Module: PROBABILITÉ ET STATISTIQUE 2016-17/SVI

Par Dr Abdelkrim MERBOUHA
Université Moulay Slimane
Faculté POLYDSCIPLINAIRE , BéniMellal

Plan du cours

- Ch 1: Statistiques descriptives et représentation graphique
- Ch 2: Probabilités
- Ch 3: Quelques distributions usuelles
- Ch 4: Échantillonnage et estimation des paramètres
- Ch 5: Tests d'hypothèses
- Ch 6: Introduction aux cartes de contrôle

Avant propos

La statistique est à la fois une science formelle, une méthode et une technique. Elle comprend la collecte, l'analyse, l'interprétation de données ainsi que la présentation de ces ressources afin de les rendre compréhensibles de tous, et en tirer des conclusions judicieuses.

Un peu d'histoire

Bien que le nom de *statistique* soit relativement récent – on attribue en général l'origine du nom au XVIIIe siècle de l'allemand *Staatskunde*

– cette activité semble exister dès la naissance des premières structures sociales. D'ailleurs, les premiers textes écrits retrouvés étaient des recensements du bétail, des informations sur son cours et des contrats divers.

On a ainsi trace de ***recensements*** en Chine au XXIIIe siècle av. J.-C. ou en Égypte au XVIIIe siècle av. J.-C.. Ce système de recueil de données se poursuit jusqu'au XVIIe siècle. En Europe, le rôle de collecteur est souvent tenu par des guildes marchandes, puis par les intendants de l'État.

Ce n'est qu'au XVIIIe siècle que l'on voit apparaître le rôle prévisionnel des statistiques avec la construction des premières tables de ***mortalité***.
Antoine Deparcieux écrit en 1746 l'*Essai sur les probabilités de la durée de vie humaine*. Elle va d'abord servir aux compagnies d'assurances sur la vie qui se créent alors.

La première application industrielle des statistiques eut lieu lors du ***recensement*** américain de 1890, qui mit en œuvre la carte perforée inventée par le statisticien Herman Hollerith. Celui-ci avait déposé un brevet au ***bureau américain des brevets***.

Une Définition préalable : Les statistiques sont le produit des analyses reposant sur l'usage de la statistique. Cette activité regroupe trois principales branches :

- *la collecte des données ;*
- *le traitement des données collectées, aussi appelé la statistique descriptive ;*
- *l'interprétation des données, aussi appelée l' inférence statistique, qui s'appuie sur la théorie des sondages et la statistique mathématique.*

Ch 1: Statistiques descriptives et représentation graphique

1. Notions fondamentales

1.1 Individu, population et échantillon

Individu: c'est l'unité sur laquelle on observe une ou plusieurs caractéristiques. Par exemple : personne, ville, plante, machine, etc.

Population: c'est l'ensemble des individus que l'on veut étudier, qui ont des propriétés communes et pour lesquelles on veut obtenir de l'information.

Échantillon: c'est un sous-ensemble de la population, soit la partie qu'on va examiner.

Taille: on appelle taille d'un échantillon le nombre d'individus dans cet échantillon. On le note par *n.*

Choix d'échantillons

La règle est que l'échantillon doit être représentatif de la population dont il est extrait, pour se faire, on procède par

- Echantillonnage aléatoire: les individus d'une population ont la même chance d'être sélectionné.
- Echantillonnage aléatoire simple: de n sujet est choisi de telle façon que chaque échantillon de taille n ait la même chance d'être choisi.

Echantillonnage systématique: On choisi un point de départ et on sélectionne chaque k ième élément de la population.

Echantillonnage opportun: On collecte simplement les résultats qui sont faciles à obtenir.

Echantillonnage stratifié: On subdivise la population en au moins deux sous-groupes différents (strates) dont les individus partagent les mêmes caractéristiques (comme le genre, la classe d'âge) et on tire un échantillon dans chaque sous groupes.

