

Travaux Dirigés d' « Introduction à l'économie ».

Licence 1^{ère} année d'Administration Economique et Sociale.

Fiche 2 : Initiation aux statistiques.

Introduction : Utilisation des statistiques.

La presse, les médias où les ouvrages sont de plus en plus riches en données chiffrées. On peut souvent lire à la vue de ces données qu'on est face à une mutation de la conjoncture... Les méthodes statistiques permettent de synthétiser l'information disponible.

I. Définitions

1. Premices à la statistique :

Avant d'étudier les principaux paramètres statistiques, il convient de savoir ce que représente un taux de variation.

L'intérêt du taux de variation réside dans le fait qu'il permet de mesurer une évolution.

$$\text{Taux de variation} = \frac{V_f - V_i}{V_i} * 100$$

Avec V_f : Valeur finale de l'observation.

V_i : Valeur initiale de l'observation.

Remarques :

- Pour calculer un taux de variation, il faut connaître la grandeur de départ ou base.
- On ne peut pas ajouter ou soustraire des taux de variation (il faut alors passer par les coefficients multiplicateurs).
- Un taux de variation positif est appelé taux de croissance.
- Il ne faut pas confondre la baisse d'une grandeur (le taux de variation est donc négatif) et le ralentissement de sa croissance (le taux de variation baisse) (autrement dit confondre une variation absolue et une variation relative)

$$\text{Coefficient multiplicateur} = \frac{\text{Taux de variation}}{100} + 1$$

Le coefficient multiplicateur est très utile pour ajouter des taux de variation.

Exemple :

Le PIB d'un pays varie de +10% puis de +5% et enfin baisse de 7 %. Quelle est la variation totale en % du Pib ?

Il faut transformer les taux de variation en coefficients multiplicateurs ainsi le Pib est multiplié par 1,1 puis par 1,05 puis par 0,93 donc au total le PIB est multiplié par $(1,1 \times 1,05 \times 0,93 = 1,074)$ donc le Pib a progressé de 7,4 % (ce qui est donc différent de $10+5-7 = 8$)

Le Taux de Croissance Annuel Moyen :

$$\text{TCAM} = n\sqrt{\text{coefficient } t \text{ multiplicateurs de la période considérée}} - 1$$

2. Définition de la statistique :

"La statistique est l'étude méthodique des faits sociaux par des procédés numériques destinés à renseigner et à aider les gouvernements".

Il faut distinguer la statistique des statistiques. Les statistiques sont les renseignements numériques pouvant être obtenus sur des individus ou des variables. La statistique est l'ensemble des méthodes qui ont pour but l'étude scientifique des données collectées.

3. Définitions des termes usuels :

a) Population :

Tout ensemble dénombrable est appelé population. Une population n'est pas obligatoirement constituée d'êtres vivants, on peut disposer d'une population de biens immatériels, événements...

b) Variable :

Une variable est une règle de mesure qui permet d'apparier un nombre à un événement simple.

On distingue les variables quantitatives des variables qualitatives.

- Un caractère quantitatif a des modalités auxquelles on peut associer un nombre ou un intervalle (appelé variable ou statistique).
- Un caractère qualitatif, son observation ne peut être traduite par un membre. Ces modalités sont repérées par un mot, une lettre traduisant un état.

Variable qualitative	Variable nominale (numéros fixés arbitrairement)
	Variable ordinale (numéros ordonnés)
Variable quantitative	Intervalle
	Rapports

Dans les variables quantitatives, nous pouvons distinguer les variables quantitatives discrètes et continues.

- Concernant les variables discrètes, le nombre de valeurs que peut prendre la variable est limité.
- Concernant les variables continues, le nombre de valeurs que peut prendre la variable est illimité.

II. Les paramètres de tendance centrale :

L'objectif est de définir une valeur de la variable susceptible de résumer l'ensemble des valeurs de la variable statistique. Il faut déterminer une valeur située au centre de la distribution, d'où la recherche d'un paramètre de tendance centrale.

Plusieurs notions : le mode, la moyenne arithmétique et la médiane.

1. Le mode :

Le mode de distribution est la valeur du caractère la plus fréquente, c'est-à-dire, la valeur la plus souvent observée dans un ensemble de données.

Exemple :

Lors d'un tournoi d'hockey, Audrey a compté 7, 5, 0, 7, 8, 5, 5, 4, 1 et 5 buts en dix parties. Le mode de son ensemble de données est 5, puisque cette valeur apparaît le plus souvent (quatre fois). Il y a fort à parier que Audrey compterait 5 buts si l'on choisissait une partie au hasard.

