Test du Khi^2

Le test du Khi^2 (khi deux ou khi carré) fournit une méthode pour déterminer la nature d'une répartition, qui peut être continue ou discrète.

Domaine d'application du test :

- Données qualitatives
- 2 ou plusieurs échantillons
- Dépendants ou indépendants
- Comparaison d'échantillons
- Recherche de liaison entre les données
- Recherche de l'influence d'une donnée autre que celle étudiée

La démarche à suivre :

- Formuler H0 (la distribution observée n'est pas différente de la distribution supposée d'après la loi que l'on souhaite tester).
- Répartir les données en classes
- déterminer le nombre de degrés de liberté à partir du nombre de classes

- fixer un risque de se tromper (la valeur 5 % est souvent choisie par défaut)
- calculer algébriquement la distance entre les ensembles d'informations à comparer
- déterminer Kh² théorique (déduire la distance critique à l'aide d'une table de χ^2)
- conclure si cette distance est supérieure à la distance critique (on conclut que le résultat n'est pas dû seulement aux fluctuations d'échantillonnage).

RQ : Le rejet de H0 est une réponse négative dans les tests de Khi² d'adéquation et d'homogénéité alors qu'il apporte une information positive dans le test de Khi² d'indépendance.

1) Test d'ajustement

Objectif: Les observations faites sur un échantillon conduisent à une certaine distribution de fréquences. Peut-on modéliser cette distribution par un modèle théorique connu.

Méthode : La méthode consiste à comparer l'histogramme des fréquences et la distribution de la loi de probabilité servant du modèle théorique.

Après avoir découpé l'intervalle d'observation en k classes, on construit un indice **d** mesurant l'écart constaté entre les effectifs réels et les effectifs théoriques.

$$d = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

où ni = effectif observé dans la classe i
n = effectif total observé
pi = probabilité d'obtenir une observation de la loi de probabilité théorique dans la classe i
npi = effectif théorique dans la classe i

Hypothèses:
Ho = {la distribution observée n'est pas significativement différente de la distribution théorique}
contre H1 = {la distribution observée est significativement différente de la distribution théorique}

Statistique de test:

$$D = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

Remarques:
Le choix et le nombre de classes est arbitraire. Cependant pour que l'approximation par la loi du χ^2 soit bonne, il est nécessaire que les effectifs théoriques dans chacune des classes soit au moins égal à 5. Si ce n'est pas le cas, il faut au préalable regrouper les classes contigües afin d'avoir un effectif suffisant. La valeur de k intervenant dans le nombre de degrés de liberté de la loi du χ^2 est celle obtenue après les éventuels regroupements.

2) Test d'indépendance

Lorsqu'on considère plusieurs populations auxquelles on associe le même ensemble de critères qualitatifs, l'hypothèse à tester est l'indépendance entre la population d'appartenance de l'individu et la valeur des critères. L'hypothèse affirme donc que le fait de connaître la population d'un individu n'influence pas la valeur des critères.

But du test : Le test d'indépendance du chi-carré vise à déterminer si deux variables observées sur un échantillon sont indépendantes ou non. Les variables étudiées sont des variables qualitatives catégorielles.

Ce test s'applique lorsqu'on souhaite démontrer l'indépendance ou la dépendance de deux critères dans une expérience.

Ce test s'effectue sur la base d'une table de contingence : plusieurs échantillons pouvant être classés selon un certain nombre de colonnes (critère 1) et de lignes (critère 2).

Objectif: Les observations de deux variables qualitatives faites sur un échantillon permettent de juger l'indépendance de ces variables.

Méthode : La méthode consiste à comparer les effectifs réels des croisements des modalités des deux variables qualitatives avec les effectifs théoriques qu'on devrait obtenir dans le cas d'indépendance de ces deux variables.

