Statistique Descriptive

N. Jégou

L2 Géographie

Statistiques en GEO

- L2 :
  - Statistique descriptive : 6-CM + 12-TD
  - R - prise en main : 6-CM + 12-TD
- M1 : Régression - Tests - ACP : 6-CM + 18-TD
- M2 : Analyse de données : 12-TD

Bibliographie[1]

Statistique descriptive, cours et exercices corrigés. Hamon, A. & Jégou, N., PUR, 2008

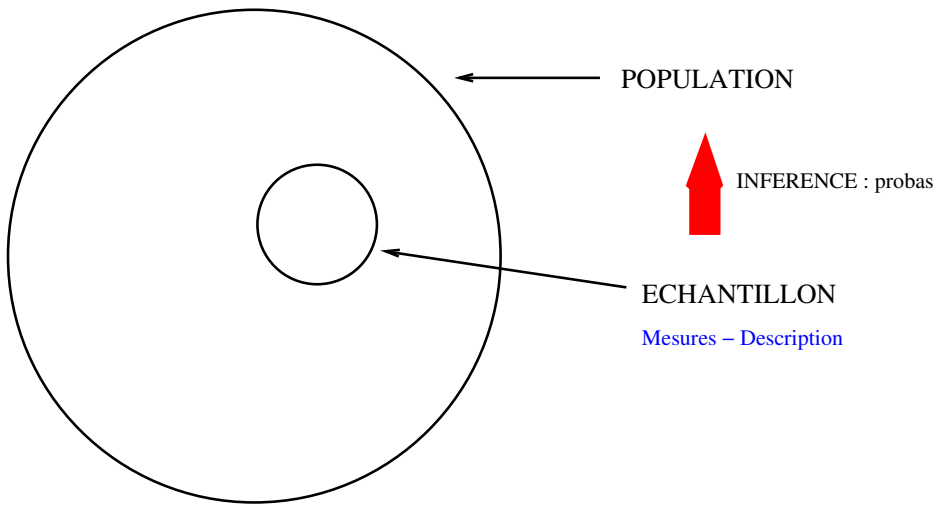
Statistique générale pour utilisateurs. Pagès, J. PUR, 2nd ed., 2010

Statistique avec R. Cornillon et al., 3ème ed. PUR, 2012

[1] pour la L2

Descriptive vs Inférence

Inférence : étendre les propriétés de l'échantillon à la population



Cadre du cours : description, sur la population ou sur un échantillon

Plan du cours

I Statistique à une variable

1. Vocabulaire
2. Graphes
3. Indicateurs

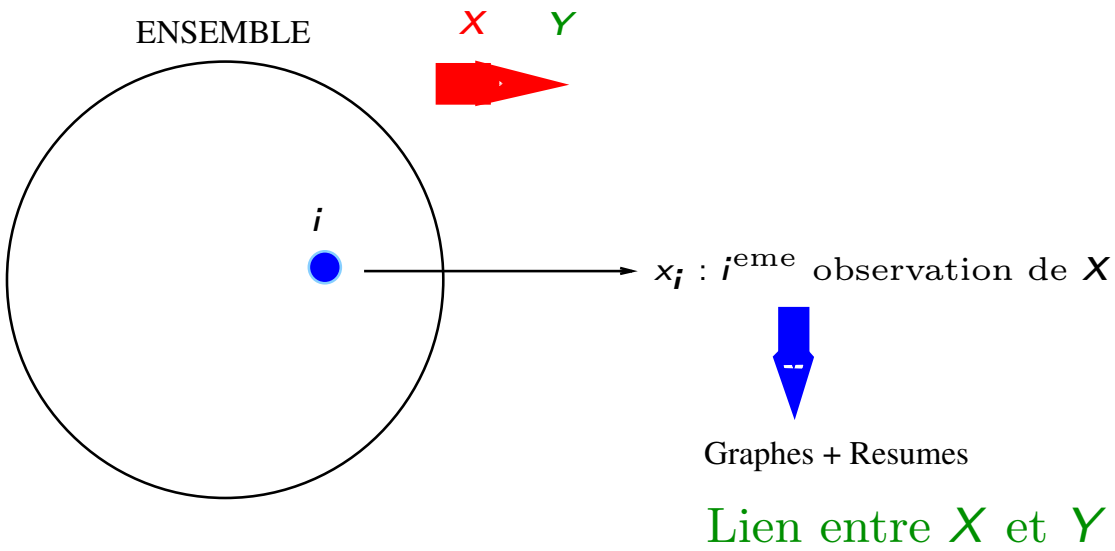
II Croisement de variables

1. Deux qualitatives
2. Qualitative × Quantitative
3. Deux quantitatives

Population - Variable(s)

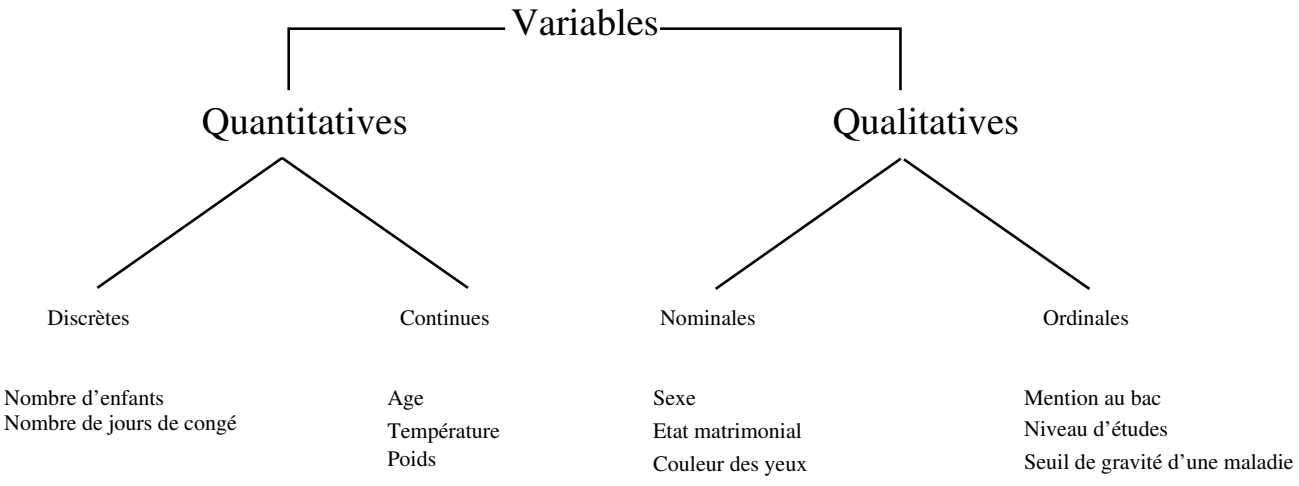
Population = Ensemble d'individus

Variable = Aléatoire (la mesure varie d'un individu à l'autre)



On note n réalisations de X : $\{x_1, \ldots, x_n\}$.

