Partie A : Statistique

Objectifs

L'objectif du partie A est de :

- ✓ Comprendre les notions de base : Population, Echantillon, Variable…
- ✓ Savoir résumer les données d'observation des variables sous forme de tableaux de distribution et de représentations graphiques adaptées.
- ✓ Connaître les principaux paramètres de position et de dispersion et savoir les calculer.
- ✓ Savoir commenter et interpréter les résultats.
- ✓ Analyser la distribution des valeurs des variables et le lien éventuel entre elles.

Prérequis

Il est recommandé aux apprenants de connaître.

- ➢ Notion des ensembles et sous-ensembles.
- ➢ Fonctions à plusieurs variables.

Mots clés :

Population, Echantillon, Variable aléatoire, Paramètres de position, Paramètres de dispersion, Corrélation, Régression.

I. Introduction

La statistique est un ensemble des méthodes qui servent à organiser les épreuves fournissant des observations, à analyser celles-ci et à interpréter les résultats. L'analyse statistique se subdivise en deux parties : Statistique descriptive et Statistique inférentielle.
Statistique descriptive : A pour but de décrire c'est à dire de résumer ou représenter les données par : représentation graphique, paramètres de position, de dispersion et de relation.

II. Définitions de base

1. **Population** : La collection d'objets ou de personnes étudiées (élèves, habitants, voitures…).

2. **Individu** : Elément de la population étudiée. (un élève, un habitant, une voiture,…).

3. **Echantillon** : Partie de la population étudiée. Nombre d'individus dans un échantillon noté n est appelé taille de l'échantillon

4. **Variable** : (Caractère) Propriété commune aux individus de la population, que l'on veut étudier.

Nous allons distinguer deux groupes de variables aléatoires : Les VA qualitatives et les VA quantitatives.

4.1. Variable qualitative : La variable est dite qualitative quand les modalités sont des catégories.

- Variable qualitative nominale : La variable est dite qualitative nominale quand les modalités ne peuvent pas être ordonnées.
- Variable qualitative ordinale : Les modalités de la variable possèdent la propriété d'ordre.

4.2. Variable quantitative : Une variable est dite quantitative si toutes ses valeurs possibles sont numériques.

- Variable quantitative discrète : Une variable est dite discrète, si l'ensemble des valeurs possibles est dénombrable (des nombres entiers).
- Variable quantitative continue : Une variable est dite continue si elle peut prendre toutes les valeurs dans un intervalle donné (borné ou non borné).

5. **Modalité** : l'une des formes particulières d'un caractère. La couleur des yeux est un caractère, ses modalités sont : bleu, vert, marron,…

III. Séries statistiques à une variable :

1. Effectif ou Fréquence absolue : (noté n_i) nombre d'apparitions de la valeur associé à un caractère dans un échantillon. ($n = \sum_{i=1}^{k} n_i$)

- Effectif cumulé croissante : (noté $N_i \nearrow$) somme cumulée du effectif avec tous ses effectifs précédents. $N_i \nearrow = \sum_{j=1}^{i} n_j \ ; \ j \leq i$

2. Fréquence relative : (noté f_i) est le rapport de cet effectif à l'effectif total de la population. $f_i = \frac{n_i}{n} \ ; i = 1, 2, \dots, k$

- Fréquence cumulé croissante : (noté $F_i \nearrow$) $F_i \nearrow = \sum_{j=1}^{i} f_j \ ; \ j \leq i$

Remarque :

- $0 \leq f_i \leq 1$
- f_i peut être exprimée en %. $(f_i \times 100)$
- $\sum_{i=1}^{k} f_i = 1$