2. La moyenne :

Il s'agit de l'ensemble des valeurs obtenues divisées par le nombre d'observations.

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

Avec μ : moyenne d'une population.
N : nombre d'unités d'observations.
X : Valeurs obtenues.

Exemple :

Le FC Montpellier organise un tournoi de football une fois par an. Au cours de la présente dernière, l'attaquant en tête de l'équipe a marqué 7, 5, 0, 7, 8, 5, 5, 4, 1 et 5 buts en dix parties. Quelle était sa moyenne de buts marqués ?

Moyenne = somme de toutes les valeurs observées / nombre d'observations
= (7 + 5 + 0 + 7 + 8 + 5 + 5 + 4 + 5 + 1) ÷ 10
= 47 ÷ 10
= 4,7

3. La médiane :

La médiane est une valeur qui sépare une série d'observations ordonnées en ordre croissant ou décroissant, en 2 parties comportant le même nombre d'observations.

En d'autres termes, la médiane divise en 2 sous parties égales la population étudiée. La première moitié des observations doit rassembler 50% des individus, et la seconde moitié les 50% restants.

On distingue 2 cas :

- Si le nombre d'observations à étudier est impair.

$$Md = \frac{(N+1)\text{ème observations}}{2}$$

Exemple :

*Supposons qu'un coureur effectue les temps suivants :
26,1, 25,6, 25,7, 25,2 et 25,0 secondes.*

Comment calcule-t-on le temps médian?

On commence d'abord en classant les valeurs dans l'ordre ascendant : 25,0, 25,2, 25,6, 25,7 et 26,1.

$$\begin{aligned}\text{Médiane} &= \text{la valeur } (n + 1) \div 2 \\ &= (5 + 1) \div 2 \\ &= 3 \qquad \text{donc} = 25.6 \text{ secondes.}\end{aligned}$$

- Si le nombre d'observations est pair.

$$\begin{aligned}\text{Md} &= \frac{(N + 1)\text{ème observations}}{2} \\ \text{Puis Md} &= \frac{(\text{valeur au desous de la médiane} + \text{valeur en deça de la médiane})}{2}\end{aligned}$$

Exemple :

Le coureur effectue sa sixième course en 24,7 secondes.

Dans ce cas, quelle est la valeur médiane?

On place, encore une fois, les données dans l'ordre ascendant : 24,7, 25,0, 25,2, 25,6, 25,7, 26,1. On utilise ensuite la même formule pour calculer le temps médian.

$$\begin{aligned}\text{Médiane} &= \text{la valeur } (n + 1) \div 2 \\ &= (6 + 1) \div 2 \\ &= 7 \div 2 \\ &= 3,5\end{aligned}$$

La médiane tombe entre les 3e et 4e valeurs, c'est-à-dire la 3,5e observation dans l'ensemble de données. On calcule donc la médiane en établissant la moyenne des deux valeurs intermédiaires, 25,2 et 25,6.

$$\begin{aligned}\text{Moyenne} &= (\text{valeur au-dessous de la médiane} + \text{valeur au-dessus de la médiane}) \div 2 \\ &= (\text{troisième valeur} + \text{quatrième valeur}) \div 2 \\ &= (25,2 + 25,6) \div 2 \\ &= 50,8 \div 2 \\ &= 25,4 \qquad \text{donc le temps médian est de 25.4 secondes.}\end{aligned}$$

Rappel : Variables qualitatives et quantitatives.

I. La variable qualitative : (la variable caractère)

Décrit l'individu étudié.

Ses éléments sont des **modalités**:

Exemples :

Sexe=2 modalités, féminin (1) ou masculin (2);

Code postal= (plusieurs modalités, tous les codes postaux de France)

Couleur des yeux, type de voiture (Berline, 4x, etc...)

II. La variable quantitative :

La variable est dite **quantitative** quand les modalités (**appelés aussi observations**) sont **mesurables**. Chaque individu est repéré par un nombre.

SERIES STATISTIQUES A UNE VARIABLE

IV. CARACTERISTIQUES D'UNE SERIE STATISTIQUE

4-1. Caractéristiques de position

Un tableau statistique ou un graphique sont parfois longs à consulter, sans permettre d'avoir une idée suffisamment concise de la distribution statistique observée.

Les caractéristiques d'une série permettent d'avoir une idée d'ensemble sur cette série. Elles servent aussi à comparer plusieurs séries statistiques.

Nous distinguerons les caractéristiques de position et les caractéristiques de dispersion.

Considérons les notes des deux épreuves des tableaux 1 et 7.

