République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de BEJAÏA

Faculté de Technologie

Département des MINES ET GEOLOGIE

Deuxième année

Par : **Madani BEZOUI** [1]

Cours de Maths 310 : Probabilité et Statistiques

Année Universitaire 2011-2012

1. madani.bezoui@gmail.com

Table des matières

Chapitre 1

Série statistique à une variable

1.1 Quelques definitions

Soient une population à "n" individus ayant les modalités $x_1, x_2, ..., x_k$, $k \leq n$

Définition 1.1 (Effectif (fréquence absolue)) noté n_i : **nombre** d'individus ayant pris la modalité x_i

Définition 1.2 (fréquence (fréquence relative)) noté f_i : **proportion** d'individus ayant pris la modalité x_i

$f_i = \frac{n_i}{n}$ en pourcentage : $f_i\% = \frac{n_i}{n} * 100$

$\sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_k = n$, $\sum_{i=1}^{k} f_i = f_1 + f_2 + ... + f_k = 1$

[H]

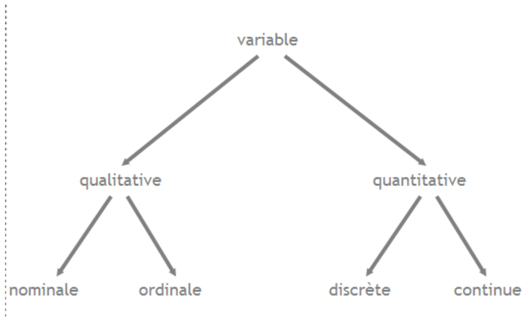


FIGURE 1.1 – Schéma récapitulatif des variables

Quelques définitions

Définition 1.3 (Série statistique) Donnée de toutes les observations

Définition 1.4 (distribution statistique) Donnée du couple $\{x_i, n_i\}_{i=\overline{1,k}}$ ou $\{x_i, f_i\}_{i=\overline{1,k}}$

1.2 I.1.2. Représentations graphiques

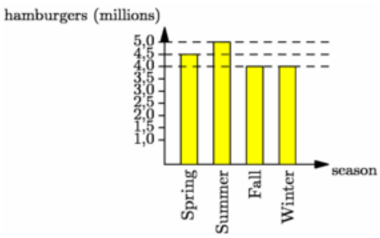
1.2.1 Cas de variables qualitative

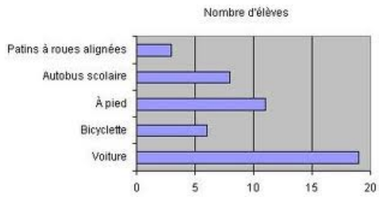
Diagramme circulaire (Cas qualitatif)

$$\alpha_i{}^\circ = f_i \times 360$$

Par exemple, sur la figure, pour trouver l'angle correspondant à la modalité "Hiver", on doit procéder de cette manière : $\alpha_{hiver} = 0.3 * 360 = 108°$

Tuyaux d'orgue ou Diagramme à bondes (Cas qualitatif)



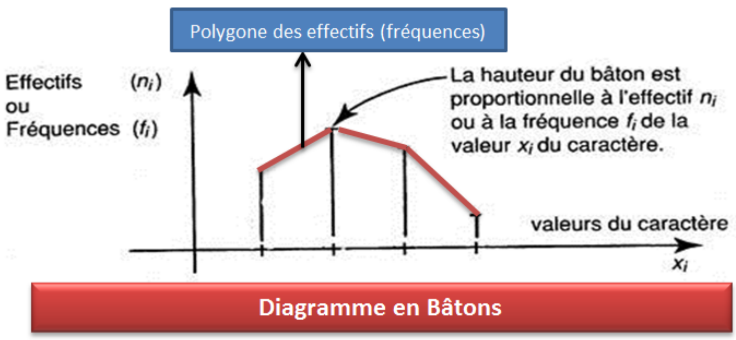


Le graphe à gauche est le plus utilisé, il suffit de trouver les fréquences ensuite de fixer l'echelle du graphe et de faire correspondre la fréquence des variables à la hauteur, (l'axe des ordonées), tandis qu'à l'axe des abscisses on attribut les modalités.

1.2.2 Cas de Série statistique Discrète

Diagramme en Bâtons (V. Quantitative Disrètes)

Pour chaques modalité x_i on représente un ségment de droite proportionnel aux effectifs (n_i) ou aux fréquences (f_i)



Remarques

Remarque 1 Le graphe a la même allure, si on représente les fréquences au lieu des effectifs en ordonnées.

Remarque 2 Les polygones d'effectifs (ou de fréquence) s'obtient on liant la somme des bâtonnets.

1.2.3 Cas de Série statistique Continue

La variable statistique continue (Rappel)

Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe.

En règle générale, on choisit des classes de même amplitude. Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille n.

— La règle de **STURGE** : k=Nombre de classes $= 1 + (3,3log(n))$,

— La règle de **YULE** : k=Nombre de classes $= 2,5.n^{\frac{1}{4}}$.

Si on a k classe $[e_0, e_1[, [e_1, e_2[, ..., [e_{k-1}, e_k[$, le centre x_i de la classe $c_i = [e_{i-1}, e_i[, i = \overline{1,k}$ est :

$$x_i = \frac{e_{i-1} + e_i}{2} = e_{i-1} + \frac{a_i}{2}$$

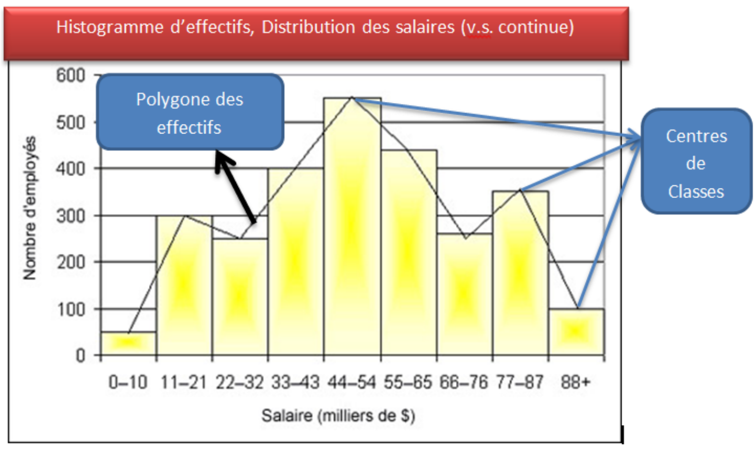
tel que

— e_{i-1} : extrémité inférieure (borne gauche de la classe c_i).

— e_i : extrémité superieure (borne droite de la classe c_i).

— a_i est l'amplitude de la i^{eme} classe, elle est égale à $e_i - e_{i-1}$.

Représentation graphique : Histogramme



La hauteur des barre $h_i = n_i$.

Représentation graphique : Histogramme (suite : Remarques)

Remarque 3 L'aire de l'histogramme est égale à la somme des aires des rectangles

$$Aire = \sum_{i=1}^{k}(a.f_i) = a.\sum_{i=1}^{k}(f_i) = a$$

L'histogramme est la représentation graphique d'une variable continue. A chaque classe de la variable, correspond la surface d'un rectangle qui a pour base l'amplitude de classe. Comme c'est la surface des rectangles qui représente les phénomènes étudiés, on remarque que :

— si les amplitudes sont égales, alors les hauteurs des rectangles sont proportionnelles aux effectifs ou aux fréquences.

— si les amplitudes sont inégales, il faudra corriger la hauteur des rectangles de manière à ce que leur surface corresponde bien à n_i (les effectifs) ou f_i (les fréquences).

