Chapitre 4

Statistiques Bivariées

Tests de distribution (khi2)

Régression linéaire & Corrélation

Cédric Wolf [a]

[a]Université de Rennes I, Unité Mixte de Recherche « ECOBIO »

Test du khi2 : Exemples

	Age	Mois	Genre
1	18	Décembre	Féminin
2	19	Juillet	Féminin
3	21	Décembre	Féminin
4	20	Juin	Féminin
5	19	Septembre	Féminin
6	19	Mai	Féminin
7	19	Janvier	Féminin
8	21	Août	Féminin
9	19	Janvier	Féminin
10	20	Septembre	Masculin

Q : Les naissances sont-elles réparties de façon homogène sur les 12 mois de l'année ?

→ Comparer la répartition observée avec une répartition théorique de 1/12 en janvier, 1/12 en février, etc...

Q : Une étude précédente à montré qu'il y a en licence de biologie 60 % de filles. Était-ce le cas en L2 BO cette année ?

→ Comparer la répartition observée avec une répartition théorique avec 60% de filles (et donc 40 % de garçons)

Test du khi2 de conformité / d'ajustement

Test du khi2 d'ajustement (ou conformité)

Cadre : On considère une VA qualitative X, et on souhaite savoir si elle est distribuée de façon conforme à une distribution théorique

Distribution (ou **profil**) de X = la répartition dans les différentes modalités

Table de contingence

Modalité	x_1	x_2	...	x_i	...	x_n
Effectif observé	N_1	N_2	...	N_i	...	N_n

H0 : La distribution observée correspond à ce qui est attendu (théorique)

Modalité	x_1	x_2	...	x_i	...	x_n
Effectif observé	N_1	N_2	...	N_i	...	N_n
Effectif théorique	$\hat{N}_1$	$\hat{N}_2$	...	$\hat{N}_i$	...	$\hat{N}_n$

$\longrightarrow$ Total = N

$\longrightarrow$ Total = N

Si l'attendu est connu sous forme de proportions (ou de pourcentages) :
Effectif théorique = proportion théorique * N

Test du khi2 : Exemples

	Age	Mois	Genre
1	18	Décembre	Féminin
2	19	Juillet	Féminin
3	21	Décembre	Féminin
4	20	Juin	Féminin
5	19	Septembre	Féminin
6	19	Mai	Féminin
7	19	Janvier	Féminin
8	21	Août	Féminin
9	19	Janvier	Féminin
10	20	Septembre	Masculin

Q : Les naissances sont-elles réparties de façon homogène sur les 12 mois de l'année ?

→ Comparer la répartition observée avec une répartition théorique de 1/12 en janvier, 1/12 en février, etc...

H0 : Les naissances sont équi-réparties sur les 12 mois de l'année

Table de contingence :

Effectifs des mois de naissance observés : (total =101)

Août	Avril	Décembre	Février	Janvier	Juillet	Juin	Mai	Mars	Novembre	Octobre	Septembre
6	11	6	10	11	7	13	13	7	2	10	5

Effectifs théorique des mois de naissance (Si H0 vraie : 101/12 pour chaque mois)

Août	Avril	Décembre	Février	Janvier	Juillet	Juin	Mai	Mars	Novembre	Octobre	Septembre
8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42

Test du khi2 d'ajustement (ou conformité)

Modalité	x_1	x_2	...	x_i	...	x_n
Effectif observé	N_1	N_2	...	N_i	...	N_n
Effectif théorique	$\hat{N}_1$	$\hat{N}_2$	...	$\hat{N}_i$	...	$\hat{N}_n$

H0 : La distribution observée correspond à ce qui est attendu(théorique)

Comment comparer les N_i et les $\hat{N}_i$?

Idée 1 : $\sum_i (N_i - \hat{N}_i) \longrightarrow$ Pb : ça fait 0

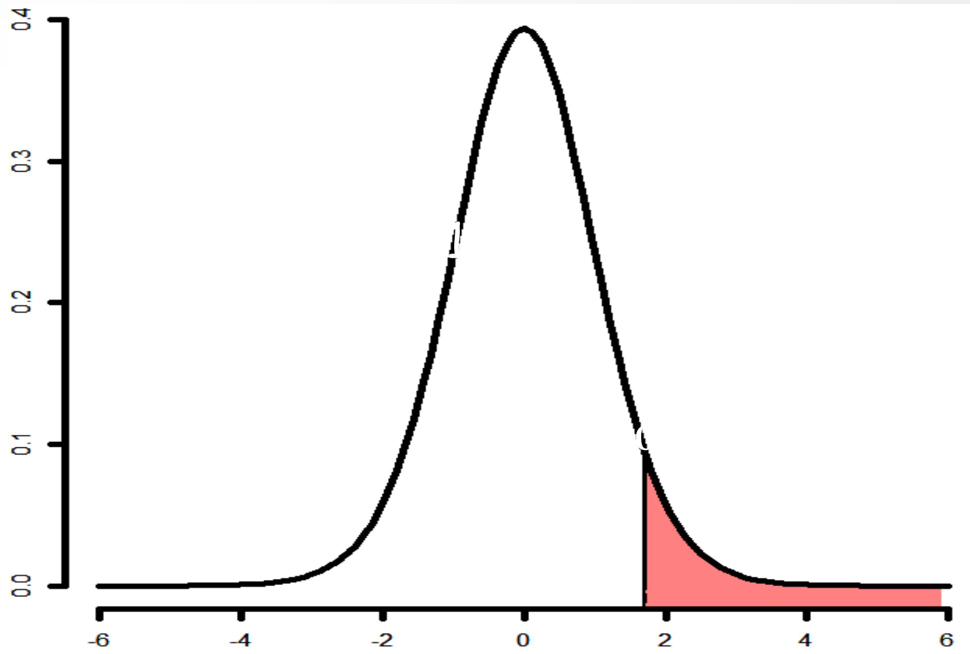
Idée 2 : $\sum_i |N_i - \hat{N}_i| \longrightarrow$ Pb : même poids pour |0-1| et |10000-9999|

Idée 3 : $\sum_i \frac{|N_i - \hat{N}_i|}{\hat{N}_i} \longrightarrow$ Pb : la distribution théorique de cette statistique est inconnue : inexploitable...

