A dark blue vertical bar runs down the left side of the page. A blue arrow points to the right from this bar, containing the date.

28/05/2018

Statistiques

Psychologie, première année

Several thin, curved lines in dark blue and light grey originate from the bottom left corner and sweep upwards and to the right.

Sugeidy Rachel Rodriguez

Table des matières

Cours 1.....	1
Cours 2.....	1
2. Représentation graphique	1
2.1 Variable quantitative.....	1
2.1.1 Variable Discrète.....	1
2.1.2 Variable continue ou assimilée	2
Cours 3.....	3
2.2 Variable qualitative.....	3
2.2.1 Diagramme en tuyaux d'orgue.....	3
2.2.2 Diagramme en camembert	4
3. Description numérique	4
3.1 Caractéristique de position	4
3.1.1 Le mode	4
3.1.2 La médiane	5
3.1.3 La moyenne.....	6
3.2 Caractéristique de dispersion	6
3.2.1 Intervalle interquartile.....	6
Cours 4.....	7
Cours 5.....	9
3.2.2 Variance et écart-type.....	9
3.3 Caractéristique de forme.....	10
3.3.1 Coefficient d'asymétrie (Skewness).....	10
3.3.2 Coefficient d'aplatissement (Kurtosis)	10
Cours 6.....	11
4. Description bidimensionnelle et mesure de liaison.....	11
4.1 Liaison entre deux variables nominales (χ^2 , C, ϕ , V)	11
Cours 7.....	13
4.2 Liaison entre deux variables ordinales.....	14
4.2.1 Gamma de Goodman-Kruskal (γ).....	15
4.2.2 Tau b de Kendall (τ_b).....	15
Cours 8.....	15
4.3 Liaison entre deux variables numériques	15
4.3.1 Etude variable de la corrélation	15
4.3.2 Coefficient de corrélation linéaire de Bravais-Pearson (r)	16
4.3.3 Coefficient de corrélation de Spearman (r_s)	16
Cours 9.....	18
5. Théorie élémentaire des probabilités.....	18
5.1 Lancer d'un dé.....	18
5.2 Lancer d'une pièce quelques fois	19
Cours 10.....	20
5.3 Lancer d'une pièce un grand nombre de fois	21
Cours 11.....	23
6. Estimation.....	23
6.1 Estimer une proportion	23
6.1.1 Estimation ponctuelle d'une proportion.....	23
6.1.2 Estimation par intervalle de confiance d'une proportion.....	24

Cours 12	25
6.2 Estimer une moyenne.....	25
6.2.1 Estimation ponctuelle d'une moyenne	25
6.2.2 Estimation par intervalle de confiance d'une moyenne	26
7. Test de proportion	26
7.1 Test sur une proportion.....	26
Cours 13	28
7.2 Test sur deux proportions	29
8. Test sur la proportion d'une distribution.....	31
8.1 Test sur une moyenne (Test de Student sur un échantillon)	31
Cours 14	33
8.2 Test de normalité de Shapiro-Wilk (W)	33
8.3 Test de Wilcoxon sur un échantillon (T+, T-)	34
8.3.1 Données sans ex-aequo	34
Cours 15	36
8.3.2 Données avec ex-aequo	36
9. Analyse des mesures paires.....	36
9.1 Approche paramétrique (Test de Student).....	36
9.2 Approche non-paramétrique (Test de Wilcoxon).....	37
Cours 16	38
10. Tests d'hypothèses appliqués à 2 échantillons indépendants	39
10.1.1 Test d'homogénéité des variances de Levene (F)	40
10.1.2 Test de Student sur 2 groupes indépendants	40
Cours 17	41
10.1.3 Test de Welch (T, ddl').....	42
10.2 Approche non-paramétrique (Mann-Whitney U: U1, U2, T1, T2).....	43
Cours 18	44
11. Tests d'hypothèses appliqués à plus de 2 groupes indépendants	45
11.1 Test de la forme de la distribution des résidus (ANOVA).....	45
11.1.1 Méthode gravure.....	45
11.1.2 Méthode inférentielle.....	45
11.2 Analyse de variances à un facteur de classification	45
11.2.1 Données et modèles.....	45
11.2.2 Le test (Sommes des carrés - SC)	46
Cours 19	47
11.2.3 Comparaisons multiples - Méthode de Bonferroni	49
Cours 20	50
11.2.4 Comparaisons multiples - Méthode de Holm.....	51
11.3 Test non-paramétrique (Kruskal-Wallis).....	51
Loi normale centrée réduite 1	53
Loi normale centrée réduite 2	53
Triangle de Pascal	54
Signification des lettres grecques et symboles divers	55

Cours 1

25.09.17

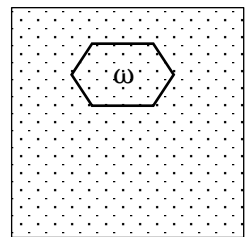
Population : population ou univers statistique (Ω)

Individu : unités statistiques ou individu (ω)*

Recensement : étude de tous les individus d'une partie de la population finie

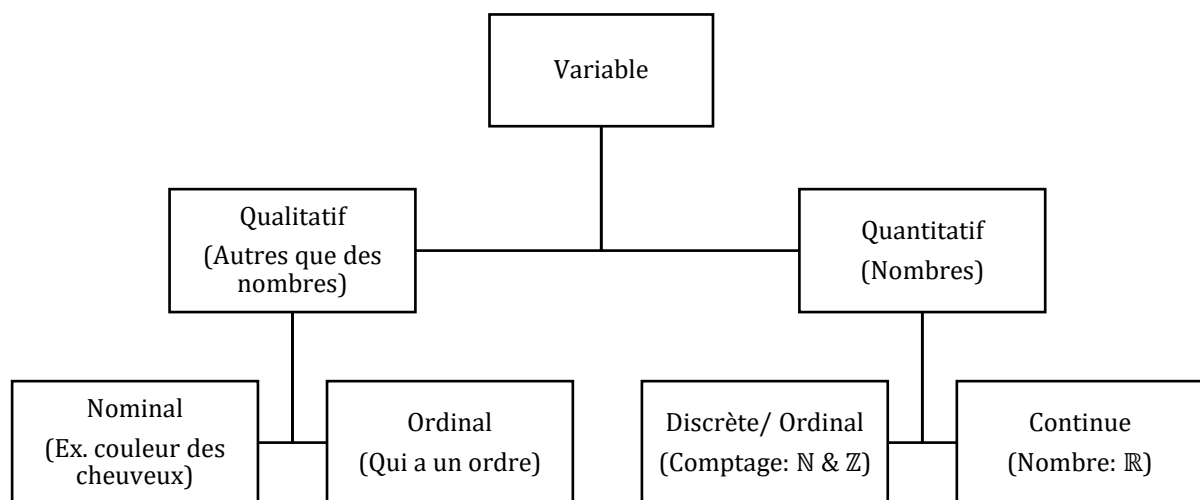
Échantillon : lors d'un sondage, le sous-ensemble étudié (E)

Variable : chaque individu d'une population est décrit par un ensemble de caractéristique appelées variables ou caractère (C)



Univers Ω

Type de variable :



Cours 2

02.10.17

2. Représentation graphique

2.1 Variable quantitative

2.1.1 Variable Discrète

Tableau des données regroupées : (douleurs chroniques)

j	1	2	3	4	5	6	7	8	9	
x_j	2	3	4	5	6	7	8	9	10	
n_j	1	10	10	6	13	19	10	2	1	= 72
F_j	$1 \div 72$	$10 \div 72$	$10 \div 72$	$6 \div 72$	$13 \div 72$	$19 \div 72$	$10 \div 72$	$2 \div 72$	$1 \div 72$	
F_j [%]	1.39%	13.89%	13.89%	8.33%	18.06%	26.39%	13.89%	2.78%	1.39%	= 100%

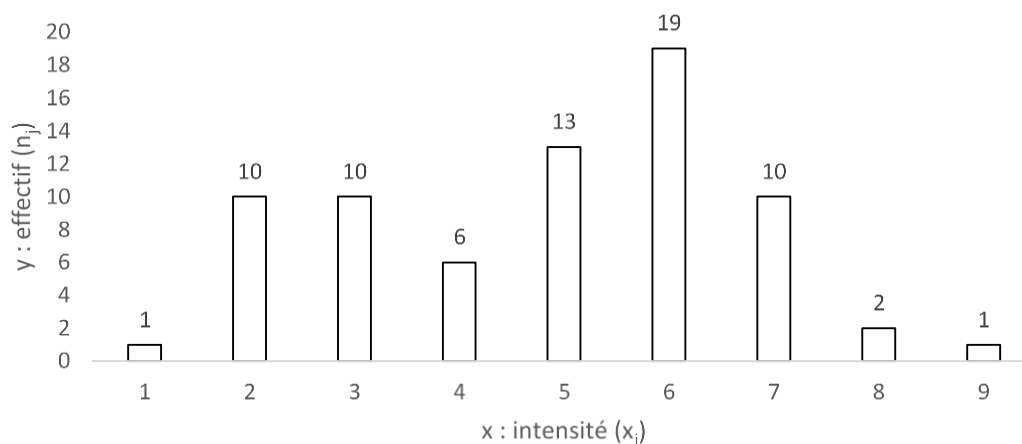
j : je numérote x_j : score ordonné n_j : effectif $F_j : n_j \div n$ = chiffre en %

Somme des fréquences :

$$\sum_{j=1}^9 f_j = \sum_{j=1}^9 \frac{n_j}{n} = \frac{1}{n} \sum_{j=1}^9 n_j = \frac{1}{n} \times n = 1$$

* Tableau des observations : ligne \rightarrow unité statistique, colonne \rightarrow variable

Diagramme en bâtons :



2.1.2 Variable continue ou assimilée

Assimilée : Variable pas continue, mais considérée comme telle pour cet usage, car il y a trop de données pour un diagramme en bâtons.

Histogramme :

Pour dessiner un histogramme, il faut :

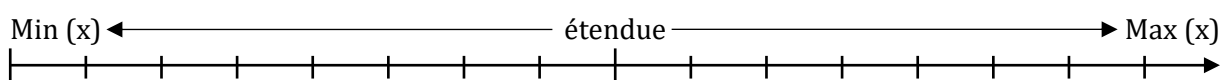
- Construire des classes (trop large : on perd trop d'info, trop étroite : peu de visibilité)

Méthode de Sturges (1926) :

- Calculer l'étendue de la variable (étendue se dit aussi empan) : $\text{étendue}(x) = \max(x) - \min(x)$
Exemple : (selon PP) : $\text{étendue}(\text{âge}) = \max(\text{âge}) - \min(\text{âge}) = 18.26 - 15.33 = 2.93$
- Estimer grossièrement le nombre de classe : $K = \lceil \log_2 n \rceil + 1$
 $\log_2 \times n$: à quelle puissance dois-je élever 2 pour obtenir n ?²

n =	5 – 8	→ K =	4
	9 – 16		5
	17 – 32		6
	33 – 64		7
	65 – 128		8
	129 – 256		9

Calcul de la largeur approximative :



$${}^3L = \frac{\text{étendue}(x)}{K}$$

On doit arrondir au multiple de la largeur : ex. série 2 : $60 < 62 < 70$, alors on choisira 60.

¹ Log arrondi vers le haut. Rappel : $\log_2 \times n = \frac{\log_{10} \times n}{\log_{10} 2}$

² De 5 – 8, il y aura 4 classes, de 9 – 16, il y aura 5 classes, etc.

³ L = largeur d'une classe

- Choisir une largeur de classe qui a la forme de 1 ; 2 ; 5 multiplié par une puissance de 10.

Exemple : $L = \frac{2,93}{8} = 0,366$

$0,2 < 0,366 < 0,5$, alors on choisit toujours la plus grande, donc $L = 0.5$.

Tableau des données regroupées en classes :

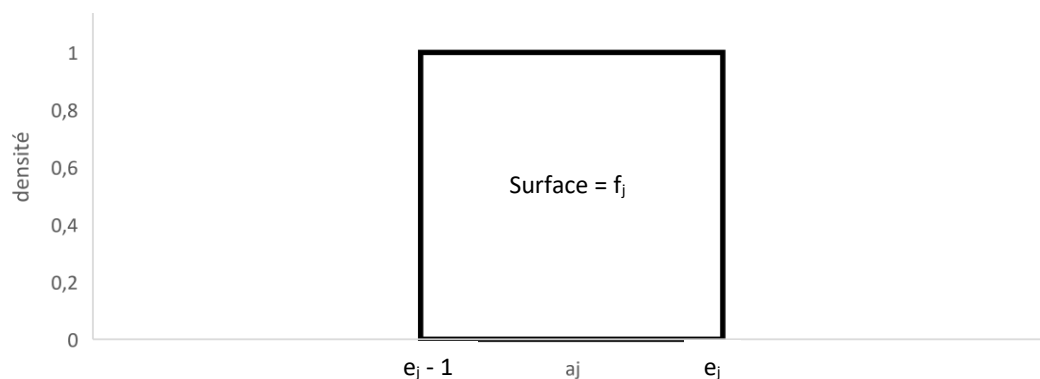
j	âge	n_j	f_j	a_j	$d_j = f_j \div a_j$
1	[15 ;15.5[1	$1 \div 72$	0.5	$2 \div 72$
2	[15.5;16[9	$9 \div 72$	0.5	$18 \div 72$
3	[16 ;16.5[23	$23 \div 72$	0.5	$46 \div 72$
4	[16.5;17[24	$24 \div 72$	0.5	$48 \div 72$
5	[17 ;17.5[9	$9 \div 72$	0.5	$18 \div 72$
6	[17.5;18[2	$2 \div 72$	0.5	$4 \div 72$
7	[18 ;18.5] ⁴	4	$4 \div 72$	0.5	$8 \div 72$

= 72

= 100%

$a_j: e_{j1} - e_j$

$f_j = a_j \times d_j \leftrightarrow d_j = f_j \div a_j$



Cours 3

9.10.17

2.2 Variable qualitative

2.2.1 Diagramme en tuyaux d'orgue

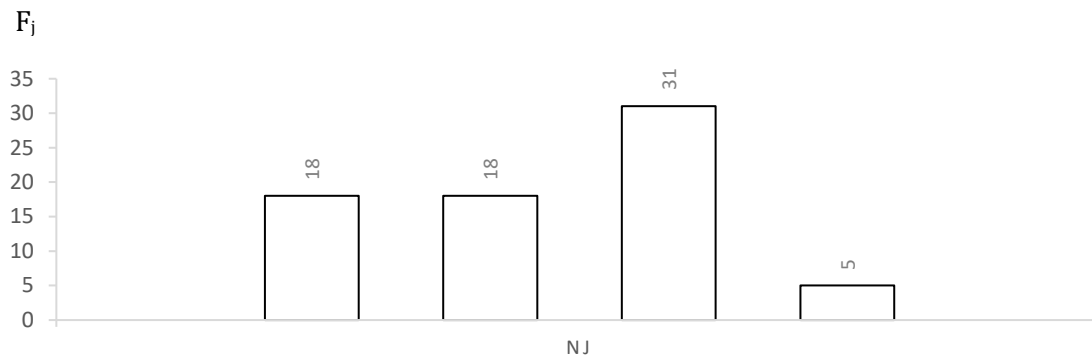
	j = 1	j = 2	j = 3	j = 4
	Tête	Ventre	Dos	Autres
n_j	18	18	31	5
F_j	$18 \div 72$	$18 \div 72$	$31 \div 72$	$5 \div 72$
F_j [%]	25%	25%	43.1%	6.9%

= 72

= 100%

$$n = \sum_{j=1}^4 n_j = 18 + 18 + 31 + 5 = 72$$

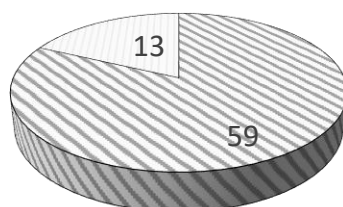
⁴ Le dernier crochet doit être « inclus »



2.2.2 Diagramme en camembert

	Filles	Garçons	
n_j	59	13	= 72
F_j	81.9%	18.1%	= 100%
	$\theta_{\text{filles}} : 360^\circ \times 18.9\%$	$\theta_{\text{garçon}} : 360^\circ \times 18.1\%$	
θ_j	295°	65°	= 360°

$\theta_j = 360^\circ \times f_j$



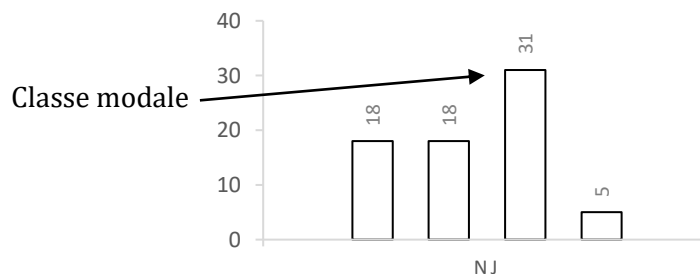
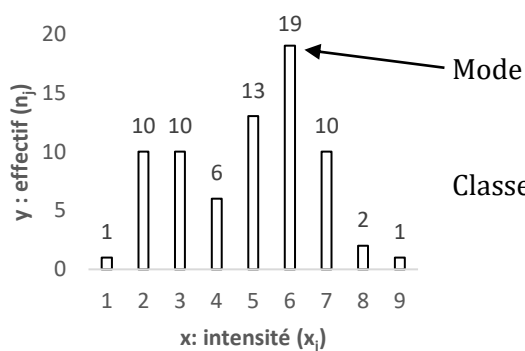
■ Filles ■ Garçons

3. Description numérique

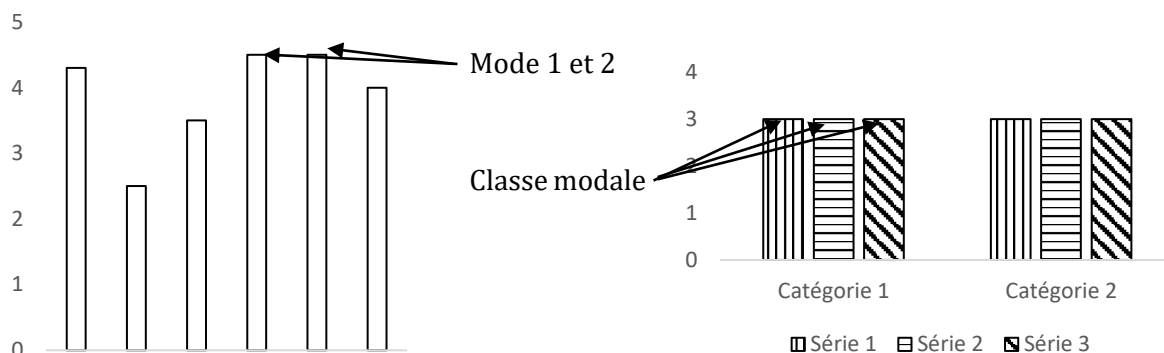
3.1 Caractéristique de position

3.1.1 Le mode

Le mode correspond au maximum de la fonction de distribution. C'est la modalité la plus fréquente d'une variable statistique. Le mode est la valeur la plus fréquente.



Certaines distributions peuvent avoir plusieurs modes et d'autres ne pas en avoir du tout :



3.1.2 La médiane

La médiane d'une variable statistique est la valeur de variable qui partage les individus supposés de manière ordonnée en deux groupes de même taille.

