2) Test d'hypothèses avec deux échantillons

Les exemples précédents ont concerné l'utilisation d'un échantillon pour obtenir une inférence à propos d'une population. Dans la réalité, il y a des situations dans laquelle il est nécessaire de comparer deux échantillons issues de deux populations afin de conduire des inférences à propos de ces populations.

2.1 Inférence sur deux moyennes : échantillons indépendants

Définition : deux échantillons sont indépendants si les valeurs d'une population ne sont pas liées aux valeurs de l'autre population.

Conditions d'application du test

1. les deux échantillons sont indépendants
2. les deux échantillons sont aléatoire simples
3. une ou les des deux conditions sont satisfaites ; les deux échantillons sont grands (n_1 >30, n_2>30) ou les deux échantillons sont issus de populations possédant des distributions normales.

Cas de variances connues

Statistique de test

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

valeurs critiques à partir de la table de loi normale

Remarque σ_1^2 et σ_2^2 sont rarement connus en réalité.

Cas de variances inconnues

Statistique de test

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Valeurs critiques

Les valeurs critiques sont obtenues à partir de la table de Student à

ddl=min(n_1-1, n_2-1)

Exemple lors d'une expérience à tester l'efficacité de la paroxétine pour traiter la maladie bipolaire, des mesures ont été réalisées sur des sujets en utilisant l'échelle de dépression de Hamilton avec les résultats donnés ci-dessous.

Utiliser un niveau de significativité de 0.05 pour tester l'affirmation que le groupe traité et le groupe placebo viennent d'une population avec la même moyenne. Interpréter le résultat.

Groupe placebo n=43, $\bar{x}$=21.57 s=3.87

Groupe traité n=33, $\bar{x}$ = 20.38 s=3.91

Solution on vérifie les conditions d'application du test ;

1. les deux échantillons sont indépendants et issus d'un tirage aléatoire simple
2. les échantillons sont de taille supérieure à 30(de grandes taille)

