REVUE DE STATISTIQUE APPLIQUÉE

W. G. COCHRAN

Analyse des classifications d'ordre

Revue de statistique appliquée, tome 14, n° 2 (1966), p. 5-17

<http://www.numdam.org/item?id=RSA_1966__14_2_5_0>

ANALYSE DES CLASSIFICATIONS D'ORDRE

W. G. COCHRAN
Professeur de Statistique, Harvard University, U.S.A.

Quand les données ont la forme d'une classification par degrés, il est souvent raisonnable de considérer que la classification représente une espèce de groupement d'une échelle de mesurage continue. On peut donc assigner à chaque classe un numéro, qui exprime notre concept de la position de la classe dans l'échelle. Ces numéros créent une variable qui suit une distribution discrète, et permet l'analyse des données au moyen des méthodes développées pour les distributions discrètes et, comme approximation, par les méthodes communes pour les distributions continues.

L'avantage de ce procédé est que les méthodes d'analyse des variables discrètes et continues sont plus flexibles et puissantes que les méthodes qu'on a pour les données qualitatives. L'exposé en présente plusieurs illustrations. Un désavantage est que le procédé est plus ou moins subjectif. Mais on trouve que des différences modérées entre les variables assignées par deux investigateurs font généralement peu de différence dans les conclusions.

L'étude décrit plusieurs méthodes moins subjectives qui ont été utilisés pour construire les numéros : comparaison avec une mesure continue qu'on considère plus précise, usage d'une population standard dans laquelle la variable associée a une distribution normale, et construction d'une échelle qui aura certaines propriétés désirées et raisonnables.

Enfin, il y a deux méthodes objectives - la méthode maximin de Abelson et Tukey, et un χ^2 test de Bartholomew en utilisant le test du rapport des vraisemblances. Cette dernière methode ne construit pas une variable discrète, mais semble être puissante si le but de l'analyse est un test de signification.

1 - INTRODUCTION

Les méthodes statistiques sont bien adaptées pour l'analyse de deux espèces d'observations. En premier lieu, on trouve les méthodes fondées sur la distribution binomiale, qui servent pour les observations classées dans l'une ou l'autre de deux catégories, comme par exemple les objets approuvés ou rejetés dans une inspection. Pour les caractères quantitatifs, comme la résistance électrique d'un fil, il y a les méthodes développées d'après la loi normale de Gauss-Laplace - l'analyse de variance, la régression, etc.

Quelquefois, cependant, les données ont la forme d'une classification par degrés, peut-être par degrés de gravité ou de préférence. Par exemple, on pourrait décrire les défauts trouvés dans un article comme : mineur, majeur ou critique. Il est évident que dans cette classification, un défaut critique est plus grave qu'un défaut majeur, et un défaut majeur est plus grave qu'un défaut mineur. De même, dans une enquête sur les accidents industriels, on pourrait décrire le degré de préjudice subi par un individu comme léger, modéré, sévère ou fatal. Dans une enquête de marché, une maison de commerce donne deux articles de toilette A et B à chaque ménagère prise dans un échantillon. Après une période d'usage, chaque ménagère exprime sa préférence en choisissant une des cinq classes : préfère nettement A, préfère A, indifférente, préfère B, préfère nettement B.

Jusqu'à présent, les méthodes les plus efficaces pour l'analyse des observations de cette espèce n'ont pas été étudiées à fond. Je veux d'abord exposer une méthode, simple et assez ancienne, qui suffit en pratique pour beaucoup de problèmes. Dans les sections 5 et 6, je décris plusieurs approches alternatives.

Ma première illustration provient d'une expérience sur l'usage des médicaments dans le traitement des lépreux. Au début de l'expérience, il y avait deux types de malades, A et B. Au cours de l'expérience, qui dura presqu'un an, les deux types reçevaient le même traitement médical. A la fin, un expert de la lèpre a examiné soigneusement chaque malade. Il a noté le progrès du malade pendant l'année, et l'a classé dans une des cinq catégories montrées dans la Table 1. La classe la plus favorable est "Amélioration importante", tandis que "Aggravation" représente la situation la moins favorable.

Table 1

Deux groupes de lépreux, classés suivant le degré d'amélioration

Degré d'amélioration	Groupe A	Groupe B	Total	Distribution A	Distribution %	Distribution B
Amélioration importante	11	7	7	8		14
moyenne	27	15	15	19		29
faible	42	16	16	29		31
nulle	53	13	13	37		25
Aggravation	11	1	1	8		2
Total	144	52	52	100		100

$\chi^2 = 6,88$ (4 d.d.l) $P = 0,16$

Conclusion : Pas de différence entre les deux groupes.

