INTRODUCTION GENERALE.

0.1. Notions.

Chaque jour, des renseignements numériques sur des personnes ou des objets sont projetés sur des écrans ou diffusés à la radio :

- Le nombre de minerais exportés de la RDC en 2014 ;

La quantité de gaz renvoyée dans l'atmosphère ; etc…
Ces données peuvent être analysées en vue de fournir des explications sur des phénomènes et d'établir des prévisions, c'est le but de la Statistique.

La statistique peut être définie comme une discipline scientifique qui consiste à collecter des données sur des ensembles nombreux, à analyser ces données en vue de prendre des décisions rationnelles sur des caractères définis dans ces ensembles nombreux. Autrement dit, la statistique comme une discipline qui étudie les méthodes de réduction de données, la variabilité et les populations.

- Les méthodes de réduction des données font partie de la statistique descriptive (ou exploration). Elles consistent à essayer de résumer un échantillon de données via des graphiques ou des caractéristiques numériques.
- L'étude de la variabilité cherche à l'expliquer. Elle fait partie de la théorie de l'échantillonnage
- L'étude des populations fait partie de la statistique inférentielle qui prend un échantillon et en tire des conclusions pour toute la population. Elle part donc de l'expérience à l'hypothèse.

Il existe également des phénomènes dont le résultat de la réalisation ne dépend pas de l'intervention de l'homme :

- En jetant un dé non truqué, l'apparition d'un des nombres 1, 2, 3, 4, 5 ou 6 est un phénomène dont on ne peut prédire la réalisation ;

Le jet d'une pièce de monnaie en l'air peut donner pile ou face et l'on ne peut prédire à l'avance le résultat ; etc…
Ces faits ou phénomènes qui se réalisent par le fait du hasard sont appelés phénomènes aléatoires et relèvent du domaine scientifique de la probabilité.

La probabilité est donc une discipline scientifique qui cherche à déterminer les chances de réalisation d'un phénomène aléatoire ou fortuit en se basant sur des observations antérieures.

0.2. Objectifs du cours.

L'objectif de la statistique descriptive est de résumer un échantillon de données. Au départ, on a échantillon et une variable X supposée quantitative. On désigne par n l'effectif de l'échantillon qui est le nombre d'objets, de sujets, des personnes,….. dans l'échantillon.
A l'issue de ce cours, l'étudiant devra être capable :

- D'utiliser les techniques de collecte des données en vue de l'analyse d'un problème posé ;
- D'organiser et de réaliser la synthèse d'une masse plus ou moins grande d'informations dans un domaine donné;
- D'utiliser les méthodes statistiques afin de tirer des conclusions valides sur la population entière en se basant sur un échantillon représentatif en rapport avec un secteur donné ;

0.3. Définition de la population à étudier : Méthodes d'observation.

Toute étude statistique commence par des observations. Ces observations permettent de recueillir des données de base qui serviront à l'étude. La collecte de données consiste à leur rassemblement et à leur enregistrement.

L'observation est directe quand l'observateur va lui-même mesurer ou compter.
L'observation est indirecte lorsqu'il se base sur les dires d'autrui (les réponses fournies par les individus).
La collecte des données par observation directe fait appel à deux procédés : le dénombrement instantané et le dénombrement continu.
Le dénombrement instantané est la mesure d'un phénomène à un moment donné.
Exemple : le recensement de la population, enquête sanitaire etc…
Le dénombrement continu est l'ensemble de plusieurs dénombrements instantanés par rapport à un même phénomène mais à des moments différents.
Exemple : le remplissement de la fiche des malades que fait l'infirmier chaque jour.
La collecte de données par observation indirecte fait souvent appel à des questionnaires.

0.4. Définitions des unités statistiques : Recensement et Sondage

L'étude complète d'une population, son recensement, c'est-à-dire l'examen de toutes les unités qui la composent n'est pas toujours possible.

Cette étude peut demander du temps, elle peut être coûteuse ou carrément impossible à réaliser. C'est pour cette raison qu'on est conduit à n'observer qu'une partie de la population c'est-à-dire procéder à un sondage (échantillonnage) ; les unités étudiées dans le sondage constitueront un échantillon. L'étude de cet échantillon fournira des informations qui pourront être étendues à la population complète.

Les principaux documents d'enregistrement sont les registres, les fiches et les questionnaires.
Sur ces documents, l'enregistrement consiste à noter avec précision l'identité de l'unité statistique et la valeur de l'observation.

Le registre sert à garder les renseignements à utiliser plus tard.

CHAPITRE I : ECHANTILLONNAGE

Dans ce chapitre, vous allez vous familiariser avec des notions d'échantillonnage aléatoire simple et de processus de sélection d'un échantillon.

I.1 Quelques définitions

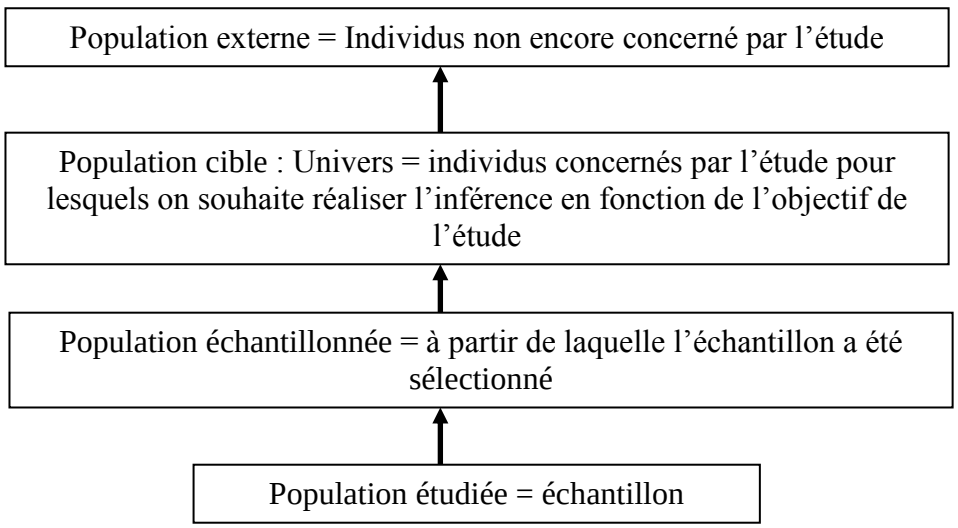
a) Population

La population est l'ensemble de tous les éléments considérés dans une étude c'est-à-dire un ensemble des éléments qui ont au moins une propriété en commun. Encore plus, nous dirons qu'une population est l'ensemble des unités ou individus sur lequel on effectue une analyse statistique.
Un recensement est une enquête sur la totalité de la population.

b) Echantillon

Un échantillon est un sous-ensemble de la population. C'est un ensemble d'individus prélevés dans une population déterminée. Il est constitué d'éléments sélectionnés dans la population et sur lesquels l'enquête sera réalisée. Cet échantillon doit être représentatif. Un échantillon est dit représentatif s'il renferme toutes les caractéristiques d'une population. Ce qui revient à considérer que chaque élément de la population a une même chance d'appartenir à un même échantillon.
La sélection de l'échantillon se fait à partir d'une liste appelée base d'échantillonnage ou de sondage. A partir des résultats obtenus dans l'échantillon, des conclusions pourront être tirées pour la population et ceux, en calculant des estimations des valeurs caractéristiques de la population ; estimation dont il sera important de pouvoir juger de la qualité en termes de précision.

Hiérarchie des populations et processus d'inférence Statistique



Les échantillons peuvent être probabilistes ou non probabilistes :

- Un échantillon probabiliste est tel que chaque élément de la population a une probabilité connue d'être inclus dans l'échantillon
- Un échantillon non probabiliste est tel que l'on ne peut définir une probabilité de sélection pour chaque élément de la population.

A la catégorie des échantillonnages non probabilistes appartient en autre la méthode de Quota dans laquelle, on laisse à l'enquêteur le soin de choisir le sujet qui l'inclus dans l'échantillon à condition qu'il respecte certains critères comme par exemple avoir des nombres précis d'hommes et femmes dans un certains groupes d'âges.

c) Taille de l'échantillon

Une étude répétée avec la même procédure sur les échantillons différents, conduit en général à un résultat différent. La précision de l'inférence dépend de la taille de l'échantillon. Pour avoir une bonne image de la population ou pour arriver à tirer des conclusions à certains paramètres de la population, il faut des effectifs suffisants. Pour atteindre les objectifs fixés, il est donc extrêmement important, avant d'entreprendre une étude, d'estimer au préalable la taille de l'échantillon à inclure.

I.2 Méthodes d'échantillonnage

Dans ce paragraphe, nous verrons comment sélectionner un échantillon à partir d'une population donnée. Pour sélectionner un échantillon probabiliste, il y a plusieurs techniques ou schémas d'échantillonnage qui sont utilisés. Nous citerons par exemple l'échantillonnage aléatoire simple et l'échantillonnage systématique.

a) Echantillonnage aléatoire simple.

Plusieurs méthodes peuvent être utilisées pour sélectionner un échantillon dans une population ; l'une des plus courantes est l'échantillonnage aléatoire simple. La définition d'un échantillon aléatoire simple et la procédure de sélection d'un échantillon aléatoire simple varient selon que la population est finie ou infinie.

a1. Cas d'une population finie.

Un échantillon aléatoire simple de taille n, issu d'une population finie de taille N, est un échantillon sélectionné de manière à ce que chaque échantillon possible de taille n ait la même probabilité d'être sélectionné.

La procédure de sélection d'un échantillon aléatoire simple, à partir d'une population finie, consiste à choisir les éléments de l'échantillon un par un, de façon à ce que les éléments restants dans la population aient la même probabilité d'être sélectionnés. Choisir n éléments de cette façon respecte la définition d'un échantillon aléatoire simple issu d'une population finie.

Admettons que nous avons la suite, 160, 149, 966, 911, 348, 321, 814, 418, 260, 267,

…..,727, 949, 503. Admettons que la taille de la population N = 648 et que l'échantillon est n= 18

1°) choisir « les yeux fermés » un point de départ de « 3 chiffres » sur la table (tableau,…)

2°) choisir (avant de démarrer) la direction selon laquelle on va se déplacer sur la table (tableau, ...)

3°) lire tous les nombres de « 3 chiffres » et garder les 18 premiers qui partent de 1 à 648.

a2. Cas d'une population infinie

De nombreuses situations dans le domaine commercial ou en économie impliquent des populations finies, mais dans certains cas, la population est soit infinie soit tellement importante qu'elle est traitée comme une population infinie pour des raisons pratiques. Dans l'échantillonnage à partir d'une population infinie, nous utilisons une nouvelle définition de l'échantillon aléatoire simple. De plus, puisque les éléments d'une population infinie ne peuvent pas être numérotés, nous devons utiliser une procédure de sélection différente des éléments de l'échantillon. En pratique une population est considérée infinie s'il est impossible de compter les éléments de la population.

Supposons que nous voulons estimer le temps moyen écoulé entre le moment où la commande est passée et le moment où elle est servie, dans un fast-food entre 11h30' et 13h30'. Si l'on considère la population formée par tous les clients potentiels, il n'est pas possible de fixer une limite finie au nombre de clients possibles. En fait, si l'on définit la population comme étant tous les clients qui peuvent éventuellement venir à l'heure de déjeuner, on peut considérer que la population est infinie. Le but sera de constitué un échantillon aléatoire simple de n clients dans cette population.

Un échantillon aléatoire simple issu d'une population infinie, est défini comme étant un échantillon sélectionné de façon à ce que les conditions suivantes soient- satisfaites :

- Chaque élément sélectionné provient de la même population
- Chaque élément est sélectionné de façon indépendante.

Remarque

1. Les populations finies sont souvent définies par des listes, telles que la liste des membres d'une organisation, les dossiers d'inscription des étudiants, les listes des comptes courants, l'inventaire des produits, etc. les populations infinies sont souvent définies par un processus qui génère les éléments de la population et qui se poursuit indéfiniment sous les mêmes conditions. Dans ce cas, il est impossible d'obtenir une liste de tous les éléments de la population. Par exemple, les populations correspondant à toutes les pièces pouvant être fabriquées, à tous les clients potentiels, à toutes les transactions bancaires possibles etc. peuvent être considérées comme des populations infinies
2. Le nombre d'échantillons aléatoires simples différents de taille n qui peuvent être sélectionnés à partir d'une population de taille N est :

$$\frac{N!}{n!\,(N-n)!}$$

b) Echantillonnage systématique

Dans certaines situations, spécialement lorsque les populations sont importantes, il est coûteux (en temps) de sélectionner un échantillon aléatoire simple en trouvant tout d'abord un nombre aléatoire et ensuite en cherchant dans la liste de la population l'élément correspondant. Une alternative à l'échantillonnage aléatoire simple est l'échantillonnage systématique. Par exemple, si l'on souhaite sélectionner un échantillon de taille 50 parmi une population contenant 5000 éléments, cela revient à sélectionner un élément tous les 5000/50 = 100 éléments de la population. Constituer un échantillon systématique dans ce cas consiste à sélectionner aléatoirement un élément parmi les 100 premiers de la liste de la population. Les autres éléments de l'échantillon sont identifiés de la façon suivante : le second élément sélectionné correspond au 100^{e} élément qui suit le premier élément sélectionné dans la liste de la population, le troisième élément sélectionné correspond au 100^{e} élément qui suit dans la liste de la population le second élément sélectionné, et ainsi de suite. En fait, l'échantillon de taille 50 est identifié en se déplaçant systématiquement dans la population et en identifiant le 100^{e}, 200^{e} et 300^{e}, etc. élément qui suit le premier élément choisi aléatoirement.

D'une manière générale, l'échantillonnage systématique consiste à choisir les éléments (unités) de l'échantillon en parcourant une base de sondage ou d'échantillonnage progressivement et en sélectionnant les éléments régulièrement à intervalle fixe (pas) ; le premier élément étant choisi aléatoirement. Voici les étapes à suivre :

1°) Calculer le pas « k » : $k = {}^{N}/_{n}$

2°) Par tirage aléatoire simple, choisir un point de départ (aléatoire) entre 1 et k. prenons i.

3°) inclure systématiquement dans l'échantillon le sujet $i,\ i+k,\ i+2k,\ i+3k,$ ...

c) *Méthode de stratification*

Lorsque la population est hétérogène face au caractère faisant l'objet de l'étude statistique, on recourt à la méthode de stratification.

Description de la méthode

- La population est subdivisée (partitionnée) en k classes : C_1, C_2, …, C_k plus ou moins homogènes face au caractère faisant l'objet de l'étude statistique. Cette partition est obtenue en faisant une étude préalable socio-économico-démographique et sanitaire.

Le nombre d'individu appartenant à chaque strate et qui devra faire partie de l'échantillon est donné par la formule suivante :

$n_i \cong$ où n_i = nombre d'individu à tirer dans la strate

N_i = taille de la strate

n = taille de l'échantillon

N = taille de la population.

- Le choix des n_i individus se fait par une méthode aléatoire dans chaque strate.

NOTA : La strate, pris dans son ensemble, doit :

- couvrir toute la zone d'enquête sans rien omettre,
- ne pas se chevaucher (empiètement). Chaque unité statistique fait partie d'une strate et une seule.

$N_i \cap N_j = \emptyset \qquad i \neq j.$

d) ***Les méthodes non aléatoires ou à choix raisonné.***

L'échantillon obtenu par choix raisonné est constitué d'unités statistiques qui n'ont pas été tirés au hasard. De ce fait, elles n'ont pas la même chance d'appartenir à un échantillon.

Nous parlerons ici de la ***Méthode des quotas*** *(utilisée surtout dans les sondages).*

On constitue un échantillon de manière à ce que certaines proportions observées dans la population se retrouvent dans l'échantillon. Dans les limites qui lui seront fixées, l'enquêteur reste libre d'interroger les unités statistiques qu'il veut. Il devra respecter les quotas qui lui sont imposés.

