Table des matières

Les méthodes sont les habitudes de l'esprit et les économies de la mémoire. Rivarol

I- De quoi s'agit-il?

La statistique est l'ensemble des analyses et des outils mathématiques permettant de déterminer les caractéristiques d'un ensemble de données.

On a principalement:

- la collecte des données ;
- le traitement des données collectées
- l'interprétation des données.

On explore d'abord les données pour avoir une idée de leurs caractéristiques, puis on fait des hypothèses de comportement. (Hypothèses qui seront à confirmer ou à infirmer ...)
Dans le cours de première, on commence par préciser le vocabulaire, puis, on effectuera des calculs sur des données déjà collectées, on définira différents outils d'observation et leurs propriétés.

II- Les rappels:

Les notions sont celles de seconde (et du collège) et concernent les séries statistiques à une variable...

II-1 Savoir le vocabulaire descriptif:

II-1-1- population:

La population est l'ensemble des éléments ou individus sur lequel porte l'étude statistique, sur lequel on effectue des observations, des classements....

II-1-2- échantillon:

Un échantillon est une partie ou un sous-ensemble de la population.

Remarque: le choix de l'échantillon est important pour la validité de l'étude statistique.

II-1-3- caractère:

Le caractère est le sujet sur lequel porte l'étude statistique. Un caractère prend différentes valeurs.

Remarque: Pour une même population, on peut définir de nombreux caractères

II-1-4- effectif:

L' effectif est le nombre d'individus prenant une ou des valeurs du caractère

II-1-5-Série statistique qualitative:

Une série statistique est qualitative lorsque le caractère prend des valeurs non numériques.

Exemples: la couleur des yeux, le modèle de voitures.

II-1-6- Série statistique quantitative:

Une série statistique quantitative lorsque les valeurs prend des valeurs numériques

Exemples: la taille des personnes dans un groupe, le nombre de pièces par appartement

II-1-6-1- Série statistique quantitative discrète:

Une série statistique quantitative est discrète lorsque les valeurs prises par le caractère étudié sont isolées.

Exemple: le nombre de pièces par appartement

II-1-6-2- Série statistique quantitative continue:

Une série statistique quantitative est continue lorsque le caractère peut prendre toutes les valeurs d'un intervalle.

Exemple: la taille des personnes dans un groupe

II-2- savoir le vocabulaire et calculer:

II-2-1- L'étendue d'une série:

L'étendue d'une série est donnée par la différence entre les valeurs extrêmes prises par le caractère.

étendue = valeur maximale – valeur minimale

II-2-2- le mode:

Le mode est la valeur du caractère ayant le plus fort effectif. (C'est la valeur du caractère la plus fréquente. Une série statistique peut avoir plusieurs modes)

II-2-3- la moyenne: (moyenne pondérée)

La moyenne (moyenne pondérée) est la somme de toutes les valeurs (comptées autant de fois qu'elle apparaît dans la série) divisée par l'effectif total. On a: $\bar{x}=\frac{n_1\times x_1+n_2 x_2+\ldots+n_p\times x_p}{n_1+n_2+\ldots+n_p}$ où les x_i représentent les valeurs prises par le caractère et les n_i leur effectif

II-2-4- la médiane:

La médiane est la valeur qui partage la population en deux parties égales: 50% de l'effectif prend une valeur du caractère inférieure ou égal à la médiane et 50% de l'effectif prend une valeur supérieure ou égale à la médiane.

Pour calculer la médiane, on range la série par ordre croissant des valeurs x_i et on a:

si l'effectif total N est impair, la médiane est donnée par la valeur de rang $\frac{N+1}{2}$

si l'effectif total N est pair, la médiane est donnée par la demi-somme (moyenne arithmétique) des valeurs de rang $\frac{N}{2}$ et $\frac{N}{2}+1$.

Complément: On définit aussi les ***quartiles,*** les ***déciles***, ... selon le même procédé. Pour le premier quartile, on partage les effectifs de la série en 25% et 75% et la valeur correspondante du caractère est le premier quartile.