Tableau 7. Notes sur 20 obtenues par des élèves de la Terminale au cours d'une épreuve 2.

13-17-08-07-08-05-09-09-07-11-07-08-10-12-07-05-15-08-07-

08-11-08-11-12-09-14-13-09-13-06-16-12-12-08-19-14-09

On souhaite en déduire une comparaison du niveau des élèves. Le plus simple est de caractériser chaque série par un seul nombre et de comparer. Plusieurs choix sont possibles pour déterminer ce nombre unique.

a. Mode (ou dominante)

On appelle *mode* ou *dominante* d'une série statistique la *valeur qui a le plus grand effectif* (ou la plus grande fréquence).

Exemple. épreuve 1 : le mode est 14

 épreuve 2 : le mode est 08

Dans le cas des séries classées, on considère des classes d'égale étendue et on appelle *classe modale* la classe qui a le plus grand effectif. Par convention, le mode est le centre de la classe modale.

Remarques :

- Il peut arriver qu'une série statistique présente plusieurs modes. On parle de série plurimodale.
- Le mode existe même lorsque le caractère est qualitatif.

b. Médiane

- Reprenons l'exemple des notes obtenues par les élèves de la Terminale au cours de l'épreuve 1, on constate que sur 40 élèves 20 ont une note < 13 et 20 ont une note > 13 . Par conséquent le "milieu" de la classe est 13. On dit que la série a une médiane 13.

- Définition : La médiane M est la valeur du caractère qui partage l'effectif total en deux parties d'effectifs égaux.

- Détermination de la médiane

i-. Si la population est peu nombreuse, la médiane est obtenue de façon immédiate :

- Cas d'une série constituée d'un nombre impair de valeurs rangées par ordre croissant.

La médiane est le nombre qui est au milieu de la série. (Exemple : notes de l'épreuve 2)

05-05-06-07-07-07-07-08-08-08-08-08-08-09-09-09-09

09-10-11-11-11-12-12-12-13-13-13-14-14-15-16-17-19

La médiane est 09

- S'il y a un nombre pair de valeurs, la médiane est un nombre arbitrairement choisi entre les deux nombres consécutifs au milieu de la série. C'est le cas des notes de l'épreuve 1.

ii-. Si la population est nombreuse, on se sert du tableau des effectifs cumulés.

- Cas d'un caractère discret : La médiane est la valeur du caractère à partir de laquelle l'effectif cumulé atteint ou dépasse la moitié de l'effectif total (ou la fréquence cumulée atteint ou dépasse 0,5)

- Cas d'une répartition en classes : On détermine d'abord la classe à laquelle appartient la médiane et l'on convient de choisir la médiane en admettant que, dans cette classe, les valeurs du caractère sont uniformément distribuées.

- Détermination graphique de la médiane

On utilise la courbe des effectifs cumulés. La médiane est l'abscisse du point de la courbe ayant pour ordonnée la moitié $N / 2$ de l'effectif total. (ou l'ordonnée 0,5 d'une courbe de fréquence)

Soit $[AB]$ le segment contenant le point d'ordonnée $N/2$. L'abscisse de ce point est la

médiane M tel que :
$$\frac{M - x_a}{\frac{N}{2} - y_a} = \frac{x_b - x_a}{y_b - y_a}.$$

c. Moyenne arithmétique

- Définition :

La *moyenne arithmétique* d'une série statistique est égale à la somme des valeurs du caractère divisées par leur nombre.

i- Cas des données énumérées : $\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N}$

ii- Cas d'une variable discrète : on utilise la moyenne pondérée

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_c x_c}{N} = \frac{\sum_{i=1}^c n_i x_i}{\sum_{i=1}^c n_i}$$

exemple : notes de l'épreuve 1

iii- Cas où les valeurs sont regroupées en classes : les n_i valeurs de la i -ème classe sont supposées regroupées au centre de la classe.

- Présentation des calculs

Exemple : reprenons les notes de l'épreuve 1

classes	Centre x_i des classes	Effectif n_i	$n_i x_i$
] 0 , 2]			
] 2 , 4]			
] 4 , 6]			
] 6 , 8]			
] 8 , 10]			

classes	Centre x_i des classes	Effectif n_i	$n_i x_i$
]10,12]			
]12,14]			
]14,16]			
]16,18]			
]18,20]			
Total			

Tableau 8.

- Propriétés de la moyenne

Changement d'origine - changement d'échelle

Soit une série statistique (x_i, n_i) . On démontre que, si $\bar{x}$ est la moyenne de cette série, alors pour tout réel a la moyenne de la série $(x_i - a, n_i)$ est $\bar{x} - a$.

