Le test d'indépendance du Khi-carré de PEARSON

Dernière mise à jour le 23 mars 2010

Le test d'indépendance du khi-carré (l'écriture anglaise est « chi-square ») a été développé par Karl PEARSON (1857-1936).

L'expression test du khi-carré recouvre plusieurs tests statistiques[1] , trois tests principalement :

- le test d'ajustement ou d'adéquation, qui compare globalement la distribution observée dans un échantillon statistique à une distribution théorique, celle du khi-carré.
- Le test d'indépendance du khi-carré qui permet de contrôler l'indépendance de deux caractères dans une population donnée.
- le test d'homogénéité du khi-carré qui teste si des échantillons sont issus d'une même population.

Le test qui nous intéresse ici est uniquement le test d'indépendance du khi-carré. Ce test sert à apprécier l'existence ou non d'une relation entre deux caractères au sein d'une population, lorsque ces caractères sont qualitatifs où lorsqu'un caractère est quantitatif et l'autre qualitatif, ou bien encore lorsque les deux caractères sont quantitatifs mais que les valeurs ont été regroupées.

À noter que ce test permet de contrôler l'existence d'une dépendance mais en aucun cas le sens de cette dépendance (sauf dans certains cas particuliers où l'existence d'une relation implique une causalité univoque comme dans l'exemple ci-après où le sexe peut avoir une influence sur le choix d'une certaine matière mais où il est impossible que le choix d'une certaine matière ait une influence sur le sexe).

À noter enfin que les différents tests du khi-carré ne doivent pas être confondus avec la distribution théorique du khi-carré, dont les valeurs tabulées servent seulement à valider ces différents tests.

Voyons comment ce test peut-être utilisé dans le cas d'une distribution à deux caractères[2].

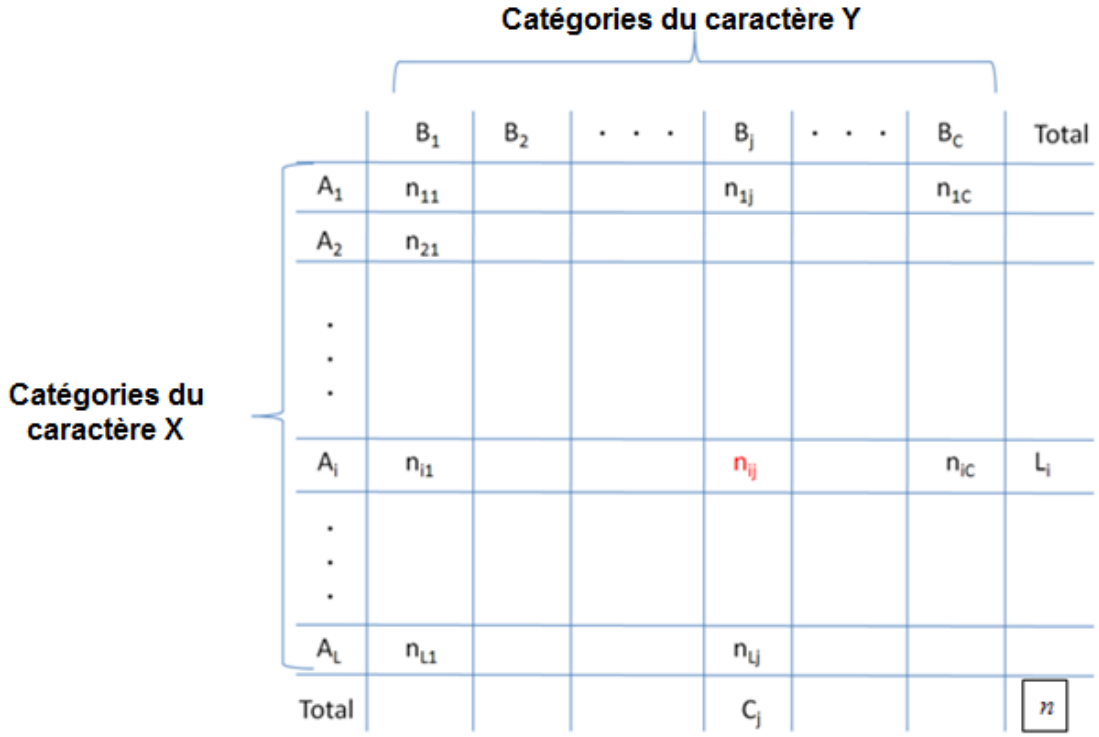
Le premier caractère, désigné par *X*, pourra être un caractère quantitatif ou qualitatif, comprenant des catégories (ou des classes) (issues généralement d'un regroupement des valeurs d'un caractère quantitatif ou des modalités d'un caractère non quantitatif). On aura ainsi les classes A_1, …, A_L

Le second caractère, désigné par *Y*, pourra être un caractère quantitatif ou qualitatif, comprenant des catégories (ou des classes) (issues généralement d'un regroupement des valeurs d'un caractère quantitatif ou des modalités d'un caractère non quantitatif). On aura ainsi les classes B_1, …, B_C.