- n est pair : garçons souffrant de maux de tête

	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$
x_i	10	4	7	5	8	6
$x_{[i]}$	4	5	6	7	8	10

$n = 6$ $x_{[i]}$: mis dans l'ordre Médiane $\uparrow = 6.5$

Si n est pair, alors :

$$M = \frac{x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}}{2}$$

Application : $n = 6$

$$x_{[\frac{n}{2}]} = x_{[3]} = 6 \qquad x_{[\frac{n}{2}+1]} = x_{[4]} = 7 \qquad M = \frac{6+7}{2} = 6.5$$

- n est impair : garçons souffrant de maux de dos

	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$
x_i	3	7	6	5	6	5	3
$x_{[i]}$	3	3	5	5	6	6	7

$n = 7$ Médiane $\uparrow = 5$

Si n est impair, alors :

$$M = x_{[\frac{n+1}{2}]}$$

Application : $n = 7$

$$M = x_{[\frac{7+1}{2}]} = x_{[4]} = 5$$

Nous constatons que la médiane est une caractéristique de tendance centrale robuste : contrairement à la moyenne, elle est peu sensible aux valeurs extrêmes et isolées.

3.1.3 La moyenne

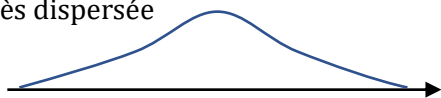
Donnée brutes (série de x_i)							Donnée regroupée ($n = \sum n_j$)				
$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$							$\bar{x} = \frac{\sum_j n_j \times x_j}{n}$				
Exemple 1							Exemple 2				
x_i	10	4	7	5	8	6	x_j	3	5	6	7
n_i	1	1	1	1	1	1	n_j	2	2	2	1
$\bar{x} = \frac{\sum_{i=1}^6 x_i}{n} = \frac{1}{6}(10 + 4 + 7 + 5 + 8 + 6) = 6.\bar{6}$							$\bar{x} = \frac{\sum_j n_j \times x_j}{n} = \frac{1}{7}(3 \times 2 + 5 \times 2 + 6 \times 2 + 7) = 5$				

Résumé :	Mode	Médiane	Moyenne
Nominal	X		
Ordinal	X	X	
Numérique	X	X	X

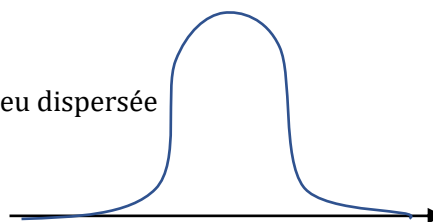
Nous constatons que la médiane est une caractéristique de tendance centrale robuste : contrairement à la moyenne, elle est peu sensible aux valeurs extrêmes et isolées.

3.2 Caractéristique de dispersion

Très dispersée



Peu dispersée

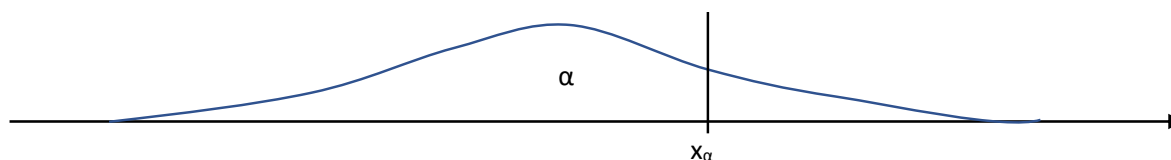


3.2.1 Intervalle interquartile

Un quantile (fractile) d'ordre α correspond à la valeur x_α telle que la proportion des individus qui obtiennent un score inférieur ou égale à x_α est α :

Proportion $P(X \leq x_\alpha) = \alpha$ et $P(X \geq x_\alpha) \geq 1 - \alpha$.

Un quantile d'ordre α est une valeur de x_α au-dessous de laquelle on trouve une proportion α de la population ou de l'échantillon.



Cours 4

16.10.17

La médiane est un quantile : $M = x_{0.5}$

Cela signifie de :

$$P(X \leq M) \geq 0.5 \text{ et } P(X \geq x_\alpha) \geq 1 - 0.5 = 0.5$$

M	4	5	6	6	7	8	10
---	---	---	---	---	---	---	----

Proportion $P(X \leq 6.5) = \frac{3}{6} \geq 0.5$ et $P(X \geq 6.5) \geq \frac{3}{6} - 0.5$

Fonction de répartition :

$$F : {}^5x \mapsto {}^6\alpha$$

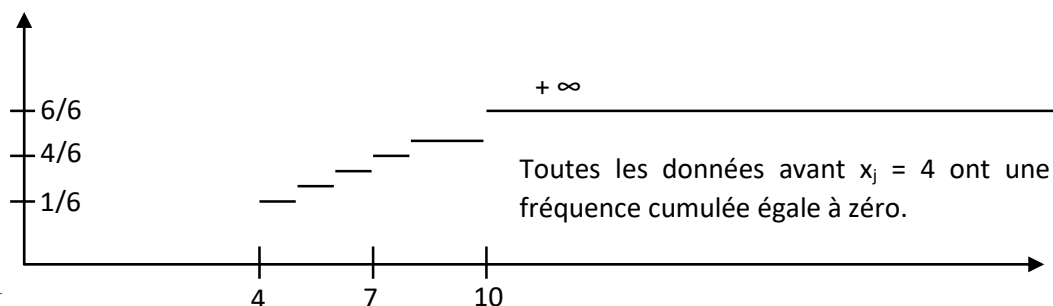
$${}^7Q : \alpha \mapsto x$$

Fréquence cumulée :

$$F(x_j) = \sum_{x_k \leq x_j} f_k$$

$$F(x_j) = \sum_{k \leq j} f_k$$

x	1	2	3	4	5	6
x_j	4	5	6	7	8	10
8f_j	$1 \div 6$	$1 \div 6$	$1 \div 6$	$1 \div 6$	$1 \div 6$	$1 \div 6$
9F_j	$1 \div 6$	$2 \div 6$	$3 \div 6$	$4 \div 6$	$5 \div 6$	$6 \div 6$
	$F_1 = f_1$	$F_2 = f_1 + f_2$	$F_3 = f_1 + f_2 + f_3$	$F_j = f_1 + f_2 + \dots + f_j = 1$		



$$F(0) = 0 ; F(4.66) = 1 \div 6 ; F(12) = 6 \div 6 = 1$$

Exemple : calculons le quantile d'ordre 0.5 : $x_{0.5}$

$$\text{Thalès: } \frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1} \Rightarrow y = y_1 + (x - x_1) \left(\frac{y_2 - y_1}{x_2 - x_1} \right)$$

Exemple 1	Exemple 2
$x_1 = 0.4$ $y_1 = 6$ $x_2 = 0.6$ $y_2 = 7$ $x = 0.5$	$x_1 = 0.2$ $y_1 = 5$ $x_2 = 0.4$ $y_2 = 6$ $x = 0.25$
$y = 6 + \frac{1}{0.2}(0.5 - 0.4) = 6 + 0.5 = 6.5$	$y = 5 + \frac{1}{0.2}(0.25 - 0.2) = 5 + 0.25 = 5.25$

⁵ Quantile

⁶ Probabilité

⁷ Fonction quantile

⁸ Fréquence

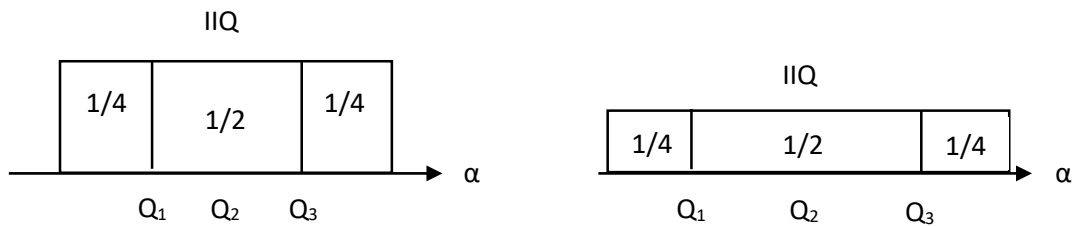
⁹ Fréquence cumulée : toujours comprise entre 0 et 1

Famille de quantiles :

Les quartiles :	Les déciles :	Les centiles :
$Q_1 = x_{0.25}$ $Q_2 = x_{0.5} = M$ (médiane) $Q_3 = x_{0.75}$	$D_1 = x_{0.1}$ $D_2 = x_{0.2}$ $D_5 = x_{0.5} = M = Q_2$ \vdots $D_9 = x_{0.9}$	$C_1 = x_{0.01}$ $C_2 = x_{0.02}$ $C_{50} = x_{0.50} = M = Q_2 = D_5$ \vdots $C_{99} = x_{0.99}$

Intervalle interquartile (IIQ ou IQR) : $IQR = Q_3 - Q_1$

L'intervalle interquartile (ou l'écart interquartile) est égal à la longueur de l'intervalle qui sépare le premier quartile du troisième.



Boîte à moustaches :

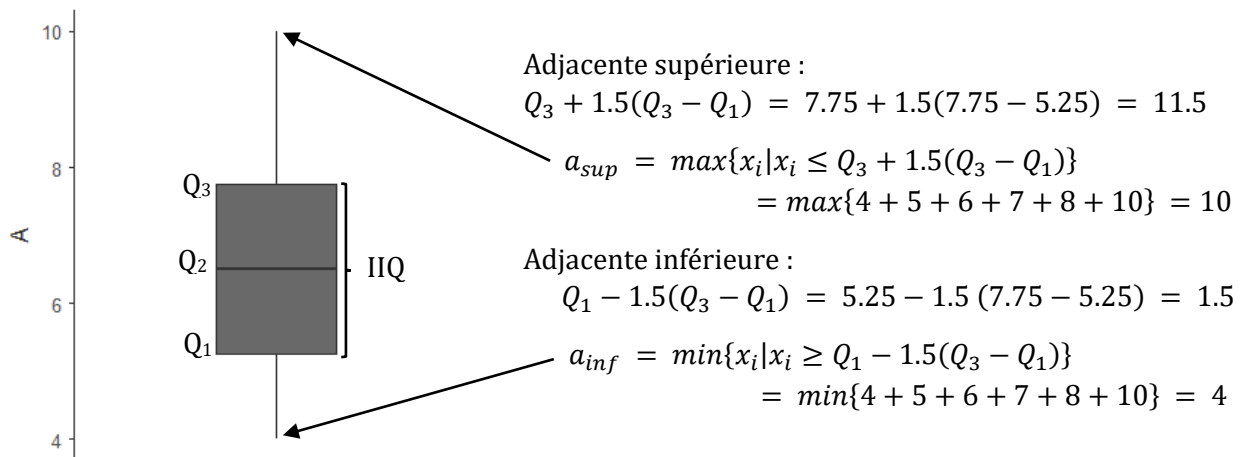
Adjacente inférieure : la plus petite valeur observée supérieure ou égale à $Q_1 - 1.5(Q_3 - Q_1)$:

$$a_{inf} = \min\{x_i | x_i \geq Q_1 - 1.5(Q_3 - Q_1)\}$$

Adjacente supérieure : la plus grande valeur observée inférieure ou égale à $Q_3 + 1.5(Q_3 - Q_1)$

$$a_{sup} = \max\{x_i | x_i \leq Q_3 + 1.5(Q_3 - Q_1)\}$$

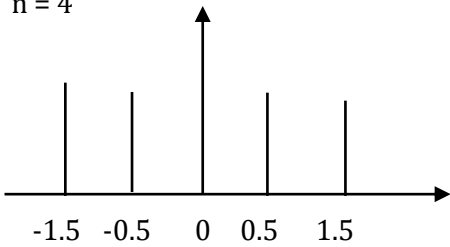
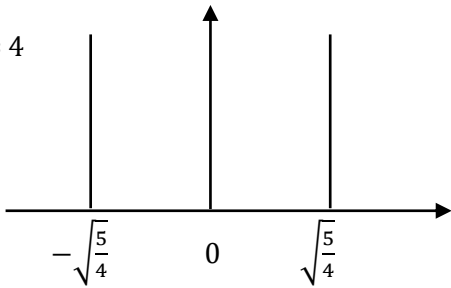
x_j	4	5	6	7	8	10
	$Q_1 = 5.25$	$Q_2 = 6.25$		$Q_3 = 7.75$		



L'intervalle interquartile (IIQ) vaut $Q_3 - Q_1 = 7.25 - 5.25 = 2.5$.

S'il existait une valeur inférieure à 4 ou supérieure à 10, dans le tableau de données, il s'agirait de données extrêmes.

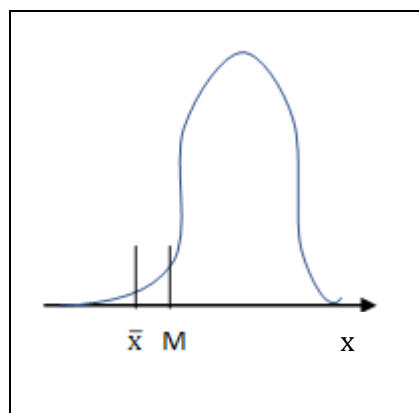
3.2.2 Variance et écart-type

Donnée brutes	Donnée regroupée
s^2 ou $s_x^2 = \text{variance}$	
$s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$ <p>Exemple 1</p>	$s_x^2 = \frac{1}{n-1} \sum_i n_j \times (x_i - \bar{x})^2$ <p>Exemple 2</p>
L'écart-type s_x est la racine carrée de la variance s_x^2 :	
<p>$n = 4$</p>  $\sigma_x^2 = \frac{1}{4} \times \left(\left(\frac{-3}{2} \right)^2 + \left(\frac{-1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 + \left(\frac{3}{2} \right)^2 \right)$ $= \frac{5}{4} = 1,25$ $\sigma_x = \sqrt{\frac{5}{4}} \cong 1.118$ <p>Scores bruts : $-1.5 ; -0.5 ; 0.5 ; 1.5$</p>	<p>$n = 4$</p>  $\sigma_x^2 = \frac{1}{4} \times 4 \left(\sqrt{\frac{5}{4}} \right)^2 = \frac{5}{4} = 1.25$ $\sigma_x = \sqrt{\frac{5}{4}} \cong 1.118$ <p>Scores bruts : $-\sqrt{\frac{5}{4}} ; -\sqrt{\frac{5}{4}} ; \sqrt{\frac{5}{4}} ; \sqrt{\frac{5}{4}}$</p>

3.3 Caractéristique de forme

3.3.1 Coefficient d'asymétrie (Skewness)

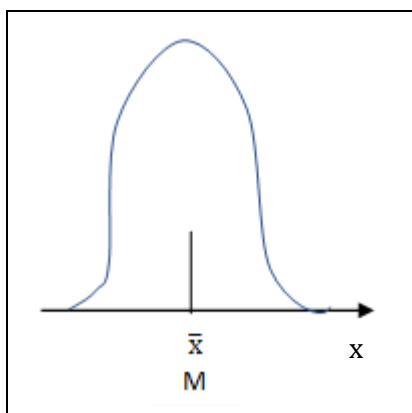
$$g_1 = \frac{m_3}{s^3} = E\left(\left(\frac{x-\mu}{\sigma}\right)^3\right) \rightarrow {}^{10}m_3 = \frac{1}{n}\sum_i(x_i - \bar{x})^3$$



Distribution qui s'étend sur la gauche

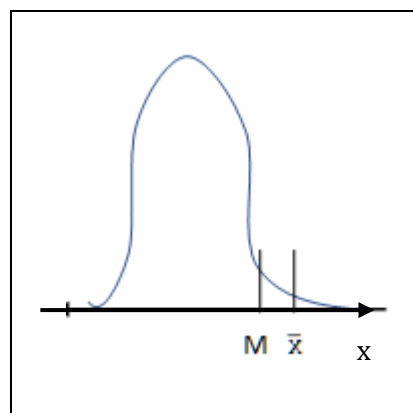
La majorité des observations se trouvent à droite de la moyenne.

$$g_1 < 0$$



Distribution symétrique

$$g_1 = 0$$



Distribution qui s'étend sur la droite

La majorité des observations se trouvent à gauche de la moyenne.

$$g_1 > 0$$

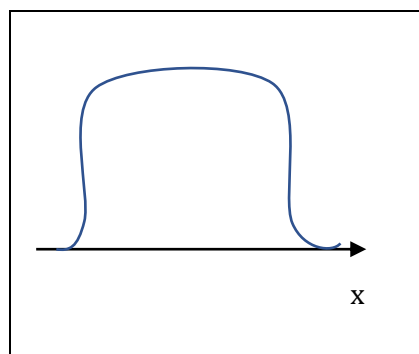
Règles d'interprétation :

Si $g_1 < -1$, alors la distribution est asymétrique à gauche.

Si $-1 \leq g_1 \leq +1$, alors la distribution est symétrique.

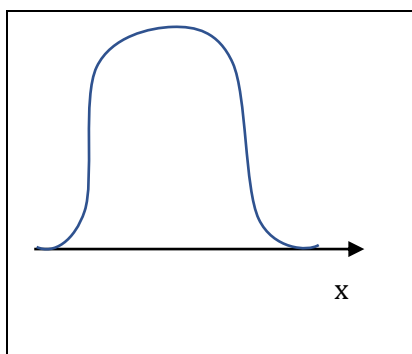
Si $g_1 > 1$, alors la distribution est asymétrique à droite.

3.3.2 Coefficient d'aplatissement (Kurtosis)



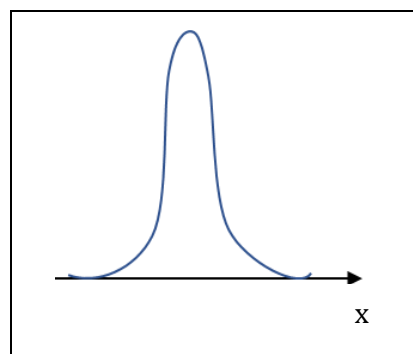
Distribution platycurtique

$$g_2 < 3$$



Distribution normocurtique ou mésocurtique

$$g_2 = 3$$



Distribution leptocurtique

$$g_2 > 3$$

Règles d'interprétation :

Si $g_2 < 2$, alors la distribution est platycurtique.

Si $2 \leq g_2 \leq 4$, alors la distribution mésocurtique.

Si $g_2 > 4$, alors la distribution est leptocurtique.

¹⁰ m_3 : moment centré d'ordre 3, moment centré d'ordre 2 = variance

4. Description bidimensionnelle et mesure de liaison

	Nominal	Ordinal	Numérique
Nominal	4.1 C, φ , V		
Ordinal		4.2 γ , τ_b	
Numérique			4.3 r, r_s

4.1 Liaison entre deux variables nominales (χ^2 , C, φ , V)

Exemple : identification du gagnant

Tableau des données observées :

	n_{ij}	x			
		Son	Vidéo	Son + vidéo	
y	Oui	10	17	9	= 36
	Non	28	18	23	= 69
		= 38	= 35	= 32	= 105

Tables de contingences (tableau croisé) :

	n_{ij}	Colonnes ¹¹			
		a_1	a_2	a_3	
Lignes	b_1	n_{11}	n_{12}	n_{13}	$n_{1\bullet}$
	b_2	n_{21}	n_{22}	n_{23}	$n_{2\bullet}$
		$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	n

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

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

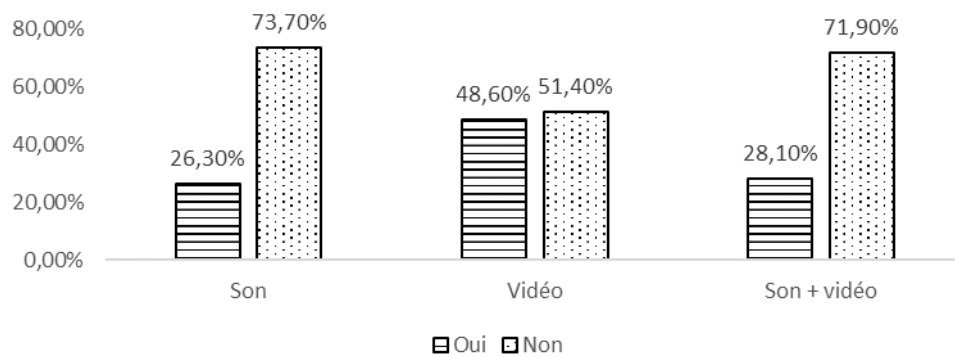
$$n_{\bullet j} = \sum_{i=1}^I n_{ij}$$

La notion de relation ou de dépendance s'exprime en termes de distributions conditionnelles. Une distribution conditionnelle de « Y », c'est une distribution de « Y » limitée à une tranche de la population. Cette tranche est définie par une modalité de la variable « X ».

