B. Sihon

Cours de Probabilités & Statistique (L1 MI)

A. Menni

8 février 2016

Chapitre1 : Statistique Descriptive Univariée

---

1 - Introduction et quelques définitions fondamentales

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données. Il ne faut pas confondre **la statistique** qui est la science qui vient d'être définie et **les statistiques** qui sont un ensemble de données chiffrées sur un sujet précis. Ces données peuvent être présentées sous forme de tableaux ou de graphiques.

L' ensemble étudié est appelé **population** et est en général noté Ω. Les éléments de la population sont appelés **individus** ou **unités statistiques**. La population est étudiée selon un ou plusieurs **caractères**. Un caractère peut être décrit de différentes manières. Chaque manière est appelée **modalité**. Un caractère est dit **qualitatif** si ses modalités ne sont pas mesurables. Il est dit **quantitatif** si ses modalités sont, par contre, mesurables. Dans un tel cas, le caractère est tout simplement dit **variable statistique**.

Un caractère qualitatif est dit **nominal** si les modalités sont exprimables par des noms et ne sont pas hiérarchisées. Un caractère nominal est dit **dichotomique** s'il ne peut prendre que deux modalités seulement. Il est dit **ordinale** si les modalités traduisent le degré d'un état caractérisant un individu sans que ce degré ne puisse être défini par un nombre qui résulte d'une mesure (les modalités sont alors hiérarchisées).

Une variable statistique est dite **discrète** si ses modalités sont des valeurs isolées. Et elle est dite **continue** si ses modalités sont quelconques dans un intervalle de valeurs continues.

Remarque

En réalité le nombre de valeurs possibles pour un caractère donné dépend de la précision de la mesure : on peut considérer comme continu un caractère discret qui peut prendre un grand nombre de valeurs et vice versa.

Pour recueillir des informations sur une population statistique, on dispose de deux méthodes :

- Le **recensement** ou la **méthode exhaustive** où chaque individu de la population est analysé selon le ou les caractères statistiques d'intérêt.
- L'**échantillonnage** ou la méthode du **sondage** qui consiste à n'examiner qu'une partie de la population appelé **échantillon**.

L'échantillonnage représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée. Ces derniers forment alors l'**échantillon** observé. Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, l'échantillon doit être représentatif de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul l'**échantillonnage aléatoire**, où les individus sont choisis au hasard, assure la représentativité de l'échantillon.

2 – Statistique Descriptive Univariée

Définition1 :

On appelle **Statistique univariée** ou **statistique à 1 dimension** une application notée X, définie d'un ensemble Ω représentant la population étudiée vers un ensemble Θ représentant l'ensemble des valeurs prises par le caractère étudié :

$$X : \Omega \longrightarrow \Theta$$

$$\omega \longmapsto X(\omega)$$

Autrement dit, pour chaque individu ω dans la population, on associe une valeur $X(\omega)$ au caractère mesuré.

- Exple1 (*caractère qualitatif nominal binaire*) : $\Omega = \{Employés\}$ et X = possession d'un véhicule.
- Exple2 (*caractère qualitatif nominal*) : $\Omega = \{Enseignants\}$ et X = la faculté employeuse.
- Exple3 (*caractère qualitatif ordinal*) : $\Omega = \{Etudiants\}$ et X = l'année de scolarsation.
- Exple4 (*caractère quantitatif discret*) : $\Omega = \{Voitures\}$ et X = le nombre de places.
- Exple5 (*caractère quantitatif continu*) : $\Omega = \{années\}$ et X = le taux de chômage en Algérie.

Définition2

On appelle **série statistique** une suite finie de modalités mesurées sur un caractère d'intérêt X. Les valeurs de X sont relevées pour un échantillon d'individus appartenant à la même population. Le nombre d'individus qui constituent l'échantillon étudié s'appelle **la taille de l'échantillon** et se note généralement n.

Exemple1

La série statistique suivante donne la répartition d'un échantillon de 46 employés d'une certaine entreprise étatique, selon qu'ils sont véhiculés (notés par 1) ou non (notés par 0).

0 0 0 1 0 0 0 0 0 0 1 1 0 0 1 0 1 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 0 1 1 0 1 1 1 1 0 1 0 0 0 1 1

Exemple2

La série statistique suivante donne la distribution de 20 enseignants selon leurs facultés d'appartenance.

Math Physique GM Math Biologie Math Chimie GC GC GM Electronique GM Physique Electronique Info Math Biologie Electronique Electronique Info

Exemple3

La série statistique suivante donne la répartition de 65 étudiants selon leur année de scolarisation.

L3 M1 L3 L1 M2 L3 L1 L3 L2 M1 M2 M1 L3 L2 L2 L1 M2
L1 L3 L2 L2 M1 L3 M2 L1 L1 L2 M1 L1 L2 L1 M2 L1 L1 M1
L3 L3 L3 M1 L2 L1 M2 L2 M1 L2 L3 M2 L2 M1 L1 L3 M2
M1 L2 L2 M2 L3 L2 M1 L3 M2 L2 M1 M1 M2

Exemple4

La série statistique suivante donne le nombre de places pour un échantillon de 50 voitures vendues par un certain constructeur automobile en l'espace d'une semaine.

