STATISTIQUE APPLIQUEE

(Outils d'aide à la décision)

Rachid MCHICH

Chap. 1: Rappels mathématiques

I - Statistique descriptive

(Mesures de tendance centrale et de dispersion, corrélation, …)

I-1 Exemples et définitions :

- Selon une enquête de Jupiter Media, 31 % des hommes adultes regardent la télévision au moins 10 heures par semaine. Cette proportion s'élève à 26 % chez les femmes adultes (The Wall Street Journal, 26-01-04).

- General Motors, leader des ristournes dans le secteur automobile, a fourni une réduction moyenne de 4 300 $ par véhicule en 2003 (USA Today, 23-01-04) .

- Plus de 40 % des managers de la société X connaissent une ascension professionnelle au sein de cette société.

Définition :

La **statistique** est l'ensemble des instruments et de recherches mathématiques permettant de déterminer les caractéristiques d'un ensemble de données (généralement vaste).

Les *statistiques* sont le produit des analyses reposant sur l'usage de la **statistique**.

Définition :

Cette activité regroupe trois principales branches :

• la collecte des données;

• le traitement des données collectées, aussi appelé: la *statistique descriptive* ;

• l'interprétation des données, aussi appelée: l'*inférence statistique*, qui s'appuie sur la théorie des *sondages* et la *statistique mathématique*.

Objectif :

Le but de la statistique est d'extraire et de résumer des informations pertinentes d'une liste de nombres difficile à interpréter par une simple lecture:

- les statistiques *exploratoires* : on explore d'abord les données pour avoir une idée qualitative de leurs propriétés ;
- les statistiques *confirmatoires*: on fait des hypothèses de comportement que l'on confirme ou que l'on infirme en recourant à d'autres techniques statistiques.

Outils de la statistique:

Les outils de la statistique descriptive sont:

1. Regrouper les observations ou mesures
2. Utiliser des représentations graphiques (histogrammes, secteurs, …etc)
3. Calcul de certains paramètres et indicateurs importants
4. Interprétation des résultats

I-2 Vocabulaire de la statistique descriptive:

1. **Population :** un ensemble de personnes, d'objets ou d'événements, base de l'étude statistique.

2. **Individu :** Un élément de cette population. (Exple: population d'employés d'une entreprise, population de produits d'une usine …etc).

3. **Echantillon :** c'est un sous-ensemble de la population, ayant les mêmes caractéristiques de la population-mère, utilisé en vue d'inférer quelque chose à propos de cette population.

4. **Caractère:** c'est une particularité ou propriété caractéristique de la population. L'étude statistique porte sur un **caractère**.

5. **Effectif d'une population:** c'est le nombre total des éléments constituant cette population, noté: **N**.

6. **Fréquence d'un caractère:** c'est le nombre d'individus possédant ce caractère divisé par l'effectif total de la population: $\mathbf{N_i}$.

Graphiques et tableaux

vs

Données numériques

On introduit plusieurs statistiques descriptives pour résumer la tendance centrale, la dispersion et la forme de la distribution d'un ensemble de données

I-3 Tableaux et Graphiques :

(i) Variables discrètes (VSD)

Dans ce cas, les modalités sont des nombres réels et qui peuvent alors être ordonnées. On parle dans ce cas de VSD. On peut regrouper les données dans un tableau comme suit :

Modalités	Effectifs	Fréquences	Pourcentages
x_1	N_1	$f_1 = N_1/N$	$p_1 = 100 f_1 \%$
.	.	.	.
.	.	.	.
.	.	.	.
x_k	N_k	f_k	p_k

On les représente alors sous forme d'histogramme ou de secteur grâce aux différentes fréquences.

