السلام عليكم ورحمة الله وبركاته

وزارة التعليم العالي و البحث العلمي

Université Badji Mokhtar-Annaba
Faculté des Sciences Médicales,
Département de Médecine

جـــامعة بـــاجي مختـــــار – عنابـــــــة
كـليـــــــــــة الـــعـلـــوم الطبية
قســـــــــــم الطب

Statistique descriptive

Dr Bouhadjar

2020/2021

1/ Historique Et Définition

Nous distinguerons trois phases importantes dans l'évolution de la statistique :

- De l'antiquité et jusqu'à la fin du 19ème siècle, la statistique est restée principalement un ensemble de techniques de dénombrement.

- De la fin du 19ème siècle aux années 1960 s'est construit la statistique mathématique surtout grâce à l'école anglaise (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman, ...).

- Depuis les années 1960, et avec le développement des outils informatique et graphique, la statistique a connu un développement considérable.

Actuellement, beaucoup de domaines utilisent les méthodes statistiques (médecine, agronomie, sociologie, industrie etc....).

Définition

La Statistique, c'est l'étude des variations observables. C'est une méthode qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à les analyser et à les interpréter.

Remarque : Il ne faut pas confondre la statistique qui est la science qui vient d'être définie et une statistique qui est un ensemble de données chiffrées sur un sujet précis.

2/ Séries Statistiques A Une Variable

a/ Terminologie

Population : ensemble concerné par une étude statistique (noté Ω).

Individu : (unité statistique) : tout élément de la population (noté $\omega \in \Omega$).

Echantillon : sous ensemble de la population sur lequel sont réalisées les observations.

Enquête : Opération consistant à observer (mesurer, questionner, ...) l'ensemble des individus d'un échantillon.

Recensement : enquête dans laquelle l'échantillon observé est la population toute entière (enquête exhaustive).

Sondage : enquête dans laquelle l'échantillon observé est un sous ensemble strict de la population (enquête non exhaustive).

Caractère : c'est une caractéristique définie sur la population et observée sur l'échantillon. Les différents types de variables statistiques sont

- Qualitatif nominal (sexe, profession, situation familiale,...)
- Qualitatif ordinal (grade militaire, grade dans l'enseignement supérieur, ...)
- Quantitatif discret (nombre d'enfants , nombre de chambre dans un appartement, ...)
- Quantitatif continu (taille, âge, vitesse, poids, taux, ...)

Modalités : les déférentes valeurs prises par chaque caractère.

b/ Comment Organiser Les Données

On regroupe toutes les données de la série statistique dans un tableau indiquant la répartition des individus selon le caractère étudié. Le regroupement s'effectue par **classes** :

• Si le caractère est qualitatif ou discontinu, une classe contient tous les individus ayant la même modalité ou la même valeur du caractère.

• Si le caractère est continu, une classe est un intervalle.

◊ Pour construire ces intervalles, on respecte les règles suivantes :

1. Le nombre de classes est compris entre 5 et 20 (de préférence entre 6 et 12)
2. Chaque fois que cela est possible, les amplitudes des classes sont égales.
3. Chaque classe (sauf la dernière) contient sa borne inférieure mais pas sa borne supérieure.

◊ Dans les calculs, une classe sera représentée par son centre, qui est le milieu de l'intervalle.

◊ Une fois la classe constituée, on considère les individus répartis uniformément entre les deux bornes (ce qui entraîne une perte d'informations par rapport aux données brutes).

Que faut-il indiquer pour chaque classe ?

1. **L'effectif :** nombre d'individus de la classe : on le note n_i(i est l'indice de la classe).

2. **La fréquence :** proportion d'individus de la population ou de l'échantillon appartenant à la classe : on la note f_i.

f_i et n_i sont liés par : $f_i = \frac{n_i}{N}$ où N est le nombre total d'individus dans la population.

Remarque : On peut remplacer f_i par $f_i \times 100$ qui représente alors un pourcentage. On a toujours :

$\sum_{i=1}^{k} n_i = N \quad 0 \leq f_i \leq 1 \qquad \sum_{i=1}^{k} f_i = 1$

où k représente le nombre de classes

3. **L'effectif** (ou la fréquence) **cumulé** (e) : effectif (ou fréquence) de la classe augmenté (e) de ceux (ou celles) des classes précédentes(lorsque la variable statistique est quantitative). La fréquence cumulée est une fonction F de la borne supérieure de la classe (dans le cas d'une variable statistique continue).

