UE4 : BIOSTATISTIQUES
PIFO

FICHE N°4 : TESTS STATISTIQUES

1. Principe général.

1.1. Avant le recueil de données.

1.1.1. Etape n°1.

L'étape n°1 consiste à définir 2 hypothèses jouant des rôles dissymétriques :

- **H_0** (choisie pour être l'hypothèse nulle).
- **H_1** (hypothèse alternative).

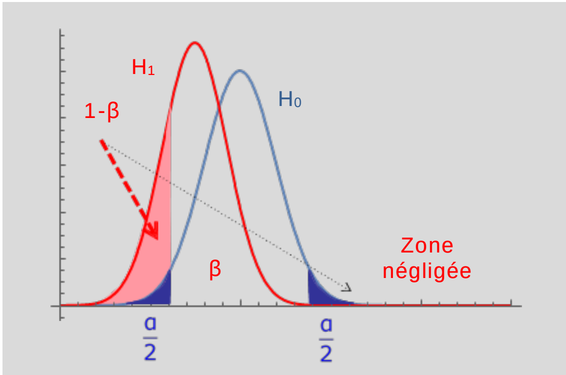
1.1.2. Etape n°2.

On suppose que H_0 est vraie.

On cherche alors à **définir une variable aléatoire (ou paramètre) Z dont on connait alors la distribution**.

1.1.3. Etape n°3.

Lors de l'étape n°3, un **seuil** est choisi.