On construit un indice d mesurant l'écart constaté entre les effectifs réels et les effectifs théoriques.

$$d = \sum_{i,j} \frac{(n_{ij} - np_{ij})^2}{np_{ij}}$$

où nij = effectif observé des individus possédant la modalité i de la 1° variable et la modalité j de la 2° variable n = effectif total observé ; pij = probabilité d'obtenir une observation possédant la modalité i de la 1ère variable et la modalité j de la 2ème variable lorsqu'elles sont indépendantes.

npij = effectif théorique des individus possédant la modalité i de la 1ère variable et la modalité j de la 2ème variable.

$$D = \sum_{i,j} \frac{(N_{ij} - np_{ij})^2}{np_{ij}}$$

Statistique de test: On utilise la statistique

$$D \longrightarrow \chi^2_{(l-1)(c-1)}$$

où l est le nombre de modalités de la 1° variable (nombre de lignes du tableau de contingence) et c est le nombre de modalités de la 2° variable (nombre de colonnes du tableau de contingence)

Remarques:

$$np_{ij} = \frac{\left(\sum_j n_{ij}\right)\left(\sum_i n_{ij}\right)}{n} = \frac{n_{i.}\, n_{.j}}{n}$$

où ni. est l'effectif des individus possédant la modalité i de la 1° variable et n.j l'effectif des individus possédant la modalité j de la 2° variable.

Exercice d'application:

La distribution du revenu des hommes est-elle différente de celle des femmes ?

Une représentation sur une table de contingence des occurrences des variables permet d'illustrer la question.

Salaire	**1000-2000**	**2000-3000**	**3000-4000**	**4000-5000**	**Total**
Hommes	50	70	110	60	290
Femmes	60	75	100	50	285
Total	110	145	210	110	**575**

On remarque que les femmes sont plus nombreuses dans les classes à bas salaires et moins nombreuses dans celles à haut salaire que les hommes.

Cette différence (c'est-à-dire cette dépendance entre les variables) est-elle statistiquement significative ?

Le test du χ^2 aide à répondre à cette question.

L-1= 4-1 = 3 ; C-1=2-1 = 1 ; **ddl =3 x 1 = 3**

Pour alpha = 5 %, la valeur critique (Khi² théorique) trouvée dans les tables de Khi² est 7,81.

Hypothèse : On suppose qu'il n'y a pas de différence entre les salaires des hommes et ceux des femmes.

Tableau des fréquences théoriques :

Hypothèse	**1000-2000**	**2000-3000**	**3000-4000**	**4000-5000**	**Total**
Hommes	55,5	73,1	105,9	55,5	290,0
Femmes	54,5	71,9	104,1	54,5	285,0
Total	110,0	145,0	210,0	110,0	**575,0**

Le calcul du χ^2 :

χ^2	**1000-2000**	**2000-3000**	**3000-4000**	**4000-5000**	**Total**
Hommes	0,54	0,13	0,16	0,37	1,20
Femmes	0,55	0,14	0,16	0,38	1,23
Total	1,09	0,27	0,32	0,75	**2,43**

Conclusion :

La distance calculée (2,43) étant inférieure à la distance critique (7,81), il n'y a pas lieu de mettre en cause l'égalité des salaires, avec un risque de se tromper égal à 5%.

---

3) Test d'homogénéité

Objectif: Les observations d'une variable qualitative sur k échantillons permettent-elles de conclure que les échantillons proviennent de la même population?

Méthode : la méthode consiste à comparer les effectifs réels de chaque modalité sur les différents échantillons, avec les effectifs théoriques qu'on devrait obtenir dans le cas où ils seraient issus d'une même population. Pour cela, on construit un **indice d** mesurant l'écart constaté entre les effectifs réels et les effectifs théoriques.

$$d = \sum_{i,j} \frac{(n_{ij} - np_j)^2}{np_j}$$

où nij = effectif observé des individus de l'échantillon i, possédant la modalité j de la variable n = effectif total observé ; pj = probabilité d'obtenir une observation possédant la modalité j de la variable lorsqu'on est en présence d'une seule population ; npj = effectif théorique des individus possédant la modalité j de la variable lorsqu'on est en présence d'une seule population.

Hypothèses:

Ho = {les k échantillons sont issus d'une seule population} contre H1 = {les k échantillons sont issus de deux populations différentes}.

Statistique de test:

$$D = \sum_{i,j} \frac{(N_{ij} - np_j)^2}{np_j}$$

On utilise la statistique

$$D \longrightarrow \chi^2_{(k-1)(r-1)}$$

où k est le nombre d'échantillons (nombre de lignes du tableau) et c est le nombre de modalité de la variable (nombre de colonnes du tableau)

Remarques:

$$np_i = \frac{\left(\sum_j n_{ij}\right)\left(\sum_i n_{ij}\right)}{n} = \frac{n_i n_j}{n}$$

où ni. est l'effectif de l'échantillon i et n.j l'effectif des individus possédant la modalité j de la variable.

