Patrie II : Fouille des données
Chapitre 1 : Rappel sur les statistiques

1. Introduction

- ✓ Objet des statistiques :
  - Etude d'un ensemble d'individus sur lesquels on observe des caractéristiques appelées variables. Selon le nombre de variables, on distingue
    - Techniques simples résumant les caractéristiques d'une variable (moyenne, médiane, etc.), permettant de détecter les valeurs atypiques.
    - Techniques s'appliquant à deux variables ou plus (corrélation, nuage de points).
- ✓ Objectifs
  - Mieux connaître la population étudiée par l'explication des variables.
  - Prévoir le comportement des individus qui ne sont non encore observés.

2. Statistiques descriptives d'une variable

- Types de variables
  - *Variable quantitative :* les valeurs prises sont numériques. Ces valeurs peuvent être
    - discrètes : c'est à dire appartenant à une liste dénombrable. **Exemple** : le nombre de pannes d'une machine, le nombre des jours de travail.
    - continues : les valeurs prises ne peuvent pas être comptées et appartiennent à un intervalle. **Exemple** : la température, la moyenne annuelle d'un étudiant.
  - *Variable qualitative :* les valeurs prises sont des labels. Ces valeurs peuvent être
    - nominales : quand elles ne sont pas ordonnables. Exemple : la couleur.
    - ordinales : quand il est possible de les ordonner selon un sens : petit < moyen < grand ; faible < normal < puissant.

2.1. Cas d'une seule variable :

- *Notions :*

Moyenne : la moyenne arithmétique est la somme des valeurs d'une variable quantitative, divisée par le nombre d'individus. Ex : la moyenne des âges 3, 12, 18 est 11.
Médiane : c'est la valeur qui sépare les valeurs d'une série statistique en deux. Ex : la médiane de la série 1 3 5 7 9 est 5.
Mode : il correspond à la valeur la plus fréquente. Ex : le mode de la série 2, 3, 3, 4, 7, 3, 2, 1, 3 est 3 avec l'effectif 4.
Variance : la variance sert à caractériser la dispersion des valeurs de la moyenne : variance de zéro ⇨ toutes les valeurs sont identiques, petite variance ⇨ les valeurs sont proches les unes des autres, variance élevée ⇨ celles-ci sont très écartées.
La formule de calcul de la variance (écart-type au carré) est la suivante.

$$\sigma^2 = \frac{\sum \left(X - \bar{X}\right)^2}{N}$$

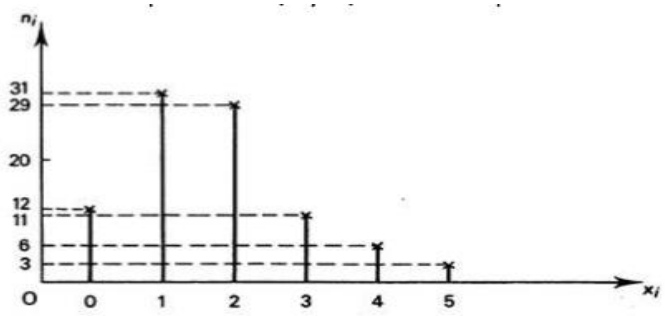
Ecart-type : C'est la racine carrée de la variance. Il permet de mesurer l'écart entre les valeurs de la série avec la même grandeur que celle des valeurs.

- Représentations graphiques

Les données statistiques peuvent être représentées graphiquement pour une meilleure interprétation et mémorisation. Ces représentations sont parfois suffisantes à elles-mêmes pour visualiser une population. Il existe différentes représentations graphiques associées au cas mono-variable.

- *Diagramme en bâtons :* il permet de représenter des effectifs d'une variable discrète sur deux axes : en abscisses les individus observés et en ordonnées représente l'effectif de chaque individu. On parle aussi de nuage de points.