Le second quartile est la médiane

Le troisième quartile est obtenu en partageant les effectifs de la série en 75% et 25%

Pour les déciles (il y a neuf déciles), on procède de même tous les 10%

II-2-5- fréquence:

La fréquence d'une valeur est le quotient de l'effectif de cette valeur par l'effectif total.

On note $f_i=\frac{n_i}{N}=\frac{\text{effectif de la valeur } x_i}{\text{effectif total}}$

Remarquer: la moyenne $\bar{x}=f_1\times x_1+\ldots+f_p\times x_p$

II-2-6- fréquence cumulée croissante

La fréquence cumulée croissante est la somme des fréquences **jusqu'à** la valeur donnée incluse.

On classe les valeurs x_i dans l'ordre croissant; $f_1+f_2+\ldots+f_k$ donne la fréquence d'individus prenant une valeur inférieure ou égale à la valeur x_k

II- 3- savoir le vocabulaire et construire:

II-3-1- diagramme en bâtons:

Le diagramme en bâtons est utilisé lorsque les valeurs sont discrètes ou qualitatives.

En abscisses, on place les valeurs du caractère, en ordonnées les effectifs.

Un bâton représente alors l'effectif de la valeur.

La **hauteur** du bâton est **proportionnelle** à l'effectif (ou à la fréquence) de la valeur du caractère représenté.

II-3-2- diagramme circulaire:

Le diagramme circulaire est construit en créant des secteurs circulaires dans un disque de façon à ce que les **angles** des secteurs soient **proportionnels** aux effectifs (ou aux fréquences)

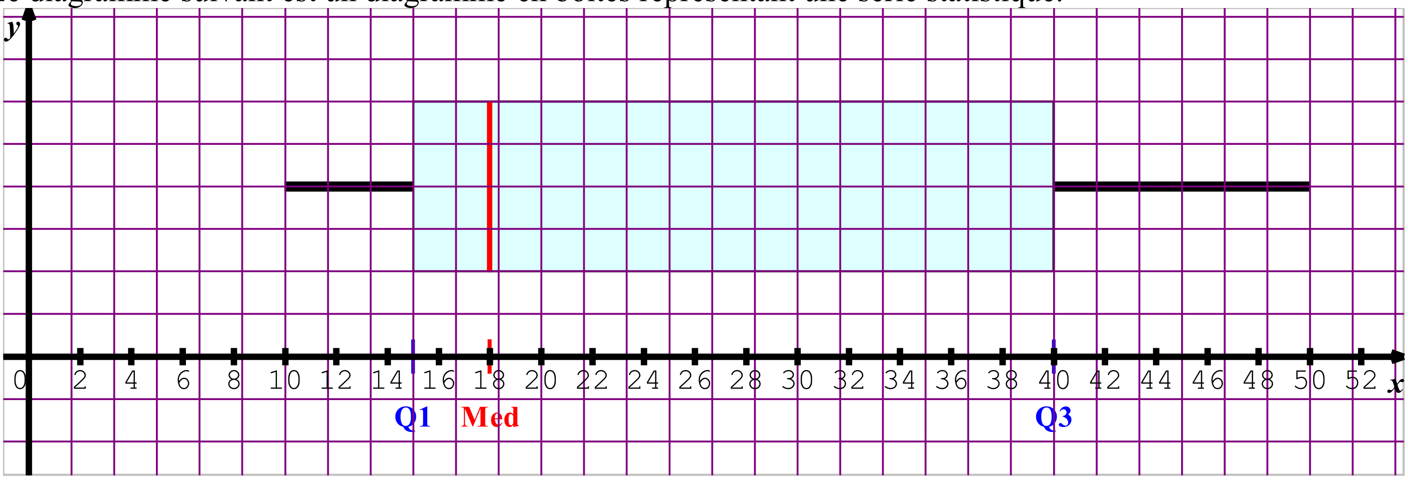
II-3-3- histogramme:

Un histogramme est construit en juxtaposant des rectangles de façon à ce que les **aires** des rectangles soient **proportionnelles** aux effectifs (ou aux fréquences)

III- Diagrammes en boîtes (à moustaches ou à pattes)

III-1- Un exemple

Le diagramme suivant est un diagramme en boîtes représentant une série statistique.



