Statistiques et Applications Informatiques

Statistiques Appliquées à la Gestion

Monday, February 18, 2019

Objectifs d'Apprentissage

- Comprendre la signification statistique des prélèvements, de l'échantillonnage et de la description statistique des données
- Apprendre a tirer profit des outils et rapports statistiques dans la prise de décisions
- Utiliser des calculs statistiques pour en tirer des conclusions sur les populations
- Habileté à analyser stratégiquement les situations et l'information obtenue, à faire la synthèse des données et à rédiger des rapports complets et concis

Définitions

- Ensemble de techniques permettant d'obtenir de l'information à partir d'observations nombreuses
- Les statistiques permettent de se renseigner sur des faits pour prendre les meilleurs décisions
- Le statisticien se penche sur les chiffres qui lui sont soumis pour obtenir des rapports numériques sensiblement indépendants du hasard et qui dénotent l'existence de causes régulières.
  - Ex : Statistiques de ventes – par représentant, statistiques géographiques, clients…

Domaine de la Statistique

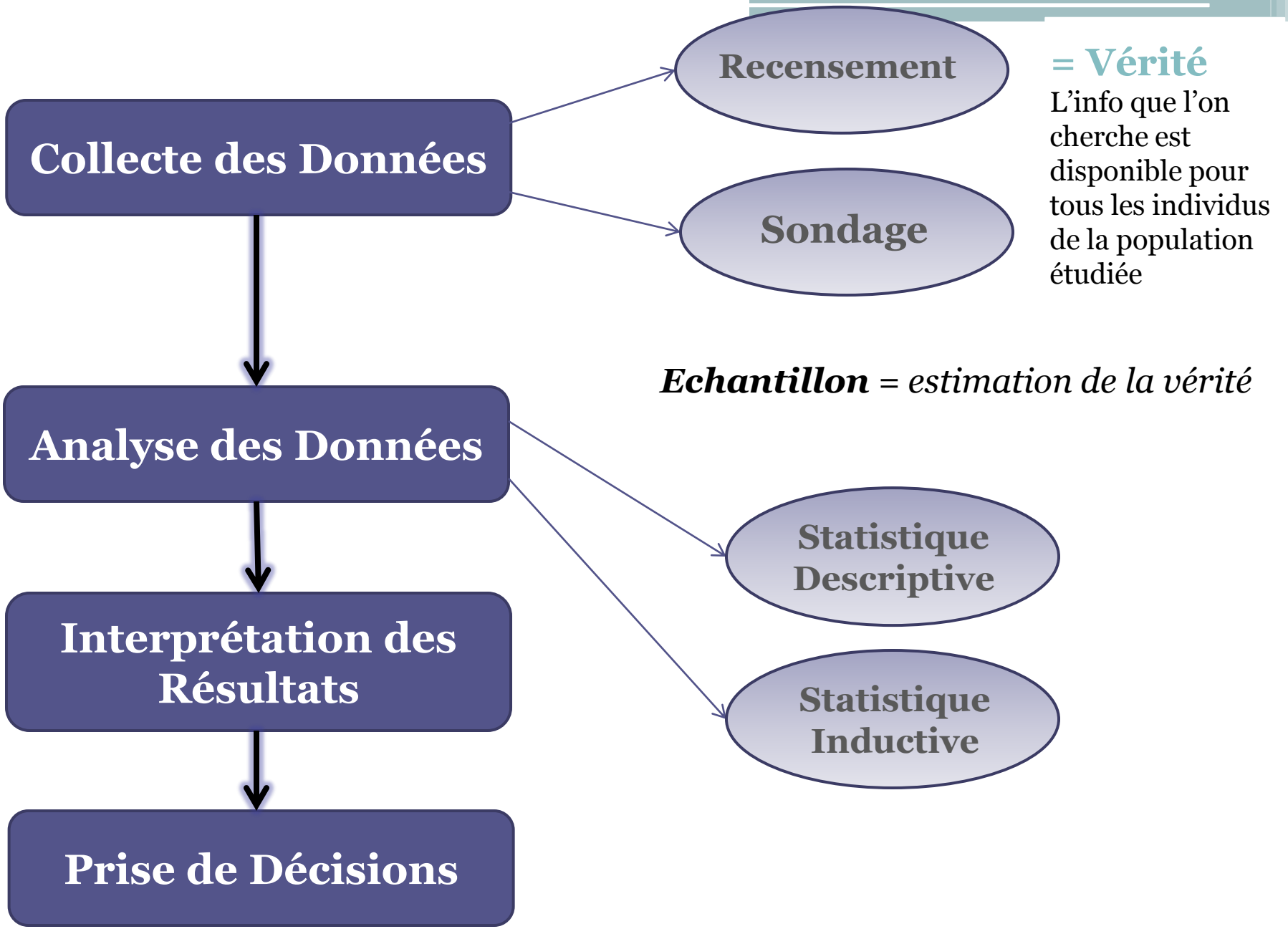
- On dénote donc 2 domaines de la statistique :
  - La **statistique descriptive** – où l'on décrit des ensembles de données complexes en opérant des réductions de ces données (*pour les rendre plus malléables/gérables/utiles*).
  - La **statistique analytique ou inductive** – où l'on débusque dans une variabilité constatée ; ce qui peut être expliqué par le hasard seulement où ce qui relève d'une autre explication.

Utilisation des Statistiques

- Comptabilité, finance
  - Séries chronologiques sur des bilans ou comptes de résultats,
  - gestion du capital,
  - trésorerie, opérations avec les banques
- Production
  - Gestion des stocks, du matériel,
  - contrôle de la qualité
- Marketing
  - Achats, ventes
  - Statistiques des ventes,
  - études de marché
- …

Les Etapes d'une Etude Statistique

- **Collecte des données** : Des observations sont effectuées au sein d'une population, relativement à un caractère ou une variable, *les résultats constituent une série statistique*.
- **Analyse des données** : Il s'agit de la détermination de paramètres statistiques qui permettent de caractériser la série statistique.
- **Interprétation des résultats** : A l'aide de propriétés mathématiques et en élaborant des tests pour une exploitation des résultats.



= Vérité

L'info que l'on cherche est disponible pour tous les individus de la population étudiée

Echantillon = *estimation de la vérité*

La statistique descriptive

- La statistique descriptive est un traitement de données qui offre des outils appropriés (Tableaux, graphiques et mesures numériques) permettant de dégager l'information essentielle qui se dissimule dans un grand nombre de données brutes.

- La Statistique Descriptive est l'ensemble des méthodes et techniques permettant de présenter, de décrire et de résumer des données numériques nombreuses et variées.

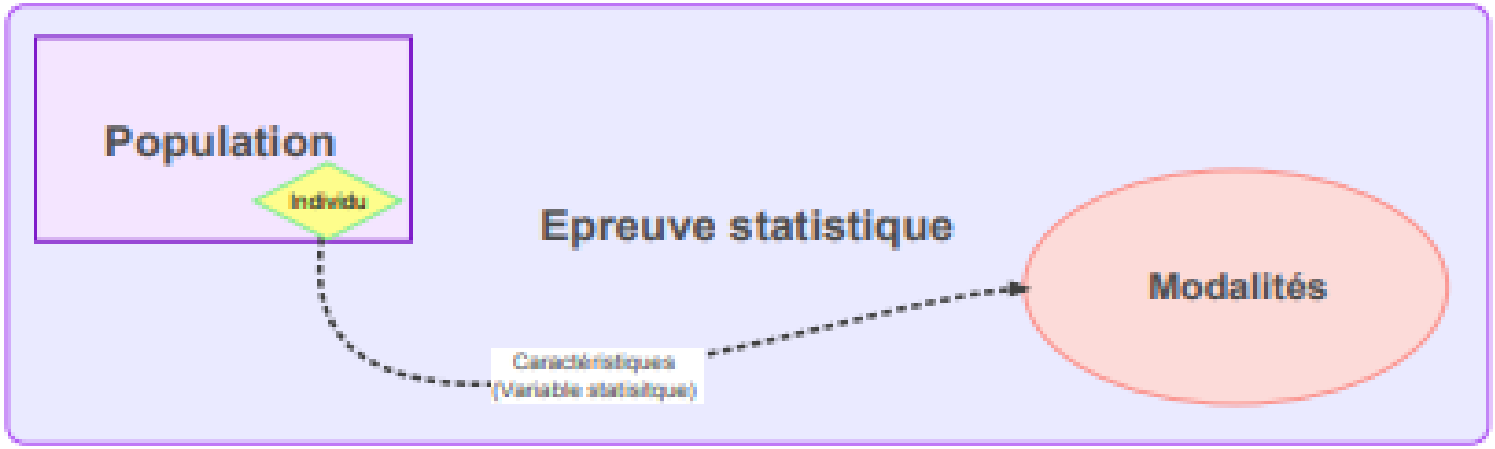
Applications

- Il existe trois types de traitements que nous pouvons effectuer sur une série statistique :
  - La synthèse de la série à l'aide d'un tableau statistique
  - Les représentations graphiques du caractère étudié
  - Le calcul des mesures caractéristiques.

Vocabulaire - Épreuve Statistique

- Épreuve statistique : une expérience que l'on provoque.
  - Exemple (La durée de vie des lampes)

    Imaginons le cas suivant : *un fabricant d'ampoules électriques ayant le choix entre 4 types de filaments se propose d'étudier l'influence de la nature du filament sur la durée de vie des ampoules fabriquées. Pour ce faire, il va faire fabriquer 4 échantillons d'ampoules identiques, sauf en ce qui concerne le filament, faire brûler les ampoules jusqu'à extinction, puis comparer les résultats obtenus.*



Vocabulaire – Population

- Population[1] : tout objet statistique étudié, qu'il s'agisse d'étudiants (d'une université ou d'un pays), de ménages ou de n'importe quel autre ensemble sur lequel on fait des observations statistiques. C'est donc l'ensemble sur lequel porte notre étude statistique, noté : Ω.
  - Exemple :
  - *On considère l'ensemble des étudiants de la section A. On s'intéresse aux nombres de frères et sœurs de chaque étudiant. Dans ce cas*

    Ω = ensemble des étudiants.
  - *Si l'on s'intéresse maintenant à la circulation automobile dans une ville, la population est alors constituée de l'ensemble des véhicules susceptibles de circuler dans cette ville à une date donnée. Dans ce cas*

    Ω = ensemble des véhicules.

Vocabulaire - Individu

- Individu[2] (unité statistique) : tout élément de la population Ω, il est noté ω (ω dans Ω). *Elément de base constitutif de la population à laquelle il appartient.*
  - Remarque : L'ensemble Ω peut être un ensemble de personnes, de choses ou d'animaux… L'unité statistique est un objet pour lequel nous sommes intéressés à recueillir de l'information.
  - Exemple :
    - *Dans l'exemple indiqué précédemment, un individu est tout étudiant de la section (A).*
    - *Si on étudie la production annuelle d'une usine de boîtes de boisson en métal (canettes). La population est l'ensemble des boîtes produites durant l'année et une boîte constitue un individu.*

Vocabulaire - Echantillon

- Échantillon : sous-ensemble construit et représentatif d'une population donnée.
- Avantages d'un échantillon :
  - Coût réduit
  - Rapidité accrue
  - Offre plus de possibilités
    - Dans certains cas il peut être impossible de faire un recensement (ex: contrôle de qualité)
  - Peut-être plus précis!
    - Cas où une main d'œuvre hautement qualifiée est requise pour la collecte des données

Illustration

Population

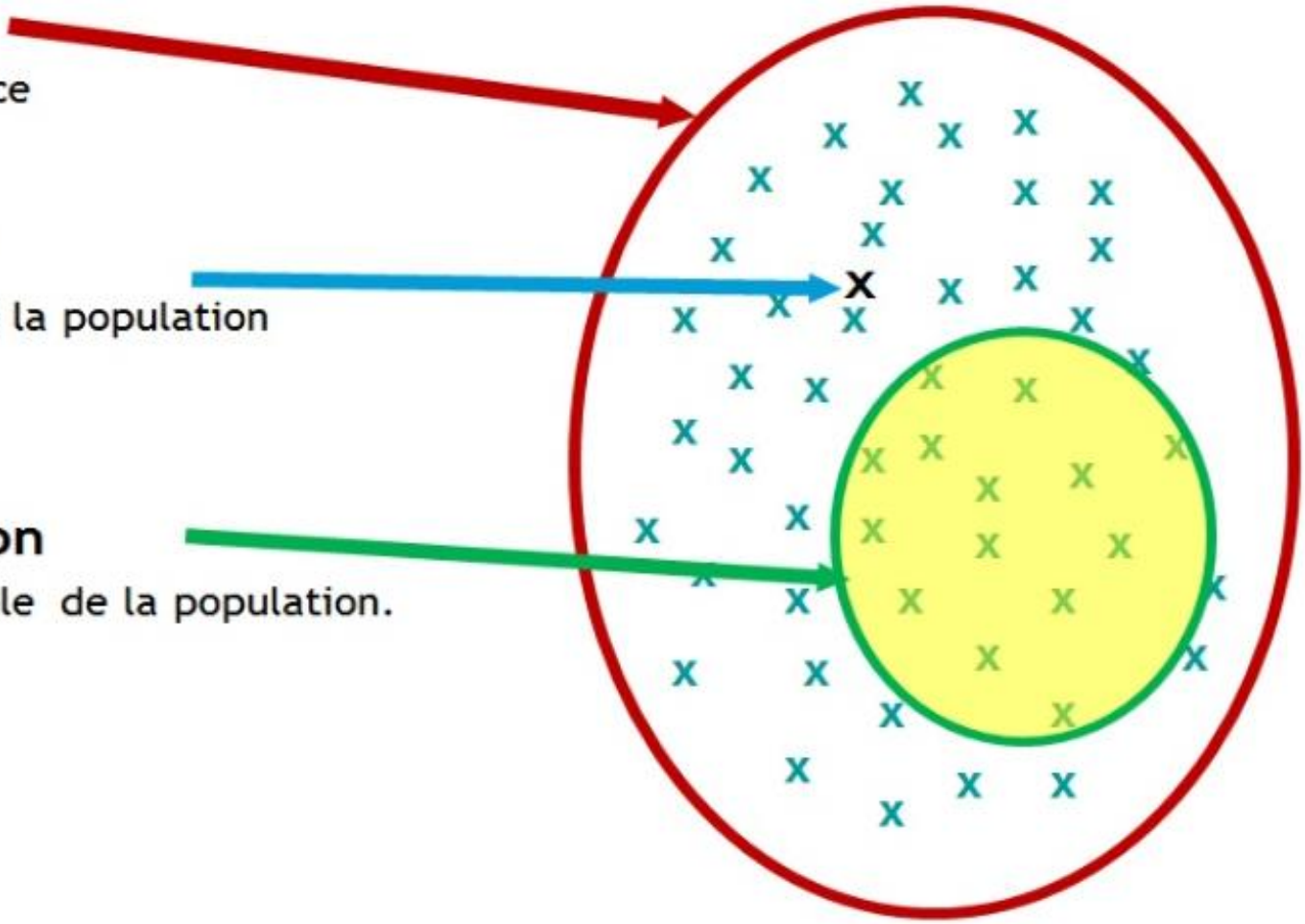
Ensemble de référence

Individu

Elément de la population

Echantillon

Sous-ensemble de la population.



Vocabulaire (suite)

- Les Variables – c'est la propriété ou l'aspect singulier que l'on se propose d'observer chez chaque individu de la population ou de l'échantillon.
  - Exemples :
    - La couleur
    - Le sexe
    - Le poids
    - La taille
    - La marque
    - Le modèle
    - L'espèce
    - Le prix
    - La surface, etc.

Vocabulaire - Caractère

- On appelle caractère (ou Variable Statistique, dénotée V.S) toute application

$$X : \Omega \rightarrow C.$$

- L'ensemble C est dit : ensemble des valeurs du caractère X (c'est ce qui est mesuré ou observé sur les individus)
  - Exemple : *Taille, température, nationalité, couleur des yeux, catégorie socioprofessionnelle …*

Remarque : Soit Ω un ensemble. On appelle et on note Card(Ω), le nombre d'éléments de Ω.

Card(Ω) := nombre d'éléments de Ω = N

Vocabulaire - Modalités

- Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci, ou :
  - Les différentes situations dans lesquelles les individus peuvent se trouver à l'égard du caractère considéré.

Exemple :

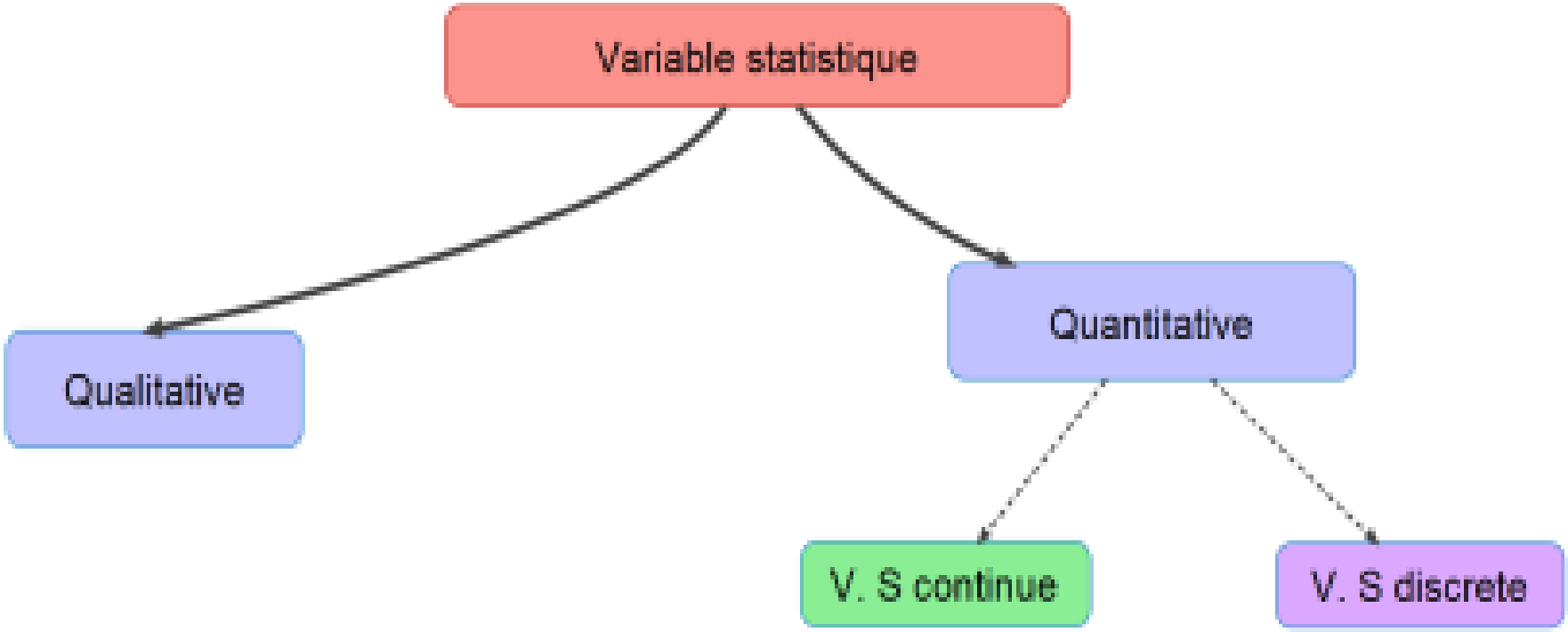
– Variable est " situation familiale "
Modalités sont " célibataire, marié, divorcé "

– Variable est" statut d'interrupteur "
Modalités sont " 0 et 1 ".

