REVUE DE STATISTIQUE APPLIQUÉE

A. VESSEREAU

Sur les conditions d'application du critérium χ^2 de Pearson

Revue de statistique appliquée, tome 6, nº 2 (1958), p. 83-96

<http://www.numdam.org/item?id=RSA_1958__6_2_83_0>

© Société française de statistique, 1958, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

SUR LES CONDITIONS D'APPLICATION DU CRITÉRIUM X^2 DE PEARSON

par

A. VESSEREAU
Ingénieur en chef des manufactures de l'Etat

On sait que le "Critérium χ^2" de Pearson permet de tester l'accord entre une "distribution observée" et une "distribution théorique".

Cette dernière est définie par les probabilités

$$p_1, p_2 \ldots p_i \ldots p_k \quad (\sum p_i = 1)$$

attachées aux k états d'une grandeur aléatoire, aux k valeurs d'une variable discrète, ou encore à k classes (égales ou inégales) découpant le champ de variabilité d'une variable continue. On supposera, dans tout ce qui suit, que cette loi théorique est "entièrement spécifiée" : les probabilités p_i sont des nombres donnés, qui n'empruntent rien aux observations.

La distribution observée comporte les effectifs :

$$n_1, n_2 \ldots n_i \ldots n_k \quad (\sum n_i = N)$$

Si elle constitue bien "un échantillon" prélevé au hasard dans la distribution théorique (N tirages indépendants et non exhaustifs au sein d'une urne à k catégories de boules en proportions $p_1, p_2 \ldots p_k$) la quantité aléatoire :

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

(qu'on désigne par X^2 pour éviter les confusions) suit approximativement la loi de χ^2 à (k - 1) degrés de liberté - c'est-à-dire la loi de la somme des carrés de (k - 1) variables normales réduites indépendantes.

Lorsque N augmente indéfiniment, la loi de X^2 converge vers cette loi de χ^2. Mais sur la rapidité de la convergence -d'où dépendent les conditions d'application pratique du "Critérium χ^2 de Pearson"- les opinions sont assez diverses. Nous citerons quelques exemples qui ont été relevés dans des ouvrages français ou étrangers.

"L'approximation est très bonne, et la distribution de χ^2 peut être appliquée avec confiance lorsque les effectifs théoriques (produits Np_i) sont au moins de 20."

"Les conditions de validité sont les suivantes :

- aucune des probabilités p_i n'est trop voisine de 0 ou de 1,
- Np_i n'est pas trop petit, pas inférieur à 5 ou même 10."

"On admet généralement que le nombre des individus d'une classe doit être au minimum de 5, et que l'effectif total de l'échantillon ne doit pas être inférieur à 50".

"Lorsqu'on applique le critérium de Pearson, on limite ordinairement le minimum des fréquences absolues pour une classe à 5 unités".

La question a fait l'objet d'études de Sukhatme, Haldane, Neyman et Pearson, Cochran. Dans un important article paru en 1954 dans Biometrics (vol. 10, N°4) Cochran estime que la condition d'un minimum de 10 ou 5 unités par classe est trop sévère. "C'est une opinion, écrit-il, car des recherches suffisantes n'ont pas encore été faites pour rendre la situation parfaitement claire". Cochran donne la recommandation pratique suivante dans le cas d'une distribution unimodale (telle que la loi normale ou la loi de Poisson) : grouper les classes de façon qu'à chaque extrémité l'effectif théorique soit au moins de 1.

Nous avons tenté d'apporter quelques éclaircissements supplémentaires sur cette question, et pour cela nous avons utilisé trois voies d'approche : (1) Méthode des moments ; (2) loi exacte de X^2 dans quelques cas simples ; (3) changements de variables.

I. MÉTHODE DES MOMENTS

S'il n'est pas possible d'obtenir l'expression mathématique de la loi de X^2 dans le cas général, on peut tout au moins calculer les premiers moments de cette variable et les comparer aux moments de la variable χ^2. On sait que, pour (k - 1) degrés de liberté, les quatre premiers moments de χ^2 sont :

Espérance mathématique	μ_1'	$= k - 1$
Moment centré du 2e ordre (variance)	μ_2	$= 2(k - 1)$
Moment centré du 3e ordre	μ_3	$= 8(k - 1)$
Moment centré du 4e ordre	μ_4	$= 12(k - 1)(k + 3)$

Remarquons immédiatement que les variables X^2 et χ^2 ont la même espérance mathématique. En effet :

$$E(X^2) = \sum_i E\left[\frac{(n_i - Np_i)^2}{Np_i}\right] = \sum \frac{Np_i(1 - p_i)}{Np_i} = \sum_i (1 - p_i) = k - 1$$

Les lois de X^2 et de χ^2 coïncident donc toujours, sans restrictions, en ce qui concerne leurs valeurs moyennes.

