2012

Etude comparative des paramètres de position dans l'étude d'une serie statistique

Ndayishimiye, Gérard

UB-IPA

UNIVERSITE DU BURUNDI

INSTITUT DE PEDAGOGIE APPLIQUEE

DEPARTEMENT DE MATHEMATIQUES

ETUDE COMPARATIVE DES PARAMETRES DE POSITION DANS L'ETUDE D'UNE SERIE STATISTIQUE

Par :

Gérard NDAYISHIMIYE

Sous la Direction de :

Monsieur Fabien KIBINDIGIRI

Mémoire présenté et défendu publiquement en vue de l'obtention du grade de **Licencié en Pédagogie Appliquée, agrégé de l'enseignement Secondaire en Mathématiques**

Bujumbura, octobre 2012

DEDICACE

A nos parents

A notre épouse : NDAYIHEREJE Jeanne

A nos enfants :

- DUHEZAGIRE DON DU CARMEL

- ITUJIMBERE Reine Bénissa

-IRIMBERE Ange d'Espoir

A la famille NIYONIZIGIYE Léonidas

A nos frères et sœurs

Nous dédions ce mémoire.

Remerciements

Au terme de ce travail de fin d'études universitaires, nous remercions certaines personnes qui ont contribué à sa réalisation.

Nos remerciements sont adressés au promoteur et directeur de ce mémoire, Monsieur KIBINDIGIRI Fabien, ses conseils et ses remarques nous ont été d'une utilité capitale pour la réalisation de ce travail.

A tous mes éducateurs depuis l'école primaire jusqu'à l'université du Burundi particulièrement ceux de l'Institut de Pédagogie Appliquée, département de mathématiques. Nous leur disons sincèrement merci.

Enfin, nous remercions toute personne qui, de près ou de loin, a contribué à l'élaboration de ce travail.

NDAYISHIMIYE Gérard

Table des matières

Introduction générale

Dans l'étude de la comparaison des paramètres de position dans une série statistique, nous avons subdivisé ce travail en trois chapitres.

Le premier chapitre parle de l'origine et de l'évolution de la statistique dans le présent et dans l'avenir.

Le deuxième chapitre définit et développe les significations de quelques termes du vocabulaire statistique, nous allons représenter graphiquement une série statistique donnée pour la mettre sous une forme simplifiée.

Le troisième chapitre nous donne des paramètres de positions qui servent à caractériser l'ordre de grandeur des observations et explique certains paramètres de dispersion qui nous seront utile pour le développement de notre sujet.

CHAPITRE I. ORIGINE ET EVOLUTION DE LA STATISTIQUE [1],[2],[3]

I.1. Origine de la statistique

L'idée première et encore fondamentale de la statistique est celle de dénombrement. On trouve aux époques les plus reculées, des exemples de dénombrements au recensement des personnes et des biens. Il y a des millions d'années, les chinois utilisaient des tables de statistique agricole.

Bien que des dénombrements de populations humaines et de terres aient été réalisés depuis la plus haute antiquité pour les besoins de la guerre et de l'impôt. C'est au cours du dix-huitième siècle que l'emploi du terme «statistique» s'est imposé en Allemagne, dans le sens limité de connaissance d'un état à la suite des travaux de **Gottfreid ACHENWALL** (1719 – 1772).

Parmi les statisticiens de cette époque, on peut citer également en Grande Bretagne, Charles BABBAGE (1792 – 1871), à qui on doit, entre autres choses, une première machine à calculer automatique et la fondation de la première société statistique, la «Statistical Society of London» devenue ultérieurement la «Royal Statistical Society» ainsi que Francis GALTON (1822 – 1911), auteur de travaux de base relatifs notamment aux notions de corrélation et de régression.

Le progrès en matière statistique s'est réalisé grâce à deux impulsions complémentaires : les mathématiciens fournissent des outils de plus en plus précis aux statisticiens et ces derniers, cessant de s'intéresser aux seules questions démographiques et économiques, pour étendre le champ de leurs études à tous les domaines où le hasard est présumé jouer un rôle, posent de nouveaux problèmes aux mathématiciens et suscitent des solutions nouvelles.

A notre époque, l'importance et la complexité des problèmes soulevés, les mathématiciens probabilistes se sont eux-mêmes divisés en deux écoles dont les bases de travail sont identiques mais dont les buts sont distincts : les uns perfectionnent le calcul de probabilités dans la voie tracée par ses fondateurs, les autres ont créé et développé «la statistique mathématique».

Actuellement des méthodes statistiques, ne se limitent plus à l'exploitation des données très nombreuses qui apparaissent dans les dénombrements démographiques, économiques ou sociologiques mais qu'elles ont étendu leur champ d'application à toutes les recherches. A ce titre, on peut citer:

- Certaines branches des sciences physiques ou physico-mathématiques : théorie cinétique des gaz, mécanique statistique.
- La biologie : génétique, hérédité, médecine.
- L'étude du milieu naturel où évoluent les êtres vivants : sciences agricoles ;
- La psychologie appliquée : comportements des individus ; sondage de l'opinion publique.
- Les problèmes industriels : contrôle des fabrications, spécification des produits.

Cette liste n'est pas limitative : certains chercheurs n'hésitent pas à s'aider de la méthode statistique dans l'étude de phénomènes qui sont en apparence tout à fait étrangers aux mesures, comme l'évolution des grands courants de la pensée.

I.2. La première moitié du vingtième siècle

La première moitié du vingtième siècle est essentiellement marquée, dans le domaine statistique, par le développement de méthodes de plus en plus nombreuses et par l'utilisation de ces méthodes dans des secteurs d'application de plus en plus diversifiés.

Les années 1920 sont dominées par la forte personnalité du statisticien britannique Ronald Aylmer FISHER (1890 – 1962), auquel on doit notamment le développement des plans d'expérience et l'analyse de la variance et de la covariance qui occupent une place prépondérante dans le domaine agronomique d'abord, et dans de nombreux autres secteurs.

Les années 1930 sont marquées par de nouvelles applications de la statistique en économie, donnant naissance à l'économétrie et par l'utilisation de l'outil statistique dans le domaine industriel, en matière de maîtrise ou de contrôle de la qualité des produits.

A partir de 1940, la statistique intervient de façon de plus en plus fréquente dans certains problèmes de gestion, en relation avec le développement de la recherche opérationnelle.

I.3. La deuxième moitié du vingtième siècle

Dans la deuxième moitié du vingtième siècle, l'histoire de la statistique est étroitement liée au développement de l'informatique.

Vers 1955, les premiers ordinateurs sont commercialisés et introduits dans les services administratifs et universitaires de statistique. Très rapidement, ces

nouveaux outils y prennent une place très considérable sur le plan pratique en ce qui concerne l'emploi des méthodes statistiques mais aussi sur le plan théorique, en matière de recherche dans le domaine statistique.

L'ordinateur a été utilisé dans un premier temps pour effectuer plus rapidement ou plus facilement que le passé, les travaux qui étaient réalisés antérieurement à l'aide de machines à calculer de bureau. Dans un deuxième stade, l'ordinateur a permis l'emploi de méthodes statistiques déjà anciennes, qui n'avaient été utilisées en pratique ou qui étaient restées sous-employées en raison de l'importance des calculs qu'elles nécessitaient. Le développement de l'informatique a provoqué la naissance de nouvelles méthodes statistiques et de nouvelles procédures de calcul.

L'ordinateur a largement influencé l'enseignement de la statistique, notamment par les facilités qu'il offre en matière de résolution d'exercices.

Le mouvement observé depuis 1955 s'est considérablement accéléré à partir de 1975 environ, du fait de l'introduction des micro-ordinateurs ou ordinateurs personnels, de l'augmentation très rapide de leurs performances.

L'informatique a été un des principaux moteurs du développement de la statistique durant la deuxième moitié du vingtième siècle.
L'importance de l'ordinateur s'est progressivement accentuée au fil du temps, la simple évolution initiale des capacités de mémoire et de vitesse de traitement de l'informatique se doublant de possibilités d'acquisition automatique de données et de liaison entre ordinateurs, sous forme de réseaux.

Une conséquence de cette évolution est la constitution et la nécessité de traiter de grandes bases de données, dont l'interconnexion permet de former de vastes ensembles parfois qualifiés d'entrepôts de données. Ces bases et ces

entrepôts de données sont souvent caractérisés, non seulement par leur volume, mais également par des structures complexes et par le caractère très incomplet des données enregistrées.

L'extraction des données a pour but d'identifier autant que possible certaines informations particulières au sein de vastes ensembles de données. De même, la méthode des réseaux de neurones artificiels a pour objet d'établir ou modéliser des relations complexes liant de nombreuses variables.

I.4. Définition de la statistique

Dérivé du substantif latin « status » (état), le mot « statistique » possède en français comme dans d'autres langues, plusieurs significations distinctes.

D'une part, utilisé le plus souvent au pluriel le terme « statistique » désigne tout ensemble cohérent de données, généralement numériques relatives à un groupe d'individus ou d'objets.

On parle par exemple de la statistique de la production industrielle (quantités produites, prix de vente, coûts de production), des statistiques démographiques (natalité, mortalité) des statistiques du chômage, des statistiques des accidents de la circulation routière

Il convient toutefois de remarquer que contrairement à une opinion communément admise, cette acceptation du terme « statistique » ne concerne pas seulement des volumes importants de données.

D'autre part, le mot « statistique » désigne l'ensemble des méthodes qui permettent de recueillir et d'analyser les données dont il convient d'être question.

Enfin, le terme « statistique »est utilisé parfois pour désigner l'un ou l'autre paramètre, tel qu'une moyenne calculée à partir d'un ensemble de données.

La statistique est la science qui se propose de rassembler, d'ordonner, de représenter et d'étudier pour en tirer des conclusions, les données numériques se rapportent à des phénomènes collectifs.

La statistique descriptive est la partie de la statistique qui se propose de rassembler, d'ordonner, de classer, de synthétiser et de représenter les données numériques.

