Matière : Méthodes statistiques

Spécialité : Chimie de l'environnement

Semestre : S2

Chapitre I : Bases de statistiques (Paramètres statistiques et estimation).

I.1 Notions de bases et domaines d'application des statistiques

- La statistique est la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation de phénomènes aléatoires ; dans lesquels le hasard intervient. L'analyse des données est utilisée pour d´écrire les phénomènes ´étudies, faire des prévisions et prendre des décisions à leur sujet.

  En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes. Les données ´étudiées peuvent être de toute nature, ce qui rend la statistique utile dans tous les champs disciplinaires et explique pourquoi elle est enseignée dans toutes les filières universitaires, de l'économie `a la biologie en passant par la psychologie, et bien sur les sciences de l'ingénieur.

Donnons quelques exemples d'utilisation de la statistique dans divers domaines.

• Economie, assurance, finance : prévisions économétriques, analyse de la consommation des ménages, fixation des primes d'assurance et franchises, études quantitatives de marchés, gestion de portefeuille, évaluation d'actifs financiers, ...

• Biologie, médecine : essais thérapeutiques, ´épidémiologie, dynamique des populations, analyse du génome

• Sciences de la terre : prévisions m´météorologiques, exploration pétrolière, ...

• Sciences humaines : enquêtes d'opinion, sondages, études de populations, ...

• Sciences de l'ingénieur : contrôle de qualité, maitrise statistique des procédés (méthode "six-sigma"), maitrise des risques industriels, évaluation des performances des systèmes complexes

L'objectif essentiel de la statistique est de maitriser au mieux cette incertitude pour extraire des informations utiles des données, par l'intermédiaire de l'analyse des variations dans les observations. Nous ne nous intéresserons pas à la collecte des données, qui est une tâche importante et difficile, mais qui ne relève pas des mathématiques. Si on omet la collecte des données, les m´méthodes statistiques se répartissent en deux classes :

• La statistique descriptive, statistique exploratoire ou analyse des données, a pour but de résumer l'information contenue dans les données de façon synthétique et efficace. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs numériques (par exemple des moyennes). Elle permet de d´dégager les caractéristiques

essentielles du phénomène ´étudie et de suggérer des hypothèses pour une ´étude ultérieure plus sophistiquée. Les probabilités n'ont ici qu'un rôle mineur.

• La statistique inférentielle va au de la de la simple description des données. Elle a pour but de faire des prévisions et de prendre des d´décisions au vu des observations. En général, il faut pour cela proposer des modèles probabilistes du phénomène aléatoire ´étudie et savoir gérer les risques d'erreurs. Les probabilités jouent ici un rôle fondamental.

I.2. La présentation des données

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données. En d'autres termes, c'est une science qui a pour objectif : la planification du projet ; la collecte, la codification, la saisie, le traitement et l'analyse des données ; la publication des résultats. Il ne faut pas confondre la statistique qui est la science qui vient d'être définie et une statistique qui est un ensemble de données chiffrées sur un sujet précis. Les premières statistiques correctement élaborées ont été celles des recensements démographiques. Ainsi le vocabulaire statistique est essentiellement celui de la démographie. Les ensembles étudiés sont appelés population. Les éléments de la population sont appelés individus ou unités statistiques. La population est étudiée selon un ou plusieurs caractères.

I.2.1. La population

C'est l'ensemble des individus (ou unités statistiques) présentant un caractère commun. Pour une thématique donnée, la population regroupe toujours la totalité des individus relatifs à cette thématique (notion d'exhaustivité). Exemples : la population congolaise, les pays de la CEMAC, les clients d'une entreprise. La population est en général notée P L'effectif total d'une population est noté N.

I.2.2. L'unité statistique ou l'individu

C'est l'élément de base constitutif de la population à laquelle il appartient. Il est indivisible et peut être un pays, un végétal, un humain ou une entreprise.

I.2.3. L'échantillon

C'est un sous-ensemble construit et représentatif d'une population donnée.

I.2.4. Le caractère ou la variable

C'est la (les) caractéristique(s) de l'individu intégrant la population étudiée. Exemple : la couleur, le sexe, le poids, la taille, la marque, le modèle, l'espèce, le prix, la surface, etc.