Distribution conditionnelle de « Y » :

	f_{ij}	x			
		Son	Vidéo	Son + vidéo	
y	Oui	$10 \div 38 = 26.3\%$	$17 \div 35 = 48.6\%$	$9 \div 32 = 28.1\%$	= 100%
	Non	$28 \div 38 = 73.7\%$	$18 \div 35 = 51.4\%$	$23 \div 32 = 71.9\%$	= 100%
		= 100%	= 100%	= 100%	= 100%

¹¹ Méthode Lincoln : lignes – colonnes



Lorsque les variables « X » et « Y » sont indépendantes :

- Les distributions conditionnelles de « Y » sont les mêmes ;
- Les distributions conditionnelles de « X » sont les mêmes ;
- L'effectif associé à la cellule (i ; j) égale à (effectif qu'on observera) :

$$\frac{n_{i\bullet} \times n_{\bullet j}}{n}$$
- La fréquence associée à la cellule (i ; j) est égale au produit des fréquences marginales $f_{\bullet j}$ et $f_{i\bullet}$:

$$f_{ij} = f_{\bullet j} \times f_{i\bullet} \text{ avec } f_{i\bullet} = \frac{n_{i\bullet}}{n} \text{ et } f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

Construisons le tableau des effectifs ou des données théoriques qui correspond à la situation d'indépendance :

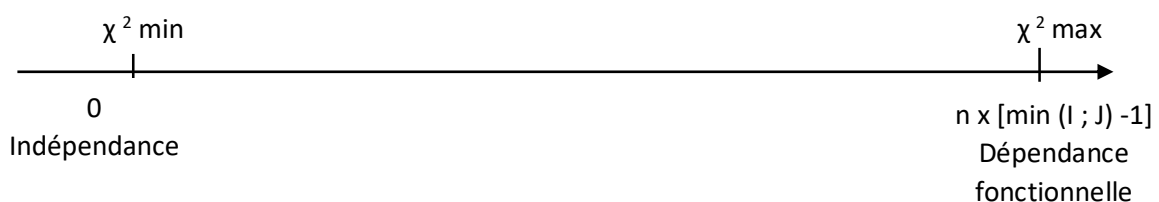
		x			
		Son	Vidéo	Son + vidéo	
y	Oui	$\frac{36 \times 38}{105} \cong 13$	$\frac{36 \times 35}{105} = 12$	$\frac{36 \times 32}{105} \cong 11$	= 36
	Non	$\frac{69 \times 38}{105} \cong 25$	$\frac{69 \times 35}{105} = 23$	$\frac{69 \times 32}{105} \cong 21$	= 69
		= 38	= 35	= 32	= 105

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

On mesure la distance entre les tableaux des n_{ij} et e_{ij} à l'aide d'un χ^2 :

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - e_{ij})^2}{n}$$

$$\chi^2 = \frac{(10 - 13)^2}{13} + \frac{(17 - 12)^2}{12} + \frac{(9 - 11)^2}{11} + \frac{(28 - 25)^2}{25} + \frac{(18 - 23)^2}{23} + \frac{(23 - 21)^2}{21} = 4.781$$



y	x		
	0		
		0	0

Application :

$$\chi^2 \max = n \times [\min^{12}(I; J) - 1] = 105 \times [\min(2; 3) - 1] = 105 \times [2 - 1] = 105$$

Tentons de construire un indice qui ne dépende plus ni de n ni de la dimension de la table de contingence :

<p>Coefficient de contingence :</p> $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$ <ul style="list-style-type: none"> - Ce coefficient = $C < 1$ - En situation d'indépendance : $C = 0$ 	<p>Coefficient phi (φ) :</p> $\varphi = \sqrt{\frac{\chi^2}{n}}$ <ul style="list-style-type: none"> - En situation d'indépendance $\varphi = 0$ - En cas de dépendance fonctionnelle $\varphi \geq 1$ 	<p>Coefficient V de Cramer :</p> $V = \sqrt{\frac{\chi^2}{\chi^2 \max}}$ <ul style="list-style-type: none"> - $\chi^2 = 0 \rightarrow V = 0$ - $\chi^2 = \chi^2 \max \rightarrow V = 1$
---	--	---

Pour que le coefficient de Cramer soit égal à 0, le tableau des données observées doit être identique au tableau des données théoriques. Pour que le coefficient de Cramer soit égal à 1 il faut que la relation entre les variables « X » et « Y » soit fonctionnelle : sur chaque ligne de la table de contingence, seule une cellule ne doit pas être vide.

Cours 7

20.11.17

Pour interpréter les résultats, on peut calculer les résidus (R) :

$$R_{ij} = n_{ij} - e_{ij} :$$

	R _{ij}	x			
		Son	Vidéo	Son + vidéo	
y	Oui	10 - 13 = -3	17 - 12 = 5	9 - 11 = -2	= 0
	Non	28 - 25 = 3	18 - 23 = -5	23 - 21 = 2	= 0
		= 0	= 0	= 0	= 0

Il est aussi possible de calculer des résidus standardisés (RS), c'est comme une distribution de poisson :

$$RS_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} = \frac{R_{ij}}{\sqrt{e_{ij}}} :$$

	RS _{ij}	x			
		Son	Vidéo	Son + vidéo	
y	Oui	$-3 \div \sqrt{13} = -0.839$	$5 \div \sqrt{12} = 1.443$	$-2 \div \sqrt{11} = -0.595$	= 0.009
	Non	$3 \div \sqrt{25} = 0.606$	$-5 \div \sqrt{23} = -1.043$	$2 \div \sqrt{21} = 0.430$	= -0.007
		= -0.233	= 0.4	= -0.165	= 0.002

Un résidu standardisé est important et mérite l'interprétation si sa valeur absolue est supérieure ou égale à 2 \rightarrow important si: $\{|x|\} \geq 2$. Ces valeurs sont celles qui se trouvent à la plus grande distance des effectifs théoriques attendus en cas d'indépendance.




¹² I : nombre de lignes, J : nombre de colonnes. Il faut choisir le plus petit nombre.


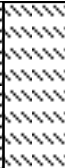

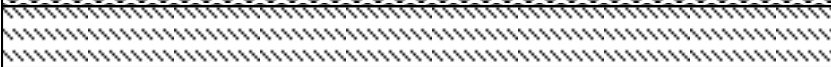
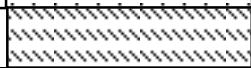



4.2 Liaison entre deux variables ordinales

Satisfaction au travail :

Revenu	Très insatisfait	Insatisfait	Satisfait	Très satisfait
<15'000	1	3	10	6
15'000 – 25'000	2	3	10	7
25'000 – 40'000	1	6	14	12
>40'000	0	1	9	11

La mesure d'association qu'on va développer s'appuie sur le nombre de discordance et de concordance observé entre les différentes paires possibles : $(x_i; y_i)$ et $(x_j; y_j)$.

Une paire est <u>concordante</u> si : $x_i < x_j$, alors $y_i < y_j$. 	Une paire est <u>discordante</u> si : $x_i < x_j$, alors $y_i > y_j$. 	Une paire est <u>ex-aequo</u> si : $x_i = x_j$, alors $y_i = y_j$. 
--	---	---

 Concordance		 Discordance
	n_i	
 Discordance		 Concordance

<p>Calculons le nombre total de concordance :</p> <p>C : $1 \times (3 + 10 + 7 + 6 + 14 + 12 + 1 + 9 + 11) = 73$</p> <p>C : $3 \times (10 + 7 + 14 + 12 + 9 + 11) = 189$</p> <p>C : $10 \times (7 + 12 + 11) = 300$</p> <p>C : $2 \times (6 + 14 + 12 + 1 + 9 + 11) = 106$</p> <p>C : $3 \times (10 + 12 + 9 + 11) = 138$</p> <p>C : $10 \times (12 + 11) = 23$</p> <p>C : $1 \times (1 + 9 + 11) = 21$</p> <p>C : $6 \times (9 + 11) = 120$</p> <p>C : $14 \times (11) = 154$</p> <p>Total : C = 1331</p>	<p>Calculons le nombre total de discordance :</p> <p>D : $3 \times (2 + 1 + 0) = 9$</p> <p>D : $10 \times (2 + 3 + 1 + 6 + 0 + 1) = 130$</p> <p>D : $6 \times (2 + 3 + 10 + 1 + 6 + 14 + 0 + 1 + 9) = 276$</p> <p>D : $3 \times (1 + 0) = 3$</p> <p>D : $10 \times (1 + 6 + 0 + 1) = 80$</p> <p>D : $7 \times (1 + 6 + 14 + 0 + 1 + 9) = 217$</p> <p>D : $6 \times (0) = 0$</p> <p>D : $14 \times (0 + 1) = 14$</p> <p>D : $12 \times (0 + 1 + 9) = 120$</p> <p>Total : D = 849</p>
<p>Calculons le nombre total d'ex-aequo :</p> <p>$n_x = 1 \times (3 + 10 + 6) + 3 \times (10 + 6) + 10 \times (6)$ $+ 2 \times (3 + 10 + 7) + 3 \times (10 + 7)$ $+ 10 \times (7) + 1 \times (6 + 14 + 12)$ $+ 6 \times (14 + 12) + 14 \times (12)$ $+ 0 \times (1 + 9 + 11) + 1 \times (9 + 11)$ $+ 9 \times (11) = 763$</p> <p>$n_y = 1 \times (2 + 1 + 0) + 3 \times (3 + 6 + 1)$ $+ 10 \times (10 + 14 + 9)$ $+ 6 \times (7 + 12 + 11)$ $+ 2 \times (1 + 0) + 3 \times (6 + 1)$ $+ 10 \times (14 + 9)$ $+ 7 \times (12 + 11) + 1 \times (0)$ $+ 6 \times (1) + 14 \times (9)$ $+ 12 \times (11) = 1221$</p>	

4.2.1 Gamma de Goodman-Kruskal ($\hat{\gamma}$)

$$\hat{\gamma} = \frac{C - D}{C + D}$$

Application : $\hat{\gamma} = \frac{1331-849}{1331+849} = 0.221$

- Si $D = 0$, alors :

$$\hat{\gamma} = \frac{C-0}{C+0} = \frac{C}{C} = 1$$

- Si $C = 0$, alors :

$$\hat{\gamma} = \frac{0-D}{0+D} = \frac{-D}{D} = -1$$

- Si $C = D$, alors :

$$\hat{\gamma} = \frac{C-D}{C+D} = \frac{C-C}{C+C} = \frac{0}{C} = 0$$

4.2.2 Tau b de Kendall (τ_b)

$$\tau_b = \frac{C - D}{\sqrt{(C + D + n_x) \times (C + D + n_y)}}$$

Application : $\tau_b = \frac{1331-849}{\sqrt{(1331+849+763) \times (1331+849+1221)}} = 0.152$

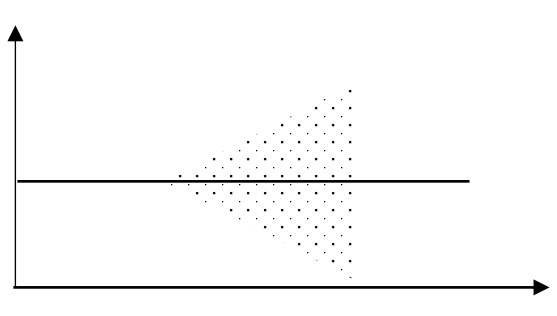
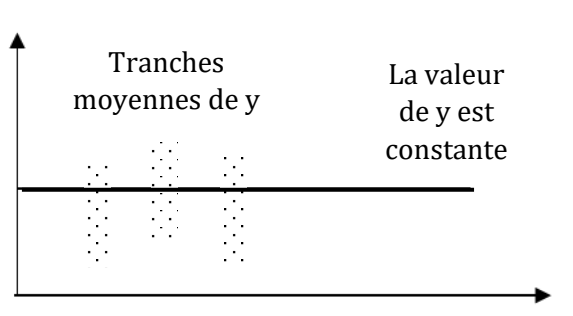
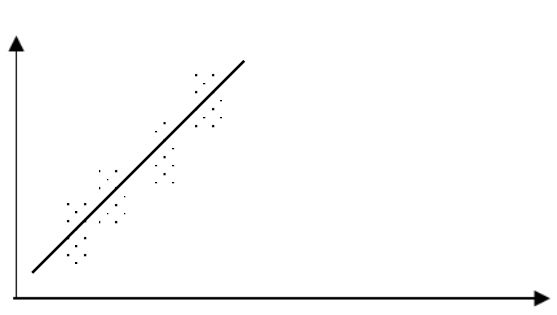
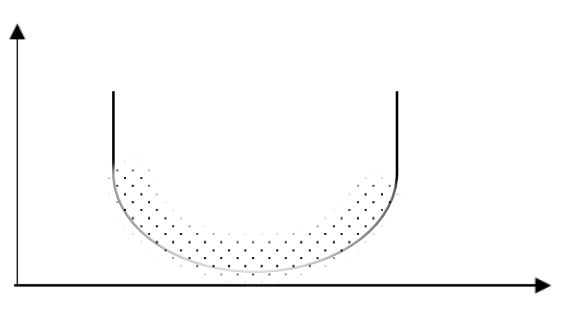
- Avec n_x est le nombre d'un ex-aequo selon « X » (mais différent selon « Y »)
- Avec n_y est le nombre d'un ex-aequo selon « Y » (mais différent selon « X »)

Cours 8

27.11.17

4.3 Liaison entre deux variables numériques

4.3.1 Etude variable de la corrélation

	
Absence de liaison en moyenne, mais pas en variance.	Absence de liaison. $E(Y X = x) = \text{constante}$
	
La dépendance ici est linéaire. $E(Y X = x) = a + bx$	La dépendance est quadratique $E(Y X = x) = a + bx + cx^2$
a : ordonnée à l'origine	b : pente

4.3.2 Coefficient de corrélation linéaire de Bravais-Pearson (r)

Ce coefficient mesure exclusivement le caractère linéaire et la dépendance entre « X » et « Y » :

$$r_{xy} = \frac{1}{n} \sum_{i=1}^n z_x(i) \times z_y(i)$$

$$z_x(i) = \frac{x_i - \bar{x}}{s_x}$$

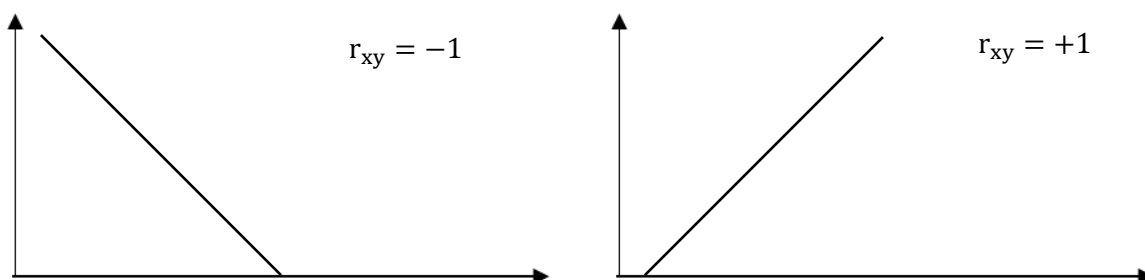
$$z_y(i) = \frac{y_i - \bar{y}}{s_y}$$

La moyenne de $z_x(i)$ vaut 0 :

$$\frac{1}{n-1} \sum (x_i - \bar{x})^2 \times \frac{1}{s_x^2} = \frac{1}{n-1} \sum \frac{(x_i - \bar{x})^2}{s_x^2} = 1$$

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n \sum x_i^2 - (\sum x_i)^2) \times (n \sum y_i^2 - (\sum y_i)^2)}}$$

Propriété :



- $-1 \leq r_{xy} \leq 1 = |r_{xy}| \leq 1$
- $|r_{xy}| = 1$: si et seulement si tous les points sont placés sur une ligne droite
- Si deux variables sont indépendantes, alors leur coefficient de corrélation est égal à 0.

ATTENTION : La réciproque n'est pas vraie.

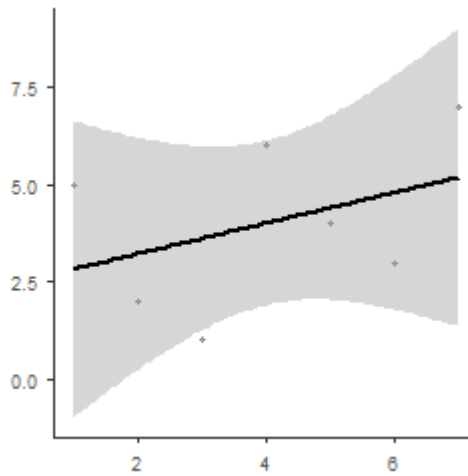
Alors que r (Bravais-Pearson) permet de quantifier la linéarité d'une relation r_s (Spearman) permet de quantifier sa monotonie.

4.3.3 Coefficient de corrélation de Spearman (r_s)

Niveau d'anxiété et dépression (sans ex-aequo) :

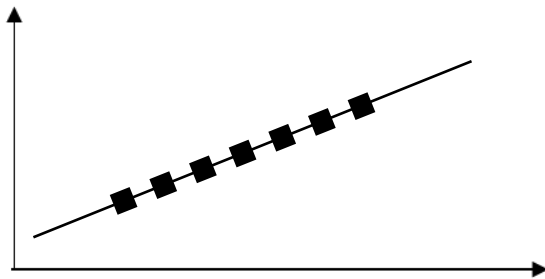
Stress (x)	Anxiété (y)	Rang (x)	Rang (y)	$d_i = \text{rg}(y) - \text{rg}(x)$	d_i^2
0	5	1	5	4	16
8	3	6	3	-3	9
9	12	7	7	0	0
7	4	5	4	-1	1
3	0	3	1	-2	4
2	2	2	2	0	0
5	6	4	6	2	4

= 34

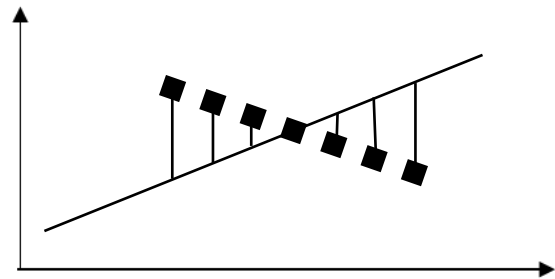


Récapitulatif des coordonnées :

- 1 : (1 ; 5)
- 2 : (6 ; 3)
- 3 : (7 ; 7)
- 4 : (5 ; 4)
- 5 : (3 ; 1)
- 6 : (2 ; 2)
- 7 : (4 ; 6)



Accord parfait entre les $rg(x)$ et $rg(y)$



Désaccord total entre les $rg(x)$ et $rg(y)$

Le degré de désaccord peut être mesuré par $\sum d_i^2$:

$$r_s = 1 - C \sum d_i^2 \text{ avec } C = \frac{6}{n(n^2 - 1)} \Rightarrow r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Application :

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 34}{7 \times (49 - 1)} = 0,39$$

Niveau d'anxiété et dépression (avec ex-aequo) :

Stress (x)	Anxiété (y)	¹³ $\tilde{r}g(x)$	$\tilde{r}g(y)$	Rang (x)	Rang (y)	d_i	d_i^2
0	1	1	3	$\frac{1+2+3}{3} = 2$	3	1	1
1	3	4	5	$\frac{4+5}{2} = 4.5$	$\frac{5+6}{2} = 5.5$	1	1
2	0	6	1	6	$\frac{1+2}{2} = 1.5$	-4.5	20.25
0	0	2	2	$\frac{1+2+3}{3} = 2$	$\frac{1+2}{2} = 1.5$	-0.5	0.25
1	4	5	7	$\frac{4+5}{2} = 4.5$	7	2.5	6.25
6	3	7	6	7	$\frac{5+6}{2} = 5.5$	-1.5	2.25
0	2	3	4	$\frac{1+2+3}{3} = 2$	4	2	4

$$r_s(X, Y) = r(rg(x); rg(y))$$

Cours 9

04.12.17

5. Théorie élémentaire des probabilités

5.1 Lancer d'un dé

$\Omega = \{1; 2; 3; 4; 5; 6\}$ Ω est l'univers des possibles.