Nature d'une variable



La nature de X oriente le type de représentation

La nature de X et Y oriente l'étude du lien : écarts à l'indépendance, corrélation,...

Exemple

PAYS	SUPERFICIE (milliers de km2)	POPULATION (millions d'hab.)	APPARTENANCE À LA C.E.E.
Allemagne	357	80	O
Autriche	83,8	7,6	N
Belgique	30,5	9,9	O
Danemark	43,1	5,1	O
Espagne	505	39,2	O
Finlande	337	4,9	N
France	552	56,5	O
Grèce	132	10	O
Irlande	70,3	3,5	O
Islande	103	0,3	N
Italie	301	58	O
Luxembourg	3,0	0,4	O
Norvège	324	4,2	N
Pays-Bas	33,9	14,9	O
Portugal	92,1	10,6	O
Royaume-Uni	244	57	O
Suède	450	8,5	N
Suisse	41,3	6,7	N

Fréquences

La fréquence d'observation de x_i est le rapport entre le nombre de fois où x_i est observée et le nombre total d'observations :

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

Ainsi

- $f_i \in [0, 1]$
- f_i peut s'exprimer en pourcentage

Fréquences

Variable qualitative :

Etat matrimonial	Fréquences f_i
Célibataires	0,452
Mariés	0,469
Veufs	0,051
Divorcés	0,028

Variable discrète :

Nombre d'enfants de 0 à 16 ans par famille	Nombre de familles (en milliers)	Fréquences f_i
0	7130	0,505
1	3201	0,227
2	2498	0,178
3	919	0,065
4	241	0,017
5	130	0,009
TOTAL	14119	1

Fréquences

Variable continue :

On regroupe les observations dans des intervalles

SUPERFICIE (km2)	Effectif	Fréquence f_i
[0; 100.000[	8	0,44
[100.000; 200.000[	2	0,11
[200.000; 300.000[	1	0,06
[300.000; 400.000[	4	0,22
[400.000; 500.000[	1	0,06
Plus de 500.000	2	0,11
TOTAL	18	1

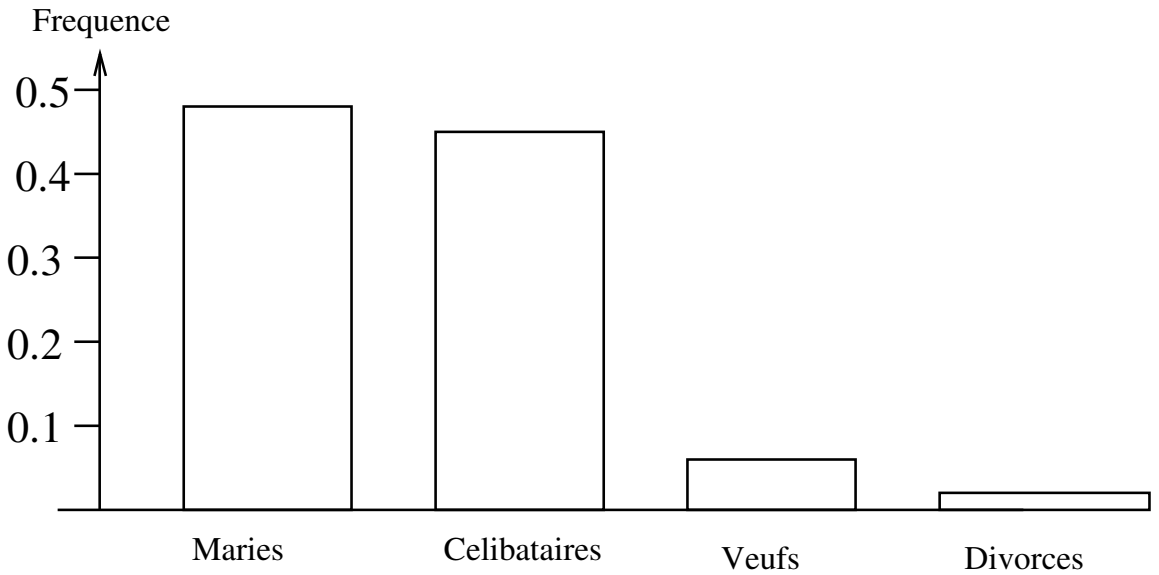
Représentations de la distribution d'une variable

Représentations qui diffèrent selon la nature de la variable

- qualitative : diagramme en barres
- quantitative discrète : diagramme en bâtons
- quantitative continue : histogramme

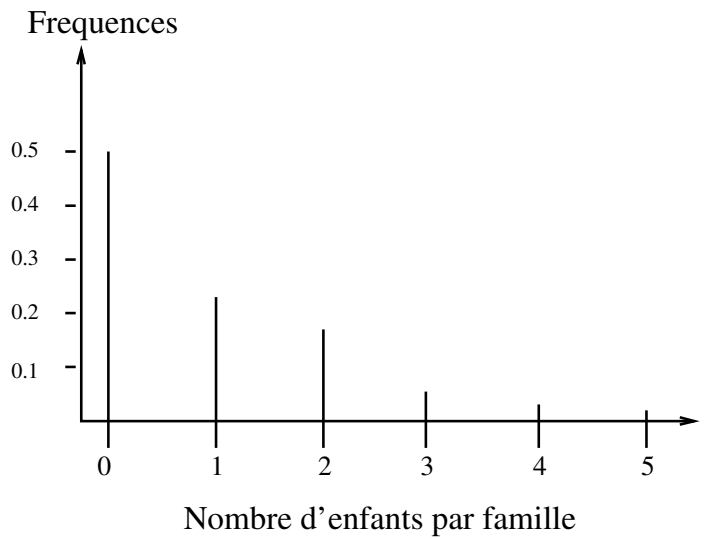
Variable qualitative : diagramme en barres

Etat matrimonial	Fréquences f_i
Célibataires	0,452
Mariés	0,469
Veufs	0,051
Divorcés	0,028