[1] Une présentation synthétique des différents tests est donnée dans Wikipédia (voir l'article « Test du khi-2 »)

[2] Pour une très bonne explication de la façon d'effectuer un test d'indépendance du khi-2 (ou chi-2), voir Charles McCREERY « The CHI-SQUARE test : A test of Association Between Categorical Variables ». Sur internet : http://www.celiagreen.com/charlesmccreery/statistics/chisquare.pdf. Voir aussi les explications très claires données sur BibMath dont nous nous sommes inspirés ci-après : http://www.bibmath.net/dico/index.php3?action=affiche&quoi=./c/chideuxtest.html .

Dans ces conditions, l'effectif *n* de la population se distribue dans un tableau croisé[3] :



Où $n_{i,j}$ représente l'effectif qui appartient simultanément à la catégorie A_i de la dimension *X* et à la catégorie B_j de la dimension *Y*.

L_i représente la somme des effectifs appartenant à la catégorie *Ai* de la dimension X. C'est donc une distribution conditionnelle (voir le chapitre 2). C_j représente la somme des effectifs de la catégorie Bj. C'est donc aussi une distribution conditionnelle.

[3] Appelé « Contingency table » en anglais et abusivement traduit en français par l'expression « Tableau de contingence »

Exemple : Sexe et préférence pour un cours au sein d'une filière « économie »

Soit le tableau ci-dessous, qui donne les résultats d'une enquête hypothétique effectuée auprès de 400 étudiants, sur leurs préférences en matière de cours. On leur a demandé : « *Parmi ces 4 matières : HPE, Droit, Micro et Macro, laquelle préférez-vous ?* » (Il était interdit de répondre : « aucune »).

	H	F	Total
HPE	50	50	100
Droit	110	25	135
Micro	40	25	65
Macro	50	50	100
Total	250	150	400

Dans cet exemple, le caractère *Y* est le sexe et comprend deux modalités (« H » et « F »). Le caractère *X* est la matière, qui comprend 4 modalités (« HPE », « Droit » « Micro » et « Macro »). On remarquera que les « catégories » des caractères X et Y ne sont pas issues d'un regroupement, mais qu'il s'agit simplement des modalités brutes de chacun des deux caractères étudiés.

Pour savoir si le sexe a une influence significative sur le choix des matières, nous allons faire un test du khi-carré. On remarque que le droit et la micro sont davantage préférés par la population masculine tandis que HPE et macro semblent ne pas être préférés plus par la population masculine que par la population féminine.

Le test du khi carré va apporter une information supplémentaire. Il va permettre de dire si les différences de préférences pour les diverses matières qui sont attribuées au sexe sont le fait du hasard du tirage ou si elles sont réelles. Elles peuvent en effet être dues au hasard de l'échantillon. Ce que le test va nous dire c'est dans quelle mesure la différence est indépendante de l'échantillon choisi (et donc se retrouverait en général si l'on prenait n'importe quel autre échantillon).

Pour cela on doit calculer l'expression suivante, que nous appellerons, faute d'une expression plus appropriée, le « khi-carré calculé »[4] :

$$\chi^2_{\text{calculé}} = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

Avec :

$$e_{ij} = \frac{L_i C_j}{n} = \frac{C_j L_i}{n}$$

Une fois que l'on connaît le khi-carré calculé, il reste à le comparer avec la valeur khi-carré issue de la distribution du khi-carré (voir le tableau ci-après).