3. Représentations tabulaires et graphiques :

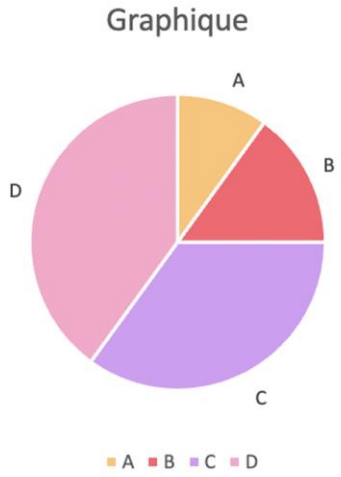
3.1. Caractère qualitatif :

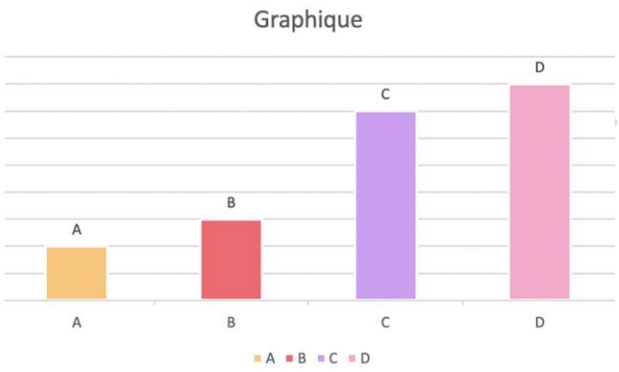
	Modalités	**Effectif n_i**	**Fréquence f_i**
	1	n_1	f_1
	2	n_2	f_2
	.	.	.
	.	.	.
	K	n_k	f_k
Total		n	1

Lorsque le caractère étudié est qualitatif on utilise un diagramme à bandes ou diagramme à secteurs.

i. **Diagramme à bandes :** C'est un diagramme qui à chaque modalité de la variable associé un rectangle de base constante et dont la hauteur est proportionnelle à l'effectif.

ii. **Diagramme à secteurs :** Ce type de graphique est formé d'un cercle divisé en secteurs, chaque secteur représente une catégorie particulière. La surface de chacun des secteurs est donné en degré par : $\theta_i = f_i .360$



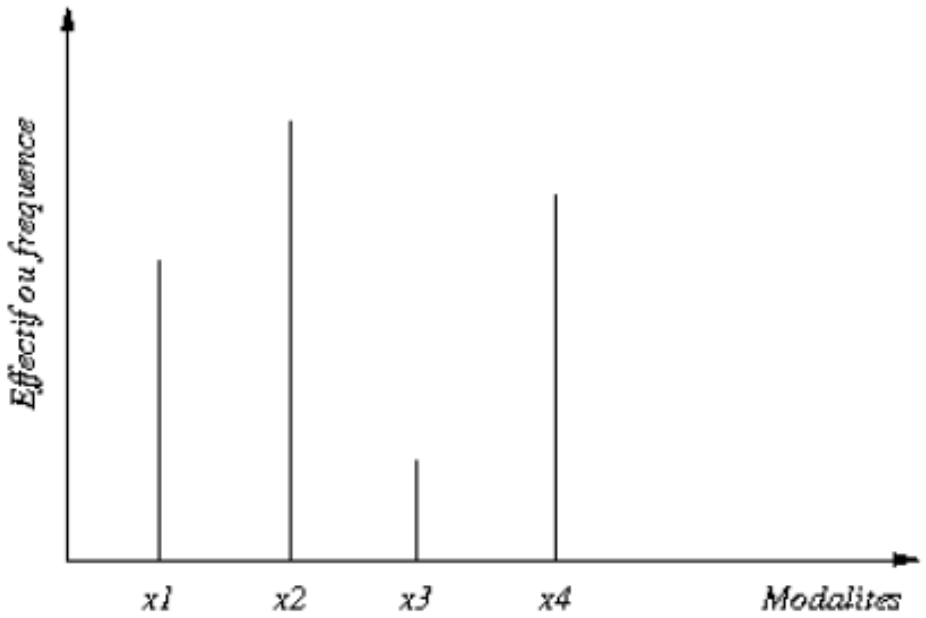


3.2. Caractère quantitatif discrète :

	Valeurs observées x_i	**Effectif n_i**	**Fréquence f_i**	**Fréquence cumulée F_i ↗**
	x_1	n_1	f_1	f_1
	x_2	n_2	f_2	$f_1 + f_2$
	.	.	.	.
	.	.	.	.
	x_k	n_k	f_k	1
Total	/	n	1	/

Lorsque le caractère étudié est quantitatif discrète on utilise un diagramme en bâton.