– Variable est " catégories socio-professionnelles "
Modalités sont " Employés, ouvriers, retraités, … "

Types des caractères

Nous distinguons deux catégories de caractères : les caractères qualitatifs et les caractères quantitatifs.



Caractère qualitatif

- Les caractères qualitatifs sont ceux dont les modalités ne peuvent pas être ordonnées, c-à-d que si l'on considère deux caractères pris au hasard, on ne peut pas dire que l'un des caractères est inférieur ou égal à l'autre.
- Définition : Les éléments de C sont représentés par autre chose que des chiffres.

  Exemple : L'état d'une maison - on peut considérer les modalités suivantes
  - Ancienne.
  - Dégradée.
  - Nouvelle.
  - Rénovée.

Caractères qualitatifs

- **Variable qualitative nominale** : Une variable statistique qualitative est dite définie sur une échelle nominale si ses modalités ne sont pas naturellement ordonnées.
  - Exemples : Situation d'activité, statut matrimonial.
- **Variable qualitative ordinale :** Une variable statistique qualitative est dite ordinale si l'ensemble de ses modalités peut être doté d'une relation d'ordre.
  - Exemple : Niveau d'instruction.

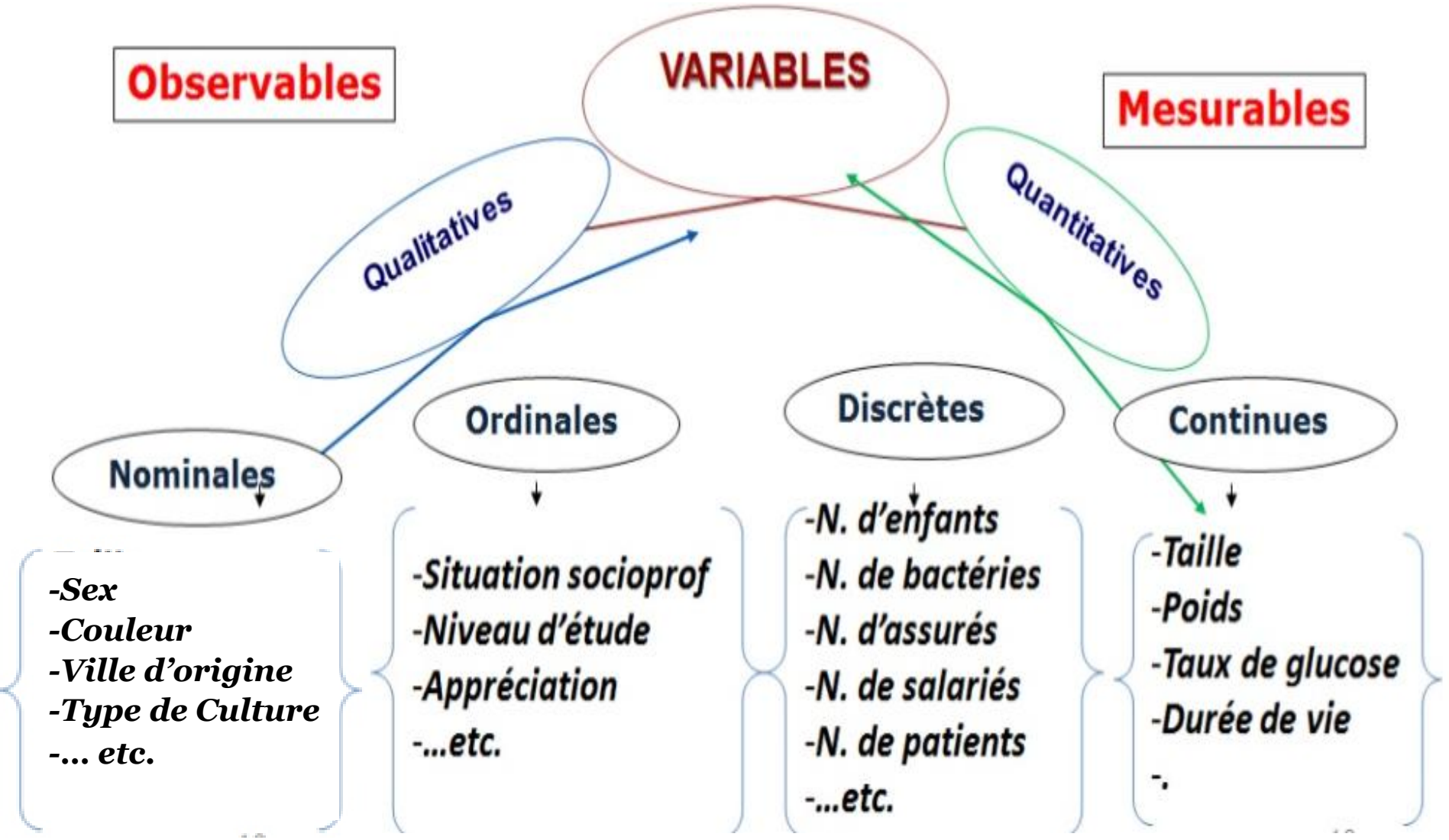
Caractère quantitatif

- Les caractères quantitatifs sont des caractères dont les modalités peuvent être ordonnées. Ainsi, l'âge, la taille de vie ou le salaire d'un individu sont des caractères quantitatifs.
- Définition : L'ensemble des valeurs est représenté par des chiffres, partagés en deux sortes de caractères, discret et continu

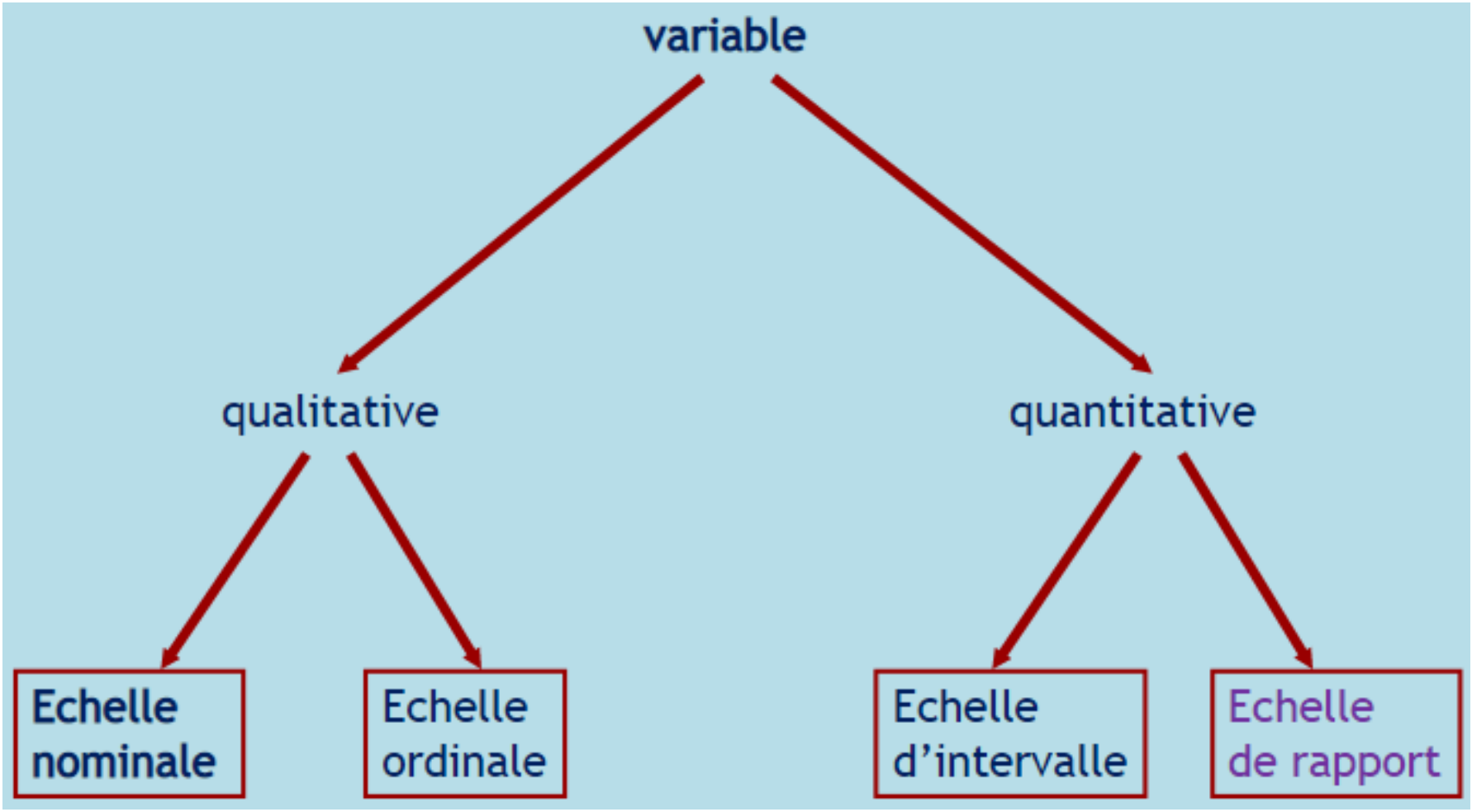
Exemples

– Le salaire d'employés d'une usine :
  Modalités : 10.000 Dhs , 20.000 Dhs… Type : Discret.
– La rigidité des ressorts. Modalités : [10, 20] N/m Type : Continu.

La statistique descriptive a pour objectif de synthétiser l'information contenue dans les jeux de données au moyen de tableaux, figures ou résumés numériques. Les variables statistiques sont analysées différemment selon leur nature (quantitative, qualitative).


VARIABLES
Observables
Mesurables
Qualitatives
Quantitatives
Nominales
Ordinales
Discrètes
Continues
-Sex
-Couleur
-Ville d'origine
-Type de Culture
-... etc.
-Situation socioprof
-Niveau d'étude
-Appréciation
-...etc.
-N. d'enfants
-N. de bactéries
-N. d'assurés
-N. de salariés
-N. de patients
-...etc.
-Taille
-Poids
-Taux de glucose
-Durée de vie
-.

Echelles de Mesure



Vocabulaire – Exemples

1. Population : l'ensemble des entreprises aux Maroc.

   Unité statistique : Maroc Telecom

   Echantillon : les entreprises recensées dans une base de données

2. Population : le parc automobile marocain: ensemble des automobiles immatriculées sur le territoire marocain.

   Unité statistique : une automobile (ex : Renault Mégane).

   Echantillon : les automobiles à 8 cf

3. Population : le parc de logements de Guelmim – ensemble des logements de la ville de Guelmim.

   Unité statistique : un logement particulier

   Échantillon : les appartements 2-pièces

Exemples d'Application

- **Exemple 1:** On a procédé au recensement des 50 salariés de la société STM en relevant les salaires horaires perçus:

✓ **Unité statistique:** un salarié de la société STM
✓ **Population:** l'ensemble des 50 salariés de la société STM
✓ **Caractère:** le salaire horaire
✓ **Type de caractère:** caractère quantitatif ou variable statistique

- **Exemple 2:** une enquête sur la nationalité des touristes visitant le Maroc a concerné un échantillon de 500 touristes.

✓ **Unité statistique:** un touriste
✓ **Population:** l'ensemble des touriste visitant le Maroc
✓ **Caractère:** nationalité
✓ **Type de caractère:** qualitatif

Excercice 1

Monday, February 18, 2019

- *La variable statistique "couleur de maisons d'un quartier" est-elle :*

☑ *qualitative*	☐ *quantitative*
☐ *discrète*	☐ *continue*

La variable statistique "revenu brut" est-elle :

☐ *qualitative*	☑ *quantitative*
☐ *discrète*	☑ *continue*

La variable statistique "nombre de maisons vendues par ville" est-elle :

☐ *qualitative*	☑ *quantitative*
☐ *discrète*	☑ *continue*

Excercice 2

- *Parmi ces assertions, préciser celles qui sont vraies, celles qui sont fausses.*

1. *On appelle variable, une caractéristique que l'on étudie.*
2. *La tâche de la statistique descriptive est de recueillir des données.*
3. *La tâche de la statistique descriptive est de présenter les données sous forme de tableaux, de graphiques et d'indicateurs statistiques.*
4. *En Statistique, on classe les variables selon différents types.*
5. *Les valeurs des variables sont aussi appelées modalités.*
6. *Pour une variable qualitative, chaque individu statistique ne peut avoir qu'une seule modalité.*
7. *Pour faire des traitements statistiques, il arrive qu'on transforme une variable quantitative en variable qualitative.*
8. *La variable quantitative poids d'automobile peut être reclassée en compacte, intermédiaire et grosse.*
9. *En pratique, lorsqu'une variable quantitative discrète prend un grand nombre de valeurs distinctes, on la traite comme continue.*

Excercice 3

- *Proposer des exemples de variable quantitative transformée en variable qualitative. Préciser les modalités de cette dernière.*

Solution : Les variables quantitatives dans le tableau ci-dessous peuvent être transformées en variables qualitatives. Les modalités de cette dernière sont précisées dans la seconde colonne.

Variable quantitative	Modalités envisageables
Hauteur	Petit, Moyen, Grand
Poids	Très léger, Léger, Moyen, Lourd, Très lourd
Rendement	Faible, Moyen, Elevé
Chiffre d'affaire	Modéré, Moyen, Important, Très important
Cylindrée	Petite, Moyenne, Grosse

Excercice 4

Monday, February 18, 2019

- Pour chacune des variables suivantes, préciser si elle est qualitative, quantitative discrète ou quantitative continue,

(a) Revenu annuel.
(b) Citoyenneté.
(c) Distance.
(d) Taille.
(e) Lieu de résidence.
(f) Âge.
(g) Couleur des yeux.
(h) Nombre de langues parlées.

Solution :

a. quantitative discrète
b. qualitative
c. quantitative continue
d. quantitative discrète
e. qualitative
f. quantitative discrète
g. qualitative
h. quantitative discrète

Excercice 5

- *Pour les sujets d'étude qui suivent, spécifier : l'unité statistique, la variable statistique et son type,*

1. *Étude du temps de validité des lampes électriques.*
2. *Étude de l'absentéisme des ouvriers, en jours, dans une usine.*
3. *Répartition des étudiants d'une promotion selon la mention obtenue sue le diplôme du Bac.*
4. *On cherche à modéliser*[1] *le nombre de collisions impliquant deux voitures sur un ensemble de* 100 *intersections routières choisies au hasard dans une ville. Les données sont collectées sur une période d'un an et le nombre d'accidents pour chaque intersection est ainsi mesuré.*

Indices de Sommation

- ***Soit***

1+2+3+4+...+n on peut écrire $\sum_{i=1}^{n} i = \sum_{1 \le i \le n} i$

$$\sum_{1 \le i \le n} a = \sum_{i=1}^{n} a = na \quad ; \quad \sum_{0 \le i \le n} a = \sum_{i=0}^{n} a = (n+1)a$$

- **Règles de calcul**

$$\sum_{i=1}^{n} (a.u_i) = a \sum_{i=1}^{n} u_i$$

$$\sum_{i=1}^{n} (u_i + v_i) = \sum_{i=1}^{n} u_i + \sum_{i=1}^{n} v_i$$

$$(\sum_{i=1}^{n} u_i).(\sum_{j=1}^{p} v_j) = \sum_{\substack{1 \le i \le n \\ 1 \le j \le p}} (u_i.v_j)$$

Indice du Produit

- ***Soit***

$\underbrace{a \times a \times \ldots \times a}_{n \text{ facteurs}}$, on peut écrire $\prod_{1 \le i \le n} a = \prod_{i=1}^{n} a = a^n$

$1 \times 2 \times 3 \times \ldots \times n$, on peut écrire $\prod_{i=1}^{n} i = \prod_{1 \le i \le n} i = n!$

- **Règles de calcul**

$$\prod_{i=1}^{n} a = a^n \; ; \; \prod_{i=0}^{n} a = a^{n+1}$$

$$\prod_{i=1}^{n} (u_i . v_i) = (\prod_{i=1}^{n} u_i).(\prod_{i=1}^{n} v_i)$$

Données Brutes

- On appelle données brutes des données qu'on rassemble sans se soucier de la notion d'ordre.
- **Exemple :** On a procédé au recensement des 50 salariés de la société STM (voir diapo 18) en relevant les salaires horaires perçus:

34	36	45	62	43	63	26	55	57	61
87	78	77	75	74	25	15	18	20	44
96	94	103	110	88	116	125	47	85	74
14	19	17	87	92	88	75	48	95	74
45	48	98	75	45	74	85	75	47	26

Etude d'une Variable Statistique Discrète

Variable à Caractère Discret

- On étudie ici un caractère statistique numérique représenté par une suite x_i décrivant la valeur du caractère avec i variant de 1 à k.
- Un caractère discret est un caractère statistique qui peut prendre un nombre fini raisonnable de valeurs (note, nombre d'enfants, nombre de pièces, …) :

$$X : \Omega \rightarrow \{x1, x2, \ldots, xn\}$$

- avec Card(Ω) := N est le nombre d'individus dans Ω

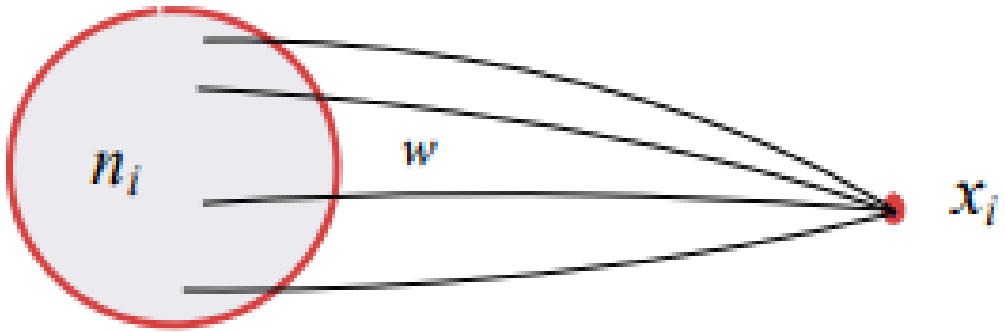
Effectif partiel

Un effectif partiel (fréquence absolue) se définit comme suit :

Pour chaque valeur x_i, on pose par définition

$$n_i = \text{Card}\{\omega \in \Omega : X(\omega) = x_i\}$$

n_i : le nombre d'individus qui ont le même x_i, ça s'appelle effectif partiel de x_i.



Le nombre d'individus qui prennent la valeur x_i

Effectif Cumulé

Pour chaque valeur xi, on pose par définition

$$N_i = n_1 + n_2 + \ldots + n_i \, .$$

L'effectif cumulé N_i d'une valeur est la somme de l'effectif de cette valeur et de tous les effectifs des valeurs qui précèdent.

N_i est le nombre d'individus dont la valeur du caractère est inférieur ou égale à x_i .