L'analyse des données recueillies sur des populations humaines, animales ou végétales ne se conçoit certainement plus sans faire un recours massif aux sciences de l'information que contiennent les données. Après avoir recueilli les données dans une étude statistique, l'étape suivante consiste à décrire l'ensemble des données sous une forme synthétique.

Ces données se présentent sous forme d'un tableau rectangulaire où les individus ou les unités statistiques se présentent en ligne et les variables en colonne. Chaque observation ou mesure effectuée sur chaque unité statistique ou individu est une des valeurs de la variable étudiée.

En statistique, une variable est une caractéristique ou un facteur susceptible de prendre une valeur différente selon les individus étudiés. La taille

d'un individu, la couleur des yeux des personnes européennes, la performance d'une personne au 100m sont des variables aux caractères.

L'analyse exploratoire des données, appelée également statistique descriptive, est une étape importante permettant de découvrir l'information contenue dans des données. Elle concerne une variable ou une caractéristique, deux variables ou deux caractéristiques à la fois, ou encore plus de deux variables ou plus de deux caractéristiques simultanément. Selon le cas, on parle de statistique descriptive à une variable ou à une dimension, de statistique descriptive à deux variables ou à deux dimensions, et de statistique descriptive à plusieurs variables ou à plusieurs dimensions (statistique multidimensionnelle).

CHAP II. SIGNIFICATION DE QUELQUES TERMES DU VOCABULAIRE STATISTIQUE [4],[5],[6],[7], [8].

II.1. Les séries statistiques

Une série statistique est la suite des observations des caractères ou des variables relevées sur les individus d'une population qui peut être constituée de véhicules automobiles dont on examine la marque où l'âge, que d'étudiants dont on étudie la nationalité, le mode de locomotion ou le nombre d'années d'études. Les éléments de la population sont des individus. Au sens statistique, un individu peut être donc une automobile, une personne, une plante,...

Il est souhaitable que la suite des observations soit accompagnée des indications utiles pour une bonne compréhension (Méthode, date, unité de mesure,...).

L'objectif est de mettre en évidence et d'étudier la distribution du caractère observé sur la population. On distingue les variables quantitatives des variables qualitatives : une variable ou un caractère, que l'on étudie est quantitative s'elle est numérique et peut faire l'objet de calculs, elle est qualitative dans le cas contraire.

Un caractère est dit quantitatif si on peut le mesurer ou le compter. Le caractère quantitatif prend le nom de variable statistique et ses différentes modalités sont les valeurs possibles de la variable.

On donne des exemples explicatifs : Le poids, le nombre de filles dans une boîte de nuit, le nombre d'heures de cours de statistique. Il existe deux types de caractères quantitatifs :

- D'une part, un caractère quantitatif est dit discret ou discontinu si les valeurs observées sont isolées et il n'existe aucune valeur intermédiaire possible : le nombre de filles dans une boîte de nuit, le nombre d'enfants par ménage et le nombre de voitures par ménage sont des variables statistique discrètes. Une variable discrète est le résultat de dénombrement.
- d'autre part, un caractère quantitatif est dit continu s'il peut prendre toute valeur d'un intervalle réel : le poids d'un individu, la taille d'une femme et l'âge d'un animal.

A côté de ces deux variables, les variables temporelles sont des variables quantitatives particulières qui utilisent les unités de temps, exprimées en secondes, minutes, heures. Là aussi, il existe deux types principaux de variables servant à définir un instant donné (début ou fin d'un événement).

A titre exemplatif, les variables « âge de la grossesse) (nombre de semaine) et incubation d'une maladie (heures, minutes) » sont des variables de durée. Les variables date de début de la grossesse (jour/mois/année) » et date de naissance » sont des variables de types « date ». Par contre, les variables « heure de début d'une maladie et heure de consommation d'un aliment sont des variables de types horaire.

Un caractère est quantitatif s'il est lié à une observation ne pouvant pas faire l'objet d'une mesure. Ses diverses modalités sont simplement constatées, repérées par un mot traduisant un état. On prend par exemples les caractères suivant : le sexe, la couleur, la situation matrimoniale, on les appelle des variables catégorielles comme pour les variables quantitatives, les variables catégorielles sont subdivisées en parties : les variables qualitatives ordinales s'exprimant en classes qui peuvent être ordonnées selon une échelle de valeurs :

le niveau d'étude (primaire, secondaire, universitaire (supérieur), le score d'appréciation d'un livre (convenable, médiocre, exécrable). Ces variables peuvent être recodées pour simplifier leur traitement informatique.

Les variables qualitatives nominales dont les classes ne peuvent être hiérarchisées. Ces dernières sont nommées, mais pas ordonnées. Leur ordre de présentation est arbitraire contrairement aux variables ordinales. le groupe sanguin (A ;B ;AB ;O), l'état-civil, (célibataire, marié, divorcé, veuf, séparé), la nationalité (burundaise, française, éthiopienne) et la religion (chrétienne, juive, musulmane,...), sont des variables qualitatives nominales.

II.2. Regroupement en classe

Une série statistique numérique pouvant prendre un grand nombre de valeurs différentes est regroupée par intervalle de valeurs. En ne retenant d'une observation que l'intervalle auquel elle appartient, on perd une certaine partie de l'information initiale, mais on rend plus aisées sa manipulation et son examen.

Le découpage doit être défini sans ambiguïté en particulier quant aux bornes des intervalles, afin de pouvoir affecter une observation à un intervalle unique.

On définit la largeur (ou l'amplitude) et le centre de chaque intervalle, qui peut être représenté par cette valeur centrale. On regroupe une série par intervalles d'amplitude égale, on préfère en général découper plus facilement les zones où se concentrent les valeurs observées et regrouper les données à l'intérieur d'intervalle d'amplitudes inégales.

II.3. Effectifs, fréquences

II.3.1 Effectifs, fréquences simples

Soit une série groupée par classes de valeurs : x_1, x_2, …, x_k éventuellement par intervalles, en retenant les valeurs centrales

L'effectif n_1 de la première classe est le nombre d'observations valant x_1

L'effectif n_2 est le nombre d'observations valant x_2, …,

L'effectif n_k est le nombre d'observations égales à x_k ;

L'effectif total est $n= n_1 + n_2 + ... + n_k$, le nombre d'individus de la population observée.

La fréquence f_i de la classe numéro i est le rapport $f_i=\frac{n_i}{n}$, la somme des fréquences est égale à 1. Comme les fréquences sont des nombres inférieurs à 1, elles sont souvent données en pourcentage.

II.3.2. Fréquences corrigées

Lorsqu'il s'agit d'une série numérique regroupée en classes d'amplitudes inégales, les fréquences ne permettent pas d'apprécier la distribution du caractère : la fréquence d'un intervalle étroit ne peut être directement comparée à celle d'un intervalle dix fois plus large.

On ramène toutes les classes à une largeur standard, en calculant par proportionnalité les fréquences corrigés correspondantes : soit l'amplitude standard choisie librement. si la classe numéro i a pour fréquence f_i et pour l'amplitude a_i, sa fréquence corrigée est : $f_i' = f_i \frac{a}{a_i}$

II.3.3. Fréquences cumulées

Lorsque les classes sont ordonnées, on définit les fréquences cumulées croissantes ,la fréquence de la classe numéro i est le rapport :

$f_i=\frac{n_1+n_2+\cdots+n_k}{n}$, qui mesure la part cumulée des i premières classes dans l'ensemble de la population. Lorsqu'on détermine les fréquences cumulées croissantes, la dernière calculée est égale à 1

II.4. Représentation graphique

Pour résumer l'information statistique contenue dans une distribution statistique quantitative ou qualitative, on utilise en première lieu des résumés graphiques. Chaque type de variable possède une ou plusieurs représentations graphiques.

II.4.1. Représentations graphiques des variables qualitatives diagramme en secteurs

Dans ce type de représentation, on utilise un disque plus communément appelé **camembert**, chacune des modalités est représentée par un secteur qui est proportionnel à l'effectif où à la fréquence.

Exemple

On considère le tableau relatif aux ventes de voitures neuves en France en Août 1998 :

Marque	Pourcentage
Renault	26.4%
PSA	25.9%
Marques étrangères	47.7%

C'est un caractère qualitatif qui prend trois valeurs ou modalités permettant de définir trois classes avec leurs fréquences :

Classes	Renault	PSA	Etrangère
Fréquence	O,264	0,259	0,477

On utilise pour présenter ces résultats un disque, divisé en secteurs. Les aires des secteurs sont proportionnelles aux effectifs :

L'angle au centre du secteur représentant la marge :

a) Renault a pour valeur en degrés $\frac{0{,}264\ x\ 360°}{1} = 95{,}04°$

b) PSA a pour valeur en degrés $\frac{0{,}259\ x\ 360°}{1} = 93{,}24°$

c) Etrangère a pour valeur en degrés $\frac{0{,}477\ x\ 360°}{1} = 171{,}72°$

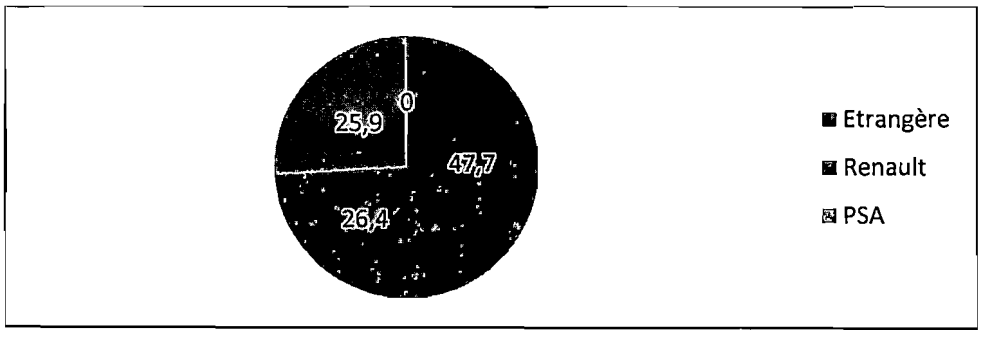


Diagramme en tuyaux d'orgue

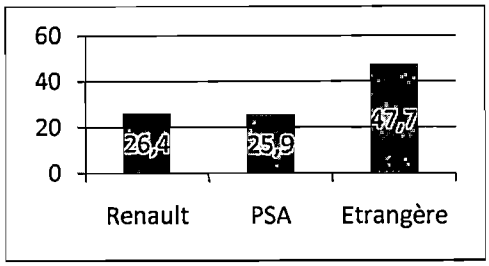
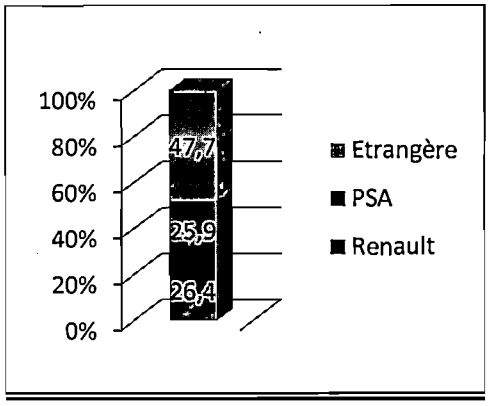


Diagramme en bandes



Dans ces deux diagrammes précédents, chaque classe est représentée par un rectangle de même largeur et de longueur proportionnelle à l'effectif, donc à la fréquence de la classe.

