LE TEST D'INDEPENDANCE DU KHI-DEUX

Le test d'indépendance du Khi^2 (χ^2) permet de déterminer si deux variables qualitatives sont indépendantes ou non. Pour cela, nous testons les deux hypothèses suivantes :

$$\begin{cases} H_0 : les\ deux\ variables\ sont\ indépendantes \\ H_1 : les\ deux\ variables\ sont\ dépendantes \end{cases}$$

Considérons l'exemple suivant : on interroge 1 000 personnes au sujet de deux caractères, la couleur des yeux (en ligne) et la couleur des cheveux (en colonne). Les résultats obtenus sont indiqués dans le tableau ci-dessous :

	Blond	Brun	Châtain	Roux	**Total**
Bleu	100	50	140	30	**320**
Marron	70	90	O_{23} =170	40	**370**
Noisette	30	40	80	20	**170**
Vert	40	40	50	10	**140**
Total	**240**	**220**	**440**	**100**	**1000**

Ce tableau est appelé **tableau des effectifs observés**[1]. Chaque case du tableau est nommée O_{ij} où l'indice i permet de repérer la ligne (ici $i = 1$ représente la couleur « Bleu », $i = 2$ la couleur « Marron », $i = 3$ la couleur « Noisette » et $i = 4$ la couleur « Vert ») et l'indice j la colonne. Ainsi, $O_{23} = 170$ signifie qu'on a dénombré 170 personnes avec des yeux « Marrons » et des cheveux « Châtains ».

Les totaux en ligne et en colonne sont appelés les « **effectifs marginaux** ». $O_{i.}$ représente le total de la ligne i (par exemple ici $O_{1.} = O_{11} + O_{12} + O_{13} + O_{14} = 100 + 50 + 140 + 30 = 320$). Le « . » signifie donc que l'on effectue une somme toutes colonnes (ou lignes) confondues. De façon générale, si la variable en colonne possède J modalités, on obtient :

$$O_{i.} = \sum_{j=1}^{J} O_{ij}$$

De manière identique, si la variable en ligne possède I modalités, alors $O_{.j}$ représente le total de la colonne j :

$$O_{.j} = \sum_{i=1}^{I} O_{ij}$$

En appliquant la logique de sommation précédente, on obtient par conséquent $O_{..}$ qui représente le total général c'est-à-dire le nombre total d'individus :

$$O_{..} = \sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij}$$

[1] Effectifs observés dans l'échantillon.

De façon générale, s'il y avait indépendance parfaite entre les deux caractères c'est-à-dire que H_0 était vérifiée, les proportions d'yeux bleus, marrons, noisettes ou verts devraient être les mêmes quelle que soit la couleur des cheveux. De même, les proportions de cheveux blonds, bruns, châtains ou roux devraient être invariables pour chaque couleur d'yeux. Ainsi, l'indépendance parfaite devrait elle se traduire dans le tableau par une proportionnalité des lignes entre elles et des colonnes entre elles.

Ainsi, dans cet exemple, $37\ \%\ (370/1\,000)$ des individus ont les yeux de couleur marron tandis qu'ils sont $22\ \%\ (220/1\,000)$ à avoir les cheveux bruns. D'après ce que nous venons de voir, si l'hypothèse d'indépendance entre les deux facteurs était vérifiée, la proportion de personnes possédant à la fois des yeux marrons et des cheveux bruns devrait être de :

$$37\ \% * 22\ \% = 8.14\ \%$$

Sachant qu'il y a dans notre échantillon 1 000 personnes, cela correspondrait donc à un effectif « théorique » de $8.14\ \% * 1\,000 = 81.4$ personnes (à comparer aux 90 personnes observées dans l'échantillon). Le calcul complet est donc le suivant :

$$T_{22} = \frac{370}{1000} * \frac{220}{1000} * 1000 = \frac{370 * 220}{1000} = 81.4$$

Cette logique permet de construire un tableau dont les effectifs – théoriques – représenteraient l'indépendance parfaite : il suffit de porter dans chaque case du tableau le produit du total en ligne et du total en colonne divisé par le total général :

$$T_{ij} = \frac{O_{i.} * O_{.j}}{O_{..}}$$

Le tableau ainsi obtenu sera appelé **tableau des effectifs théoriques**. Dans l'exemple précédent, on obtient (valeurs arrondies au dixième) :

	Blond	Brun	Châtain	Roux	**Total**
Bleu	76.8	70.4	140.8	32.0	**320**
Marron	88.8	$T_{22} = 81.4$	162.8	37.0	**370**
Noisette	40.8	37.4	74.8	17.0	**170**
Vert	33.6	30.8	61.6	14.0	**140**
Total	**240**	**220**	**440**	**100**	**1000**

Vous noterez que, compte tenu de la construction, les totaux sont les mêmes dans le tableau des effectifs observés et dans celui des effectifs théoriques.

