REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Note sur le χ^2 de Karl Pearson

Revue de statistique appliquée, tome 16, n° 3 (1968), p. 65-74

<http://www.numdam.org/item?id=RSA_1968__16_3_65_0>

NOTE SUR LE χ^2 DE KARL PEARSON

par P. THIONET

Faculté des Sciences de Poitiers

1 - INTRODUCTION

On désignera par (p) la distribution multinomiale ($p_1\ p_2\ ..\ p_N$) autrement dit (p_j) ; (q), q_j auront une signification semblable. L'échantillon a pour taille n ; (n) désigne sa composition (n_1, n_2, .. n_N) suivant les N classes-j ; (n) est le résultat de n épreuves indépendantes multinomiales, relatives à une loi inconnue.

Faisant l'hypothèse H° que la loi inconnue est (p), on calcule une sorte de distance entre (n) et (p) appelée χ^2 de Karl Pearson, -ou statistique de Pearson, - que nous désignerons plus précisément par χ^2_p, à savoir :

$$\chi^2_p = \sum_j \frac{(n_j - n\,p_j)^2}{n\,p_j}$$

autrement dit :

$$\sum_j n_j^2/n\,p_j - n\ ;$$

on posera :

$$n_j - n\,p_j = x_j .\ \sqrt{n}$$

d'où :

$$\chi^2_p = \sum_j x_j^2/p_j$$

De la même façon, avec $n_j - n\,q_j = y_j\ \sqrt{n}$, on posera :

$$Y^2_q = \sum_j y_j^2/q_j\ ,$$

puis :

$$\chi^2_q = \sum_j x_j^2/q_j \quad , \quad Y^2_p = \sum_j y_j/p_j\ .$$

On désignera par (G) l'hypothèse suivant laquelle les n_j sont assimilables à des variables de Laplace-Gauss, ce qui veut dire que l'échantillon n'a pas une taille trop petite et que chaque classe-j a une probabilité (p_j ou q_j) assez grande. Avec les hypothèses jointes H° et G, on sait (depuis le début du siècle) que χ^2_p est une variable aléatoire khi-carré à N-1 degrés de liberté. On trouve la Table du khi-carré dans tous les livres de statistique ; le rapprochement de la valeur numérique de la statistique avec la ligne υ = N-1 de la Table conduit à rejeter H° quand χ^2_p parait trop grand (: correspond à une probabilité d'être dépassé α (disons) de 5 % ou de 1 %.

On reproche à ce test d'être peu puissant, c'est-à-dire de ne pas faire rejeter H° dans bien des cas où la loi multinomiale inconnue n'est pas (p) mais (q). Pour remédier à cet inconvénient divers auteurs ont modifié le test du khi-carré : citons Neyman [1] 1937, Cochran [2] 1954, 1955, Evelyn Fix, Hodges et Lehmann [3] 1959.

D'autres auteurs ont adapté à la distribution multinomiale le test le plus puissant, c'est-à-dire (d'après un résultat général de Neyman et Egon Pearson) le ratio des vraisemblances. C'est notamment le cas de Lerman, dont le test n'est applicable que si n est petit [4] 1967.

La présente note, qui suppose n assez grand, essaie de combiner les deux points de vue ; elle était annoncée à la fin de [4].

2 - RAPPELS

On sait que χ^2 (khi-carré) est la somme de υ carrés de variables de Laplace-Gauss normées et indépendantes ; d'où il suit que :

$$E(\chi^2) = \upsilon \quad , \quad V(\chi^2) = 2\,\upsilon$$

Sous l'hypothèse (p), autrement dit H°, on sait que :

$$E\,x_j = 0, \quad E\,x_j^2 = V\,x_j = p'_j\,p_j, \quad E\,x_i\,x_j = -\,p_i\,p_j \ . \ (p'_j = 1 - p_j)$$