- **Diagramme en bâton :** On associe un segment vertical dont la hauteur est proportionnelle à la valeur (effectif ou fréquence) connue.



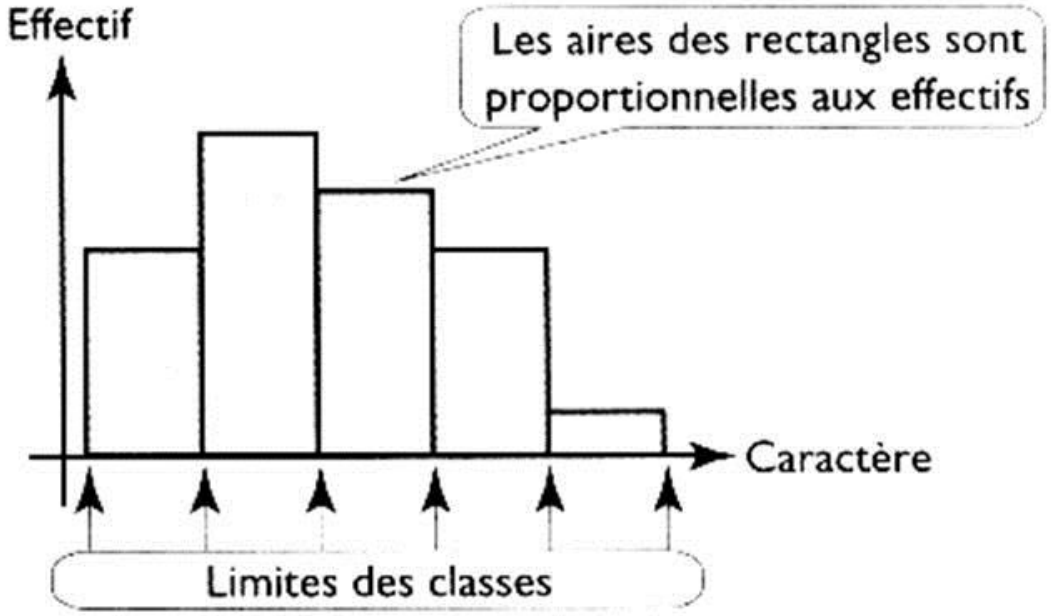
3.3. Caractère quantitatif continue :

Classes $[e_i , e_{i+1}[$	Centre c_i	Effectif n_i	Fréquence f_i	Fréquence cumulée F_i ↗
$[e_1 , e_2[$	c_1	n_1	f_1	f_1
$[e_2 , e_3[$	c_2	n_2	f_2	$f_1 + f_2$
	.	.	.	.
	.	.	.	.
$[e_k , e_{k+1}[$	c_k	n_k	f_k	1
Total	/	n	1	/

- Une classe est un intervalle de type $[e_i , e_{i+1}[$
- Le centre de classe c_i est : $\frac{e_i + e_{i+1}}{2}$
- L'amplitude d'une classe est : $a_i = e_{i+1} - e_i$

Lorsque le caractère étudié est quantitatif discrète on utilise un histogramme.

