\alpha)$ de μ sera $m \pm t_{n-1;\alpha} \sqrt{\frac{s^2}{n}}$

Il est exprimé de la manière suivante :

$$IC(1-\alpha) = \left[m - t_{n-1;\alpha} \sqrt{\frac{s^2}{n}} \; ; m + t_{n-1;\alpha} \sqrt{\frac{s^2}{n}}\right]$$

Pour α = 5 % et n = 25 $\Rightarrow$ $$IC(95) = \left[m - t_{n-2,064} \sqrt{\frac{s^2}{n}} \; ; m + 2,064 \sqrt{\frac{s^2}{n}}\right]$$

Exemple 3

En reprenant les données de l'exemple 1 concernant le dosage d'un paramètre biologique B sur un échantillon de sujets issus d'une population présentant une allergie particulière, calculer l'intervalle de confiance de la moyenne du paramètre B pour la population de sujets présentant cette allergie.

On rappelle les données : n = 120, m_B = 35,0 IU et s_B = 15,08 UI.

- On ne connaît pas la loi de distribution de la variable B mais l'échantillon étant de grande taille (n = 120 > 30), l'intervalle de confiance de la moyenne (μ_B) au risque d'erreur α = 5 %, noté intervalle de confiance à 95 %, est égal à :

$$IC(1-\alpha) = \left[m - Z_\alpha \sqrt{\frac{s^2}{n}} ; m + Z_\alpha \sqrt{\frac{s^2}{n}} \right]$$

$$IC(95) = \left[35,0 - 1,96 \sqrt{\frac{227,27}{120}} ; 35,0 + 1,96 \sqrt{\frac{227,27}{120}} \right] = [32,3 ; 35,7]$$

Exemple 4

- On a effectué le dosage du cholestérol sanguin chez 20 sujets appartenant à la population générale et sélectionnés par tirage au sort. La moyenne de la cholestérolémie des 20 sujets est égale à 5,0 mmoles/l, et la variance des 20 mesures est égale à 6,5 (mmoles/l)^2. Sachant que la cholestérolémie (X) dans la population générale suit une loi normale, quel est l'intervalle de confiance de la moyenne μ de X ?

  Sachant que la cholestérolémie suit une loi normale dans la population géné-rale et que l'échantillon, d'effectif < 30, est représentatif de la population car constitué par ti-rage au sort, l'intervalle de confiance à 95% de la moyenne ( ) exprimé en mmoles/l est égal à :

$$IC(1-\alpha) = \left[m - t_{n-1;\alpha} \sqrt{\frac{s^2}{n}} ; m + t_{n-1;\alpha} \sqrt{\frac{s^2}{n}} \right]$$

Avec $t_{n-1;\alpha} = t_{19\ ;0,05} = 2,093$

$$IC(95) = \left[5,0 - 2,093 \sqrt{\frac{6,5}{20}} ; 5,0 + 2,093 \sqrt{\frac{6,5}{20}} \right] = [3,81 ; 6,19]$$

IV.3. Intervalle de confiance d'une variance

Le calcul de l'intervalle de confiance d'une variance est plus complexe et s'appuie sur la loi du Khi-deux. On montre que $(n-1)\frac{S^2}{\sigma^2} \rightarrow \chi^2_{n-1}$

Le calcul des deux bornes de l'intervalle de confiance d'une variance σ^2 ne sera pas abordé dans ce cours.

IV.4.Intervalle de confiance d'une proportion

Distribution d'échantillonnage d'une proportion

La distribution d'échantillonnage d'une proportion π est une loi bino-miale $B(n,p)$ pour un échantillon de taille n . La valeur de p est généralement esti-mée à partir des données de l'échantillon.

Dans certaines conditions on peut faire une approximation de la loi binomiale par la loi normale. En effet, on peut montrer que si $n \to \infty$

$B(n,p) \to N\left(np, \sqrt{np(1-p)}\right)$ si l'on considère le nombre de succès

ou $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ si l'on considère la proportion de succès

En pratique on pourra effectuer cette approximation si $n > 30$ et p voisin de 0,5 (on considère en général $0,10 < p < 0,90$) que l'on exprime sous la forme: $np > 5$ et $n(1-p) > 5$.

