1 - Statistiques descriptives

Introduction

Définition

« La **statistique** a pour objet l'étude, à l'aide de traitements mathématiques, de nombreux faits correspondant à l'observation d'un phénomène, dans le but de rendre compte de la réalité, d'essayer de l'expliquer et d'aider à la prise de décision » (J. Hubler, 1996)

Définition

Les statistiques : données chiffrées ou les résultats numériques de la statistique

La collecte des données statistiques

Deux principales sources de données statistiques

Les recensements

Sont des opérations, issues du dénombrement, qui consistent à étudier de façon exhaustive et en fonction de plusieurs critères tous les éléments d'une population

Le dénombrement : comptage des individus d'une population
Le recensement : chiffrer les données selon plusieurs aspects (âge, sexe, chiffre d'affaires, etc.)

Les enquêtes

Portent sur un sous-ensemble d'une population appelé échantillon

La qualité de l'enquête et donc des résultats dépend du choix de l'échantillon

Exemple !

Domaines d'application

Démographie
Sciences économiques et sociales
Sociologie
Marketing
Géophysique
Physique
Médecine
Sciences politiques
Etc.

Vocabulaire statistique (définitions)

Population : ensemble des unités statistiques ou individus sur lesquels on effectue une analyse statistique

Exemple Production de pièces dans une usine, ensemble des étudiants d'une université,…

Unité statistique (individu) : élément de la population sur lequel porte l'observation

Exemple Pièce mécanique, étudiant

Echantillon : sous-ensemble d'individus prélevés dans une population déterminée

Exemple Étudiants de moins de 20 ans,
ensemble de pièces prélevées au hasard dans la production

5

Explorer, décrire et inférer (aperçu)

Des questions importantes se posent et débouchent souvent sur des décisions prises sur la base d'une information limitée sous forme de données.

Analyse exploratoire et descriptive

Décrire le contenu de ses données

Représenter les données de façon à voir des tendances et à découvrir des structures

Exemple Dans quel intervalle la majorité des données est située
Quelles sont les valeurs les plus fréquentes

Inférence statistique

1ère étape : Une analyse exploratoire **suggère** des hypothèses de travail et des modèles qui peuvent être formalisés, confirmés ou refusés
2ème étape : *analyse inférentielle* qui utilise des méthodes de *test et d'estimation.*

Exemple : le salaire des hommes est supérieur à celui des femmes

6

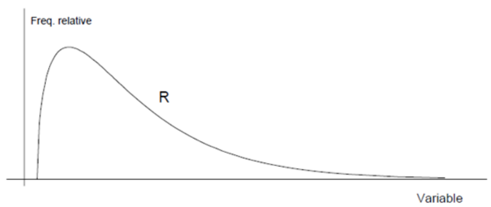
Analyse descriptive

Description graphique des distributions

Distribution numérique des distributions

Exemples

- L'âge moyen des habitants d'une ville
- Proportion d'individus atteints d'une maladie dans une région



7

Analyse descriptive univariée

Variable statistique X (caractère)

Une variable est une caractéristique étudiée pour une population donnée

Ω Population sous étude $\quad X : \Omega \rightarrow E$

E : ensemble des modalités que peut prendre X

$\forall$ i=1,…n, $X(\omega_i) \in E$, Valeur de X prise par l'individu ω_i de la population

- E quelconque, sans structure : variable qualitative (ou nominale)
- E muni d'une structure d'ordre : variable qualitative ordonnée (ou ordinale)
- E muni d'une structure d'espace vectoriel

  (en pratique : $E \in R$) variable quantitative (ou cardinale)

8

Variable statistique X

X qualitative

Population : population française

$X(\omega) = 1 \Leftrightarrow$ Sexe de ω : masculin

$X(w) = 2 \Leftrightarrow$ Sexe de ω : féminin

X quantitative

Population : population française

$X(\omega)$ = âge de ω. $X(\omega) \in$ IN

$X(\omega)$ = salaire mensuel de ω. $X(\omega) \in$ IR

X ordinale

Population : population française

$X(\omega) = 1 \Leftrightarrow$ niveau d'études de ω : certificat d'études primaires