5 3 8 5 5 2 7 4 5 6 5 2 5 5 4 5 6 5 2 5 4 5 5 5 5 6 5 2 5 2 5 3 8 5 5 2 3 7 5 8 5 5 7 5 5 8 2 4 7 7

Exemple5

La série statistique suivante donne le taux de chômage (%) en Algérie relevé entre 2001 et 2010.

34 31 26,2 25,4 17,1 15,7 11,8 12,5 10,2 9,9

Définition3 [variable statistique discrète]

Soit X une statistique univariée avec $X(\Omega) = \{x_1,\ x_2,\ \ldots,\ x_p\}$

1. on appelle **effectif** de la valeur x_i, le nombre de fois que cette valeur se répète dans l'échantillon étudié. Ce nombre est noté n_i. Et on a $\sum_{i=1}^{p} n_i = n$
2. on appelle **effectif cumulé** en x_i, la somme de tous les effectifs n_j avec $j \leq i$. Ce nombre est noté $\tilde{n}_i$:

$$\tilde{n}_i = \sum_{j=1}^{i} n_j = n_1 + n_2 + \cdots + n_i$$

3. on appelle **fréquence** de la valeur x_i, le taux de répétition de cette valeur dans l'échantillon étudié. Cette fréquence est notée f_i :

$$f_i = \frac{n_i}{n}$$

Remarques :

a) Parfois on peut utiliser le terme de **fréquence relative** pour désigner les fréquences.

b) La fréquence est un nombre compris entre 0 et 1

c) Le pourcentage est une fréquence exprimée en pour cent. Il est égal à $100 \times f_i$

4. on appelle **fréquence cumulée** en x_i, le nombre noté $\tilde{f}_i$:

$$\tilde{f}_i = \frac{\tilde{n}_i}{n} = \sum_{j=1}^{i} f_j.$$

Définition4 [**variable statistique continue**]

Soit X une statistique univariée et supposons que $X(\Omega) = [a, b]$ tel que cet intervalle soit divisé en k classes :

$$[a, b] = [a_0, a_1[\cup [a_1, a_2[\cup \ldots \cup [a_{k-1}, a_k]$$

avec $a_0 \leq a$ et $a_k \geq b$

1. on appelle **effectif** de la *ième* classe $[a_{i-1}, a_i[$, le nombre de valeurs du caractère appartenant à cette classe. Ce nombre se note toujours n_i
2. on appelle **effectif cumulé** en a_i le nombre $\tilde{n}_i = \sum_{j=1}^{i} n_j$.
3. on appelle **fréquence** de la *ième* classe $[a_{i-1}, a_i[$, le nombre $f_i = \frac{n_i}{n}$.
4. on appelle **fréquence cumulée** en a_i le nombre $\tilde{f}_i = \frac{\tilde{n}_i}{n} = \sum_{j=1}^{i} f_j$

Propriétés :

— Dans le cas où $X(\Omega) = \{x_1, x_2, \ldots, x_p\}$, on a : $\sum_{i=1}^{p} n_i = n$ et $\sum_{i=1}^{p} f_i = 1$

— Dans le cas où $X(\Omega) = [a, b] = [a_0, a_1[\cup [a_1, a_2[\cup \ldots \cup [a_{k-1}, a_k]$, on a $\sum_{i=1}^{k} n_i = n$ et $\sum_{i=1}^{k} f_i = 1$

3 Tableaux statistiques

Le tableau de distribution des effectifs ou des fréquences est un mode synthétique de présentation des données. Sa constitution est immédiate dans le cas d'un caractère discret mais nécessite en revanche une transformation des données dans le cas d'un caractère continu.

3.1 Cas d'un caractère discret

X	x_1	x_2	...	x_p
n_i	n_1	n_2	...	n_p
f_i	f_1	f_2	...	f_p
n_i	n_1	n_2	...	$n_p = n$
f_i	f_1	f_2	...	$f_p = 1$

Tableau 1 - Tableau statistique

Exemple

3.2 Cas d'un caractère continu

Classes	$[a_0, a_1[$	$[a_1, a_2[$	...	$[a_{k-1}, a_k]$
Centres des classes C_i	C_1	C_2	...	C_k
n_i	n_1	n_2	...	n_k
f_i	f_1	f_2	...	f_k
$\tilde{n}_i$	$\tilde{n}_1$	$\tilde{n}_2$	...	$\tilde{n}_k = n$
$\tilde{f}_i$	$\tilde{f}_1$	$\tilde{f}_2$	...	$\tilde{f}_k = 1$