- **Variable qualitative**

Une variable statistique est dite de nature qualitative si ses modalités ne sont pas mesurables. Les modalités d'une variable qualitative sont les différentes catégories d'une nomenclature. Ces catégories doivent être exhaustives (chaque individu est affecté à une modalité) et incompatibles (un individu ne peut être affecté à plusieurs modalités) de façon à créer une

partition. Le sexe, la profession, l'état matrimonial sont quelques exemples de variables qualitatives. Pour ses enquêtes auprès des ménages, l'Insee utilise la nomenclature des Professions et catégories socioprofessionnelles (PCS-2003).

Les modalités d'une variable qualitative peuvent être classées sur deux types d'échelle : nominale ou ordinale. À ces deux types d'échelle correspondent deux types de variables qualitatives

- Variable qualitative nominale : Une variable statistique qualitative est dite définie sur une échelle nominale si ses modalités ne sont pas naturellement ordonnées. Exemples : Situation d'activité, statut matrimonial.
- Variable qualitative ordinale : Une variable statistique qualitative est dite ordinale si l'ensemble de ses modalités peut être doté d'une relation d'ordre. Exemple : Niveau d'instruction.

- **Variable quantitative**

Toute variable qui n'est pas qualitative ne peut être que quantitative. Les différentes modalités d'une variable quantitative constituent l'ensemble des valeurs numériques que peut prendre la variable. Une variable statistique est dite de nature quantitative si ses modalités sont mesurables. Les modalités d'une variable quantitative sont des nombres liés à l'unité choisie, qui doit toujours être précisée.

Il existe deux types de variables quantitatives : les variables discrètes et les variables continues. Ces variables ont en commun des modalités clairement ordonnées, pour lesquelles l'écart entre les valeurs possède une signification, et sur lesquelles il est possible de réaliser des opérations mathématiques telles que des calculs de moyennes, etc. Néanmoins, elles ont des propriétés et des traitements spécifiques qui nécessitent une étude séparée.

1) Variable quantitative discrète : Lorsque les modalités sont des valeurs numériques isolées, comme le nombre d'enfants par ménage, on parle de variable discrète. Exemples : Âge, salaire, nombre de lit dans un hôpital.
2) Variable quantitative continue : Lorsque la variable, par exemple la taille d'un individu, peut prendre toutes les valeurs d'un intervalle, ces valeurs peuvent alors être regroupées en classes, et on parle dans ce cas de variable continue. Exemples : Poids, taux du sucre, taille, taux du sel

I.2.5. La modalité

C'est la valeur qualitative ou quantitative que peut prendre le caractère précédemment défini. Exemple : sexe féminin ou masculin, poids 45 kg, couleur verte, etc. Attention, les modalités sont exhaustives et mutuellement exclusives. Chaque individu doit pouvoir être classé dans une et une seule modalité.

I.2.6. La discrétisation

Lorsque les modalités sont des valeurs numériques isolées, comme le nombre d'enfants par ménage, on parle de variable discrète. Ce découpage en classes pose de nombreuses questions

: choix des amplitudes, amplitudes constantes ou variables, nombre de classes, etc. Nous ne rentrerons pas ici dans le détail de ces opérations.

I.3.caractéristiques de tendance centrale des données

Les paramètres de tendance centrale ou « mesures de tendance centrale » sont des grandeurs susceptibles de représenter au mieux un ensemble de données. L'appellation «tendance centrale » vient du fait que ces paramètres donnent une idée de ce qui se passe au centre d'une distribution, d'un ensemble de données. On distingue trois mesures de tendance centrale :

- La moyenne ;
- Le mode ;
- La médiane.

Tous trois ne décrivent par la même chose et sont, de ce fait, complémentaires dans la description et l'analyse d'une distribution. Ces statistiques ne se calculent que dans le cas où nous avons à faire à des variables quantitatives. Dans le cas où nous avons des variables qualitatives, on procède aux fréquences. Avant d'analyser ces trois indicateurs de position, nous allons d'abord aborder la notion de la fréquence.