Tableau Statistique
Variable discrète (Caractère quantitatif discret)

Effectif cumulé croissant à la modalité x_i est

$$n_{croissant} = n_1 + n_2 + \cdots + n_i$$

Effectif cumulé décroissant à la modalité x_i est

$$n_{décroissant} = n_k + n_{k-1} + \cdots + n_i$$

Remarque

- X; le caractère étudié,
- n; la taille de l'échantillon,
- k; le nombre de modalités du caractère X,
- $x_1, \ldots, x_k$ les modalités d'effectifs respectifs $n_1, \ldots, n_k$.

Tableau Statistique

Variable discrète (Caractère quantitatif discret)

Exemple 1. On étudie le nombre d'enfants dans une famille dans une cité habitée par 100 familles. On a obtenu les résultats suivants :

01	02	03	04	……	98	99	100
2	0	4	2	…….	6	3	1

X Nombre d'enfants	Effectif n_i	Fréquence f_i	Pourcentage p_i	Effectif cumulé Croissant $n_i\ c\nearrow$	Effectif cumulé décroissant $n_i\ c\searrow$
0	11	0,11	11%	11	100
1	16	0,16	16%	27	89
2	21	0,21	21%	48	73
3	25	0,25	25%	73	52
4	17	0,17	17%	90	27
5	8	0,08	8%	98	10
6 et plus	2	0,02	2%	100	2
Total	100	1	100%	-	-

Tableau statistique
Variable continue (Caractère quantitatif continu)

Comment former les classes:

La répartition des données brutes en classes nécessite donc de la part du statisticien de faire un choix sur le nombre de classes et donc sur l'amplitude. Ce choix doit être suffisamment judicieux pour permettre la représentation graphique des données sans perdre pour autant trop d'information initialement contenue dans la série statistique.

On ordonne les données par ordre croissant et on calcule la longueur de la série statistique :

$$\boldsymbol{étendue = e = x_n - x_1}$$

x_1 est la plus petite observation

x_n est la plus grande observation

n étant le nombre d'observations.

Pour construire le tableau statistique on doit regrouper les données dans des intervalles qu'appelle **classes**, pour cela on doit d'abords déterminer les nombre de classes qui sera défini par

$$n_c = 1 + \frac{10}{3}\log_{10} n$$

où n est la taille des valeurs recueillie

Par la suite on détermine la longueur des classes qu'on appelle

telle que

$$a = \frac{e}{n_c}$$

Tableau statistique
Variable continue (Caractère quantitatif continu)

Exemple 2. On étudie le poids en Kg de 150 nouveaux né dans une maternité. On a obtenu les résultats suivants :

01	02	03	04	…….	148	149	150
2,356	3,102	2,212	4,125	…….	3,256	3,894	2,352

Pour notre exemple on a; $x_1 = 2.212$ et $x_n = 4.593$.

alors

$$e = 4.593 - 2.212 = 2.381$$

$$n_c = 1 + \frac{10}{3}\log_{10} 150 = 8.2536 \approx 8$$

$$a = \frac{e}{n_c} = \frac{2{,}381}{8} \approx 0.2976 \approx ,03$$

Tableau statistique
Variable continue (Caractère quantitatif continu)

X Poids des nouveaux né	Centre
[2,2; 2,5[	2,35
[2,5; 2,8[	2,65
[2,8; 3,1[	2,95
[3,1; 3,4[	3,25
[3,4; 3,7[	3,55
[3,7; 4,0[	3,85
[4,0; 4,3[	4,15
[4,3; 4,6[	4,45
TOTAL	-

X_i Poids des nouveaux né	Centre x_i	n_i	f_i	p_i	n_i c↗	n_i c↘
[2,2; 2,5[	2,35	5	0,0333	3,33%	5	150
[2,5; 2,8[	2,65	11	0,0733	7,33%	16	145
[2,8; 3,1[	2,95	21	0,1400	14%	37	134
[3,1; 3,4[	3,25	39	0,2600	26%	76	113
[3,4; 3,7[	3,55	35	0,2333	23,33%	111	74
[3,7; 4,0[	3,85	20	0,1333	13,33%	131	39
[4,0; 4,3[	4,15	13	0,0867	8,67%	144	19
[4,3; 4,6[	4,45	6	0,0400	4%	150	6
TOTAL	-	150	1	100%	-	-

PARAMETRES STATISTIQUES

Paramètres de position et de dispertion

On distingue deux catégories de valeurs typiques :

- Les paramètres du $\mathbf{1}^{er}$ ordre ou paramètres de position : **moyenne arithmétique, mode et médiane**.
- Les paramètres du $\mathbf{2}^{ème}$ ordre ou paramètres de dispertion : **écart-type, coefficient de variation** et **étendue interquartile**.