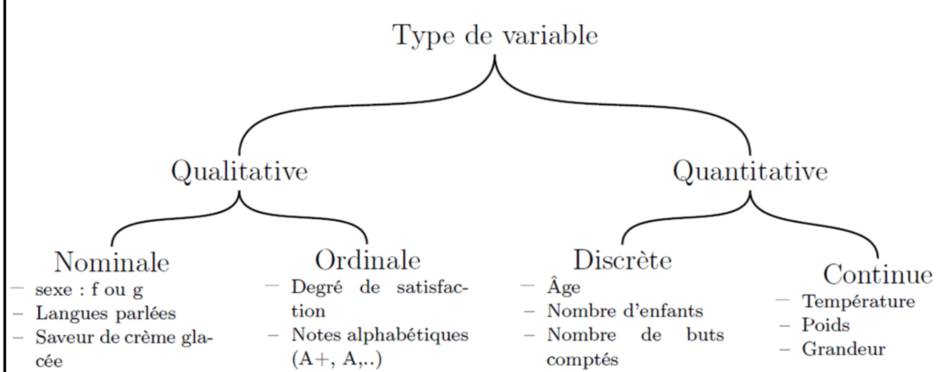
$X(\omega) = 2 \Leftrightarrow$ niveau d'études de ω : brevet d'études secondaires

$X(\omega) = 3 \Leftrightarrow$ niveau d'études de ω : baccalauréat

$X(\omega) = 4 \Leftrightarrow$ niveau d'études de ω : études supérieures

9

Variable statistique X



10

Exemple d'étudiants :   n=45

Sexe (Homme, Femme) , Taille [120, 210] , Poids (40,200) ,
Nbre Frères et Sœurs {0,1?....10} , Couleur des yeux (Brun, Bleu, Vert, Noir Gris}

T	P	S	F	C
180	70	h	2	brun
177	57	h	3	brun
180	60	h	1	bleu
180	66	h	0	brun
183	62	h	6	vert
184	68	h	0	brun
185	65	h	1	noir
184	72	h	2	brun
174	65	h	3	noir
180	72	h	1	brun
168	52	h	3	brun
180	75	h	0	bleu
183	75	h	2	brun
181	68	h	0	bleu
180	65	h	4	brun

T	P	S	F	C
190	66	h	1	brun
183	78	h	0	bleu
167	60	h	4	bleu
181	67	h	0	brun
179	98	h	2	brun
173	75	h	1	vert
170	68	h	1	gris
170	59	h	3	brun
183	72	h	2	bleu
179	73	h	3	vert
180	72	h	3	bleu
188	70	h	2	brun
176	65	h	1	vert
178	72	h	1	brun
185	71	h	1	bleu

T	P	S	F	C
168	52	f	0	brun
157	47	f	1	vert
167	53	f	2	vert
168	57	f	4	bleu
163	65	f	1	brun
167	60	f	2	brun
166	68	f	2	bleu
164	49	f	7	vert
172	57	f	3	brun
165	59	f	2	bleu
158	62	f	0	brun
161	65	f	1	brun
160	61	f	1	bleu
162	58	f	2	brun
165	58	f	5	brun

11

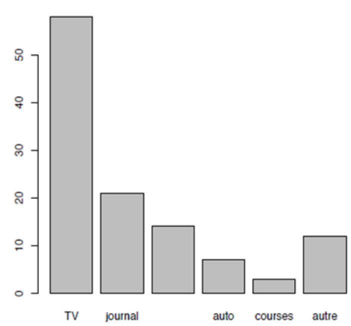
Variable qualitative

Les modalités d'une variable qualitative ne sont pas numériques. On ne peut pas calculer les paramètres statistiques : moyenne, écart-type, …etc.

On peut représenter les données dans un tableau en indiquant pour chaque modalité l'effectif, la fréquence relative,…

Exemple : Circonstances pendant lesquelles les étudiants se rongent les ongles

activité	*fréquence*
regarder la télévision	*58*
lire un journal	*21*
téléphoner	*14*
conduire une auto	*7*
faire ses courses	*3*
autre	*12*



12

Distribution d'une variable qualitative

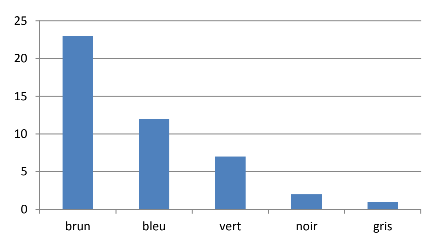
$\{x_1, x_2, \ldots, x_k\}$ l'ensemble des modalités de X

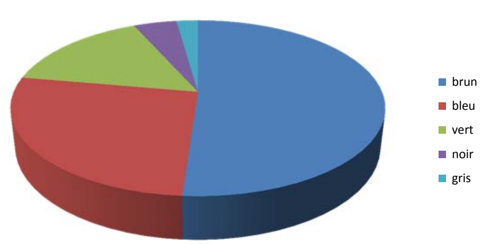
– *fréquence absolue de* x_i : n_i
– *fréquence relative de* x_i : $f_i = n_i/n$
– *distribution de fréquence de X, l'ensemble des couples* (x_i, n_i) *ou* (x_i, f_i).