Tableau 2 – Tableau statistique

Le centre d'une classe $[a_{i-1}, a_i[$ se calcule comme :

$$C_i = \frac{a_{i-1} + a_i}{2}$$

3.2.1 Choix du nombre de classes

Lorsque le caractère étudié est quantitatif continu, le nombre de modalités est théoriquement infini. Pour permettre une étude commode de la série statistique, on regroupe en classes ces modalités, en divisant l'étendue de la série statistique par un certain nombre "k". Ce nombre, qui est simplement le nombre de classes, dépend de la taille de l'échantillon étudié. Plusieurs formules ont été proposées dans la litérature pour le calcul de k. On adoptera dans ce cours la formule suivante :

$$k \simeq 5 \times \log_{10}(n)$$

On donne les classes de manière à ce qu'elles soient fermées à gauche et ouvertes à droite (sauf éventuellement pour la dernière qui peut être des fois une classe fermée). L'amplitude des classes est donc systématiquement obtenue en divisant l'étendue de la série statistique par le nombre de classes :

$$l = \frac{E}{k} = \frac{x_{max} - x_{min}}{k}$$

Attention !

Quand on fait la répartition en classes, il faut s'assurer que toutes les observations sont incluses dans ces classes. Pour ce faire, une manière simple et naïve est de jouer sur la valeur de a_0 et/ou de a_k.

Exemple

Remarque

Dans le cas où le caractère étudié est qualitatif nominal, le tableau statistique des données donne uniquement les effectifs et/ou les fréquences. Les effectifs et fréquences cumulés n'ont pas de sens dans ce cas, contrairement au cas où le caractère est ordinal.

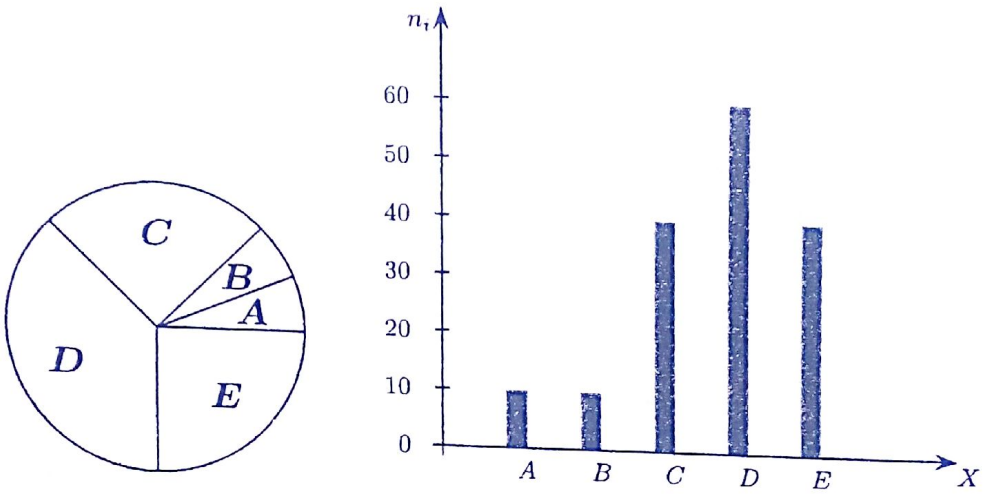
Exemples

4 Représentations graphiques

Les représentations graphiques ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution des données. Elles facilitent donc l'interprétation des données recueillies.

4.1 Caractère qualitatif

Les représentations graphiques les plus utilisées d'une série statistique correspondant à un caractère qualitatif sont le diagramme en barres et le diagramme en secteurs circulaires.



(a) *Secteurs circulaires*

(b) *Diagramme en barres*

Figure 1 – Représentations graphiques d'une série statistique à caractère qualitatif

4.2 Caractère quantitatif discret

Pour les caractères quantitatifs discrets, la représentation graphique est le **diagramme en bâtons** (Voir Figure2) où la hauteur du $i^{ème}$ bâton correspond à l'effectif n_i (resp. la fréquence f_i) associé(e) à la modalité x_i du caractère .