D'où le test : $\boxed{X^2 = \sum_i \frac{(N_i - \hat{N}_i)^2}{\hat{N}_i}}$ À comparer avec $X^2_{n-1;\alpha}$ (loi du khi2)

Table du khi2

dl	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.1}$	$\chi^2_{0.9}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
1	.0000	.0002	.0010	.0039	.0158	2.706	3.841	5.024	6.635	7.879
2	.0100	.0201	.0506	.1026	.2107	4.605	5.991	7.378	9.210	10.60
3	.0717	.1148	.2158	.3518	.5844	6.251	7.815	9.348	11.34	12.84
4	.2070	.2971	.4844	.7107	1.064	7.779	9.488	11.14	13.28	14.86
5	.4117	.5543	.8312	1.145	1.610	9.236	11.07	12.83	15.09	16.75
6	.6757	.8721	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	.9893	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	32.80
16	5.142	5.812	6.908	7.962	9.312	23.54	26.30	28.85	32.00	34.27
17	5.697	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.434	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57	40.00
22	8.643	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
24	9.886	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2
120	83.85	86.92	91.57	95.70	100.6	140.2	146.6	152.2	159.0	163.6



α : Probabilité de dépasser $\chi_{\nu;\alpha}$

ν : Degré de liberté (dl)

Ex : Pour un dl 11 et un seuil de 0.05, on aura :

$$\chi^2_{11\,;0.05} = 19.68$$

Test du khi2 : Exemples

H0 : Les naissances sont équi-réparties sur les 12 mois de l'année

Table de contingence :

Effectifs des mois de naissance observés : (total =101)

Août	Avril	Décembre	Février	Janvier	Juillet	Juin	Mai	Mars	Novembre	Octobre	Septembre
6	11	6	10	11	7	13	13	7	2	10	5

Effectifs théorique des mois de naissance (Si H0 vraie : 101/12 pour chaque mois)

Août	Avril	Décembre	Février	Janvier	Juillet	Juin	Mai	Mars	Novembre	Octobre	Septembre
8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42	8.42

D'où :

$$\chi^2 = \frac{(6-8.42)^2}{8.42} + \frac{(11-8.42)^2}{8.42} + \frac{(6-8.42)^2}{8.42} + \cdots + \frac{(5-8.42)^2}{8.42} = 15.32$$

À comparer avec $\chi^2_{11;0.05}$

$\chi^2 < \chi^2_{11;0.05} = 19.68$ Donc test non significatif

La répartition observée des mois de naissances ne montre pas de différence significative avec une équipartition

Test du khi2 : Exemples

	Age	Mois	Genre
1	18	Décembre	Féminin
2	19	Juillet	Féminin
3	21	Décembre	Féminin
4	20	Juin	Féminin
5	19	Septembre	Féminin
6	19	Mai	Féminin
7	19	Janvier	Féminin
8	21	Août	Féminin
9	19	Janvier	Féminin
10	20	Septembre	Masculin

Q : Une étude précédent à montré qu'il y a en licence de biologie 60 % de filles. Est-ce le cas en L2 BO cette année ?

→ Comparer la répartition observée avec une répartition théorique avec 60% de filles

H0 : La répartition observée correspond à 60% de filles

Table de contingence :

Féminin	Masculin
64	37

Ne pas oublier les garçons !
On compare toute la répartition

Effectifs théorique

Féminin	Masculin
60.6	40.4

$$\chi^2 = \frac{(64-60.6)^2}{60.6} + \frac{(37-40.4)^2}{40.4} = 0.48$$

À comparer à $\chi^2_{1;0.05} = 3.84$

Donc test non significatif

La L2 BO 2014 est en cohérence avec une répartition comportant 60% de filles

Test du khi2 : Exemples

	Age	Mois	Genre
1	18	Décembre	Féminin
2	19	Juillet	Féminin
3	21	Décembre	Féminin
4	20	Juin	Féminin
5	19	Septembre	Féminin
6	19	Mai	Féminin
7	19	Janvier	Féminin
8	21	Août	Féminin
9	19	Janvier	Féminin
10	20	Septembre	Masculin

Q : La répartition en âges est-elle la même pour les filles et les garçons ?

→ Comparer les répartitions observées respectivement pour les filles et les garçons

Test du khi2 d'homogénéité / d'indépendance

Rem : C'est le même test (et même résultat) que si l'on se demandait si la répartition en genre est la même pour les différents âges (symétrie des 2 variables considérées)

Test du khi2 d'indépendance (ou d'homogénéité)

Cadre : Une VA qualitative Y à expliquer par une VA qualitative X

H0 : Il y a indépendance (ou homogénéité), c'est à dire que les distributions de chacune des modalités Y selon les modalités de X sont identiques
(ou vice-versa ; c'est symétrique !)

L'idée est de calculer des effectifs théoriques pour chaque modalité croisée Si H0 est vraie, puis de regarder la conformité de l'ensemble des données observées à ces données théoriques

<u>Etape 1</u>: Etablir le tableau de contingence des effectifs observés

	X_1	X_2	...	X_j	...	X_n
Y_1	N_{11}	N_{12}	$\cdots$	N_{1j}	$\cdots$	N_{1n}
Y_2	N_{21}	N_{22}	$\cdots$	N_{2j}	$\cdots$	N_{2n}
...	$\vdots$	$\vdots$				$\vdots$
Y_i	N_{i1}	N_{i2}	$\cdots$	N_{ij}	$\cdots$	N_{in}
...	$\vdots$	$\vdots$				$\vdots$
Y_p	N_{p1}	N_{p2}	$\cdots$	N_{pj}	$\cdots$	N_{pn}

Nij : Effectif de la modalité Yi pour Y de de modalité Xj pour X

Test du khi2 d'indépendance (ou d'homogénéité)

Cadre : Une VA qualitative Y à expliquer par une VA qualitative X

H0 : Il y a indépendance (ou homogénéité), c'est à dire que les distributions de chacune des modalités Y selon les modalités de X sont identiques
(ou vice-versa ; c'est symétrique !)

L'idée est de calculer des effectifs théoriques pour chaque modalité croisée Si H0 est vraie, puis de regarder la conformité de l'ensemble des données observées à ces données théoriques

Etape 1bis: Tableau de contingence des effectifs observés puis calculer les marges

N_{ij}	X_1	X_2	...	X_j	...	X_n	marge
Y_1	N_{11}	N_{12}	$\cdots$	N_{1j}	$\cdots$	N_{1n}	N1.
Y_2	N_{21}	N_{22}	$\cdots$	N_{2j}	$\cdots$	N_{2n}	N2.
...	$\vdots$	$\vdots$				$\vdots$	
Y_i	N_{i1}	N_{i2}	$\cdots$	N_{ij}	$\cdots$	N_{in}	Ni.
...	$\vdots$	$\vdots$				$\vdots$	
Y_p	N_{p1}	N_{p2}	$\cdots$	N_{pj}	$\cdots$	N_{pn}	Np.
marge	N.1	N.2		N.j		N.n	N

(Somme des lignes et colonnes)

Nij : Effectif de la modalité Yi pour Y de de modalité Xj pour X

Ni. : Somme des effectifs de la ligne i

N.j : Somme des effectifs de la colonne j

Test du khi2 : Exemples

	Age	Mois	Genre
1	18	Décembre	Féminin
2	19	Juillet	Féminin
3	21	Décembre	Féminin
4	20	Juin	Féminin
5	19	Septembre	Féminin
6	19	Mai	Féminin
7	19	Janvier	Féminin
8	21	Août	Féminin
9	19	Janvier	Féminin
10	20	Septembre	Masculin

Q : La répartition en âges est-elle la même pour les filles et les garçons ?