D'où :

$$E\,\chi_p^2 = \sum_j p_j\ (1 - p_j)/p_j = N - 1$$

De même

$$x_j = y_j + \sqrt{n}\,(q_j - p_j) = y_j + a_j$$

en posant :

$$a_j = \sqrt{n}\,(q_j - p_j)$$

soit encore :

$$D_p = \sum a_j^2/p_j$$

$$Y_p^2 = X_p^2 - 2\,S + D_p$$

avec

$$S = \sum_j a_j\ x_j/p_j$$

d'où

$$E\ S = 0$$

$$E\ Y_p^2 = E\ X_p^2 + D_p = N\text{-}1 + D_p$$

Sous l'hypothèse (q), on échangera les lettres p et q, x et y.

Ajoutons l'hypothèse G à (p) (ou H°).

Proposition : X_p^2 est une variable khi-carré et Y_p^2 une variable khi-carré décentrée dont D_p est le paramètre de décentrage, N-1 étant le nombre de degrés de liberté des deux variables. De même, sous les hypothèses (q) et G, Y_q^2 est un khi-carré et X_q^2 un khi-carré décentré, avec les paramètres D_q et N-1.

Explications

On trouve la démonstration relative au khi-carré dans les manuels : voir Fourgeaud et Fuchs, Statistique, (Dunod 1967) p. 304.

Celle relative au khi-carré décentré, χ_c^2, qui fait défaut, serait parallèle. Tout d'abord :

χ_c^2 est la somme de υ carrés de variables de Laplace-Gauss réduites mais non toutes centrées, indépendantes. Le centre C de la distribution dans R^{υ} n'est plus l'origine 0 ; supposons les axes orthonormés, faisons-les tourner pour que l'un d'eux passe par C : il vient

$$\chi_c^2 = (z + b)^2 + \chi'^2$$

en désignant par χ'^2 un khi-carré centré, indépendant de z, et ayant υ-1 degrés de liberté. C'est dire que χ_c^2 dépend du seul paramètre de décentrage b^2 (ou b) et des υ degrés de liberté.

Il résulte immédiatement de cela que, dans le plan d équation $\sum x_j = 0$, le point (y) de coordonnées $y_j = (x_j - a_j)$ a une distribution de Gauss de centre C $(-a_j)$; on vérifiera que $\sum a_j = 0$. Si les p_j sont égaux (à 1/N), le résultat sur y^2 p est évident. Au contraire, avec des p_j inégaux, le fait parait moins net. On se ramène alors au premier cas en posant :

$$x'_j = x_j / \sqrt{p_j} \quad , \quad y'_j = y_j / \sqrt{p_j} \quad , \quad a'_j = a_j / \sqrt{p_j}$$

avec un plan d'équation :

$$\sum_j x'_j \sqrt{p_j} = 0.$$

Nous pouvons donc accepter la Proposition avancée.

3 - ALTERNATIVE H' à H°

On ne peut définir la puissance d'un test (et lui reprocher de ne pas être puissant) si l'on omet de préciser quelle alternative H' est opposée à H°. Si l'alternative est une hypothèse (q) entièrement spécifiée, on dit qu'on a un test d'hypothèses simples ; c'est le cas retenu dans [4].

Il est certain que χ_p^2 ne fait pas intervenir explicitement une alternative (q). Faut-il en conclure que H' désigne tout (q) autre que (p) ? aussi voisin de (p) qu'on veut ? A ce compte là, tout test serait mauvais.

Commençons donc par nous fixer le risque de 1ère espèce α, que le test χ_p^2 fasse rejeter (p), alors que (p) est vrai.

$$\text{Prob} \left[\sum_j x_j^2/p_j > \chi^2 \right] = \alpha \ , \ \sum_j x_j = 0$$

Ceci définit un ellipsoïde à l'intérieur duquel doit se trouver le point (x_j) (et un autre pour le point (n_j))

La figure se trouve réduite à un ellipsoïde (P) et au point (x) qui polarise les alternatives, créant une direction privilégiée. En particulier, si $q_j = n_j/n \ \forall_j$, on a $y_j = 0$ et $Y_p^2 = Y_q^2 = 0$

Le point observation est tombé au centre de la distribution alternative.

