Statistique descriptive

Présenté par :
Pr.Abdelaziz Qaffou

EST-Beni Mellal – Université Sultan Moulay Slimane

DUT-Première année AI-Session d'automne 2020

Plan du cours

Sommaire

Introduction

La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données.
Il ne faut pas confondre la statistique qui est une science qui vient d'être définie et une statistique qui est un ensemble de données chiffrées sur un sujet précis.

Introduction

Le but de la statistique descriptive est de décrire des données (faire parler les chiffres), c'est un résumé d'information sous forme de tableaux statistiques, de représentations graphiques ou tirer des informations sur des données statistiques à partir des indicateurs de position, de dispersion, de forme et de concentration.
Mais le résumé néglige certaines aspects, c'est comme ça né la statistique inférentielle (que d'autres auteurs l'appellent inférence statistique ou statistique mathématique ou statistique paramétrique), pour savoir à quel point l'on peut résumer sans perdre des informations essentielles et quel est le meilleur résumé avec le moins d'erreurs.
Dans la statistique descriptive, on étudie une population qui est un ensemble d'individus (ou unités statistiques) sur lesquels on effectue une analyse statistique. La population est étudiée selon un ou plusieurs caractères.

Vocabulaires statistiques

- Population : ensemble des unités statistiques ou d'individus sur lesquels on effectue une analyse statistique ;
- Unité statistique (ou individu) : élément de la population sur lequel porte l'observation ;
- Echantillon : ensemble d'individus prélevés dans une population déterminée ;
- Caractère (critère ou variable statistique) : permet de décrire et de classer la population.

La collecte des données

La première étape de la statistique descriptive est la collecte des données sur une population qui ce fait par deux méthodes :

- La méthode exhaustive ou le recensement, où chaque individu de la population est étudié ou dénombrer de façon exhaustive selon le ou les caractères étudiés de la population sans exception.
- La méthode d'enquête par sondages ou échantillonnage qui conduit à n'examiner qu'un sous-ensemble (ou une partie) de la population appelé échantillon, elle n'est pas exhaustive, car elle n'interroge pas tous les éléments d'une population.

L'échantillonnage représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.

La deuxième étape de la statistique descriptive est de décrire, d'étudier et de classer les individus de la population, chaque individu est caractérisé par un caractère (ou critère) qu'on appelle variable statistique. Cette variable statistique peut-être qualitative ou quantitative.

Variable qualitative

Les variables qualitatives sont des variables qui ne sont pas représentées par des nombres, on les appelle des caractères qui prennent des modalités, on distingue trois type de variables qualitatives :

- Nominale : celles qui prennent un nom (exemples : couleur des yeux, profession, nationalité,...).
- Ordinale : celles qui peuvent être classées, ordonnées ou hiérarchisées (exemples : mention obtenu, niveau d'étude, niveau se satisfaction,...).
- Dichotomique : celles qui prennent seulement deux modalités (exemples : sexe (masculin ou féminin), vrai ou faux, échecs ou succès, oui ou non,...).

Variable quantitative

Les variables quantitatives sont des variables représentées par des nombres et sur lesquels on peut faire des opérations arithmétique, elles prennent des valeurs. On distingue deux types de variables quantitatives :

- Discrète : celle qui ne prend qu'un nombre fini de valeurs, isolées (exemples : nombre d'enfants par ménage, nombre de visites,...).
- Continue : celle qui peut prendre une infinité de valeurs, ou bien toutes les valeurs d'un intervalle. Généralement, tout ce qui est mesurable est une variable quantitative continue, ces valeurs sont regroupées en classes (exemples : la taille, le poids, la surface, la température,...).

On appelle une **série statistique** la suite des valeurs prises par une variable statistique x sur les unités d'observations. Le nombre d'unités d'observation est noté n. Les individus qu'on note par x_i sont présentés sous forme d'une série statistique qu'on note par $S = \{x_1, x_2, ..., x_n\}$, le nombre total d'individus est n.
Dans une série statistique, les individus ne sont pas regroupés, ne sont pas classés. Si on les regroupe par valeurs ou par modalités, on obtient une **distribution statistique**. Le regroupement peut-être par modalités (pour les variables qualitatives), par valeur (pour les variables quantitatives discrètes) ou par classes (pour les variables quantitatives continues).
Lorsqu'on effectue une distribution par classes de valeurs, on peut choisir des classes d'amplitudes égales ou d'amplitudes inégales.