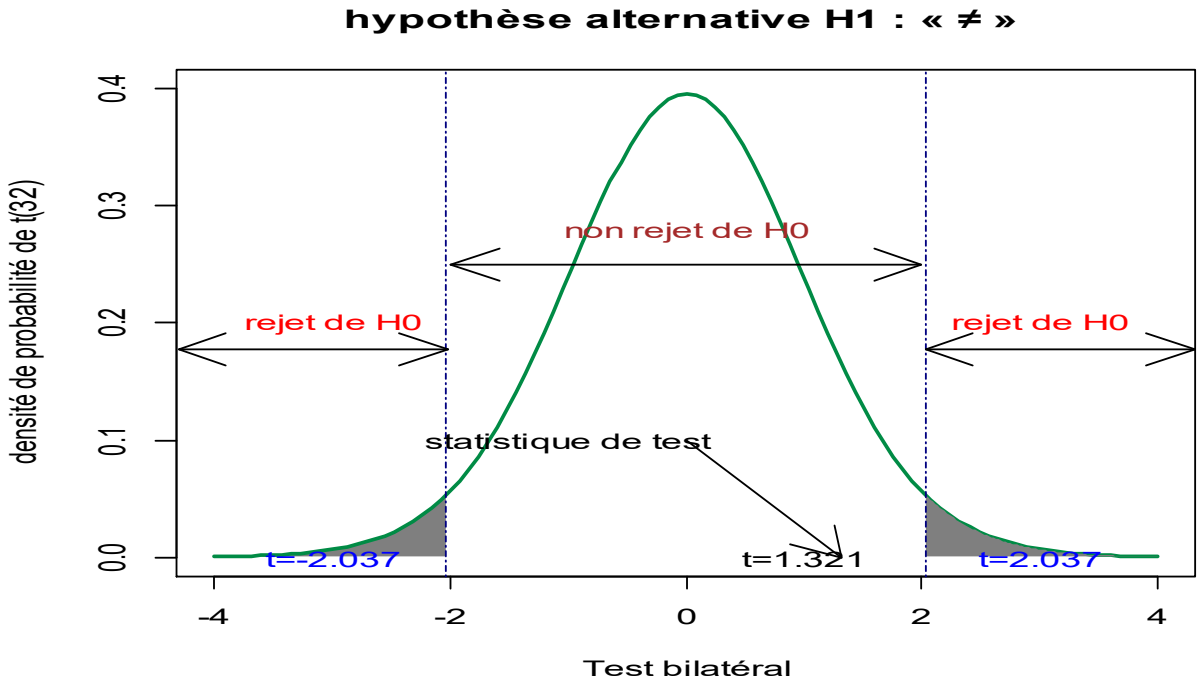
H_0 : $\mu_1 = \mu_2$

H_1 : $\mu_1 \neq \mu_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(21.57 - 20.38)}{\sqrt{\frac{3.87^2}{43} + \frac{3.91^2}{33}}} = 1.321$$

valeurs critiques

nous utilisons la loi de Student $t_{\frac{\alpha}{2},33\text{-}1}$=2.037



Conclusion la statistique de test ne se trouve pas dans la région critique, nous ne pouvons pas rejeter l'hypothèse nulle $\mu_1 = \mu_2$

Interprétation il n'y a pas suffisamment de preuves pour garantir le rejet de l'affirmation que les patients ayant reçu un placebo et ceux traités par la paroxétine ont la même moyenne. Comme les moyennes ne sont pas significativement différentes le traitement ne semble pas avoir d'effet significatif et cette substance n'est pas un bon traitement pour la maladie bipolaire

Cas de variances égales (inconnues)

Même quand les valeurs spécifiques des écart types ne sont pas connus, s'il est possible de considérer qu'ils ont la même valeur, on peut avoir une estimation de la variance commune

Conditions d'application

- les deux populations ont le même écart type
- les deux échantillons sont indépendants.
- Les deux échantillons sont aléatoires simples

- Une des deux conditions suivantes sont satisfaites ; les deux échantillons sont tous les deux grands ou les deux viennent de populations dont la distribution est normale.

Test d'hypothèse : échantillons indépendants et $\sigma_1 = \sigma_2$

Statistique de test

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}}$$

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

ddl= $(n_1 - 1) + (n_2 - 1)$

Remarque si on veut utiliser cette méthode, comment déterminer si les deux ecart types sont égaux? Une approche est utilisée est celle de test de comparaison de deux variances qui sera traitée vers la fin de ce chapitre.

3) Inférences à partir de données appariés Avec les données appariées, il existe une relation telle que chaque valeur d'un échantillon correspond à une valeur de l'autre échantillon.

Exemples

- lors d'une expérience de test d'efficacité d'un régime pauvre en matière grasse, le poids de chaque sujet est mesuré avant et après le régime
- Dans le test des effets d'un engrais sur la hauteur d'arbre, les arbres de l'échantillon sont plantées par paires, un arbre recevant le traitement et l'autre pas.

Conditions d'application

- les données sont des données appariées
- les échantillons sont aléatoires simples.
- Une ou les deux conditions sont satisfaites ; le nombre de paires est grand ou les paires de valeurs proviennent de populations dont la distribution est approximativement normale

Notations

d : différence individuelle entre les deux valeurs d'une paire

μ_d : valeur moyenne des différence d pour la population de toutes les paires

$\bar{d}$: valeur moyenne des différences

s_d : ecart type des différences d pour les données appariées de l'échantillon.

n : nombre de paires.

Statistique de test

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

valeurs critiques à partir de la table de Student et ddl= n-1.

Exemple « efficacité de l'hypnose pour réduire la douleur»

une étude a cherché à mesurer l'efficacité de l'hypnose pour réduire la douleur. Les résultats pour les sujets aléatoire sont données dans le tableau ci-dessous. Les valeurs concernent des mesures avant et après hypnose
L'hypnose semble t-elle être un bon traitement pour réduire la douleur ?