I.3.1. Les fréquences absolues, relatives et cumulées

A chaque modalité de variable X, peut correspondre un ou plusieurs individus dans l'échantillon de taille n. On appelle effectif de la modalité le nombre. Il est aussi appelé fréquence Absolue. La fréquence relative est le nombre fi tel que $fi = ni/n$

La fréquence cumulée croissante est cependant le nombre tel que

$$F_i = \sum_{p=1}^{i} f_p$$

Exemple : Représentons la fréquence relative et la fréquence cumulée du tableau ci-dessous

Xi	ni
1	8
2	18
3	14
4	10
Total	50

Solution

Xi	ni	fi	FCC	FCD
1	8	8/50=0,16	0,16	1
2	18	18/50=0,36	0,16+0,36=0,52	1-0,16=0,84
3	14	14/50=0,28	0,52+0,28=0,8	0,84-0,36=0,48
4	10	10/50=**0,2**	0,8+0,2=**1**	0,48-0,28=**0,2**
Total	50	50/50=**1**		

I.3.2. La moyenne

La moyenne constitue l'un des paramètres fondamentaux de tendance centrale mais non suffisant pour caractériser une distribution. Complémentaire du mode et surtout de la médiane, la moyenne constitue à n'en point douter, la mesure la plus calculée et la plus utilisée lors de la description de séries statistiques. Il existe plusieurs types de moyennes, chacun adapté à des situations précises :

DESIGNATION	NOTATION COURANTE
Moyenne arithmétique	$\overline{X}$
Moyenne géométrique	$\overline{G}$ ou $\overline{x}_G$
Moyenne harmonique	$\overline{H}$ ou $\overline{x}_H$
Moyenne quadratique	$\overline{Q}$ ou $\overline{x}_Q$

- La moyenne arithmétique

C'est la plus simple et la communément utilisée et ce, pas toujours à bon escient. Elle se note la plupart du temps par .

Elle peut être simple ou pondérée.

La moyenne arithmétique simple Sa version simple correspond à une somme de résultats divisée par le nombre de résultats et s'écrit

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

La moyenne arithmétique pondérée La moyenne arithmétique pondérée, autant le dire tout de suite, donne, dans son utilisation classique (c'est-à dire lorsque tous les individus ont le même poids), le même résultat que la moyenne arithmétique simple. Sa formule est cependant différente puisqu'elle introduit la notion de poids via un terme supplémentaire qui peut s'avérer

utile dans certaines situations, notamment lorsque justement les individus composant une population n'ont pas le même poids ou coefficient :

Certains individus, pour diverses raisons, ont davantage d'influence dans ladite population que les autres. Ce peut être le cas par exemple lorsque l'on a affaire à une série de notes dont le coefficient n'est pas le même. Cette moyenne s'écrit de la manière suivante :

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

- **La moyenne géométrique**

Sa définition purement mathématique est un peu rébarbative mais son utilité est grande comme nous allons le démontrer. La moyenne géométrique de n valeurs positives xi est la racine nième du produit de ces valeurs. Elle est notée t :La moyenne géométrique est un instrument permettant de calculer des taux moyens notamment des taux moyens annuels. Son utilisation n'a un sens que si les valeurs ont un caractère multiplicatif.

- **La moyenne quadratique**

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

$$\overline{Q} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i{}^2)}$$

I.3.3. Le mode

Le mode, d'une série statistique est la valeur du caractère la plus fréquente ou dominante dans l'échantillon. Autrement dit, c'est la valeur qui a la fréquence (absolue ou relative) la plus élevée. Lorsque la distribution a plus d'un mode, on parle d'une distribution « multimodale » (bimodale, trimodale , etc). Par contre, si l'on est en présence de données groupées en classes, le mode se rapportera à la classe comportant le plus grand nombre d'individus : on parlera alors de classe modale. Cependant, il peut y arriver que l'on s'intéresse à avoir la valeur approchée ou exacte de ce mode. Par conséquent, il est recommandé d'appliquer la démarche suivante :

- Pour avoir une valeur approximative du mode, on calcule la moyenne de la classe qui a la fréquence la plus élevée ; - Pour avoir une valeur exacte, le mode se calcule de la manière suivante :

$$M_o = x_m + \frac{i\Delta i}{\Delta s + \Delta i}$$

Avec Xm : limite inférieure de la classe modale ; i : amplitude de la classe modale ; Δi : écart d'effectif entre la classe modale et la classe inférieure la plus proche Δs : écart d'effectif entre la classe modale et la classe supérieure la plus proche

I.3.4. La médiane

Dans le calcul de la médiane, on distingue deux cas :

- Si la variable est discrète On désigne par n le nombre d'observations. Si n est pair : la médiane est alors égale à la moyenne des valeurs encadrant le milieu de la série.

