

STATISTIQUE DESCRIPTIVE

Benchikh Tawfik

Faculté de Médecine, UDL, SBA

1^{ère} année Médecine

21/28 Septembre 2016



PLAN DU COURS

1 INTRODUCTION

2 TERMINOLOGIE DE BASE

3 VARIABLE QUANTITATIVE DISCRÈTE

4 VARIABLE QUANTITATIVE CONTINUE

5 VARIABLE QUALITATIVE

6 EXERCICE



PLAN DU COURS

- 1 INTRODUCTION
- 2 TERMINOLOGIE DE BASE
- 3 VARIABLE QUANTITATIVE DISCRÈTE
- 4 VARIABLE QUANTITATIVE CONTINUE
- 5 VARIABLE QUALITATIVE
- 6 EXERCICE



PLAN DU COURS

- 1 INTRODUCTION
- 2 TERMINOLOGIE DE BASE
- 3 VARIABLE QUANTITATIVE DISCRÈTE
- 4 VARIABLE QUANTITATIVE CONTINUE
- 5 VARIABLE QUALITATIVE
- 6 EXERCICE



PLAN DU COURS

- 1 INTRODUCTION
- 2 TERMINOLOGIE DE BASE
- 3 VARIABLE QUANTITATIVE DISCRÈTE
- 4 VARIABLE QUANTITATIVE CONTINUE
- 5 VARIABLE QUALITATIVE
- 6 EXERCICE



PLAN DU COURS

- 1 INTRODUCTION
- 2 TERMINOLOGIE DE BASE
- 3 VARIABLE QUANTITATIVE DISCRÈTE
- 4 VARIABLE QUANTITATIVE CONTINUE
- 5 VARIABLE QUALITATIVE
- 6 EXERCICE

PLAN DU COURS

- 1 INTRODUCTION
- 2 TERMINOLOGIE DE BASE
- 3 VARIABLE QUANTITATIVE DISCRÈTE
- 4 VARIABLE QUANTITATIVE CONTINUE
- 5 VARIABLE QUALITATIVE
- 6 EXERCICE



INTRODUCTION

- La statistique descriptive a pour but de résumer l'information contenue dans les données de façon à en dégager les caractéristiques essentielles sous une forme simple et intelligible.
- Les deux principaux outils de la statistique descriptive sont les représentations graphiques et les indicateurs statistiques.

STATISTIQUE: DÉFINITION

- On appelle statistique l'ensemble des méthodes (ou encore techniques) permettant d'analyser (on dira plutôt de traiter) des ensembles d'observations où données.

- ① **Population statistique** Ω : ensemble (au sens mathématique de terme) concerné par une étude statistique: ensemble d'objets ou de personnes homogènes étudiées.
- ② Individu ou unité statistique: tous élément de la population ($\omega \in \Omega$).
- ③ **Échantillon**: groupe restreint ou sous-ensemble, issu de la population sur lequel sont effectivement réalisées les observations.
- ④ **Taille de l'échantillon** n : cardinal du sous-ensemble correspondant.

TERMINOLOGIE DE BASE

- ① **Enquête statistique:** opération consiste à observer (ou mesurer, ou questionner, ... etc) l'ensemble des individus d'un échantillon.
- ② **Recensement:** enquête dans laquelle l'échantillon observé est la population tout entière (enquête exhaustive).
- ③ **Sondage:** enquête dans laquelle l'échantillon observé est un sous-ensemble strict (une partie) de la population (enquête non exhaustive)

VARIABLE STATISTIQUE: DÉFINITION

- c'est une caractéristique définie sur la population et observé sur l'échantillon, par exemple: âge, salaire, taille, sexe, ... etc.
- Mathématiquement, il s'agit d'une application définie sur l'échantillon:

$$X : \Omega \rightarrow \begin{cases} \mathbb{R} & \text{si quantitative} \\ \mathcal{E} & \text{si qualitative.} \end{cases}$$

TYPES DES VARIABLE STATISTIQUE

- Si la variable est à valeurs dans \mathbb{R} (celle prenant des valeurs numériques), ou partie de \mathbb{R} , comme \mathbb{N} ou \mathbb{Z} , ou un ensemble de partie de \mathbb{R} , elle est dite quantitative, par exemple: âge, salaire, taille, poids, nombre d'enfant dans une famille, ...etc.
Sinon elle est dite qualitative (celles prenant des valeurs non numériques), par exemple: sexe, couleur des yeux, catégorie socio-professionnelle, ... etc.

NOTATIONS

- Une variable statistique (ou aléatoire) est notée par une lettre majuscule X , Y , et les valeurs qu'elle prend par des lettres minuscules x_1 , x_2 , ..., y_1 , y_2 , ...
- un seul caractère étudié: série numérique à une dimension,
- deux caractères étudiés: série numérique à deux dimensions,
- plus de deux caractères: on doit utiliser les techniques de l'analyse multidimensionnelle.

TYPES DES VARIABLES QUANTITATIVES

- Une variable quantitative est appelée:
 - **discrète** si elle prend un nombre fini (dénombrable) de valeurs souvent entières, exemple: nombre d'enfants d'un ménage,
 - **continue** si elle prend toutes les valeurs d'un intervalle fini ou infini (les données sont en général regroupées en classes), Exemple: taille. En effet, entre une personne mesurant 160cm et 161cm, on peut imaginer une infinité de valeurs (ce qui n'existe pas entre 1 et 2 enfants par exemple). Ce sont la précision des instruments de mesure et les conventions qui font que la taille est traitée comme une variable discrète.

TYPES DES VARIABLES QUALITATIVES

- Une variable qualitative **ordinaire** prend des valeurs qui sont ordonnées, hiérarchisées. On peut classer les modalités les unes par rapport aux autres mais on ne peut pas dire à partir de cet ordre de "combien" est la différence entre deux modalités.
Exemple: Les réponses à un sondage, du type "pas du tout", "un peu", "assez", "beaucoup".
- Sinon la variable est dite **nominale**.

MODALITÉS: DÉFINITION

- Si la variable est qualitative, on appelle modalités les valeurs possibles de cette variable. L'ensemble des modalités est noté $\mathcal{E} = \{e_1, \dots, e_m\}$.
- Par exemple, si la variable est la couleur des yeux d'un individu, l'ensemble des modalités est $\mathcal{E} = \{vert, bleu, brun, gris, noir, \dots\}$.