I.2 RAPPELS SUR L'ANALYSE COMBINATOIRE

Principe fondamental

Si une procédure quelconque peut être représentée de n_1 façons différentes, si après cette procédure, une seconde peur être représentée de n_2 façons différentes, et si ensuite une troisième procédure peut être représentée de n_3 façons différentes, et ainsi de suite, alors le nombre de façons différentes permettant d'exécuter les procédures dans l'ordre indiqué est égal au produit $n_1.n_2.n_3\ldots$

I.2.1 Notion

- Soit à choisir deux livres parmi cinq : un livre de statistique (s), un livre de mathématiques (m), un livre d'économie politique (e), un livre d'anglais (a) et un livre de français (f).

  On a les possibilités suivantes :

  a-e
  a-f e-f
  a-s e-s f-s
  a-m e-m f-m s-m

  On constate qu'il y a 10 manières différentes possibles de choisir deux livres parmi les cinq proposés.
- Si maintenant nous voulons déterminer les différentes manières de former les quintés de livres parmi 25 livres, nous constatons tout de suite qu'il n'est pas aisé de constituer ces différents groupements.

Les différentes manières dont on a groupé les livres constituent ce que l'on appelle **dénombrement**.
L'analyse combinatoire joue un grand rôle dans l'étude de la statistique et de probabilité car elle offre les principes et les lois pour déterminer les nombres de groupements des éléments donnés.

I.2.2 Arrangement et permutation

I.2.1. Arrangement sans répétition

Définition On appelle arrangement sans répétition de n éléments pris p à p, une injection f de {1, 2, …, p} vers {1, 2, …, n}
Avant de donner la formule de l'arrangement, considérons d'abord un petit d'exemple.

Exemple :

Calculer le nombre d'arrangements de 6 objets a,b,c,d,e et f pris trois à trois. En d'autres termes, calculer le nombre de « mots de trois lettres » avec des lettres distinctes que l'on peut former à partir des six lettres précédentes.

On peut choisir la première lettre de 6 façons différentes, la seconde de 5 façons différentes et la dernière de 4 façons différentes. Ainsi d'après le principe fondamental de l'analyse combinatoire, il y a 6.5.4 = 120 mots de trois lettres possibles sans répétition, à partir de six lettres ou encore 120 arrangements de 6 objets pris trois à trois.

Le premier élément d'un arrangement de p objets pris dans un ensemble de n objets peut être choisi de n façons différentes. Il y a ensuite n – 1 façons de choisir le deuxième élément de l'arrangement, et n – 2 façons différentes de choisir le troisième élément. En continuant de cette façon, on voit qu'il y a n – (p – 1) = n – p + 1 façons différentes de choisir le p-ième (ou dernier) élément.

Notation et Calcul : $A_n^p = n(n-1)(n-2)\dots(n-p+1) = \frac{n!}{(n-p)!}$

Exemple : Le nombre d'arrangement sans répétition de 3 éléments {x,y,z} pris 2 à 2 est :

$$A_3^2 = \frac{3!}{(3-2)!} = 6$$

Ces 6 arrangements sont : (x,y) ; (x,z) ; (y,x) ; (y,z) ; (z,x) et (z,y)

REMARQUES

- Deux arrangements sans répétition se diffèrent :
  - soit par la nature des éléments : (x,y) ≠ (y,z)
  - soit par l'ordre des éléments : (x,y) ≠ (y,x)
- on interprète parfois un arrangement sans répétition de n éléments pris p à p comme un tirage sans remise dans une urne.

I.2.3. Arrangement avec répétition

Définition : On appelle arrangement avec répétition de n éléments pris p à p, une application f de {1, 2, …, p} vers {1, 2, …, n}

Notation et Calcul : $\alpha_n^p = n^p$

Exemple : Soit {1, 2}. Le nombre d'arrangements avec répétition des éléments 1 et 2 pris 3 à 3 est : $\alpha_2^3 = 2^3 = 8$
Ces huit arrangements avec répétition sont : (1,1,1) ; (1,1,2) ; (1,2,1) ; (1,2,2) ; (2,1,1) ; (2,1,2) ; (2,2,1) et (2,2,2)

REMARQUES

- Deux arrangements avec répétition se diffèrent :
  - soit par la nature des éléments : (1,1,1) ≠ (1,1,2)
  - soit par l'ordre des éléments : (1,2,1) ≠ (2,1,1)
- on interprète parfois un arrangement avec répétition de n éléments pris p à p comme un tirage avec remise dans une urne. (plus de précision au chapitre consacré à la probabilité).

I.2.3. Permutation

I.2.3.1 Permutation sans répétition

Définition : On appelle permutation sans répétition de n éléments distincts une bijection d'un ensemble à n éléments vers un autre ensemble à n éléments.

N.B : dans ce cas nous pourrons dire que la permutation est un cas particulier de l'arrangement sans répétition où p = n

Notation et calcul : $P_n = n!$

Exemple : Le nombre de permutations sans répétition des éléments de {x,y,z} est $P_3 = 3! = 3 * 2 * 1 = 6$

Ces 6 permutations sont : (x,y,z) ; (x,z,y) ; (y,x,z) ; (y,z,x) ; (z,x,y) et (z,y,x)

I.2.3.2 Permutation avec répétition

Très souvent, on désire connaitre le nombre de permutations qu'il y a parmi des objets dont certains sont semblables.

Définition-Notation et calcul : Soit F un ensemble fini à n éléments dont certains se répètent n_1, n_2, …, n_k fois. Le nombre de permutations avec répétition de ces n éléments est donné par :

$$P_n^n = \frac{n!}{n_1!\, n_2! \ldots n_k!}$$

La formule ci-dessus donne le résultat du nombre de permutations de n objets dont n_1 sont semblables, n_2 sont semblables, ……., n_p sont semblables.

Exemple : le nombre de permutations avec répétition des chiffres du nombre 7299828522 est : $P_{10}^{10} = \frac{10!}{2!2!4!} = 37800$

(2 se répète 4 fois, 9 se répète 2 fois et 8 se répète 2 fois)

Exemple2 : Combien de signaux différents, chaque signal étant constitué de 8 pavillons alignés verticalement, peut-on former d'un ensemble de 4 pavillons rouges indiscernables, 3 pavillons blancs indiscernables et un pavillon bleu ?

$$P_8^8 = \frac{8!}{4!\, 3!} = 280$$

I.2.4. COMBINAISONS

I.2.4.1 Combinaison sans répétition

Définition : Soit F_n un ensemble fini et p un naturel tel que p≤n. On appelle combinaison sans répétition de n éléments de F_n pris p à p, une partie à p éléments de F_n.

Notation et calcul : $C_n^p \; ou \; \binom{p}{n} = \frac{n!}{p!(n-p)!}$

Exemple : pour choisir trois livres parmi sept, nous avons : $C_7^3 = \frac{7!}{3!(7-3)!} = 35$ possibilités de le faire.

REMARQUE

Deux combinaisons sans répétition ne diffèrent que par la nature des éléments (x,y)≠(x,z) ; l'ordre des éléments ne compte pas (x,y)=(y,x)

1.2.4.2 Combinaison avec répétition

Définition : Soit F_n un ensemble fini et p un naturel tel que p ≤ n. Le nombre de combinaisons avec répétition de n éléments de F_n pris p à p, est le nombre noté D_n^p et défini par : $D_n^p = C_{n+p-1}^p$

Exemple : Soit à tirer successivement 3 cartes d'un jeu de 52 cartes en remettant la première avant de tirer la suivante.

Nous avons $D_{52}^3 = C_{52+3-1}^3 = 24\,804$ possibilités car il s'agit d'une combinaison (l'ordre des cartes ne compte pas) et une carte peut être tirée plusieurs fois.

Propriétés de la combinaison

1) Quels que soient deux naturels n et p tels que p ≤ n, on a : $C_n^p = C_n^{n-p}$.
2) Pour tout entier naturel n, on a :
$$C_n^0 = 1 = C_n^n \quad \text{et} \quad C_n^1 = n = C_n^{n-1}$$
3) Quels que soient deux naturels n et p tels que 1 ≤ p ≤ n-1, on a :
$$C_{n-1}^{p-1} + C_{n-1}^p = C_n^p$$

Exercices

1) Calculer 4! , 5! , 6! , 7! , 8!

2) Calculer
   i) $\frac{13!}{11!}$
   ii) $\frac{7!}{10!}$

3) Simplifier
   i) $\frac{n!}{(n-1)!}$ ii) $\frac{(n+2)!}{n!}$

4) En supposant qu'il n'y a pas de répétitions
   i) Combien de nombres de 3 chiffres peut-on former à l'aide des six chiffres 2, 3, 5, 6, 7 et 9 ?
   ii) Combien de ces nombres sont inférieurs à 400 ?
   iii) Combien sont pairs
   iv) Combien sont impairs
   v) Combien sont des multiples de 5 ?

Réponse
i) 120 ; ii) 40 ; iii) 40 iv) 80 ; v) 20

5) Combien de façons différentes peut-on répartir un groupe de 7 personnes
   i) Sur une rangée de 7 chaises
   ii) Autour d'une table ronde

Réponse
i) 7! ii) 6!

6) Supposons qu'unr urne contient 8 boules. Déterminer le nombre d'échantillons de taille 3
   i) Non exhaustifs,
   ii) Exhaustifs

Réponse
i) 512 ii) 336

7) De combien de manières peut-on former un jury de 3 hommes et 2 femmes parmi 7 hommes et 5 femmes ?

Réponse
350 manières

8) Une délégation de 4 lycéens est choisie chaque année pour suivre le congrès annuel de l'Association des Parents d'Elèves.
   i) De combien de manières peut-on former la délégation s'il y a 12 lycées éligibles ?
   ii) De combien de manières, si deux des lycées éligibles refusent de suivre le congrès ensemble ?
   iii) De combien de manières si deux des lycées éligibles sont des frères jumeaux et ne pourront suivre le congrès qu'ensemble ?

Réponse
i) 495 manières ii) 450 manières iii) 255

9) A l'oral d'un examen, un étudiant doit répondre à 8 questions sur un total de 10.
   i) Combien de choix possibles y-at-il ?
   ii) Combien de choix y a-t-il s'il doit répondre aux 3 premières questions ?
   iii) Combien de choix y a-t-il s'il doit répondre au moins à 4 des 5 premières questions ?

Réponse
i) 45 ii) 21 iii) 35

10) On veut former un comité comprenant 3 des 20 personnes d'un groupe. Combien y a-t-il de ces comités ?

11) De combien de manière peut-on asseoir 8 personnes en rang si :
- Aucune restriction n'est mise ;
- Les personnes A et B veulent être ensemble ;
- Les hommes ne doivent avoir que des voisines et inversement, en supposant qu'il y a 4 hommes et 4 femmes ;
- Les hommes, qui sont au nombre de 5, doivent rester ensemble
- Les personnes forment 4 couples de gens mariés et si chaque couple doit rester réuni ?

12) Une personne a 20000 dollars à placer sur 4 affaires potentielles. Chaque investissement doit être un nombre entier de milliers de dollars et il existe un engagement minimum pour chaque affaire que l'on retiendra. Ces minima sont

respectivement 2, 2, 3 et 4 milliers de dollars. Combien de stratégies d'investissement y a-t-il si :

- Un investissement doit être fait sur chaque affaire ;
- Au moins 3 des 4 affaires doivent être couvertes ?

13) Huit nouveaux professeurs vont être envoyés dans 4 écoles.

- Combien y a-t-il d'affectations possibles
- Qu'en est – il si l'on impose que chaque école recevra deux professeurs ?

14) Développer $(3x^2 + y)^5$

CHAPITRE II. ELABORATION ET PRESENTATION DES DONNEES STATISTIQUES.

Une population peut être présentée sous forme de tableau et éventuellement sous forme graphique ou diagramme.

II.1 Dépouillement

Les données brutes de recensement, de sondage ou de mesure directe sont d'abord reçues en vrac. On appelle données brutes, les données qu'on a rassemblé sans se soucier de la notion d'ordre. Citons par exemple la suite de taille de 120 étudiants pris par ordre alphabétique.

Une suite ordonnée est l'arrangement des données numériques par valeur croissante ou décroissante. La différence entre la plus grande et la plus petite valeur s'appelle « étendue des données ». Si parmi les 100 étudiants précédents le plus élancé mesure 196 cm et le plus petit 156 cm, l'étendue est de $196 - 156 = 40\ cm$

Le dépouillement consiste à passer des renseignements reçus en vrac en un groupement ordonné des données dans des tableaux, état qui prétend à l'analyse et à l'interprétation facile.

Pour les enquêtes peu importantes, on pratique le dépouillement manuel, par contre pour les enquêtes plus importantes de nos jours, on pratique le dépouillement à l'aide d'une machine appelée « ordinateur ».

II.2. Tableaux statistiques.

1. *Tableau du premier ordre.*

   Ce type de tableau comprend une seule variable.

Tableau n°1.1 : Causes de décès chez les malades hospitalisés dans un centre hospitalier :

N°	Causes de décès	Nombre de décès
1.	Rougeole	20
2.	Gastro-entérite	18
3.	Tuberculose	15
4.	Malaria	14
5.	Accouchement dystocique	10
6.	Malnutrition	9
7.	Tétanos	9
8.	Hernie étranglée	8
9.	Trypanosomiase	7
10.	Accident de circulation	6
11.	Malformation congénitale	3
12.	Autres causes	16
	Total	129

2. *Tableau du deuxième ordre ou tableau à double entrée.*

C'est un tableau comprenant deux variables.

Tableau n°1.2. Distribution par âge et par sexe des cas de cancer de poumons au cours d'une année dans les hôpitaux de Kindu.

Age (ans)	Sexe		Total
	Masculin	Féminin	
1-5	14	5	19
6-10	15	17	32
11-15	24	23	47
16-20	42	18	60
21 et plus	43	36	79
Total	138	99	237

3. *Tableau du troisième ordre ou tableau à triple entrée.*

C' est un tableau contenant les données relatives à trois variables.

Tableau n°1.3. Répartition par âge, par sexe et par groupe de maladie des patients ayant reçu des soins ambulatoires dans un Centre de Santé.

	Tranches d'âges						
	0-15		16-45		>45		Total
	M	F	M	F	M	F	
Maladies infectieuses	10	20	12	13	16	2	73
Troubles nutritionnelles	22	25	40	10	15	8	120
Traumatisme	30	25	11	14	18	16	114
Autres affections	12	10	13	15	6	15	71
Total	74	80	76	52	55	41	378

4. Tableau de contingence.

C'est un tableau où sont présentées 2 variables comportant chacune un certain nombre des classes.

Tableau n°1.4. Etat nutritionnel et résultats scolaires de 70 élèves d'une école secondaire de Lubumbashi.

Résultats scolaires	Etat nutritionnel		Total
	Bon	Médiocre	
Bons	11	15	26
Médiocres	8	26	34
Total	19	41	70

II.3 Groupement en série

Le groupement (la distribution) en série consiste à placer par ordre croissant ou décroissant les valeurs ou individus de l'échantillon.

Exemple

Soit un échantillon de 40 individus

49	61	55	48	59	49	56	55	50	59
51	51	56	53	58	57	50	50	53	55
52	55	50	57	54	54	51	56	54	53
56	53	52	51	51	53	52	56	52	53

Comment présenter cette donnée d'une façon ordonnée ?

48	49	49	50	50	50	50	51	51	51
51	51	52	52	52	52	53	53	53	53
53	53	54	54	54	55	55	55	55	56
56	56	56	56	57	57	58	59	59	61

On constate que la plus grande valeur est 61 et la plus petite 48.

II.4 La méthode de pointage

On classe les données dans un tableau à deux colonnes où chaque valeur sera représentée par un certain nombre de traits. Selon cette pratique un 5^{e} trait barre les 4 premiers de façon à

former les groupes de 5 unités qui faciliteront le dénombrement final. Il existe deux types de pointage : « pointage en bâton et en carreaux

II.5. Distribution des fréquences.

La distribution des fréquences est le tableau construit après le pointage des données. Elle permet de constater immédiatement certains phénomènes qui caractérisent l'ensemble des données. On l'enrichit parfois avec des fréquences cumulées, des fréquences relatives et des fréquences relatives cumulées.

5.1. *Tableau de distribution des fréquences d'une variable discrète.*

Le tableau est construit en mettant dans la première colonne les diverses valeurs $(x_1 , x_2 , \cdots x_n)$ que prend la variable, en ordre croissant ; et dans la seconde colonne, les effectifs correspondants (ou fréquences).