La Table 1 à gauche donne les résultats. Le but de l'analyse est de déterminer s'il y a une différence entre les degrés moyens d'amélioration montrés par les deux types de malades.

2 - L'EMPLOI DE χ^2

Le procédé le plus connu pour tester l'hypothèse que les deux dis-

tributions A et B sont les mêmes, à l'exception des erreurs d'échantillonnage, est le test χ^2, ou :

$$\chi^2 = \sum \frac{(\text{Observés} - \text{Théoriques})}{\text{Théorique}}$$

La somme s'étend sur les 10 cellules de la Table 1 (5 lignes et 2 colonnes). Les nombres théoriques sont les nombres de malades qu'on trouverait si les deux distributions étaient réellement les mêmes. Par exemple, on attendrait (18) (144)/196 = 13.0 malades de type A avec amélioration importante.

La valeur de χ^2 est 6.88, avec 4 d.d.l. Si l'hypothèse nulle est vraie, la probabilité d'obtenir une valeur de χ^2 aussi grande est 0.16. Selon le test χ^2, on dirait que les distributions, et par conséquent les degrés moyens d'amélioration, ne sont pas différents entre A et B.

Mais le χ^2 test n'est pas approprié dans ce cas, car le test ne spécifie pas une hypothèse alternative particulière, et ne tient pas compte du fait que nous avons une classification par degrés de gravité. Si on change l'ordre des classes dans Table 1, la valeur de χ^2 reste la même. La classification représente une tentative de mesure de la quantité d'amélioration de chaque malade. Notre hypothèse alternative est que, en moyenne, le degré de progrès diffère entre les deux types A et B. Si cette hypothèse alternative est vraie, nous nous attendons à trouver que les malades d'un type montrent une plus grande proportion de cas classés "Amélioration importante" ou "Amélioration moyenne" et une plus petite proportion classée "Amélioration nulle" ou "Aggravation".

La Table 1, à droite, donne ces proportions. Notez que le type B contient 43 % de malades dans les deux meilleures classes, contre 27 % pour le type A. Dans les deux classes qui ne présentent pas d'amélioration, le type B contient 27 % de malades, et le type A 45 %. Ces résultats suggèrent que les malades de type B ont fait plus de progrès que ceux de type A.

3 - VARIABLES ASSOCIEES

La méthode que je veux exposer se base sur le point de vue que

1/ si les instruments de mesure étaient assez bons, on pourrait mesurer le progrès de chaque malade sur une échelle continue,

2/ la classification faite par l'expert représente une espèce de groupement de cette échelle continue en 5 classes.

De ce point de vue, il semble raisonnable d'assigner à chaque classe un numéro, qui exprime notre concept de la position de la classe dans l'échelle. Ces numéros créent une variable, associée à la classification, et qui suit une distribution discrète. Ayant construit la variable associée, on peut analyser les données au moyen des méthodes développées pour les distributions discrètes. Il n'est pas nécessaire de faire une hypothèse, peu réaliste, que l'expert a groupé sans erreurs. En pratique, il y aura des cas où la mesure idéale placerait le malade dans la deuxième catégorie, mais où l'expert l'a placé dans la première. En anglais, les numéros qui produisent la variable associée s'appellent généralement "scores".

Comment assigner les numéros ? Ceci dépend de l'information qu'on a concernant la construction des classes, et également des buts de l'enquête. Cette question sera discutée plus tard. Dans l'exemple étudié actuellement, j'ai affecté le numéro 0 à la classe "Amélioration nulle", et les numéros 1 et 2 aux classes "Amélioration faible" et "Amélioration moyenne", puisque l'expert semblait considérer ces classes comme également séparées. Ma première idée était d'assigner le numéro $\underline{4}$ à la classe "Amélioration importante" et le numéro $\underline{-2}$ à la classe "Aggravation", parce que l'expert a examiné un malade plus soigneusement et plus longtemps avant de la placer dans une de ces deux classes extrêmes. Mais j'ai décidé que, peut-être, ceci n'était qu'une impression subjective, et j'ai donné les numéros 3 et -1.

Lorsque les numéros sont assignés, on a deux échantillons de la variable associée. On teste la signification de la différence entre les moyennes des deux échantillons par le test t de Student. La Table 2 contient le calcul, le symbole x représentant la variable associée. La méthode est la méthode élémentaire d'application du test t. La valeur de t est 2,616, avec une probabilité un peu moindre qu'une chance sur cent. On peut donc en conclure que les malades de type B ont fait, en moyenne, plus de progrès que ceux de type A.