DONNÉES STATISTIQUES: 1

- Il désigne l'ensemble des individus observés (ceux de l'échantillon), l'ensemble des variables considérées et les observations de ces variables sur ces individus.
- Les données sont en général présentées sous forme de tableaux (individus en lignes et variables en colonnes) et stockées dans un fichier informatique.

DONNÉES STATISTIQUES: 2

	X_1	X_2	\dots	X_j	\dots	X_p
<i>individu1</i> ω_1	x_{11}	x_{12}	\dots	x_{1j}	\dots	x_{1p}
<i>individu2</i> ω_2	x_{21}	x_{22}	\dots	x_{2j}	\dots	x_{2p}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
<i>individu<i>i</i></i> ω_i	x_{i1}	x_{i2}	\dots	x_{ij}	\dots	x_{ip}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
<i>individun</i> ω_n	x_{n1}	x_{n2}	\dots	x_{nj}	\dots	x_{np}

$$x_{ij} = X_j(\omega_i).$$

DONNÉES STATISTIQUES: EXEMPLE

- Voici un tout petit exemple de données:

	sexe	âge	revenu mensuel net
<i>individu 1</i>	1	55	20680
<i>individu 2</i>	1	41	46870
<i>individu 3</i>	1	28	12350
<i>individu 4</i>	2	64	19410
<i>individu 5</i>	2	32	24560

SÉRIE STATISTIQUE: DÉFINITION

- On appelle série statistique la suite des valeurs prises par une variable X sur les unités d'observation. Le nombre d'unités d'observation est noté n (ou N). Les valeurs de la variable X sont notés

$$x_1, \dots, x_i, \dots, x_n.$$

SÉRIE STATISTIQUE: EXEMPLE

- On s'intéresse à la variable "état-civil" notée X () et à la série statistique des valeurs prises par X sur 20 personnes. La codification est:

C: célibataire

M: marié(e)

V: veuf(ve)

D: divorcée

SÉRIE STATISTIQUE: EXEMPLE

- Les modalités de la variable X est $\{C, M, V, D\}$. Considérons la série statistique suivante:

M	M	D	C	C	M	C	C	C	M
C	M	V	M	V	D	C	C	C	M

Ici, $n = 20$, $x_1 = M$, $x_2 = M$, $x_3 = D$, $x_4 = C$, $x_5 = C$, \dots , $x_{20} = M$.

VARIABLE QUANTITATIVE DISCRÈTE: DÉFINITION

- On appelle variable quantitative discrète une variable quantitative ne prenant que des valeurs entières, plus rarement décimale ($\in \mathbb{N}$).
- Le nombre des distinctes (nombre des modalités) d'une telle variable est assez faible (moins de vingtaine).

VARIABLE QUANTITATIVE DISCRÈTE: EXEMPLE

- Le nombre d'enfants dans une population de familles;
- le nombre d'années d'études après le bac dans une population d'étudiants.
- Le nombre de défauts relevés sur une pièce.

VARIABLE QUANTITATIVE DISCRÈTE: REMARQUE

- On a noté l'âge (arrondi à l'année près) des 25 salariés d'une entreprise; la statistique brute est donné ci dessous:
43; 29; 57; 45; 50; 29; 37; 59; 46; 31; 46; 24; 33; 38;
49; 31; 62; 60; 52; 38; 38; 26; 41; 41; 52; 60.
- La variable "âge" est discrète, mais le nombre des modalités est 16
⇒ considérer comme variable continue.

PLAN DE COURS

1 INTRODUCTION

2 TERMINOLOGIE DE BASE

3 VARIABLE QUANTITATIVE DISCRÈTE

- **Tableau statistique:**
- Représentations graphiques usuelles:
- Caractéristique numérique

4 VARIABLE QUANTITATIVE CONTINUE

5 VARIABLE QUALITATIVE

6 EXERCICE



FRÉQUENCES ABSOLUES, RELATIVES, CUMULÉES: 1

- Si on interroge $n = 200$ personnes, les données brutes se présenteront sous la forme d'une suite (appelé série statistique) du type: brun, vert, vert, bleu, ..., gris, vert.
- Cette suite n'est pas lisible. La meilleure manière de représenter ces données est d'utiliser le tableau statistique (les fréquences absolues et relatives).

FRÉQUENCES ABSOLUES, RELATIVES, CUMULÉES: 2

- Dans le cas des variables discrètes X , avec r observations distinctes de la variable X (r : nombre des modalités), on appelle:
 - **Fréquence absolue** n_i ou **effectif**, associée à une valeur x_i de la variable aléatoire X , le nombre d'apparitions de cette variable dans la population ou dans l'échantillon.
 - **Fréquence relative**, associée à la valeur x_i de la variable aléatoire X , le nombre

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

où n_i est la fréquence absolue et n le nombre total (taille de l'échantillon) de données.

FRÉQUENCES ABSOLUES, RELATIVES, CUMULÉES: 3

- **Fréquence cumulée absolue**, associée à une valeur x_i de la variable, le nombre d'individus dont la mesure est inférieure ou égale à x_i .

$$N_i = \sum_{k=1}^i n_k$$

On définit la **fréquence cumulée relative**:

$$F_i = \sum_{k=1}^i f_k.$$

- On notera que $N_r = \sum_{j=1}^r n_j = n_1 + n_2 + \dots + n_r = n$ et

$$F_r = \sum_{j=1}^r f_j = f_1 + f_2 + \dots + f_r = 1.$$

FRÉQUENCES ABSOLUES, RELATIVES, CUMULÉES:

REMARQUES

- Les fréquences relatives et les fréquences cumulées relatives peuvent être utilisées pour comparer deux ou plusieurs populations.
- Dans le cas d'une distribution continue, les données sont en général regroupées en classes. Les fréquences absolues, relatives et cumulées sont définies par rapport aux classes et non par rapport aux valeurs de la variable.

