CSI- Afrique – Renforcement des interventions dans le domaine de la Politique économique et sociale

Atelier de développement des compétences des chercheurs des syndicats affiliés, Lomé, 3 au 7 mai 2010

Statistique pour les syndicalistes

Par
Kokou Baninganti
Ph.D. Statistics

La statistique est la science qui élabore et applique où besoin est, les méthodes de collecte, de mise en forme, de traitement et d'analyse des données en vue d'éclairer une prise de décision. L'approche statistique fait quelquefois figure de loi, ce qui donne parfois l'impression que l'on peut utiliser les données pour leur faire dire ce qu'elles ne disent pas. Cependant, il faut reconnaître que, dans la société, le Gouvernement, le Patronat, et les Syndicats se servent de la statistique pour orienter leurs actions dans le sens de leurs intérêts. C'est pour cette raison que la production statistique se situe au centre des luttes de classes. Par conséquent, la maîtrise des compétences statistiques constitue un enjeu de taille pour les organisations syndicales.

En effet, dans leurs activités d'organisation, de revendication et de négociation et de prévision, les responsables et chercheurs syndicalistes collectent, manipulent, traitent et présentent des résultats qui sont de nature statistique. L'aboutissement des revendications et négociations basées sur ces résultats dépend incontestablement de la pertinence et de la fiabilité des méthodes utilisées. Une augmentation salariale par exemple ne se fait pas dans n'importe quelle proportion car elle doit être fondée sur le pouvoir d'achat dont la détermination est conçue à partir des méthodes statistiques. Dans une entreprise, l'inégalité de la répartition des salaires (la concentration des salaires) ne peut être établie et prouvée que par des méthodes statistiques. Les taux de croissance économique, de chômage et d'inflation sont des indicateurs macroéconomiques qui ont impact sur les conditions de vie et de travail des populations. Il est donc tout-à-fait indiqué que dans leurs démarches et suggestions, les syndicalistes se prononcent de façon avertie et éclairée sur les évolutions de ces agrégats.

Il est également évident que pour utiliser toutes ces approches statistiques, il faut d'abord disposer de données statistiques, ce qui pose le problème de la nature et des sources des données à collecter. La méconnaissance des sources appropriées peut conduire à une utilisation erronée ou inappropriée des données statistiques.

Ainsi, pour faire un usage éclairé et averti des résultats des traitements statistiques, les syndicalistes, doivent maîtriser les éléments de base de collecte, d'analyse, d'interprétation et de présentation des données statistiques qui, en réalité, sont devenus depuis longtemps, un minimum de connaissances professionnelles pour tout spécialiste.

La présente session intitulée « Statistique Elémentaire pour Syndicalistes » a donc toute sa place dans cet atelier de développement des compétences des chercheurs des syndicats affiliés au CSI. Elle vise à mettre à la disposition des chercheurs syndicalistes des approches et méthodes simples mais efficaces de collecte, de traitement et d'analyse des données pour soutenir une recommandation ou une politique lors d'une prise de décision.

La session est structurée en 6 modules :

- Module 1- Concept et rôle de la statistique
- Module 2- Les types de données et les méthodes de leur collecte
- Module 3- Les sources des données

- Module 4- Le traitement et l'analyse des données
- Module 5- Traduction des résultats d'analyse en recommandation ou politique
- Module 6-Travaux pratiques

Le premier module définit le concept de la statistique, son rôle et son langage, c'est-à-dire sa terminologie de base.
Le deuxième module définit les types de données et les méthodes de leur collecte.
Le troisième module décrit et présente les sources essentielles des données secondaires.
Le module quatre aborde les méthodes de base de traitement et d'analyse des données en présentant les différentes synthèses des données : synthèse par tableaux, synthèse par graphiques, synthèse par caractéristiques numériques, synthèse par carte et synthèse par photographie. Il examine également les trois méthodes d'analyse des données, à savoir : les méthodes univariées, les méthodes bivariées et les méthodes multivariées.
Le cinquième module traite des procédures permettant de traduire les résultats des traitements et analyses réalisés en recommandations et politiques.
Le dernier module est consacré aux travaux pratiques.
Deux documents disponibles en version électronique supportent l'animation de cette session :

- Un bref exposé de la présentation
- Une présentation Power Point

Module 1 : Concept, rôle et langage de la statistique

1.1. Concept de la statistique

Née des besoins de l'Homme, la statistique est l'une des sciences que l'on peut qualifier d'universelle. Dans la vie quotidienne, chaque individu réalise une multitude d'activités qui renvoient à la notion de statistique.
Lorsqu'une personne se fixe comme objectif de traverser la chaussée pour se rendre par exemple à son lieu de travail elle s'arrête, regarde à droite et à gauche, évalue le nombre de voitures, de motos, de cyclistes et de personnes qui arrivent. Mieux encore, elle mesure les vitesses auxquelles ces voitures, motos et vélos roulent et puis calcule la probabilité de traverser la chaussée sans être écrasée et lorsque cette probabilité est suffisamment grande, elle décide de traverser. Tout individu procède de cette manière sans que personne ne fasse de commentaire spécial. Mais en réalité, la personne a procédé à la collecte, au traitement à l'analyse de toute une masse de données avant de prendre la décision de traverser la chaussée.
Lorsqu'un patient se présente devant un médecin, celui-ci, après les salutations d'usage et après avoir écouté l'histoire de la maladie, il prend la température, la tension artérielle, le poids, la taille, le lieu de résidence, écoute le battement du cœur et le fonctionnement des poumons et demande au besoin des analyses qui, en réalité, complètent la procédure de collecte des données. C'est seulement après analyse des données collectées que le médecin prend la décision de prescrire un produit précis pour soigner le patient.
Le Syndicat National des Enseignants du Togo constate que le salaire mensuel (toutes catégories confondues) de ses membres ne permet plus de couvrir leurs besoins, que les effectifs des classes sont pléthoriques et rendent difficiles les acquisitions des élèves et que faute de formation professionnelle initiale, les enseignants nouvellement engagés ont des faiblesses pédagogiques et qu'au total le système éducatif est négativement affecté. Suite à une telle situation, le Syndicat décide de lancer un mot d'ordre de grève pour attirer l'attention des autorités en charge de l'éducation sur les conditions de vie et travail des enseignants.

Nous remarquons que dans les exemples ci-dessus, il y a collecte, traitement et analyse de données et puis prise de décision. Ce sont là, les mots clés de la définition du concept de statistique.

La statistique est la science qui élabore et applique où besoin est, des méthodes de collecte, de traitement et d'analyse de données pour éclairer une décision en terme de recommandation et de politique.

1.2. Rôle de la Statistique

Sans regarder à droite et à gauche pour collecter et traiter les données avant de traverser la chaussée, la personne dont il a été question plus haut va plutôt, dans le meilleur de cas, se retrouver à la traumatologie du Centre Hospitalier.
Le médecin qui, sans collecte de données sur le patient, lui tend une ordonnance court un grand risque d'aggraver l'état du malade.
Et de la même manière, lorsque les syndicats ne disposent pas de statistiques pour justifier leur assertion sur le niveau des salaires, les effectifs pléthoriques et l'absence de formation professionnelle initiale des enseignants nouvellement engagés, ils auront des difficultés à négocier avec les autorités.
C'est certainement la manière la plus pratique de montrer le rôle de la statistique.
Aujourd'hui, le rôle de statistique a grandi au point où elle est devenue universelle car utilisée dans tous les domaines de l'activité de l'Homme. C'est pour cela que l'on pense que la statistique peut être considérée comme un minimum de connaissances professionnelles pour tout spécialiste.

1.3. Langage de la statistique

Comme toute science, la statistique possède son langage, sa terminologie qui véhicule les connaissances de ce domaine.

La population statistique

Généralement lorsqu'un problème se pose et nécessite comme dans bien de cas la collecte des données, il faut tout d'abord identifier la population statistique concernée.
On appelle population statistique (collectivité statistique ou univers statistique) un ensemble d'individus (au sens large), d'objets, de choses ou de phénomènes qui possèdent des caractéristiques communes.
Problème : Mesure des acquisitions des élèves du cours primaire au Togo
Population Statistique : Ensemble des élèves du cours primaires au Togo
Problème : Etude du niveau de vie des travailleurs au Togo
Population statistique : Ensemble des travailleurs du Togo
Problème : Mesure de la pauvreté au Togo
Population statistique : Ensemble des ménages au Togo

Unité statistique

C'est un élément constitutif de la population statistique. L'unité statistique est encore appelée observation statistique. L'unité statistique est l'élément porteur d'information
Problème : Mesure des acquisitions des élèves du cours primaire au Togo
Population Statistique : Ensemble des élèves du cours primaires au Togo
Unité statistique : Chaque élève du cours primaire au Togo constitue une unité statistique

Problème : Etude du niveau de vie des travailleurs au Togo
Population statistique : Ensemble des travailleurs du Togo
Unité statistique : Chaque travailleur togolais constitue une unité statistique

Problème : Mesure de la pauvreté au Togo
Population statistique : Ensemble des ménages au Togo
Unité statistique : chaque ménage au Togo est une unité statistique

Enquête statistique par sondage

Une enquête par sondage est une opération statistique de collecte des données sur une partie seulement de la population statistique. Les différentes manières de choisir ces unités définissent les méthodes de sondage.

Enquête exhaustive ou recensement

Une enquête statistique est dite exhaustive si elle porte sur toutes les unités la population statistique. Ce genre d'enquête est encore appelé recensement. L'exemple le plus connu est le recensement général de la population et de l'habitat.

Échantillon statistique

Lorsque l'on choisit seulement certaines unités statistiques d'une population, on parle d'échantillon statistique.
Un échantillon statistique est l'ensemble des d'unités statistiques d'une enquête par sondage. Le nombre de ces unités constitue la taille de l'échantillon. Le problème fondamental d'un échantillon statistique est le degré de sa représentativité par rapport à la population statistique. Deux moments importants méritent d'être mis en exergue lorsque l'on parle d'échantillon statistique : la sélection des unités statistiques à enquêter et la précision de l'enquête statistique.
Pour le choix des unités à interroger lors d'une enquête on se réfère à deux méthodes : la méthode empirique et la méthode probabiliste.
Lorsque le degré d'homogénéité de la population au regard de la question posée est jugé suffisant, l'on peut se contenter d'un échantillon aléatoire simple. Dans le cas contraire, il est vivement recommandé de recourir à la stratification de l'échantillon afin d'en accroître la représentativité. Bien que chaque strate puisse être considérée comme un échantillon aléatoire, la procédure en elle-même reste empirique.
Il est important d'avoir toujours à l'esprit que l'objectif que l'on poursuit en échantillonnant les données est d'obtenir une entité la plus représentative de la population statistique avec la moindre marge d'erreur ou le moindre risque d'erreur et surtout au moindre coût. Mais dans la pratique, ces contraintes ne sont pas faciles à concilier. Le plus souvent la contrainte financière force le statisticien à la réduction du seuil de signification.
Dans la pratique des enquêtes statistiques, au Togo au cours de la dernière décennie, l'on a réalisé un certain nombre d'enquêtes échantillonnées : Les Enquêtes Démographiques et Sociales (EDS1 et EDS2), l'enquête MICS et l'enquête QUIBB. En l'absence de données statistiques actuelles, ce sont les données de ces enquêtes qui constituent les données statistiques de référence au Togo.