(ii) Variables continues (VSC)

Dans ce cas, les valeurs du caractère appartiennent à des intervalles, qu'on regroupe en général dans des classes adjacentes, d'amplitudes pas forcément égales :

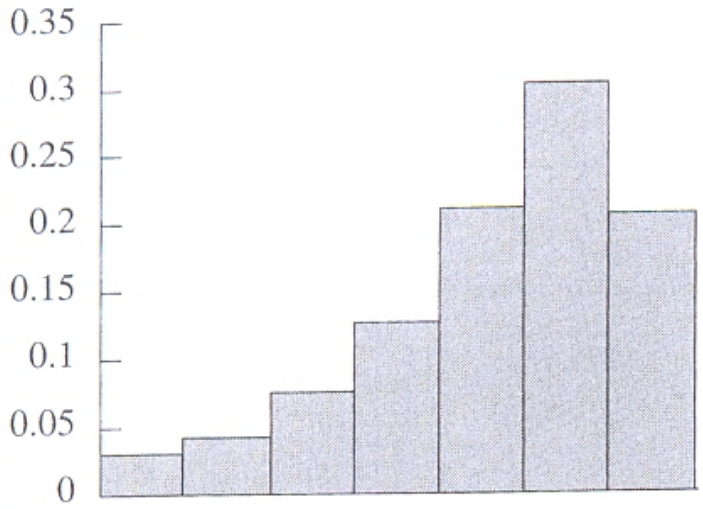
Classes	Centre des classes	Effectifs	Fréquences
$[X_0, X_1[$	$x_1=(X_0+X_1)/2$	n_1	$f_1=n_1/N$
.	.	.	.
.	.	.	.
.	.	.	.
$[X_{p-1}, X_p[$	.	n_k	f_k

La représentation se fait alors grâce à un histogramme dont les rectangles sont de largeur égale à l'amplitude de la classe.

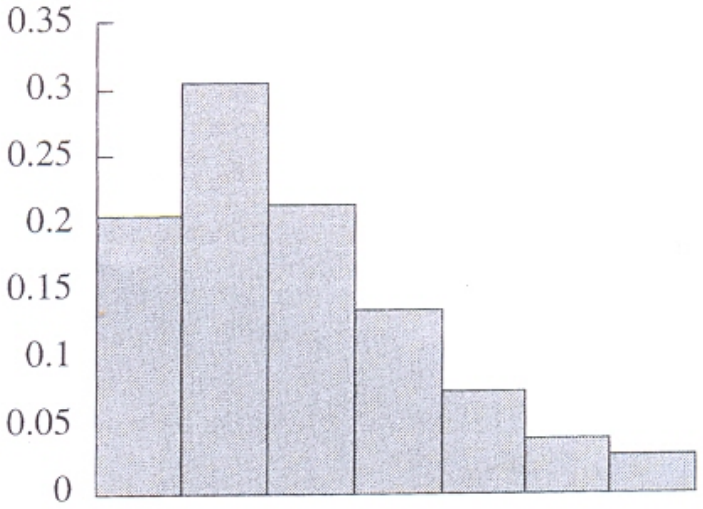
Exemple : Considérons les données quantitatives indiquant le temps nécessaire (en jours) pour effectuer l'audit de 20 clients par le cabinet d'un expert comptable.

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

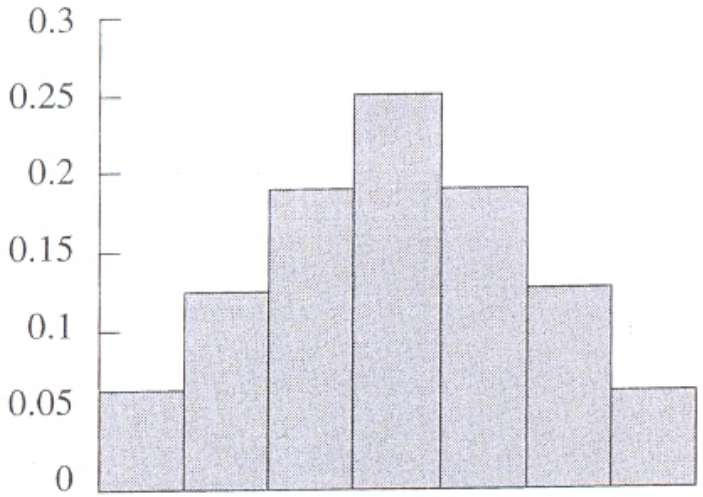
Histogramme A : Distribution modérément asymétrique à gauche
Degré d'asymétrie = –0,85



