Statistique descriptive

M1 du Master MMAS

James Ledoux

Dépt de mathématiques, Univ. Poitiers

11 juillet 2009

Statistique

Ensemble de méthodes permettant d'analyser des ensembles d'observations (ou de données)

- Méthodes relevant des mathématiques :
  Calcul des probabilités, Algèbre linéaire, Optimisation, Calcul scientifique, ...
- Utilisation intensive de l'outil informatique :
  SGBD, Calcul scientifique, Calculs distribués (GRID), ...

« Deux » classes de méthodes en statistique :

1. Statistique descriptive ou **exploratoire** : contenu du cours
2. Statistique **inférentielle** : module au second semestre

Statistique exploratoire

Nettoyage des données :

- valeurs manquantes, aberrantes,
- modalités trop rares,
- incohérences, liaison non linéaire,
- ...

Pré-traitement des données

- transformations, recodages,
- réduction de dimension,
- classification, typologie des données,
- ...

Objectifs : présenter / résumer / structurer | les données

Statistique inférentielle

Statistique exploratoire

- mise en évidence des propriétés de la population étudiée
- suggérer des hypothèses

Objectif de la statistique inférentielle : étendre à une population globale des phénomènes observés sur un « échantillon »

- proposer un modèle probabiliste sur la base de la phase exploratoire
- valider ou infirmer les hypothèses émises
  - Construction d'estimateurs, d'intervalles de confiance
  - Test d'hypothèses, ...
- Prévision et prise de décision

➔ les deux démarches sont complémentaires

Une bibliographie succinte

C. Bishop.
Pattern Recognition and Machine Learning.
Springer, 2006.

A. Morineau.
Statistique exploratoire multidimensionnelle.
Dunod, 1995.

G. Saporta.
Probabilités, Analyse de Données et Statistique.
Technip, 2006.

S. Tufféry.
Data Mining et statistique décisionnelle.
Technip, 2006.

Logiciels généralistes de statistique

- SAS édité par SAS Institute : noyau de base avec une galerie de module complémentaires.
- SPSS
- Excel par Microsoft, XLstat, ...
- R hérité de S : gratuit et installable sur n'importe quelle plateforme
- Minitab, SPAD, Splus version commerciale de S, Statgraphics Centurion, ...

Une étude statistique de données :

1. Identification du groupe d'individus soumis à l'étude : l'ensemble des individus ≡ population :
   - individu au sens large : personnes, animaux, matériels, ...
2. L'étude porte sur :
   - la population complète : recensement
   - un sous-ensemble de la pop. : un échantillon
     - Théorie des sondages : méthodes pour choisir et collecter un échantillon (pas ici)
     - Objectif : un échan. représentatif de la population pour extrapoler les résultats obtenus sur l'échan.
     - La remontée à la population s'appuie sur la Stat. Inférentielle (2e semestre)
     - (Z) la notion d'échantillon a un sens précis en calcul des probabilités

3. Étude par :
   - Observation sur chaque individu d'un certain nombre de variables notées avec des lettres majuscules : $X_1, X_2, \ldots$:
   - X_j est à valeurs dans $\mathcal{E}$
     la « valeur » observée de X_j pour l'individu N° i : x_{ij}
   - les données ≡ l'ensemble des valeurs des variables observées pour l'échantillon

Exemple

- Population : l'ensemble des étudiants inscrits à l'Univ. de Poitiers
- Variables : X_1 ≡Année de naissance, X_2 ≡nationalité, X_3 ≡sexe, X_4 ≡nbre de frères et soeurs, ...

Typologie des variables

- À valeurs **numériques** : variable quantitative
  - l'ensemble des valeurs $\mathcal{E}$: domaine
  - si $\mathcal{E}$ est un intervalle : var. (quantitative) continue
  - dans le cas contraire : var. (quantitative) discrète

  Exemple : la var. taille est continue, les var. nbre de frères et soeurs ou année de naissance sont discrètes
- À valeurs **non-numériques** : var. qualitative ou catégorielle
  - chaque « valeur » est appelée une modalité
    Exemple : sexe avec deux modalités (m,f)
    - naturellement ordonnée : ordinale (mention au bac)
    - dans le cas contraire : nominale

  sexe (m,f) : après un recodage binaire, $\mathcal{E} = \{0, 1\}$, mais le caractère numérique est artificiel.