Si n est impair alors il est possible d'identifier simplement la valeur qui partage la population en deux effectifs égaux. Le rang central étant égal à [(n+1)/2].

- Si la variable est continue et qu'elle est groupée en classe, On cherche la classe contenant le individu de l'échantillon. Cette classe est appelée la classe médiane. En supposant que tous les individus de cette classe sont uniformément répartis à l'intérieur, la médiane se calcule de la façon suivante par interpolation linéaire :

$$Me = xm + a\,[\ 2 - Ni\ ni\ ]$$

xm : limite inférieure de la classe médiane ;

a : amplitude de la classe médiane ;

ni : effectif de la classe médiane

Ni. : Effectif cumulé inférieur à xm n : taille de l'échantillon

Exemples : a) Calculer la médiane des séries suivantes : 1 = {9, 10, 9, 9, 11, 10, 11, 11, 11, 9} 2 = {2, 3, 5, 5, 6, 7, 7, 8, 8, 9, 2, 2, 1, 3, 3, 4, 2, 1, 1} b), Calculer la médiane des données se trouvant dans le tableau suivant :

Xi	ni	Ni	fi	FCC
0-10	48	48	0,24	0,24
. 10-15	40	88	0,2	0,44
15-20	56	144	0,28	0,72
20-30	32	176	0,16	0,88
30-50	24	200	0,12	1
Total	200		1	

I.4. Mesure de la dispersion

Les indices de tendance centrale définissent le comportement général des données. Mais les données peuvent varier beaucoup autour de cette tendance.

On doit donc définir un indice qui caractérise la variabilité des données dans l'échantillon. Cet indice est appelé indice de dispersion parce qu'il renseigne sur la dispersion ou l'éparpillement des données autour notamment des paramètres de tendance centrale. Nous étudierons quatre paramètres de dispersion parmi les principaux, en mettant plus particulièrement l'accent sur la variance et l'écart-type :

L'étendue et le rapport de variation ,l'intervalle interquartile
la variance et l'écart-type
Le coefficient de variation

I.4.1. L'étendu et le rapport de variation

Le Minimum et le maximum d'une série statistique correspondent respectivement, comme leur nom l'indique, aux valeurs minimale et maximale rencontrées dans une série. Ces deux paramètres ont une double utilité. Ils permettent de calculer :

L'étendue de la distribution, également appelée Intervalle de Variation (IV), c'est-à dire l'écart entre le minimum et le maximum.

$$\textbf{Etendu} = \textbf{Maximum} - \textbf{Minimum}$$

Le Rapport de Variation (RV), c'est-à-dire le rapport de la valeur maximale de la distribution à

$$\textbf{Rapport de variation} = \textbf{Maximum} \quad / \textbf{Minimum}$$

la valeur minimale de la même distribution.

I.4.2. La variance et l'ecart-type

Considérons une distribution pour laquelle on a calculé les paramètres de tendance centrale comme la médiane ou la moyenne. Comme leurs noms l'indiquent, et comme mentionné plus haut, ces mesures caractérisent le centre de la distribution. Parmi celles-ci, considérons la moyenne comme une référence. Que pensez-vous de l'écart entre chaque valeur de la distribution et cette moyenne ?

$$(\mathbf{x}\boldsymbol{i} - \bar{\mathbf{x}})$$

Plus cet écart sera faible, plus la valeur xi sera proche de la moyenne et donc du centre de la distribution. A contrario, plus l'écart sera important et plus xi sera éloignée du centre de la distribution. La prise en compte de la somme de l'ensemble des écarts à la moyenne, c'est-àdire de la somme de tous les écarts entre les xi et la moyenne donne logiquement 0

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Si l'on veut tenir compte de l'ensemble des distances à la moyenne sans pâtir d'une somme nulle, résultat de la compensation entre écarts négatifs et écarts positifs, il est nécessaire d'élever au carré chaque écart de telle sorte que l'on est :

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$$

Que pensez-vous alors de la moyenne calculée de ces écarts élevés au carré?

