Tests du χ^2 d'indépendance et d'homogénéité

Cours de Tests paramétriques

4 novembre 2024

4.1 Introduction

Tests du χ^2 :

Tests paramétriques basés sur une statistique de test suivant approximativement une loi du χ^2 sous l'hypothèse nulle.

Objectifs :

- Tests d'indépendance
- Tests d'homogénéité

4.2 Test du χ^2 d'indépendance - Variables observées

- X : variable aléatoire qualitative ou quantitative discrète à K modalités, notées $a_1, \ldots, a_K$.
- Y : variable aléatoire qualitative ou quantitative discrète à L modalités, notées $b_1, \ldots, b_L$.
- n données : $(x_1, y_1), \ldots, (x_n, y_n)$ réalisations de n couples de variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendantes et de même loi que le couple (X, Y).

Objectif du test

On veut tester l'hypothèse

$$(H_0) : X \text{ et } Y \text{ sont indépendantes}$$

contre

$$(H_1) : X \text{ et } Y \text{ ne sont pas indépendantes}$$

Exemple 1

On souhaite savoir si le temps écoulé depuis la vaccination contre une maladie donnée a ou non une influence sur le degré de gravité de la maladie lorsque celle-ci se déclare.

- Gravité de la maladie : légère (L), moyenne (M) ou grave (G).
- Durée écoulée depuis vaccination : moins de 10 ans (A), entre 10 et 25 ans (B), plus de 25 ans (C).
- 1 574 malades.

	A	B	C	Total
G	1	42	230	273
M	6	114	347	467
L	23	301	510	834
Total	30	457	1087	1574

Exemple 1 (suite)

D'un point de vue descriptif on peut étudier la distribution **conditionnelle** de la gravité de la maladie **conditionnellement** à la durée écoulée depuis vaccination :

	A	B	C
G	0.03	0.09	0.21
M	0.20	0.25	0.32
L	0.77	0.66	0.47

Qu'en pensez-vous ?

Principe du test d'indépendance

Justification heuristique du test.

La loi du couple de variables (X, Y) est caractérisée par $P[X = a_k, Y = b_l]$, pour $k = 1, \ldots, K$, $l = 1, \ldots, L$

Réécriture mathématique des hypothèses H_0 et H_1 :

$$(H_0)\ \forall k = 1, \ldots, K, \forall l = 1, \ldots, L,$$
$$P[X = a_k, Y = b_l] =$$

$$(H_1)\ \exists k \in \{1, \ldots, K\}, \exists l \in \{1, \ldots, L\}\ :$$
$$P[X = a_k, Y = b_l] \neq$$

Principe du test d'indépendance

On introduit, pour $1 \leq k \leq K$ et $1 \leq l \leq L$, les variables aléatoires :

- N_{kl}, nombre de couples de variables (X_i, Y_i), pour $1 \leq i \leq n$, tels que $X_i = a_k$ ET $Y_i = b_l$.
- $N_{k\bullet} = \sum_{l=1}^{L} N_{kl}$, nombre de variables X_i, $1 \leq i \leq n$, qui prennent la valeur a_k.
- $N_{\bullet l} = \sum_{k=1}^{K} N_{kl}$, nombre de variables Y_i, pour $1 \leq i \leq n$, qui prennent la valeur b_l.

Principe du test d'indépendance

Etant donnée une réalisation $(x_1, y_1), \ldots, (x_n, y_n)$ de $(X_1, Y_1), \ldots, (X_n, Y_n)$, on note respectivement $n_{kl}, n_{k\bullet}$ et $n_{\bullet l}$ les réalisations correspondantes de $N_{kl}, N_{k\bullet}$ et $N_{\bullet l}$, qui peuvent être représentées dans le **tableau de contingence** ci-dessous.

$X \setminus Y$	b_1	$\ldots$	b_l	$\ldots$	b_L	Total
a_1	n_{11}	$\ldots$	n_{1l}	$\ldots$	n_{1L}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	n_{k1}	$\ldots$	n_{kl}	$\ldots$	n_{kL}	$n_{k\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_K	n_{K1}	$\ldots$	n_{Kl}	$\ldots$	n_{KL}	$n_{K\bullet}$
Total	$n_{\bullet 1}$	$\ldots$	$n_{\bullet l}$	$\ldots$	$n_{\bullet L}$	n

Principe du test d'indépendance

On estime alors, pour $1 \leq k \leq K$ et $1 \leq l \leq L$,

- $P(X = a_k \text{ et } Y = b_l)$ par
- $P(X = a_k) \times P(Y = b_l)$ par

Sous (H_0), pour tous $1 \leq k \leq K$, $1 \leq l \leq L$, l'écart entre **fréquence observée** et **fréquence théorique sous** (H_0) est censé être proche de 0, ou encore l'écart entre **effectif observé** et **effectif théorique sous** (H_0) est censé être proche de 0.