Variable discrète : diagramme en bâtons

Nombre d'enfants de 0 à 16 ans par famille	Nombre de familles (en milliers)	Fréquences f_i
0	7130	0,505
1	3201	0,227
2	2498	0,178
3	919	0,065
4	241	0,017
5	130	0,009
TOTAL	14119	1

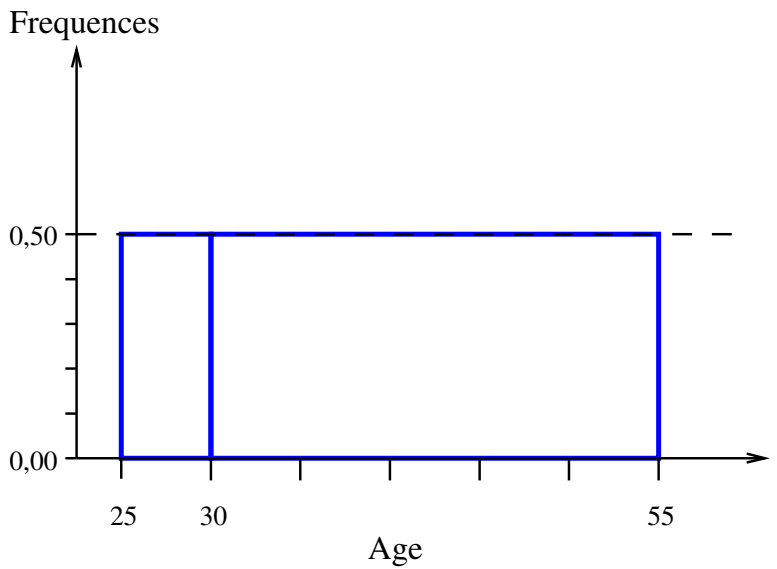


Variable continue : histogramme

Exemple introductif :

Classe d'âge	Effectifs n_i	Fréquences f_i
[25, 30[	25	0,5
[30, 55[	25	0,5
Total	50	1

Figure en "trompe l'œil" :

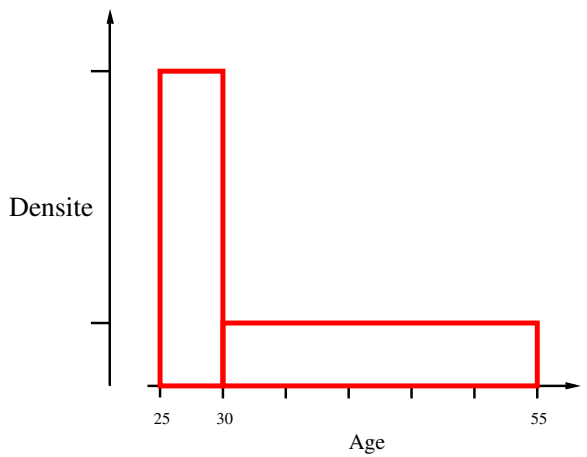


Variable continue : histogramme

Exemple introductif :

Classe d'âge	Effectifs n_i	Fréquences f_i	Densités $n_i/(e_{i+1} - e_i)$
$[25, 30[$	25	0,5	5
$[30, 55[$	25	0,5	1
Total	50	1	

Histogramme : effectifs $\Leftrightarrow$ aires



Tendance centrale - Dispersion

- Evident : réservé aux variables quantitatives
- Tendance centrale :
  moyenne, médiane (quartiles), mode
- Dispersion :
  variance, écart-type, écarts inter-quartiles

Tendance centrale

Comment définir le centre ?

- Milieu (moitié avant, moitié après) : Médiane
- Centre de gravité : Moyenne
- Observation la plus fréquente : Mode

La médiane

Définition : La médiane est une valeur possible de la variable telle qu'au moins la moitié des observations lui sont supérieures ou égales et au moins la moitié des observations lui sont inférieures ou égales

Exemple

Pays	Superficie (milliers de km^2)
Luxembourg	3,00
Belgique	30,5
Pays-Bas	33,9
Suisse	41,3
Danemark	43,1
Irlande	70,3
Autriche	83,8
Portugal	92,1
Islande	103
Grèce	132
Royaume-Uni	244
Italie	301
Norvège	324
Finlande	337
Allemagne	357
Suéde	450
Espagne	505
France	552

$$\text{Médiane} = \frac{103 + 132}{2} = 117.5$$

Variable discrète

Nombre d'enfants de 0 à 16 ans par famille	Fréquences	Fréq. cumulées
0	0,505	0,505
1	0,227	0,732
2	0,178	0,91
3	0,065	0,975
4	0,017	0,992
5	0,009	1

$$M = 0$$

Variable continue agrégée

Lorsque l'on ne dispose que d'intervalles qui contiennent les valeurs on utilise la définition suivante :
Soit la fonction cumulative

$$\begin{cases} \mathbb{R} & \to & [0,1] \\ x & \mapsto & F(x) = \text{ proportion d'observations } \leq x \end{cases}$$

La médiane M est la solution de

$$F(M) = 0.5$$

Répartition de l'âge des hommes

Age	Fréquences (%)	Fréq. cumulées (%)
De 15 à moins de 20 ans	5,8	5,8
De 20 à moins de 30 ans	24,8	30,6
De 30 à moins de 40 ans	20,5	51,1
De 40 à moins de 50 ans	14,8	65,9
De 50 à moins de 60 ans	14,2	80,1
De 60 à moins de 70 ans	10,7	90,8
De 70 à moins de 95 ans	9,2	100

$F(x) = 0.5$ pour $x \in [39, 40[$

Plus précisément $F(x) = 0.5$ pour

$$x = 30 + \frac{50 - 30.6}{51.1 - 30.6} \times (40 - 30) \approx 39.5$$

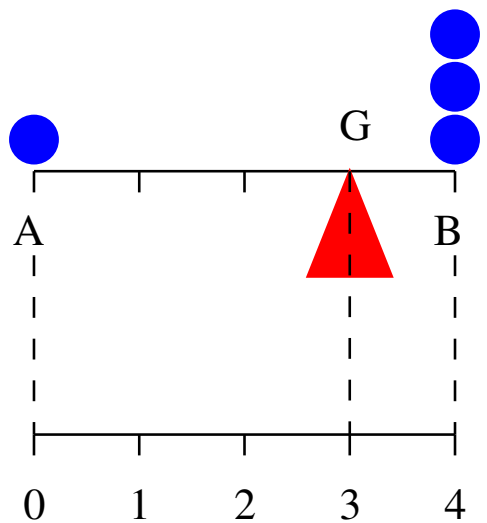
donc $M = 39.5$

La moyenne

Soit $x_1, \ldots, x_n$ les observations de X. La moyenne est

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

Exemple : $x_1 = 0$, $x_2 = x_3 = x_4 = 4$



La moyenne

Nombre d'enfants par famille :