Puissance

Sous l'hypothèse (q) et G, χ^2_p n'a pas une distribution simple ; c'est seulement en confondant les p_j et les q_j qu'on peut y voir un khi-carré décentré. On se reportera au Tableau suivant :

Vraie loi multinomiale :		(p)	(q)
Loi asymptotique	χ^2	χ^2_p	Y^2_q
	χ^2 décentré	Y^2_p	χ^2_q

En somme, la théorie classique est doublement asymptotique :

- les x_j ou y_j sont supposés finis ;
- ils sont de l'ordre de leurs écarts-types, soit $\sqrt{p_j}$ et $\sqrt{q_j}$ (si N n'est pas petit, le facteur $\sqrt{1-p_j}$ ou $\sqrt{1-q_j}$ ne modifie pas les ordres de grandeur) ;
- ils sont donc de l'ordre de $1/\sqrt{N}$;
- les p_j et q_j sont de l'ordre de $1/N$;
- posons : $\varepsilon_j = \frac{q_j - p_j}{p_j}$, donc : $\frac{1}{p_j} = \frac{1 + \varepsilon_j}{q_j}$

Assimiler χ^2_p à un khi-carré décentré équivaut à tenir ε_j pour négligeable à côté de 1 ($\forall_j$). Les $q_j - p_j$ sont donc beaucoup plus petits que les p_j ou q_j.

C'est dire que la théorie n'est valable que dans un cas sans intérêt, où les deux ellipsoïdes (P) et (Q) ont leurs centres presque confondus. Dans les cas réalistes, la puissance du test khi-carré est inconnue.

4 - TEST LE PLUS PUISSANT

Contre une alternative bien spécifiée, le Lemme de Neyman et Pearson (Egon) conduit à la statistique :

$$\lambda = \text{Log}\,(V(p\ ;\ n)/V(q\ ;\ n)) = \sum_j n_j \text{ Log } (p_j/q_j)$$

Avec les définitions précédentes*:

$$\gamma_j = \text{Log}\,(p_j/q_j) = -\text{ Log }(1 + \varepsilon_j) = -\varepsilon_j + \frac{\varepsilon_j^2}{2} + \ldots$$

Sous l'hypothèse G : λ est (asymptotiquement) variable de Laplace-Gauss ; c'est (à un facteur près) la projection sur l'axe (γ) du point $(n_j/\sqrt{n})$ qui admet une distribution (dégénérée) de Gauss, soit dans l'hypothèse (p), soit dans l'hypothèse (q).

---

* Nota : $\gamma_j \# -\varepsilon_j = \frac{p_j - q_j}{p_j}$. Mais on peut aussi approcher les γ_j par $(p_j^2 - q_j^2)/2\,p_j\,q_j$.

Rapprochement entre λ et les statistiques khi-carré

Sous les hypothèses G et (p), il vient :

$$\lambda = n \sum_j p_j \gamma_j + \sqrt{n} \sum_j \gamma_j x_j$$

et, avec l'approximation

$$\gamma_j = -\varepsilon_j = (p_j - q_j)/p_j ,$$

$$\lambda = \text{constante} - S \quad , \quad \text{où} \quad S = \sum_j a_j x_j / p_j .$$

D'où $2\lambda = \text{constante} + Y_p^2 - X_p^2$

De même, sous les hypothèses G et (q), on aura intérêt à écrire

$$2\lambda = \text{constante} + Y_q^2 - X_q^2$$

en remplaçant γ_j par $(p_j - q_j)/q_j$.

5 - LE TEST S

S est très proche du test le plus puissant.

S a une distribution de Laplace-Gauss, avec $ES = 0$, $VS = D_p$, sous G + (p).