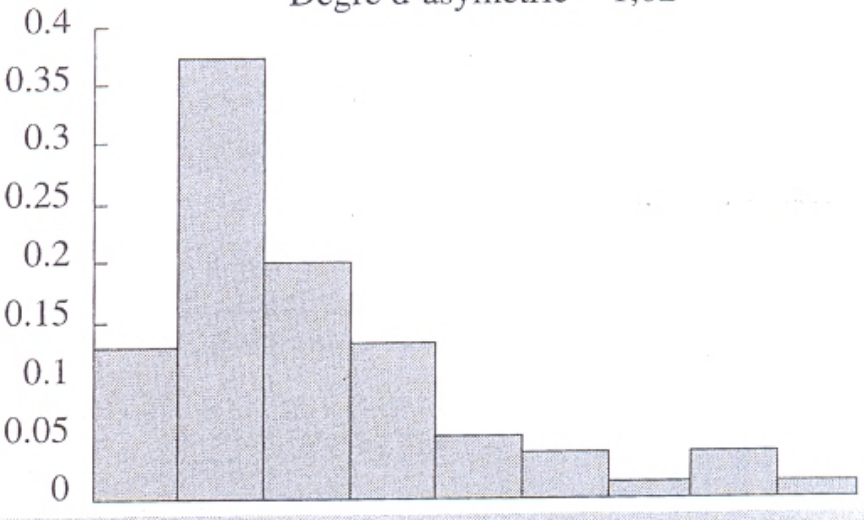
Histogramme B : Distribution modérément asymétrique à droite
Degré d'asymétrie = 0,85



Histogramme C : Distribution symétrique
Degré d'asymétrie = 0



Histogramme D : Distribution fortement asymétrique à droite
Degré d'asymétrie = 1,62



Effectifs et fréquences cumulées croissants et décroissants

- Pour une VSD :

Soit m_i une modalité d'une VSD. L'effectif cumulé croissant à gauche (resp. décroissant à droite) de m_i est le nombre d'individus pour lesquels la VSD prend des valeurs <= m_i (resp. >=m_i).

Ce nombre est donné par $N_{cc}= N_1+\ldots+N_i$ (resp. $N_{cd}= N_i+\ldots+N_k$).

- Fréquence cumulée croissante à gauche de *mi* : fcc= N_{cc}/N .
- Fréquence cumulée décroissante à droite de *mi* : $f_{cd}= N_{cd}/N$.

- <u>Pour une VSC :</u> *Les modalités sont des intervalles .*
- Effectif cumulé croissant à gauche de x_i
- Effectif cumulé décroissant à droite de x_i
- Fréquence cumulée croissante relative à une classe I_i
- Fréquence cumulée décroissante relative à une classe I_i

(iii) Caractère qualitatif

Dans ce cas, les modalités sont des qualités, qui ne peuvent pas être ordonnées. En général, on fait une représentation en secteurs.

I-4 Valeurs numériques :

- Lorsque les valeurs numériques sont issues d'un échantillon, on parle alors de **statistiques d'échantillon**.

- Lorsque les valeurs numériques sont issues d'une population, on parle de **paramètres de la population**.

- **Statistique d'échantillon** : Valeur numérique utilisée comme mesure d'un échantillon
- **Paramètre de la population** : Valeur numérique utilisée comme mesure de la population
- **Estimateur ponctuel** : Statistique d'échantillon utilisée pour estimer le paramètre correspondant de la population

Ci-dessous certaines notations utilisées:

	Statistiques d'échantillon	Paramètres de la population
Moyenne	$\bar{x}$	μ
Variance	s^2	σ^2
Ecart type	s	σ
Covariance	s_{xy}	σ_{xy}
Corrélation	r_{xy}	ρ_{xy}

Mesures de tendance centrale

- Moyenne : Elle est obtenue en sommant la valeur des observations et en divisant par le nombre d'observations.
  - Moyenne d'échantillon :