Variable qualitative nominale

Soit x la variable statistique qui désigne "état civil". La série statistique des valeurs prises par x sur 20 personnes est codifiée ainsi, C : Célébataire, M : Marié(e), D : Divorcé(e) et V : Veuf(ve). Considérons la série statistique suivante :

$$S = \{M, M, D, C, C, M, C, C, C, M, C, M, V, M, V, D, C, C, C, M\}.$$

Ici, $n = 20$, $x_1 = M, x_2 = M, x_3 = D, \dots, x_{19} = C, x_{20} = M$.
La distribution statistique est :

Valeurs (x_i)	Effectifs (n_i)
C	9
M	7
V	2
D	2
Total	20

Variable qualitative ordinale

On interroge 50 personnes sur leur dernier diplôme obtenu (variable statistique x,) la codification a été faite selon le tableau suivant :

Dernier diplôme	x_i
Sans diplôme	Sd
Primaire	P
Secondaire	S
Technicien	T
Universitaire	U

Variable qualitative ordinale

Considérons la série statistique suivante :
S=(Sd,Sd,Sd,Sd,P,P,P,P,P,P,P,P,P,P,P,S,S,S,S,S,S,S,S,S,S,
S,S,S,S,T,T,T,T,T,T,T,T,T,U,U,U,U,U,U,U,U,U,U,U,U)
La distribution statistique est :

Valeurs (x_i)	Effectifs (n_i)
Sd	4
P	11
S	14
T	9
U	12
Total	50

Variable quantitative discrète

Un quartier est composé de 50 ménages, la variable x représente le nombre de personnes par ménage, les valeurs de la variable sont données dans la série statistique suivante :

S=(1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,
3,3,3,3,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,6,6,6,8,8)

La distribution statistique est :

Valeurs (x_i)	Effectifs (n_i)
1	5
2	9
3	15
4	10
5	6
6	3
8	2
Total	50

Variable quantitative continue

On considère les quantités de lait (en litre) vendues par le laitier du quartier durant les 20 derniers jours d'un mois de Ramadan :
La série statistique est la suivante :
S=(10,68-21,23-20,12-24,90-16,47-23,12-10,65-12,07-12,90-17,08-19,00-14,04-17,94-17,09-13,25-16,21-22,10-18,14-21,31-11,25)
On classe les quantités de lait par ordre croissant, puis on regroupe les individus en 5 classes d'amplitudes égales à 3, on obtient la distribution statistique suivante :

Valeurs (x_i)	Effectifs (n_i)
[10 ;13[	5
[13 ;16[	2
[16 ;19[	7
[19 ;22[	3
[22 ;25[	3
Total	20

Sommaire

Le tableau de distribution des effectifs (ou fréquences absolues) et des fréquences (ou fréquences relatives) est un mode de présentation des données. Sa constitution est immédiate dans le cas d'un caractère qualitatif et quantitatif discret, mais nécessite en revanche une transformation des données dans le cas d'un caractère quantitatif continu.
A chaque modalité ou valeur du caractère x, peut correspondre un ou plusieurs individus dans l'échantillon de taille n.

Effectif

On appelle effectif de la modalité x_i, le nombre n_i (le nombre de fois de x_i), parfois, on peut rencontrer le terme de fréquence absolue pour les effectifs.

Fréquence

On appelle fréquence de la modalité ou valeur x_i, le nombre f_i tel que $f_i = \frac{n_i}{n}$, parfois, on peut rencontrer le terme de fréquence relative pour les fréquences.
Le pourcentage est une fréquence exprimée en pour cent, il est égale à $f_i \times 100$.

Effectifs cumulés

Les effectifs cumulés sont exprimés par : $N_i = \sum_{p=1}^{i} n_p$.

Fréquences cumulés

Les fréquences cumulées sont exprimées par : $F_i = \sum_{p=1}^{i} f_p$.

Notation

$n = \sum_{i=1}^{k} n_i$. et $\sum_{i=1}^{k} f_i = 1$.

Variable qualitative nominale

On prend le même exemple précédent de l'état civil, le tableau statistique des effectif, des fréquences et des pourcentages est le suivant :

Valeurs (x_i)	Effectifs (n_i)	Fréquences (f_i)	Pourcentage
C	9	0,45	45
M	7	0,35	35
V	2	0,10	10
D	2	0,10	10
Total	20	1	100

Remarque : On ne peut pas calculer les effectifs cumulés et les fréquences cumulées pour les variables qualitatives nominales.
Interprétation : 45% des individus sont célibataires.