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

Ce paramètre # est la variance. La variance satisfait à toutes les exigences énoncées plus haut relativement à la mesure de la dispersion d'une distribution. La variance pose toutefois le problème de proposer un résultat en unité élevée au carré. Si les données xi sont en euros, la moyenne sera en euros, de même que l'écart (xi−x) alors que la variance sera en euros carrés. Il faut noter que la valeur de # est la variance de la population; la variance de l'échantillon est de ce fait :

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

Pour revenir à l'unité initiale, il faut extraire la racine carrée de la variance. C'est ce qui nous donne l'écart-type. Comme ce fut le cas pour le calcul de la moyenne de données groupées, pour calculer la variance des données groupées, il faut prendre en compte le centre de chaque classe et considérer que les individus d'une même classe ont tous la même valeur, celle du centre de leur classe.

I.4.3. Le coefficient de variation l'ecart-type,

Malgré sa pertinence dans la mesure de la dispersion d'une distribution, possède un inconvénient majeur : il est exprimé dans l'unité de la variable à laquelle il se rapporte. Il est alors impossible de comparer les dispersions de deux distributions ayant un lien entre elles (lien de causalité ou autre) et dont les valeurs s'expriment dans des unités différentes. Pour comparer la dispersion de deux séries qui ne sont pas exprimées dans les mêmes unités, on utilise le coefficient de variation. Cette statistique est une mesure neutre qui s'exprime la plupart du temps en pourcentage. Il se calcule en divisant l'écart-type par la moyenne et s'écrit donc :

$$CV = \frac{\sigma}{\bar{x}}$$

*Plus grand est le coefficient de variation, plus grande est la dispersion.

Exemple : Trouver la variance et l'écart-type de la distribution suivante :

Classes	n_i	Centre des classes (c_i)	$n_i * c_i$	c_i-moyenne	$n_i * (c_i\text{-moyenne})^2$
[4-6[	2	(4 + 6) / 2 = 5	2*5 = 10	5-15,5 = -10,5	$2*(-10,5)^2 = 220,5$
[6-10[	5	(6 + 10) / 2 = 8	5*8 = 40	8-15,5 = -7,5	$5*(-7,5)^2 = 281,25$
[10-20[	8	(10 + 20) / 2 = 15	8*15 = 120	15-15,5 = -0,5	$8*(-0,5)^2 = 2$
[20-30[	4	(20 + 30) / 2 = 25	4*25 = 100	25-15,5 = 9,5	$4*(9,5)^2 = 361$
[30-50[	1	(30 + 50) / 2 = 40	1*40 = 40	40-15,5 = 24,5	$1*(24,5)^2 = 600,25$
Total	**20**		**310**		**1465**

Solution

La moyenne de ces données est : = 310/ 20 = 15,5

La variance est $ 2 = 1/ ∗ ∑i=1n [∗ (& −) 2] = 1465/ 20 = 73,25

L'écart-type est $ = √$ 2 = √73,25 = 8,56

I.5. Représentation des données

Les graphiques sont les corollaires d'une bonne analyse et d'une interprétation la plus complète possible de séries statistiques ou de résultats sur des traitements de données. Ces modes de représentation de la donnée participent à la compréhension des phénomènes, au même titre que les tableaux simples ou élaborés, apportant une information certes agrégée, synthétique mais très visuelle et en cela plus facile à aborder et à interpréter que ne le ferait un tableau de chiffres.

Chaque type de graphique est adapté à une ou plusieurs situations ou façon de représenter l'information. Selon la nature des données, la nature de variable, le nombre de variables et ce que l'on souhaite montrer, il sera judicieux de choisir la représentation graphique la mieux adaptée