[4] Pour obtenir directement le khi-2 calculé, voir le calculateur en ligne (très pratique pour vérifier ses résultats) : http://www.seuret.com/biostat/chi.php

	H	F	Total
HPE	50	50	100
Droit	110	25	135
Micro	40	25	65
Macro	50	50	100
Total	250	150	400

Et en effectuant les multiplications comme indiqué par la formule $e_{i,j} = \frac{C_{i,j} L_{i,j}}{n}$:

(250 x 100)/400 = 62,5 (150 x 100)/400 = 37,5
(250 x 135)/400 = 84,375 (150 x 135)/400 = 50,625
(250 x 65)/400 = 40,625 (150 x 65)/400 = 24,375
(250 x 100)/400 = 62,5 (150 x 100)/400 = 37,5

	Hommes	Femmes	Total
HPE	62,5	37,5	100
Droit	84,375	50,625	135
Micro	40,625	24,375	65
Macro	62,5	37,5	100
Total	250	150	400

Ensuite calculons les $n_{i,j} - e_{i,j}$:

$n_{i,j}$

	Hommes	Femmes
HPE	50	50
Droit	110	25
Micro	40	25
Macro	50	50
Total	250	150

Moins

$e_{i,j}$:

	Hommes	Femmes
HPE	62,5	37,5
Droit	84,375	50,625
Micro	40,625	24,375
Macro	62,5	37,5
Total	250	150

égal

$n_{i,j} - e_{i,j}$:

	Hommes	Femmes
HPE	-12,5	12,5
Droit	25,625	-25,625
Micro	-0,625	0,625
Macro	-12,5	12,5

Les chiffres de ce tableau au carré

Puis calculons $(n_{i,j} - e_{i,j})^2$:

	Hommes	Femmes
HPE	156,25	156,25
Droit	656,640625	656,640625
Micro	0,390625	0,390625
Macro	156,25	156,25

Ensuite calculons $(n_{i,j} - e_{i,j})^2 / e_{i,j}$:

	Hommes	Femmes
HPE	2,5	4,16666667
Droit	7,78240741	12,970679
Micro	0,00961538	0,01602564
Macro	2,5	4,16666667

Ensuite effectuons la somme des huit chiffres obtenus :

$$\chi^2_{\text{calculé}} = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}} = 2,5 + 7,782 + 0,00961 + 2,5 + 4,166 + 12,97 + 0,016 + 4,1666 = 34,11 \text{ environ}$$

Pour trouver cette valeur dans le tableau, nous devons prendre en compte deux informations supplémentaires :

- Le nombre de « degrés de liberté » qui se calcule ainsi :

Degrés de liberté
=
{(Nb de catégories [ou modalités ou valeurs] de X) – 1}
x
{(Nb de catégories [ou modalités ou valeurs] de Y) – 1}

Ici, il y a 4 modalités pour *X* (les 4 matières) et 2 modalités pour *Y* (les deux sexes). Donc, le nombre de degrés de liberté est égal à :

(4 - 1) x (2 – 1) = 3 x 1 = 3.

- Ensuite, nous devons choisir la probabilité de fiabilité du test : 5% de chances de se tromper, 1% ou 1 pour 1000. Nous allons choisir 5%, soit P = 0,05.

Nous avons donc 3 degrés de liberté et une probabilité de fiabilité du test de P=0,05. Par conséquent, nous voyons dans la table que le khi-carré théorique est égal à :

$$\chi^2_{0,05} = 7,82$$

Il nous reste maintenant à comparer le khi carré théorique issu de la table (7,82) avec le khi-carré calculé (34,11 environ) :

$$\chi^2_{0,05} = 7,82 < \chi^2_{\text{calculé}} = 34,11$$

La règle est la suivante :

- Si le khi-carré calculé est inférieur au khi-carré théorique : indépendance
- Si le khi-carré calculé est supérieur au khi-carré théorique : dépendance

Etant donné que le chi-carré calculé est supérieur au khi carré théorique, nous pouvons conclure que le sexe a une influence sur le choix de la matière. Notre observation initiale sur la base de l'échantillon est donc probablement vraie à l'extérieur de l'échantillon (avec cependant 5% de chances de nous tromper).