L'effectif total est donné par $N = \text{card}\{\Omega) = \sum_{i=1}^{n} n_i$

Exemple – Effectif partiel Vs. Effectif Cumulé

Prenons l'exemple, décrit dans le tableau ci-dessous, des nombres d'écrans par ménage à Casablanca (sur un échantillon de 200 familles)

x_i	0	1	2	3	4	5	6
n_i (Effectif)	18	32	66	41	32	9	2

L'effectif partiel du nombre de familles qui ont 2 écrans est 66

50 est le nombre de familles qui ont un nombre d'enfant inférieur ou égal à 1 (effectif cumulé). Nous le regardons dans le tableau :

x_i	0	1	2	3	4	5	6
n_i (Effectif)	18	50	116	157	189	198	200

Fréquence Partielle

- f_i s'appelle la fréquence partielle de x_i quand pour chaque valeur x_i, on a

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

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

- On peut remplacer f_i par $f_i \times 100$ qui représente alors un pourcentage.

- fi = est le pourcentage des ω tel que $X(\omega) = x_i$.

$$\Rightarrow \quad \sum_{i=1}^{n} f_i = 1$$

Fréquence Cumulée

- F_i s'appelle la fréquence cumulée de x_i quand pour chaque valeur x_i, on a

$$F_i = f_1 + f_2 + \ldots + f_i$$

- Fi = est le pourcentage des ω tel que la valeur X(ω) est inférieure ou égale à x_i .
- Dans l'exemple précédent, 0.785 représente 78,5% de familles dont le nombre d'écrans est inférieur ou égale à 3.

x_i	0	1	2	3	…
n_i (Effectif)	18	50	116	157	…
Ni (Effectif)	0.09	0.25	0.58	0.785	…

Exemples – Fréquences Partielle vs Cumulée

Dans l'exemple précédent, 0, 33 : il y a 33% de familles avec un nombre d'écrans égal à 2. Ce pourcentage est calculé de la façon suivante (N = 200) :

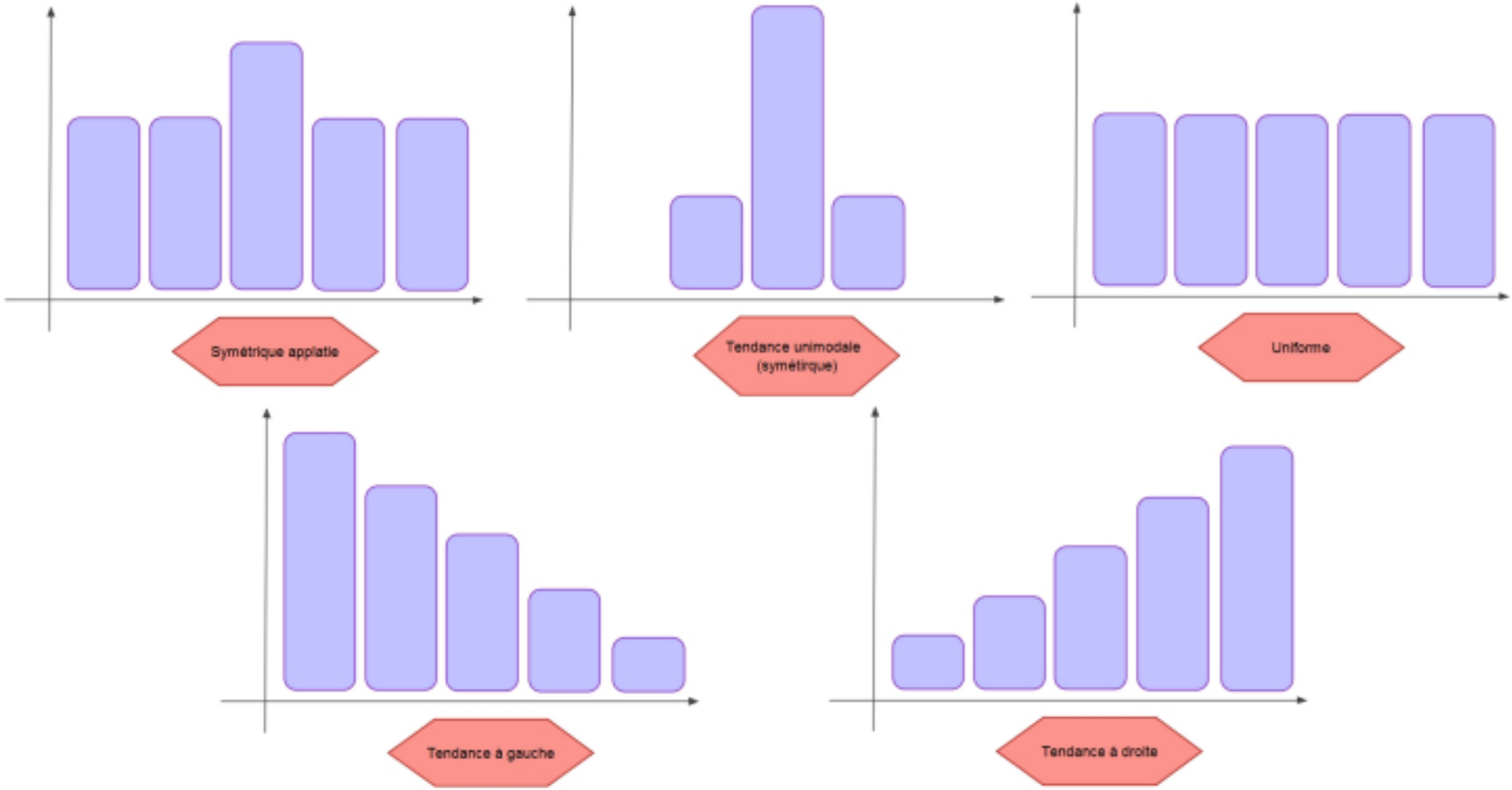
x_i	$\cdots$	2	$\cdots$
n_i (Effectif)	$\cdots$	66	$\cdots$
N_i (Effectif)	$\cdots$	$\frac{66}{200} = 0.33$	$\cdots$

Le tableau à droite représente un exemple de fréquences cumulées d'erreurs d'assemblage sur un ensemble d'appareils : où 94% des appareils ont une nombre d'erreurs inférieur ou égal à 3

Nombre d'erreurs	Nombre d'appareils	Fréquences cumulées
0	101	0.26
1	140	0.61
2	92	0.84
3	42	0.94
4	18	0.99
5	3	1

Monday, February 18, 2019

Représentations Graphiques



Quelques Caractéristiques du Graphique

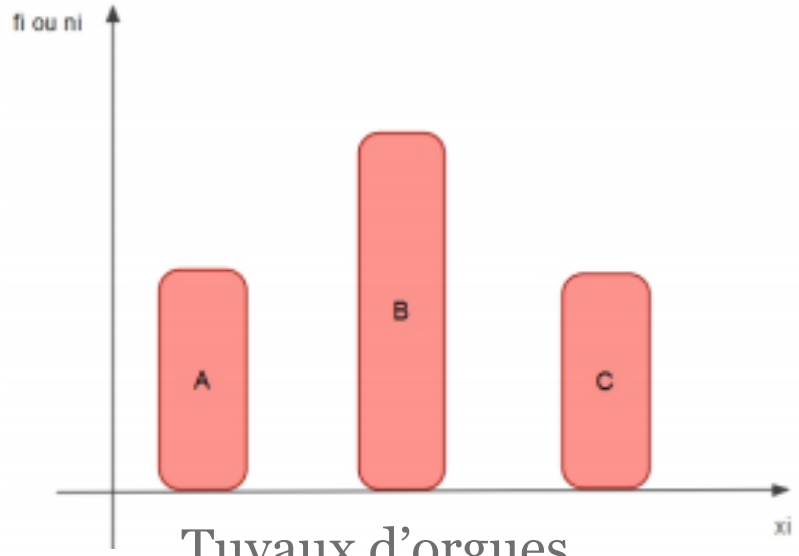
Représentation graphique des séries statistiques

- On distingue les méthodes de représentation d'une variable statistique en fonction de la nature de cette variable (qualitative ou quantitative).
- Le graphique est un support visuel qui permet :
  - **La synthèse :** visualiser d'un seul coup d'œil les principales caractéristiques (mais on perd une quantité d'informations),
  - **La découverte :** met en évidence les tendances.
  - **Le contrôle :** on aperçoit mieux les anomalies sur un graphique que dans un tableau.
  - **La recherche des régularités :** régularité dans le mouvement, répétition du phénomène.

Distribution à Caractère Qualitatif

Deux diagrammes permettent de représenter une variable qualitative :

- le diagramme en bandes (dit tuyaux d'orgue) et
- le diagramme à secteurs angulaires (dit camembert ou diagramme circulaire).



Tuyaux d'orgues

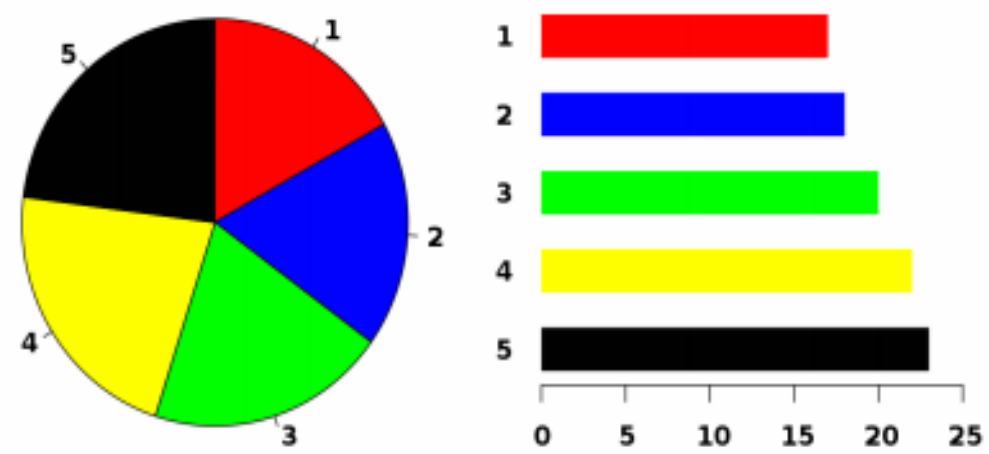
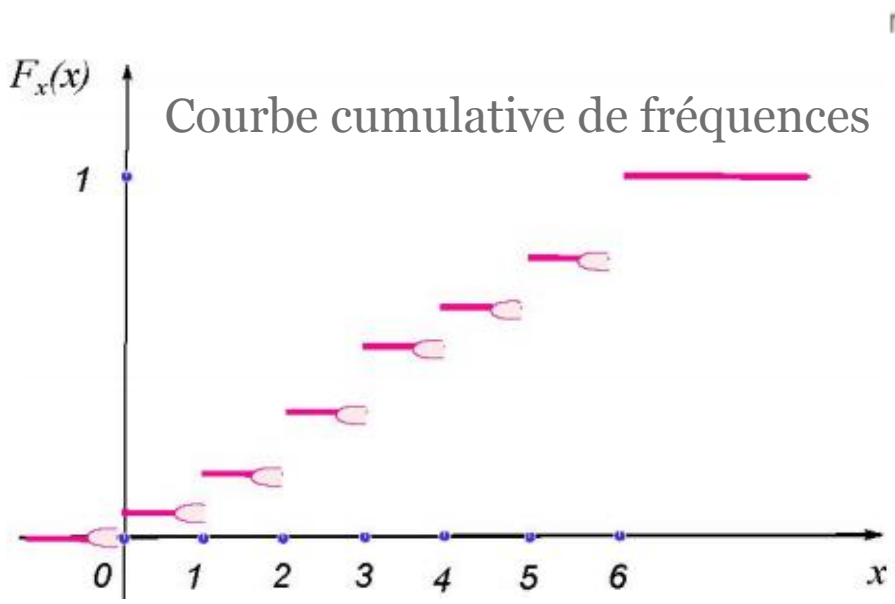


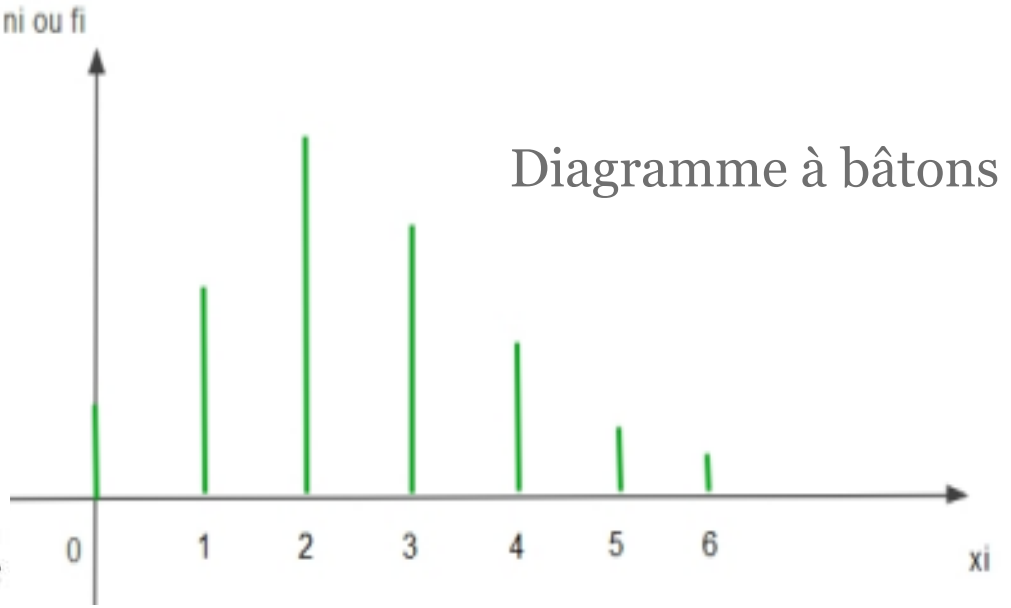
Diagramme circulaire

Distribution à Caractère Quantitatif Discret

Deux diagrammes permettent de représenter une variable quantitative discrète :

- le diagramme en bâtons et
- le diagramme cumulatif





Série Statistique

- Une série statistique est une simple énumération des observations

$$x_1, x_2, x_3, \ldots\ldots, x_i, \ldots\ldots, x_n$$

- Ces observations étant rangées par ordre croissant:

$$x_1 \leq x_2 \leq x_3 \leq \ldots\ldots \leq x_i \leq \ldots\ldots \leq x_n$$

- **n** est le nombre total des observations appelé aussi effectif. Une même observation peut se répéter plusieurs fois. La différence entre la valeur la plus grande et la valeur la plus petite est appelée étendue.

$$Etendue = X\max - X\min$$

Caractéristiques d'1 Série Statistique

- On caractérise souvent une série statistique par 2 types de paramètres :
  - Les paramètres dits de position ou d'ordre 1 :
    - Moyenne
    - Mode
    - Médiane
    - Quartiles…
  - Les paramètres de dispersion ou d'ordre 2 :
    - Variance
    - Ecart type
    - Coefficient de variation…

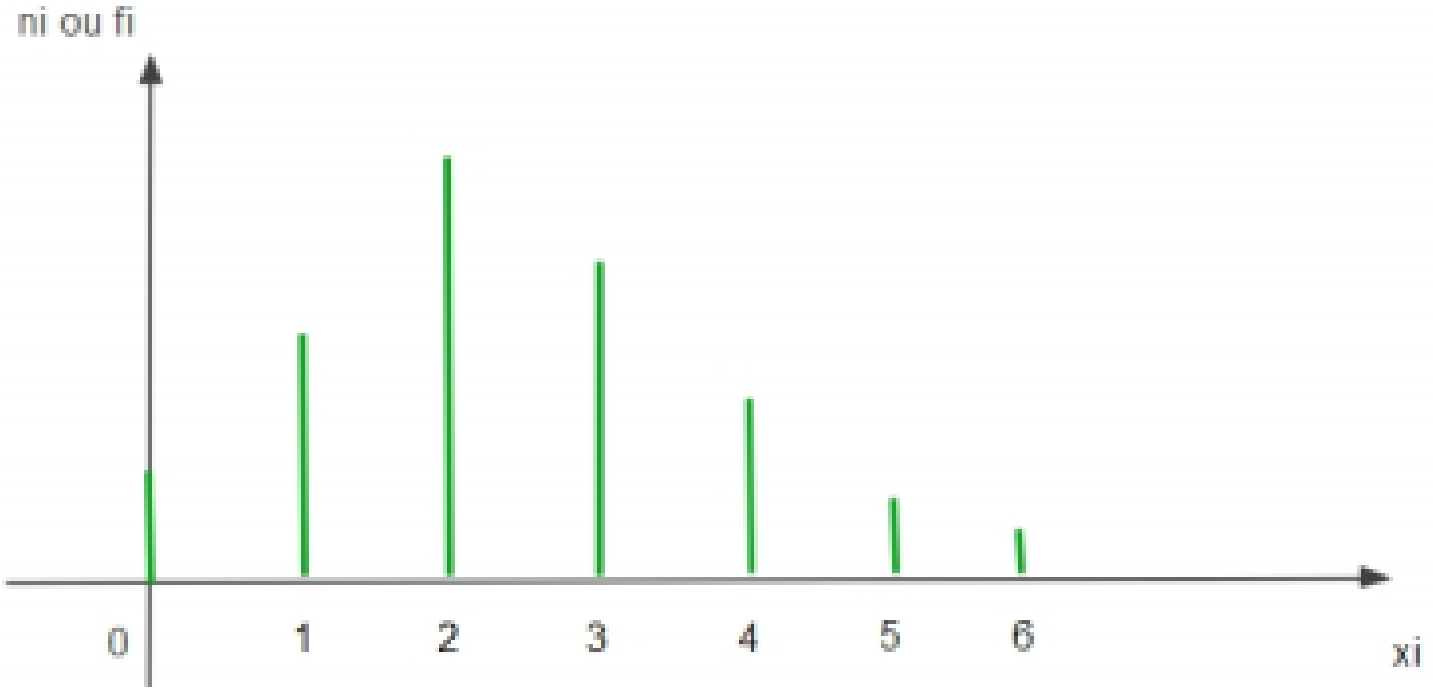
Paramètres de Tendance Centrale

Les Paramètres de Position

- Les paramètres de position sont : le mode, la moyenne et la médiane.
- Mode : modalité d'effectif maximal, le mode d'une V.S est la valeur qui a le plus grand effectif partiel (ou la plus grande fréquence partielle) et il est dénoté par M_o.
- Classe modale : est une classe de densité maximale – une classe pour laquelle le quotient effectif/ amplitude (qui s'appelle la densité d'effectif) est maximal.
  - Pour des classes d'amplitudes égales ou pour les variables discrètes, les classes modales ou les modes correspondent aux effectifs maxima.
  - ***Remarque :*** *Il peut exister plusieurs modes ou plusieurs classes modales*.

Exemple de Mode

- *Dans l'exemple ci-dessous, le mode est égal à 2 correspondant ainsi au plus grand effectif.*



Remarque : on peut avoir plus d'un mode

La Médiane

On appelle médiane la valeur Me de la V.S X qui vérifie la relation suivante :

$$F_x(Me^-) < 0.5 \leq F_x(Me^+) = F_x(Me)$$

La médiane partage la série statistique en deux groupes de même effectif.