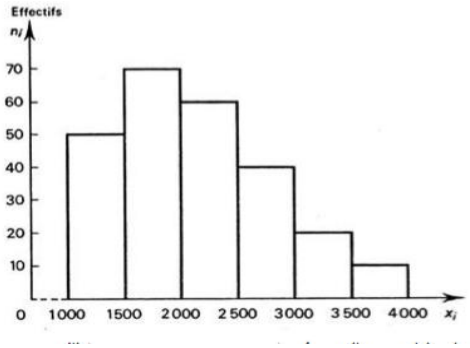
Exemple : soit la série suivante qui représente le nombre de foyers selon le nombre de personnes par foyer (tableau). Le diagramme à gauche représente graphiquement ce tableau.



Nombre d'enfants par foyer « x_i »	Nombre de foyer concernés f_i
0	12
1	31
2	29
3	11
4	6
5	3

- *Histogrammes* : Ils s'adaptent aux cas d'une variable continue quantitative dont les valeurs peuvent être classées en intervalles. L'axe des abscisses représente les classes et l'axe des ordonnées représente les valeurs sous forme de rectangles.

Exemple : soit la série suivante qui représente le nombre de personne selon la tranche de salaire (tableau). Le diagramme à gauche représente graphiquement ce tableau.

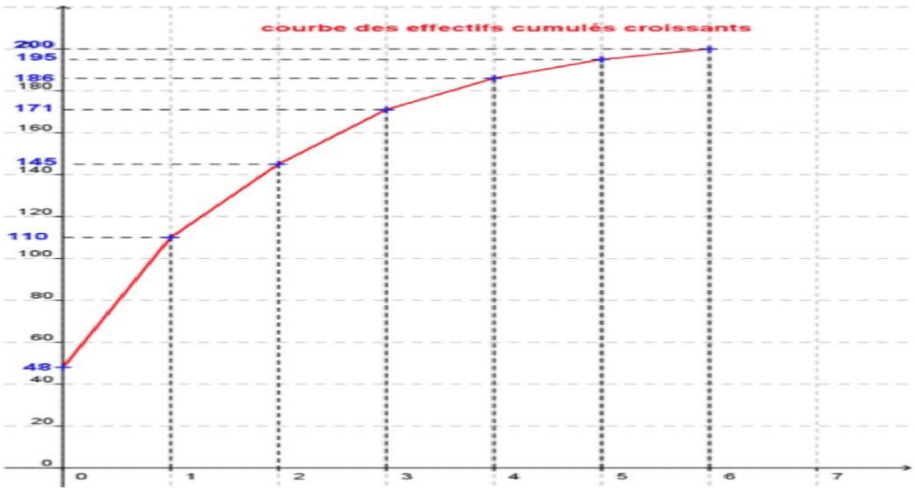


Salaire en € : x_i	Effectifs n_i
1000 à 1500	50
1501 à 2000	70
2001 à 2500	60
2501 à 3000	40
3001 à 3500	20
3501 à 4000	10
	250

- *Graphiques cumulatifs* : ils permettent de représenter les cumuls d'effectifs d'une série ordonnée. Si les effectifs initiaux ne correspondent pas à des cumuls, on les calcule d'abord avant de représenter le diagramme.

Exemple : le tableau suivant représente l'évolution de salaire selon le grade. La deuxième ligne représente le montant d'évolution et la troisième ligne représente le cumul de salaire

Modalités	0	1	2	3	4	5	6
Effectifs	48	62	35	26	15	9	5
ECC	48	110	145	171	186	195	200

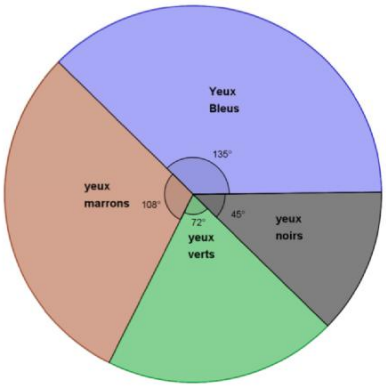


- *Diagrammes en secteur* : ce diagramme permet de représenter des proportions d'effectifs par rapport à la totalité. Avant de le dessiner, on calcule la proportion de chaque valeur de variable (individu). On calcule ensuite l'angle de chaque valeur de variable.

Exemple : soit le tableau suivant qui donne les effectifs et fréquences des couleurs d'objets observés. Les modalités sont les couleurs bleu, marron, vert et noir. Les fréquences sont traduites en angles en les multipliant par 3,6.