Paramètres de position: Le Mode

Le mode, noté **Mo**, d'un caractère (qualitatif ou quantitatif) est la modalité la plus observée, c'est à dire celle qui a le plus grand effectif ou la plus grande fréquence. Le mode peut ne pas exister et s'il existe, ne pas être unique.

- L'ensemble **2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18** a pour mode **Mo = 9**
- L'ensemble **3, 5, 8, 10, 12, 15, 16** n'a pas de mode.
- L'ensemble **2, 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18** a deux modes $\boldsymbol{Mo_1}$ **= 2** et $\boldsymbol{Mo_2}$ **= 9**, on dit que c'est une série bimodale.

La Moyenne arithmétique

La moyenne arithmétique d'une série statistique $x_1, x_2, \ldots, x_k$, d'un caractère quantitatif X, et d'effectifs respectifs $n_1, n_2, \ldots, n_k$ est donné par le nombre réel X défini par

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

où x_i est la modalité *i* de la variable X dans le cas d'une variable discrète et c'est le centre de la classe *i* dans le cas d'une variable continue et $n = \sum_{i=1}^{k} n_i$

La Médiane

Soit une série statistique d'une variable X ayant pour modalités (ordonnées par ordre croissant) $x_1 < x_2 < \cdots < x_n$. On appelle médiane de X, le nombre **Me**, s'il existe, qui partage la série statistique en deux parties d'égal effectif.

La Médiane (Cas d'une variable discrète)

Considérons les notes de 19 étudiants obtenues à l'examen de statistiques :

$$9, 15, 4, 8, 16, 10, 11, 10, 5, 12, 9, 10, 7, 12, 15, 11, 6, 13, 12.$$

On ordonne la série par ordre croissant, on a alors

$$4, 5, 6, 7, 8, 9, 9, 10, 10, 11, 11, 11, 12, 12, 12, 13, 15, 15, 16.$$

La moitié de la série est 9 alors

$$\{4, 5, 6, 7, 8, 9, 9, 10, 10\}, \mathbf{11}, \{11, 11, 12, 12, 12, 13, 15, 15, 16\}$$

$9\ valeurs$ $9\ valeurs$

la médiane est le nombre qui partage la série en deux parties de même effectif, la valeur qui réalise ceci est 11 donc **Me = 11**.

On ajoute maintenant la note 7 à la série précédente, on a alors une série de 20 notes :

$$4, 5, 6, 7, 8, 9, 9, 10, 10, 10, 11, 11, 11, 12, 12, 12, 13, 15, 15, 16.$$

La moitié de la série est 10 alors

$$\{4, 5, 6, 7, 7, 8, 9, 9, 10, 10\}, \{11, 11, 12, 12, 12, 13, 15, 15, 16\}$$

$10\ valeurs$ $10\ valeurs$

la médiane est le nombre qui partage la série en deux parties de même effectif, elle se trouve entre le dernier 10 et le premier 11, dans ce cas on prendra la valeur moyenne de ces deux notes, alors $Me = \frac{10+11}{2} = 10,5.$

D'une manière générale soit X est une variable discrète prenant les valeurs ordonnées $x_1, x_2, \dots, x_P, x_{p+1}, \dots, x_n$, alors si

$$n = 2p \Rightarrow Me = \frac{x_p + x_{p+1}}{2}$$

$$n = 2p + 1 \Rightarrow Me = x_{p+1}.$$

La Médiane (Cas d'une variable continue)

On considère l'exemple des nouveaux né

X Poids des nouveaux né	Centre x_i	n_i	n_i c↗	$n_i\, x_i$
[2,2; 2,5[	2,35	5	5	11,75
[2,5; 2,8[	2,65	11	16	29,15
[2,8; 3,1[	2,95	21	37	61,95
[3,1; 3,4[	3,25	39	76	126,75
[3,4; 3,7[	3,55	35	111	124,25
[3,7; 4,0[	3,85	20	131	77
[4,0; 4,3[	4,15	13	144	53,95
[4,3; 4,6[	4,45	6	150	26,7
Total		150	-	511,5