Modalité	Fréquence absolue	Fréquence relative
brun	23	0.511
bleu	12	0.267
vert	7	0.156
noir	2	0.044
gris	1	0.022
Totaux	45	1.000

13

Distribution d'une variable qualitative





14

Variable quantitative

Une variable quantitative possède des modalités mesurables.

Elle peut être discrète ou continue.

Variable quantitative discrète

Les valeurs peuvent être isolées. Le nombre des valeurs est généralement petit.
Exemple : Nombre d'enfants par famille, nombre de visites d'un musée par jour

Variable quantitative continue

Elle peut prendre un nombre **infini** de valeurs dans son intervalle de définition
Les valeurs sont très nombreuses et leur énumération est fastidieuse.

Exemple : Taille des étudiants.
Entre 1,70m et 1,73m on peut avoir un grand nombre de valeurs (une infinité en théorie)

Il est préférable de découper la variable en classes.

15

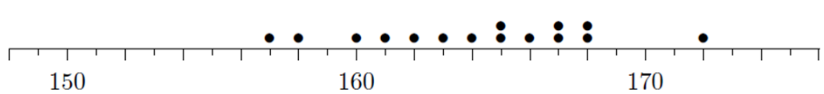
Distribution d'une variable quantitative

Etudier la forme d'une distribution
S'intéresser à l'endroit où se situe la distribution et à son étalement

1 - Cas où le nombre n d'observations est petit (n < 20)

Représenter les observations sur l'axe

Exemple Taille d'un échantillon d'étudiants



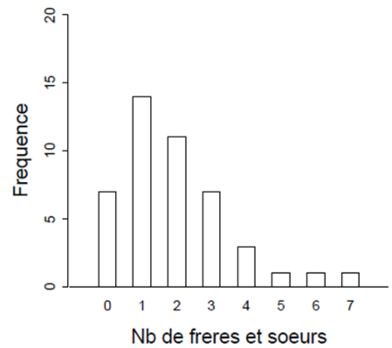
Mise en évidence les valeurs extrêmes ou aberrantes (*outliers)*

16

2 - Cas où le nombre d'observations différentes est petit comparativement à n

Exemple Distribution du nombre de frères et sœurs dans l'échantillon de 45 étudiants

Modalité (Nb de frères et sœurs) x_i	Fréquence absolue n_i	Fréquence relative f_i
0	7	0.156
1	14	0.311
2	11	0.244
3	7	0.156
4	3	0.067
5	1	0.022
6	1	0.022
7	1	0.022

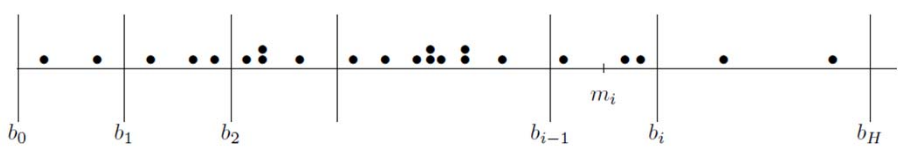


La majorité des étudiants ont 0, 1, 2, ou 3 frères et sœurs

17

3 - Cas où le nombre de modalités et celui des observations est grand

Regrouper les données en classes



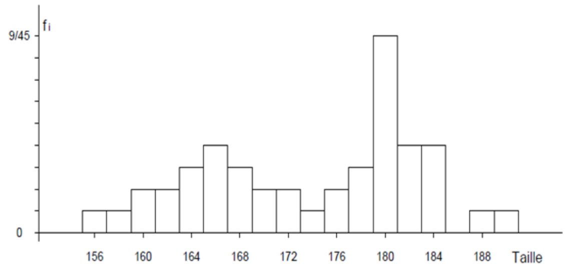
Recommandations pour réaliser un histogramme

- Nombre de classes entre 5 et 20
- plus n est grand, plus le nombre de classes peut être grand
- presque toutes les classes contiennent un nombre élevé d'observations
- Les classes sont de largeurs égales (Sauf classes à très grand effectif)
- Dans la mesure du possible on évitera d'avoir des classes ouvertes

