Chapitre 6

TESTS D'AJUSTEMENT ET D'INDÉPENDANCE

1. TEST D'AJUSTEMENT DU χ^2 DE PEARSON

1.1 Cas discret.

- H_0 : les probabilités théoriques de chaque valeur i sont égales à p_i ;
- H_1 : au moins une probabilité théorique est différente de la valeur supposée p_i.

Définition : statistique X^2 choisie pour comparer les proportions observées aux probabilités théoriques :

$$X^2 = n \sum_{i=1}^{k} (n_i/n - p_i)^2 / p_i = \sum_{i=1}^{k} (n_i - n\,p_i)^2 / n\,p_i$$

- X^2 grand : n_i/n et p_i trop différents pour que l'égalité soit vraisemblable. On rejette H_0.
- X^2 faible : les différences ne sont pas contraires à l'hypothèse nulle.
- Région critique et risque α :

$$P(X^2 > x_\alpha^2) = \alpha$$

Théorème : si les termes $n \times p_i$ sont tous supérieurs ou égaux à 5, la loi de probabilité de la v.a. X^2 est approximativement la loi du χ^2 de degré de liberté $\nu = k - l - 1$, où k est le nombre de valeurs possibles des observations et l le nombre de paramètres calculés à l'aide des données.

Définition: on appelle **probabilité critique** (en anglais « p-value ») la probabilité que la valeur observée x^2 de la statistique X^2 soit dépassée.

1.2 Cas d'une variable continue.

Considérons maintenant le cas des v.a. continues :

- L'hypothèse nulle est définie par une densité théorique de X ;
- L'hypothèse alternative est une loi non précisée de X.

La procédure est la suivante :

- On définit des intervalles I_i, i = 1, …, k, donc on calcule les probabilités théoriques p_i.
- On répartit les n observations de la v.a. X dans ces intervalles.
- On en déduit la densité observée égale à la suite des proportions n_i/n, où n_i est le nombre d'observations classées dans l'intervalle I_i.
- On applique la procédure précédente pour les comparer.

2. TEST D'INDÉPENDANCE DU χ^2 DE PEARSON.

Données étudiées : tableau de contingence donnant la répartition d'un ensemble d'u.s. suivant deux critères X et Y (p lignes et q colonnes).

	Cheveux blonds (j = 1)	*Cheveux bruns (j = 2)*	*Autre couleur (j = 3)*	*Effectifs marginaux*
Masculin (i = 1)	*25*	*51*	*17*	$n_{1.} = 93$
Féminin (i = 2)	*62*	*31*	*14*	$n_{2.} = 107$
Effectifs marginaux	$n._1 = 87$	$n._2 = 82$	$n._3 = 31$	*200*

2.1 Tableau des effectifs théoriques.

- Hypothèse nulle : indépendance des deux questions ;
- Hypothèse alternative : non précisée ;

Définition : on appelle répartition théorique des unités statistiques d'un échantillon suivant deux critères la répartition que l'on aurait si ces deux critères étaient indépendants.

$$n_{i,j}' = n\, p_{i\bullet}\, p_{\bullet j} = n_{i\bullet}\, n_{\bullet j} / n$$

dans laquelle les termes $n_{i\bullet}$, $n_{\bullet j}$ sont les effectifs marginaux calculés sur le tableau de données et $p_{i\bullet}$ et $p_{\bullet j}$ les proportions marginales.

S'il y a indépendance entre le sexe et la couleur des cheveux, la répartition théorique des étudiants est la suivante :

	Cheveux blonds	*Cheveux bruns*	*Autre couleur*
Masculin	*40.46*	*38.13*	*14.42*
féminin	*46.54*	*43.87*	*16.58*

Par exemple, l'effectif théorique d'étudiantes au cheveux blonds est 107 x 87 / 200 = 46.54.

2.2 Test d'indépendance du χ^2 de Pearson.

Définition : la statistique X^2 utilisée pour comparer les répartitions théoriques et observées est définie par :

$$X^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} (n_{i,j} - n_{i,j}')^2 / n_{i,j}'$$

La loi de X^2 sous l'hypothèse d'indépendance est loi du χ^2 de degré de liberté ν :

$$\nu = (p - 1)\,(q - 1)$$

Définition: région critique $= [\chi_\alpha^2 , +\infty [$, χ_α^2 étant le nombre auquel une proportion α de X^2 est supérieure si l'hypothèse d'indépendance est vraie.

Exemple *: chaque terme du tableau ci-dessous indique la valeur du terme correspondant dans la somme donnant le X^2 appelé parfois « contribution au X^2 » :*

$5.907 = (25-40.46)^2 / 40.46$	$4.344 = (51-38.13)^2 /38.13$	$0.462 = (17-14.42)^2/14.42$
$5.136 = (62-46.54)^2/46.54$	$3.776 = (31-43.87)^2/43.87$	$0.401 = (14-16.58)^2/16.58$

La valeur de X^2 est la somme des termes du tableau. On obtient :

$$X^2 = 20.02$$

Une liaison entre la couleur des cheveux et le sexe n'étant pas du tout invraisemblable, nous choisissons un risque raisonnable α égal à 5%.

$$RC = [5.991, +\infty[$$

3. TEST SUR LE COEFFICIENT DE CORRÉLATION LINÉAIRE.

Les données sont ici quantitatives, et la liaison entre les deux variables (X,Y) est mesurée par le coefficient de corrélation linéaire.

4.1 Hypothèses et erreurs.

Soit ρ le coefficient de corrélation théorique des variables X et Y supposées normales.

- Hypothèse nulle H_0 : $\rho = 0$.
- Hypothèse alternative H_1 : $\rho \neq 0$.

Définition : l'estimateur empirique du coefficient de corrélation théorique ρ est la v.a. notée R dont la valeur observée sur un échantillon de couples est le coefficient de corrélation observé r.

4.2 Région critique.

On pose :

$$F = (n-2)\, R^2/ (1 - R^2)$$

La loi théorique de F est la loi de Fisher Snedecor de degrés de libertés ν_1, ν_2 :

$$\nu_1 = 1, \nu_2 = n-2.$$

Exemple *: on admet qu'après l'exclusion des clients retraités 25, 31 et 43, l'âge et le logarithme du revenu suivent la loi normale.*

Risque de première espèce α = 0.05	*région critique RC = [4.05, +∞ [.*
r = 0.6846	*f = 27.027.*
f > 4.05	*on rejette l'hypothèse nulle.*