Table 2

Analyse avec variables associées

Degré d'amélioration	x	Fréquences A f	Fréquences B f		A	B
Amélioration importante	3	11	7	Σf	144	52
moyenne	2	27	15	Σfx	118	66
faible	1	42	16	$\bar{x}$	0,819	1,269
nulle	0	53	13	Σfx^2	260,0	140,0
Aggravation	-1	11	1	$(\Sigma fx)^2/\Sigma f$	96,7	83,8
				$\Sigma f(x-\bar{x})^2$	163,3	56,2
		144	52	d. d. l.	143	51
				s^2	1,150	1,102

$$\bar{s}^2 = 219,5/194 = 1,131$$

$$t = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{s^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{0,450}{\sqrt{(1,131)\left(\frac{1}{144} + \frac{1}{52}\right)}} = 2,616$$

$$P = 0,009$$

Conclusion : le Groupe B montre plus d'amélioration, en moyenne, que le Groupe A.

Ce calcul s'étend facilement à la comparaison de plus de deux groupes A, B, C, D, ... On fait d'abord un test F (test de Snedecor) de l'hypothèse nulle que les moyennes de tous les groupes sont égales. Si le test montre qu'il y a des différences, on peut ensuite faire des tests t plus détaillés. Si les groupes représentent des niveaux croissants

de quelque ingrédient, on peut obtenir la courbe de régression de la moyenne de chaque groupe sur les niveaux. Par exemple, on pourrait préparer quatre plats de pommes de terre, contenant quatre quantités différentes de sel. Chaque sujet goûte tous les plats et les classe : "Beaucoup trop de sel", "Trop de sel", "Assez de sel" etc. Le but est d'obtenir la relation entre la quantité de sel et la réponse du sujet.

On peut avoir une double classification rangée par degrés. Si on étudie la relation entre les conditions de travail des ouvriers dans une usine et la qualité de leur production, les conditions sont classées par exemple comme supérieure, satisfaisante, mauvaise, et la qualité comme excellente, bonne, pauvre. Dans ce cas on assigne deux variables associées, y et x. L'analyse de données de cette espèce a été présentée par Yates (1948) [8]. Dans certains cas il y a plus d'intérêt à comparer la variabilité des groupes qu'à comparer leurs moyennes. L'exemple dans la Table 3 est tiré des résultats d'une ancienne enquête conduite dans les écoles de Londres. On a essayé de classer selon leur intelligence cent garçons et cent filles pris dans les mêmes écoles.

Table 3

Intelligence des élèves à Londres

	Supérieure	au-dessus de la moyenne	Moyenne	au-dessous de la moyenne	Faible	Total
Garçons	7	22	48	19	4	100
Filles	5	18	58	16	3	100
	2	1	0	-1	-2	

$s_g^2 = 76,9$ $s_f^2 = 62,4$

$F = 76,9/62,4 = 1,23$, d. d. l. = (99,99)

La table ne suggère pas une différence entre le degré moyen d'intelligence des garçons et celui des filles. Mais il semble que les filles sont moins variables entre elles que les garçons. Notez que 58 % des filles ont l'intelligence moyenne, contre 48 % des garçons.

Si on peut assigner une variable associée aux classes d'intelligence, on peut calculer la variance de la mesure d'intelligence pour les garçons et pour les filles. Avec les numéros que j'ai donnés, on trouve $s_g^2 = 76.9$ (garçons) : $s_f^2 = 62.4$ (filles). La valeur de F est $76.9/62.4 = 1,23$, avec 99 et 99 d. d. l. Puisque cette valeur n'est pas significative au niveau 5 % de probabilité, la variabilité supérieure qu'on constate pour les garçons peut être expliquée par les erreurs d'échantillonnage.

4 - CRITIQUE DE LA METHODE

Comme nous venons de le voir, la construction d'une variable associée a plusieurs avantages. Elle permet l'emploi de tests qui sont plus puissants que le test χ^2, parce qu'on peut les utiliser pour tester le type particulier d'hypothèse alternative qu'on attend dans les observations. En outre, on peut, par exemple, examiner si la différence moyenne

entre le groupe A et le groupe B vaut deux fois la différence moyenne entre le groupe B et le groupe C, ce qui n'est pas possible avec une analyse qualitative. Plus généralement, la méthode est une tentative d'obtenir la plupart des avantages d'une échelle continue de mesure.