6 éléments = événement A est un sous-ensemble de Ω : $A \subset \Omega$.

Exemple : $A\{1; 3; 5\}$, $B\{2; 4\}$, $C\{1; 6\}$

Comme $A \cap B = \emptyset$, A et B sont incompatibles.

Non-C, \bar{C} ou C^c est l'ensemble contraire à C : $\bar{C} = \{\omega | \omega \in \Omega \wedge \omega \notin C\}$

$$A \cup B = \{\omega | \omega \in \Omega, \omega \in A \vee \omega \in B\}$$

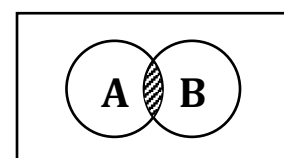
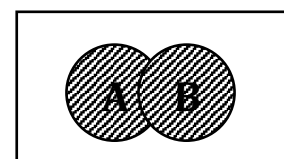
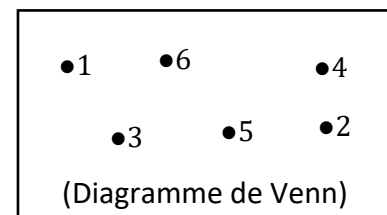
Exemple : $A \cup B = \{1; 3; 5\} \cup \{2; 4\} = \{1; 2; 3; 4; 5\}$

$$A \cap B = \{\omega | \omega \in \Omega, \omega \in A \wedge \omega \in B\}$$

Exemple : $A \cap B = \{1; 3; 5\} \cap \{2; 4\} = \emptyset$

Axiomes :

- $P(A) \geq 0$
- $P(\Omega) = 1$
- Si A et B sont incompatibles ($A \cap B = \emptyset$)
- $P(A \cup B) = P(A) + P(B)$ ¹⁴

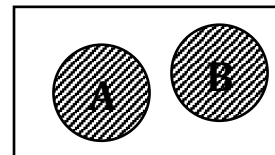


¹³ $\tilde{r}g$: Pseudo rang

¹⁴ Loi du « ou »

Probabilité conditionnelle :

$$P(A|C)^{15} = \frac{P(A \cap C)}{P(C)} \text{ avec } P(C) \neq 0$$



Exemple : $P(A|C)$: calculons la probabilité de : $A\{1; 3; 5\}, C\{1; 6\} = A \cap C = (\{1\})$

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{1/6}{2/6} = \frac{1}{2}$$

Classiquement la probabilité se calcule en rapportant le nombre de cas favorables, aux nombres de cas possibles.

$$P(A) = \frac{\# \text{ éléments favorables}}{\# \text{ éléments possibles}} = \frac{|A|}{|\Omega|} = \frac{3}{6}$$

Deux éléments sont indépendants, lorsque $P(A|B) = P(A)$.

Exemple : $P(A|B) : A\{1; 3; 5\}, B\{2; 4\} \Rightarrow P(A \cap B) = \emptyset$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{\emptyset\})}{P(\{2; 4\})} = \frac{0/6}{2/6} = 0 \text{ et } P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6}$$

Comme $P(A|B) \neq P(A)$, cela signifie qu'ils ne sont pas indépendants. Lorsque leurs événements sont indépendants :

$$P(A \cap B) = P(A) \times P(B)^{16}$$

En effet, si A et B sont indépendants :

$$P(A|B) = P(A), \text{ alors } P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ et donc } P(A \cap B) = P(A|B) \times P(B) = P(A) \times P(B)$$

5.2 Lancer d'une pièce quelques fois

$n = 1$ (1 événement) :

Pile ou Face : face est le succès

$X = x_i$	0	1
$P(X = x_i)$	$1 - \pi$	π

$$P(F: \text{Face}) = P(\text{"succès"}) = \pi \text{ et } P(P: \text{Pile}) = P(\text{"échec"}) = 1 - \pi$$

Calculons l'espérance de X :

$$E(X) = \sum_i x_i \times P(X = x_i)$$

Application :

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = 0 \times (1 - \pi) + 1(\pi) = \pi$$

¹⁵ A « sachant que » C

¹⁶ Loi du « et »

Calculons la variance de X :

$$Var(X) = \sum (x_i - E(X))^2 \times P(X = x_i)$$

Application :

$$Var(X) = (0 - \pi)^2 \times (1 - \pi) + (1 - \pi)^2 \times \pi = \pi(1 - \pi)[\pi + (1 - \pi)] = \pi(1 - \pi)$$

n = 2 (2 événements) :

	FF	PF	FP	PP
X = « succès »	2	1	1	0

3 possibilités :

- $P(X = 0) = P(PP) = P(P) \times P(P) = (1 - \pi)^2$
- $P(X = 1) = P(PF \text{ ou } FP) = P(PF) + P(FP) = \pi \times (1 - \pi) + (1 - \pi) \times \pi = 2\pi \times (1 - \pi)$
- $P(X = 2) = P(FF) = \pi \times \pi = \pi^2$

Application de l'espérance et de la variance :

$$\begin{aligned} E(X) &= 0 \times P(x = 0) + 1 \times P(x = 1) + 2 \times P(x = 2) \\ &= 0 \times (1 - \pi)^2 + 1 \times (2\pi \times (1 - \pi)) + 2 \times \pi^2 = 2\pi \end{aligned}$$

$$\begin{aligned} Var(X) &= (0 - 2\pi)^2(1 - \pi)^2 + 2(1 - 2\pi)^2\pi(1 - \pi) + (2 - 2\pi)^2(\pi^2) \\ &= 2\pi(1 - \pi)[2\pi(1 - \pi) + (1 - 2\pi)^2 + 2(1 - \pi)\pi] = 2\pi(1 - \pi) \end{aligned}$$

La loi qui décrit le lancer d'une pièce est une loi binomiale de paramètre $n = 1$ et π (probabilité du succès) :

$$X \sim B(1; \pi)$$

La loi qui décrit le lancer d'une pièce 2 fois est la loi binomiale des paramètres $n = 2$ et π :

$$X \sim B(2; \pi)$$

La loi qui décrit le lancer d'une pièce n fois est la loi binomiale des paramètres n et π :

$$X \sim B(n; \pi)$$

$${}^{17}P(X = k) = C_k^n \times \pi^k \times (1 - \pi)^{n-k} \Leftrightarrow C_k^n = \frac{n!}{k!(n-k)!}$$

$${}^{18}k \in \{0, 1, 2, 3, \dots, n\}$$

Cours 10

11.12.17

Moyenne binomiale :

$$E(X) = n \times \pi$$

Variance binomiale :

$$Var(X) = n \times \pi \times (1 - \pi)$$

¹⁷ Voir le « triangle de Pascal »

¹⁸ Toujours commencer par 0

5.3 Lancer d'une pièce un grand nombre de fois

Lorsque $n \rightarrow \infty$; $X \sim B(n; \pi)$ alors $E(X) = n \times \pi$ et $Var(X) = n \times \pi \times (1 - \pi)$ tendent vers l'infini. Pour pallier cet inconvénient, on centre et on réduit nos distributions :

$$Y_n = \frac{X_n - E(X_n)}{\sqrt{Var(X_n)}} = \frac{X_n - n \times \pi}{\sqrt{n \times \pi(1 - \pi)}}$$

On associe donc à chaque $\left\{y = \frac{x - n \times \pi}{\sqrt{n \times \pi(1 - \pi)}}\right\}$ la probabilité $\{P_y = P_x = C_x^n \times \pi^x \times (1 - \pi)^{n-x}\}$ avec $x \in \{0, 1, \dots, n\}$.

Définition :

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \times e^{-\frac{x^2}{2}} \quad \varphi(x) \geq 0$$

$$\int_{-\infty}^{+\infty} \varphi(t) dt = 1$$

Ces propriétés définissent une densité.

$$E(X) = \int_{-\infty}^{+\infty} x \varphi(x) dx = 0$$

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 \times \varphi(x) dx = 1$$

$$X \sim N(0; 1)$$

La courbe de la loi normale se déplace vers la droite ce qui signifie qu'elle tend vers l'infini.

Application : Que vaut la densité d'une loi normale centrée et réduite en 0 ?

$$\varphi(0) = \frac{1}{\sqrt{2\pi}} \times e^{-\frac{0^2}{2}} = \frac{1}{\sqrt{2\pi}} \times e^0 = \frac{1}{\sqrt{2\pi}} \cong 0,399$$

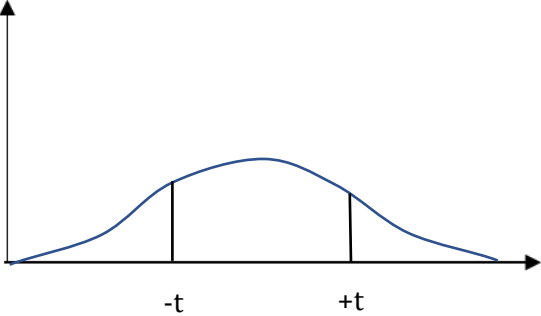
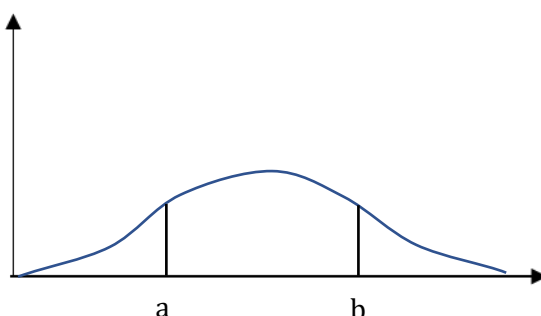
Fonction de répartition :

$$\Phi(x) = P(X \leq x) = \int_{-\infty}^{+\infty} \varphi(t) dt$$

$$\Phi(-\infty) = 0$$

$$\Phi(0) = \frac{1}{2}$$

$$\Phi(+\infty) = 1$$

J'ai un quantile, je cherche une probabilité.	J'ai une probabilité, je cherche un quantile.
	
$\Phi(t) + \Phi(-t) = 1 \Leftrightarrow \Phi(-t) = 1 - \Phi(t)$	$P(a \leq X \leq b) = \Phi(b) - \Phi(a) \Leftrightarrow \Phi(a) + P(a \leq X \leq b) = \Phi(b)$
${}^{19}\Phi(1.52) = 0.9357$ $\Phi(-0.48) = 1 - \Phi(0.48) = 1 - 0.6844$ $= 0.3156 = 31.56\%$	$P(-1 \leq X \leq 1) = \Phi(1) - \Phi(-1)$ $= \Phi(1) - [1 - \Phi(1)]$ $= 2\Phi(1) - 1$ $= 2 \times 0.8413 - 1 = 0.6826$ $\cong 68\%$ $P(-2 \leq X \leq 2) = \Phi(2) - \Phi(-2)$ $= 2\Phi(2) - 1$ $= 2 \times 0.9772 - 1 = 0.9544$ $= 95.44\%$

Cas général :

$$X \sim N(\mu, \sigma^2)$$

$$f: x \rightarrow \frac{1}{\sqrt{2\pi} \times \sigma} \times e^{\frac{-1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

$$P(a \leq X \leq b) = P(a - \mu \leq X - \mu \leq b - \mu) = P\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

Application : $QI \sim N(100; 15^2)$

$\mu = 100$ et $\sigma = 15$

$$P(QI > 130) = P(130 < QI < +\infty) = \Phi(+\infty) - \Phi\left(\frac{130 - 100}{15}\right) = 1 - \Phi(2) = 1 - 0.9772$$

$$= 0.228 \cong 2.3\%$$

Loi normale : (moyenne, écarte-type²)

et

Loi binomiale : (fréquence, succès)

¹⁹ Voir le tableau « la loi normale réduite »

6. Estimation

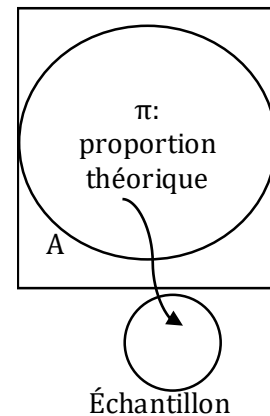
6.1 Estimer une proportion

6.1.1 Estimation ponctuelle d'une proportion

$$\pi = \frac{|A|}{|\Omega|} \Leftrightarrow \hat{\pi} = f_A \Leftrightarrow f_A = \frac{n_A}{n}$$

ID	Nombre de quartiers	Nombre de quartiers avec pépins
1	10	2
2	10	6
3	10	1
4	10	1
5	11	0
6	8	4
7	9	5
8	10	5
9	9	0
10	10	2

n (taille) = 97

n_A (sous-ensemble) = 26

Les individus statistiques ici sont les quartiers :

$$f_A = \frac{26}{97} = 0.268 \cong 26.8\%$$

$$\hat{\pi} = f_A \cong 26.8\%$$

Cette estimation est-elle sans biais ?

$$^{20}X_1 \sim \text{Ber}(\pi)$$

$$X_2 \sim \text{Ber}(\pi)$$

$$X_n \sim \text{Ber}(\pi)$$

$$x = x_1 + x_2 + \dots + x_n = \sum_{i=1}^n x_i$$

$$E(X) = n \times \pi$$

$$X \sim B(n, \pi)$$

$$\text{Var}(X) = n \times \pi \times (1 - \pi)$$

$$F = \frac{X}{n}$$

$$E(F) = E\left(\frac{X}{n}\right) = \frac{1}{n} \times E(X) = \frac{1}{n} \times n \times \pi = \pi$$

L'estimation est-elle précise ?

Sans biais → précis	Sans biais → imprécis	Avec biais → précis	Avec biais → imprécis

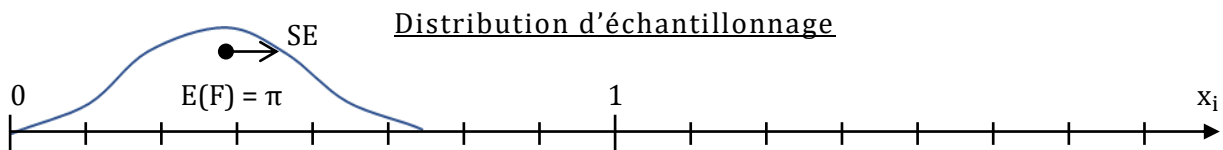
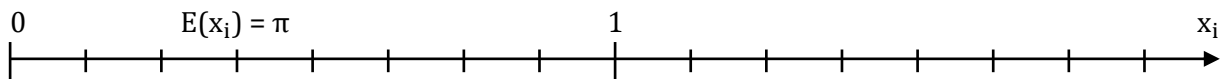
²⁰ Loi de Bernoulli, pour un seul lancer

Calculons la variance de $F = \frac{X}{n}$:

$$Var(F) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2} \times Var(X) = \frac{1}{n^2} \times n \times \pi \times (1 - \pi) = \frac{\pi(1 - \pi)}{n}$$

La précision est inversement proportionnelle à la variance de F , donc pour accroître la précision il suffit d'augmenter de n la taille de l'échantillon → plus l'échantillon augmente et plus l'écart-type diminue (la variance aussi).

Fonction de distribution



SE : Erreur standard (écart-type de la distribution d'échantillonnage)

6.1.2 Estimation par intervalle de confiance d'une proportion

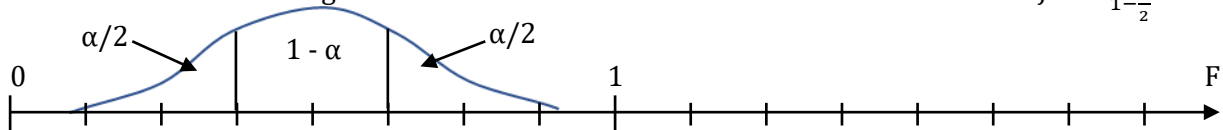
Méthode asymptotique sans correction de continuité :

1. On estime ponctuellement la valeur nominale - JAMOV → Frequencies : : $\hat{\pi} = f = \frac{n_A}{n}$
2. On calcule l'erreur standard : $SE = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{f(1-f)}{n}}$
3. On choisit un niveau de confiance :

α	$1 - \alpha$	$u_{1-\alpha/2}$
1%	99%	2.576
5%	95%	1.960
10%	90%	1.645

4. On calcule la marge d'erreur :

$$\Delta f = u_{1-\frac{\alpha}{2}} \times SE$$



5. On calcule les deux bornes de l'intervalle de confiance (IC) :

$$IC_{1-\alpha}[f - \Delta f; f + \Delta f]$$

Application :

$$n = 97 ; n_A = 26 ; f = 0.268$$

IC à 95%	IC à 90%
1. $\hat{\pi} = f = 0.268$	1. $\hat{\pi} = f = 0.268$
2. $SE = \sqrt{\frac{0.268(1-0.268)}{97}} = 0.045$	2. $SE = \sqrt{\frac{0.268(1-0.268)}{97}} = 0.045$
3. $1 - \alpha : 95\%$	3. $1 - \alpha : 90\%$
4. $\Delta f = u_{1-\frac{\alpha}{2}} \times SE = u_{1-\frac{5\%}{2}} \times 0.045$ $= u_{0.975} \times 0.045 = 1.96 \times 0.045 = 0.088$	4. $\Delta f = u_{1-\frac{\alpha}{2}} \times SE = u_{1-\frac{10\%}{2}} \times 0.045$ $= u_{0.95} \times 0.045 = 1.645 \times 0.045 = 0.074$
5. $IC_{95\%}[f - \Delta f ; f + \Delta f]$ $= IC_{95\%}[0.180 ; 0.356]$	5. $IC_{95\%}[f - \Delta f ; f + \Delta f]$ $= IC_{95\%}[0.194 ; 0.342]$

On perd en précision lorsqu'on gagne en confiance.

Nous constatons qu'une valeur n'appartient pas à l'intervalle de confiance lorsqu'elle n'est pas incluse. Par exemple 0.35 est incluse dans l'intervalle de gauche, mais pas à droite.