II.4.2 Représentations graphiques des variables quantitatives.

-Diagramme bâtons

Le diagramme bâtons est la représentation d'une distribution statique discrète (x_i, n_i) ou (x_i,f_i)ou $(x_i, f_i\ \%)$ où chaque bâton a une hauteur proportionnelle à l'effectif n_i de la modalité x_i.

Le rôle majeur est de représenter graphiquement la série statistique donnée sous forme simplifiée. Les distributions de

fréquences non cumulées absolues ou relatives peuvent être représentées graphiquement de trois façons différentes : par des diagrammes en bâtons, par des histogrammes et par des polygones de fréquences.

Dans des trois cas, les valeurs observées des classes sont portées en abscisses et les fréquences en ordonnées.

II.4.2.1. Les diagrammes en bâtons des effectifs

Les diagrammes en bâtons sont établis en traçant parallèlement à l'axe des ordonnées et en regard de chaque valeur observée x_i, un segment de longueur égale à la fréquence de cette valeur.

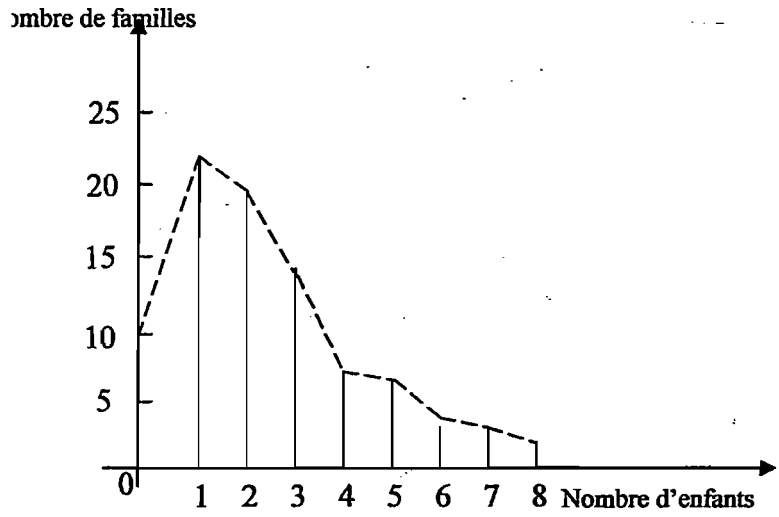
Exemple

Soit une distribution des 86 familles d'un village selon le nombre d'enfants.

Nombre d'enfants	0	1	2	3	4	5	6	7	8
Nombres de familles	10	21	19	14	9	7	3	2	1

On porte en abscisses les diverses valeurs du caractère étudié :

0, 1, 2,…., 8 enfants et en ordonnées le nombre de familles correspondant. Les segments verticaux obtenus constituent un diagramme en bâtons (ou en colonnes).



Les polygones de fréquences sont construits en joignant par une ligne brisée, les extrémités des segments voisins des diagrammes en bâtons relatifs aux distributions non groupées.

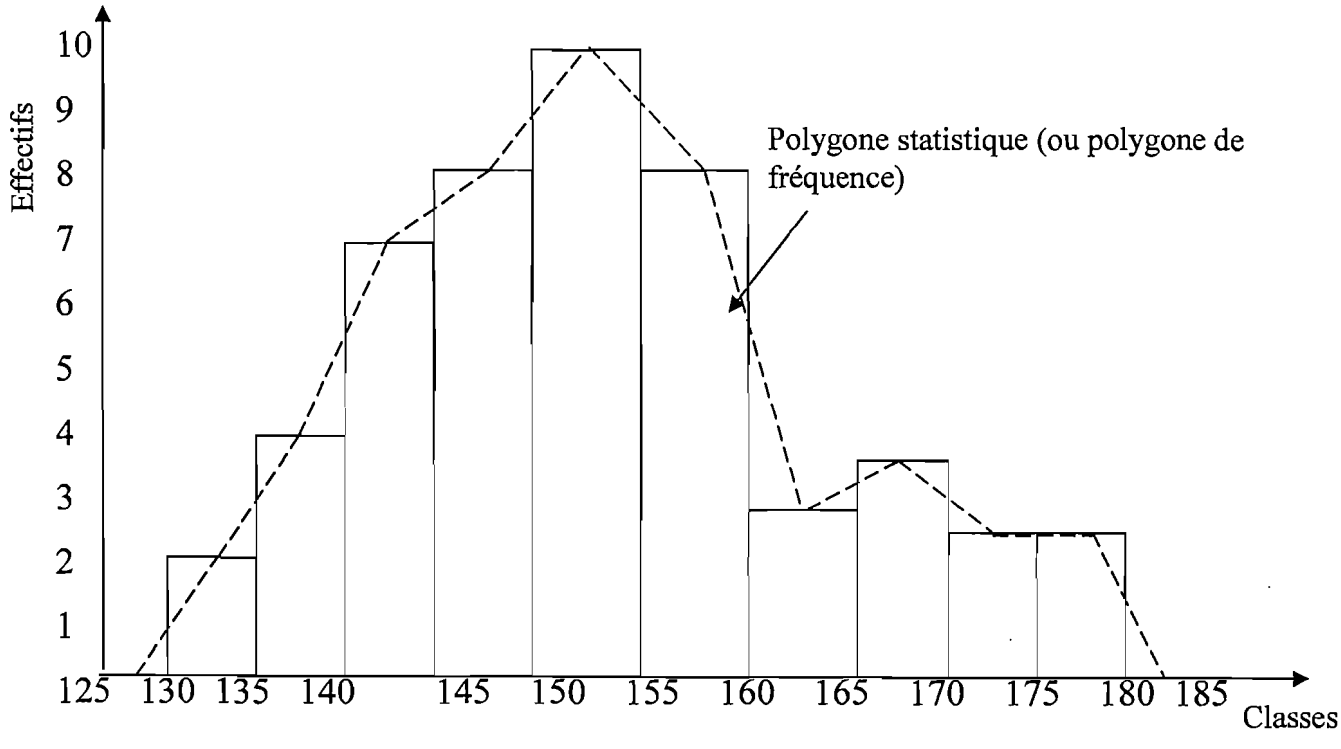
II.4.2.2. L'histogramme

Remarquons qu'il est aisé de représenter le polygone de fréquences et l'histogramme dans un même système d'axes orthogonaux. En effet, considérons un échantillon de 50 unités exprimées en cm, données par la série statistique suivante.

Construire le polygone des fréquences et l'histogramme dans un même système d'axes orthogonaux.

139	171	134	155	144	153	175	139	149	154
137	140	142	168	152	149	148	155	168	144
134	144	137	156	153	149	169	158	147	160
152	140	161	153	177	146	152	140	145	152
151	145	157	156	160	170	165	158	158	161

Classe	Effectifs	Effectifs cumulés	Fréquences relatives	Fréquences relatives cumulées
[130,135]	2	2	0.04	0.04
[135,140]	4	6	0.08	012
[140,145]	7	13	0.14	0.26
[145,150]	8	21	0.16	0.42
[150,155]	10	31	0.20	0.62
[155,160]	8	39	0.16	0.78
[160,165]	3	42	0.06	0.84
[165,170]	4	46	0.08	0.92
[170,175]	2	48	0.04	0.96
[175,180]	2	50	0.14	1.00



Dans un même système d'axes orthogonaux nous venons de construire le polygone des fréquences et l'histogramme à l'aide des éléments du tableau donné ci-haut. Le polygone de fréquences est tracé en pointillés et

Dans un même système d'axes orthogonaux nous venons de construire le polygone des fréquences et l'histogramme à l'aide des éléments du tableau donné ci-haut. Le polygone de fréquences est tracé en pointillés et l'histogramme des fréquences est la surface délimitée par ce polygone de fréquences.

CHAP III. COMPARAISON DES PARAMETRES DE POSITION DANS UNE SERIE STATISTIQUE [9],[10],[11],[12].

III.1. La moyenne arithmétique

III.1.1. La moyenne arithmétique simple

La moyenne arithmétique simple d'une série de n nombres ou observations $x_1,x_2,\ldots,x_n$ est le nombre $\bar{x}$ donné par : $\bar{x}_s=\frac{x_1+x_2+\cdots+x_n}{n}$

$$=\frac{1}{n}\sum_1^n x_i$$

On parle de simple car les x_i se représentent une seule fois (ont la pondération 1)

Exemple 1

Les nombres d'enfants de huit familles sont les suivants :0 ;0 ;1 ;1 ;1 ;2 ;3 ;4.