La critique principale est, bien entendu, que le procédé est arbitraire. Deux investigateurs peuvent assigner des numéros différents à la même classification. Mais on trouve que dans la majorité des cas, des différences modérées entre les numéros assignés produisent peu de différence entre les conclusions qu'on tire de l'analyse. Par exemple, dans la Table 1 j'ai calculé le test t avec la série de numéros 4, 2, 1, 0, -2 que j'avais d'abord considérée. La valeur de $\underline{t}$ devient 2.46, avec P = 0.015, contre t = 2,62, P = 0.009 produit par la série alternative 3, 2, 1, 0, -1. Les deux analyses mènent à la même conclusion.

Néanmoins, il y a quelquefois de grandes difficultés avec les classes extrêmes. Aux Etats-Unis on a mené beaucoup d'enquêtes sur le degré de dommage souffert par les personnes dans les accidents d'automobiles. Les classes de dommage sont, par exemple, mineur, modéré, sévère, rendant incapable, et fatal. Comment assigner des numéros aux classes : rendant incapable et fatal, afin de les placer dans la même échelle que les autres classes ? De même, un défaut dans une pièce mécanique, tel que la machine ne marche pas, semble être d'un autre degré que celui du défaut qui, quoique gênant, permet à la machine de fonctionner. Dans ces cas, la difficulté peut provenir du fait que la classification ne représente pas une seule échelle. Il faut peut-être deux ou trois variables pour décrire assez exactement la classification.

La variable associée a une distribution discrète non normale : en effet, elle est quelquefois assez dissymétrique. Peut-on vraiment employer les méthodes standard fondées sur la distribution normale, comme le test t et le test F ? Pour examiner cette question dans des cas simples, on peut calculer la distribution exacte de $\underline{t}$ et de $\underline{F}$ par la méthode de "randomization" de Fisher. A mon avis, en pratique, les méthodes normales suffisent généralement, étant entendu que le niveau 5 pour-cent de $\underline{t}$ ou de F trouvé par ces méthodes n'a pas une probabilité exactement égale à 5 %. Il correspond, peut-être, à un niveau entre 4 et 7 pour-cent. Il faut noter aussi deux points :

1/ Les variances peuvent différer d'un groupe à l'autre, spécialement quand les distributions sont biaisées. Si on prépare des pommes de terre avec beaucoup de sel, il peut arriver que tous les sujets les classifient comme "Beaucoup trop de sel", et la variance est nulle, tandis qu'un plat contenant moins de sel peut recevoir tous les degrés de la classification. Aussi, avant de combiner les variances pour obtenir une seule estimation, il est sage de vérifier que les variances ne diffèrent pas, et il faut fréquemment utiliser des méthodes d'analyse qui n'assument pas l'égalité des variances.

2/ Avec des échantillons très petits, c'est-à-dire contenant moins de 20 unités, je suggère l'emploi d'une correction de continuité dans le test t, afin de tenir compte du fait que nous avons une distribution discrète.

5 - AUTRES METHODES POUR CONSTRUIRE LES VARIABLES ASSOCIEES

Jusqu'à présent j'ai donné une seule méthode, principalement sub-

jective, pour assigner les numéros. En fait, différents investigateurs ont utilisé plusieurs méthodes, moins subjectives, pour construire les variables associées. Il n'existe pas beaucoup de comparaisons entre les méthodes, de manière qu'on ne peut juger de leurs avantages et désavantages relatifs. Mais, puisque le champ d'application diffère d'une méthode à l'autre, le manque de comparaison n'est pas trop gênant. Je vais décrire brièvement quelques-unes de ces méthodes.

5. 1 - Comparaison au moyen de mesures plus précises.

Quelquefois il est possible de comparer la classification par degrés de gravité avec une autre mesure du même caractère qu'on considère comme étant plus exacte. Cette possibilité existe quand la classification représente une méthode de mesure rapide et bon marché, tandis que la méthode plus exacte est coûteuse. Par exemple, un observateur qui marche autour d'un arbre peut rapidement classer le dommage causé aux fruits par les insectes, comme nul, léger, modéré, sévère. D'autre part, on peut cueillir le fruit dans un échantillon d'arbres et mesurer le pourcentage de fruits "gâtés ou touchés" sur chaque arbre de l'échantillon par l'inspection des fruits individuels. On peut donc examiner si la classification correspond à un groupement des arbres selon le pourcentage de fruits touchés. Si oui, on peut assigner à chaque classe, comme variable associée, la moyenne des pourcentages pour les arbres qui tombent dans cette classe. S'il y a plusieurs observateurs, il est possible :

1/ de comparer la précision de différents observateurs,

2/ d'examiner si on peut utiliser la même variable associée pour tous les observateurs.