Pour tout réel h , la moyenne de la série (hx_i, n_i) est $h\bar{x}$.

x_i	10	14	19	20
n_i	5	45	30	20

Exemple : considérons la série statistique

Posons $u_i = x_i - 14$. On a le tableau :

x_i	n_i	$u_i = x_i - 14$	$n_i u_i$
10	5	- 4	- 20
14	45	0	0
19	30	5	150
20	20	6	120
	100		260

La moyenne $\bar{u} = \frac{260}{100} = 2,6$ d'où $\bar{x} = 2,6 + 14 = 16,6$

d. Quantiles

Ce sont des caractéristiques de position de même nature que la médiane. Il s'agit des valeurs du caractère qui correspondent aux effectifs cumulés.

- *Quartiles* : Pour une série statistique dont les valeurs sont classées par ordre croissant, la médiane partage la série des résultats en deux parties de même effectif.

On peut aussi partager en 4 parties de même effectif. On obtient ainsi 3 valeurs :

Q_1 qui correspond à $\frac{N}{4}$, Q_2 à $\frac{2N}{4}$, et Q_3 à $\frac{3N}{4}$ appelées quartiles. Q_2 est la médiane.

L'intérêt des quartiles consiste à diminuer l'importance des extrémités d'une série statistique.

- *Déciles, centiles* : Les déciles et les centiles partagent en dix ou cent parties d'effectifs égaux.

Les déciles correspondent à $\frac{N}{10}$, $\frac{2N}{10}$, ..., $\frac{9N}{10}$. Le 5^{ème} décile est la médiane.

Et, pour les centiles : $\frac{N}{100}$, $\frac{2N}{100}$, ..., $\frac{99N}{100}$ Le 5^{ème} centile est la médiane.

On les utilise lorsque l'effectif total est élevé.

Les quartiles, déciles et centiles sont appelées *quantiles*.

Remarque : La détermination graphique des quantiles est analogue à celle de la médiane. On peut aussi les déterminer par un calcul d'interpolation linéaire.

e. Comparaison du mode, de la médiane et de la moyenne

Chacun des trois paramètres de position présente des avantages et des inconvénients et le choix dépend de l'usage que l'on veut en faire.

- Le **mode** est facile à déterminer. Il est défini même lorsque le caractère est qualitatif mais il apporte une information réduite sauf lorsqu'une valeur du caractère l'emporte nettement sur les autres.
- La **moyenne** donne une information plus riche qui permet certaines décisions. D'autre part, si on connaît la moyenne et l'effectif de deux populations, on peut en déduire la moyenne de la population totale. Mais la moyenne est influencée par les valeurs extrêmes.
- La **médiane** est facile à déterminer. Et, elle est très peu influencée par les valeurs extrêmes.

4-2. Caractéristiques de dispersion

Considérons les notes obtenues en mathématiques par deux élèves :

Elève A	10	11	05	12	17
Elève B	11	10	11	10	13

Elles ont même moyenne (11) et même médiane (11). Cependant elles diffèrent profondément, la première série est beaucoup plus dispersée que la deuxième. Ceci montre que la moyenne (ou tout autre paramètre de position) ne suffit pas pour caractériser une série statistique. Il est donc important de résumer une série statistique non seulement par des caractéristiques de tendance centrale, mais par des caractéristiques de dispersion. Un paramètre de dispersion est d'autant plus élevé que les valeurs de la série seront dispersées.

On en définira de deux sortes : celles liées à la moyenne : écart absolu moyen et écart type ; celles liées à la médiane : écart interquartile, écart interdécile, etc.

a. Etendue

Nous avons déjà défini l'étendue d'une série statistique. Ce paramètre est d'un intérêt limité car les valeurs extrêmes sont souvent accidentelles.

c. Ecart moyen

- L'écart de x_i à la moyenne est $|x_i - \bar{x}|$

- On appelle *écart moyen* d'une série statistique la moyenne des écarts à la moyenne.

$$e_m = \frac{\sum_{i=1}^N |x_i - \bar{x}|}{N}$$

Dans le cas de données regroupées où les valeurs $x_1, x_2, \dots, x_c$ ont des effectifs $n_1, n_2, \dots, n_c$

$$e_m = \frac{\sum_{i=1}^c n_i |x_i - \bar{x}|}{N}$$

Avantages et inconvénients :

L'écart moyen a l'avantage de faire intervenir tous les éléments de la série statistique, mais sa détermination nécessite beaucoup de calculs quand N est grand. C'est un paramètre peu utilisé.

d. Ecart interquantiles

On appelle *écart interquartile* le réel $Q_3 - Q_1$. Il représente l'étendue d'une partie de la série initiale centrée sur la médiane, et contenant 50% de l'effectif total.