Puisque le tableau des effectifs théoriques a été calculé en supposant que H_0 est vraie, plus ce tableau et celui des effectifs observés seront « proches » et moins on aura tendance à rejeter H_0. Dans ce cas, on considérera que les caractères sont indépendants[2]. Il faut donc mettre au point une façon de mesurer la « distance » entre les deux tableaux et une valeur de référence permettant de décider si elle est « faible » ou

[2] Plus précisément, on considèrera que l'on n'a pas suffisamment d'informations pour pouvoir rejeter l'hypothèse d'indépendance.

« forte ». Pour chaque cellule des deux tableaux, on va s'intéresser à la différence entre les effectifs observés et les effectifs théoriques correspondants c'est-à-dire que l'on va calculer $(O_{ij} - T_{ij})$. Comme nous devons avoir un indicateur global, on va ensuite faire la somme de ces différences pour toutes les cases du tableau.

Deux remarques avant de procéder au calcul :

1) il faut rendre toutes les différences positives afin que les écarts positifs et négatifs ne se compensent pas ; on y parvient en élevant les différences $(O_{ij} - T_{ij})$ au carré[3].
2) une différence de 1 entre 10 et 11 n'a pas la même importance qu'entre 1 000 et 1 001 ; en conséquence, il faut utiliser la notion de « différence relative », c'est-à-dire que l'on va diviser la différence par l'effectif théorique correspondant.

Compte tenu de ces remarques, la statistique du Khi^2 calculée que l'on va noter χ_c^2 est telle que :

$$\chi_c^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

Ce qui donne le calcul suivant pour l'exemple :

$$\chi_c^2 = \frac{(100 - 76.8)^2}{76.8} + \frac{(50 - 70.4)^2}{70.4} + \cdots + \frac{(10 - 14.0)^2}{14.0} = 29.725$$

En observant la formule du χ_c^2 on constate aisément que sa valeur minimale est 0 puisqu'il s'agit d'une somme de carrés divisés par des effectifs et que ces valeurs sont nécessairement positives. Cette valeur nulle correspond au cas où $O_{ij} = T_{ij}\ \forall\, i, j$ c'est-à-dire que les effectifs observés et théoriques sont rigoureusement identiques quelle que soit la ligne et la colonne. Si tel est le cas, il est évident que H_0 est vérifiée.

Cette situation d'égalité parfaite étant très particulière, on démontre de façon plus générale que si les deux caractères sont indépendants, la statistique χ_c^2 suit une loi du χ^2. De même que la loi normale est définie par deux paramètres – sa moyenne et son écart-type – la loi du χ^2 possède un argument qui est le nombre de degrés de liberté. Dans le cas qui nous intéresse, ce nombre de degrés de liberté[4], noté ddl, est déterminé comme suit :

$$ddl = (Nombre\ de\ colonnes\ du\ tableau - 1) \times (Nombre\ de\ lignes\ du\ tableau - 1)$$

Nous venons de voir que l'hypothèse H_0 est parfaitement vérifiée lorsque la statistique χ_c^2 est nulle. De façon plus générale, pour rejeter – ou pas – l'hypothèse d'indépendance, on compare la valeur calculée du test à la valeur tabulée qui sera au seuil α[5] le fractile d'ordre $1 - \alpha$ d'une loi du χ^2 dont le nombre de ddl

---

[3] L'utilisation des valeurs absolues serait également possible mais ne permettrait pas les développements à venir.

[4] Techniquement, le nombre de degré de liberté correspond au nombre de cases du tableau dont on peut « fixer librement » les valeurs pour des effectifs marginaux donnés.

[5] α représente le risque de rejeter H_0 alors qu'elle est vraie (**risque de première espèce**). Par conséquent, α doit être

a été déterminé comme indiqué précédemment. Si la valeur calculée par le test est supérieure à la valeur tabulée, on rejette l'hypothèse d'indépendance H_0 au profit de celle de dépendance H_1. Dans le cas contraire, on ne rejette[6] pas H_0.

Les logiciels de statistique fournissent une autre possibilité – équivalent à la précédente – de lire les résultats du test. Pour cela, ils associent aux tests statistiques ce que l'on nomme une « **probabilité critique** » qui correspond au seuil α' telle que :

$$\chi^2_{1-\alpha'}(ddl) = \chi^2_c$$

On rejette alors H_0 si la probabilité est inférieure au seuil usuel de $5\ \%$.

Dans l'exemple, $\chi^2_c = 29.725$ et le fractile (cf. annexe pour consulter la table[7] de la loi du χ^2) pour un seuil α de $5\ \%$ vaut $\chi^2_{0.95}(ddl = 9) = 16.919$. Dans la mesure où $\chi^2_c > \chi^2_{0.95}(ddl = 9)$, on rejette H_0 et on conclut qu'il existe un lien de dépendance entre la couleur des cheveux et la couleur des yeux. De même, la probabilité critique associée à ce test est égale à $0.05\ \%$ valeur qui est largement inférieure au seuil de $5\ \%$. La conclusion est bien évidemment identique.

