Chapitre IV. Tests du chi-deux

Cours de Tests paramétriques

Deuxième Année - IUT STID - Olivier Bouaziz

2018-2019

Introduction

Tests du chi-deux :

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

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 $(X_1, Y_n), \ldots, (X_n, Y_n)$ 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

.....

Réécriture mathématique des hypothèses H_0 et H_1 :

(H_0) .....

(H_1) .....

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 $t_n \geq c_\alpha$, alors on rejette l'hypothèse d'indépendance entre X et Y.
- si $t_n < c_\alpha$, 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

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) : Les variables $X_1, \ldots, X_L$ suivent toutes la même loi

contre

(H_1) : Les variables $X_1, \ldots, X_L$ 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 ?

Principe du test d'homogénéité

Justification heuristique du test.

Réécriture mathématique des hypothèses H_0 et H_1 :

(H_0) .....

(H_1) .....

Principe du test d'homogénéité

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

- N_{kl}, nombre de variables parmi $(X_{1l}, X_{2l}, \ldots, X_{n_l l})$ qui prennent la valeur a_k.
- $N_{k\bullet} = \sum_{l=1}^{L} N_{kl}$, nombre de variables X_{il}, $1 \leq i \leq L$, $1 \leq i \leq n_l$, qui prennent la valeur a_k.

Principe du test d'homogénéité

On note respectivement n_{kl} et $n_{k\bullet}$ des réalisations de N_{kl} et $N_{k\bullet}$ qui peuvent être représentées dans le **tableau de contingence** ci-dessous. On note également $n = n_1 + n_2 \ldots + n_L$.

Modalités de X\ Population	1	...	l	...	L	Total
a_1	n_{11}	...	n_{1l}	...	n_{1L}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	n_{k1}	...	n_{kl}	...	n_{kL}	$n_{k\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_K	n_{K1}	...	n_{Kl}	...	n_{KL}	$n_{K\bullet}$
Total	n_1	...	n_l	...	n_L	n

Principe du test d'homogénéité

Sous (H_0), pour $1 \leq k \leq K$, on peut estimer $P(X = a_k)$ par :

.....

Le test consiste alors à comparer, pour tous $1 \leq k \leq K$ et $1 \leq l \leq L$:

- l'effectif observé pour la modalité a_k dans la l^{e} population :

  .....

à

- l'effectif théorique sous (H_0) pour la modalité a_k dans la l^{e} population :

  .....

Principe du test d'homogénéité

Statistique de test

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

Principe du test d'homogénéité

Proposition 2

Si les conditions suivantes sont satisfaites

- le nombre d'observations $n = \sum_{l=1}^{L} n_l$ est « grand »,
- $n_{k\bullet} n_l / n \geq 5$ 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'homogénéité

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 $t_n \geq c_\alpha$, alors on rejette l'hypothèse d'homogénéité des L populations.
- si $t_n < c_\alpha$, alors on ne rejette pas l'hypothèse d'homogénéité des L populations.

Retour à l'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	