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

- Moyenne de la population :

$$\mu = \frac{\sum x_i}{N}$$

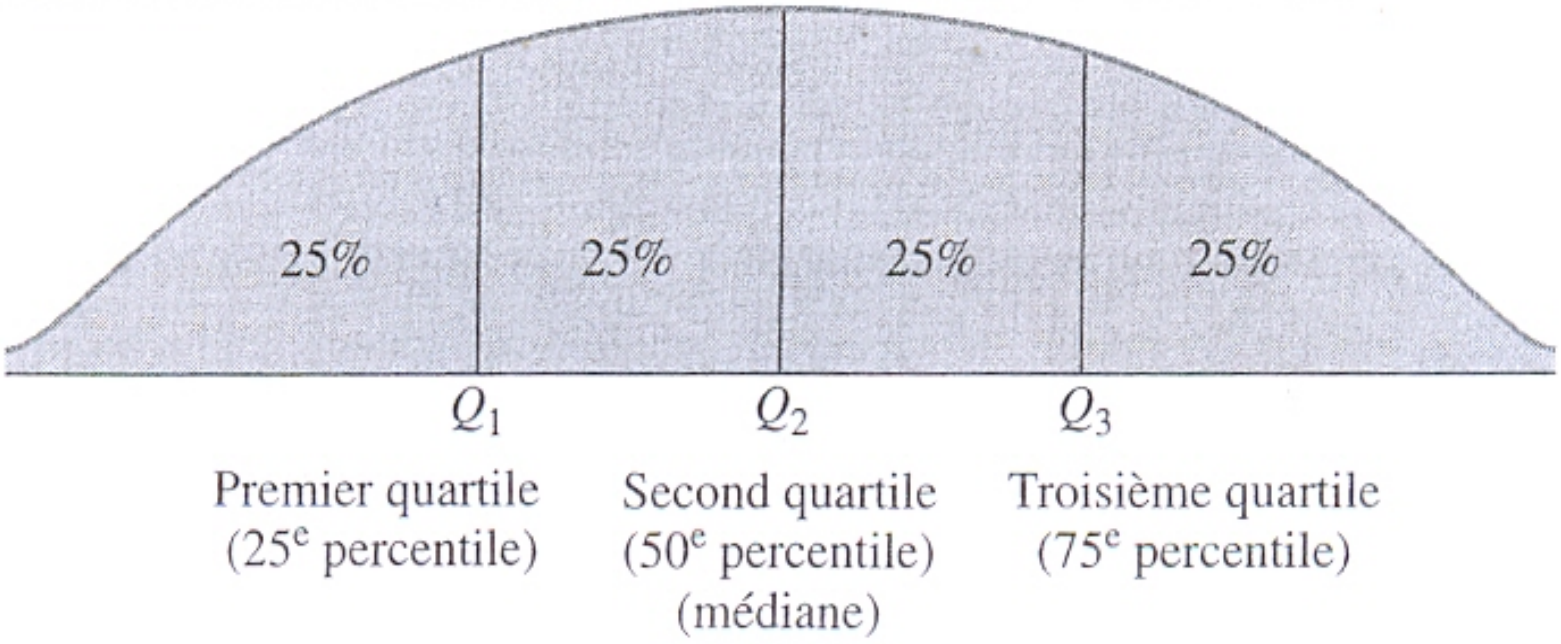
Mesures de tendance centrale

- Médiane : Il s'agit de la valeur centrale de l'ensemble des données, classés en ordre croissant.

- Mode : Défini comme la valeur de l'observation la plus fréquente.

- Percentile : Valeur telle que, au moins *p* pour cent des observations ont une valeur <= à cette valeur et au moins *(100 - p)* pour cent des observations ont une valeur >= à cette valeur. La médiane correspond au 50^{e} percentile.

- Quartile : Les 25e, 50e et 75e percentiles sont appelés respectivement premier quartile, deuxième quartile (médiane) et troisième quartile. Les quartiles divisent l'ensemble des données en quatre parties, chacune contenant environ 25% des données.