Principe du test d'indépendance

Statistique de test

$$T_n = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(N_{kl} - \frac{N_{k\bullet} N_{\bullet l}}{n}\right)^2}{\frac{N_{k\bullet} N_{\bullet l}}{n}}$$

Principe du test d'indépendance

Proposition 1

Si les conditions suivantes sont satisfaites

- le nombre d'observations n est « grand »,
- $n_{k\bullet} n_{\bullet l}/n \geq 5$ pour tous $k = 1, \ldots, K$ et $l = 1, \ldots, L$,

alors sous (H_0),

$$T_n \text{ suit approximativement la loi } \chi^2\left((K-1)(L-1)\right)$$

Principe du test d'indépendance

Zone de rejet au niveau α

$$R_{n,\alpha} = \{T_n \geq c_\alpha\},$$

où c_α est le quantile d'ordre $1-\alpha$ d'une loi $\chi^2\left((K-1)(L-1)\right)$.

Règle de décision :

- si , alors on rejette l'hypothèse d'indépendance entre X et Y.
- si , alors on ne rejette pas l'hypothèse d'indépendance entre X et Y.

Retour à l'exemple 1

On souhaite savoir si le temps écoulé depuis la vaccination contre une maladie donnée a ou non une influence sur le degré de gravité de la maladie lorsque celle-ci se déclare.

- Gravité de la maladie : légère (L), moyenne (M) ou grave (G).
- Durée écoulée depuis vaccination : moins de 10 ans (A), entre 10 et 25 ans (B), plus de 25 ans (C).
- 1 574 malades.

	A	B	C	Total
G	1	42	230	273
M	6	114	347	467
L	23	301	510	834
Total	30	457	1087	1574

Retour à l'exemple 1

On introduit :

- X la variable gravité de la maladie. C'est une variable qualitative à trois modalités $a_1 =$"Grave", $a_2 =$"Moyenne", $a_3 =$"Légère" ($K = 3$),
- Y la variable durée écoulée depuis vaccination. C'est une variable qualitative à trois modalités $b_1 =$"Moins de 10 ans", $b_2 =$"Entre 10 et 25 ans", $b_3 =$"Plus de 25 ans" ($L = 3$).

On veut alors tester

$$(H_0) : X \text{ et } Y \text{ sont indépendantes}$$

contre

$$(H_1) : X \text{ et } Y \text{ ne sont pas indépendantes.}$$

On dispose d'un échantillon de n couples de variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendantes et de même loi que le couple (X, Y). $n = 1\,574$.

Retour à l'exemple 1

On introduit, pour $k = 1, 2, 3$ et $l = 1, 2, 3$, les variables aléatoires :

- N_{kl}, nombre de couples de variables (X_i, Y_i), pour $1 \leq i \leq n$, tels que
- $N_{k\bullet} = \sum_{l=1}^{L} N_{kl}$, nombre de variables X_i, $1 \leq i \leq n$, qui prennent la valeur.
- $N_{\bullet l} = \sum_{k=1}^{K} N_{kl}$, nombre de variables Y_i, pour $1 \leq i \leq n$, qui prennent la valeur

La statistique de test est :

$$T_n = \sum_{k=1}^{3} \sum_{l=1}^{3} \frac{\left(N_{kl} - \frac{N_{k\bullet} N_{\bullet l}}{n}\right)^2}{\frac{N_{k\bullet} N_{\bullet l}}{n}}.$$

Retour à l'exemple 1

On calcule les effectifs théoriques $n_{k\bullet}n_{\bullet l}/n \geq 5$ pour tous $k = 1, \ldots, K$ et $l = 1, \ldots, L$:

	A	B	C
G	$273 \times 30/1574 \approx \mathbf{5.20}$	$273 \times 457/1574 \approx \mathbf{79.26}$	$273 \times 1087/1574 \approx \mathbf{188.53}$
M	$467 \times 30/1574 \approx \mathbf{8.90}$	$467 \times 457/1574 \approx \mathbf{135.59}$	$467 \times 1087/1574 \approx \mathbf{322.51}$
L	$834 \times 30/1574 \approx \mathbf{15.90}$	$834 \times 457/1574 \approx \mathbf{242.15}$	$834 \times 1087/1574 \approx \mathbf{575.96}$

Ils sont bien tous plus grand que 5. De plus la taille d'échantillon est très grande ($n = 1\,574$), on en déduit donc que T_n suit approximativement une loi sous H_0.