Le calcul des moments centrés est beaucoup plus laborieux. Nous l'avons entrepris et mené à bonne fin jusqu'au 4e ordre. Peu après, consultant la bibliographie, nous avons constaté que J. B. S. Haldane avait, dès 1937 -par une méthode différente de la nôtre- effectué le même travail. Celui-ci a été publié dans Biometrika, volume XXIX, parts I et II, Juin 1937, sous le titre "The exact values of the moments of the distribution of χ^2 used as a test of goodness of fit when expectations are small".

Nous passerons complètement sur le détail des calculs qui sont relativement simples pour la variance, mais singulièrement fastidieux pour les moments du 3e et surtout du 4e ordre. On obtient finalement les expressions suivantes :

TABLEAU I - Moments de $X^2 = \sum_{i=1}^{i=k} \frac{(n_i - Np_i)^2}{Np_i}$

$$\mu_1' = k - 1$$

$$\mu_2 = 2(k-1) - \frac{1}{N}\left[k^2 + 2k - 2 - \sum \frac{1}{p_i}\right]$$

$$\mu_3 = 8(k-1) - \frac{1}{N}\left[18k^2 + 36k - 32 - 22\sum\frac{1}{p_i}\right] + \frac{1}{N^2}\left[2k^3 + 18k^2 + 28k - 24 - (3k + 22)\sum\frac{1}{p_i} + \sum\frac{1}{p_i^2}\right]$$

$$\mu_4 = 12(k-1)(k+3) - \frac{12}{N}\left[k^3 + 25k^2 + 44k - 32 - (k + 31)\sum\frac{1}{p_i}\right] + \frac{1}{N^2}\left[3k^4 + 132k^3 + 984k^2 + 1.464k - 1.140 - (6k^2 + 236k + 1.316)\sum\frac{1}{p_i} + 3\left(\sum\frac{1}{p_i}\right)^2 + 112\sum\frac{1}{p_i^2}\right] - \frac{1}{N^3}\left[6k^4 + 120k^3 + 696k^2 + 960k - 720 - (12k^2 + 224k + 944)\sum\frac{1}{p_i} + 3\left(\sum\frac{1}{p_i}\right)^2 + (4k + 112)\sum\frac{1}{p_i^2} - \sum\frac{1}{p_i^3}\right].$$

On constate que les moments centrés de X^2 diffèrent des moments de χ^2 par des termes en $1/N$, $1/N^2$ ou $1/N^3$ qui dépendent :

pour la variance, de k (nombre de classes), et $\sum \frac{1}{p_i}$

pour le moment du 3e ordre, de k (nombre de classes), $\sum \frac{1}{p_i}$ et $\sum \frac{1}{p_i^2}$

pour le moment du 4e ordre, de k (nombre de classes), $\sum \frac{1}{p_i}$, $\sum \frac{1}{p_i^2}$ et $\sum \frac{1}{p_i^3}$.

On peut effectuer sur X^2 une transformation simple, de la forme $X^2 + A$, ou MX^2, A et M étant choisis de façon que les deux premiers moments de la variable transformée (μ_1' et μ_2) coïncident avec les deux premiers moments d'une variable χ^2 (dont le nombre de degrés de liberté sera modifié, et cessera généralement d'être un entier, ce qui ne constitue pas un inconvénient grave). Nous reviendrons sur ce point au § III.

En conservant la valeur habituelle de X^2, on constate, à partir des expressions ci-dessus, que l'étroitesse de l'accord entre la loi de X^2 et celle de χ^2 à (k - 1) degrés de liberté -jugée par la comparaison de leurs quatre premiers moments- dépend :

- de l'importance de l'échantillon (N) ;
- du nombre de classes (k) ;
- de la loi théorique - $\left(\text{par les expressions} \sum\frac{1}{p_i}, \sum\frac{1}{p_i^2}, \sum\frac{1}{p_i^3}, \ldots\right)$.

Il ne semble donc pas possible d'énoncer une règle unique et simple in-

diquant dans quelles conditions l'approximation à la loi de χ^2 est satisfaisante ou acceptable.

Dans chaque cas particulier, on peut essayer de juger de ce degré d'approximation par la comparaison des moments exacts de X^2 et de ceux du χ^2 approprié. Il ne semble pas, cependant, que cette méthode soit vraiment intéressante. Outre que les calculs numériques seront fastidieux, on se trouvera embarassé pour décider si l'écart entre les valeurs observées peut être considéré comme acceptable ou non.

Enfin la concordance approximative des moments ne constitue pas un critère suffisant puisque la loi de X^2 est discontinue tandis que celle de χ^2 est continue.