Comparer les répartitions observées respectivement pour les filles et les garçons

Test du khi2 d'homogénéité / d'indépendance

H0 : Les répartitions des Filles et des Garçons en âge sont les mêmes
(<=> Les sex-ratio des différents âges sont identiques)
(<=> il y a indépendance de la variable « Genre » avec la variable « Age »)

Table de contingence :

	18	19	20	21&+	Σ
Fille	5	39	14	6	64
Garçon	3	14	11	9	37
Σ	8	53	25	15	N=101

Test du khi2 d'indépendance (ou d'homogénéité)

Etape 2 : Déterminer les effectifs théoriques Si H0 est vraie

	X_1	X_2	...	X_j	...	X_n	marge
Y_1							N1.
Y_2							N2.
...							
Y_i							Ni.
...							
Y_p							Np.
marge	N.1	N.2		N.j		N.n	N

S'il y avait indépendance des deux variables, quels seraient les effectifs dans chaque « case » - en supposant que les effectifs des différentes modalités (donc les N.j et les Ni.) restent identiques ?

Test du khi2 d'indépendance (ou d'homogénéité)

<u>Etape 2</u> : Déterminer les effectifs théoriques Si H0 est vraie

H0 : Les répartitions des Filles et des Garçons en âge sont les mêmes
(<=> Les sex-ratio des différents âges sont identiques)
(<=> il y a indépendance de la variable « Genre » avec la variable « Age »

Table de contingence :

	18	19	20	21&+	Σ
Fille	5	39	14	6	64
Garçon	3	14	11	9	37
Σ	8	53	25	15	N=101

Si H0 est vraie, la proportion de fille de 18 ans serait la même que la proportion globale de filles
Donc la proportion théorique de fille de 18 ans devrait être de 64/101 = 0.63

Puisqu'il y a 8 individus de 18 ans, il y aurait donc 0.63 * 8 = 5.07 filles

De même pour chaque âge et de même pour les garçons

Test du khi2 : Exemples

Table de contingence :

N_{ij}	18	19	20	21&+	Σ
Fille	5	39	14	6	64
Garçon	3	14	11	9	37
Σ	8	53	25	15	N=101

Tableau des effectifs théoriques

$\hat{N}_{ij}$	18	19	20	21&+	Σ
Fille	$\frac{8\times64}{101}$	$\frac{53\times64}{101}$	$\frac{25\times64}{101}$	$\frac{15\times64}{101}$	64
Garçon	$\frac{8\times37}{101}$	$\frac{53\times37}{101}$	$\frac{25\times37}{101}$	$\frac{15\times37}{101}$	37
Σ	8	53	25	15	N=101

$\hat{N}_{ij}$	18	19	20	21&+
Fille	5.07	33.58	15.84	9 .50
Garçon	2.93	19.42	9.16	5.50

Test du khi2 d'indépendance (ou d'homogénéité)

Ni. : Somme des effectifs de la ligne i

N.j : Somme des effectifs de la colonne j

Si H0 est vérifiée, la proportion de la modalité Xj est la même pour chaque modalité Yi ; et donc la même que la proportion globale, soit N.j /N

L'effectif de la modalité (Yi,Xj) sera donc (N.j /N) * Ni.

On définit ainsi le tableau de contingence des effectifs théoriques par (les marges sont conservées)

$$\hat{N}_{ij} = \frac{N_{i.} N_{.j}}{N}$$

$\hat{N}_{ij}$	X_1	X_2	...	X_j	...	X_n	marge
Y_1	$\hat{N}_{11}$	$\hat{N}_{12}$	$\cdots$	$\hat{N}_{1j}$	$\cdots$	$\hat{N}_{1n}$	N1.
Y_2	$\hat{N}_{21}$	$\hat{N}_{22}$	$\cdots$	$\hat{N}_{2j}$	$\cdots$	$\hat{N}_{2n}$	N2.
...	$\vdots$	$\vdots$				$\vdots$	
Y_i	$\hat{N}_{i1}$	$\hat{N}_{i2}$	$\cdots$	$\hat{N}_{ij}$	$\cdots$	$\hat{N}_{in}$	Ni.
...	$\vdots$	$\vdots$				$\vdots$	
Y_p	$\hat{N}_{p1}$	$\hat{N}_{p2}$	$\cdots$	$\hat{N}_{pj}$	$\cdots$	$\hat{N}_{pn}$	Np.
marge	N.1	N.2		N.j		N.n	N

D'où le test :

$$X^2 = \sum_i \sum_j \frac{\left(N_{ij} - \hat{N}_{ij}\right)^2}{\hat{N}_{ij}}$$

à comparer avec $X^2_{(n-1)(p-1);\alpha}$

Tests du khi2 : Compléments

Condition de validité du test : Règle de Cochran :

$\hat{N}_{ij} > 0$

Au moins 80% des $\hat{N}_{ij}$ sont >5

→ Sinon ? Regrouper des modalités

Un test du khi2 s'effectue toujours sur des <u>effectifs</u> (même si au final il compare des proportions)

Dans le cas d'une comparaison de 2 variables à 2 modalités (tableau de contingence 2x2), on appliquera la correction (de continuité) de Yates (pour compenser un biais liés au fait que l'on approche une loi binomiale – donc discontinue- par une loi continue -celle du khi2 -)

N_{ij}	X_1	X_2
Y_1	N_{11}	N_{12}
Y_2	N_{21}	N_{22}

$$X^2 = \sum_i \sum_j \frac{\left(|N_{ij} - \hat{N}_{ij}| - 0.5\right)^2}{\hat{N}_{ij}}$$

Test du khi2 : Exemples

Table de contingence :

N_{ij}	18	19	20	21&+
Fille	5	39	14	6
Garçon	3	14	11	9

Tableau des effectifs théoriques :

$\hat{N}_{ij}$	18	19	20	21&+
Fille	5.07	33.58	15.84	9 .50
Garçon	2.93	19.42	9.16	5.50

D'où :

$$\chi^2 = \frac{(5-5.07)^2}{5.07} + \frac{(39-33.58)^2}{33.58} + \cdots + \frac{(9-5.50)^2}{5.50} = 6.50$$