1.2.4 Cas de d'amplitude de classes inégales

Exemple : Cas de classes d'amplitudes inégales

Exemple de l'étude de la taille en cm dans une classe de terminale :

Classes x_i	Effectifs n_i	Fréquences f_i %	Amplitudes a_i
[150 – 160[	3	8,57 %	10
[160 – 170[	8	22,86 %	10
[170 – 175[	13	37,14 %	5
[175 – 180[	7	20 %	5
[180 – 200[	4	11,43 %	20
total	35	100 %	

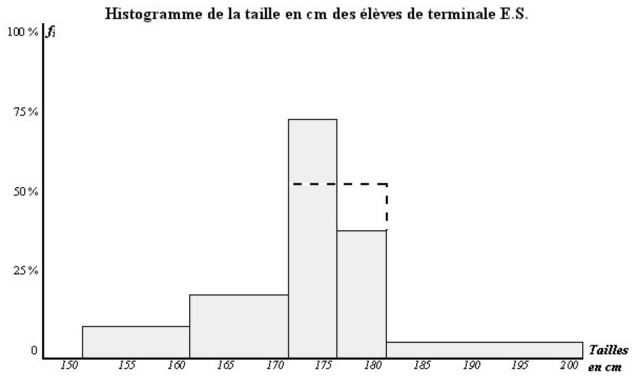
L'amplitude de la classe [170 - 175[est de 5. Si l'amplitude de référence est 10, il faudra multiplier la hauteur par deux puisque la base de cette clase est égale à la moitié de l'amplitude de référence.

Pour simplifier le calcul, on peut ajouter deux colonnes au tableau de départ (Amplitude de référence = 10) :

Classes x_i	Effectifs n_i	Fréquences f_i %	Amplitudes a_i	c = coefficient correcteur = 10 /a_i	Hauteur corrigée $h_i = f_i . c$
[150 – 160[	3	8,57 %	10	1	8,57
[160 – 170[	8	22,86 %	10	1	22,86
[170 – 175[	13	37,14 %	5	2	74,28
[175 – 180[	7	20 %	5	2	40
[180 – 200[	4	11,43 %	20	0,5	5,57
total	**35**	**100 %**			

Histogramme : Cas de classes d'amplitudes inégales

Pratiquement, on prend $h_{corige} = \frac{74.28+40}{2} = 57.14$. comme sur cette figure :



L'aire de l'histogramme est conservé après correction des fréquences.

1.2.5 Fonction de répartition et courbes cumulative

Fonction de Répartition (Cas Discret)

Soit une variable statistique discrète ayant k modalité $x_1, x_2, ..., x_k$ données par ordre croissant. La fonction de répartition estt définie de R dans [0,1], comme suitt :

$$F(x) = \begin{cases} 0 & x < x_1; \\ f_1 & x_1 \leq x < x_2; \\ f_1 + f_2 & x_2 \leq x < x_3; \\ \vdots & \vdots \\ f_1 + f_2 + ... + f_i & x_2 \leq x < x_3; \\ \vdots & \vdots \\ f_1 + f_2 + ... + f_k = 1 & x \geq x_k. \end{cases}$$

On remarque que cette la fonction F est :

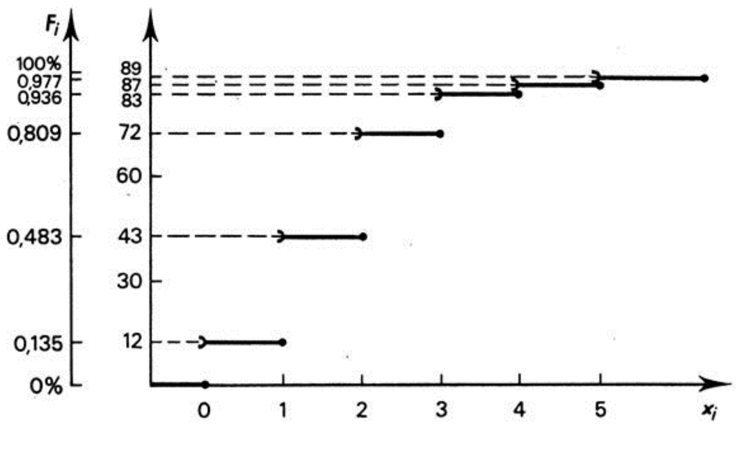
— Positive : $0 \leq F(x) \leq 1$,

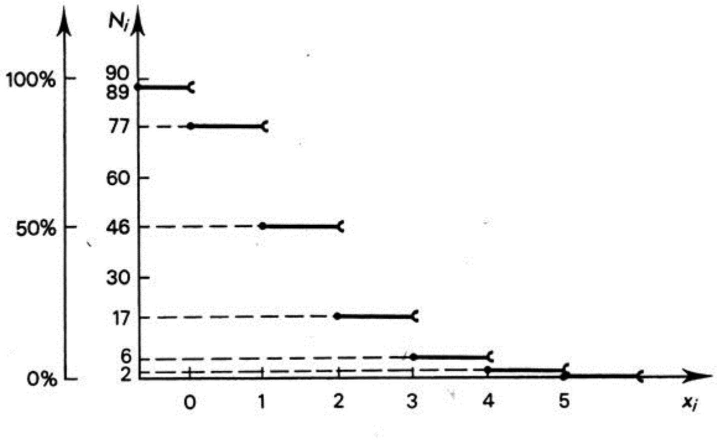
— Monotone (croissante), Discontinue aux points $x_1, x_2, ..., x_k$.

— $\lim_{x \to -\infty} = 0$, $\lim_{x \to +\infty} = 1$

Courbe commutative (Pour le cas discret)

C'est une courbe en secalier comportant des palliers horizontaux. Il y a similitude entre la courbe commutative et la courbe de la fonction de répartition.

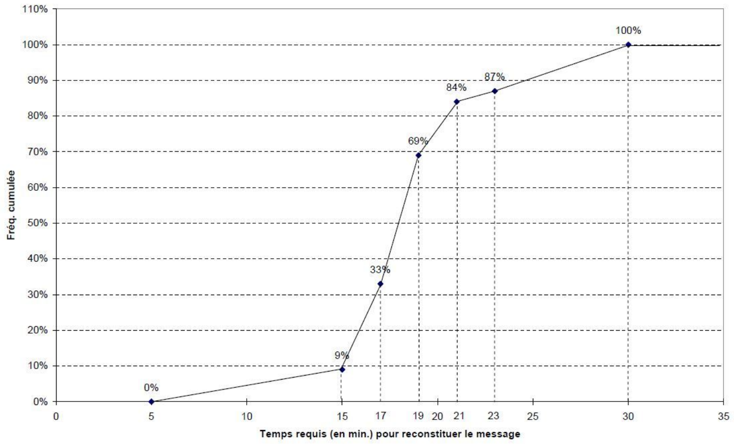




Fonction de répartition (Cas continu)

$$F(x) = \begin{cases} 0 & x < e_0; \\ F(e_1) = f_1 & ; \\ F(e_2) = f_1 + f_2 & ; \\ \vdots & \\ F(e_3) = f_1 + f_2 + ... + f_i & ; \\ \vdots & \vdots \\ F(e_k) = f_1 + f_2 + ... + f_k = 1 & \\ F(x) = 1 & si x \geq e_k \end{cases}$$

on a présenté un message mutilé à 60 personnes et mesuré le temps nécessaire à la reconstitution exacte du message. Les résultats (compris entre 5 et 30 minutes) de ces 60 personnes ont donné lieu à la courbe cumulative des fréquences suivantes :



Le pourcentage de personnes qui ont mis strictement plus de 21 min. est donné par 100% -84%=16%.