Remarque:

Dans certains ouvrages les conditions sont : $np > 10$ et $n(1-p) > 10$

Expression de l'intervalle de confiance d'une proportion

L'intervalle de confiance peut toujours être calculé en utilisant la loi binomiale, ce qui peut s'avérer compliquer quand on ne dispose pas de logiciel ou de tables appropriés.

Le calcul est simplifié si les conditions d'approximation par une loi normale sont vérifiées.

Nous présenterons uniquement cette situation.

$$IC(1-\alpha) = \left[p - Z_\alpha \sqrt{\frac{p(1-p)}{n}} \; ; p + Z_\alpha \sqrt{\frac{p(1-p)}{n}}\right]$$

Pour α = 5 % $\Rightarrow$ $$IC(95) = \left[p - 1,96 \sqrt{\frac{p(1-p)}{n}} \; ; p + 1,96 \sqrt{\frac{p(1-p)}{n}}\right]$$

Exemple 5

• En reprenant les données de l'exemple 2 sur l'estimation du pourcen-tage d'enfants obèses ou en surcharge pondérale chez les enfants de 10 à 16 ans sur un échantillon représen-tatif, calculer l'intervalle de confiance à 95 % de .

On rappelle les données : n= 300, P = 0,15.

• Pour calculer l'intervalle de confiance à 95 % on peut utiliser la loi normale car les conditions d'approximation de la loi binomiale sont vérifiées. n est grand, $0,10 < p < 0,90$ (nP = 300 x 0,15 = 45 et $n(1-p)$ = 300 x 0,85 = 255 > 30).

$$IC(1-\alpha) = \left[p - Z_\alpha \sqrt{\frac{p(1-p)}{n}} \; ; p + Z_\alpha \sqrt{\frac{p(1-p)}{n}}\right]$$

$$IC(95) = \left[0,15 - 1,96 \sqrt{\frac{0,15 \times 0,85}{300}} \; ; 0,15 + 1,96 \sqrt{\frac{0,15 \times 0,85}{300}}\right] = [0,11 \; ; 0,19]$$

L'intervalle de confiance à 95 % est compris entre 11 % et 19 %.

La corrélation est positive (les filles les plus grandes tendent à être les plus lourdes), mais de très mauvaise qualité (*r* proche de zéro).

Références

1- F. Carrat, A. Mallet, V. Morice , Biostatistique PACES - UE4 Université Pierre et Marie Curie,2013 – 2014.

2- Sylvie Rousseau, Enquêtes et sondages, manuel d'exercices, Année scolaire 2011–2012.

3- Licence de psychologie L3,U.F.R. S.P.S.E. PLPSTA02 Bases de la statistique inférentielle, université paris X Nanterre 2005-2006

4- Abdi-Basid ADAN, DEVOIR DE Statistique, Inferentielle, Master Professionnel en Méthodes Statistiques et Econométriques (MSE)

5- F. J. Silva, Introduction `a la statistique, Université de Limoges, Septembre 2016

6- Bardin Bahouayila. Cours de statistique descriptive. DEUG. Congo-Brazzaville. 2016. cel-01317598.

7- Jérôme BASTIEN, introduction a` la statistique inferentielle 2015-2016 .

8- Yves Tillé, Résume du Cours de Statistique Descriptive ,15 d´décembre 2010.

9- Jean-Yves DAUXOIS,CTU, Licence de Mathématiques Statistique Inférentielle, Université de Franche-Comté, Année scolaire 2011-2012.

10- Adil ELMARHOUM , echantillonnage et estimations ,université Mohamed V – AGDAL 2013.

11- Abdennasser Chekroun, Statistiques descriptives et exercices, Université Abou Bekr Belkaid Tlemcen, 2017 – 2018.