Sujet	A	B	C	D	E	F	G	H
Avant	6.6	6.5	9	10.3	11.3	8.1	6.3	11.6
Après	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Solution les données sont liées par paires car se sont les mesures prises sur les mêmes individus (avant et après l'hypnoses) et on suppose que les échantillons sont issus de populations distribuées normalement)

$H_0 : \mu_d = 0$

$H_1 : \mu_d > 0$

On introduit une nouvelle variable des différences notée d

Les valeurs de d sont

-0.2 4.1 1.6 1.8 3.2 8.0 2.9 9.6

On calcule

$$\bar{d} = 3.875$$

$s_d = 3.324$

statistique de test

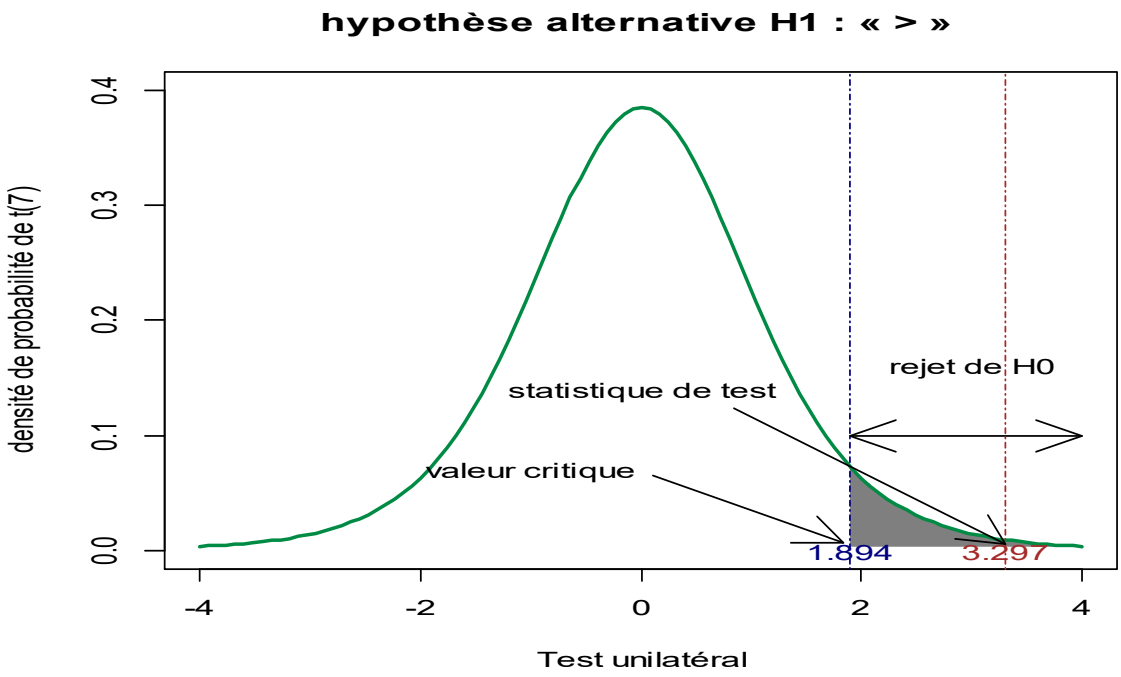
$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{3.875}{\frac{3.324}{\sqrt{8}}} = 3.297$$

valeurs critique à partir de la table de student ddl=8-1=7 et la colonne 2*0.05=0.1

$t_{\alpha,7}$= 1.894

t > $t_{\alpha,7}$= 1.894 donc on rejette H_0

Interprétation : il y a suffisamment de preuves pour confirmer que les mesures de douleur sont plus basse après hypnose. L'hypnose semble être un bon traitement pour réduire la douleur.



4) Inférence sur deux proportions dans les médias ainsi que dans la littérature scientifique on est confronté à des comparaisons des proportions de deux

populations. Les méthodes présentées dans cette section traite ce genre de problème sous certaines conditions.

Conditions requises

- Les proportions sont issues de deux échantillons aléatoires simples indépendants.
- $n_1p_1 > 5$, $n_1(1- p_1)$ et $n_2p_2 > 5$, $n_2(1- p_2) > 5$.

Notations

n_1= taille d'échantillon 1

$\hat{p}_1=\frac{x_1}{n}$ (proportion d'échantillon 1)

p_1= proportion de la population 1

On attribue un sens similaire à n_2, $\hat{p}_2$ et p_2 qui proviennent de la population 2.