5. 2 - Usage d'une population standard.

Quand on étudie des échantillons qui proviennent de populations différentes, il peut exister une population standard à laquelle on désire comparer les autres populations. La population standard peut représenter la méthode habituelle utilisée pour fabriquer un article, les autres populations représentant des changements de méthode. Supposons :

1/ que nous avons un grand échantillon tiré de la population standard,

2/ qu'il est raisonnable d'admettre que si le caractère pouvait être mesuré exactement, il aurait approximativement une distribution normale dans la population standard.

Dès lors, dans la population standard on regarde la classification comme un groupement de la distribution normale avec moyenne zéro et écart-type 1. On peut facilement déterminer les bornes de chaque classe dans l'échelle normale. La variable associée est la moyenne de la distribution normale entre ces bornes. Cette méthode est ancienne : le statisticien Karl Pearson l'employait beaucoup.

Pour illustrer les calculs j'ai utilisé les données de la Table 1, en combinant les deux groupes de malades pour construire une "population standard" artificielle. La Table 4 donne les détails.

Table 4

Illustration de la méthode de K. Pearson

Degré d'amélioration	% en classes	% cumulés P = F(u)	Borne supérieure de u	Moyenne de u	Variable associée
Aggravation	9	9	-1,34	-1,8	-0,9
Amélioration nulle	21	30	-0,52	-0,9	0
faible	30	60	+0,25	-0,1	0,8
moyenne	34	94	+1,55	+0,8	+1,7
importante	6	100	∞	+2,0	+2,9

On cumule d'abord les pourcentages dans les classes. Ensuite, on trouve les bornes, en usant la table de la distribution normale cumulative. Par exemple, la table normale montre que 9 % de la distribution reste entre $u = -\infty$ et $u = -1,34$. La moyenne de u entre les bornes a et b est $[Z(a) - Z(b)]/P$, ou Z(a) et Z(b) sont les ordonnées de la distribution pour $u = a$ et $u = b$. Enfin, j'ai ajouté 0.9 à chaque valeur de la variable associée, de manière que la valeur devienne zéro pour la classe "Amélioration nulle". Dans cet exemple, la variable associée diffère peu de la variable initialement utilisée dans la Table 1.

Dans une version alternative de cette méthode, due à Bross (1958), [5] la variable associée est la probabilité cumulative jusqu'au milieu de la classe. Les propriétés de la méthode de Bross, qui évite toute supposition de normalité, sont décrites dans la référence donnée [5] .

5.3 - Construction d'une échelle qui aura certaines propriétés désirées.

Plusieurs investigateurs ont suggéré cette approche. Dans l'application décrite par Ipsen (1955) [6] une série de niveaux Z_1 , Z_2 , Z_3... d'un certain traitement ont été appliqués aux sujets, chaque sujet recevant un seul niveau. Plus tard, les sujets sont classés dans des catégories ordonnées qui expriment le degré de succès du traitement. Ayant assigné une variable associée x à la classification, on peut calculer la moyenne $\bar{x}_i$ pour tous les sujets qui ont reçu le niveau Z_i. On peut penser que la variable $\bar{x}_i$ doit avoir une relation linéaire avec la variable Z_i. Ipsen détermine la valeur x_j qu'on donne à la classe j de manière que le variance due à la régression linéaire entre les $\bar{x}_i$ et les Z_i contienne la proportion maximum de la variance totale de x entre les sujets.

La solution est très simple. La valeur de x assignée à la classe j est la moyenne des valeurs de Z_i pour les sujets qui tombent dans cette classe. Donc, la méthode est presque la même que celle décrite dans la section 5.1, si on considère le variable Z comme la mesure la plus exacte. La Table 5 présente un exemple donné par Ipsen. La variable Z représente le niveau d'immunisation contre le typhus, et les classes A, B, C, D, F sont les degrés de gravité de thyphus. Les moyennes $\bar{Z}$, dans la partie inférieure de la table, sont les valeurs de la variable associée. Dans la dernière ligne, Ipsen a transformé les valeurs de manière que leurs bornes soient 0 et 100.

