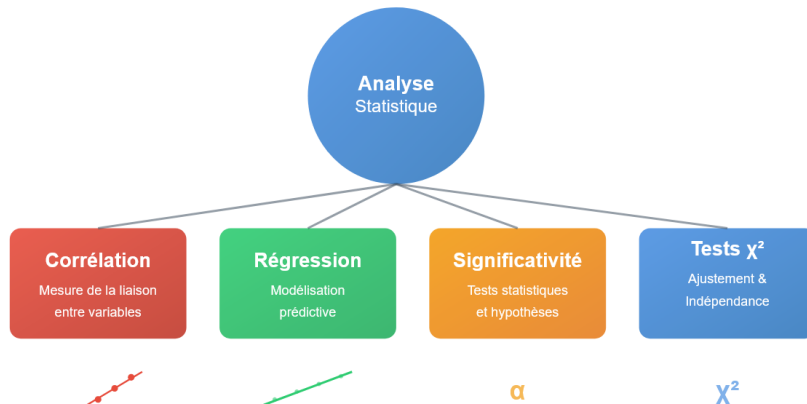


PASS : UE8

Statistiques Inférentielles



Support de Cours Introduction aux Biostatistiques Partie 3

Plan

2/69

Je me permets d'attirer votre attention
sur le fait que ce diaporama ne
constitue en aucun cas un polycopié du
cours dispensé en direct et ne saurait,
en aucune manière, s'y substituer



Plan

3/69

 **Corrélation**

 Régression Linéaire

 Significativité

 Khi-deux d'ajustement

 Khi-deux d'indépendance

 Exercices

Diagramme de Dispersion

4/69

Corrélation vs Régression Simple

CORRÉLATION

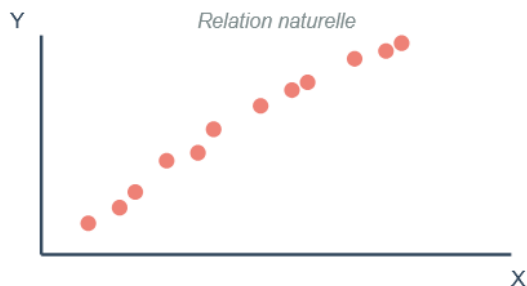
Mesure de la force du lien

Variables aléatoires



Objectif

Étudier la **force du lien**
entre X et Y



RÉGRESSION SIMPLE

Prédiction de Y en fonction de X

X fixé par l'expérimentateur



Objectif

Prédire les valeurs de Y
en fonction de X

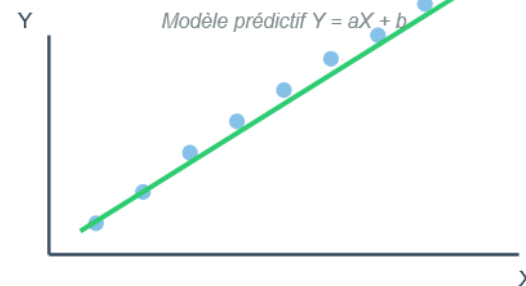
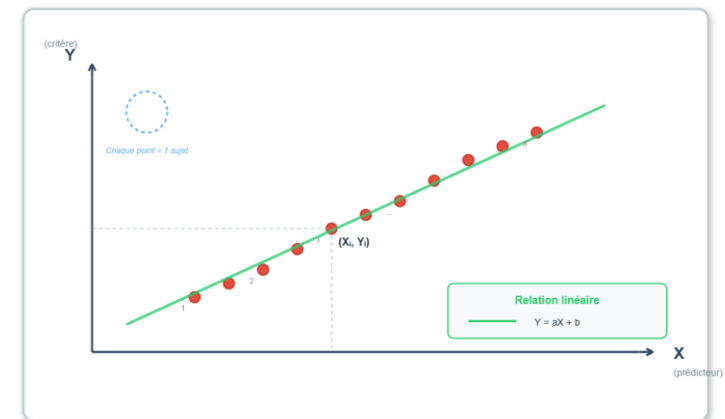


Diagramme de Dispersion

5/69

- Le **diagramme de dispersion** représente chaque sujet expérimental de l'étude par un point dans un espace bidimensionnel. Les coordonnées de ces points (X_i, Y_i) correspondent aux scores obtenus par l'individu pour les variables X et Y, respectivement.
- Important** : on suppose qu'il existe une **relation linéaire**.
- La variable utilisée comme **prédicteur** est toujours placée en abscisse (axe X), tandis que la variable à prédire (le **critère**) est placée en ordonnée (axe Y).
- Dans le contexte d'une **corrélation**, l'emplacement de X ou Y importe peu.

Diagramme de Dispersion



REGRESSION

X en abscisse (prédicteur) → Y en ordonnée (critère)

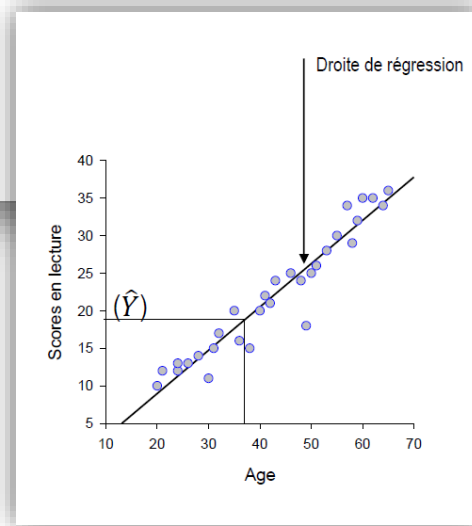
CORRELATION

L'emplacement de X ou Y importe peu

Diagramme de Dispersion

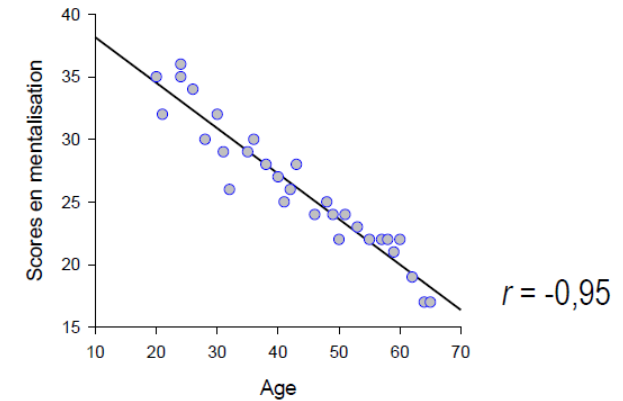
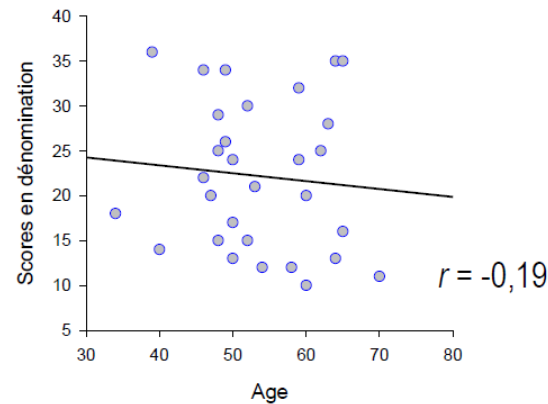
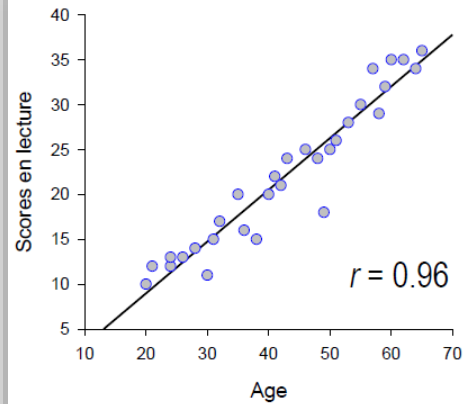
6/69

- La **droite de régression** est la ligne qui s'ajuste le mieux aux données et représente la meilleure prédiction de Y_i en fonction de X_i .
- Le symbole $\hat{Y}$ (« Y chapeau ») représente la meilleure prédiction de Y pour une valeur donnée de X.
- Le degré de regroupement des points autour de la droite de régression est lié à la **corrélation r** entre X et Y.
- Si tous les points étaient parfaitement alignés sur la droite, la corrélation serait parfaite et équivaldrait à +1.
- Le **coefficient de corrélation** se situe toujours entre -1 et +1.



La Covariance

7/69



- ❖ Pour calculer une corrélation, nous avons besoin d'une statistique appelée la **covariance** (cov_{XY}). La covariance est un nombre qui reflète le degré auquel deux variables varient ensemble.
- ❖ D'un point de vue formel :

$$\text{COV}_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1}$$

La Covariance : Exemple

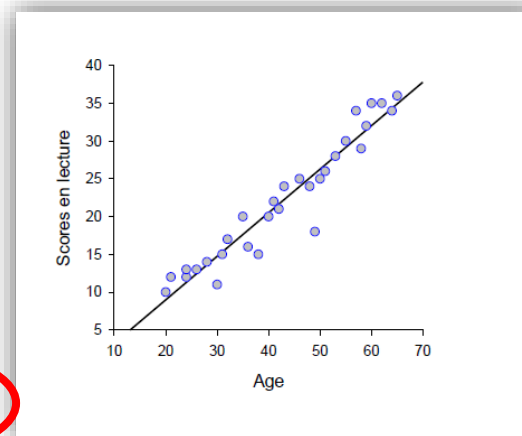
8/69

n sujets	Age (X)	Denomination (Y)	(X - $\bar{X}$)	(Y - $\bar{Y}$)	(X - $\bar{X}$)(Y - $\bar{Y}$)
1	20	10	-22,93	-12,20	279,79
2	21	12	-21,93	-10,20	223,72
3	24	12	-18,93	-10,20	193,12
4	24	13	-18,93	-9,20	174,19
5	26	13	-16,93	-9,20	155,79
6	28	14	-14,93	-8,20	122,45
7	30	11	-12,93	-11,20	144,85
8	31	15	-11,93	-7,20	85,92
9	32	17	-10,93	-5,20	56,85
10	35	20	-7,93	-2,20	17,45
11	36	16	-6,93	-6,20	42,99
12	38	15	-4,93	-7,20	35,52
13	40	20	-2,93	-2,20	6,45
14	41	22	-1,93	-0,20	0,39
15	42	21	-0,93	-1,20	1,12
16	43	24	0,07	1,80	0,12
17	46	25	3,07	2,80	8,59
18	48	24	5,07	...	...
19	49	18	6,07	...	...
20	50	25	7,07	...	...
21	51	26	8,07	...	...
22	53	28	10,07	...	...
23	55	30	12,07	...	...
24	57	34	14,07	...	...
25	58	29	15,07	...	...
26	59	32	16,07	...	...
27	60	35	17,07	...	...
28	62	35	19,07	...	...
29	64	34	21,07	...	...
30	65	36	22,07	...	...
Moyenne	42,93	22,20			
ET	13,79	8,28			

$$COV_{XY} =$$



$$COV_{XY} = \frac{3177,40}{30-1} = 109,56$$



La Corrélation de Pearson « r »

9/69

n sujets	Age (X)	Denomination (Y)	(X - $\bar{X}$)	(Y - $\bar{Y}$)	(X - $\bar{X}$)(Y - $\bar{Y}$)
1	20	10	-22,93	-12,20	279,79
2	21	12	-21,93	-10,20	223,72
3	24	12	-18,93	-10,20	193,12
4	24	13	-18,93	-9,20	174,19
5	26	13	-16,93	-9,20	155,79
6	28	14	-14,93	-8,20	122,45
7	30	11	-12,93	-11,20	144,85
8	31	15	-11,93	-7,20	85,92
9	32	17	-10,93	-5,20	56,85
10	35	20	-7,93	-2,20	17,45
11	36	16	-6,93	-6,20	42,99
12	38	15	-4,93	-7,20	35,52
13	40	20	-2,93	-2,20	6,45
14	41	22	-1,93	-0,20	0,39
15	42	21	-0,93	-1,20	1,12
16	43	24	0,07	1,80	0,12
17	46	25	3,07	2,80	8,59
18	48	24	5,07	1,80	9,12
19	49	18	6,07	-4,20	-25,48
20	50	25	7,07	2,80	19,79
21	51	26	8,07	3,80	30,65
22	53	28	10,07	5,80	58,39
23	55	30	12,07	7,80	94,12
24	57	34	14,07	11,80	165,99
25	58	29	15,07	6,80	102,45
26	59	32	16,07	9,80	157,45
27	60	35	17,07	12,80	218,45
28	62	35	19,07	12,80	244,05
29	64	34	21,07	11,80	248,59
30	65	36	22,07	13,80	304,52
Moyenne	42,93	22,20			3177,40
ET	13,79	8,28			Somme

Le calcul du coefficient de Pearson r s'effectue de la façon suivante: $r = \frac{COV_{yx}}{S_x S_y}$

Ce coefficient est toujours situé entre 1 et -1:

- Plus on tend vers 1 ou -1, plus la relation est forte ;
- Plus on s'approche de 0, plus la relation est faible ;
- le signe + ou le signe - indique le sens de la relation ;

$$COV_{xy} = \frac{3177,40}{30-1} = 109,56$$

$r =$



La Corrélation de Pearson : Les Outils

10/69

The image shows the STATISTICA software interface. The main window displays a data spreadsheet with the following columns: ID sujet, Age, Lecture, Dénomination, Mentalisation, Var5, Var6, and Var7. The data is organized into rows, with the first row being the header and subsequent rows containing numerical values.