À comparer à $\chi^2_{(4-1)(2-1);0.05} = \chi^2_{3;0.05} = 7.81$

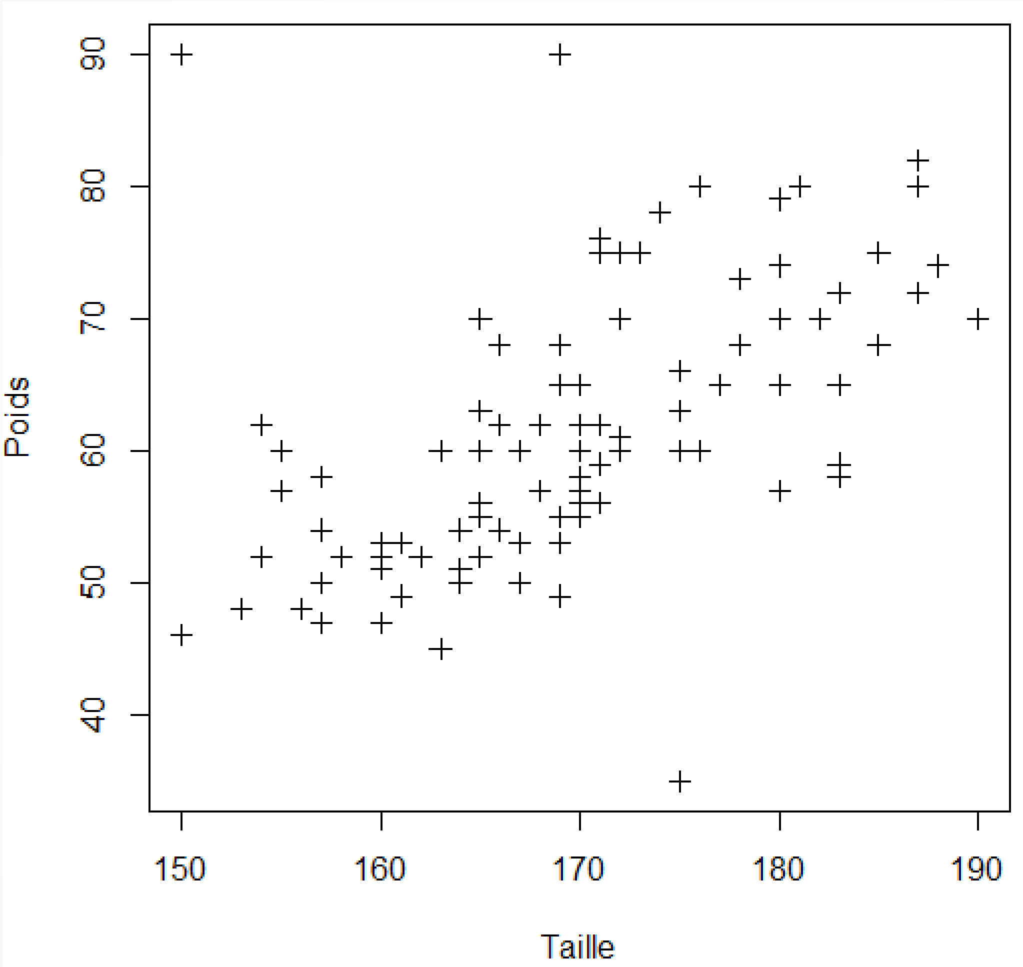
Donc test non significatif

Les répartitions des âges des filles et des garçons ne sont pas significativement différents

Rem : on s'aperçoit que l'on dépasse tout de même $\chi^2_{3;0.1} = 6.25$, donc avec un seuil d'erreur de 10 % on rejetterait l'hypothèse (R donne une p_value de 0.09) : Dans ce cas (non significatif), on parle de <u>tendance</u> statistique (pour les p_values comprises entre 0.05 et 0.1).

Corrélation et Régression linéaire : Cadre

	Taille	Poids
1	167	60
2	169	65
3	169	68
4	157	58
5	157	50
6	166	62
7	167	50
8	171	56
9	165	60
10	183	59

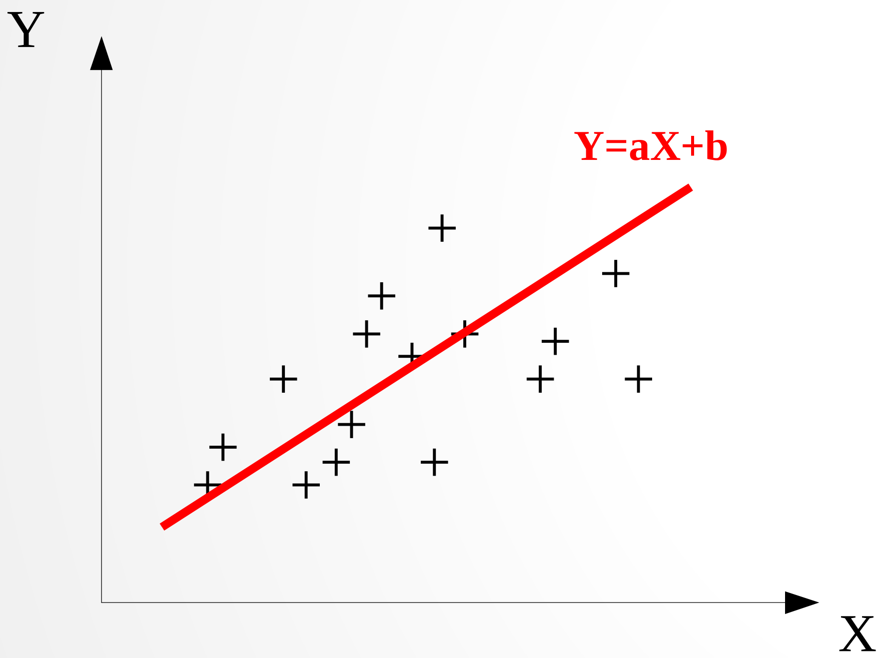


But(s) : Y-a-t-il un lien entre la Taille et le Poids ?
Quel est ce lien ?
Peux-t-on prédire le Poids en fonction de la Taille ?

Régression linéaire & Corrélation

Cadre : Une VA quantitative Y à expliquer par une VA quantitative X

But : Identifier et si possible quantifier la liaison statistique entre X et Y



Peut-on prévoir la variation de Y en fonction de celle de X ?

→ Notion de **corrélation**

Corrélation **positive** si Y augmente lorsque X augmente (et inversement)

Corrélation **négative** si Y diminue lorsque X augmente (et inversement)

Peut-on quantifier cette relation ?