Cours 12

26.02.18

6.2 Estimer une moyenne

6.2.1 Estimation ponctuelle d'une moyenne

Estimation de la moyenne des valeurs théoriques :

$$\hat{\mu} = \bar{x} = \frac{x_i}{n}$$

Cette estimation est sans biais :

$$E(\bar{x}) = \hat{\mu}$$

La précision de cette estimation dépend de la taille de l'échantillon :

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{n}$$

Application :

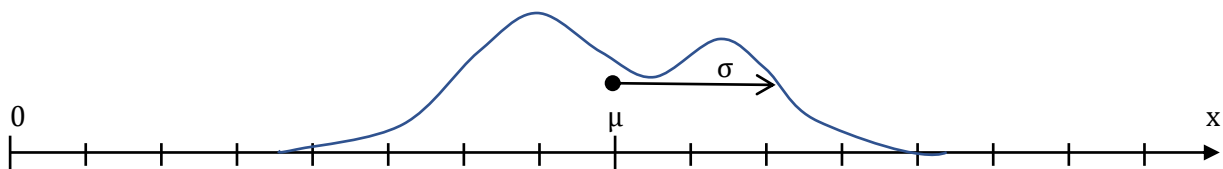
$$n \text{ (nb de mandarines)} = 10$$

$$x_i \text{ (nb de quartiers)} = 97$$

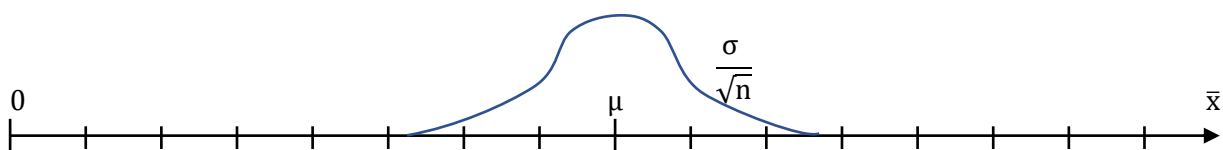
$$\hat{\mu} = \bar{x} = \frac{97}{10} = 9.7$$

$$s = 0.823$$

Fonction de distribution :



Distribution d'échantillonnage :



6.2.2 Estimation par intervalle de confiance d'une moyenne

1. Estimer la valeur de la moyenne des valeurs numériques - JAMOV → T-Test : $\hat{\mu} = \bar{x} = \frac{x_i}{n}$
2. Calculer l'erreur standard : $SE = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$
3. Choisir un niveau de confiance : $1 - \alpha$
4. Calculer la marge d'erreur : ²¹ $\Delta\bar{x} = t_{1-\frac{\alpha}{2}} \times (n - 1) \times SE$
5. Calculer les bornes de l'intervalle : $IC_{1-\alpha}[\bar{x} - \Delta\bar{x}; \bar{x} + \Delta\bar{x}]$

Application :

$x_i = 97, n = 10, \bar{x} = 9.7$

IC à 95%	IC à 90%
<ol style="list-style-type: none"> 1. $\hat{\mu} = \bar{x} = 9.7$ 2. $SE = \frac{s}{\sqrt{n}} = \frac{0.823}{\sqrt{10}} = 0.260$ 3. $1 - \alpha : 95\%$ 4. $\Delta\bar{x} = t_{1-\frac{\alpha}{2}} \times (n - 1) \times SE$ $= t_{1-\frac{5\%}{2}} \times (10 - 1) \times 0.260$ $= t_{0.975} \times (9) \times 0.260$ $= 2.262 \times 0.260 = 0.588$ 5. $IC_{1-\alpha}[\bar{x} - \Delta\bar{x}; \bar{x} + \Delta\bar{x}]$ $= IC_{95\%}[9.111; 10.288]$ 	<ol style="list-style-type: none"> 1. $\hat{\mu} = \bar{x} = 9.7$ 2. $SE = \frac{s}{\sqrt{n}} = \frac{0.823}{\sqrt{10}} = 0.260$ 3. $1 - \alpha : 90\%$ 4. $\Delta\bar{x} = t_{1-\frac{\alpha}{2}} \times (n - 1) \times SE$ $= t_{1-\frac{10\%}{2}} \times (10 - 1) \times 0.260$ $= t_{0.95} \times (9) \times 0.260$ $= 1.833 \times 0.260 = 0.477$ 5. $IC_{1-\alpha}[\bar{x} - \Delta\bar{x}; \bar{x} + \Delta\bar{x}]$ $= IC_{95\%}[9.223; 10.177]$

Nous constatons que les deux intervalles ont leur centre à la même position mais ils n'ont pas la même étendue : l'intervalle de confiance de niveau 95% est plus large que celui de niveau 68%. Lorsque le niveau de confiance augmente la précision de l'estimation diminue.

7. Test de proportion

7.1 Test sur une proportion²²

Exemple : mère du prof distingue différentes manières de préparer du thé

Essai	1	2	3	4	5	6	7	8
Réponse	✓	✓	✗	✓	✓	✓	✓	✓
Point	1	1	0	1	1	1	1	1

$x_{emp} = 7$
 $n = 8$
 $f_{emp} = 7/8$

²¹ Utiliser le programme R : $\rightarrow qt(1 - \frac{\alpha}{2}, n - 1)$: précis pour un petit échantillon. Pour un grand échantillon utiliser plutôt la loi normale centrée réduite (dès 100 sujets).

²² Pour parler d'une proportion, il faut : un nombre de participants et un pourcentage

Deux possibilités :

- H_0 : hypothèse nulle "ma mère répond au hasard" $\rightarrow \pi = \frac{1}{2}$
 \rightarrow Il s'agit d'un phénomène peu probable qui s'est réalisé.
- H_1 : hypothèse alternative "ma mère ne répond pas au hasard" $\rightarrow \pi > \frac{1}{2}$
 \rightarrow Ma mère a un certain pouvoir discriminant.

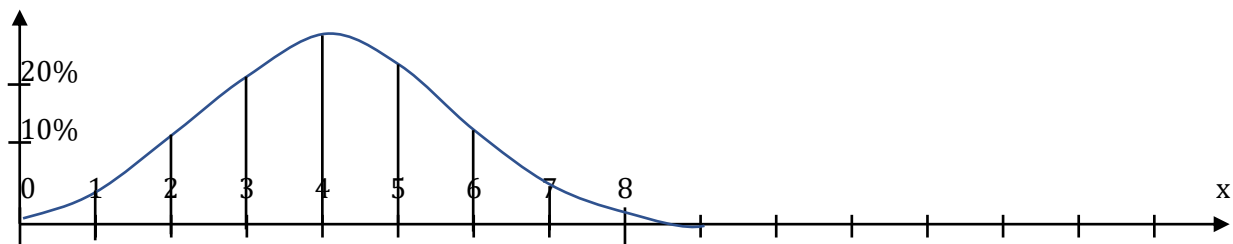
Sous H_0 nommons X la variable de décision. X représente le nombre de succès parmi les n essais.

$$X \sim B(n = 8, \pi = \frac{1}{2})$$

$$P(x = k) = C_k^n \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{n-k}$$

k	0	1	2	3	4	5	6	7	8
C_k^n	1	8	28	56	70	56	28	8	1
$P(x = k)$	$\frac{1}{256}$	$\frac{8}{256}$	$\frac{28}{256}$	$\frac{56}{256}$	$\frac{70}{256}$	$\frac{56}{256}$	$\frac{28}{256}$	$\frac{8}{256}$	$\frac{1}{256}$

= 0.4 = 3.1 = 10.9 = 21.9 = 27.3 = 21.9 = 10.9 = 3.2 = 0.4



Calculons la probabilité d'obtenir un score supérieur ou égal à $x_{\text{emp}} = 7$. Cette probabilité se nomme « probabilité critique » :

$$P(X \geq 7|H_0) = P(x = 7|H_0) + P(x = 8|H_0) = \frac{8}{256} + \frac{1}{256} = 0.035 \cong 3.5\%$$

Cela correspond à la probabilité/compatibilité de H_0 avec les données. On choisit un seuil de signification : $\alpha = 5\%$. La règle de décision est la suivante :

Si p-valeur < α , alors on rejette H_0

Dans notre exemple : p-valeur = 0.035, $\alpha = 0.05$. Comme $p < 0.05$, on rejette H_0 .

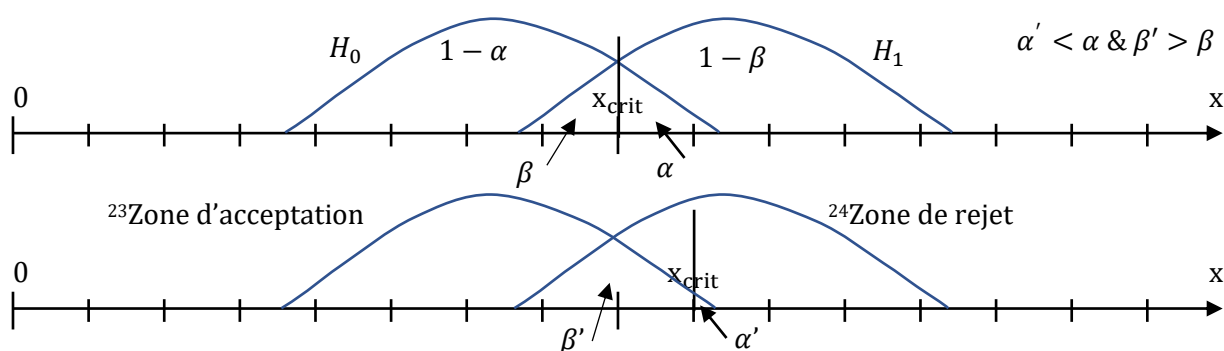
Démarche générale :

1. Choisir H_0 et H_1 .
2. Choisir α , le seuil de signification.
3. Déterminer la valeur empirique de la variable de décision.
4. Calculer la probabilité critique avec $p = P(X > x_{\text{emp}}|H_0)$.
5. Comparer α et la valeur p, puis conclure.

Décision	H_0	H_1
$A \rightarrow$ Acceptation	$1 - \alpha$	β
	Bonne conclusion	Erreur de 2 ^{ème} espèce
$R \rightarrow$ Rejet	α	$1 - \beta$
	Erreur de 1 ^{ère} espèce	Puissance

$\alpha = P(R|H_0)$: probabilité de rejeter H_0 , même s'il est vrai.

$\beta = P(A|H_1)$: probabilité d'accepter H_0 , alors que H_1 est vrai.



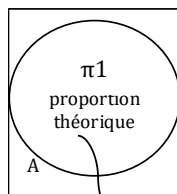
Trois formes de test :

Test unilatéral à gauche	Test bilatéral	Test unilatéral à droite
$H_0: \pi = \pi_0$	$H_0: \pi = \pi_0$	$H_0: \pi = \pi_0$
$H_1: \pi < \pi_0$	$H_1: \pi \neq \pi_0$	$H_1: \pi > \pi_0$
$X \sim B(n, \pi_0)$		
$p = P(X \leq x_{obs} H_0) = P -$	$p = 2 \times \min(P+; P-)$	$p = P(X \geq x_{obs} H_0) = P +$
Si $p < \alpha$, alors on rejette H_0		

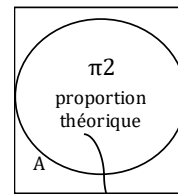
²³ Avant x_{crit} : Zone d'acceptation

²⁴ Après x_{crit} : Zone de rejet

7.2 Test sur deux proportions



$$f_1 = \frac{x_1}{n_1}$$



$$f_2 = \frac{x_2}{n_2}$$

Test bilatéral :

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 \neq \pi_2$$

Exemple : 3^{ème} étude de Guéguen

Talons plats	Talons hauts
$n_- = 60$	$n_+ = 60$
Ramasse le gant	
$n_- = 37$	$n_+ = 56$
$f_- = \frac{x_1}{n_1} = \frac{37}{60} \cong 61.7\%$	$f_+ = \frac{x_2}{n_2} = \frac{56}{60} \cong 93.3\%$

Posons $\alpha = 5\%$

$$H_0 : \pi_- = \pi_+$$

$$H_1 : \pi_- \neq \pi_+$$

	n_{ij}	x		
		Plats	Hauts	
y	Ramasse	37	56	= 93
	Ramasse pas	23	4	= 27
		= 60	= 60	= 120

Correction de continuité de Yates :

	e_{ij}	x		
		Plats	Hauts	
y	Ramasse	46.5	46.5	= 93
	Ramasse pas	13.5	13.5	= 27
		= 60	= 60	= 120

Méthode 1 :

$$\chi^2 = \sum_i \sum_j \frac{(|n_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}$$

Application :

$$\chi^2 = \frac{9^2}{46.5} + \frac{9^2}{46.5} + \frac{9^2}{13.5} + \frac{9^2}{13.5} = 15.484$$

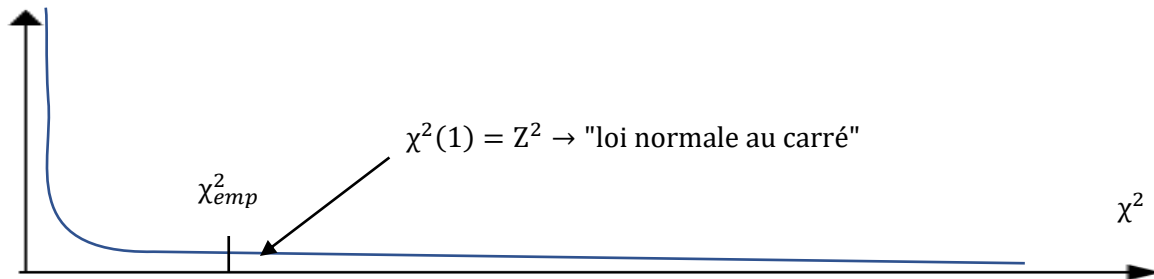
Méthode 2 (plus précise) :

$$\chi^2 = \frac{n \times \left(|AD - BC| - \frac{n}{2} \right)^2}{(A + B) \times (C + D) \times (A + C) \times (B + D)}$$

A	B	= A+B
C	D	= C+D
= A+C	= B+D	= n

Application :

$$\chi^2 = \frac{120 \times \left(|37 \times 4 - 23 \times 56| - \frac{120}{2} \right)^2}{(93) \times (27) \times (60) \times (60)} = \frac{139\,968\,000}{9\,039\,600} = 15.484$$

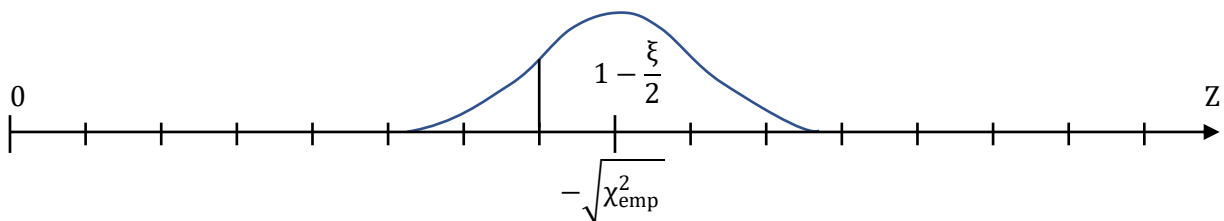


La distribution théorique sous H_0 correspond à une distribution de χ^2 à 1 degré de liberté.

$p = P(\chi^2(1) \geq \chi^2_{\text{emp}})$, on calcule p en utilisant le programme R :

$$\rightarrow \text{pchisq}(15.484, 1, \text{lower.tail} = \text{FALSE}) = 8.32 \times 10^{-5}$$

Ici, $p < \alpha$, alors on rejette H_0 .

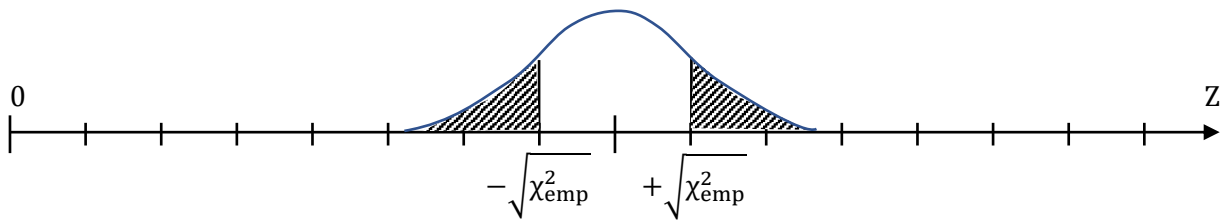


Si $\hat{\pi}_1 \geq \hat{\pi}_2$, alors $Z_{\text{emp}} = -\sqrt{\chi^2_{\text{emp}}}$ et $p = 1 - \frac{\xi}{2}$.

	Chapitre :	1 échantillon 1 mesure	1 échantillon 2 mesures	2 échantillons 1 mesure	k échantillons 1 mesure
	Nominale	7.1 Test sur une proportion		7.2 Test sur 2 proportions	
Test non- paramétrique	Ordinale	8.3 Test de Wilson sur 1 échantillon	9.2 Test de Wilcoxon	10.2 Test de Mann Whitney	11.3 Test de Kruskal et Wallis
Test paramétrique	Numérique	8.1 Test de Student sur 1 échantillon	9.1 Test de Student sur des mesures pairees	10.1 Test de Student sur 2 groupes indépendants	11.1 ANOVA à 1 facteur de classification

On réalise une mesure pour chaque individu concernant ces tests.

$k \geq 2$



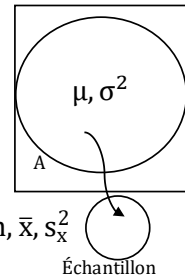
8. Test sur la proportion d'une distribution

8.1 Test sur une moyenne (Test de Student sur un échantillon)

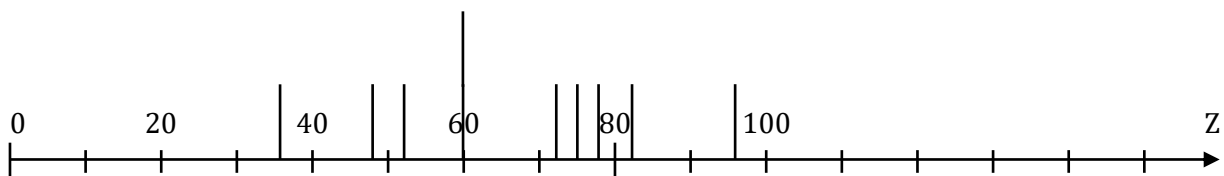
Population : on veut savoir la valeur de μ donc on extrait un échantillon, une taille moyenne et un s_x^2 .

Test bilatéral en 3 formes : $H_0: \mu = \mu_0$ et $H_1: \mu \neq \mu_0$

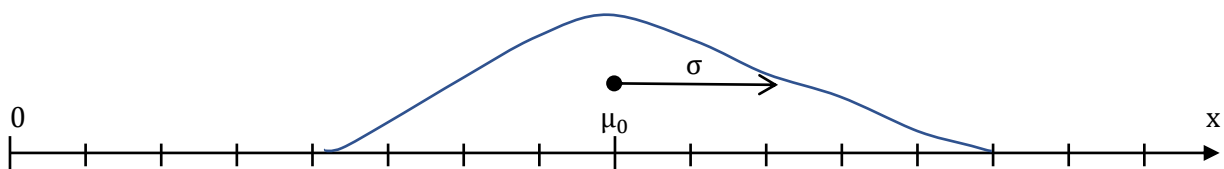
Exemple : il nous donne une photo de sa petite fille et on doit estimer son âge en mois (calculs à faire sur Jamovi) :



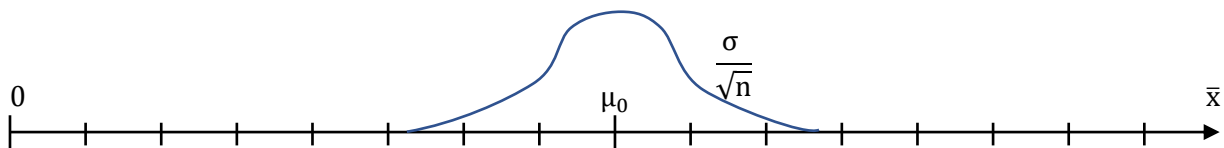
i	1	2	3	4	5	6	7	8	9	10
x_i	72	96	48	52	82	60	36	75	60	78
	$n = 10$			$\bar{x} = 65.9$			$s = 18$			$u_0 = 40$



Fonction de distribution :



Distribution d'échantillonnage :



On transforme \bar{x} en une valeur centrée et réduite :

$$u = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu_0}{\sigma} \times \sqrt{n}$$

Comme σ n'est pas connu, on doit l'estimer, mais comment ? $\hat{\sigma} = s$ donc l'écart-type, alors la variable de décision devient :

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu_0}{s} \times \sqrt{n}$$

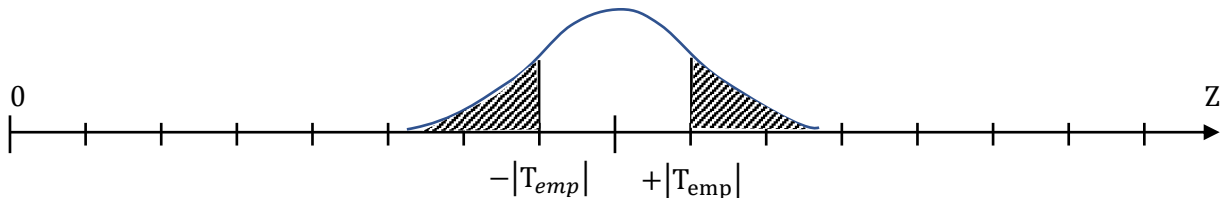
Sous H_0 , la variable T suit une loi de Student à $n-1$ degré de liberté :

$$E(T(\gamma)) = 0$$

$$Var(T(\gamma)) = \frac{\gamma}{\gamma-2}$$

Caractéristiques :

- Il y a autant de distribution de Student que de valeurs possibles pour γ .
- $T(\gamma)$ est symétrique par rapport à 0.
- Lorsque γ tend vers l'infini, la loi de Student $T(\gamma)$ tend vers une loi normale.
- Les lois de Student des distributions de probabilité.