Exemple 1

Classes	n_i	f_i	$N_i \uparrow$	$N_i \downarrow$	$F_i \uparrow$	$F_i \downarrow$
$[61, 65[$	4	0.2	4	20	0.2	1
$[65, 69[$	4	0.2	8	16	0.4	0.8
$[69, 73[$	5	0.25	13	12	0.65	0.6
$[73, 77[$	3	0.15	16	7	0.8	0.35
$[77, 81[$	4	0.2	20	4	1	0.2
TOTAL	20	1				

1.3 Mode

Le mode et la classe modale (définitions)

Noté "M_o",

Cas discret V.S qui a le plus grand effectif (ou fréquence),

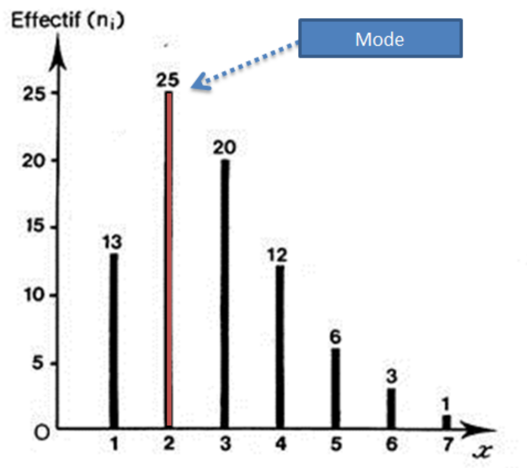
Cas Continu (On parle de classe modale), c'est la classe qui a le plus grand effectif (ou fréquence), le mode se calcule de manière approchée :

— Le mode représente le centre de classe modale. $M_o = x_i$,

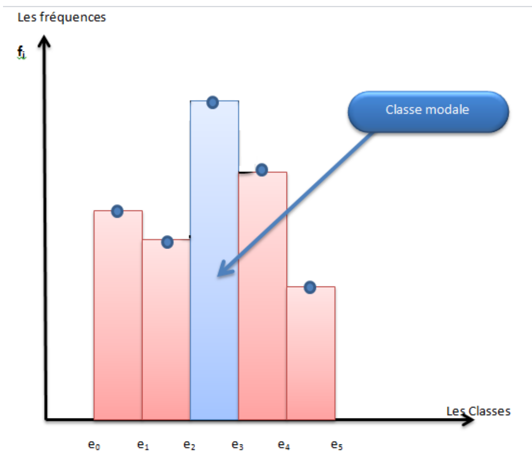
— $M_o = e_{i-1} + a_i \frac{\Delta_1}{\Delta_1 \Delta_2}$, tq :

- Δ_1 : L'excès de la classe modale par rapport à la classe précédente.
- Δ_2 : L'excès de la classe modale par rapport à la classe précédente.

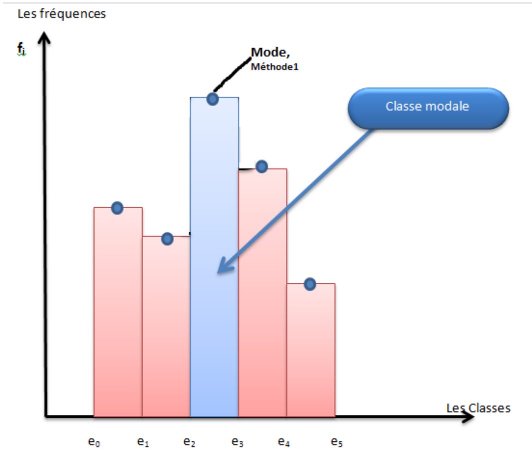
Mode (Cas discret)



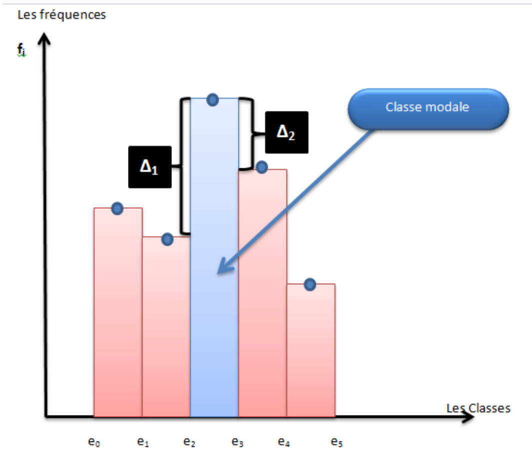
Classe Modale (Cas continu)

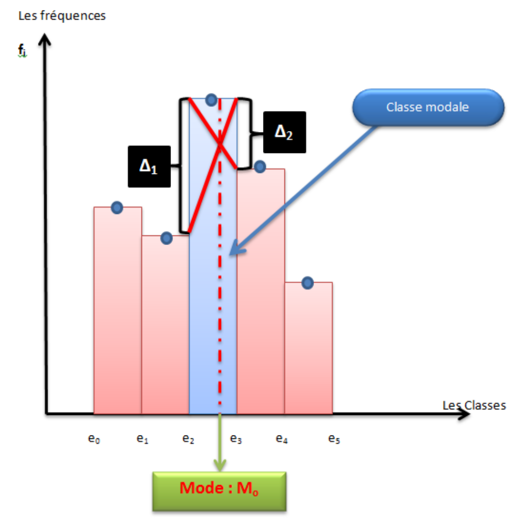


Classe Modale (Cas continu, Méthode N°1)



Classe Modale (Cas continu, Methode N°2)





Remarque 4 Une série peut être unimodale, bimodale ou multimodale

1.4 Médiane

Médiane (définition)

Soit X, une variable statistique ayant k modalité ordonnée par ordre croissant : $x_1, x_2, ..., x_k$.

Définition 1.5 La médiane, noté M_e, est la valeur qui partage une série ordonnée, en deux parties d'effectif égale. En générale, M_e est la v.s, telle que, $F(M_e) = \frac{1}{2}$ ou, $N(Me) = \frac{n}{2}$.

Soient les observations (données brutes) suivantes, préalablement ordonnées.

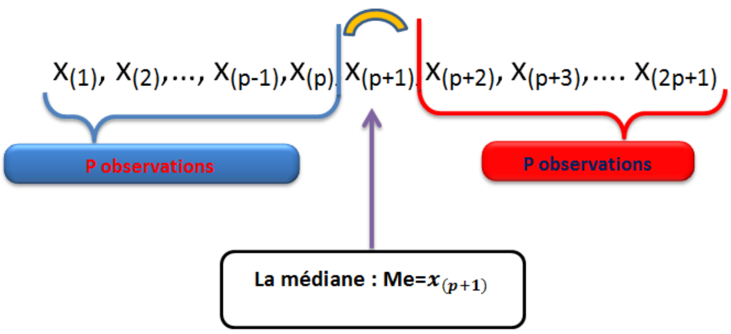
$$x_{(1)}, x_{(2)}, ..., x_{(n)}$$

Dans la recherche de la médiane, on distigue deux cas :

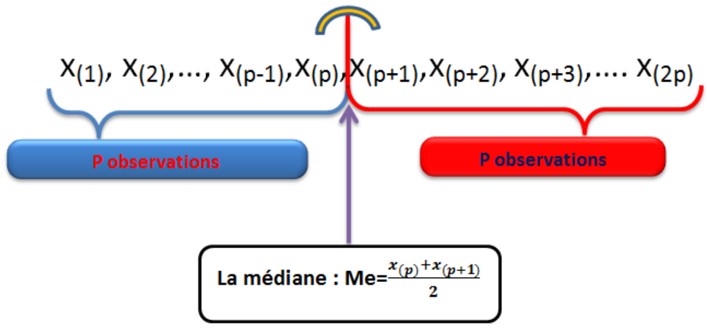
— Nombre d'observation est impair,

— Nombre d'observation est pair.