Nombre d'enfants de 0 à 16 ans par famille	Nombre de familles (en milliers)	Fréquences f_i
0	7130	0,505
1	3201	0,227
2	2498	0,178
3	919	0,065
4	241	0,017
5	130	0,009
TOTAL	14119	1

$$\bar{x} = \frac{7130 \times 0 + \ldots + 130 \times 5}{14119} \approx 0.9$$

Age des hommes :

Age	Fréquences (%)
De 15 à moins de 20 ans	5,8
De 20 à moins de 30 ans	24,8
De 30 à moins de 40 ans	20,5
De 40 à moins de 50 ans	14,8
De 50 à moins de 60 ans	14,2
De 60 à moins de 70 ans	10,7
De 70 à moins de 95 ans	9,2

$$\bar{x} = \frac{17.5 \times 5.8 + \ldots + 82.5 \times 9.2}{100} \approx 43.4$$

Est-ce raisonnable ?

Le Mode

Définition Le mode est la valeur la plus souvent observée

- Unicité ?
- Variable continue : intervalle modal = intervalle de plus forte densité

Mesures de dispersion

Définitions

Etendue	=	écart entre les observations extrêmes
Variance	=	dispersion autour de la moyenne
	=	Moyenne de carrés des écarts à la moyenne
Quartiles	=	Découpage en 4 de la série comme pour la médiane

Variance, écart-type

Variance = Moyenne des carrés des écarts à la moyenne

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

ou

Variance = Moyenne de carrés - carré de la moyenne

$$V = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2$$

Variance, écart-type

L'écart-type (penser "écart-typique à la moyenne") est la racine carrée de la variance :

$$\sigma = \sqrt{V}$$

L'écart-type a la même unité que la variable

Variance - Exemples

Pays	Superficie (milliers de km²)	Pays	Superficie (milliers de km²)
Luxembourg	3,00	Grèce	132
Belgique	30,5	Royaume-Uni	244
Pays-Bas	33,9	Italie	301
Suisse	41,3	Norvège	324
Danemark	43,1	Finlande	337
Irlande	70,3	Allemagne	357
Autriche	83,8	Suéde	450
Portugal	92,1	Espagne	505
Islande	103	France	552

La moyenne est $\bar{x} = 205.7$ donc

$$V = \frac{(3 - 205.7)^2 + \ldots + (522 - 205.7)^2}{18} \approx 30600$$

et

$$\sigma = \sqrt{V} \approx 175$$

Variance - Exemples

Nombre d'enfants de 0 à 16 ans par famille	Nombre de familles (en milliers)
0	7130
1	3201
2	2498
3	919
4	241
5	130
TOTAL	14119

La moyenne est $\bar{x} = 0.9$ donc

$$V = \frac{(0-0.9)^2 \times 7130 + \ldots + (5-0.9)^2 \times 130}{14119} \approx 1.2$$

et

$$\sigma \approx 1.1$$

Ecart inter-quartiles

Selon le même principe que l'on définit la médiane, on définit le 1er quartile Q_1 et le 3ème quartile Q_3 :

- Q_1 (resp. Q_3) : valeur possible de la variable telle que au moins 25% (resp. 75%) des observations lui sont inférieures ou égales et au moins 75% (resp. 25%) lui sont supérieures ou égales
- $Q_2 = M$
- L'écart inter-quartiles est $Q_3 - Q_1$

Quartiles - Exemples

Pays	Superficie (milliers de km^2)
Luxembourg	3,00
Belgique	30,5
Pays-Bas	33,9
Suisse	41,3
Danemark	43,1
Irlande	70,3
Autriche	83,8
Portugal	92,1
Islande	103
Grèce	132
Royaume-Uni	244
Italie	301
Norvège	324
Finlande	337
Allemagne	357
Suéde	450
Espagne	505
France	552

$Q_1 = 43,1$

$Q_3 = 337$

Quartiles - Variable discrète

Nombre d'enfants de 0 à 16 ans par famille	Fréquences	Fréq. cumulées
0	0,505	0,505
1	0,227	0,732
2	0,178	0,91
3	0,065	0,975
4	0,017	0,992
5	0,009	1

$$Q_1 = 0 \qquad M = 0 \qquad Q_3 = 2$$

Variable continue agrégée

Comme pour la médiane, on revient à la fonction cumulative :

$$\begin{cases} \mathbb{R} & \to & [0,1] \\ x & \mapsto & F(x) = \text{ proportion d'observations } \leq x \end{cases}$$

- Q_1 tel que $F(Q_1) = 0.25$
- $M = Q_2$ tel que $F(M) = 0.5$
- Q_3 tel que $F(Q_3) = 0.75$

Répartition de l'âge des hommes

Age	Fréquences (%)	Fréq. cumulées (%)
De 15 à moins de 20 ans	5,8	5,8
De 20 à moins de 30 ans	24,8	30,6
De 30 à moins de 40 ans	20,5	51,1
De 40 à moins de 50 ans	14,8	65,9
De 50 à moins de 60 ans	14,2	80,1
De 60 à moins de 70 ans	10,7	90,8
De 70 à moins de 95 ans	9,2	100

