Cours n° 7 : Le test de khi deux de contingence

Introduction

Voyons maintenant comment analyser la relation statistique entre deux variables qualitatives grâce au test de khi-deux de contingence que l'on appel aussi le test de khi-deux d'indépendance.

Le khi-deux, noté aussi χ^2, chi-carré ou chi-square en anglais est une mesure des écarts à l'indépendance c'est à dire une mesure de la distance numérique entre des effectifs observés et des effectifs théoriques.

En bref, le test du khi-deux de contingence teste l'indépendance entre les lignes et les colonnes d'un tableau croisé. Autrement dit, le test de khi-deux de contingence permet de se prononcer sur la « significativité » du lien entre deux variables qualitatives.

Le test de khi deux de contingence : à quoi ça sert ?

Le test de khi deux est un test statistique qui va permettre de se prononcer sur le lien entre deux variables qualitatives. Comment affirmer que les deux variables de notre distribution ne sont pas indépendante ? Autrement dit comment savoir si deux variables n'ont aucune relation statistique entre elles ?

Définition:

Le khi2 sert á comparer les fréquences théoriques aux fréquences données ou calculer a partir d´un échantillon .

L´application de la loi du khi2 est un moyen d´un procédé appelé test d´indépendance, aussi appelé le test du Khi2 est un test d´hypothèse visant á valider ou invalider l´hypothèse d´indépendance entre deux variables a l´aide de la loi du khi2.

Conditions d'application du test

Un test de khi deux s'applique uniquement sur des tableaux de contingence :

- ✓ ayant au moins 2 lignes et 2 colonnes.
- ✓ contenant des valeurs positives entières.
- ✓ ayant au minimum 5 observations par cases du tableau et/ou dans le tableau des effectifs théoriques

Les grandes étapes d'application du test de khi-deux de contingence

1 - Identifier si le test de khi-deux est applicable sur le tableau.

2 - Si oui, formuler l'*hypothèse d'indépendance* : "On fait l'hypothèse qu'il y a une relation d'indépendance entre les lignes et les colonnes du tableau".

3 - Calculer l'*indicateur de khi-deux* en calculant le tableau des *effectifs théorique* et le tableau des *écarts à l'indépendance*.

4 - Confronter l'indicateur de khi-deux à la *table de loi de khi-deux* ce qui nous permet d'accepter ou de rejeter l'hypothèse d'indépendance.

5 - Interpréter les résultats du test en examinant le tableau d'origine (ou tableau des effectifs observés), le tableau des effectifs théoriques et le tableau des écarts à l'indépendance.

La loi du khi2 se calcule selon la formule suivante:

$$\chi^2 = \frac{\sum (t\,x - t'x)^2}{\sum t'x}$$

D´ou:

t x: Les effectifs dans les différentes case d´un tableau statistique.

t´x: Se calcule selon la formule suivante:

$$t'x = \frac{\text{(effectifs cumulé de la colonne Xi)x(effectifs cumulé de la ligne Xi)}}{\text{Ensemble des effectifs}}$$

La formulation des hypothèses:

Comme dans tous les tests d´hypothèses, il faut formuler deux hypothèses pour appliquer le test d´indépendance: l´hypothèse nulle et hypothèse alternative.

Soient *X*1 et *X*2 deux variables qualitatives. Sous *H*0, la distribution de *X*1 devrait être *indépendante* de celle de *X*2. *Au contraire*, si la distribution de *X*1 est *liée* a celle de *X*2, on rejette *H*0 au prote de *H*1, les deux variables *X*1 et *X*2 sont liées.

H0 : Les variables *X*1 et *X*2 sont indépendantes.

H1 : Il existe une liaison entre *X*1 et *X*2.

d´ou:

- ✓ L´hypothèse nulle **(H0):** est l´hypothèse de départ selon laquelle, il n´existe aucun lien de dépendance entre les deux variables.
- ✓ L´hypothèse alternative **(H1):** aussi appelé hypothèse du chercheur, c.à.d. qu´il existe un lien de dépendance entre les deux variables.

Les étapes pour calculer le test du Khi2:

a) Emettre l´hypothèse concernant la conformité entre les fréquences calculées á partir d´un échantillon et les fréquences théoriques.

b) Comparer (χ^2) calculer a (χ^2) théorique, déterminé a partir de la table statistique du khi2 á un seuil de signification **alpha** (∞) et un degré de liberté (ddl).