Dans l'exemple ci-après

x_i	0	1	2	3	4	5	6
n_i	18	32	66	41	32	9	2

0 ne satisfait pas la formule : $F_x(0^-) = 0 < 0.5 \leq F_x(0^+) = 0.09$

Par contre 2 si : $Fx(2^-) = 0.25 < 0.5 \leq Fx(2^+) = F(2) = 0.58$

La Moyenne

On appelle moyenne de X la quantité :

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

avec N = Card(Ω). On peut donc exprimer et calculer la moyenne dite "arithmétique" avec des effectifs ou avec des fréquences.

il existe d'autres moyennes que la moyenne arithmétique

Exemple :

x_i	0	1	2	3	4	5	6
n_i	18	32	66	41	32	9	2

Si $\overline{x}$ = 2.46, alors nous avons en moyenne une famille de quartier a 2.46 écrans dans leur maison.

Paramètres de Dispersion

- Les indicateurs statistiques de dispersion usuels sont l'étendue, la variance et l'écart type.
  - **L'étendue** (de la V.S X) est la différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité :

$$e = x_{max} - x_{min}$$

  - *L'étendue donne une première idée de la dispersion des observations. C'est un indicateur très rudimentaire et il existe des indicateurs de dispersion plus élaborés*

Paramètres de Dispersion

La variance d'une série statistique X est le nombre :

$$Var(X) = \sum_{i=1}^{n} f_i(\overline{x} - x_i)^2$$

On dit que la variance est la moyenne des carrés des écarts à la moyenne x.

Les « écarts à la moyenne » sont les (x–xi), les « carrés des écarts à la moyenne » sont donc les $(x–xi)^2$.

En faisant la moyenne de ces écarts, on trouve la variance.

Une meilleure représentation de la variance est la suivante :

Soit (xi , ni) une série statistique de moyenne x et de variance Var(X). Alors,

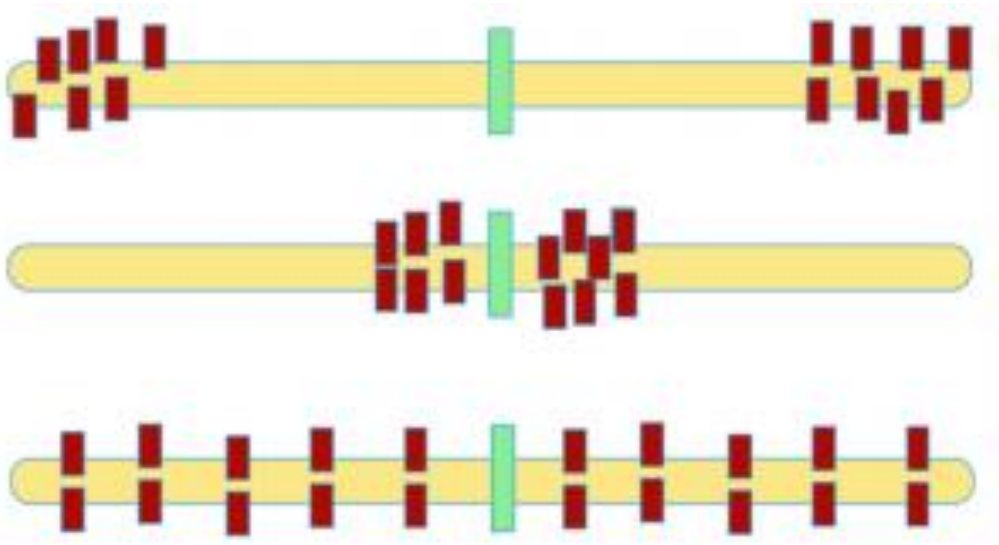
$$V\ ar(X) = \sum_{i=1}^{n} f_i x_i^2 - \overline{x}^2$$

L'Ecart-Type

La quantité

$$\sigma_X = \sqrt{Var(x)}$$

s'appelle l'écart type de la variable statistique X.



La dispersion d'une série statistique autour de sa moyenne

Le paramètre σ_x mesure la distance moyenne entre x et les valeurs de X (voir Figure ci-dessus). Il sert à mesurer la dispersion d'une série statistique autour de sa moyenne.

- Plus il est petit, plus les caractères sont concentrés autour de la moyenne (on dit que la série est homogène).
- Plus il est grand, plus les caractères sont dispersés autour de la moyenne (on dit que la série est hétérogène).

Exercice 6

- Le tableau suivant donne la répartition selon le groupe sanguin de 40 individus pris au hasard dans une population,

Groupes sanguins	*A*	*B*	*AB*	*O*
L'effectif	*20*	*10*	n_3	*5*

1. *Déterminer la variable statistique et son type.*
2. *Déterminer l'effectif des personnes ayant un groupe sanguin AB.*
3. *Donner toutes les représentations graphiques possibles de cette distribution.*

Solution

1 - La population dans cette étude est les 40 personnes. Donc N = 40. La variable statistique est le groupe sanguin des individus et elle est qualitative.

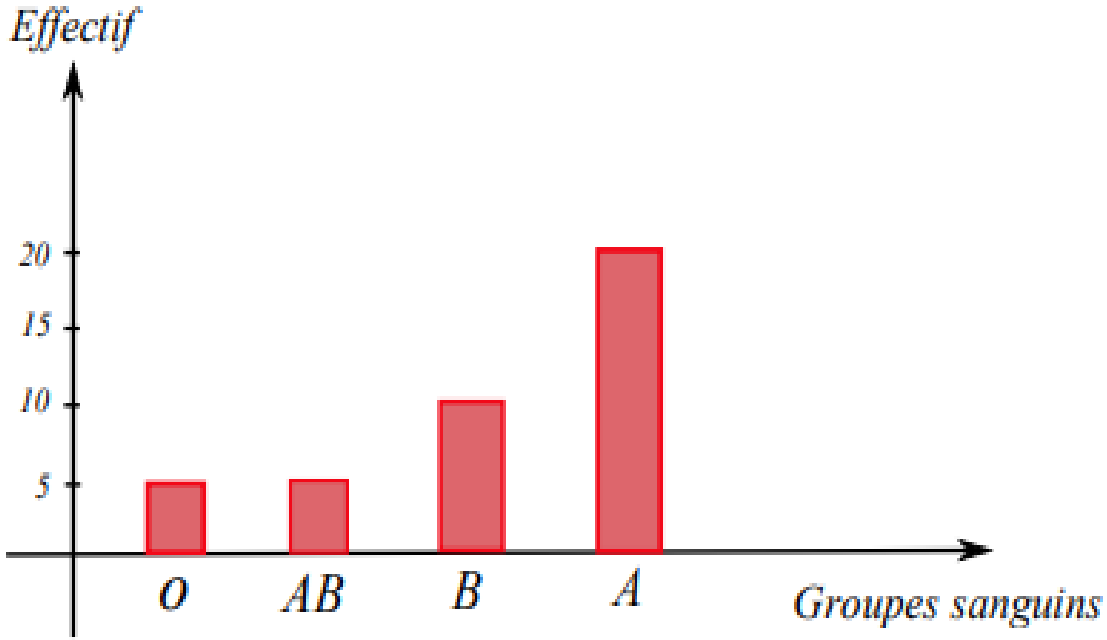
2 - L'effectif total est égal à 40. Par conséquent, $N = 40 = \sum_{i=1}^{4} n_i$

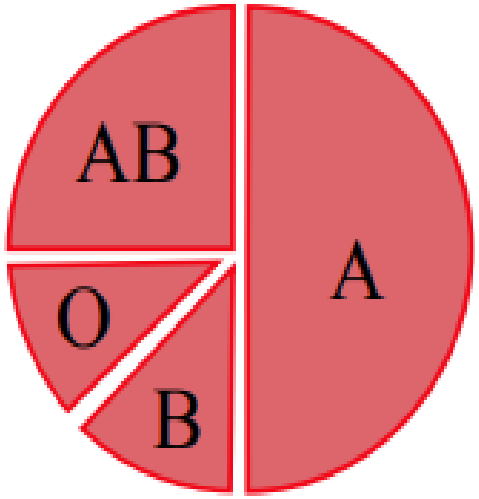
Alors, $20 + 10 + n_3 + 5 = 40$.

Ce qui implique que $n_3 = 5$.

Exercice 6 - Solution

- Nous avons deux représentations possibles "Tyaux d'orgue" et "Diagramme en secteur".





A gauche "Tyaux d'orgue" *et à droite "Diagramme en secteur"*

Exercice 7

- *On considère deux groupes d'étudiants. Nous relevons leurs notes d'examens dans les deux tableaux suivants :*

Note (groupe A)	8	9	10	11
Effectif	2	2	1	1

Note (groupe B)	6	8	9	13	14
Effectif	2	2	2	1	1

Calculer la moyenne et l'écart type de chaque groupe. Comparer les deux groupes.

Exercice 7 - Solution

Solution Dans un premier temps, nous remarquons que l'effectif total du groupe A est égal à 6 et celui du groupe B est égal à 8.

En utilisant la formule de la moyenne, nous obtenons

$$\overline{x}_A = 9.2 \quad \text{et} \quad \overline{x}_B = 9.1.$$

On remarque que les moyennes sont très proches. Peut-on pour autant conclure que ces deux groupes ont des niveaux identiques ?

Nous répondons à cette question après le calcule des écarts type. Ils sont donnés par

$$\sigma_X^A = 1.11 \quad \text{et} \quad \sigma_X^B = 2.8.$$

Nous remarquons que même si les deux groupes ont des moyennes quasiment identiques, le groupe B est beaucoup plus dispersé que le groupe A car $\sigma_X^B > \sigma_X^A$. Les étudiants de ce groupe ont des notes plus irréguliers. On peut dire donc que le groupe B est moins homogènes que le groupe A. En observant les valeurs du tableau, on voit que c'est cohérent.

Exercice 8

Monday, February 18, 2019

- Le gérant d'un magasin vendant des articles de consommation courante a relevé pour un article particulier qui semble connaître une très forte popularité, le nombre d'articles vendus par jour. Son relevé a porté sur les ventes des mois de Mars et Avril, ce qui correspond à 52 jours de vente. Le relevé des observations se présente comme suit :

7 13 8 10 9 12 10 8 9 10 6 14 7 15 9 11 12 11 12 5 14 11 8 10 14 12 8

5 7 13 12 16 11 9 11 11 12 12 15 14 5 14 9 9 14 13 11 10 11 12 9 15.

1. *Quel type est la variable statistique étudiée.*
2. *Déterminer le tableau statistique en fonction des effectifs, des fréquences, des effectifs cumulés et des fréquences cumulés.*
3. *Tracer le diagramme des bâtonnés associé à la variable X.*
4. *Soit F_x la fonction de répartition. Déterminer F_x.*
5. *Calculer le mode Mo et la moyenne arithmétique $\overline{x}$.*
6. *Déterminer à partir du tableau puis à partir du graphe, la valeur de la médiane Me.*
7. *Calculer la variance et l'écart-type.*

Exercice 8 - Solution

Monday, February 18, 2019

1 - La population est les 52 jours et la variable stat. étudiée est le nombre d'articles vendus par jour
Son type est : quantitatif discret (nombre).

2 - Le tableau statistique est donné par :

x_i	5	6	7	8	9	10	11	12	13
n_i	3	1	3	4	7	5	8	8	3
f_i	3/52	1/52	3/52	4/52	7/52	5 /52	8 /52	8/52	3 /52
N_i	3	4	7	11	18	23	31	39	42
F_i	3/52	4/52	7/52	11/52	18/52	23/52	31/52	39/52	42/52

14	15	16	$\sum$
6	3	1	$N = 52$
6 /52	3/52	1/52	1
48	51	52	$\varnothing$
48/52	51/52	1	$\varnothing$

Exercice 8 - Solution

3 - L'élaboration du diagramme des bâtonnets de X,

4 - La fonction de répartition est donnée par

$$F_x(x) = \begin{cases} 0, & si \quad x < 5, \\ 3/52, & si \quad 5 \leq x < 6, \\ 4/52, & si \quad 6 \leq x < 7, \\ 7/52, & si \quad 7 \leq x < 8, \\ \ddots & \ddots \\ 1, & si \quad x \geq 16. \end{cases}$$

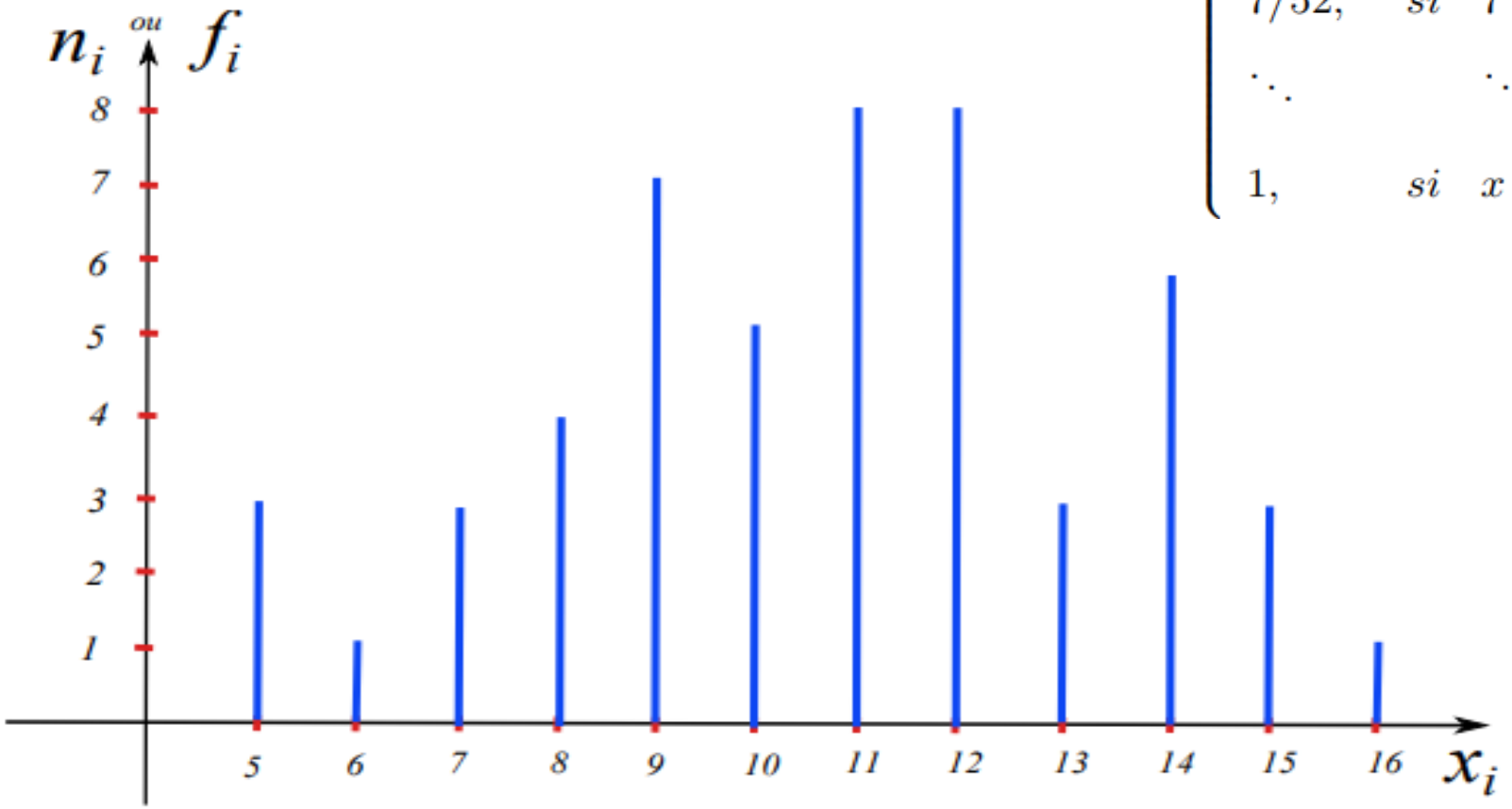


Diagramme à bâtons

Exercice 8 - Solution

5 - Le mode est la valeur de la variable qui a le plus grand effectif, c'est à dire, $n_i = 8$.
Donc, $M_o = 11$ et Mo = 12.

La moyenne arithmétique est donnée par : $$\overline{x} = \frac{1}{N}\sum_{i=1}^{12} n_i x_i = \sum_{i=1}^{12} f_i x_i$$

Par conséquent : $$\overline{x} = \frac{1}{52}(3 \times 5 + 1 \times 6 + 5 \times 7 + \ldots + 1 \times 16) = \frac{555}{52} = 10.67$$

6 - La médiane est la valeur de la variable qui divise la population de la série statistique en deux parties égales :

$$F_x(11^-) = \frac{23}{52} < 0.5 \leq F_x(11^+) = F(Me) = \frac{31}{52}$$

Donc, Me = 11.

7 - Nous commençons par la variance, $$Var(X) = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \overline{x}^2$$

Après calcul, on trouve $$Var(X) = 7.64$$

Par conséquent, l'écart type est calculé comme : $$\sigma_X = \sqrt{Var(x)} = 2.76$$

Exercice 9 (with Solution)

Un quartier résidentiel comprend 99 unités d'habitation ayant une valeur locative moyenne de 10000 Da. Deux nouvelles unités d'habitation sont construites dans le quartier: l'une a une valeur locative de 7000 Da et l'autre, une villa luxueuse, a une valeur locative de 114000 Dhs.

– Quelle est la nouvelle moyenne de valeur locative pour le quartier ?

– Pouvait-on s'attendre à de tel résultat ?

Solution - *Le nouveau total des mesures de valeur locative est*

$$(99 \times 10000) + 7000 + 114000 = 1111000$$

Le nouveau total d'individus statistiques est $99 + 2 = 101$

La nouvelle moyenne est donc $\dfrac{1111000}{101} = 110000$

- On pouvait s'attendre à une augmentation de la moyenne car l'une des deux nouvelles valeurs est très nettement au dessus de la moyenne initiale.

Exercice 10 (Supplémentaire)

- Pour déterminer le type de logement $(F2, F3, ...)$ *à construire, on étudie* 20 *familles selon leur nombre d'enfants. Durant l'expérience, on note les résultats suivants :*

$$1, 3, 5, 5, 3, 2, 4, 4, 7, 0, 2, 4, 3, 7, 0, 5, 4, 2, 3, 2$$

- *Déterminer, la population, l'unité (individu), la variable statistique et les modalités.*
- *Déterminer le tableau statistique avec* x_i, n_i, f_i *et* F_i.

Modalité	0	1	2	3	4	5	6	7	$\sum$
n_i									
N_i									
f_i									
F_i									

Exercice 11 (Supplémentaire)

- *Au poste de péage, on compte le nombre de voitures se présentant sur une période de* 5 *min. Sur* 100 *observations de* 5 *min, on obtient les résultats suivants :*

Nombre de voitures	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Nombre d'observations	*2*	*8*	*14*	*20*	*19*	*15*	*9*	*6*	*2*	*3*	*1*	*1*

1. *Construire la table des fréquences et le diagramme en bâtons en fréquences de la série du nombre de voitures.*
2. *Calculer la moyenne et l'écart-type de cette série.*
3. *Déterminer la médiane.*

Etude d'une Variable Statistique Continue

Variable à Caractère Continu

- On appelle V.S. continue (ou à caractère continu) toute application de Ω et à valeurs réelles et qui prend un nombre "important" de valeurs (*Les caractères continus sont ceux qui ont une infinité de modalités*).