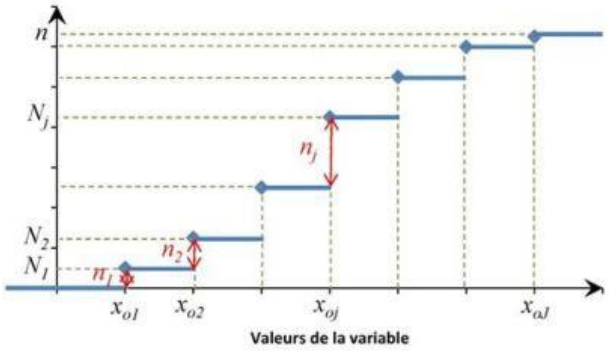
- **Histogramme :** C'est un diagramme composé des rectangles adjacents, chaque rectangle associé à chaque classe ayant une surface proportionnelle à l'effectif ou à la fréquence de cette classe.

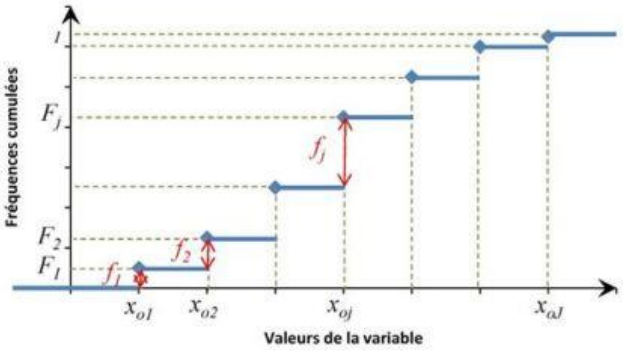


Remarque :

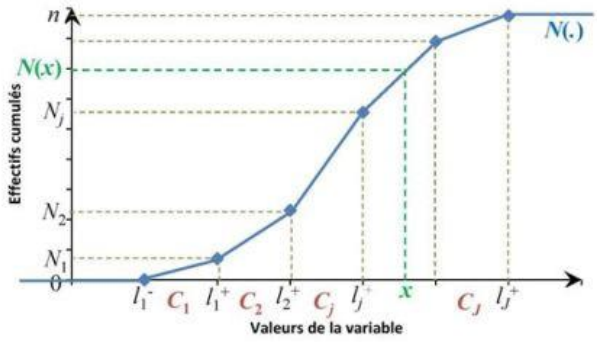
Diagramme cumulatif est le diagramme représentatif de la fréquence cumulée ou d'effectif cumulé.

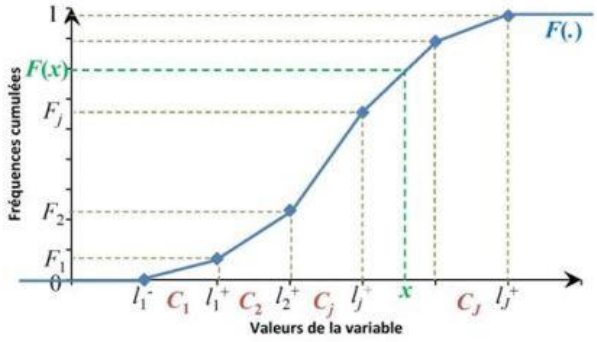
- Diagramme cumulatif pour une variable discrète : formé à l'aide d'un diagramme en escalier.





- Diagramme cumulatif pour une variable continue : formé à l'aide de segment de droite.





4. Caractéristiques de position :

Les paramètres de position ou « mesures de tendance centrale » sont des grandeurs susceptibles de représenter au mieux un ensemble de données. L'appellation «tendance centrale » vient du fait que ces paramètres donnent une idée de ce qui se passe au centre d'une distribution d'un ensemble de données. Les paramètres de position permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.
On distingue trois mesures de tendance centrale :

4.1. Moyenne : La moyenne constitue l'un des paramètres fondamentaux de tendance centrale, la mesure la plus calculée et la plus utilisée lors de la description de séries statistiques mais non suffisant pour caractériser une distribution.

- Pour une variable discrète : $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$

- Pour une variable continue : $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$

4.2. Mode « Mo »:

- Pour une variable discrète : Le mode est la valeur x de la sèrie ayant l'effectif le plus élevé.