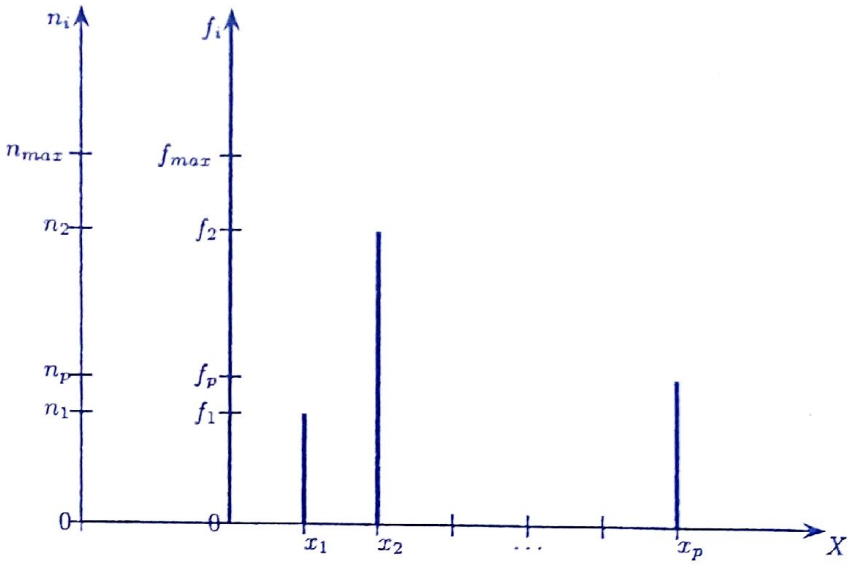


Figure 2 – Structure générale d'un diagramme en bâtons

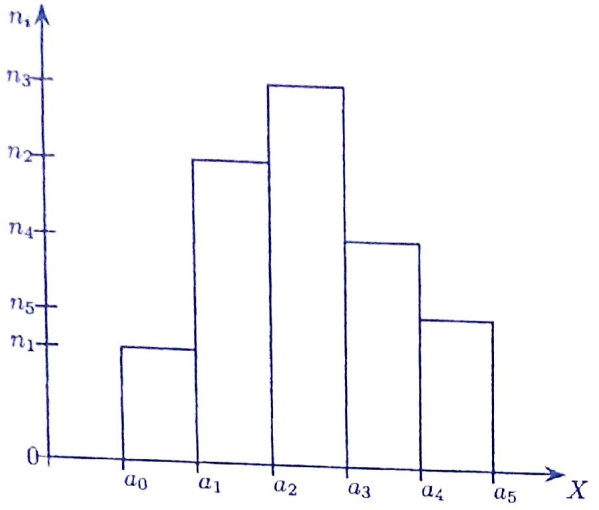
On appelle **polygone des effectifs** la ligne brisée qui joint les sommets des bâtonnets dans le diagramme. Le polygone des fréquences est identique à celui des effectifs car il s'obtient de ce dernier par un simple changement d'échelle sur l'axe des ordonnées qui consiste à remplacer les effectifs n_i par les fréquences f_i.

4.3 Caractère quantitatif continu

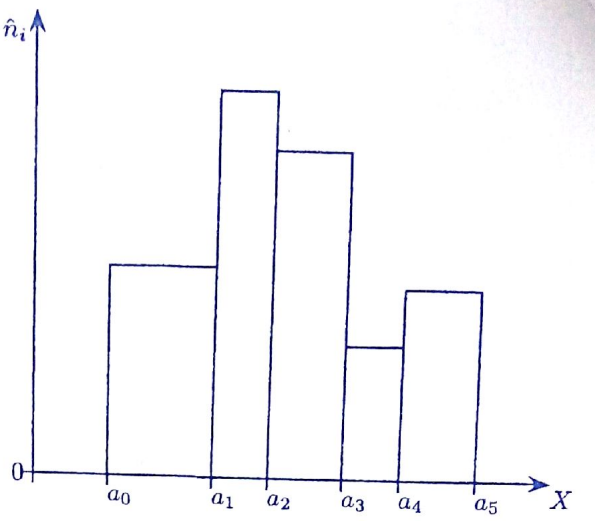
Pour les caractères quantitatifs continus, la représentation graphique est l'**histogramme** où la hauteur du rectangle est proportionnelle à l'effectif n_i (resp. à la fréquence f_i) . Ceci n'est vrai que si les intervalles $[a_{i-1}, a_i[$ $(i = 1, \ldots, k)$ sont de longueur constante. Dans ce cas l'aire comprise sous l'histogramme s'avère proportionnelle à l'effectif total.

En revanche lorsque les intervalles de classe sont inégaux, des modifications s'imposent pour conserver cette proportionnalité. Dans ce cas, en ordonnée, au lieu de porter l'effectif, on indique le rapport de la fréquence sur l'intervalle de classe. Ainsi, l'histogramme devient l'ensemble des rectangles tels que le *ième* a pour base inférieure l'amplitude de la classe $l_i = a_i - a_{i-1}$ et pour hauteur l'effectif corrigé $\hat{n}_i = \dfrac{n_i}{\hat{a}_i}$.

Avce, $\hat{a}_i = \dfrac{l_i}{\mu}$ et μ est appelé **unité d'amplidute**. Elle est en général égale à la plus petite amplitude.



(a) *longueur des intervalles égale*



(b) *longueur des intervalles inégale*

Figure 3 – Deux exemples d'histogrammes avec 5 interalles de classe

Le polygone des effectifs est la ligne brisée qui joint les milieux des bases supérieures des rectangles lorsque les classes sont de même amplitude.