Calculons la valeur empirique de la variable de décision :

$$T_{emp} = \frac{\bar{x} - \mu_0}{s} \times \sqrt{n} = \frac{65.9 - 40}{17.978} \times \sqrt{10} = 4.556$$

Programme R :

$$\rightarrow p = 2 * pt(abs(4.556), 9, lower.tail = FALSE) = 0.00137$$

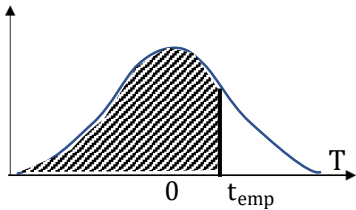
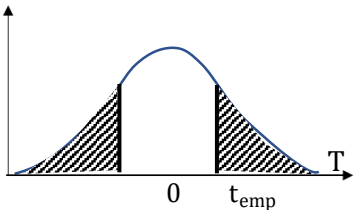
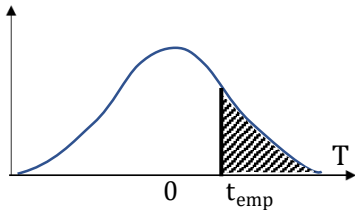
Selon les normes APA (Association des psychologues américains), le résultat est le suivant :

$$T(9) = 4.556, p = 0.01$$

Règle de décision : si $p < \alpha$, on rejette H_0 . Ici, comme $p < \alpha$, alors on rejette H_0 .

$p > \alpha$	$p = \alpha$	$p < \alpha$
$ T_{emp} < T_{crit}$	$ T_{emp} = T_{crit}$	$ T_{emp} > T_{crit}$
$T_{crit} = t_{1-\alpha/2}$		

3 formes de test :

Test unilatéral à gauche	Test bilatéral	Test unilatéral à droite
		
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu < \mu_0$	$H_1: \mu \neq \mu_0$	$H_1: \mu > \mu_0$
$t_{\text{emp}} = \frac{\bar{x} - \mu_0}{s} \times \sqrt{n}$		
$p = P(T < t_{\text{emp}})$	$p = P(T > \text{ou} < t_{\text{emp}})$	$p = P(T > t_{\text{emp}})$

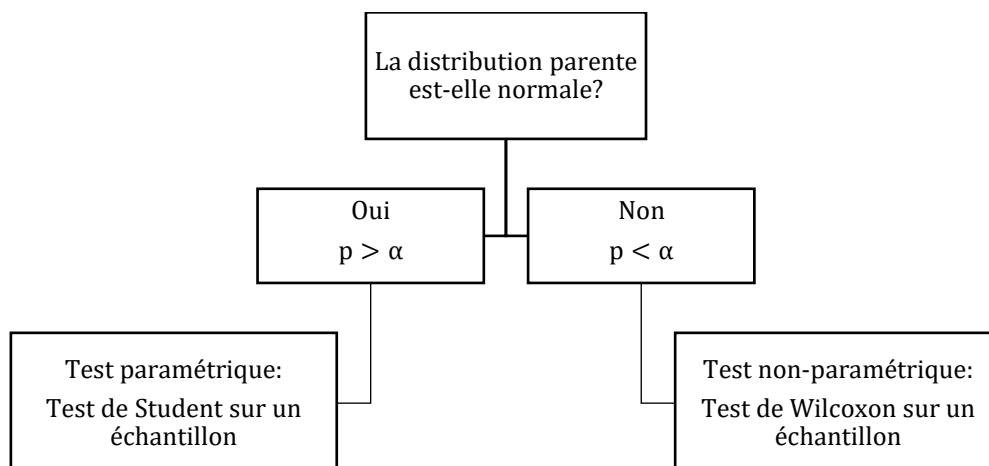
Si la probabilité critique est plus petite que l'hypothèse nulle, alors on la rejette : si $p < \alpha$, alors on rejette H_0 .

8.2 Test de normalité de Shapiro-Wilk (W)

Nous utiliserons, pour tester la normalité d'une distribution, la procédure de Shapiro-Wilk sur Jamovi. La variable de décision est un W :

$$W = \frac{(\sum_{i=1}^n a_i \times x_{[i]})^2}{\sum (x_i - \bar{x})^2}$$

On calcule W , puis la probabilité critique associée. Si $p < \alpha$, on rejette H_0 qui affirme que la population parente suit une loi normale.



Le test non paramétrique et le test paramétrique ne conduisent pas à la même conclusion. Cette contradiction apparente est due au fait que les deux tests n'ont pas la même puissance. Le test paramétrique est plus puissant : si les conditions d'application du test paramétrique sont satisfaites, alors le test paramétrique permet de rejeter plus facilement une hypothèse nulle qui est fausse.

8.3 Test de Wilcoxon sur un échantillon (T^+ , T^-)

8.3.1 Données sans ex-aequo

$$H_0: P(X > \mu_0) = P(X < \mu_0)$$

$$H_1: P(X > \mu_0) > P(X < \mu_0)$$

$$\rightarrow \tilde{\mu} = \mu_0$$

$$\rightarrow \tilde{\mu} > \mu_0$$

Exemple : fromager qui aurait la main lourde, $\mu_0 = 500g$:

Morceaux (n)	x_i	$d_i: x_i - \mu_0$	$ d_i $	R_i
1	576	76	76	6
2	491	-9	9	4
3	485	-15	15	5
4	499	-1	1	1
5	606	106	106	8
6	493	-7	7	3
7	506	6	6	2
8	585	85	85	7

$$T^+ = \sum_{d_i > 0} R_i = 6 + 8 + 2 + 7 = 23$$

$$T^- = \sum_{d_i < 0} R_i = 4 + 5 + 1 + 3 = 13$$

$$T^+ + T^- = 23 + 13 = 36$$

$$\frac{n \times (n+1)}{2} = \frac{8 \times (8+1)}{2} = 36$$

La variable de décision du test de Wilcoxon est $T^+ = 23$. Dans notre exemple, la variable de décision empirique de la variable de décision vaut $T^+ = 23$.

Le nombre de motifs vaut $2^n = 2^8 = 256$. Construisons la distribution de T^+ sous H_0 :

Rang/ Motif	1	2	3	4	5	6	7	8	T^+	T^-
1)	-	-	-	-	-	-	-	-	0	36
2)	+	-	-	-	-	-	-	-	1	35
3)	-	+	-	-	-	-	-	-	2	34
4)	+	+	-	-	-	-	-	-	3	33
5)	+	-		-	-	-	-	-	4	32
...
256)	+	+	+	+	+	+	+	+	36	0

T^+	T^-		#
0	36	(0)	1
1	35	(1)	1
2	34	(2)	1
3	33	(3) (1,2)	2
4	32	(4) (1,3)	2
5	31	(5) (3,2) (1,4)	3
6	30	(6) (1,5) (2,4) (1,2,3)	4
...
36	0	(1,2,3,4,5,6,7,8)	1

$$p = P(T^+ \geq t^+ | H_0)$$

Application :

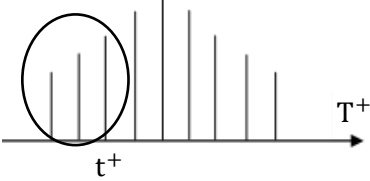
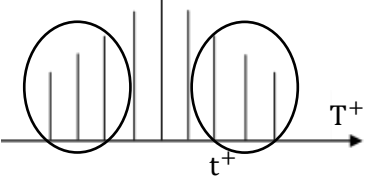
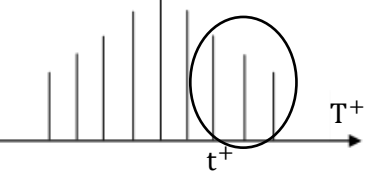
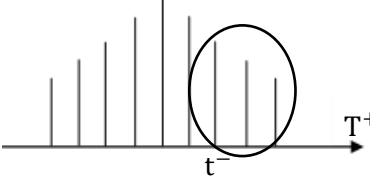
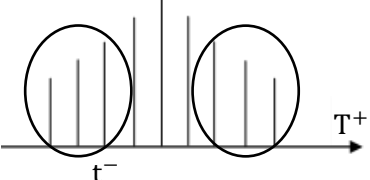
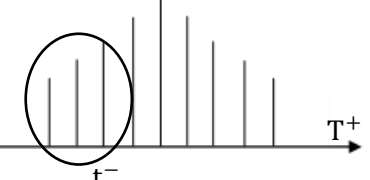
$$p = P(T^+ \geq t^+ | H_0) = \frac{11 + 10 + 9 + 7 + 6 + 5 + 4 + 3 + 2 + 2 + 1 + 1 + 1}{256} = \frac{70}{256} = 0.273$$

Comme $p > \alpha$, alors on ne rejette pas H_0 . Notation APA :

$$T^+ = 23, p = .273$$

Lorsque la p-valeur est plus grande que 0.01, la notation est celle indiquée ci-dessus. Lorsque la p-valeur est inférieure à 0.01, on note « $< .001$ »

Trois formes de test :
(seulement les différences sont mentionnées)

Test unilatéral à gauche	Test bilatéral	Test unilatéral à droite
$H_1: \hat{\mu} < \mu_0$	$H_1: \mu \neq \mu_0$	$H_1: \mu > \mu_0$
		
$p = P(T^+ \leq t^+)$	$p = 2 \times \min(P(T^+ \leq t^+); P(T^+ \geq t^+))$	$p = P(T^+ \geq t^+)$
		
$p = P(T^- \leq t^-)$	$p = 2 \times \min(P(T^- \leq t^-); P(T^- \geq t^-))$	$p = P(T^- \geq t^-)$

8.3.2 Données avec ex-aequo

Exemple : fromager qui aurait la main lourde, $\mu_0 = 500\text{g}$:

Morceaux (n)	x_i	$d_i: x_i - \mu_0$	$ d_i $	rg	rg moyen
1	490	-10	10	1	1.5
2	470	-30	30	4	4.5
3	460	-40	40	6	6.5
4	500	0	0	²⁵ ×	×
5	510	10	10	2	1.5
6	470	-30	30	5	4.5
7	460	-40	40	7	6.5
8	610	110	110	8	8
9	520	20	20	3	3
$n = 9 \quad n' = 8 \quad \tilde{\mu} = \mu_0 = 500\text{g}$					

$$T^+ = 1.5 + 8 + 3 = 12.5$$

$$T^- = 1.5 + 4.5 + 6.5 + 4.5 + 6.5 = 23.5$$

$$T^+ + T^- = 12.5 + 23.5 = 36 \quad n' = 8 \Rightarrow \frac{n' \times (n' + 1)}{2} = \frac{8 \times (8 + 1)}{2} = 36$$

$p = P(T^+ \geq t^+ | H_0) = P(T^+ \geq 12.5 | H_0) = 0.8$, $p > \alpha$, alors on ne rejette pas H_0 .

9. Analyse des mesures pairées

9.1 Approche paramétrique (Test de Student)

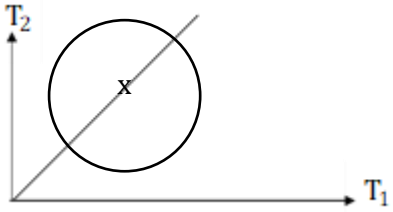
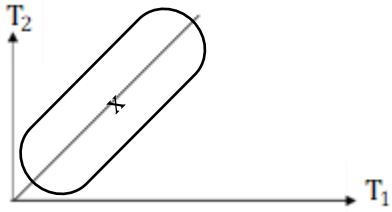
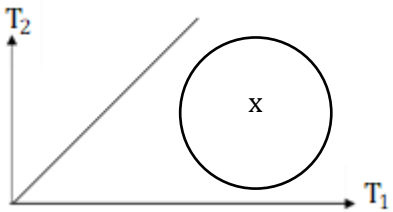
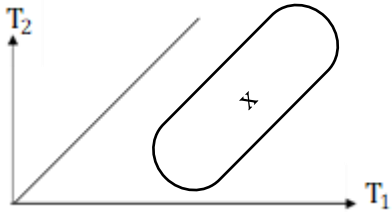
ID	T_1	T_2	$D = T_1 - T_2$
1	15	16	-1
2	10	4	6
3	7	13	-6
4	10	33	-23
5	12	9	3
6	2	11	-9
7	14	26	-12
8	3	0	3
9	20	11	9
10	14	16	-2
11	31	43	-12
12	14	24	-10
$n = 12 \quad {}^{26}S_d = 9.1998 \quad \bar{d} = -4.5$			

$$H_0: \mu_1 = \mu_2 \rightarrow H_0: \mu_1 - \mu_2 = \mu_D = 0$$

$$H_1: \mu_1 < \mu_2 \rightarrow H_1: \mu_1 - \mu_2 = \mu_D < 0$$

²⁵ Il ne faut pas tenir compte des différences qui égalent 0 dans les rangs.

$${}^{26}S_d = \sqrt{s_d^2} = \sqrt{\left(\frac{1}{n-1} \sum_{i=1}^n d_i^2\right) - \frac{n}{n-1} \times \bar{d}^2} = \sqrt{\frac{1174}{12-1} - \frac{12}{12-1} \times (-4.5)^2} \cong \sqrt{84.636} \approx 9.1998$$

	$^{27}\mu_1 = \mu_2$ $^{28}\rho_{12} = 0$		$\mu_1 = \mu_2$ $\rho_{12} \neq 0$
	$\mu_1 \neq \mu_2$ $\rho_{12} = 0$		$\mu_1 \neq \mu_2$ $\rho_{12} \neq 0$

$$\alpha = 5\% ; t_{emp} = \frac{\bar{x} - \mu_0}{s_x} \times \sqrt{n}$$

$$X \rightarrow D$$

$$s \rightarrow s_d$$

$$\bar{x} \rightarrow \bar{d}$$

$$\mu_0 \rightarrow 0$$

$$t_{emp} = \frac{\bar{x} - \mu_0}{s} \times \sqrt{n} \Leftrightarrow \frac{\bar{d} - 0}{s_d} \times \sqrt{n}$$

Application :

$$t_{emp} = \frac{\bar{d} - 0}{s_d} \times \sqrt{n} = \frac{-4.5 - 0}{9.1998} \times \sqrt{12} = -1.694$$

Programme R :

$$\rightarrow p = P(T(n-1) \leq t_{emp}) = P(T(11) \leq -1.694) = 0.059$$

Comme $p > \alpha$, on ne rejette pas H_0 . Résultats APA :

$$t(11) = -1.694, p = .059, \text{"unilatéral" ou "one tail"}$$

Pour pouvoir réaliser le test de Student sur une mesure paillée (test paramétrique), les différences $D = X_1 - X_2$ devraient être issues d'une distribution normale.

$$D \sim N(\mu_D, \sigma_D^2)$$

9.2 Approche non-paramétrique (Test de Wilcoxon)

Avec $D = X_1 - X_2$:

$$H_0: P(X > \mu_0) = P(X < \mu_0) \rightarrow H_0: \tilde{\mu}_D = 0$$

$$H_1: P(X > \mu_0) > P(X < \mu_0) \rightarrow H_1: \tilde{\mu}_D < 0$$

²⁷ « = » signifie que la forme passe par l'axe, « ≠ » signifie que la forme ne passe pas par l'axe.

²⁸ « = » signifie qu'il n'y a pas de lien entre μ_1 et μ_2 , « ≠ » signifie qu'il existe un lien entre μ_1 et μ_2 .

$\alpha = 5\%$

$D = T_1 - T_2$	$ D $	rg moyen ($ D $)
-1	1	1
6	6	5.5
-6	6	5.5
-23	23	12
3	3	3.5
-9	9	7.5
-12	12	10.5
3	3	3.5
9	9	7.5
-2	2	2
-12	12	10.5
-10	10	9

$$T^+ = 5.5 + 3.5 + 3.5 + 7.5 = 20$$

$$T^- = 1 + 5.5 + 12 + 7.5 + 10.5 + 2 + 10.5 + 9 = 58$$

$$T^+ + T^- = 20 + 58 = 78$$

$$n' = 12 \Rightarrow \frac{12 \times (12+1)}{2} = 78$$

Programme R :

$$\rightarrow p = P(T^+ \leq 20) \cong 0.073$$

Résultats APA :

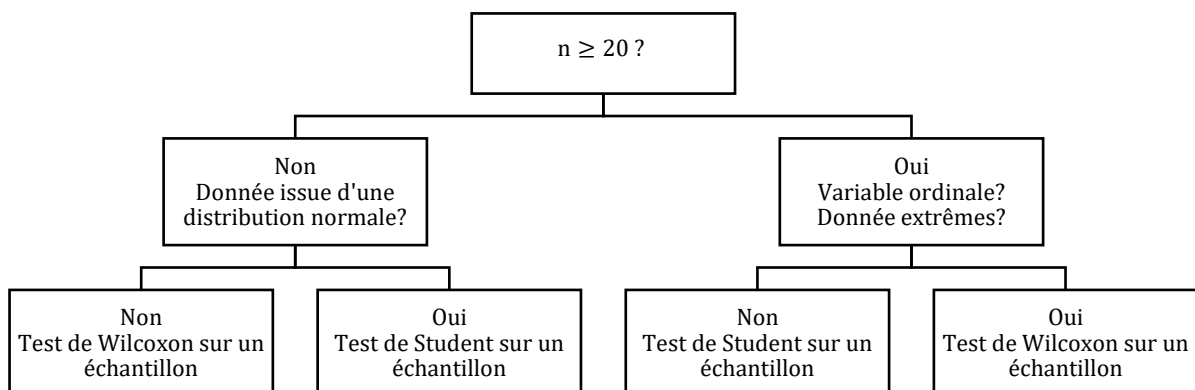
$$T^+ = 20, p = .073, \text{unilatéral}$$

Cours 16

16.04.18

Puissance = nombre de fois où on rejette un test hypothèse nulle.