Variable qualitative ordinale

On prend le même exemple précédent du dernier diplôme obtenu, le tableau statistique des effectif, des fréquences, des effectifs cumulés et des fréquences cumulées est le suivant, avec les notations suivantes :

- Valeurs : x_i
- Effectifs : n_i
- Effectifs cumulés croissants : N_i^+ (au plus-moins de)
- Effectifs cumulés décroissants : N_i^- (au moins-plus de)
- Fréquences : f_i
- Fréquences cumulées croissantes : F_i^+ (au plus-moins de)
- Fréquences cumulées croissantes : F_i^- (au moins-plus de)

x_i	n_i	N_i^+	N_i^-	f_i	F_i^+	F_i^-
Sd	4	4	50	4/50=0,08	0,08	1
P	11	4+11=15	50-4=46	0,22	0,30	0,92
S	14	15+14=29	46-11=35	0,28	0,58	0,70
T	9	29+9=38	35-14=21	0,18	0,76	0,42
U	12	38+12=50	21-9=12	0,24	1	0,24
Total	50	—	—	1	—	—

Interprétation

- Il y a **15** individus qui ont **au plus** un diplôme des études primaires, ou bien **15** personnes qui ont **moins de** un diplôme secondaire.
- Il y a **35** personnes qui ont **au moins** un diplôme secondaire ou bien **35** personnes **plus de** ont un diplôme primaire.
- 58% ont **au plus** un diplôme secondaire, ou bien 58% ont **moins d**'un diplôme primaire.
- 70% ont **au moins** un diplôme secondaire, ou bien **plus de** 70% ont un diplôme technicien.

Variable quantitative discrète

On prend l'exemple de nombre de personne par ménage dans un quartier, le tableau statistique des effectif, des fréquences, des effectifs cumulés et des fréquences cumulées est le suivant :

x_i	n_i	N_i^+	N_i^-	f_i	F_i^+	F_i^-
1	5	5	50	0,10	0,10	1
2	9	14	45	0,18	0,28	0,90
3	15	29	36	0,30	0,58	0,72
4	10	39	21	0,20	0,78	0,42
5	6	45	11	0,12	0,90	0,22
6	3	48	5	0,06	0,96	0,10
8	2	50	2	0,04	1	0,04
Total	50	—	—	1	—	—

Variable quantitative continue

On prend l'exemple de quantités de lait (en litre) vendues par le laitier du quartier, le tableau statistique des effectif, des fréquences, des effectifs cumulés et des fréquences cumulées est le suivant :

x_i	n_i	N_i^+	N_i^-	f_i	F_i^+	F_i^-
[10 ;13[	5	5	20	0,25	0,25	1
[13 ;16[	2	7	15	0,10	0,35	0,75
[16 ;19[	7	14	13	0,35	0,70	0,65
[19 ;22[	3	17	6	0,15	0,85	0,30
[22 ;25[	3	20	3	0,15	1	0,15
Total	20	—	—	1	—	—

Sommaire

Les représentations graphiques ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies. Les graphiques apparaissent comme plus parlants que les tableaux, les graphiques sont spécifiques à un type de variables ou de caractères (qualitatif, quantitatif discret ou quantitatif continue).

Caractère qualitatif

Il y a trois types de représentations graphiques pour un caractère qualitatif soit nominal ou ordinal.

- **Diagramme en bâtons** : à chaque modalité x_i, on associe un bâton de longueur h_i proportionnelle à l'effectif n_i ou à la fréquence f_i.
- **Diagramme en barres** : même principe que le diagramme en bâtons.
- **Diagramme circulare** : à chaque modalités x_i, on associe un angle α_i avec $\alpha_i = f_i \times 360$.

Variable quantitative discrète

On a deux représentations, une correspondante à des effectifs (ou à des fréquences), l'autre aux effectifs cumulés (ou aux fréquences cumulées). Pour les fréquences (ou les effectifs) cumulées, ne sont que la fonction de répartition $F(x)$ donnée par :

$$F(x) = \begin{cases} 0, & \text{si } x < x_1 \\ f_1, & \text{si } x_1 \leq x < x_2 \\ f_1 + f_2, & \text{si } x_2 \leq x < x_3 \\ . & \\ . & \\ . & \\ \sum_{i=1}^{n-1} f_i, & \text{si } x_{n-1} \leq x < x_n \\ 1, & \text{si } x \geq x_n \end{cases}$$

Les courbes cumulatives dans le cas discret, sont en escalier.

Variable quantitative continue

- Les fréquences (ou les effectifs) sont représentées par les histogrammes, il y a deux cas :
  - **Cas d'amplitudes égales** : sur l'axe des abscisses, on met l'amplitude des classes, sur l'axe des ordonnées, on met les fréquences (ou les effectifs).
  - **Cas d'amplitudes inégales** : on effectue une correction sur les fréquences (ou les effectifs), $h_i = f_i \frac{a_0}{a_i}$ (ou $h_i^{'} = n_i \frac{a_0}{a_i}$) est la hauteur du rectangle de l'histogramme, avec a_0 est la plus petite amplitude commune (PGCD).
- Les fréquences (ou les effectifs) cumulées (les courbes cumulatives sont continues), on relie les points extrêmes de chaque classe par un segment de droite.