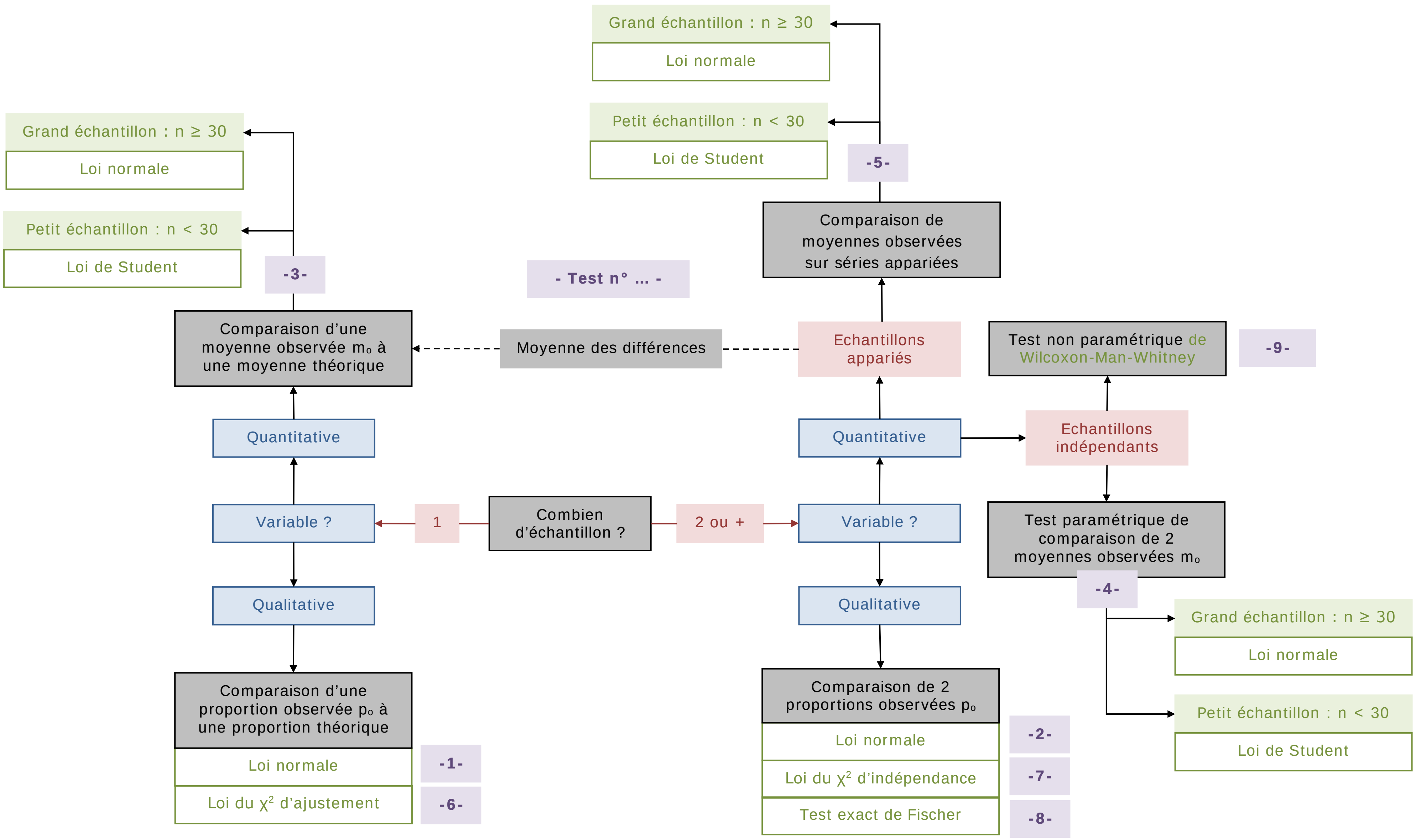


Typiquement $\alpha = 0{,}05$.

Ensuite un intervalle de pari est construit pour le paramètre Z de niveau de confiance « $1 - \alpha$ » noté $IP_{1-\alpha}$.

	H_0 vraie / H_1 fausse	H_0 fausse / H_1 vraie
Rejet de H_0	α : Risque de 1ère espèce	$1 - \beta$: Puissance
Non rejet de H_0	$1 - \alpha$: Confiance	β : Risque de 2ème espèce

2. Quel test ?


- Test n° … -
Combien d'échantillon ?
1
2 ou +
Variable ?
Quantitative
Qualitative
Comparaison d'une moyenne observée m₀ à une moyenne théorique
-3-
Grand échantillon : n ≥ 30
Loi normale
Petit échantillon : n < 30
Loi de Student
Comparaison d'une proportion observée p₀ à une proportion théorique
Loi normale
-1-
Loi du χ² d'ajustement
-6-
Variable ?
Quantitative
Qualitative
Echantillons appariés
Moyenne des différences
Comparaison de moyennes observées sur séries appariées
-5-
Grand échantillon : n ≥ 30
Loi normale
Petit échantillon : n < 30
Loi de Student
Echantillons indépendants
Test non paramétrique de Wilcoxon-Man-Whitney
-9-
Test paramétrique de comparaison de 2 moyennes observées m₀
-4-
Grand échantillon : n ≥ 30
Loi normale
Petit échantillon : n < 30
Loi de Student
Comparaison de 2 proportions observées p₀
Loi normale
-2-
Loi du χ² d'indépendance
-7-
Test exact de Fischer
-8-

2.1. Test n°1.

Il s'agit de comparer une proportion observée p_o d'un échantillon à une proportion théorique p d'une population.

<table>
<tr><th>Loi</th><th>Paramètres connus</th><th>Conditions de validité</th><th>Paramètre</th><th colspan="3">Décision</th></tr>
<tr><td rowspan="3">Normale centrée réduite</td><td rowspan="3">Proportion obs. de l'éch. : p_o
Proportion théo. de la pop. : p
Taille de l'éch. : n</td><td rowspan="3">$n \cdot p \geq 5$
$n \cdot (1-p) \geq 5$</td><td rowspan="3">$Z = \dfrac{p_o - p}{\sqrt{\dfrac{p \cdot (1-p)}{n}}}$</td><td>Méthode de la différence</td><td>De l'α choisie dans votre énoncé (si ce n'est pas précisé α = 5%), on détermine un u_α à partir de la table de la loi normale centrée réduite.</td><td>Si $Z \in [-u_\alpha \dots u_\alpha]$, NRH_0
Si $Z \notin [-u_\alpha \dots u_\alpha]$, RH_0</td></tr>
<tr><td>Méthode du « petit p »</td><td>De Z, vous déterminez le degré de signification p lu dans le tableau de la loi normale centrée réduite.</td><td>Si $p \geq \alpha$, NRH_0
Si $p < \alpha$, RH_0</td></tr>
<tr><td>Méthode de l'IC</td><td>$IC_{1-\alpha} = \left[p_o \pm u_\alpha \cdot \sqrt{\dfrac{p_o \cdot (1-p_o)}{n}}\right] \Rightarrow$</td><td>Si $p \in IC_{1-\alpha}$, NRH_0
Si $p \notin IC_{1-\alpha}$, RH_0</td></tr>
</table>