TABLEAU STATISTIQUE: RANGEMENT DES DONNÉES PAR VALEURS NON DÉCROISSANTES

- C'est un tableau dont la première colonne comporte l'ensemble des r observations distinctes de la variable X (r : nombre des modalités); ces observations sont rangées par ordre croissant et non répétées; nous les noterons $\{x_j; j = 1, \dots, r\}$. Dans une second colonne, on dispose, en face de chaque valeurs x_j le nombre de réplcation qui lui sont associées (fréquence absolu ou l'effectif de x_j), Les effectifs n_j sont souvent remplacés par les fréquences relatives $f_j = \frac{n_j}{n} \times 100$ (colonne 3).
- Dans le 4 et 5 colonnes ont trouve les effectifs commulés et les fréquences cumulées.

TABLEAU STATISTIQUE: DÉFINITION

Valeurs de la de la variable	Fréquences (effectifs) absolues	Fréquences relatives relatives	Fréquences cumulées absolues	Fréquences cumulées relatives
x_i	n_i	f_i	N_i	F_i

TABLEAU STATISTIQUE: EXEMPLE 1

- Enquête sur 64 familles. Le caractère à étudié est le nombre d'enfants par famille.

modalité	Effectif	Fréquence	Effectif	Fréquence
x_j	n_j	relatif f_j	cumulé N_j	cumulé F_j
0	16	0.25	16	0.25
1	18	0.28	34	0.53
2	14	0.22	48	0.75
3	12	0.19	60	0.94
4	4	0.06	64	1

Tableau des distribution des fréquences: répartition de nombre d'enfants dans 64 familles.

TABLEAU STATISTIQUE: EXEMPLE 1

- 16 familles ont 0 enfant.
- 18 familles ont 1 enfant.
- 14 familles ont 2 enfants.
- 12 familles ont 3 enfants.
- 4 familles ont 4 enfants.
- $N_1 = n_1 = 16$; $N_4 = n_1 + n_2 + n_3 + n_4 = 60$.

TABLEAU STATISTIQUE: EXEMPLE 2

- Prenons l'exemple: l'âge des 25 salariés d'une entreprise:

$$\{x_1 = 24, x_2 = 26, \dots, x_{16} = 59\}.$$

modalité	Effectif	Fréquence	Effectif	Fréquence
x_j	n_j	relatif f_j	cumulé N_j	cumulé F_j
24	1	$1/200 * 100 = 4.55$	1	4.55
26	1	4.55	2	9.1
\vdots	\vdots	\vdots	\vdots	\vdots
57	1	4.55	21	95.45
59	1	4.55	22	100

Tableau des distribution des fréquences (répartition d'âge des

PLAN DE COURS

1 INTRODUCTION

2 TERMINOLOGIE DE BASE

3 VARIABLE QUANTITATIVE DISCRÈTE

- Tableau statistique:
- Représentations graphiques usuelles:
- Caractéristique numérique

4 VARIABLE QUANTITATIVE CONTINUE

5 VARIABLE QUALITATIVE

6 EXERCICE



DIAGRAMME EN BATÔNS OU EN BARRE

- Il permet de donner une vision d'ensemble des observations réalisées.
- Exemple: nombre d'enfants dans 64 familles.

Diagramme en bâtonnets des effectifs

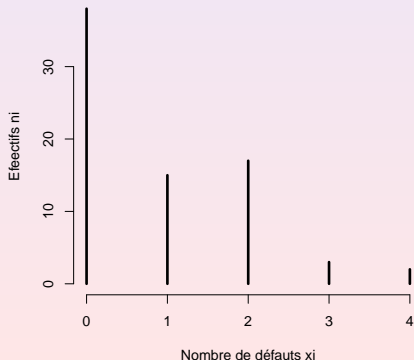
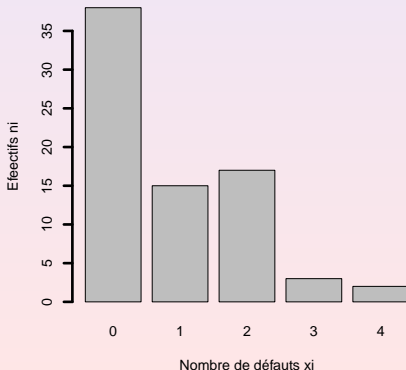


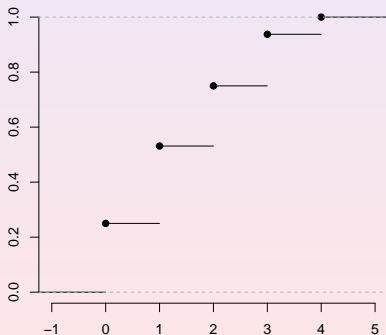
Diagramme en bâtonnets des effectifs



LE DIAGRAMME CUMULATIF:

- Il permet de déterminer le nombre (où la proportion) d'observation inférieures ou égales à une valeurs données de la série.
- Exemple: nombre d'enfants dans 64 familles

Le diagramme cumulatif ou la fonction en escalier



PLAN DE COURS

1 INTRODUCTION

2 TERMINOLOGIE DE BASE

3 VARIABLE QUANTITATIVE DISCRÈTE

- Tableau statistique:
- Représentations graphiques usuelles:
- Caractéristique numérique

4 VARIABLE QUANTITATIVE CONTINUE

5 VARIABLE QUALITATIVE

6 EXERCICE



- Les caractéristiques (ou résumés) numériques introduit ici servent à synthétiser la série statistique étudiée d'un petit nombre de valeurs numériques classées en 2 grandes catégories:
 - les caractéristiques de tendance centrale,
 - les caractéristiques de dispersion.

CARACTÉRISTIQUE DE LA TENDANCE CENTRALE OU DE POSITION

- Leur objectif est de fournir un ordre de grandeur de la série étudiée, c-à-d d'en situer le centre, le milieu. Les principales caractéristiques de tendance centrale sont la moyenne arithmétique, la médiane, le mode et les quantiles.

PARAMÈTRE DE POSITION: MOYENNE

- On appelle la **moyenne** d'une série statistique le nombre:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^r n_i x_i ,$$

où r est le nombre de modalités de la variable X .

- Propriétés**

- La moyenne arithmétique permet de résumer par un seul nombre la série statistique.
- Elle prend en compte toutes les valeurs de la série et elle est facile à calculer.
- Elle est sensible aux valeurs extrêmes, il est parfois nécessaire de supprimer des valeurs extrêmes ou "aberrantes".
- La quantité $e_i = x_i - \bar{X}$ est l'écart de la valeur x_i à la moyenne arithmétique.