Recherche de la médiane



Cas discret, n est pair $(n = 2p)$



Exemple Numérique (cas discret)

Soient les observations suivantes :

Données brutes $1, 4, 0, 1, 0, 1, 4, 3, 2, 1$

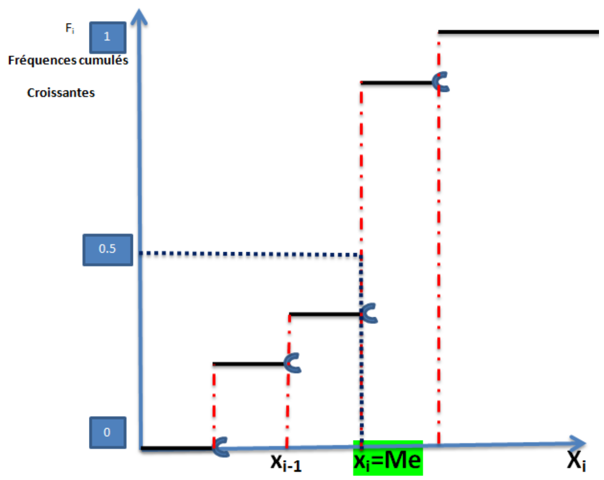
Observations ordonnées $0, 0, 1, 1, \mathbf{1}, \mathbf{1}, 2, 3, 4, 4$

Es que n est pair ? Oui, $n = 10 = 5 \times 2 \Longrightarrow p = 5$

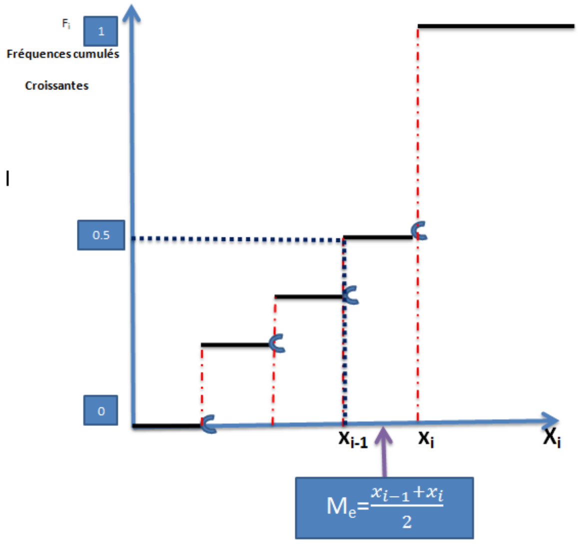
Trouver la médiane $M_e = \frac{x_p + x_{p+1}}{2} = \frac{1+1}{2} = 1$

Méthode graphique

Cas discret, Aucun pallier horizontal n'a $O.5$ comme ordonnée



Cas Discret, Un pallier horizontal a $O.5$ comme ordonnée



La médiane : Cas continu

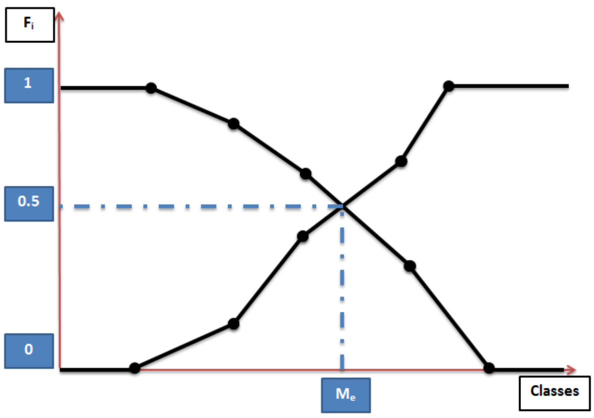
$$M_e = e_{i-1} + a_i \frac{\frac{1}{2} - F(e_{i-1})}{F(e_i) - F(e_{i-1})}$$

Tels que :

— $[e_{i-1}, e_i[$: est la classe modale,

— a_i est l'amplitude de la classe modale ($a_i = e_i - e_{i-1}$)

Graphiquement :



1.5 Moyenne arithmétique

La moyenne arithmétique

Noté $\overline{x}$, c'est la somme de toutes les observations divisée par n.

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

Cas discret $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$

Cas continu $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$

x_i : Centres de classes.

k : Nombre de classes.

1.5.1 Propriétés de la moyenne

Propriétés

1. $\frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x}) = 0$
2. Changement d'origine : Soit :$x'_i = x_i - x_0$, tel que x_0 : constante appelée origine.

   Alors la nouvelle moyenne sera : $\overline{x} = \overline{x'} + x_0$
3. Changement d'origine et d'echelle : $x'_i = \frac{x_i - x_0}{a}$
   - x_0 : origine (mode, médiane)
   - a : echelle (différence entre deux valeurs successives, ou l'amplitude unité dans le cas continu)

   La nouvelle moyenne sera alors : $\overline{x} = a\overline{x'} + x_0$

1.6 Paramètres de dispersion

1.6.1 Etendu

l'étendue d'une distribution est égale à la différence entre la plus grande et la plus petite valeur de la distribution :

$$e = max_{i=\overline{1,k}}(x_i) - min_{i=\overline{1,k}}(x_i)$$

1.6.2 Variance et Ecart type

La variance, notée $V(x)$ est la moyenne du carré des écarts à la moyenne.

$$V(x) = \frac{1}{n}\sum_{i=1} kn_i(x_i - \overline{x})^2 = \sum_{i=1}^{k} f_i(x_i - \overline{x})^2$$

L'écart type, noté σ_x est la racine carrée de la moyenne du carré des écarts à la moyenne, c'est à dire la racine carrée de la variance.

$$\sigma_x = \sqrt{V(x)} = \sqrt{\sum_{i=1}^{k}(x_i - \overline{x})^2}$$

Propriétés de la variance

1. $V(x) \geq 0$
2. $V(x) = \frac{1}{n}\sum_{i=1}^{k} n_i.x_i^2 - \overline{x}^2$
3. $V(x) = \sum_{i=1}^{k} f_i.x_i^2 - \overline{x}^2$
4. Soit $x_i' = \frac{x_i - b}{a} \Longrightarrow x_i = ax_i' + b$ alors

   (a) $V(x) = a^2 V(x') + b$

   (b) $\sigma_x = |a|\sigma_{x'}$

Remarque 5 La série qui a un ecart type petit, est moins dispersée.

1.6.3 Les quartiles

Les quarrtiles d'ordre α, noté Q_α ; sont les valeurs de la v.s. qui sont solution de l'equation $F(x) = \alpha$,

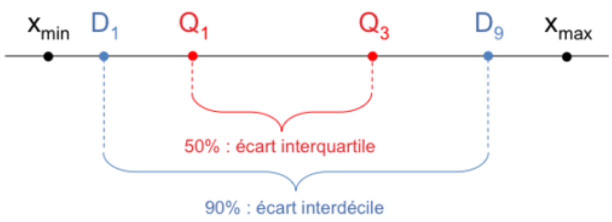
$$F(Q_1) = \frac{1}{4},\ F(Q_2) = F(M_e) = \frac{1}{2},\ F(Q_3) = \frac{3}{4}$$
$$N(Q_1) = \frac{1}{4}n,\ N(Q_2) = N(M_e) = \frac{1}{2}n,\ N(Q_3) = \frac{3}{4}n$$

1.6.4 Les déciles

Les 9 déciles sont les nombres réels qui partagent l'étendue en dix intervalles de même effectif.