  Exemple : Soit Ω l'ensemble des nouveaux nés à la clinique Ghandi pendant les 3 premiers mois de 2018. Nous désignons par X le poids (en grammes) des nouveaux nés.

  On suppose que $x_{min} = 2.701$ et $x_{max} = 5.001$

- L'étude de ce caractère se fait en partageant les valeurs prises par X en classes de valeurs.

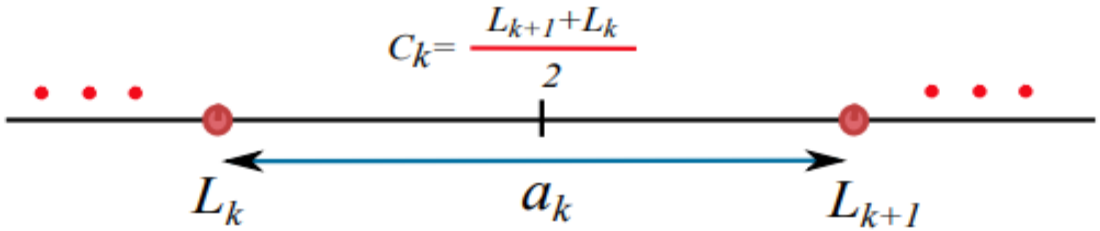
Classe de Valeurs

On appelle classe de valeurs de X un intervalle de type [a, b[tel que $X \in [a, b[$ si et seulement si $a \leq X(w) < b$

(*c'est à dire, que les valeurs du caractère sont dans la classe [a, b[*)

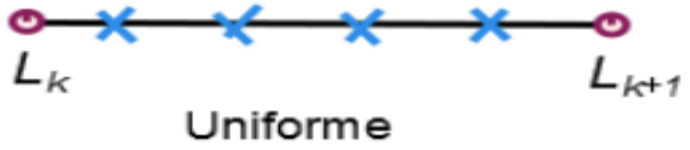
Dès qu'un caractère est identifié en tant que continu, ces modalités ${}^1C_k = [L_k, L_{k+1}[$ sont des intervalles avec

- L_k : borne inférieure.
- L_{k+1} : borne supérieure.
- $a_k = L_{k+1} - L_k$: son amplitude, son pas ou sa longueur.
- $C_k = x_k = (L_{k+1} + L_k)/2$: son centre.



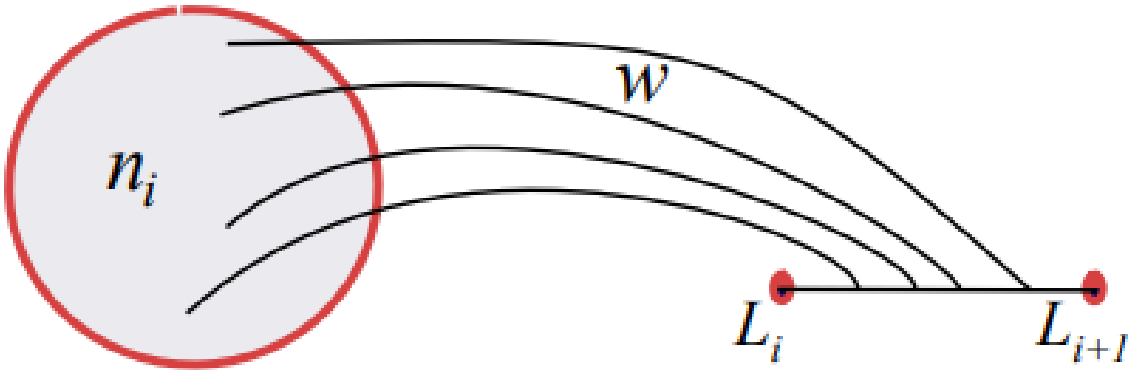
Nombre de Classes

- En combien de classes partageons-nous les valeurs ? la réponse n'est pas unique. Dans la section « Discrétisation » on verra une multitude de méthodes de classification



Effectif et Fréquence d'une Classe

- La quantité $n_i = \text{Card}\{w \in \Omega : X(w) \in C_i\}$ s'appelle effectif partiel de C_i.



Le nombre d'individus qui prennent des valeurs xi dans Ci .

- Le nombre $f_i := \frac{n_i}{N}$

est appelé la fréquence partielle de C_i.

L'Effet Cumulé d'une Classe

On appelle l'effectif cumulé de C_i la quantité

$$N_i := \sum_{j=1}^{i} n_j$$

On appelle la fréquence cumulée de C_i la quantité

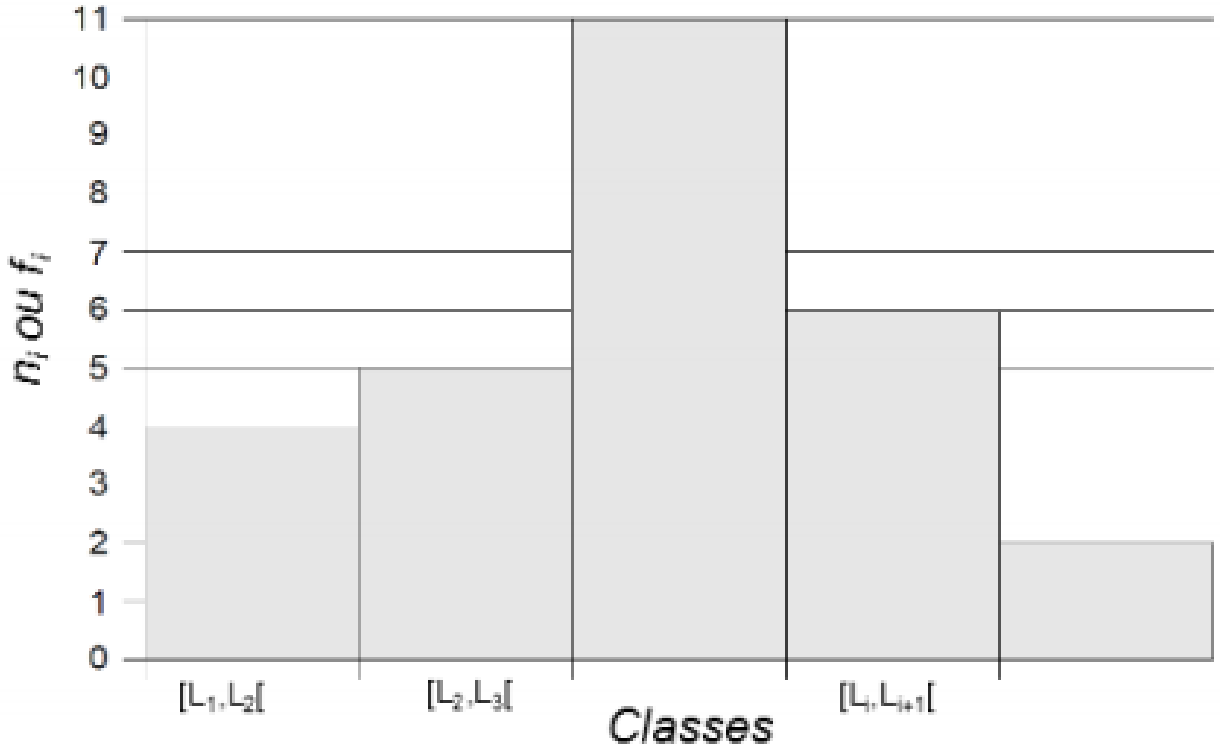
$$F_i := \sum_{j=1}^{i} f_j$$

Nous avons les interprétations suivantes :

- n_i : est le nombre d'individus dont les valeurs des caractères sont dans la classe C_i,
- f_i : est le pourcentage des w tel que $X(w) \in C_i$,
- N_i : est égale au $Card\{w : X(w) \in C_1 \cup C_2 \cup \ldots \cup C_i\}$,
- F_i : est le pourcentage des w tel que $X(w) \in C_1 \cup \ldots \cup C_i$.

Représentation Graphique

- Histogramme des fréquences (ou des effectifs)



Fonction de Répartition

La fonction $F_x : R \rightarrow [0, 1]$ définie par $F_x(x)$ représente le pourcentage des individus tel que la valeur de leur caractère est inférieure ou égale à x.

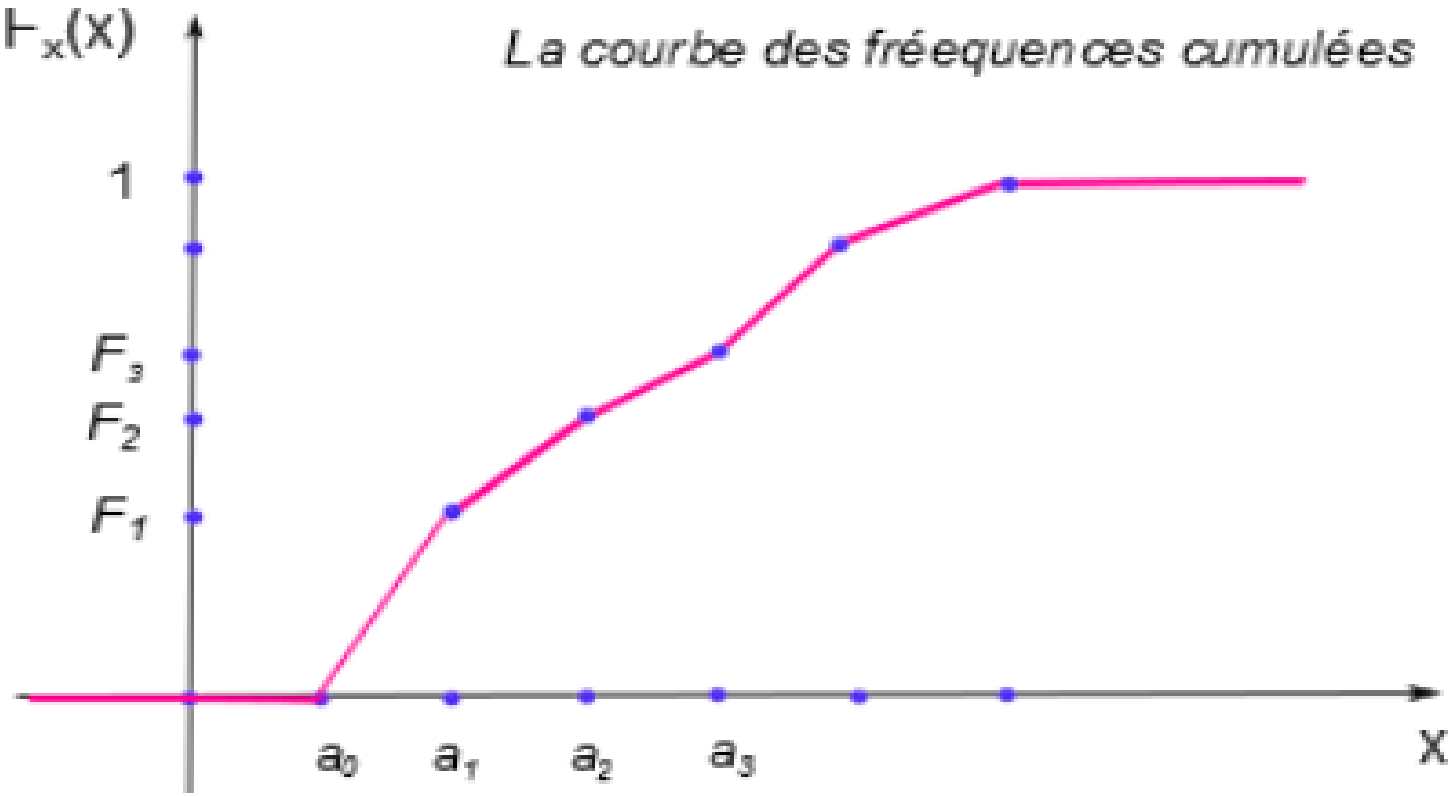
Elle est donnée par :

Et elle s'appelle la fonction de Répartition de X.

$$F_x(x) = \begin{cases} 0, & \text{si } x < a_0 \\ \frac{f_i}{h}(x - a_0), & \text{si } a_0 \leq x < a_1, \\ f_i + \frac{f_{i+1}}{h}(x - a_i), & \text{si } a_i \leq x < a_{i+1} \\ 1, & \text{si } x \geq a_n \end{cases}$$

Courbe des Fréquences Cumulées

La courbe de F_x est nulle avant a_0, constante égale à 1 après a_n et joint les points $(a_0, 0)$, (a_1, F_1),...,$(a_n, 1)$ par des segments de droites



La Moyenne (et le Centre de la Classe)

On note par c_i le centre de la classe C_i et nous considérons f_i la fréquence partielle de C_i .

La quantité $\overline{x} = \sum_{i=1}^{n} f_i C_i$ s'appelle la moyenne de X.

Le Mode et la Classe Modale

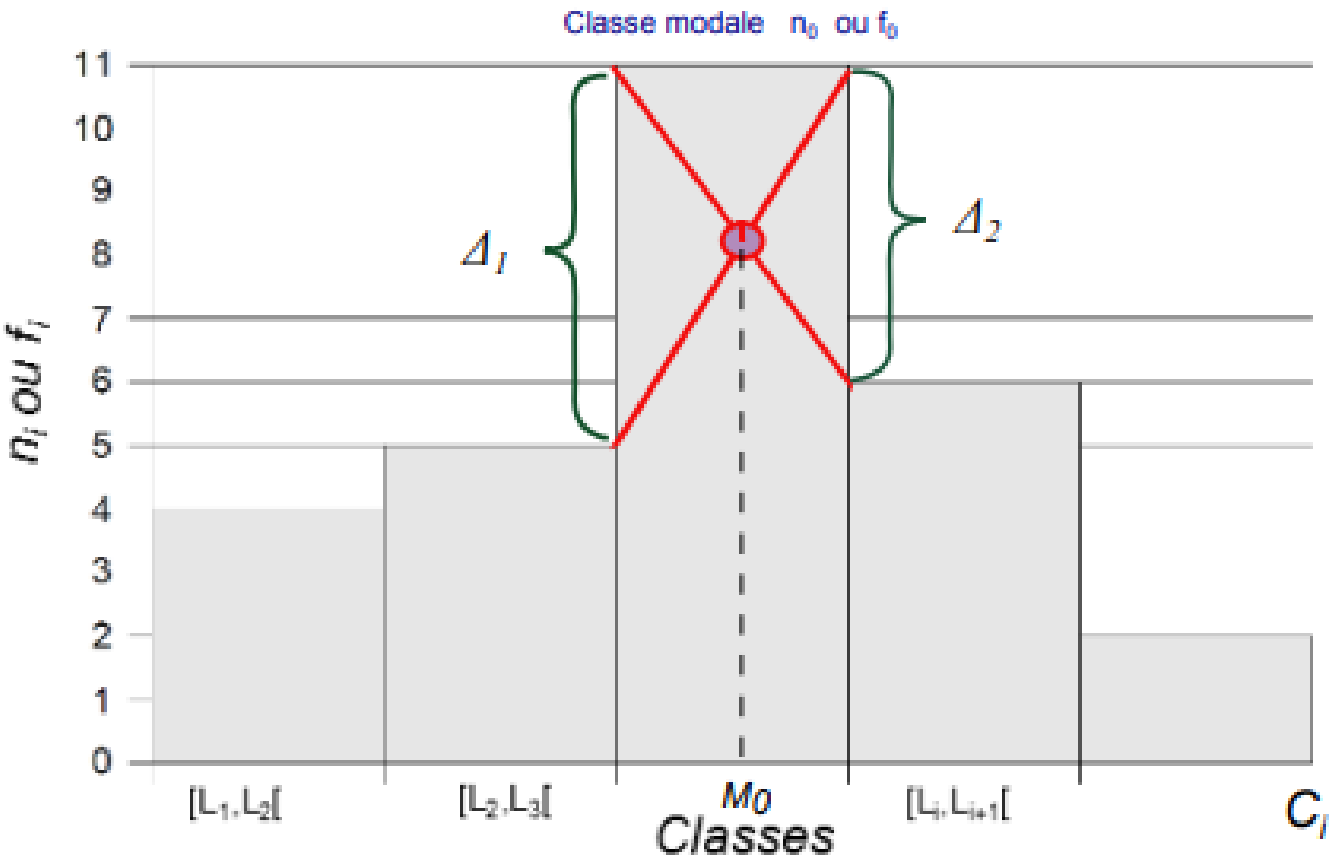
Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif partiel (ou la plus grande fréquence partielle). La quantité

$$M_0 = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} a_i$$

s'appelle le mode avec

- L_i *: la borne inférieure de la classe modale.*
- a_i *: le pas de la classe modale.*
- $\Delta_1 = n_0 - n_1$, $\Delta_2 = n_0 - n_2$ *ou bien* $\Delta_1 = f_0 - f_1$, $\Delta_2 = f_0 - f_2$.
- n_0 *et* f_0 *sont l'effectif et la fréquence associés à la classe modale.*
- n_1 *et* f_1 *sont l'effectif et la fréquence de la classe qui précède la classe modale.*
- n_2 *et* f_2 *sont l'effectif et la fréquence de la classe qui suit la classe modale.*

L'expression du mode est déterminée à partir de l'intersection des deux segments représentés ci-dessous *(Cette notion n'est pas unique)*.



Représentation ou détermination graphique du mode (cas continu)

La Médiane

C'est la valeur M_e telle que $F(M_e) = 0.5$ *(Cette valeur est unique)*.
Nous pouvons la déterminer graphiquement ou par calcul.

1. **Première méthode** : Graphiquement à partir de la formule

$$\tan(\alpha) = \frac{F(L_{i+1}) - F(L_i)}{L_{i+1} - L_i} = \frac{0.5 - F(L_i)}{Me - L_i}.$$

Plus précisément, dans la figure , nous mettons $F(x) = 0.5$ et $x = Me$.

2. **Deuxième méthode** : En utilisant directement la fonction de répartition donnée par

$$F(x) = \frac{f_{i+1}}{h}(x - L_i) + F_i.$$

Nous retrouvons donc

$$0.5 = \frac{f_{i+1}}{h}(Me - L_i) + F_i.$$

Paramètres de Dispersion

Monday, February 18, 2019

La variance est la quantité

$$Var(x) = \sum_{i=1}^{n} f_i(\overline{x} - C_i)^2$$

Pour le calcul, on utilise

$$Var(x) = \sum_{i=1}^{n} f_i C_i^2 - \overline{x}^2$$

La quantité

$$\sigma_X = \sqrt{Var(x)}$$

s'appelle l'écart type de la V.S X.

- La notion de médiane peut être généralisée pour donner la définition des quartiles (ou quantiles) :
  - Pour $i \in \{1, 2, 3\}$, la quantité Q_i tel que $F(Q_i) = i\ 4$ s'appelle le i^{em} quartile.
  - **Exemple :** Pour $i = 2$, Q_2 tel que $F(Q_2) = 2/4 = 0.5$. Donc, $Q_2 = M_e$.
- La détermination ou le calcul de Qi se fait exactement comme le calcul de la médiane (graphiquement ou analytiquement).