MOYENNE: EXEMPLE

- L'exemple "Nombre d'enfants par famille": $\bar{X} = 1.53 \simeq 2$ enfants par famille.
- Soit la série: 1,1,1,39,39,39. $\bar{X} = 20$ (non significatif).

CHANGEMENT D'ORIGINE ET CHANGEMENT D'ÉCHELLE

- On pose pour toutes les données, $y_i = ax_i + b$, a et b étant des constantes; on obtient:

$$\bar{Y} = a\bar{X} + b$$

.

PARAMÈTRE DE POSITION: MODE

- On appelle **mode** d'une série statistique le nombre que l'on rencontre le plus fréquemment (plus observé, plus fréquent), c-à-d celui qui a le plus grand effectif.
- **Propriétés:**
 - Le mode peut ne pas exister, et s'il existe, il peut ne pas être unique.
 - Une série statistique n'ayant qu'un seul mode est appelé unimodale.

MODE: EXEMPLES

- ① La s.s l'âge des salariés: $mode = 38 = x_7$.
- ② La s.s.: nombre d'enfants par famille: $mode = 1 = x_2$.
- ③ La s.s. 3,5,8,10,12,15,16 n'a pas de mode.
- ④ La s.s. 2,3,4,4,4,5,5,7,7,7,9 a deux mode 4 et 7: on l'appelle série bimodale.

PARAMÈTRE DE POSITION: MÉDIANE

- Si les valeurs d'une série statistique sont ordonnées par ordre de grandeurs croissantes (ou décroissantes), la **médiane** est la valeur, observée ou possible; qui se situe au centre de la série ainsi ordonnée, c-à-d, la plus petite valeur m_e telle qu'il y ait au moins autant d'observation inférieures ou égales que d'observation supérieures à elle.

MÉDIANE: CALCUL

- On distingue deux cas:
 - Si la série possède un nombre impair de valeurs ($N = n = 2k + 1$), la médiane sera la $(k + 1)^{ime}$ valeur (la valeur de l'observation qui a le rang $k + 1$).
 - Si la série compte un nombre pair de valeurs et vaut $2k$, la médiane sera l'une ou l'autre des valeurs centrales ou n'importe quelle valeur intermédiaire entre ces deux valeurs, par exemple, la moyenne arithmétique de ces deux valeurs, mais, dans ces conditions, ce n'est pas une valeur observée.

MÉDIANE: PROPRIÉTÉS

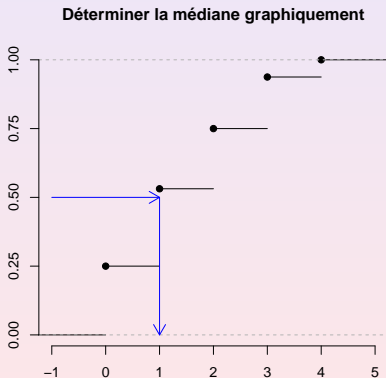
- La médiane est moins sensible que la moyenne à la présence de valeurs extrêmes.
- La médiane est influencée par le nombre des données mais non par leurs valeurs, elle ne peut donc pas être utilisée en théorie de l'estimation.
- Si la variable statistique est discrète, la médiane peut ne pas exister; elle correspond seulement à une valeur possible de cette variable.
- La médiane est le point d'intersection des courbes cumulatives croissante et décroissante.

MÉDIANE: EXEMPLES

- Soit s.s.: 1,2,4,4,4,5,6,7,8,8,9,9,10,11,12. On a $n = 15 = 2 * 7 + 1$ (impair), donc $m_e = M_e = x_{k+1} = x_8 = 7$.
- Soit s.s.: 4;5;8;8;9;11;12;14;17;19. $N = 10$, donc $M_e = \frac{x_5+x_6}{2} = 10$ (\notin s.s.), alors m_e n'existe pas (théoriquement).
- Soit s.s.: 2,2,2,3,3,3,4,4,4,5,5. $N = 11$, donc $M_e = x_6 = 3$, n'existe pas (il y a plusieurs 3).
- L'exemple "nombre d'enfants par famille", $m_e = 1$.

DÉTERMINER LA MÉDIANE GRAPHIQUEMENT

- On peut calculer la médiane graphiquement en utilisant le diagramme cumulatif:



PARAMÈTRE DE POSITION: QUANTILES

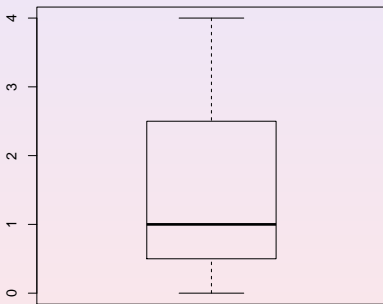
- Les **quantiles** sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales.
 - ① Pour $k = 4$, les quantiles, appelés **quartiles**, sont trois nombres Q_1 , Q_2 , Q_3 tels que:
 - 25% des valeurs prises par la série sont inférieures à Q_1 ,
 - 25% des valeurs prises par la série sont supérieures à Q_3 ,
 - Q_2 est la médiane M_e ,
 - $Q_3 - Q_1$ est l'**intervalle interquartile**, il contient 50% des valeurs de la série.
 - ② Pour $k = 10$, les quantiles sont appelés **déciles**, il y a neuf déciles D_1, D_2, \dots 10% des valeurs de la série sont inférieures à D_1, \dots
 - ③ Pour $k = 100$, les quantiles sont appelés **centiles**, il y a 99 centiles, chacun correspondant à 1% de la population.

QUARTILES: LE DIAGRAMME EN BOÎTE À MOUSTACHES OU BOX-PLOT (TUKEY)

- Il permet de représenter schématiquement les principales caractéristiques d'une distribution en utilisant les quartiles. La partie centrale de la distribution est représentée par une boîte de largeur arbitraire et de longueur la distance interquartile, la médiane est tracée à l'intérieur. La boîte rectangle est complétée par des moustaches correspondant aux valeurs suivantes:
 - valeur supérieure: x_1 .
 - valeur inférieure: x_n .

BOXPLOT: EXEMPLE

- L'exemple: "nombre d'enfant par famille":



CARACTÉRISTIQUES DE DISPERSION

- Ces caractéristiques quantifient les fluctuations des valeurs observées autour de la valeur centrale et permettent d'apprécier l'étalement de la série. Les principales sont: l'écart-type ou son carré appelé variance, le coefficient de variation et l'étendue.