On détermine d'abord la classe médiane qui correspond à la moitié des effectifs $\left(\frac{n}{2} = 75\right)$ d'où $M \in [3,1;\ 3,4[$.

$$Me = x_i + \Delta x_i \frac{\frac{n}{2} - n_{i-1}^{c} \nearrow}{n_i^{c} \nearrow - n_{i-1}^{c} \nearrow}$$

Remarque:

Le mode dans le cas continue (classe modale) est égale à

$$Mo = x_i + \Delta x_i \frac{d_1}{d_1 + d_2}$$

$d_1 = n_i - n_{i-1}$

$d_2 = n_i - n_{i+1}$

Ou $\Delta x_i = x_{i+1} - x_i$

n_i effectif de la classe modale

n_{i-1} effectif de la classe précédente

n_{i+1} effectif de la classe suivante

Paramètres de dispersion: L'écart type

L'écart type d'une série statistique $(x_1, x_2, \dots, x_k)$, d'un caractère X, et d'effectif respectif respectif $(n_1, n_2, \dots \quad, n_k)$ est donné par le nombre réel δ_X défini par

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} n_i (x_i - \bar{X})^2}$$

$$Ou\ \sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} n_i {x_i}^2 - \bar{X}^2}$$

Ou $n = \sum_{i=1}^{n} n_i$, x_i est la modalité i de la variable X dans le cas d'une variable discrète et c'est le centre de la classe i dans le cas d'une variable continue.

Remarque. On appelle variance de la variable X, noté $Var\ (X)$, le carré de l'écart type. $Var\ (X) = {\sigma_X}^2$

Paramètres de dispersion: Coefficient de variation

Le coefficient de variation est un paramètre de dispersion relative exprimé en pourcentage et défini par

$$CV_{\mathrm{X}} = 100\frac{\sigma_X}{\bar{X}}$$

Paramètres de dispersion: Etendue interquartile

Les quartiles d'une série statistique sont les valeurs qui partagent la série en quatre parties de même effectif. Alors il existe trois quartiles, le premier quartile Q_1, le deuxième quartile Q_2 et le troisième quartile Q_3. Le deuxième quartile Q_2 étant la médiane M.

L'étendue interquartile:

$$IQR = Q_3 - Q_1$$

Cas d'une variable discrète

Les quartiles

- $Q_1 = \frac{n}{4}$
- $Q_2 = \frac{n}{2}$
- $Q_3 = \frac{3n}{4}$

$$IQR = Q_3 - Q_1$$

Cas d'une variable continue

- $Q_1 = \frac{n}{4}$

$$Q_1 = x_i + \Delta x_i \frac{\frac{n}{4} - n_{i-1}^{c} \nearrow}{n_i^{c} \nearrow - n_{i-1}^{c} \nearrow}$$

- $Q_2 = Me$

- $Q_3 = \frac{3n}{4}$

$$Q_3 = x_i + \Delta x_i \frac{\frac{3n}{4} - n_{i-1}^{c} \nearrow}{n_i^{c} \nearrow - n_{i-1}^{c} \nearrow}$$

GRAPHIQUES

VARIABLE QUALITATIVE

Nominale

Ordinale

Effectifs ou Fréquences

Diagramme en barres

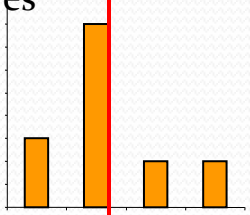


Diagramme en barres

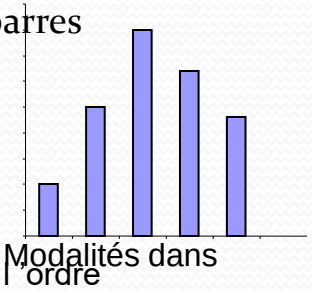
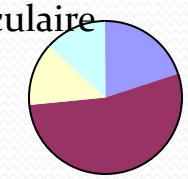


