Université de Tlemcen
Faculté SNV &STU
bahidja_sahel @yahoo.fr

Test du khi – deux

Introduction :

Un test du khi-deux est utilisé pour :

- Un test d'adéquation ou d'ajustement :
  C'est une méthode qui permet de comparer une distribution observée sur un échantillon à une distribution théorique.
- Un test d'homogeneité :
  C'est une méthode qui permet de comparer deux ou plusieurs distributions observées.
- Un test d'indépendance :
  C'est une méthode qui permet de voir s'il n'y a pas de relation ou d'association entre deux variables

1.Test d'adéquation (ou ajustement) :

Ce test permet de juger la qualité de l'ajustement d'une distribution expérimentale à une distribution théorique en d'autres termes il permet de tester l'hypothèse que les fréquences observées pour les différentes catégories (classes) sont en adéquation avec une distribution donnée.

- Notations :

  O_i : représentent les fréquences observées des résultats
  T_i : représentent les fréquences attendues (théoriques) des résultats
  K : représente le nombre de classes
  N : représente le nombre total d'essais

- Conditions d'application du test :

  - Les données sélectionnées aléatiorement
  - Pour chaque catégorie , la fréquence attendue est supérieure ou égale à 5($Ti \geq 5$)

- Hypothèses à tester :

  H0 : les observations suivent la distribution théorique
  H1 : les observations ne suivent pas la distribution théorique

- Statistique du test :

$$X^2C = \sum_{i=1}^{K} \frac{(Oi-Ti)^2}{Ti}$$

- Valeur critique :

  Utiliser la table de la loi du khi – deux avec un d.d.l ϑ

Remarques :

1. Les fréquences attendues doivent être supérieurs ou égaux à 5 sinon on regroupe deux ou plusieurs classes (modalités)
2. Le test d'adéquation est toujours unilatéral à droite
3. Le d.d.l ϑ = K-1-R avec R est le nombre de paramétres à estimer éventuellement pour caractériser la distribution théorique

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.

Letableau suivant donne les fréquences des génotypes pour 90 malades choisis aléatoirement

génotype	AA	Aa	aa
Fréquences observées Oi	22	55	13

Tester au niveau de significativité α = 0.01 , l'hypothèse que ces fréquences observées peuvent être ajustées aux fréquences attendues de la distribution théorique

Solution :

- H0 : la distribution des génotypes des enfants est adéquate avec la distribution donnée p1=0.25,

  p2 =0.50 ,p3 = 0.25

  H1 : la distribution des génotypes des enfants n'est pas adéquate avec la distribution donnée

- Vérifions que les conditions du test sont satisafaites

  1. Les données sont sélectionnées aléatoirement
  2. Les fréquences attendues sont supérieures ou égales à 5 pour cela on doit d'abord les calculer Ti = N* pi

  Calcul des Ti et les différences entre les Oi et les TI

génotype	AA	Aa	aa
Fréq .obs .Oi	22	55	13
Fréq.the Ti	90*0.25=22.5	90*0.50=45	90*0.25=22.5
Oi-Ti	-0.50	10	-9.50
$(Oi-Ti)^2$	0.25	100	90.25
$(Oi-Ti)^2 / Ti$	0.0111	2.2222	4.0111

- Statistique du test :

$$X^2C = \sum_{i=1}^{K} \frac{(Oi-Ti)^2}{Ti} = 0.0111+2.2222+4.0111=6.2444$$

- Valeur critique :

Sur la table du khi-deux ,on lit $X^2 \alpha = X^2(0.01 ; 3-1) = 9.210$

- Décision :

comme $X^2C < X^2\alpha$ aors on accepte H0

- Conclusion :

Avec un risque de 0.01 on peut dire que la distribution observée est conforme avec la distribution donnée

2.Test d'homogeneité :

On constitue deux ou plusieurs échantillons surlesquels on a observé les distributions selon les modalités d'une variable qualitative ou quantitative.

Le test d'homogeneité permet de tester si les distributions sont identiques ou homogènes .