Echantillonnage en grappes: La population est subdivisé en sections (grappes), puis on prend tous les membres des grappes sélectionnées

1.2 Variable statistique

- Une ***variable statistique*** est une caractéristique susceptible de variations observables.
- Les ***modalités*** d'une variable statistique sont les valeurs possibles à priori de la variable.
- Les ***données*** sont les valeurs observées a posteriori de la variable. C'est les valeurs de la variable sur les individus de l'échantillon

$x_1, x_2, \ldots, x_n$.

1.3 Type de variables

- ***Variable quantitative*** : elle fait référence à un système de mesure numérique. On en distingue deux types :

- ***Variable quantitative discrète***: l'ensemble des modalités est dénombrable. Par exemple, la variable nombre d'enfants par ménage.

- ***Variable quantitative continue :*** l'ensemble des modalités est un intervalle. Par exemple, la durée de vie d'un instrument électronique.

- ***Variable qualitative :*** elle ne fait pas référence à un système de mesure numérique. On distingue deux cas :

  - *Ordinale* : l'ensemble des modalités correspond à une échelle avec un ordre. Par exemple: niveau de satisfaction, niveau de scolarité.

  - *Nominale* : l'ensemble des modalités ne correspond à aucune échelle. Par exemple: sexe, nationalité.

pour les biologistes

La terminologie est un peu différente chez les biologistes, où l'on définit alors 4 niveaux de mesure

- Le niveau nominal de mesure. (oui/non; couleurs…)
- Niveau ordinal de mesure

- Le niveau intervalle de mesure: qui est semblable au niveau ordinal avec la propriété supplémentaire que la différence entre deux valeurs a un sens, (exemple: la température, les années de naissance)
- Le niveau rapport de mesure: est semblable au niveau intervalle de mesure avec la propriété de mesure qu'il y a un zéro naturel pour lequel aucune quantité n'est présente.

  (exemple: poids)

2. Représentation des données

2.1 Variables quantitatives continues

- ***Données brutes :*** sont les données ‘’en vrac ‘’.

Exemple 1: les données ci-après représentent les durées de vie (10 3 en heures) d’un certain type d’appareils électroniques.

78.9	83.4	90.0	88.2	89.3	60.8
75.0	88.0	92.3	73.1	73.1	76.3
60.3	67.4	84.2	84.2	70.2	97.8
92.1	80.0	77.0	77.0	77.2	84.5
93.7	78.5	65.0	56.5	56.5	84.2

- ***Données groupées :*** sont des données obtenues en regroupant les données brutes ordonnées dans des classes sous la forme d'un *tableau de fréquences.*

- ***Tableau de fréquences:*** *les données brutes ci-dessus sont résumées dans le tableau de fréquence ci après :*

Exemple 2 :

Classe	Fréquence *n*	Valeur centrale *m*	Fréquence relative *f*
[55,65[	4	60	4/30
[65,75[	5	70	5/30
[75,85[	13	80	13/30
[85,95[	7	90	7/30
[95,105[	1	100	1/30
	n = 30		F = 1

Définitions

- L'amplitude d'une classe $[e_i, e_{i+1}[$ est $e_{i+1} - e_i$

- Son centre est $\dfrac{e_i + e_{i+1}}{2}$

Nombre de classes à retenir

Sturges suggère

$$1+\frac{10}{3}\log(n)$$

Lorsque la répartition des données semble symétrique , il convient de construire des classes de même amplitude. Si elle n'est pas symétrique ou si une classe a un effectif nul ou très faible, alors il convient de la regrouper avec la suivante (resp. précédente) si cette classe se trouve à droite (resp. gauche) de la classe la plus fréquente. Cette technique permet d'obtenir une répartition régulière.

2.2 Variables quantitatives discrètes

- ***Données brutes.***

Exemple 3: nombre de pièces défectueuses.