5 Fonction de répartition

5.1 Cas d'un caractère discret

Les fréquences cumulées sont représentées au moyen de la fonction de répartition. Cette fonction, présentée en Figure 4, est croissante et définie de $\mathbb{R}$ dans $[0,1]$. Elle vaut :

$$F(x) = \begin{cases} 0 & si \quad x < x_1 \\ \tilde{f}_{j-1} & si \quad x_{j-1} \leq x < x_j \ ; \ j = 2,3,\ldots,p \\ 1 & si \quad x \geq x_p \end{cases}$$

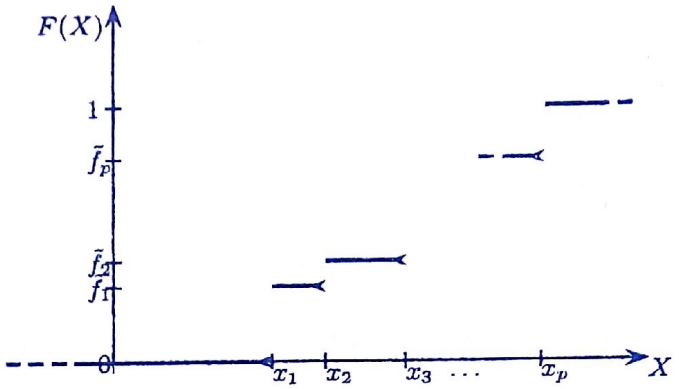


Figure 4 – Fonction de répartition d'une variable statistique discrète

$F_x(x) =$ 0 si $x < 2$
0,14 si $\leq x <$
0,2 si $\leq x <$
0,28 si $\leq x <$
0,76 si $\leq x <$
0,82 si $\leq x <$
0,92 si $\leq x <$
1 si $x \geq 8$

5.2 Cas d'un caractère continu

La fonction de répartition $F(x)$, présentée en Figure 5, est une fonction croissante définie de $\mathbb{R}$ dans $[0,1]$ et vaut :

$$F(x) = \begin{cases} 0 & si \quad x < a_0 \\ \tilde{f}_{j-1} + \frac{f_j}{a_j - a_{j-1}}(x - a_{j-1}) & si \quad a_{j-1} \leq x < a_j \ ; \ j = 1, 2, \ldots, k \\ 1 & si \quad x \geq a_k \end{cases}$$

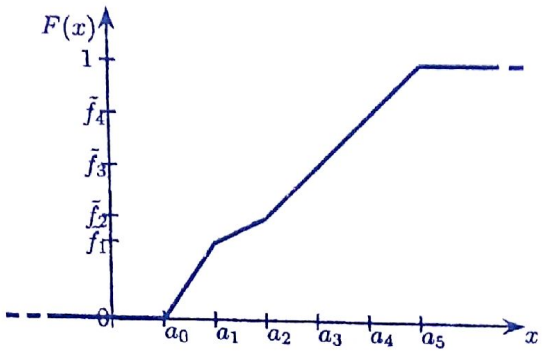


Figure 5 – Fonction de répartition d'une variable statistique continue

6 Paramètres caractéristiques d'une série statistique

6.1 Paramètres de position

6.1.1 Cas d'un caractère discret

(a) La moyenne arithmétique

Soit un échantillon de n valeurs observées X_1, X_2, ..., X_i,..., X_n d'un caractère quantitatif discret X (cet échantillon est également appelé **série statistique brute**). On définit la moyenne observée $\bar{x}$ de X comme la moyenne arithmétique des n valeurs :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Si la moyenne est calculée à partir d'un tableau d'effectifs alors

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i$$

On constate que

$$\bar{x} = \sum_{i=1}^{p} f_i x_i$$

(b) Le mode

Le mode, noté M_o, d'une série statistique discrète est donné par la valeur du caractère la plus fréquente ou dominante dans l'échantillon. Naturellement, une série statistique peut avoir plusieurs modes.

(c) Les quantiles

Soit une série statistique brute rangée dans l'ordre croissant des obsérvations $X_{(1)}$, $X_{(2)}$, ..., $X_{(i)}$,..., $X_{(n)}$. On appelle **quantile** d'ordre $\alpha \in [0,1]$ de la série statistique, le nombre réel q_α dont $100 \times \alpha\%$ des obsérvations lui sont inférieures et $100 \times (1-\alpha)\%$ lui sont supérieures :

$$q_\alpha = \begin{cases} \frac{X_{(n\alpha)} + X_{(n\alpha+1)}}{2} & si \quad n\alpha \in \mathbb{N} \\ X_{([n\alpha]+1)} & si \quad n\alpha \notin \mathbb{N} \end{cases}$$

Les principaux quantiles sont :

c.1 La médiane.

C'est le quantile d'ordre $\alpha = 0.5$:

$$M_e = q_{0.5} = \begin{cases} \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & si \quad n \ est \ pair \\ X_{([\frac{n}{2}]+1)} & si \quad n \ est \ impair \end{cases}$$

c.2 Les quartiles.