2.2. Test n°2.

Il s'agit de comparer 2 proportions observées p_o de 2 échantillons.

<table>
<tr><th>Loi</th><th>Paramètres connus</th><th>Conditions de validité</th><th>Paramètre</th><th colspan="3">Décision</th></tr>
<tr><td rowspan="2">Normale centrée réduite</td><td rowspan="2">Proportion obs. de l'éch. 1 : p_1
Proportion obs. de l'éch. 2 : p_2
Taille de l'éch. 1 : n_1
Taille de l'éch. 2 : n_2</td><td rowspan="2">$n_1 \cdot p \geq 5$
$n_1 \cdot (1-p) \geq 5$
$n_2 \cdot p \geq 5$
$n_2 \cdot (1-p) \geq 5$
avec
$p = \dfrac{n_1 \cdot p_1 + n_2 \cdot p_2}{n_1 + n_2}$</td><td rowspan="2">$Z = \dfrac{p_1 - p_2}{\sqrt{p \cdot (1-p) \cdot \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$</td><td>Méthode de la différence</td><td>De l'α choisie dans votre énoncé (si ce n'est pas précisé α = 5%), on détermine un u_α à partir de la table de la loi normale centrée réduite.</td><td>Si $Z \in [-u_\alpha \dots u_\alpha]$, NRH_0
Si $Z \notin [-u_\alpha \dots u_\alpha]$, RH_0</td></tr>
<tr><td>Méthode du « petit p »</td><td>De Z, vous déterminez le degré de signification p dans le tableau de la loi normale centrée réduite.</td><td>Si $p \geq \alpha$, NRH_0
Si $p < \alpha$, RH_0</td></tr>
</table>

2.4. Test n°4.

Il s'agit de comparer 2 moyennes sur échantillons indépendants.

Loi	Paramètres connus	Conditions de validité	Paramètre	Décision		
Normale centrée réduite	Moyenne obs. de l'éch. 1 : m_1 Moyenne obs. de l'éch. 2 : m_2 Taille de l'éch. 1 : n_1 Taille de l'éch. 2 : n_2 Variance obs. de l'éch. 1 : s_{o1}^2 Variance obs. de l'éch. 2 : s_{o2}^2 NB. on doit estimer les variances théo. des populations dont sont issus les 2 échantillons : $s_1^2 = \frac{n_1}{n_1 - 1} \cdot s_{o1}^2$ et $s_2^2 = \frac{n_2}{n_2 - 1} \cdot s_{o2}^2$	Grand échantillon n_1 et $n_2 \geq 30$ (Théorème central limite)	$Z = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Méthode de la différence	De l'α choisie dans votre énoncé (si ce n'est pas précisé α = 5%), on détermine un u_α à partir de la table de la loi normale centrée réduite.	Si $Z \in [-u_\alpha \ldots u_\alpha]$, NRH_0 Si $Z \notin [-u_\alpha \ldots u_\alpha]$, RH_0
				Méthode du « petit p »	De Z, vous déterminez le degré de signification p dans le tableau de la loi normale centrée réduite.	Si $p \geq \alpha$, NRH_0 Si $p < \alpha$, RH_0
		Petit échantillon (n_1 et/ou n_2 < 30) X doit suivre une loi gaussienne NB. $S_1 \sim S_2$	$Z = \frac{m_1 - m_2}{\sqrt{s^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ avec $s^2 = \frac{(n_1 - 1) \cdot s_1^2 - (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}$	Méthode de la différence	De l'α choisie dans votre énoncé (si ce n'est pas précisé α = 5%), on détermine un $t_{\alpha,ddl}$ à $n_1 + n_2$ -2 ddl à partir de la table de Student.	Si $Z \in [-t_{\alpha,n-1ddl} \ldots t_{\alpha,n-1ddl}]$, NRH_0 Si $Z \notin [-t_{\alpha,n-1ddl} \ldots t_{\alpha,n-1ddl}]$, RH_0
				Méthode du « petit p »	De Z, vous déterminez le degré de signification p dans le tableau de la loi de Student.	Si $p \geq \alpha$, NRH_0 Si $p < \alpha$, RH_0