1. Une seule variable :

$$X = (x_1, \ldots, x_n)^\top \quad \text{Vecteur colonne}$$

Statistique descriptive unidimensionnelle

2. Au moins deux variables ($p \geq 2$) :

$$X = (x_{ij})_{i=1,j=1}^{i=n,j=p} \quad \text{Matrice rectangulaire : } n \times p$$

   - lignes : représentent les n individus ou unités statistiques
   - colonnes : représentent les p variables

Données sont identifiées à un tableau rectangulaire noté X

Statistique descriptive multidimensionnelle

3. Cette table est le concept de base dans le logiciel SAS (pour les TP)

L'ensemble $\mathcal{E} = \{m_1, \ldots, m_k\}$ des k modalités de la variable

Cette variable a été observée sur un échantillon de n individus

Définition 1

*On appelle **fréquence absolue** de la modalité m_j le nombre total (**effectif**) n_j d'individus de l'échantillon pour lesquels la variable a pris la modalité m_j :*

$$n_j := \sum_{i=1}^{n} 1_{m_j}(x_i)$$

*On appelle **fréquence relative** de la modalité m_j, la proportion d'individus à présenter cette modalité*

$$f_j := \frac{n_j}{n}$$

Exemple 1 (Couleur des cheveux)

On observe la couleur des cheveux de 592 femmes.
4 modalités ont été retenues : $\mathcal{E} = \{brune, châtain, rousse, blonde\}$.
L'information est résumée dans le tableau suivant :

Modalité	brune	châtain	rousse	blonde	
n_j	108	286	71	127	592
f_j **(arrondi)**	0.1824	0.4831	0.12	0.2145	$\sum_j f_j = 1$
$f_j * 100$	18.24	48.31	12	21.45	100

La 2e ligne fait apparaître une loi de probabilité sur l'ensemble des modalités $\mathcal{E}$
Le vecteur des fréquences relatives est parfois appelé le **profil** de la variable

Exemple 2 (Élection européenne 2004)

Les résultats de l'élection européenne de 2004 sont directement présentés sous la forme d'une table des « fréquences relatives »

Listes	Ext. G.	PC	PS	Verts	Div. G	Div. D.	UDF	UMP	FN	Ext. D.
% **des voix**	3.3	5.2	28.9	7.4	5.9	10.7	11.9	16.6	9.8	0.3

Diagramme en colonnes ou en bâtons ou « bar-plot »

À chaque modalité m_j, on associe un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative f_j

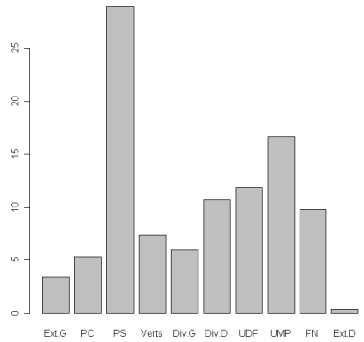


FIGURE 1: Diagramme en bâtons pour les résultats de l'élection

Diagramme en camembert ou « pie-chart »

À chaque modalité m_j, on associe un secteur de disque dont l'aire (ou l'angle au centre) est proportionnelle à la fréquence relative f_j

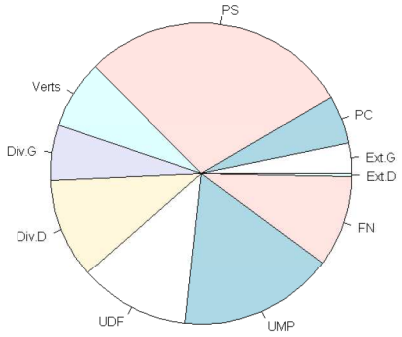


FIGURE 2: Diagramme en camembert pour les résultats de l'élection

Variable quantitatives discrètes

- Lorsque le nombre de valeurs prises par la variable est « modeste » : les mêmes représentations à l'aide des fréquences absolues et relatives
- Pour une représentation graphique, le diagramme en bâtons est pertinent mais pas le camembert car la différence fondamentale avec une variable qualitative est que l'ensemble des valeurs (réelles) est naturellement ordonnées

Exemple 3 (Nombre d'enfants)

Dans le cadre d'une enquête auprès de 1000 couples, on a collecté le nombre d'enfants du foyer.

Nbre d'enfants	0	1	2	3	4	5	6
n_j	235	183	285	139	88	67	3
f_j	0.235	0.183	0.285	0.139	0.088	0.067	0.03

Le vecteur $(f_1, \ldots, f_7)$ définit une probabilité sur $\mathcal{E} = \{0, 1, \ldots, 6\}$

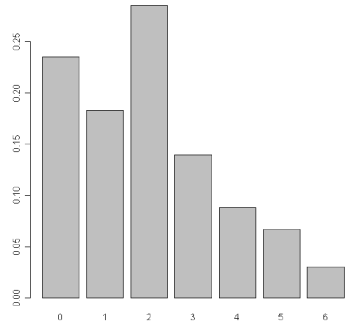


FIGURE 3: Diagramme en bâtons du nbre d'enfants pour 1000 couples

- Choix d'un modèle probabiliste : pas une loi géométrique, étudier loi binomiale, Poisson, ...
- Pb du choix du paramètre des lois candidates : $p = ?$, $\lambda = ?$,...