Caractère statistique

Une unité statistique peut être vue et décrite de plusieurs manières. Ces différents aspects des unités statistiques constituent des caractères statistiques. Ces facettes selon lesquelles l'on peut décrire et analyser une unité statistique sont appelées caractères statistiques.
Pour la population des ménages au Togo, chaque ménage peut par exemple être décrit par sa taille (le nombre de personnes vivant dans le ménage), la région de localisation du ménage, le sexe du chef de ménage, le revenu du chef de ménage, affiliation syndicale du chef de ménage etc.

Modalité statistique

La modalité d'un caractère statistique est une situation possible dans laquelle peut se trouver ce caractère. C'est l'une des réponses possibles à une question reflétant la nature d'un caractère statistique.
Si nous prenons par exemple le caractère le sexe du chef de ménage, nous pouvons identifier deux modalités de ce caractère (Masculin et Féminin). S'il s'agit de la région de localisation du ménage, il y a cinq modalités : Maritime, Plateaux, Centrale, Kara et Savanes.

Variable statistique

Lorsque l'on associe une modalité à un caractère statistique, on obtient une variable statistique.
Dans la pratique, de nos jours, caractère et variable statistiques sont assimilables. Tout caractère statistique est donc considéré comme une variable statistique.

Caractère quantitatif ou variable quantitative

Une variable statistique est dite quantitative si ses modalités sont quantifiables et mesurables. Elle évoque immédiatement une unité physique de mesure. Exemple l'âge est une variable quantitative et son unité de mesure est l'année. Le revenu du chef de ménage est également une variable quantitative et son unité de mesure est le francs CFA.

Variable continue

C'est une variable quantitative dont les modalités prennent des valeurs de l'ensemble R des réels. Les modalités d'une variable continue sont généralement données sous forme d'intervalles. L'exemple le plus connu est l'âge qui est souvent présenté avec des modalités d'intervalles ou de classes de 5 ans. (0-5 ; 5-10 ; 10-15 etc.)

Variable discrète

C'est une variable quantitative dont les modalités prennent les valeurs dans l'ensemble N des entiers naturelles. C'est le cas par exemple du nombre de frère et sœurs où les valeurs de la variable sont 0, 1, 2 etc.

Variable qualitative ou caractère qualitatif

Une variable statistique qualitative est une variable dont les modalités ne sont pas mesurables. Exemple : le sexe d'un étudiant, la région de provenance d'un ouvrier etc.

Pour la collecte des données, il est plus commode de coder les modalités des variables qualitatives. Dans le cas du sexe, masculin prend le code 1 et féminin le code 2. Les codes sont choisis arbitrairement. Il ne faut surtout pas confondre le code d'une variable qualitative avec la valeur d'une variable quantitative.

Variable ordinale

Une variable ordinale est une variable qualitative dont les modalités évoquent un certain ordre. C'est par exemple le cas de la mention du baccalauréat d'un étudiant : passable, assez-bien, bien et très-bien. Le prix d'un téléphone portable peut être : abordable, cher, très cher.

Variable nominale

Une variable nominale est une variable qualitative dont les modalités ne possèdent pas une relation d'ordre. Toute variable qualitative est considérée comme une variable nominale.

Variable date

C'est une variable dont les modalités sont représentées par des dates sous un format précis.
Exemple :JJ/mm/aa ou JJ/mm/aaaa ou encore 13 Avril 1988.
01/02/98 ou 01/02/1998 ou 1er Février 1998

Variable dichotomique

On appelle variable dichotomique une variable dont les modalités sont donnés par les alternatives de type Vrai/faux, Oui/non

Distribution/Répartition statistique

Lorsque l'on associe dans un tableau les modalités d'un variable à ses effectifs ou à ses effectifs cumulés on obtient une distribution ou une répartition statistique.

Effectif

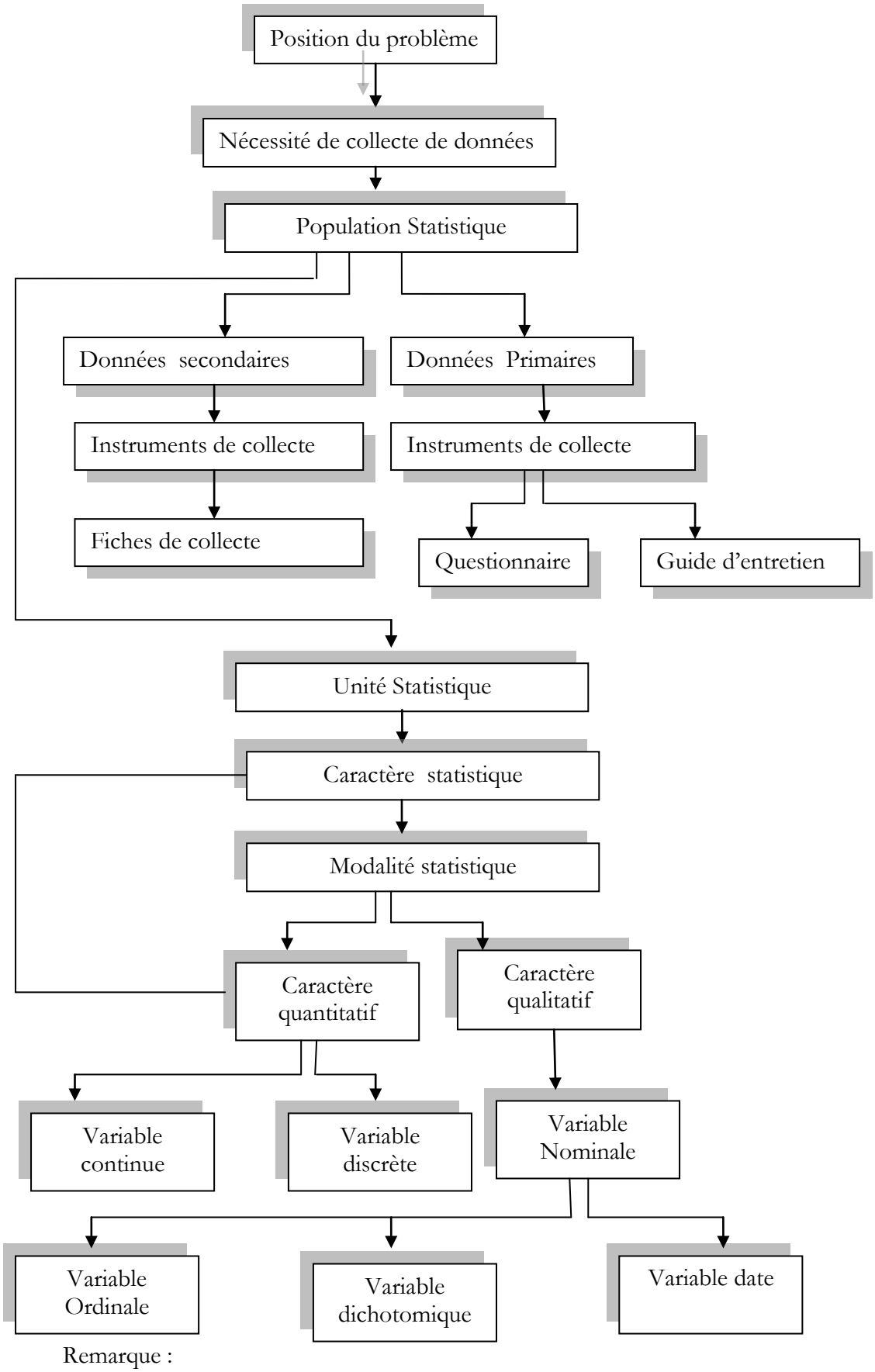
On appelle effectif le nombre d'unités statistiques possédant une modalité donnée. Ils sont généralement notés ni
Fréquence

C'est la valeur en % (notée fi) du rapport de l'effectif au nombre total d'observations.

Effectifs cumulés et fréquences cumulées

Ils sont obtenus à partir du cumul des effectifs tandis que les fréquences cumulées sont obtenues à l'aide du cumul des fréquences.
Diagramme : Définition et éléments de terminologie de base de la statistique

Diagramme : Définition et éléments de terminologie de base de la statistique



Remarque :

Le terme statistique au singulier signifie deux choses : i) la science statistique et ii) un test statistique ou une caractéristique statistique numérique.

Module 2 : Les types de données et les méthodes de leur collecte

2.1. Deux types de données : les données secondaires et les données primaires

Dans la pratique, on distingue deux types de données statistiques : les données dites secondaires et les données primaires.
Les données secondaires sont les données qui ne sont pas produites par le consommateur ou l'utilisateur, tandis que les données primaires sont produites par lui.
Lorsqu'un chercheur syndicaliste fait recours à des données du Ministère de la Fonction Publique pour analyser l'évolution des travailleurs qui y émargent, il utilise des données secondaires. Mais lorsqu'une équipe de chercheurs syndicalistes organise une enquête auprès des ménages pour recueillir les données sur leur niveau de vie, elle dispose de données primaires.
Les données secondaires doivent avoir une source précise et doivent être vérifiées avant toute utilisation. Ces données sont généralement moins coûteuses, mais ont l'inconvénient de ne pas toujours satisfaire les besoins de l'utilisateur.
Les données primaires sont plus difficiles à collecter et sont liées à des coûts plus élevés, mais elles ont l'avantage de répondre aux besoins de l'utilisateur.

2.2. Les instruments de collecte des données

Pour la collecte des données secondaires, qui sont des données qui existent déjà sous une forme précise, généralement l'instrument de collecte est choisi en fonction du format existant.
Pour la collecte des données primaires, l'on utilise souvent un questionnaire. Le questionnaire est administré aux répondants de plusieurs manières : directement par un agent de collecte de données (questionnaire directe), par poste (questionnaire postal), par téléphone (questionnaire téléphonique) ou par Internet (questionnaire électronique). Le choix de l'instrument dépend de la taille de l'échantillon, des moyens matériels investis dans la collecte des données, de la géographie concernée et des délais impartis pour disposer des données. Naturellement chaque type de questionnaire a ses avantages et ses inconvénients.
L'instrument le plus utilisé dans la collecte des données est le questionnaire qui doit viser un objectif précis, être structuré avec des questions claires (qui n'indisposent pas ou qui ne choquent pas et ne sèment pas de confusion dans l'esprit du répondant) et être le plus court possible.