Si les a_j sont multipliés par un même facteur, il en est de même de la variable S et de son écart-type ; la loi réduite de S reste invariante. C'est seulement la puissance de S qui s'améliore à mesure que (q) s'éloigne de (p). Celle-ci est donnée par la loi de S sous G + (q) ; c'est une seconde loi de Laplace-Gauss, avec $ES = D_p$, et :

$$VS = \sum_j a_j^2 q_j / p_j - \left(\sum_j a_j^2 q_j / p_j \right)^2$$

Nota : Si l'on avait respecté la symétrie entre les p_j et les q_j qui existait dans les γ_j, la première variance n'aurait pas été aussi simple.

Emploi du test S

Supposons la seconde variance du même ordre que la première ; pour avoir $\alpha = 5\,\%$ environ et β un peu plus grand, D_p devrait être de l'ordre de 3 écarts-types, soit :

$$\sigma^2 = 3\sigma , \quad \text{d'où} \quad \sigma^2 = 9 = D_p ,$$

ou bien 16 pour 4 écarts-types. Mais D_p et la moyenne quadratique des ε_j, soit ε, sont liés par :

$$D_p = \sum a_j^2 / p_j = n \sum (p_j - q_j)^2 / p_j = n \varepsilon^2$$

Conséquence :

$$\varepsilon = 0,1 \Longrightarrow n = 900 \text{ (ou } 1600)$$

$$\varepsilon = 0,5 \Longrightarrow n = 36 \text{ (ou } 64)$$

c'est-à-dire de l'ordre de 1000 dans le 1er cas, et de 50 dans le second.

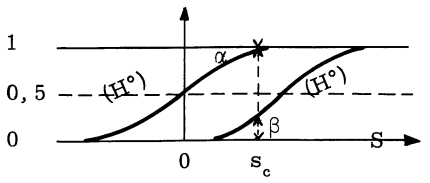

Fig. 1

Interprétation

Le vecteur $(q_j - p_j)$ est projeté sur $(n_j/n - p_j)$ non pas orthogonalement mais suivant 2 directions conjuguées :

L'équation :

$$\sum_j \frac{a_j \, x_j}{p_j} = 0$$

définit des directions (a), (x), conjuguées, par rapport à l'ellipsoïde (P) du point (p). Le test S sera d'autant meilleur que les points (p), (q) et (n) seront mieux alignés, - que les vecteurs (a) et (x) seront davantage colinéaires.

S'ils le sont rigoureusement, S est proportionnel à D_p ; S^2/VS est proportionnel à D_p.

1/ Cas où les p_j sont égaux (à 1/N) : il vient :

$$S^2 / D_p = \frac{\left(\sum a_j \, x_j\right)^2}{(1/N) \sum a_j^2}$$

contre $\chi^2 = N \sum x_j^2$; avec $\sum x_j = \sum a_j = 0$

Si les vecteurs (a) et (x) sont colinéaires, leur corrélation est de carré unité ; et les 2 statistiques sont rigoureusement égales ; mais S est significative bien avant X^2, comme on le voit, en comparant les 2 lignes de la Table du χ^2 à 1 et à N-1 degrés de liberté.

Exemple :

		Point α = 5 %
υ	= 1	3, 84
	= 10	18, 31

Il est clair que le contraire se produira quand (a) se sera beaucoup écarté de la direction privilégiée (x).

Remarque : λ donnerait des résultats comparables à ceux de S.

Exemple : N = 10, $p_j = 1/10$. n = 100 , $\sqrt{n} = 10$.