1.6.5 Intervalle interquartiles

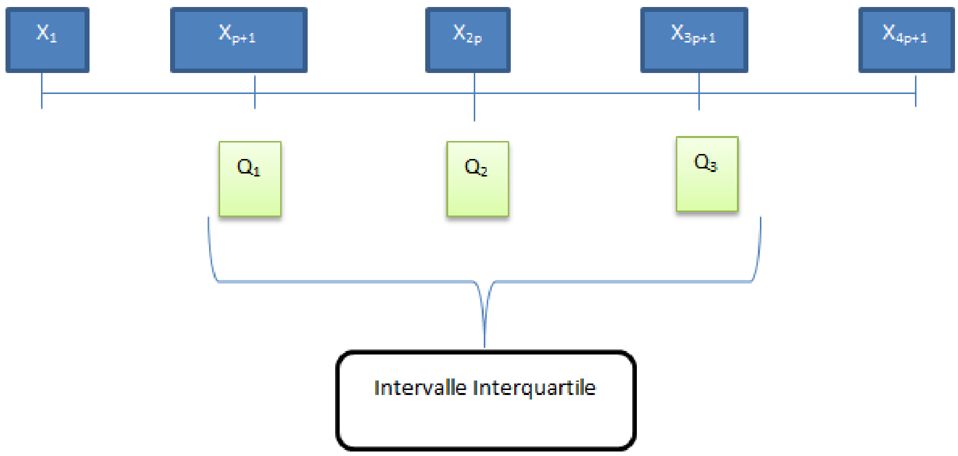
Intervalle qui contient 50% de l'effectif, laissant à gauche et à droite 25%

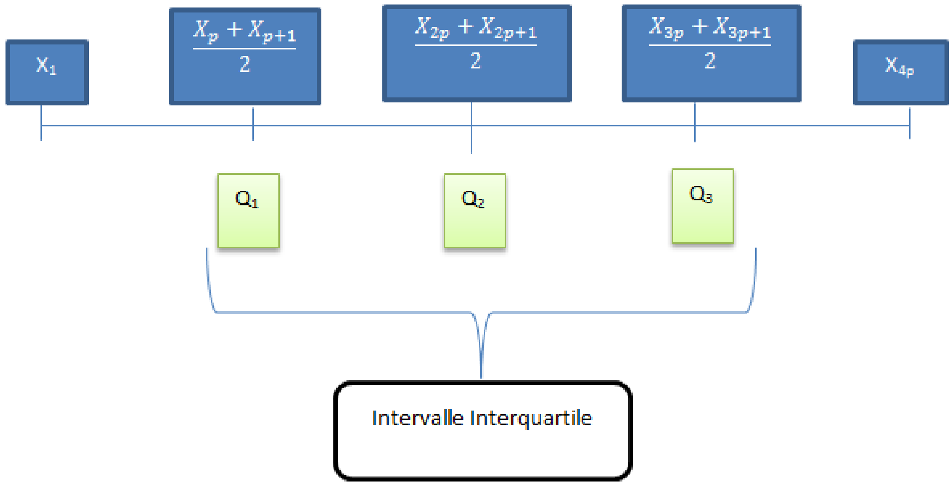


Pour cela, il faut déterminer Q_1, Q_3 avec la même méthode qu'on a utiliser pour trouver la médiane.

Exemple : Cas discret

Si on peu ecrire n, sous la forme $n = 4p$ ou $n = 4p + 1$, alors on détermine la médiane, ensuite on détermine Q_1, Q_3, tel que Q_1 est la médiane de la partie gauche, ett Q_3, celle de la partie droite (comme dans la figure)





1.6.6 Moment simple d'ordre r

$$m_r = \frac{1}{n} \sum_{i=1} k n_i x_i^r$$

1.6.7 Moment centré d'ordre r

$$\mu_r = \frac{1}{n} \sum_{i=1} k n_i (x_i - \overline{x})^r$$

1. $\overline{x} = m_1$
2. $V(x) = m_2 - m_1^2$

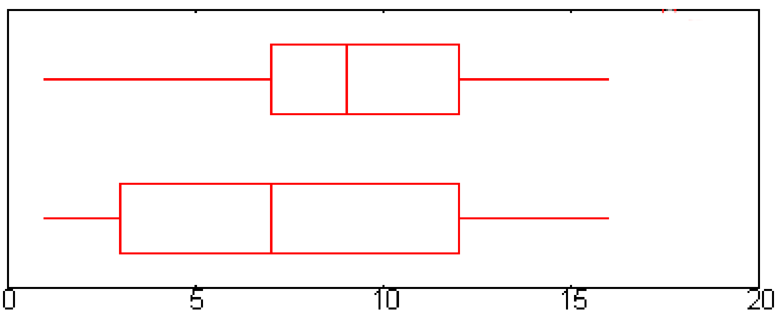
1.7 La boite à moustache ou Box-Plot

1.7.1 Présentation

Dans les représentations graphiques de données statistiques, la boîte à moustaches ou diagramme en boîte est un moyen rapide de figurer le profil essentiel d'une série statistique quantitative. Elle a été inventée en 1977 par John Tukey, mais peut faire l'objet de certains aménagements selon les utilisateurs. Son nom est la traduction de Box and Whiskers Plot.

1.7.2 Principe

La boîte à moustaches résume seulement quelques caractéristiques de position du caractère étudié (médiane, quartiles, minimum, maximum ou déciles). Ce diagramme est utilisé principalement pour comparer un même caractère dans deux populations de tailles différentes. Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. Ce rectangle suffit pour le diagramme en boîte. On ajoute alors des segments aux extrémités menant jusqu'aux valeurs extrêmes, ou jusqu'aux premier et neuvième déciles (D1 / D9), voire aux 5e et 95e centiles. On parle alors de diagramme en boîte à moustaches ou de diagramme à pattes.



Comparaison de deux diagrammes en boîte à moustaches D_1 / D_9 avec
- pour la boîte inférieure : Q1 = 3, M = 7, Q3=12, $D_1 = 1$, $D_9 = 16$
- pour la boîte supérieure : Q1 = 7, M = 9, Q3=12, $D_1 = 1$, $D_9 = 16$

1.8 Introduction

La statistique à deux variables, c'est l'étude du lien entre deux variables (caractères) sur une même population (taille et point d'un nouveau né, vitesse et consommation du véhicule,...). On appelera ces deux caractères X et Y, définis sur une population de taille

n (n individus), on s'interesse à étudier le lien entre X et Y. Les données brutes sont les couples $(x_i, y_j), i = \overline{1, n}$.

1.9 Tableau de contingence

Le tableau de contingence est un moyen particulier de représenter simultanément deux caractères observés sur une même population, s'ils sont discrets ou bien continus et regroupés en classes. Les deux caractères sont X et Y, la taille de l'échantillon est n. Les modalités ou classes de x seront notées $x_1, x_2, \cdots, x_p$, celles de Y sont notées $y_1, y_2, \cdots, y_q$.

On note :

1. n_{ij} : l'effectif conjoint de x_i et y_j : c'est le nombre d'individus pour lesquels X prend la valeur x_k et Y la valeur y_j,
2. $n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$: l'effectif marginal de x_i : c'est le nombre d'individus pour lesquels X prend la valeur x_i,
3. $n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$: l'effectif marginal de y_j : c'est le nombre d'individus pour lesquels Y prend la valeur y_j.

On représente ces valeurs dans un tableau à double entrée, dit tableau de contingence.

$X \setminus Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_q	Total
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2q}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\vdots$	n_{ij}	$\vdots$	n_{iq}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_p	n_{p1}	n_{p2}	$\vdots$	n_{pj}	$\vdots$	n_{pq}	$n_{p\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet q}$	$n_{\bullet\bullet} = n$

1.10 Distribution conjoint

fréquence jointe notée f_{ij}, c'est la proportion d'individus ayant pris simultanément la modalité x_i de X et la modalité y_j de Y.

$$f_{ij} = \frac{n_{ij}}{n}$$

Définition 1.6 On appelle distribution conjointe du couple (X, Y) ou encore série statistique à deux dimensions, les données des couples $(x_i, y_j),\ i = \overline{1,p},\quad j = \overline{1,q}$ et les fréquences jointes $f_{ij},\ i = \overline{1,p},\quad j = \overline{1,q}$.

1.11 Distribution marginale

Définition 1.7 On appelle fréquence marginale associée à la modalité x de X notée $f_{i\bullet}$ la proportion :

$$f_{i\bullet} = \frac{n_{i\bullet}}{n} = \sum_{i=1}^{q} f_{ij}$$

$$\sum_{i=1}^{p} f_{i\bullet} = 1$$

Définition 1.8 On appelle distribution marginale de la variable X (resp. Y), les données des couples $(x_i, f_{i\bullet}),\ i = \overline{1,p}$, (resp. $(y_j, f_{\bullet j}), j = \overline{1,q}$).