Arbre de décision :

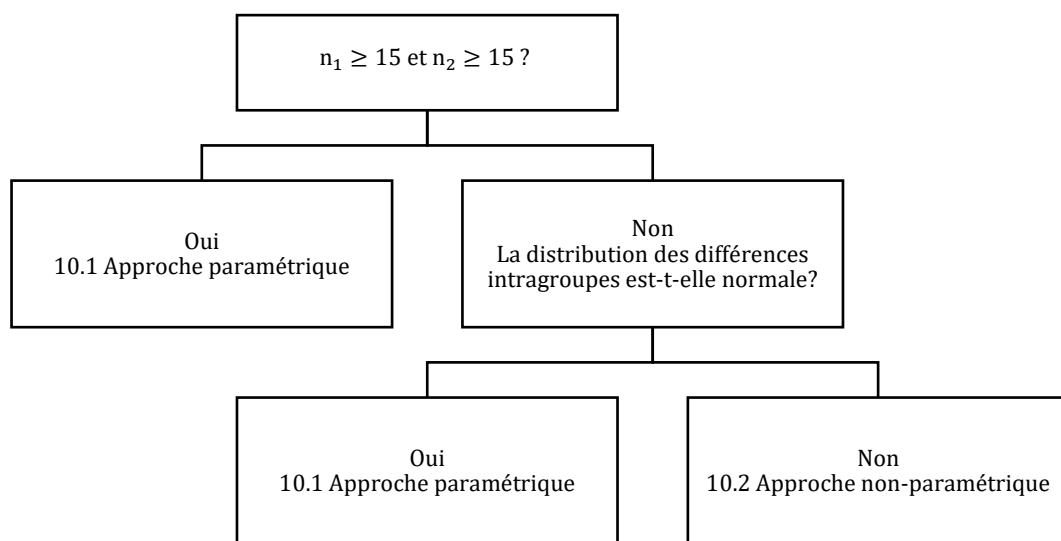


10. Tests d'hypothèses appliqués à 2 échantillons indépendants



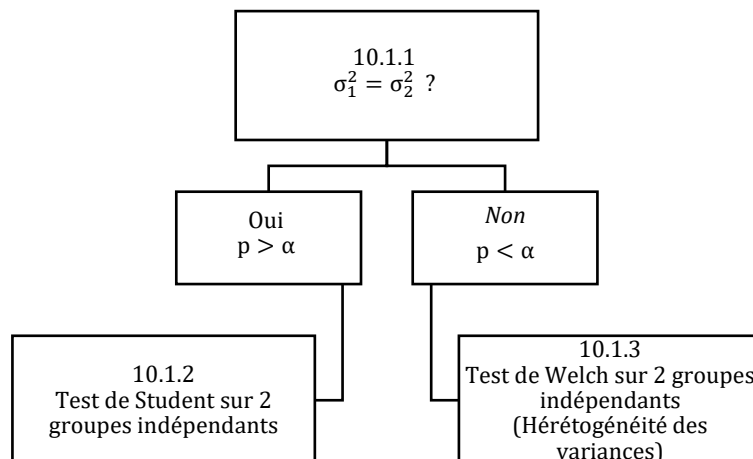
On peut comparer μ_1 et μ_2 , pour cela on extrait les échantillons. Afin de caractériser la dispersion, on utilise la variance.

2 échantillons indépendants :



Approche paramétrique :

Illustration (Réf. [7]) : Est-ce possible de raccourcir le temps d'endormissement ? Design : $n_1 = 29$ et $n_2 = 23$, alors on fait un test paramétrique.



10.1.1 Test d'homogénéité des variances de Levene (F)

Statistique descriptive :

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

	Completed	ToDo
n_j	29	28
\bar{x}_j	25.1	15.8
s_j	15.9	14.1

Dans JAMOVI : « Assumption Checks » : ☐ Normality

☒ Equality of variance (Levene's)

$^{29}F(m-1, n-m) : F(1,55) = 0.365, p = .548$. Comme $p > \alpha$, alors on ne rejette pas H_0 .

m : nombre de groupe et n : taille de l'échantillon

10.1.2 Test de Student sur 2 groupes indépendants

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right); \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

Estimation :

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

Variance :

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right))$$

Sous $H_0 : \mu_1 - \mu_2 = 0 \Rightarrow E(\bar{X}_1 - \bar{X}_2) = 0$

$$\text{Sous } H_0 : u = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

σ^2 n'est pas connue, cette variance doit donc être estimée :

échantillon	σ_1^2
1	s_1^2
2	$\frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{n_1 + n_2 - 2}$

²⁹ F de Fisher-Snedecor

Sous H_0 , la variable de décision prend la forme d'une loi de Student :

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{n_1 + n_2 - 2}}}$$

Sous H_0 , cette loi de Student suit une distribution du T à $n_1 + n_2 - 2$ degré de liberté (ddl) :

$$T \sim T(n_1 + n_2 - 2)$$

Il ne reste plus qu'à calculer la valeur empirique de la variable de décision t_{emp} , puis la p-valeur :

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}}}$$

$$p = P(T > |t_{emp}| | H_0)$$

Si $p < \alpha$, alors on rejette H_0 .

Illustration :

1. Hypothèse : $H_0: \mu_1 = \mu_2$ et $H_1: \mu_1 \neq \mu_2$
2. Choisir un seuil de confiance : $\alpha = 5\% = 0.05$
3. Calculer la valeur empirique de la variable de décision :

$$t_{emp} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}}} = \frac{25.1 - 15.8}{\sqrt{\left(\frac{1}{29} + \frac{1}{28}\right) \times \frac{(29 - 1) \times 15.9^2 + (28 - 1) \times 14.1^2}{29 + 28 - 2}}} = \frac{9.3}{3.986} = 2.33$$
4. Calculer la probabilité critique (dans R) : $p = 2 \times P(T(55) > |2.33|) = 0.24$
5. Comparaison et conclusion : Comme $p < \alpha$, on rejette H_0 .
 On s'endort mieux, lorsqu'on fait une ToDo liste.

Selon les normes APA, le résultat est le suivant :

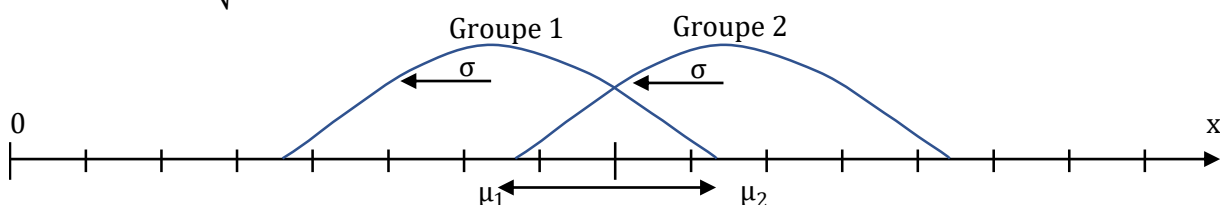
$$\rightarrow t(55) = 2.330, p = .024$$

Cours 17

23.04.18

Taille de l'effet :

$$d = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{n_1 + n_2 - 2}}} = \frac{|\mu_1 - \mu_2|}{\sigma} = \frac{|25.1 - 15.8|}{15.043} = 0.616$$



Interprétation de la taille d'effet (d de Cohen) :

0 ← Négligeable → 0.3 ← Petit → 0.5 ← Moyen → 0.8 ← Grande → 2.0

10.1.3 Test de Welch (T, ddl')

Sur JAMOVI → List specificity

	Completed	ToDo
n_j	29	28
\bar{x}_j	14.103	15.964
s_j	5.741	9.143

On peut faire un test paramétrique parce qu'on a plus de 15 participants par groupe.

Test d'homogénéité des variances (Levene) :

$F(1,55) = 8.843, p = .004$

Comme $p < \alpha$, on rejette H_0 et l'on opte pour $H_1 : \sigma_1^2 \neq \sigma_2^2$. Construisons la variable de décision pour le test de Welch :

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right); \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right) \Rightarrow \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Sous $H_0 : \mu_1 = \mu_2$: $u = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}} \sim N(0,1)$. Comme σ_1^2 et σ_2^2 ne sont pas connus, elles doivent être

estimées : $\hat{\sigma}_1^2 = s_1^2$; $\hat{\sigma}_2^2 = s_2^2$. En substituant σ_1^2 et σ_2^2 par leur estimation, on modifie la variable u qui devient une loi de Student :

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

T suit une loi de Student à degré de liberté prime (ddl') :

$$ddl' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Application :

1. Hypothèses : $H_0 : \mu_1 = \mu_2$ et $H_1 : \mu_1 \neq \mu_2$
2. Choisir le seul : $\alpha = 5\% = 0.05$
3. Valeur empirique de la variable de décision :

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{14.103 - 15.964}{\sqrt{\frac{5.741^2}{29} + \frac{9.143^2}{28}}} = \frac{-1.861}{2.030} = -0.9166$$

$$ddl' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{5.741^2}{29} + \frac{9.143^2}{28}\right)^2}{\left(\frac{5.741^2}{29}\right)^2 + \left(\frac{9.143^2}{28}\right)^2} = \frac{16.9912}{0.0461 + 0.3301} = 45.157$$

4. Calculer la p-valeur :
 $p = 2 \times P(T(ddl') > |t_{emp}| | H_0) = 2 \times P(T(45.157) > |-0.9166|) = .364$
 → Programme R :
 $2 * pt(abs(-0.9166), 45.157, lower.tail = FALSE) = .364$
5. Comparaison et conclusion : Comme $p \geq \alpha$, on ne rejette pas H_0 .
 La longueur moyennes des listes est la même dans les deux conditions expérimentales.

10.2 Approche non-paramétrique (Mann-Whitney U: U_1, U_2, T_1, T_2)

Sur JAMOV → Mann-Whitney U

	Hommes	Femmes
n_j	15	10

L'échantillon est trop petit, car il y a moins de 15 participants par groupe.

Vérifions l'hypothèse de normalité (test de Shapiro-Wilk) : $W = 0.866, p = .004$. Les données ne se distribuent pas normalement, nous réaliserons un test non-paramétrique.

Scores ordonnés Hommes (H)	7	11	15	16	17	18	18	18	21	23	25	25	26	29	50
Score ordonnés Femmes (F)	4	6	7	8	9	11	12	12	12	13					

Hypothèses :
 $H_0 : P(X_F < X_H) = P(X_F > X_H) | \tilde{\mu}_F = \tilde{\mu}_H$
 $H_1 : P(X_F < X_H) > P(X_F > X_H) | \tilde{\mu}_F < \tilde{\mu}_H$

Sous H_1 , on affirme que X_F est stochastiquement (=aléatoire) inférieure à X_H .

	H	7	11	15	16	17	18	18	18	21	23	25	25	26	29	50	
F	Rg	3.5	7.5	13	14	15	17	17	17	19	20	21	22	23	24	25	= 258
4	1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
6	2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
7	3.5	=	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
8	5	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
9	6	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
11	7.5	-	=	+	+	+	+	+	+	+	+	+	+	+	+	+	
12	10	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	
12	10	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	
12	10	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	
13	12	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	
= 67		+: "victoire" = +1					=: "égalité" = +0.5					-: "défaite" = -1					

$$U_1 = |\{(x_{1i}, x_{2j}) | x_{1j} < x_{2j}\}| = \text{"défaite" et "égalité"} = 11 \times 1 + 2 \times 0.5 = 12$$

$$U_2 = |\{(x_{1i}, x_{2j}) | x_{1j} > x_{2j}\}| = \text{"victoire" et "égalité"} = 138$$

$$U_1 + U_2 = 12 + 138 = 150$$

$$n_1 \times n_2 = 15 \times 10 = 150$$

La variable empirique de la variable de décision d'un test de Mann-Whitney est égal à U_1 .

$$T_1 = 3.5 + 7.5 + 13 + 14 + 15 + 17 + 17 + 17 + 19 + 20 + 21 + 22 + 23 + 24 + 25 = 258$$

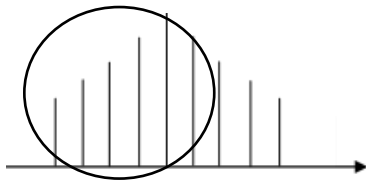
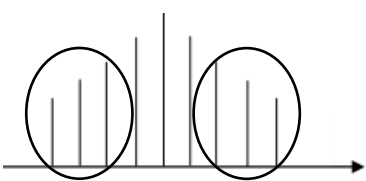
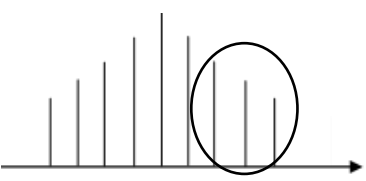
$$T_2 = 1 + 2 + 3.5 + 5 + 6 + 7.5 + 10 + 10 + 10 + 12 = 67$$

$$U_1 = n_1 \times n_2 + \frac{n_1 \times (n_1 + 1)}{2} - T_1 = 15 \times 10 + \frac{15 \times 16}{2} - 258 = 12$$

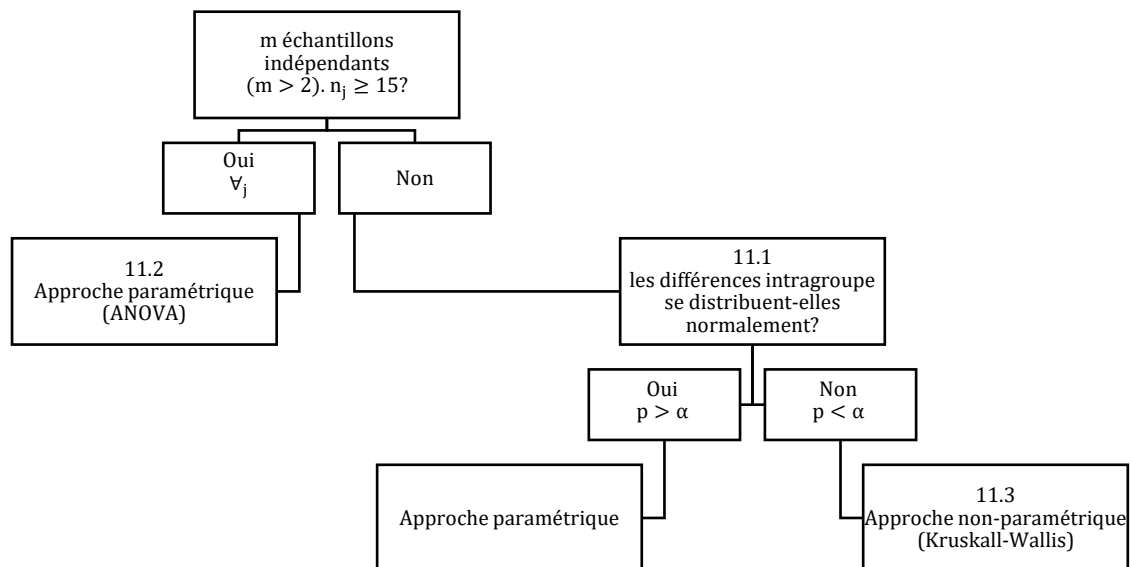
$$U_2 = n_1 \times n_2 + \frac{n_2 \times (n_2 + 1)}{2} - T_2 = 15 \times 10 + \frac{10 \times 11}{2} - 67 = 138$$

Cours 18

30.04.18

Test unilatéral à gauche	Test bilatéral	Test unilatéral à droite
$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$	$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$	$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$
$H_1: \tilde{\mu}_1 < \tilde{\mu}_2$	$H_1: \tilde{\mu}_1 \neq \tilde{\mu}_2$	$H_1: \tilde{\mu}_1 > \tilde{\mu}_2$
		
$-(X_{1i} < X_{2j}); +(X_{1i} > X_{2j})$		
$U_1: \#_-$ $U_2: \#_+$		
$p = P(U \leq U_2) = P(U \geq U_1)$	$p = 2 \times P(U \leq \min(U_1; U_2))$	$p = P(U \geq U_2) = P(U \leq U_1)$
Selon JAMOV, $U_{\text{emp}} = \min(U_1; U_2)$		

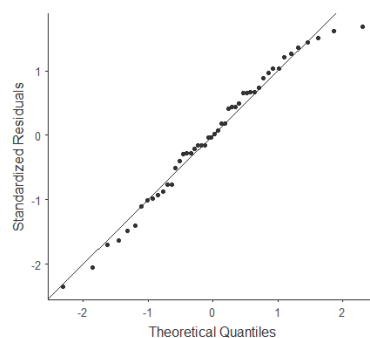
11. Tests d'hypothèses appliqués à plus de 2 groupes indépendants



11.1 Test de la forme de la distribution des résidus (ANOVA)

11.1.1 Méthode gravure

Q-Q plot : graphique des quantiles-quantiles



11.1.2 Méthode inférentielle

Démarche sur JAMOV I :

1. Calculer la moyenne des groupes (n, moyenne, écart-type)
2. Créer la variable moyenne « M » avec la fonction « IF »
3. Créer la variable résidus « R » en faisant la colonne : Scores – M
4. Réaliser le test de normalité de Shapiro-Wilk sur les résidus

Dans notre exemple, $W = 0.987$, $p = .918$.

11.2 Analyse de variances à un facteur de classification

11.2.1 Données et modèles

On dispose de m échantillon de tailles respectives $n_1, n_2, \dots, n_j, \dots, n_m$ correspondant aux m modalités de facteur de classification A. on pose :

$$n = \sum_{j=1}^m n_j$$

On a le tableau suivant :

	$a_1, a_2, \dots, a_j, \dots, a_m$
	$X_{11}, X_{21}, \dots, X_{j1}, \dots, X_{m1}$
	$X_{12}, X_{22}, \dots, X_{j2}, \dots, X_{m2}$
	$\vdots \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$
	$X_{1n}, X_{2n}, \dots, X_{jn}, \dots, X_{mn}$
\bar{X}_j	$\bar{X}_1, \bar{X}_2, \dots, \bar{X}_j, \dots, \bar{X}_m$
s_j^2	$s_1^2, s_2^2, \dots, s_j^2, \dots, s_m^2$

La moyenne générale vaut :

$$\bar{x} = \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^{n_j} x_{ji} = \frac{1}{n} \sum_{j=1}^m n_j \times \bar{x}_j$$

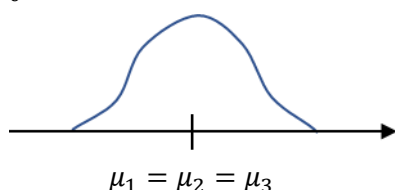
On considère souvent que les échantillons sont issus de distributions normales :

$$X_j \sim N(\mu_j, \sigma^2)$$

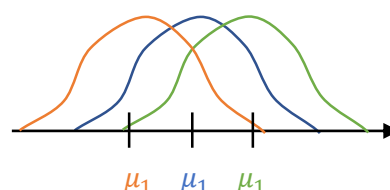
Le problème est de tester : $H_0: \mu_1 = \mu_2 = \dots = \mu_j = \dots = \mu_m$

$$H_1: \exists k \exists l \mu_k \neq \mu_l$$

Sous H_0 :



Sous H_1 :



11.2.2 Le test (Sommes des carrés - SC)

$$\bar{X} = \frac{1}{n} \sum n_j \times \bar{X}_j$$

$$(X_{ji} - \bar{X}) = (X_{ji} - \bar{X}_j) + (\bar{X}_j - \bar{X})$$

$$\underbrace{\sum_{j=1}^m \sum_{i=1}^{n_j} (X_{ji} - \bar{X})^2}_{SC_T \text{ Somme des carrés totale}} = \underbrace{\sum_j \sum_i (X_{ji} - \bar{X}_j)^2}_{SC_R \text{ Somme des carrés intragroupes}} + \underbrace{\sum_j n_j (\bar{X}_j - \bar{X})^2}_{SC_A \text{ Somme des carrés intergroupes}}$$

Rappel :

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ji} - \bar{X})^2 \Rightarrow \sum (X_{ji} - \bar{X}_j)^2 = (n_j - 1) \times s_j^2 \Rightarrow SC_R = \sum_{j=1}^m (n_j - 1) \times s_j^2$$

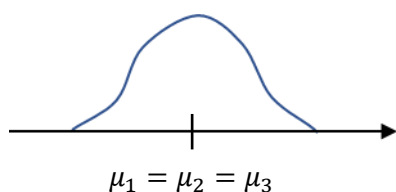
$$SC_T = SC_A + SC_R \text{ avec :}$$

$$SC_T = \sum_{j=1}^m \sum_{i=1}^{n_j} (X_{ji} - \bar{X})^2$$

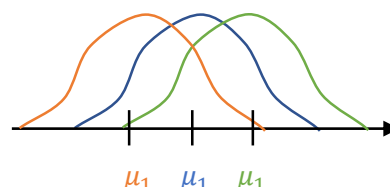
$$SC_A = \sum_j n_j \times (\bar{X}_j - \bar{X})^2$$

$$SC_R = \sum_{j=1}^m (n_j - 1) \times s_j^2$$

Sous H_0 : (positions similaires)



Sous H_1 : (positions différentes)



La dispersion est similaire pour les 2 graphes = la même variance	
$\hat{\sigma}^2 = \frac{SC_R}{n - m}$	$\frac{SC_R}{n - m} < \frac{SC_A}{m - 1}$
$\hat{\sigma}^2 = \frac{SC_A}{m - 1}$	La variable de décision sera donc le rapport entre $\frac{SC_R}{n-m}$ et $\frac{SC_A}{m-1}$

$$F = \frac{\frac{SC_A}{m-1}}{\frac{SC_R}{n-m}}$$

F devrait prendre, sous H_0 , une valeur proche de 1. Plus précisément, sous H_0 , la variable F suit une loi de Fisher-Snedecor à $m - 1$ et $n - m$ degré de liberté (ddl). La p-valeur vaut :

$$p = P(F \geq F_{emp} | H_0) \text{ avec } F \sim F(m - 1, n - m)$$

Table de l'ANOVA :

	SC	ddl	CM	F_{emp}
A	SC_A	$m - 1$	CM_A	$F_{emp} = \frac{CM_A}{CM_R}$
R	SC_R	$n - m$	CM_R	
T	SC_T	$n - 1$	\emptyset	

Application :

	Enfants	Parents	Adultes
n_j	14	13	13
\bar{x}_j	113.2143	85	74.0769
s_j	25.6970	32.0702	29.9902
s_j^2	660.336	1028.4977	899.4121

$$\bar{x} = \frac{1}{40} \sum_{j=1}^3 n_j \times \bar{x}_j = \frac{1}{40} (14 \times 113.2143 + 13 \times 85 + 13 \times 74.0769) = 91.325$$

$$SC_A = \sum_{j=1}^3 n_j \times (\bar{X}_j - \bar{X})^2 = 14(113.2143 - 91.325)^2 + 13(85 - 91.325)^2 + 13(74.0769 - 91.325)^2 = 11095.495$$

$$SC_R = \sum_{j=1}^m (n_j - 1) \times s_j^2 = 13 \times 660.336 + 12 \times 1028.4977 + 12 \times 899.4121 = 31717.2902$$

Table de l'ANOVA :

	SC	ddl	CM	F _{emp}
A	11095.5	3 - 1 = 2	$\frac{11095.5}{2} = 5547.2$	$F_{\text{emp}} = \frac{5547.2}{857.27} = 6.471$
R	31719.28	40 - 3 = 37	$\frac{31719.28}{37} = 857.27$	
T	11095.5 + 31719.28 = 42814.78	40 - 1 = 39	∅	

$$p = P(F(2,37) \geq 6.471) = .004$$

Programme R :

$$pf(6.471, 2, 37, \text{lower.tail} = FALSE) = 0.003891457$$

Résultats avec les normes AP :

$$F(2,37) = 6.471, p = .004$$

On rejette H_0 , parce que $p < \alpha$. Ces 3 groupes n'ont pas les mêmes capacités.