Mesures de dispersion

- Etendue : égale à la différence entre la plus grande et la plus petite valeurs.

- Etendue interquartile (EIQ): égale à la différence entre le 3e et le 1e quartiles :

$$EIQ = Q_3 - Q_1$$

- Variance : basée sur les écarts au carré des observations par rapport à la moyenne :

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Mesures de dispersion

- Ecart type : égal à la racine carrée de la variance

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

- Coefficient de variation : Mesure de dispersion relative, égale au rapport de l'écart type à la moyenne, multiplié par 100

$$\frac{Ecart\ type}{Moyenne} * 100$$

I-4 Détection des valeurs singulières :

Définition :

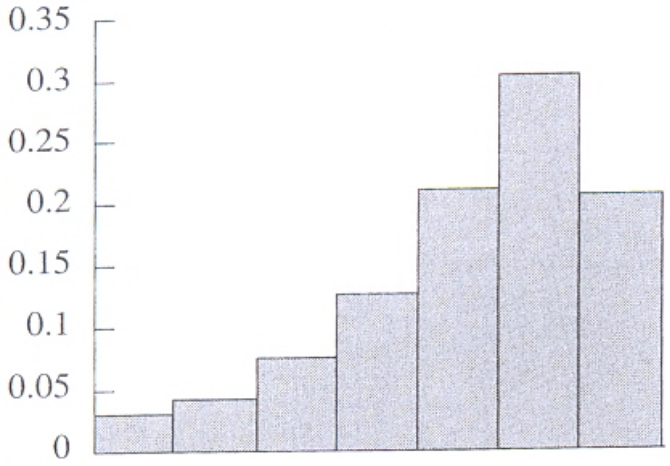
- Valeur singulière : Observation anormalement grande ou petite
  - Erreur d'enregistrement : à corriger avant toute analyse
  - Observation pas correctement incluse dans l'ensemble des données : à supprimer
  - Valeur inhabituelle, correctement enregistrée et qui appartient à l'ensemble des données: à conserver.

- **Forme d'une distribution**

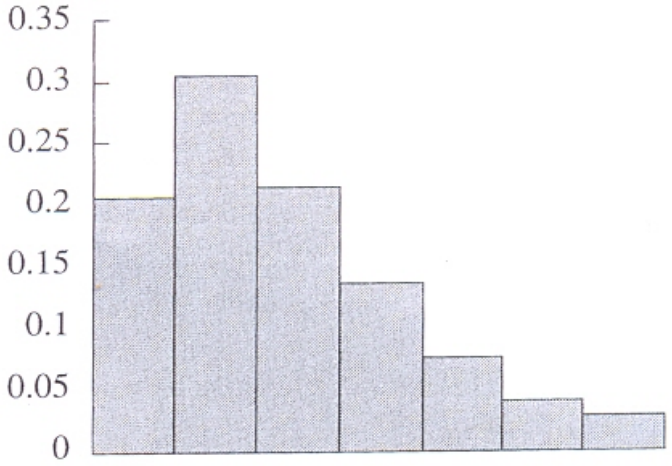
Degré d'asymétrie : Mesure de la forme d'une distribution de données.

- Des données biaisées à gauche sont caractérisées par un degré d'asymétrie négatif.
- Des données comportant un biais à droite sont caractérisées par un degré d'asymétrie positif.

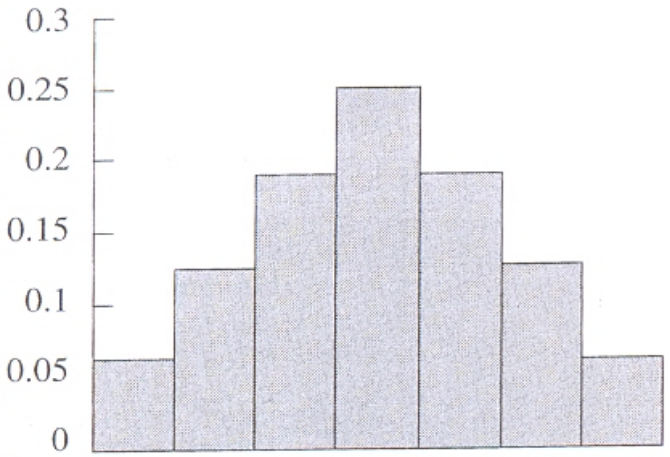
Histogramme A : Distribution modérément asymétrique à gauche
Degré d'asymétrie = –0,85