Si la variable X admet p valeurs distinctes $\{e_1, \dots, e_p\}$:

Définition 2 (Fréquence absolue cumulée)

$N_0 := 0$, *pour* $j = 1, \dots, p$ $N_k := \sum_{j=1}^{k} n_j$ *et* $N_k = n$ *pour* $k \geq p$

Définition 3 (Fréquence relative cumulée)

$F_0 := 0$, *et pour* $j = 1, \dots, p$ $F_k := \sum_{j=1}^{k} f_j$ *et* $F_k = 1$ *pour* $k \geq p$

À partir des ces fréquences cumulées, on peut alors tracer la **fonction de répartition empirique** correspondante :

$$F_X(x) := \begin{cases} 0 & \text{si } x < e_1 \\ F_k & \text{si } e_j \leq x < e_{j+1} \\ 1 & \text{si } x \geq e_p \end{cases} \tag{1}$$

qui est ici une fonction en escalier continue à droite.

Variables quantitatives continues

- En général, le diagramme en bâton est inutile car chaque valeur apparaît une seule fois dans la série ($n_j = 1$)
- Pour les deux représentations graphiques traditionnelles, on a besoin d'ordonner les données de l'échantillon :

Définition 4 (Statistique d'ordre)

Pour un échantillon $X = (x_1, \dots, x_n)^\top$*, on range la liste des valeurs dans l'ordre croissant :* $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$*. Le vecteur* $X_{()} = (x_{(1)}, \dots, x_{(n)})^\top$ *est appelée la* ***statistique d'ordre*** *de l'échantillon* X

Exemple 4 (Ampoule)

On a mesuré la durée de vie en h de 10 ampoules identiques :

$$\begin{array}{rccccccccccc} X = (& 91.6 & 35.7 & 251.3 & 24.3 & 5.4 & 67.3 & 170.9 & 9.5 & 118.4 & 57.1 &)^\top \\ X_{()} = (& 5.4 & 9.5 & 24.3 & 35.7 & 57.1 & 67.3 & 91.6 & 118.4 & 170.9 & 251.3 &)^\top \end{array}$$

Histogramme

1. Choisir $a_0 < x_{(1)}$ et $x_{(n)} < a_k$: l'intervalle $]a_0, a_k]$ recouvre l'ensemble des valeurs prise par la variable.
2. Faire une partition de $]a_0, a_k]$ en k intervalles $]a_{j-1}, a_j]$:
   - chaque intervalle $]a_{j-1}, a_j]$ s'appelle une **classe**
   - la **longueur de la classe** $]a_{j-1}, a_j]$ vaut $h_j := a_j - a_{j-1}$
   - Si toute les classes ont même longueur alors on construit un **histogramme à pas fixe**. Dans le cas contraire, on parle d'un **histogramme à pas variable**
   - On appelle **effectif de la classe**, le nombre de valeurs de l'échantillon contenues dans la classe

$$n_j := \sum_{i=1}^{n} 1_{]a_{j-1}, a_j]}(x_i)$$

et $f_j := n_j/n$ la **fréquence de la classe**

Définition 5

L'histogramme est la figure constituée des rectangles dont les ***bases*** *sont les classes et dont les* ***aires*** *sont égales aux fréquences de ces classes*

- La construction d'un histogramme dépend des paramètres k, a_0, a_k et des h_j. Les variations de ces paramètres peuvent entraîner des fortes variations dans « l'allure » de l'histogramme
- Quelques « conseils » standards :
  - Il est recommandé d'avoir entre 5 et 20 classes. **La règle de Sturges** préconise de choisir
    $$k \approx 1 + \log_2 n = 1 + \ln n / \ln 2.$$
  - Le choix des bornes a_0 et a_k doit respecter une certaine homogénéité dans les longueurs des classes. Un choix classique est
    $$a_0 := x_{(1)} - 0.025(x_{(n)} - x_{(1)}) \quad a_k := x_{(n)} + 0.025(x_{(n)} - x_{(1)})$$

Exemple 5 (Ampoule)

La règle de Sturges $n = 10$: $k = 5$ classes.
Comme $x_{(1)} = 5.4$ et $x_{(10)} = 251.3$, on obtient via la règle précédente : $a_0 = -0.747$ et $a_5 = 257.4$, qu'on peut arrondir à $a_0 = 0$ et $a_5 = 260$.
Pour un histogramme à pas fixe avec 5 classes, on a $h := 260/5 = 52$.

exemple suite

Classes	$]0,52]$	$]52,104]$	$]104,156]$	$]156,208]$	$]208,260]$
Effectifs n_j	4	3	1	1	1
Fréq. f_j	0.4	0.3	0.1	0.1	0.1
Haut. f_j/h	0.0077	0.0058	0.0019	0.0019	0.0019