Notations :

L : nombre d'échantillons indépendants de tailles respectives n_1, n_2 , .., n_L prélevés de L populations

Oij : fréquence observée pour la modalité inscrite en colonne j et pour l'échantillon i (ligne)

pour le tableau de contingence

Tij : fréquence attendue pour l'échantillon i et pour la modalité inscrite en colonne j

N : nombre total de fréquences observées

Un tableau de contingence est un tableau à double entrée :

X / échan.	a_1	……	aj	……	a_c	Total en ligne
Echan.1	O_{11}	…..	O_{1j}	……	O_{1c}	O_1.
…….	…….	…..	…….	……	……	…..
Echan.i	O_{i1}	….	Oij	…..	O_{ic}	Oi.
……	……	…..	…..	…..	…….	…….
Echan.L	O_{L1}	…..	O_{Lj}	……	O_{Lc}	O_L.
Total en colonne	$O._1$	……	$O._j$	……	$O._c$	O..

La fréquence attendue (théorique) pour un tableau de contingence est :

$$Tij = \frac{(total\ de\ la\ ligne\ i)*(total\ de\ la\ colonne\ j)}{total\ général\ N}$$

- Conditions d'application du test :
  - Les données sont sélectionnées aléatoirement.
  - Pour chaque case du tableau de contingence,la fréquence attendue Tij est au moins5

- **H0 :** les distributions sont homogènes (identiques)
  H1 : les distributions ne sont pas homogènes (identiques)

- Statistique du test :

$$X^2_C = \sum_{i=1}^{L} \sum_{j=1}^{C} \frac{(Oij - Tij)^2}{Tij}$$

- Valeur critique : utiliser la table de la loi du khi-deux avec un d.d.l ϑ = (L-1)*(C-1)
  Où L est le nombre de lignes et C est le nombre de colonnes
  Dans un test d'homogeneité , la région critique est située à droite de la valeur critique.

Exemple :

Un pré-test est effectué pour évaluer la préférence d'une pâte dentifrice.200 personnes ont été choisis au hasard respectivement dans deux régions et on a remis à chaque personne deux tubes de pâte de dentifrice, l'un étant la nouvelle pâte , l'autre une pâte d'un concurrent, on a obtenu les préférences suivantes :

	Préfère la nouvelle pâte	Préfère la pâte du concurrent	Indifférent	total
Région 1	90	50	60	200
Région 2	105	60	35	200
total	195	110	95	400

Au risque de 5% , la préférence de la pâte suivant les 3 modalités retenues se répartit –elle –de façon identique (homogènes) dans les deux régions ?

Solution :

- **H0 :** la préférence du dentifrice, suivant les 3 modalités retenues, se répartit de façon homogène dans les deux régions.
  H1 : la préférence du dentifrice ne se répartit pas de façon homogène dans les deux régions.

- Vérifions si les conditions du test sont satisfaites :
  - Les données sont choisies aléatoirement
  - Pour la deuxième condition : Tij ≥ 5 il faut d'abord calculer les Tij

Sur le tableau ci -dessous,les nombres mis entre parenthèses sont les fréquences attendues Tij calculés avec la formule suivante :

$$Tij = \frac{(total\ de\ la\ ligne\ i)*(total\ de\ la\ colonne\ j)}{total\ général\ N}$$

	Préfère la nouvelle pâte	Préfère la pâte du concurrent	Indifférent	total
Région 1	90 (97.5)	50 (55)	60 (47.5)	200
Région 2	105 (97.5)	60 (55)	35 (47.5)	200
total	195	110	95	400

On constate que toutes les fréquences attendues Tij dépassent 5 donc la deuxième condition du test est satisfaite.

- Statistique du test :

$$X^2_c = \sum_{i=1}^{L} \sum_{j=1}^{C} \frac{(Oij - Tij)^2}{Tij} = 8.64$$

- Valeur critique :

  d.d.l ϑ = (L-1)*(C-1) = (2-1)*(3-1) = 2

  sur la table du khi-deux on lit $X^2\alpha = X^2(0.05 ; 2) = 5.99$

- Décision :

  Comme $X^2c > X^2\alpha$ donc on rejette H0

- Conclusion :

  Au risque de 5%, on ne peut pas dire que les deux régions ont un comportement homogène en ce qui concerne la préférence du dentifrice.