2.3. L'échantillonnage

Lorsqu'une population statistique est définie et qu'il faut collecter les données, dans beaucoup de cas, pour des raisons précises, il n'est pas possible d'étudier toute la population entière. On est souvent contraint à l'échantillonnage. En tirant un échantillon, l'on doit pouvoir extrapoler les résultats de l'échantillon à la population entière. Pour ce faire, l'échantillon doit être représentatif de la population entière. Et pour que l'échantillon soit représentatif de la population entière, il doit être aléatoire, c'est-à-dire qu'il doit être obtenu par l'intermédiaire d'un mécanisme probabiliste qui définit clairement une probabilité unique qu'une unité donnée de la population soit incluse dans

l'échantillon. Un échantillon aléatoire peut être construit à l'aide d'une table de nombres aléatoires. La méthode d'échantillonnage aléatoire simple ne peut être utilisé que si les unités de la population peuvent être facilement identifiées et si cette population est relativement petite et homogène. Dans le cas d'une grande population, la méthode devient fastidieuse et onéreuse.
Dans bien de cas, lorsque l'échantillon aléatoire simple n'est pas approprié, l'on peut recourir à d'autres types d'échantillons : échantillon stratifié, l'échantillon systématique ou à l'échantillon par grappes.
Pour obtenir un échantillon stratifié, il faut diviser la population en sous-groupes plus ou moins homogènes appelés strates et tirer dans chaque strate un échantillon aléatoire.
L'échantillon systématique est constitué d'unités choisies dans la population à des intervalles fixes en termes de temps, d'espace ou d'ordre d'occurrence. (Exemple : choisir 10 syndicalistes sur une liste de 110 syndicalistes, dans le contrôle de la qualité on choisit un boulon après chaque lot de 50 boulons). L'échantillonnage systématique est utilisé lorsque les unités de la population sont déjà arrangées dans un certain ordre.
Comme nous pouvons le remarquer, l'échantillonnage aléatoire simple ou stratifié exige que toutes les unités statistiques soient listées, ce qui n'est pas toujours possible. C'est dans ces conditions que l'on fait appel à l'échantillonnage par grappes qui consiste à :

- subdiviser la population en sous-groupes appelés grappes
- tirer un échantillon aléatoire de grappes
- inclure toutes les unités des grappes retenues dans l'échantillon

Mais un échantillon non aléatoire est composé d'éléments choisis sur la base des connaissances et de l'expérience personnelles de l'analyste. Dans ces conditions, on parle de choix raisonné dans la constitution de l'échantillon.

Module 3 : Les sources des données

L'effectif des syndicalistes du Syndicat National des Enseignants du Togo est de 3548 adhérents au titre de l'année 2009.
Telle que fournie, une telle donnée ne peut pas être considérée comme fiable. Il s'agit d'une donnée qui n'informe pas suffisamment. Par contre si nous disons : Selon, les Statistiques de la cellule d'étude et de recherche du Syndicat National des Enseignants du Togo au 31 Décembre 2009, l'effectif des adhérents est de 3548, nous avons une information complète et plus fiable parce que la source des données est clairement précisée.
Ainsi, une utilisation avisée des données statistiques doit obligatoirement en donner la source. Lorsque les sources ne sont pas indiquées, l'on suppose que l'utilisateur est le producteur des statistiques auxquelles il fait référence.
La non mention des sources des données statistiques peut avoir des conséquences plus fâcheuses, surtout quand il s'agit de statistiques politiques. Certains journalistes ont fait les frais de tels usages des données statistiques.
La source des données statistiques c'est l'institution ou l'individu qui a produit ces statistiques et qui en est l'auteur, qui en a la paternité et donc qui en est responsable.
Au Togo, différentes données statistiques sont produites par plusieurs institutions nationales.
Selon les directives des organes spécialisés des Nations unies pour les questions de population, dans chaque pays, tous les 10 ans on doit organiser un recensement général de la population et de l'habitat. Le dernier recensement de la population et de l'habitat date de 1981. Le prochain recensement est en cours de préparation.
Outre le recensement de la population et de l'habitat, des enquêtes nationales spécialisées sont organisées, dans le cadre d'études telles que les Enquêtes Démographiques et

Sociales (EDS), Enquêtes QUIBB, Enquêtes MICS. Ces Enquêtes, réalisées par la Direction Générale de la Statistique et de la Comptabilité Nationale (DGSCN) ont été financées par des organismes internationaux et constituent des sources précieuses de données démographiques, économiques et sociales sur le pays.

Tableau : Source des données statistiques

Types de données	**Ministère**	**Direction**
Santé	Santé	Direction de la Santé
Education	Education	Direction Générale de la Planification de l'Education
Agriculture	Agriculture, de l'élevage et de la Pêche	Agriculture, de l'élevage et de la Pêche
Industrie	Economie	Economie
Economie	Economie	Direction de l'économie, du Budget
Travail et emploi	Travail et Fonction publique	Direction de la Fonction publique, observatoire de l'emploi
Commerce et tourisme	Commerce	Commerce

Depuis quelques décennies, certains organismes internationaux (Banque mondiale, Fonds Monétaire Internationale, Banque Africaine de Développement, Union Africaine, Organisation Internationale du Travail etc.) en collaboration avec les services statistiques des pays, ont commencé à publier des bases de données, structurés couvrant de longues périodes et plusieurs indicateurs macroéconomiques. Le format le plus connu de ces publications est le CD Room que l'on installe et qui permet de rechercher, visualiser, imprimer et copier les données désirées. Les sites de ces organismes sont faciles à identifier.

On parle beaucoup ces derniers temps de bases de données internationales. Il s'agit en réalité de données colligées par certains organismes internationaux mises en ligne et diffusées sur Internet et accessibles aux utilisateurs. Ces bases de données constituent les seules sources de données généralement actuelles disponibles pour les chercheurs. Il s'agit essentiellement de :

WDI-World Development Indicators
ADI-African Development Indicators
IFS-International Financial Statistics site:http\www.imf.org

Module 4- Le traitement et l'analyse des données

La collecte des données constitue la toute première étape de toute approche statistique. Les données obtenues ne sont pas forcement sous le format approprié. Ces données peuvent être en vrac, sous forme de tableaux ou de tables. La forme la plus usuelle des données est le fichier de données qui est un tableau dont les colonnes représentent les variables et les lignes observations ou les unités statistiques. Au croisement des lignes et des colonnes on retrouve soit les valeurs des variables (pour les variables quantitatives) ou soit les codes des modalités (pour les variables qualitatives).

Le traitement des données consiste à synthétiser les données du fichier de données. Cette synthèse est multiforme : tableaux, graphiques, numérique, photographie, cartes.

4.1. La synthèse par tableau simple : Cas d'une variable statistique quantitative continue

Les données

Dans le cadre d'un programme de prise en charge des dépenses liées à la santé, le service socio médical du Syndicat des enseignants du Supérieur a collecté les données ci-après sur la taille (en centimètres) d'un échantillon de 50 syndicalistes.

158	172	166	170	168	175	152	190	191	157
163	160	149	186	188	172	173	184	181	180
171	169	171	173	171	180	198	167	175	177
170	173	168	167	169	180	181	178	166	164
160	168	166	162	170	182	183	190	167	169

Questions

1. Que peut-on dire de ces données ?
2. Synthétiser ces données en élaborant un tableau de distribution avec des amplitudes de 5 cm [145 à 150]…….. [195 à 200]
3. Que peut-on conclure ?

1. Sous cette forme, l'on ne peut rien dire sur les données recueillies.
2. Synthèse des données par tableau

Tableau : Distribution d'un échantillon de syndicalistes selon la taille (cm)

Classes	**Comptage**	**Effectifs**	**Effectifs cumulés**	**Fréquences**	**Fréquences cumulées**
[145-150]	I	1	1	0,02	0,02
[150-155]	I	1	2	0,02	0,04
[155-160]	II	2	4	0,04	0,08
[160-165]	IIIII	5	9	0,10	0,18
[165-170]	IIIII IIIII II	12	21	0,24	0,42
[170-175]	IIIII IIIII I	11	32	0,22	0,64
[175-180]	IIII	4	36	0,08	0,72
[180-185]	IIIII III	8	44	0,16	0,88
[185-190]	II	2	46	0,04	0,92
[190-195]	III	3	49	0,06	0,98
[195-200]	I	1	50	0,02	1,00
Total		50		1,00	

Que peut-on dire sur le tableau ?
- Il y a relativement peu de syndicalistes de taille très courtes [145-160]. Ils ne sont 4 (soit 8%) à avoir une taille égale ou inférieure à 160 cm.
- Ils sont relativement plus nombreux 40 (soit 80%) à avoir une taille comprise entre 160 et 185 cm
- On constate également qu'ils ne sont pas du nombreux à avoir de très grandes tailles (180 à 200 cm). Ils ne sont que 6 (soit 12%)

4.2. Synthèse pat tableau simple: Cas d'une variable statistique quantitative discrète

Les données

La taille des ménages d'un groupe de syndicalistes (Personnes)

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

Questions
1-Que peut-on dire de ces données ?
2-Etablir une synthèse par tableau de ces données en considérant la variable comme une variable statistique quantitative discrète.
3-Que peut-on dire de cette synthèse ?

Tableau
Synthèse par tableau simple d'un groupe de syndicalistes selon la taille du ménage (Personnes)

Taille du ménage	Comptage	Effectifs	Effectifs cumulés	Fréquences (en %)	Fréquences cumulées (%)
2	IIIII	5	5	10,4	10,4
3	IIIII I	6	11	12,5	22,9
4	IIIII III	8	19	16,7	39,6
5	IIIII IIIII II	12	31	25,0	64,6
6	IIIII III	8	39	16,7	81,3
7	III	3	42	6,2	87,5
8	IIII	4	46	8,3	95,8
9	II	2	48	4,2	100,0
Total		48		100,0	

1. Comme précédemment, les données en vrac sont muettes et donc ne permettent de faire aucune analyse.
3. Par contre la synthèse par tableau donne l'occasion de se rendre compte qu'il y a très peu de syndicalistes (22,9%) ayant des ménages de petites taille (2 à 3 personnes) et également très peu de syndicalistes (12,5%) avec des ménages de grande taille (8 à 9 personnes). Une très large majorité des chefs de ménage (58,4%) vivent dans des ménages comptant 4 à 6 personnes.