II. LOI EXACTE DE X^2 DANS QUELQUES CAS SIMPLES

Il est théoriquement possible d'établir la loi exacte de X^2 lorsqu'on se donne les valeurs des p_i et l'importance N de l'échantillon. Toutefois, les calculs ne sont abordables que dans des cas très particuliers. Nous les avons effectués dans les cas suivants :

- nombre de classes k allant de 2 à 10 ;
- probabilités p_i toutes égales, donc égales à I/k, les effectifs théoriques étant $Np_i = N/k = n$.

Remarquons que les expressions des quatre premiers moments de X^2 prennent alors des formes assez simples.

TABLEAU 2 - Moments de $X^2 = \sum_{i=1}^{i=k} \frac{(n_i - Np_i)^2}{Np_i}$ quand $p_i = \frac{1}{k}$

$$\mu_1' = k - 1$$

$$\mu_2 = 2(k - 1)(1 - 1/N)$$

$$\mu_3 = 8(k - 1)\left[1 + \frac{k - 8}{2N} - \frac{k - 6}{2N^2}\right]$$

$$\mu_4 = 12(k - 1)(k + 3)\left[1 + \frac{2(3k - 19)}{N(k + 3)} + \frac{2k^2 - 81k + 285}{3N^2(k + 3)} - \frac{2(k^2 - 30k + 90)}{3N^3(k + 3)}\right]$$

La loi exacte de X^2 s'obtient à partir des calculs suivants :

Soit $(n_1, n_2, \ldots n_k)$ -avec $\Sigma n_i = N$- l'une des distributions observées. Il lui correspond une probabilité :

$$\frac{N!}{n_1!\, n_2! \ldots n_k!}\left(\frac{1}{k}\right)^N = \left(\frac{1}{k}\right)^N F$$

et une valeur :

$$X^2 = \sum_{i=1}^{i=k} \frac{(n_i - n)^2}{n} = -N + \frac{\Sigma n_i^2}{n}$$

La même probabilité et la même valeur de X^2 peuvent s'obtenir un nombre de fois égal à :

$$R = \frac{A_N^r}{s_1!\ s_2!\ \ldots\ s_i!\ \ldots},$$

r désignant le nombre d'effectifs n_i non nuls, A_N^r le nombre d'arrangements de N objets r à r, s_i le nombre de fois que l'effectif n_i se trouve répété dans le groupe $(n_1, n_2 \ldots n_k)$.

Enumérant toutes les distributions observées qui diffèrent par la valeur d'au moins un effectif, quelle que soit la classe où les effectifs sont placés, on dresse le tableau suivant :

Effectifs $n_1, n_2 \ldots n_k$	R	F	$P^e = (1/k)^N$ RF	Valeur de X^2
- - -	- - -	- - -	- - -	- - -
- - -	- - -	- - -	- - -	- - -

On additionne enfin les probabilités correspondant à une même valeur de X^2. La distribution exacte de X^2 a ainsi été établie dans les cas suivants :

k	n	N
2	3 à 10	6 à 20
3	2 - 8	6 - 24
4	2 - 5	8 - 20
5	1 - 3	5 - 15
6	1 - 3	6 - 18
7	1 - 2	7 - 14
8	1 - 2	8 - 16
9	1	9
10	1	10

A titre d'exemple, nous donnons ci-dessous (tableau 3) la distribution de X^2 pour (k = 5, n = 3), (k = 8, n = 2), (k = 10, n = 1).

Sur le graphique qui suit, on a tracé, pour ces mêmes valeurs de k et n, la courbe en escalier représentant, en fonction des valeurs de X^2, la probabilité que ces valeurs soient atteintes ou dépassées. Le même graphique montre la courbe continue correspondante pour la loi de χ^2 à (k - 1) dégrés de liberté. On constate que la courbe continue traverse correctement la ligne en escalier. L'accord entre les deux lignes est particulièrement étroit pour les faibles valeurs de la probabilité- celles utilisées dans les tests- et il est naturellement d'autant plus étroit que n est plus élevé.

La même constatation a été faite dans tous les cas étudiés. -Elle se trouve résumée dans le tableau 4 qui s'interprète de la façon suivante :

Loi de $X^2 = \sum_{i=1}^{i=k} \frac{(n_i - Np_i)^2}{Np_i}$ avec $p_i = \frac{1}{k}$
A
Loi de X^2 avec k = 5, n = 3, N = 15
Loi de χ^2 à 4 degrés de liberté
B
Loi de X avec k = 8, n 2, N = 16
Loi de χ^2 à 7 degrés de liberté
C
Loi de X^2 avec k 10, n = 1, N = 16
Loi de χ^2 à 9 degrés de liberté
100
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0
A
B
C
4,1%
3,8%
4.0%
5
9.49
10
14.07
15
16.92
X^2
χ^2

TABLEAU 3 - Loi exacte de X^2.