PARAMÈTRE DE DISPERSION: ETENDUE

- L'étendue est la quantité: $E = x_{max} - x_{min}$.
- **Propriétés**
 - L'étendue est facile à calculer.
 - Elle ne tient compte que des valeurs extrêmes de la série; elle ne dépend ni du nombre, n_i des valeurs intermédiaires; elle est très peu utilisée dès que le nombre de données dépasse 10.
 - Elle est utilisée en contrôle industriel où le nombre de pièces prélevées dépasse rarement 4 ou 5; elle donne une idée appréciable de la dispersion.
 - Cependant, dès que cela est possible, on préfère prélever 15 à 20 unités et utiliser l'écart-type pour apprécier la dispersion.

ETENDUE: EXEMPLE

- Soit la s.s.: 0,10,11,12,13,14,14. $E = 14 - 0 = 14$ qui est une différence très grande, alors que la s.s. est concentrée.

PARAMÈTRE DE DISPERSION: INTERVALLE INTERQUARTILE

- $[Q_1 - Q_3]$: on a une première idée de la dispersion de la série statistique.

PARAMÈTRE DE DISPERSION: VARIANCE ET ÉCART-TYPE

- La **variance** d'un échantillon, notée $s^2 = \text{Var}(X)$, est appelée aussi écart quadratique moyen ou variance empirique.
- La racine carrée de la variance est appelée **écart-type**. C'est la moyenne de la somme des carrés des écarts par rapport à la moyenne arithmétique.

$$s^2 = \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{X})^2 = \frac{1}{n} \sum_{i=1}^r n_i x_i^2 - (\bar{X})^2 .$$

- Ecart-type= $s = \sqrt{s^2}$.

VARIANCE: PROPRIÉTÉS

- La moyenne arithmétique \bar{X} et l'écart-type s s'expriment avec la même unité que les valeurs observées x_i .
- L'écart-type s caractérise la dispersion d'une série de valeurs. Plus s est petit, plus les données sont regroupées autour de la moyenne arithmétique \bar{X} et plus la population est homogène; cependant avant de conclure, il faut faire attention à l'ordre de grandeur des données.
- L'écart-type permet de trouver le pourcentage de la population appartenant à un intervalle centré sur la moyenne.
- La variance tient compte de toutes les données, c'est la meilleure caractéristique de dispersion (nombreuses applications en statistique).

CØEFFICIENT DE VARIATION

- S'exprime, sous la forme d'un pourcentage, par l'expression suivante:

$$CV = \frac{s}{\bar{X}} \times 100$$

(ne dépend pas des unités choisies)

- Propriétés:
 - Il permet d'apprécier la représentativité de la moyenne arithmétique \bar{X} par rapport à l'ensemble des données.
 - Il permet d'apprécier l'homogénéité de la distribution, une valeur du coefficient de variation inférieure à 15% traduit une bonne homogénéité de la distribution.
 - Il permet de comparer deux distributions, même si les données ne sont pas exprimées avec la même unité ou si les moyennes arithmétiques des deux séries sont très différentes.

VARIABLE QUANTITATIVE CONTINUE: DÉFINITION

- Une variable quantitative continue peut prendre une infinité de valeurs possibles. Le domaine de la variable est alors \mathbb{R} ou un intervalle de \mathbb{R} .

VARIABLE QUANTITATIVE CONTINUE: REMARQUE

- La taille peut être mesurée en centimètres, voire en millimètres. On peut alors traiter les variables continues comme des variables discrètes.
- Cependant, pour faire des représentations graphiques et construire le tableau statistique, il faut procéder à des regroupements en classes. Le tableau regroupé en classe est souvent appelé distribution groupée.

TABLEAU STATISTIQUE

- Les données sont regroupées en r classes.
- Une **classe** est définie par ses extrémités e_{i-1} , e_i et son effectif n_i .
- **Effectif** d'une classe ou **fréquence absolue**: le nombre n_i de valeurs de la variable X telles que: $e_{i-1} \leq X < e_i$ (l'effectif de la classe i).
- **Amplitude** d'une classe: la quantité $e_i - e_{i-1}$.
- **Centre de la classe**: milieu de la classe $\frac{e_i + e_{i-1}}{2}$.
- N_i l'effectif cumulé de la classe i .
- f_i la fréquence relative de la classe i ,
- F_i la fréquence cumulée relative de la classe i : $F = \sum_{j=1}^i f_j$, avec $F_1 = f_1$. Elle donne la proportion des individus tels que $X < e_i$.

TABLEAU STATISTIQUE:

Classes	Effectifs	Fréquences	Effectifs	Fréquences cumulés
variable	absolues	relatives	cumulés	relatives
$e_{i-1} \leq x_i < e_i$	n_i	f_i	N_i	F_i

NOMBRE DE CLASSES: RÈGLES-CONSEILS

- Le nombre de classes ne doit pas être trop petit (perte d'informations), ni trop grand, le regroupement en classes est alors inutile et de plus, certaines classes pourraient avoir des effectifs trop faibles. En général, le nombre de classes est compris entre 5 et 15 (ou 20); il dépend du nombre n d'observations et de l'étalement des données.
- e_0 est plus petit que la plus petite donnée. e_r est plus grand que la plus grande donnée. Chaque donnée appartient à une seule classe.

NOMBRE DE CLASSES: RÈGLES-CONSEILS

- Les classes sont de largeurs égales. La largeur est de préférence choisie de manière à ce que les milieux soient des nombre entiers (ou faciles avec très peu de décimales). Il peut arriver qu'une ou deux classe aient des fréquences très grandes par rapport aux autres classes. On peut alors utiliser des classes de largeur inégale.
- Dans la mesure du possible, on évitera des classes ouvertes, c'est-à-dire sans borne.

NOMBRE DE CLASSES: RÈGLES-CONSEILS

- Il existent des formules qui nous permettent d'établir le nombre de classes r et l'intervalle de classe (l'amplitude) pour une série statistique de n observations.
 - La règle de Sturge: $r = 1 + (3.3 \log_{10}(n))$.
 - La règle de Yule: $r = 2.5 \sqrt{\sqrt{n}}$.
 - Règle générale: $r = \sqrt{n}$.