Overlaid on the spreadsheet is the 'Basic Statistics and Tables: Spreadsheet22' dialog box. The 'Quick' tab is active, and the 'Correlation matrices' option is selected. Other options visible include Descriptive statistics, t-test independent by groups, t-test independent by variables, t-test dependent samples, t-test single sample, Breakdown & one-way ANOVA, Breakdown: non-factorial tables, Frequency tables, Tables and benners, Multiple response tables, Difference tests: t, %, means, and Probability calculator. The dialog box also features buttons for OK, Cancel, Options, and Open Data.

	1	2	3	4	5	6	7	8
	ID sujet	Age	Lecture	Dénomination	Mentalisation	Var5	Var6	Var7
1	1	20	10	60	35			
2	1	21	12	54	32			
3	1	24	12	58	36			
4	1	24	13	64	35			
5	1	25	13	50	34			
6	1	28	14	40	30			
7	1	30	11	70	32			
8	1	31	15	52	29			
9	1	32	17	50	25			
10	1	35	20	60	29			
11	1	36	16	65	30			
12	1	38	15	48	28			
13	1	40	20	47	27			
14	1	41	22	46	25			
15	1	42	21	53	26			
16	1	43	24	59	28			
17	1	45	25	62	24			
18	1	48	24	50	25			
19	1	49	18	34	24			
20	1	50	25	48	22			
21	1	51	26	49	24			
22	1	53	28	63	23			
23	1	55	30	52	22			
24	1	57	34	49	22			
25	1	58	29	48	22			
26	1	59	32	59	21			
27	1	60	35	64	22			
28	1	62	35	65	19			
29	1	64	34	46	17			
30	1	65	36	39	17			

La Corrélation de Pearson : Les Outils

11/69

The screenshot displays the STATISTICA software interface. The main window shows a spreadsheet with data for 11 subjects across 8 variables. The 'Product-Moment and Partial Correlations: Spreadsheet22' dialog box is open, showing the 'One variable list' tab. The 'First list' is set to 'none' and the 'Second list' is set to 'none'. The 'Quick' tab is selected, and the 'Display format for correlation matrices' section is visible. The 'Display simple matrix (highlight p's)' option is selected. The 'p-level for highlighting' is set to 0.05. The 'Include means and std. devs. in square matrices' option is checked. The 'Weighted moments' section is also visible, with 'DF' set to 'W-1' and 'N-1'. The 'MD deletion' section is set to 'Casewise'. The 'Select one or two variable lists' dialog box is also open, showing the 'First variable list' as '2' and the 'Second variable list (optional)' as '2-5'. The 'Show appropriate variables only' option is unchecked.

	1	2	3	4	5	6	7	8
	ID sujet	Age	Lecture	Dénomination	Mentalisation	Var5	Var6	Var7
1	1	20	10	60	35			
2	1	21	12	54	32			
27	1	60	35	64	22			
28	1	62	35	65	19			
29	1	64	34	46	17			
30	1	65	36	39	17			

La Corrélation de Pearson : Les Outils

12/69

STATISTICA - Workbook7* - [Correlations (Spreadsheet22)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Workbook Window Help

Data: Spreadsheet22* (11v by 30c)

	1	2	3	4	5	6	7	8
	ID sujet	Age	Lecture	Dénomination	Mentalisation	Var5	Var6	Var7
1	1	20	10	60	35			
2	1	21	12	54	32			
3	1	24	12	68	36			

Workbook7* - Correlations (Spreadsheet22)

Correlations (Spreadsheet22)
Marked correlations are significant at $p < .05000$
N=30 (Casewise deletion of missing data)

Variable	Age	Lecture	Dénomination	Mentalisation
Age	1.00	0.96	-0.19	-0.96

La Corrélation de Pearson : « r Ajusté »

13/69

- Le **coefficient de corrélation** que nous avons calculé est généralement utilisé. Toutefois, il s'agit d'une **estimation biaisée** du coefficient de corrélation de la population (puisque nous travaillons sur un échantillon).
- Nous devons donc **ajuster la corrélation**, avec la formule suivante :

$$r_{\text{ajusté}} = \sqrt{1 - \frac{(1 - r^2)(N - 1)}{N - 2}}$$

- En reprenant notre exemple précédent :

$$r_{\text{ajusté}} =$$

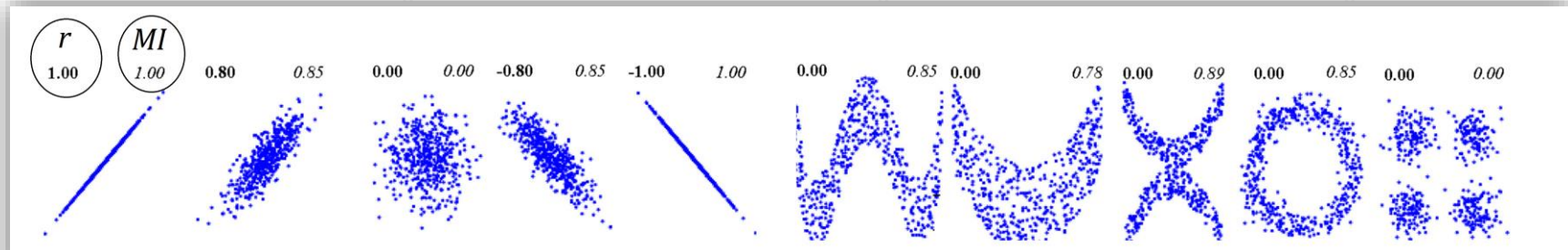
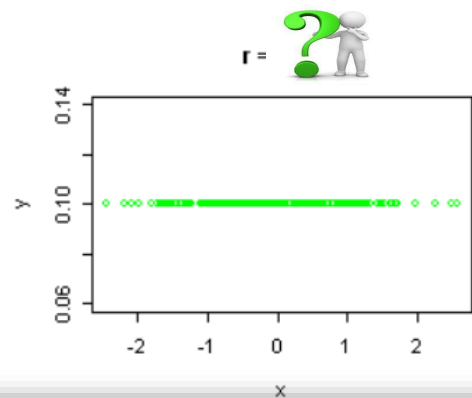
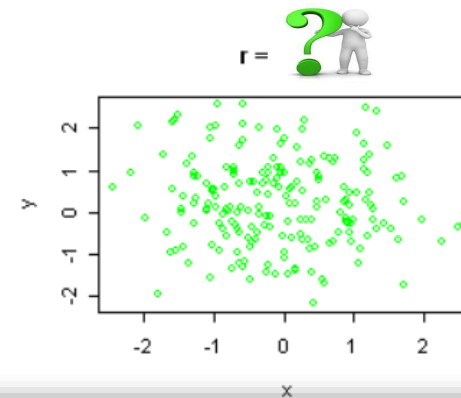
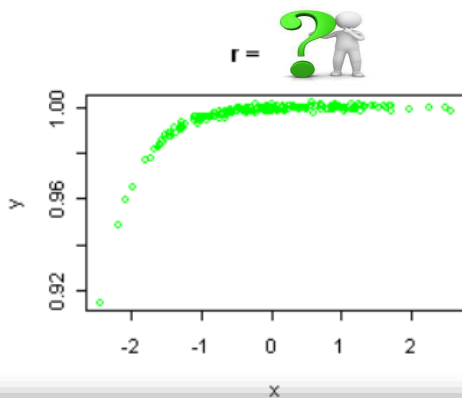
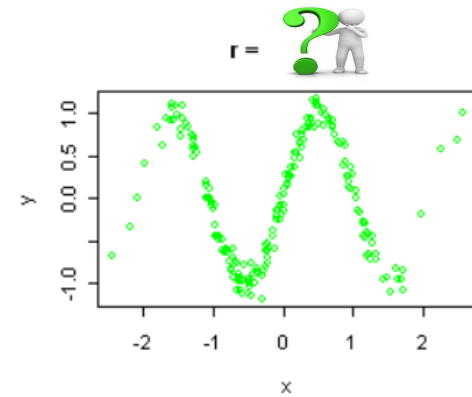
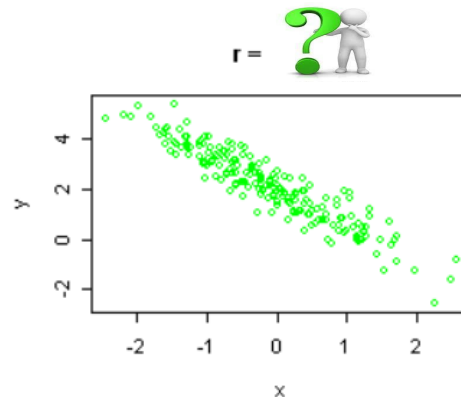
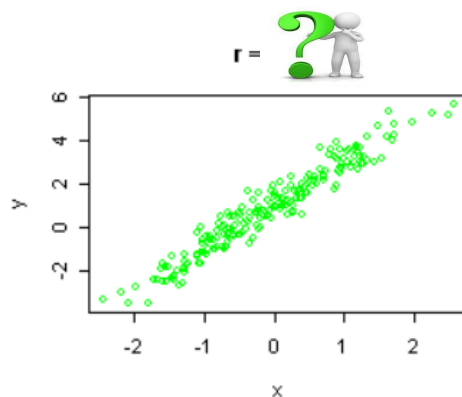

- Si n était plus petit et la corrélation plus faible :