Test d'hypothèses

Nous testons l'affirmation $p_1 = p_2$ et nous utiliserons l'estimation pondérée de p_1 et p_2 notée $\bar{p}$

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Statistique de test

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

valeurs critiques à partir de α et la table de la loi normale.

Exemple : « test de l'efficacité d'un vaccin »

Un article d'une revue américaine rapportait les résultats expérimentaux relatifs à un vaccin administré à des enfants. Sur 1070 enfants ayant reçu la vaccin, 14 ont développé la grippe , sur les 523 enfants qui ont reçu un placebo, 95 ont développé la grippe .

Utiliser un seuil de significativité de 0.05 pour tester l'affirmation que la proportion d'enfants vaccinés qui développent la grippe est inférieure à celle des enfants qui ont reçu un placebo.

Enfants vaccinés (échantillon1)	Enfants non vaccinés (échantillon2)
n_1= 1070 $\hat{p}_1=\frac{14}{1070}$ = 0.0131	n_2 =532 $\hat{p}_2=\frac{95}{532}$ = 0.1786

Solution

On vérifie que les conditions requises sont satisfaites

- les deux échantillons sont aléatoires simples et
- 14 > 5, 1056 > 5 et 95 > 5, 437 > 5

L'affirmation d'un taux de grippe plus faible chez les enfants vaccinés peut être représentée par $p_1 < p_2$

Nous testons les hypothèses

$H_0 : p_1 = p_2$

$H_1 : p_1 < p_2$

$$\bar{p} = \frac{14 + 95}{1070 + 532} = 0.068$$

Statistique de test

$$z = \frac{\frac{14}{1070} - \frac{95}{532}}{\sqrt{0.068(1 - 0.068)(\frac{1}{1070} + \frac{1}{532})}} = -12.39$$

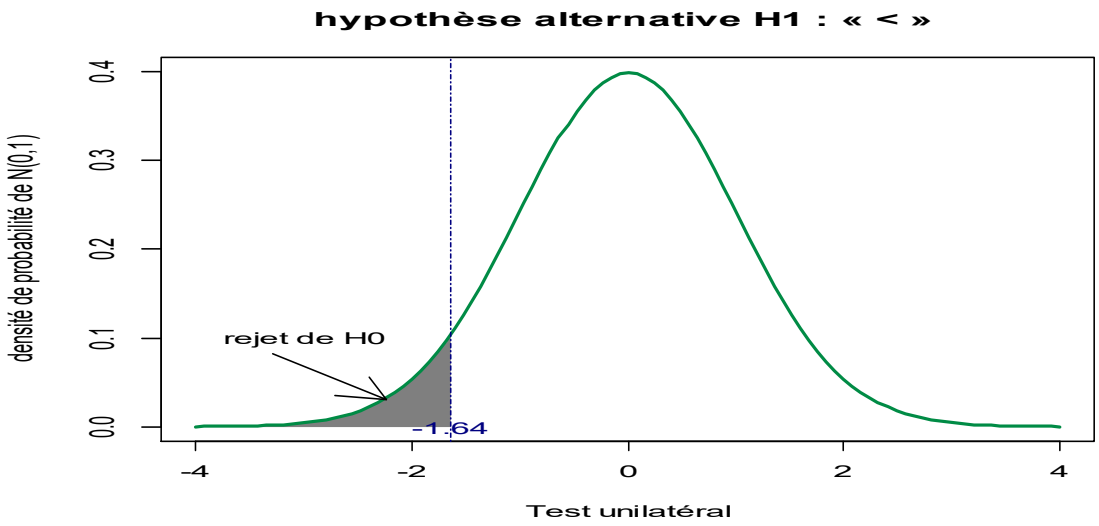
la valeur critique à partir de la table de la loi normale est

$-t_\alpha$ = -1.65

z =-12.39 < - 1.65 nous rejetons H_0.

Interprétation

Nous devons considérer l'affirmation originale que les enfants qui ont reçu le vaccin ont développé la grippe avec un taux inférieur à celui de ceux qui ont reçu un placebo.



Comparaison de la dispersion de deux échantillons cette section présente une méthode qui permet de comparer les variances de deux populations. Les calculs seront simplifiés si nous considérons les deux échantillons de telle façon que s_1^2 est la plus grande des deux variances.