Dans les 3 cas, on arrondit à l'entier le plus proche (nombre de classes est un entier). Par commodité, on peut aussi arrondir la valeur obtenue de l'amplitude de classe.

NOMBRE DE CLASSES: RÈGLES-CONSEILS

- Ensuite on doit trouver a : amplitude de la classe: elle est égale à E/k où $E = x_{max} - x_{min}$ est l'étendue de la série des observations (si les classes sont d'égale amplitude).
- Si au contraire, on commence par définir l'amplitude des classes, on ne doit pas choisir cette amplitude trop faible, ni trop grande.
- Les valeurs d'une classe sont assimilées à la valeur centrale (centre de la classe).
- Le regroupement en classes fait perdre aux individus leur caractère propre ainsi que les détails fins des distributions.
- À partir de la plus petite valeur observée, on obtient les bornes de classes en additionnant successivement l'intervalle de classe (l'amplitude).

EXEMPLES

- Dans le cadre d'un contrôle de la croissance des enfants, on a mesuré la taille d'un échantillon de 50 enfants. La toise a donnée les les résultats (en cm) suivants:

152	152	152	153	153	154	154	154	155	155
156	156	156	156	156	157	157	157	158	158
159	159	160	160	160	161	160	160	161	162
162	162	163	164	164	164	164	165	166	167
168	168	168	169	169	170	171	171	171	171

EXEMPLES: LA SUITE

- On construit le tableau statistique:

$[e_{i-1} - e_i[$	n_i	f_i	N_i	F_i
$[151.5; 155.5[$	10	0.20	10	0.20
$[155.5; 159.5[$	12	0.24	22	0.44
$[159.5; 163.5[$	11	0.22	33	0.66
$[163.5; 167.5[$	7	0.14	40	0.80
$[167.5; 171.5[$	10	0.20	50	1.00
	50		1	

REPRÉSENTATION GRAPHIQUES: HISTOGRAMME

- L'histogramme consiste à représenter les effectifs (resp. les fréquences relatives) des classes par des rectangles contigus dont la surface (et non la hauteur) représente l'effectif (resp. la fréquence).

HISTOGRAMME: REMARQUE 1

- L'histogramme est un outil statistique facile à utiliser, donnant rapidement une image du comportement d'un procédé industriel et l'allure globale de la distribution; il montre l'étalement des données et apporte ainsi des renseignements sur la dispersion et sur les valeurs extrêmes; il permet de déceler, éventuellement, des valeurs aberrantes.

HISTOGRAMME: REMARQUE 2

- Pour un histogramme des effectifs, la hauteur du rectangle correspondant à la classe j est donc donnée par:

$$h_j = \frac{n_j}{a_j}$$

- On appelle h_j la densité d'effectif.
- L'aire de l'histogramme est égale à l'effectif total n , puisque l'aire de chaque rectangle est égale à l'effectif de la classe j : $a_j \times h_j = n_j$.

HISTOGRAMME: REMARQUE 3

- Pour un histogramme des fréquences relatives on a:

$$d_j = \frac{f_j}{a_j}.$$

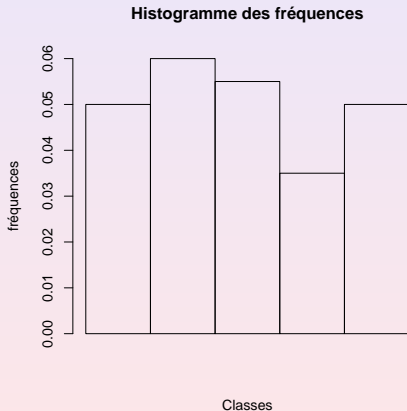
- On appelle d_j la densité de fréquence.
- l'aire de l'histogramme est égale à 1, puisque l'aire de chaque rectangle est égale à la fréquence de la classe j : $a_j \times d_j = f_j$.

HISTOGRAMME: REMARQUE

- Dans le cas de classes de même amplitude certains auteurs et logiciels représentent l'histogramme avec les effectifs (resp. les fréquences relatives) reportés en ordonnée, l'aire de chaque rectangle étant proportionnelle à l'effectif (resp. la fréquence) de la classe.

HISTOGRAMME: EXEMPLE

- Taille des enfants:



REPRÉSENTATION GRAPHIQUE: POLYGONE DE FRÉQUENCES

- Il permet de représenter sous forme de courbe, la distribution des fréquences absolues ou relatives.
- Il est obtenu en joignant, par des segments de droite, les milieux des côtés supérieurs de chaque rectangle de l'histogramme. Pour fermer ce polygone, on ajoute à chaque extrémité une classe de fréquence nulle.

LA COURBE CUMULATIVE

- **Courbe cumulative croissante:** on joint les points ayant pour abscisses la limite supérieure des classes et pour ordonnées les fréquences cumulées croissantes correspondant à la classe considérée (pour le premier point, on porte la valeur 0). Elle donne le nombre d'observations inférieures à une valeur quelconque de la série.
- **Courbe cumulative décroissante:** la construction de cette courbe est analogue à la précédente. Les points ont pour abscisses, les limites inférieures des classes et pour ordonnées, les fréquences cumulées décroissantes (pour le dernier point, la valeur est 0). Elle donne le nombre d'observations supérieures à une valeur quelconque de la série.

LA FONCTION DE RÉPARTITION: DÉFINITION

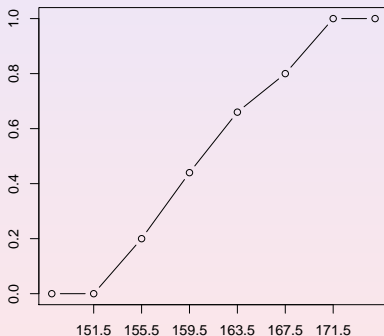
- La fonction de répartition $F(x)$ (le graphe de la courbe cumulative) est une fonction de \mathbb{R} dans $[0, 1]$, qui est définie par:

$$F(X) = \begin{cases} 0 & x < e_1 \\ F_{i-1} + \frac{f_i}{e_i - e_{i-1}}(x - e_{i-1}) & \\ 1 & x \geq e_r \end{cases}$$

COURBE CUMULATIVE: EXEMPLES

- Taille des enfants:

Fonction de répartition d'une distribution groupée



PARAMÈTRES STATISTIQUES

- **Classe modale:** c'est la classe dont l'effectif est relativement le plus élevé et on attribue au mode la valeur centrale de cette classe.
- **Classe médiane:** la classe qui divise l'échantillon en 2 parties de même effectifs. La médiane c'est le centre de cette classe.
- **Les quantiles:** Les quantiles d'une variable continue peuvent être déterminer de façon directe à partir de la courbe cumulative. Cela signifie que, par le calcul, on doit commencer par déterminer la classe dans laquelle se trouve le quantile cherché, puis le déterminer dans cette classe par interpolation linéaire.