Exemple 1 :

Existe t-il un lien entre le nombre de grossesse et le décès des bébés ?

1/ Ranger les valeurs sous forme de tableau : Il s'agit des fréquences observées

Age du décès	**Nombre de grossesses inférieur à 3**	**Nombre de grossesses supérieur à 3**
Inférieur à 3 mois	18	6
Supérieur à 3 mois	17	19

2/ Calculer la somme des colonnes et des rangées :

2/ Dans notre exemple :

Age du décès	**Nbre de grossesses inférieur à 3**	**Nbre de grossesses supérieur à 3**	**Total lignes**
Inférieur à 3	18	6	**24**

mois			
Supérieur à 3 mois	17	19	**36**
Total colonnes	**35**	**25**	**60**

3/ Calculer les fréquences théoriques : Il s'agit du produit de la somme des colonnes par la somme des rangées, respectif, divisé par la somme totale

3/ Fréquences théoriques :

Age du décès	**Nombre de grossesses inférieur à 3**	**Nombre de grossesses supérieur à 3**
Inférieur à 3 mois	35 X 24 = 840/60 = **14**	25 X 24 = 600/60 = **10**
Supérieur à 3 mois	35 X 36 = 1260/60 = **21**	25 X 36 = 900/60 = **15**

4/ Calculer Khi² :

$$\chi^2 = \frac{(14-18)^2}{14} + \frac{(10-6)^2}{10} + \frac{(21-17)^2}{21} + \frac{(15-19)^2}{15}$$

$$\chi^2 = \frac{16}{14} + \frac{16}{10} + \frac{16}{21} + \frac{16}{15}$$

$$\chi^2 = 4{,}57$$

5/ Calculer le degré de liberté :

(2-1)*(2-1) = 1

6/ Comparer le Khi² calculé au Khi² de la table

Khi²calculé = 4,57

Khi² de la table est de 3,84 pour un ddl = 1 et un seuil de confiance à 95%.

Khi² calculé = 4,57 est supérieur au **Khi²** de la table = 3,84

7/ Conclusions

Khi² calculé est supérieur au **Khi²** de la table, donc il existe une différence significative entre l'âge du décès du bébé et le nombre de grossesses.

Exemple 2 :

On a lancé un dé 90 fois et on a obtenu les issues 1 à 6 (k=6) avec les effectifs suivants: 12, 16, 20, 11, 13, 18
Nous vérifions que 90 lancers sont suffisants: n (1/6) ≥5 implique que n ≥36).

Si le dé n'est pas tronqué (notre hypothèse), on attend comme effectifs moyens théoriques 15 pour toutes les issues.

$$Q = \frac{(12-15)^2}{15} + \frac{(16-15)^2}{15} + \frac{(20-15)^2}{15} + \frac{(11-15)^2}{15} + \frac{(13-15)^2}{15} + \frac{(18-15)^2}{15} = \frac{64}{15} = 4.266$$

Pour k-1=5 degrés de liberté et un seuil de tolérance de 5%, la valeur $\chi^2_{k-1,\alpha}$ du tableau est 11.1. Cela signifie que la probabilité que Q soit supérieur à 11.1 est de 5%.

Comme 4.266 < 11.1, on accepte l'hypothèse selon laquelle le dé est régulier.

Exemple 3:

On cherche à déterminer si le fait de fumer est indépendant du sexe des individus. Les deux variables à étudier sont des **variables qualitatives catégorielles** qui comptent deux catégories chacune :

- pour la variable "sexe" : M et F,
- pour la variable "fumer" : "fume" ou "ne fume pas".

La table de contingence obtenue à partir d'un échantillon de 100 individus (n = 100) est la suivante :

		Fumer		
		"fume"	"ne fume pas"	Total
Sexe	M	21	44	65
	F	10	25	35
	Total	31	69	100

Notons n_{ij} (i = 1, 2, j = 1, 2) les fréquences observées. Nous allons estimer les fréquences de chaque case du tableau sous l'hypothèse de l'indépendance entre les deux variables.