4.3. Synthèse pat tableau croisé: Cas de deux variables qualitatives

Dans la pratique de l'analyse statistique des données, afin d'approfondir l'analyse d'un phénomène, l'on procède à des synthèses croisées, appelées encore tableaux croisés ou tableaux à plusieurs entrées. Il s'agit d'une synthèse simultanée de deux ou plusieurs variables qui peuvent être qualitatives ou quantitatives. Le développement des tableaux croisés a donné naissance à des méthodes très performantes d'analyse statistique des phénomènes économiques. Les exemples les plus connus sont les tableaux input output du professeur Wassily Léontiev et les matrices de comptabilités sociales qui sont à l'origine de l'élaboration des Modèles Calculables d'Equilibre Général.
Signalons que, d'une façon générale, le croisement de deux variables vise à confirmer (par un test approprié) une relation de dépendance entre ces variables.
Ici nous allons nous contenter d'un cas très de croisement de deux variables qualitatives : la zone de résidence (Ville, Campagne) et le type d'habitation (maison en dur, maison en banco)

Les données

Dans une enquête de routine sur les conditions de vie des populations d'une localité du Togo, le syndicat national des ouvriers et des paysans a collecté des données suivantes sur la zone de résidence (R) et le type d'habitation (H) d'un groupe de travailleurs.

Tableau

La zone de résidence (R) et le type d'habitation (H) d'un groupe de travailleurs.

R	v	v	c	v	c	c	c	v	v	v	c	v	c	c	v
H	d	b	b	d	B	b	d	d	b	d	b	d	b	d	d

Les variables

R- la zone de résidence du travailleur

H- Le type d'habitat du travailleur

Les modalités

v-Ville

c-Campagne

d-Habitat en dur

b-Habitat en Banco

Tableau de comptage

	v	**c**	**Total**
d	IIIII	II	8
b	III	IIII	7
Total	9	6	15

Tableau de répartition d'un groupe de travailleurs selon la zone de résidence et le type d'habitat

	v	**c**	**Total**
d	6	2	**8**
b	3	4	**7**
Total	**9**	**6**	**15**

Que peut-on dire de ce tableau ?

Sur les 15 travailleurs enquêtés, 9 résident en ville et 6 en campagne. Parmi les 9 qui résident en ville, 6 habitent dans des maisons en dur et 3 dans des maisons en banco.

Sur le nombre total de travailleurs enquêtés 8 vivent dans des maisons en dur et 7 dans des maisons en banco. Parmi ceux qui vivent dans de maisons en dur 6 résident en ville et deux en campagne, tandis que parmi ceux qui vivent dans des maisons en banco, 3 résident en ville et 4 en campagne.

Notons que dans chaque valeur dans une cellule d'un tableau croisé peut être interprétée de plusieurs manières. Prenons par exemple 6 au croisement de d et v.

Cette valeur signifie qu'il y a 6 travailleurs résidant en ville et qui vivent dans des habitations en dur.

Nous pouvons rapporter cette valeur à 8 (nombre total des travailleurs vivant dans des maisons en dur), nous obtenons alors la proportion des travailleurs des villes vivant dans des maisons en dur dans l'ensemble des travailleurs vivant dans des maisons dur.

Nous pouvons rapporter cette valeur 6 à 9 (nombre total des travailleurs résidant en ville), ainsi nous obtenons la proportion des travailleurs résidant en ville et habitant dans des maisons en dur, dans l'ensemble des travailleurs des villes.

Enfin, nous pouvons rapporter cette valeur 6 à 15 (Au nombre total des travailleurs enquêtés), ce qui veut dire que nous l'interprétons comme étant la proportion des travailleurs résidant en ville et habitant dans des maisons en dur dans l'ensemble dans travailleurs.

Pour faire simple, nous pouvons affirmer que chaque valeur a quatre significations :

- effectif
- pourcentage par rapport au total de la ligne
- pourcentage par rapport au total de la colonne
- pourcentage par rapport au total des totaux

Règle générale pour la synthèse par tableaux croisés

Comme nous pouvons le constater, la synthèse par tableaux croisés permet d'approfondir l'analyse des données. Cependant, il faut remarquer que l'une des conditions primordiales de ce croissement est le nombre de modalités des variables en jeu. Il n'y a aucune difficulté se rendre compte qu'il n'est pas du tout utile de croiser deux variables quantitatives dont les modalités sont très nombreuses. Mais lorsqu'une variable quantitative est recodée pour prendre la forme d'une variable continue avec un nombre de classes déterminées, on peut l'utiliser dans la synthèse par tableaux croisés.
Le recodage peut donc constituer une méthode de transformation d'une variable quantitative en variable qualitative.

4.4. Synthèse par graphique

L'une des méthodes les pus anciennes de synthèse des données est la méthode graphique qui représentent les données sous forme de figures géométriques (surfaces, points, lignes, courbes etc.). Avec l'avènement et le développement des technologies d'information et de communication, l'élaboration des graphiques a connu des progrès très rapides. Il n'y a plus pratiquement de graphiques manuels. Un graphique professionnel approprié exprime et reflète toujours mieux la réalité.
A l'heure des nouvelles technologies de traitement et de présentation de l'information, il n'y a plus de graphique manuel. Désormais toutes les représentations graphiques sont élaborées de façon informatisée et automatisée. Les diagrammes ci-dessous présentées ont été obtenus à l'aide de logiciels statistiques professionnels.

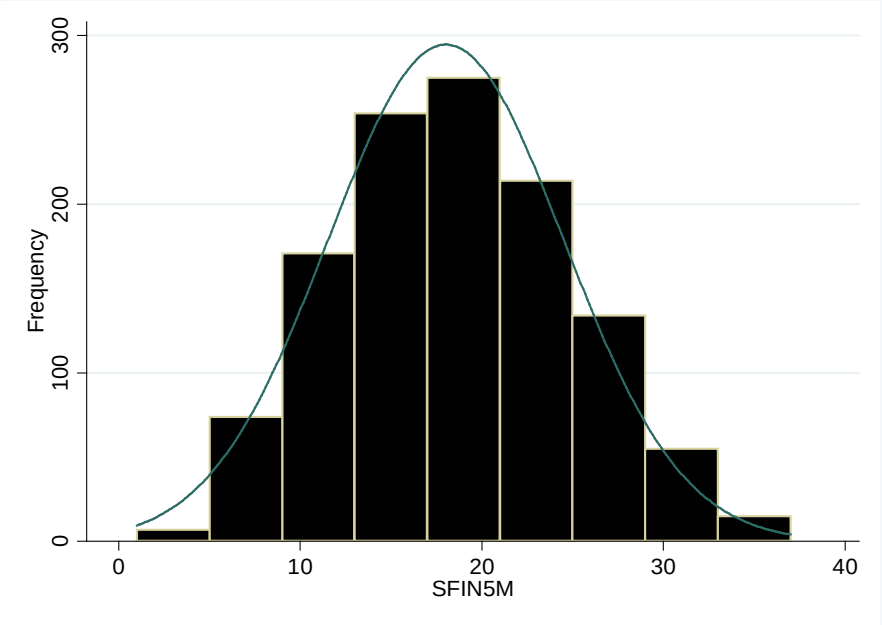
Histogramme

En statistique appliquée, l'on utilise beaucoup la synthèse graphique par histogramme. Elle constitue la représentation la plus indiquée des variables statistiques quantitatives continues. L'élaboration d'un histogramme vise également à comparer la distribution empirique à la distribution théorique correspondante. La distribution théorique la plus connue est la distribution de la Place- Gauss appelée encore la courbe en cloche ou la courbe normale. Il s'agit d'une distribution qui présente des effectifs relativement peu élevés dans les modalités inférieures, des effectifs assez élevés dans les modalités intermédiaires et des modalités également faibles dans les modalités supérieures. (Peu, beaucoup, peu).
La courbe normale est une courbe de référence en statistique appliquée. Lorsqu'une variable est le résultat de l'interaction d'une multitude de facteurs, elle peut être décrite par une loi de la Place-Gauss. Beaucoup de phénomènes se moulent dans cette loi. C'est le cas par exemple des scores enregistrés par un groupe d'apprenants lors d'une évaluation.
L'exemple présenté ci-dessous est le cas d'une distribution normale symétrique. Dans la réalité, les distributions sont étalées soit à droite ou à gauche. Ces écarts par rapports à la distribution normale mesure le degré d'asymétrie (étalée à droite ou à gauche) et degré d'aplatissement (plus aplatie ou plus pointue que la distribution normale). Les mesures

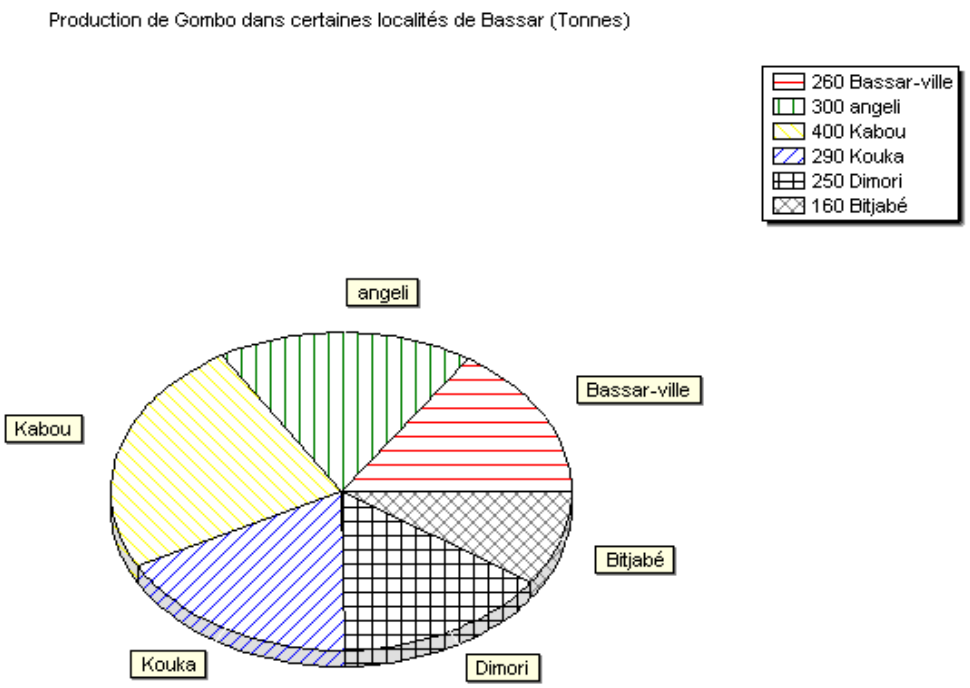
d'asymétrie et d'aplatissement constituent les caractéristiques de forme d'une distribution statistique.

Graphique

Histogramme de la répartition des élèves du CM1 au Togo selon le score final de calcul lors de l'évaluation du PASEC en 2000.