Modalités	bleu	marron	vert	noir	total
Effectifs	15	12	8	5	40
Fréquences	0,375	0,3	0,2	0,125	1
Fréquences en %	37,5	30	20	12,5	100
Angle (en °)	**135**	**108**	**72**	**45**	360

x 3,6



2.2. Cas de deux variables : il arrive souvent qu'on ait besoin d'analyser deux variables à la fois et qu'on cherche la relation entre elles. Par exemple : la relation entre la taille des enfants et leur âges. On note ce deux variables x et y.

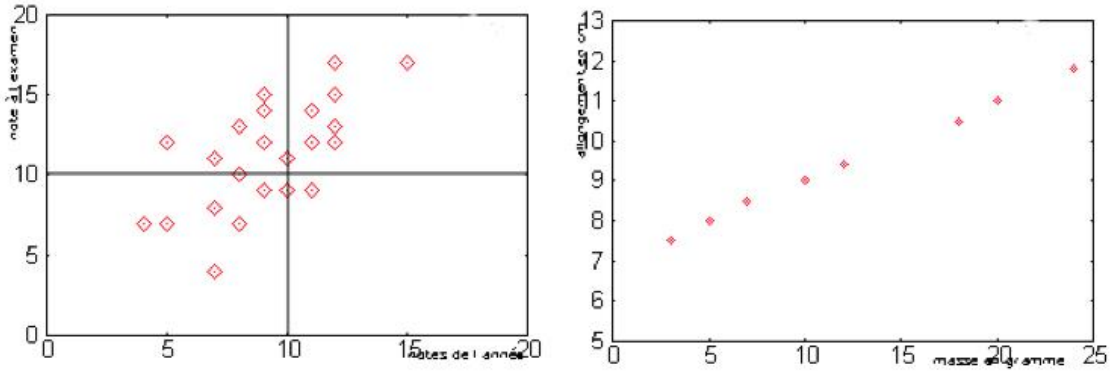
Notions : parmi les notions qui font intervenir les deux variables à la fois sont

- La covariance : $\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

- Le coefficient de corrélation linéaire: $r = \frac{cov(x,y)}{\sigma_x \sigma_y}$

La covariance décrit la relation entre les changements des deux variables. Si elle est positive, aux grands écarts d'une variable correspondent de grands écarts de la deuxième et vice-versa. Par contre, une covariance négative signifie qu'aux grands écarts d'une variable correspondent de petits écarts de la deuxième.

- Représentation graphique : chaque individu est représenté par un point dans un plan. Les valeurs d'une variable sont placées sur un axe et les valeurs de la deuxième variable sur l'autre axe. Cette représentation est appelée nuage de points. Elle permet de déduire visuellement s'il y a une relation entre les valeurs des deux variables. Dans les graphiques ci-dessous, le nuage de points à gauche témoigne que les valeurs des d'une variable sont indépendantes des valeurs de l'autre. Le nuage à droite montre qu'il se peut qu'il y ait une relation entre les valeurs.



Lorsqu'il peut exister une relation entre les valeurs des deux variables, on procède à l'ajustement

2.3. Cas multidimensionnel : le cas multidimensionnel concerne plusieurs variables à la fois. La représentation graphique n'est pas adaptée à l'être humaine. Le traitement de ces valeurs fait intervenir les méthodes d'analyse de données.

2. Analyse de données

Définition : l'analyse de données regroupe une famille de méthodes pour décrire un grand nombre de données avec comme objectif de faire ressortir les relations entre elles ou de comprendre ce qui les rend homogènes.

Exemples de méthodes

- Analyse en composantes principales (ACP) : réduire p variables corrélées en q variables non corrélées.
- Analyse factorielle des correspondances (AFC) : trouver des liens ou correspondances entre deux variables qualitatives (nominales) dans des tableaux de contingence.
- Analyse des correspondances multiples (ACM) : L'ACM est l'équivalent de l'ACP pour les variables qualitatives et elle se réduit à l'AFC lorsque le nombre de variables qualitatives est égal à 2.