K = 5	n = 3	N = 15	K = 8	n = 2	N = 16	K 10	n = 1	N = 10
X^2	P	ΣP	X^2	P	ΣP	X^2	P	ΣP
0,000	0,00551	-	0	0,00029	-	0	0,00036	-
0,666	0,08266	0,00551	1	0,01084	0,00029	2	0,01633	0,00036
1,333	0,09299	0,08817	2	0,05420	0,01113	4	0,11431	0,01669
2,000	0,13639	0,18116	3	0,07528	0,06533	6	0,21228	0,13100
2,666	0,10126	0,31755	4	0,12717	0,14062	8	0,22386	0,34328
3,333	0,14878	0,41881	5	0,11443	0,26779	10	0,19337	0,56714
4,000	0,03513	0,56759	6	0,13882	0,38222	12	0,08255	0,76051
4,666	0,11531	0,60272	7	0,09463	0,52104	14	0,08573	0,84306
5,333	0,07067	0,71803	8	0,10986	0,61567	16	0,03175	0,92879
6,000	0,04519	0,78870	9	0,06576	0,72553	18	0,01570	0,96054
6,666	0,02745	0,83389	10	0,05819	0,79129	20	0,01016	0,97624
7,333	0,04357	0,86134	11	0,03872	0,84948	22	0,00762	0,98640
8,000	0,02072	0,90491	12	0,03918	0,88820	24	0,00285	0,99402
8,666	0,02763	0,92563	13	0,01666	0,92738	26	0,00139	0,99687
9,333	0,00531	0,95326	14	0,01831	0,94404	28	0,00018	0,99826
10,000	0,00816	0,95857	15	0,01086	0,96235	30	0,00063	0,99844
10,666	0,00779	0,96673	16	0,00721	0,97311	32	0,00072	0,99907
11,333	0,00956	0,97452	17	0,00569	0,98042	34	0,00005	0,99979
12,000	0,00351	0,98408	18	0,00496	0,98611	36	0,00006	0,99984
12,666	0,00337	0,98759	19	0,00425	0,99107	40	0,00000	0,99990
13,333	0,00266	0,99096	20	0,00228	0,99353	42	0,00006	0,99991
14,000	0,00071	0,99362	21	0,00069	0,99581			
14,666	0,00080	0,99433	22	0,00082	0,99650	> 42	0,00003	0,99997
15,333	0,00142	0,99513	23	0,00086	0,99732			
16,000	0,00095	0,99655	24	0,00077	0,99818			
16,666	0,00118	0,99750	25	0,00033	0,99895			
18,000	0,00044	0,99868	26	0,00018	0,99928			
18,666	0,00024	0,99912	27	0,00015	0,99946			
20,666	0,00018	0,99936	28	0,00005	0,99961			
21,333	0,00018	0,99954	29	0,00007	0,99966			
22,000	0,00012	0,99972	30	0,00007	0,99973			
22,666	0,00006	0,99984	31	0,00007	0,99977			
24,000	0,00003	0,99990	32	0,00007	0,99984			
26,666	0,00001	0,99993	33	0,00001	0,99991			
27,333	0,00003	0,99994	34	0,00001	0,99992			
28,000	0,00001	0,99997	35	0,00001	0,99993			
28,666	0,00001	0,99998	> 35	0,00006	0,99994			
> 28,666	0,00001	0,99999						
	1,00000	1,00000		1,00000	1,00000		1,00000	1,00000

TABLEAU 4 - Risques réels en % dans l'application du critérium χ^2 . (Cas particulier $p_i = 1/k$).

	α = 5 %									α = 1 %								
K degrés de liberté k - 1	2 1	3 2	4 3	5 4	6 5	7 6	8 7	9 8	10 9	2 1	3 2	4 3	5 4	6 5	7 6	8 7	9 8	10 9
$\chi^2(\alpha)$	3,84	5,99	7,81	9,49	11,07	12,59	14,07	15,51	16,92	6,64	9,21	11,34	13,28	15,09	16,81	18,47	20,09	21,67
n = 1	-	-	-	3,4	5,9	3,5	2,0	4,0	4,0	-	-	-	0,16	0,40	0,86	1,00	0,85	1,36
n = 2	-	5,4	3,4	4,0	3,3	4,3	3,8	-	-	-	0,41	0,67	0,50	0,86	1,08	0,89	-	-
n = 3	3,1	5,1	5,0	4,1	4,7	-	-	-	-	-	0,29	0,80	0,90	0,97	-	-	-	-
n = 4	7,0	7,0	4,3	-	-	-	-	-	-	0,78	1,16	0,90	-	-	-	-	-	-
n = 5	2,1	4,3	4,2	-	-	-	-	-	-	0,20	0,67	0,68	-	-	-	-	-	-
n = 6	3,9	5,7	-	-	-	-	-	-	-	0,64	1,01	-	-	-	-	-	-	-
n = 7	5,7	5,5	-	-	-	-	-	-	-	1,29	0,69	-	-	-	-	-	-	-
n = 8	7,7	5,0	-	-	-	-	-	-	-	0,42	1,15	-	-	-	-	-	-	-
n = 9	3,1	-	-	-	-	-	-	-	-	0,75	-	-	-	-	-	-	-	-
n = 10	4,1	-	-	-	-	-	-	-	-	1,18	-	-	-	-	-	-	-	-