Conditions requises

- Les deux populations sont indépendantes
- Les deux populations ont une distribution normale. (une condition très importante)

Notations

S_1^2 la plus grande des variances des deux échantillons.

n_1 taille de l'échantillon de la plus grande variance

σ_1^2 la variance théorique de la population de laquelle est issu l'échantillon à plus grande variance.

Les symboles S_2^2, n_2 et σ_2^2 sont utilisé pour l'autre échantillon et population.

Statistique de test

$$F = \frac{S_1^2}{S_2^2} \qquad (> 1)$$

Valeurs critiques à partir de α et la table Fisher (n_1-1, n_2-1)

Exemple : « Calcium et pression sanguine »

Des données ont été collectées au cours d'une étude sur les suppléments calciques et leurs effets sur la pression sanguine. Un groupe placebo et un groupe calcium ont commencé l'étude par une mesure de pression sanguine.

On a obtenu les résultats suivants

	Effectif	Ecart type
Placebo	n=13	s_1= 9.46
Calcium	n=15	s_2 = 8.469

A un niveau de significativité de 0.05, tester l'affirmation que les deux échantillons sont issus de populations de mêmes écart-type.

Solution

Nous vérifions si les conditions sont satisfaites ;

Les deux échantillons sont indépendants.

Les échantillons viennent de populations normales.

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

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

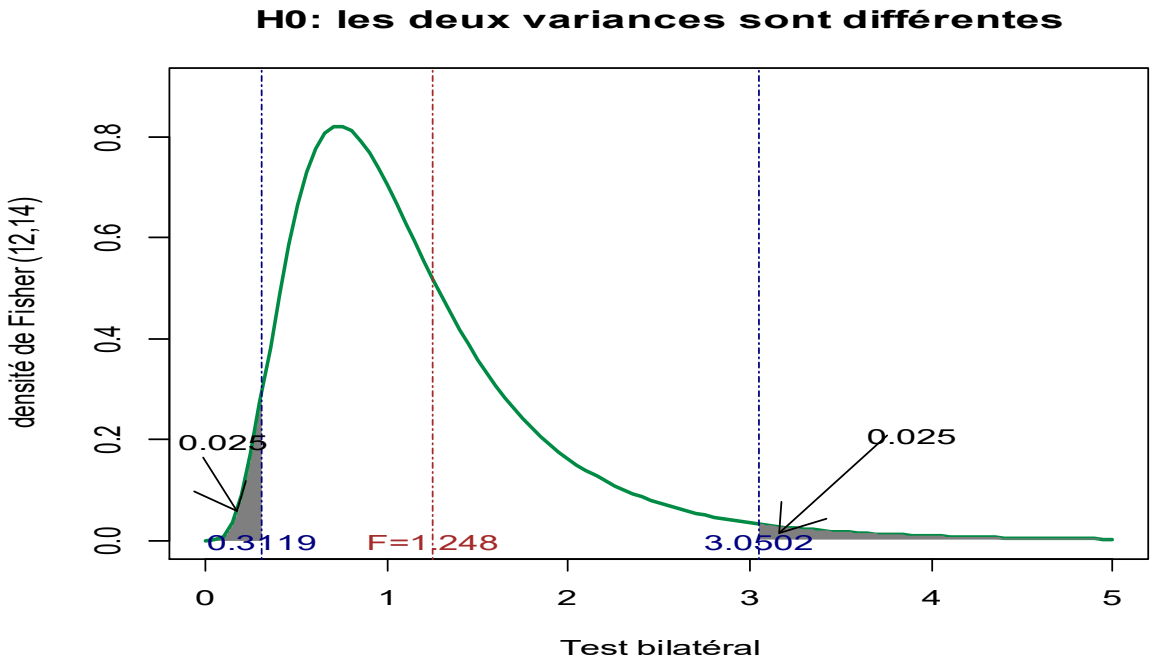
Statistique de test

$$F = \frac{9.468^2}{8.469^2} = 1.248$$

Valeurs critiques il s'agit d'un test bilatéral avec une aire de 0.025 (0.05/2), on compare F a la valeur critique qui se trouve à droite qui correspond à 3.0502

(table de Fisher avec α = 0.025, ddl1=12 , ddl2=14, F(12, 14))

Conclusion F < 3.0502 , F=1.248 ne se situe pas dans la région critique. Ainsi nous ne pouvons pas rejeter H_0.