Table 5

Degré de gravité de la maladie de 71 malades atteints de typhus, classés selon le niveau d'immunisation

Niveau d'immunisation (z)	Nombre de malades — Degré de gravité du typhus				
	A	B	C	D	F
1			5	2	3
2		4	8	4	1
3		2	8	1	
4		4	3		
5	11	20	5		
f	11	30	29	7	4
Σfz	55	130	82	13	5
$\bar{z}$	5,0	4,3	2,8	1,9	1,2
valeur associée	0	18	58	84	100

Fisher (1947) [7] a décrit une autre application du même type. Chacun des douze serums a été injecté dans le sang de chacun des douze sujets. On a marqué les réàctions des sujets par les symboles -, ?, faible, (+), et ++(1), qui forment une classification rangée par degré de force ou de puissance. On sait que la force de réaction doit différer entre les sérums et aussi entre les sujets. Dès lors, Fisher a assigné les valeurs 0, x_2, x_3, x_4, et 1 aux différentes classes, et a déterminé les valeurs de x_2, x_3, et x_4 qui rendent maximum le rapport de la somme des carrés entre séra et sujets à la somme totale des carrés dans l'analyse de la variance. Ici, le calcul n'est pas si simple : il faut trouver les racines caractéristiques d'un déterminant.

6 - DEUX METHODES OBJECTIVES

6.1 - La méthode "maximin".

Toutes les méthodes précédentes supposent, en effet, que le chercheur connaît un peu plus que l'ordre simple des classes. Par exemple, avec les lépreux, j'ai jugé que la distance entre les classes "Amélioration nulle" et "Amélioration faible" est approximativement la même que la distance entre "Amélioration faible" et "Amélioration moyenne".

Au contraire, la méthode maximin de Abelson et Tukey (1963, 1959) [2] [1] est une méthode objective qui suppose seulement ce qu'on connaît réellement concernant l'ordre des classes. Ces auteurs supposent qu'il y a une variable associée x_j, vraie mais inconnue, pour laquelle $x_1 \leqslant x_2 \leqslant x_3 \ldots$ Ils excluent le cas $x_1 = x_2 = x_3 \ldots$; c'est-à-dire qu'il y a au moins une inégalité. Si y est la variable qu'on analyse (par exemple, proportion de lépreux dans le groupe A), la relation la plus simple entre y_j et x_j est

$$y = \alpha + \beta x + e$$

---

(1) Le symbole (?) indique une réaction probablement faible.

où la variable $\underline{e}$ est une variable de moyenne nulle, de variance σ^2, la covariance entre e_i et e_j $(i \neq j)$ étant nulle. Pour examiner si y_j change d'une façon monotone quand x_j change, il faut estimer β.

Puisque x_j est inconnu, on construit une variable associée c_j. On peut avoir la convention $\Sigma c_j = 0$. Le numérateur du coefficient de régression de y_j par rapport à c_j est $c = \Sigma c_j y_j$. On en déduit :

$$E(x) = \beta \Sigma c_j x_j = \beta \Sigma c_j (x_j - \bar{x})$$

$$V(c) = \sigma^2 \Sigma c_j^2$$

Si on veut tester la quantité c, pour découvrir si $\beta \neq 0$, la puissance du test est déterminée par $E(c)/\sqrt{V(c)}$.

$$\frac{E(c)}{\sqrt{V(c)}} = \frac{\beta \Sigma c_j (x_j - \bar{x})}{\sigma \sqrt{\Sigma c_j^2}} = \left\{\frac{\beta}{\sigma} \sqrt{\Sigma (x_j - \bar{x})^2}\right\} \left\{\frac{\Sigma c_j (x_j - \bar{x})}{\sqrt{\Sigma c_j^2 \; \Sigma (x_j - \bar{x})^2}}\right\}$$

Le premier terme de l'expression à droite ne dépend pas du choix des c_j. Le second terme est le coefficient de corrélation r entre c_j et x_j. On veut donc rendre r maximum.

Il y a une infinité de valeurs de x_j satisfaisant les relations $x_1 \leqslant x_2 \leqslant x_3 \ldots$ Pour chaque choix des c_j, on peut déterminer les x_j qui rendent $\underline{r}$ minimum. Abelson et Tukey recommandent l'utilisation des c_j pour lesquelles le minimum $\underline{r}$ est maximum. Ainsi le nom "maximin".

Pour 2 jusqu'à 8 classes, la Table 6 donne les valeurs des "maximin" c_j (ajustées pour que les valeurs centrales soient -1, 0, et 1.