Sur ce diagramme apparaît la valeur minimale (ici: 10), la valeur maximale (ici: 50) prises par le caractère, la valeur médiane (ici: 18) de la série et les deux quartiles Q_1 (ici: 15) et Q_3 (ici: 40).
Par définition de ces quartiles, on sait que 50 % de l'effectif prend des valeurs situés entre ces deux quartiles.
Il reste 25 % de l'effectif prenant une valeur inférieure à Q_1 et 25% de l'effectif prenant une valeur supérieure à Q_3 .

III-2- Les calculs

La série statistique représentée par le diagramme en boîtes du §III-1 est définie par les données (récoltées) suivantes:

Valeur	10	17	15	50	40	18	30
Effectif	3	5	10	2	9	7	6

III-2-1- Médiane

III-2-1-1- Calcul de la médiane

On commence par ordonner la série dans l'ordre croissant des valeurs du caractère.
On détermine alors la valeur (du caractère) médiane qui partage la série en deux parties de même effectif.
Dans cet exemple, l'effectif total vaut 42.
Comme l'effectif est pair, on fait la moyenne entre la 21ème et la 22ème valeur.
Série ordonnée:

Valeur	10	15	17	18	30	40	50
Effectif	3	10	5	7	6	9	2
Effectif cumulé	3	13	18	25	31	40	42

La 21ème valeur est 18 et la 22ème valeur est 18: la médiane M = $\dfrac{18+18}{2}$ = 18

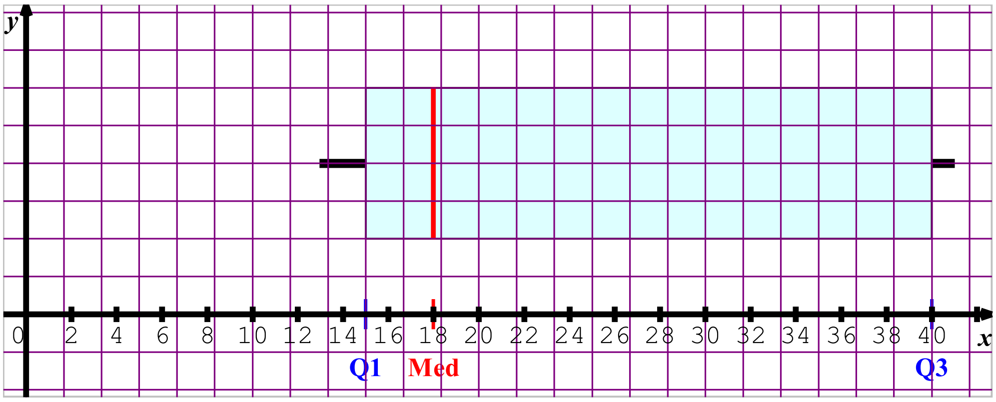
III-2-1-2- Propriété de la médiane

On ne change pas la valeur de la médiane en modifiant les valeurs extrêmes du caractère.

Par exemple, la série

Valeur	13	15	17	18	30	40	41
Effectif	3	10	5	7	6	9	2

a la même valeur médiane que la précédente.
Le diagramme en boîtes:



III-2-2- Les quartiles.

III-2-2-1- Calculs des quartiles

Le premier quartile Q_1 est défini par la valeur du caractère tel que 25 % de l'effectif au moins prend une valeur inférieure ou égale à Q_1 .
Le quart de 42 est 10,5. On prend donc la 11ème valeurs, soit $Q_1 = 15$
Le troisième quartile Q_3 est défini par la valeur du caractère tel que 75 % de l'effectif au moins prend une valeur inférieure ou égale à Q_3 .
Les trois-quarts de 42 font 31,5. On prend donc la 32ème valeur: $Q_3 = 40$

III-2-2-2- Intervalle interquartile

L'intervalle $]Q_1 ; Q_3[$ est l'intervalle interquartile

III-2-2-3- Écart interquartile

Le nombre $Q_3 - Q_1$ est l'écart interquartile.