0	1	0	2	0
0	1	2	0	0
1	0	1	3	0
1	2	1	0	0

- ***Tableau de fréquences.***

On peut présenter les données brutes précédentes selon les modalités dans le tableau de fréquence suivant:

Modalité x	Fréquences *n*	Fréquences relatives *f*
0	10	10/20
1	6	6/20
2	3	3/20
3	1	1/20
	n = 20	F = 1

3. Caractères de position et de dispersion

3.1 Caractères de position: moyenne d'un échantillon

3.1.1 Moyenne arithmétique

- ***Cas de données brutes :***

  La moyenne d'un échantillon est donnée par la formule :

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

Exemple: Mesure du taux de plomb dans l'air

On liste ci-dessous des valeurs mesurées de plomb dans l'air (en microgrammes par mètre cube). L'Agence de protection de l'environnement a établi un standard de qualité de l'air pour le plomb: un maximum de 1.5. Les mesures indiquées ci-dessous

Ont été enregistrées dans l'immeuble numéro 5 du World Trade Center dans les jours qui ont suivi immédiatement sa destruction par l'attaque du 11 septembre 2001.

Trouver la moyenne pour cet échantillon de mesure de plomb dans l'air

5.4 1.10 0.42 0.73 0.48 1.10

$$\bar{x} = 1.538$$

- ***Cas de données groupées :***

  La moyenne des données groupées s'exprime par :

$$\bar{x} = \frac{n_1 m_1 + \cdots + n_p m_p}{n}$$

Propriétés de la moyenne arithmétique

Soit $x_1, x_2, \ldots, x_n$ un échantillon.