Retour à l'exemple 1

- La zone de rejet pour un test de niveau 5% est

$$R_{n,0.05} = \{T_n \geq c_{0.05}\},$$

où $c_{0.05}$ est le quantile d'ordre $1 - 0.05 = 0.95$ d'une loi $\chi^2(4)$. On trouve $c_{0.05} = 9.49$.

- Calcul de t_n :

$$t_n = \frac{(1-5.20)^2}{5.20} + \frac{(42-79.26)^2}{79.26} + \frac{(230-188.53)^2}{188.53} + \frac{(6-8.90)^2}{8.90} + \cdots + \frac{(301-242.15)^2}{242.15} + \frac{(510-575.96)^2}{575.96}.$$

$t_n \approx 61.31$

- **Conclusion** : on rejette H_0 au risque 5%.

Retour à l'exemple 1

Calcul de la p-valeur :

$$\text{p-valeur} = P_{H_0}[T_n > t_n] \approx P[\chi^2(4) > t_n] = 1 - P[\chi^2(4) \leq 61.31] \approx 0.$$

On est sûr, avec un risque quasiment nul de se tromper, que la gravité de la maladie est associée à la durée écoulée depuis vaccination.

Remarques sur le test du χ^2

- Les conditions $n_{k\bullet} n_{\bullet l}/n \geq 5$ pour tous $k = 1, \ldots, K$ et $l = 1, \ldots, L$, sont **primordiales** et doivent toujours être vérifiées.
- La condition n "grand" est subjective mais est aussi importante car il s'agit d'un test **asymptotique**. Personnellement je considère n "grand" pour $n \geq 50$.
- Si les conditions d'application du test du χ^2 ne sont pas vérifiées, on pourra
  - soit regrouper des classes entre elles. Par exemple, on pourra regrouper les classes (B) et (C) (entre 10 ans et 25 ans et plus de 25 ans) en la classe plus de 10 ans dans l'exemple sur la vaccination.
  - soit opter pour le **test exact de Fisher** (voir slides suivants). Ce test est **exact** et n'a aucune condition d'application. Il peut cependant être long à implémenter en pratique si la taille d'échantillon n est grande.

Remarques sur le test du χ^2

- Le test du χ^2 et le test de proportion à deux échantillons (voir chapitre 3) sont **identiques** quand X et Y n'ont que deux modalités chacune.
  - La statistique de test du test de proportion suit (asymptotiquement) une loi normale centrée réduite. Si on l'élève au carré on obtient un χ^2 à un degrés de liberté. Cette statistique de test au carré **a la même valeur** que la statistique de test pour le test du χ^2 et suit bien la même loi. La p-valeur est donc **identique** pour les deux tests.
  - Les conditions d'application du test de proportion (en particulier $n_1 p_n \geq 5, n_1(1-p_n) \geq 5$ et $n_2 p_n \geq 5, n_2(1-p_n) \geq 5$, voir chapitre 3) sont **identiques** à celles du χ^2.
  - En pratique, pour deux variables X et Y à deux modalités chacune, il n'y a donc aucune différence entre utiliser le test de proportion ou le test du χ^2.
- Le test du χ^2 est donc une **généralisation** du test de proportion au cas où l'une des deux variables (X ou Y) a plus de deux modalités.

Le test exact de Fisher

- Ce test s'utilise exactement dans le même contexte que pour le test du χ^2 d'indépendance (pour deux variables qualitatives ou quantitatives discrètes, avec les mêmes hypothèses (H_0) et (H_1)).
- C'est un test **exact**, il n'a aucune condition d'application. L'idée du test est de calculer la probabilité sous (H_0) d'obtenir un tableau de contingence tel que celui observé quand les marges sont fixées. Il se base sur la loi hypergéométrique. Pour calculer la statistique de test, il faut lister tous les tableaux de contingence possibles qui ont les mêmes marges, ce qui peut prendre du temps (même sous logiciel).
- Il peut toujours s'utiliser à condition que le logiciel arrive à faire tous les calculs.
- Il est très utile quand les conditions d'application du test du χ^2 ne sont pas respectées.