2.6. Test n°6.

Il s'agit de comparer une proportion observée p_o d'un échantillon à une proportion théorique p d'une population.

<table>
<tr><th>Loi</th><th>Paramètres connus</th><th>Hypothèses</th><th>CV</th><th>Principe</th><th>Paramètre</th><th>Décision</th></tr>
<tr>
<td>X^2 d'ajustement</td>
<td>Répartition théorique : On connait les proportions théoriques des k modalités d'une variable qualitative.

Répartition observée : On connait les effectifs observés des k modalités d'une variable qualitative sur un échantillon de taille n.</td>
<td>H_0 : Les répartitions sont identiques.

H_1 : Les répartitions sont différentes.</td>
<td>$C_i \geq 5$</td>
<td>La clé du principe consiste au mélange des 2 populations pour calculer une pseudo-répartition théorique si elle n'est pas donnée dans l'énoncé.

1. Construction d'un tableau de contingence qui contient les résultats expérimentaux.
<table>
<tr><th>Modalités</th><th>1</th><th>...</th><th>k</th><th>Total</th></tr>
<tr><td>Ech.</td><td>O_1</td><td>...</td><td>O_k</td><td>n</td></tr>
</table>
Et si on connait les proportions théoriques :
<table>
<tr><th>Modalités</th><th>1</th><th>...</th><th>k</th><th>Total</th></tr>
<tr><td>Ech.</td><td>O_1</td><td>...</td><td>O_k</td><td>n</td></tr>
<tr><td>Pop.</td><td>C_1</td><td>...</td><td>C_k</td><td>n</td></tr>
</table>
2. Si on ne connait pas les proportions théoriques, il faut construire une pseudo-répartition à partir des résultats expérimentaux :

$C_i = n \cdot p_i$ avec p_i, proportions théoriques.

3. Calcul du paramètre du test.</td>
<td>$Q = \sum_{j=1}^{\text{nb cases tableau}} \frac{(O_j - C_j)^2}{C_j}$</td>
<td>Si $Q \in [0 \ldots K_{\alpha,ddl}]$, NRH_0
Si $Q \notin [0 \ldots K_{\alpha,ddl}]$, RH_0</td>
</tr>
</table>

NB_1. Le nombre de ddl correspond à k-1.

NB_2. De l'α choisie dans votre énoncé (si ce n'est pas précisé α = 5%) et du nombre de degré de liberté (c'est-à-dire (k-1)), on détermine un $K_{\alpha,\,ddl}$ à partir de la table du χ^2 :

ddl \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588

Pour α = 5% et 10 modalités, on lit dans le tableau la valeur pour laquelle on a α = 0,05 et 9 degrés de liberté (ddl) c'est-à-dire k-1, k étant le nombre de modalité. On obtient ici $K_{5\%,9ddl}$ = 16,919.

Exemple à faire avec calculette.

2.7. Test n°7.

Il s'agit de comparer 2 proportions observées p_o de 2 échantillons.