Soit x_i les valeurs observées ; n_i l'effectif absolu des valeurs observées ; f_i la fréquence relative des valeurs observées avec : $f_i = \frac{n_i}{n}$; n : la taille de l'échantillon.

En reprenant le tableau de l'exemple précédent, la distribution de fréquence peut être donnée par :

x_i	n_i	f_i	f_{ca}	f_{cd}
48	1	0,025	1	40
49	2	0,05	3	39
50	4	0,1	7	37
51	5	0,125	12	33
52	4	0,1	16	28
53	6	0,15	22	24
54	3	0,075	25	18
55	4	0,1	29	15
56	5	0,125	34	11
57	2	0,05	36	6
58	1	0,025	37	4
59	2	0,05	39	3
60	0	0	39	1
61	1	0,025	40	1
	40	1		

Exercice

Soit à déterminer la distribution de fréquence de l'échantillon ci-après : (respecter les étapes)

51	77	64	66	62	53	67	68	67	73
73	76	73	50	76	70	74	75	52	54
56	63	65	69	61	61	60	72	70	75
71	72	70	74	65	71	72	73	71	55
58	50	57	58	67	72	59	62	60	56

5.2. *Tableau de distribution des fréquences d'une variable continue.*

Dans le cas d'une variable continue, la présentation sous forme de tableau requiert de longs calculs car le nombre de valeur est élevé.

On évite cette situation en effectuant un groupement des données en classes.

Groupement des données en classes

Pour résumer une grande quantité des données brutes, il est nécessaire de les distribuer en classe ou catégories, et de déterminer le nombre d'individus appartenant à chaque classe, que l'on appelle aussi « fréquence ou effectif de la classe ».

Le groupement des données en classe consiste à remplacer toutes les valeurs ou caractères situés dans un intervalle borné par une valeur unique dite « centre de classe » à laquelle on attribue un effectif égale à la somme des effectifs des valeurs, des caractères appartenant à cet intervalle.

Sous forme mathématique on a : $x \in [a\ ,b[$

a) Découpage en classe

a1) l'intervalle des classes : c'est l'intervalle dans lequel doit être compris une donnée qui appartient à cette classe. Il peut être ouvert ou fermé à gauche ou à droite.

a2) Les bornes d'une classe : une classe donnée comporte toujours une borne ou limite. La limite inférieure Li et la limite supérieure Ls qui sont respectivement la plus petite valeur de la classe et la valeur supérieure appartenant à la classe. Elles sont dites « limites brutes ».

A cause des incertitudes des mesures, les bornes ou limite des classes ne sont réelles ou vraies que si l'on tient compte de l'étendue de précision. Dans ce cas, on soustrait à Li 0,5 et on ajoute à Ls 0,5. L'amplitude augmente d'une unité.

Exemple

$[48\ ,50[$: les limites vraies sont : $[47{,}5\ ;50{,}5[$

a3) L'amplitude d'une classe, l'intervalle ou longueur ou dimension : c'est la différence entre la limite supérieure et la limite inférieure de la classe lorsque l'intervalle est ouvert. Par contre, il faut ajouter une unité à cette différence si l'intervalle est fermé.

a4) L'étendue de la série : c'est la différence entre la plus grande valeur et la plus petite valeur de la distribution.

Exemple : $61 - 48 = 13$

a5) Le centre des classes : c'est le point correspondant au milieu de cette classe

$$[48\,,50[\qquad x_i = 49$$

Tableau n°2 : Distribution des fréquences des ouvriers d'une clinique suivant leur âge :

Age (classes)	Effectif ni	Centre de classe	Effectif cumulé N_i	Fréquence relative cumulée F_i
[20,25[	9	22,5	9	9/150
[25,30[	27	27,5	36	36/150
[30,35[	36	32,5	72	72/150
[35,40[	45	37,5	117	117/150
[40,45[	18	42,5	135	135/150
[45,50[	9	47,5	144	144/150
[50,55[	6	52,5	150	150/150
$\sum n_i = 150$				

b) Exercices

On demande de procéder au groupement en classe des données des exemples ci-dessus sachant que l'intervalle est fixé à la valeur 2.

1°) En tenant compte des limites brutes

2°) En tenant compte des limites réelles.

- *Détermination du nombre de classes*

La formule de Sturges ou de Liorzou donne le nombre de classes k à considérer :

$k = 1 + \frac{10 \log n}{3}$ où n est la taille de l'échantillon.

k peut être donné à l'avance.

1. On calcule d'abord l'étendue de la série

$$d = X_{max} - X_{min}$$

2. On calcule ensuite l'intervalle des classes (amplitude ou dimension)

$$a = \frac{d}{k-1}$$

3. On détermine l'étendue de travail W.

$$W = a.k$$

4. On détermine la limite inférieure des classes

$$L_i = X_{min} - \frac{a}{2}$$

5. On détermine la limite supérieure des classes

$$L_i = L_i + W$$

Exemple : Utiliser la méthode de sturges pour représenter les données de l'exercice précédent en classe.

5.3. *Tableau des effectifs cumulés (Distribution des effectifs cumulés)*

- *Cas d'une valeur discrète*

Les effectifs cumulés sont les nombres d'observations inférieures ou égales à une valeur donnée de la variable.

- *Cas d'une variable continue*

Les effectifs cumulés sont les nombres d'observations inférieures ou égales aux limites supérieures d'une classe.

La distribution cumulée des effectifs correspondant aux cas discret et continue se présente de la manière suivante :

Centre de classes	Fréquences relatives cumulées f_i
x_1	f_1
	$f_1 + f_2$
x_2	$f_1 + f_2 + f_3$
x_3	$f_1 + f_2 + \ldots + f_k$
.	
.	
.	
x_k	

Les N_i sont les effectifs cumulés ou fréquence absolue cumulée.

Tableau n°1.7. Tableau statistique général de présentation des données

Centre de classe	Fréquences ou effectifs (n_i)	Fréquences relatives (f_i)	Fréquences relatives cumulées (Fi)
x_1	n_1	$f_1 =$	$F_1 = f_1 =$
		$f_2 =$	$F_2 = f_1 + f_2 = \ =$
x_2	n_2	$f_3 =$	$F_3 = f_1 + f_2 + f_3 =$
		$f_k =$	$F_k = f_1 + f_2 + \ldots + f_k =$
x_3	n_3		
	.		
x_k	n_k		

II.6. Représentation graphique des données

Une population peut être présentée sous forme de tableau et éventuellement sous forme de graphique ou diagramme.

La distribution des fréquences est souvent représentée sous forme graphique. On distingue plusieurs types de graphiques entre autre le polygone des fréquences, l'histogramme, la courbe de fréquence cumulée et la courbe de pourcentage (l'ogive)

Pour la représentation graphique on trace d'abord les axes de référence ou axes coordonnés appelé encore axe rectangulaire dont l'un horizontal est l'axe des x ou encore l'axe des abscisses et l'autre vertical est l'axe des y ou encore l'axe des ordonnées. Généralement les abscisses représentent les valeurs ou les intervalles des classes tandis que les ordonnées sont des fréquences.

II.6.1.Cas d'une variable discrète

a). Le diagramme en bâtonnets

En considérant l'exemple du tableau 5.1. En chacun des points traçons parallèlement à l'axe des ordonnées (y) un bâton de longueur proportionnel à l'effectif correspondant. Nous obtenons un diagramme en bâton.

b). Polygone des fréquences.

On l'obtient en reliant les sommets successifs du diagramme en bâtons.

II.6.2.Cas d'une variable continue.

a) Histogramme des effectifs

L'histogramme est la représentation de la population lorsqu'on a procédé au groupement des scores par classe. Il est constitué des rectangles contigus dont les surfaces sont proportionnelles aux fréquences. Chaque classe est représentée par un rectangle dont la base est l'intervalle de classe et dont la hauteur est le nombre des scores dans la classe. L'ensemble de tous ces rectangles adjacents constitue l'histogramme.
Reprenons l'exemple portant sur l'âge des ouvriers d'une clinique. Prenons un repère cartésien orthogonal sur l'axe des abscisses. Plaçons les points aux limites de chaque classe, point qui détermine ici le segment d'égale longueur. Sur chacun de ces segments, construisons des rectangles de hauteur proportionnelle à l'effectif de la classe considérée.

b).Polygone statistique des effectifs

En joignant par des segments de droite les milieux des côtés supérieurs des rectangles constituant l'histogramme, On obtient le polygone statistique (ou polygone des fréquences).

a) *Autre représentation graphique :* Graphique circulaire ou à secteurs circulaires

Problème : Sur les 250 étudiants entrés à la suite d'un concours à la Faculté des Sciences : 13, soit 5,2 % ont préparé le concours pendant 3 ans ; 96, soit 38,4 % ont préparé le concours pendant 2 ans ; 192, soit 44,8 %, ont préparé le concours pendant 1 ans et 29, soit 11,6 % n'ont pas préparé le concours d'entrée.
Représenter, graphiquement à l'aide d'un graphique circulaire, la ventilation, en % de la durée de préparation des étudiants admis au concours.
Le secteur angulaire correspondant à chaque % est par exemple pour 5,2% sera :

29
13
96
192

II.7 Classification des variables ou caractères

La nature des variables (propriétés ou phénomènes communs à tous les individus) est un élément décisif pour un choix adéquat des méthodes statistiques. Ces caractères peuvent

adopter différentes modalités ou valeurs. Et généralement une discussion de la statistique descriptive débute par cette classification.

Les variables peuvent être regroupées en deux : les variables quantitatives et les variables qualitatives (ou en catégories).

II.7.1 Les variables quantitatives

Elles expriment une quantité ; leurs valeurs définissent un ordre et les différences ou les rapports des valeurs d'une variable quantitative sont des nombres qui ont un sens. Les valeurs quantitatives peuvent être comptées ou mesurées et exprimées par des chiffres. Ces variables quantitatives sont subdivisées en deux :

a) Les variables quantitatives discrètes (ou discontinues)

Elles peuvent prendre certaines valeurs numériques et correspondent à un dénombrement. Elles peuvent être comptées. Ces valeurs sont des nombres entiers (sans décimal) ou encore un caractère est dit discret ou discontinu lorsqu'il ne peut prendre que les valeurs isolées dans un intervalle donné.

Exemple

- Le nombre d'accident par chauffeur
- Le nombre d'épisode de diarrhée
- Le nombre de pile et face lorsqu'on lance 30 fois une pièce de monnaie.

b) Les variables quantitatives continues

Un caractère est dit continu lorsqu'il est susceptible de prendre toutes les valeurs numériques dans un intervalle donné. Les valeurs possibles sont limitées par l'instrument de mesure. Ces variables possèdent en général une unité et peuvent comporter des nombres entiers avec décimal. Ces variables peuvent être mesurées et même exprimées par des fractions.

Exemple :

Le poids, la taille, l'âge (le temps), la distance, la vitesse, la température, le volume…

II.7.2 Les variables qualitatives (ou en catégories)

Elles qualifient les individus, expriment une qualité. Elles ne peuvent être chiffrées mais bien classées, répertoriées selon certaines valeurs, certaines particularités. Ces variables en catégories ou qualitatives peuvent être subdivisées en deux groupes :

a) Les variables du type qualitatif non ordonné

Ce type qualitatif non ordonné peut comprendre deux classes :

- Type qualitatif dichotomique (échelle 0 à 1) : par exemple un enfant décédé ou vivant à la sortie de l'hôpital , sexe masculin ou féminin, étranger ou pas, race noir ou pas
- Type qualitatif multichotomique : par exemple les groupes sanguins (A, B, AB, O), origine de l'eau de boisson (la source, rivière, le puits, la pluie), Etat civil (marié, célibataire, veuf, divorcé), la province d'origine au Congo démocratique…

b) Les variables du type qualitatif ordonné

Les valeurs occupent un ordre prédéterminé. Par exemple la cotation des étudiants (GD, D, S, A, R), l'évolution radiologique de la tuberculose pulmonaire avant et 6 mois après un traitement (amélioration considérable, amélioration modérée, statuquo, aggravation modérée, aggravation considérable, décès).

Remarque

Pour désigner les caractères quantitatifs donc mesurable, le terme variable est utilisé. Une variable est donc tout phénomène observable susceptible de prendre n'importe quelle valeur numérique du moins enter certaines bornes ou sur une échelle continue (cas d'une variable continue) ou discrète (cas d'une variable discontinue).

CHAPITRE III. DESCRIPTION SYNTHETIQUE DES VARIABLES QUANTITATIVES

III.1 Distribution des fréquences :

La distribution des fréquences est un tableau dans lequel à chaque valeur de la variable est associée le nombre des sujets (fréquences) ou la proportion des sujets (fréquence relative) ayant la valeur considérée (pour synthétiser et présenter les données). Lors de la construction d'une distribution des fréquences, il convient de respecter quelques règles (de bon sens en général)

1. Le nombre des classes devrait avoisiner 10 à 20
2. Les limites des classes doivent correspondre à la précision des données (ex : une taille de 65,5 cm et une limite de classe de 65,59 cm sont inappropriées)
3. Les intervalles des classes de largeurs égales sont souhaitées mais non essentiels
4. Les intervalles des classes doivent être mutuellement exclusifs
5. Eviter les classes ouvertes si possible, leurs graphiques sont difficiles à réaliser
6. Le centre des classes est essentiel à déterminer de fois, sur base de la limite vraie au lieu de la limite tabulée des classes.

Exemple

Limite tabulée	Limite vraie	Centre des classes
3.0 – 3.4	2.95 – 3.45	3.2
50 – 52	49.5 – 52.5	51
20 – 24	$[20 - 25[$	22.5

Voici quelques conseils pratiques pour la présentation d'une distribution des fréquences et pour d'autres données :

1. Les titres doivent être libellés clairement et avec précision (tables, colonnes, les lignes, totaux)
2. Les totaux doivent être indiqués
3. Si des pourcentages sont obtenus, la base du pourcentage doit être clairement indiquée
4. Les unités de mesure doivent être clairement indiquées.
5. Les tables expriment souvent clairement et avec concision que la prose
6. Les tables excessivement complexes doivent être évitées

Exemple : décès provoqués par la diphtérie en Angleterre pendant l'année 1900

Age (an)	Nombre de décès (Fréquence)	Fréquence relative % de décès
0 – 4	49479	62,93
5 – 9	23349	29,69
10 – 14	4092	5,20
15 – 19	1123	1,43
20 – 24	585	0,74
Total	78627	100

III.2 Exploration graphique

III.2.1 Stem end leaf

C'est un graphique fait des nombres. Il est constitué d'un tronc (stem) et des feuilles (leaf)

Exemple

Soient les nombres ci-après :

110	145	140	170	130	150	120	155	135	133
140	145	130	127	150	123	110	149	165	150
140	116	135	165	145	135	130	130	138	175
140	125	155	155	100	110	120	160	135	155
155	145	153	145	140	122	120	100	130	190
150	117	110	145	110	115	130	140	145	145
103	185	130	150	105	140	140	165	135	150
130	140	130	130	165	125	135	140		

Construction du « stem and leaf »

Le stem and leaf a l'avantage de combiner à la fois la représentation graphique et le tableau de distribution des fréquences.

III.2.2 L'histogramme

L'histogramme est constitué des rectangles (barres contigües) de bases égales à la largeur de classe, centrée sur les centres des classes. Sa surface est proportionnelle à la fréquence observée dans la classe. Avec les classes de même largeur, la hauteur des rectangles peut alors être égale à la fréquence. Lorsque les largeurs des classes sont variables, on peut porter comme hauteur du rectangle, la densité de fréquence (fréquence divisée par largeur des

classes)

- L'ordonnée débute à 0
- Le nombre de 5 à 20 classes est régulièrement cité. Il est en général choisi ampiriquement (bon sens) on choisira des classes commodes en fonction des valeurs et des unités de la variable.
- Un graphique ne doit pas masquer les caractéristiques de la distribution.

III.2.3 Polygone des fréquences

C'est une courbe qui relie tous les points situés au centre de chaque largeur supérieur de rectangle qui compose l'histogramme.

III.2.4 Le diagramme à bâtonnet (discontinue)

Il représente des caractères quantitatifs discrets (discontinus) et est caractérisé par le fait qu'on place des points régulièrement espacés sur l'abscisse. Ces points correspondent aux valeurs des caractères étudiés.