$F(x) = 0.25$ pour

$$x = 20 + \frac{25 - 5.8}{30.6 - 5.8} \times (30 - 20) \approx 27.7$$

donc

$$Q_1 = 27.7$$

$F(x) = 0.75$ pour

$$x = 50 + \frac{75 - 65.9}{80.1 - 65.9} \times (60 - 50) \approx 27.7$$

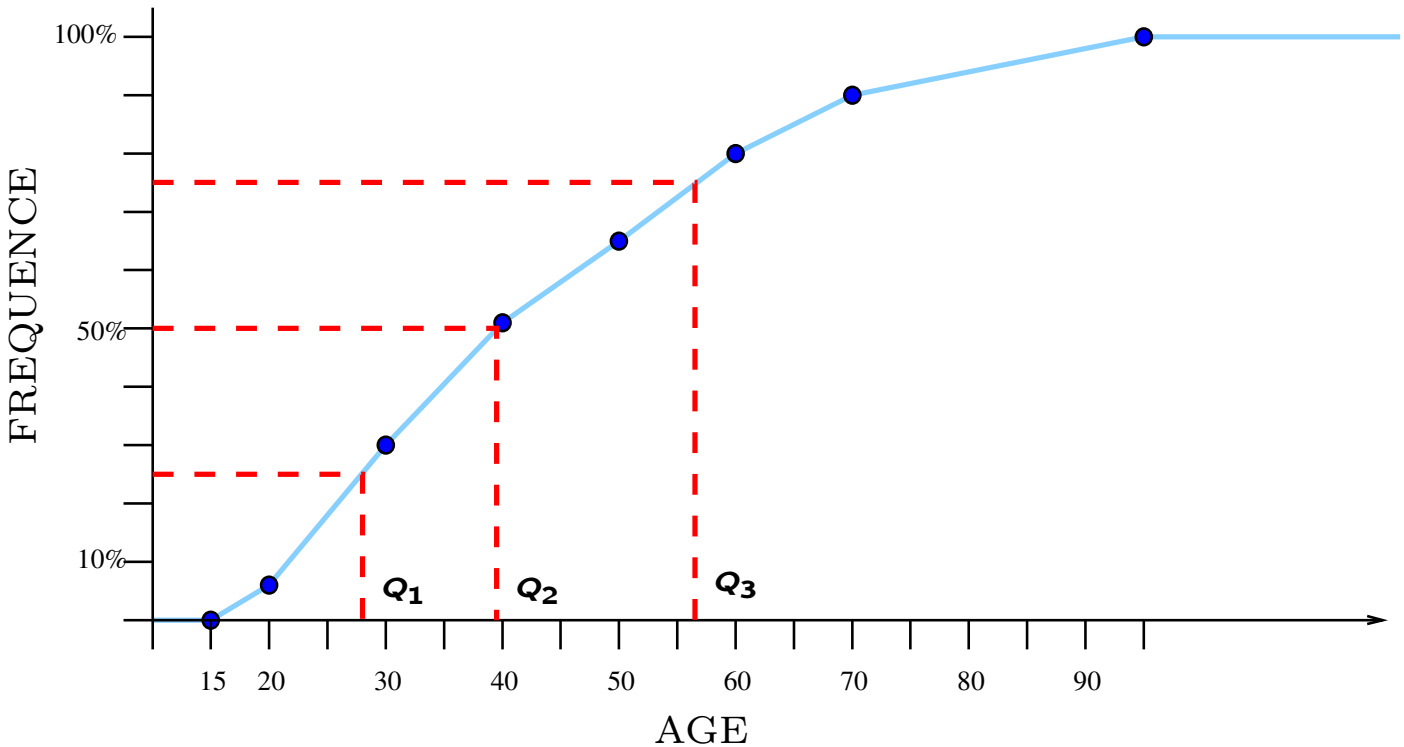
donc

$$Q_1 = 56.4$$

A partir de la courbe des fréquences cumulées

La courbe des fréquences cumulées est la courbe de la fonction cumulative F

Exemple : répartition de l'âge des hommes



Exemple

- On interroge $n = 10$ personnes
- X : sexe
- Y : fréquence de lecture d'un quotidien (trois modalités : 0 pour "ne lit jamais le journal" ; 1 pour "de temps en temps" ; 2 pour "tous les jours")

Individu	Variable 1 X	Variable 2 Y
1	H	1
2	H	1
3	F	0
4	H	2
5	F	0
6	F	1
7	F	0
8	H	0
9	F	2
10	F	1

Question : Indépendance des variables ?

Tableau de contingence

On regroupe les observations par croisements de modalités :

		Y			Total
		0	1	2	
X	F	3	2	1	6
	H	1	2	1	4
Total		4	4	2	10

Tableau de contingence : notations

	y_1	y_2	y_3	$\cdots$	y_s	Total
x_1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2s}	$n_{2\bullet}$
x_3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3s}	$n_{3\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_r	n_{r1}	n_{r2}	n_{r3}	$\cdots$	n_{rs}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	$\cdots$	$n_{\bullet s}$	n

- i : indice de ligne ; j : indice de colonne
- n_{ij} : nombre d'individus dans la $i$ème modalité de X et la $j$ème modalité de Y (effectif conjoint)
- $n_{i\bullet}$: nombre d'individus dans la $i$ème modalité de X (effectif marginal de la $i$ème modalité de X)

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$$

- $n_{\bullet j}$ est le nombre d'individus dans la $j$ème modalité de Y (effectif marginal de la $j$ème modalité de Y)

$$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$$

Distributions conditionnelles

Conditionnement par les modalités de Y : distributions conditionnelles de X

		Y			Total
		0	1	2	
X	F	$= 3/4 = 0,75$	0,5	0,5	0,6
	H	$= 1/4 = 0,25$	0,5	0,5	0,4
Total		1	1	1	1

$$f_{i|j} = f_{X=i|Y=j} = \frac{n_{ij}}{n_{\bullet j}}$$

Distributions conditionnelles

Conditionnement par les modalités de X : distributions conditionnelles de Y

		Y			Total
		0	1	2	
X	F	0,50	0,33	0,17	1
	H	0,25	0,5	0,25	1
Total		0,4	0,4	0,2	1

$$f_{j|i} = f_{Y=j|X=i} = \frac{n_{ij}}{n_{i\bullet}}$$

Distributions attendues sous l'indépendance

Indépendance : les distributions conditionnelles correspondent aux distributions marginales c'est-à-dire

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

ou

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

Ainsi l'indépendance équivaut à

$$n^{\star}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

Tableau théorique sous l'indépendance

En utilisant $n_{ij}^{\star} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$, il vient

		Y			Total
		0	1	2	
X	F	2,4	2,4	1,2	6
	H	1,6	1,6	0,8	4
Total		4	4	2	10

Ecart à l'indépendance : χ^2

Tableau réel : n_{ij}

		Y			Total
		0	1	2	
X	F	3	2	1	6
	H	1	2	1	4
Total		4	4	2	10

Tableau théorique : $n^{\star}_{ij}$

		Y			Total
		0	1	2	
X	F	2,4	2,4	1,2	6
	H	1,6	1,6	0,8	4
Total		4	4	2	10

Ecart entre les tableaux :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^{\star}_{ij})^2}{n^{\star}_{ij}}$$

Contributions au χ^2

Brutes :