- Pour une variable continue : Le mode se rapportera à la classe comportant le plus grand nombre d'individus : on parlera alors de classe modale.

4.3. Médiane « Me » : Valeur centrale sur l'axe x divisant l'échantillon en 2 groupes égaux d'individus. Pour calculer Me, il faut d'abord ordonner la série.

- Pour une variable discrète : On désigne par n le nombre d'observations.

  - Si n est impair : Il est possible d'identifier simplement la valeur qui partage la population en deux effectifs égaux. Le rang central étant égal à $\frac{n+1}{2}$ alors $Me = x_{\frac{n+1}{2}}$.

- Si n est pair : La médiane est alors égale à la moyenne des valeurs encadrant le milieu de la série alors $Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$

- Pour une variable continue : On cherche la classe contenant le $\frac{n}{2}$ individu de l'échantillon. Cette classe est appelée la classe médiane. En supposant que tous les individus de cette classe sont uniformément répartis à l'intérieur et $Me \in [e_j , e_{j+1}[$ la médiane se calcul de la façon suivante par interpolation linéaire :

$$Me = e_j + \frac{\frac{n}{2} - N_{j-1}\nearrow}{n_j} \times a_j$$

e_j : limite inférieure de la classe médiane ;
a_j : amplitude de la classe médiane ;
n_j : effectif de la classe médiane ;
$N_{j-1}\nearrow$: effectif cumulé inférieur à la classe médiane ;
n : taille de l'échantillon.

5. Caractéristiques de dispersion :

Les paramètres de dispersion nous renseignent sur la dispersion des valeurs autour de la valeur centrale de référence.

5.1. Etendue « E » : Etendue d'une série statistique quantitative est la différence entre la plus grande valeur de x et la plus petite valeur. $E = x_{max} - x_{min}$

5.2. Variance « V(x) » : Il est souvent intéressant de savoir si les valeurs sont très dispersées ou si elles sont proches de la moyenne. La variance est la caractéristique de dispersion la plus utilisée.

- Pour un variable discrète : $V(x) = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2$

- Pour une variable continue : $V(x) = \frac{1}{n} \sum_{i=1}^{k} n_i (c_i - \bar{x})^2$

5.3. Ecart-type « σ(x) » : La racine carrée de la variance : $\sigma(x) = \sqrt{V(x)}$

5.4. Coefficient de variation « Cv » : Si on veut comparer plusieurs séries statistiques ayant des moyennes très différentes, il vaut mieux se référer au coefficient de variation plutôt que l'écart-type. $Cv = \frac{\sigma_x}{\bar{x}} \times 100$

Remark :

- Plus la valeur du coefficient de variation est faible, plus la dispersion autour de la moyenne est petite, plus la population est homogène.