Interprétation : il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle d'égalité des variances.

3) Test d'adéquation

Dans cette section, nous présentons une méthode pour tester l'hypothèse que les fréquences observées pour les différentes catégories (classes) sont en adéquation avec une distribution donnée.

Notations

no_i représente les fréquences observés de résultats.

nt_i représente les fréquences attendu de résultats.

k représente le nombre de classes.

n représente le nombre total d'essais.

Conditions d'application

- Les données sélectionnées aléatoirement
- Pour chaque catégorie, la fréquence attendu est au moins 5 ($nt_i \leq 5$).

Les hypothèses à tester

H_0 : les observations suivent la distribution p_i.

H_1 : les observations ne suivent pas la distribution p_i.

Statistique de test

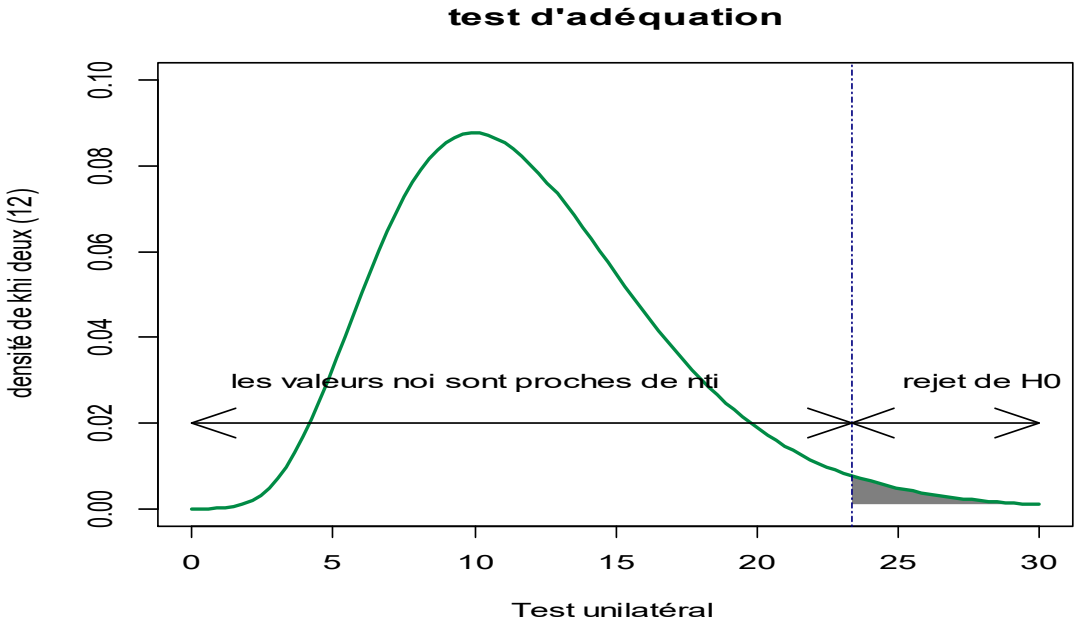
$$\chi_c^2 = \sum \frac{(noi - nti)^2}{nti}$$

Valeurs critiques

Les valeurs critiques sont les dans a table de khi deux avec ddl=k-1

Les tests d'hypothèses d'adéquation sont toujours unilatéraux à droite.

Représentation graphique



Interprétation

Le non rejet de H_0 signifie que les valeurs noi et nti sont proches.

Le rejet de H_0 signifie que les valeurs noi et nti sont éloignées.

Exemple

A partir du génotype des parents, on s'attend à ce que les enfants aient des génotypes répartis comme suit : 25% de génotype AA, 50% de génotype Aa et 25% de génotype aa. Pour une maladie particulière, AA représente un enfant sain, Aa un enfant porteur et aa un enfant malade.

Le tableau suivant donne les fréquences des génotypes pour 90 malades sélectionnés aléatoirement.