$$r_{\text{ajusté}} = \sqrt{1 - \frac{(1 - 0,55^2)(10 - 1)}{10 - 2}} = 0,464$$

La Corrélation de Pearson : Le Piège

14/69

$r = 0.97$
 $r = 0.59$
 $r = -0.93$
 $r = \text{NA}$
 $r = -0.07$
 $r = 0.0057$



Plan

15/69

👉 Corrélation

👉 Régression Linéaire

👉 Significativité

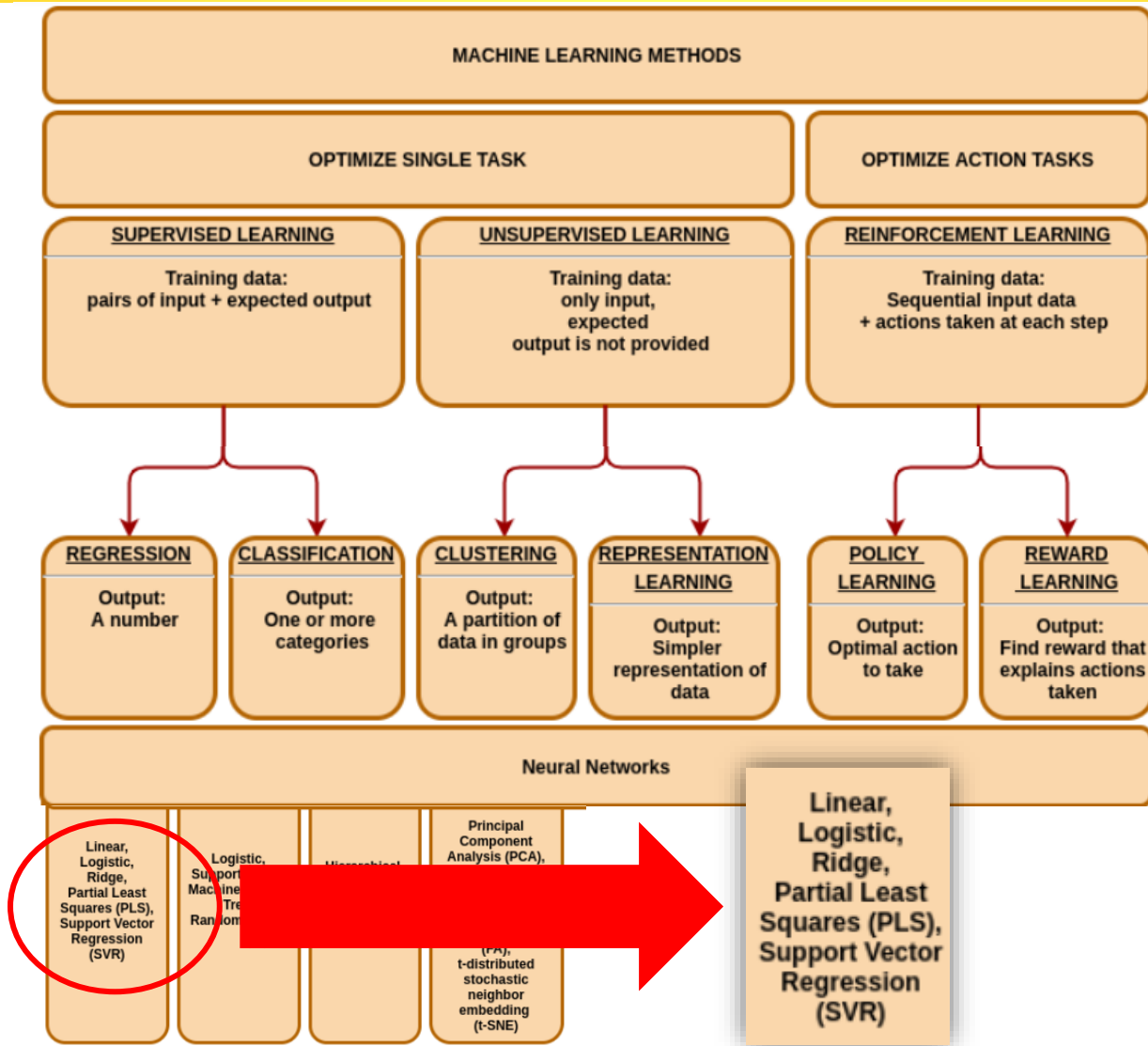
👉 Khi-deux d'ajustement

👉 Khi-deux d'indépendance

👉 Exercices

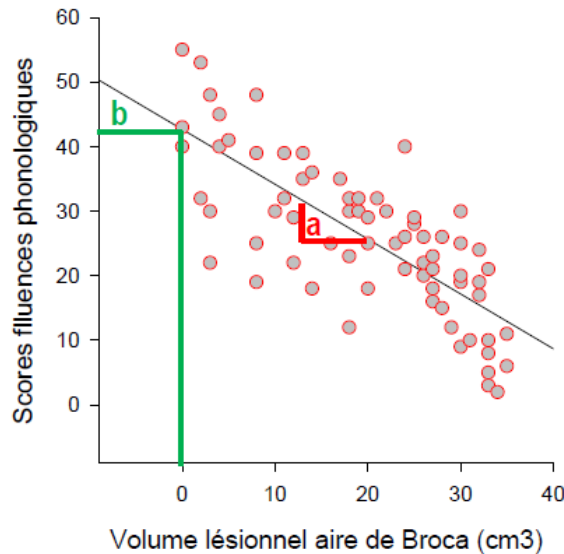
Régression et Apprentissage Automatique

16/69



Droite de Régression

17/69



- Supposons que nous souhaitons prédire le score en fluence phonologiques de patients atteints d'une lésion de l'aire de Broca. Le graphique ci-contre représente les scores en fluence en fonction du volume lésionnel dans l'aire de Broca.

- La droite de régression linéaire va s'écrire : $\hat{Y} = aX + b$.

Avec $\hat{Y}$ = la valeur prédite de Y

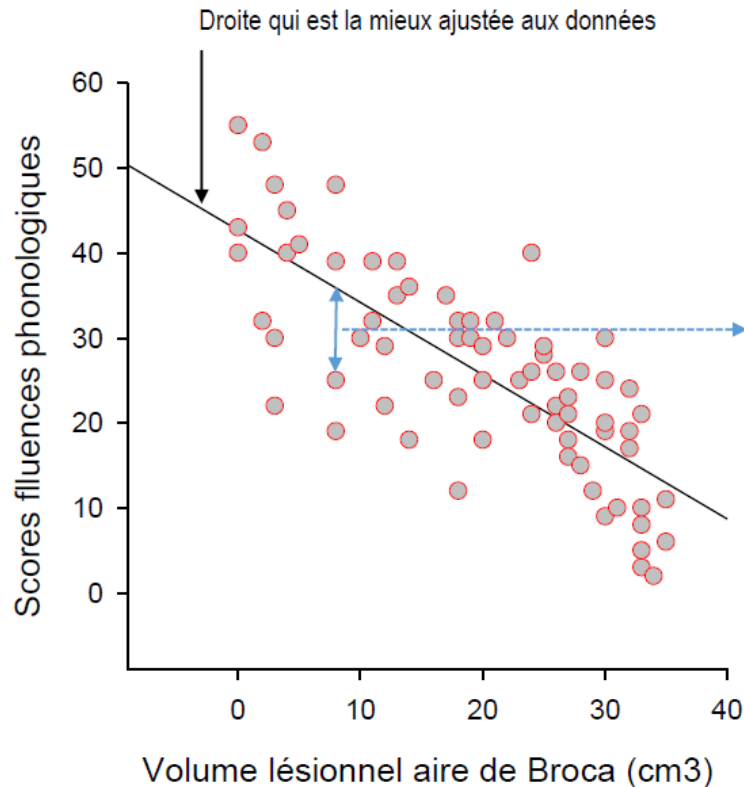
a = **la pente** de la droite de régression

b = **l'ordonnée à l'origine** (la valeur de $\hat{Y}$ lorsque X = 0)

X = la valeur du prédicteur

Droite de Régression

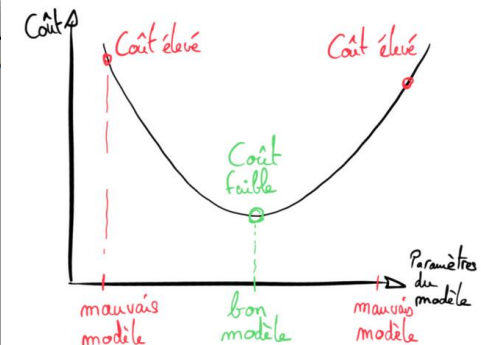
18/69



- Equations normales:

$$b = \bar{Y} - a\bar{X}$$

$$a = \frac{cov_{XY}}{S_X^2}$$

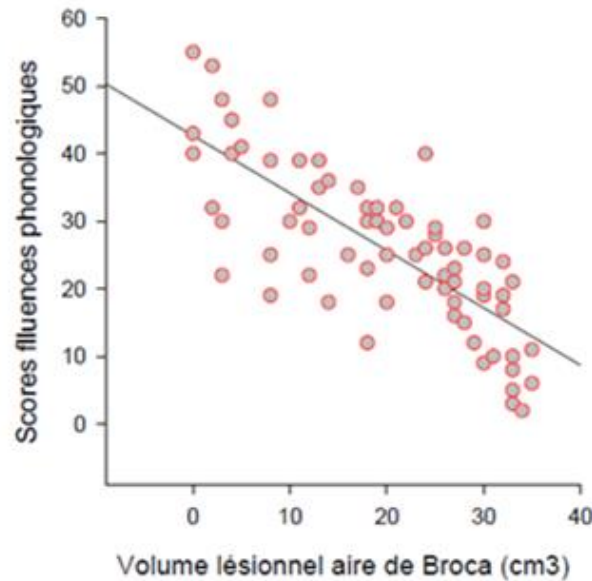


La droite de régression est la droite qui utilise comme paramètres les valeurs optimales de a et de b en recherchant les valeurs de a et de b qui minimisent

$$\sum (Y - \hat{Y})^2$$

Droite de Régression : Exemple

19/69



- Dans cet exemple :

$$\bar{Y} = 19,65$$

$$\bar{X} = 26$$

$$s_X = 10,62$$

$$cov_{xy} = -95,92$$

$$a =$$

$$=$$

$$b =$$

$$=$$

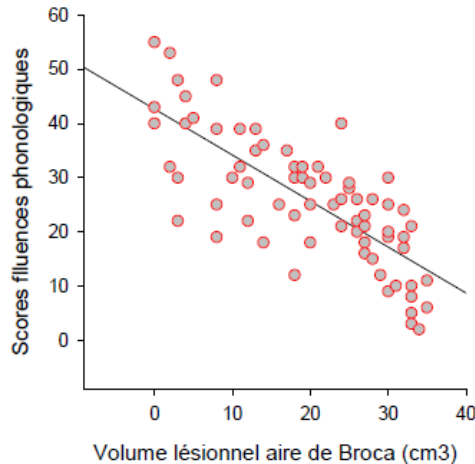
$$=$$

$$\hat{Y} =$$



Droite de Régression : Prédiction

20/69



$$\hat{Y} = -0,85X + 41,75$$

Pour $X = 1$; $\hat{Y} =$,

Pour $X = 10$; $\hat{Y} =$,

Pour $X = 30$; $\hat{Y} =$,

Pour $X = 40$; $\hat{Y} =$,



Remarque :

- La pente de la régression est une donnée particulièrement utile.
- Elle constitue une mesure du taux de changement prédit au niveau de Y. On sait ici que quand le volume lésionnelle dans l'aire de Broca augmente de 1 CM³ le score en dénomination diminue de 0,85 point.

Droite de Régression : Outils Numériques

21/69

Data: Spreadsheet22* (3v by 71c)

	1	2	3
	ID sujet	Lésion de l'air de Broca (CM3)	Score fluence phonologique
1	1	0	55
2	2	0	40
3	3	0	43
4	4	2	32
5	5	2	53
6	6	3	22
7	7	3	30
8	8	3	48
9	9	4	45
10	10	4	40
11	11	5	41
12	12	8	39
13	13	8	48
14	14	8	19
15	15	8	25
16	16	10	30
17	17	11	32
18	18	11	39
19	19	12	22
20	20	12	29
21	21	13	35
22	22	13	39
23	23	14	36
24	24	14	18
25	25	16	
26	26	17	
27	27	18	
28	28	18	
29	29	18	

→

General Linear Models (GLM): Spreadsheet22

Quick | Options | OK | Cancel | Options | Open Data | SELECT CASES | Weighted moments | DF = W-1 N-1

Type of analysis:

- One-way ANOVA
- Main effects ANOVA
- Factorial ANOVA
- Nested design ANOVA
- Huge balanced ANOVA
- Repeated measures ANOVA
- Simple regression**
- Multiple regression
- Factorial regression
- Polynomial regression
- Response surface regression
- Mixture surface regression
- Analysis of covariance
- Separate-slopes model
- Homoogeneity-of-slopes model
- General linear models**

Specification method:

- Quick specs dialog**
- Analysis Wizard
- Analysis syntax editor

Use General linear models to analyze designs with any combination of categorical independent variables (factors), continuous predictor variables (covariates), or repeated measures.

Multiple dependent variables can be specified for any type of analysis. Both univariate and multivariate results are available when multiple dependent variables are specified.

Sélectionner régression simple

↓

GLM Simple Regression: Spreadsheet22

Quick | Options | OK | Cancel | Options | Syntax editor

Variables

Dependent variables: none

Continuous regressor: none

Between effect: none

Placement des variables

Sélectionner les variables

Droite de Régression : Outils Numériques

22/69

Select dependent variables and a continuous predictor (covariate):

1 - ID sujet
2 - Lésion de l'air de Broca (CM3)
3 - Score fluence phonologique

Select All Spread Zoom

Dependent variable list:
3

1 - ID sujet
2 - Lésion de l'air de Broca (CM3)
3 - Score fluence phonologique

Select All Spread Zoom

Predictor variable:
2

OK Cancel [Bundles]...