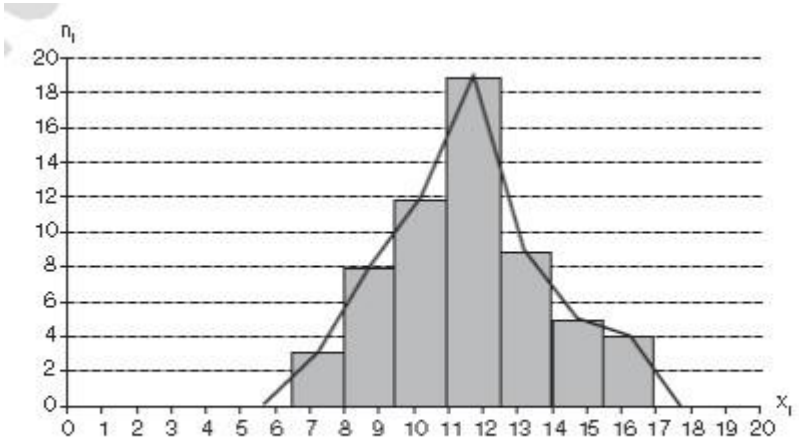
- **Le cas des variables continues**

Quand on a une variable sous forme de classe (une variable discrétisée), on ne peut que faire l'histogramme. Un histogramme est un diagramme composé de rectangles contigus dont les aires sont proportionnelles aux effectifs (ou aux fréquences) et dont les bases sont déterminées par les intervalles de classes.

Exemple : Le responsable des ressources humaines d'une entreprise a relevé la distribution statistique suivante correspondant à l'ancienneté du personnel cadre dans l'entreprise, exprimée en années :

Classes	Effectifs
[6,5 ; 8[	3
[8 ; 9,5[	8
[9,5 ; 11[	12
[11 ; 12,5[	19
[12,5 ; 14[	9
[14 ; 15,5[	5
[15,5 ; 17[	4
Total	60

Solution :



- **Le cas des variables discretes**

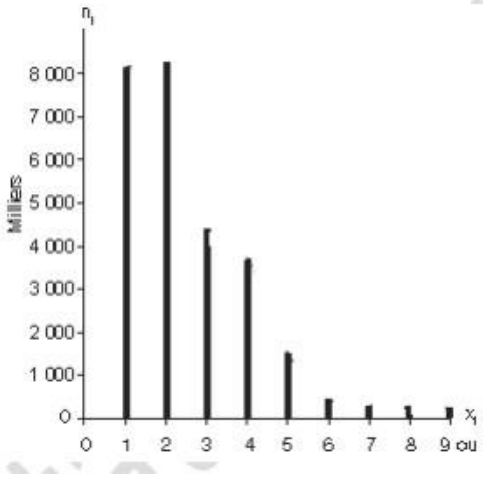
Quand on a une variable discrète, on peut faire le diagramme en bâton, le diagramme en ligne ou le nuage de points.

Diagramme en bâton On appelle diagramme en bâtons un graphique qui à chaque modalité d'une variable quantitative discrète associe un segment (bâton) dont la hauteur est proportionnelle à l'effectif (ou à la fréquence).

Exemple : Faites le diagramme en bâtons du nombre de personnes par ménage en France 1999 se trouvant dans le tableau ci-dessous.

X_i	1	2	3	4	5	6	7	8	9 ou plus
n_i	8000	8100	4500	3500	1500	500	300	200	300

Solution :



- **Le cas des variables qualitatives**

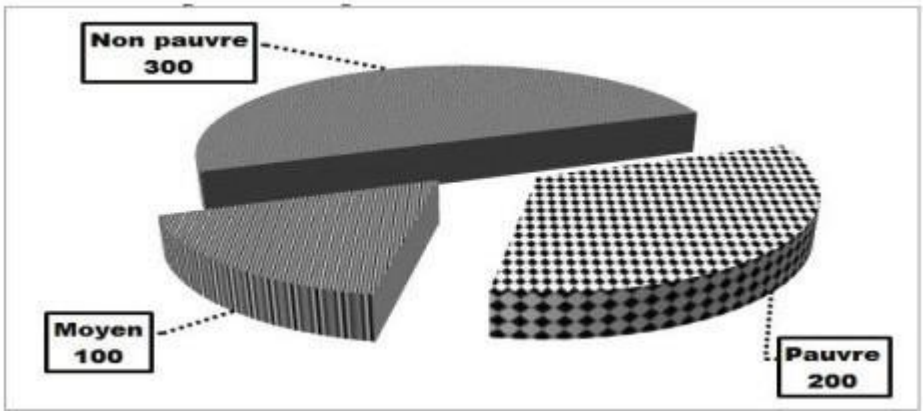
1) Diagramme circulaire (camembert)

