TEST DU KHI^2

UFR14

2011

TD9

1 LA LOI DU χ^2

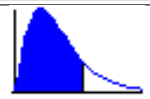
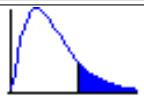
1. Définition : cette loi attribuable à Karl Pearson se déduit de la loi normale centrée réduite.

   Si Z_1, Z_2, ..., Z_ν sont ν(nu) variables aléatoires normales centrées réduites indépendantes, alors la somme des carrés de ces ν variables aléatoires:

   $S = Z_1^2 + Z_2^2 + .. + Z_\nu^2 = \sum_{i=1}^{i=\nu} Z_i^2$ suit une loi de χ^2à ν degrés de liberté.

2. Valeurs tabulées de Khi-deux

   La table du Khi-deux donne les valeurs du khi-deux qui dépendent du degré de liberté ν et du seuil de signification α.
   Exemple : pour un degré de liberté $\nu = 8$ et pour un seuil de signification α de 0.05 (5%), la table donne un $\chi^2 = 15.5073$, ce qui signifie que $P\left(\chi^2 > 15.5073\right) = 0.05$,donc que $P\left(\chi^2 \leq 15.5073\right) = 0.95 = 1 - \alpha$

	
$\Pr\left(X \leq \chi^2\right) = 1 - \alpha$	$\Pr\left(X \geq \chi^2\right) = \alpha$

2 Test du Khi-deux

1. Introduction

   Le test du Khi-deux s'inscrit dans la théorie générale des tests d'hypothèses .Il s'agit de construire une démarche qui va fournir une règle de décision permettant, sur la base de résultats d'un échantillon, de faire un choix entre deux hypothèses statistiques. On appelle **hypothèse nulle**, notée H_0 (hypothèse de différence nulle), l'hypothèse que l'on effectue sur la population parente (deux caractères sont indépendants, le nombre de clients suit une loi de Poisson,etc.) ; toute la démarche du test s'effectue en considérant cette hypothèse nulle comme vraie ; le rejet éventuel de l'hypothèse nulle conduit à l'acceptation de l'hypothèse alternative (contre hypothèse) H_1.

   On doit noter que même si l'hypothèse nulle est vérifiée sur l'échantillon tiré, les fluctuations d'échantillonnage peuvent conduire à une mauvaise conclusion. On doit donc établir des règles de décision qui conduisent sans équivoque au non-rejet ou au rejet de H_0. La décision de favoriser telle hypothèse est basée sur les résultats d'un échantillon et donc élaborée à partir d'une information très partielle ; il est impossible d'être sûr de prendre la bonne décision, on peut seulement limiter la probabilité de prendre une décision erronée.
   La décision prise à l'issue du test comporte deux risques : rejeter H_0, alors que cette hypothèse est vraie et "accepter" H_0 alors que cette hypothèse est fausse. On notera "qu'accepter" H_0 ne signifie pas que l'on a prouvé qu'H_0 est vraie, mais uniquement que les données de l'échantillon ne sont pas suffisamment contradictoires avec H_0 pour pouvoir rejeter H_0.Le cas de la justice est éloquent, le principe de présomption d'innocence stipule que tout accusé est présumé innocent (H_0) ; "accepter" ou plutôt ne pas rejeter H_0, c'est acquiter faute de preuves.

2. Les deux risques

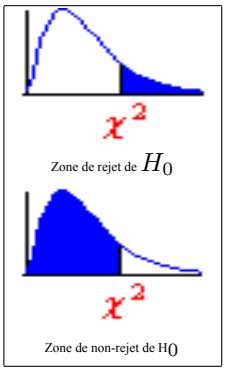
   Le premier risque, noté α est appelé le risque de première espèce : c'est le rique, consenti à l'avance, de rejeter à tort l'hypothèse nulle ; la démarche des tests va permettre de contrôler α, c'est à dire de rejeter à tort une hypothèse nulle vraie dans une faible proportion de cas ; α s'appelle le seuil de signification, les seuils les plus utilisés étant $\alpha = 0.05$ et $\alpha = 0.01$.

   Ce risque, celui en justice de condamner un innocent est grave, mais néanmoins inévitable ; peut-on le réduire ? bien sûr, mais on comprend bien que si l'on veut se tromper très rarement, et ne prendre aucun risque de rejeter à tort H_0,alors on va accepter H_0 dans tous les cas et augmenter le risque d'accepter H_0 alors qu'elle est fausse.
   Reprenons le cas de la justice, on voit bien que si l'on refuse de prendre le moindre risque de condamner un innocent, alors on doit accepter le risque de relaxer des coupables et augmenter ainsi l'autre risque, noté β et appelé le risque de seconde espèce, celui de ne pas rejeter H_0, alors que cette hypothèse est fausse. On voit donc que l'on ne peut pas trop diminuer α ; on prend le plus souvent $\alpha = 0.05$

3. Arbre de décision

a.