La moyenne arithmétique ou tout simplement, la moyenne

$$\bar{x}_s=\frac{0+0+1+1+1+2+3+4}{8}=1,5$$

III.1.2. La moyenne arithmétique pondérée

Elle peut être calculée à partir des valeurs distinctes et des effectifs $\bar{x}=\frac{1}{n}\sum_{i=1}^{p} n_i\, x_i$ Soient les différentes valeurs observées ou les points centraux $x_1,\ldots..x_i,\ldots\ldots\ldots..x_p$ et $n_1,\ldots\ldots\ldots n_i,\ldots\ldots..n_p$ les fréquences correspondantes.

Dans le cas des distributions non groupées, la deuxième expression est strictement équivalente à la première. Pour les distributions groupées, on commet en général une certaine erreur en remplaçant chacune des valeurs réellement observées par le point central de la classe correspondante.

De l'exercice précédent, on peut le résoudre en faisant les calcules avec les valeurs distinctes et les effectifs, soit le tableau suivant :

x_j	n_j
0	2
1	3
2	1
3	1
4	1
	8

Donc $\bar{x}=\frac{2.0+3.1+1.2+1.3+1.4}{8}$

$= \mathbf{1,5}$

Exemple 2

Supposons que les notes soient pondérées par le nombre de crédits, et que les notes de l'étudiant soient les suivantes

Note :	5	4	3	6	5
Crédit :	6	3	4	3	4

La moyenne pondérée des notes par crédits est alors :

$$\bar{x}_P = \frac{6.5 + 3.4 + 4.3 + 3.6 + 4.5}{6 + 3 + 4 + 3 + 4}$$

$$= \frac{30+12+12+18+20}{20}$$

$$= \frac{92}{20}$$

$$= 4{,}6$$

III.2. Moyenne géométrique

La moyenne Géométrique $\bar{x}_g$ d'une série statistique composée de n valeurs positives x_1, ..., x_n est par définition, la racine $n^{ième}$ du produit de ces n valeurs positives et on a :

$$\bar{x}_g = \sqrt[n]{x_1 . x_2 \dots x_p}$$

$$= (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$$

Par l'extension, la moyenne géométrique d'une distribution de fréquences, de valeurs positives x_i et de fréquences respectives f_i(i=1,...,p) peut être définie de la manière suivante : $\bar{x}_g = \sqrt[n]{x_1^{n_1}.x_2^{n_2}\ldots x_p^{n_p}}$

$$= (\prod_{i=1}^{p} x_i^{ni})^{\frac{1}{n}}$$

D'une façon générale la moyenne géométrique est l'antilogarithme de la moyenne arithmétique des logarithmes puisque

$\log(\bar{x}_g) = \frac{1}{n}\sum_{i=1}^{n} log x_i$

Cette moyenne est utilisée parfois quand on veut calculer la moyenne de taux d'intérêt.

Exemple

Supposons que les taux d'intérêt pour 4 années consécutives soient respectivement de 5, 10, 15 et 10% Que va-ton obtenir après 4 ans si on place 100francs.

Après un an, on a : 100 x 1,05F =105F

Après 2 ans, on a : 100 x 1,05F x 1,1=115,5F

Après 3 ans, on a : 100 x 1,05F x 1,1 x 1,15=132,825F

Après 4 ans, on a : 100 x 1,05F x 1,1 x 1,15 x 1,1 = 146,1075F

En calculant la moyenne arithmétique des taux, on obtient :

$$\bar{x} = \frac{1,05+1,10+1,15+1,1}{4}$$

$$= 1,1$$

En calculant la moyenne géométrique des taux, on obtient :

$$\bar{x}_g=(\frac{1{,}05.1{,}10.1{,}15.1{,}1}{1})^{\frac{1}{4}}$$

$$= 1{,}099431377$$

Le bon taux moyen est bien $\bar{x}_g$ et non $\bar{x}$ car si on applique 4 fois le taux moyen de la $\bar{x}_g$ aux 100 francs on obtient :

$$100\text{F x } \bar{x}_g^4 = 100.(1{,}099431344)^4$$

$$= 146{,}1075\text{F}$$

III.3. La moyenne harmonique

Si $x_i > 0$, on appelle moyenne harmonique la quantité :

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n}(\frac{1}{x_i})}$$

$$= \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$$

Pour extension, pour les distributions de fréquences, cette moyenne se définit comme suit :

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{p}(\frac{n_i}{x_i})}$$

Tandis que la moyenne quadratique ($\bar{x}_q$), est la racine carré de la moyenne, c'est-à-dire respectivement pour les séries statistiques et pour les distributions de fréquences :

$\bar{x}_q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$ pour une série non pondérée

ou

$\bar{x}_q = \sqrt{\frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{p} (n_i x_1^2)}$ Pour une série pondérée

Bien qu'elle puisse être calculée pour les valeurs quelconques positives, nulles ou négatives. Cette moyenne n'a un sens comme indice de position, que si tous les x_i sont positifs ou négatifs.

D'une façon plus générale, la moyenne de degré k, d'une série statistique ou d'une distribution de fréquences est définie comme suit dans le cas de valeurs x_i positives :

$\bar{x}_k = (\frac{1}{n} \sum_{i=1}^{n} x_i^k)^{\frac{1}{k}}$ (cas simple)

$= [\frac{1}{n} \sum_{i=1}^{p} (n_i x_i^k)]^{\frac{1}{k}}$ (cas pondérée)

L'exposant k pouvant prendre n'importe quelle valeur positive ou négative. Quand k est égal à -1, on trouve la moyenne harmonique, quand k est égal à 1, la moyenne arithmétique, quand k est égal à 2, la moyenne quadratique, et quand k tend vers 0 la moyenne géométrique :

$\bar{x} = \bar{x}_1$

$\bar{x}_h = \bar{x}_{-1}$

$\bar{x}_q = \bar{x}_2$

et $\bar{x}_g = \lim_{k \to 0} \bar{x}_k$

Pour une valeur du paramètre k, égale à 3 on dit que c'est une moyenne cubique.

$$\bar{x}_c = \bar{x}_3$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_1^3\right)^{\frac{1}{3}}$$

$$= \left[\frac{1}{n}\sum_{i=1}^{p}(n_i x_i^3)\right]^{\frac{1}{3}}$$

Pour k tendant vers l'infini, on retrouve selon la valeur maximum observée

$$\lim_{k \to -\infty} \bar{x}_k = x_{minimum}$$

$$\lim_{k \to +\infty} \bar{x}_k = x_{maximum}$$

Exemple : Un cycliste parcourant 4 étapes de 100km. Les vitesses respectives pour ces étapes sont de 10km/h, 30km/h, 40km/h, 20km/h.

Calculer : a) La vitesse moyenne

b) La moyenne arithmétique des vitesses

c) La moyenne harmonique des vitesses

d) Dites une remarque sur le calcul de ces vitesses

Résolution

La résolution nous montre qu'on a parcouru

- la première étape en 10h
- la deuxième étape en 3h20min
- la troisième étape en 2h30min
- la quatrième étape en 5h.

Il a donc parcouru le total de 400km

en: 10h + 3h20min + 2h30min + 5h=20h50min

=20,8333h

La moyenne arithmétique des vitesses :

$$\bar{x} = \frac{\frac{30+10+40+20}{4}km}{h}$$

$$= 25km/h$$

La moyenne harmonique des vitesses :

$$\overline{x_h} = \frac{\frac{4}{\frac{1}{30}+\frac{1}{10}+\frac{1}{40}+\frac{1}{20}}km}{h}$$

$$= 19.2km/h$$

On constate que la moyenne harmonique est la moyenne appropriée de calculer la vitesse moyenne pondérée.

III.4. Généralisation de la moyenne

a) La moyenne arithmétique : $\bar{x}$, pour n observations $x_1, x_2, \ldots, x_n$; on a

- $\bar{x}_s = \frac{\sum_{i=1}^{n} x_i}{n}$, n=$\sum n_i$(Moyenne arithmétique simple)
- $\bar{x}_p = \frac{\sum_{i=1}^{k} n_i x_i}{n}$, k=Nombre de modalité (Moyenne arithmétique pondérée)

b) La moyenne géométrique G

Cas simple : $\bar{x}_{G_s} = \sqrt[n]{x_1 . x_2 \dots x_n}$

$= \sqrt[n]{\prod_{i=1}^{n} x_i}$

Cas pondéré : $\bar{x}_{G_p} = \sqrt[n]{x_1^{n^1} . x_2^{n^2} . x_3^{n^3} \dots x_k^{n^k}}$

$= \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}$

c) La moyenne harmonique

Cas simple : $\bar{x}_{h_s} = \frac{1}{\frac{1}{n}(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n})}$

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

$= \frac{n}{\sum_{i=1} \frac{1}{x_i}}$ (Moyenne harmonique simple)

Cas pondéré : $\bar{x}_{h_p} = \frac{1}{\frac{1}{n}(\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_k}{x_n})}$

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

$= \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$ (Moyenne harmonique pondéré)

d) La moyenne quadratique Q

Cas simple : $\bar{x}_{q_s} = \sqrt{\frac{x_1^2+x_2^2+x_3^3+\cdots+x_n^K}{n}}$

$$= (\frac{1}{n}\sum_{i=1}^{n} x_i^2)^{1/2}$$

Cas Pondéré : $\bar{x}_{q_p} = \sqrt{\frac{n_1x_1^2+n_2x_2^2+\cdots n_kx_k^2}{n}}$

$$= (\frac{1}{n}\sum_{i=1}^{k} n_ix_i^2)^{1/2}$$

III.5. La médiane

La médiane est un paramètre de position tel que la moitié des observations lui sont inférieures ou égales et la moitié supérieures ou égales.

Pour les séries statistiques et pour les distributions non groupées, quand le nombre d'observation n est impair, la médiane est l'observation de rang (n+1)/2

Quand n est pair, tout nombre compris entre $x_{n/2}$ et $x_{n/2+1}$ répond à la définition et on convient généralement de prendre, comme valeur de la médiane, la moyenne arithmétique de ces deux observations.

$\bar{x} = (x_{n/2} + x_{n/2+1})/2$, notée Me qui est le nombre partageant une série des valeurs observées en deux séries de même taille. La Me est déterminée après avoir rangé par ordre croissant les individus de la population.