Implémentation sous R

- Le test du χ^2 s'implémente avec la commande : `chisq.test`
  Il faut vérifier à la main que les conditions d'application du test sont bien vérifiées (en particulier que les effectifs théoriques sont tous bien supérieurs à 5). Si ce n'est pas le cas, R implémente quand même le test mais renvoie un message d'erreur.
- Le test exact de Fisher s'implémente avec la commande : `fisher.test`
- Les deux tests prennent en entrée un tableau de contingence qui se construit, par exemple, en utilisant la fonction `matrix`

Implémentation sous R - exemple 1

- Création du tableau de contingence :

  > mat=matrix(c(1,6,23,42,114,301,230,347,510),ncol=3)

- Test du χ^2 :

  > chisq.test(mat,correct = FALSE)

  Pearson's Chi-squared test

  data : mat
  X-squared = 61.311, df = 4, p-value = 1.538e-12

- Test exact de Fisher :

  > fisher.test(mat)

  Fisher's Exact Test for Count Data

  data : mat
  p-value = 2.676e-13
  alternative hypothesis : two.sided

4.3 Test du χ^2 d'homogénéité - Lien avec le test du χ^2 d'indépendance

Le test que nous voyons dans cette section, le test du χ^2 d'homogénéité, est en fait **identique** au test du χ^2 d'indépendance. Il n'est donc pas nécessaire de l'apprendre, il suffit juste de comprendre le lien entre les deux tests.

- Pour le test du χ^2 d'homogénéité il n'y a qu'**une seule variable** et le but est de tester si la loi de cette variable est la même dans différentes populations.
- L'hypothèse nulle consiste à dire que la loi de cette variable est la même dans toutes les différentes populations.
- L'hypothèse alternative consiste à dire que la loi de cette variable n'est pas la même dans toutes les différentes populations.

Si l'on reprend l'exemple 1, on peut se poser la question (test d'homogénéité) si la loi de la gravité de la maladie est la même dans les 3 populations constituées par la durée écoulée depuis vaccination.

- Si la gravité de la maladie est distribuée pareille selon chaque population, cela revient à dire que la durée écoulée depuis vaccination n'a pas d'effet sur la gravité de la maladie. Et donc, il y a indépendance entre gravité de la maladie et durée écoulée depuis vaccination.
- Si la gravité de la maladie n'est pas distribuée pareille selon chaque population (par exemple il y a plus de chances d'avoir la gravité (G) dans la population (A) que dans les autres populations), cela revient à dire que la durée écoulée depuis vaccination est liée à la gravité de la maladie. Selon la durée écoulée depuis vaccination, la gravité de la maladie ne sera pas la même.

Faire un test d'homogénéité ou d'indépendance répond donc exactement à la même question. En pratique, on pourra toujours reformuler un test d'homogénéité en test d'indépendance et réciproquement.

Variables observées

- X : variable aléatoire qualitative ou quantitative discrète à K modalités, notées $a_1, \ldots, a_K$.
- Comparaison de la distribution de X dans L populations différentes.
- Pour chaque $1 \leq l \leq L$, on dispose d'un échantillon de n_l données $x_{1l}, \ldots, x_{n_l l}$ réalisations de n_l variables $X_{1l}, \ldots, X_{n_l l}$ indépendantes et de même loi que X_l.
- On suppose que les L échantillons $(X_{11}, \ldots, X_{n_1 1}), (X_{12}, \ldots, X_{n_2 2}), \ldots, (X_{1L}, \ldots, X_{n_L L})$ sont indépendants.

Objectif du test

On veut tester l'hypothèse

$$(H_0) : \text{Les variables } X_1, \ldots, X_L \text{ suivent toutes la même loi}$$

contre

$$(H_1) : \text{Les variables } X_1, \ldots, X_L \text{ ne suivent pas toutes la même loi}$$

Exemple 2

On a mesuré les groupes sanguins dans 2 populations de 1032 Pygmées et 484 Esquimaux. Au vu de ces résultats, peut-on dire que la distribution des groupes sanguins est la même dans les deux populations ?

Groupe sanguin\ Pop.	Pygmées	Esquimaux	
AB	103	7	
B	300	17	
A	313	260	
O	316	200	
Total	1032	484	

Exemple 2 (suite)

D'un point de vue descriptif on peut étudier la distribution **conditionnelle** du groupe sanguin **conditionnellement** au type de population (Pygmées ou Esquimaux) :

Groupe sanguin\ Pop.	Pygmées	Esquimaux
AB	0.10	0.01
B	0.29	0.04
A	0.30	0.54
O	0.31	0.41

- Qu'en pensez-vous ?
- Faire le test au risque 5%.