Table 6

Valeurs des maximins c_j et de r^2

j	n = 2	n = 3	n = 4	n = 5	n = 6	n = 7	n = 8
1	-1	-1	-6.5	-4.4	-13.0	-8.1	-20.8
2	1	0	-1	-1	- 3.5	-2.4	- 6.4
3		1	1	0	- 1	-1	- 3.2
4			6.5	1	1	0	- 1
5				4.4	3.5	1	1
6					13.0	2.4	3.2
7						8.1	6.4
8							20.8
min r^2	1.000	.750	.651	.596	.557	.530	.510
min r^2 linéaire	1.000	.750	.600	.500	.429	.375	.333
min r^2 lin.-2-4	1.000	.750	.649	.588	.549	.522	.493

L'échelle c_j n'est pas une fonction linéaire de $\underline{j}$. Elle diffère d'une échelle linéaire en ce que les valeurs extrême sont beaucoup plus grandes. Pour n = 8, par exemple, une échelle linéaire aurait les valeurs -7, -5, -3, -1, 1, 3, 5, 7, tandis que les c_j sont -20.8, -6.4, -3.2, -1, 1, 3.2, 6.4, 20.8. En effet, les auteurs notent qu'une échelle

qu'ils appellent "linéaire-2-4" est une bonne approximation des maximin c. Pour construire le "linéaire-2-4", on commence avec une échelle linéaire, et on multiplie les valeurs extrêmes par 4, et les valeurs voisines par 2. Pour n = 8, ceci donne -28, -10, -3, 1, 1, 3, 10, 28. La valeur du maximin r^2 est 0,510, et pour le linéaire-2-4 la valeur minimum de r^2 est 0,493, (valeurs données en dessous de la Table 6). Si le lecteur n'a pas la table des c_j, les auteurs remarquent qu'il peut employer l'échelle linéaire-2-4.

Les auteurs ont montré que les valeurs les plus défavorables de la vraie échelle x_j (les valeurs qui donnent les minima r^2) sont de l'espèce 0, 1, 1, 1, ... ou 0, 0, 1, 1, 1 : c'est-à-dire les valeurs qui ont une seule inégalité. En dessous de la Table 6 on trouve également les r^2 minimum donnés par une variable associée qui est linéaire en j. Pour $n \geqslant 5$, ces r^2 sont notablement plus petits que les r^2 maximin. Si une échelle comme la linéaire-2-4 semble un peu bizarre au lecteur, il faut remarquer que la méthode de Abelson et Tukey fût développée pour le chercheur qui ne connaît rien de plus que $x_1, \leqslant x_2 \leqslant x_3 \ldots$

La méthode s'étend à d'autres types d'information initiale. Par exemple, les auteurs construisent les c_j dans un exemple avec 6 classes, où on sait que

$$x_1 \leqslant x_2 \leqslant x_3 \leqslant x_5 \leqslant x_6 \qquad \text{et} \qquad x_1 \leqslant x_4 \leqslant x_5$$

mais où on ne connaît pas la relation entre x_4 et x_2, x_3. On peut également utiliser la méthode quand on sait quelque chose sur les distances entre les x_j, mais les détails pour ce cas n'ont pas encore été publiés.

6.2 - Un χ^2 test alternatif.

Dans cette méthode, due à Bartholomew (1961, 1959) [4], [3], les conditions sont les mêmes que pour la méthode maximin. Pour les membres de l'échantillon qui tombent dans la classe j, la variable qu'on analyse peut être ou une proportion p_j ou la moyenne $\bar{y}_j$ d'une variable continue. On suppose que $E(p_j) = x_j$ (ou $E(\bar{y}_j) = x_j$), les x_j représentant l'échelle vraie mais inconnue. Le problème posé par Bartholomew est de tester l'hypothèse nulle $x_1 = x_2 = x_3 = \ldots$ contre l'alternative $x_1 \leqslant x_2 \leqslant x_3 \leqslant \ldots$

Il construit un test par le critère du rapport des vraisemblances. Pour illustrer cette méthode, dans le cas le plus simple où $\bar{y}_j$ est la moyenne de n_j observations normales, avec σ^2 connu, nous avons :

$$-2\sigma^2 \ln L = \Sigma n_j (\bar{y}_j - x_j)^2$$

où L représente la vraisemblance. Pour l'hypothèse nulle (H. N.), $x_j = x$, et L est maximum lorsque $x = \bar{y} = \Sigma n_j \bar{y}_j / \Sigma n_j$. Donc :

$$-2\sigma^2 \ln L_{max} (\text{h. N.}) = \Sigma n_j (\bar{y}_j - \bar{y})$$

Pour l'alternative (H.A.), il faut choisir les x_j qui rendent L maximum, compte tenu des relations