Remarque importante :

Pour mener un test du Khi^2 de manière correcte, il faut que les effectifs soient suffisamment grands. On exigera notamment que les effectifs théoriques soient tous strictement supérieurs[8] à 5. Si ce n'est pas le cas, il faudra procéder à des regroupements (« logiques ») de lignes et/ou de colonnes dans le tableau des effectifs observés. Si de tels regroupements sont effectués, ils modifieront également le tableau des effectifs théoriques ainsi que le nombre de ddl de la loi du Khi^2 utilisée pour conclure le test.

« faible ». Il est donc usuel de retenir pour α un seuil de $5\ \%$ ou de $1\ \%$ (position plus conservatrice). Cela dit, plus le seuil est faible et moins on rejette H_0. Par conséquent, il est possible sur des données réelles de tolérer un seuil de $10\ \%$.

[6] Vous noterez que H_0 n'est pas rejetée ce qui ne veut pas dire pour autant qu'elle est acceptée. En effet, à côté du risque de première espèce, il existe un **risque de seconde espèce** qui représente la probabilité d'accepter H_0 alors qu'elle est fausse. On notera que les erreurs de première et de seconde espèce évoluent en sens contraire, minimiser l'une conduit à maximiser l'autre. Il n'existe pas de lien simple entre les deux types d'erreur.

[7] Sous Excel, les fonctions permettant de calculer le fractile d'ordre $1-\alpha$ d'une loi du χ^2 et la probabilité critique sont respectivement : $\chi^2_{1-\alpha}(ddl) = LOI.KHIDEUX.INVERSE(1-a; ddl)$ et $\alpha' = 1 - LOI.KHIDEUX.N(\chi^2_c; ddl; VRAI)$.

[8] Les effectifs théoriques apparaissant au dénominateur de la statistique du Khi^2 calculé, une valeur faible d'un ou plusieurs effectifs théoriques conduit à des valeurs fortes de χ^2_c et donc à rejeter H_0 fréquemment.

Annexe : table de la loi du Khi^2.

$1-\alpha$ / n	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995
1	0.708	1.074	1.642	2.706	3.841	5.024	6.635	7.879
2	1.833	2.408	3.219	4.605	5.991	7.378	9.210	10.597
3	2.946	3.665	4.642	6.251	7.815	9.348	11.345	12.838
4	4.045	4.878	5.989	7.779	9.488	11.143	13.277	14.860
5	5.132	6.064	7.289	9.236	11.070	12.833	15.086	16.750
6	6.211	7.231	8.558	10.645	12.592	14.449	16.812	18.548
7	7.283	8.383	9.803	12.017	14.067	16.013	18.475	20.278
8	8.351	9.524	11.030	13.362	15.507	17.535	20.090	21.955
9	9.414	10.656	12.242	14.684	**16.919**	19.023	21.666	23.589
10	10.473	11.781	13.442	15.987	18.307	20.483	23.209	25.188
11	11.530	12.899	14.631	17.275	19.675	21.920	24.725	26.757
12	12.584	14.011	15.812	18.549	21.026	23.337	26.217	28.300
13	13.636	15.119	16.985	19.812	22.362	24.736	27.688	29.819
14	14.685	16.222	18.151	21.064	23.685	26.119	29.141	31.319
15	15.733	17.322	19.311	22.307	24.996	27.488	30.578	32.801
16	16.780	18.418	20.465	23.542	26.296	28.845	32.000	34.267
17	17.824	19.511	21.615	24.769	27.587	30.191	33.409	35.718
18	18.868	20.601	22.760	25.989	28.869	31.526	34.805	37.156
19	19.910	21.689	23.900	27.204	30.144	32.852	36.191	38.582
20	20.951	22.775	25.038	28.412	31.410	34.170	37.566	39.997
21	21.991	23.858	26.171	29.615	32.671	35.479	38.932	41.401
22	23.031	24.939	27.301	30.813	33.924	36.781	40.289	42.796
23	24.069	26.018	28.429	32.007	35.172	38.076	41.638	44.181
24	25.106	27.096	29.553	33.196	36.415	39.364	42.980	45.559
25	26.143	28.172	30.675	34.382	37.652	40.646	44.314	46.928
26	27.179	29.246	31.795	35.563	38.885	41.923	45.642	48.290
27	28.214	30.319	32.912	36.741	40.113	43.195	46.963	49.645
28	29.249	31.391	34.027	37.916	41.337	44.461	48.278	50.993
29	30.283	32.461	35.139	39.087	42.557	45.722	49.588	52.336
30	31.316	33.530	36.250	40.256	43.773	46.979	50.892	53.672
40	41.622	44.165	47.269	51.805	55.758	59.342	63.691	66.766
50	51.892	54.723	58.164	63.167	67.505	71.420	76.154	79.490
60	62.135	65.227	68.972	74.397	79.082	83.298	88.379	91.952
70	72.358	75.689	79.715	85.527	90.531	95.023	100.425	104.215
80	82.566	86.120	90.405	96.578	101.879	106.629	112.329	116.321
90	92.761	96.524	101.054	107.565	113.145	118.136	124.116	128.299
100	102.946	106.906	111.667	118.498	124.342	129.561	135.807	140.169