Caractère nominal

On prend le caractère état civil de l'exemple précédent :

Valeurs (x_i)	Effectifs (n_i)	Fréquences (f_i)	Pourcentage
C	9	0,45	45
M	7	0,35	35
V	2	0,10	10
D	2	0,10	10
Total	20	1	100

Figure 1

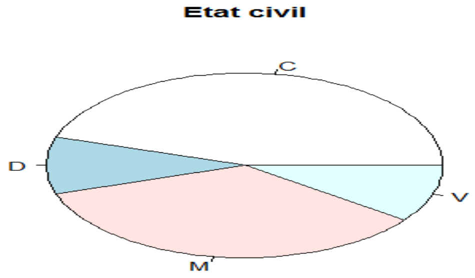

Figure 1 – Diagramme en secteur.

Caractère ordinal

On considère l'exemple du dernier diplôme obtenu :

Valeurs (x_i)	Effectifs (n_i)
Sd	4
P	11
S	14
T	9
U	12
Total	50

Figure 2

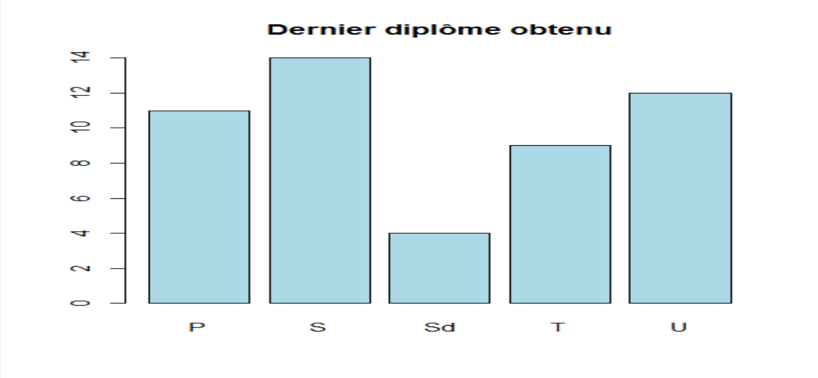

Figure 2 – Diagramme en batôns.

Variable quantitative discrète

On considère l'exemple du dernier diplôme obtenu :

Valeurs (x_i)	Effectifs (n_i)
1	5
2	9
3	15
4	10
5	6
6	3
8	2
Total	50

Figure 3

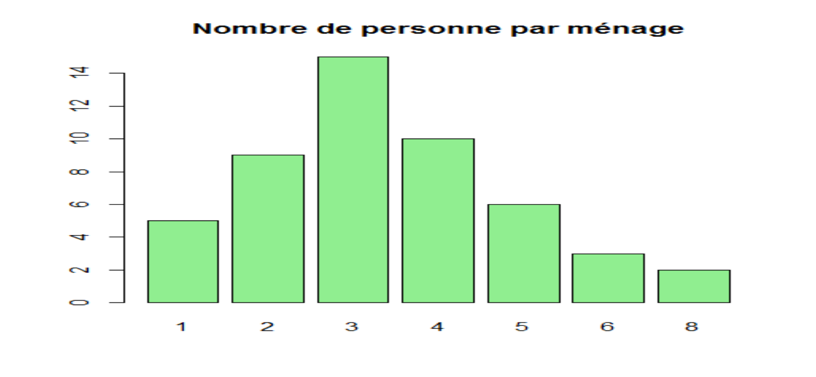

Figure 3 – Diagramme en batôns.

Pour la fonction de répartition ou bien les fréquences cumulées croissantes, on a :

Figure 4

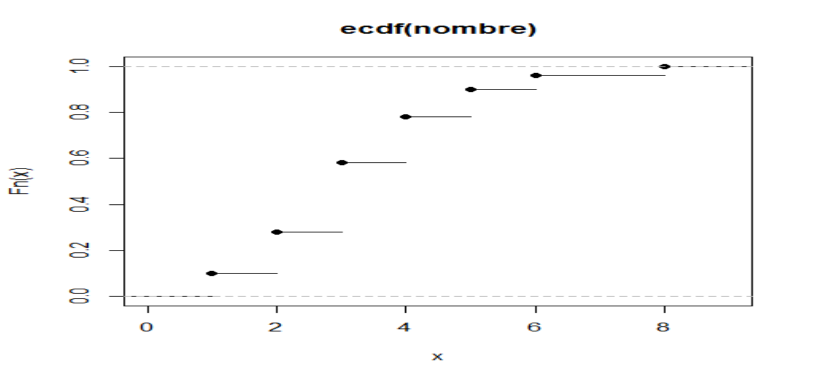

Figure 4 – Fonction de répartition d'une variable discrète.