		Y			Total
		0	1	2	
X	F	0,15	0,07	0,03	0,25
	H	0,225	0,1	0,05	0,375
Total		0,375	0,17	0,08	0,625

En pourcentages :

		Y			Total
		0	1	2	
X	F	0,24	0,112	0,048	0,4
	H	0,36	0,16	0,08	0,6
Total		0,6	0,272	0,128	1

Exemple : les œufs de coucou

Espèce 1		Espèce 2	Espèce 3	Espèce 4	Espèce 5	Espèce 6
19.65	22.25	21.05	20.85	21.05	21.05	19.85
20.05	22.25	21.85	21.65	21.85	21.85	20.05
20.65	22.25	22.05	22.05	22.05	21.85	20.25
20.85	22.25	22.45	22.85	22.05	21.85	20.85
21.65	22.25	22.65	23.05	22.05	22.05	20.85
21.65	22.25	23.25	23.05	22.25	22.45	20.85
21.65	22.45	23.25	23.05	22.45	22.65	21.05
21.85	22.45	23.25	23.05	22.45	23.05	21.05
21.85	22.45	23.45	23.45	22.65	23.05	21.05
21.85	22.65	23.45	23.85	23.05	23.25	21.25
22.05	22.65	23.65	23.85	23.05	23.45	21.45
22.05	22.85	23.85	23.85	23.05	24.05	22.05
22.05	22.85	24.05	24.05	23.05	24.05	22.05
22.05	23.05	24.05	25.05	23.05	24.05	22.05
22.05	23.25	24.05		23.25	24.85	22.25
22.05	23.25			23.85		
22.05	23.45					
22.05	23.65					
22.05	23.85					
22.05	24.25					
22.25	24.25					
22.25						

Questions

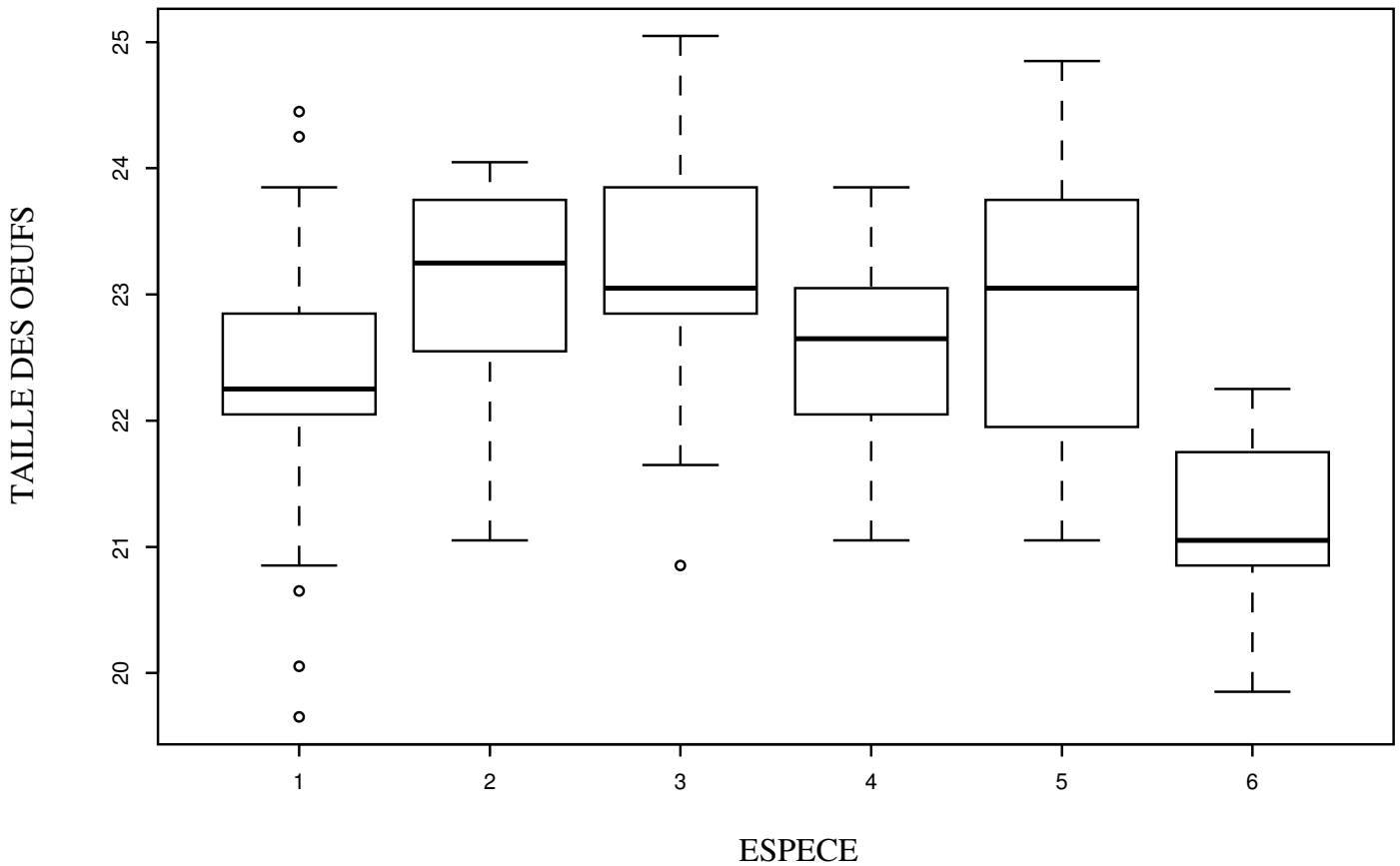
La taille des œufs diffère-t-elle selon l'espèce hôte ?

Espèce i	Effectifs n_i	Moyennes $\bar{y}_i$	Médianes M_i	Ecart-type σ_i
1	45	22,3	22,25	0,91
2	15	23,09	23,25	0,87
3	14	23,12	23,05	1,03
4	16	22,575	22,55	0,66
5	15	22,9	23,05	1,03
6	15	21,13	21,05	0,72
Total	120	22,46	22,35	1,07

Y : taille des œufs ; X : espèce hôte

La variabilité de Y est-elle expliquée par X ?