	100 p_j	=	10	10	10	10	10	10	10	10	10	10
	100 q_j	=	8	9	10	11	12	13	11	10	8	8
	a_j	=	-0,2	-0,1	0	0,1	0,2	0,3	0,1	0	-0,2	-0,2
$D_p = 2,8$	n_j	=	6	8	11	0	14	15	12	9	6	8
	10 x_j	=	-4	-2	+1	-1	+4	+5	+2	-1	-4	-2
	100 x_j^2	=	16	4	1	1	16	25	4	1	16	4
	10 a_j	=	-2	-1	0	1	2	3	1	0	-2	-2
	100 $a_j x_j$	=	+8	+2	0	-1	+8	+15	+2	0	+8	+4

On trouve :

$$X_p^2 = 10 \sum x_j^2 = 8,8$$

qui n'est pas significatif ; avec 9 degrés de liberté les seuils de signification du khi-carré sont :

α =	5 %	2,5 %	1 %
χ^2 =	16,92	19,01	21,67

En revanche le test S est significatif :

$$S = 10 \sum a_j \, x_j = 4,6$$

La variance de S étant $D_p = 2,8$, son écart-type est 1,67.

Sa valeur normée est donc : 4,6/1,67 = 2,76 correspondant à α = 0,3 %.

Conclusion

D'après Karl Pearson, on peut accepter l'hypothèse (p). Mais, d'après Neyman et Egon Pearson, on doit rejeter (p) pour lui préférer (q).

2/ Cas où les p_i sont inégaux : l'inégalité de Schwarz s'étend facilement :

On écrira que le trinôme en t, $\sum (x_j \, t - a_j)^2/p_j$, est positif, nul si et seulement si les x et les a sont proportionnels. D'où :

$$S^2 / D_p < X_p^2$$

avec égalité stricte quand les vecteurs x et a sont colinéaires. Alors le test S possède sur le test X^2 toute sa supériorité.

6 - REGIONS CRITIQUES DES TESTS :

Bien entendu nous ne sommes pas le premier à découvrir que deux tests peuvent donner des résultats différents avec les mêmes données (n). Ceci s'explique assez bien ici par les régions critiques des tests.

Pour celui de Karl Pearson, on a vu qu'il s'agit de l'extérieur d'un

certain ellipsoïde ; alors que pour λ ou S, il s'agit du domaine (ouvert) situé d'un certain côté d'un plan.

Nous raisonnerons dans le cas où $p_j = 1/N$, $\forall$ j, de façon que χ^2 soit le carré d'une distance euclidienne. Mais cette restriction n'est pas essentielle.

Le khi-carré χ^2 à N-1 degrés de liberté sera la somme d'un khi-carré à 1 degré de liberté (c'est S^2/D_p) et d'un autre, indépendant (orthogonal) à N-2 degrés de liberté. Dans le plan (à 2 dimensions), les régions critiques sont délimitées :

- pour χ^2 par des cercles de centre 0.
- pour S, par des parallèles (fig. 2).

α =	5 %	2,5 %	1 %	0,1 %
1 d.l.				
χ^2 =	3.84	5.02	6.63	10.83
χ =	1.96	2.24	2.57	3.29
10 d.l.				
χ^2 =	18.31	20.48	23.21	29.59
χ =	4.28	4.53	4.53	5.44
cos θ =	0.458	0.494	0.534	0.605
θ =	62°45	60°20	57°45	52°45

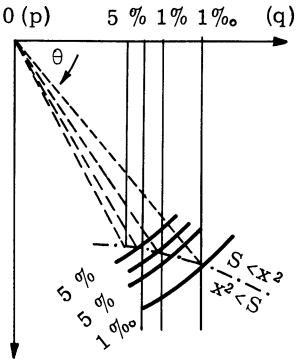

Fig. 2

Le calcul ci-dessus donne la position des intersections cercles - droites verticales, pour un même α ; autrement dit les intersections des régions critiques des 2 tests χ^2 et S ayant même mesure. On a ainsi 4 points de la frontière entre la zône où S est meilleur que χ^2, et la zône où c'est l'inverse.

Remarque - Pour comparer valablement deux tests, il conviendrait sans doute de tenir compte de plusieurs autres éléments, -sans oublier la commodité des calculs. Ce qui suit doit être applicable à des problèmes plus généraux.