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

☐ Show appropriate variables only

→

GLM Results 1: Spreadsheet22

Quick Summary Profiler Resids Matrix Report

All effects/Graphs All effects

Effect sizes

Between effects

Design terms Whole model R

Alpha values

Confidence limits: .950

Significance level: .050

Sélectionner les variables: dépendante versus prédictive

Workbook11* - Parameter Estimates (Spreadsheet22)

Effect	Score fluence phonologique Param.	Score fluence phonologique Std.Err	Score fluence phonologique t	Score fluence phonologique p	-95.00% Cnf.Lmt	+95.00% Cnf.Lmt	Score fluence phonologique Beta (B)	Score fluence phonologique St.Err.B	-95.00% Cnf.Lmt	+95.00% Cnf.Lmt
Intercept	41.75000	1.939691	22.01303	0.000000	38.82787	46.56906				
Lésion de l'air de Broca (CM3)	-0.84949	0.086945	-9.77043	0.000000	-1.02298	-0.67599	-0.764199	0.078215	-0.920275	-0.608122

Coefficient de Détermination « r^2 »

23/69

❖ Calcul du r^2

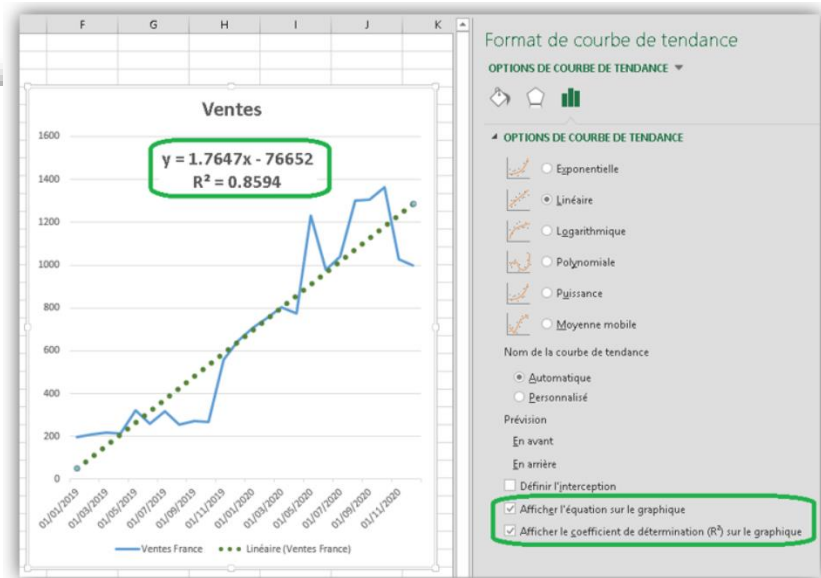
$$r^2 = \frac{SC_Y - SC_{\text{résiduelle}}}{SC_Y} \quad \text{ou} \quad \frac{SC_{\hat{Y}}}{SC_Y}$$

Avec,

$SC_Y = \sum (Y - \bar{Y})^2$ correspond à la variabilité des scores en dénomination

$SC_{\hat{Y}} = \sum (\hat{Y} - \bar{Y})^2$ correspond à la variabilité des scores en dénomination imputable à la variabilité du volume lésionnel dans l'aire de Broca

$SC_{\text{résiduelle}} = SC_Y - SC_{\hat{Y}}$ correspond à la variabilité des scores en dénomination qui n'est pas imputable à la variabilité du volume lésionnel dans l'aire de Broca



Plan

24/69

 Corrélation

 Régression Linéaire

 Significativité

 Khi-deux d'ajustement

 Khi-deux d'indépendance

 Exercices

Utilité de « r^2 »

25/69

Importance du r^2 (dans notre exemple, $r = 0,764$, $r^2 = 0,584$)

- Dans notre exemple, grâce au r^2 , nous pouvons dire que $0,764^2 = 58,4\%$ de la variabilité dans les scores de dénomination peut être prédite grâce (donc est liée) au volume lésionnel dans l'aire de Broca, ce qui implique que 41,6 % de la variabilité restante est imputable à d'autres facteurs. Notre modèle semble donc excellent.
- Si notre r était de 0,13, seulement $0,13^2 = 0,0169 = 1,69\%$ de la variabilité des scores en dénomination serait expliquée par la variabilité de notre prédicteur. Dès lors, notre modèle serait inadéquat.

r^2 : Outils Numériques

26/69

GLM Results 1: Spreadsheet22

Quick | Summary | Profiler | Resids | Matrix | Report

All effects/Graphs | All effects

Effect sizes

Between effects

Design terms | **Whole model R**

Sélectionner « *whole model R* »

Vous obtenez les valeurs de r et de r^2

Workbook11* - Test of SS Whole Model vs. SS Residual (Spreadsheet22)

Dependent Variable	Multiple R	Multiple R ²	Adjusted R ²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Score fluence phonologique	0.764199	0.584000	0.577882	5622.747	1	5622.747	4005.253	68	58.90078	95.46134	0.000000

Significativité de « r » : Test de Student

27/69

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

❖ Dans le cadre de la régression linéaire, nous allons calculer la signification de r et de α de la même façon (ce qui n'est pas le cas si l'on se place dans le cadre des régressions multiples).

❖ Nous pouvons montrer que

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad (DL=N-2)$$

Dans notre exemple,

$t =$  $(N=70)$

H_0 =pas de Corrélacion

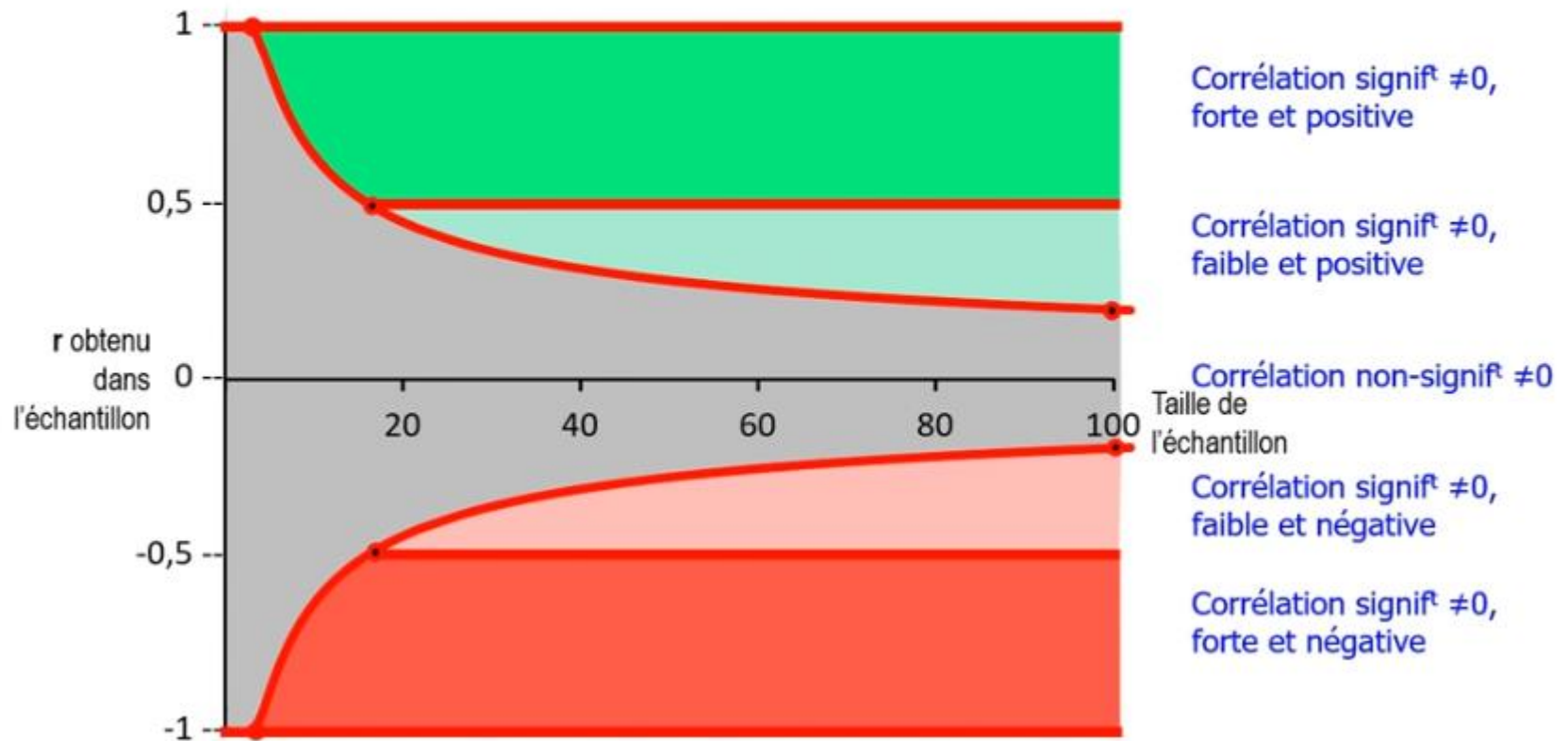
H_1 =Corrélacion entre les 2 variables

Si $t > t_{\text{table}}$ Alors on rejette H_0

Significativité de « r » : Synthèse

28/69

Interprétation en fonction de r (en Y) et de l'effectif n (en X)
(graphe obtenu d'après la table de Student, test bilatéral, $\alpha=5\%$):



Significativité de « r » : Exemple

29/69

STATISTICA - Spreadsheet22

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help

Data: Spreadsheet22* (5v by 71c)

	1 ID sujet	2 Lésion de l'air de Broca (CM3)	3 Score fluence phonologique	4 Score en fluence sémantique	5 Score en lecture
1	1	18	55	43	20
2	2	20	40	22	19
3	3	22	43	29	15
4	4	25	32	41	20
5	5	30	53	23	18
6	6	34	22	12	17
7	7	28	30	32	15
8	8	40	48	25	18
9	9	41	45	39	20
10	10	42	40	44	12
11	11	45	41	22	18
12	12	48	39	15	17
13	13	50	48	53	13
14	14	51	19	19	18
15	15	52	25	42	17
16	16	53	30	52	14
17	17	58	32	29	11
18	18	60	39	35	18
19	19	62	22	30	17
20	20	65	29	31	25
21					
22					
23					
24					
25					
26					
27					
28					
29					

Nouvelles données

Basic Statistics and Tables: Spreadsheet22

Quick

- Descriptive statistics
- Correlation matrices
- t-test independent by groups
- t-test independent by variables
- t-test dependent samples
- t-test single sample
- Breakdown & one-way ANOVA
- Breakdown; non-factorial tables
- Frequency tables
- Tables and banners
- Multiple response tables
- Difference tests: r, %, means
- Probability calculator

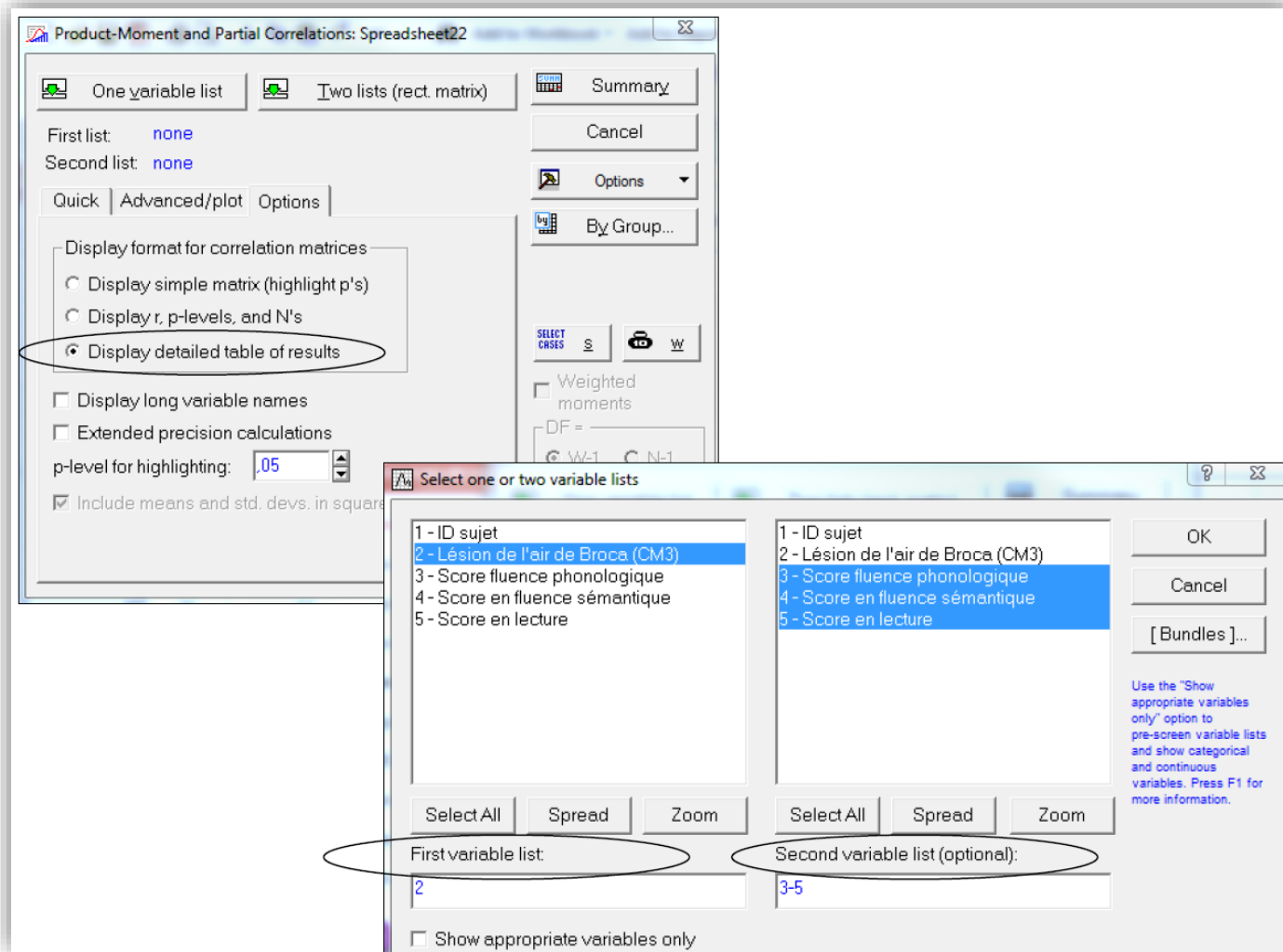
OK Cancel Options

Open Data

SELECT CHRS

Significativité de « r » : Exemple

30/69



Significativité de « r » : Exemple

31/69

Qu'en déduisez-vous ?