Interprétation des Quartiles

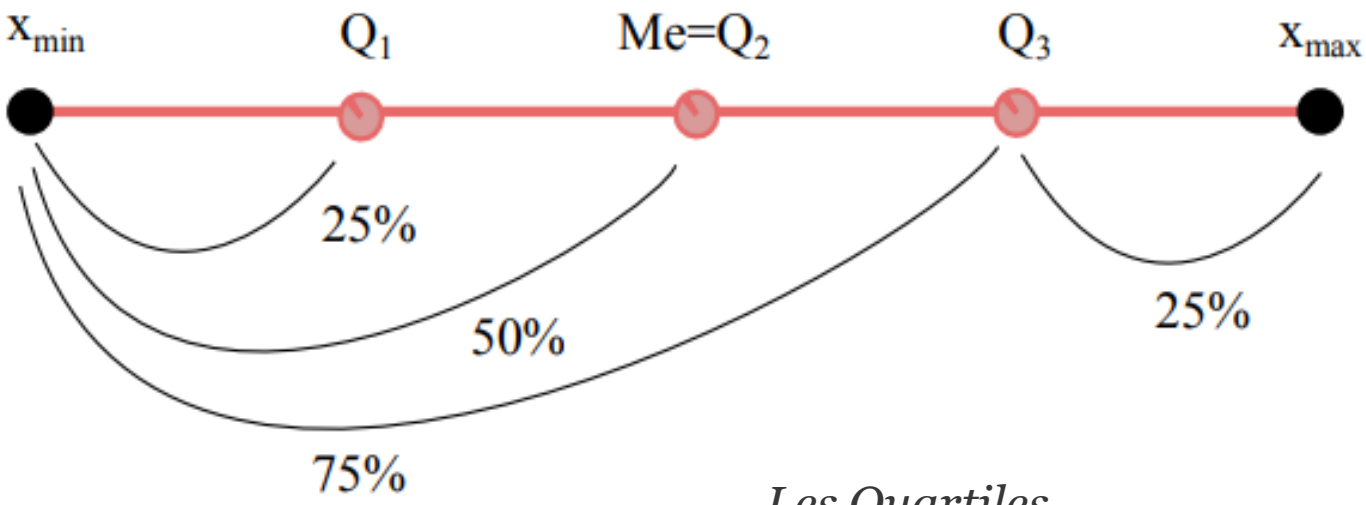
Interprétation des quartiles: Il y a 25 % d'individus dont la valeur du caractère est dans l'intervalle $[a_0, Q_1]$. De même pour les autres quartiles.

Ces intervalles s'appellent "intervalles interquartiles".

$Q_1 \longrightarrow 25\%,$

$Q_2 \longrightarrow 50\%,$

$Q_3 \longrightarrow 75\%.$



Les Quartiles

Exercice 12

- *Classer ces statistiques selon leurs natures (indicateur de position ou de dispersion)*

	Minimum	Moyenne	Écart-type	Mode	Médiane	Premier quartile
Position						
Dispersion						

Solution

Les natures des statistiques sont classées comme suit :

Position	Minimum, Moyenne, Médiane, Mode, Premier quartile
Dispersion	Écart-type

Exercice 13

Monday, February 18, 2019

- Chez un fabriquant de tubes de plastiques, on a prélevé un échantillon de 100 tubes dont on a mesuré le diamètre en décimètre.

1.94	2.20	2.33	2.39	2.45	2.50	2.54	2.61	2.66	2.85
1.96	2.21	2.33	2.40	2.46	2.51	2.54	2.62	2.68	2.87
2.07	2.26	2.34	2.40	2.47	2.52	2.55	2.62	2.68	2.90
2.09	2.26	2.34	2.40	2.47	2.52	2.55	2.62	2.68	2.91
2.09	2.28	2.35	2.40	2.48	2.52	2.56	2.62	2.71	2.94
2.12	2.29	2.36	2.41	2.49	2.52	2.56	2.63	2.73	2.95
2.13	2.30	2.37	2.42	2.49	2.53	2.57	2.63	2.75	2.99
2.14	2.31	2.38	2.42	2.49	2.53	2.57	2.65	2.76	2.99
2.19	2.31	2.38	2.42	2.49	2.53	2.59	2.66	2.77	3.09
2.19	2.31	2.38	2.42	2.50	2.54	2.59	2.66	2.78	3.12

Exercice 13

1. *Identifier la population, les individus, le caractère et son type.*
2. *En utilisant la méthode de Yule puis de Sturge, établir le tableau statistique (Faites débuter la première classe par la valeur 1.94).*
3. *Tracer l'histogramme de cette variable statistique.*
4. *Déterminer par le calcul la valeur du diamètre au-dessous de laquelle se trouvent 50% des tubes de plastique. Que représente cette valeur.*
5. *Déterminer par le calcul le pourcentage de tubes ayant un diamètre inférieur à 2.58.*

Solution

1 - Identification de cette épreuve statistique :

- Population : les tubes.
- Individus : le tube.
- Caractère : le diamètre.
- Type : quantitative continue.
- Modalités : 1.94,..., 3.12.

Exercice 13 - Solution

Solution

2 - Par la méthode de Yule, nous avons

$$k = 2.5\sqrt[4]{N} = 2.5\sqrt[4]{100} = 7.9 \simeq 8$$

Par la méthode de Sturge, nous avons

$$k = 1 + 3.3\log_{10}(N) = 1 + 3.3\log_{10}(100) = 7.6 \simeq 8$$

Nous avons donc l'amplitude qui égale

$$a_i = \frac{x_{\max} - x_{\min}}{k} \simeq 0.15$$

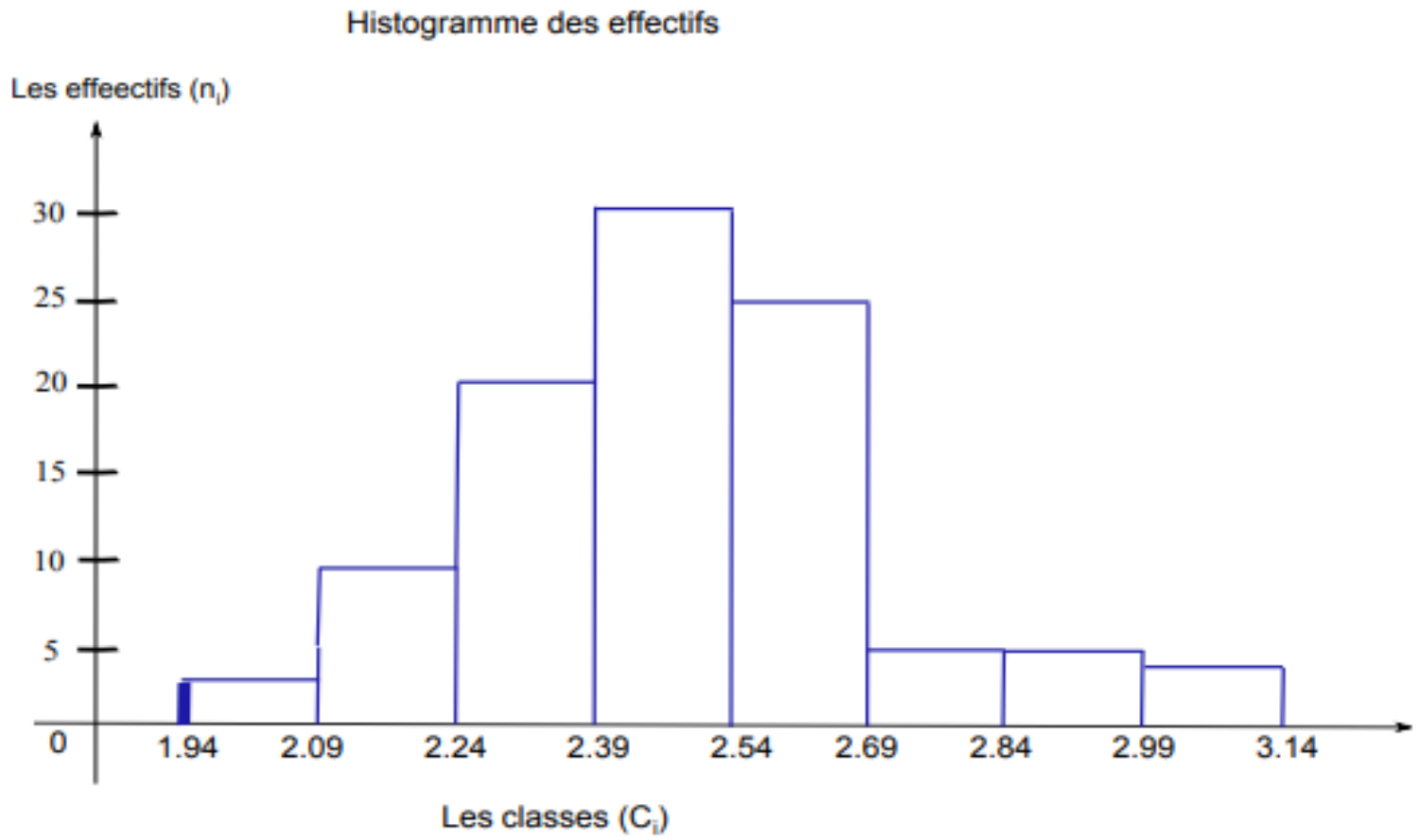
Nous obtenons le tableau statistique suivant :

X	n_i	f_i	N_i	F_i
$[1.94, 2.09[$	3	0.03	3	0.03
$[2.09, 2.24[$	9	0.09	12	0.012
$[2.24, 2.39[$	18	0.18	30	0.3
$[2.39, 2.54[$	29	0.29	59	0.59
$[2.54, 2.69[$	25	0.25	84	0.84
$[2.69, 2.84[$	6	0.06	90	0.90
$[2.84, 2.99[$	6	0.06	96	0.96
$[2.99, 3.14[$	4	0.04	100	1
$\sum$	100	1	\	\

Exercice 13 - Solution

Solution

3 - Nous établissons l'histogramme de cette variable, comme suit :



Exercice 13 - Solution

Solution

4 - Cette valeur représente la médiane. Le calcule se fait par extrapolation

$$\tan(\alpha) = \frac{0.59 - 0.3}{2.54 - 2.39} = \frac{0.5 - 0.3}{Me - 2.39}.$$

Nous trouvons $Me = 2.5$.

5 - Le calcule du pourcentage de tubes ayant un diamètre inférieur à 2.58 se fait de la même manière et nous avons

$$\tan(\alpha) = \frac{0.84 - 0.59}{2.69 - 2.54} = \frac{x - 0.59}{2.58 - 2.54}.$$

Nous trouvons que la valeur cherché est égale à 0.66 (66%).

Exercice 14

- Une étude sur le budget consacré aux vacances d'été auprès de ménages a donné les résultats suivants

Budget X	Fréquence cumulée	Fréquences
$[800, 1000[$	0.08	
$[1000, 1400[$	0.18	
$[1400, 1600[$	0.34	
$[1600, \beta[$	0.64	
$[\beta, 2400[$	0.73	
$[2400, \alpha[$	1	

Le travail demandé :

- *Certaines données sont manquantes. Calculer la borne manquante α sachant que l'étendue de la série est égale à 3200.*
- *Calculer les fréquences dans le tableau.*
- *Calculer la borne manquante β dans les deux cas suivants :*
  1. *Le budget moyen est égal à 1995.*
  2. *Le budget médian est égal à 1920.*

Exercice 14 - Solution

Solution

- On sait que l'étendue est égale au maximum moins le minimum. Ainsi, $3200 = x_{max} - x_{min} = \alpha - 800$, et donc $\alpha = 4000$.

- Nous complétons le tableau comme suit :

Budget X	Fréquence cumulée	Fréquences
$[800, 1000[$	0.08	0.08
$[1000, 1400[$	0.18	0.1
$[1400, 1600[$	0.34	0.16
$[1600, \beta[$	0.64	0.3
$[\beta, 2400[$	0.73	0.09
$[2400, \alpha[$	1	0.27

Exercice 14 - Solution

- *Le calcul de la borne manquante β dans le cas où le budget moyen est égal à 1995 se fait comme suit :*

$$\overline{x} = 1995 = 0.08 \times 900 + 0.1 \times 1200 + 0.16 \times 1500 + 0.3 \times \frac{1600 + \beta}{2} + 0.09 \times \frac{\beta + 2400}{2} + 0.27 \times 3200$$

- *Ce qui implique que*

$$1644 + 0.195 \times \beta = 1995 \qquad donc \qquad \beta = 1800$$

- *Le calcule la borne manquante β dans le cas où le budget médian est égal à 1920, c'est à dire, Me = 1920 se fait comme suit : il faut raisonner par interpolation linéaire sur l'intervalle [1600 – β[.*

  On pose le rapport des distances :

$$\frac{1920 - 1600}{\beta - 1600} = \frac{0.5 - 0.34}{0.64 - 0.34} \qquad donc \qquad \beta = 2200$$

La section suivante sera complétée sous peu

- Pour mesurer le degré d'homogénéité d'une population, certaines techniques utilisent la notion de variance.
- La variance se calcule en calculant les écarts (de chaque observation) par rapport à la moyenne, en élevant ces écarts au carré et en divisant par le nombre d'observations.
- La formule de la variance est :
- $= {}^{1}\sum^{n} (x_i - x)^2 \, {}^{n}\, i=1$
- On peut appréhender la variance comme étant une surface. Plus elle est importante, plus la distribution s'éloigne de la moyenne. Si on considère cette surface comme étant un carré, la racine carrée de la variance représentera un coté de ce carré. Ce sera l'écart-type qui sera lui aussi une mesure de la dispersion autour de la moyenne.
- Comme la distance euclidienne, la variance permet de découper une population en sous ensembles homogènes.

Exemple de Variance

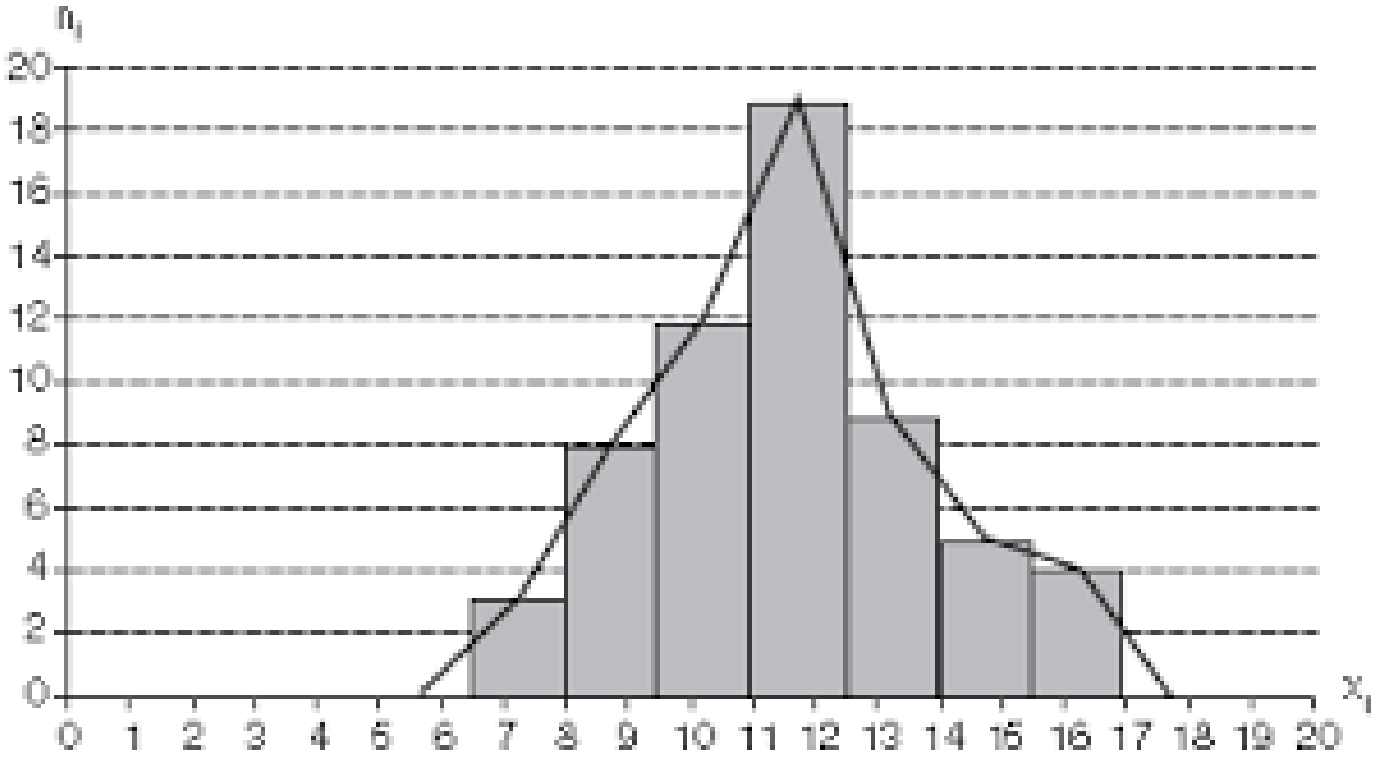
- Considérons les notes en math et en français obtenues par des élèves d'une classe :

Dans cet exemple, la variance des notes de maths est de 39,43, celles des notes de français de 8,28.

De même, l'écart type des notes de maths est de 6,27 et celui des notes de français de 2,87.

Ainsi, le professeur de math construit une échelle de différenciation plus importante que le professeur de français.

	Maths	Français
Elève 1	3	7
Elève 2	4	8
Elève 3	6	9
Elève 4	11	11
Elève 5	16	13
Elève 6	18	14
Elève 7	19	15
Moyenne	11	11


n_i
20
18
16
14
12
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
x_i


Note en Statitique en 1993
Poulation de 50 étudiants
Fréquences absolues
ou effectifs
18
15
12
9
6
3
0
5
10
15
20
Polygone des fréquences absolues
Classes de notes
en Statistique

Analyse des Données

Analyse des données

- L'analyse des données est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives.
- Certaines méthodes aident à faire ressortir les relations pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon à faire apparaitre clairement ce qui les rend homogènes.
- L'analyse des données permet de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci. Le succès de cette discipline dans les dernières années est dû, dans une large mesure, aux représentations graphiques fournies. Ces graphiques peuvent mettre en évidence des relations difficilement saisies par l'analyse directe des données ; mais surtout, ces représentations ne sont pas liées à une opinion « a priori » sur les lois des phénomènes analysés contrairement aux méthodes de la statistique classique.

- L'Analyse de Données est un ensemble de méthodes mathématiques qui permettent de traiter des "items" dans un tableau. Ces items sont décrit par un ensemble de variables.

- L'objectif de l'Analyse de Données est de traiter des informations du type :
  - quels sont les items identiques ou dissemblables,
  - quelles sont les relations entre les items et les variables associées.
- Le plus souvent les résultats de ces analyses permettent d'ajuster et d'optimiser les politiques ou stratégies commerciales des entreprises.

- L'analyse des données est un processus qui permet de transformer une masse d'informations en information structurée permettant la prise de décision marketing.

  L'analyse des données est utilisée dans un grand nombre de contexte marketing.

  L'analyse de données est révolutionnée par l'émergence du phénomène big data.

Représentation des Données

- Plusieurs niveaux de description statistique :
  - présentation brute des données,
  - présentation par tableaux numériques,
  - représentations graphiques
  - résumés numériques fournis par un petit nombre de paramètres caractéristiques.

Les **données brutes** ou **tableau élémentaire** le tableau relevant pour chaque unité statistique, la modalité de la variable étudiée.