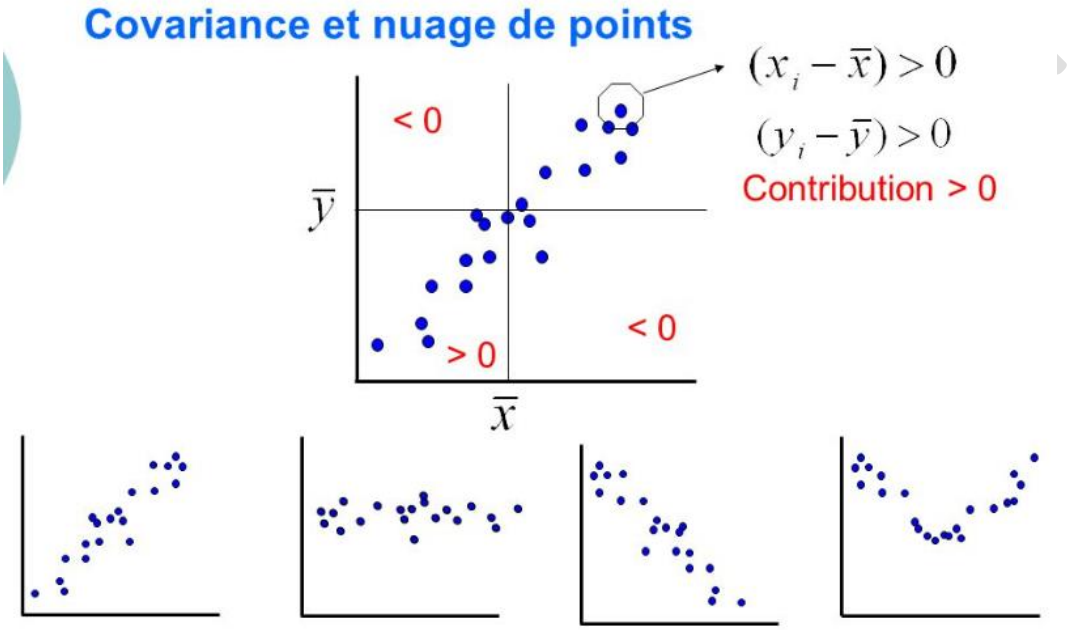
- On considère qu'une distribution de données est homogène, lorsque Cv. est égal ou inférieur à 15%.

IV. Séries statistiques à deux variables :

L'objectif est d'analyser la distribution des valeurs des variables x et y et le lien éventuel entre elles.

1. Nuage des points :

Un graphique qui traduise les deux séries statistiques à l'aide de diagramme à 2 dimensions. Soit x et y deux variables statistiques numériques observées sur k individus. Dans un repère orthogonal$(o, \vec{\imath}, \vec{\jmath})$, l'ensemble des k points de coordonnées (x_i ; y_i) forme le nuage de points associé à cette série statistique.



2. Covariance :

On appelle covariance de la série statistique double de variables x et y le nombre réel

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{k} x_i y_i - \bar{x}\bar{y}$$

3. Coefficient de corrélation linéaire :

Le coefficient de corrélation linéaire est un nombre permettant de déterminer l'intensité d'un lien linéaire entre deux variables quantitatives.

Le coefficient de corrélation linéaire d'une série statistique de variables x et y est le nombre r défini par : $\rho = \frac{Cov\,(x\,,y)}{\sigma_x \sigma_y}$

Remarque :

- Le coefficient de corrélation est une valeur qui n'a pas d'unité et qui est toujours comprise entre -1 et +1
- Un coefficient de corrélation linéaire est positif indique un lien linéaire positif, alors que, si ρ est négatif, le lien linéaire entre les deux variables est négatif.
- Plus la valeur de ρ est près de -1 ou +1 plus le lien linéaire ente les deux variables est fort.

4. Droite de régression :

Une droite de régression est la droite qui s'ajuste le mieux à un nuage de points présentant une corrélation linéaire. La droite de régression sert à faire des prévisions. On parle de corrélation linéaire lorsque les points d'un nuage ont tendance à s'aligner. Plus la tendance est forte, plus la corrélation linéaire est forte.

La droite D d'équation y = ax + b est appelée droite de régression de y en x de la série statistique si la quantité suivante est minimale : $S = \sum_{i=1}^{k}[y_i - (ax_i + b)]^2$

Pour définir les coefficients a et b; on développe S et on considère successivement comme un trinôme en b; puis, b étant déterminé, comme un trinôme en a: On trouve :

$$a = \frac{Cov\,(x\,,y\,)}{\sigma_x^2}$$

$$b = \bar{y} - a\,\bar{x}$$

Modèle d'une régression linéaire simple

- Le modèle de la régression:

$$Y_i = \alpha + bX_i + \varepsilon_i$$

- alors, toutes les régressions linéaires simples sont décrites par deux paramètres, l'ordonnée à l'origine (α) et la pente (b)