Remarque : c'est un paramètre facile à déterminer. Il a l'avantage par rapport à l'étendue, d'écarter les valeurs accidentelles. Mais il ne renseigne pas sur la répartition des valeurs à l'intérieur de son domaine.

Ce sont des caractéristiques de dispersion.

De même, l'intervalle dans lequel se trouve 80% de la population (les 10% extrêmes étant éliminés des deux côtés) s'appelle écart interdécile : $D_9 - D_1$.

On emploie parfois d'autres caractéristiques. Par exemple :

L'écart semi-interquartile : $\frac{Q_3 - Q_1}{2}$

L'écart interquartile relatif : $\frac{Q_3 - Q_1}{M}$.

e. Variance. Ecart type

- Définition

- La *variance* d'une série est la moyenne des carrés des écarts à la moyenne.

$$V = \frac{\sum_{i=1}^c n_i |x_i - \bar{x}|^2}{\sum_{i=1}^c n_i}$$

- L'*écart type* d'une série est la moyenne quadratique des écarts à la moyenne. C'est la racine carrée de la variance.

$$\sigma = \sqrt{V}$$

Exemple : *calculez la variance et l'écart type des notes de l'élève A*

- Remarques.

Plus la variance est grande, plus la série est dispersée.

Plus la variance est petite (voisin de 0), plus la série est resserrée autour de la moyenne.

La variance est une quantité positive ou nulle.

- Méthode de calcul.

Avec des valeurs observées x_i très simples, il arrive souvent que la moyenne $\bar{x}$ soit un nombre décimal. Dans ce cas, le calcul de la variance V nécessite des calculs fastidieux.

On démontre que : $V = \frac{\sum n_i x_i^2}{N} - (\bar{x})^2$

Cette expression de la variance permet de simplifier les calculs.

- Présentation des calculs
Elève A

Elève B

x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$	x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$
5	1				10	2			
10	1				11	2			
11	1				13	1			
12	1								
17	1								
total					total				

Tableau 9.

On constate que l'élève A ayant des notes beaucoup plus dispersées que celles de l'élève B, l'écart type de A est nettement plus élevé que l'écart type de B.

Ainsi la variance et l'écart type renseignent sur l'étalement de la série de part et d'autre de la moyenne.

L'écart type mesure une grandeur de même nature que le caractère. La variance mesure le "carré" d'une telle grandeur.

- Propriétés de la variance

- La variance et l'écart type de la série $(x_i - a, n_i)$, où a est un réel, sont indépendants de a : ce sont respectivement la variance et l'écart type de (x_i, n_i) .

- On démontre que si V et σ sont respectivement la variance et l'écart type de (x_i, n_i) alors la série $(h x_i, n_i)$:

a pour variance $V' = h^2 V$,

et pour écart type $\sigma' = |h| \sigma$

- Avantages et inconvénients de la variance et de l'écart type

La variance et l'écart type sont deux bons paramètres de dispersion. Ils font intervenir toutes les valeurs observées. De plus, connaissant les variances de deux populations, on sait évaluer directement la variance de la population totale. Ceci explique que ce sont les paramètres de dispersion les plus employés.

On ne peut pas comparer les dispersions de 2 séries statistiques de même caractère qu'en utilisant dans les deux cas la même unité.

Remarque

Dans les études statistiques, on est amené souvent à considérer deux intervalles $[\bar{x} - \sigma; \bar{x} + \sigma]$ et $[\bar{x} - 2\sigma; \bar{x} + 2\sigma]$. Les pourcentages de valeurs (du caractère) se trouvant dans ces intervalles servent en particulier à la définition de normes techniques (appareils de mesure, machines industrielles ...)

e. Coefficient de variation

Les paramètres de dispersion que nous venons de définir dépendent tous de l'unité de mesure choisie pour les observations. Ceci est très gênant pour comparer deux séries qui n'utilisent pas la même unité de mesure (ex. : pour un groupe d'individus les poids sont-ils plus dispersés que la taille)

Pour éliminer l'influence de l'unité de mesure, on définit des coefficients de dispersion indépendants de l'unité de mesure. Le plus usuel est le coefficient de variation qui est défini

par $\frac{\sigma}{\bar{x}}$. Ce nombre est souvent exprimé en pourcentage.