IV-Histogramme

IV-1- Définition

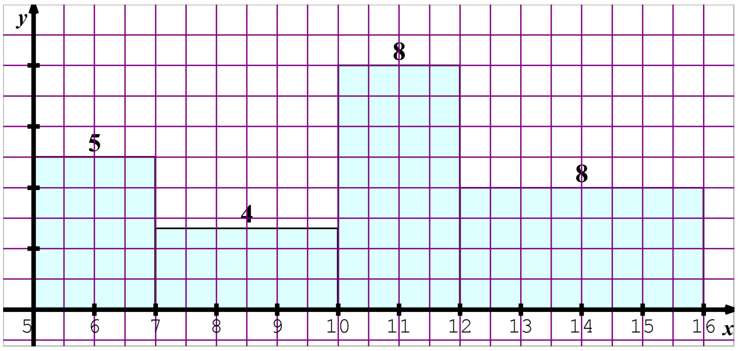
Un histogramme est construit en juxtaposant des rectangles de façon à ce que les **aires** des rectangles soient **proportionnelles** aux effectifs (ou aux fréquences)

IV-2- Exemple

On donne la série statistique suivante

Classe	[5;7[	[7;10[	[10;12[	[12;16]
Effectifs	5	4	8	8

L'histogramme de cette série



Les calculs:

Classe	[5;7[	[7;10[	[10;12[	[12;16]
Effectifs	5	4	8	8
Étendue de la classe	2 = 4 carreaux	3=6 carreaux	2 = 4 carreaux	4= 8 carreaux
Aire en petits carreaux	20 (lecture sur le graphique)	16 (proportionnalité)	32 (proportionnalité)	32 (proportionnalité)
hauteur du rectangle (en carreau)	5	$\frac{16}{6}=\frac{8}{3}$	$\frac{32}{4}=8$	$\frac{32}{8}=4$

V- Moyennes mobiles. Lissage d'une série chronologique

V-1- Objectif

Une série chronologique est une série de valeurs d'un caractère observé à des dates périodiques (chaque jour ou chaque mois ou chaque année ...).

Les caractères peuvent avoir des variations très irrégulières mais sur le long terme, on peut observer une tendance plus régulière.

Pour dégager cette tendance, une méthode consiste à lisser la série chronologique pour atténuer les fluctuations brusques par la méthode des moyennes mobiles.

On lisse la série chronologique par les moyennes mobiles d'ordre 3 en remplaçant dans la série les valeurs par la moyenne sur trois dates consécutives.

Pour une série de *n* valeurs, on remplace les x_i à partir de x_2 jusqu'à x_{n-1} par la moyenne

$$x'_i = \frac{x_{i-1}+x_i+x_{i+1}}{3}$$

On "perd" les deux valeurs extrêmes.

On peut faire le lissage par moyennes mobiles d'ordre 5 sur le m^me principe.

Pour une série de *n* valeurs, on remplace les x_i à partir de x_3 jusqu'à x_{n-2} par la moyenne

$x'_i = \frac{x_{i-2}+x_{i-1}+x_i+x_{i+1}+x_{i+2}}{5}$. On "perd" alors les quatre valeurs extrêmes.

V-2- Un exemple

Voici une série chronologique donnant l'indice d'un bien de consommation

année	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
indice	105	102	102	106	114	119	115	114	116	114

année	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
indice	118	125	129	121	122	126	130	134	125	132

Lissage de la série par moyennes mobiles d'ordre 3. (Utilisation d'un tableur)