Soient $y_1, y_2, \ldots, y_n$ tels que: $y_i = ax_i + b$

où a et b sont deux nombres réels, alors

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

3.1. 2 Moyenne harmonique

La moyenne harmonique est donnée par

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

3.1.3 Moyenne géométrique

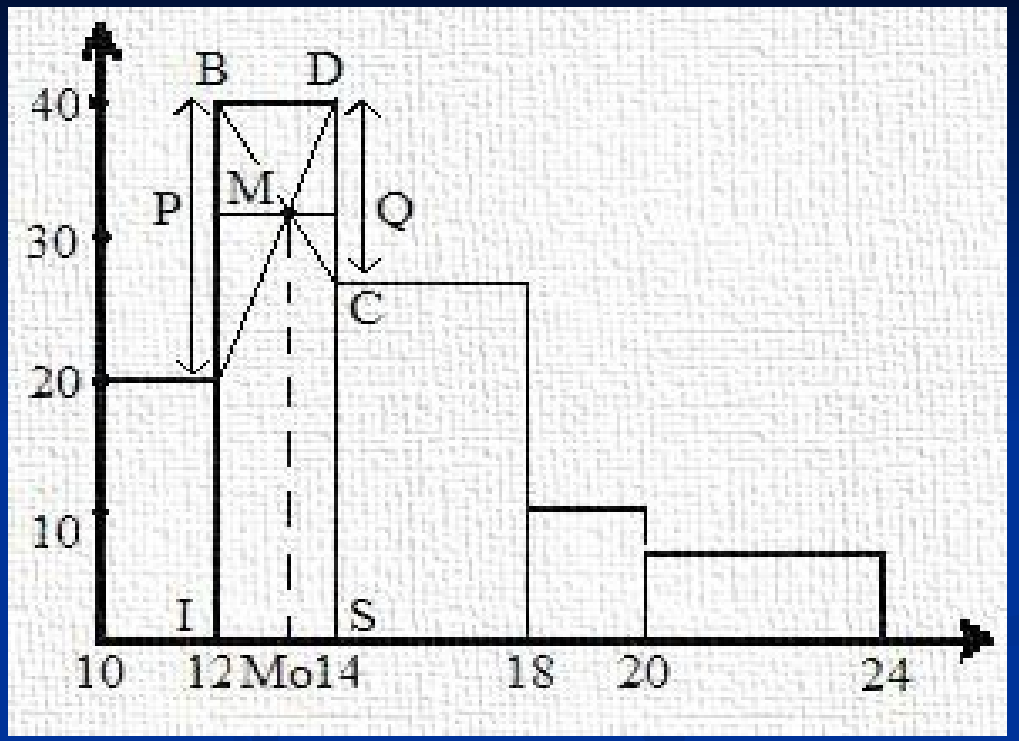
Elle est donnée par

$$G = \left(x_1 . x_2 \ldots . x_n\right)^{1/n}$$

3.1.4 Le mode

Le mode est la valeur de la variable étudiée correspondant à l'effectif le plus élevé.
Il renseigne sur la valeur la plus probable. On peut le déterminer par simple lecture du tableau des effectifs ou des fréquences pour les caractères quantitatifs discrets et qualitatifs.
Dans le cas d'un caractère continu, on détermine la classe modale, puis on cherche le mode à l'intérieur de la classe modale graphiquement.

Classe	n_i	a_i	$n_i' = \frac{\alpha\, n_i}{a_i}$
[10,12]	20	2	20
[12,14]	40	2	40
[14,18]	52	4	26
[18,20]	10	2	10
[20,24]	8	4	4



a_i classe i

α est la plus petite amplitude, ici c'est 2.

Les deux triangles (MAB) et (MCD) sont semblables, alors

$$\frac{AB}{MP}=\frac{CD}{MQ}$$ ce qui entraine que

$$\frac{\Delta_1}{M_0-I}=\frac{\Delta_2}{S-M_0}=\frac{\Delta_2}{I+a-M_0}$$

Et qui donne

$$M_0=I+\frac{a\Delta_1}{\Delta_1+\Delta_2}$$

Δ_1 et Δ_2 sont respt les excédents d'effectifs de la classe CM sur la classe qui la précède et la succède.

Application à l'exemple:

$$M_0 = 12 + \frac{2 \times 20}{20 + 14}$$

3.1.4 Le midrange

$$\text{Midrange} = \frac{\text{Minimum} + \text{Maximum}}{2}$$

Exemple: Mesure du taux de plomb dans l'air

$$\text{Midrange} = \frac{1.10 + 5.40}{2} = 3.25$$

3.2 Caractères de dispersion

3.2.1 Variance d'un échantillon

- ***Cas de données brutes :***

La variance d'un échantillon $x_1, \ldots, x_n$ *est* donnée par:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right]$$

S est appelé écart-type empirique

- ***Cas de données groupées :***

La variance des données s'exprime par :

$$S^2 = \frac{1}{n}\left[\left(\sum_{i=1}^{p} n_i m_i^2\right) - n\bar{x}^2\right]$$

Recette de l'étendu

Pour estimer une valeur de l'écart-type s: pour une approximation rapide de l'écart-type, utiliser:

$$s \approx \frac{\text{étendu}}{4}$$

Règle empirique (dite des 68-95-99.7) pour une distribution en « cloche ».

Une autre règle utile pour interpréter les valeurs d'un écart-type est la règle empirique. Cette règle dit que pour des données avec une distribution approximativement en cloche, les propriété suivantes s'appliquent:

- À peu près 68 % des valeurs sont situées à moins d'un écart-type la moyenne

- À peu près 95 % des valeurs sont situées à moins de deux écart-type la moyenne.

- À peu près 99.7 % des valeurs sont situées à moins de trois écart-type la moyenne

Théorème de Chebychev

La proportion d'un ensemble de données situés à moins de K écart-type de la moyenne est au moins de $1-\frac{1}{K^2}$ où K est une nombre positif

Plus grand que 1.

Pour k = 2 et k = 3 nous obtenons les proposition suivantes:

- Au moins 75 % de toutes les données sont à moins de 2 écarts-type de la moyenne.

- Au moins 89 % de toutes les données sont à moins de 3 écarts-type de la moyenne.

Propriétés de la variance empirique

Soit $x_1, x_2, \ldots, x_n$ un échantillon.