<table>
<tr><th>Loi</th><th>Paramètres connus</th><th>Hypothèses</th><th>CV</th><th>Principe</th><th>Paramètre</th><th>Décision</th></tr>
<tr>
<td>X^2 d'indépendance</td>
<td>On a 2 variables qualitatives à k et m modalités.

1 seul échantillon est étudié pour lequel on relève, pour chacun des sujets, la valeur des 2 variables.</td>
<td>H_0 : Les 2 variables sont indépendantes.

H_1 : Les 2 variables sont liées.</td>
<td>$C_{ij} \geq 5$</td>
<td>La clé du principe consiste au mélange des 2 populations pour calculer une pseudo-répartition théorique si elle n'est pas donnée dans l'énoncé.

1. Construction d'un tableau de contingence qui contient les résultats expérimentaux.
<table>
<tr><th>Modalités</th><th>1</th><th>…</th><th>k</th><th>Total</th></tr>
<tr><td>1'</td><td>$O_{1\ (mod.\ 1')}$</td><td>…</td><td>$O_{k\ (mod.\ 1')}$</td><td>n_1</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>m</td><td>$O_{1\ (mod.\ m)}$</td><td>…</td><td>$O_{k\ (mod.\ m)}$</td><td>n_m</td></tr>
</table>
2. Construction d'une pseudo-répartition en mélangeant les résultats expérimentaux c'est-à-dire en oubliant leur origine.
<table>
<tr><th>Modalités</th><th>1</th><th>…</th><th>k</th><th>Total</th></tr>
<tr><td>Mélange</td><td>$O_{1\ (mod.\ 1')}$
+ …+
$O_{1\ (mod.\ m)}$</td><td>…</td><td>$O_{k\ (mod.\ 1')}$
+ …+
$O_{k\ (mod.\ m)}$</td><td>n_i
+ … +
n_m</td></tr>
<tr><td>Fréq^{ce}</td><td>$p_1 = \frac{\sum_{i=1'}^{m} O_{1(mod.\ i)}}{\sum n_i}$</td><td>…</td><td>$p_k = \frac{\sum_{i=1'}^{m} O_{k(mod.\ i)}}{\sum n_i}$</td><td>1</td></tr>
</table>
3. Construction du tableau des effectifs attendus
<table>
<tr><th>Modalités</th><th>1</th><th>…</th><th>k</th><th>Total</th></tr>
<tr><td>1'</td><td>$C_{1\ (mod.\ 1')} = n_1 \cdot p_1$</td><td>…</td><td>$C_{k\ (mod.\ 1')} = n_1 \cdot p_k$</td><td>$n_1$</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>m</td><td>$C_{1\ (mod.\ m)} = n_m \cdot p_1$</td><td>…</td><td>$C_{k\ (mod.\ m)} = n_m \cdot p_k$</td><td>$n_m$</td></tr>
</table>
4. Calcul du paramètre du test.</td>
<td>$Q = \sum_{i,j}^{\text{nb cases tableau}} \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$</td>
<td>Si $Q \in [0 \ldots K_{\alpha,ddl}]$, NRH_0
Si $Q \notin [0 \ldots K_{\alpha,ddl}]$, RH_0</td>
</tr>
</table>

NB₁. Le nombre de ddl correspond à $(k-1)\cdot(m-1)$.

- Etape n°2 : On doit construire une pseudo-répartition en mélangeant les résultats expérimentaux c'est-à-dire en oubliant leurs origines.

	aa	aA	AA	Total
Mélange	34	52	24	110
Fréquence	0,31	0,47	0,22	1

- Etape n°3 : On doit construire un tableau des effectifs attendus. Ces valeurs sont nos « C_i ».

	aa	aA	AA	Total
Catégorie 1	13,3	20,3	9,4	43
Catégorie 2	9,3	14,2	6,5	30
Catégorie 3	11,4	17,5	8,1	37
Total	34	52	24	110

Les conditions du test sont validées car toutes les valeurs de ce tableau sont supérieures ou égales à 5.