Boxplots



Décomposition de la variance

La variance totale σ^2 s'écrit

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}\right)^2 + \frac{1}{n}\sum_{i=1}^{r} n_i \sigma_i^2$$

- **Vintra** $= \frac{1}{n}\sum_{i=1}^{r} n_i \sigma_i^2$ mesure la variabilité au sein de chaque groupe
- **Vinter** $= \frac{1}{n}\sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}\right)^2$ est la variabilité expliquée par X

Rapport de corrélation

Le rapport de corrélation mesure la part de variabilité expliquée par la variable qualitative :

$$\eta^2 = \frac{\textbf{Vinter}}{\sigma^2} = \frac{\frac{1}{n}\sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}\right)^2}{\sigma^2}$$

Dans l'exemple : L'espèce hôte explique 31% de la variabilité des œufs de coucous :

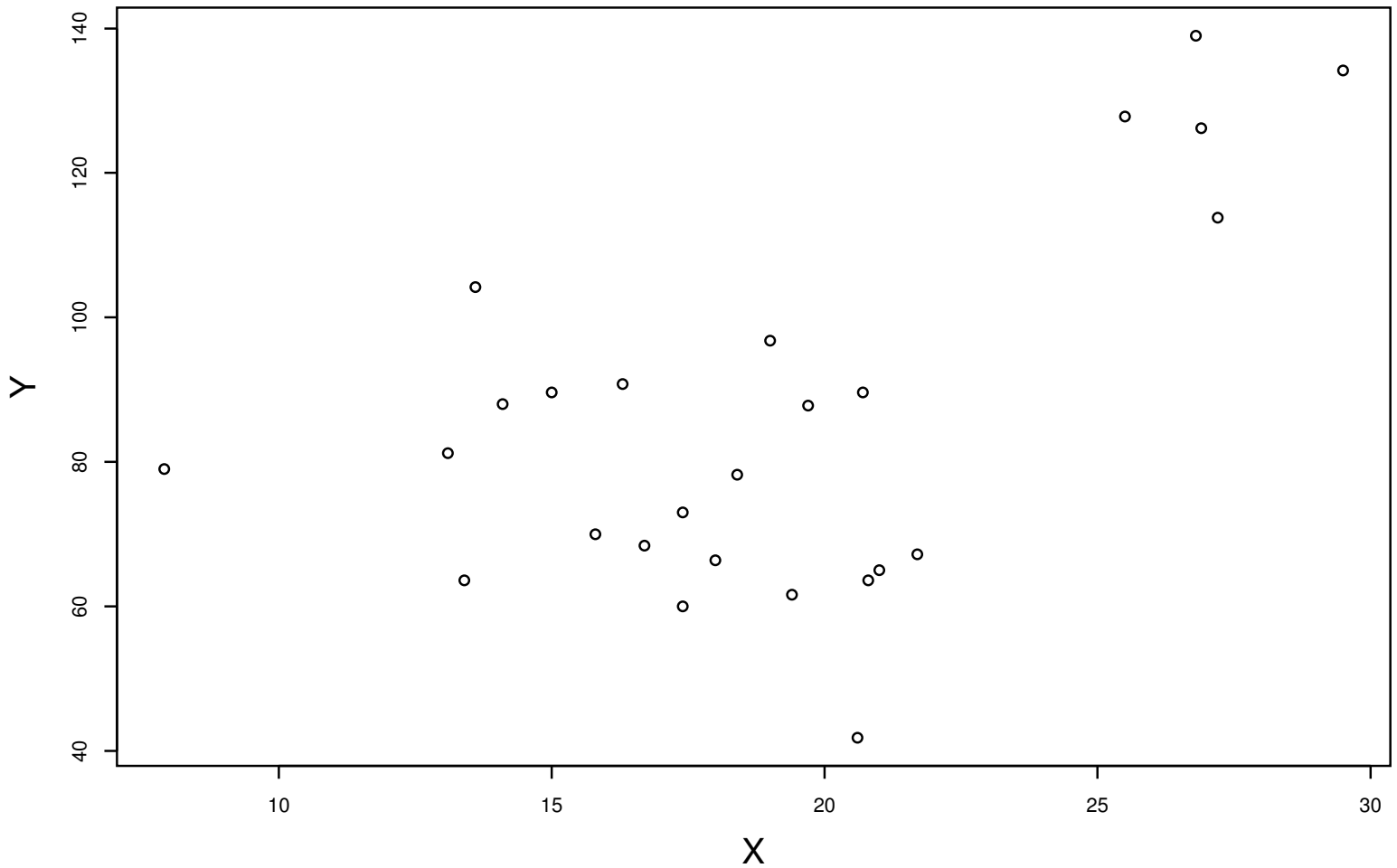
$$\eta^2 = 0.31$$

Exemple

identifiant de la mesure	X Température (Celsius)	Y Teneur en O_3 (μg/ml)
1	13,4	63,6
2	15,0	89,6
3	7,9	79,0
4	13,1	81,2
5	14,1	88,0
6	16,7	68,4
7	26,8	139,0
8	18,4	78,2
9	27,2	113,8
10	20,6	41,8
11	21,0	65,0
12	17,4	73,0
13	26,9	126,2
14	25,5	127,8
15	19,4	61,6
16	20,8	63,6
17	29,5	134,2
18	21,7	67,2
19	19,7	87,8
20	19,0	96,8
21	20,7	89,6
22	18,0	66,4
23	17,4	60,0
24	16,3	90,8
25	13,6	104,2
26	15,8	70,0

La température explique-t-elle la pollution de l'air ?

Nuage de points



Modèle linéaire

- Modelisation : On cherche $f : \mathbb{R} \to \mathbb{R}$ telle que $Y \approx f(X)$
- Linéaire : on suppose l'existence de réels a et b et d'une variable aléatoire ε tels que

$$Y = aX + b + \varepsilon$$

- Les paramètres a et b du modèle sont inconnus : on utilise les données pour les estimer