Supposons que l'on veuille tester la distribution observée au seuil α (α = probabilité de conclure que la distribution théorique ne s'applique pas alors qu'elle s'applique), en adoptant pour X^2 la valeur $\chi^2(\alpha)$ lue dans la table de χ^2 à (k - 1) degrés de liberté. Les distributions de X^2 et χ^2 n'étant pas confondues, le risque réel n'est pas égal à α : c'est ce risque réel qui est indiqué dans le tableau 4 pour α = 5 % et α = 1 %.

Sur ce tableau on a tracé des lignes de séparation qui indiquent à partir de quelle valeur de N il semble que le risque correspondant à 5 % reste compris entre 4 % et 6 %, et le risque correspondant à 1 % entre 0,8 % et 1,2 % ; tout au moins est-il probable que pour N plus élevé, les risques réels ne sortent pas beaucoup de ces intervalles. L'allure générale des lignes semble montrer que, dès que N dépasse 15 ou 20, les risques réels sont suffisamment voisins de 5 % et 1 % pour que l'erreur commise en utilisant la distribution de χ^2 puisse être considérée comme pratiquement sans importance. Cette constatation, qui se traduit par un minimum imposé à N, et non aux Np_i, s'applique, rappelons-le, au cas particulier où les probabilités sont égales dans toutes les classes. Elle n'est pas tellement étonnante, puisque alors la variance de X^2 est :

$$2(k - 1)\ [1 - 1/N]$$

Elle ne diffère de la variance de χ^2 que par le coefficient (1 - 1/N). Par ailleurs, l'effet de la discontinuité de X^2, dans la zone de cette distribution intéressée par les tests, n'aurait pas grande importance pratique.

Ainsi, avec 10 classes d'égale probabilité et 10 observations, une série telle que :

$$4\ 1\ 0\ 0\ 3\ 0\ 1\ 0\ 0\ 1 \qquad (X^2 = 18)$$

est improbable au seuil 5 %, et une série telle que :

$$3\ 0\ 0\ 2\ 0\ 3\ 0\ 0\ 2\ 0 \qquad (X^2 = 16)$$

est déjà très suspecte.

III. CHANGEMENTS DE VARIABLE

On ne peut évidemment généraliser à toute situation les conclusions de l'étude précédente. On observe en effet, dans ce cas particulier, deux circonstances favorables :

- la variance de X^2 égale à $2(k - 1)(1 - 1/N)$ est toujours inférieure à celle du χ^2 utilisé dans le critérium de Pearson,
- la quantité $\Sigma(1/p_i)$ qui intervient dans l'expression générale de la variance et des moments d'ordre supérieur est minimum quand tous les p_i sont égaux.

On peut considérer que, pour un nombre fini d'observations, le test par χ^2 à (k - 1) degrés de liberté est inexact pour deux raisons :

1/ La loi de χ^2 est continue alors que celle de X^2 est discontinue. L'étude du cas particulier précédent semble montrer (sans que cela soit effectivement démontré) que l'effet de la discontinuité n'est pas très important, si l'on n'est pas très exigeant sur la valeur exacte du seuil de signification.

2/ A l'exception de la valeur moyenne, les distributions de X^2 et de χ^2 n'ont pas exactement les mêmes moments.

En ce qui concerne ce deuxième point, on peut obtenir l'égalité des deux

premiers moments (moyenne et variance) en prenant une "variable X^2 transformée".

a/ Soit $$X'^2 = X^2 + A \text{ avec } \frac{1}{2N}\left[\sum \frac{1}{p_i} - k^3 - 2k + 2\right].$$

La valeur moyenne de X'^2 est $d' = (k-1)+A$ et sa variance $2d' = 2(k-1)+2A$

La loi de X'^2 coïncide par ses deux premiers moments avec une loi de χ^2 à d' degrés de liberté. Tous les moments centrés de X'^2 sont naturellement les mêmes que ceux de X^2, mais ils doivent être comparés aux moments d'une variable χ^2 dont le nombre de degrés de liberté est modifié.

b/ Soit $$X''^2 = MX^2 \text{ avec } M = 2(k-1)\Big/2(k-1) + \frac{1}{N}\left[\sum \frac{1}{p_i} - k^2 - 2k + 2\right].$$

La valeur moyenne de X''^2 est $d'' = M(k-1)$ et sa variance $2d'' = 2M(k-1)$.