Génotype	AA	Aa	aa
Fréquences observées (noi)	22	55	13

Tester au niveau de significativité $\alpha = 0.01$ l'hypothèse que ces fréquences observées correspondent aux fréquences attendues.

Solution

Les données ont été sélectionnées aléatoirement.

Il reste à vérifier que les fréquences attendues sont toutes d'au moins 5 et cela en calculant les nti.

H_0 : la distribution des génotypes des enfants est p1=0.25(AA), p2=0.5(Aa), p3= 0.25 (aa).

H_1 : au moins une des proportions ci-dessus est différente des valeurs supposées.

Calcul des nti.

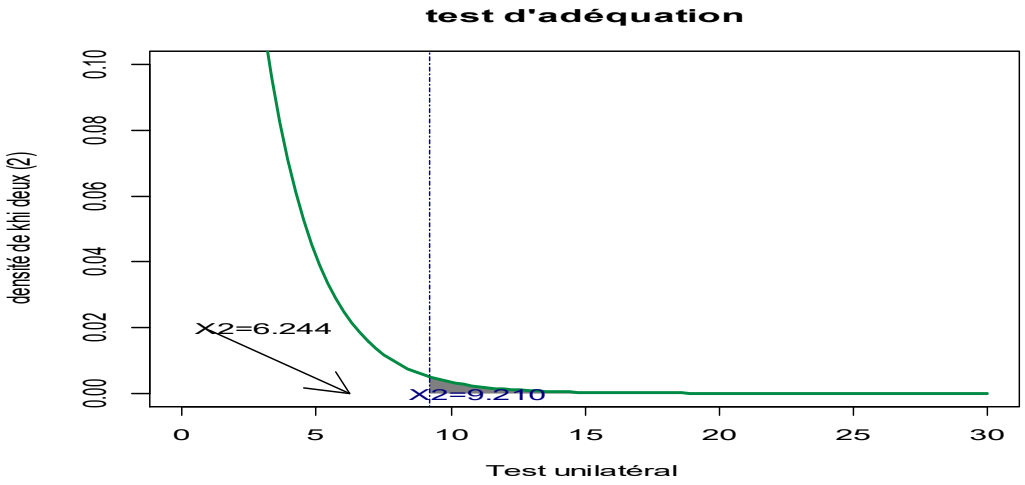
Génotype	AA	Aa	aa
Fréquences observées (noi)	22	55	13
nti = n*pi	90*0.25=22.5	90*0.5=45	90*0.25=22.5
noi-nti	-0.5	10	-9.5
(noi-nti) ^{2}	0.25	100	90.25
$\frac{(noi-nti)^2}{nti}$	0.0111	2.2222	4.0111

X^2_c = 0.111+0.2222+4.0111= 6.2444.

La valeur critique est lu dans la table de khi deux ddl=3-1=2 et α=0.01 on a alors 9.210.

Décision : X^2_c ne tombe pas dans la région critique, il n'y a pas suffisamment de preuves pour garantir le rejet de H_0.

On ne peut pas rejeter l'hypothèse que la distribution est p1=0.25, p2=0.5, p3=0.25



5) Test d'indépendance

Tableaux de contingence : indépendance et homogénéité

un tableau de contingence est un tableau dans lequel les fréquences correspondent à deux variables : une variable est utilisée en ligne et l'autre en colonne.

Un test d'indépendance teste l'hypothèse nulle qu'il n'y a pas d'association entre la variable en ligne et celle en colonne dans un tableau de contingence.

Conditions requises.

- Les données d'échantillon ont été sélectionnées aléatoirement.
- H_0 est l'affirmation que les variables de ligne et de colonne sont indépendantes, H_1 est l'affirmation que les variables de ligne et de colonne sont dépendantes.

- Pour chaque case du tableau de contingence, la fréquence attendu nt_{ij} est au moins 5.

Statistique de test

$$\chi_c^2 = \sum \frac{(no_{ij} - nt_{ij})^2}{nt_{ij}}$$

Valeurs critiques

la valeur est lue dans la table de khi deux et le degrés de liberté ddl= (r-1)(k-1)

où r est le nombre de lignes et k le nombre de colonnes.

Dans un test d'indépendance la région critique est située à droite de la valeur critique.