Il en existe trois. Ceux sont les quantiles d'ordre $\alpha = \frac{1}{4}$, $\alpha = \frac{2}{4} = \frac{1}{2}$ *et* $\alpha = \frac{3}{4}$:

$$Q_j = q_{\frac{j}{4}} = \begin{cases} \frac{X_{(\frac{jn}{4})} + X_{(\frac{jn}{4}+1)}}{2} & si \quad \frac{jn}{4} \in \mathbb{N} \\ X_{([\frac{jn}{4}]+1)} & si \quad \frac{jn}{4} \notin \mathbb{N} \end{cases} \quad ; \; j \in \{1,2,3\}$$

Remarquons que $Q_2 = M_e$

c.3 Les déciles.

Il en existe neuf. Ceux sont les quantiles d'ordre $\alpha = \frac{1}{10}$, $\alpha = \frac{2}{10}$ $\cdots$ $\alpha = \frac{9}{10}$:

$$D_j = q_{\frac{j}{10}} = \begin{cases} \frac{X_{(\frac{jn}{10})} + X_{(\frac{jn}{10}+1)}}{2} & si \quad \frac{jn}{10} \in \mathbb{N} \\ X_{([\frac{jn}{10}]+1)} & si \quad \frac{jn}{10} \notin \mathbb{N} \end{cases} \quad ; \; j \in \{1,2,\ldots,9\}$$

Remarquons que $D_5 = M_e$

c.4 Les centiles.

Il en existe quatre-vingt-dix-neuf. Ceux sont les quantiles d'ordre $\alpha = \frac{1}{100}$, $\alpha = \frac{2}{100}$ $\cdots$ $\alpha = \frac{99}{100}$:

$$C_j = q_{\frac{j}{100}} = \begin{cases} \frac{X_{(\frac{jn}{100})} + X_{(\frac{jn}{100}+1)}}{2} & si \quad \frac{jn}{100} \in \mathbb{N} \\ X_{([\frac{jn}{100}]+1)} & si \quad \frac{jn}{100} \notin \mathbb{N} \end{cases} \quad ; \; j \in \{1,2,\ldots,99\}$$

Remarquons que :

$$D_1 = C_{10}$$
$$D_2 = C_{20}$$
$$Q_1 = C_{25}$$
$$D_3 = C_{30}$$
$$D_4 = C_{40}$$
$$M_e = D_5 = C_{50}$$
$$D_6 = C_{60}$$
$$D_7 = C_{70}$$
$$Q_3 = C_{75}$$
$$D_8 = C_{80}$$
$$D_9 = C_{90}$$

6.1.2 Cas d'un caractère continu

(a') La moyenne arithmétique

Lorsqu'il s'agit d'une série statistique regroupée en classes $([a_{i-1}, a_i[, n_i)$; $i = 1, \ldots, k$, on considère que chaque observation X_i est représentée par le centre de classe c_i à laquelle elle appartient. La moyenne est ainsi calculée par :

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{k} n_i c_i$$

Ou encore

$$\bar{\mathbf{x}} = \sum_{i=1}^{k} f_i c_i$$

(b') Le mode

On le calcule par la formule d'interpolation linéaire suivante :
Soit $[a, b[$ la classe modale. C'est à dire celle qui correspond à l'effectif (ou fréquence) le plus élevé. Alors

$$M_o = a + \frac{\delta_p}{\delta_p + \delta_s} \times l_m$$

Avec,

δ_p= l'écart d'effectifs entre la classe modale et la classe précédente.
δ_s= l'écart d'effectifs entre la classe modale et la classe suivante.
l_m= l'amplitude de la classe modale.

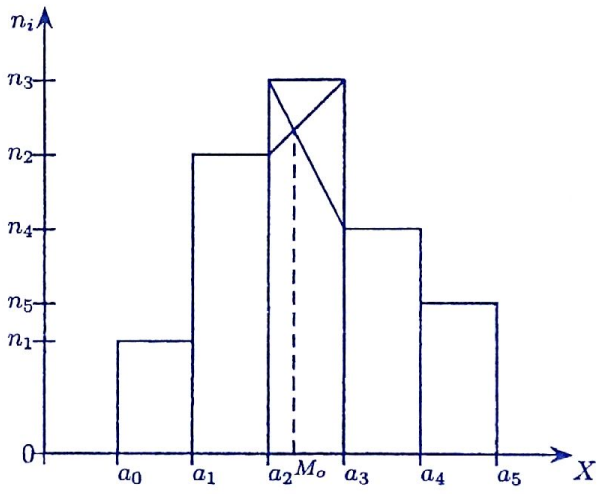


Figure 6 – Détermination graphique du mode

(c') Les quantiles

Dans le cas où la série statistique étudiée est regroupée en classes, les quantiles gardent leur définition mais se calculent d'une autre manière :
On définit d'abord la classe à laquelle appartient le quantile q_α : C'est la classe $[a_{i-1}, a_i[$ tel que $\tilde{f}_{i-1} \leq \alpha < \tilde{f}_i$ (ou bien tel que $\tilde{n}_{i-1} \leq n\alpha < \tilde{n}_i$).