18

3 - Cas où le nombre de modalités et celui des observations est grand

Exemple Taille des étudiants

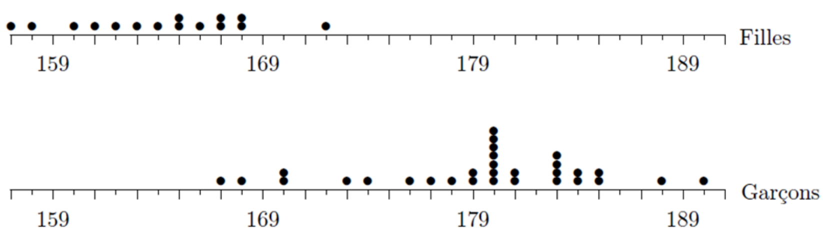


Classe	Fréq. n_i	Fréq.rel. f_i
155-157	1	1/45
157-159	1	1/45
159-161	2	2/45
161-163	2	2/45
163-165	3	3/45
165-167	4	4/45
167-169	3	3/45
169-171	2	2/45
171-173	2	2/45
173-175	1	1/45
175-177	2	2/45
177-179	3	3/45
179-181	9	9/45
181-183	4	4/45
183-185	4	4/45
185-187	0	0/45
187-189	1	1/45
189-191	1	1/45

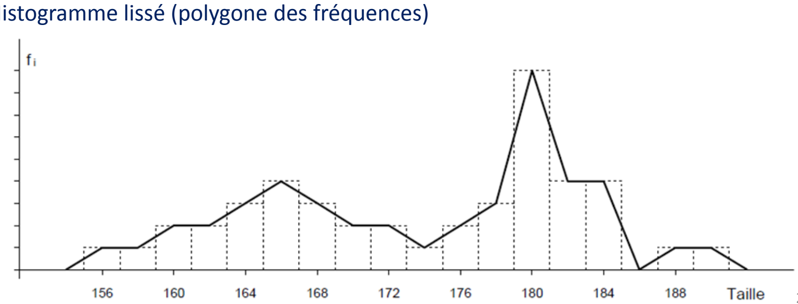
L'histogramme des tailles est bimodale

→ l'échantillon peut être réparti en deux groupes

3 - Cas où le nombre de modalités et celui des observations est grand



Histogramme lissé (polygone des fréquences)

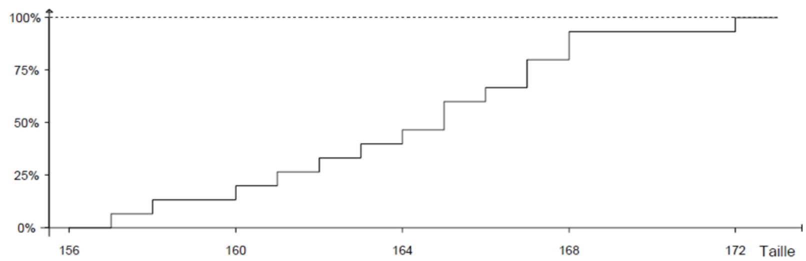


3 - Cas où le nombre de modalités et celui des observations est grand

Fonction de distribution cumulative empirique

$x_1, \ldots, x_n$: ensemble d'observations *d'une variable X*

$F_n(x)$ = (nombre des $x_i \le x$) / n

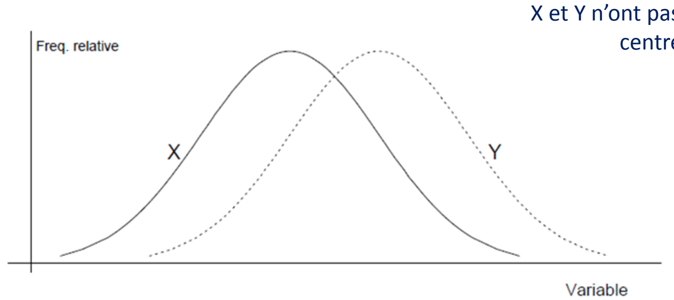


21

Caractéristiques principales d'une distribution

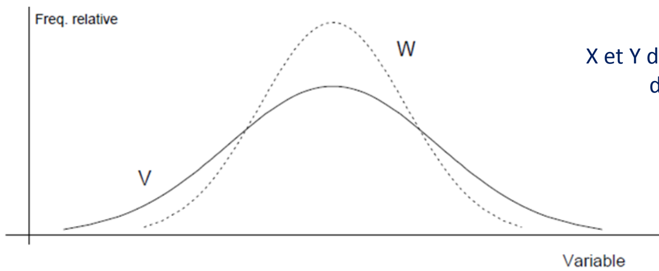
1. le centre (et toute autre caractéristique qui détermine la position)
2. la dispersion (étalement)
3. la symétrie ou dissymétrie par rapport au centre;
4. le nombre de modes (bosses).