→ Notion de modèle de **régression linéaire** : Modèle Y = a X +b

Régression linéaire & Corrélation : Hypothèses

Les hypothèses d'applications sont :

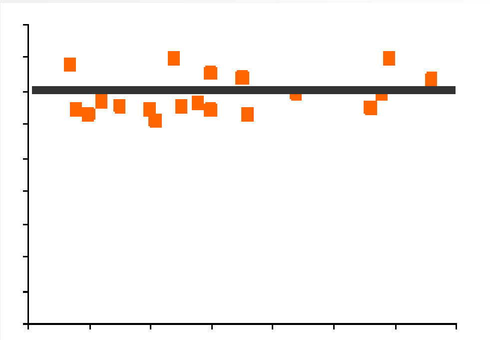
- Indépendance des observations

Contre exemple : Même individus mesurés à des dates différentes

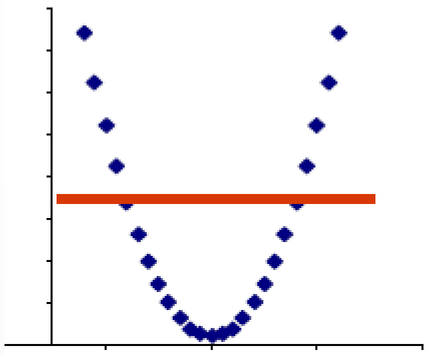
- Distribution normale et de variance identique pour chaque modalité de X (et de Y)

Difficilement vérifiable ! En pratiqe : nuage de forme pas trop “bizarre”

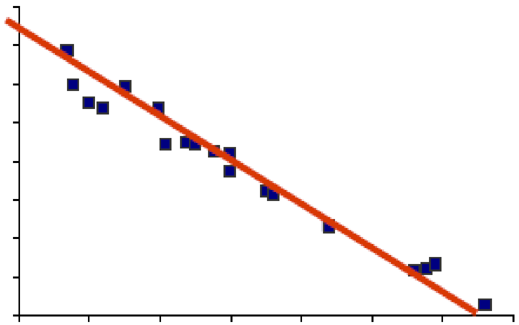
- La liaison est linéaire (vérification graphique)



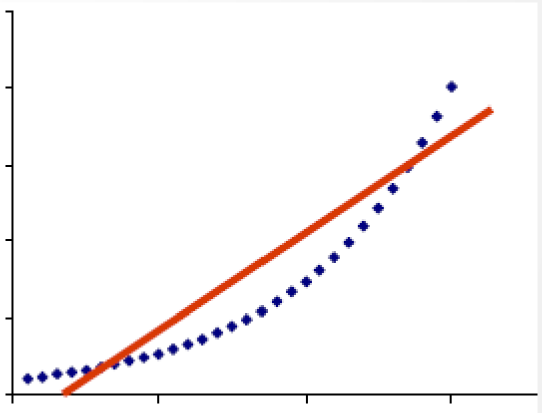
OK



Pas OK



OK



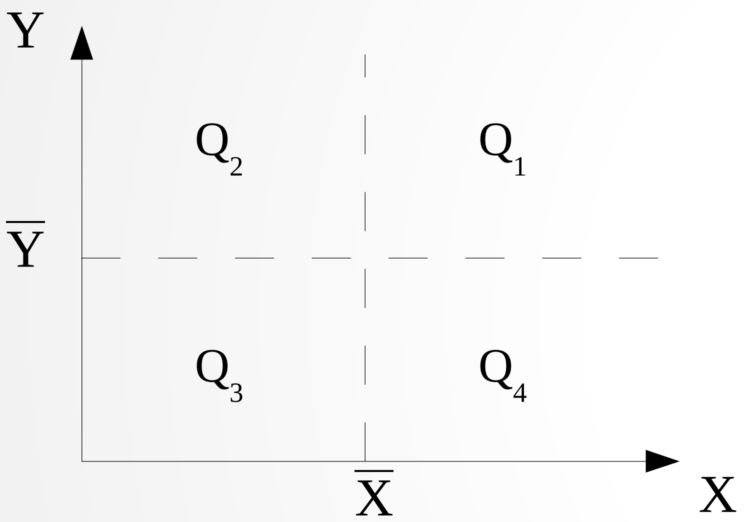
Pas OK

Cf test41() et test42() qui illustrent des situations ou les données sont théoriquement très corrélées (mais pas linéairement)

La covariance

La **covariance** entre X et Y se définit par : $s_{XY}=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\overline{X})(Y_i-\overline{Y})$

(covariance estimée pour la population, celle de l'échantillon s'obtient en remplaçant 1/(N-1) par 1/N)



	Q_1	Q_2	Q_3	Q_4
$(X_i-\overline{X})$	+	-	-	+
$(Y_i-\overline{Y})$	+	+	-	-
$(X_i-\overline{X})(Y_i-\overline{Y})$	+	-	+	-

Points plutôt dans les zones Q_1 et Q_3 : $s_{XY}>0$, covariance positive

Points plutôt dans les zones Q_2 et Q_4 : $s_{XY}>0$, covariance négative

Répartition équilibrée : $s_{XY}\approx 0$, covariance nulle – pas de lien

Propriétés :

$$s_{XX}=s_X^2$$

$$s_{X+Y}^2=s_{X-Y}^2=s_X^2+s_Y^2+2s_{XY}$$

La covariance varie de – ∞ à +∞ et dépend des unités des Vas (ex : Kg.cm)

La corrélation

La **corrélation** entre X et Y se définit par : $r = \frac{s_{XY}}{s_X \, s_Y}$

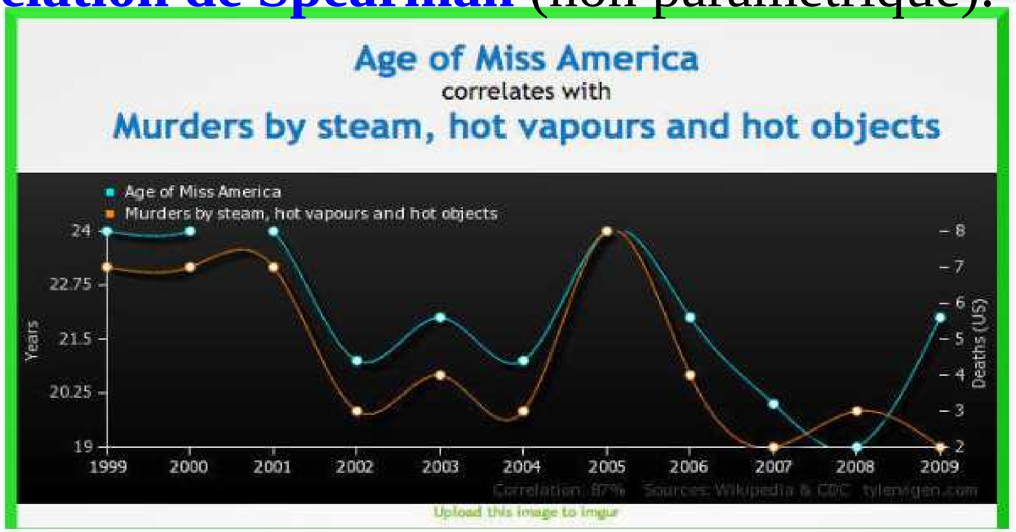
(Elle est égale pour l'échantillon et le population)