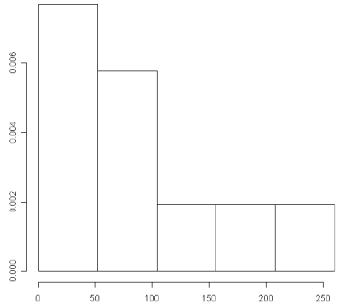


FIGURE 4: Ampoule, histogramme à pas fixe

- sur le précédent hist., fort déséquilibre entre les effectifs de classes : alternative à l'hist. à pas fixe, construire un **histogramme à effectifs de classe constants**

exemple suite : $n_j = 2$ et $f_j = 0.2$

Classes	$]0,17]$	$]17,46]$	$]46,79]$	$]79,145]$	$]145,260]$
Larg. h_j	17	29	33	66	115
Haut. f_j/h_j	0.0118	0.0069	0.0061	0.003	0.0017

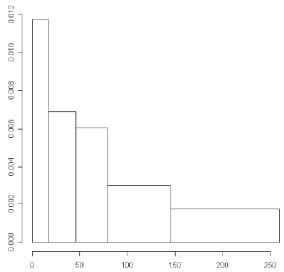


FIGURE 5: Ampoule, histogramme à effectifs de classe fixes $n_j := 2$

Polygone des fréquences

1. Tracer la ligne brisée reliant les milieux des sommets des rectangles
2. Prolonger cette ligne de part et d'autre des bornes de l'histogramme de sorte que l'aire sous le polygone soit égale à 1 (comme une densité de probabilité)

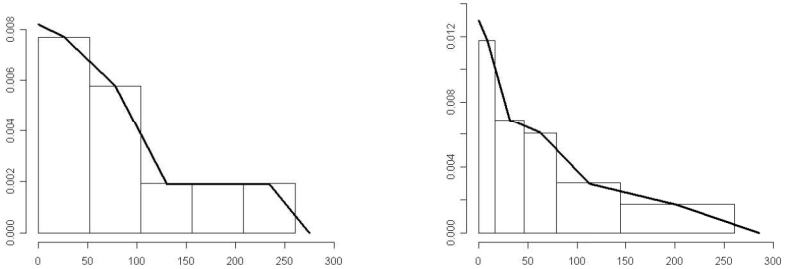


FIGURE 6: Ampoule : Polygones sur histogrammes à h_j puis n_j constant

Histogramme comme estimation de la densité

- Soit $\widehat{f}$ la fonction en escalier correspondant à un histogramme :

$$\widehat{f}_X(x) = f_j/h_j \quad x \in]a_{j-1}, a_j]$$

Alors

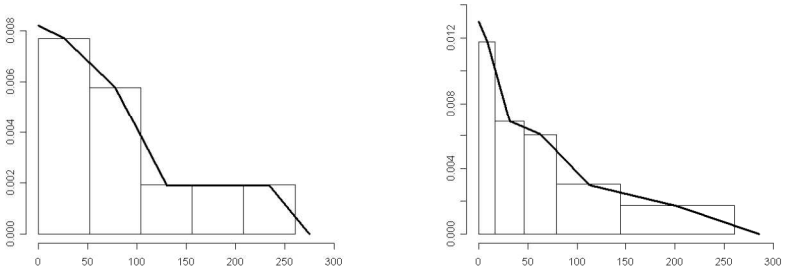
$$\text{Aire du rectangle } j = f_j = \int_{a_{j-1}}^{a_j} \widehat{f}_X(x)\,dx$$

La fraction f_j d'observations dans l'intervalle $]a_{j-1}, a_j]$ apparaît comme une estimation raisonnable de la « probabilité »

$$\mathbb{P}\{X \in]a_{j-1}, a_j]\} = \int_{a_{j-1}}^{a_j} f_X(x)\,dx$$

qu'une observation appartienne à cette classe

- Un histogramme fournit une estimation de la densité des observations ☛ proposer un modèle de densité, ...



Ampoule : h_j puis n_j constant

- L'histogramme à effectifs de classe constant décrit toujours plus finement la distribution
- Histogrammes distincts pour les mêmes données : relativiser leur pouvoir d'estimation de la densité

  ⇨ Donne une allure générale de cette densité

  C'est encore mieux avec le polygone.

- Ici : pas une loi gaussienne, une loi exponentielle ?

$$f_X(x) = \lambda \exp(-\lambda) 1_{\mathbb{R}_+}(x) \quad \lambda = ?$$

- Faire appel à d'autres outils que l'histogramme ou le polygone : courbe plus régulière

  méthodes d'estimation de densité : méthode du noyau

- Par exemple :

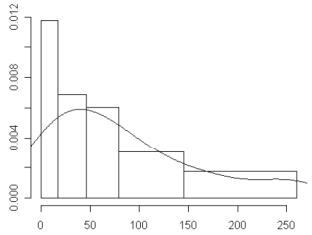


FIGURE 7: Ampoule et estimation de densité par noyau

- Si on remplace les effectifs de classe n_j par les effectifs cumulés :

$$N_k := \sum_{j=1}^{k} n_j$$

⇨ Histogramme cumulé et polygone des fréquences cumulés

- Estimation de la **fonction de répartition** de la variable étudiée

Définition 6 (Fonction de répartition empirique)