Taille d'effet :

$$\eta^2 = \frac{SC_A}{SC_T}$$

Interprétation du η :

$$0 \leftarrow \text{Small} \rightarrow 0.10 \leftarrow \text{Medium} \rightarrow 0.25 \leftarrow \text{Large} \rightarrow 0.40$$

11.2.3 Comparaisons multiples - Méthode de Bonferroni

Cette méthode est utilisée pour comparer tous les groupes les uns aux autres. Si on a m groupes, on va devoir faire c avec :

$$c = C_2^m = \frac{m!}{(m-2)! \times 2!} = \frac{m \times (m-1)}{2}$$

Rappel :

$$t_{\text{emp}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{n_1 + n_2 - 2}}} \rightarrow \hat{\sigma}^2 = \frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{n_1 + n_2 - 2}$$

# échantillons	$\hat{\sigma}^2$
1	s_1^2
2	$\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}$
3	$\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2 + (n_3 - 1) \times s_3^2}{n_1 + n_2 + n_3 - 3}$
\vdots	\vdots
m	$\frac{\sum (n_i - 1) \times s_i^2}{n - m} = CM_R$

$$\hat{\sigma}^2 = \frac{\sum (n_i - 1) \times s_i^2}{n - m} = \frac{SC_R}{n - m} = CM_R$$

La valeur empirique d'une comparaison réalisée selon la méthode de Bonferroni vaut :

$$t_{jk} = \frac{\bar{x}_j - \bar{x}_k}{\sqrt{\left(\frac{1}{n_j} + \frac{1}{n_k}\right) \times CM_R}}$$

$$H_0 [jk]: \mu_j = \mu_k$$

$$H_0 [jk]: \mu_j \neq \mu_k$$

Sous $H_0 [jk]$, T_{jk} suit une loi de Student à $n - m$ degré de liberté (ddl) :

$$p = P(T(n - m) \geq |t_{jk}|)$$

Comment éviter l'explosion de l'erreur de première espèce ? Taux d'erreur par comparaison (EC) : α^* ; Taux d'erreur sur l'ensemble (EE) : α . Lorsqu'on se place sous $H_0: \mu_1 = \mu_2 = \dots = \mu_m$ un lien existe entre α et α^* :

$$\alpha = 1 - \underbrace{(1 - \alpha^*)^c}_{\substack{\text{Probabilité de ne} \\ \text{jamais se tromper}}} = \underbrace{1 - (1 - \alpha^*)^c}_{\substack{\text{Probabilité de se} \\ \text{tromper 1 fois au moins}}}$$

$$\alpha^* = 0.05$$

m	$c = \frac{m \times (m-1)}{2}$	α
2	1	0.05
3	3	0.14
4	6	0.26
5	10	0.40
6	15	0.54

Rappel :

$$(1-x)^a = C_0^a \times x^0 - C_1^a \times x^1 + C_2^a \times x^2 - \dots + C_a^a \times x^a (-1)^a$$

Si x est petit par rapport à 1 : $(1-x)^a \cong C_0^a \times x^0 - C_1^a \times x^1 = 1 - a \times x$

$$\alpha = 1 - (1 - \alpha^*)^c = 1 - c \times \alpha^* = 1 - 1 + c \times \alpha^* = c \times \alpha^*$$

$$\alpha = c \times \alpha^* \rightarrow \alpha^* = \frac{\alpha}{c}$$

Cours 20

14.05.18

$$p^* = \frac{P}{c} \Leftrightarrow \alpha^* = \frac{\alpha}{c}$$

La probabilité critique d'une comparaison selon Bonferroni vaut donc :

$$p^* = \min(1, c \times p) \text{ avec } c = \frac{m \times (m-1)}{2} \rightarrow p = 2 \times P(T(n-m) \geq |t_{emp}|)$$

Règle de décision : si $p^* < \alpha$, on rejette H_0 .

Application :

	Enfants (1)	Parents (2)	Adultes (3)
n_j	14	13	13
\bar{x}_j	113.2143	85	74.0769
s_j	25.6970	32.0702	29.9902
s_j^2	660.336	1028.4977	899.4121
CM_R	857.278		

(1) vs (2) :

$$t_{[12]} = \frac{\bar{x}_j - \bar{x}_k}{\sqrt{\left(\frac{1}{n_j} + \frac{1}{n_k}\right) \times CM_R}} = \frac{28.214}{\frac{11.277}{SE_{sur JAMOVI}}} = 2.502$$

$$p = 2 \times P(T(n-m) > t_{emp}) = 2 \times P(T(37) > 2.502) = 0.017$$

Programme R :

$$2 * pt(t_{emp}, (n-m), lower.tail = FALSE) = 2 * pt(2.502, 37, lower.tail = FALSE) = 0.017$$

$$\text{Bonferroni : } p^* = \min(1, c \times p) = p^* = \min(1, 3 \times p) = \min(1, 3 \times 0.017) = \min(1, 0.051) = .051$$

11.2.4 Comparaisons multiples - Méthode de Holm

Procédure itérative :

I	II	III
Les p-valeurs sont ordonnées	Les p-valeurs sont ajustées selon leur position	Calcul du maximum cumulatif
P[1]	$c \times P[1]$	$\min(1, \max(c \times P[1]))$
P[2]	$c \times P[1], (c - 1) \times P[2]$	$\min(1, \max(c \times P[1], (c - 1) \times P[2]))$
P[3]	$c \times P[1], (c - 1) \times P[2], (c - 2) \times P[3]$	$\min(1, \max(c \times P[1], (c - 1) \times P[2], (c - 2) \times P[3]))$

Exemple :

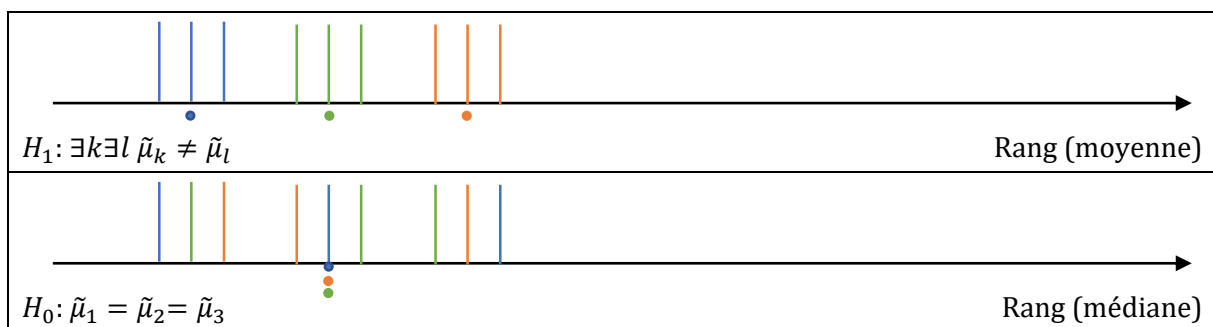
I	II	III
0.001	0.003	0.003
0.017	0.034	0.034
0.348	0.348	0.348

Holm			Bonferroni			
Enfant	a		Enfant	a		
Parent		b	Parent		ab	
Adulte		b	Adulte			b

11.3 Test non-paramétrique (Kruskal-Wallis)

Exemple : nous pensons aussi avec les mains. < 15, alors on fait un test d'hypothèse Kruskal-Wallis :

Le principe est simple : il suffit de réaliser une analyse de variance sur les rangs :



$$H \propto \text{Var}(\bar{R}_j) \text{ avec } \text{Var}(\bar{R}_j) = \frac{1}{n} \sum n_j (\bar{R}_j - \bar{R})^2 \text{ avec } \bar{R}_j = \frac{T_j}{n_j}$$

$$\bar{R}_j = \frac{1}{n} \sum n_j \times \bar{R}_j = \frac{n+1}{2} \text{ avec } n = \sum_{j=1}^m n_j$$

Gestes encouragés	Rg	Gestes permis	Rg	Gestes interdits	Rg
2	1	6	7	10	18.5
3	2.5	8	11.5	12	21.5
3	2.5	8	11.5	14	28
5	4	9	15.5	14	28
6	7	9	15.5	14	28
6	7	9	15.5	14	28
6	7	9	15.5	14	28
6	7	11	20	15	32
8	11.5	13	23.5	16	33
8	11.5	13	23.5	20	34
10	18.5	14	28	21	35
12	21.5	14	28	26	36

$T_j :$ $= 101$ $= 215$ $= 350$

	Gestes encouragés	Gestes permis	Gestes interdits
T_j	101	215	350
$R_j = \frac{T}{n_j}$	8.4167	17.9167	29.1667
$\bar{R} = \frac{n+1}{2}$	18.5		

Valeur empirique :

$$H' = \frac{12}{n \times (n+1)} \times \sum n_j \times (\bar{R}_j - \bar{R})^2$$

Lorsqu'il y a ex-aequo, H' doit être corrigé :

$$H = \frac{\frac{12}{n \times (n+1)} \times \sum n_j \times (\bar{R}_j - \bar{R})^2}{1 - \frac{\sum (t_k^3 - t_k)}{n^3 - n}}$$

TIES (ex-aequo)	2	3	5	6	8	9	10	11	12	13	14	15	16	20	21	26
t_k	1	2	1	5	4	4	2	1	2	2	7	1	1	1	1	1

$$H' = \frac{12 \times 12}{36 \times 37} \left[\left(\frac{101}{12} - 18.5 \right)^2 + \left(\frac{215}{12} - 18.5 \right)^2 + \left(\frac{350}{12} - 18.5 \right)^2 \right] = 23.3288$$

$$1 - \frac{\sum (t_k^3 - t_k)}{n^3 - n} = 1 - \frac{6 + 120 + 60 + 60 + 6 + 6 + 6 + 336}{46620} = 0.9871$$

$$H = \frac{23.3288}{0.9871} = 23.633$$

Sous H_0 , $H \sim \chi^2(m-1)$. La variable de décision H suit une loi du khi carré à $m-1$ degré de liberté (ddl). La probabilité critique vaut donc :

$$p = P(\chi^2(m-1) > H) = P(\chi^2(2) > 23.633) < .001$$

Loi normale centrée réduite 1

J'ai une probabilité, je cherche un quantile. (Φ)

6.5.1 Fonction de répartition de la loi normale réduite

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9983	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Loi normale centrée réduite 2

J'ai un quantile, je cherche une probabilité. (z, u)

6.5.2 Fractiles de la loi normale réduite

p	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.50	0.0000	0.0003	0.0005	0.0006	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012
0.51	0.0025	0.0028	0.0030	0.0033	0.0035	0.0038	0.0040	0.0043	0.0045	0.0048
0.52	0.0050	0.0053	0.0055	0.0058	0.0060	0.0063	0.0065	0.0068	0.0070	0.0073
0.53	0.0075	0.0078	0.0080	0.0083	0.0085	0.0088	0.0090	0.0093	0.0095	0.0098
0.54	0.0100	0.0103	0.0105	0.0108	0.0111	0.0113	0.0116	0.0118	0.0121	0.0123
0.55	0.126	0.128	0.131	0.133	0.136	0.138	0.141	0.143	0.146	0.148
0.56	0.151	0.154	0.156	0.159	0.161	0.164	0.166	0.169	0.171	0.174
0.57	0.176	0.179	0.181	0.184	0.187	0.189	0.191	0.194	0.197	0.199
0.58	0.202	0.204	0.207	0.210	0.212	0.215	0.217	0.220	0.222	0.225
0.59	0.228	0.230	0.233	0.235	0.238	0.240	0.243	0.246	0.248	0.251
0.60	0.253	0.256	0.259	0.261	0.264	0.266	0.269	0.272	0.274	0.277
0.61	0.279	0.282	0.285	0.287	0.290	0.292	0.295	0.298	0.300	0.303
0.62	0.305	0.308	0.311	0.313	0.316	0.319	0.321	0.324	0.327	0.329
0.63	0.332	0.335	0.337	0.340	0.342	0.345	0.348	0.350	0.353	0.356
0.64	0.358	0.361	0.364	0.366	0.369	0.372	0.375	0.377	0.380	0.383
0.65	0.385	0.388	0.391	0.393	0.396	0.399	0.402	0.404	0.407	0.410
0.66	0.412	0.415	0.418	0.421	0.423	0.426	0.429	0.432	0.434	0.437
0.67	0.440	0.443	0.445	0.448	0.451	0.454	0.457	0.459	0.462	0.465
0.68	0.466	0.470	0.473	0.476	0.479	0.482	0.485	0.487	0.490	0.493
0.69	0.496	0.499	0.502	0.504	0.507	0.510	0.513	0.516	0.519	0.522
0.70	0.524	0.527	0.530	0.533	0.536	0.539	0.542	0.545	0.548	0.550
0.71	0.553	0.556	0.559	0.562	0.565	0.568	0.571	0.574	0.577	0.580
0.72	0.583	0.586	0.589	0.592	0.595	0.598	0.601	0.604	0.607	0.610
0.73	0.613	0.616	0.619	0.622	0.625	0.628	0.631	0.634	0.637	0.640
0.74	0.643	0.646	0.650	0.653	0.656	0.659	0.662	0.665	0.668	0.671
0.75	0.674	0.678	0.681	0.684	0.687	0.690	0.693	0.697	0.700	0.703
0.76	0.706	0.710	0.713	0.716	0.719	0.722	0.726	0.729	0.732	0.736
0.77	0.739	0.742	0.745	0.749	0.752	0.755	0.759	0.762	0.765	0.769
0.78	0.772	0.776	0.779	0.782	0.786	0.789	0.793	0.796	0.800	0.803
0.79	0.806	0.810	0.813	0.817	0.820	0.824	0.827	0.831	0.834	0.838
0.80	0.842	0.845	0.849	0.852	0.856	0.860	0.863	0.867	0.871	0.874
0.81	0.878	0.882	0.885	0.889	0.893	0.896	0.900	0.904	0.908	0.912
0.82	0.915	0.919	0.923	0.927	0.931	0.935	0.938	0.942	0.946	0.950
0.83	0.954	0.958	0.962	0.966	0.970	0.974	0.978	0.982	0.986	0.990
0.84	0.994	0.999	1.003	1.007	1.011	1.015	1.019	1.024	1.028	1.032
0.85	1.036	1.041	1.045	1.049	1.054	1.058	1.063	1.067	1.071	1.076
0.86	1.080	1.085	1.089	1.094	1.098	1.103	1.108	1.112	1.117	1.122
0.87	1.126	1.131	1.136	1.141	1.146	1.150	1.155	1.160	1.165	1.170
0.88	1.175	1.180	1.185	1.190	1.195	1.200	1.206	1.211	1.216	1.221
0.89	1.227	1.232	1.237	1.243	1.248	1.254	1.259	1.265	1.270	1.276
0.90	1.282	1.287	1.293	1.299	1.305	1.311	1.317	1.323	1.329	1.335
0.91	1.341	1.347	1.353	1.359	1.365	1.372	1.379	1.385	1.392	1.398
0.92	1.405	1.412	1.419	1.426	1.433	1.440	1.447	1.454	1.461	1.468
0.93	1.476	1.483	1.491	1.499	1.506	1.514	1.522	1.530	1.538	1.546
0.94	1.555	1.563	1.572	1.580	1.589	1.598	1.607	1.616	1.626	1.635
0.95	1.645	1.655	1.665	1.675	1.685	1.695	1.706	1.717	1.728	1.739
0.96	1.751	1.762	1.774	1.787	1.799	1.812	1.825	1.838	1.852	1.866
0.97	1.881	1.896	1.911	1.927	1.943	1.960	1.977	1.995	2.014	2.034
0.98	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257	2.290
0.99	2.326	2.366	2.409	2.457	2.512	2.576	2.652	2.748	2.878	3.090

Triangle de Pascal

Toujours commencer à compter de 0

$$5 = C_4^5$$

$$C_7^{15} = 6435$$

[illegible]

Signification des lettres grecques et symboles divers

\bar{d} : moyenne des différences	Λ : et	α : alpha : taux pour probabilité
r_s : corrélation de Spearman	\vee : ou	$\alpha^* : p^* : p$ -critique
r_{xy} : coefficient de Bravais-Pearson	\cap : intersection	β : beta
s_x : écart-type	\cup : union	δ : delta min
s_x^2 : variance	∞ : infini	Δ : delta maj : une différence (soustraction)
\bar{x} : moyenne ou espérance	c : coefficient de contingence	θ : thêta : degré d'un cercle (360°)
$\hat{\gamma}$: gamma : Gamma de Goodman-Kruskal	c_i : centre de la classe i	λ : lambda
$\hat{\mu}$: moyenne des valeurs théoriques	D : différence	μ : mu : moyenne
$\tilde{\mu}$: mu til : médiane	d_i : densité de fréquence de la classe i	π : pi : probabilité
$\hat{\pi}$: probabilité empirique	f : fréquence	ρ : rhô
$\hat{\sigma}$: estimation d'écart-type	f_i : fréquence relative de la classe i	Σ : sigma maj : la somme (addition)
σ^2 : variance	i : numéro de la classe	σ : sigma min : écart-type
τ_b : tau : Tau b de Kendall	m : nombre de groupe	Ω : omega maj : univers, population
χ^2 : χ carré : correction de continuité de Yates	n ou n_i : effectif de la classe i , taille	ω : omega min : individu
\forall : quelque soit	P : probabilité	Φ : phi maj
\exists : il existe	R : résidus	φ : phi min : coefficient φ
\propto : est proportionnel	r : coefficient de corrélation	
\int : intégrale	RS : résidus standardisés	