Variable quantitative continue

On considère l'exemple de quantité de lait vendue :

Valeurs (x_i)	Effectifs (n_i)
[10 ;13[	5
[13 ;16[	2
[16 ;19[	7
[19 ;22[	3
[22 ;25[	3
Total	20

Figure 5

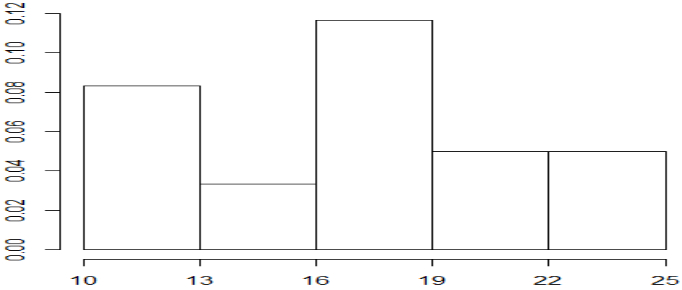

Figure 5 – Histogramme des effectifs.

La courbe des fréquences cumulées croissantes est données par :

Figure 6

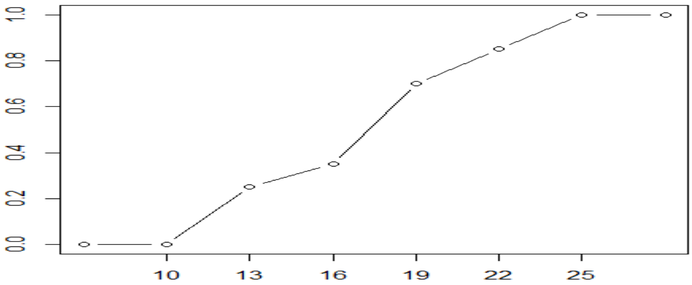

Figure 6 – Fonction de répartition d'une distribution groupée.

Sommaire

Le mode (les cas : normal, ordinal et discret)

Le mode est la valeur la plus fréquente, est la valeur distincte correspodante au plus grand effectif. Le mode peut être calculé pour tous les types de variables, qualitatives et quantitatives. Le mode n'est pas nécessairement unique. Quand une variable continue est découpée en classes, on peut définir une classe modale (classe correspondante à l'effectif le plus élevé).

- Si on prend la variable qualitative nominale "état civil", on a le mode est la modalité C.
- Si on prend la variable qualitative ordinale "dernier diplôme obtenu", le mode est la modalité T.
- Si on prend la variable quantitative discrète "nombre de personne par ménage", le mode est 3.

Le mode (le cas continu)

Si on prend la variable quantitative continue "quantité de lait vendue", la variable est découpée en 5 classes :

Valeurs (x_i)	Effectifs (n_i)
[10 ;13[	5
[13 ;16[	2
[16 ;19[	7
[19 ;22[	3
[22 ;25[	3
Total	20

La classe modale est celle qui contient le plus grand effectif, c'est la classe [16 ;19[, puis on utilise la formule suivante pour déterminer le mode :

$$Mo = x_i + a_i \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$$

Le mode (le cas continu)

avec
$[x_i; x_{i+1}[$: la classe modale,
a_i : l'amplitude de la classe modale,
n_i : l'effectif de la classe modale,
n_{i-1} : l'effectif de la classe qui précède la classe modale,
n_{i+1} : l'effectif de la classe qui suit la classe modale.
Dans notre exemple, on a $[x_i; x_{i+1}[= [16; 19[$ est la classe modale avec $n_i = 7$, $n_{i-1} = 2$, et $n_{i+1} = 3$, $a_i = 3$ et $x_i = 16$.
Donc $Mo = 16 + 3\frac{7-2}{(7-2)+(7-3)}$.
D'où $Mo = 17,66$.

Remarque :
Le mode n'est pas affecté par les valeurs extrêmes, il peut qu'une série ne contient pas de mode, ou bien peut contenir plusieurs modes.

La moyenne (le cas discret)

La moyenne ne peut être définie que sur une variable quantitative, la moyenne est la somme des valeurs observées divisée par leur nombre, elle est notée $\bar{x}$. il existe plusieurs types de moyenne ; la moyenne arithmétique, géométrique, harmonique, quadratique,...Dans ce cours on va utilser uniquement la moyenne arithmétique, soit simple, soit pondérée. Pour la moyenne arithmétique simple, elle est calculée quand on a une série statistique par la façon suivante :

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

Considérons le nombre d'enfant de 8 familles : $S = \{0, 0, 1, 1, 1, 2, 3, 4\}$.