$$x_1 \leqslant x_2 \leqslant x_3 \quad \ldots$$

La solution dépend de l'ordre des $\bar{y}_j$. Si

$$\bar{y}_1 \leqslant \bar{y}_2 \leqslant \bar{y}_3 \leqslant \ldots$$

en accord avec l'alternative, la solution est clairement $x_j = \bar{y}_j$, de manière que

$$-2\sigma^2 \ln L_{max}(H.A.) = 0$$

Dans ce cas, le critère pour tester est, ignorant le $2\sigma^2$,

$$[\ln L_{max}(H.A.) - \ln L_{max}(H.N.)] = \Sigma n_j (\bar{y}_j - \bar{y})^2.$$

Mais supposons, avec 3 classes, que

$$\bar{y}_1 \leqslant \bar{y}_3, \quad \bar{y}_2 \leqslant \bar{y}_3, \quad \bar{y}_1 > \bar{y}_2.$$

Bartholomew montre que la solution pour maximiser la H.A. est

$$x_1 = x_2 = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} = \bar{y}_{12} \qquad x_3 = \bar{y}_3$$

Ainsi

$$-2\sigma^2 \ln L_{max}(H.A.) = n_1 (\bar{y}_1 - \bar{y}_{12})^2 + n_2 (\bar{y}_2 - \bar{y}_{12})^2$$

Dans les cas plus compliqués, il donne les règles simples pour trouver la solution et ensuite le critère. En général, on combine les valeurs des $\bar{y}_j$ voisins pour produire une série de valeurs qui s'accordent avec l'hypothèse alternative. Le test exige des tables spéciales, décrites dans la référence 1959 [3].

Bartholomew (1961) [4] a comparé la puissance de son test avec celle due à la méthode maximin. Bien entendu, sa puissance relative dépend des valeurs des vrais x_j. Ni l'une ni l'autre des méthodes n'est supérieure sur toute l'étendue des alternatives. Mais dans les cas les plus défavorables aux maximin c_j, le test de Bartholomew gagne matériellement s'il y a plus de quatre classes. La raison semble être qui le critère du rapport des vraisemblances s'adapte à la nature vraie de l'alternative, tandis que les c_j restent fixes pour un nombre donné de classes. La conclusion pratique est que :

1/ si on sait seulement que $x_1 \leqslant x_2 \leqslant x_3 \ldots$ et

2/ si le but est un test de signification, le test de Bartholomew semble être préférable.

Comme la méthode maximin, ce test s'étend :

1/ aux cas dans lesquels l'information sur l'ordre est incomplète ; par exemple, si on sait que x_4 est le plus grand, mais l'ordre des x_1, x_2, x_3, est inconnu,

2/ aux cas où l'on a une certaine information imprécise sur les distances entre les x_j (ces résultats ne sont pas encore publiés).

J'ai préparé cet article, à la demande du Professeur Vessereau, pour une des Réunions d'Etudes sur les applications de la statistique dans les entreprises, au temps ou j'étais Professeur à l'Institute de Statistique sur invitation du Professeur Dugué de l'Institut. Je désire remercier le Professeur Dugué, le Professeur Vessereau et leurs collègues pour la généreuse hospitalité avec laquelle ils m'ont accueilli.

REFERENCES

[1] ABELSON R. P. and TUKEY J. W. (1959) - Efficient conversion of nonmetric information into metric information. Proc. Social Statist. Section, Amer. Statist. Assoc., 226-230.

[2] ABELSON R. P. and TUKEY J. W. (1963) - Efficient utilization of non-numerical information in quantitative analysis. Ann. Math. Statist., 34, 1347-1369.

[3] BARTHOLOMEW D. J. (1959) - A test of homogeneity for ordered alternatives. Biometrika, 46, 36-48 and 328-335.

[4] BARTHOLOMEW D. J. (1961) - A test of homogeneity of means under restricted alternatives. Jour. Roy. Statist. Soc. B, 23, 239-281.

[5] BROSS I. D. (1958) - How to use Ridit analysis. Biometrics, 13, 18-38.

[6] IPSEN J. (1955) - Appropriate scores in bio-assays using death times and survivor symptoms.

[7] FISHER R. A. (1947) - Les Méthodes Statistiques. Presses Universitaires de France. Sections 49.2 et 49.3.

[8] YATES F. (1948) - The analysis of contingency tables with groupings based on quantitative characters.