Exemple

Fréquence	5	17	31	20	11	4	1
Nombre d'enfant à charge	0	1	2	3	4	5	6

Quelques commentaires généraux sur les graphiques

- Un graphique doit aider le lecteur à mieux comprendre le messager (sans ambigüité)
- Les axes doivent porter des titres clairs et les unités de mesure doivent être indiquées
- Les échelles sont extrêmement importantes ; celles qui ne débutent pas à 0 doivent être interprétées avec précaution.

III.2.5 Le Box plot

Le Box plot est une boite rectangulaire dont les limites inférieures et supérieures sont respectivement égales au percentile 25 et au percentile 75 ; la médiane est placée dans la boite. Du centre des bords inférieurs et supérieurs de la boite partent deux segments verticaux limités par des valeurs « adjacentes ».

Le box plot permet de déterminer une asymétrie éventuelle de la distribution, de visualiser rapidement des valeurs extrêmes, de voir où se situe certains percentiles particuliers et de comparer la distribution d'une variable dans plusieurs distributions.

Quelques définitions

1. Les quartiles : les quartiles sont des groupes égaux constitués par un quart de sujet sur base des valeurs ordonnées par ordre croissant.

- Le premier quartile sont des sujets dont les valeurs sont inférieures au percentile 25
- Le troisième quartile sont des sujets dont les valeurs sont comprises entre P_{50} et P_{75}

2. Ecart interquartile (EIQ) = $P_{75} - P_{25}$

3. Valeur adjacente supérieure : c'est la plus petite valeur observée inférieure ou égale à $(P_{75} + 1{,}5 * EIQ)$

4. Valeur adjacente inférieure : c'est la plus grande valeur observée supérieure ou égale à $(P_{25} - 1{,}5 * EIQ)$

5. Les Outliers : sont des sujets dont les valeurs sont situées au-delà des valeurs adjacentes

6. Les extrêmes : sont des sujets dont les valeurs sont situées à plus de trois écarts interquartiles des bords de la boite (P_{25} et P_{75}).

III.3 Statistique de réduction et paramètre

Objectif : Résumer l'information contenue dans une série statistique. Pour se faire, on calcule des valeurs statistiques ou des statistiques.

Souvent, il est intéressant de décrire (résumer) les données avec un ou deux chiffres au lieu de se contenter des graphiques ou des distributions des fréquences. Les mesures synthétiques d'une variable dans la population s'appelle « paramètre ». Par contre, les mesures synthétiques calculées à partir des données d'un échantillon sont appelées « statistiques de réduction ». Les mesures synthétiques les plus couramment utilisés sont de quatre type :

- *Paramètres de tendance centrale ou de position ou de localisation.*

Ce sont des paramètres qui ont tendance à se positionner au centre d'une distribution statistique (moyenne arithmétique, mode, médiane, quartiles,…)

- *Paramètres de dispersion*

Ce sont des paramètres qui donnent les écarts entre les différentes valeurs et la moyenne arithmétique (Etendue, variance, écart-type, coefficient de variation,…)

- *Paramètres de forme*

Ce sont des paramètres qui caractérisent la forme de la courbe de fréquence (symétrie,

asymétrie)
- Kurtosis, peakedness, aplatissement

III.3.1. Paramètres de tendance centrale (ou de position ou de localisation)

Une tâche importante de la statistique est celle de faciliter l'intelligence d'un ensemble des nombres en synthétisant tous ces nombres en une valeur unique pour chaque façon d'envisager cet ensemble. La réduction des données a pour objet de donner l'image de l'ensemble des observations à l'aide des paramètres ou valeurs les plus représentatives de l'ensemble des observations. Ainsi, un seul nombre représentera la variabilité ; ou encore un seul nombre représentera la valeur moyenne.

III.3.1.1 Les moyennes

Généralement, on retient 4 types de moyenne : arithmétique, géométrique, harmonique et quadratique.

1. La moyenne arithmétique () ou moyenne

Parmi les moyennes des grandeurs, c'est la moyenne arithmétique qui est la mieux adaptée au calcul des nombreuses mesures : la corrélation, écart-type, la variance, la régression…

C'est une notion fondamentale en statistique. Cette mesure est maniable mathématiquement (d'où sa popularité) mais elle est affectée par les valeurs extrêmes (aberrantes).

Exemple

L'âge de décès de 5 personnes 34, 64, 68, 70 et 74 ans.

La moyenne est de 62 ans à cause de la petite valeur ; (4 sont plus grands que cette moyenne).

La moyenne est une valeur de tendance centrale qui dépend de la grandeur des éléments.

La moyenne arithmétique est égale à la somme des valeurs divisée par le nombre des valeurs ou encore la moyenne arithmétique est la somme des valeurs observées divisée par le nombre d'observation.

Remarque

Cette définition est valable pour les données non groupées en classes.

a) Distribution non groupée en classe

Soit une série statistique $X = \{x_1, x_2, \cdots, x_n\}$, la moyenne arithmétique notée $\bar{x}$ s'obtient par la formule :

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

avec n : le nombre d'observations des valeurs

Cette formule sous forme condensée s'écrit : $\sum x_i = n\bar{x}$; c'est la somme des éléments en

fonction de la moyenne. Cette forme d'équation montre que chacun des scores variables des individus pouvait se remplacer par le même score constant, celui de la moyenne et la somme totale des scores resterait inchangée. C'est là une interprétation importante de la notion de moyenne arithmétique.

Exemple

Soit la série suivante :

5 ; 8 ; 4 ;5 ;2 ; 10

n = 6

$\bar{x} = \frac{5+8+4+5+2+10}{6} = \frac{34}{6} = 5{,}6666$

Remarque

Si les valeurs observées x_i ont des fréquences absolues n_i.

On utilise la moyenne arithmétique pondérée donnée par : $\frac{\sum x_i \,.n_i}{n}$

ou encore $n\bar{x} = \sum x_i \,.n_i$

Exemple (cfr. Ex. suivant)

x_i	n_i	$n_i\, x_i$
2	1	2
4	1	4
5	2	10
8	1	8
10	1	10
	N = 6	34

$$= \frac{34}{6} = 5{,}6666$$

b) Distribution groupée en classe

Si x_i représente le centre des classes et n_i représente les effectifs correspondants, on définit la moyenne arithmétique comme étant le rapport entre la somme des produits de x_i par n_i et la somme de n_i. D'où :

$$\bar{x} = \frac{\sum_{i=1}^{c} x_i \,.n_i}{\sum_{i=1}^{c} n_i} = \frac{\sum_{i=1}^{c} x_i \,.n_i}{n}$$

Où c : est le nombre des classes

n : la taille de l'échantillon

Exemple d'une distribution d'âges

Soit une distribution dont on dispose les classes ci-après :

Age (ans)	Centre x_i	Fréq. Abs.n_i	$n_i . x_i$
$[20-25[$	22,5	9	202,5
$[25-30[$	27,5	27	742,5
$[30-35[$	32,5	36	1170
$[35-40[$	37,5	45	1687,5
$[40-45[$	42,5	18	765
$[45-50[$	47,5	9	427,5
$[50-55[$	52,5	3	157,5
$[55-60[$	57,5	3	172,5
		150	5325

= 5325/150= 35,5 ans

c) Méthode de calcul simplifiée de la moyenne arithmétique

A. Usage d'une variable auxiliaire

Ex : Calculer la moyenne arithmétique de : 12, 15, 18, 21, 24, 27

1°) $\bar{x} = \frac{12+15+18+21+24+27}{6} = \frac{117}{6} = 19{,}5$

2°) Choisissons une variable auxiliaire $x_0 = 18$, appelée moyenne provisoire et retranchons-la à toutes les valeurs

x_i	$x_i - x_0$
12	- 6
15	- 3
18	0
21	3
24	6
27	9

$\sum x_i = 117$ $\sum x_i - x_0 = 9$

La somme des nouvelles valeurs est 9, soit une moyenne de $\frac{9}{6} = 1{,}5$

En ajoutant 18, on retrouve la moyenne réelle c'est-à-dire 18 + 1,5 = 19,5

Cette méthode est dite simplification des calculs par soustraction.

Soit une série d'observation $x_1, x_2, \cdots, x_n$; on se fixe une valeur quelconque x_0 de cette série et on exprime que $x_1 = x_0 + (x_1 - x_0)$; $x_2 = x_0 + (x_2 - x_0)$; … ; $x_n = x_0 + (x_n - x_0)$

La somme $x_1 + x_2 + \cdots + x_n$ peut s'écrire sous la forme :

$$\sum_{i=1}^{n} x_i = nx_0 + \sum_{i=1}^{n} (x_i - x_0)$$

En posant $z_i = (x_i - x_0)$, on a $\sum_{i=1}^{n} x_i = nx_0 + \sum_{i=1}^{n} z_i$

En divisant les deux membres par l'effectif total n de la série, on obtient

$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{nx_0}{n} + \frac{\sum_{i=1}^{n} z_i}{n}$, or $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ et $\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n}$

D'où $\bar{x} = x_0 + \bar{z}$

N.B : Si la moyenne arithmétique $\bar{z}$ peut s'obtenir rapidement que $\bar{x}$, il y a intérêt d'utiliser la moyenne provisoire $\bar{x}_0$ pour déterminer la moyenne arithmétique $\bar{x}$ cherchée.

Exemple.

Soit à calculer la moyenne de la série suivante :

1255, 1355, 1455, 1555, 1655 et 1755

Soit $x_0 = 1455$ est la variable auxiliaire, on a $z_i = x_i - x_0 = x_i - 1455$

$z_1 = 1255 - 1455 = -200$

$z_2 = 1355 - 1455 = -100$

$z_3 = 1455 - 1455 = 0$

$z_4 = 1555 - 1455 = 100$

$z_5 = 1655 - 1455 = 200$

$z_6 = 1755 - 1455 = 300$

$\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n} = \frac{300}{6} = 50$

$\bar{x} = x_0 + \bar{z} = 1455 + 50 = 1505$

B. Usage d'une variable auxiliaire avec changement d'échelle.

Si x_0 est une moyenne provisoire et a l'amplitude des classes, toute valeur x_i de la variable concernée peut s'écrire sous la forme :

$x_i = x_0 + \frac{a(x_i - x_0)}{a}$ (1)

En posant $z_i = \frac{x_i - x_0}{a}$ on obtient $x_i = x_0 + az_i$ (2)

Cette méthode est dite simplification par soustraction et division

Exemple : Soit la série 12, 15, 18, 21, 24, 27

On choisit $x_0 = 18$. Retranchons 18 à chaque valeur on a : - 6, - 3, 0, 3, 6, 9. Ensuite divisons par 3, on a : - 2, - 1, 0, 1, 2, 3 = z_i

La somme est : $3 = \sum z_i$

$\frac{somme}{n} = \frac{3}{6} = 0{,}5$

$x_i = x_0 + a.z_i = 18 + 0{,}5\ .3 = 19{,}5$

B.1 Cas d'une série simple

La moyenne d'une série simple se note :

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

Si on remplace x_i par son expression (2), on obtient :

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_0 + az_i)}{n}$$

Or $\sum_{i=1}^{n}(x_0 + az_i) = \sum_{i=1}^{n} x_0 + a\sum_{i=1}^{n} z_i$

D'où $\bar{x} = \frac{nx_0 + a\sum_{i=1}^{n} z_i}{n} = \frac{nx_0}{n} + \frac{a\sum_{i=1}^{n} z_i}{n}$

$\bar{x} = x_0 + a.\bar{z}$

B.2 Cas d'une série groupée

La moyenne arithmétique est notée

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

Remplaçons x_i par son expression en z_i on a :

$x_i = x_0 + az_i$; $\bar{x} = \frac{\sum_{i=1}^{k} n_i(x_0 + az_i)}{n}$

Or : $\sum_{i=1}^{k} n_i(x_0 + a\ z_i) = \sum_{i=1}^{k} n_i x_0 + a\sum_{i=1}^{k} n_i z_i = nx_0 + a\sum_{i=1}^{k} n_i z_i$

$$\bar{x} = \frac{nx_0 + a\sum_{i=1}^{k} n_i z_i}{n} = x_0 + a\frac{\sum_{i=1}^{k} n_i z_i}{n} = x_0 + a\ \bar{z}$$

$\bar{x} = x_0 + a\bar{z}$

Exemple

Classes	n_i	x_i	$x_i - x_0$	z_i	$n_i z_i$
$[155 - 185[$	37	170	- 60	-2	-74
$[185 - 215[$	19	200	-30	-1	-19

$[215-245[$	26	230	0	0	0
$[245-275[$	11	260	30	1	11
$[275-305[$	7	290	60	2	14
	100				-68

$z_i = \frac{x_i - x_0}{a}$ avec a =30

$x_0 = 230$

$\bar{z} = \frac{-68}{100} = -0{,}68$

$\bar{x} = 230 + 30(-0{,}68) = 209{,}6$

Propriétés de la moyenne arithmétique

a) La moyenne arithmétique est sensible aux valeurs extrêmes de la série

Soient deux séries :

X = 1, 4, 5, 7, 8, 10, 15, 20, 25, 35

Y= 1, 4, 5, 7, 8, 10, 15, 20, 25, 155

$\bar{x} = \frac{130}{10} = 13$ et $\bar{y} = \frac{250}{10} = 25$

b) La moyenne arithmétique d'une série unique composée de deux séries dont on connait les moyennes arithmétiques est une moyenne arithmétique pondérée des moyennes arithmétiques de deux séries individuelles

Preuve

Soit $\bar{x}$ la moyenne arithmétique d'une série dont les termes sont x_i et les effectifs n_i. $\bar{y}$ la moyenne arithmétique d'une série dont les termes sont y_i et les effectifs m_j.

$\bar{x} = \frac{\sum n_i x_i}{\sum n_i}$ et $\bar{y} = \frac{\sum m_j y_j}{\sum m_j}$

La moyenne arithmétique $\bar{z}$ de la série unique composée de deux séries précédentes est donnée par :

$$\bar{z} = \frac{\sum n_i x_i + \sum m_j y_j}{\sum n_i + \sum m_j}$$

Prenons $\sum n_i = n$ et $\sum m_j = m$

$$\bar{z} = \frac{\frac{n\sum n_i x_i}{n} + \frac{m\sum m_j y_j}{m}}{n+m} = \frac{n\bar{x} + m\bar{y}}{n+m}$$

Exemple

x: 11, 9, 15, 25, 30

y: 12, 13, 18, 27, 35, 39

z: 11, 9, 15, 25, 30, 12, 13, 18, 27, 35, 39

$\bar{z} = \frac{5.18+6.24}{5+6} = \frac{90+144}{11} = \frac{234}{11} \rightarrow \bar{z} = 21{,}2727$

c) La somme algébrique des écarts des termes de la série par rapport à la moyenne arithmétique est nulle.

Preuve

Soit $\{x_1, x_2, \cdots, x_n\}$ les valeurs d'une série dont la moyenne est notée $\bar{x}$, la somme des écarts par rapport à $\bar{x}$.

$s = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})$

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

Or $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \implies n\bar{x} = \sum_{i=1}^{n} x_i$

D'où $s = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$

d) La somme des carrés des écarts de chaque terme de la série par rapport à la moyenne arithmétique est minimum, elle est inférieure à la somme des carrés des écarts par rapport à toute autre valeur.

Hypothèse : Soit la variable x pouvant prendre les valeurs x_i.

$\bar{x}$ la moyenne des x_i

x_0 une valeur quelconque de la variable x différente des $\bar{x}$.

Thèse :

$\sum(x_i - \bar{x})^2 < \sum(x_i - x_0)^2$

Démonstration

On sait que $(x_i - x_0)$ peut s'écrire $(x_i - \bar{x}) + (\bar{x} - x_0)$

$$\sum(x_i - x_0)^2 = \sum[(x_i - \bar{x}) + (\bar{x} - x_0)]^2$$

$$= \sum(x_i - \bar{x})^2 + 2\sum(x_i - \bar{x})(\bar{x} - x_0) + \sum(\bar{x} - x_0)^2$$

Or $\sum(x_i - \bar{x}) = 0$ en vertu de la 3e propriété

$= \sum(x_i - \bar{x})^2 + \sum(\bar{x} - x_0)^2 = \sum(x_i - \bar{x})^2 + n(\bar{x} - x_0)^2$

$\sum(x_i - \bar{x})^2 = \sum(x_i - x_0)^2 - n(\bar{x} - x_0)^2$

La quantité $(\bar{x} - x_0)^2$ étant toujours positif à cause du carré on a :

$\sum(x_i - \bar{x})^2 < \sum(x_i - x_0)^2$

5. Avantage

La moyenne arithmétique est un paramètre de position le plus utilisé et le plus représentatif de l'ensemble des observations.