Le diagramme à secteurs

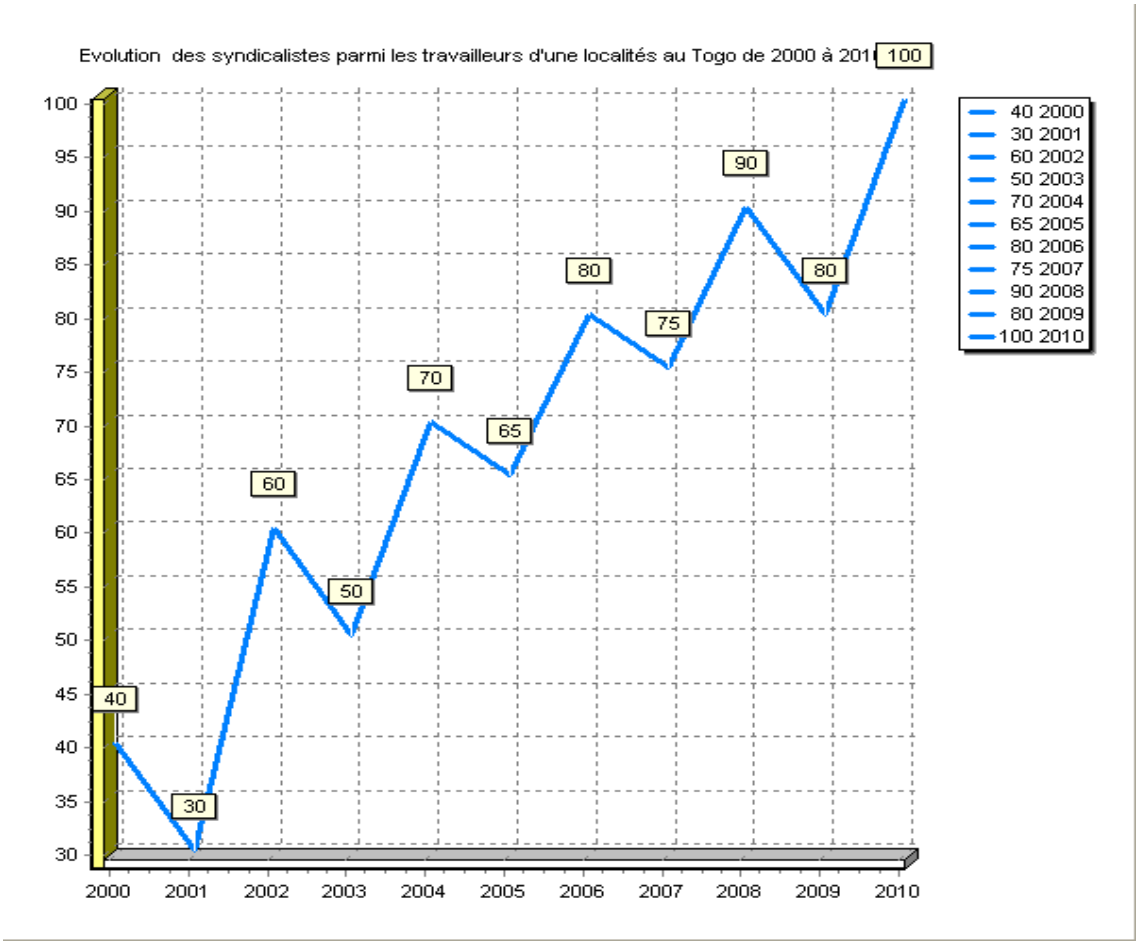


Lorsque l'on dispose d'un tableau de distribution, ce tableau peut être transformé en diagramme à secteurs. Chaque secteur a une valeur en degrés proportionnelle à la valeur correspondante dans le tableau.

$$S_i^o = \frac{360^o}{\sum v_i}.v_i$$

<u>La courbe de tendance</u>

La courbe de tendance est utilisée pour représenter graphiquement l'évolution d'un phénomène dans le temps



4.5. Synthèse par caractéristiques numériques

La synthèse numérique des données statistiques se fait à l'aide des caractéristiques statistiques numériques. On appelle caractéristique numérique toute valeur algébrique construite à partir des observations. Les plus traditionnelles et les plus connues de ces caractéristiques sont les caractéristiques de tendances centrales représentées par les moyennes. On distingue les moyennes classiques et les moyennes structurelles. Les moyennes classiques sont : la moyenne harmonique, la moyenne géométrique, la moyenne arithmétique et la moyenne quadratique. Les moyennes structurelles sont : le mode ou la moyenne des fréquences et la médiane ou la moyenne de position.
En 1945, le statisticien anglais Yule a établi une liste de 6 propriétés souhaitables pour qu'une caractéristique statistique numérique soit considérée comme intéressante.
1. La caractéristique doit être définie de façon objective
2. La caractéristique doit dépendre de toutes les observations
3. La caractéristique doit avoir une signification concrète
4. La caractéristique doit être simple à calculer
5. La caractéristique doit se prêter aux calculs algébriques
6. La caractéristique doit être peu sensible aux fluctuations d'échantillonnage.

Mais, en réalité il n'y a pas beaucoup de caractéristiques statistiques qui possèdent simultanément toutes ces propriétés ou remplissent simultanément ces six conditions.
Les moyennes peuvent être calculées soit sur des données non regroupées pour obtenir les moyennes dites simples ou sur des données regroupées pour donner les moyennes pondérées. Les données sont souvent regroupées sous formes de variables discrètes ou de variables continues.

Les moyennes classiques

Définition de la moyenne

La moyenne d'une variable statistique est la valeur de cette variable qui, en même temps qu'elle donne une idée générale de la population statistique, peut remplacer chacune des valeurs de cette variable sans toutefois modifier la dimension objective du phénomène étudié.

Les moyennes simples et les moyennes pondérées

La formule générale de la moyenne simple est la suivante

$$\bar{x} = \sqrt[a]{\frac{\sum x_i^a}{n}}$$

Tandis que les moyennes pondérées peuvent être calculées par la formule ci-dessous

$$\bar{x} = \sqrt[a]{\frac{\sum x_i^a n_i}{\sum n_i}}$$

Selon la valeur attribuée à la constante a, on obtient un type donné de moyenne

Tableau

Ordre des moyennes

Valeur de a	-1	0	+1	+2
Type de moyenne	Harmonique	géométrique	arithmétique	Quadratique

La moyenne harmonique simple

La moyenne harmonique, d'une façon générale est rarement utilisée en statistique. Elle intervient surtout dans le cas des grandeurs inversement proportionnelles. Elle définie par la formule :

$$\bar{x}_h = \frac{n}{\sum \frac{1}{x_i}}$$

Exemple 1

Chaque samedi un commerçant zamarma se rend à pieds de Notsè à Wahala à une vitesse de 6km/h. Le soir il refait le même parcourt mais en sens inverse et à une vitesse de 3 km/h. Quelle sa vitesse moyenne ?

Remarquons tout de suite que la vitesse moyenne est égale à la distance totale parcourue divisée par le temps.

Soient

d- la distance entre les deux localités

t_1- le temps mis pour aller de Notsè à Wahala

t_2- le temps mis pour aller de Wahala à Notsè

v_m- la vitesse moyenne pour faire l'aller retour

$$v_m = \frac{2d}{t_1 + t_2} = \frac{2d}{\frac{d}{v_1} + \frac{d}{v_2}} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}} = \frac{2}{\frac{1}{6} + \frac{1}{3}} = 4km/h$$

Exemple 2 (Exercice à faire)

Une entreprise Ghanéenne de production de boisson fabriquée à base de plantes médicinales locales consacre un budget fixe B, chaque trimestre, à sa campagne publicitaire. Le tableau ci-dessous indique les prix de l'affiche au cours des quatre trimestres de l'année 2007.

Tableau

Les prix des affiches au cours des 4 trimestres de l'année 2007 d'une entreprise ghanéenne (en Cedis)

Trimestre	T1	T2	T3	T4
Prix de l'affiche	350	380	400	440

Calculer le prix moyen de l'affiche

La moyenne harmonique pondérée

La moyenne harmonique pondérée est de la forme analytique suivante :

$$\bar{x}_h = \frac{\sum n_i}{\sum \frac{n_i}{x_i}}$$

Exemple 3

Afin de se rendre à Takoradi (Ghana) pour assister à une réunion des syndicats de l'Afrique de l'Ouest, Monsieur Salami fait une opération de change au marché noir. Il dispose de Francs CFA et achète le Cedis Ghanéen. Il échange 18000 francs CFA au taux de 450 francs CFA pour un Cedi, 13800 francs CFA au taux de 460 francs pour un cedi et 4800 francs CFA au taux de 480 francs pour un cedi.

Quel est le taux de change moyen ?

La théorie économique définit le taux de change moyen (Tcm) comme étant égal au rapport du montant total de francs CFA échangés (M_{te}) au montant total de Cedis obtenus (M_{to}).

$$T_{cm} = \frac{M_{te}}{M_{to}}$$

$$T_{cm} = \frac{18000 + 13800 + 4800}{\frac{18000}{450} + \frac{13800}{460} + \frac{4800}{480}} = \frac{36600}{40 + 30 + 10} = 457,5$$

Ici, on constate facilement que c'est la forme pondérée de la moyenne harmonique qui a été utilisée.

La moyenne géométrique simple

La moyenne géométrique simple

Elle est donnée par la formule : $\bar{x}_g = \sqrt[n]{x_1.x_2.x_3.....x_n}$

La moyenne géométrique est essentiellement utilisée dans la pratique pour caractériser la tendance générale d'une série chronologique, c'est-à-dire d'un phénomène évoluant dans le temps. Dans ces conditions, cette moyenne prend la dénomination de taux de croissance moyen. Mais pour une utilisation appropriée de ce type de moyenne, comme l'indique son appellation, il faut que l'évolution constatée suive une tendance qui s'apparente à une progression géométrique.

Exemple 4

Tableau
Évolution des ventes d'arachide d'un paysan de Nassablé de 2004 à 2007 (en francs CFA)

Année	2004	2005	2006	2007
Notation 1	v_0	v_1	v_2	v_3
Notation 2	v_1	v_2	v_3	v_4
Vente	47 288	59 130	70 262	74 081
Taux de croissance	1,00	1,24	1,19	1,06

Calculer le taux de croissance moyen annuel des ventes d'arachides de 2004 à 2007

$$T_{cm} = \sqrt[3]{\frac{59130}{47288} * \frac{70262}{59130} * \frac{74081}{70262}} = \sqrt[3]{1,24 * 1,19 * 1,06} = 1,16$$

Pour calculer ce taux, nous venons d'utiliser la formule de la moyenne géométrique simple. Mais en réalité, en utilisant la notation 1, le calcul est fait de la manière suivante :

$$T_{cm} = \sqrt[3]{\frac{v_1}{v_0} * \frac{v_2}{v_1} * \frac{v_3}{v_2}} = \sqrt[3]{\frac{v_3}{v_0}}$$

De façon générale, en adoptant toujours la notation 1 qui attribue l'indice 0 à la première période, la formule utilisée pour calculer le taux moyen de croissance est la suivante :

$r = \sqrt[n]{\frac{y_n}{y_0}}$ Où n est l'indice de la dernière valeur.