Soient $y_1, y_2, \ldots, y_n$ tels que: $y_i = ax_i + b$

où a et b sont deux nombres réels, alors

$$S_y^2 = aS_x^2 + b$$

Le coefficient de variation :

Le coefficient de variation est le rapport de l'écart-type par rapport à la moyenne.

$$CV = \frac{S}{\bar{x}}$$

En particulier, le coefficient de variation permet de comparer la variabilité relative de plusieurs distributions qui diffèrent fortement par leur ordre de grandeur et éventuellement même par leur unité de mesure.

3.2.2 Etendu

On appelle étendu d'un échantillon l'écart en valeur absolue entre la plus grande valeur et la plus petite valeur de l'échantillon.

c

$$x_{\max} - x_{\min}$$

3.2.3 Intervalle interquartile

- **Médiane d'un échantillon :** c'est la valeur qui partage l'échantillon en deux parties égales.

Calcul de la médiane :

Soit $x_1, \cdots, x_n$ un échantillon ordonné.

Pour calculer la médiane M, on distingue deux cas :

Cas 1 : n est impaire:

$$M = x_{\frac{n+1}{2}}$$

Cas 2 : n est paire:

$$M = \frac{x_{n/2} + x_{1+n/2}}{2}$$

Quartiles

Quartile d'ordre 1, noté Q_1

Q_1 est la médiane des observations qui sont strictement plus petites que la médiane.

Quartile d'ordre 3, noté Q_3

Q_3 est la médiane des observations qui sont strictement plus grandes que la médiane.

Cas d'approximation de calcul de quartiles, cas continu

Quand on ne dispose que de tableau des fréquences cumulées croissantes, on approche les quartiles graphiquement grâce au polygone des fréquences cumulées croissantes, et par interpolation linéaire grâce au tableau correspondant.

Exemple de calcul

Utilisation du tableau des fréquences cumulées croissantes

Le tableau des fréquences cumulées croissantes est :

mi	8	12	16	20	30	40	60
fréquences cumulées Croissantes	7	12.3	21.1	48.1	81.7	94.7	100