6. Inconvénient

a) La moyenne arithmétique n'est pas une donnée mais plutôt une statistique déduite des données conservées et qui peut comporter parfois des valeurs décimales insensées.
Exemple : on dispose de 5 familles ayant respectivement 5, 3, 2, 3 et 6 enfants. Calculer le nombre moyen d'enfants.
$\bar{x} = \frac{5+3+2+3+6}{5} = \frac{19}{5} = 3{,}8 \approx 4$ enfants.

b) Lorsqu'une distribution est plurimodale ou bimodale (distribution à plusieurs modes), la valeur de la moyenne arithmétique perd sa signification et il convient de l'abandonner.

c) Lorsque la série comporte les valeurs extrêmes anormales, il est préférable de prendre la médiane qui est une valeur moins influençable.

Autres moyennes

A. La moyenne pondérée

$\bar{x} = \frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w_1+w_2+\cdots+w_n}$ ou $\bar{x} = \frac{\sum_{i=1}^{n} w_i\, x_i}{\sum_{i=1}^{n} w_i}$

On appelle poids ou coefficient de pondération d'une donnée un nombre positif ou nul attaché à cette donnée de façon à indiquer son importance relative dans l'ensemble.

Exemple

Soit un examen comportant 5 épreuves dont les pondérations respectives sont : 15 pour la statistique, 15 pour la probabilité, 8 pour l'informatique, 8 pour la mathématique et 4 pour l'anglais.

Trouver la moyenne d'un étudiant dont les côtes sont respectivement : 10, 11, 8, 10 et 13

$\bar{x} = \frac{15.10+15.11+8.8+8.10+4.13}{15+15+8+8+4} = \frac{511}{50} = 10{,}22$

$\bar{x}\% = \frac{10{,}22}{20}.100 = 51{,}1\%$

B. Moyenne géométrique

La moyenne géométrique de n termes observés $x_1, x_2, \cdots, x_n$ est la racine nième du produit de ces termes ou antilogarithmique de la moyenne arithmétique des logarithmes des valeurs observées.

B1. Cas d'une distribution simple

1. La forme générale est donnée par :

$G = \sqrt[n]{x_1.x_2 \ldots x_n} = (\prod_{i=1}^{n} x_i)^{1/n}$

$\log(G) = (\sum \log x_i)^{1/n}$

$G = 10^{(\sum \log x_i)/n}$

Soit à calculer la moyenne géométrique simple de 5 termes : 2, 5, 6, 10 et 13

$$G = \sqrt[5]{2.5.6.10.13} = 6{,}003699$$

2. moyenne géométrique pondérée

$G_p = \sqrt[w]{x_1^{w_1}.x_2^{w_2}.x_3^{w_3} \ldots x_k^{w_k}} = \sqrt{\prod_{i=1}^{k} x_i^{w_i}}$

$avec\ w = w_1 + w_2 + \cdots + w_k$

Exemple : Soit la série 2, 5, 6, 10 et 13 dont les pondérations sont respectivement 30, 25, 20, 15 et 10 ; trouver la moyenne géométrique pondérée correspondante.

$G_p = \sqrt[100]{2^{30}.5^{25}.6^{20}.10^{15}.13^{10}} = 4{,}809235952$

B2. Cas d'une distribution groupée en classe

La moyenne géométrique d'une série de k termes $x_1, x_2, \cdots, x_k$ de fréquence absolue $n_1, n_2, \cdots, n_k$ est donnée par :

$G = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}}$ avec $N = n_1 + n_2 + \cdots + n_k$

$x_i =$ centre de classe

C. Moyenne quadratique

La moyenne quadratique est la racine carrée de la moyenne arithmétique des carrés des termes

C1. Cas d'une distribution non groupée

1. La forme générale est donnée par :

$$Q = \bar{x}_q = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{N}} = \left(\sum_{i=1}^{n} \frac{x_i^2}{N}\right)^{1/2}$$

Exemple : Calculer la moyenne quadratique de la série suivante 2, 5, 6, 10 et 13.

$$Q = \sqrt{\frac{2^2 + 5^2 + 6^2 + 10^2 + 13^2}{5}} = \sqrt{66{,}8} = 8{,}1731$$

2. Moyenne quadratique pondérée

$$Q_p = \bar{x}_{Q_p} = \sqrt{\frac{w_1 x_1^2 + w_2 x_2^2 + \cdots + w_n x_n^2}{w_1 + w_2 + \cdots + w_n}} = \sqrt{\frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i}}$$

C2. Cas d'une série groupée en classe

$$Q_p = \bar{x}_Q = \sqrt{\frac{n_1 x_1^2 + n_2 x_2^2 + \cdots + n_k x_k^2}{n_1 + n_2 + \cdots + n_k}} = \sqrt{\frac{\sum_{i=1}^{k} n_i x_i^2}{\sum_{i=1}^{k} n_i}}$$

D. Moyenne harmonique

La moyenne harmonique est l'inverse de la moyenne arithmétique des inverses des termes

D1. Cas d'une distribution non groupée.

1. La forme générale est donnée par :

$$\bar{x}_H = H = \frac{N}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Exemple : Calculer la moyenne harmonique de la série 2, 5, 6, 10 et 13.

$$H = \frac{5}{\frac{1}{2} + \frac{1}{5} + \frac{1}{6} + \frac{1}{10} + \frac{1}{13}} = 4{,}79115$$

2. Moyenne harmonique pondérée

$$H = \bar{x}_{H_p} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

D2. Cas d'une série groupée en classe

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

Comparaison des différentes moyennes

$$H < G < \bar{X} < Q$$

III.3.1.2 Le Mode (Mo)

2.1. Définition.

C'est la valeur observée la plus fréquente (c'est la valeur de la variable à laquelle correspond l'effectif le plus grand). On parle de « classe modale » pour une série groupée qui correspond à la plus grande fréquence. Le mode est encore peu maniable mathématiquement que la médiane. De fois une série peut manquer de mode ou posséder plus d'un mode. (exemple distribution bimodale ou plurimodale).

Exemple. Soit le résultat de 15 élèves à un examen :

16, 16, 15, 14, 12, 12, 10, 9, 8, 8, 8, 7, 6, 5 et 4

Le mode est 8

Cas d'une distribution groupée en classe

Procédure

b.1. Déterminer la classe, celle qui admet la plus grande fréquence absolue.

b.2. Utiliser la formule d'interpolation suivante :

$$M_0 = L_i + h\frac{\Delta_1}{\Delta_1 + \Delta_2}$$

Où : l_i est la limite inférieure de la classe modale

h est l'amplitude

Δ_1= fréquence classe modale- fréquence classe précédente

Δ_2= fréquence classe modale – fréquence classe suivante.

Tableau 2.2. Exemple de calcul du mode.

Age (ans)	Fréq. Abs. n_i
20-25	9
25-30	28
30-35	36
35-40	45
40-45	18
45-50	9
50-55	3
55-60	3

n= 150

$l_i = 35$
$h = 40-35=5$
$\Delta_1 = 45-36= 9$
$\Delta_2 = 45-18= 27$

Mo = 35 + 5 (9/ (9+27))
= 35 + 45/36
= 36,25 ans = 36 ans

III.3.1.3 Médiane (Me ou $\tilde{x}$)

1. Définition

C'est la valeur observée se situant au milieu d'une série statistique rangée par ordre croissant. Elle est telle que la moitié des observations lui sont inférieures et la moitié lui sont supérieures. Notons qu'elle est peu affectée par les observations extrêmes. C'est ainsi que qu'elle est préférée dans le cas des distributions asymétriques.

2. Détermination de la Médiane

a. Cas d'une distribution non groupée

On examine la parité de n (n= taille de l'échantillon).

Dans le cas d'un nombre impair (2n +1) des termes, la médiane est le terme de rang n+1

Exemples

Ex_1. 10 ;12 ;13 ;14 ;18

n= 5 (impair)

= (5+1)/2 = 3^e

Me = 13

Lorsque le nombre de termes est pair (2n), la médiane est la moyenne arithmétique de deux termes du milieu

Ex_2. 10 ;12 ;13 ;14 ;18 ;20

n= 6 (pair)

n/2= 3^e

(n/2) + 1= 4^e

Me = (3^e +4^e) /2 = (13+14)/2= 13,5

b. Cas d'une distribution groupée en classe (continue)
Il faut d'abord déterminer la classe médiane. Celle-ci est la première classe dans l'ordre du tableau (colonne de la fréquence cumulée ascendante) dont les effectifs cumulés sont égales ou dépassent la moitié de l'effectif total.

Procédure

b.1. Calculer les fréquences absolues cumulées (Ni)

b.2. Déterminer la classe médiane (= classe contenant l'observation de rang n/2)
b.3. Utiliser la formule d'interpolation suivante :

$$\tilde{x} = L_i + a\frac{\left(\frac{n}{2} - F_{cai}\right)}{n_i}$$

Où
L_i = est la limite inférieure de la classe médiane
a = est l'amplitude de la classe médiane
n_i est la fréquence absolue de la classe médiane ou l'effectif de la classe médiane
$\frac{n}{2}$ est la moitié des effectifs (la moitié de la taille de l'échantillon)
Fcai = est la fréquence absolue cumulée de la classe précédant la classe médiane

Exemple.

Age (ans)	Fréq.abs. n_i	Fréq. abs.cum N_i	Marquage
20-25	9	9	1^e ;2^e ;…9e
25-30	27	36	10^e….36^e
30-35	36	72	37^e …..72^e
35-40	45	117	73^e …117e
40-45	18	135	118^e …135e
45-50	9	144	136^e …144e
50-55	3	147	145^e …147e
55-60	3	150	148^e …150e

n= 150

n/2 =150/2=75^{e}
Li = 35
a = 5
n = 150
Fcai = 72
$\tilde{x} = 35 + 5\frac{(75-72)}{45} = 35{,}3\ ans$
Me= 35 ans

III.3.2 Les quantiles

Les quantiles sont des valeurs de la variable aléatoire qui partagent la distribution en n parties ayant le même effectif. Il faut noter que pour un partage en n parties égales, il y a n – 1 quantiles. En pratique on utilise les quantiles suivants :

- La médiane : est la valeur de la variable qui divise la série en deux parties égales.
- Les quartiles : sont les trois valeurs de la variable qui divisent la série en quatre parties égales.
- Les déciles : sont les neufs valeurs de la variable qui divisent la série en dix parties égales.
- Les centiles : sont les 99 valeurs de la variable qui divisent la série en 100 parties égales.

n	Désignation	Symbole
2	Médiane	$\tilde{x}$
4	Quartiles	Q1, Q2, Q3
10	Déciles	D1, D2, …, D9
100	centiles	C1, C2,……..,C99

III.4. Paramètres de dispersion

Rôle : déterminer les écarts des différentes valeurs de la série statistique vis-à-vis de la moyenne.

Les paramètres de position ou de localisation ne sont pas suffisants pour caractériser une série des données ; en effet, la moyenne donne une certaine idée de ce qui est la population mais elle n'en donne pas la physionomie complète. On peut par exemple disposer de deux séries des données ayant la même moyenne mais leurs valeurs se dispersent différemment autour de cette moyenne ; dispersion plus ou moins grande pour l'une que pour l'autre. Voici

deux populations très différentes, elles ont pourtant la même moyenne de 10.

A : 8 8 9 9 10 10 11 11 12 12

B : 1 7 8 8 9 10 11 12 13 18

Dans le groupe A, les données sont assez autour de la moyenne tandis que dans le groupe B, les données sont très dispersées.

Il existe différentes façons d'exprimer la dispersion :

1. Le Domaine de variation (ou étendue)

Il est la différence entre le plus grand nombre d'une série des valeurs et le plus petit.

$$W = X_{max} - X_{min}$$

Avantage : calcul simple

Inconvénient : ne tient compte que de 2 valeurs extrêmes.

2. L'espace inter-décile : est la différence entre le premier et le neuvième décile

$$D = D_9 - D_1$$

3. Espace interquartile : est la différence entre le premier quartile et le troisième quartile

4. *Ecart moyen* (ou écart à la moyenne) est défini par :

$$EM = \frac{\sum|x_i - \bar{x}|}{N}$$

Dans le cas d'une série groupée en classe l'écart moyen est donné par :

$$EM = \sum n_i|x_i - \bar{x}|$$

5. L'écart type et la variance (σ^2)

Ecart type ou déviation standard

Nous allons chercher une mesure de variation dans laquelle toutes les observations sont employées. On considère l'écart de chaque observation par rapport à la moyenne.

- La variance est la somme des carrés des écarts par rapport à la moyenne arithmétique divisée par le degré de liberté (n -1) . elle utilise les unités de mesure élevées au carré.

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

En d'autres mots, nous dirons que la variance est la moyenne des carrés des écarts des observations par rapport à la moyenne arithmétique.

- L'écart type (ou déviation standard)
  Elle est la racine carrée de la variance

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

Pour une série groupée en classe

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

L'écart type est le paramètre de dispersion le plus employé ; on distingue l'écart type d'une population de l'écart type d'un échantillon tiré dans la population. Une faible valeur de l'écart type indique une forte accumulation des observations autour de la moyenne arithmétique et une grande valeur de l'écart type indique un étalement autour de la moyenne arithmétique.

6. Cœfficient de variation (C.V.)

C'est la déviation standard ou écart type exprimée en pourcentage de la moyenne arithmétique.

$$C.V\% = \frac{\sigma}{\bar{x}}$$

N.B. Si C.V. < 17 %, on parle d'une dispersion faible
Si C.V. > 17 %, on parle d'une dispersion forte

Exemple de calcul de la variance, l'écart-type et le coefficient de variation.

Age (ans)	Fréq. ni	Centre xi	ni.xi			
20-25	9	22,5	205,5	-13	169	1521
25-30	27	27,5	742,5	-8	64	1728
30-35	36	32,5	1170	-3	9	324
35-40	45	37,5	7687,5	2	4	180
40-45	18	42,5	765	7	49	882
45-50	9	47,5	427,5	12	144	1296
50-55	3	52,5	157,5	17	289	867
55-60	3	57,5	172,5	22	484	1452
	n= 150		5325			8250

a) Variance = 8250/149= 55,37 (ans) ²

b) Ecart-type = 7,4 7ans.

c) Coefficient de variation (C.V)

C.V x 100= 7,4/35,5 x 100 = 20,8 %

Nous sommes en présence d'une dispersion forte.

III.5. Les paramètres de forme

Ils permettent d'étudier l'asymétrie et l'aplatissement.

Pour une distribution symétrique unimodale, les trois mesures de localisation (moyenne arithmétique, médiane et mode) sont confondues.

Pour présenter une variable quantitative, il faut décrire au moins le nombre d'observations sur lequel porte l'analyse, une mesure de localisation et une mesure de dispersion :

Exemple

	n	Moyenne	Ecart type
Périmètre brachial	1055	132,8	21,7

1) *Coefficients d'asymétrie.*

Une courbe de fréquence présente deux aspects : l'asymétrie et le degré de convexité. Au premier point de vue une courbe est soit symétrique, soit positivement dissymétrique, soit négativement dissymétrique.

$m_3 = \frac{1}{n}\sum(x_i - \bar{x})^3$: moment centré (sur la moyenne $\bar{x}$) d'ordre 3.

$m_2 = \sigma^2$: la variance

Coefficient : $a_3 = \frac{m_3}{\sigma^3} = \frac{m_3}{(\sqrt{m_2})^3}$

Si $a_3 < 0$ la courbe est étalée à gauche

Si $a_3 = 0$ la courbe est symétrique

Si $a_3 > 0$ la courbe est étalée à droite

Une distribution est dite symétrique si les observations sont également dispersées de chaque côté de l'axe central.

Dans le cas contraire, la distribution est dite dissymétrique ou asymétrique. De plus, dans une distribution symétrique les quartiles Q1 et Q3, puis les déciles D1 et D9 sont équidistants de la valeur centrale.