Avec la deuxième notation, nous obtenons :

$$T_{cm} = \sqrt[3]{\frac{v_2}{v_1} * \frac{v_3}{v_2} * \frac{v_4}{v_3}} = \sqrt[3]{\frac{v_4}{v_1}}$$

Avec cette deuxième notation, la formule générale de calcul du taux de croissance moyen est la suivante :

$r = \sqrt[n-1]{\frac{y_n}{y_1}}$ Où n est l'indice de la dernière valeur

Indépendamment de la notation, le résultat du calcul est le même.

La moyenne géométrique pondérée $\overline{x}_g = \sqrt[\sum n_i]{x_1^{n_1}.x_2^{n_2}.....x_m^{n_m}}$

La moyenne arithmétique

La moyenne arithmétique est la moyenne la plus utilisée. D'une manière générale, lorsque l'on parle de moyenne sans préciser le type, il s'agit de la moyenne arithmétique.

La moyenne arithmétique simple $\bar{x} = \frac{\sum x_i}{n}$

La moyenne arithmétique pondérée $\bar{x} = \frac{\sum x_i n_i}{n}$

De plus, la moyenne arithmétique possède des propriétés mathématiques intéressantes. Notamment, la somme des écarts des valeurs d'une variable par rapport à sa moyenne arithmétique est nulle.

Propriété 1

La somme des écarts des valeurs d'une variable par rapport à sa moyenne arithmétique est nulle.

$\sum (x_i - \bar{x}) n_i = 0$

En fait, en développant la partie gauche de l'équation, nous pouvons écrire :

$\sum (x_i - \bar{x}) n_i = \sum x_i n_i - \bar{x} \sum n_i$ Or par définition, $\bar{x} = \frac{\sum x_i n_i}{\sum n_i} \rightarrow \sum x_i n_i = \bar{x} \sum n_i$ d'où finalement

$\sum (x_i - \bar{x}) n_i = \sum x_i n_i - \bar{x} \sum n_i = \bar{x} \sum n_i - \bar{x} \sum n_i = 0$

Propriété 2

Calcul de la moyenne arithmétique à l'aide de la variable auxiliaire de type : $x'_i = x_i - x_0$

$\bar{x}' = \frac{\sum x'_i n_i}{\sum n_i} = \frac{\sum (x_i - x_0) n_i}{\sum n_i} = \frac{\sum x_i n_i}{\sum n_i} - \frac{x_0 \sum n_i}{\sum n_i} = \bar{x} - x_0 \Rightarrow \bar{x} = \bar{x}' + x_0$

Exemple 5

Un groupe de cultivateurs d'Aouda a produit au cours de l'année 2006, les quantités suivantes d'ignames :
5003, 5006, 5004, 5007, 5002
Calculer la production moyenne par cultivateur
Soit $x'_i = x_i - x_0$ avec xo = 5000

Tableau de calcul

xi	5003	5006	5004	5007	5002
x'i	3	6	4	7	2

$\bar{x} = \bar{x}' + x_0 = \frac{22}{5} + 500 = 4,4 + 5000 = 5004,4$

Propriété 3

On peut également calculer la moyenne arithmétique en utilisant la variable auxiliaire de type

$$x_i^{'} = \frac{x_i}{k}$$

La moyenne de cette variable auxiliaire est égale à :

$$\overline{x'} = \frac{\sum x_i^{'} n_i}{\sum n_i} = \frac{1}{k}\left[\frac{\sum x_i n_i}{\sum n_i}\right] = \frac{1}{k}\overline{x}, \text{ alors} \qquad \overline{x} = k\overline{x'}$$

Exemple 6

Soit le salaire mensuel d'un groupe d'ouvrier d'une entreprise de la place (en francs cfa)
10000 20000 40000 30000 50000
Calculer le salaire mensuel moyen de ce groupe d'ouvriers
k=10000 $\overline{x} = k\overline{x'} = 10000 * 3 = 30000$

Propriété 4

Méthode de calcul simplifié de la moyenne arithmétique à l'aide de la variable auxiliaire de type :

$$x_i^{'} = \frac{(x_i - x_0)}{k}$$

Cette méthode combine les deux dernières propriétés de la moyenne arithmétique

$$\overline{x'} = \frac{\sum x_i^{'} n_i}{\sum n_i} = \frac{\sum\left(\frac{x_i - x_0}{k}\right) n_i}{\sum n_i} = \frac{\frac{1}{k}\left[\sum\left(x_i - x_0\right)\right] n_i}{\sum n_i} = \frac{1}{k}\left[\frac{\sum x_i n_i}{\sum n_i} - \frac{x_0 \sum n_i}{\sum n_i}\right] = \frac{1}{k}\left(\overline{x} - x_0\right) \Rightarrow \overline{x} = k\overline{x'} + x_0$$

Exemple 7

Tableau
Les ventes annuelles d'un groupe de paysans de Bitjabé (milliers de francs CFA)

Vente	x_i	n_i	$x_i n_i$	$x'_i = x_i - 35$	$x'_i n_i$	$x'_i = (x_i - 35)/5$	$x'_i n_i$	$x_i^2 n_i$
00 - 10	5	3	15	-30	-90	-6	-18	75
10 - 20	15	5	75	-20	-100	-4	-20	1125
20 - 30	25	14	350	-10	-140	-2	-28	8750
30 - 40	35	20	700	0	0	0	0	24500
40 - 50	45	18	810	10	180	2	36	36450
50 - 60	55	8	440	20	160	4	32	24200
60 - 70	65	2	130	30	60	6	12	8450
Total	-	70	2520	-	70	-	14	103550

Le choix de x_0 est arbitraire, mais il est plus bénéfique de prendre les valeurs qui permettent une simplification effective des calculs.

$$\overline{x} = \frac{\sum x_i n_i}{\sum n_i} = \frac{2520}{70} = 36 \quad \overline{x} = k\overline{x'} + x_0 = \frac{14}{70}.5 + 35 = 36 \quad \overline{x} = \overline{x'} + x_0 = \frac{70}{70} + 35 = 36$$

La moyenne quadratique

La moyenne quadratique est utilisée pour calculer la moyenne des carrés des valeurs d'une variable, par exemple dans le calcul de la variance qui est définie comme le carré de la moyenne moins la moyenne des carrés. Liée à la notion de moment, cette moyenne est également rarement utilisée dans la pratique.

La moyenne quadratique simple $\bar{x}_q = \sqrt{\frac{\sum x_i^2}{n}}$

La moyenne quadratique pondérée $\bar{x}_q = \sqrt{\frac{\sum x_i^2 n_i}{\sum n_i}}$

Exemple 7 bis

En prenant le cas des ventes d'un groupe de paysans de Bitjabé, la moyenne quadratique est égale à :

$$\bar{x}_q = \sqrt{\frac{\sum x_i^2 n_i}{\sum n_i}} = \sqrt{\frac{103550}{70}} = 38,461$$

La moyenne quadratique des ventes des paysans est égale à 38461 francs

Comme nous pouvons le remarquer, il existe toute une gamme variée de moyennes. Dans la pratique, l'on semble avoir privilégié la moyenne arithmétique. Mais en réalité, le choix du type de moyenne à utiliser dans telle ou telle circonstance, demeure un problème statistique délicat.

Les moyennes structurelles

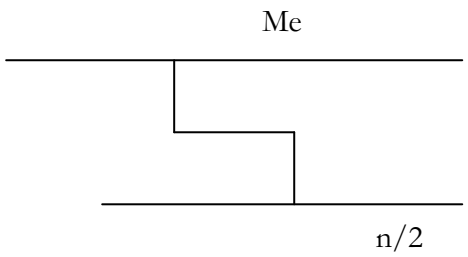
Il s'agit essentiellement de la médiane, du mode et des quantiles (Quartiles, Déciles, Percentiles)

La médiane

La médiane d'une variable statistique est la valeur de cette variable qui divise le nombre total d'observations en deux parties égales. Schématiquement cela peut s'exprimer de la manière suivante :

Définition de la médiane

Graphique



La médiane des données non regroupées

Pour déterminer la médiane dans le cas des données non regroupées, il faut :

- voir si le nombre d'observations est pair ou impair

- ranger la série des données de la variable (par ordre croissant par exemple)
- identifier la valeur qui occupe la position médiane

Ensuite établir que :

- dans le cas d'un nombre impair d'observations, il y a une seule valeur de la variable qui occupe la position médiane et qui est égale à la médiane
- dans le cas d'un nombre pair d'observations, deux valeurs de la variable occupent la position médiane et leur demi-somme est égale à la médiane.

Exemple 8

Soit le chiffre d'affaires de 7 entreprises de la place au titre de l'année de 2007 (CA en millions de francs CFA)

CA 2 5 6 4 3 9 1

Le nombre d'observation est impair

Rangement par ordre croissant :

CA 1 2 3 4 5 6 9

Le chiffre d'affaires qui occupe la position médiane dans la série des valeurs rangées est 4. La médiane de cette série impaire d'observations non regroupées est égale à 4 millions de francs

Exemple 9

Soit la taille d'un groupe de 6 ménages recensés au cours d'une enquête budget consommation

(TAILM en personnes)

TAILM 8 4 3 10 9 5

Le nombre d'observations est pair

Rangement par ordre croissant

TAILM 3 4 5 8 9 10

Les deux chiffres d'affaires qui occupent la position médiane sont 5 et 8, leur demi-somme 6,5 est égale à la valeur médiane de cette série d'un nombre pair de données non regroupées.

La médiane d'une variable statistique discrète

La médiane d'une variable statistique discrète est déterminée par le tableau de répartition de la variable dans lequel sont indiqués les valeurs de la variable et les effectifs cumulés (ou les fréquences cumulées). La médiane d'une telle variable correspond au premier effectif cumulé supérieur à la moitié du nombre total d'observations. Si l'on utilise les fréquences cumulées, alors la médiane d'une variable discrète correspond à la première fréquence cumulée supérieure à 50.

Exemple 10

Tableau

La distribution d'un échantillon de pièces d'une entreprise de fabrication de pièces détachées selon la taille (en centimètres)

Taille	12	13	14	15	16	17	18
ni	6	20	25	34	20	12	3
Ni	6	26	51	85	105	117	120

Le nombre total d'observations est égal à 120. La moitié de ce nombre est égale à 60. Le premier effectif cumulé supérieur à 60 est égal à 85. A cet effectif cumulé correspond la valeur médiane de la distribution discrète. Elle est égale à 15 cm. La médiane de cette distribution discrète est donc égale à 15 cm.