X et Y n'ont pas le même centre

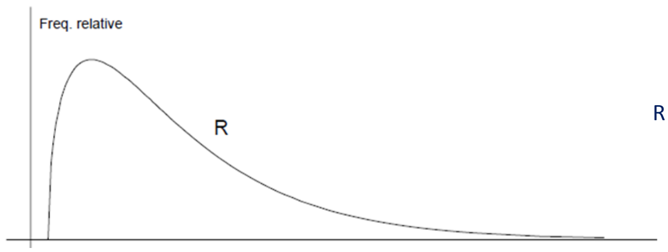


22

Caractéristiques principales d'une distribution



X et Y diffèrent par leur dispersion



R présente une dissymétrie

23

Description numérique des distributions

Principales synthèses

- *position*
- *dispersion*
- dissymétrie

Moyenne arithmétique

$$m(X) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$$ (Pondérations)

Moyenne géométrique

Le logarithme des observations de X est moins asymétrique que les observations de X

Si Y=ln(X) y_i=ln(x_i) $$m(Y) = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) = \ln(x_1.x_2....x_n)^{1/n}$$

$$\boxed{g(X) = (x_1.x_2....x_n)^{1/n}}$$ s'appelle moyenne géométrique

24

Exemple d'application

A l'issue d'une manifestation, la police annonce 10 000 manifestants, et les organisateurs 100 000.

Quel est le nombre de manifestants?

Police	Manifestants	Moyenne	Remarque
10 000	100 000	55 000	Surestime l'importance du chiffre donné par les organisateurs par rapport à la police
1000	100 000	50 500	Changement faible

La police et les organisateurs trichent de la même façon

Soit **x** est le nombre réel de manifestants,
Si la police annonce **2** fois moins, les manifestants annoncent **2** fois plus

Si **k** est le coeff. multiplicateur, la police annonce **x/k** manifestants, et les organisateurs **kx**

Une meilleure approximation : la moyenne géométrique

$$g(X) = \sqrt{(10\,000) * (100\,000)} = 31622$$

25

Moyenne harmonique

C'est l'inverse de la moyenne arithmétique de l'inverse des x_i

$$h(X) = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots.. + \frac{1}{x_n}\right)}$$

Exemple

Escalader une côte de 1km à 20 km/h, la redescendre à 30 km/h

Vitesse moyenne : 25 km/h?

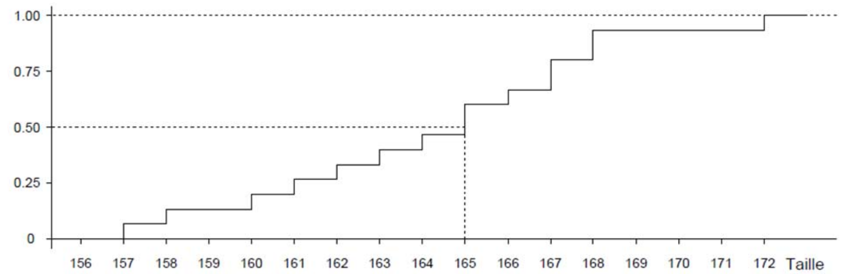
$$x = v.t \qquad x = x_1 + x_2 = 1 + 1 = 2 \qquad t = t_1 + t_2 = \frac{1}{v_1} + \frac{1}{v_2}$$

Vitesse moyenne : $v = \frac{x}{t} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}} = 24\,\text{km/h}$

26

La médiane (Me)

Valeur de la variable qui partage la série en deux parties égales



$F_n(x) \leq 0.5$ si $x < Me$
$F_n(x) \geq 0.5$ si $x > Me$

27

Moyenne ou médiane ?

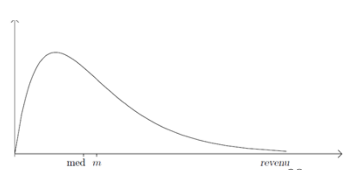
- Si la distribution est symétrique, m(X)=Me(X)
- La moyenne se laisse influencer par les valeurs exceptionnelles, atypiques ou erronées on dit que Me(X) est plus robuste que m(X)

Exemple 27, 29, 31, 31, 31, 34, 36, 39, 42.
n = 9, (n+ 1)/2 = 5 et Me= x[5] = 31 (cinquième valeur)

27, 29, 31, 31, 31, 34, 36, 39, 42, 45.
n = 10, n/2 = 5, n/2 + 1 = 6 et Me = (x[5] + x[6])/2 = (31 + 34)/2 = 32.5

Lorsque la distribution de la majorité des données est symétrique mais il y a des outliers, la médiane est généralement préférable.

• Pour un salarié, il est intéressant de connaître la médiane dans le but de situer son propre revenu dans la “moitié riche” ou dans la “moitié pauvre”
• Pour le département des finances il est préférable de déterminer la moyenne de cette distribution, car elle permet d’estimer le bénéfice attendu des impôts



28

Les quantiles

Classer les observations (et non pas les modalités observées) par ordre croissant, puis de repérer des points de coupure dans la suite ainsi définie, selon un pourcentage que l'on s'est fixé.