2. Coefficients d'aplatissement

Du point de vue de la convexité, une courbe peut être mésokurtique, platykurtique (aplatie ou

surbaissée) ou leptokurtique (pointue).

$m_4 = \frac{1}{n}\sum(x_i - \bar{x})^4$ moment centré (sur la moyenne $\bar{x}$) d'ordre 4 ;

$m_2 = \sigma^2$: la variance

Coefficient $a_4 = \frac{m_4}{\sigma^4} = \frac{m_4}{(\sqrt{m_2})^4}$

Si $a_4 > 3$ la distribution est leptokurtique

Si $a_4 = 3$ la distribution est normale (mésokurtique)

Si $a_4 < 3$ la distribution est platikurtique

CHAPITRE IV. DESCRIPTION DES VARIABLES QUALITATIVES OU EN CATEGORIES

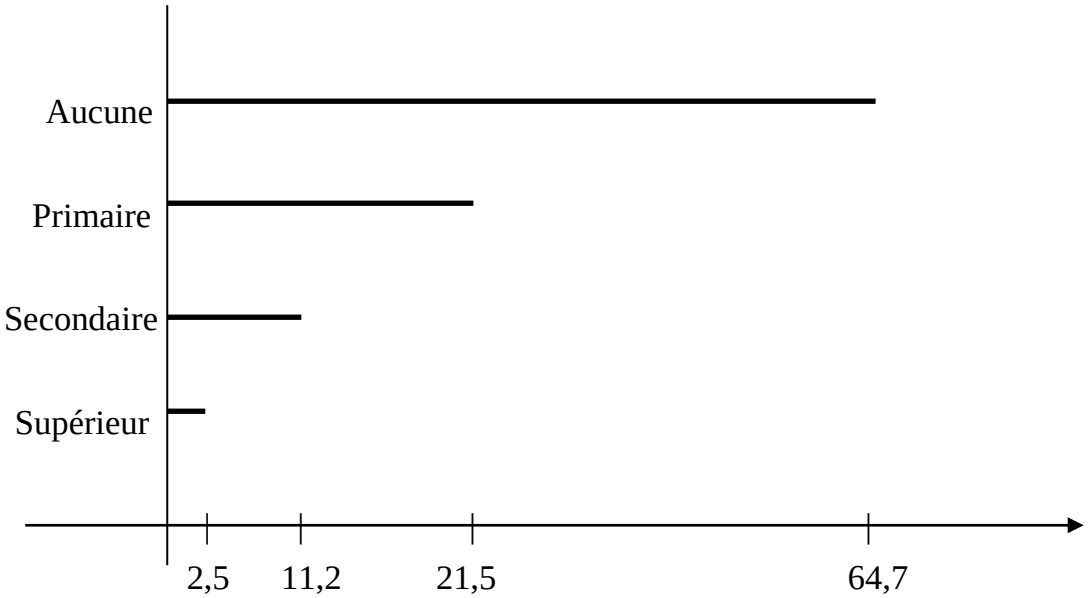
IV.1 Distribution des fréquences

Exemple : La distribution de niveau d'études des mères d'enfants recensés

Niveau d'études de la mère	**Nombre de mère = n**	**% des mères**
Aucun	559	64,7
Primaire	186	21,5
Secondaire	97	11,2
Supérieur	22	2,5
Total	**864**	**100**

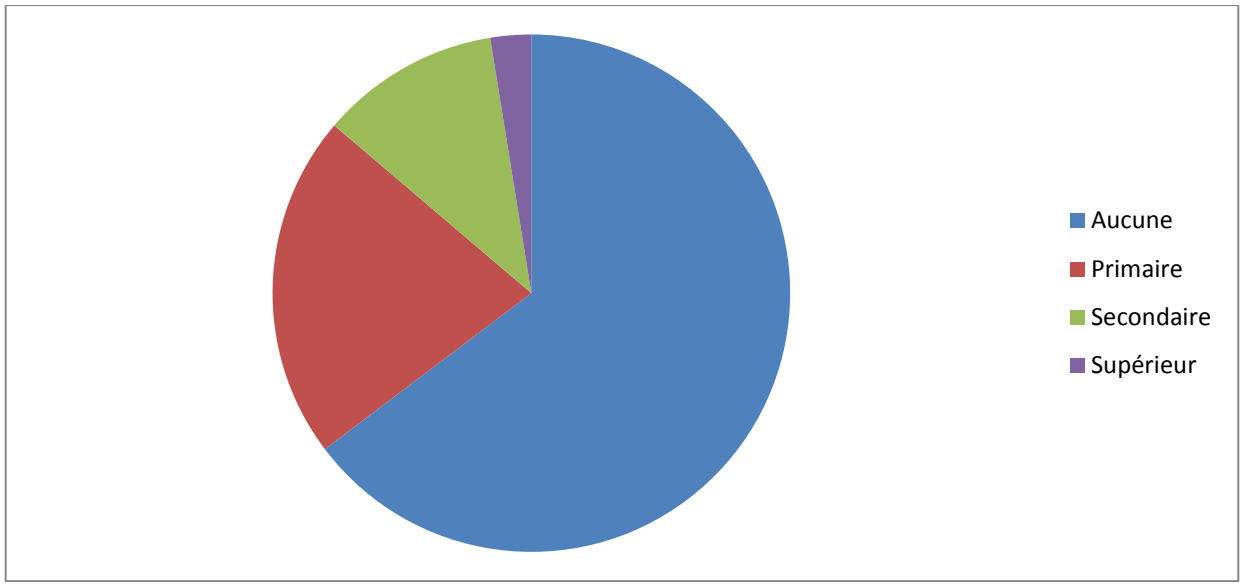
IV.2 Représentations graphiques

IV.2.1 Diagramme à colonnes ou barres



IV.2.2 Cercle, « tarte », « Camembert »

$$Angle\ au\ centre\ en\ degrés = \frac{valeur\ en\ \%\ \times 360}{100}$$



IV.3 La courbe de probabilité

IV.3.1 La courbe normale de distribution

A la suite de nombreuses mesures, des traits physiques ou traits mentaux, des rendements scolaires, phénomènes économiques, biologiques, physiques,… plusieurs chercheurs sont parvenus à des répartitions numériques ou graphiques qui avaient une grande ressemblance entre elles. Ces répartitions tendaient à se rapprocher d'une courbe dont la forme suggérée le profil d'une cloche.

Divers noms s'attachent à cette courbe : courbe de Gauss, courbe de Bernouilli, courbe de Laplace, courbe de Gauss-Laplace, courbe de Moivre, courbe de Quetelet, courbe des erreurs, courbe normale des erreurs, courbe normale de probabilité, courbe des fréquences idéales, courbes de distribution etc…

L'appellation la plus répandue est « courbe normale ».

N.B : Dans une distribution normale, nous avons les relations suivantes :

a) $E.M = \frac{4}{5}\,\sigma$ avec $E.M = \frac{1}{n}\sum|x_i - \bar{x}|$ ou $E.M = \frac{1}{n}\sum n_i|x_i - \bar{x}|$

b) L'intervalle interquartile

$EIQ = \frac{Q_3 - Q_2}{2} = \frac{2}{3}\sigma$

c) 68,27% des observations sont compris dans l'intervalle $[\bar{x} \pm \sigma]$

95% des observations sont compris dans l'intervalle $[\bar{x} \pm 2\sigma]$

99,73% des observations sont compris dans l'intervalle $[\bar{x} \pm 3\sigma]$

Ainsi la courbe de Gauss est définie complètement par deux paramètres : la moyenne et l'écart type. La plupart des mesures biologiques, sociologiques, psychologiques, scolaires, académiques, etc… sont des variables qui se distribuent normalement c'est-à-dire selon la loi de Gauss.

IV.3.2 La variable aléatoire normale réduite

Appelée encore variable centrée, elle mesure l'écart de la moyenne arithmétique en unité d'écart type :

$$\mu = \frac{x_i - \bar{x}}{\sigma}$$

La variable normale réduite est caractérisée par une moyenne nulle et une variance égale à l'unité ($\bar{x} = 0\ ;\ \sigma^2 = 1$). Cette variable a permis à Gauss d'établir les tables de la loi normale par l'équation $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}}$

Exemple 1

Un étudiant a obtenu la côte de 84 à un examen de comptabilité pour lequel la moyenne était de 76 et l'écart type de 10. A l'examen de statistique, dont la moyenne était de 82, l'écart type de 16, il a obtenu 90. Dans quel cours cet étudiant est-il relativement plus doué ?

$$\mu_1 = \frac{84-76}{10} = 0{,}8 \quad \mu_2 = \frac{90-82}{16} = 0{,}5$$

Nous dirons que l'étudiant est plus doué en comptabilité où la moyenne est de 0,8 fois l'écart type au dessus de la moyenne.

Exemple 2

Si une distribution est normale avec une moyenne $\bar{x} = 40$ et un écart type $\sigma = 8$. On demande :

a) L'étendue approximative de la distribution
b) Quelle sera la plus forte côte de la distribution ?
c) Entre quelle côte peut –on s'attendre à trouver le 68% de la distribution.

Solution

a) $99{,}73 \implies \bar{x} \pm 3\sigma = 40 \pm 3 \times 8 = [16\,;64]$
   Etendue : 64 – 16 = 48

b) 64
c) $68{,}27 \in [\bar{x} \pm \sigma] = [40 \pm 8]$ d'où 68 % $\in$ $[32\,;48]$

IV.3.3 Estimation du défaut de normalité

Une courbe de fréquence présente deux aspects : l'asymétrie et le degré de convexité. A partir du coefficient $a_3 = \frac{m_3}{\sigma^3}$ nous pouvons connaitre si la courbe est soit symétrique au dissymétrique positivement ou négativement

Du point de vue convexité, le coefficient $a_4 = \frac{m_4}{\sigma^4}$, nous permet de savoir si la distribution est leptokurtique, normale ou platikurtique.

IV.3.4 Usager de la courbe normale

La courbe de Gauss est appliquée à toutes les activités humaines où les données sont nombreuses et variables à cause de la concordance étroite qui existe entre cette courbe théorique et les distributions de fréquences obtenues. Ainsi la courbe théorique se vérifie en biologie pour diverses formules de la génétique, en anthropométrie pour la répartition des tailles et des poids ; en science sociale, en psychologie pour la répartition des intelligences, en pédagogie pour la répartition des scores (note, côte, résultat des tests d'instruction, en sciences physiques pour la mesure des erreurs d'observation etc…

EXERCICES

1) Une partie d'une distribution de fréquence relative est donnée ci-dessous

Classe	**Fréquence relative**
A	0.22
B	0.18
C	0.40
D	

a) Quelle est la fréquence relative de la classe D ?
b) La taille de l'échantillon est égale à 200. Quelle est la fréquence de la classe D ?
c) Donnez la distribution de fréquence.
d) Donnez la distribution de fréquence en pourcentage.

2) On a demandé aux étudiants de première année, entrant à l'école de commerce de l'université de Lubumbashi, d'indiquer leur matière préférée. Les réponses suivantes ont été obtenues.

Matière	**Nombre**
Management	55
Comptabilité	51
Finance	28
Marketing	82

Résumez les données en construisant :
a) Les distributions de fréquence relative et en pourcentage
b) Un diagramme en barres
c) Un diagramme circulaire

3) Les positions d'un échantillon de 55 membres de club de base-ball Hall of Famers de Cooperstown, dans l'Etat de New Yorkdans, sont représentées ci-dessous. Chaque observation indique la position principale occupée par les Hall of Famers : lanceur (L), receveur (R), 1ère base (1), 2e base (2), 3e base (3), bloqueur (B), champ gauche (G), champ droit (D) et milieu de terrain (M)

G R M L 2 R 1 B B 1 G R
R R R D M G D R M M R R
2 3 R L G 1 M R R R B 1

D 1 2 L B L 2 G R D D G
R R D

a) Utilisez les distributions de fréquence absolue et relative pour résumer les données
b) Quelle est la position la plus occupée par les Hall of Framers ?
c) Quelle est la position la moins occupée par les Hall of framers ?

4) Considérerons les données suivantes :
14 21 23 21 16 14 21 23 21 16 19 22 25 16 16 24 24 25 19
16 19 18 19 21 12 16 17 18 23 25 20 23 16 20 19 24 26 15
22 24 20 22 24 22 20

a) Développer une distribution de fréquence en utilisant les classes 12 – 14 ; 15 – 17 ; 18 – 20 ; 21 - 23 ; et 24 – 26
b) Développer une distribution de fréquence relative et une distribution de fréquence en pourcentage en utilisant les mêmes classes.

5) Considérer la distribution de fréquence suivante.

Classe	**Fréquence**
10 – 19	10
20 – 29	14
30 – 39	17
40 – 49	7
50 – 59	2

a) Construisez une distribution de fréquence cumulée absolue et une distribution de fréquence cumulée relative
b) Construisez un histogramme et une ogive

6) La compagnie Hewa Bora accepte les réservations de vol par téléphone. Les données suivantes correspondent à la durée d'un appel (en minutes) pour un échantillon de 20 réservations par téléphone. Construisez les distributions de fréquence absolue et relative pour ces données. Fournissez également un histogramme.
2,1 4,8 5,5 10,4
3,3 3,5 4,8 5,8
5,3 5,5 2,8 3,6
5,9 6,6 7,8 10,5
7,5 6,0 4,5 4,8

7) Selon le service du jeune personnel de chez Roth, les salaires annuels des assistants-managers d'un grand magasin varient entre 28000 et 57000 francs congolais. Supposez que les données suivantes correspondent à un échantillon des salaires annuels de 40 assistants-managers de grands magasins (les données sont exprimées en milliers de francs)

48 35 57 48 52 56 51 44
40 40 50 31 52 37 51 41
47 45 46 42 53 43 44 39
50 50 44 49 45 45 50 42
52 55 46 54 45 41 45 47

a) Quel est le salaire le plus élevé de l'échantillon ? Le plus faible ?
b) Utilisez la largeur de classe de 5000 francs et résumez les données sous forme de tableau
c) Quelle est la proportion de salaire annuel inférieur ou égal à 35000 Fc ?
d) Quel est le pourcentage de salaire annuel supérieur à 50000 Fc ?
e) Construisez un histogramme

8) Les données suivantes correspondent au nombre d'unités produites par un employé au cours des 20 derniers jours.

160 170 181 156 176
148 198 179 162 150
162 156 179 178 151
157 154 179 148 156

Résumez les données en construisant :

a) Une distribution de fréquence
b) Une distribution de fréquence relative
c) Une distribution de fréquence cumulée
d) Une distribution de fréquence cumulée relative
e) Une ogive

9) Le rapport Nielsen sur la technologie à la maison (20 février 2016) concernait la technologie domestique et son usage par des personnes âgées de 12 ans et plus. Les données suivantes correspondent aux heures d'utilisation d'un ordinateur au cours d'une semaine pour un échantillon de 50 personnes.

4,1 1,5 10,4 5,9 3,4 5,7 1,6 6,1 3,0 3,7
3,1 4,8 2,0 14,8 5,4 4,2 3,9 4,1 11,1 3,5
4,1 4,1 8,8 5,6 4,3 3,3 7,1 10,3 6,2 7,6
10,8 2,8 9,5 12,9 12,1 0,7 4,0 9,2 4,4 5,7
7,2 6,1 5,7 5,9 4,7 3,9 3,7 3,1 6,1 3,1

Résumez les données en construisant :

a) Une distribution de fréquence (en utilisant une largeur de classe de 3 heures)
b) Une distribution de fréquence relative
c) Un histogramme
d) Une ogive
e) Commentez les résultats quant à l'usage d'un ordinateur à la maison

10) Dans une étude concernant la satisfaction sur le plan professionnel, une série de tests a été effectuée par 50 personnes. Les données suivantes ont été obtenues ; des scores élevés correspondent à une insatisfaction importante

87 76 67 58 92 59 41 50 90 75 80 81 70
73 69 61 88 46 85 97 50 47 81 87 75 60
65 92 77 71 70 74 53 43 61 89 84 83 70
46 84 76 78 64 69 76 78 67 74 64

Construisez un diagramme « stem-and-leaf » pour ces données

11) Considérez un échantillon avec les observations suivantes : 10, 20, 12, 17 et 16.

a) Calculer la moyenne et la médiane
b) Calculer l'étendue et l'étendue interquartile

12) Considérez un échantillon avec les observations suivantes : 10, 20, 21, 17, 16 et 12.

a) Calculer la moyenne et la médiane
b) Calculez la variance et l'écart-type

13) Considérez un échantillon avec les observations suivantes : 53, 55, 70, 58, 64, 57, 53, 69, 57, 68 et 53. Calculer la moyenne, la médiane et le mode.