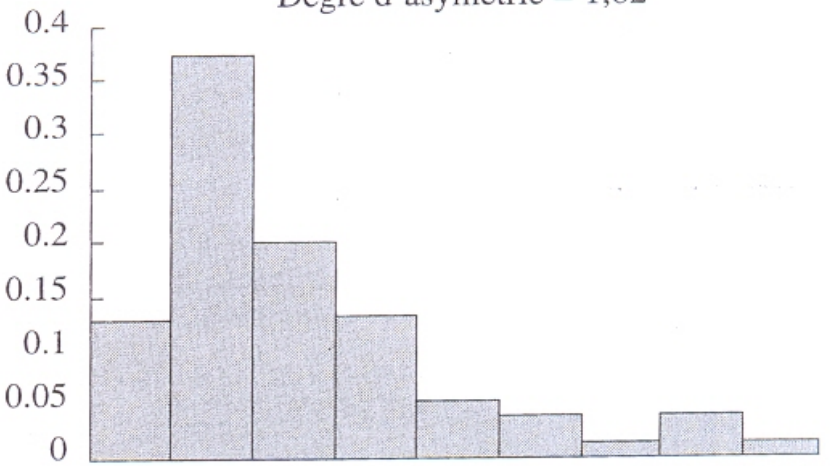
Histogramme B : Distribution modérément asymétrique à droite
Degré d'asymétrie = 0,85



Histogramme C : Distribution symétrique
Degré d'asymétrie = 0



Histogramme D : Distribution fortement asymétrique à droite
Degré d'asymétrie = 1,62



$$\gamma_1 = \frac{\mu_3}{\sigma^3} \quad avec \quad \mu_3 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^3$$

- **Variable centrée réduite z** : Valeur obtenue en divisant l'écart par rapport à la moyenne $(x_i - \overline{x})$ par l'écart type s.

$$z_i = \frac{x_i - \overline{x}}{s}$$

La variable centrée réduite mesure la distance, en nombre d'écarts type, entre l'observation x_i et la moyenne.

Exemple :

Nbr d'étudiants dans la classe	Ecart par rapport à la moyenne	Valeur de la variable centrée réduite
46 54 42 46 32		

$\overline{x} = ??$ $s = ??$

- **Théorème de Chebyshev :**

Théorème utilisé pour déduire le pourcentage d'observations qui se situent dans un intervalle de z écarts type de part et d'autre de la moyenne:

Théorème de Chebyshev :

« Au moins $(1 - \frac{1}{z^2})$ *des observations doivent se situer au plus à* ***|z|*** *écarts types de part et d'autre de la moyenne (càd. dans l'intervalle* $\left[\overline{x} - zs, \overline{x} + zs\right]$ *), avec* ***z >1 ».***

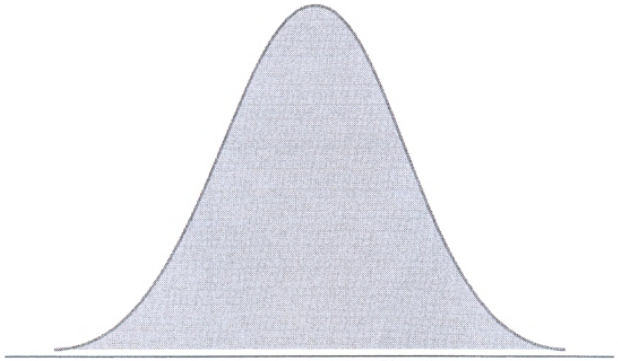
Exemple:

Supposons que la moyenne des notes de 100 étudiants de l'ENCGT soit égale à 70 et que l'écart type est de 5.