$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet}x_i = \sum_{i=1}^{p} f_{i\bullet}x_i$

$V(x)\ \ \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet}x_i^2 - \overline{x}^2 = \sum_{i=1}^{p} f_{i\bullet}x_i^2 - \overline{x}^2$

$\sigma(x) = \sqrt{V(x)}$

$\overline{y} = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j}y_j = \sum_{j=1}^{q} f_{\bullet j}y_j$

$V(y)\ \ \frac{1}{n}\sum_{j=1}^{p} n_{\bullet j}y_j^2 - \overline{y}^2 = \sum_{j=1}^{q} f_{\bullet j}y_j^2 - \overline{y}^2$

$\sigma(y) = \sqrt{V(y)}$

Exemple illustratif

Soit le tableau de contingence suivant :

$X \setminus Y$	0	1	2	$n_{i\bullet}$
1	6	3	1	$n_{1\bullet} = 10$
2	5	5	5	$n_{2\bullet} = 15$
$n_{\bullet j}$	$n_{\bullet 1} = 11$	$n_{\bullet 2} = 8$	$n_{\bullet 3} = 6$	$n = 25$

Le tableau de fréquences correspondant est :

$X \setminus Y$	0	1	2	$f_{i\bullet}$
1	0.24	0.12	0.04	$f_{1\bullet} = 0.4$
2	0.44	0.32	0.24	$f_{2\bullet} = 0.6$
$f_{\bullet j}$	$f_{\bullet 1} = 0.44$	$f_{\bullet 2} = 0.32$	$f_{\bullet 3} = 0.24$	$f_{\bullet\bullet} = 1$

Distribution marginale de X

x_i	$n_{i\bullet}$	$f_{i\bullet}$
1	10	0.4
2	15	0.6
Total	25	1

Distribution marginale de X

y_i	$n_{i\bullet}$	$f_{i\bullet}$
0	11	0.44
1	8	0.32
2	6	0.24
Total	25	1

1.12 Distribution conditionnelle

On peut Considérer la distribution de X (resp. Y) sur une sous population ayant pris la modalité y_j de Y (resp. x_i de X).

On les appelle distributions conditionnelles, notées : $X/Y = y_j$. (resp. $Y/X = x_i$).

$X/Y = y_j$ prend les modalitées $x_1, x_2, \cdots, x_p$.

$Y/X = x_i$ prend les modalitées $y_1, y_2, \cdots, y_q$.

L'effectif de la population ayant pris la modalité y_j de Y (resp. x_i de X) est $n_{\bullet j}$ (resp. $n_{i\bullet}$)

La fréquence conditionnelle de $X/Y = y_j$ (resp. $Y/X = x_i$) est définie par :

$$f_{i/j} = \frac{n_{ij}}{n\bullet j} = \frac{f_{ij}}{f_{\bullet j}},\ i = \overline{1,p}$$

$$f_{j/i} = \frac{n_{ij}}{ni\bullet} = \frac{f_{ij}}{f_{i\bullet}},\ j = \overline{1,q}$$

Tableau de contingence :

X\Y	0	1	2
1	6	3	1
2	5	5	5

Distribution de $X/(Y = y_1 = 0)$

x_i	n_{i1}	$f_{i/j=1}$
1	6	0.55
2	5	0.45
Total	11	1

Distribution de $X/(Y = 1)$

x_i	n_{i2}	$f_{i/j=2}$
1	3	0.38
2	5	0.62
Total	8	1

Distribution de $X/(Y = 2)$

x_i	n_{i3}	$f_{i/j=3}$
1	1	0.17
2	5	0.62
Total	6	1

Moyennes conditionnelles

$\overline{x} = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} x_i n_{ij} = \sum_{i=1}^{p} x_i f_{i/j},\ j = \overline{1,q}$

$\overline{y} = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} y_j n_{ij} = \sum_{j=1}^{p} y_j f_{j/i},\ i = \overline{1,p}$

Variances conditionnelles

$V_j(x) = \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} n_{ij}(x_i - \overline{x_j})^2,\ j = \overline{1,q}$

$V_i(y) = \frac{1}{n_{i\bullet}} \sum_{j=1}^{p} n_{ij}(y_j - \overline{y_i})^2,\ i = \overline{1,p}$

Ecart type

$\sigma_j(x) = \sqrt{V_j(x)}$

$\sigma_i(y) = \sqrt{V_i(y)}$

Moyenne des moyenne

$\overline{x} = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j}\overline{x_j}$

$\overline{y} = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet}\overline{y_i}$

1.13 Indépendance

On dit que deux variables X, Y sont indépendantes si les variations de l'une ne dépendent pas des variations de l'autre. Ou bien que les fréquences conditionnelles $f_{i/j}$ ne dépendent pas de j et $f_{j/i}$ ne dépendent pas de i.

Si $f_{i/j} = f_{i\bullet}, i = \overline{1,p}$ et $f_{j/i} = f_{\bullet j}, j = \overline{1,q}$

est équivalent à dire que ($\Leftrightarrow$) X et Y sont indépendantes.

$\Leftrightarrow f_{ij} = f_{i\bullet} \times f_{\bullet j}$

$\Leftrightarrow n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$

1.14 Nuage de points

Les couples $\{(x_i, y_j), i = \overline{1,p}, j = \overline{1,q}\}$ peuvent être représentés graphiquement comme dans la figure 1.2.

Le nauge de point c'est l'ensemble des points $\{M_{ij}, i = \overline{1,p}, j = \overline{1,q}\}$. Le nuage de points est un bon indicateur sur l'existance d'une liaison entre X et Y.

Remarque 6 Si les points du nuage sont répartis sur une bonde régulière et allongées, on peut proposer un ajustement linéaire pour expliquer Y en fonction de X ou inversement nuage.png

Figure 1.2 – Nuage de points

1.15 La covariance

On appelle covariance entre "X et Y" la quantité :

$$\begin{aligned} cov(x,y) &= \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i-\overline{x})(y_i-\overline{y}) \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_iy_j(\overline{x}.\overline{y}) \end{aligned}$$

Remarque 7 La covariance peut être positive ou négative, elle joue un rôle analogue à celui de la variance.

1.16 Droite de regression

Soit, (Δ), la droite de regression de Y en X :

$(\Delta) : y = ax + b$

Soit le point $M = (x_i, y_j)$,

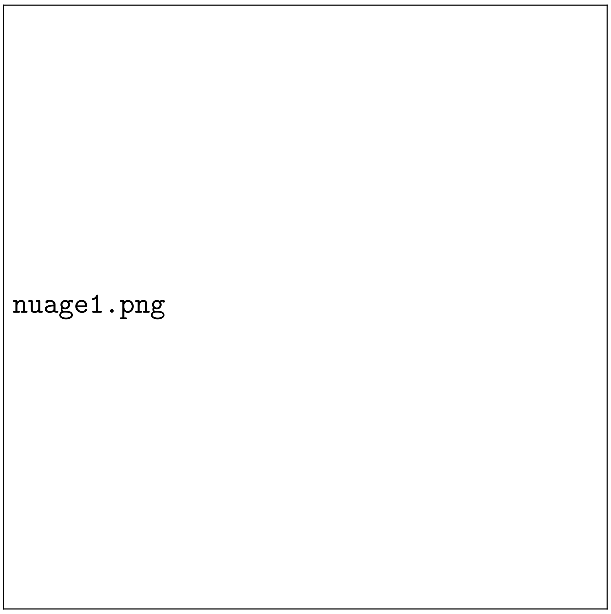


FIGURE 1.3 – Abscence de liaison linéaire

$y'_j = ax_i + b$

y_j : Valeur obbservée, et y'_j : valeur ajustée.

Soit, l'erreur entre y_j et y'_j, $e_i j = y_j - y'_j = y_j(a.x_i + b)$

Chercher la droite (Δ) revient à chercher les paramètres a et b ?