14) Le salaire moyen de débutant pour des jeunes diplômés en comptabilité en 1996-1997 était de 30 393 dollars. Un échantillon des salaires de départ est donné ci-dessous. Les données sont en milliers de dollars.

30,7 28,8 29,1 31,1 30,1
29,7 30,7 30,0 30,6 30,5
31,2 32,1 30,2 30,3 32,9
32,2 29,9 28,9 30,6 31,8
32,2 30,3 30,4 32,3 33,3
32,7 29,3 30,3 30,9 30,3

a) Quel est le salaire de départ moyen ?
b) Quelle est la médiane ?
c) Quel est le mode ?
d) Quels sont les premier et troisième quartiles ?
e) Est-ce que ces données sont compatibles avec le salaire moyen cité, égal à 30 393 dollars ?

15) L'individu moyen écoute de la musique 45 minutes par jour. Les données concernant le nombre de minutes passées à écouter de la musique, ont été collectées auprès de 30 individus.

88,3 4,3 4,6 7,0 9,2
0,0 99,2 34,9 81,7 0,0
85,4 0,0 17,5 45,0 53,3
29,1 28,8 0,0 98,9 64,5
4,4 67,9 94,2 7,6 56,6
52,9 145,6 70,4 65,1 63,6

a) Calculez la moyenne
b) Est-ce que ces données sont compatibles avec la moyenne présentée ?
c) Calculez la médiane
d) Calculez le premier et le troisième quartile
e) Calculez et interprétez le 40^{e} percentile.

16) Pour tester la consommation d'essence, 13 automobiles ont parcouru 300 milles en ville et à la campagne. Les données sur la consommation, en milles par gallon, sont présentées ci-dessous.

Ville : 16,2 16,7 15,9 14,4 13,2 15,3 16,8 16,0 16,1 15,3 15,2 15,3 16,2

Campagne : 19,4 20,6 18,3 18,6 19,2 17,4 17,2 18,6 19,0 21,1 19,4 18,5 18,7

Utilisez la moyenne, la médiane et le mode pour étudier les différences de performance entre la conduite en ville et à la campagne.

17) La Pennsylvanie est le cinquième producteur de sapins de Noël, avec une récolte de 1,5 millions de sapins en 2014. Le prix d'un sapin varie entre 3,50 et 5,50 dollars. Supposez que les données suivantes correspondent au prix de 16 sapins vendus dans la région de Philadelphie.

3,90 4,20 3,90 5,10

4,20 4,50 4,10 5,10

4,30 4,20 4,00 5,20

4,50 4,20 4,50 5,10

a) Calculer la moyenne, la médiane et le mode
b) Calculer le 20^{e} et le 90^{e} percentile
c) Calculez les quartiles.

18) Un département de production utilise une procédure d'échantillonnage pour tester la qualité des nouvelles pièces produites. Le département utilise la règle de décision suivante : si un échantillon de 14 pièces a une variance supérieure à 0,005, la ligne de production doit être fermée pour réparation. Supposez que les données suivantes aient été collectées :

3,48 3,45 3,43 3,48 3,52 3,50 3,39

3,48 3,41 3,38 3,49 3,45 3,51 3,50

La ligne de production doit-elle être fermée ? pourquoi ?

19) Les scores obtenus par un lanceur au cours de six jeux étaient 182, 168, 184, 190, 170 et 174. Utilisez ces données comme un échantillon pour calculer les statistiques descriptives suivantes :

a) L'étendue, la variance, l'écart-type et le coefficient de variation.

20) Considérer un échantillon avec les observations suivantes : 10, 20, 12, 17 et 16. Calculez les valeurs de la variable centrée réduite z pour chacune des cinq observations.

21) Considérez un échantillon de moyenne 500 et d'écart-type 100. Quelle est la valeur de la variable centrée réduite z pour les observations suivantes : 520, 650, 500, 450 et 280 ?

22) Des données, distribuées en forme de cloche, ont une moyenne de 30 et un écart-type de 5. Utilisez la règle empirique pour déterminer le pourcentage d'observations comprises entre :

a) 20 et 40
b) 15 et 45
c) 25 et 35

23) Considérez la distribution de fréquence ci-dessous, obtenue à partir d'un échantillon de données.

Classe	$\mathbf{n_i}$
3 – 7	4
8 – 12	7
13 – 17	9
18 – 22	5

a) Calculer la moyenne, la médiane et le mode
b) Calculer la variance et l'écart-type de l'échantillon

24) Une station service a construit la distribution de fréquence suivante concernant le nombre de gallons d'essence vendus par voiture, pour un échantillon de 680 voitures.

Essence (gallons)	**Fréquence**
0 – 4	74
5 – 9	192
10 – 14	280
15 – 19	105
20 – 24	23
25 – 29	6
Total	**680**

Calculer la moyenne, la variance et l'écart-type pour ces données groupées. Si la station service prévoit de servir 120 voitures en un jour, à combien estimez-vous le nombre total de gallons d'essence vendus ?

25) Lors d'une enquête sur les abonnés au magazine Fortune, la question suivante a été posée : « parmi les quatre derniers numéros, combien en avez-vous lu ou parcouru ? ». Les 500 réponses à cette question sont résumées par la distribution de fréquence suivante :

Nombre de numéros lus	**Fréquence**
0	15
1	10
2	40
3	85
4	350
Total	**500**

a) Quel est le nombre moyen de numéros lus par un abonné à Fortune ?

b) Quel est l'écart-type du nombre de numéros lus ?

26) L'américain moyen dépense 65,88 dollars par mois au restaurant. Les dépenses en restaurant d'un échantillon de jeunes adultes au cours du mois dernier sont notées ci-dessous.

253 101 245 467 131 0 225

80 113 69 198 95 129 124

11 178 104 161 0 118 151

55 152 134 169

a) Calculer la moyenne, la médiane et le mode

b) Aux vues des résultats de la question a), ces jeunes adultes semblent-ils dépenser le même montant que l'Américain moyen ?

c) Calculer le premier et le troisième quartile

d) Calculer l'étendue et l'écart interquartile

e) Calculer la variance, l'écart-type et le coefficient de variation

27) Ci-dessous est présenté le rendement d'un échantillon de 10 titres échangés sur le marché boursier de New York

Détenteur	**Rendement (%)**	**Détenteur**	**Rendement (%)**
Argosy	12,6	Caterpillar	6,3
Chase Manahattan	6,7	Dow	6,8
IBM	7,0	Lucent	6,7
Mobil	7,3	Pacific Bell	6,7
RJR Nabisco	8,1	Service Mdse	8,6

Calculer les statistiques descriptives suivantes :

a) Moyenne, médiane et mode
b) Premier et troisième quartile
c) Etendue et écart interquartile
d) Variance et écart-type
e) Coefficient de variation

28) Une étude a été menée concernant la capacité des constructeurs d'ordinateurs à résoudre des problèmes rapidement. Les résultats suivants ont été obtenus.

Entreprise	**Nombre de jours nécessaires pour résoudre un problème**	**Entreprise**	**Nombre de jours nécessaires pour résoudre un problème**
Compaq	13	Gateway	21
Packard Bell	27	Digital	27
Quantex	11	IBM	12
Dell	14	Hewlett-Packard	14
NEC	14	AT&T	20
AST	17	Toshiba	37
Acer	16	Micron	17

a) Quel est le nombre moyen de jours nécessaires pour résoudre un problème? Quelle est la médiane ?
b) Quelle est la variance ? quel est l'écart-type ?
c) Quel est le meilleur fabricant, en terme de délai ?
d) Quelle est la valeur de la variable centrée réduite pour Packard Bell ?
e) Quelle est la valeur de la variable centrée réduite pour IBM ?
f) Y-a-t-il des valeurs singulières ?

29) Le prix d'un repas au restaurant La Maison French a la distribution de fréquence suivante. Calculer la moyenne, la variance et l'écart-type

Prix du repas 5$)	**Fréquence**
25 – 34	2
35 – 44	6
45 – 54	4
55 – 64	4
65 – 74	2
75 – 84	2
Total	**20**

30) Un système de radar de la police d'Etat contrôle la vitesse des automobiles roulant sur la voie express de l'Etat de New York. La distribution de fréquence des vitesses est présentée ci-dessous.

Vitesse (milles par heure)	**Fréquence**
45 – 49	10
50 – 54	40
55 – 59	150
60 – 64	175
65 – 69	75
70 – 74	15
75 – 79	10
Total	475

a) Quelle est la vitesse moyenne des automobiles roulant sur la voie express de l'Etat de New York ?
b) Calculez la variance et l'écart-type

CHAPITRE V. AJUSTEMENT D'UNE DISTRIBUTION.

V.1 Liaison ou dépendance statistique

Très souvent dans la pratique, on constate que deux ou plusieurs variables sont en liaison fonctionnelle. Ces variables ont une certaine relation entre-elles sans que l'on puisse l'exprimer sous forme mathématique.

Exemple

- Le poids et la taille des hommes sont liés d'une certaine façon.
- La circonférence d'un cercle dépend de son rayon.
- La pression d'un gaz de masse donnée dépend de sa température et de son volume.

Il faudra alors trouver le type des courbes qui lie graphiquement ces différentes variables et exprimer cette liaison sous forme mathématique à l'aide d'une opération.

Exemple

- La relation entre le périmètre d'un cercle et son rayon est $c = 2\pi R$
- La relation entre périmètre d'un carré et son côté est : $P = 4C$
- La relation entre l'espace parcouru par un corps en chute libre et le temps est :

$$e = \frac{1}{2} g t^2$$

V.2 Nuage des points

Pour déterminer l'équation liant les variables

a) Une première étape consiste à réunir les couples (ou n-uples) des valeurs correspondants à ces variables. Soit les variables X, Y des valeurs respectives : $x_1, x_2, \cdots, x_n$ et $y_1, y_2, \cdots, y_n$.

b) L'étape suivante consiste à placer les points (x_1, y_1); (x_2, y_2); $\cdots$; (x_n, y_n) dans un système d'axe rectangulaire. L'ensemble des points ainsi obtenu est appelé « diagramme de dispersion » et se présente sous forme d'un nuage des points.

c) A partir du diagramme de dispersion, on peut représenter une courbe continue approchant les données. Une telle courbe est appelée « courbe d'ajustement ».

Par exemple la fig1 montre que les données semblent être parfaitement approchées par une ligne droite. On dit alors qu'il y a une relation linéaire entre les variables. Tandis que pour les figures 2 et 3 nous parlerons d'une relation non linéaire.

V.3 Méthodes d'ajustement

Pour référence. Voici plusieurs types communs de courbe d'ajustement et leurs équations.

Modèle linéaire

1) $y = a_0 + a_1 x$: la droite ou polynôme du premier degré.
2) $y = a_0 + a_1 x + a_2 x^2$: la parabole ou polynôme du second degré.
3) $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$: courbe cubique ou polynôme du troisième degré.
4) $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n$: courbe du nième degré.

Modèle non linéaire

5) $y = ab^x$ modèle géométrique qui peut se ramener à : $\log y = \log a + x \log b = a_0 + a_1 x$
6) $y = a\, e^{bx}$ modèle exponentiel
7) $y = a\, x^b$ modèle puissance qui peut se réduire à $\log y = \log a + b \log x = a_0 + a_1 x$
8) $y = a\, b^{c^x}$ modèle Gompertz $= \log y + c^x \log b$
9) $y = a_0 + a_1 \ln x$ modèle logarithmique
10) $y = \frac{1}{a_0 + a_1 x}$ ou $\frac{1}{y} = a\, b^x + c$ fonction logistique.

V.3.1 La méthode graphique

La droite est le type le plus simple des courbes d'approximation. Pour tracer la droite, deux points sont nécessaires (x_1, y_1) et (x_2, y_2). Son équation est de la forme $y = a_0 + a_1 x$.
Si le nuage des points évoque une droite, on lui cherche la position qui adapte le mieux la forme du nuage. On choisit deux points A et B de la droite assez éloignés l'un de l'autre et on repère les coordonnées (x_1, y_1) et (x_2, y_2). Les paramètres de cette droite sont donnés par :

- L'équation d'une droite passant par les deux points A et B

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

- Ou par un système de deux équations à deux inconnues qu'il faudra résoudre pour trouver les valeurs de a_0 et a_1

$$\begin{cases} y_1 = a_0 + a_1 x_1 \\ y_2 = a_0 + a_1 x_2 \end{cases}$$

a_1 est la pente de la droite et représente la variation de y sur la variation correspondante de x
a_0 est la valeur de y quand x = 0 ; c'est l'ordonnée à l'origine.

Exemple

Représenter graphiquement les points correspondants aux coordonnées ci-après et tracer la droite qui ajuste le mieux le nuage des points obtenus. Quelle est l'équation de la droite pour les deux points extrêmes ?

Soient x_i la variable poids en gramme et y_i la variable taille en cm.

X	12 11 10 7 5 4 2 1
y	25 23 21 15 11 9 5 3

X	8 7 6 5 5 4 3 2
Y	10 6 7 8 6 4 5 2

X	9 8 7 5 5 3 2 1
Y	8 3 5 6 1 8 7 2

X	5	10	15	20	25	30	35	40	45
Y	103	106	109	112	117	121	124	127	131

V.3.2 Méthode des moyennes

Soit $y = ax + b$ l'équation de la droite ajustant les nuages des points. L'expression $y - ax - b$ prend la valeur nulle lorsque les coordonnées (x, y) appartiennent à la droite $y = ax + b$.

Comme les points représentant les données expérimentales ne sont pas tous sur la même droite, cette condition est rarement vérifiée. Par conséquent $y - ax - b$ n'est pas toujours nulle. Dans cette éventualité, on divise les observations en deux sous-ensembles et on suppose que l'expression $y - ax - b$ s'annule par la moyenne $\bar{x}$ de chacune de ces deux sous-ensembles.

Exemple

Considérer les données de l'exemple précédent.

V.3.3 Méthode des moindres carrés

La méthode des moindres carrés consiste à identifier les valeurs de a_0 et a_1 qui rendent minime la somme des carrés de la différence des coordonnées des points réels et des points substitués sur la droite $y = a_0 + a_1 x$

Cette somme est :

$$z = (a_0 + a_1x_1 - y_1)^2 + (a_0 + a_1x_2 - y_2)^2 + \cdots + (a_0 + a_1x_n - y_n)^2$$
$$= \sum_{i=1}^{n} (a_0 + a_1x_i - y_i)^2$$

Pour que cette fonction soit minimale, il faut que les dérivées partielles par rapport à a_0 et a_1 soient nulles.

$\frac{\partial z}{\partial a_0} = a_1 \sum_{i=1}^{n} x_i + na_0 - \sum_{i=1}^{n} y_i = 0 \quad \rightarrow \quad a_0N + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$

$\frac{\partial z}{\partial a_1} = a_1 \sum_{i=1}^{n} x_i^2 + a_0 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0 \quad \rightarrow \quad a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$

On obtient deux équations appelées « équations normales de la droite moindre carrés dont la résolution donne :

$a_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}$

$a_1 = \frac{N \sum xy - \sum x \sum y}{N \sum x^2 - (\sum x)^2}$

Exemple

X	Y	XY	X^2	Y^2
12	25	300	144	625
11	23	253	121	529
10	21	210	100	441
7	15	105	49	225
5	11	55	25	121
4	9	36	16	81
2	5	10	4	25
1	3	3	1	9
52	112	972	460	2056

$a_1 = \frac{8.972 - 52.112}{8.460 - (52)^2} = 2$

$a_0 = \frac{460.112 - 52.972}{976} = 1$

$y = 2x + 1$

V.4. REGRESSION

On peut estimer la valeur de y à partir d'une droite des moindres carrés qui ajuste les données de l'échantillon. La courbe ainsi définie est la droite de régression (estimation) de y en x car l'on estime y à partir de x. Dans le cas où l'on désire estimer la valeur de x à partir d'une valeur donnée en y, on utilisera la courbe de régression de x en y. c'est-à-dire $x = b_0 + b_1 y$

Il faudra alors minimiser l'équation $z = \sum(b_0 + b_1 y - x)^2$ en annulant les dérivées premières par rapport à $b_0\ et\ b_1$, on obtient le système d'équation normale suivante :

$b_1 \sum y^2 + b_0 \sum y - \sum xy = 0 \qquad b_1 \sum y^2 + b_0 \sum y = \sum xy$

$b_1 \sum y + N b_0 - \sum x = 0 \qquad b_1 \sum y + N b_0 = \sum x$

Après résolution on a :

$b_0 = \frac{\sum x \sum y^2 - \sum y \sum xy}{N \sum y^2 - (\sum y)^2}$

$b_1 = \frac{N \sum xy - \sum x \sum y}{N \sum y^2 - (\sum y)^2}$

$\begin{cases} y = a_0 + a_1 x \\ x = b_0 + b_1 y \end{cases}$ sont appelées « les deux droites de régression.