Prenons un exemple d'un série statistique simple concernant la note obtenue par sept étudiants sur 15

5 7 9 10,5 11 12 13

Ici la note médiane Me est 10,5

Dans le cas des distributions non groupées la médiane peut également être déterminée graphiquement en cherchant sur le polygone de fréquences cumulées, l'abscisse du point d'ordre $\frac{n}{2}$ ou $\frac{1}{2}$, suivant qu'on considère les fréquences absolues ou les fréquences relatives.

Quand l'ordonnée $\frac{n}{2}$ ou $\frac{1}{2}$ correspondant exactement a niveau d'une marche du polygone de fréquences, la médiane n'est pas complètement définie. La convention établie ci-dessus revient à prendre alors, comme valeur de la médiane, l'abscisse du milieu de la marche

Exemple 1

Soit à déterminer la médiane d'une série statistique ordonnée :

1 2 3 4 5 6

On a Me $= \frac{3+4}{2}$

$= 3,5$

Exemple 2

La cote sur 20 de 15 étudiants :

12	7	12	10	14
6	16	10	9	12
11	12	10	8	15

Déterminer la médiane et sa position

On commence d'abord à tirer les données par l'ordre croissant

6 7 8 9 10 10 10 11 12 12 12 12 14 15 16

Donc la médiane est la valeur classée en 8ème Position :

$$\frac{2n+2}{4} = \frac{2.15+2}{4}$$

$$=8$$

La médiane est 11.

Pour les distributions groupées, la classe médiane est celle qui convient la médiane. Si on admet que les observations appartenant à cette classe y sont reparties uniformément, la médiane peut être estimée grâce à la relation :

$$Me = Lo + a_m\left(\frac{0,5\_F_{m-1}}{f_m}\right)$$

Où

- Lo :limite inférieure de la médiane
- a_m :étendue
- 0,5 : Milieu de la série statistique
- F_{m-1} :Fréquence relative cumulée qui précède la fréquence relative cumulée de » la médiane
- f_{im} :Fréquence relative médiane

Soit l'exemple dont la série statistique classée

Cette méthode de calcul revient à faire une interpolation linéaire entre les limites de la classe médiane, ce qui peut être réalisé graphiquement en recherchant comme avant sur le polygone de fréquence cumulées, l'abscisse du point d'ordonnée $\frac{n}{2}$ ou $\frac{1}{2}$.

III.6. Le mode

Le mode est toute valeur du caractère pour laquelle l'effectif ou la fréquence est maximum. Lorsque les valeurs du caractère sont regroupées en classe, supposées de même amplitude, toute classe d'effectif maximum est appelé « classe modale ».

Le mode d'une variable statistique ne satisfait pas les conditions générales de Jule, son estimation est difficile sauf si on a ajusté une courbe à la série. Il n'y a pas de formule analytique de calcul de mode, dans le cas d'une variable dont la distribution est unimodale et modérément asymétrique.

Exemple 1

Dans les séries : a) {23 ;24 ;25 ;28 ;25 ;30 ;31}

b) {3 ;8 ;9 ;14 ;15}

c) {2 ;4 ;4 ;9 ;8 ;4 ;10 ;14 ;11 ;14 ;16 ;14}

La réponse en :

a) Le mode est 25

b) Le mode n'existe pas

c) On a deux modes : 4 et 14.

Pour le cas d'une distribution pondérée (médiane), on calculer la médiane de la façon suivante :

Exemple

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

Dans notre cas n $=\sum_{i=1} n_i$

$= 28$

$2k=n$

$= 28$; donc k=14

La médiane $= \frac{x_k+x_{k+1}}{2}$

$= \frac{x_{14}+x_{15}}{2}$

{8 ;8 ;8 ;8 ;8 ;7 ;7 ;7 ;7 ;7 ;7 ;7 ;6 ;6 ;6 ;4 ;4 ;4 ;4 ;4 ;4 ;4 ;4 ;4 ;2 ;2 ;2 ;2}

La classe médiane est [6,6[

Dans ce cas la médiane est : $\frac{6+6}{2} = 6$

Remarque :Une distribution n'ayant qu'un seul mode est une distribution uni modale, si elle a deux modes, elle est dite bimodale et si elle a plusieurs modes elle est plurimodale.

III.7. Quartiles, intervalle interquartile

Pour limiter, l'effet des valeurs les plus marginales, on préfère à l'étendue l'intervalle interquartile, qui est l'étendue de la série privée de ses deux quarts extrêmes.

Il contient la moitié centrale des observations plus précisément, soit une série statistique numérique ordonnée par valeurs croissantes, on définit d'abord les quartiles.

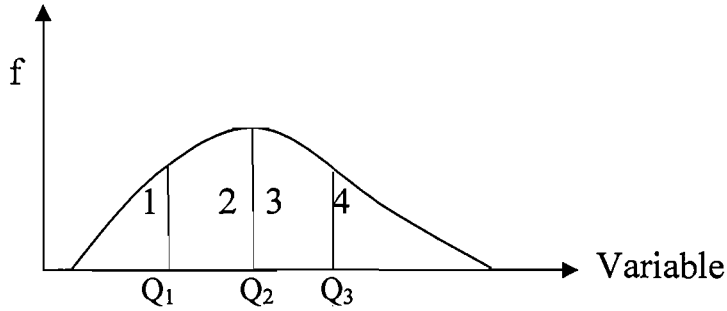
Le quartile, noté Q_1, est la valeur de la série telle qu'il y ait un quart des observations qui lui soient inférieures et les trois quarts qui lui soient supérieures.

Le second quartile, noté Q_2, est la valeur de la série qui sépare les deux premiers des deux derniers, c'est la médiane.

Le troisième quartile, noté Q_3, est la valeur de la série telle qu'il ait les trois quarts des observations qui lui soient inférieures et un quart qui lui soit supérieure.

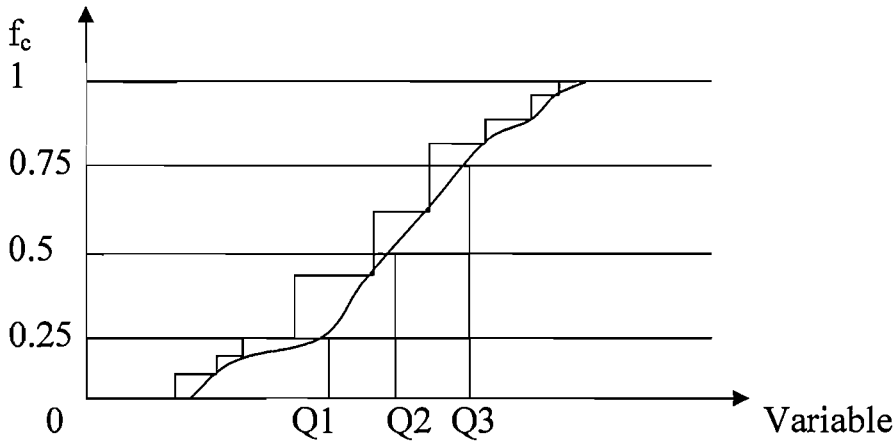
Soit une variable statistique X, l'intervalle interquartile est la différence entre le quartile d'ordre 3 et le quartile d'ordre 1 : Q_3-Q_1. L'intervalle interquartile représente 50% de valeurs de la variable X, en laissant de part et d'autre de cet intervalle 25% des valeurs de la variable.

L'intervalle interquartile est l'écart : Q_3-Q_1



Les surfaces: 1, 2, 3 et 4 sont égales.

Le Plus souvent, les séries étant données en classes, on estime les quartiles Q_1 et Q_3 par l'interpolation. Ce sont les points d'intersection des lignes horizontales et le polygone des fréquences cumulées.



Quand les observations sont nombreuses, il est utile de les condenser sous la forme d'une distribution de fréquences ou d'une distribution statistique (observée).

Le nombre d'occurrences d'une même valeur observée est par définition sa fréquence absolue ou plus simplement sa fréquence. La distribution de fréquence est formée des différentes valeurs observées x_1, x_2, ..., x_i, ..., x_p, rangées par ordre croissant, et des fréquences absolues sont évidemment telles que :$\sum_{i=1}^{p} n_i = n$

On utilisera la même notation x_i pour désigner à la fois n valeurs observées (x_1, x_2, ..., x_n) et les p valeurs différentes les unes les autres qui apparaissent dans la distribution de fréquences (x_1, x_2, ..., x_p). La disjonction entre les deux séries de valeurs se fait uniquement en fonction des limites de l'indice i avec i allant de 1 à n dans le premier cas et de 1à p dans le deuxième cas. Une autre solution aurait été d'utiliser deux symboles différents pour les deux séries de valeurs x_i avec

i=1, ..., n dans le premier cas et x_j avec j

j=1, ..., p dans le second cas.

Les fréquences peuvent être exprimées en valeur relative, en proportions ou éventuellement en pourcentages du nombre total d'observations. En désignant ces fréquences relatives par le symbole

n_i', on a : $n_i' = \frac{n_i}{n} (ou\ 100\ n_i/n)$ et $\sum_{i=1}^{p} n_i' = 1 (ou\ 100)$

Les fréquences observées peuvent être additionnées de proche en proche, de manière à établir des distributions de fréquences cumulées. Qu'il s'agisse de fréquences absolues ou fréquences relatives, la fréquence cumulée d'une valeur observée x_i est la somme des fréquences correspondant à cette valeur et à l'ensemble des valeurs inférieures.

III.8. Déciles et centiles

La génération de la notion de médiane porte le nom de quartile. Parmi les quartiles utilisés, nous trouverons les quartiles, les déciles et les centiles. Le quartile Q_2 d'une variable statistique est égal à la médiane. Les calculs faits pour la médiane sont les mêmes pour la recherche des trois quartiles.

Les déciles, notés D_1, D_2, ..., D_9 respectivement les centiles ou percentiles, souvent notés c_1, c_2, ...,c_{99} partagent l'effectif total d'une série ou d'une distribution statistique rangée par ordre croissant (ou décroissant) en dix (respectivement cent parties égales).