Le choix du seuil de signification (ou seuil alpha):

Le seuil de signification est un paramètre du test d´hypothèse et sa valeur est fixée par l´utilisateur avant la réalisation du test.

Les tests d´hypothèses sont des instrument imparfait qui servent á prendre des décessions s´appuyant sur des probabilités .

Exemple: **alpha** (∞)= 5% → 5/100= 0,05.

alpha (∞)= 1% → 1/100= 0,01.

Nombre de degré de liberté: (DDL):

Le nombre de degré de liberté (ddl) est un paramètre de la loi du khi2,est égale au nombres de cellules de la partie centrale du tableau de contingence.

Le degré de liberté (ddl) se calcule comme suite:

ddl= **(Nombre de modalités x - 1) x (Nombre de modalités y - 1)**

Commentaire et interprétation du Khi2:

- Si χ^2calcule < χ^2théorique, on peut conclure que la différence entre la réparation des données dans l´échantillon et celle de la population **n´est pas significative**, elle est due au hasard á un seuil de signification **alpha** (∞).

- Si χ^2calcule > χ^2théorique, on peut conclure que la différence entre la réparation des données dans l´échantillon et celle de la population **est pas significative** ,elle est due a une cause systématique á un seuil de signification **alpha** (∞).

- Si χ^2calcule = χ^2théorique, le khi2 **n´est pas utile** pour la comparaison.

Exemple:

soit un tableau de contingences d´effectifs observées avec deux variables qualitatives (Genre et tabagisme) qui se traduit dans le tableau suivant.

Tabagisme / Genre	Fumeur	Non fumeur	Ensemble
Masculin	36	16	52
Féminin	18	28	46
Ensemble	54	44	98

Question: tester au seuil de signification de 5% l´hypothèse d´indépendance entre le sexe et tabagisme.

1. Emettre l´hypothèse:
   Ho: il n´existe aucun lien entre le genre et le tabagisme.
   H1: il existe un lien entre le genre et le tabagisme.
2. On calcule les effectifs selon la formule t´x:

$$t'x1 = \frac{54 \times 52}{98} = 28,65 \qquad t'x2 = \frac{44 \times 52}{98} = 23,34$$

$$t'x3 = \frac{54 \times 46}{98} = 25,34 \qquad t'x4 = \frac{44 \times 46}{98} = 20,65$$

t x	t´x	(t x-t´x)	(t x-t´x)²	$\frac{(t\ x-t´x)^2}{t´x}$
36	28,65	7,35	54,02	1,88
16	23,34	7,34	53,87	2,30
18	25,34	7,34	53,87	2,12
28	20,65	7,35	54,02	2,61
/	/	/	/	**8,91**

3. On calcule le degré de liberté (ddl):
   ddl= **(Nombre de modalités x- 1) x (Nombre de modalités y-1)**
   ddl=(2-1)x(2-1)= **1**
4. Seuil de signification:
   ∞= 5/100= 0,05
   Donc on aura la valeur du X² théorique = **3,84** (voir la table du Khi2)
5. La comparaison entre les deux valeurs
   χ^2calcule = **8,91**
   χ^2théorique = **3,84**
6. Commentaire:
   On déduit : le X² calcule > X² théorique.
   La différence entre la distribution de données dans l´échantillon étudie et la population dont il provient est significative et due á une cause systématique á (100-0,05) = 95% de chance.
   Donc on accepte l´hypothèse alternative (H1),qu'il existé un lien entre le genre et le tabagisme, et on refuse l´hypothèse nulle (H0).

Extrait de la table du khi deux

ddl \ ∞	0,1	0,05	0,025	0,01
1	2,72	3,84	5,02	6,63
2	4,61	5,99	7,38	9,21
3	6,25	7,81	9,35	11,34
4	7,78	9,49	11,14	13,28

Exercice : On veut étudier la différence entre le nombre de résidents accédant a deux types de formation médicale.

Formation A / Formation B	Admis	Refuses	Ensemble
Admis	42	48	90
Refuses	78	112	190
Ensemble	120	160	280

Question: La différences entre ces deux nombres est-elle significative a **5 %** ?