La valeur du pourcentage détermine le(s) paramètres :

50% pour la médiane

25% pour les quartiles

10% pour les déciles

1% pour les centiles

Le concept de quantile nécessite de travailler avec une variable ordinale ou cardinale

29

Le mode

Le mode est la modalité observée la plus fréquente

Le mode est toujours calculable, quelque soit le type de la variable

fumeur_3

		Fréquence	Pour cent	Pourcentage valide	Pourcentage cumulé
Valide	jamais fumé	446	44,4	44,5	44,5
	ancien fumeur	187	18,6	18,6	63,1
	fumeur	370	36,8	36,9	100,0
	Total	1003	99,8	100,0	
Manquante	Système manquant	2	,2		
Total		1005	100,0		

Il peut y avoir plusieurs modes

Si les données sont groupées en classes, on parle de **classe modale**

30

Variance et écart-type

La variance mesure donc la dispersion de la distribution autour de sa moyenne : plus la distribution est dispersée autour de la moyenne, plus la variance est grande et inversement.

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(xi - \overline{x})^2 \qquad V(X) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$$

V(X) = 0 $\Leftrightarrow$ les observations sont toutes égales et donc égales à la moyenne.

$$\sigma = \sqrt{V(X)}$$

Avantage de l'écart-type : même échelle que la variable X

31

Variance et écart-type

Distribution 1

Note	Effectif
8	15
10	20
12	15

$\overline{X}$ = 10

σ^2 = 2,4

Distribution 2

Note	Effectif
4	15
10	20
16	15

$\overline{X}$ = 10

σ^2 = 21,6

Même moyenne

Variances différentes

32

Variance et écart-type

Distribution 1	
Note	Effectif
8	15
10	20
12	15

Distribution 2	
Note	Effectif
4	15
10	20
16	15

Observations resserrées autour de la moyenne

Observations étalées autour de la moyenne

Même moyenne

$\sigma^2 = 2{,}4$

Variances différentes

$\sigma^2 = 21{,}6$

33

Coefficient de variation

$$C_v = \frac{\sigma}{\bar{x}}$$

Mesure le dispersion relative (sans dimension)

MAD median absolute deviation

MAD(X) = Me (X - Me(X))

Ecart interquartile

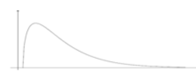
$$I_q = q_{0.75} - q_{0.25}$$

34

Le coefficient de dissymétrie de Pearson

Une distribution de fréquences est positivement dissymétrique (dissymétrique à droite) si la portion de sa courbe située à droite du sommet (mode) est plus longue que l'autre

positivement dissymétrique : mode < Me < m



négativement dissymétrique : mode > Me > m

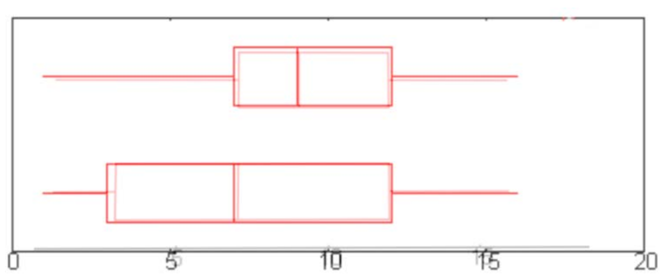
Le coefficient de dissymétrie de Pearson : $\dfrac{3\left(m(x)-Me(X)\right)}{\sigma(X)}$

35

Le Box-plot (Boîte à moustaches)

Représentation graphique simple mais puissante qui permet de juger la position, la dispersion et la symétrie des données.

Ce diagramme est utilisé souvent pour comparer un même caractère dans deux populations de tailles différentes.

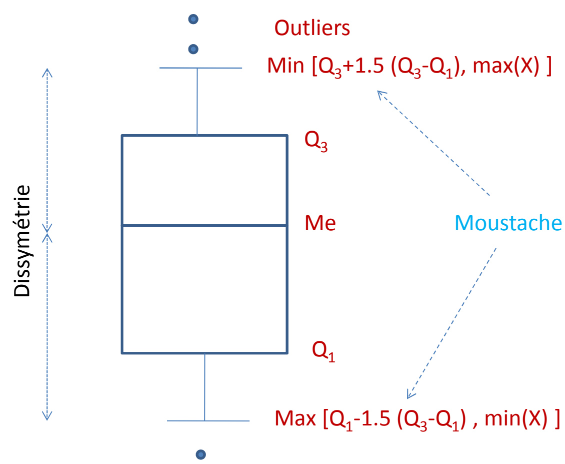


36

Le Box-plot (Boîte à moustaches)

Cinq synthèses numériques (médiane, quartiles, limites, outliers)

visualiser les informations essentielles (position, dispersion, asymétrie)



37

Relation entre deux variables

$x_1, \ldots, x_n$ et $y_1, \ldots, y_n$ valeurs deux variables quantitatives X et Y observées sur un échantillon.