Par exemple : $C_{50}=D_5=Q_2$

$C_{10}=D_1$

$C_{90}=D_9$

Les déciles séparent les observations ordonnées en deuxièmes successif tandis que les centiles, les observations de la population sont ordonnées en centièmes.

Les quartiles permettent de construire d'autres coefficients permettant dans certaines conditions la comparaison des séries statistiques d'échelle ou de nature différentes.

L'intervalle interquartile $\frac{Q_3-Q_1}{Q_2}$ donne une mesure de la dispersion d'une série, indépendante de l'unité employée.

Le coefficient $\frac{Q_3-Q_2}{Q_2-Q_1}$ donne une mesure de l'asymétrie, également indépendante de l'unité employée.

Avant de faire l'étude comparative des paramètres de position, nous sommes obligé de dire un mot sur les paramètres de dispersion qui permettent de savoir comment une masse de d'observation se répartit autour d'une valeur centrale.

On les appelle aussi valeurs ou paramètres de la variabilité. Les plus importantes de ces paramètres sont : la variance et l'écart-type en raison du fait qu'ils prennent en considération toutes les valeurs de la distribution.

Ils s'adaptent mieux à la réalité observée qu'ils intègrent totalement.

a) Etendue ou intervalle de variation d'une série

L'étendue est la différence entre la plus grande et la plus petite des observations faites sur une variable statistique quantitative.

Soit la variable statistique quantitative x, la distribution est (x_i, n_i) avec

$i \in \{i, \ldots, r\}$ l'étendue est le nombre : $x_r - x_1$, les x_i doivent être classés par ordre croissant. Cette amplitude $x_r - x_1$ est souvent désignée par le symbole w.

Exemple

Soit une série statistique, ordonnées par ordre croissant : -5 ; -3 ; -1 ; 1 ; 3 ; 5.

L'étendue : $w = 5 - (-5)$

$= 10$

L'étendue est la mesure la plus simple de la dispersion (variabilité ou étalement) des observations faites sur une variable. L'étendue ne dépend que très indirectement de l'ensemble des valeurs x_i de la variable x. L'étendue est

très influencée par les valeurs extrêmes de la variable statistique qui sont parfois aberrantes, ce qui fait une mesure peu utilisée.

b) Ecart absolu moyen

Soit une distribution statistique (x_i,n_i) où i∈ $\{1, \ldots, r\}$, on appelle écart absolu moyen, le nombre, noté $\bar{x}_e$ et défini par :

$$\bar{x}_e = \frac{1}{n}\sum_{i=1}^{i=r} n_i \, |x_i - \bar{x}|$$

L'utilisation de l'écart absolu moyen comme la différence à la moyenne d'une variable statistique est nulle.

En effet, le nombre $\bar{x}_e$ est tout simplement une moyenne arithmétique pondérée des valeurs absolues de la variable centrée sur sa moyenne arithmétique. L'écart absolu moyen est un indicateur de dispersion difficilement maniable puisqu'il contient des valeurs absolues, elle est la somme des valeurs absolues des écarts à la moyenne divisée par le nombre d'observations :

$$\sigma_{moyenne} = \frac{1}{n}\sum^{n} |x_i - \bar{x}|$$

c) Variance

La variance d'une distribution x_1 ;x_2 ;x_3 ;… ;x_n d'une variable x est donnée pour le cas :

Simple : $V(x)=\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$

Pondéré :$V(x) =\frac{1}{n}\sum_{j=1}^{n} n_i(x_j - \bar{x})^2$;

On appelle variance la somme des carrés des écarts à la moyenne divisée par le nombre d'observations

d) Ecart-type

Elle est la racine carrée de la variance et on la note « σ »

On a donc : $\sigma = \sqrt{V(x)}$

$$= \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}.$$

e) Formule développée de la variance

$$V(x) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

$$= \frac{1}{n}\sum_{j=1}^{n}({x^2}_j - 2x_j\bar{x} + \bar{x}^2)$$

$$= \frac{1}{n}\sum_{j=1}^{n}{x^2}_j - \frac{2\bar{x}}{n}\sum_{j=1}^{n}x_j + \bar{x}^2$$

$$= \frac{1}{n}\sum_{j=1}^{n}{x^2}_j - 2\bar{x}\bar{x} + \bar{x}^2$$

$$= \frac{1}{n}\sum_{j=1}^{n}{x^2}_j - \bar{x}^2$$

III.9. Etude Comparative des paramètres de Position

La comparaison de paramètres de dispersion absolue de deux caractères n'a de sens que si les caractères sont de même nature et de même ordre de grandeur. Dans le cas contraire, la comparaison n'est possible qu'en ayant recours à des mesures de dispersion relative, c'est-à-dire en effectuant le rapport entre un paramètre de dispersion absolue et la valeur centrale qui lui tient de référence. Un paramètre de dispersion relative est une mesure de l'écart relatif des valeurs d'unes distribution absolue divisé par une valeur qui peut être exprimé en %.

Une distribution qui donne une valeur centrale ne nous renseigne pas sur la dispersion des valeurs autour de cette valeur centrale, c'est-à-dire sur la tendance qu'elles ont à se concentrer ou se disperser autour de celle-ci.

Exemple

Si l'on considère deux professeurs x et y chargés de noter 9 élèves, peut on apprécier leur manière de noter simplement en regardant la moyenne, la médiane ou le mode de leurs notes ?

Etudiant	**Notes du Professeur x**	**Notes du professeur y**
A	7	0
B	8	5
C	9	9
D	10	10
E	10	10
F	10	10
G	11	11
H	12	15
I	13	20
Mode	10	10
Moyenne	10	10
Médiane	10	10

A s'en tenir à l'analyse des valeurs centrales on serait amené à conclure que les deux professeurs x et y notent rigoureusement de la même manière : la moyenne, la médiane et le mode est dix, mais on sent bien intuitivement que ce n'est pas le cas et qu'il existe une différence, tient au fait que le professeur x concentre ses notes autour de 10 alors que le professeur y disperse d'avantage ses notes autour de la valeur de référence.

Il est donc utile de compléter les valeurs centrales par un paramètre de dispersion absolue qui donne un ordre de grandeur de l'écart des valeurs à la valeur centrale de référence.

En prenant l'exemple précédent, la dispersion des notes du professeur x est de 13-7=6 points alors que celles du professeur y est de 20 – 0=20 points. L'écart maximum entre deux notes est donc plus élevé chez le professeur y que chez le professeur x.

III.9.1. Mesures de la dispersion statistique en référence à la médiane

Quand à la dispersion statistique en référence à la médiane, l'intervalle interquartile est l'étendue de la distribution sur laquelle se trouvent concentrée la moitié des éléments dont les valeurs de x sont les moins différentes de la médiane. On exclut alors de la distribution les 25% des valeurs les plus faibles et les 25% des valeurs les plus fortes de x.

L'intervalle interquartile des notes au professeur x est les deux points puisque la moitié 50% de ses notes sont comprises dans l'intervalle [9,11] une fois qu'on retiré les 25% des notes les plus faibles et les 25% des notes les plus fortes. Il en va de même pour le professeur y qui concentre également 50% de ses notes dans l'intervalle [9,11].

Pour ce critère, la dispersion des deux distributions est donc équivalente.

III.9.2. Mesure de la dispersion statistique à la moyenne arithmétique

L'écart absolu moyen est la moyenne arithmétique de la valeur absolue des écarts à la moyenne. C'est donc la distance moyenne à la moyenne.E.A.M de x= $\frac{1}{n}\sum_{i+1}^{n}|x_i - \bar{x}|$

III.9.3. Calcul de l'écart absolu moyen des notes du professeur x

i	x_i	$\lvert x_i - \bar{x} \rvert$
A	7	3
B	8	2
C	9	1
D	10	0
E	10	0
F	10	0
G	11	1
H	12	2
I	13	3
Total	90	12
Moyenne	10	$\frac{12}{9} \cong 1{,}33$

L'écart moyen de la notation du professeur x est de 1,3, ce qui signifie que les notes s'écartent en moyenne de 1,3 de la moyenne. Il n'y a pas en moyenne, de gros écarts à la moyenne. Si on effectue le même calcul pour le professeur y on trouve un écart absolu moyen de 3,6. Ce qui signifie que ses notes s'écartent généralement beaucoup plus de la moyenne. On dit donc que pour ce critère, la dispersion des notes du professeur y est plus forte que celle du professeur x.

Une distribution de fréquences a plusieurs modes si on veut mettre en évidence le fait qu'elle a plusieurs classes non contiguës dont la fréquence est nettement plus élevée que celle des autres classes.

Dans ce cas des distributions symétrique, la médiane est normalement inférieure ou supérieure à la moyenne, selon que la dissymétrie est gauche, décentré vers la gauche ou vers à droite et la différence entre les deux paramètres est autan plus importante, en valeur absolue que la dissymétrie est plus prononcée.

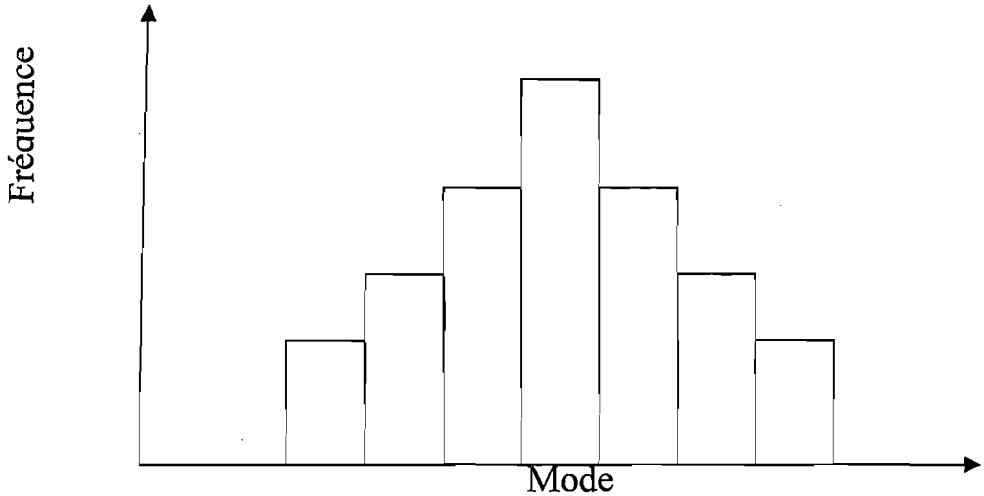
III.9.4. Coefficient d'asymétrie

Le coefficient d'asymétrie d'une distribution est donné par :

$$\alpha_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n\sigma^3(x)}\left(\sqrt{\frac{n}{n-1}}\right)^3$$

Où n est le nombre d'observations et $\sigma(x)$ l'écart-type on la déviation standard une distribution est symétrique si α_3=0.