Correlations (Spreadsheet22)

Marked correlations are significant at $p < ,05000$
(Casewise deletion of missing data)

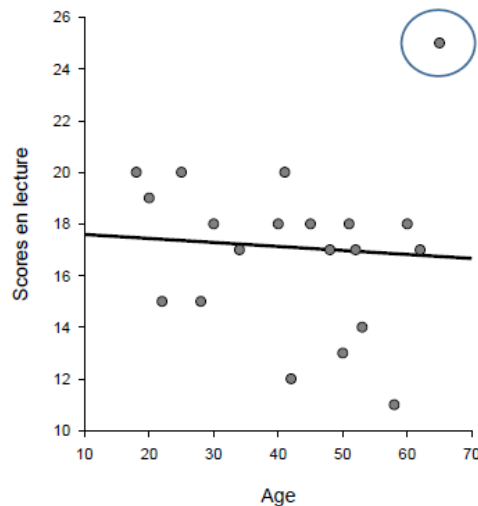
Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	r ²	t	p	N	Constant dep: Y	Slope dep: Y	Constant dep: X	Slope dep: X
Lésion de l'air de Broca (CM3)	42,20000	14,64887									
Score fluence phonologique	36,60000	10,49010	-0,448814	0,201434	-2,13082	0,047147	20	50,16295	-0,321397	65,13889	-0,626746
Lésion de l'air de Broca (CM3)	42,20000	14,64887									
Score en fluence sémantique	31,90000	11,59809	0,099254	0,009851	0,42319	0,677171	20	28,58378	0,078583	38,20095	0,125362
Lésion de l'air de Broca (CM3)	42,20000	14,64887									
Score en lecture	17,10000	3,19374	-0,071323	0,005087	-0,30337	0,765085	20	17,75621	-0,015550	47,79412	-0,327141

Corrélation et Outliers

32/69

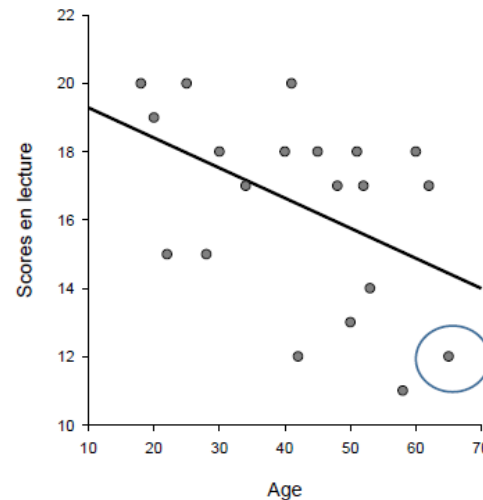
❖ Les valeurs extrêmes.

Les scores extrêmes sur une distribution influencent considérablement les résultats.



$$r = 0,0713, r^2 = 0,00509, t = -0,303, p = 0,765$$

Le modèle est non-significatif



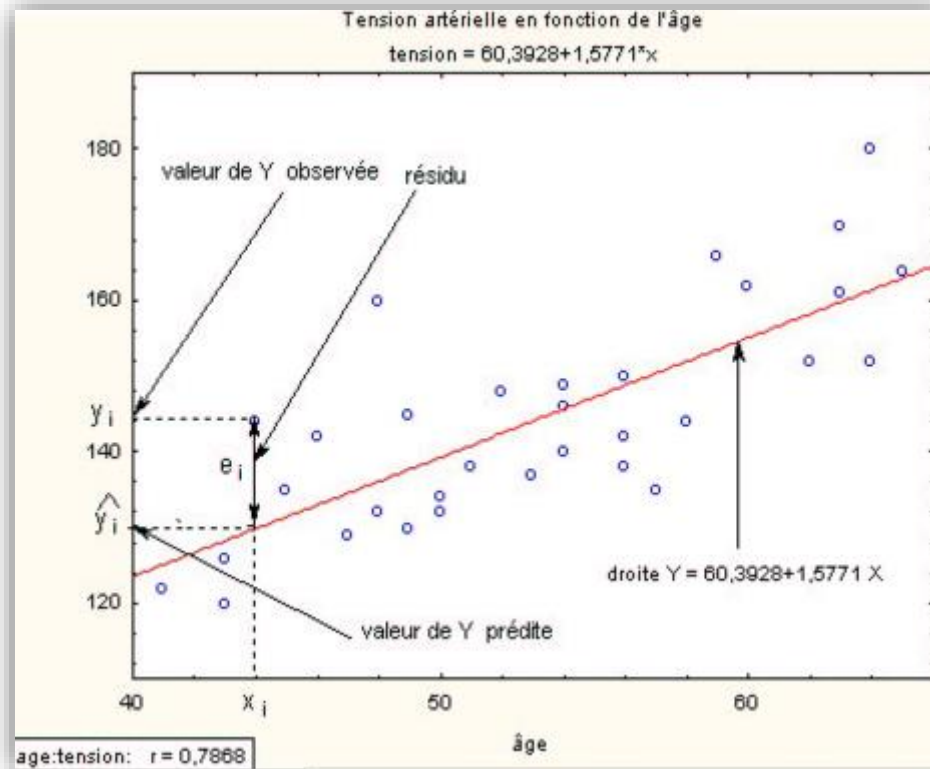
$$r = 0,462, r^2 = 0,213, t = -2,208, p = 0,040$$

Prendre la valeur absolue pour la comparaison avec t_{table}

Le modèle est significatif

Régression et Résidus

33/69



On Pose :

$e_i = y_i - \hat{y}_i$ s'appelle le résidu de l'observation i .

la somme des carrés des résidus est $SCR = \sum e_i^2$.

$$\hat{\sigma}^2 = SCR / (n - 2)$$

On Obtient :

Exemple numérique :

l'individu n° 5 a pour âge $x_5 = 44$.

La tension observée est $y_5 = 144$.

La tension prédite (ou estimée) par le modèle est

$\hat{y}_5 =$? ? ? ? ? ? ? ? ? ?

Le résidu pour l'observation n° 5 est $e_5 =$? ? ? ? ? ? ? ? ? ?

Le résidu est positif (point au dessus de la droite).

$$SCR = 2605,569$$

$$\hat{\sigma}^2 =$$
 ? ? ? ? ? ? ? ? ? ? (N=34)

$$r^2 =$$
 ? ? ? ? ? ? ? ? ? ?

Le modèle de régression explique ? ? ? ? ? % de la variation totale.

Régression : Tests sur les Paramètres

34/69

Soit le Modèle $\hat{Y} = b_1X + b_0$

- (a) est-ce-que le coefficient b_1 est non nul, autrement dit la variable X a-t-elle réellement une influence sur Y ?
- (b) est-ce-que le coefficient b_0 est non nul, autrement dit faut-il une constante dans le modèle ?

Nous pouvons expliciter les questions précédentes en termes de test d'hypothèse :

- (a) correspond à $H_0 : b_1 = 0$, contre $H_1 : b_1 \neq 0$
- (b) correspond à $H_0 : b_0 = 0$, contre $H_1 : b_0 \neq 0$

statistique T de Student

On utilise la statistique, notée T définie par la formule, pour $j = 0, 1$:

$$T = \frac{\hat{b}_j}{\hat{\sigma}_{\hat{b}_j}}$$

La statistique T suit la loi de Student à $n - 2$ ddl.

Significativité des Coef. de Régression: Outils Numériques

35/69

Workbook11* - Parameter Estimates (Spreadsheet22)

Workbook11*

- Basic Statistics/Ta
 - Descriptive stat
 - Descriptive
 - General Linear Mo
 - GLM Results 1:
 - Test of SS V
 - Univariate T
 - Test of SS V
 - Column Lal
 - Parameter
 - General Linear Mo
 - GLM Results 1:
 - Column Lal
 - Column Lal
 - Test of SS V

Parameter Estimates (Spreadsheet22) Sigma-restricted parameterization										
Effect	Score fluence phonologique Param.	Score fluence phonologique Std.Err	Score fluence phonologique t	Score fluence phonologique p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt	Score fluence phonologique Beta (B)	Score fluence phonologique St.Err.B	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Intercept	41,75000	1,939691	22,01303	0,000000	38,82787	46,56906				
Lésion de l'air de Broca (CM3)	-0,84949	0,086945	-9,77043	0,000000	-1,02298	-0,67599	-0,764199	0,078215	-0,920275	-0,608122

Régression : Exemple (Tension Artérielle)

36/69

Soit le Modèle $\hat{Y} = b_1X + b_0$ avec $\hat{\sigma} = \sqrt{2605,569/32} = 9,0235$

Synthèse de la Régression; Variable Dép. : tension (donnees-bol)						
R= ,78678955 R²= ,61903780 R² Ajusté = ,60713273						
F(1,32)=51,998 p<,00000 Err-Type de l'Estim.: 9,0235						
N=34	Bêta	Err-Type de Bêta	B	Err-Type de B	t(32)	niveau p
OrdOrig.			60,39282	11,87925	5,083893	0,000016
age	0,786790	0,109110	1,57709	0,21871	7,210952	0,000000

Pour La pente :

$$t_{obs} = \left| \frac{b_1}{\sqrt{\frac{\hat{\sigma}^2}{n \cdot S_{XX}}}} \right|$$

(a) $H_0 : b_1 = 0$, contre $H_1 : b_1 \neq 0$

$t_{obs} =$ 

$\alpha_{obs} = 0$, on  donc H_0 au risque 5%.

(b) $H_0 : b_0 = 0$, contre $H_1 : b_0 \neq 0$

$t_{obs} =$ 

$\alpha_{obs} \approx 0$, on  donc H_0 au risque 5%.

Régression : Intervalle de Prédiction

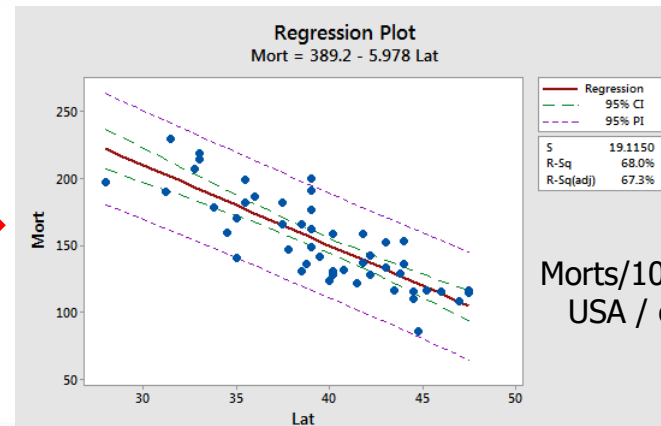
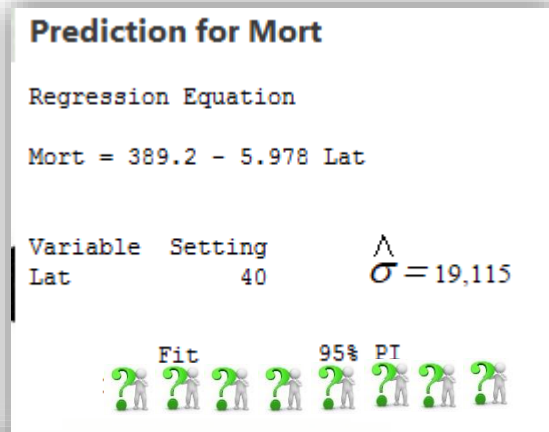
37/69

95% des prévisions seront comprises entre les bornes :

$$y \in \left[\hat{y} \pm t_{0.975}(n-2) \times \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \right]$$

On se facilite la tâche en considérant la formule :

$$y \in [\hat{y} \pm 2\hat{\sigma}] \quad \text{soit} \quad |e| = |y - \hat{y}| < 2\hat{\sigma}$$



Morts/10Millions d'Habitants
USA / cancer de la peau

Une observation *i* est un *outlier* si son résidu standardisé $\frac{e_i}{\hat{\sigma}}$ est supérieur à 2 en valeur absolue.