(x_i, y_i) ⟶ Individu

Nuage des points (diagramme de dispersion)

Représentation de l'ensemble des individus dans le plan formé par X et Y

Exemple

Repérer les outliers (fautes ou exceptions)

légère association entre taille et poids

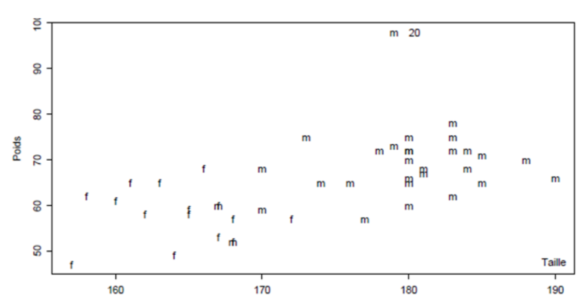


Diagramme Taille/Poids pour l'échantillon de 45 étudiants

Relation entre deux variables

Covariance

$$COV(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m(X))(y_i - m(Y))$$

Exemple

x_i : −9 −5 +3 +7 −1 −7
y_i : +4 +3 −1 −3 +0 +3

	x_i	y_i	$x_i - m(X)$	$y_i - m(Y)$	$(x_i - m(X))(y_i - m(Y))$
	−9	+4	−7	+3	−21
	−5	+3	−3	+2	−6
	+3	−1	+5	−2	−10
	+7	−3	+9	−4	−36
	−1	+0	+1	−1	−1
	−7	+3	−5	+2	−10
Tot	−12	6	0	0	−84
Tot/6	−2	1	0	0	−14

Cov(X,Y) dépend des unités de mesure de X et de Y

Relation entre deux variables

Coefficient de corrélation linéaire

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma(X).\sigma(Y)}$$

Exemple

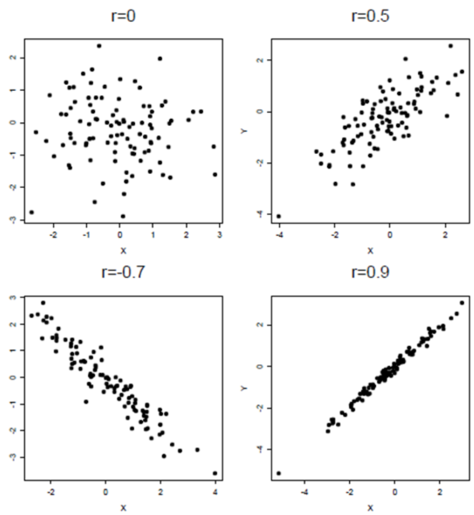
$$r(X,Y) = \frac{-14}{\sqrt{31.67}\sqrt{6.33}} = -0.989$$

Le coefficient de corrélation linéaire mesure le degré de corrélation linéaire entre X et Y

$-1 \leq r(X,Y) \leq 1$

Relation entre deux variables

Coefficient de corrélation linéaire



Relation entre deux variables

Régression simple

Y : variable à expliquer X : variable explicative

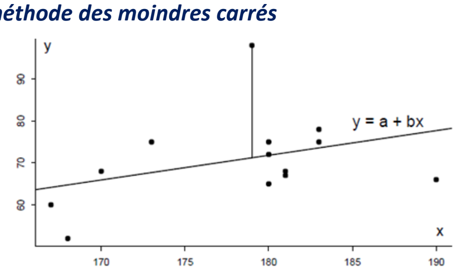
Exemple *X = dose d'un médicament* *Y = une mesure d'amélioration*
X = poids; *Y = taux de cholestérol*

choisir une forme géométrique simple : droite d'équation *y = a x + b*

Ajustement d'une droite par la méthode des moindres carrés

Minimiser l'expression :

$$\sum_i (y_i - a x_i - b)^2$$



Relation entre deux variables

Régression simple

Y : variable à expliquer X : variable explicative

Y = a X + b **droite de régression de Y en X**

$$\hat{a} = \frac{\mathrm{Cov}(X,Y)}{V(X)} \qquad \hat{b} = \overline{y} - \hat{a}\overline{x}$$