Notons e_{ij}, les fréquences estimées : $e_{ij} = \frac{1}{n_{..}} \sum_{i=1}^{p} \sum_{k=1}^{n_i} x_{ik}$

Nous obtenons : $e_{11} = \frac{65*31}{100} = 20.15$ $e_{12} = \frac{65*69}{100} = 44.85$ et $e_{21} = \frac{35*31}{100} = 10.85$ $e_{22} = \frac{35*69}{100} = 24.15$

En définitive, le tableau de fréquences estimées est le suivant :

		Fumer		
		"fume"	"ne fume pas"	Total
Sexe	M	20.15	44.85	65
	F	10.85	24.15	35
	Total	31	69	100

Si l'hypothèse nulle H_0 est vraie, la statistique Khi $^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$

suit une loi du chi-carré avec (l-1)(c-1) degrés de liberté, où l est le nombre de lignes et c le nombre de colonnes du tableau. Dans notre exemple, nous avons donc (2-1)(2-1) = 1 degré de liberté et Khi ^{2} = 0.036 + 0.016 + 0.066 + 0.030 = 0.148. En choisissant un seuil de signification de 5%, la valeur Khi² de la table de chi-carré est égale à 3.84. Comme la valeur de Khi ^{2} calculée est nettement inférieure à la valeur trouvée dans la table du chi-carré, nous ne rejetons pas l'hypothèse nulle et concluons que les deux variables étudiées sont indépendantes.

Exemple 4:

Au cours d'une enquête, on interroge (1) 369 mères d'enfants nés avec une malformation et (2) 968 mères d'enfants nés sans malformation. On constate que 35,06% des mères d'enfants nés avec malformations et 33,02% des mères d'enfants nés sans malformations fumaient.

Hypothèse initiale (hypothèse nulle H0): le fait d'être mal formé à la naissance ne dépend pas du fait que la mère soit fumeuse ou non. Les deux critères sont indépendants.

Hypothèse alternative (H1): Les 2 critères "enfants malformés ou non" et "avoir une mère fumeuse ou non" sont dépendants (liés)

Remarque: Dans cet exemple, "être un enfant mal formé ou normal" constitue 2 états du critère 1 et "être issu d'une mère non fumeuse ou fumeuse" constitue 2 états du critère 2.

Voici le tableau de contingence:

valeurs observées:	**enfant malformé**	**enfant "normal"**	**Total**
mère fumeuse	480	980	1460
mère non fumeuse	889	1988	2876
Total	1369	2968	4337

Dans ce tableau, on retrouve les fréquences expérimentales mais aussi les totaux par lignes (1460 enfants sont nés de mères fumeuses contres 2876 de mères non fumeuses) et par colonnes (1369 enfants sont nés avec des malformations contre 2968 enfants normaux) ainsi que le nombre total d'individus analysés dans l'expérience (4337).

Voici le tableau de fréquences théoriques:

valeurs théoriques:	**enfant malformé**	**enfant "normal"**	Total
mère fumeuse	= 1369 *1460 / 4337 = 460,85	= 2968 *1460 / 4437 = 999,15	**1460**
mère non fumeuse	= 1369 * 2876 / 4437 = 908,14	= 2968 * 2876 / 4437 = 1968,86	**2876**
Total	**1369**	**2968**	**4437**

Pour déterminer χ^2 observé pour chaque échantillons de l'expérience, on applique la formule suivante:

χ^2 = (fréquence observée - fréquence théorique) ^{2} / fréquence théorique

On obtient alors le tableau suivant:

	enfant malformé	**enfant "normal"**	
mère fumeuse	[480 - 460,85]^2 / 460,85	[(980 - 999,15)^2 / 999,15	**1460**
Mère non fumeuse	(889 - 908,14)^2 / 908,14	(1988 - 1968,86)^2 / 1968,86	**2876**
	1369	**2968**	**4437**

ou encore:

	enfant malformé	**enfant "normal"**
mère fumeuse	0,783	0,361
mère non fumeuse	0,397	0,183

Khi² exp = 0,783 + 0,361 + 0,397 + 0,183 = **1,72**

Khi² th $_{1dl;0,95}$= 3,84

La valeur de **Khi²** observé est inférieure à 3,84 ; alors on accepte H0 ; cela implique que les mères fumeuses n'ont pas plus ou moins de chance de donner naissance à un enfant anormal qu'une mère non fumeuse. Les deux critères sont indépendants, je peux l'affirmer avec seulement 5 chances sur 100 de me tromper.