Pour un échantillon $x_1, \ldots, x_n$ donné, il s'agit de la fonction F_n définie par

$$\forall x \in \mathbb{R}, \quad F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \leq x\}} = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \frac{i}{n} & \text{si } x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{si } x \geq x_{(n)} \end{cases}$$

Autrement dit $F_n(x)$ est la fraction d'observations inférieures à x

- $F(x)$, avec F fonction de répartition d'une loi de probabilité, représente la probabilité pour qu'une observation soit inférieure à x
- $F_n(x)$ apparaît comme une estimation naturelle de $F(x)$
- On peut montrer que cette estimation est d'excellente qualité lorsque la taille de l'échantillon n est assez grande
  (Th. Glivenko-Cantelli, 2e semestre)
- F_n représente la fonction de répartition de la **mesure empirique** $\mathbb{P}_n$ (mesure de proba. sur $\mathbb{R}$) :

$$\mathbb{P}_n(\cdot) := \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i\}}(\cdot) \tag{2}$$

Exemple 6

Cette définition coïncide avec la formule (1) données pour les variables quantitatives discrètes

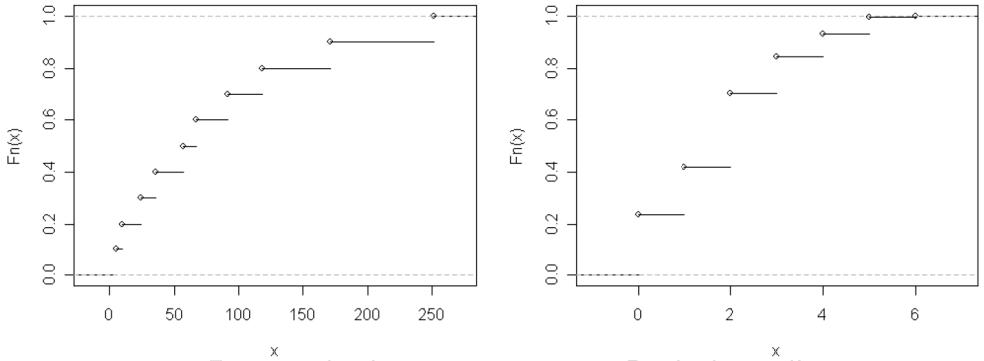


FIGURE 8: Fonction de répartition empirique : Durée de vie d'une ampoule et Nbre enfants par couple

Indicateurs « statistiques »

Pour une variable quantitative X, compléter la description des données $x_1, \ldots, x_n$ par des « résumées » numériques :

Indicateurs de localisation

- Un indicateur de localisation c se veut être un résumé des données
- Choisir un critère d'erreur commise $J(c)$ en remplaçant la série de donnée par c :
  - Quantifier l'erreur locale commise ≡ pour chaque observation :

$$d(x_i, c) \quad \text{pour une certaine « distance » } d$$

  - Puis prendre un critère global du type :

$$J(c) := \frac{1}{n}\sum_{i=1}^{n} d(x_i, c)$$

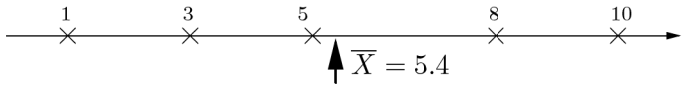
  - Minimiser la fonction $c \mapsto J(c)$ ou encore $c \mapsto J(c) := -\sum_{i=1}^{n} d(x_i, c)$

1er choix : $d(x_i, c) := (x_i - c)^2$
Une étude élémentaire de fonction donne alors comme unique minimum de $J(\cdot)$:

Définition 7 (Moyenne empirique)

(3)
$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Géométriquement : $\overline{X}$ centre de gravité des n points $\{x_i\}_{i=1}^n$ affectés du même poids $1/n$



- Sensibilité aux valeurs aberrantes

- **Une généralisation** : au lieu d'un poids uniforme de chaque individu dans la définition du critère $J(\cdot)$, on choisit une famille de poids $(p_1, \ldots, p_n)$ et alors le critère devient

$$J(c) := \sum_{i=1}^{n} p_i d(x_i, c)$$

et on obtient

(4)
$$c_{\min} = \sum_{i=1}^{n} p_i\, x_i$$

- $\overline{X}$ ou $c_{\min}$ sont appelées le résumé de l'échantillon au sens des « moindres carrées »

- Rappelons que

$$\mathbb{P}_n(f) := \int f(x) d\mathbb{P}_n(x) = \frac{1}{n}\sum_{i=1} f(x_i)$$

D'où $\overline{X} = \mathbb{P}_n(f)$ avec $f(x) = x$ et représente donc le moment d'ordre 1 de la mesure empirique

- « Alors sans surprise » : une propriété fondamentale de linéarité : pour $a, b \in \mathbb{R}$