3.Test d'indépendance :

Un tableau de contingence est un tableau à double entrée 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 de relation entre la variable ligne et celle en colonne du tableau de contingence.

Notations :

X , Y variables aléatoires qualitatives

C : le nombre de modalités aj (j=1,…C de la variable X)

L : le nombre de modalités bi (i =1 ;…L de la variable Y)

N : la taille de l'échantillon

Oij : fréquence observée pour les modalités inscrites en ligne i et la colonne j

Tij : fréquence attendue pour les modalités inscrites en ligne i et la colonne j

On a le tableau de contingence :

Y / X	a_1	…….	a_C	total
b_1	O_{11}	……	O_{1C}	$O_1.$
…..	….	……	……	…..
b_L	O_{L1}	…….	O_{LC}	$O_L.$
total	$O._1$	…….	$O._C$	O..

Une fréquence attendue (théorique) pour un tableau de contingence est :

Tij = $\frac{(total\ de\ la\ ligne\ i)*(total\ de\ la\ colonne\ j)}{total\ général\ N}$

- Conditions d'application du test :
  - Les données d'échantillon sont sélectionnées aléatoirement.
  - Pour chaque case du tableau de contingence , la fréquence attendue Tij est au moins égale à 5.

- **H0** : les variables sont indépendantes
- **H1** : les variables ne sont pas indépendantes

- Statistique du test :

$$X^2_C = \sum_{i=1}^{L} \sum_{j=1}^{C} \frac{(Oij-Tij)^2}{Tij}$$

- Valeur critique :

Utiliser la loi du khi-deux avec un d.d.l ϑ = (C-1)*(L-1)

- Prise de décision :

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

Exemple :

Dans une population ,on étudie la liaison entre les variables qualitatives « couleur des yeux « (X) et « couleur des cheveux » (Y) .Pour cela, on constitue aléatoirement un échantillon de 200 individus et on note les observations suivantes :

Y \ X	Yeux bleus	Yeux marrons	Yeux verts
Cheveux blonds	25	15	10
Cheveux bruns	30	70	20
Cheveux roux	10	10	10

Peut-on conclure à l'indépendance de ces deux variables ?

Solution :

- **H0 :** la couleur des cheveux et la couleur des yeux sont indépendantes
- **H1 :** la couleur des cheveux et la couleur des yeux ne sont pas indépendantes

- Vérifions si les conditions du test requises sont satisfaites :

  - Les données sont choisies aléatoirement
  - Il reste à vérifier la 2eme condition les fréquences attendues Tij ≥ 5,pour cela on doit d'abord les calculer par la formule :

$$Tij = \frac{(total\ de\ la\ ligne\ i)*(total\ de\ la\ colonne\ j)}{total\ général\ N}$$

Dans le tableau suivant , les fréquences attendues Tij sont mises entre parenthèses

Y \ X	Yeux bleus	Yeux marrons	Yeux verts	Total(loi marginale)
Cheveux blonds	25 (16.25)	15 (23.75)	10 (10)	50
Cheveux bruns	30 (39)	70 ((è)	20 (24)	120
Cheveux roux	10 (9.75)	10 (14.25)	10 (6)	30
total	65	95	40	200

Les Tij sont supérieurs à 5 donc la condition est satisfaite

- Statistique du test :

$$X^2_C = \sum_{i=1}^{L}\sum_{j=1}^{C}\frac{(Oij-Tij)^2}{Tij} = 17.584$$

- Valeur critique :

La table du khi-deux nous donne $X^2(\alpha ; \vartheta) = X^2(0.05 ; 4) =9.49$

Le d.d.l ϑ =(C-1)*(L-1) = (3-1)*(3-1) =2*2=4

- Prise de décision :

comme $X^2_C > X^2(\alpha ; \vartheta)$ donc on rejette H0

- Conclusion :

Au risque de 5% , on ne peut pas dire que la couleur des cheveux et la couleur des yeux sont indépendantes.