Exemple de Tableau de Données Brutes

NUMERO	SALAIRE	SEXE	AGE	ANC	NIVEAU
1	129472	F	42	3	B
2	212696	M	54	10	B
3	210888	M	47	10	A
4	213692	M	47	1	B
5	202408	M	44	5	B
6	196132	M	42	10	A
7	97580	M	30	5	A
8	97580	F	52	6	A
9	172496	M	48	8	A
10	95900	F	58	4	A
11	212696	M	46	4	C
12	234060	M	36	8	C
13	225176	M	49	10	B
14	197532	F	55	10	B
15	179536	M	41	1	A
16	213716	F	52	5	B
17	186296	M	57	8	A
18	235872	F	61	10	B
19	212696	M	50	5	A
20	214508	M	47	10	B
21	196132	M	54	5	B
22	219924	M	47	7	A
23	250120	M	50	10	B
24	110100	F	38	3	A
25	97580	M	31	5	A
26	227536	M	47	10	A

Tri à Plat

- Le tri à plat est l'une des opérations statistiques les plus communes – on compte le nombre d'individus par modalité ou valeur :
  - Ce nombre est l'effectif ou la fréquence absolue de chaque modalité
  - On appelle effectif de la modalité x_i, le nombre n_i de fois que cette modalité est observée. N est l'effectif total
- Le tri à plat est la transformation qui permet de passer du tableau des données brutes au tableau de la distribution statistique présentant les modalités et les effectifs, les modalités étant classées par ordre croissant. (si la variable est ordinale ou si elle est quantitative)

Tableaux de Distribution

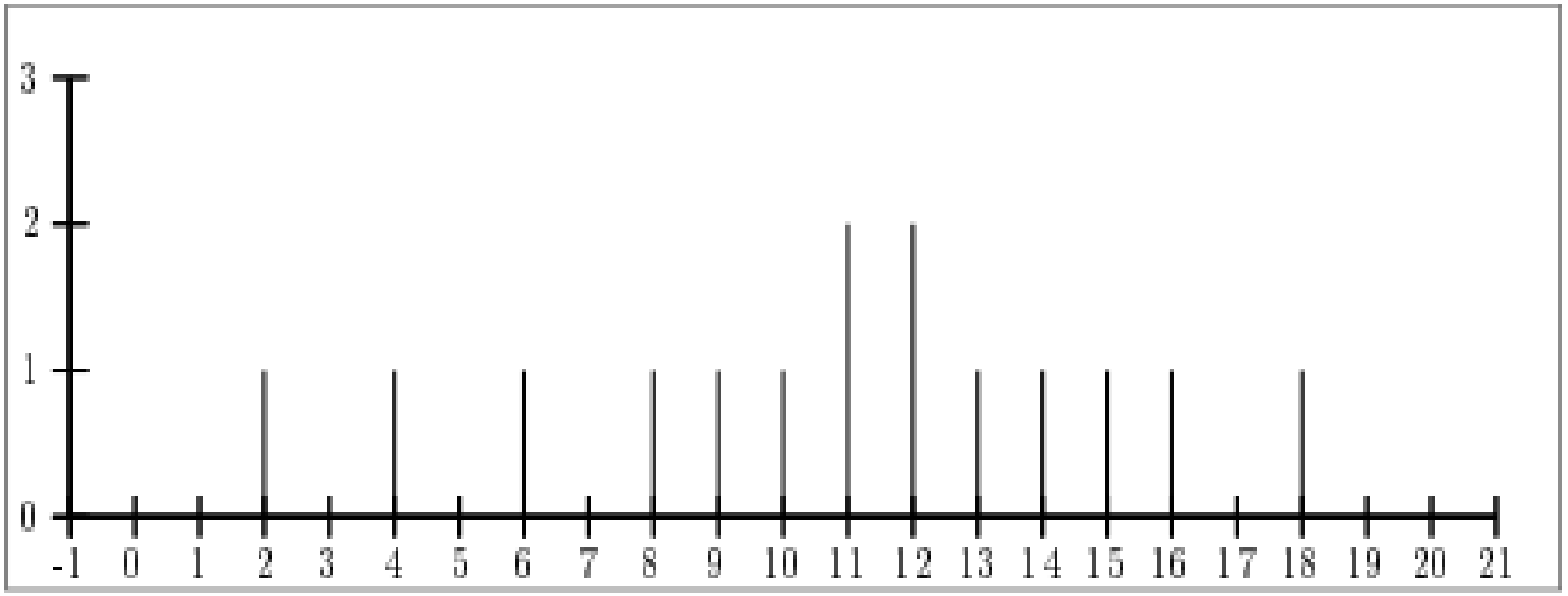
- Le tableau de distribution de fréquences est un mode synthétique de présentation des données. Sa constitution est immédiate dans le cas d'un caractère discret mais nécessite en revanche une transformation des données dans le cas d'un caractère continu.

- Certaines techniques statistiques cependant ne fonctionnent qu'avec des variables qualitatives; d'où la nécessité de recourir à la discrétisation.

Tableau de Répartition de la Variable Note à l'Examen de Statistiques

Note à l'Examen de Statistique	Effectifs	Fréquences
k=0	0	0
k=1	0	0
k=2	1	1/15
k=3	0	0
k=4	1	1/15
k=5	0	0
k=6	1	1/15
k=7	0	0
k=8	1	1/15
k=9	1	1/15
k=10	1	1/15
k=11	2	2/15
k=12	2	2/15
k=13	1	1/15
k=14	1	1/15
k=15	1	1/15
k=16	1	1/15
k=17	0	0
k=18	1	1/15
k=19	0	0
k=20	0	0

De façon générale, pour représenter le tableau précédent, on pourrait utiliser un diagramme en bâtons



Néanmoins cette forme se prête difficilement à l'interprétation.

Interprétation et Représentation des Données

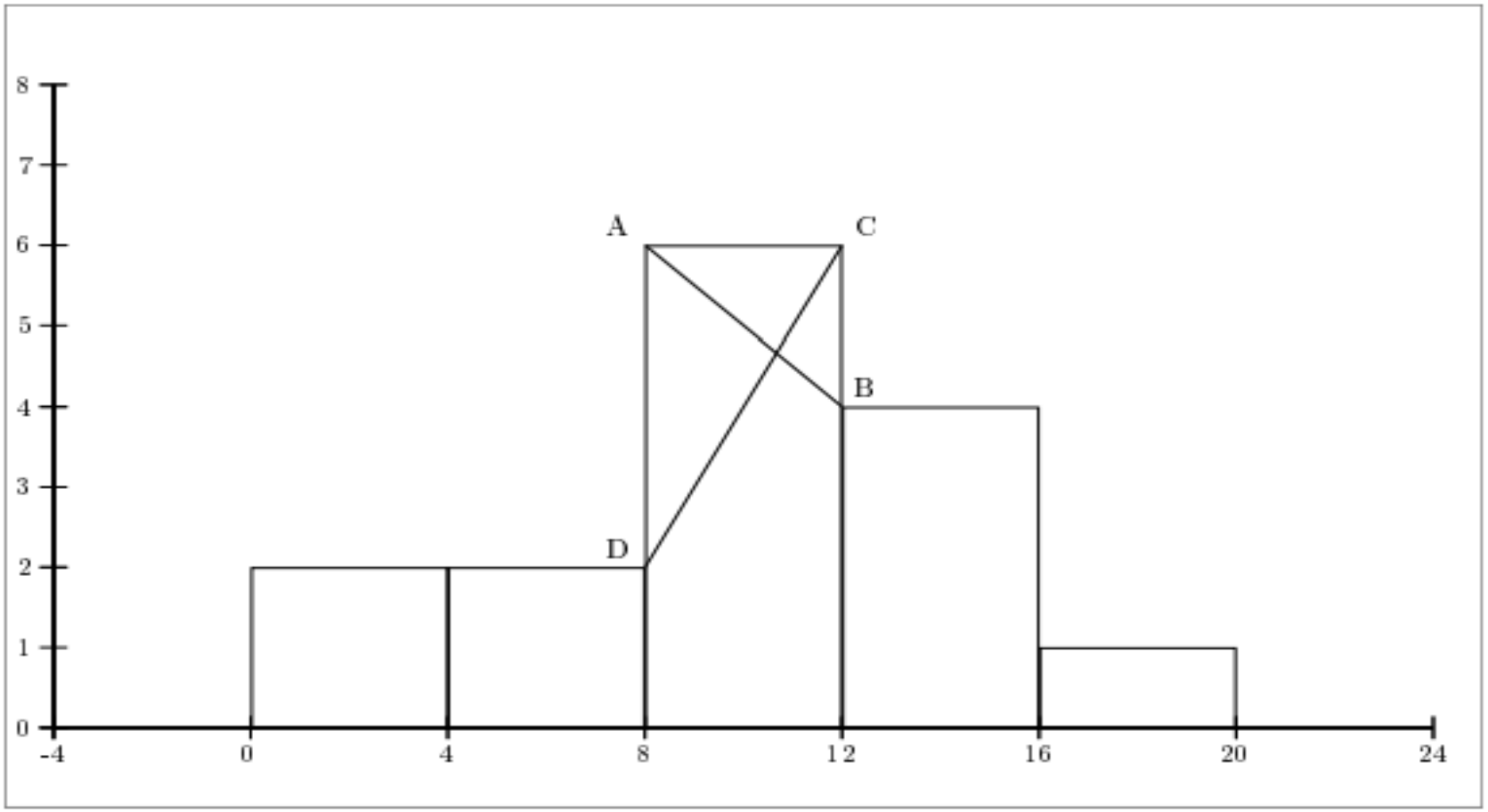
Pour remédier à cette difficulté d'interprétation, il faut créer des classes de notes[1] (nombre d'individus ayant obtenu des notes comprises entre 0 et 4, entre 4 et 8, . . .).

Cette approche nous permet d'obtenir une variable dite classée[2].

Tableau de Répartition de la variable classée Note à l'Examen de Statistique

variable classée	Effectifs	Fréquences
$[0, 4]$	2	$2/15$
$]4, 8]$	2	$2/15$
$]8, 12]$	6	$6/15$
$]12, 16]$	4	$4/15$
$]16, 20]$	1	$1/15$

Histogramme des Effectifsde la variable classée Note à l'Examen de Statistique



La représentation graphique des effectifs de chaque classe s'appelle l'histogramme des effectifs; on peut de la même façon réaliser l'histogramme des fréquences.

Discrétisation

Discrétisation

- Exprimer graphiquement une information implique souvent sa simplification. Ainsi en découpant en classes une série de données, la discrétisation réduit une variable quantitative en variable ordonnée.
- Pour réaliser une discrétisation, il faut choisir le nombre de classes et les bornes de classe. Pour réaliser une bonne discrétisation, il faut justifier à la fois le nombre de classes et les bornes de classe, le terme "bonne" faisant référence à des critères explicitement définis.
- La discrétisation soulève deux questions:
  - Combien de classes créer?
  - Où placer les bornes des classes?

Discrétisation

- La discrétisation soulève deux questions:
  - Combien de classes créer?
  - Où placer les bornes des classes?

- Pour conserver au mieux l'information et rendre compte de la répartition géographique du phénomène, les bornes doivent être choisies de manière à créer des classes homogènes et distinctes entre elles.
  - La forme de la distribution statistique influe sur le choix de la méthode de discrétisation.

Discrétisation – Nombre de Classes

- Le nombre de classes doit évidemment être en rapport avec le nombre d'individus de la série de valeurs: sa longueur. Ainsi pour des valeurs relatives aux 80 étudiants de la filière Mgt de l'UIC, inutile de discrétiser en 2 classes (appauvrissement excessif) ou en 20 classes (risque de classes vides).
- Outre la longueur de la série, d'autres règles imposent certaines limites. L'expression graphique ne supporte qu'un nombre de classes limité.
  - Si on veut aisément distinguer les classes (perception sélective), leur nombre doit se limiter à 5 - 7 (selon le nombre d'individus).
  - Si, au contraire, le but est de montrer une progression plus ou moins continue, un gradient, le nombre de classes peut être nettement supérieur. Tel est le cas des cartes de diffusion.

Discrétisation – Nombre de Classes

Il existe quelques formules "toute faites" pour déterminer, à l'aveugle, le nombre **k** de classes à partir du nombre **n** de données *(dans cet exemple – n = 20, max = 50, min = 10)* :

Appellation	Formule
Brooks-Carruthers	$5 \times \log_{10}(n)$
Huntsberger	$1 + 3.332 \times \log_{10}(n)$
Sturges	$\log2(n + 1)$
Scott	$(max - min)/(3.5 \times \sigma \times n^{-1/3})$
Freedman-Diaconis	$(max - min)/(2 \times IQ \times n^{-1/3})$

Les 2 dernière approches exploitent plus d'informations en provenance des données

Méthodes de Regroupement en Classes

Les méthodes de regroupement en classes les plus utilisées sont:

- **Regroupement en classes par équidistance** : appelé aussi **discrétisation en classes d'amplitude égale**. L'écart entre les classes est constant. Il suffit de diviser la dynamique de la série (maximum – minimum) par le nombre de classes souhaité.

  Avantages et inconvénients: Facile à réaliser, mais la qualité du regroupement en classes dépend de la distribution des valeurs de la série.

  Cette méthode risque de produire des classes vides si la distribution est asymétrique ou comporte de fortes discontinuités pour une étendue donnée :

  - exemple : Une classe où il y a un étudiant excellent qui obtient de très bonnes notes tandis que les autres ont pour la plupart des notes moyennes – la densité de la population se concentre au niveau de la moyenne => *problème de dispersion*)

Méthodes de Regroupement en Classes 2

- **Regroupement en classes par progression arithmétique** : contrairement au regroupement en classes par équidistance, l'amplitude des classes augmente selon une progression arithmétique à raison de R.

  R étant la dynamique (étendue) de la série divisée par l'addition des classes.
  - Exemple pour 5 classes:
    R = (max – min) / (1 + 2 + 3 + 4 + 5)
- L'amplitude des classes: classe n°1: min à min+R, classe n° 2: min+R à min+2R, etc.
- Avantages et inconvénients: La méthode est conçue pour les distributions asymétriques avec beaucoup de valeurs faibles et peu de valeurs fortes.

  Si, au contraire, la distribution s'apparente à la loi normale, la méthode n'est pas adaptée.

Méthodes de Regroupement en Classes 3

- **Regroupement en classes par équifréquence** : appelée également **discrétisation selon les quantiles** ou **par équipopulation**. Toutes les classes ont, dans la mesure du possible, le même nombre d'individus.

  La réalisation repose sur un tri par ordre croissant des individus de la série et le calcul du nombre idéal d'individus par classe (nombre d'individus / nombre de classes)[1].

  Les bornes des classes peuvent être définies par les valeurs extrêmes de chaque classe.

  Avantages et inconvénients : La discrétisation selon les quantiles diminue le poids des valeurs extrêmes. Elle les nivelle en les regroupant avec des valeurs plus proches de la moyenne pour équilibrer les classes.

  Par conséquent, la carte est visuellement la plus riche, la plus équilibrée mais pas forcément la plus juste car elle masque les fortes discontinuités (par exemple l'étudiant brillant sera regroupé avec les autres étudiants moyens).

  Plus généralement, la méthode peut regrouper des individus assez éloignés et, par conséquent, ne pas respecter les seuils dans la distribution.

Méthodes de Regroupement en Classes 4

- **Regroupement en classes selon les moyennes emboîtées** : la moyenne arithmétique divise la série en deux classes primaires. Leurs moyennes respectives les divisent à leur tour en deux et ainsi de suite. Le nombre de classes est toujours un multiple de 2 ce qui peut s'avérer contraignant.

- Avantages et inconvénients : La méthode est basée sur la moyenne considérée comme centre de gravité de la distribution ainsi que sur les moyennes des sous-ensembles.

  L'inconvénient majeur est la rigidité de la méthode en nombre de classes: multiple de 2.

  La moyenne peut également découper des ensembles de valeurs proches (non respect des seuils naturels de la distribution).

Méthodes de Regroupement en Classes 5

- **Regroupement en classes selon la méthode de Jenks** : Cette méthode se base sur le principe de ressemblance / dissemblance en calculant la distance paramétrique entre toutes les valeurs de la série[1].

Avantages et inconvénients : Il s'agit d'une méthode respectant l'allure de la série mais elle ne permet pas de rendre comparables plusieurs cartes d'une même série.

Cette méthode est certainement la plus adaptée des méthodes statistiques. Sa mise en œuvre nécessite l'emploi d'un logiciel statistique.

Méthodes de Regroupement en Classes 6

- **Regroupement en classes manuel** : appelé aussi méthode des seuils observés.

  Contrairement aux autres méthodes, il s'agit d'une méthode manuelle où l'opérateur découpe la série visualisée graphiquement (histogramme de fréquence) selon les discontinuités (seuils). Le nombre de classes résultant est un compromis entre l'allure de la série (seuils, discontinuités) et le nombre de classes initialement projeté.

- Avantages et inconvénients : La méthode convient d'autant plus que la distribution est «typée» (série asymétrique, plurimodale). De ce point de vue il s'agit de la méthode opposée à la discrétisation par équifréquence notamment préconisée en cas de répartition uniforme des valeurs.

  Le regroupement en classes manuel s'appuyant sur des seuils observés est certainement la méthode de bon sens dans la mesure où elle produit des cartes équilibrées graphiquement tout en épousant la distribution de la série. Elle a l'avantage d'obliger l'opérateur à analyser graphiquement la série de valeurs[1].

  La significativité des discontinuités observées dépend néanmoins de la longueur de la série. Ainsi, le constat d'un seuil dans une population de 500 individus est autrement plus significatif que l'observation de la même discontinuité dans une population nettement plus réduite.