La corrélation varie entre -1 et 1, et est indépendante des unités :

- $r > 0$: Y augmente avec X
- $r < 0$: Y diminue lorsque X augmente
- $r \approx 0$: Indépendance entre X et Y
- $r = 1$ ou $r = -1$: Les points sont parfaitement alignés

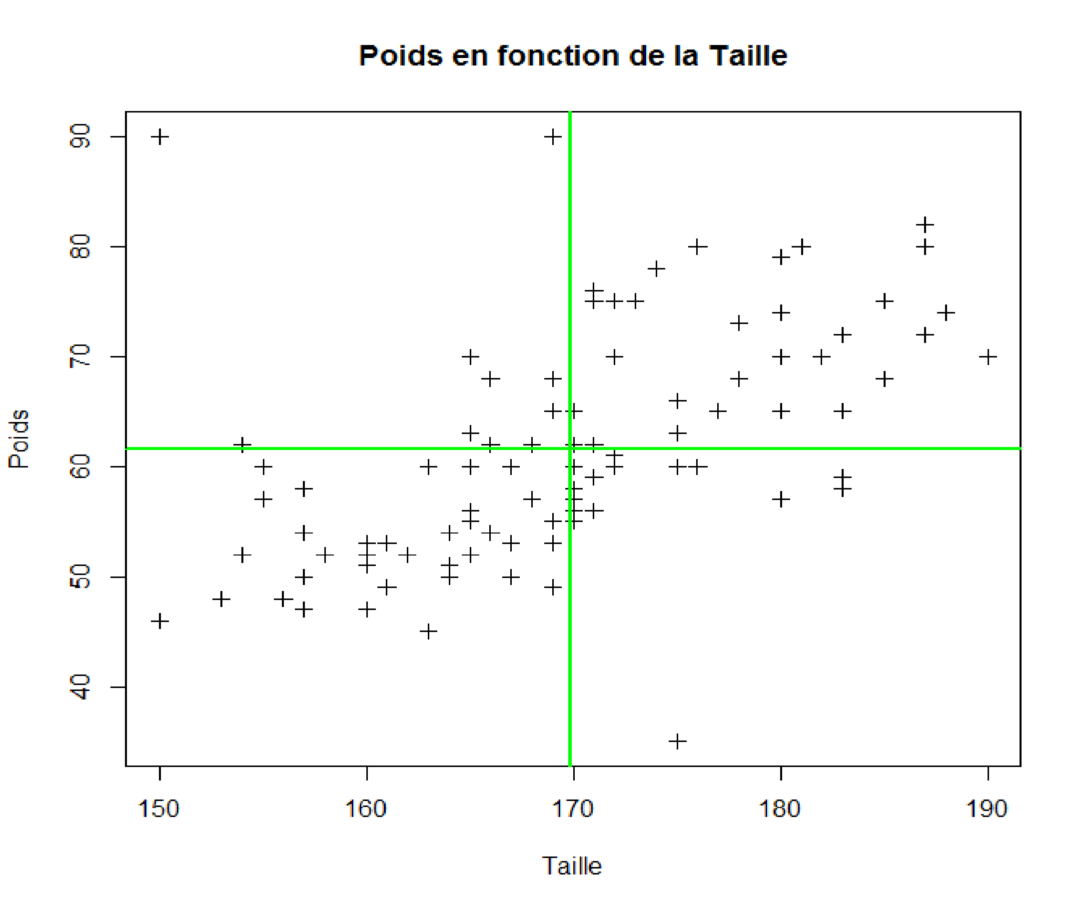
Des tests permettent de tester l'hypothèse H0 “Il y a indépendance entre X et Y” :
Test de corrélation de Pearson (paramétrique, nécessite la normalité des données et ne pas avoir de valeurs extrêmes) ou **test de corrélation de Spearman** (non paramétrique).

⚠ Un test de corrélation, même significatif, ne démontre pas un lien de cause à effet



Covariance / corrélation : exemple

	Taille	Poids
1	167	60
2	169	65
3	169	68
4	157	58
5	157	50
6	166	62
7	167	50
8	171	56
9	165	60
10	183	59



$$\bar{X} = 169.85\,cm$$

$$s_X = 9.16\,cm$$

$$\bar{Y} = 61.6\,Kg$$

$$s_Y = 10.28\,cm$$

$$s_{XY} = \frac{1}{100}\left((167-169.85)(60-61.6)+(169-169.85)(65-61.6)+\ldots\right) = 50.69$$

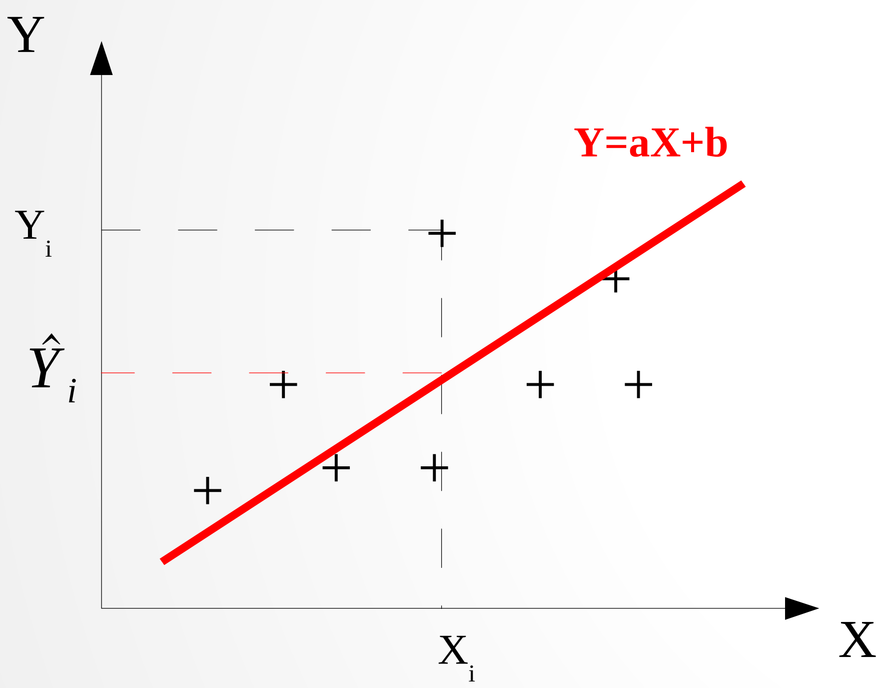
(en Kg.cm !)