La médiane d'une variable statistique continue

Dans le cas d'une variable continue, la médiane est obtenue par l'une des formules suivantes :

$$Me = L_1 + k\frac{N_0 - N_1}{N_e - N_1}$$

Où :

Me- La médiane de la variable statistique continue

L_1- La borne inférieure de l'intervalle médian

L'intervalle médian correspond au premier effectif cumulé supérieur à la moitié du nombre total d'observations ou à la première fréquence cumulée supérieure à 50.

k- l'amplitude de la classe médiane

Ne- l'effectif cumulé de la classe médiane

N_1- l'effectif cumulé de la classe précédant la classe médiane

No=0,5.n (n est le nombre total d'observations)

On peut aussi utiliser la formule équivalente suivante :

$$Me = L_1 + k\frac{0,5\sum n_i - N_1}{n_e}$$

Où :

Me- La médiane de la variable statistique continue

L_1- La borne inférieure de l'intervalle médian

L'intervalle médian correspond au premier effectif cumulé supérieur à la moitié du nombre total d'observations ou à la première fréquence cumulée supérieure à 50.

k- L'amplitude de la classe médiane

ne- L'effectif de la classe médiane

N_1- l'effectif cumulé de la classe précédant la classe médiane

On peut également utiliser la borne supérieure de l'intervalle médian pour calculer la médiane.

Étant donné que $L_2 - L_1 = k \rightarrow L_1 = L_2 - k$

$$Me = L_1 + k\frac{0,5\sum n_i - N_1}{n_e} = L_2 - k + \frac{k(0,5\sum n_i - N_1)}{n_e} = L_2 - \frac{k(n_e - 0,5\sum n_i + N_1)}{n_e}$$

Or $N_1 + n_e = Ne$,

D'où

$$Me = L_2 - \frac{k(n_e - 0,5\sum n_i + N_1)}{n_e} = L_2 - k\frac{N_1 + n_e - 0,5\sum n_i}{n_e} = L_2 - k\frac{N_e - 0,5\sum n_i}{n_e}$$

$$Me = L_2 - k\frac{N_e - 0,5\sum n_i}{n_e}$$

Exemple 11

Les ventes annuelles d'un groupe de paysans de Bitjabé au titre de l'année 2005 (en milliers de francs CFA)

Ventes	0-10	10-20	20-30	30-40	40-50	50-60	60-70	Total
ni	3	5	14	20	18	8	2	70
Ni	3	8	22	42	60	68	70	

Détermination de l'intervalle médian ou de la classe médiane

Le nombre total d'observations est égal à 70. La moitié de ce nombre est égale à 35. La classe ou l'intervalle qui contient la médiane correspond au premier effectif cumulé supérieur à 35. Dans notre cas l'intervalle médian correspond à l'effectif cumulé 42. Cet intervalle est 30-40

Détermination de la médiane d'une variable continue à l'aide de la formule :

$$Me = L_1 + k\frac{N_0 - N_1}{N_e - N_1}$$

$$L_1 = 30 \quad k = 10 \quad N_0 = 35 \quad N_1 = 22 \quad N_e = 42$$

$$Me = 30 + 10\frac{35 - 22}{42 - 22} = 36,5$$

Détermination de la médiane d'une variable continue à l'aide de la formule :

$$Me = L_1 + k\frac{0,5\sum n_i - N_1}{n_e}$$

$$L_1 = 30 \quad k = 10 \quad N_1 = 22 \quad n_e = 20$$

$$Me = 30 + 10\frac{35 - 22}{20} = 36,5$$

Détermination de la médiane d'une variable continue à l'aide de la borne supérieure

$$Me = L_2 - k\frac{N_e - 0,5\sum n_i}{n_e}$$

$$L_2 = 40 \quad k = 10 \quad N_e = 42 \quad n_e = 20$$

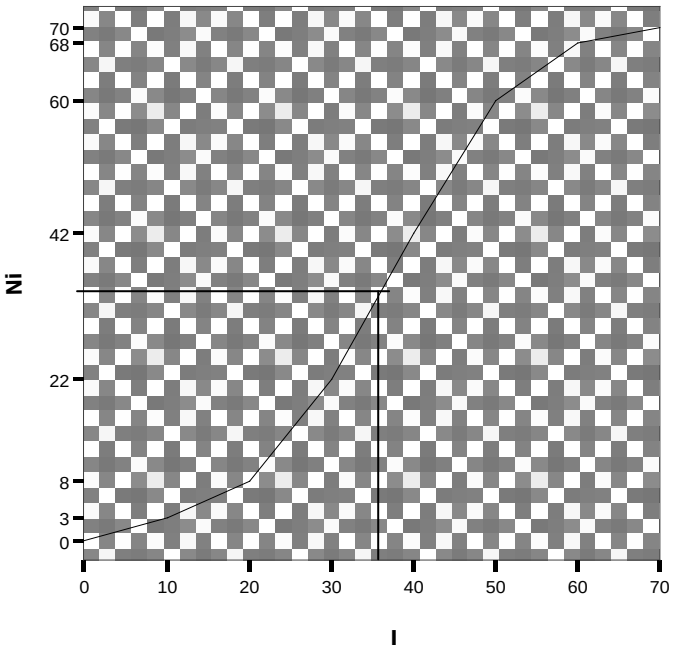
$$Me = 40 - 10\frac{42 - 35}{20} = 36,5$$

Détermination graphique de la médiane

On peut déterminer graphiquement la médiane d'une variable statistique continue. Il suffit de se référer à la courbe des effectifs cumulés ou des fréquences cumulées et de localiser sur l'axe des ordonnées le point A de coordonnées (0; 0,5Σni) ou (0; 50) et de tracer à partir de ce point une droite parallèle à l'axe des abscisses qui va couper la courbe des fréquences cumulées en un point B dont l'abscisse est égale à la valeur de la médiane.

Graphique

Détermination graphique de la médiane



Interprétation de la médiane

En divisant l'effectif total des observations rangées en deux parties égales, la médiane se situe au centre de cette suite. C'est tout comme si l'on passait de distribution initiale à une distribution simplifiée qui se présente comme suit :

Tableaux.
Distributions simplifiées

Classes	Fréquences (%)
x_{min} - Me	50
Me - x_{max}	50

Classes	Fréquences (%)
0 – 36,5	50
36,5 – 70	50

On peut donc aisément interpréter la médiane en concluant que les 50 % des paysans les moins studieux (les moins productifs) ont réalisé des ventes comprises entre 0 et 36 500 francs CFA, alors que les 50 autres pourcents des paysans les plus studieux (les plus productifs) ont réalisé des ventes annuelles comprises entre 36 500 et 70 000 francs CFA.

Le mode

Le mode est la valeur la plus observée de la variable statistique. La plupart des observations ont une valeur égale ou proche de la valeur modale.

Le mode des données non regroupée

Le mode des données non regroupées est égal à la valeur la plus observée, c'est-à-dire celle qui se répète le plus grand nombre de fois.

Exemple 12

Soient les données suivantes représentant le nombre de cigarettes fumées par un groupe de 10 étudiants au cours d'une journée.

4 5 4 2 6 7 8 3 4 1

A la lecture de ces données, on remarque tout simplement que la valeur 4 est la plus observée (3 fois). Le mode de ces données est égal à 4 cigarettes.

Le mode d'une variable statistique discrète

Le mode d'une variable statistique discrète est la valeur de cette variable qui correspond à l'effectif le plus élevé (ou à la fréquence la plus élevée) de la distribution.

Exemple 12 bis

En considérant l'exemple sur la distribution des pièces d'une entreprise selon la taille, l'effectif le plus élevé est égal à 34. A cet effectif correspond la valeur modale qui égal à 15 cm.
Mo = 15 cm.

Le mode d'une variable statistique continue

Le mode d'une distribution statistique continue se calcule à l'aide de la formule suivante :

$$M_o = L_1 + k\frac{n_o - n_1}{(n_o - n_1) + (n_o - n_2)} = L_1 + k\frac{d_1}{d_1 + d_2}$$

Où :
Mo- Le mode de la variable statistique continue
L_1- La borne inférieure de l'intervalle modal
La classe modale correspond à l'effectif le plus élevé (ou à la fréquence la plus élevée) dans la distribution
k - L'amplitude de la classe modale
no- L'effectif de la classe modale
n_1 – L'effectif de la classe qui précède la classe modale
n_2 - L'effectif de la classe qui suit immédiatement la classe modale

Comme précédemment, l'on peut utiliser la borne supérieure de l'intervalle modal pour déterminer le mode

$$M_o = L_2 - k\frac{n_o - n_2}{(n_o - n_1) + (n_o - n_2)} = L_2 - k\frac{d_2}{d_1 + d_2}$$

Exemple 13

Calcul du mode d'une variable continue

En se référant toujours à l'exemple des ventes annuelles des paysans de Bitjabé, nous avons :

La classe modale correspond à l'effectif le plus élevé. Ici, il s'agit de la classe 30-40 qui contient 20 observations.

$L_1=30$ $k=10$ $no=20$ $n_1=14$ $n_2=18$

$$M_o = L_1 + k\frac{n_0 - n_1}{(n_0 - n_1)+(n_0 - n_2)} = 30 + 10\frac{20-14}{(20-14)+(20-18)} = 37,5$$

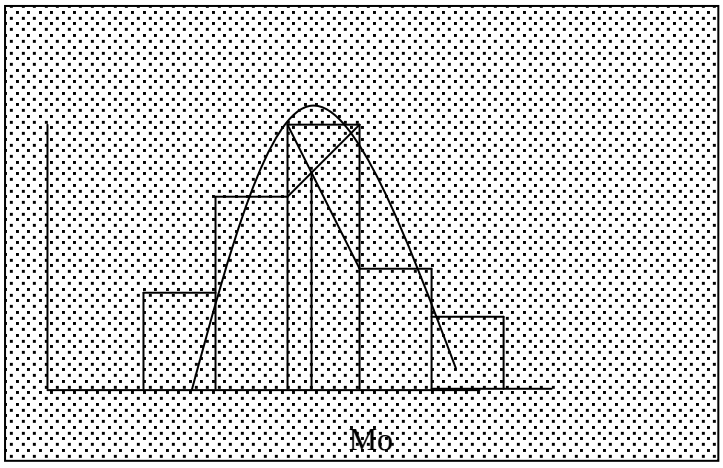
$$M_o = L_2 - k\frac{n_o - n_2}{(n_o - n_2)+(n_o - n_1)} = 40 - 10\frac{20-18}{(20-18)+(20-14)} = 37,5$$

Détermination graphique du mode

Le mode peut être déterminé graphiquement en utilisant l'intervalle modal d'un histogramme (Voir graphique ci-dessous). On distingue les distributions unimodales et les distributions plurimodales. L'existence de plusieurs modes est le plus souvent signe de l'hétérogénéité des données traitées.

Graphique

Détermination graphique du mode d'une variable continue



Interprétation du mode

En prenant le cas de la distribution continue, le mode est égal à 37,5 soit 37500 francs CFA. Ce qui veut dire qu'au cours de l'année 2005 la plupart des paysans de Bitjabé ont vendu des produits pour un montant de 37500 francs

Les quantiles

Par analogie avec la médiane, on peut déterminer 3 valeurs de la variables (Quartiles-Qi) qui divisent le nombre total d'observations en 4 parties égales ou 9 valeurs (Di-déciles) qui divisent le nombre total d'observations en 10 parties égales ou encore 99 valeurs de la variable (Pi-les percentiles) qui divisent le nombre total d'observations en 100 parties égales. Toutes ces valeurs sont appelées quantiles et les plus remarquables sont donc : les quartiles, les déciles et les percentiles.