La loi de X''^2 coïncide par ses deux premiers moments avec une loi de χ^2 à d" degrés de liberté. Les moments centrés de X''^2 se déduisent de ceux de X^2 en les multipliant par M^2, M^3, M^4, ... ils doivent être comparés aux moments d'une loi de χ^2 à d" degrés de liberté.

En opérant une de ces transformations -qui fait intervenir une loi de χ^2 ayant un nombre de degrés de liberté non entier- l'ajustement est amélioré et l'on peut penser que le test à partir de la nouvelle loi de χ^2 sera plus exact.

Etude d'un cas particulier

Nous avons étudié l'effet de ces changements de variable dans le cas particulier suivant :

k classes
effectifs théoriques (Np_i) égaux à 1 dans C de ces k classes ;
effectifs théoriques (Np_i) tous égaux -donc égaux à $(N - C)/(k - C)$ dans les $(k - C)$ autres classes.

La variance de X^2 prend alors la forme suivante :

$$\mu_2(X^2) = 2(k-1) + C + \frac{(k-C)^2}{N-C} - \frac{k^2+2k-2}{N}$$

Si N est suffisamment grand pour que les deux derniers termes soient petits par rapport à C, la variance de X^2 dépasse d'environ C unités la variance de χ^2 à $(k - 1)$ degrés de liberté, ce qui peut entraîner une erreur importante dans le test.

Les moments du 3e et 4e ordre prennent les formes suivantes :

$$\mu_3(X^2) = 8(k-1) + 23C + (\text{termes en } 1/N \text{ et } 1/N^2).$$

$$\mu_4(X^2) = 12(k-1)(k+3) + 3C^2 + (12k+373)C + 112 + (\text{termes en } 1/N, 1/N^2, 1/N^3).$$

Ces moments peuvent différer considérablement des moments de χ^2 à $(k - 1)$ degrés de liberté.

Pour continuer l'étude, nous avons fait choix des valeurs particulieres suivantes :

$$k = 10,\ C = 1, 2, \ldots 9,\ N = 20, 50, 100.$$

Par exemple, pour N = 50 et C = 5, la distribution théorique est la suivante :

1 1 1 1 1 9 9 9 9 9

Il ne s'agit donc plus de distributions d'équi-probabilité comme au § II, mais au contraire de distributions fortement ou très fortement "déséquilibrées" (certaines probabilités étant très petites par rapport aux autres).

I. Effet des changements du variable sur les moments

a/ La variance de χ^2 à $(k - 1) = 9$ degrés de liberté est 18. Avec les valeurs numériques choisies, la variance de X^2 varie suivant le cas de 17,36 à 25,83. Les transformations X'^2 et X''^2 annulent cette discordance importante.

b/ Le calcul numérique des moments μ_3 et μ_4 de X^2 conduit à la constatation suivante. Ces moments s'écartent de plus en plus des moments d'une variable χ^2 à 9 degrés de liberté à mesure :

- que le nombre de classes à effectif théorique égal à 1 est plus élevé,
- que le nombre total N d'observations est plus grand.

La différence dans les cas extrêmes (C = 9, N = 100) devient considérable. On peut dire sommairement que l'approximation de la loi de X^2 à une loi de χ^2 à (k - 1) degrés de liberté est d'autant moins bonne que la distribution théorique est plus "déséquilibrée".

c/ Les changements de variable $X'^2 = X^2 + A$, $X''^2 = MX^2$, avec :

$$A = \frac{C}{2} + \frac{(k - C)^2}{2(N - C)} - \frac{k^2 + 2k - 2}{2N},$$

$$M = 2(k - 1) \Big/ \left[2(k - 1) + C + \frac{(k - C)^2}{N - C} - \frac{k^2 + 2k - 2}{N} \right],$$

rapprochent tous les deux les moments du 3e et 4e ordre des moments de la variable χ^2 comportant les degrés de liberté appropriés. On constate (avec les valeurs numériques choisies pour k, C et N), que le rapprochement est meilleur avec la variable X''^2. C'est cette variable que nous retiendrons seule dans ce qui suit.