1. Emettre l´hypothèse:
   Ho: il n´existe aucun lien entre les deux formations.
   H1: il existe un lien entre les deux formations.
2. On calcule les effectifs selon la formule t´x:

$$t'x1 = \frac{120 \times 90}{280} = 38,57 \qquad t'x2 = \frac{160 \times 90}{280} = 51,42$$

$$t'x3 = \frac{120 \times 190}{280} = 81,42 \qquad t'x4 = \frac{160 \times 190}{280} = 108,57$$

t x	t´x	(t x-t´x)	(t x-t´x)²	$\frac{(t\,x-t'x)^2}{t'x}$
42	38,57	3,43	11,76	0,30
48	51,42	3,42	11,69	0,22
78	81,42	3,42	11,69	0,14
112	108,57	3,,43	11,76	0,10
/	/	/	/	**0,76**

3. On calcule le degré de liberté (ddl):
   ddl= **(Nombre de modalités x- 1) x (Nombre de modalités y-1)**
   ddl=(2-1)x(2-1)= **1**
4. Seuil de signification:
   ∞= 5/100= 0,05
   Donc on aura la valeur du X^2 théorique = **3,84** (voir la table du Khi2)
5. La comparaison entre les deux valeur
   χ^2calcule = **0,76**
   χ^2théorique = **3,84**
6. Commentaire:
   On déduit : le X^2 calcule < X^2 théorique.
   on peut conclure que la différence entre la réparation des données dans l´échantillon et celle de la population **n´est pas significative**, elle est due au hasard á un seuil de signification **alpha** (∞).
   Donc on accepte l´hypothèse nulle (H0),qu'il n´existé aucun lien entre les deux formations, et on refuse l´hypothèse alternative (H1).

Extrait de la table du khi deux

ddl \ ∞	0,1	0,05	0,025	0,01
1	2,72	3,84	5,02	6,63
2	4,61	5,99	7,38	9,21
3	6,25	7,81	9,35	11,34
4	7,78	9,49	11,14	13,28

Série d'exercices sur le khi 2.

Exercice 1: On a vaccine contre la grippe 300 personnes, reparties en fonction de l'âge et la maladie, retrace dans le tableau suivant.

Maladie / Age	Ont la grippe	n'ont pas la grippe	Ensemble
Moins de 55 ans	38	82	120
Plus de 55 ans	73	107	180
Ensemble	111	189	300

Question: peut-on au risque de **5%** considère qu'elle existe une corrélation entre l'efficacité du vaccin et l'âge des personnes vaccinées?

Exercice 2: une cite connait depuis quelques temps des problèmes importants d'alcoolismes. Le comite du quartier a décider de procéder a une enquête pour mieux cerner le problème par sondage auprès des jeunes .

Genre / Alcoolisme	Masculin	Féminin	Ensemble
Oui	18	14	32
Non	23	36	59
Ensemble	41	50	91

Question: Calculer a l'aide d'un test du khi2 l'existence d'une relation entre l'alcoolisme et les variables explicatives au seuil de **1%** .

Exercice 3: Lors d'une enquête faite sur le degré de motivation et la lecture en fonction du genre a donner les résultats suivant.

Motivation / Genre	motive	moyennement motive	démotive	Ensemble
Fille	13	7	8	28
Garçon	11	8	5	24
Ensemble	24	15	13	52

Question: La motivation varie t-elle en fonction du genre au seuil de signification de **5%** ?

Exercice 4: Une enquête faite sur l'influence du milieu géographique et les résultats scolaires, retrace dans le tableau suivant.

Milieu / Résultats	Rural	Urbain	Ensemble
Réussite	145	122	267
Echec	150	153	303
Ensemble	295	275	570

Question: Calculer le khi au seuil de signification de **1%** .

Exercice 5: Soit un tableau de contingence d'effectifs observes avec deux variables qualitatives (revenu et consommation de drogue) qui se distribuent ainsi .

Revenu / Alcoolisme	Haut	Bas	Ensemble
Oui	28	24	52
Non	33	46	79
Ensemble	61	70	131

Question: Tester au seuil de **5%** l'hypothèse d'indépendance entre le revenu et la consommation de drogue.