Estimateur des moindres carrés

La droite la plus proche des points (au sens de la mesure quadratique) s'obtient en minimisant

$$S(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

La solution est

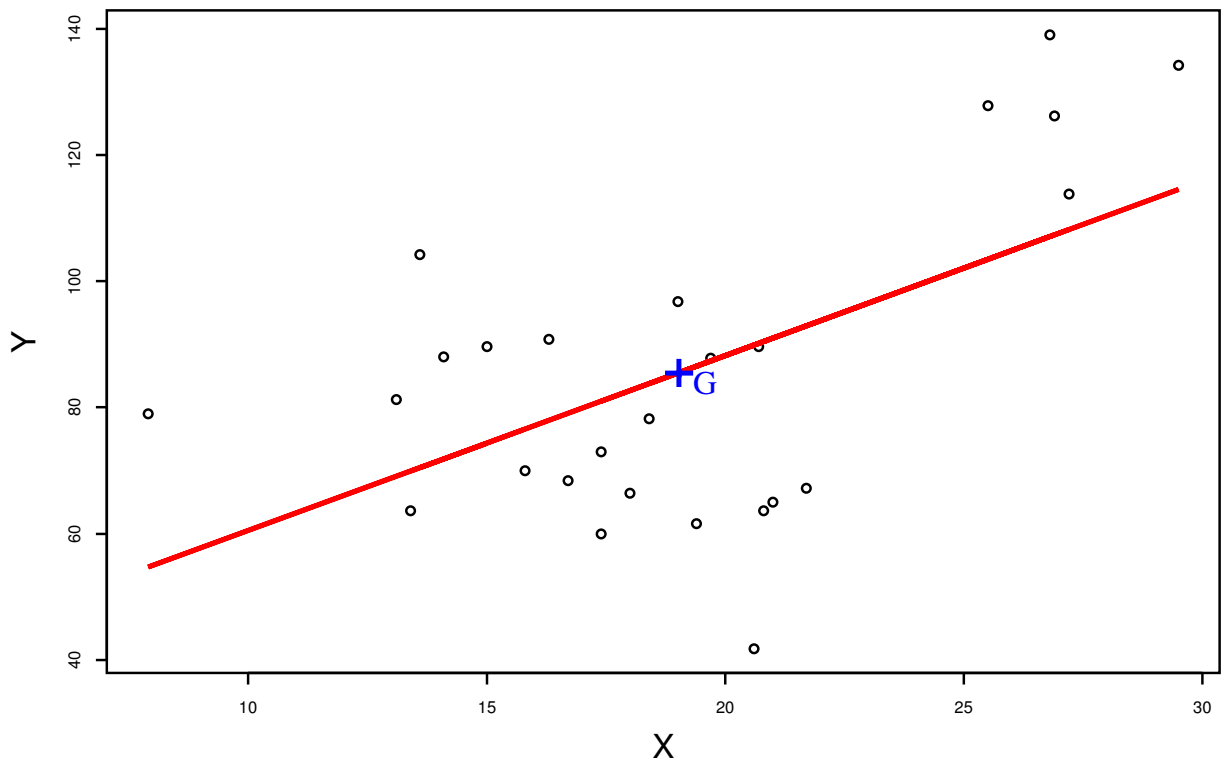
$$\hat{a} = \frac{\mathbf{cov}(X,Y)}{\sigma_X^2} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

ou

$$\mathbf{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

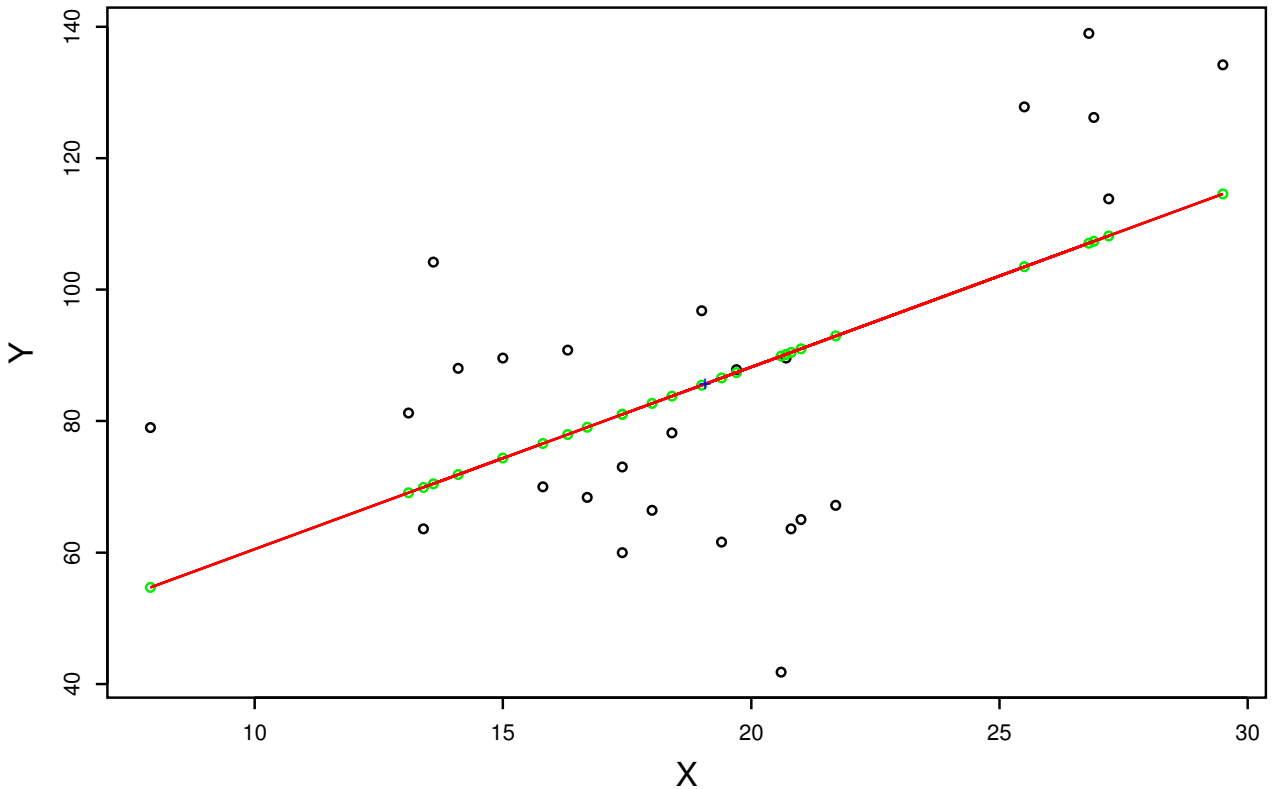
Droite des moindres carrés

Dans l'exemple : $\hat{a} = 2.8$ et $\hat{b} = 32.8$



Ajustement

Points ajustés : $(x_i; \hat{y}_i = \hat{a}x_i + \hat{b})$



Mesure de la qualité d'ajustement : le R^2

Rapport de la variance des valeurs ajustées à la variance des observations de Y :

$$R^2 = \frac{\mathbf{var}(\hat{Y})}{\mathbf{var}(Y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Dans l'exemple : le modèle explique 31% de la variabilité de Y

$$R^2 = 0.31$$