II. Effet du changement du variable $X''^2 = MX^2$ sur le seuil de signification

Nous admettrons que la variable $X''^2 = MX^2$ (avec l'expression de M donnée ci-dessus) suit la loi de χ^2 à $d'' = M(k - 1)$ degrés de liberté : ce n'est encore qu'une approximation puisque X''^2 et χ^2 ne coïncident que par leur espérance mathématique et leur variance, et que X''^2 est discontinu alors que χ^2 est continu.

a/ avec les valeurs numériques choisies (k = 10), si l'on admet que X^2 est distribué en loi de χ^2 à k - 1 = 9 degrés de liberté, le test conduit à rejeter au seuil 5 % l'hypothèse concernant la loi théorique lorsque X^2 atteint ou dépasse la valeur 16,92.

La loi de χ^2 à d'' degrés de liberté donne, au même seuil de 5 %, une valeur limite pour X''^2, d'où l'on déduit la valeur limite de $X^2 = X''^2/M$. Le tableau 5 contient, dans sa partie gauche, les valeurs limites de X^2 ainsi obtenues, pour les différentes combinaisons (C, N) étudiées. Ces valeurs limites varient de 16,76 à 18,64.

b/ Inversement, à la limite 16,92 pour X^2 correspond pour X''^2 dans la loi de χ^2 à d'' degrés de liberté une probabilité qui n'est plus de 5 %, et qui -sous les réserves faites précédemment- peut être considérée comme le risque réel

encouru lorsqu'on se réfère à la loi de χ^2 à k - 1 = 9 degrés de liberté. Ces probabilités sont données dans la partie droite du tableau 5, pour les différentes combinaisons (C, N) étudiées. Elles varient de 4,8 % à 8,1 %.

TABLEAU 5 - Effet du changement de variable $X''^2 = MX^2$.

C	Valeur limite pour X^2 au seuil 5 % (a)			Probabilité correspondant à X^2 = 16,92 (b)			C
	N = 20	N = 50	N = 100	N = 20	N = 50	N = 100	
1	16,76	16,99	17,06	4,8%	5,1%	5,3%	1
2	16,83	17,14	17,26	4,9	5,4	5,7	2
3	16,92	17,32	17,44	5,0	5,8	6,1	3
4	17,00	17,49	17,66	5,2	6,2	6,5	4
5	17,10	17,66	17,84	5,4	6,5	6,8	5
6	17,20	17,82	18,06	5,6	6,8	7,1	6
7	17,33	18,02	18,26	5,9	7,1	7,5	7
8	17,49	18,21	18,45	6,2	7,4	7,8	8
9	17,65	18,40	18,64	6,5	7,7	8,1	9

Revenant aux valeurs inscrites dans la partie gauche du tableau 5, on constate que la plus élevée 18,64 correspond, dans la loi de χ^2 à 9 degrés de liberté, à une probabilité d'environ 3 %.

Dans les cas particuliers étudiés, et sous les réserves déjà faites (non concordance des moments d'ordre supérieur à 2, discontinuité), on peut donc dire que le risque réel ne dépassera pas 5 % si l'on prend pour valeur limite de X^2 la valeur lue au seuil 3 % dans la table de χ^2 à 9 degrés de liberté.

CONCLUSION

Les trois tentatives que nous avons faites pour essayer de mieux préciser les conditions d'application du critérium χ^2 de Pearson nous conduisent aux conclusions suivantes :

1. La méthode des moments

fournit un moyen très général pour reconnaître si la loi de X^2 est voisine de la loi de χ^2 à (k - 1) degrés de liberté. Mais ce moyen n'est guère utilisable dans la pratique pour les raisons suivantes :

- il nécessite des calculs numériques laborieux,
- on ne saura comment interpréter les écarts constatés entre les moments de X^2 et de χ^2 ,
- la comparaison des moments ne tient pas directement compte de la discontinuité de X^2 .

La conclusion de cette étude est plutôt négative : il ne semble pas qu'on puisse définir les conditions d'application correcte du critérium χ^2 par une règle simple et précise applicable en toute circonstance.

(a) Cette valeur est 16,92 dans la loi de χ^2 à 9 degrés de liberté.

(b) Cette valeur est 5 % dans la loi de χ^2 à 9 degrés de liberté.

II. L'étude d'un cas très particulier (tous les p_i égaux)

montre que, dans certaines circonstances, le test de χ^2 est pratiquement valable même si les effectifs sont petits, jusqu'à être égaux à l'unité dans toutes les classes. Dans ce cas particulier, c'est le nombre total d'observations N qui intervient, beaucoup plus que les effectifs par classe.

D'autre part, cette étude semble montrer que, dans la région de la distribution de X^2 qui est intéressée par les tests habituels (seuils de 5 % ou 1 %) la discontinuité n'introduit pas une erreur très importante.