$$Q_1 = 16 + 4\frac{25 - 21.1}{48.1 - 21.1} = 16.57$$

$$Q_2 = 20 + 10\frac{50 - 48.1}{81.7 - 48.1} = 20.56$$

$$Q_3 = 20 + 10\frac{75 - 48.1}{81.7 - 48.1} = 28$$

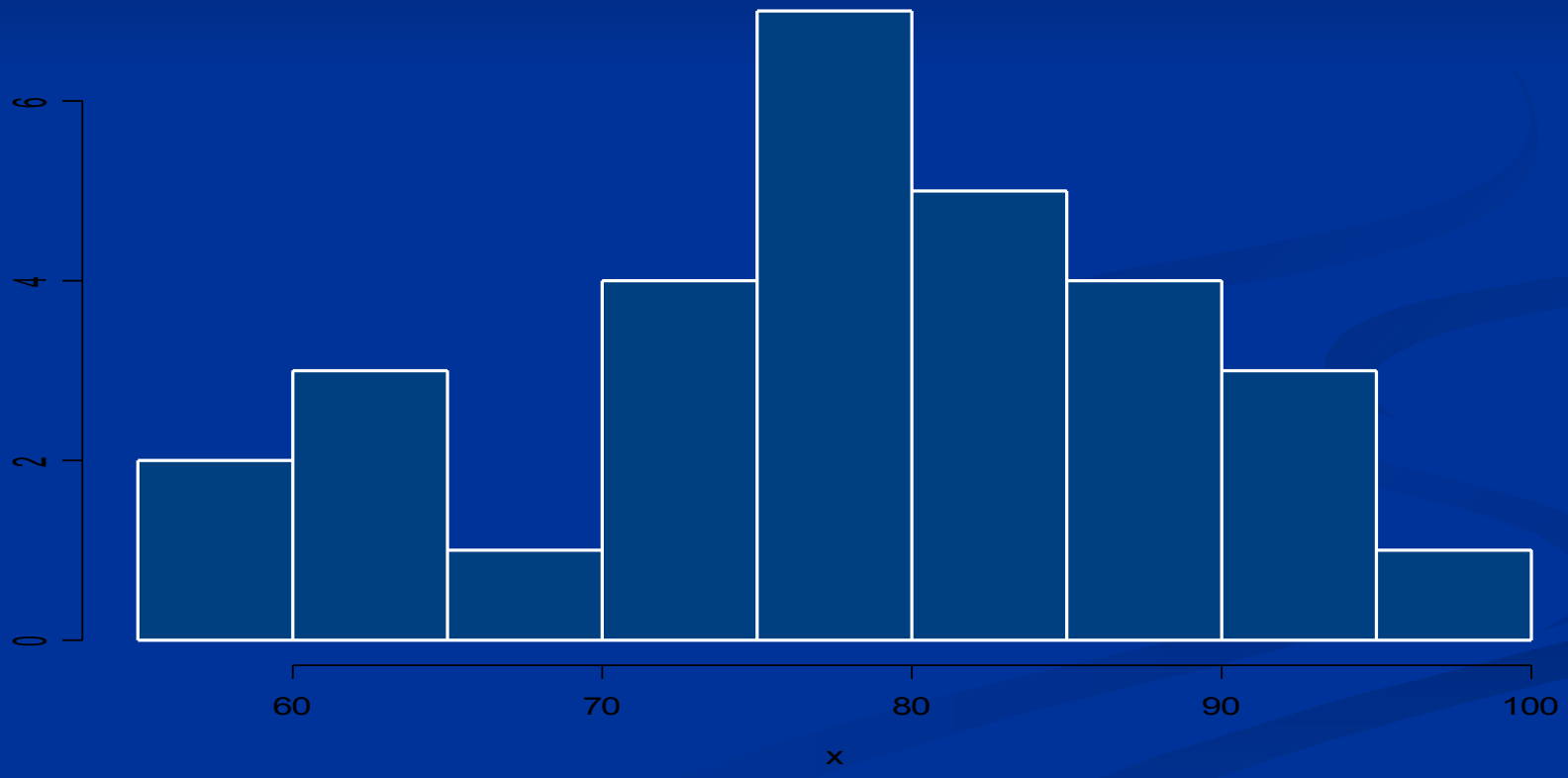
Intervalle interquartile

L'intervalle interquartile est

$$[Q_1, Q_3]$$

4. Représentations graphiques des données

4.1 Représentation graphique d'une variable continue: Histogramme



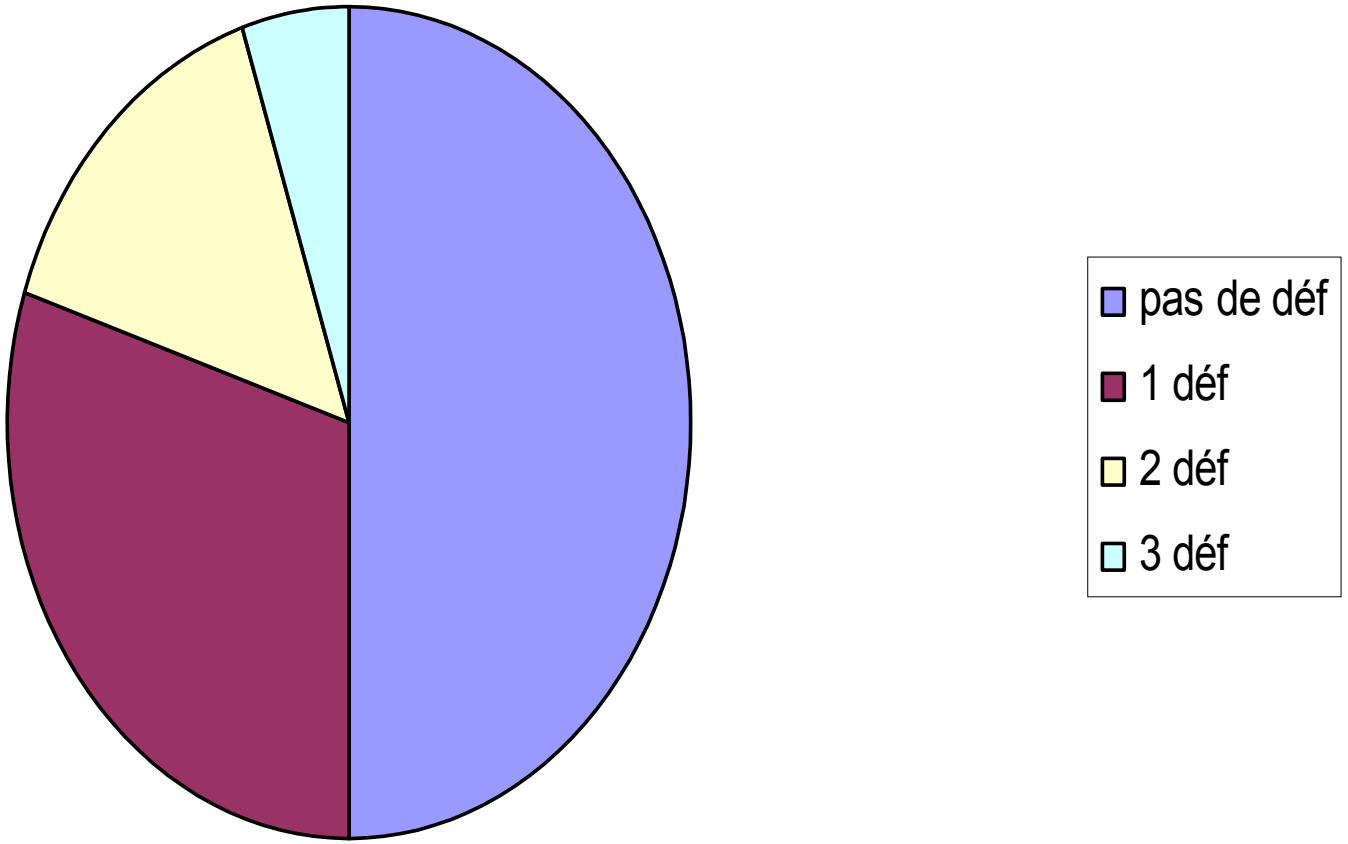
L’histogramme donne une idée sur le comportement de la distribution théorique des données.

Par exemple, cet histogramme laisse croire que la distribution théorique des observations suit une loi normale.

4.2 Représentation graphique d'une variable qualitative ou quantitative discrète:

Diagramme à secteurs circulaires.

Le diagramme à secteurs circulaires visualise les fréquences des modalités dans le cas d'une variable qualitative ou quantitative discrète. Les données discrètes de l'exemple 3 sont reprises dans le diagramme circulaire ci-après.


pas de déf
1 déf
2 déf
3 déf

5. Diagramme en boîte (Box PLot).

Le diagramme en boîte est une représentation graphique qui nous informe sur la symétrie et la dispersion des données. Il permet aussi d'identifier les valeurs aberrantes (les outliers) et de comparer plusieurs populations.

Afin de construire le diagramme en boîte, nous utiliserons la médiane et les quartiles d'un échantillon.

Jeu : détectez l'outlier dans la photo ci-dessous

Construction du diagramme en boîte

1) On calcule Q_1, M et Q_3.