Les deux droites de régression ne se coïncident pas, mais se coupent au point médian $M(\bar{X}, \bar{Y})$ appelé « barycentre » ou « centre de gravité du nuage ».

Exemple

Trouver la droite de régression de x en y des données des exemples précédents et montrer que les deux droites de régression se coupent au point médian $M(\bar{X}, \bar{Y})$

La représentation d'un ensemble des points est fournie par la droite dont la somme des carrés des distances des divers points est minimale. Cette droite est appelée droite de régression de y en x lorsque ses distances sont comptées parallèlement à y.

Lorsque la régression linéaire se fait en prenant les abscisses comme critère d'ajustement, on obtient la droite de régression de x en y. en d'autres termes, si au lieu d'être une variable indépendante, la variable x est une variable « expliquée », l'équation devient :

$$x = b_0 + b_1 y$$

Pour obtenir une droite des moindres carrés, on peut raccourcir les calculs en transformant les données par un changement des variables suivant :

$x = X - \bar{x}$

$y = Y - \bar{y}$

L'équation de la droite peut alors s'écrire :

$y = \left(\frac{\sum xy}{\sum x^2}\right) x$ ou $x = \left(\frac{\sum xy}{\sum y^2}\right) y$

Avec $a_1 = \frac{\sum xy}{\sum x^2}$ et $b_1 = \frac{\sum xy}{\sum y^2}$

Les équations de régression sont identiques si et seulement si tous les points du diagramme de dispersion sont sur une même droite. Dans ce cas, on dit que la corrélation linéaire entre x et y est alors maximale.

V.5. CORRELATION LINEAIRE

1. Existence d'une liaison statistique entre deux variables

On peut étudier les variations de deux caractères X et Y. Elles peuvent être indépendantes ou liées ; si la mesure de l'une est connue, on peut calculer l'autre pour la même unité statistique. On dit alors que les variables X et Y sont en liaison fonctionnelle.
Sans être liées rigoureusement, les deux variables peuvent être plus ou moins indépendantes. Leurs valeurs varient dans le même sens (croissant ou décroissant) ou en sens contraire.

On dit qu'elles sont en corrélation, positive ou négative, selon le cas :
- Le point de départ est le diagramme de dispersion où chaque sujet est représenté par un point dont les coordonnées x et y sont ses valeurs aux deux variables analysées.
Ex : Taille (y) et Poids (X)
- L'analyse du type des données envisagée peut avoir deux objectifs différents :
* Etudier l'association entre les deux variables (corrélation)
* Prédire une de deux variables par l'autre (régression).
Au premier de ces objectifs correspond l'analyse de corrélation dans laquelle les deux variables se placent sur le même pied. Il s'agit de déterminer dans quelle mesure elles sont liées.
Au second objectif, correspond l'analyse de régression dans laquelle les variables ont une position asymétrique. L'une est la variable « dépendante » ou « réponse » Y et l'autre est variable « indépendante » ou « prédicteur » X.

Coefficient de corrélation
Il se donne par la formule ci-après :

$r = \frac{covariance\ (x,y)}{\sigma_x \sigma_y}$

$covariance\ (x, y) = \left(\frac{1}{n}\sum xy\right) - \bar{x}.\bar{y}$

σ_x = écart type de x

σ_y = écart type de y.

Ou encore $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$

Les propriétés des coefficients de corrélation sont les suivantes :

a) r varie entre – 1 et 1 $(r \in [-1,\ 1])$

b) r est une mesure d'association normée :
0 indique l'absence d'association et 1 indique l'association parfaite

c) le signe de r indique le sens de l'association :
« + » indique une association positive, quand x augmente y augmente aussi et vice versa ;
« - » indique une association négative, quand x augmente y diminue et vice versa.

Géométriquement, une association parfaite se traduit par des points exactement en ligne droite ; la connaissance des valeurs d'une variable permet de prédire de façon univoque les valeurs correspondantes de l'autre variable.

Plus le coefficient de corrélation diminue et se rapproche de 0, plus les points se dispersent en ellipse autour de la droite ; cette ellipse devient un cercle en absence de corrélation

Relation entre coefficient de régression et coefficient de corrélation

a) La droite de régression de y en x est donnée par $y = a_0 + a_1 x$ ou par sa forme simplifiée $y = ax$ où a_0
$a = r\frac{\sigma_y}{\sigma_x}$ ou la droite de régression devient $y = r\frac{\sigma_y}{\sigma_x}x$

b) La droite de régression de x en y est donnée par $x = b_0 + b_1 y$ ou par sa forme simplifiée $x = by$ où $b_0 = 0$
$b = r\frac{\sigma_x}{\sigma_y}$ la droite devient $x = r\frac{\sigma_x}{\sigma_y}y$

Exercices

1) La production de cuivre en million de tonnes durant les années 1985 à 1994 est donnée dans le tableau ci-après :

Année	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Production	98,2	92,3	80,0	89,1	83,5	68,9	69,2	67,1	58,3	61,3

a) Reporter les données sur un graphique
b) Trouver l'équation de la droite des moindres carrés ajustant les données
c) Estimer la production du cuivre pour l'année 1995

2) Le tableau ci-après donne les notes finales en Algèbre et physique obtenues par 10 étudiants choisis au hasard parmi un vaste ensemble d'étudiants
   a) Représenter graphiquement les données
   b) Trouver l'équation de deux droites de régression
   c) Si la note d'algèbre est 75, quelle sera sa note en physique ?
   d) Si la note en physique d'un étudiant est 95, à quelle note doit-il s'attendre en Algèbre ?

Algèbre X	75	80	93	65	87	71	98	68	84	77
Physique Y	82	78	86	72	91	80	95	72	89	74

3) Ci-dessous sont présentées les données concernant cinq observations de deux variables *x* et *y*

x_i	1	2	3	4	5
y_i	3	7	5	11	14

   a) Représentez le nuage de points associé à ces données
   b) Quelle relation entre les deux variables le nuage de points indique-t-il ?
   c) Essayez de décrire la relation entre *x* et *y* en traçant une ligne droite à travers le nuage de points.
   d) Construisez l'équation estimée de la régression en calculant les valeurs de a_0 et a_1
   e) Utilisez l'équation estimée de la régression pour prévoir la valeur de *y* lorsque $x = 4$
   f) Calculez le coefficient de détermination r^2. Commentez l'adéquation de la régression aux données.
   g) Calculez le coefficient de corrélation de l'échantillon.

4) Ci-dessous sont présentées les données concernant cinq observations de deux variables *x* et *y*

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a) Représentez le nuage de points associé à ces données
b) Quelle relation entre les deux variables le nuage de points indique-t-il ?
c) Essayez de décrire la relation entre *x* et *y* en traçant une ligne droite à travers le nuage de points.
d) Construisez l'équation estimée de la régression en calculant les valeurs de a_0 et a_1
e) Utilisez l'équation estimée de la régression pour prévoir la valeur de *y* lorsque $x = 6$
f) Calculez le coefficient de détermination r^2. Commentez l'adéquation de la régression aux données.
g) Calculez le coefficient de corrélation de l'échantillon.

5) Ci-dessous sont présentées les observations collectées lors d'une analyse de la régression de deux variables.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a) Représentez le nuage de points associé à ces variables.
b) Construisez l'équation estimée de la régression correspondant à ces données.
c) Utilisez l'équation estimée de la régression pour prévoir la valeur de *y* lorsque $x = 4$
d) Quel est le pourcentage de la somme des carrés totale attribuable à l'équation estimée de la régression ?
e) Quelle est la valeur du coefficient de corrélation de l'échantillon ?

6) Les données suivantes correspondent à la taille (en inches) et au poids (en livres) de différentes nageuses.

Taille	68	64	62	65	66
Poids	132	108	102	115	128

a) Représentez le nuage de points associé à ces données en utilisant la taille comme variable indépendante.
b) Quelle relation entre les deux variables le nuage de points indique-t-il ?
c) Essayez de décrire la relation entre la taille et le poids en traçant une ligne droite à travers ces données.
d) Construisez l'équation estimée de la régression en calculant les valeurs de a_0 et a_1
e) Si une nageuse mesure 63 inches, à combien estimeriez-vous son poids ?

7) Les données suivantes correspondent aux dépenses publicitaires (en millions de dollars) et les ventes de caisses de boisson (en millions) de sept grandes marques de boisson non-alcoolisée.

8)

Marque	**Dépenses publicitaires ($)**	**Ventes de caisses**
Coca-Cola classique	131,3	1929,2
Pepsi-Cola	92,4	1384,6
Coca Light	60,4	811,4
Sprite	55,7	541,5
Dr. Pepper	40,2	536,9
Mountain Dew	29,0	535,6
7-Up	11,6	219,5

a) Représentez le nuage de points associé à ces données en utilisant les dépenses publicitaires comme variable indépendante.
b) Quelle relation entre les deux variables, le nuage de points indique-t-il ?
c) Tracez une droite à travers les données pour décrire une relation linéaire entre les dépenses publicitaires et les ventes de caisses de boisson.

d) Utilisez la méthode des moindres carrés pour estimer l'équation de la régression.
e) Interprétez la pente de l'équation estimée de la régression.
f) Prévoyez les ventes de caisses de boisson d'une marque qui dépense 70 millions de dollars de publicité.

36) The Wall Street Journal Almanac 2008 fournissait des informations sur les « performances » des compagnies aériennes américaines. Les données sur le pourcentage des vols arrivant à l'heure et le nombre de plaintes déposées pour 100000 passagers sont reproduites ci-dessous.

Compagnie	**Pourcentage d'arrivées à l'heure**	**Plaintes**
Southwest	81,8	0,21
Continental	76,6	0,58
Northwest	76,6	0,85
US Airways	75,7	0,68
United	73,8	0,74
American	72,2	0,93
Delta	71,2	0,72
America West	70,8	1,22
TWA	68,5	1,25

a) Représentez le nuage de points associé à ces données
b) Quelle relation entre les deux variables le nuage de points indique-t-il ?
c) Construisez l'équation estimée de la régression montrant la relation entre le pourcentage des vols arrivant à l'heure et le nombre de plaintes déposées pour 100000 passagers
d) Interprétez la pente de l'équation estimée de la régression.
e) Quel est le nombre estimé de plaintes pour 100000 passagers si le pourcentage des vols arrivant à l'heure est de 80% ?

9) Fréquemment, les grands hôtels proposent des tarifs spéciaux aux hommes d'affaire. On observe les tarifs les plus bas lorsque les réservations sont faites 14 jours à l'avance. Le tableau ci-dessous présente les tarifs « Affaires » et les tarifs réduits

obtenus suite à une réservation faite 14 jours à l'avance, pour une nuit, pour un échantillon de six hôtels du groupe ITT Sheration

Hôtel	**Tarif « Affaire » ($)**	**Tarif réduit ($)**
Birmingham	89	81
Miami	130	115
Atlanta	98	89
Chicago	149	138
Nouvelle Orléans	199	149
Nashville	114	94

a) Représentez le nuage de points associé à ces données en utilisant les tarifs « Affaire » comme variable indépendante.
b) Estimez par la méthode des moindres carrés l'équation de la régression.
c) L'hôtel ITT Sheraton de Tampa propose un tarif « Affaire » de 135 dollars par nuit. Estimez le tarif réduit pratiqué par cet hôtel lorsqu'on réserve 14 jours à l'avance.

10) Un responsable des ventes a collecté les données suivantes sur les années d'expérience et le montant des ventes annuelles de différents vendeurs.

Vendeur	**Années d'expérience**	**Ventes annuelles (milliers de dollars)**
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a) Représentez le nuage de points associé à ces données, en utilisant le nombre d'années d'expérience comme variable indépendante.
b) Estimez l'équation de la régression qui peut être utilisée pour prévoir les ventes annuelles sachant le nombre d'années d'expérience du vendeur.
c) Utilisez l'équation estimée de la régression pour prévoir les ventes annuelles d'un vendeur qui a neuf années d'expérience.

11) Les données suivantes correspondent aux revenus des hôtels et des casinos, en millions de dollars, de 10 hôtels-casinos de Las Vegas

Société	**Revenus des hôtels**	**Revenus des casinos**
Boyd Gaming	303,5	548,2
Circus Circus Entreprises	664,8	664,8
Grand Casinos	121,0	270,7
Hilton Corp. Gaming Div	429,6	511,0
MGM Grand, Inc	373,1	404,7
Mirage Resorts	670,9	782,8
Primadonna Resorts	66,4	130,7
Rio Hotel & Casino	105,8	105,5
Sahara Gaming	102,4	148,7
Station Casinos	135,8	358,5

a) Représentez le nuage de points associé à ces données en utilisant les revenus des hôtels comme variable indépendante.
b) Existe-t-il une relation linéaire entre les deux variables ?
c) Estimez l'équation de la régression qui lie les revenus des casinos aux revenus des hôtels.
d) Supposez que le revenu d'un hôtel soit de 500vmillions de dollars. Quel est le revenu estimé du casino ?

12) Le tableau suivant fournit le pourcentage de femmes travaillant dans chaque société (x) et le pourcentage de postes à responsabilité occupés par des femmes dans chaque

société (y), les sociétés échantillonnées font parties du secteur commercial

Société	x_i	y_i
Federated Department Stores	72	61
Kroger	47	16
Marriott	51	32
McDonald's	57	46
Sears	55	36

a) Représentez le nuage de points associé à ces données.
b) Quelle relation entre x et y le nuage de points indique-t-il ?
c) Estimez l'équation de la régression obtenue avec ces données.
d) Prévoyez le pourcentage de postes à responsabilité occupés par des femmes dans une société comptant 60% de femmes parmi ses employés.
e) Utilisez l'équation estimée de la régression pour prévoir le pourcentage de postes à responsabilité détenus par des femmes dans une société où 55% des emplois sont occupés par des femmes. Comparez cette valeur aux 36% observés chez Sears, une société qui compte 55% de femmes parmi ses employés.

13) Dans l'administration fiscale, on suppose que le montant raisonnable des déductions détaillées totales est fonction du revenu brut ajusté du contribuable. D'importantes déductions, dont les déductions pour dons et frais médicaux, sont probables lorsqu'il s'agit de contribuables dont les revenus bruts ajustés sont importants. Si un contribuable demande des déductions supérieures à la moyenne pour un niveau de revenu donné, la probabilité d'un contrôle fiscal s'accroît. Les données sur le revenu brut ajusté et le montant moyen ou raisonnable de déductions détaillées sont présentées dans le tableau suivant. Les données sont en milliers de dollars.

Revenu brut ajusté (en milliers de dollars)	Déductions détaillées totales (en milliers de dollars)
22	9,6
27	9,6
32	10,1
48	11,1
65	13,5
85	17,7
120	25,5

a) Représentez le nuage de points associé à ces données, en utilisant le revenu brut ajusté comme variable indépendante.
b) Utilisez la méthode des moindres carrés pour estimer l'équation de la régression.
c) Prévoyez le niveau raisonnable des déductions pour un contribuable dont le revenu brut ajusté s'élève à 52 500 dollars. Si ce contribuable demande une déduction totale de 20400 dollars, est-ce que la demande d'un contrôle faite par un agent de l'administration fiscale apparaît justifiée ? expliquez.