(ou 506.9 g.m !)

$$r = \frac{50.69}{9.16 \times 10.28} = 0.54$$

Régression linéaire

La **droite de régression** de Y par rapport à X, d'équation Y = aX+b est la droite s'ajustant le mieux aux données par la méthode des moindres carrés



a et b déterminés de sorte que $\sum_{i=1}^{N} |Y_i - \hat{Y}_i|^2$ soit minimal, avec $\hat{Y}_i = a\,X_i + b$

Les $|Y_i - \hat{Y}_i|$ sont les **résidus** du modèle

On montre que *lorsque les résidus suivent une loi Normale*, a et b sont donnés par :

$$\boxed{a = \frac{s_{XY}}{s_X^2}} \qquad b = \overline{Y} - a\,\overline{X}$$

L'hypothèse H0 “il n'y a pas de lien linéaire entre X et Y” peut être testée au moyen d'un **test de pente nulle** (H0 est équivalent à “a=0”).
Ce test est **équivalent** à un test de corrélation de Pearson (si applicable).

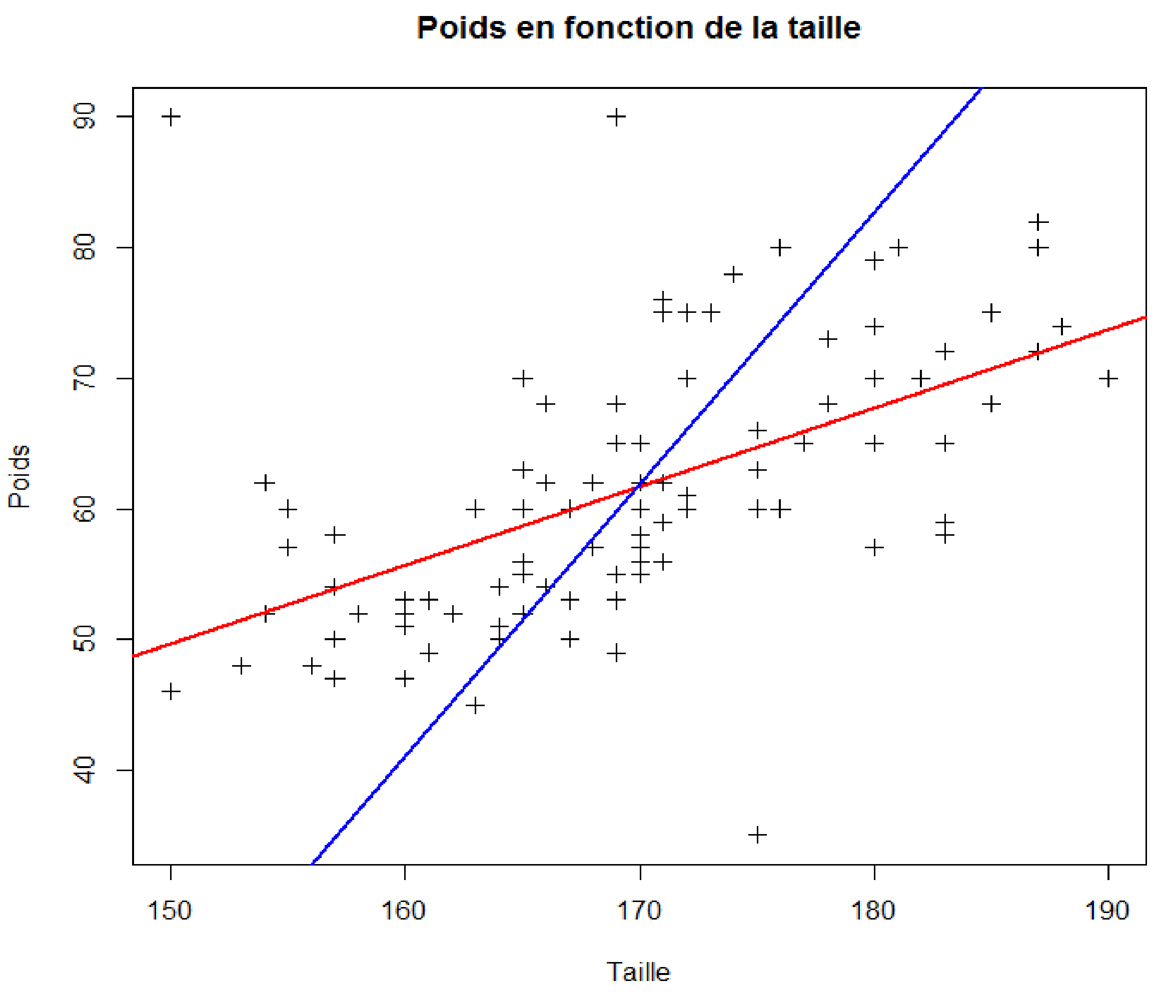
Rem : On peut également tester l'hypothèse H0 : « La pente est égale à a_0 » par : $t_p = \dfrac{|a - a_0|\sqrt{N-2}}{\sqrt{\dfrac{s_Y^2}{s_X^2} - a^2}}$

à comparer avec $t_{N-2;\alpha}$

Régression linéaire : non symétrique !

La régression linéaire de Y par rapport à X, ne donne pas le même résultat que la régression linéaire de X par rapport à Y

Contrairement à la corrélation, il n'y a pas symétrie



Régression linéaire de Y par rapport à X :

Y = aX+b

$$a = \frac{s_{XY}}{s_X^2} = \frac{50.69}{9.16^2} = 0.60$$

puis $b = \bar{Y} - a\,\bar{X} = -40.9$

Régression linéaire de X par rapport à Y :

X = a' Y+b'

$$a' = \frac{s_{XY}}{s_Y^2} = \frac{50.69}{10.28^2} = 0.48$$

puis $b' = \bar{X} - a'\,\bar{Y} = 140.3$

Corrélation et Régression linéaire : Comparatif

Corrélation	Régression linéaire : Y = a X + b
Lien entre X et Y ?	Lien entre X et Y ? + quantification
Rôle symétrique de X et Y	Rôle asymétrique de X et Y
Pas de prédiction possible	Prédiction possible

Lien entre les deux notions ?