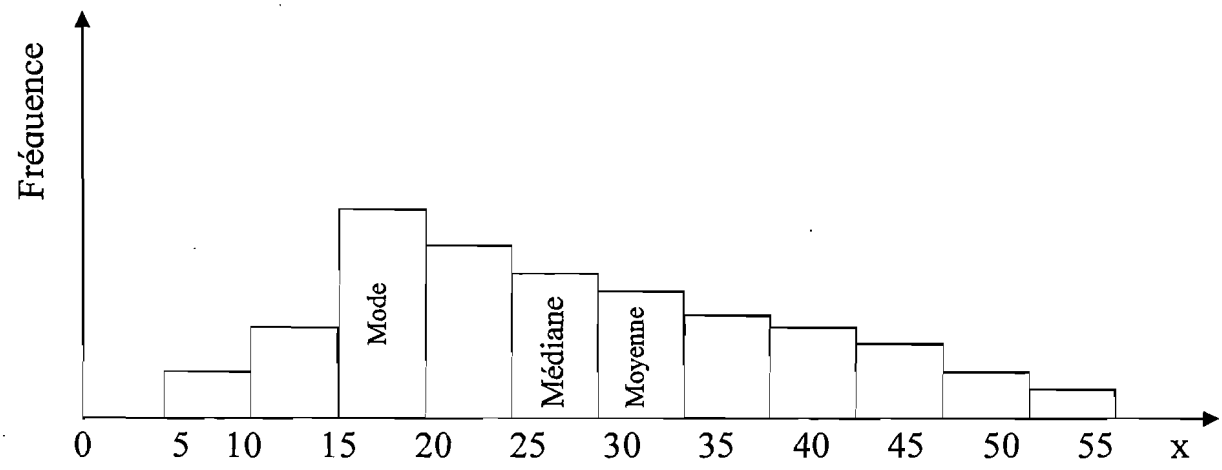
Les valeurs d'une variable statistique sont alors également dispersées de départ et d'autre d'une valeur centrale.



Une distribution est asymétrique à droite si $\alpha_3 > 0$

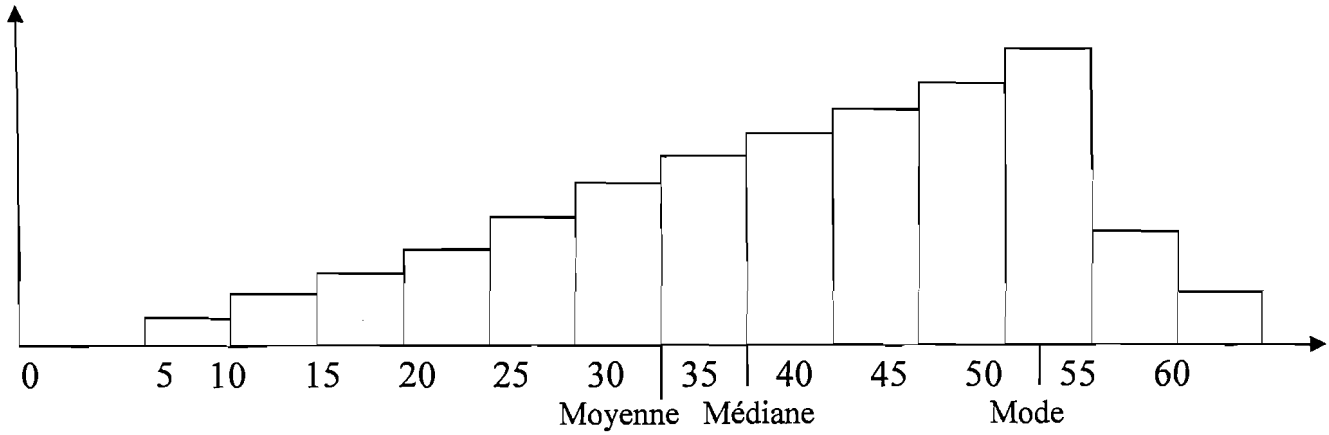
Les valeurs d'une variable statistique sont alors plus dispersées à droite.

Autrement dit, il s'agit d'une distribution dissymétrique avec des valeurs très élevées (queue à droite).



Une distribution est asymétrique à gauche si $\alpha_3 < 0$. Les valeurs d'une variable statistique sont alors plus dispersées à gauche.

Il s'agit d'une distribution dissymétrique vers les valeurs basses (queue à gauche).



III.9.5. Coefficient d'aplatissement

Une distribution est plus ou moins aplatie selon que les fréquences des valeurs voisines des valeurs centrales différent peu ou beaucoup les unes par rapport aux autres. L'aplatissement d'une distribution est donné par :

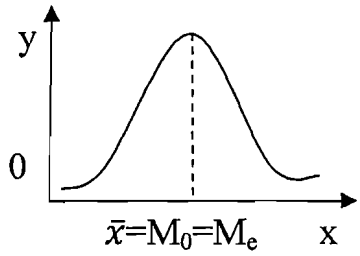
$$\alpha_4 = n \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{(n-1)^2 \sigma^4(x)}$$

Où n est le nombre d'observation et $\sigma(x)$ l'écart type ou la déviation standard. L'aplatissement est nul ($\alpha_4 = 0$) pour une distribution normale, positif $\alpha_4 > 0$ lorsque la distribution est moins aplatie, négatifs ($\alpha_4 < 0$) lorsque la distribution est plus aplatie.

III.9.6. Relation entre les trois mesures de tendance centrale

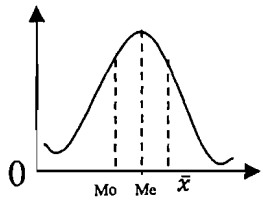
Trois types de relation apparaissent entre les trois mesures de tendance centrale.

- Première situation : les trois valeurs coïncident : M_0 ;M_e et $\bar{x}$. Dans ce cas, on dit que les valeurs observées forment une courbe symétrique de forme :

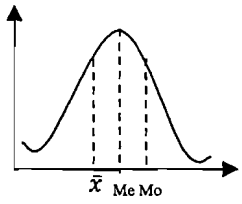


Ces paramètres ne sont pas exactement égaux mais sont très proches

- La deuxième situation : le mode se trouve à gauche de la médiane et de la moyenne : « c'est l'asymétrie négative : il y a plus de fréquence dans le côté gauche



- La troisième situation : Le mode se trouve à droite de la médiane et l a moyenne : c'est l'asymétrie à droite



On remarque que la moyenne arithmétique est la plus préférée de ces trois mesures de tendance centrale : elle est considérée comme étant la plus représentative du fait qu'elle intègre toutes les valeurs de la distribution à la

différence du mode et de la médiane mais sa faiblesse est qu'elle est fortement influencée par les valeurs extrêmes dans le cas où les données en contiennent

III.9.7. Tableau comparatif des paramètres de position

Mode	Médiane	Moyenne
• Le mode peut ne peut pas exister	• La médiane existe toujours	• La moyenne existe toujours
• Il peut y avoir plus d'un mode dans une distribution	• Elle provient d'une conception simple	• Elle est une mesure la plus familière
• il est facile à déterminer	• Elle est difficile à exprimer algébriquement que la moyenne	• Elle s'exprime de façon algébrique
• Il ne tient pas compte de toutes les valeurs	• Elle tient compte de la position des valeurs	• Elle tient compte de toutes les valeurs
• il peut être influencé par le choix des classes	• Elle peut être influencée par le choix des classes	• Elle est influencée par le choix des classes
• Il est significatif si la fréquence correspondante est nettement supérieure à celle des autres	• Elle est souvent utilisée lorsque la distribution des fréquences est dissymétrique	• Elle se prête facilement à la manipulation algébrique à cause de son expression mathématique simple
• il est peut stable	• Elle est moins stable par rapport à la moyenne	• Elle est plus stable par rapport aux autres paramètres de position
• il est le moins utilisé	• Elle est moins utilisée que la moyenne mais plus utilisée que le mode	• Elle est la plus utilisée que les autres

Exercice commenté

On donne la liste des notes obtenues par les élèves d'une classe à un devoir

18	8	10	10	12	15	12	10	15	18
5	7	10	11	10	9	10	10	8	9
9	10	5	11	7	10	9	4	10	12
13	15	12	15	12	13	13	12	16	

1°. Présenter cette série statistique sous forme d'un tableau en indiquant pour chaque note possible de 0 à 20 l'effectif correspondant n_i

2°. Tracer le diagramme en bâtons correspondant

3°. Calculer la moyenne $\bar{x}$ de cette série, déterminer son mode, sa médiane et ses deux quartiles

4°. On se propose de répartir les notes pour classes de la façon suivante :