La fréquence attendue pour un tableau de contingence

$$\mathrm{nt_{ij}} = \frac{(some\ de\ la\ ligne\ i)(somme\ de\ la\ colonne\ j)}{n},$$

fréquence attendue pour la case ij

Exemple : « test de l'efficacité de vaccin de Salk »

Dans une expérience, on a donné à des enfants le vaccin de Salk contre la polyométrie et à d'autres enfants un placebo. Les résultats de l'expérience sont résumés dans le tableau suivant

	Oui	non
Groupe traitement vaccin	33	200712
Groupe placebo	115	201114

Utiliser un niveau de significativité de 0.05 pour tester l'hypothèse d'indépendance entre les groupes et le traitement.

Solution

Les données sont des comptages de fréquences indépendants.

Il reste à vérifier la 2ème condition ; $nt_{ij} \geq 5$.

On teste les hypothèses suivantes :

H_0 : recevoir le vaccin Salk est indépendant d'avoir la polio

H_1 : recevoir le vaccin Salk et avoir la polio sont dépendants.

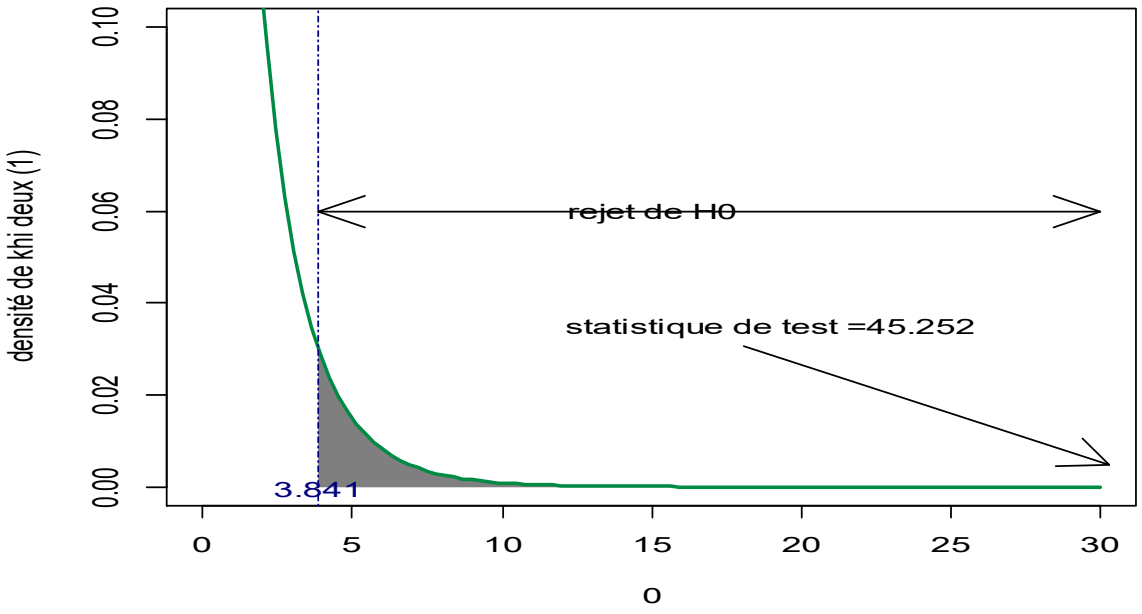
Calcul des ntij

	Oui	non
Groupe traitement vaccin	33 (73.911)	200712(200671.089)
Groupe placebo	115(74.089)	201114 (201154.911)

$$\chi_c^2 = \frac{(33-73.911)^2}{73.911} + \frac{(200712-200671.089)^2}{200671.089} + \frac{(115-74.089)^2}{74.089} + \frac{(201114-201154.911)^2}{201154.911}$$

$$= 22.645 + 0.008 + 22.591 + 0.008 = 45.252$$

La valeur critique est χ^2 = 3.841 trouvée dans la table de khi deux (ligne ddl=(r-1)(k-1)=(2-1)(2-1)=1 et colonne 0.05)



Interprétation comme la statistique de test tombe dans la région critique, on rejette H_0 qui est « recevoir le vaccin Salk est indépendant d'avoir la polio ». il apparaît que les variables sont dépendantes.