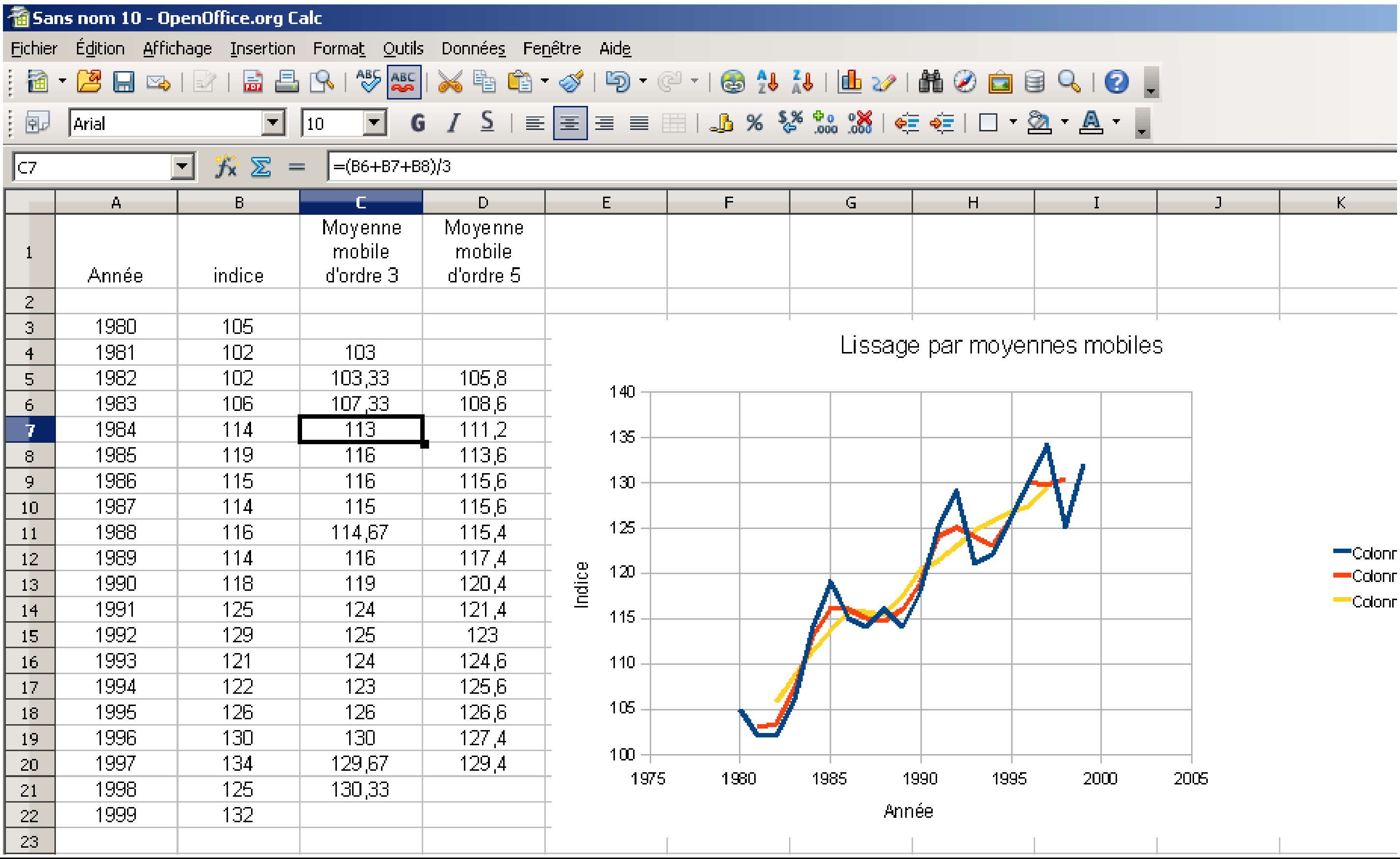
Pour 1981: $x'_2 = \dfrac{105+102+102}{3} = 103$, pour 1982: $x'_3 = \dfrac{102+102+106}{3} \approx 103{,}33$

année	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
moyenne		103	103,33	107,33	113	116	116	115	114,67	116

année	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
moyenne	119	124	125	124	123	126	130	129,67	130,33	

V-3- Représentation graphique

La représentation graphique fait apparaître une tendance à l'augmentation de ce bien de consommation. Le lissage par moyennes mobiles d'ordre 5 montre encore davantage cette tendance.