Mal 0 à 3	Moyen 11 à 12
Insuffisant 4 à 7	Bien 13 à 16
Possible 8 à 10	Très bien 17 à 20

a) Donner l'effectif de chaque classe

b) Tracer l'histogramme correspondant

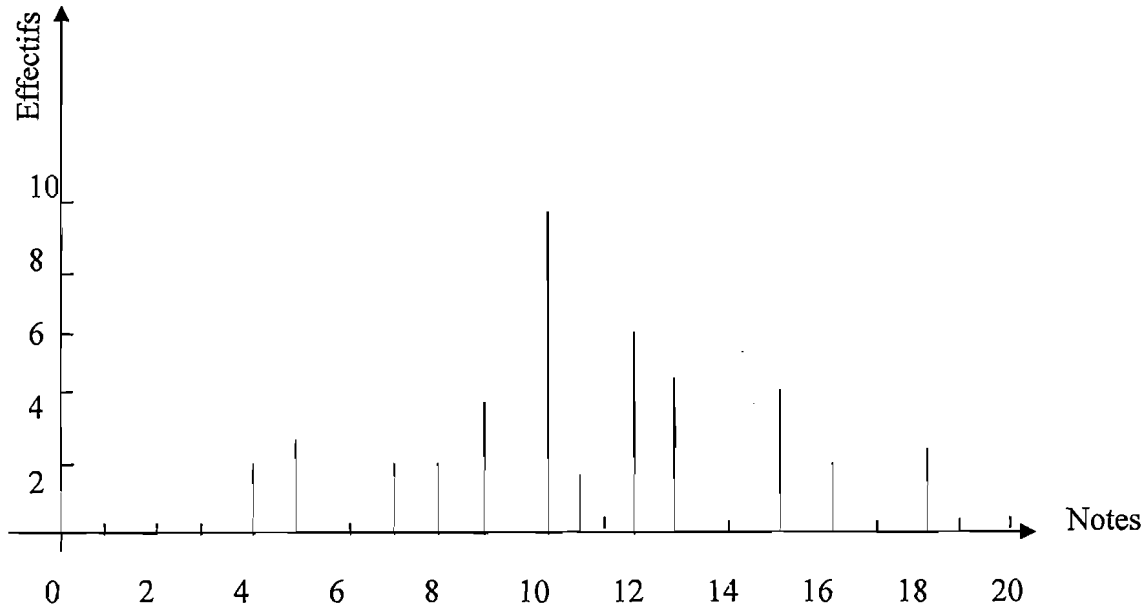
c) Donner la moyenne correspondante.

1° Réponse

x_i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n_i	0	0	0	0	1	2	0	2	2	4	10	2	6	4	0	4	1	0	1	0	0

Ce tableau est relativement facile à obtenir.

On vérifie systématiquement que la somme des effectifs de chaque classe donne bien l'effectif total (39)



3° La moyenne $\bar{x}$ est donnée par la formule

$$\frac{\sum_{i=1}^{p} x_i n_i}{\sum_{i=1}^{p} n_i}$$

Ici, on trouve $\bar{x} \cong 10{,}77$

Le mode de la série est la valeur de caractère pour laquelle l'effectif est maximum. Il est donc 10.

Nous avons onze notes inférieures ou égales à 9.

Dix-huit notes supérieures ou égales à 11.

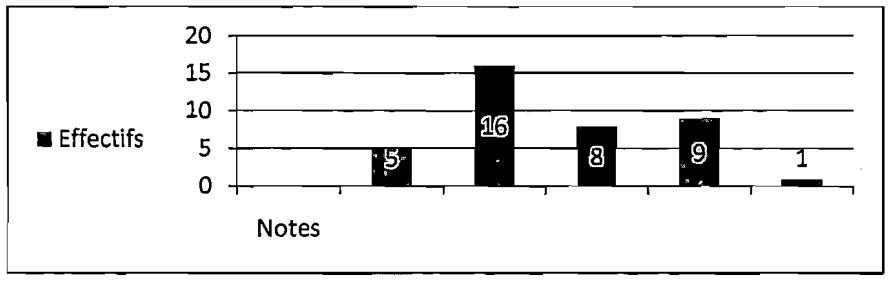
La médiane est donc 10 : Elle permet bien de partager la population totale en deux groupes :

19 élèves ont une note inférieure ou égale à 10 et les 20 autres ont une note supérieure ou égale à 10.

Le premier quartile est 9 et le 3ème quartile est 13. On a effectivement 9 élèves ayant des notes ≤ à 9, 10 autres ont des notes comprises au sens large entre 10 et 13, et les 10 derniers élèves ont une note supérieure ou égale à 13

4° a)

Classes	$0 \leq N \leq 3$	$4 \leq N \leq 7$	$8 \leq N \leq 10$	$11 \leq N \leq 12$	$13 \leq N \leq 16$	$17 \leq N \leq 20$
Effectifs	0	5	16	8	9	1



Les amplitudes des classes ne sont pas constantes, donc dans l'histogramme il n'y a pas proportionnalité entre hauteur du rectangle dessiné et effectif de la classe correspondante. La proportionnalité existe entre aire et effectif. Dans la figure précédente, on a pris1 cm² pour représenter un effectif de 4 élèves.

d) Pour calculer la moyenne, on utilise le centre c_i de chaque classe. La moyenne est alors :

$\bar{x} = \sum_{i=1}^{p} \frac{c_i n_i}{\sum_{i=1}^{p} n_i}$, ici $\bar{x}$=10,58

Exemple 2

Calculez ces différentes sortes de moyenne pour les données suivantes

1) {2 ;3 ;5}

2)

x_i	n_i
9	2
8	3
7	4

Résolution

1.a) $\bar{x}_s = \frac{\sum_{i=1}^{n} x_i}{n}$,

$= \frac{2+3+5}{3}$

$= 3{,}3$

2.a) $\bar{x}_p = \frac{\sum_{i=1}^{k} n_i x_i}{n}$

$= \frac{2.9+3.8+4.7}{2+3+4}$

$= 7{,}7$

1.b) : $\bar{x}_{G_s} = \sqrt[n]{x_1 . x_2 \dots x_n}$

$= \sqrt[n]{\prod_{i=1}^{n} x_i}$

$= \sqrt[3]{2.3.5}$

$= 3{,}10$

2.b) $\bar{x}_{G_p} = \sqrt[n]{x_1^{n1} . x_2^{n2} x_3^{n3} \dots x_k^{nk}}$

$= \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}$

$= \sqrt[9]{9^2 . 8^3 . 7^4}$

$= 7{,}13$

1.c) $\bar{x}_{h_s} = \frac{1}{\frac{1}{n}(\frac{1}{x_1}+\frac{1}{x_2}+\dots+\frac{1}{x_n})}$

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

$= \frac{n}{\sum_{i=1}\frac{1}{x_i}} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\dots+\frac{1}{x_n}\right)}$

2.c) $\bar{x}_{h_p} = \frac{1}{\frac{1}{n}(\frac{n_1}{x_1}+\frac{n_2}{x_2}+\dots+\frac{n_k}{x_n})}$

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

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

$$= \frac{1}{\frac{1}{3}(\frac{1}{2}+\frac{1}{3}+\frac{1}{5})}$$

$$= 2{,}93$$

$$= \frac{9}{\frac{2}{9}+\frac{3}{8}+\frac{4}{7}}$$

$$= 7{,}70$$

1.d) : $\bar{x}_{q_s} = \sqrt{\frac{x_1^2+x_2^2+x_3^3+\cdots+x_n^K}{n}}$

$$= (\frac{1}{n}\sum_{i=1}^{n} x_i^2)^{1/2}$$

$$= \sqrt{\frac{2^2.3^2.5^2}{3}}$$

$$= 4$$

2.d) $\bar{x}_{q_p} = \sqrt{\frac{n_1x_1^2+n_2x_2^2+\cdots n_kx_k^2}{n}}$

$$= \sqrt{\frac{2(9)^2.3(8)^2.5(7)^2}{9}}$$

$$= 1{,}8$$

Conclusion générale

Notre travail nous a permis de faire une étude comparative des paramètres de position dans une série statistique.

Nous avons vu que l'histoire de la statistique est liée étroitement au développement de l'informatique qui à son tour a provoque la naissance de nouvelles méthodes statistiques et de nouvelles procédures de calcul.

Nous avons passé en revue les principales définitions de certains termes de vocabulaire statistique.

En définitive, avant de faire la comparaison de ces paramètres, nous avons défini et donné des exemples de la moyenne, de la médiane et du mode.

Nous ne pensons pas avoir tout fait, nous invitons aux chercheurs s'intéressant à cette comparaison ou à ces paramètres de me compléter.

Bibliographies

Ouvrage généraux

[1]. Alain PILLER, *Statistique descriptive*, Manuel d'exercices corrigés avec rappels de cours, Paris, 33, rue Galilée, 75116, 1957.

[2]. André VESEREAU, *La statistique « que sais-je ? »* le point des connaissances actuelles n°281, Presses Universitaires de France 108, Boulevard SAINT-GERMAIN PARIS, 1962.

[3]. B. Verlant et G Saint-Pierre, *Statistique et habilités*, Editions FOUCHER, Paris 1997

[4]. GUYBONTEMPS, Bernard RANDE, René SEROUX, Pierre COMPAGNON, *Mathématiques Géométrie 1ère S et E*, Editions BORDAS, Paris 1988.

[5]. JAAK VERLOOY, *Statistique descriptive*, de Neder Landsche Boekhandel, 1972.

[6]. Michel JAMBU, *Méthodes de base et l'analyse des données*, collection technique et scientifique des télécommunications, Editions Eyrolles et France Telecom-Cent 1999.

[7]. Pierre DAGNELIE, *Statistique théorique et applique*, Tome 1 statistique descriptive et bases de l'inférence statistique, 2ème Ed. Paris et Bruxelles, De BOECK et Larcier, 2007

[8]. CHRISTIAN GAUTIER, JEAN CLAUDE MARTIN, *Mathématiques première A_1 et B* Editions HACHETTE, Paris 1991

[9]. Martine QUINIO BENAMO, probabilités et statistique aujourd'hui, pourquoi faire ? Comment faire ? l'harmatton, Paris 2007

[10]. Jean CHRISTOPHE François et Claude GRASLAND, la statistique et la cartographie en géographie, Université Paris VII. Deug de Géographie 1ère année, 1999 – 2000

Mémoires

[11]. Benoît SAKUBU, *Enseignement de la statistique de 3ème des humanités générales du Burundi*, U.B, I.P.A ; A/A 2007

[12]. NITUNGA Aimée Aline, Régine NIYOYUNGURUZA, *Enseignement de la statistique au secondaire*, U.B, I.P.A ; A/A 2000 - 2001