Méthodes de Discrétisation

Méthode de discrétisation	Principe	Forme de distribution
Par seuils observés	les bornes sont crées par observation de la distribution	dissymétrique bimodale
Amplitudes égales	les classes possèdent la même amplitude	uniforme symétrique
Effectifs égaux (quantiles)	les bornes sont construites en réalisant des classes d'effectifs égaux	dissymétrique bimodale
Discrétisation standardisée	déterminées selon une fraction d'écart-type par rapport à la moyenne	symétrique
Progression géométrique	les classes sont découpées selon une progression géométrique	dissymétrique
Progression arithmétique	les classes sont établies selon une progression arithmétique	dissymétrique
Jenks	maximise la variance inter-classe et minimise la variance intra-classe	dissymétrique bimodale
Q6	variante de la discrétisation selon les quantiles avec isolement des classes extrêmes de la série	dissymétrique bimodale

Discrétisation -- Exemples

X
14
15
17
19
21
23
25
26
27
28
29
31
38
39
41
44
45
49
51
53

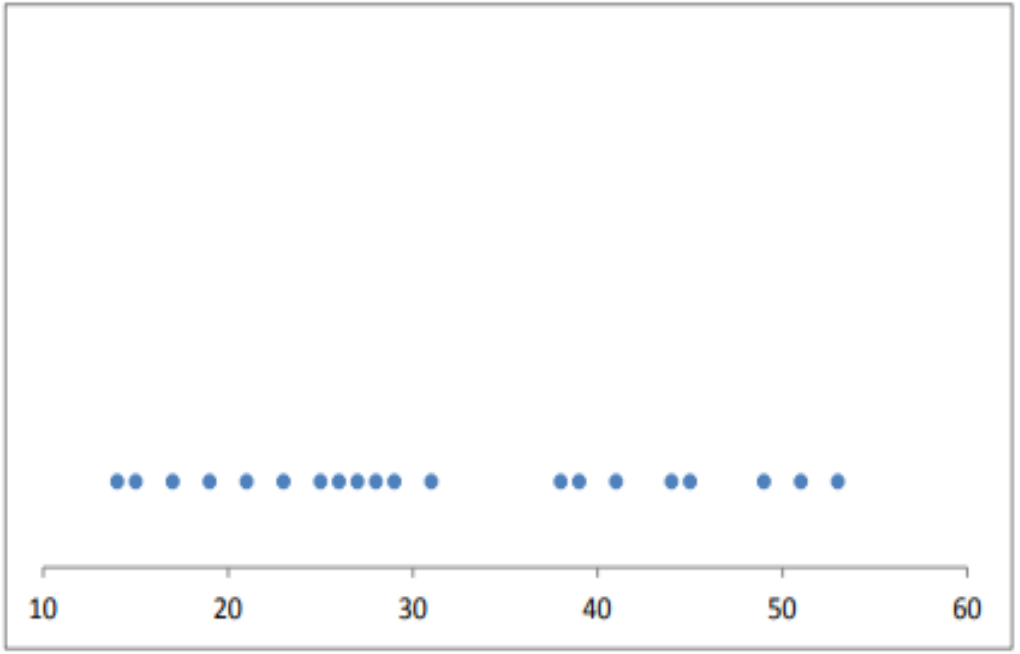
n	20

Min	14
Max	53

1er quartile	22.5
Mediane	28.5
3e quartile	41.75

Moyenne	31.75
Ecart-type	12.07

Quelques caractéristiques disponibles :



Ex : Intervalles de largeurs (amplitudes) égales

K : nombre d'intervalles est fixé (comment ?)

Construire des intervalles d'amplitudes égales

$$a = \frac{\max - \min}{K}$$

Calcul de l'amplitude à partir des données (a)

On en déduit les (K-1) bornes (b_1, b_2, etc.)

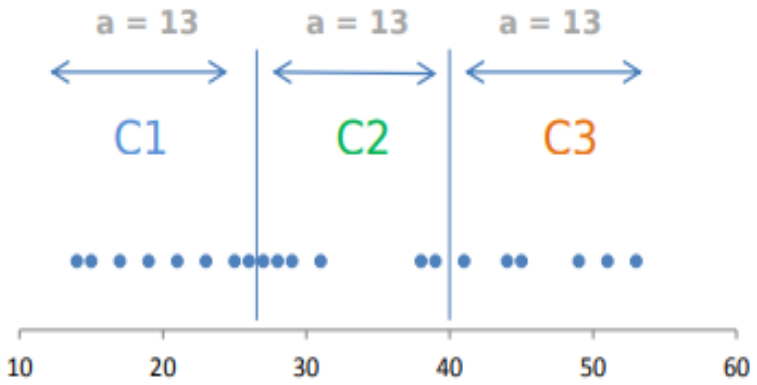
$$b_1 = \min + a$$

$$b_2 = b_1 + a = \min + 2 \times a$$

$$\ldots$$

X	Classes
14	C1
15	C1
17	C1
19	C1
21	C1
23	C1
25	C1
26	C1
27	C2
28	C2
29	C2
31	C2
38	C2
39	C2
41	C3
44	C3
45	C3
49	C3
51	C3
53	C3

K	3
max	53
min	14
a	13
b1	27.0
b2	40.0



C1 : x < 27

C2 : 27 ≤ x < 40

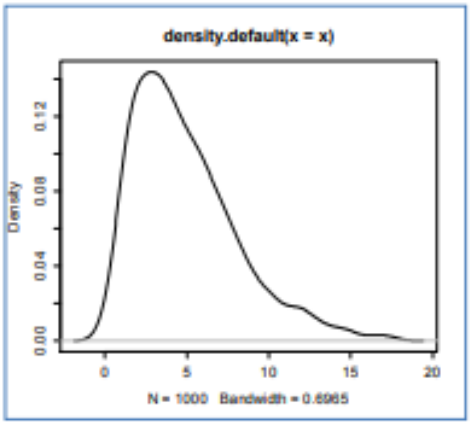
C3 : x ≥ 40

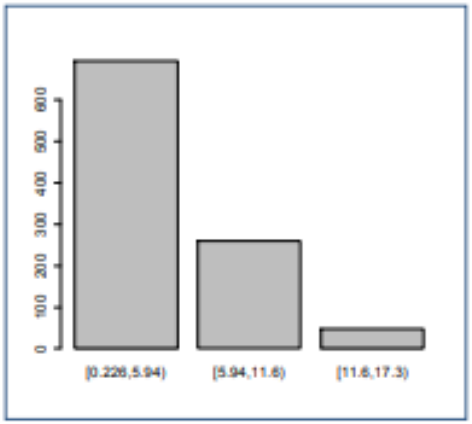
Le sens de l'inégalité est arbitraire

Ex : Intervalles de largeurs (amplitudes) égales

+ Rapidité de calcul et simplicité (facile à expliquer)

Ne modifie pas la forme de la distribution des données





- Choix de K arbitraire, pas toujours évident

Sensibilité aux points extrêmes (min ou max)

Possibilité d'avoir des intervalles avec très peu d'individus voire vides

K nombre d'intervalles est fixé (comment ?)

Construire des intervalles de fréquence égales

Ex. quantile d'ordre 0.25 = 1er quartile ; quantile d'ordre 0.5 = médiane ; etc.

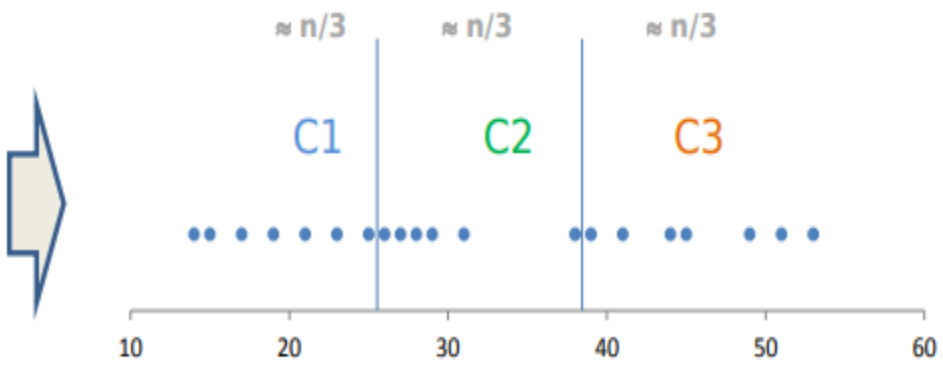
Calcul des quantiles à partir des données (q_1, q_2)

Quantiles = bornes (q_1, q_2, etc.)

Ex : Intervalles de largeurs (amplitudes) égales

X	Classes
14	C1
15	C1
17	C1
19	C1
21	C1
23	C1
25	C1
26	C2
27	C2
28	C2
29	C2
31	C2
38	C2
39	C3
41	C3
44	C3
45	C3
49	C3
51	C3
53	C3

q(0.33)	25.33
q(0.66)	38.67



C1 : x < 25.33

C2 : 25.33 ≤ x < 38.67

C3 : x ≥ 38.67

Le sens de l'inégalité est arbitraire

Ex : Intervalles de largeurs (amplitudes) égales

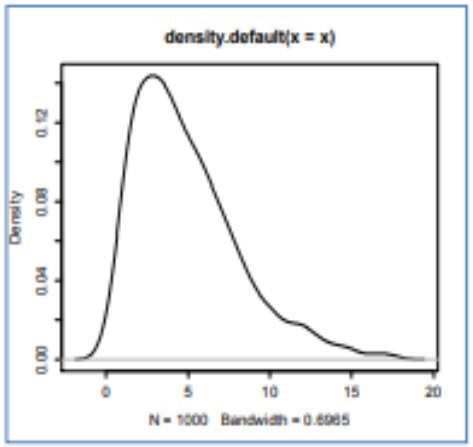
+

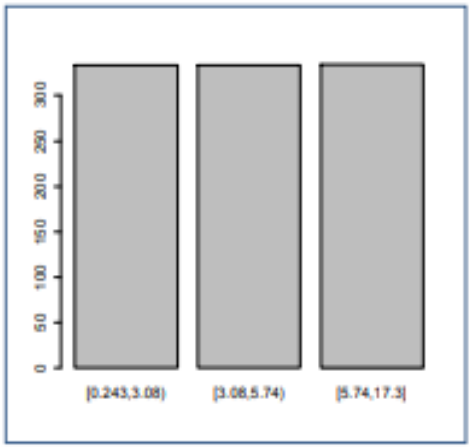
Rapidité (~) de calcul et simplicité (facile à expliquer)

« Lissage » des points extrêmes

Intervalles avec un nombre déterminé d'individus

Egalise la distribution des données





-

Choix de K arbitraire, pas toujours évident

Seuils ne tenant pas compte des proximités entre les valeurs

Nombre de Classes par Méthode

Appellation	Formule	K « calculé » (sans arrondi)
Brooks-Carruthers	$5 \times \log_{10}(n)$	6.51
Huntsberger	$1 + 3.332 \times \log_{10}(n)$	5.34
Sturges	$\log2(n + 1)$	4.39
Scott	$(max - min)/(3.5 \times \sigma \times n^{-1/3})$	2.50
Freedman-Diaconis	$(max - min)/(2 \times IQ \times n^{-1/3})$	2.75

σ: écart-type

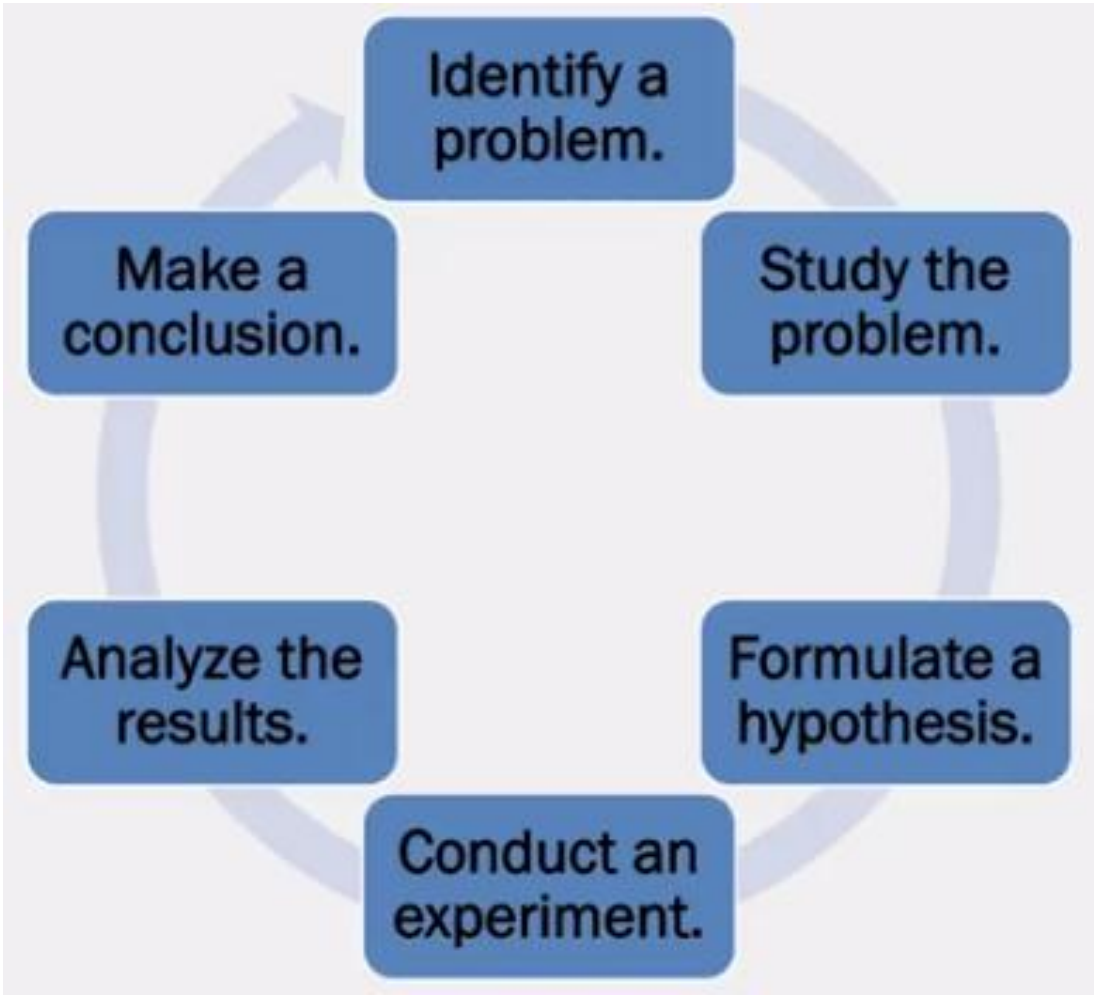
IQ : intervalle interquartiles

Monday, February 18, 2019

Scientific Study Method

Steps in Scientific Method

Monday, February 18, 2019



- There are two types of hypotheses :
  - Null hypothesis (H_o)
    - What you expect to happen
    - Ex : the research in West Virginia University expected that the result would lead to VW emissions would be the same as reported by the car manufacturer
  - alternate hypothesis ($H_a$or H_1)
    - Everything else that can happen
    - Ex : the car emissions were higher than those reported by the VW

- H_o and H_a must be mutually exclusive. Both cannot be true and collectively they cover all possibilities.

- H_o is a belief that we try to reject (or confirm) using sample evidence.

Exemple – Moyenne de Population

- Exemple contrôle de la qualité[1] : vérification du poids imprimé sur un sac de chocolat fabriqué par l'entreprise X.
- Le contrôle de la qualité prélève des échantillons de 50 sacs pour en vérifier l'exactitude.
- Le travail de l'ingénieur qualité est de s'assurer que les spécifications (fournies par l'entreprise) sont respectées. Ainsi, chaque fois que des contrôles de qualité sont effectués, on effectue des tests d'hypothèses.

  Dans ce cas, notre conviction actuelle est que les sacs sont remplis correctement. Et l'alternative serait qu'ils ne le soient pas.

$H_0 = 300g$

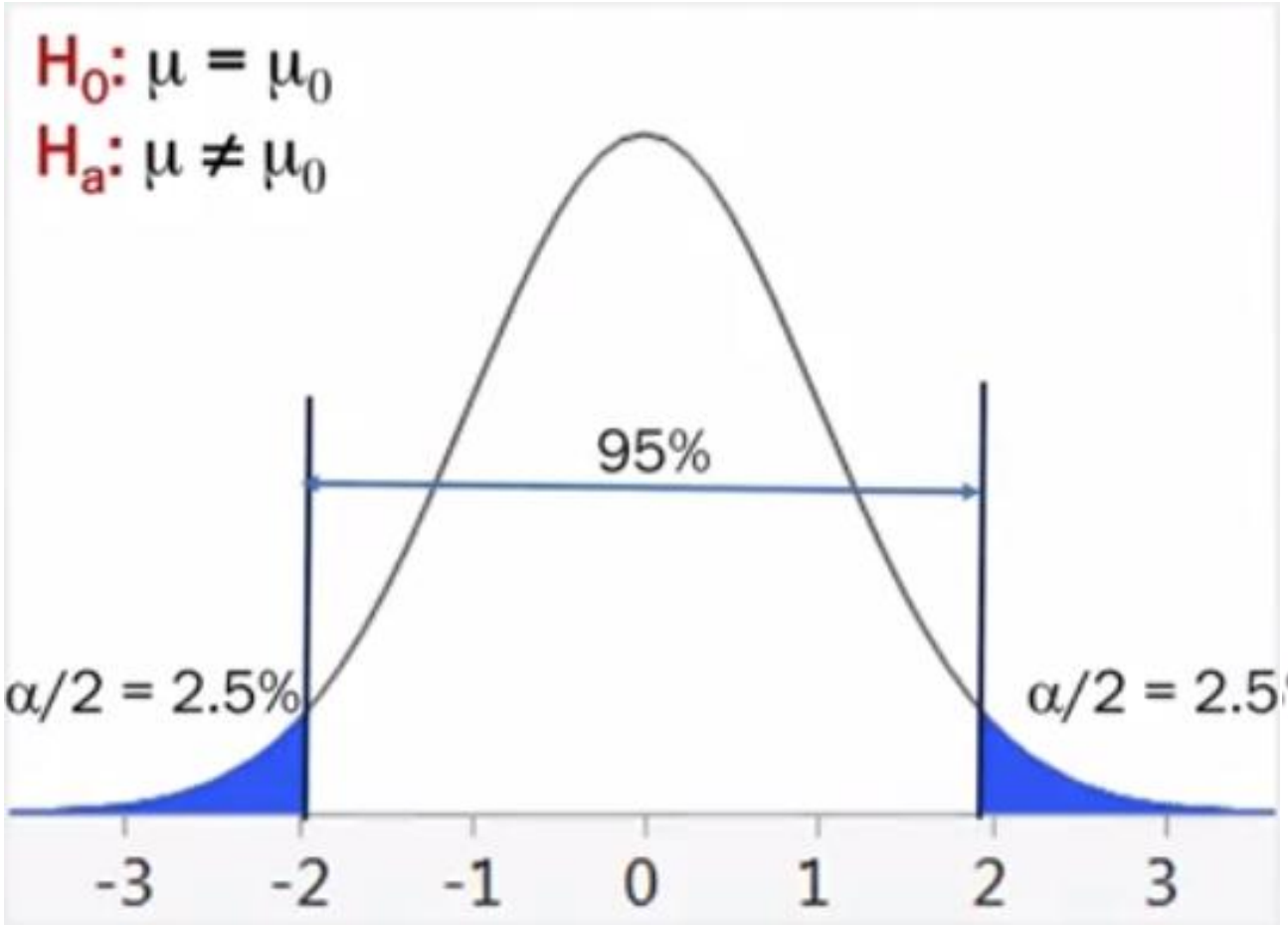
$H_1 \neq 300g$

Exemple – Moyenne d'Admission à l'UIC

- Un autre exemple : La moyenne au bac des étudiants qui intègrent l'UIC en 1è année.
- L'université estime n'accepter que les étudiants ayant une mention au bac
- L'hypothèse est donc :

  H0 ≥ 12 *(les étudiants ont une moyenne minimale de 12)*

  H1 < 12 *(les étudiants ont une moyenne inférieure à 12)*


$H_0: \mu = \mu_0$
$H_a: \mu \neq \mu_0$
95%
α/2 = 2.5%
α/2 = 2.5%
-3
-2
-1
0
1
2
3

Analyse et Types D'Erreurs

Quand on entreprend une analyse de données, on peut commettre 2 types d'erreurs :

- Erreur de type 1 (ou α) – Rejeter H_0 incorrectement (dans les processus de fabrication, on l'appelle « le risque du fabricant »)
- Erreur de type 2 (ou β) – Confirmer H_0 incorrectement (dans les processus de fabrication, on l'appelle « le risque du consommateur »)

Ho = 300g

H1 ≠ 300g

Decision		Reality	
		H_0: μ = 300 is True	H_0: μ = 300 is False
Reject	H_0: μ = 300	Type I Error	√
Retain	H_0: μ = 300	√	Type II Error