Donc $\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{1}{8}(0+0+1+1+2+3+4) = 1,5$.

La moyenne (le cas discret)

Pour la moyenne pondérée, elle est calculée quand on a une distribution statistique, elle est définie par :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, \text{ avec k : nombre de modalités}$$

Considérons la série statistique S, sa distribution statistique est :

Valeurs (x_i)	Effectifs (n_i)
0	2
1	3
2	1
3	1
4	1
Total	8

Donc $\bar{x} = \frac{1}{8}\sum_{i=1}^{5} n_i x_i = \frac{1}{8}(2\times 0 + 3\times 1 + 1\times 2 + 1\times 3 + 1\times 4) = 1,5.$

La moyenne (le cas continu)

Pour une variable continue découpée en classes, on définit la moyenne ainsi : $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$ avec k : nombre de classes et c_i : le centre de la classe i.

Prenant l'exemple de la variable continue "quantité de lait vendue", on a :

Valeurs (x_i)	Effectifs (n_i)	Centre de classe c_i	$n_i c_i$
[10 ;13[	5	11,5	57,5
[13 ;16[	2	14,5	29
[16 ;19[	7	17,5	122,5
[19 ;22[	3	20,5	61,5
[22 ;25[	3	23,5	70,5
Total	20		341

Donc $\bar{x} = \frac{1}{20}\sum_{i=1}^{5} n_i c_i = \frac{341}{20} = 17,05$.

Remarque : la moyenne est affectée par les valeurs extrêmes.

La médiane (le cas discret)

La première des choses, on classe les unités statistiques de la population en ordre croissant, après on voit si la taille n de la population est paire ou impaire.
Si n est paire, c'est à dire : $n = 2k$, dans ce cas, la médiane qu'on note par Me est donnée par :

$$Me = \frac{x_{(k)} + x_{(k+1)}}{2}$$

La médiane (le cas discret)

Prenons la série statistique suivante $S = \{0, 1, 1, 2, 4, 2, 3, 4, 5, 2\}$
On a n est pair, $n = 10 = 2 \times 5$, donc $k = 5$.
On classe la série en ordre croissant :

Série classée en ordre croissant	Rang
0	1
1	2
1	3
2	4
2	5
2	6
3	7
4	8
4	9
5	10

Donc la médiane est $Me = \frac{x_{(5)} + x_{(6)}}{2} = \frac{2+2}{2} = 2$.

La médiane (le cas discret)

Si n est impaire, la médiane est égale à la valeur de la série classée qui a $\frac{n+1}{2}$ comme rang, ou bien $Me = x_{(\frac{n+1}{2})}$.

Si on prend la série suivante $S = \{0, 1, 3, 4, 2, 3, 4, 5, 2\}$. On a n est impair, $n = 9 = 2 \times 4 + 1$, donc $k = 4$. On classe la série en ordre croissant

Série classée en ordre croissant	Rang
0	1
1	2
2	3
2	4
2	5
3	6
4	7
4	8
5	9

Donc la médiane est $Me = x_{(5)} = 2$.

La médiane (le cas continu)

On cherche les fréquences cumulées croissantes, après, on cherche où se situe la valeur $0,5$ de la fréquence cumulée croissante. La classe correspondante à $0,5$ s'appelle la classe médiane $[x_i; x_{i+1}[$ et la médiane est donnée par

$$Me = x_i + a_i \frac{0,5 - F(x_i)}{F(x_{i+1}) - F(x_i)}.$$

La médiane (le cas continu)

On prend la variable continue "quantité de lait vendue", on a le tableau statistique suivant :

Valeurs (x_i)	Effectifs (n_i)	Fréquences f_i	F_i^+
[10 ;13[	5	0,25	0,25
[13 ;16[	2	0,10	0,35
[16 ;19[	7	0,35	0,70
[19 ;22[	3	0,15	0,85
[22 ;25[	3	0,15	1
Total	20	1	—

On a la valeur $0,5$ se trouve entre $0,35$ et $0,70$ dans la colonne des fréquences cumulées croissantes, donc la classe médiane est $[x_i; x_{i+1}[= [16; 19[$, $F(x_{i)} = 0,35$ et $F(x_{i+1}) = 0,70$, $a_i = 3$ d'où :

$$Me = 16 + 3\frac{0,5 - 0,35}{0,70 - 0,35} = 17,28.$$

Les quartiles

Les quartiles sont les valeurs d'une série ou d'une distribution statistique rangés par ordre croissant qui partagent l'effectif total en quatre parties égales, il existe trois quartiles, le premier quartile Q_1, le deuxième Q_2, qui égale à la médiane et le troisième quartile Q_3.