Formule générale des quantiles

$$Q_L = L_1 + k.\frac{\frac{j}{p}\sum n_i - N_1}{n_e}$$

Pour P=4 et j=1, 2,3, on obtient les quartiles Q_j
Pour p=10 et j=1, 2,3….9, on a les déciles D_j
Pour p=100 et j=1, 2,3,….99, on obtient les percentiles P_J

Les Quartiles

Détermination des quartiles dans le cas d'une variable statistique discrète
Dans le cas d'une variable statistique discrète, le quartile d'ordre j correspond au premier effectif cumulé supérieur à (j/4)*∑ni ou à la première fréquence cumulée supérieure à (j/4)*100%
Ainsi, le premier quartile correspond au premier effectif cumulé supérieur à (1/4) ∑ni.
Le deuxième quartile ou médiane correspond au premier effectif cumulé supérieur à (2/4) ∑ni = (1/2) ∑ni.
Le troisième quartile correspond au premier effectif cumulé supérieur à (3/4) ∑ni.
Exemple 14
Tableau
Répartition d'un groupe d'entreprises industrielles selon le chiffre d'affaires (CA en millions de francs CFA)

CA	10	11	12	13	14	15	16
ni	6	20	25	34	20	12	3
Ni	6	26	51	85	105	117	120

Pour synthétiser, utilisons le tableau ci-dessous

Formule	(1/4) ∑ni	(2/4) ∑ni	(3/4) ∑ni
Valeur avec ∑ni=120	30	60	90
Quartiles	Q_1=12	Q_2=13	Q_3=14

Calcul des quartiles des variables statistiques continues

Formule générale des quartiles

$$Q_1 = L_1 + k\frac{\frac{j}{4}\sum n_i - N_1}{n_e}$$

Le premier quartile est donné par la formule suivante :

$$Q_1 = L_1 + k\frac{\frac{1}{4}\sum n_i - N_1}{n_e}$$

Où :
Q_1 = la valeur du premier quartile
L_1 = la borne inférieur de l'intervalle ou la classe qui contient le premier quartile.
L'intervalle qui contient le premier quartile correspond au premier effectif cumulé supérieur au ¼ du nombre total d'observations (ou à la première fréquence cumulée supérieure à 25%)
ni = l'effectif de la classe i
N_1 = l'effectif cumulé de la classe qui vient avant la classe qui contient le premier quartile
n_e = l'effectif de la classe qui contient le premier quartile.

Exemple 15
Les ventes annuelles d'un groupe de paysans de Bitjabé au titre de l'année 2005 (en milliers de francs CFA)

Ventes	0-10	10-20	20-30	30-40	40-50	50-60	60-70	Total
ni	3	5	14	20	18	8	2	70
Ni	3	8	22	42	60	68	70	

Les trois quartiles sont calculés à l'aide des éléments du tableau ci-dessous

Tableau de calcul des quartiles

	J	P	J/P	Σn_i	J/P Σn_i	Classe	L_1	k	N_1	n_e	Valeur
Q1	1	4	0,25	70	17,5	20-30	20	10	8	14	26,8
Q2	2	4	0,50	70	35,0	30-40	30	10	22	20	36,5
Q3	3	4	0,75	70	52,5	40-50	40	10	42	18	45,8

Les déciles

Détermination des déciles d'une variable statistique discrète

Tableau de détermination des deuxième, sixième et huitième déciles d'une distribution de pièces selon la taille (en centimètres).

Formule	(2/10) $\sum$ni	(6/10) $\sum$ni	(8/10) $\sum$ni
Valeur avec $\sum$ni=120	24	72	96
Déciles	D_2=13	D_6=15	D_8=16

Calcul des déciles des variables continues
Formule générale de calcul des déciles (j variant de 1 à 9)

$$D_j = L_1 + k\frac{\frac{j}{10}\sum n_i - N_1}{n_e}$$

Le 6ième décile (D6) est calculé par la formule suivante

$$D_6 = L_1 + k\frac{\frac{6}{10}\sum n_i - N_1}{n_e}$$

Où : D6 est la valeur du 6ième Décile
L_1 est la borne inférieure de l'intervalle qui contient le 6ième décile. Cet intervalle correspond au premier effectif cumulé supérieur au 6/10 du nombre total d'observations (ou à la première fréquence cumulée supérieure à 60 %). N_1 est l'effectif cumulé de la classe qui vient avant la classe qui contient le 6ième décile ; ne = l'effectif de la classe qui contient le 6ième décile.
Les autres déciles sont calculés de façon analogue.

Tableau de calcul des déciles D1, D5 et D8

	J	P	J/P	Σn_i	J/P Σn_i	Classe	L_1	k	N_1	n_e	Valeur
D1	1	10	0,10	70	7,0	10-20	10	10	3	5	18,0
D5	5	10	0,50	70	35,0	30-40	30	10	22	20	36,5
D8	8	10	0,80	70	56,0	40-50	40	10	42	18	47,8

Les percentiles

Détermination des percentiles d'une variable discrète

Tableau de détermination du premier, 36ième et 90ième percentile de la distribution de pièces selon la taille (centimètres).

Formule	(1/100) ∑ni	(36/100) ∑ni	(90/100) ∑ni
Valeur avec ∑ni=120	1,2	43,2	108
Déciles	P_1=12	P_{36}=14	P_{90}=17

Calcul des percentiles d'une variable continue

Les percentiles d'une variable continue sont calculés à l'aide de la formule suivante :

$$P_j = L_1 + k\frac{\frac{j}{100}\sum n_i - N_1}{n_e}$$ Avec j variant de 1 à 99

Pour calculer, par exemple le 15ième percentile (P_{15}), on utilise la formule suivante :

$$P_{15} = L_1 + k\frac{\frac{15}{100}\sum n_i - N_1}{n_e}$$

Où P_{15} est la valeur du 15ième percentile.
L_1 est la borne inférieure de l'intervalle qui contient le 15ième percentile. Cet intervalle correspond au premier effectif cumulé supérieur au 15/100 du nombre total d'observations (ou à la première fréquence cumulé supérieure à 15 %). N_1 est l'effectif cumulé de la classe qui vient avant la classe qui contient le 15ième percentile ; ne = l'effectif de la classe qui contient le 15ième percentile.
Les autres percentiles sont calculés de la même manière.
Pour plus de détails sur le calcul des quantiles voir la séquence de calcul des quantiles.

Tableau de calcul du premier (P_1), vingt cinquième (P_{25}) et soixante dixième percentiles (P_{70})

	J	P	J/P	Σn_i	J/P Σn_i	Classe	L_1	k	N_1	n_e	Valeur
P1	1	100	0,01	70	0,7	0-10	0	10	0	3	2,3
P25	25	100	0,25	70	17,5	20-30	20	10	8	14	26,8
P70	70	100	0,70	70	49,0	40-50	40	10	42	18	43,9

Séquences de calcul des quantiles

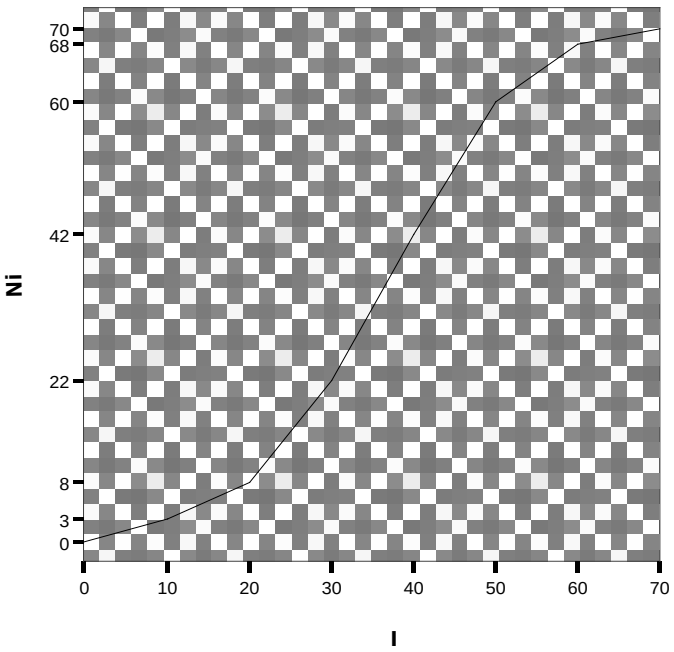
Pour calculer un quantile, il est plus prudent de procéder par les étapes suivantes :

1) Définition de p afin d'identifier le quantile à calculer
2) Définition de J pour savoir le quellième quantile calculer
3) J/P, intervenant dans l'identification de l'intervalle qui contient le quantile
4) Σni, intervenant dans l'identification de l'intervalle qui contient le quantile
5) J/P (Σni), intervenant dans l'identification de l'intervalle qui contient le quantile
6) Identification de L1, la borne inférieure de l'intervalle qui contient le quantile
7) K, l'amplitude de la classe qui contient le quantile
8) N1, l'effectif cumulé de la classe avant la classe qui contient le quantile

9) ne, l'effectif de la classe qui contient le quantile
10) Calcul du quantile

Détermination graphique des quantiles

Il est possible, par analogie avec la médiane, de déterminer graphiquement les quantiles des variables statistiques continues à l'aide de la courbe des effectifs cumulés ou la courbe des fréquences cumulées.
Détermination graphique du troisième quartile (Q3)
Sur l'axe des ordonnées de la courbe des effectifs cumulés, on localise le point A de coordonnées (0; 3/4Σni) et on trace à partir de ce point une droite parallèle à l'axe des abscisses qui va couper la courbe des fréquences cumulées en un point B dont l'abscisse est égale au troisième quartile (Q3). Cette procédure peut être utilisée pour la détermination graphique de tout quantile d'une variable statistique continue.



Interprétation des quantiles

Tout comme dans le cas de l'interprétation de la médiane, pour interpréter un quantile, il est beaucoup plus simple de recourir à la reconstruction de la distribution en définissant deux modalités faisant intervenir :

- la borne inférieure du premier intervalle de la distribution
- la valeur du quantile
- la borne supérieure du dernier intervalle de la distribution

En prenant l'exemple du premier quartile, nous avons la distribution suivante
Tableau
Interprétation du premier quartile

Classes	fi
0-26,8	25
26,8-70	75
Total	100

Les 25 % des paysans les moins performants ont réalisé des ventes annuelles comprises ente 0 et 26 800 francs CFA, alors que les 75 autres pourcents des paysans (les plus performants) ont réalisé des ventes comprises entre 26 800 et 70 000 francs CFA.