- Etape n°4 : On calcule le paramètre.

$$Q = \underbrace{\frac{(15-13,3)^2}{13,3}}_{Cat1-aa} + \underbrace{\frac{(19-20,3)^2}{20,3}}_{Cat1-aA} + \underbrace{\frac{(9-9,4)^2}{9,4}}_{Cat1-AA} + \underbrace{\frac{(11-9,3)^2}{9,3}}_{Cat2-aa} + \underbrace{\frac{(13-14,2)^2}{14,2}}_{Cat2-aA} + \underbrace{\frac{(6-6,5)^2}{6,5}}_{Cat2-AA} + \underbrace{\frac{(8-11,4)^2}{11,4}}_{Cat3-aa} + \underbrace{\frac{(20-17,7)^2}{17,5}}_{Cat3-aA} + \underbrace{\frac{(9-8,1)^2}{8,1}}_{Cat3-AA} = 2,183$$

- Etape n°5 : Le nombre de ddl = (3-1)·(3-1) = 4 degrés de liberté. Soit $K_{0,05\ ;\ 4\ ddl}$ = 9,488.

Comme Q appartient à [0 ; $K_{0,05\ ;\ 4\ ddl}$], on ne peut pas conclure à un lien entre génotype et sous-type.

2.8. Test n°8.

Parfois, on souhaite comparer 2 répartitions observées mais les effectifs calculés sont trop petits pour utiliser le test du Chi-2. Pour contourner cette difficulté des petits effectifs, on peut utiliser un test qui utilise directement la loi binomiale (ou multinomiale) : En pratique on rentre les effectifs dans un tableau de contingence et l'ordinateur donne directement le p.

Loi	Paramètres connus	Conditions de validité	Paramètre	Décision
Fischer	le petit p	Aucune	Aucun calcul relatif à ce test n'est au programme : le petit p vous sera donné. NB$_1$. Test exact non paramétrique. NB$_2$. Loi binomiale ou multinomiale.	Si $p \geq \alpha$, NRH_0 Si $p < \alpha$, RH_0

Remarque suite à la résolution des exercices de votre enseignant :

- **Lors du croisement de 2 variables qualitatives dont l'une est à 2 classes, la seconde, à 3 classes, un test non paramétrique de Wilcoxon-Mann-Withney n'est pas possible.**
- **Lors du croisement d'une variable qualitative et d'une variable qualitative à 2 classes, un test non paramétrique de Wilcoxon-Mann-Withney est possible.**

Exemple pour l'étape n°2 :

ENONCE : Sur les 12 premiers patients traités par X et les 6 premiers patients traités par le traitement habituel, on a mesuré la diminution de la masse tumorale 1 an après l'initiation du traitement.

Les résultats sont présentés ci-dessous (en cm, une valeur positive correspond à une augmentation de la taille). On utilise le test de Mann-Whitney-Wilcoxon.

X	**-7,2**	**-3,3**	**-3,1**	**-2,2**	**-1,6**		**-0,9**		
Traitement habituel						**-1,5**		**-0,8**	**-0,5**

X	**0,0**	**0,2**	**0,8**	**1,4**		**2,3**	**3,9**		
Traitement habituel					**2,2**			**4,1**	**6,3**

CORRECTION :

RANG	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
X (B)	-7,2	-3,3	-3,1	-2,2	-1,6		-0,9		
Traitement habituel (A)						-1,5		-0,8	-0,5

RANG	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
X (B)	0,0	0,2	0,8	1,4		2,3	3,9		
Traitement habituel (A)					2,2			4,1	6,3

On en déduit S = 6 + 8 + 9 + 14 + 17 + 18 = 72.

2.10. Test n°10 : Méthode de Bootstrap ou strapboot.

Son objectif est de déterminer un intervalle de confiance d'un coefficient (moyenne, proportion, RR, OR, différence, etc…) pour regarder si la valeur de comparaison est contenu dans ce dernier.

Méthodologie :