On peut relier a et r : $a = \frac{s_{XY}}{s_X^2} = \frac{s_{XY}}{s_X s_Y} \times \frac{s_Y}{s_X} = r \frac{s_Y}{s_X}$

Indépendant du lien entre X et Y

Rem : $aa' = r^2$

a=1/a' (même droite) ⇔ $r^2=1$ ⇔ points alignés

Régression linéaire : le r²

Le coefficient de détermination, plus généralement appelé **r²** du modèle (c'est effectivement le carré du coefficient de corrélation) permet de juger de la qualité du modèle : Il représente la proportion de variance de Y qui est expliquée par le modèle (donc par la variance sur X)

Plus le r² est proche de 1, meilleur est le modèle. Un r² égal à 1 signifierait que la valeur de Y peut être exactement prédite par celle de X (sans marge d'erreur)

Dem : Si $\hat{Y}_i = a X_i + b$ alors $s_{\hat{Y}}^2 = a^2 s_X^2 = \left(r^2 \frac{s_Y^2}{s_x^2}\right) s_x^2 = r^2 s_Y^2$ car $a = r \frac{s_Y}{s_x}$

Ainsi : $s_Y^2 = r^2 s_Y^2 + (1 - r^2) s_Y^2$

Variance totale pour Y

$= s_{\hat{Y}}^2$

Variance expliquée par le modèle

Variance inexpliquée

Régression linéaire : Prédictions

Pour une valeur x_0 de la variable X, le modèle prédit pour Y la valeur $y_0 = a\ x_0 + b$

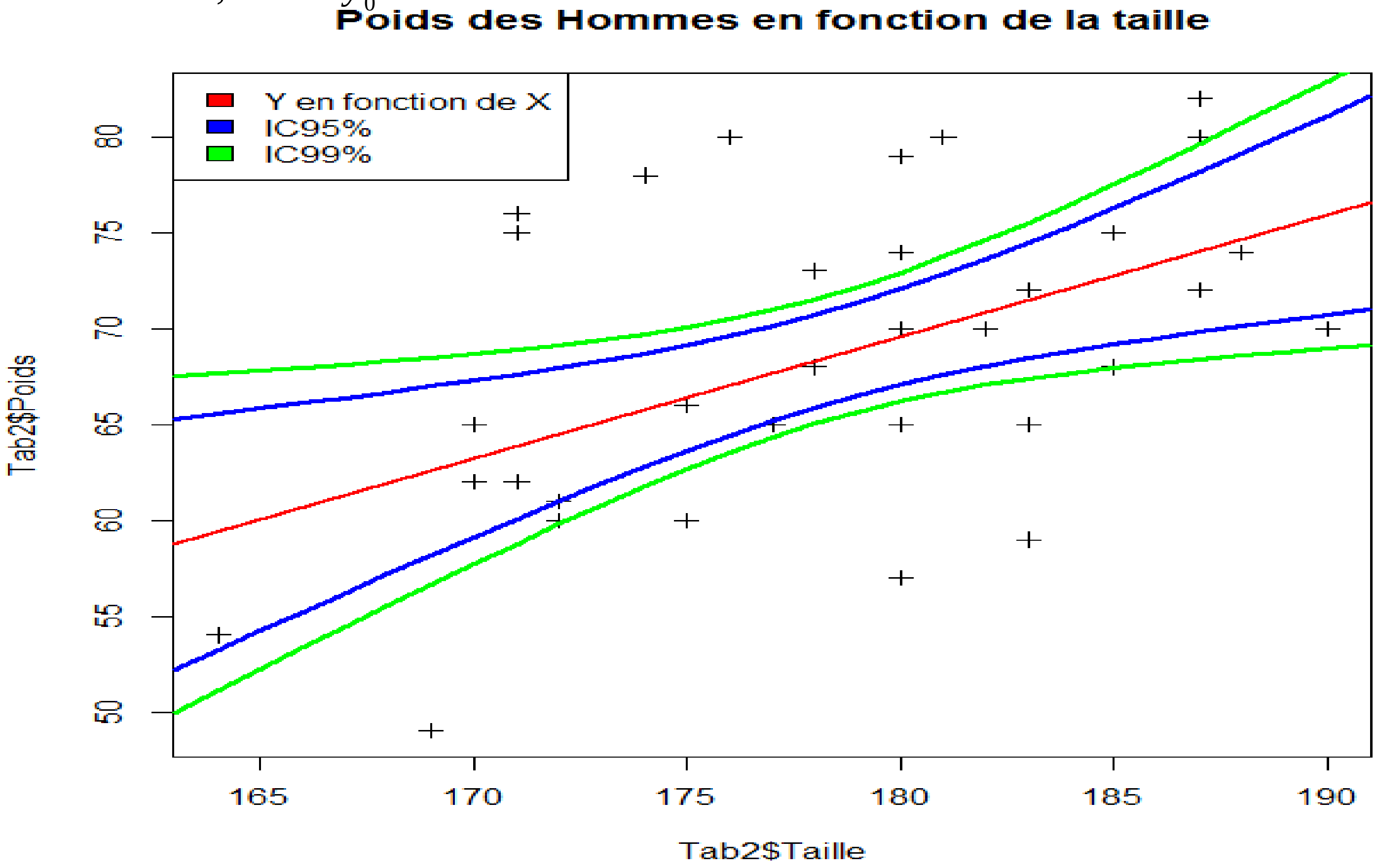
Il est possible de calculer un intervalle de confiance autour de cette valeur par :

$$IC_{95\%} = [\, y_0 - t_{N-2\,;\,0.05}\, s_{y_0}\,;\, y_0 + t_{N-2\,;\,0.05}\, s_{y_0}\,]$$

où $$s_{y_0}^2 = s^2 \left(\frac{1}{N} + \frac{(x_0 - \bar{X})^2}{(N-1)\, s_X^2} \right)$$

$$s^2 = \frac{1}{N-2} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

et $$\hat{Y}_i = a\, X_i + b$$



Corrélation et Régression linéaire : Démarche

Etape 1 : Vérifier les conditions d'application (visuellement...) :

Etape 2 : Réaliser le modèle de régression linéaire et récupérer les résidus

```
mod = lm(Y ~ X)
res = mod$residuals
```

rem : $res suffit

Etape 3 : S'assurer de la distribution Normale des résidus

```
shapiro.test(res)
```

Si normalité des résidus

Etape 4 : Interpréter la régression : `summary(mod)` (ou `anova(mod)`)

Ligne Intercept : **b** p_value peu représentative et peu importante

Ligne X : donne **a** p_value associé au test de pente (H0 : a=0)
(la même que pour le test de corrélation !)

Sont également décrit dans les lignes suivantes le r² et la p_value du test de corrélation de Pearson, que l'on peut retrouver par `cor.test(X, Y)`

Etape 5 : Tracer la droite de régression : `abline(mod)`

Si non-normalité des résidus : Etape 4bis : Test de corrélation non paramétrique de Spearman

```
cor.test(X, Y, method="spearman")
```