Plan

38/69



Corrélation



Régression Linéaire



Significativité



Khi-deux d'ajustement



Khi-deux d'indépendance



Exercices

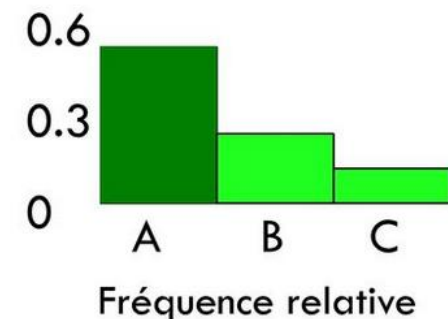
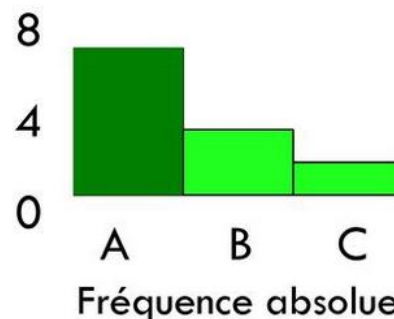
Khi-carré : Variables Qualitatives

39/69

- Permet de faire des tests d'hypothèses lorsque les deux variables sont qualitatives.
- On se sert de données exprimées en terme de fréquences.



- Le nombre d'observations se trouvant dans une classe (p.ex., nb de loups ayant le poil roux)
- Fréquence absolue (le nombre d'observations)
- Fréquence relative (la proportion des observations)



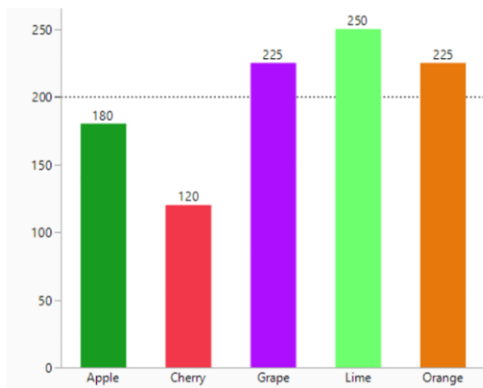
Khi-carré : Qualité d'Ajustement

40/69

Le test du khi-deux de qualité de l'ajustement vérifie si les données d'échantillon sont susceptibles de provenir d'une distribution théorique spécifique. Nous avons un jeu de valeurs de données et une idée concernant la manière dont les valeurs de données sont distribuées. Le test nous donne le moyen de décider si les valeurs de données sont « suffisamment bonnes » pour s'adapter à notre idée ou si notre idée est discutable.

Pour appliquer le test de qualité de l'ajustement au jeu de données, nous avons besoin des éléments suivants :

- Des valeurs de données qui ne sont pas un simple échantillon aléatoire issu de l'ensemble de la population.
- Données catégorielles ou nominales. Le test du khi-deux de qualité de l'ajustement n'est pas adapté aux données continues.
- Un jeu de données qui est suffisamment large pour qu'au moins cinq valeurs soient attendues dans chacune des catégories de données observées.



Exemple du dénombrement
des saveurs présentes dans
10 sacs de 100 bonbons

Khi-carré : Méthodologie Calculatoire

41/69

Saveur	Nombre de bonbons (10 sacs)	Nombre attendu de bonbons	Observés-Attendus
Pomme	180	200	180-200 = -20
Citron	250	200	250-200 = 50
Cerise	120	200	120-200 = -80
Orange	225	200	225-200 = 25
Raisin	225	200	225-200 = 25

H_0 : les proportions de saveurs de bonbons sont égales

H_1 : les proportions de saveurs ne sont pas les mêmes

REGLE => Si χ^2 calculé > χ^2 table alors on rejette H_0

$$\chi^2_{k-1} = \sum_{i=1}^k \frac{(Valeur.observée_i - Valeur.théorique_i)^2}{Valeur.théorique_i} =$$



Khi-carré : ddl et Risque alpha limite

42/69

Saveur	Nombre de bonbons (10 sacs)	Nombre attendu de bonbons	Observés-Attendus	Différence au carré	Différence au carré / Nombre attendu
Pomme	180	200	180-200 = -20	400	400 / 200 = 2
Citron	250	200	250-200 = 50	2500	2500 / 200 = 12,5
Cerise	120	200	120-200 = -80	6400	6400 / 200 = 32
Orange	225	200	225-200 = 25	625	625 / 200 = 3,125
Raisin	225	200	225-200 = 25	625	625 / 200 = 3,125

$$\text{ddl} = (5-1) \times (2-1) = 4$$

Pour les données des bonbons, nous fixons un seuil de significativité de 5 % ($\alpha = 0,05$), acceptant ainsi un risque de 5 % de conclure à une inégalité des saveurs entre les sacs alors qu'elles sont en réalité égales.

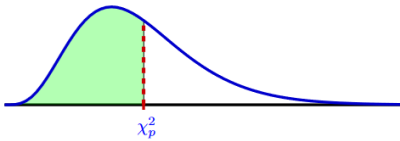
Khi-carré : Table

43/69

Tableau C

Percentiles de la distribution du χ^2 . Valeurs de χ^2_p correspondant à p

$$\alpha = 1 - p$$



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

La valeur du khi-deux avec
 $\alpha = 0,05$
et 4 degrés de liberté est 9,488

Comme notre statistique de test
(52,75) est supérieure à la valeur
critique du khi-deux (9,488), nous
rejetons l'hypothèse nulle d'égalité
des proportions de saveurs de
bonbons.

Khi-carré : Conclusions

44/69

Les sacs de bonbons ne contiennent pas un nombre égal de chaque saveur. Si vous préférez le Citron, vous en aurez probablement plus, mais si vous aimez la Cerise, vous risquez d'en trouver moins.

Plan

45/69



Corrélation



Régression Linéaire



Significativité



Khi-deux d'ajustement



Khi-deux d'indépendance



Exercices

Khi-carré : Test d'hypothèses

46/69

- Test khi-deux d'indépendance entre deux variables qualitatives
- Hypothèses :
 - H_0 : Les variables A et B sont indépendantes
(Les proportions sont identiques pour chaque modalité de A quelle que soit B)
 - H_1 : Les variables A et B ne sont pas indépendantes
(Les proportions diffèrent selon les modalités de A et B)
- Statistique du test :

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

où O_i est l'effectif observé et E_i l'effectif attendu.

- Interprétation :
 - Si χ^2 est supérieur à la valeur critique, on rejette H_0 (association entre A et B).
 - Si χ^2 est inférieur, on ne rejette pas H_0 (indépendance).

Khi-carré : Exemple Simple

47/69

- **Hypothèse** : Le sex-ratio dans la population est de 50:50.
 - **Données** : Sexe à la naissance de 10 000 enfants.
 - **Résultats** : Fréquences observées — 5 185 garçons et 4 815 filles.
- **Analyse** : Fréquence attendue selon l'hypothèse — 5 000 garçons et 5 000 filles.

Khi-carré : Exemple Simple

48/69

- Calcul du Khi-carré :

$$\chi_{k-1}^2 = \sum_{i=1}^k \frac{(\text{Valeur observée}_i - \text{Valeur théorique}_i)^2}{\text{Valeur théorique}_i}$$

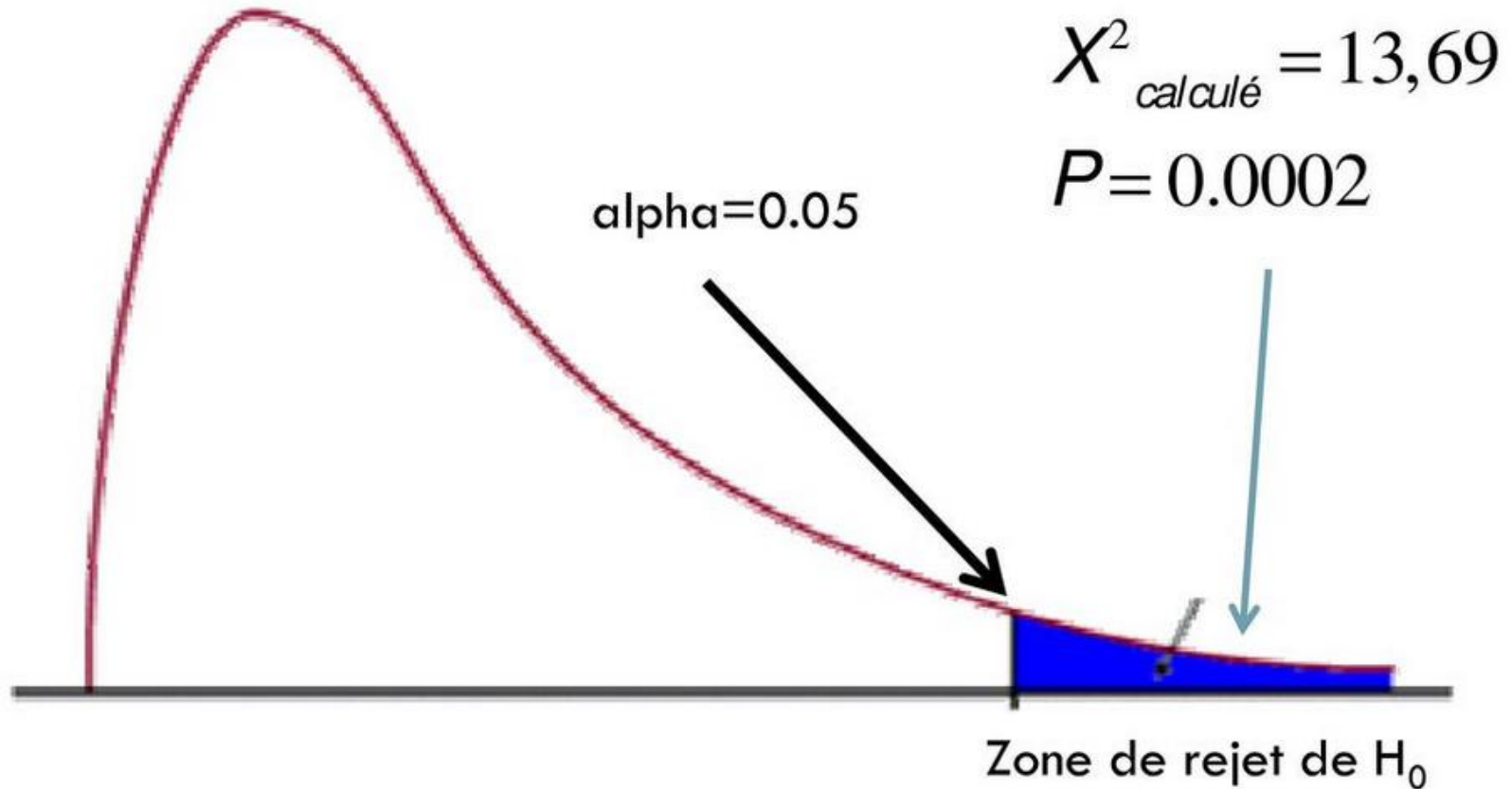
Pour notre exemple :

$$\chi_1^2 = \frac{185^2}{5000} + \frac{185^2}{5000} = 13,69$$

- Degrés de liberté : $n - 1$, où n correspond au nombre de classes.
Ici, $2 - 1 = 1$ degré de liberté.