q_α est alors estimé par la formule d'interpolation linéaire suivante :

$$q_\alpha = a_{i-1} + \frac{a_i - a_{i-1}}{n_i}\left(n\alpha - \tilde{n}_{i-1}\right)$$

Il en découle les expressions des principaux quantiles suivants :

c'.1. La médiane.
C'est le quantile d'ordre $\boldsymbol{\alpha = 0.5}$:

$$M_e = q_{\frac{1}{2}} = a_{i-1} + \frac{a_i - a_{i-1}}{n_i}\left(\frac{n}{2} - \tilde{n}_{i-1}\right)$$

avec $\tilde{n}_{i-1} \leq \frac{n}{2} < \tilde{n}_i$

c'.2. Les quartiles.
Ceux sont les quantiles d'ordre $\boldsymbol{\alpha = \frac{1}{4}}$, $\boldsymbol{\alpha = \frac{2}{4}}$ *et* $\boldsymbol{\alpha = \frac{3}{4}}$:

$$Q_j = q_{\frac{j}{4}} = a_{i-1} + \frac{a_i - a_{i-1}}{n_i}\left(\frac{jn}{4} - \tilde{n}_{i-1}\right) \ , \ j \in \{1, 2, 3\}$$

avec $\tilde{n}_{i-1} \leq \frac{jn}{4} < \tilde{n}_i$
Remarquons que $Q_2 = M_e$

c'.3. Les déciles.
Ceux sont les quantiles d'ordre $\boldsymbol{\alpha = \frac{1}{10}}$, $\boldsymbol{\alpha = \frac{2}{10}}$ $\boldsymbol{\ldots}$ $\boldsymbol{\alpha = \frac{9}{10}}$:

$$D_j = q_{\frac{j}{10}} = a_{i-1} + \frac{a_i - a_{i-1}}{n_i}\left(\frac{jn}{10} - \tilde{n}_{i-1}\right) \ , \ j \in \{1, 2, \ldots, 9\}$$

avec $\tilde{n}_{i-1} \leq \frac{jn}{10} < \tilde{n}_i$

Remarquons que $D_5 = M_e$

c'.4. Les centiles.
Ceux sont les quantiles d'ordre $\boldsymbol{\alpha = \frac{1}{100}}$, $\boldsymbol{\alpha = \frac{2}{100}}$ $\boldsymbol{\ldots}$ $\boldsymbol{\alpha = \frac{99}{100}}$:

$$C_j = q_{\frac{j}{100}} = a_{i-1} + \frac{a_i - a_{i-1}}{n_i}\left(\frac{jn}{100} - \tilde{n}_{i-1}\right) \ , \ j \in \{1, 2, \ldots, 99\}$$

avec $\tilde{n}_{i-1} \leq \frac{jn}{100} < \tilde{n}_i$
Les remarques faites pour les centiles dans le cas discret restent valables ici.

6.1.3 Détermination graphique d'un quantile

' Illustrée en cours '

6.1.3 Détermination graphique d'un quantile

' Illustrée en cours '

6.2 Paramètres de dispersion

6.2.1 Cas d'un caractère discret

(a) La variance

On définit la variance $V(X)$ d'une série statistique brute $X_1, X_2, \ldots, X_n$ par

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

Cette formule est appelée **formule de définition**.

On montre facilement que $V(X) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{x}^2$ appelée formule développée de la variance.

Lorsqu'il s'agit de calculer la variance à partir d'un tableau d'effectifs (ou de fréquences), la formule suivante est utilisée :

$$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2$$

(b) L'écart-type

Il est en général difficile d'utiliser la variance comme mesure de dispersion car le recours au carré conduit à un changement d'unités. Elle n'a donc pas de sens direct dans beaucoup de domaines, contrairement à l'écart-type σ_X qui est défini comme étant la racine carrée de la variance et qui s'exprime dans la même unité que la moyenne :

$$\sigma_X = \sqrt{V(X)}$$

(c) Le coefficient de variation

Il se note CV_X et est défini par :

$$CV_X = \frac{\sigma_X}{\bar{x}}$$

Le coefficient de variation permet de faire une comparaison entre deux ou pluseurs séries statistiques différentes. Il est souvent exprimé en pourcentage.

Si Remi.

6.2.2 Cas d'un caractère continu

(a') La variance

Elle est définie par

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \bar{x})^2 \;=\; \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 \;-\; \bar{x}^2$$

Remarque importante : De part sa définition même, la variance est toujours un nombre positif. Il est inacceptable et strictement interdit que sur la feuille d'examen figure une variance affectée d'une valeur négative!! Sinon...