H_0 **Vraie** ($\nearrow \alpha$)	**Rejet** H_0	**Erreur de** $1^{ère}$ **espèce** $\alpha = P_{H_0}$ (rejeter H_0)
H_0 **Vraie** ($\searrow 1-\alpha$)	Non-rejet de H_0	Bonne décision
H_1 **Vraie** ($\nearrow \pi=1-\beta$)	Rejet H_0	Bonne décision $\beta = P_{H_1}$ (ne pas rejeter H_0)
H_1 **Vraie** ($\searrow \beta$)	**Non-rejet de** H_0	**Erreur de 2ème espèce**



4. Puissance du test :

   La probabilité, notée π, de rejeter H_0 quand cette hypothèse est fausse est appelée la puissance du test, et est donnée par $\pi = 1 - \beta$: c'est la capacité d'un test à réfuter une hypothèse fausse. Minimiser β revient à maximiser la puissance du test.

5. Test d'indépendance

   Le test du Khi-deux sert notamment pour tester l'indépendance de deux caractères qualitatifs, quand on dispose d'un tableau de contingence.

   Le principe est de "mesurer " la distance entre une distribution observée et une distribution théorique (celle de l'indépendance).

   On note O_i les effectifs observés et C_i les effectifs calculés ou théoriques, ceux de l'indépendance et

   on utilise la quantité suivante :

   $\boxed{\chi^2_{cal} = \frac{(O_1 - C_1)^2}{C_1} + \frac{(O_2 - C_2)^2}{C_2} + ... + \frac{(O_n - C_n)^2}{C_n}}$ qui suit approximativement une loi du khi-deux avec ν degrés de liberté, si l'échantillon est assez grand ($n > 30$). Il reste à préciser le degré de liberté.

   Le calcul du degré de liberté est donné par : $\nu = (l-1)(c-1)$ si l désigne le nombre de lignes et c le nombre de colonnes du tableau . L'indépendance est basée sur l'indépendance probabiliste : A et B sont indépendants si et seulement si : $P_B(A) = P(A)$.

6. Plan :
   a. Formulation des hypothèses H_0 et H_1.
   b. Choix du seuil de signification : 5%
   c. Calcul des effectifs théoriques ou calculés que nous noterons C_i (avec comme condition que chaque C_i doit être au moins égal à 5) et calcul du Khi^2. $\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$.
   d. Déterminer le ddl (degré de liberté) : $\nu = (c-1) * (l-1)$ où l et c désignent respectivement le nombre de lignes et de colonnes des données de l'échantillon, c'est à dire le nombre de modalités de chaque caractère.
   e. Lire le Khi^2 de la table ; cette valeur est dite "critique" et permet de définir les régions de rejet et d'acceptation de H_0.
   f. Règle de décision basée sur les valeurs observées de l'échantillon :
      Si la valeur du Khi^2 calculé est inférieure ou égale au Khi^2 de la table (valeur critique : seuil limite de la région de non rejet de H_0), on ne peut rejeter H_0, par contre si $\chi^2_{calc} > \chi^2_{table}$, on rejettera l'hypothèse d'indépendance statistique des deux caractères.
   g. Décision : on applique la règle de décision définie précédemment, et on conclut à partir des données de l'échantillon à l'existence ou à l'abscence d'un lien statistique entre les caractéres.
   h. Valeur p ($p - value$) ou **degré de signification** :

      Le choix d'un seuil de signification de 5% peut sembler arbitraire ; On prolonge le test par une information précieuse : le degré de signification. Dans l'exemple proposé, on trouve $\chi^2_{calc} \simeq 4.11$ et $\chi^2_{0.05;1} \simeq 3.84$; $\chi^2_{calc} > \chi^2_{0.05;1}$, on est donc conduit, au seuil de 5%, à rejeter H_0 et à accepter l'hypothèse H_1 d'une liaison significative entre le sexe et la durée du chômage. La question se trouve alors posée de l'arbitraire du seuil de 5% et on peut rechercher la plus petite valeur du risque d'erreur qui conclut à cette liaison significative : on peut avec Excel utiliser la fonction $LOI.KHIDEUX\,(4.11;1)$ qui nous renvoie la probabilité critique p de 4.26% que l'on appelle le degré de signification ou valeur p. Si l'on a fixé un seuil de signification α, on rejette H_0 si $p < \alpha$. Plus p est proche de zéro, plus forte est la contradiction entre H_0 et les données de l'échantillon.