Degrés de liberté	P=0,05	P=0,01	P=0,001	Degrés de liberté	P=0,05	P=0,01	P=0,001
1	3.84	6.64	10.83	50	67.51	76.15	86.66
2	5.99	9.21	13.82	51	68.67	77.39	87.97
3	7.82	11.35	16.27	52	69.83	78.62	89.27
4	9.49	13.28	18.47	53	70.99	79.84	90.57
5	11.07	15.09	20.52	54	72.15	81.07	91.88
6	12.59	16.81	22.46	55	73.31	82.29	93.17
7	14.07	18.48	24.32	56	74.47	83.52	94.47
8	15.51	20.09	26.13	57	75.62	84.73	95.75
9	16.92	21.67	27.88	58	76.78	85.95	97.03
10	18.31	23.21	29.59	59	77.93	87.17	98.34
11	19.68	24.73	31.26	60	79.08	88.38	99.62
12	21.03	26.22	32.91	61	80.23	89.59	100.88
13	22.36	27.69	34.53	62	81.38	90.80	102.15
14	23.69	29.14	36.12	63	82.53	92.01	103.46
15	25.00	30.58	37.70	64	83.68	93.22	104.72
16	26.30	32.00	39.25	65	84.82	94.42	105.97
17	27.59	33.41	40.79	66	85.97	95.63	107.26
18	28.87	34.81	42.31	67	87.11	96.83	108.54
19	30.14	36.19	43.82	68	88.25	98.03	109.79
20	31.41	37.57	45.32	69	89.39	99.23	111.06
21	32.67	38.93	46.80	70	90.53	100.42	112.31
22	33.92	40.29	48.27	71	91.67	101.62	113.56
23	35.17	41.64	49.73	72	92.81	102.82	114.84
24	36.42	42.98	51.18	73	93.95	104.01	116.08
25	37.65	44.31	52.62	74	95.08	105.20	117.35
26	38.89	45.64	54.05	75	96.22	106.39	118.60
27	40.11	46.96	55.48	76	97.35	107.58	119.85
28	41.34	48.28	56.89	77	98.49	108.77	121.11
29	42.56	49.59	58.30	78	99.62	109.96	122.36
30	43.77	50.89	59.70	79	100.75	111.15	123.60
31	44.99	52.19	61.10	80	101.88	112.33	124.84
32	46.19	53.49	62.49	81	103.01	113.51	126.09
33	47.40	54.78	63.87	82	104.14	114.70	127.33
34	48.60	56.06	65.25	83	105.27	115.88	128.57
35	49.80	57.34	66.62	84	106.40	117.06	129.80
36	51.00	58.62	67.99	85	107.52	118.24	131.04
37	52.19	59.89	69.35	86	108.65	119.41	132.28
38	53.38	61.16	70.71	87	109.77	120.59	133.51
39	54.57	62.43	72.06	88	110.90	121.77	134.74
40	55.76	63.69	73.41	89	112.02	122.94	135.96
41	56.94	64.95	74.75	90	113.15	124.12	137.19
42	58.12	66.21	76.09	91	114.27	125.29	138.45
43	59.30	67.46	77.42	92	115.39	126.46	139.66
44	60.48	68.71	78.75	93	116.51	127.63	140.90
45	61.66	69.96	80.08	94	117.63	128.80	142.12
46	62.83	71.20	81.40	95	118.75	129.97	143.32
47	64.00	72.44	82.72	96	119.87	131.14	144.55
48	65.17	73.68	84.03	97	120.99	132.31	145.78
49	66.34	74.92	85.35	98	122.11	133.47	146.99
50	67.51	76.15	86.66	99	123.23	134.64	148.21
				100	124.34	135.81	149.48

Source de la table : http://www.apprendre-en-ligne.net/random/tablekhi2.html

Etapes du test d'indépendance du khi-carré

Pour résumer, les principales étapes du test d'indépendance du Khi-carré sont :

1) Si ce n'est pas déjà fait, distribuer la population statistique dans un tableau à deux caractère où les modalités et/ou les valeurs sont regroupées par catégories.

2) Calculer le khi-carré dans l'hypothèse d'indépendance des deux caractères :

$$\chi^2_{\text{calculé}} = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

3) Calculer le nombre de degrés de liberté par la formule :

(Nombre de lignes -1) x (Nombre de colonnes – 1)

4) Définir une probabilité d'erreur (en pratique 5%, 1% ou 1 pour mille)

5) Utiliser le nombre de degrés de liberté et la probabilité d'erreur pour déterminer le khi-carré théorique à partir de la table fournie.

6) comparer la valeur khi-carré calculée avec la valeur khi-carrée théorique (issue de la table) et appliquer la règle suivante :

- Si le khi-carré calculé est inférieur au khi-carré théorique : indépendance

Si le khi-carré calculé est supérieur au khi-carré théorique : dépendance

Voir aussi :

Estimation d'une fonction de demande par la méthode MCO
Coefficient de détermination
Estimation d'une fonction de demande par la méthode MCO
Estimation de la loi d'OKUN par la méthode MCO