Diagramme circulaire



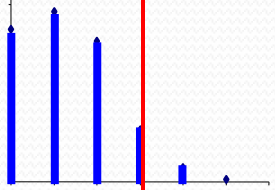
VARIABLE QUANTITATIVE

Discrète

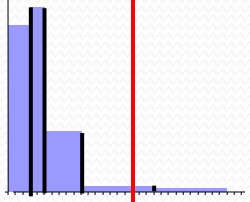
Continue

Effectifs ou Fréquences

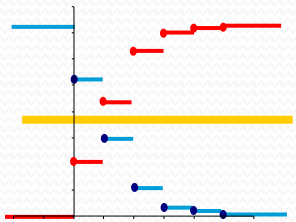
Diagramme en bâtons

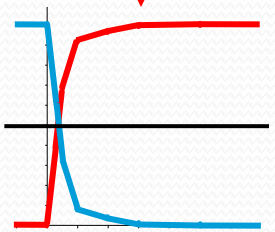


Histogramme



Courbes cumulatives des effectifs ou des fréquences





Séries statistiques doubles

Introduction

Soit $\mathcal{P}$ une population d'effectif total n, sur laquelle on étudie deux caractères quantitatif X et Y , on s'intéresse à la liaison entre ces deux variables. On définit la série statistique double de $\mathcal{P}$ pour les caractères X et Y

$$\mathcal{P} \longrightarrow \mathbb{R}^2$$

$$e_{ij} \rightarrow \left(X_i, Y_j\right)$$

On essaye de mettre en évidence la liaison entre X et Y consiste à tracer le nuage de points associé à la série statistique double.

Tableau de contingence

X \ Y	y_1	y_2	…	y_j	…	y_l	Effectif marginal
x_1	n_{11}	n_{12}		n_{1j}		n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2l}	$n_{2\bullet}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{il}	$n_{i\bullet}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
x_k	n_{k1}	n_{k2}		n_{kj}		n_{kl}	$n_{k\bullet}$
Effectif marginal	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet l}$	n

n_{ij} est l'effectif partiel du couple (x_i , y_i) , $n_{i\bullet}$ est l'effectif marginal de x_i et $n_{j\bullet}$ est l'effectif marginal de y_j

$n_{i\bullet} = \sum_{j=1}^{l} n_{ij}$ et $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$

Définition Le couple (X, Y) est statistiquement indépendant si on a $\forall i = 1, \dots, k;\ j = 1, \dots, l$

$$f_{ij} = \frac{n_{ij}}{n} = f_{i\bullet} f_{\bullet j} = \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n}$$

- **Définition** On appelle covariance des variables X et Y et on note $Cov\,(X, Y)$, le nombre

$$Cov\,(X, Y) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} (x_i - \bar{X})(y_j - \bar{Y})$$

$$= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}\, x_i y_j - \bar{X}\bar{Y}$$

Remarque

*Les moyennes marginales $\bar{X}$ et $\bar{Y}$ sont données par

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\, x_i \text{ et } \bar{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}\, y_j$$

*Si les variables X et Y sont statistiquement indépendantes alors $Cov\,(X,Y\,) = 0$. Mais la réciproque n'est pas vraie.

Définition

Si $Cov\,(X,Y\,) = 0$ on dit que les variables X et Y sont non corrélées.

Définition

On appelle coefficient de corrélation linéaire des variables X et Y le nombre

$$r = \rho\,(\mathrm{X},\mathrm{Y}\,) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Remarque

- On a $-1 \leq \rho(X,Y\,) \leq 1$
- Si $\rho(X,Y\,) = 0$, les variables X et Y sont non corrélées.
- Soient X et Y deux variables statistiques liées par la relation $Y = \mathrm{a}X + b$

On a $cov(X,Y) = a\sigma^2{}_X$ et comme $\sigma^2{}_Y = a^2\sigma^2{}_X$ alors

ρ (X,Y) =1 si a > 0 et ρ (X,Y) $= -1$ si a < 0

- Quand ρ^2 (X,Y) $> 0{,}9$ il existe une relation linéaire entre X et Y de la forme $Y = aX + b$ qu'on appelle droite de régression de Y en X et qui rend minimum la somme

$$S = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}\left(y_j - ax_i - b\right)^2$$

Théorème

La droite de régression de Y en X est la droite de la forme

$Y = \mathrm{a}X + b$ avec

$$\mathrm{a} = \frac{cov(X,Y)}{\sigma^2{}_X} \text{ et } b = \bar{Y} - \frac{cov(X,Y)}{\sigma^2{}_X}\bar{X} = \bar{Y} - \mathrm{a}\bar{X}$$


Merci pour votre
Attention