Année	indice	Moyenne mobile d'ordre 3	Moyenne mobile d'ordre 5
1980	105		
1981	102	103	
1982	102	103,33	105,8
1983	106	107,33	108,6
1984	114	113	111,2
1985	119	116	113,6
1986	115	116	115,6
1987	114	115	115,6
1988	116	114,67	115,4
1989	114	116	117,4
1990	118	119	120,4
1991	125	124	121,4
1992	129	125	123
1993	121	124	124,6
1994	122	123	125,6
1995	126	126	126,6
1996	130	130	127,4
1997	134	129,67	129,4
1998	125	130,33	
1999	132		

VI- Variance- Écart-type

la **variance** et l'**écart-type** servent à caractériser la dispersion des valeurs autour de la moyenne

VI-1- Définition de la variance et calcul de la variance

La **variance** est la moyenne des carrés des écarts à la moyenne

On calcule la moyenne $\overline{x}$, puis, on fait l'écart $x_i - \overline{x}$ au carré pour chaque valeur x_i.

Chaque $(x_i - \overline{x})^2$ a le même effectif n_i que la valeur x_i.

On calcule la moyenne de ces écarts au carré.

Un exemple de calcul

On a relevé les notes d'un devoir et obtenu le tableau suivant:

notes x_i	5	7	12	13	15	16	Total
effectif n_i	3	4	2	3	4	2	18
$n_i \times x_i$	15	28	24	39	60	32	198
$e_i^2 = (x_i - \overline{x})^2$	36	16	1	4	16	25	XXXXXX
$n_i \times e_i^2$	108	64	2	12	64	50	300

La moyenne vaut: $\overline{x}$ = 198/18 = 11

La variance vaut: V = 300/18 = $\frac{50}{3}$ ≈ 16,67

Formules

La formule pour calculer la variance V d'après la définition est:

$$V = \frac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + \ldots + n_p(x_p - \overline{x})^2}{N}$$

Il existe une autre formule qui se déduit de la précédente en développant:

$$V = \frac{n_1 x_1^2 + n_2 x_2^2 + \ldots + n_p x_p^2}{N} - \overline{x}^2$$

Application de la deuxième formule:

On reprend l'exemple précédent

notes x_i	5	7	12	13	15	16	Total
effectif n_i	3	4	2	3	4	2	18
$n_i \times x_i$	15	28	24	39	60	32	198
$n_i \times x_i^2$	75	196	288	507	900	512	2478

La moyenne vaut: $\overline{x}$ = 198/18 = 11

La variance vaut: V = 2478/18 – 11² = $\frac{413}{3}$ – 121 = $\frac{50}{3}$

VI-2- Définition de l'écart-type et calcul

L'écart-type σ d'une série statistique est la racine carrée de la variance.

L'écart-type est mesuré dans la même unité que la variable.

$\sigma = \sqrt{V}$.

Dans l'exemple précédent, on a:

L'écart-type σ vaut: $\sigma = \sqrt{\frac{50}{3}} \approx 4{,}08$

VI-3- Lien avec la moyenne: propriété de la moyenne

On prend dans ce paragraphe toutes les valeurs de la variable une à une.

On a: x_1, x_2, ..., x_N où N est l'effectif total. (Certaines valeurs peuvent être répétées)

La variance est définie par $V = \frac{f(\overline{x})}{N}$

où f est la fonction définie par $f(x) = (x_1 - x)^2 + (x_2 - x)^2 + ... + (x_N - x)^2$

Cette somme est la somme des carrés des distances d'un nombre x aux N valeurs x_1, x_2, ..., x_N de la variable statistique.

Propriété:

La moyenne $\overline{x}$ est la valeur pour laquelle f atteint son minimum.

ou encore,

Pour tout x réel, $f(\overline{x}) \leqslant f(x)$

VII- Effet de structure

Il faut être prudent avec les statistiques

Exemple 1

Dans une entreprise de 100 employés, on a en l'année 2008:

10 agents d'entretien au salaire mensuel moyen de 1 000 €, 70 ouvriers spécialisés au salaire mensuel moyen de 1 500 € et 20 cadres au salaire mensuel moyen de 2 500€.

Du fait de la conjoncture, l'entreprise licencie en 2009 les agents d'entretien et sous-traite l'entretien à une entreprise spécialisée, garde les ouvriers et les cadres mais baisse leur salaire de 50€ par mois.

Calculer le salaire moyen en 2008, en 2009.

Exemple 2

Dans une entreprise de 100 employés, on a en l'année 2008:

10 agents d'entretien au salaire mensuel moyen de 1 000 €, 70 ouvriers spécialisés au salaire mensuel moyen de 1 500 € et 20 cadres au mensuel salaire moyen de 2 500€.