La droite des moindres carrés **Y = a X + b** passe par le point ($\overline{x}$, $\overline{y}$)

$$R^2 = \frac{\sigma^2(\hat{Y})}{\sigma^2(Y)}$$

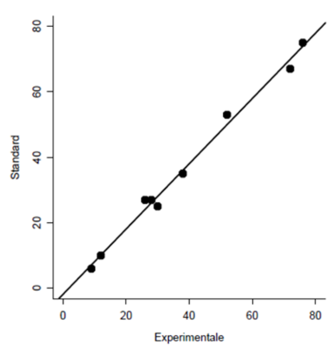
- *$0 \le R^2 \le 1$,*
- *Si R^2 est proche de 1 (par exemple $R^2 = 0.8$), X explique très bien la variation de Y*
- *Si R^2 est proche de 0, la variable X n'est pas une bonne variable explicative.*
- *$R^2 = [r(X, Y)]^2$*

Relation entre deux variables

Régression simple

Exemple pression intracrânienne (en mm Hg) chez le chien
Avec ou sans perforation du crâne

X	Y
Mesure expérimentale	Mesure standard
9	6
12	10
28	27
72	67
30	25
38	35
76	75
26	27
52	53



m(X) = 38.11 *m(Y) = 36.11,*
σ^2(X) = 513.43 *σ^2(Y) = 514.54*
Cov(X, Y) = 511.77
r(X,Y)=0.996
R^2=0.992

$\hat{a} = 0.997$ $\hat{b} = -1.876$

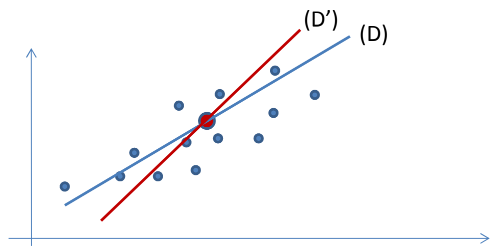
Relation entre deux variables

Droite de régression de X en Y

X = a' Y + b' **droite de régression de X en Y**

$$\hat{a} = \frac{Cov(X,Y)}{V(Y)} \qquad \hat{b}' = \overline{x} - \hat{a}'\overline{y}$$

La droite des moindres carrés **X = a' Y + b'** passe par le point $(\overline{x}, \overline{y})$



Application 1

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

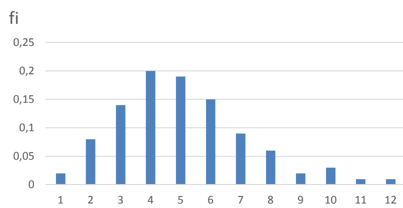
Nombre de voitures	1	2	3	4	5	6	7	8	9	10	11	12
Nombre d'oservations	2	8	14	20	19	15	9	6	2	3	1	1

- Construire la table des fréquences et le diagramme en bâtons en fréquences de la série du nombre de voitures.
- Calculer la moyenne et l'écart-type de cette série.
- Déterminer la médiane, les quartiles et tracer le box-plot.

46

Application 1

xi	1	2	3	4	5	6	7	8	9	10	11	12
ni	2	8	14	20	19	15	9	6	2	3	1	1
fi	0,02	0,08	0,14	0,2	0,19	0,15	0,09	0,06	0,02	0,03	0,01	0,01
Fi	0,02	0,1	0,24	0,44	0,63	0,78	0,87	0,93	0,95	0,98	0,99	1
nixi	2	16	42	80	95	90	63	48	18	30	11	12
$nixi^2$	2	32	126	320	475	540	441	384	162	300	121	144



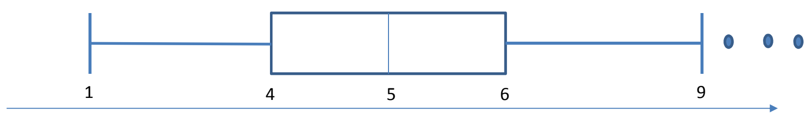
Moyenne : 5,07

Variance : 4,76

Ecart-type : 2,18

Médiane : 5

Q1: 4 Q3 : 6



47

Application 2

- On cherche à étudier la relation entre le nombre d'enfants d'un couple et son salaire. On dispose de la série bidimensionnelle suivantes :

Salaire en euros (X)	510	590	900	1420	2000	600	850	1300	2200
Nombre d'enfants (Y)	4	3	2	1	0	5	6	7	8

- Calculer le coefficient de corrélation linéaire entre ces deux variables statistiques. Conclusion ?
- Un expert en démographie affirme que les deux caractéristiques sont indépendantes. Qu'en pensez-vous ?
- Tracer la droite de régression de Y en X

48

Application 2

	X	Y
moyenne	1152,2	4
variance	344284	7
Ecart-type	586,76	2,58
Covariance(X,Y)	38,89	
Coeff de corrélation	0,026	
a	0,00011	
b	3,86985	