Les quartiles (le cas discret)

On classe la série en ordre croissant, selon la taille n de la population :

- si $n = 4p + 0$ alors $\begin{cases} Q_1 = \frac{x_{(p)} + x_{(p+1)}}{2} \\ Q_2 = \frac{x_{(2p)} + x_{(2p+1)}}{2} \\ Q_3 = \frac{x_{(3p)} + x_{(3p+1)}}{2}. \end{cases}$
- si $n = 4p + 1$ alors $\begin{cases} Q_1 = \frac{x_{(p)} + x_{(p+1)}}{2} \\ Q_2 = x_{(2p+1)} \\ Q_3 = \frac{x_{(3p+1)} + x_{(3p+2)}}{2}. \end{cases}$
- si $n = 4p + 2$ alors $\begin{cases} Q_1 = x_{(p+1)} \\ Q_2 = \frac{x_{(2p+1)} + x_{(2p+2)}}{2} \\ Q_3 = x_{(3p+2)}. \end{cases}$
- si $n = 4p + 3$ alors $\begin{cases} Q_1 = x_{(p+1)} \\ Q_2 = x_{(2p+2)} \\ Q_3 = x_{(3p+3)}. \end{cases}$

Exemple

Si on prend la varaible discrète "nombre d'enfants par ménage", avec la série statistique est :

$$S = \{4,0,1,1,2,2,2,3,3,4,2,3,4,5,2,1,3,3,4,5\}$$

On a $n = 20 = 4 \times 5 = 4p$ donc $p = 5$
En classant cette série, on obtient le tableau suivant :

Exemple

Nombre d'enfants	Rang
0	1
1	2
1	3
1	4
2 (p ème valeur)	5 (rang p)
2 (p+1 ème valeur)	6 (rang p+1)
2	7
2	8
2	9
3 (2p ème valeur)	10 (rang 2p)

Exemple

Nombre d'enfants	Rang
3 (2p+1 ème valeur)	11 (rang 2p+1)
3	12
3	13
3	14
4 (3p ème valeur)	15 (rang 3p)
4 (3p+1 ème valeur)	16 (rang 3p+1)
4	17
4	18
5	19
5	20

Par conséquent, on a : $\begin{cases} Q_1 = \frac{x_{(p)}+x_{(p+1)}}{2} = \frac{2+2}{2} = 2 \\ Q_2 = \frac{x_{(2p)}+x_{(2p+1)}}{2} = \frac{3+3}{2} = 3 \\ Q_3 = \frac{x_{(3p)}+x_{(3p+1)}}{2} = \frac{4+4}{2} = 4. \end{cases}$

Les quartiles (le cas continu)

Pour le calcul du quartile Q_p d'ordre p, on calcule les fréquences cumulées croissantes (FCC), après on cherche la valeur de p où se trouve parmi les valeurs des FCC. Le quartile d'ordre p appartient à la classe correspondante à p.
D'une manière générale :

$$\text{Si } Q_p \in]x_i, x_{i+1}] \text{ alors } Q_p = x_i + a_i \frac{p - F(x_i)}{F(x_{i+1} - F(x_i))}$$

avec a_i : l'amplitude de la classe $]x_i, x_{i+1}]$.

- Pour le premier quartile (le quartile d'ordre 0,25), noté $Q_{0,25}(Q_1$ ou $Q_{\frac{1}{4}})$ car il représente 25% (un quart $(\frac{1}{4})$) de la population, on a $p = 0,25$ et $Q_{0,25} = Q_1 = x_i + a_i \frac{0,25 - F(x_i)}{F(x_{i+1} - F(x_i))}$
- Pour le deuxième quartile (le quartile d'ordre 0,5), noté $Q_{0,5}(Q_2$ ou $Q_{\frac{1}{2}} = Me)$ car il divise la population en deux effectif égaux, on a $p = 0,5$ et $Q_{0,5} = Q_2 = x_i + a_i \frac{0,5 - F(x_i)}{F(x_{i+1} - F(x_i))}$

Les quartiles (le cas continu)

Pour le troisième quartile (le quartile d'ordre 0,75), noté $Q_{0,75}$(Q_3 ou $Q_{\frac{3}{4}}$) car il représente 75% ($\frac{3}{4}$) de la population, on a $p = 0,75$ et $Q_{0,75} = Q_3 = x_i + a_i \frac{0,75 - F(x_i)}{F(x_{i+1} - F(x_i))}$

Exemple

En prenant la variable continue "quantité de lait vendu," on a :