Fixons α (risque de 1° espèce) et considérons les régions critiques des deux tests ; sous l'hypothèse H°, elles ont même mesure α (fig. 3). Comme leurs frontières se croisent, les régions hachurées ont même mesure (sous H°) δ.

Si le point échantillon (n) a autant de chances de se trouver dans l'une ou dans l'autre, il apparait cependant qu'il est beaucoup plus fâcheux pour nous de le voir en A qu'en B.

Si (n) est en A : le test χ^2 fait rejeter H°, le test S n'en fait rien. Si (n) se trouve en B, c'est juste le contraire.

Or la région où se trouve B est marginale ; il suffit de prendre (disons) α = 3 % pour le test S, contre (disons) 5 % pour χ^2 pour être assuré de rejeter les mêmes (n) aberrants avec S ou χ^2.

Au contraire la région où se trouve A sera toujours à l'abri de la manoeuvre précédente (si naturelle qu'on la fait inconsciemment). Là S est mauvais et nous préférons X^2 : on peut penser que l'alternative (q) a été mal choisie.

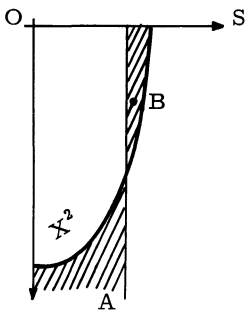

Fig. 3

7 - HYPOTHESES "COMPOSITES"

En étudiant le cas de deux hypothèses simples, (p) et (q), on a constaté que S concernait toute une trajectoire rectiligne d'hypothèses (q) issues de (p) : En s'éloignant de (p) les régions critiques sont conservées et c'est seulement la puissance du test qui s'améliore.

Quant aux orientations des trajectoires (vecteurs a), sont équivalentes celles situées sur un cône de révolution d'axe (p), (n). Bien entendu on se heurte à la doctrine établie suivant laquelle on doit choisir le test sans connaitre les données.

8 - RESUME

Il s'agit d'un prolongement de l'article "Sur certains succédanés du test de Neyman et Pearson" [4]. On désignera par (p) et (q) des lois multinomiales. On rappelle les définitions du X^2 de K. Pearson, du χ^2 et du χ_c^2 (décentré) et la convergence (en loi) de X^2 vers χ^2. On étudie la puissance du test de X^2 : on montre que, sous une alternative à l'hypothèse zéro, ce n'est pas la statistique $X^2 = X_p^2$ mais une autre, X_q^2, qui converge vers un χ^2. Les confondre n'est valable que si l'alternative (q) diffère très peu de H° ou (p).

On donne le test asymptotiquement le plus puissant et un test S qui lui ressemble beaucoup et semble plus simple. S est défini pour une alternative donnée (q), mais vaut pour toute une trajectoire linéaire d'alternatives issues de (p). On compare S à X^2 : on établit la supériorité de S si la trajectoire d'alternatives passe par le point observation. On constate que la comparaison reste à l'avantage de S dans toute une zone de positions de (q). L'avantage revient à X^2 dans une autre zône. On esquisse (sur un exemple) le tracé de la ligne de démarcation entre les zones.

REFERENCES

[1] J. NEYMAN - Smooth test for goodness of fit, Skandin. Aktuar. 20 (1937), 150-199.

J. NEYMAN - Contribution to the theory of χ test, Proceedings first Berkeley Symposium 1949, 239-273.

[2] W.G. COCHRAN - Some methods for strengthening the common χ^2 test. Biometrics 10 (1954) 417-451.

W.G. COCHRAN - A test of a linear function of the deviations between observed and expected numbers J.A.S.A. 50 (1955) 377-397.

[3] Evelyn FIX, J.L. HODGES et E.L. LEHMANN - The restricted chi-square test, The Harald Cramer Volume, 1959, 91-107.

[4] P. THIONET - Sur certains succédanés du test de Neyman et Pearson, Revue de Statistique Appliquée, 15, (1967) n° 2, 19-38.