1. Combien d'étudiants ont obtenu une note entre 60 et 80?

2. Combien d'étudiants ont obtenu une note entre 58 et 82?

- **Règle empirique :** Règle qui donne le pourcentage d'observations situées dans les intervalles de un, deux ou trois écarts type autour de la moyenne, pour une distribution en forme de cloche (distribution dite normale)



Pour des données ayant une distribution en forme de cloche:

- Environ 68% des observations se situent dans $\left[\overline{x}-s,\overline{x}+s\right]$
- Environ 95% des observations se situent dans $\left[\overline{x}-2s,\overline{x}+2s\right]$
- Presque toutes les observations se situent dans $\left[\overline{x}-3s,\overline{x}+3s\right]$

- **Analyse exploratoire des données**

Résumé en cinq chiffres : Technique d'analyse exploratoire des données qui utilise cinq chiffres pour résumer les données: la plus petite valeur, le 1e quartile, la médiane, le 3e quartile et la plus grande valeur. Par exemple,

7710 7755 7850 7880 7880 7890 7920 7940 7950 8050 8130 8325

1) 7710
2) Q_1= 7865
3) Q_2=7905
4) Q_3= 8000
5) 8325

A peu près 25% des données sont comprises entre 2 valeurs adjacentes

II - Statistique bivariée

II-1 Mesures de la relation entre 2 variables

- **Nuage de points :** A chaque couple de données (x_i, y_i) est associé un point M dans le plan. On obtient ainsi ce qu'on appelle un nuage de points représentant la série statistique.

- **Point moyen :** $G(x_G, y_G)$ où :

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

$$y_G = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Ajustement affine :

- Ajustement à la règle (en passant par le point moyen) : méthode très approximative.

- Méthode de Mayer : 2 sous-nuages, puis 2 points moyens formant la droite de Mayer (passant aussi par le point moyen) : méthode assez approximative.

II-2 Mesures par la covariance :

Covariance : Mesure de la relation **linéaire** entre deux variables.

- Des valeurs positives indiquent une relation linéaire positive.
- Des valeurs négatives indiquent une relation linéaire négative.

- Covariance population : $\sigma_{xy} = \dfrac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$

- Covariance échantillon : $s_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

II-3 Mesures par le coefficient de corrélation :

- **Coefficient de corrélation :** Mesure de la relation **linéaire** entre deux variables, dont les valeurs sont comprises entre -1 et +1:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad \text{ou} \qquad \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

  - Des valeurs proches de +1 indiquent une forte relation linéaire positive.
  - Des valeurs proches de -1 indiquent une forte relation linéaire négative.
  - Des valeurs proches de 0 indiquent l'absence de relation linéaire.

- **Méthode des moindres carrés :**

- Droite de régression de Y en X , $(D_{Y/X}): y = ax + b$

où $a = \dfrac{\sigma_{xy}}{(\sigma_x)^2}$ et $b = \overline{Y} - a\overline{X}$

- Droite de régression de X en Y , $(D_{X/Y}): x = a'y + b'$

où $a' = \dfrac{\sigma_{xy}}{(\sigma_y)^2}$ et $b' = \overline{X} - a'\overline{Y}$

(A noter que les deux droites se coupent au point moyen)

Remarques :

1. Il est possible qu'un lien fort (mais non linéaire) entre *X* et *Y* conduise à une valeur faible de *r*. C'est pour cela que « *r* » est appelé des fois : *coefficient de corrélation de la partie linéaire entre X et Y.*

2. Deux variables dont « *r* » est proche de 0 sont dites décorrélées (à ne pas confondre avec indépendantes).

3. Un fort « *r* » n'implique pas forcément une relation de causalité entre *X* et *Y* (Existence possible d'une troisième variable *Z*).

4. Il existe plusieurs types d'ajustements non linéaires. Certains types peuvent être ramenés au cas de l'ajustement linéaire en utilisant la fonction logarithme népérien. Par exemple :

$$y = Cx^m \quad ou \quad y = Ca^x$$

5. On peut aussi faire un ajustement pour des V.S.C. en utilisant les centres des intervalles de modalité.