PARAMÈTRES STATISTIQUES: LA MOYENNE ET L'ÉCART-TYPE

- La **moyenne**, la **variance** et l'**écart-type** d'une variable continue se déterminent de la même manière que dans le cas discret, dans les formules, on doit prendre pour les x_i les centres de classes au lieu des observations.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n c_i = \frac{1}{n} \sum_{i=1}^r n_i c_i,$$

$$s^2 = \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (c_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n c_i^2 - (\bar{X})^2$$

Ecart-type= $s = \sqrt{s^2}$.

VARIABLE QUALITATIVE NOMINALE

- Une variable qualitative nominale a des valeurs distinctes qui ne peuvent pas être ordonnées. On note r le nombre de valeurs distinctes ou modalités.
- Les valeurs distinctes sont notées $x_1, \dots, x_i, \dots, x_r$.
- On appelle **effectif** d'une modalité ou d'une valeur distincte, le nombre de fois que cette modalité apparaît. On note n_j l'effectif de la modalité x_i .
- La **fréquence** d'une modalité est l'effectif divisé par le nombre d'unités d'observation:

$$f_i = \frac{n_i}{n}, i = 1, \dots, r.$$



VARIABLE QUALITATIVE NOMINALE: EXEMPLE

- Avec la série de l'exemple 17: "état-civil", on obtient le tableau statistique

x_i	n_i	f_i
C	9	0.45
M	7	0.35
D	2	0.10
V	2	0.1
$n = 20$		1

REPRÉSENTATION GRAPHIQUE

- Le tableau statistique d'une variable qualitative nominale peut être représenté par deux types de graphique. Les effectifs sont représentés par un diagramme en barres et les fréquences par un diagramme en secteurs (ou camembert ou piechart en anglais).

DIAGRAMME EN SECTEURS

- Exemple: "État-civil":

Diagramme en secteurs des fréquences

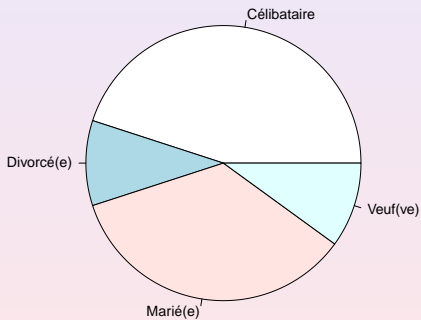
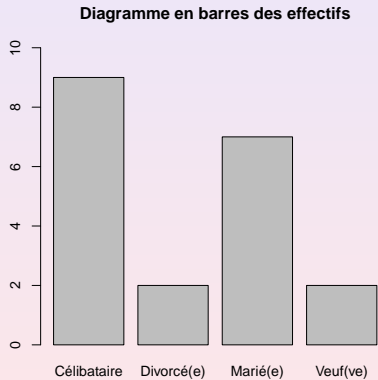


DIAGRAMME EN BARRES

- Exemple: "État-civil":



VARIABLE QUALITATIVE ORDINALE

- Les valeurs distinctes d'une variable ordinale peuvent être ordonnées, ce qu'on écrit

$$x_1 < x_2 < \dots < x_i < \dots < x_r .$$

- La notation $x_1 < x_2$ se lit x_1 précède x_2 .

TABLEAU STATISTIQUE:

- Si la variable est ordinale, on peut calculer les effectifs cumulés:

$$N_j = \sum_{i=1}^j n_i, j = 1, \dots, r$$

EXEMPLE

- On interroge 50 personnes sur leur dernier diplôme obtenu (variable Y). La codification a été faite selon le Tableau:

Dernier diplôme obtenu	x_i
Sans diplôme	Sd
Primaire	P
Secondaire	Se
Supérieur non-universitaire	Su
Universitaire	U

EXEMPLE

- On a obtenu la série:

Sd	Sd	Sd	Sd	P	P	P	P	P	P	P	P	P	P
Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Se	Su	Su
Su	Su	Su	Su	U	U	U	U	U	U	U	U	U	U

EXEMPLE

- Finalement, on obtient le tableau statistique complet:

Tableau statistique complet

x_i	n_i	N_i	f_i	F_i
Sd	4	4	0.08	0.08
P	11	15	0.22	0.30
Se	14	29	0.28	0.58
Su	9	38	0.18	0.76
U	12	50	0.24	1.00
	50		1.00	

REPRÉSENTATION GRAPHIQUE: DIAGRAMME EN SECTEURS

- Les fréquences d'une variable qualitative ordinale sont représentées au moyen d'un diagramme en secteurs: exemple "dernier diplôme obtenu "

Diagramme en secteurs des fréquences

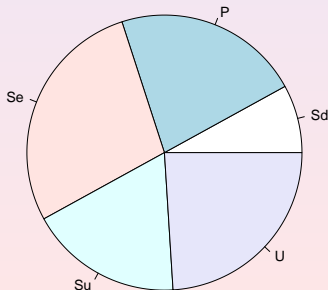


DIAGRAMME EN BARRES DES EFFECTIFS

- Les effectifs d'une variable qualitative ordinale sont représentés au moyen d'un diagramme en barres: exemple "dernier diplôme obtenu "

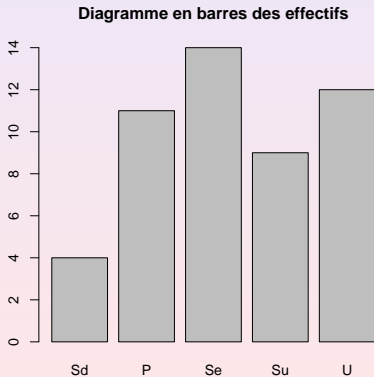
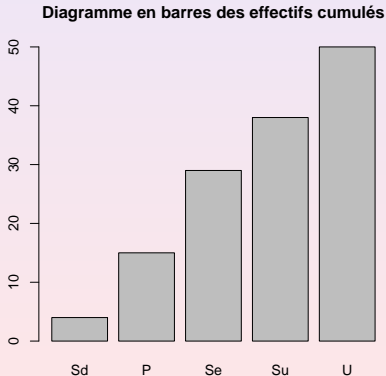