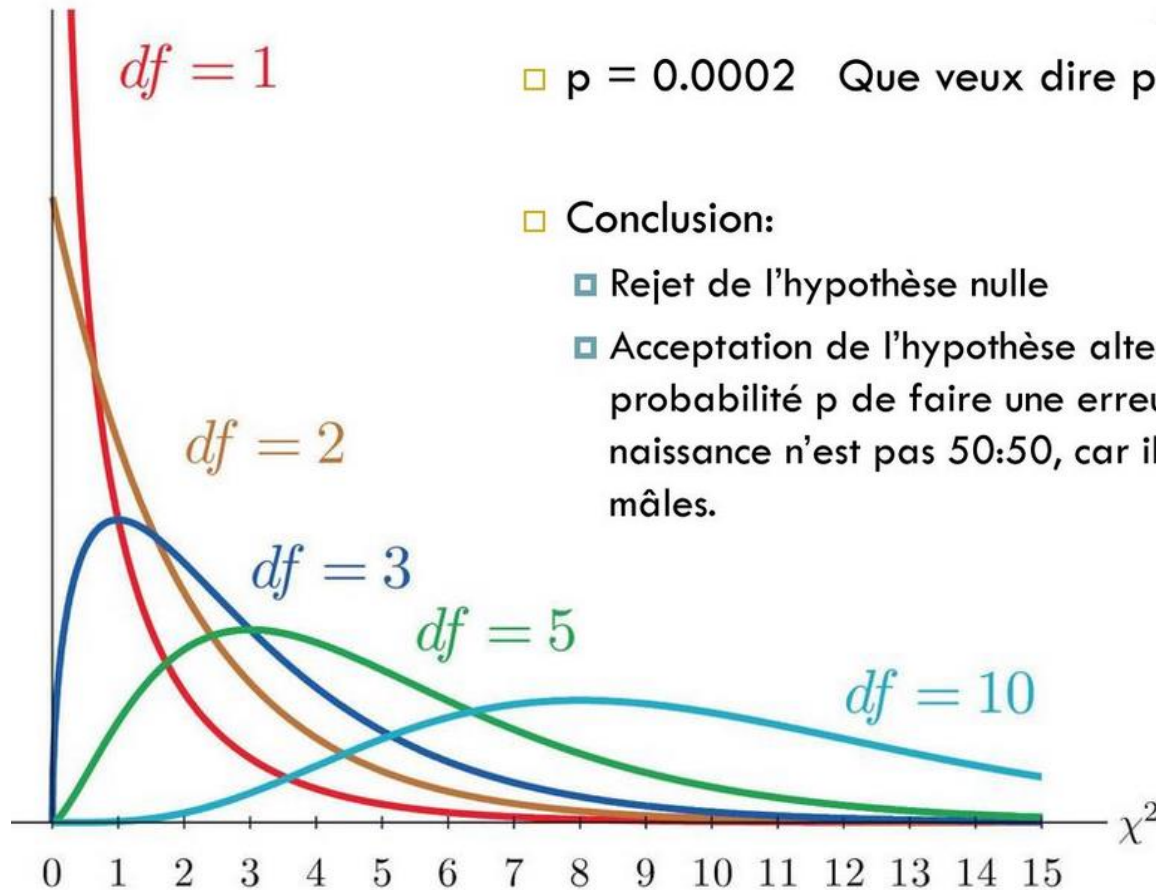
Khi-carré : Exemple Simple et zone de rejet

49/69



Khi-carré : ddl et Significativité

50/69



□ $p = 0.0002$ Que veux dire p ?

□ Conclusion:

- Rejet de l'hypothèse nulle
- Acceptation de l'hypothèse alternative avec une probabilité p de faire une erreur: Le sex-ratio humain à la naissance n'est pas 50:50, car il y a une prédominance de mâles.

Exemple 2

51/69

- **Question** : La fréquence des sexes des cerfs de Virginie est-elle la même dans différents secteurs ?
 - H_0 : La fréquence des sexes des cerfs de Virginie est la même dans différents secteurs.
 - H_1 : La fréquence des sexes des cerfs de Virginie diffère dans différents secteurs.

Secteur	Mâles	Femelles
Anticosti	30	34
Cantons de l'Est	55	25
Laurentides	12	4

Quelle est la fréquence des femelles dans les cantons de l'Est?

Exemple 2

52/69

Secteur	Mâles	Femelles	Total
Anticosti	30	34	64
Cantons de l'Est	55	25	80
Laurentides	12	4	16
Total	97	63	160

Calcul de la **fréquence attendue** pour chaque cellule :

Pour chaque cellule :

$$\text{Fréquence attendue} = \frac{\text{Somme ligne} \times \text{Somme colonne}}{\text{Abondance totale (N)}}$$

Exemple : Pour les mâles attendus à Anticosti :

$$97 \times \frac{64}{160} = 38,8 \text{ mâles à Anticosti}$$

Exemple 2

53/69

Observation

Secteur	Mâles	Femmes
Anticosti	30	
Cantons de l'Est	55	
Laurentides	12	
Total	97	



Théorie

	Mâles	Femelles
Anticosti	?	?
Cantons de l'Est	?	?
Laurentides	?	?

- Le test du Khi-carré nécessite des effectifs raisonnables dans chacune des catégories (pas moins de 5 observations par cellule)

$$\chi^2 = 8.67 > 5.991$$

Degrés de liberté = $(n-1)(m-1)$

Dans l'exemple présent: $(3-1)(2-1) = 2$

$p = 0.0131$ on rejette?

moins un.

- Si on test la relation entre deux variables catégoriques (ex. région vs sexe des cerfs) le nombre de degrés de liberté devient $(n-1)(m-1)$
 - Où n est le nombre de catégories d'une des variables et m le nombre de catégories de l'autre.

Plan

54/69

- 👉 Corrélation
- 👉 Régression Linéaire
- 👉 Significativité
- 👉 Khi-deux d'ajustement
- 👉 Khi-deux d'indépendance
- 👉 Exercices

Exercices Sur les Probabilités

55/69

Dans une population de 1000 individus, 120 sont atteints d'une maladie :

- Quelle est la probabilité pour qu'un individu tiré au hasard soit atteint de cette maladie?
- On observe un petit échantillon de 5 sujets (tiré au hasard avec remise).
Quelle la probabilité que les 5 sujets soient indemnes de la maladie ?



$$- P(M) = \frac{120}{1000} = 0,12 = p$$

$$- P(\bar{M} \cap \bar{M} \cap \bar{M} \cap \bar{M} \cap \bar{M}) = (1 - p)^5 = 0,528$$

Exercices Sur les Probabilités

56/69

La probabilité de naissance d'un garçon est de 0,51. Un couple décide d'avoir 4 enfants. Il désire avoir 3 filles et 1 garçon. En considérant l'absence de possibilité de grossesse multiple, quelles sont les probabilités que le couple réalise son vœu :

- dans le cas où le garçon est le dernier de la fratrie
- dans le cas où le garçon occupe un rang quelconque de la fratrie



Soit $P(G) = 0,51$ et $P(F) = 0,49$

- $P_1 = P(F \cap F \cap F \cap G) = P(F) \times P(F) \times P(F) \times P(G) = 0,49^3 \times 0,51 = 0,06$
- $P_2 = P(G \cap F \cap F \cap F) + P(F \cap G \cap F \cap F) + P(F \cap F \cap G \cap F) + P(F \cap F \cap F \cap G)$
 $= 4 \times P_1 = 0,24$

Exercices Sur les Probabilités

57/69

Un dépistage systématique concernant un éventuel trouble de l'audition est effectué à la naissance. On sait que 2 % des nouveaux-nés présentent des troubles de l'audition. Ce dépistage commence par un test donnant 95 % de résultats positifs pour les nouveaux-nés atteints de ces troubles et 6 % de résultats positifs pour les bébés indemnes de ces troubles.

Quelle est la probabilité qu'un nouveau-né pris au hasard soit atteint de ces troubles sachant que le test a donné un résultat positif ?

Quelle est la probabilité qu'un nouveau-né pris au hasard soit indemne de ces troubles sachant que le test a donné un résultat négatif ?



A : le nouveau-né est atteint d'un trouble de l'audition. $P(A) = 0,02$

B : le test est positif. On ne connaît pas $P(B)$ mais on connaît les deux probabilités conditionnelles suivantes : $P(B/\bar{A}) = 0,06$ et $P(B/A) = 0,95$

On cherche $P(A/B)$. Pour cela, on utilise la formule de Bayes :

$$\text{VPP} \Rightarrow P(A/B) = \frac{P(B/A) \times P(A)}{P(B)} = \frac{P(B/A) \times P(A)}{P(B/A)P(A) + P(B/\bar{A})P(\bar{A})} = \frac{0,95 \times 0,02}{0,95 \times 0,02 + 0,06 \times 0,98} = 0,244$$

$$\text{VPN} \Rightarrow \text{De la même façon, } P(\bar{A}/\bar{B}) = \frac{P(\bar{B}/\bar{A}) \times P(\bar{A})}{P(\bar{B}/\bar{A})P(\bar{A}) + P(\bar{B}/A)P(A)} = \frac{0,94 \times 0,98}{0,94 \times 0,98 + 0,05 \times 0,02} = 0,998$$

Exercice Sur la Corrélation

58/69

On mesure le poids de $N = 200$ pandas vivant à différentes altitudes.



Poids (kgs)	70-80	80-90	90-100	100-110	110-120
Altitude (m)					
1500-2000	28	5			
2000-2500	13	29	24	12	
2500-3000	6	12	21	12	5
3000-3500		2	10	16	5
	47	48	55	40	10



En notant X l'altitude et Y le poids, les calculs donnent

$$\bar{x} = 2472.5 \quad s_X^2 = 226743.75 \quad \bar{y} = 90.9 \quad s_Y^2 = 141.19$$

$$s_{XY} = 3587.34 \quad r = 0.634 \quad r^2 = 0.402$$

Existe-t-il un lien entre le poids et l'altitude où vivent les pandas ?

$$\bar{y} = \frac{75 \times 47 + 85 \times 48 + 95 \times 55 + 105 \times 40 + 115 \times 10}{200} = 90,9 \text{ kg}$$

$$s_Y^2 = \frac{47 \times (75 - 90,9)^2 + \dots + 10 \times (115 - 90,9)^2}{200} = 141,19 \text{ kg}^2$$

$$s_{XY} = \frac{28 \times (1750 - 2472,5) \times (75 - 90,9) + \dots + 5 \times (3250 - 2472,5) \times (115 - 90,9)}{200} = 3587,34$$

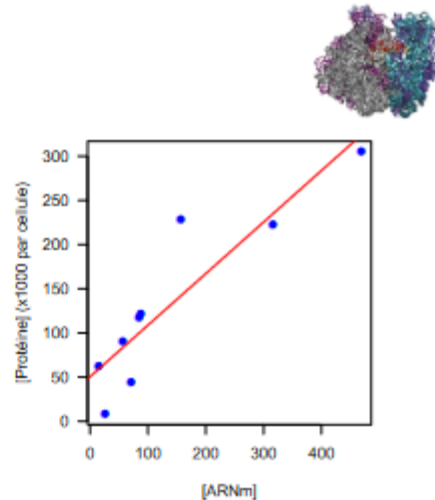
$$t_{obs} = \frac{0.634 \sqrt{198}}{\sqrt{1 - 0.402}} = 11.536$$

Pour 198 ddl et un risque $\alpha = 0.05$, $t_{seuil} = 1.96$, on rejette H_0 et on accepte H_1 : il existe une relation linéaire entre le poids des pandas et l'altitude à laquelle ils vivent.

Exercice sur la Régression

59/69

ARNm X (/cellule)	Protéine Y (x1000/cellule)
87	121
156	228
56	90
70	44
14	62
25	8
468	305
84	117
315	222



$$n = 9, \bar{x} = 141.667, \bar{y} = 133$$

$$s_X^2 = 20604.667, s_{XY} = 12022.889$$

$$\hat{a} = 0.584, \hat{b} = 50.266$$

$$Y = 0.584X + 50.266$$

$$\hat{\sigma}_R^2 = \frac{1}{N-2} \sum_{i=1}^N e_i^2 = 2032.5$$

$$t_{obs} = \left| \frac{0.584}{\sqrt{\frac{2032.5}{9 \times 20604.667}}} \right| = 5,579$$

$$t_{obs} > t_{7,0.05} = 2.365$$

Donc Significativité +++
(pour un seuil de 0,05)

Donner l'équation de la droite de régression
reliant X à Y

Calculer la somme des carrés des résidus

Faire un test de la pente

Exercice sur la Comparaison des Moyennes

60/69

Hommes	Femmes
1,88	1,91
1,85	3,02
2,66	2,75
1,42	1,79
1,57	1,98
1,45	1,48
2,53	1,41
2,74	2,11
2,97	2,72
2,47	2,13
2,73	2,78
2,00	2,71
1,70	1,37
2,94	1,97
2,90	2,82
2,98	1,49
3,03	1,55
2,21	1,86
2,38	2,47
2,97	2,37
1,38	1,52
1,80	2,19
1,98	2,17
1,72	3,08
2,15	2,12
2,56	2,31
2,80	2,69
2,28	2,73
2,85	1,74
2,40	1,43
2,59	3,01
3,07	1,87
1,76	
1,45	

Comparer les concentrations de cholestérol des hommes et des femmes

H0 = pas de différence => on utilise un test de Student non apparié

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Cholesterol	N	Moyenne	Ecart type
Femmes	32	2,18	0,54
Hommes	34	2,30	0,55

$$t = \frac{|2,18 - 2,3|}{\sqrt{\frac{0,54^2}{32} + \frac{0,55^2}{34}}} = 0,894 \ll 1,96$$

On ne peut pas rejeter H0 => pas significatif

Seuil 0,05		
Student	TEST.STUDENT(matrice1,matrice2,uni/bilatéral,type)	
0,357	p_value=T.TEST(B5:B38;C5:C36;2;2)	On Garde H0, moyennes identiques
0,357	p_value=T.TEST(B5:B38;C5:C36;2;3)	
0,186	p_value=T.TEST(B5:B36;C5:C36;2;1)	
Fisher	TEST.F(matrice1;matrice2)	On Garde H0, variances identiques
0,879	p_value=F.TEST(B5:B38;C5:C36)	
Correlation	COEFFICIENT.CORRELATION(matrice1, matrice2)	Pas significatif
0,148	r=COEFFICIENT.CORRELATION(B5:B36;C5:C36)	
	Regression y=m.x+b	DROITEREG(y_connus, [x_connus], [constante], [statistiques])
m value	0,148363387=DROITEREG(B5:B36;C5:C36;VRAI;VRAI)	
b value	0,181062081=DROITEREG(B5:B36;C5:C36;VRAI;VRAI)	

Exercice sur Le χ^2 d'Ajustement

61/69

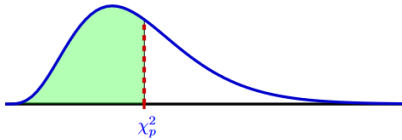
Le tableau suivant montre la distribution des unités 0,1,2, ..., 9 d'une table de nombres aléatoires comportant 250 nombres.