Pour cela, on utilise le critère de minimisation de la somme des erreurs e_{ij} au carré, on mminimise le critère :

$$S(a,b) = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} e_{ij}^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(y_j - (ax+b))^2$$

***Théorme** 1 Soit $M_{ij}, i = \overline{1,p}, j = \overline{1,q}$, un nuage de point de coordonnées (x_i, y_j).*

L'équation de la droite de regression de Y en X est écrite sous la forme suivante $y = ax+b$ est définie par :

$$a = \frac{cov(X,Y)}{V(X)}$$

$$b = \overline{y} - a\overline{x}$$

nuage2.png

Figure 1.4 – Faible liaison linéaire

Soit $(\Delta') : x = \alpha y + \beta$

Avec la même procédure appliquée pour trouver la droite de Y en X, on trouve que la droite de X en Y s'écrit sous la forme $x = \alpha y + \beta$ tels que :

$$a = \frac{cov(X, Y)}{V(Y)}$$

$$b = \overline{x} - a\overline{y}$$

— $\overline{x}$: moyenne marginale de X,

— $\overline{y}$: moyenne marginale de Y,

— $V(X)$: variance marginale de X,

— $V(X)$: variance marginale de X,

Remarque 8 Le point $(\overline{x}, \overline{y})$ est le point d'intersection des deux droites de regression (Δ) et (Δ'). Il est applé "**Centre de gravité du nuage de point**" ou "**Point moyen**" `nuage3.png`

FIGURE 1.5 – Forte liaison linéaire

1.17 Coeffecient de corrélation

Pour mesurer l'intensité de la liaison linéaire entre X et Y, on calcule le coefficient de corrélation noté $R(\rho, R_{xy}), \rho_{xy}$ définie par $R = \frac{cov(X,Y)}{\sigma_x \sigma y}$ où,

$\sigma_x = \sqrt{V(x)}, \sigma_y = \sqrt{V(y)}$.

— $R^2 = a\alpha$

— $R^2 \leq 1 \Leftrightarrow -1 \leq R^2 \leq 1$

— $R^2 = 1 \Rightarrow$ les points sont allignés et les droites Δ et Δ' sont confondues

$|R| = 1$ **:** forte liaison linéaire

$|R| \backsimeq 0$ abscence de liaison linéaire

Plus (Δ) **et** (Δ') **se rapproche** plus la correlation est forte.

X**,** Y **deux variables statistiques continues** x_i, y_j représentent les centres de classes eij.png

Figure 1.6 – Droite de Y en X

Dans le cas de données brutes (x_i, y_j)**,** $i = \overline{1,p}, j = \overline{1,q}$

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

,

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

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

.

Si $R \backsimeq 0$ Il faut rejeter l'ajustement linéaire.

Si $R \backsimeq 1$ Il faut compléter par un examen graphique.

Dans le cas d'abscence de renseignements complémentaires On admet que $|R| \geq 0.75$ justifie l'ajustment linéaire.

R1.png

Figure 1.7 – Cas où R=-1

R11.png

Figure 1.8 – R=+1

R0.png

Figure 1.9 – R=0

R00.png

Figure 1.10 – $R < 0$

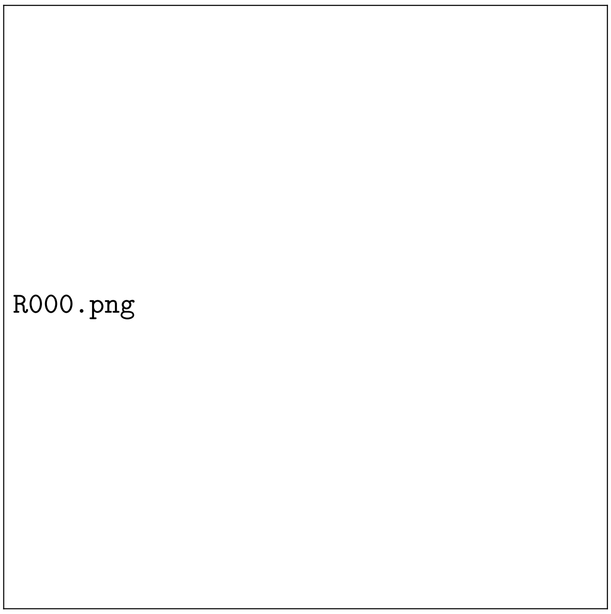


FIGURE 1.11 – $R > 0$

1.18 Droite de Mayer

On partage l'ensemble des points (x, y) en deux sous ensemble E_1 et E_2 ayant presque le même nombres de points , on cherche les centres de gravité G_1 et G_2, des sous ensembles E_1 et E_2. La droite de Mayer est celle qui lie les deux points G_1, G_2

1.19 Notions sur les ensembles

Soit A, un ensemble qui contient les évènements a, b, c, noté : $A = \{a, b, c\}$.

Si un élément a **appartient** à A, on écrit alors : $a \in A$.

Si un élément a n'appartient pas à A, on écrit alors : $a \notin A$.

Soit $B = \{a, b\}$, On dit que B est un **sous-ensemble** de A, on peut aussi écrire : $B \subseteq A$.

1.19.1 Opérations sur les ensembles

$$A \subseteq B \Leftrightarrow [\forall x \in A \Longrightarrow x \in B]$$

mayer.png

FIGURE 1.12 – Droite de Mayer

$$A \cup B \Leftrightarrow [x \in A \; ou \; x \in B]$$

$$A \cap B = \{x \in A \; et \; x \in B\}$$

1.20 Analyse combinatoire

Définition 1.9 (Analyse combinatoire) L'analyse est le domaine de la mathématique qui s'occupe de l'étude de l'ensemble des issues, événements ou faits (distinguables ou non tous distinguables) avec leurs arrangements (combinaisons) ordonnés ou non selon certaines contraintes données.

Définition 1.10 (suite Ordonnée) Une suite d'objets (événements, issues, objets,...) est dite "ordonnée" si chaque suite composée d'un ordre particulier des objets est comptabilisée comme une configuration particulière.

Définition 1.11 (famille non ordonnée) Une suite est donc "non ordonnée" si et seulement si, nous nous intéressons à la fréquence d'apparition des objets indépendamment de leur ordre.

Définition 1.12 (Evénement distinct) Des objets (d'une suite) sont dits "distincts" si leurs caractéristiques ne permettent pas de les confondre avec des autres objets.

Remarque 9 Nous avons choisi de mettre l'analyse combinatoire dans ce chapitre car lorsque nous calculons des probabilités, nous avons également assez souvent besoin de savoir quelle est la probabilité de tomber sur une combinaison ou un arrangement d'événements donnés sous certaines contraintes.

1.20.1 Principe de multiplication

Permet de compter le nombre de résultats d'expériences qui peuvent se décomposer en une succession de sous-expériences.

On suppose qu'une expérience est la succession de m sous-expériences. Si la i^{eme} expérience a ni résultats possibles pour i = 1 ; : : : ; n, alors le nombre total de résultats possibles de l'expérience globale est

$$n = \prod_{i=1}^{m} n_i$$

1.20.2 Arrangement

Définition 1.13 Etant donné un ensemble E de n objets, on appelle arrangements de p objets toutes suites ordonnées de p objets pris parmi les n objets. Le nombre d'arrangements de p objets pris parmi n est noté : A_n^p, tel que $p \leq n$ et $n > 0$

Il existe deux type d'arrangements,

Arrangement avec répétition : $a_n^p = n^p$

Arragmenet sans répétition : $A_n^p = \frac{n!}{(n-p)!}$

Remarque 10 Dans le cas d'arrangement sans répétition, si $n = p$, on parlera alors de permutation, dans le nombre est égale à $n!$

tri.png

1.20.3 Combinaisons

Définition 1.14 Un combinaison (notée (C_n^p)) de k éléments pris dans un ensemble à n éléments distincts est un sous-ensemble à k éléments de cet ensemble. Les éléments sont pris sans répétition et ne sont pas ordonnés.

Le nombre de combinaisons est donné par la formule suivante :

$$C_n^p = \frac{n!}{(n-p)!p!}$$

L'une des plus importante propriété des ombinaisons est $C_n^p = C_n^{n-p}$, ce qui a donnée comme conséquence le triangle de Pascal.