(5) $$\overline{aX + b} = a\overline{X} + b$$

ou plus généralement pour deux variables X_1 et X_2 observées sur le même ensemble d'individus :

(6) $$\overline{aX_1 + bX_2} = a\overline{X}_1 + b\overline{X}_2$$

Se vérifient trivialement « à la main »

Cas des données groupées

1. Si comme pour les données sur la variable discrète de l'Ex. 3 on dispose seulement d'un tableau d'effectifs ou de fréquences

Nbre d'enfants	0	1	2	3	4	5	6
n_j	235	183	285	139	88	67	3
f_j	0.235	0.183	0.285	0.139	0.088	0.067	0.03

alors $$\overline{X} = \frac{1}{n}\sum_{j=1}^{k} n_j\, x_j^* = \sum_{j=1}^{k} f_j\, x_j^*$$

si on a observé k valeurs différentes $\{x_1^*, \ldots, x_k^*\}$ de la variable
Si on interprète f_j comme la probabilité pour que la variable prenne la valeur x_j^* alors

(7) $$\overline{X} \equiv \sum_{j=1}^{k} \mathbb{P}\{X = x_j^*\}\, x_j^* = \mathbb{E}[X]$$

Les propriétés de linéarité de l'espérance mathématique se transposent ainsi aux versions empiriques et on retrouvent les formules (5) ou (6)

2. Pour une var. continue, la seule donnée d'un tableau du type :

Classes	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
Effectifs n_j	4	3	1	1	1
Fréq. f_j	0.4	0.3	0.1	0.1	0.1
Haut. f_j/h	0.0077	0.0058	0.0019	0.0019	0.0019

ne permet pas d'obtenir de manière exacte la moyenne empirique

On peut alors utiliser la formule (7) en prenant, par exemple, le milieu de chaque classe $(a_{j-1} + a_j)/2$ comme valeur x_j^* de la variable

Pour le tableau ci-dessus, on obtiendrait :

$$\overline{X} \overset{\text{exacte}}{=} 83.15 \qquad \overline{X} \overset{(7)}{\approx} 88.4$$

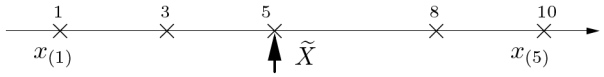
2e choix : $d(x_i, c) := |x_i - c|$

Définition 8 (Médiane empirique)

*La **médiane empirique**, notée $\widetilde{X}$ ou $\widetilde{X}_{1/2}$, est définie comme un réel partageant l'échantillon en deux groupes de même effectif.*

1. *Si n est impair alors la médiane est l'observation au centre de l'échantillon ordonné :* $\widetilde{X} := x_{((n+1)/2)}$
2. *Si n est pair alors on peut choisir le milieu de l'intervalle* $]x_{(n/2)}, x_{(n/2+1)}[$:

$$\widetilde{X} := \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$



- La moitié des observations sont supérieures (inférieures) à $\widetilde{X}$.

- **Données groupées** : utiliser une interpolation linéaire. Choisir la classe $]a_{j-1}, a_j]$ telle que $F_j \geq 1/2$ puis :

$$\widetilde{X}_{app} := a_{j-1} + \frac{h_j}{f_j}(1/2 - F_{j-1}) \quad (8)$$

- Indicateur peu sensible aux valeurs aberrantes :

Données 1					Données 2				
1	3	5	8	10	1	3	5	8	10000
$\overline{X} = 5.4$, $\widetilde{X} = 5$					$\overline{X} = 2003.4$, $\widetilde{X} = 5$				

- Pour l'exemple des ampoules :

$$\widetilde{X} = (57.1 + 67.3)/2 = 62.2 << 83.15 = \overline{X}$$

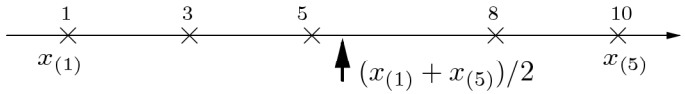
Une ampoule sur deux est tombée en panne avant 62.2h malgré une durée de vie moyenne de 83.15.
Ainsi une petite partie des ampoules aura une durée de vie largement supérieure à la majeure partie des autres
⇨ répartition non symétrique et « queue lourde »

Autre choix : $J(c) = \sup_{i=1,\dots,n} |x_i - c|$

Moyenne des valeurs extrêmes

$$\frac{x_{(1)} + x_{(n)}}{2}$$

- Géométriquement : point milieu du domaine de variation $[x_{(1)}, x_{(n)}]$ des observations



- Sensibilité aux valeurs aberrantes

1 Moindres carrés

La moyenne minimise l'erreur

$$J(c) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{(x_i - c)^2}_{\text{écart à } c}$$

D'où $J(c_{\min})$ on peut choisir comme indicateur de dispersion autour de $c_{\min}$

Définition 9 (Variance empirique)

La valeur positive

$$s_X^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 \quad (9)$$

est appelée la variance empirique de l'échantillon.