Unités	0	1	2	3	4	5	6	7	8	9
Fréq Obs	17	31	29	18	14	20	35	30	20	36
Fréq Est.	25	25	25	25	25	25	25	25	25	25

Tableau C

Percentiles de la distribution du χ^2 . Valeurs de χ^2_p correspondant à P

$\alpha = 1 - p$



df	$\chi^2_{0.005}$	$\chi^2_{0.01}$	$\chi^2_{0.025}$	$\chi^2_{0.05}$	$\chi^2_{0.1}$	$\chi^2_{0.9}$	$\chi^2_{0.95}$	$\chi^2_{0.975}$	$\chi^2_{0.99}$	$\chi^2_{0.995}$
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

Est-ce que la distribution observée est significativement différente de la distribution théorique?



$$\chi^2 = \frac{(17-25)^2}{25} + \dots + \frac{(36-25)^2}{25} = 23.3$$

$23,3 > 16,92 \Rightarrow$ Cette table de nombres aléatoires est suspecte / H_0 est rejeté

Exercice 1 sur Le χ^2 d'Indépendance

62/69

Le tableau suivant montre les fréquences observées concernant la guérison d'une maladie cutanée pour des patients ayant utilisés un traitement spécifique (sérum)

	Guérit	Ne guérit pas	Total
Groupe A (serum)	75	25	100
Groupe B (sans sérum)	65	35	100
Total	140	60	200

Le traitement utilisé est-il utile à la guérison ?



Fréquences attendues sous H_0	Guérit	Ne guérit pas	Total
Groupe A (serum)	70	30	100
Groupe B (sans sérum)	70	30	100
Total	140	60	200

$$\chi^2 = \frac{(75-70)^2}{70} + \frac{(65-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(35-30)^2}{30} = 2.38$$

$$\nu = (h-1)(k-1) = 1; \chi_{0.95}^2 = 3.84$$

Impossibilité de rejeter H_0 / le sérum n'est pas significativement efficace

Exercice 2 sur Le χ^2 d'Indépendance

63/69

Tableau de contingence du nombre de joueurs de hockey de différentes nationalités utilisant différentes marques de bâtons de hockey.

	Canadien	Américain	Russe	Tchèque	Suédois	Finlandais	Somme
Sherwood	32	14	9	5	13	6	79
Titan	14	6	6	4	11	3	44
Koho	15	7	9	7	12	6	56
Nike	8	18	7	3	4	2	42
Easton	22	17	11	2	13	1	66
Somme	91	62	42	21	53	18	287

Le choix de la marque du bâton de hockey que les joueurs utilisent est-il influencé par l'origine du joueur?

H_0 : il n'y a pas de préférence de marque de bâton de hockey chez les joueurs de différentes nationalités

H_1 : les joueurs de différentes nationalités ont des préférences différentes au niveau de la marque de bâton de hockey qu'ils utilisent

$$\chi^2_{(0,05, 20)} = 31,41$$

Exercice 2 sur Le χ^2 d'Indépendance

64/69

	Canadien	Américain	Russe	Tchèque	Suédois	Finlandais	Somme
Sherwood	32	14	9	5	13	6	79
Titan	14	6	6	4	11	3	44
Koho	15	7	9	7	12	6	56
Nike	8	18	7	3	4	2	42
Easton	22	17	11	2	13	1	66
Somme	91	62	42	21	53	18	287

$$ddl = 4 \times 5 = 20$$



exemple, la première cellule : $f_{th(1,1)} = \frac{91 \times 79}{287} = 25,05$ $\chi^2 = \sum_{i=1}^l \sum_{j=1}^c \frac{(f_{i,j} - f_{th(i,j)})^2}{f_{th(i,j)}} = 30,694$

	Canadien	Américain	Russe	Tchèque	Suédois	Finlandais	Somme
Sherwood	25,05	17,07	11,56	5,78	14,59	4,95	79
Titan	13,95	9,51	6,4	3,22	8,13	2,76	44
Koho	17,76	12,1	8,2	4,1	10,34	3,51	56
Nike	13,32	9,07	6,15	3,07	7,76	2,63	42
Easton	20,93	14,26	9,66	4,83	12,19	4,14	66
Somme	91	62	42	21	53	18	287

On ne rejette pas H_0
au seuil $\alpha = 0,05$ car
 $\chi^2_{\text{calc}} < \chi^2_{(0,05, 20)}$

Pas de liens entre
nationalités et
marque de bâtons

Calcul des fréquences théoriques

Exercice 3 sur Le χ^2 d'Indépendance

65/69

On se demande si la localisation d'un cancer de l'estomac est liée ou non au sexe du malade. Pour cela on dispose des renseignements suivants :

localisation	pylore	estomac	cardia
Hommes	53	66	75
Femmes	48	33	23

La localisation et le sexe sont-ils liés ?

Localisation	Pylore	Estomac	Cardia	Tot
Hommes	53 (66)	66 (65)	75 (64)	194
Femmes	48 (35)	33 (34)	23 (34)	104
Tot	101	99	98	298



ddl=2

Calcul des fréquences théoriques

$$\chi^2 = 2,56 + 0,015 + 1,89 + 4,83 + 0,03 + 3,56 = 12,9 > \chi^2_{(0.05,2)} (=5,91)$$

On rejette H_0 au seuil $\alpha = 0,05$ car $\chi^2_{\text{calc}} > \chi^2_{(0.05,2)}$

On conclut que la localisation précise d'un cancer de l'estomac est lié au sexe du malade

Exercice de Synthèse

66/69

L'ion thiocyanate (SCN^-) est un marqueur biologique reconnu de l'intoxication chronique par le tabac. Sa concentration sérique en ($\mu\text{mol/L}$) a été évaluée chez 35 volontaires non fumeurs (X) et chez 35 volontaires fumeurs (Z). Les résultats sont les suivants :

$$\sum x_i = 1347,5 \quad \sum x_i^2 = 54713,75 \quad \sum z_i = 4550 \quad \sum z_i^2 = 653240$$

Calculer les moyennes et écart-types de ces deux échantillons.

Non fumeurs	Fumeurs
n=35	n=35
$\mu_x = 1347,5/35 = 38,5 \mu\text{mol/L}$	$\mu_z = 4550/35 = 130 \mu\text{mol/L}$
$s_x = \sqrt{\frac{54713,75}{35} - 38,5^2} = 9 \mu\text{mol/L}$	$s_z = \sqrt{\frac{653240}{35} - 130^2} = 42 \mu\text{mol/L}$

König-Huygens $\rightarrow s_x^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2$



Les deux moyennes diffèrent-elles statistiquement ?

H_0 : il n'y a pas de différence entre fumeur et non fumeur

H_1 : il y a une différence

$$t = \frac{|\mu_x - \mu_z|}{\sqrt{\frac{s_x^2}{n} + \frac{s_z^2}{n}}} = \frac{38,5 - 130}{\sqrt{\frac{81}{35} + \frac{1764}{35}}} = 12,6 \gg 1,96$$

Rejet de H_0 avec un risque très faible ($p < 10^{-9}$) $\Rightarrow$ différence significative

Estimer l'IC95% de la moyenne μ de la concentration en SCN^- dans la population des fumeurs.

$$\mu_z \pm 1,96 \cdot \frac{s_z}{\sqrt{n}} = 130 \pm 13,9 \rightarrow IC[95\%] = [116 ; 143,9]$$

Exercice de Synthèse

67/69

L'ion thiocyanate (SCN^-) est un marqueur biologique reconnu de l'intoxication chronique par le tabac. Sa concentration sérique en ($\mu\text{mol/L}$) a été évaluée chez 35 volontaires non fumeurs (X) et chez 35 volontaires fumeurs (Z). Les résultats sont les suivants :

$$\sum x_i = 1347,5 \quad \sum x_i^2 = 54713,75 \quad \sum z_i = 4550 \quad \sum z_i^2 = 653240$$

Le taux (Y) de cet ion a été évalué dans la salive des 35 fumeurs. Les résultats sont tels que :

$$\sum y_i = 144,55 \quad \sum y_i^2 = 710,4 \quad \sum y_i z_i = 20730,3$$

Les variables Z et Y sont-elles corrélées ?

$$\mu_Y = 4,13$$

$$s_Y = 1,8$$

H_0 : il n'y a pas de corrélation entre Y et Z
 H_1 : il y a une corrélation

König-Huygens $\rightarrow \text{cov}(Y, Z) = \left(\frac{1}{n} \sum_{i=1}^n y_i \cdot z_i\right) - \bar{y}\bar{z}$

$$r = \frac{\text{cov}(Y, Z)}{\sqrt{\text{var}(Y) \cdot \text{var}(Z)}} = \frac{\frac{20730,3}{35} - 4,13 \times 130}{1,8 \times 42} = 0,733$$

On fait un test de Student sur r avec $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,733 \times 5,74}{\sqrt{1-0,733^2}} = 6,2$



La lecture table pour un $\alpha=0,05$ vaut $t_{\text{table}}=2,042$ (ddl=33 ~30)

$t \gg t_{\text{table}}$ donc on rejette H_0 , et il y a bien une corrélation entre ces deux grandeurs ($p < 0,05$)

Exercice de Synthèse

68/69

L'ion thiocyanate (SCN^-) est un marqueur biologique reconnu de l'intoxication chronique par le tabac. Sa concentration sérique en ($\mu\text{mol/L}$) a été évaluée chez 35 volontaires non fumeurs (X) et chez 35 volontaires fumeurs (Z). Les résultats sont les suivants :

$$\sum x_i = 1347,5 \quad \sum x_i^2 = 54713,75 \quad \sum z_i = 4550 \quad \sum z_i^2 = 653240$$

Le taux (Y) de cet ion a été évalué dans la salive des 35 fumeurs. Les résultats sont tels que :

$$\sum y_i = 144,55 \quad \sum y_i^2 = 710,4 \quad \sum y_i z_i = 20730,3$$

Quelle est l'équation de la droite des moindres carrés ($Y=b_0+b_1.Z$) qui s'ajuste aux couples de points (z_i, y_i) ?

$$b_1 = \frac{\text{cov}(Y,Z)}{s_Z^2} = \frac{55,4}{1764} = 0,0314 \text{ et } b_0 = \mu_Y - b_1 \cdot \mu_Z = 4,13 - 0,0314 \times 130 = 0,048 \rightarrow Y = 0,048 + 0,0314xZ$$

Faire la prévision de concentration salivaire en SCN^- d'un fumeur ayant une concentration sérique de $200\mu\text{mol/L}$

$$Y = 0,048 + 0,0314 \times 200 = 6,33\mu\text{mol/L}$$

Proposer un intervalle de prévision à 95% de cette concentration salivaire si la variance de l'erreur de prévision (résidu) est 0,75 ?

$$IC[95\%] = 6,33 \pm 2\sqrt{0,75} = [4,6; 8,06]$$



MERCI POUR VOTRE ATTENTION

69/69