(b') L'écart-type

C'est la racine carrée de la variance :

$$\sigma_X = \sqrt{V(X)}$$

(c') Le coefficient de variation

$$CV_X = \frac{\sigma_X}{\bar{x}}$$

6.2.3 Calcul de $\bar{x}$ et $V(X)$ par changement de variable

Etant donnée une série statistique groupée en classes à valeurs trop importantes et afin d'éviter des calculs trop voluméneux, on procède au changement de variable comme suit :
On pose $y_i = \dfrac{c_i - c^*}{l}$; $i = 1, ..., k$
avec, c^* est le centre de la classe centrale et l est l'amplitude.
Cela implique que $c_i = ly_i + c^*$; $i = 1, ..., k$
On montre alors que :

— $\bar{x} = l\bar{y} + c^*$ avec $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{k} n_i y_i$

— $V(X) = l^2 V(Y)$ avec $V(Y) = \dfrac{1}{n}\sum_{i=1}^{k} n_i y_i^2 \;-\; \bar{y}^2$

6.3 Les moments et moments centrés

1. On appelle **moment** d'ordre r ($r \geq 1$), la quantité

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

Ou bien dans le cas ou le caractère est continu

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

2. On appelle **moment centré** d'ordre r ($r \geq 1$), la quantité

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

Ou bien dans le cas ou le caractère est continu

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

Remarques

- Le moment centré de premier ordre est toujours nul.
- Le moment centré de second ordre est lui même la variance.
- Si une disribution est symétrique alors tous ses moments centrés d'ordres impairs sont nuls.
- Si un changement de variable est appliqué, le moment centré d'ordre r s'exprime alors comme suit :

$$\mu_r(X) = l^r \mu_r(Y)$$

6.4 Paramètres de forme

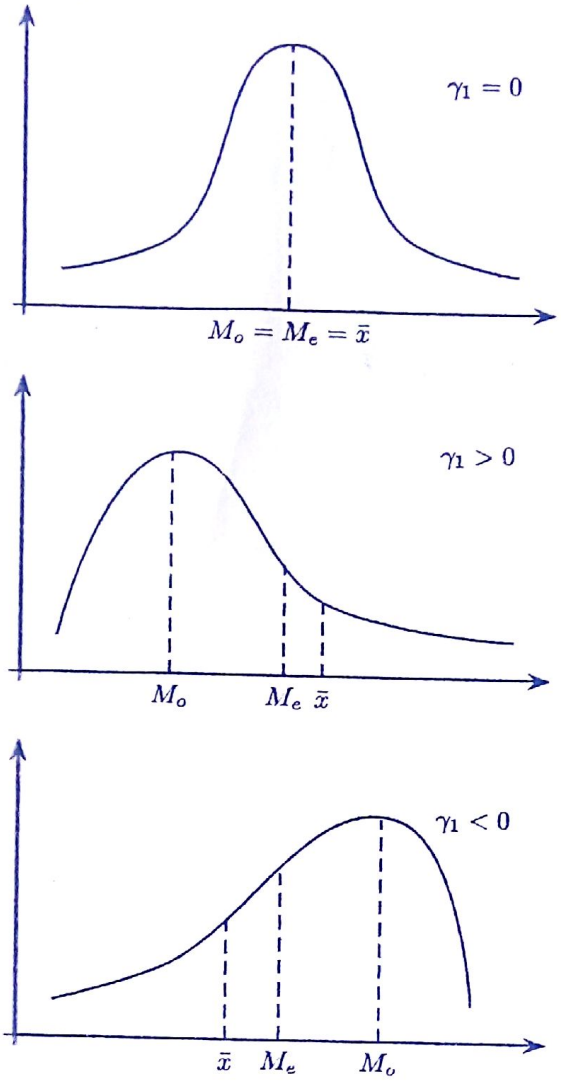
6.4.1 Le coefficient d'asymétrie

On définit dans ce cours le coefficient d'asymétrie de Fisher :

$$\gamma_1 = \frac{\mu_3}{\sigma_X^3} = \frac{\mu_3}{(\mu_3)^{3/2}} = \frac{\mu_3}{V(X)^{3/2}}$$

Comme son nom l'indique, ce coefficient permet de mesurer l'asymétrie d'une distribution statistique de la manière suivante :

- Si $\gamma_1 = 0$, la distribution est parfaitement symétrique. dans ce cas, la moyenne arithmétique, le mode et la médiane sont tous les trois confondus.
- Si $\gamma_1 > 0$, la courbe de la distribution présente une oblique à gauche avec un étalement à droite. dans ce cas, on a $M_o < M_e < \bar{x}$.
- Si $\gamma_1 < 0$, la courbe de la distribution présente une oblique à droite avec un étalement à gauche. dans ce cas, on a $M_o > M_e > \bar{x}$.



6.4.2 Le coefficient d'aplatissement

On définit également dans ce cours le coefficient d'aplatissement de Fisher :

$$\gamma_2 = \frac{\mu_4}{\sigma_X^4} - 3 = \frac{\mu_4}{V(X)^2} - 3$$

- Si γ_2 prend la valeur zéro, on dit que la distribution est normale ou Gaussienne (ou mésokurtique).
- Si γ_2 est strictement positif, on dit que la distribution est leptokurtique.
- Si γ_2 est strictement négatif, on dit que la distribution est platykurtique.