- Comme pour la moyenne, on peut interpréter s_X^2 comme

$$\int (x - \overline{X})^2 d\mathbb{P}_n(x) = \int (x - \mathbb{P}_n(f))^2 d\mathbb{P}_n(x) \quad \text{avec } f(x) = x.$$

moment centré d'ordre 2 de la mesure empirique $\mathbb{P}_n$
ou dans le cas de données groupées pour une variable discrète :

$$s_X^2 := \sum_{j=1}^{k} f_j \, (x_j^* - \overline{X})^2 \equiv \sigma^2(X)$$

- D'où les formules (se vérifient trivialement « à la main »)

$$s_{aX+b}^2 = a^2 \, s_X^2 \quad s_X^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \overline{X}^2 = \overline{X^2} - \overline{X}^2 \quad (10)$$

- Moralement : toute démonstration de formule à partir de calculs d'espérance de variables aléatoires admet sa contrepartie en « empirique »

- **L'écart-type empirique** est défini par $s_X := \sqrt{s_X^2}$ (même unité que les données)
- La dispersion doit toujours se comparer à la valeur moyenne : une dispersion de 10 pour $\overline{X} = 12$ ou $\overline{X} = 200$ n'a pas la même signification

Définition 10 (Coefficient de variation empirique)

*Le **coefficient de variation empirique** est le rapport écart-type sur moyenne :*

$$CV := \frac{s_X}{\overline{X}} \tag{11}$$

  - Il s'agit d'un indicateur sans dimension
  - En pratique, on utilise le seuil de 0.15 pour établir une faible/forte variabilité

2. L'**étendue** est la différence entre les deux valeurs extrêmes : $E := x_{(n)} - x_{(1)}$
   - C'est un indicateur très sensible aux valeurs aberrantes « outliers »
   - Utilisé en contrôle de qualité pour détecter de tels « outliers »

3 Quantile

Définition 11 (Quantile)

Pour $\alpha \in]0,1[$, le quantile d'ordre α de F_n, noté $Q(\alpha)$, est défini comme la valeur réelle vérifiant

$$Q(\alpha) := \inf\{x \in \mathbb{R} \mid F_n(x) \geq \alpha\}. \tag{12}$$

- La fonction $\alpha \mapsto Q(\alpha)$ est appelée la fonction quantile associée à la fonction de répartition F_n.
- Elle est également connue sous le nom d'inverse généralisé, ici de la fonction F_n. Voir les techniques de simulation de lois de probabilité
- **Calcul de la fonction quantile à partir d'un échantillon**

$$\forall \alpha \in \left]\frac{i-1}{n}, \frac{i}{n}\right], \quad Q(\alpha) = x_{(i)}$$

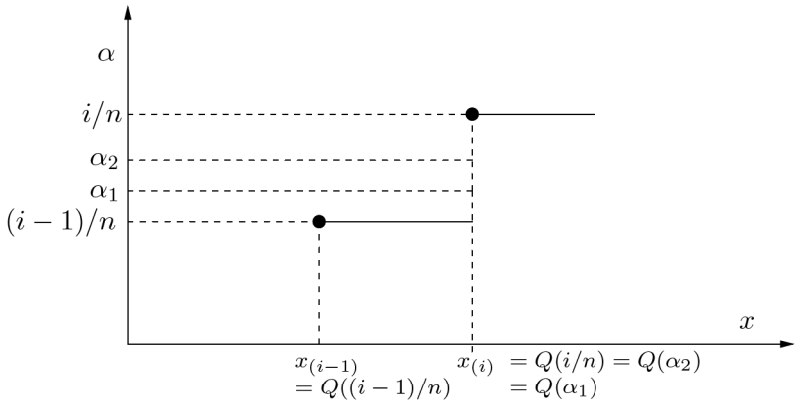


FIGURE 9: Fonction quantile pour F_n

ou encore

$$\forall \alpha \in]0,1[, \quad Q(\alpha) = \begin{cases} x_{([n\alpha])} & \text{si } n\alpha \text{ est un entier} \\ x_{([n\alpha]+1)} & \text{si } n\alpha \text{ n'est pas un entier} \end{cases} \tag{13}$$

où $[\cdot]$ désigne la fonction partie entière.

- **Calcul de $Q(\cdot)$ à partir de données groupées** : comme pour la médiane (voir (8)) utiliser une interpolation linéaire.

  Choisir la classe $]a_{j-1}, a_j]$ telle que $F_j \geq \alpha$ puis :

$$\widetilde{X}_{app} := a_{j-1} + \frac{h_j}{f_j}(\alpha - F_{j-1}) \tag{14}$$

- **Quantiles empiriques** : valeurs partageant l'échantillon ordonné en un certain nombre de parties de même effectif
  - en deux parties : la médiane (ou $Q(1/2)$)
  - en quatre parties : les quartiles, notés $Q(1/4)$, $Q(1/2)$, $Q(3/4)$
  - en dix parties : les déciles , notés $Q(1/10)$, ..., $Q(9/10)$
  - ...