1.20.4 Binôme de Newton

La formule de Newton est une formule mathématique donnée par Isaac Newton pour trouver le développement d'une puissance entière quelconque d'un binôme. Elle est aussi appelée formule du binôme de Newton, ou plus simplement formule du binôme.

$$(x+y)^n = \sum_{k=0}^{n} C_n^p x^{n-k} y^k$$

1.21 Introduction au calcul des probabilités

Cette première notion de la théorie des probabilités s'est imposée au 17^{eme} siècle dans l'étude des jeux de hasard (jeux de dés, de cartes, etc...). Une expérience aléatoire se décrit mathématiquement par la donnée de l'ensemble des résultats possibles de l'expérience en question. Il est de tradition de noter Ω un tel résultat (parfois appelé ; éventualité" dans la suite) et de désigner par l'espace de tous ces résultats possibles.

1.22 Terminologie des probabilités

Espace ou ensemble fondamental

Définition 1.15 On appel espace fondamental noté Ω d'une expérience (E), tous les résultats attendus de cette expérience.

— Pour l'experience E_1 : lancement d'un dé numéroté de 1 jusqu'a 6, l'espace fondamental est :

$$\Omega_1 = \{1, 2, 3, 4, 5, 6\}$$

— E_2 : "Jet d'une pièce", cette éxpèrience aura pour espace fondamental :

$$\Omega_2 = \{P, F\}$$

— A l'instant $t = 0$ on met en fonctionnement un appareil et l'on s'intèresse au temps de bon fonctionnement :

$$\Omega_3 = \Re_+$$

— On lance deux dés et l'on s'intèresse à la somme des deux faces supérieures :

$$\Omega_4 = \{2, 3, 4, ..., 12\}$$

1.22.1 Evénement (d'une expèrience aléatoire)

Définition 1.16 Un événement est un ensemble de résultats ω d'une expérience aléatoire possédant une même proprité.

Dans l'expérience du jet d'un dé, les formulations suivantes représentent des évenements :

— ω_1 : "Avoir le chiffre 1"

— ω_2 : "Avoir une chiffre impair"

— ω_3 : "Avoir un chiffre inférieur à 3"

Opérations sur les événements

— A tout événement ω, on associée son contraire $\overline{\omega}$, qui représente son complémentaire dans l'ensemble fondammentale.

— L'événement impossible (qui ne se réalise jamais), sera noté $\emptyset$. L'équation $A_1 \cap A_2 = \emptyset$, A_1 et A_2 sont disjoints, et signifie que A_1 et A_2 sont incompatibles.

— L'événement certain (qui se réalise toujours) sera noté Ω.

1.22.2 Opération sur les évènements

oper.png

oper1.png

1.23 Le concept de probabilité

Considérons une expérience aléatoire dont l'ensemble fondamental des résultats contient un nombre fini n de résultats, c'est-à-dire $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$.

A chacun des n événements simples (événements ne correspondant qu'à un seul résultat) qu'on peut définir pour cette expérience aléatoire correspond une probabilité de réalisation ; c'est-à-dire à chacun des événements ω_i correspond une probabilité de réalisation $(P\{\omega_i\}), (i = 1, 2, ..., n)$. Ces n probabilités sont telles que : pour $i = 1, 2, ..., n, 0 \leq P(\omega_i) \leq 1$ et $\sum_{i=1}^{n} P(\omega_i) = 1$

Soit A un événement défini à partir de cette expérience, c'est-à-dire soit A un sous-ensemble de Ω. La probabilité de réalisation de l'événement A peut se noter $P(A)$ et, de façon générale,

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}) \tag{1.1}$$

Si les n éléments de l'ensemble Ω sont équiprobables, c'est-à-dire si $P(\omega_i) = \frac{1}{n}, i = 1, 2, ..., n$, alors :

$$P(A) = \frac{\#A}{\#\Omega} = \frac{\#A}{\#n} \tag{1.2}$$

où le symbole # désigne la cardinalité (le nombre d'éléments) de l'ensemble qui le suit.

1.23.1 Concept de probabilité conditionnelle

Définition 1.17 (probabilité conditionnelle) Soit A et E deux événements où E a une probabilité de réalisation non-nulle. La probabilité conditionnelle de réalisation de l'événement A étant donné que l'événement E s'est réalisé est notée $P(A|E)$. Parfois, la probabilité conditionnelle $P(A|E)$ se calcule ou se déduit directement. Lorsque ce n'est pas le cas, on utilise la définition : $P(A|E) = P(A \cap E)/P(E)$.

1.24 Quelques propriétés des probabilités conditionnelles

Les propriétés des probabilités de réalisation de deux événements A et B, s'étendent à la probabilité conditionnelle.

Soit A, B et E trois événements, où E a une probabilité de réalisation non nulle.

$0 \leq P(A|E) \leq 1 \quad 0 \leq P(B|E) \leq 1$

$P(\overline{A}|E) = 1 - P(A|E)$ et $P(\overline{B}|E) = 1 - P(B|E)$

1.25 Concept d'indépendance en probabilité

Deux événements possibles A et B sont dits indépendants en probabilité si la réalisation de l'un ne modifie en rien la probabilité de réalisation de l'autre. On peut définir ce concept d'indépendance en probabilité de 2 événements possibles de 3 façons équivalentes :

1. $P(A|B) = P(A)$;

2. $P(A|B) = P(A)$;

3. $P(A \cap B) = P(A).P(B)$

1.26 Formule de Probabilité totale et théorème de Bayes

Soit $E_1, E_2, ..., E_k$ une suite d'événements qui forme une partition de Ω et dont les probabilités de réalisation $P(E_i), i = 1, 2, ..., k$, sont connues et non nulles.

Soit A un événement quelconque pour lequel il est possible de calculer directement les probabilités conditionnelles de réalisation $P(A|E_i), i = 1, 2, ..., k$.

Formule des probabilités totales

$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + ... + P(A|E_k)P(E_k)$

Loi de Bayes $P(E_i|A) = P(A \cap E_i)/P(A) = P(A|E_i)P(E_i)/P(A), i = 1, 2, ..., k$.

1.27 Quelques exercices de Probabilité

Exercice $N°1$

Soit A l'ensemble des nombres de quatre chiffres, le premier étant non nul.

1. Calculer le nombre d'éléments de A.
   (a) Dénombrer les éléments de A :
   (b) composés de quatre chiffres distincts
   (c) composés d'au moins deux chiffres identiques
   (d) composés de quatre chiffres distincts autres que 5 et 7

Exercice $N°2$

Quatre garçons et deux filles s'assoient sur un banc.

1. Quel est le nombre de dispositions possibles ?
2. Même question si les garçons sont d'un côté et les filles de l'autre.

3. Même question si chaque fille est intercalée entre deux garçons.

4. Même question si les filles veulent rester l'une à côté de l'autre

Exercice $N°3$

A l'Université de Bejaia, la probabilité pour que l'étudiant X en première année abandonne est de $\frac{1}{5}$ et la probabilité pour que l'étudiant Y de deuxième année abandonne est de $\frac{1}{8}$. En supposant les deux événements indépendants, calculer la probabilité que :

1. X et Y abandonnent l'Université de Béjaia.

2. Ni X, ni Y ne quittent l'Université.

3. Y seulement quitte l'UB.

4. L'un des deux quitte l'UB.

Exercice $N°4$

On jette deux dés l'un après l'autre, on note les valeurs obtenues de chaque dé.

1. Citer les éléments de l'espace des épreuves (l'ensemble fondamental).

2. Citer les éléments des événements suivants :

   (a) A = {la somme des valeurs obtenues ≥ 5}

   (b) B = {la valeur du premier > la valeur du deuxième}

   (c) C = {la valeur du premier = 4}

3. Citer les éléments des événements suivants :

   (a) $A \cap B$

   (b) $B \cup C$

   (c) $A \cap (B \cup C)$