Du fait de la conjoncture, l'entreprise se réorganise:

l'entreprise sous-traite l'entretien à une entreprise spécialisée et pour ne pas licencier de personnels, les agents d'entretien passent dans la catégorie des ouvriers spécialisés, et, 5 ouvriers spécialisés deviennent des cadres.

Le salaire des ouvriers baisse de 50 € par mois et celui des cadres de 100 € par mois.

Calculer le salaire moyen en 2008, en 2009.

Commentaires:

Malgré la baisse des salaires, le salaire moyen a augmenté. Ce phénomène est dû à la structure de l'entreprise qui a une forte influence sur le calcul de la moyenne.

VIII- Tableau à double entrée: Fréquence conditionnelle

Exemple:

En France en 1999, 36 % des ménages étaient équipés d'un lave-vaisselle et 62 % d'un magnétoscope. 25% sont équipés de ces deux biens.

Tableau à double-entrée

	Lave-vaisselle	Sans lave-vaisselle	Total
Magnétoscopes	25	37	62
Sans magnétoscopes	11	27	38
Total	36	64	100

La fréquence des personnes équipées d'un lave-vaisselle sachant qu'ils possèdent un magnétoscope est $\frac{25}{62}$

La fréquence des personnes équipées d'un magnétoscope sachant qu'ils possèdent un lave-vaisselle est $\frac{25}{36}$

La fréquence des personnes équipées d'un magnétoscope sachant qu'ils ne possèdent pas de lave-vaisselle est $\frac{37}{64}$

La fréquence des personnes ne possédant pas de magnétoscope sachant qu'ils possèdent un lave-vaisselle est: $\frac{11}{36}$

On note A l'ensemble des personnes possédant un lave-vaisselle, B celui des personnes possédant un magnétoscope.

$\overline{A}$ est l'ensemble des personnes ne possédant pas de lave-vaisselle et $\overline{B}$ celui des personnes ne possédant pas de magnétoscopes.

$f_B(A)$ est la fréquence des personnes équipées d'un lave-vaisselle sachant qu'ils possèdent un magnétoscope.

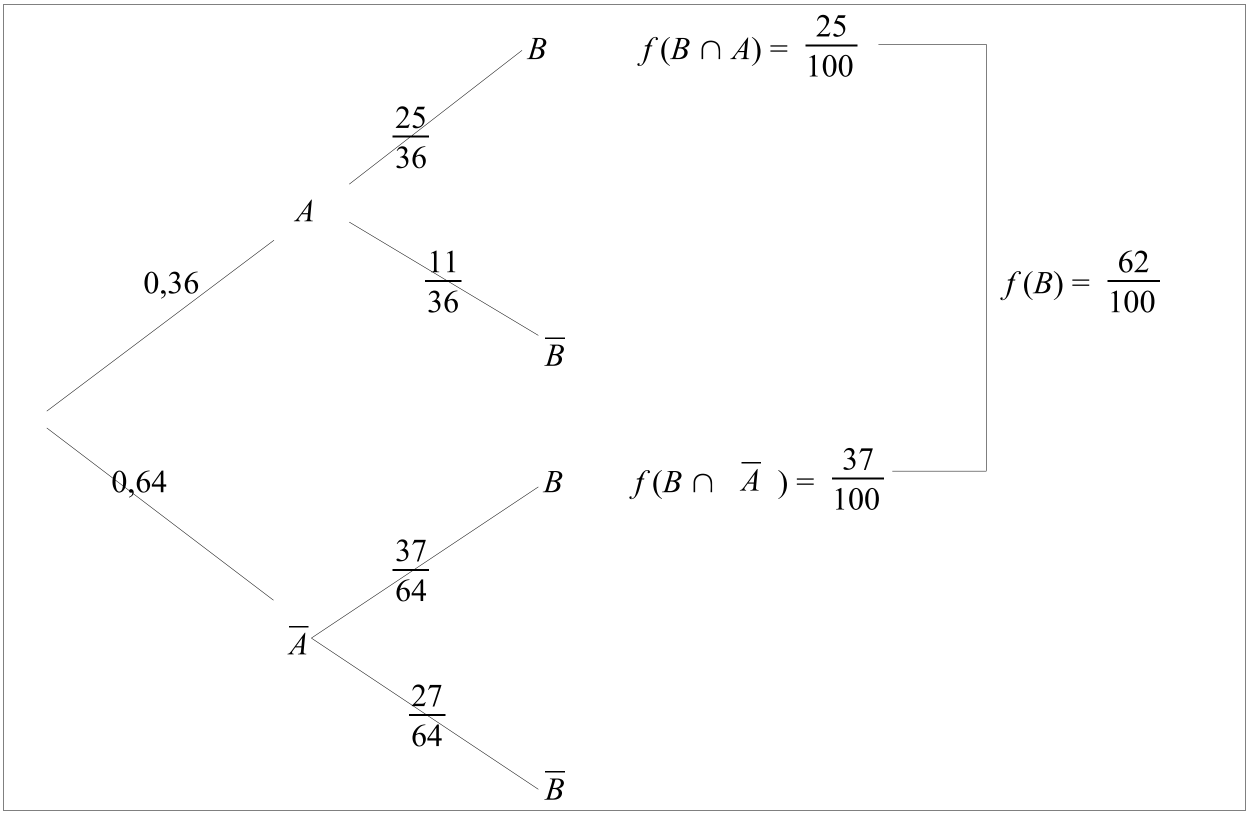
Définition:

Soient deux caractères A et B étudiés sur une même population P.

La fréquence de A sachant B est la proportion des individus ayant les deux caractères A et B par rapport à ceux possédant le caractère B.

On note: $f_B(A) = \frac{\text{effectif de } A \text{ et } B}{\text{effectif de } B}$

Arbre pondéré



IX- Résumé

IX-1 Diagrammes

IX-1-1- Diagrammes en bâtons

On utilise ces diagrammes pour représenter des séries discrètes.

La hauteur du bâton est proportionnelle à l'effectif ou à la fréquence.

IX-1-2- Diagrammes en boîtes

On utilise ces diagrammes pour représenter des séries où les valeurs extrêmes ont peu d'influence.

Ils permettent de comparer des séries en mettant en évidence leurs valeurs médianes et la dispersion autour de la médiane grâce à l'écart inter-quartile.

IX-1-3- Histogrammes

On utilise ces diagrammes pour représenter des séries continues réparties en classe.

L'aire des rectangles est proportionnelle à l'effectif ou à la fréquence.

IX-2- Les paramètres

IX-2- 1- Les valeurs centrales

IX-2- 1- 1- La moyenne

La moyenne est la valeur du caractère telle que si tous les individus d'une population statistique prennent la même valeur alors la somme totale de toutes ces valeurs est inchangée.

N effectif total: $N\ \overline{x} = n_1 x_1 + n_2 x_2 + \ldots + n_k x_k$

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

IX-2- 1- 2- La médiane

Le nombre d''individus dont le caractère statistique est inférieur à la médiane *Me* est égal au nombre d'individus dont le caractère statistique est supérieur à *Me*.

Le diagramme des fréquences cumulées permet de déterminer rapidement la valeur médiane en prenant en abscisse la valeur du caractère qui correspond à 50% de la fréquence cumulée.

IX-2- 2- Les paramètres de dispersion

IX-2- 2- 1- L'écart-type

L'écart-type qui est la racine carrée de la variance est un outil pour mesurer la dispersion autour de la moyenne

$$V = \frac{n_1\left(x_1-\overline{x}\right)^2+n_2\left(x_2-\overline{x}\right)^2+\ldots+n_p\left(x_p-\overline{x}\right)^2}{N} = \frac{n_1 x_1^2+n_2 x_2^2+\ldots+n_p x_p^2}{N} - \overline{x}^2$$

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

IX-2- 2- 2- L'écart interquartile

L'écart interquartile permet de situer 50% de l'effectif autour de la médiane.

La médiane partage la série en deux sous-séries.

Le premier quartile est la médiane de la première sous-série et le troisième quartile est la médiane de la seconde sous-série.