Valeurs (x_i)	Effectifs (n_i)	Fréquences f_i	F_i^+
[10 ;13[	5	0,25	0,25
[13 ;16[	2	0,10	0,35
[16 ;19[	7	0,35	0,70
[19 ;22[	3	0,15	0,85
[22 ;25[	3	0,15	1
Total	20	1	—

Pour le premier quartile Q_1, on a 0,25 se trouve par coincidence dans la première classe [10;13[, donc $Q_1 \in [10;13[= [x_i, x_{i+1}[$, alors $x_i = 10, a_i = 13 - 10 = 3, F(x_i) = F(10) = 0$ et $F(x_{i+1}) = F(13) = 0,25$, on remplace dans la formule, on obtient :

$$Q_1 = 10 + 3\frac{0,25 - 0}{0,25 - 0} = 13.$$

Exemple

- Pour le deuxième quartile $Q_2 = Me$, on a 0,5 se trouve entre 0,35 et 0,70, donc la classe correspondante à 0,5 est $[16;19[$, alors $Q_2 \in [16;19[= [x_i, x_{i+1}[$, d'où $x_i = 16, a_i = 19 - 16 = 3, F(x_i) = F(16) = 0,35$ et $F(x_{i+1}) = F(19) = 0,70$, on remplace dans la formule, on obtient :

$$Q_2 = 16 + 3\frac{0,5 - 0,35}{0,70 - 0,35} = 17,28.$$

- Pour le troisième quartile Q_3, on a 0,75 se trouve entre 0,70 et 0,85 donc la classe correspondante est $[19;22[$, donc $Q_3 \in [19;22[= [x_i, x_{i+1}[$, alors $x_i = 19, a_i = 22 - 19 = 3, F(x_i) = F(19) = 0,70$ et $F(x_{i+1}) = F(22) = 0,85$, on remplace dans la formule, on obtient :

$$Q_3 = 19 + 3\frac{0,75 - 0,70}{0,85 - 0,70} = 19.$$

L'étendue

L'étendue est la différence entre la plus grande et la plus petite valeur de la série statistique :

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

il mesure la dispersion de la population, sensible aux valeurs extrêmes.

L'intervalle interquartileL'étendue

L'intervalle interquartile (IIQ) est la différence entre le troisième et le premier quartile :

$$IIQ = Q_3 - Q_1$$

il mesure la dispersion de la population en se concentrant autour de la médiane, il élimine les valeurs extrêmes.

La variance

la variance est la somme des carrées des écarts à la moyenne divisée par le nombre d'observations :

- Pour une série statistique : $V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ variance simple.
- Pour une distribution statistique :

$$V(x) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2$$ variance pondérée, k : nombre de modalités.

- Pour une variable continue découpée en classes de centres c_i, on a :

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

Plus la variance est élevée, plus la dispersion autour de la moyenne est élevée. Elle sert à calculer l'écart-type et le coefficient de variation.

L'écart-type

L'écart-type est la racine carrée de la variance :

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

Si l'écart-type est faible alors les valeurs sont assez centrées.
Si l'écart-type est élevé alors les valeurs sont assez dispersées.

Le coefficient de variation

Le coefficient de variation est le rapport entre l'écart-type et la moyenne :

$$CV_x = \frac{\sigma_x}{\bar{x}} \times 100$$

il est exprimé toujours en pourcentage, il indique la dispersion par rapport à la moyenne, peut être utilisé pour comparer plusieurs séries de données exprimées en différentes unités.

L'écart absolu moyen

L'écart absolu moyen est la somme des valeurs absolues des écarts à la moyenne divisée par le nombre d'observation :

$$e_{moy} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|.$$

Lorsque les observations sont groupées par classes, on a :

$$e_{moy} = \frac{1}{n}\sum_{i=1}^{n} n_i |c_i - \bar{x}|.$$

Les moments

On appelle moment d'origine d'ordre $r \in \mathbb{N}$, le paramètre : $m_r \frac{1}{n}\sum_{i=1}^{n} x_i^r$.

Lorsque les observations sont groupées par classes, on a : $m_r \frac{1}{n}\sum_{i=1}^{n} n_i c_i^r$.

On appelle moment centré d'ordre $r \in \mathbb{N}$, le paramètre : $m_r \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^r$.

Lorsque les observations sont groupées par classes, on a :
$m_r \frac{1}{n}\sum_{i=1}^{n} n_i(c_i - \bar{x})^r$.

Les moments généralisent la plupart des paramètres, on a en particulier :
pour $r = 1$: $m_1 = \bar{x}$
pour $r = 2$: $m_2 = V(x)$
pour $r = 3$ et 4, on va les utiliser pour mesurer la symétrie et l'aplatissement.

Merci pour votre attention