2) On calcule l'écart interquartile $E = Q_3 - Q_1$

3) On calcule les bornes normales :

$$b_i = Q_1 - 1.5E$$

$$b_s = Q_3 + 1.5E$$

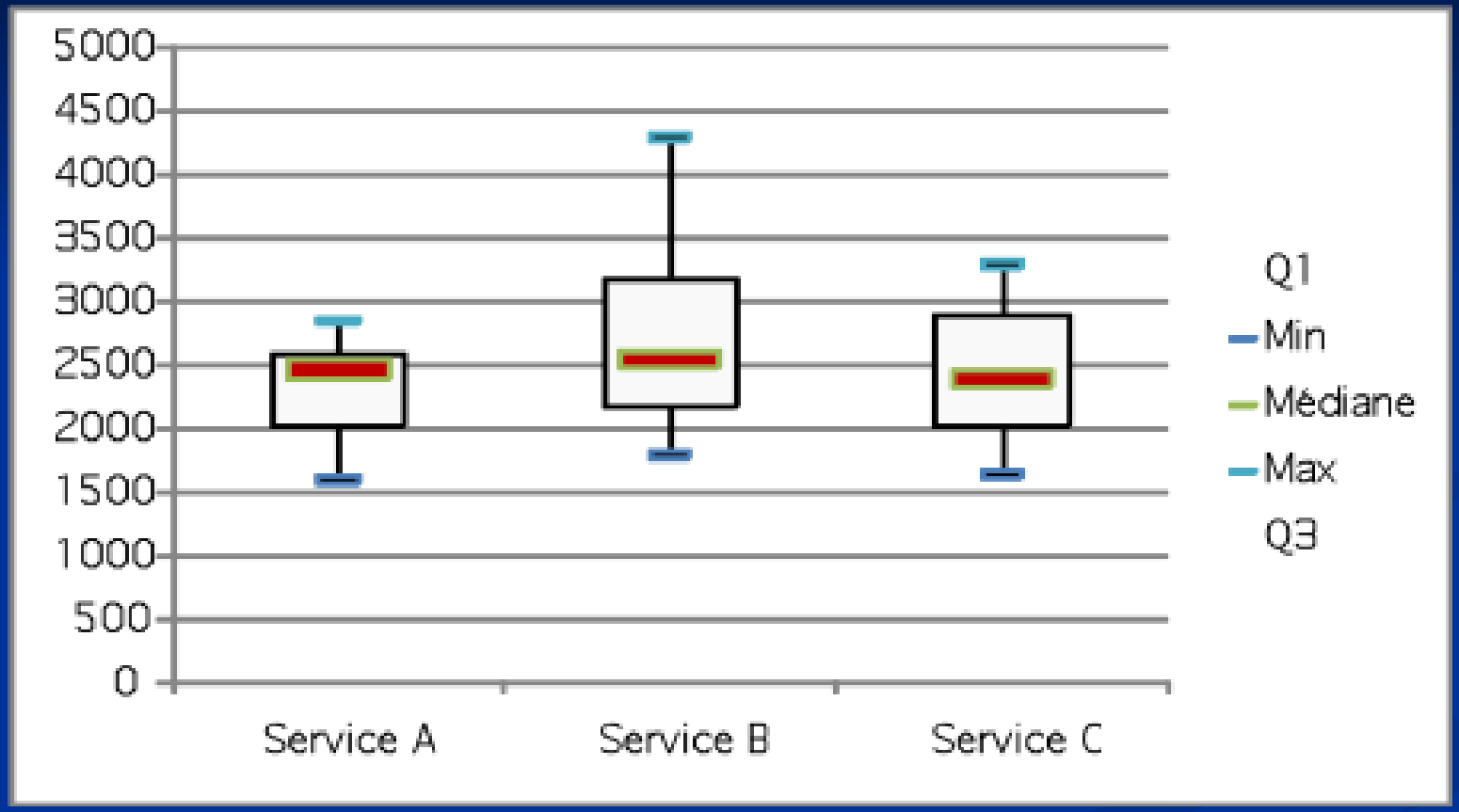
4) On identifie les données aberrantes qui sont les données qui sont en dehors de l'intervalle (b_i ,b_s).


X
X X
b_i
Q_1
M
Q_3
b_s

Remarque

Les boites à moustaches permettent d'aider à comparer les distributions dans différentes populations au vue d'une caractéristique donnée.

Le schéma ci-dessous le montre:


5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Service A
Service B
Service C
Q1
Min
Médiane
Max
Q3