DIAGRAMME EN BARRES DES EFFECTIFS CUMULÉS

- Les effectifs cumulés d'une variable qualitative ordinale sont représentés au moyen d'un diagramme en barres: exemple "dernier diplôme obtenu "



EXERCICE

- Afin d'établir un rapport éventuel entre l'âge et les loisirs, un psychosociologue enquête auprès d'une population de 20 personnes et obtient les informations suivantes:

EXERCICE

Sujet	1	2	3	4	5	6	7	8	9	10
Âge x	12	14	40	35	26	30	30	50	75	50
Loisir y	S	S	C	C	S	T	T	L	L	L

Sujet	11	12	13	14	15	16	17	18	19	20
Âge x	30	45	25	55	28	25	50	40	25	35
Loisir	T	C	C	C	S	L	L	C	T	T

Notations: S: Sport, C: Cinéma, T: Théâtre, L: Lecture.

EXERCICE

1. Que représente la première ligne de ce tableau? Sur combien de sujets l'enquête a-t-elle porté?
2. Combien a-t-on de variables? Quel est le type de chaque variable ?
3. Pour chaque variable, dresser le tableau statistique complet.
4. Représenter les fréquences des variables et à l'aide de diagrammes en bâtons. Quel est le mode.

EXERCICE

5. Tracer le graphe de la fonction de répartition de x . Peut-on tracer celui de y ? En déduire la valeur de la médiane.
6. Quel est le pourcentage des sujets:
 - (A) âgés de moins de 30 ans ?
 - (B) qui ne préfèrent pas la lecture ?
7. Calculer la moyenne, la variance et l'écart-type (empiriques) de x . Commenter brièvement ces résultats. ($\sum x_i = 720$; $\sum x_i^2 = 30200$).

EXERCICE: SOLUTION 1

1. Premier ligne représente les sujets (les individus) de l'échantillon.
La taille de l'échantillon: $n = 20$ individus.
2. On a 2 variables: L'âge: X et le Loisir: Y .
 - X est une variable quantitative discrète (car les valeurs sont entières et le nombre de modalités égal à $12 \leq 15$)).
 - Y est une variable qualitative nominale (les modalités ne sont pas ordonnées).

3.1 Tableau des distribution des fréquences: répartition de l'âge (X)

x_j	n_j	relatif f_j	cumulé N_j	cumulé F_j
12	1	0.05	1	0.05
14	1	0.05	2	0.10
25	3	0.15	5	0.25
26	1	0.05	6	0.30
28	1	0.05	7	0.35
30	3	0.15	10	0.50
35	2	0.10	12	0.60
40	2	0.10	14	0.70
45	1	0.05	15	0.75
50	3	0.15	18	0.90
55	1	0.05	19	0.95
75	1	0.05	20	1.00

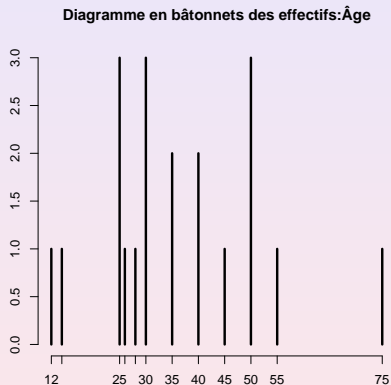
EXERCICE: SOLUTION 3

3.2 Tableau des distribution des fréquences: répartition de 20 individus suivant leurs loisir (Y).

x_i	n_i	f_i
C	6	0.30
L	5	0.25
S	4	0.20
T	5	0.25
$n = 20$		1

EXERCICE: SOLUTION 4

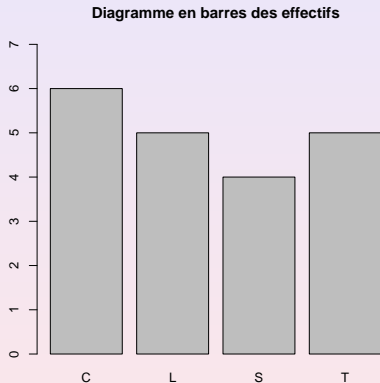
4.1 Diagramme en bâton des effectifs: Âge: X .



Le mode: $Mo = 25, 30, 50$.

EXERCICE: SOLUTION 5

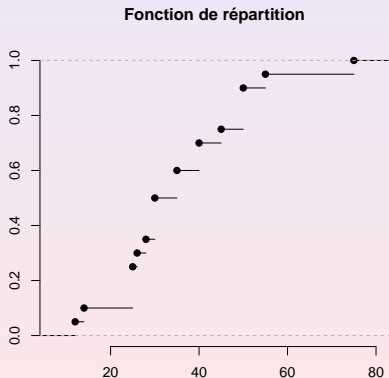
4.1 Diagramme en bâton des effectifs: Loisir *Y*.



Le mode: *Mo* = "cinéma".

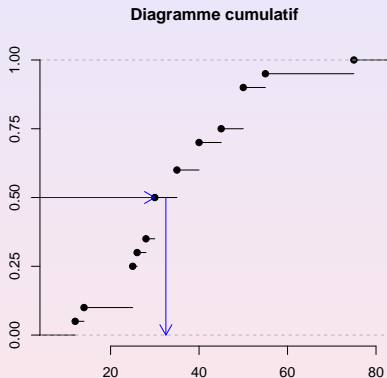
EXERCICE: SOLUTION 5

5.1 Diagramme cumulatif ou le graphe de la fonction de répartition: Âge X.



EXERCICE: SOLUTION 5

5.2 La médiane est égale à 32.5:



5.3 On ne peut tracer le diagramme cumulatif pour une variable qualitative: Loisir *Y*.

EXERCICE: SOLUTION 5

7.1 La moyenne de X "Âge":

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^r n_i x_i = 720/20 = 36.$$

7.2 La variante:

$$s^2 = \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{X})^2 = \frac{1}{n} \sum_{i=1}^r n_i x_i^2 - (\bar{X})^2 = 214$$

Écart-type= $s = 11.57$.