Un diagramme circulaire est un graphique constitué d'un cercle divisé en secteurs dont les angles au centre sont proportionnels aux effectifs (ou aux fréquences). De fait, les aires des secteurs sont proportionnelles aux effectifs. L'angle ' d'une modalité d'effectif ni est donné en degrés par :

Exemple

	ni	fi	αi
Pauvre	200	0,3	120
Moyen	100	0,2	60
Non pauvre	300	0,5	180
Total	600		

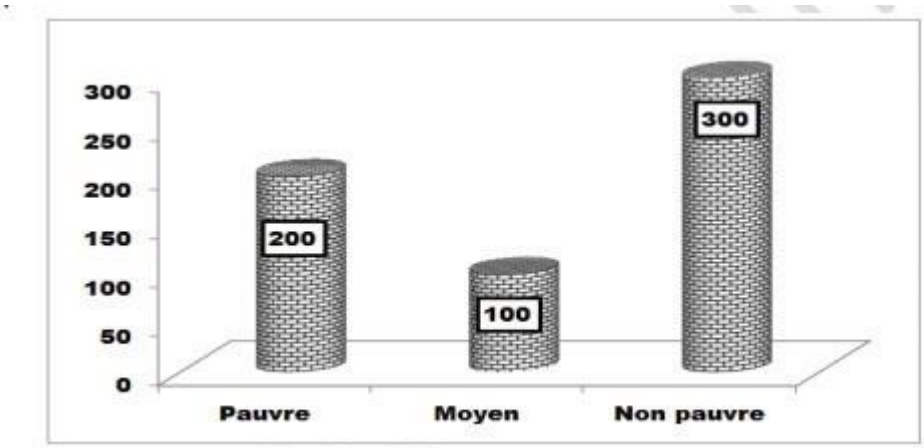
Solution :



2) Diagramme en tuyau d'orgue

Un diagramme en tuyaux d'orgue est un graphique qui à chaque modalité d'une variable qualitative associe un rectangle de base constante dont la hauteur est proportionnelle à l'effectif (ou à la fréquence). De fait, les aires des secteurs sont proportionnelles aux effectifs. Les rectangles sont en général disjoints, verticaux ou horizontaux.

Exemple :

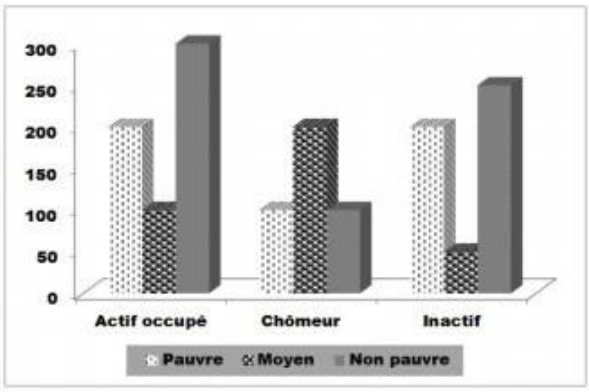


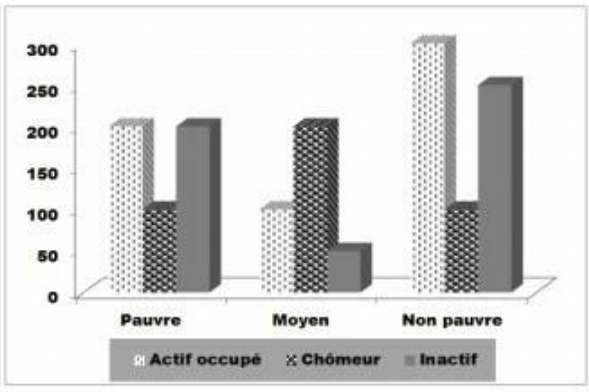
3) Diagramme en barre multiple

Exemple :

	Actif occupé	Chômeur	Inactif	Total
Pauvre	200	100,0	200	500
Moyen	100	200,0	50	350
Non pauvre	300	100,0	250	650
Total	600	400	500	1500

Solution :


300
250
200
150
100
50
0
Actif occupé
Chômeur
Inactif
Pauvre
Moyen
Non pauvre


300
250
200
150
100
50
0
Pauvre
Moyen
Non pauvre
Actif occupé
Chômeur
Inactif