Définition 12 (Quantile empirique)

*Pour $\alpha \in]0,1[$, le **quantile empirique** d'ordre α, noté $Q(\alpha)$, est défini par*

$$Q(\alpha) := \begin{cases} \frac{x_{(n\alpha)}+x_{(n\alpha+1)}}{2} & \text{si } n\alpha \text{ est un entier} \\ x_{([n\alpha]+1)} & \text{si } n\alpha \text{ n'est pas un entier} \end{cases} \quad (15)$$

- Quand $n\alpha$ n'est pas un entier, les formules (15) et (13) coïncident. Si $n\alpha$ est un entier alors la formule (15) revient à prendre le point milieu de l'intervalle $[x_{(n\alpha)}, x_{(n\alpha+1)}]$ au lieu de la valeur théorique $x_{(n\alpha)}$.

- La largeur, $Q(3/4) - Q(1/4)$, de l'intervalle $[Q(1/4), Q(3/4)]$ est dite la **distance inter-quartile** qui est insensible aux valeurs aberrantes

Exemple 7 (Ampoules)

On avait déjà la médiane : $Q(1/2) = 62.2$
Les quartiles : $Q(1/4) = x_{(3)} = 24.3$ et $Q(3/4) = x_{(8)} = 118.4$
distance inter-quartile : 94.1

Exemple 8 (VaR)

En finance, la variabilité dans les séries de données est appelée **la volatilité**. Son étude est à la base de l'analyse de risque financier. Les quantiles empiriques sont très utilisés pour étudier les phénomènes extrêmes. En finance, la **value at Risk (VaR)** est la plus utilisée des mesures de risque de marché. Elle représente la perte potentielle maximale d'un investisseur sur la valeur d'un portefeuille d'actifs, compte-tenu d'un horizon de détention et d'un niveau de confiance $1-\alpha$ donnés. En général, l'obtention de la VaR est celle d'un quantile d'ordre α de la distribution des variations de valeurs du portefeuille sur la période donnée.

- Représentation graphique : **la boîte à moustaches** ou « box-plot »

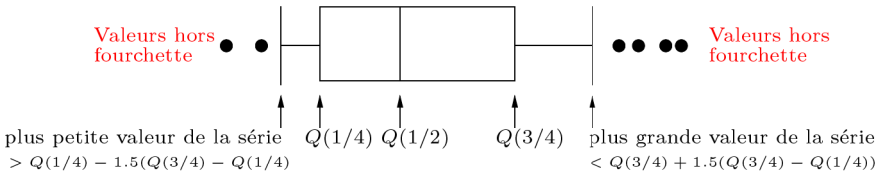


FIGURE 10: Principe de construction du box-plot

La boîte correspond à l'intervalle de valeurs $[Q(1/4), Q(3/4)]$.
On adjoint la valeur de la médiane Q(1/2) (et parfois celle de la moyenne)

4 **Caractéristiques de forme**

- **Le coefficient d'asymétrie** ou de « skewness » (Fisher)

$$\gamma_1 := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{s_X}\right)^3$$

  - Il vaut 0 pour une série dont la répartition est symétrique (e.g. loi normale), > 0 si la queue de la distribution est à droite (e.g. loi de exponentielle) et < 0 si la queue de la distribution est à gauche

- **Le coefficient d'aplatissement** ou de « kurtosis » (Fisher)

$$\gamma_2 := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{s_X}\right)^4 - 3$$

  - Sert à comparer la concentration des valeurs à celle d'une loi normale centrée pour laquelle kurtosis$= 3$
  - Si > 0 alors plus concentrée ou plus pointue que la gaussienne, si < 0 alors plus aplatie que la gaussienne (en général)

- Tous sont sensibles aux valeurs extrêmes de la distribution comme la moyenne et la variance

Statistical Analyses Software

- Procédures :
  (a) graphiques : GCHART/PIE,VBAR,HBAR GPLOT/PLOT, ...
  (b) statistiques : MEANS, UNIVARIATE, SUMMARY, BOXPLOT
  ...
- Les calculs de la variance, des coefficients d'asymétrie et d'aplatissement sont réalisés à l'aide de formules quelque peu différentes de celles reportées ici (explication plus tard)
- la notion d'histogramme dans SAS n'est pas celle utilisée en mathématique (en fait des diagrammes en bâton) : aucune procédure de base ne réalise un histogramme traditionnel à partir de données brutes
  - pas de réel pb pour des histogrammes à pas fixe mais pour les autres....
- Référence de base : poly collectif stocké à l'adresse http ://www-math.sp2mi.univ-poitiers.fr/∼jledoux/saspdf.pdf