3/ Après avoir montré que les distributions de X^2 et χ^2 ont tendance à être d'autant plus différentes que la loi théorique (caractérisée par les probabilités p_i) est plus "déséquilibrée", la troisième étude propose un changement de variable faisant coïncider par leurs deux premiers moments (espérance mathématique et variance) la variable transformée et une variable χ^2 dont le nombre de degrès de liberté est modifié.

Cette transformation de variable n'est pas à conseiller dans la pratique, car elle impose des calculs ennuyeux. Mais, compte tenu des constatations résumées en 1/ et 2/, elle suggère finalement une règle un peu grossière, mais simple, qui est la suivante :

- lorsque tous les effectifs théoriques sont au moins égaux à quelques unités, appliquer le critérium χ^2 sous sa forme usuelle,
- lorsque certains effectifs sont petits -de l'ordre de l'unité- considérer que la valeur limite pour X^2 au seuil 5 % est celle que l'on trouve dans la table de χ^2 à (k - 1) degrés de liberté au seuil 2 % ou 3 %.

A notre avis, il n'y a pas lieu d'être extrêmement exigeant sur la valeur exacte du seuil de probabilité : il nous parait pratiquement sans importance, dans la plupart des cas, que le seuil réel soit compris entre 4 % et 6 %, alors qu'on se figure qu'il est exactement de 5 %.

On peut se demander s'il n'est pas plus simple de procéder à des "groupements", de sorte que les classes, suffisamment garnies, permettent d'effectuer le test χ^2 sous sa forme usuelle. En fait, cette façon d'opérer peut masquer des différences essentielles.

On a constaté au § II que la série observée :

4 1 0 0 0 1 0 3 1 0

diffère significativement (au seuil 5 %) de la série théorique

1 1 1 1 1 1 1 1 1 1

Si l'on fait des groupements (en admettant qu'ils aient une base logique) de façon que les effectifs théoriques soient au moins de 5, il ne reste que 2 classes, et les distributions théorique et observée deviennent identiques (5 - 5).

Faire participer les petits effectifs au test -quitte à perdre un peu sur la rigueur de celui-ci- peut être très utile pour examiner ce qui se passe aux "queues" des distributions.

Il arrive par exemple qu'une distribution s'ajuste bien à une loi normale dans sa partie centrale, mais que cela ne soit plus vrai aux extrémités. Le groupement des classes peut alors faire disparaître arbitrairement des anomalies qui sont redoutables lorsque les valeurs correspondantes de la variable interviennent dans des questions de sécurité ou de santé.

SUMMARY

on the application of Pearson's fest

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

In order to see what happens at the "tails" of distributions, it may be useful to introduce the small theorical values in Pearson's χ^2 test, even if it means losing some of the strictness of this test.

It happens, for example, that a distribution is well ajusted to a normal law in its central part, but not in its extremities. Grouping some classes may then make disappear, in an arbitrary way, anomalies which are very dangerous when the corresponding values interfere in security or health matters.

The three attempts we have made in trying to define the conditions of application of Pearson's χ^2 test lead us to the following conclusions.

(1) The method of moments gives a very general method of recognizing if the law of X^2 is close to the law of χ^2 with (K - 1) degrees of freedom. But this method is not very practical because it entails rather tedious numerical calculations, we do not know how the variations found out between moments of X^2 and moments of χ^2 should be interpreted, comparison of moments does not take directly into account the discontinuity of X^2 .

The conclusion of this study is rather negative : it does not seem possible to define the conditions of a correct application of the χ^2 test with an easy and accurate rule applicable to every case.

(2) The study of a very particular case (when all the p_i are equal) shows that under certain circumstances, the test of χ^2 is of practical value, even if the numbers are small, even so small as to equal one in each class. In that particular case, it is the total amount of observations N which matters rather than the numbers in each class.

Besides, this study seems to prove that in the area of the distribution of X^2 which is covered by the usual tests (levels of 5 or 1 %), the discontinuity does not introduce a very important error.

(3) After having pointed out that the distributions of X^2 and χ^2 tend to be the more different as the theorical law (characterized by probabilities p_i) is more "unbalanced", the third study proposes a transformation of the variable making the transformed variable and a χ^2 variable, of which the number of degrees of freedom is modified, coincide by their first two moments (expected value and variance).

This transformation is not to be recommended in practice, for it imposes tedious calculations. But even taking into account the statements summarized in (1) and (2), it suggests finally a rule, rather broad but easy, which is :

When all theorical numbers are at least equal to a few unities, apply test of χ^2 under its usual form ; when some of them are small -ranging about unity- consider that the limit value for X^2 at the level of 5 % is the value found in the table of χ^2 with (K - 1) degrees of freedom at the level of 2 or 3 per cent.