REVUE DE STATISTIQUE APPLIQUÉE

J. LELLOUCH
D. SCHWARTZ

Association de 2 variables en tenant compte de l'influence d'une troisième (variables qualificatives)

Revue de statistique appliquée, tome 9, n° 2 (1961), p. 89-102
<http://www.numdam.org/item?id=RSA_1961__9_2_89_0>

ASSOCIATION DE 2 VARIABLES EN TENANT COMPTE DE L'INFLUENCE D'UNE TROISIÈME
(variables qualificatives)

J. LELLOUCH et D. SCHWARTZ

Unité de Recherches Statistiques de l'Institut National d'Hygiène (à l'Institut Gustave Roussy)

Dans de nombreux domaines, l'approche expérimentale est impossible ou n'a pas de sens, et les données de base résultent de la seule observation. La mise sous contrôle de certains facteurs parasites, qu'apporte habituellement le schéma expérimental, doit alors être recherchée a posteriori par le calcul.

Soit à étudier la liaison entre cigarette et cancer du poumon. Des raisons matérielles et morales évidentes interdisent de constituer par tirage au sort 2 groupes comparables qui recevront l'ordre de fumer ou de ne pas fumer ; on doit se contenter d'observer les sujets qui se sont choisis fumeurs ou non fumeurs ; mais ils peuvent alors différer par leur situation sociale, leur milieu d'habitation, etc., facteurs dont il faudra tenir compte en comparant les taux de cancers. Cette démarche est générale dans le domaine de l'étiologie [1].

S'agit-il d'un problème d'ordre pronostique -un signe x entraînant une évolution fâcheuse- la répartition du signe x selon le mode expérimental est dépourvue de sens ; mais les sujets avec et sans x peuvent alors différer par d'autres caractères, dont il faut tenir compte si on veut connaître le rôle propre du signe.

Un autre problème bien connu est celui de la comparaison des taux de mortalité dans 2 populations dont la structure d'âge est différente. Les démographes désirent neutraliser cette différence et comparer les mortalités à âge égal.

Ces quelques exemples ne font qu'illustrer une difficulté commune à toutes les sciences d'observation. Cependant, même quand l'expérimentation est possible, il arrive que certains facteurs, non controlés dans le schéma expérimental, méritent d'être pris en considération dans l'analyse statistique.

Il s'agit finalement d'un problème très général, primordial dans certains secteurs de recherche, et qu'on peut définir mathématiquement ainsi :

"étudier l'association de certaines variables en éliminant l'influence d'autres variables (dites "variables de structure") qui peuvent être liées aux précédentes".

Les techniques statistiques appropriées sont très variées, selon la formulation exacte du problème et la nature, qualitative ou quantitative, des diverses variables. Elles englobent par exemple la corrélation partielle et l'analyse de la covariance. Nous nous proposons d'étudier en détail le cas de l'association de 2 variables qualitatives X et Y avec élimination de l'influence d'une seule variable de structure Z, elle-même qualitative.

Diverses méthodes vont être indiquées ; on les comparera ensuite dans le traitement d'un exemple numérique.

I - UN CAS SIMPLE -

X et Y qualitatifs à 2 états $\bar{x}$ et x, $\bar{y}$ et y

Z qualitatif à k états z_1, z_2 ... z_k.

Les données de base se présentent sous la forme de k tableaux 2 × 2 indépendants (cellules), le i^e étant le suivant :

X / Y	$\bar{x}$	x	Total
$\bar{y}$	a_i^{11}	a_i^{12}	m_i^1
y	a_i^{21}	a_i^{22}	m_i^2
Total	n_i^1	n_i^2	N_i
Pourcentages	$p_i^1 = \frac{a_i^{21}}{n_i^1}$	$p_i^2 = \frac{a_i^{22}}{n_i^2}$	

Cellule $Z = z_i$.

Soit par exemple à comparer les taux de mortalité de 2 populations P_1 et P_2 de structure d'âge Z différente, la variable Z, divisée en tranches d'âge, étant considérée comme qualitative selon un usage courant ; on constitue avec tous les sujets qui appartiennent à la tranche d'âge z_i les tableaux analogues à celui-ci-dessous :

Population	P_1	P_2	Total
Non décédés dans l'année	a_i	b_i	m_i^1
Décédés dans l'année	c_i	d_i	m_i^2
Total	n_i^1	n_i^2	N_i

L'hypothèse H_0 à tester est la suivante : X et Y ne sont liés dans aucune des k cellules.

On peut évidemment tester l'association dans chacune des diverses cellules par les méthodes habituelles (χ^2 ou calcul de la probabilité) ; mais les effectifs peuvent y être faibles, d'où manque de puissance ; en outre, si k

est élevé, on risque de trouver des significations dont on ne sait, sans test complémentaire, si elles sont réelles ou dues au hasard (risque de 1ère espèce). Il vaut mieux utiliser un test unique. Il existe différents tests de l'hypothèse H_0, qui ne peuvent être comparés qu'en fonction de l'hypothèse alternative H_1. Nous allons les passer en revue.

DIFFERENTS TESTS DE L'HYPOTHESE H_0

A - Combinaison des tests.

Ce test est utilisé quand on ignore ce que peut être l'hypothèse alternative H_1, ce qui correspond souvent à un problème mal posé.

La méthode, extrêmement générale, est due à Fisher [2]. Le principe en est le suivant : sous H_0, la probabilité de trouver dans une cellule déterminée une signification au seuil α ou moins est par définition même α. La fonction de répartition du seuil de probabilité p est donc $F = p$, et sa densité est $dF = dp$, soit une densité uniforme. Pour se ramener à des lois tabulées, on pose :

$$u = - \log p, \quad \text{ou } p = e^{-u}$$

La densité de probabilité de u est alors $e^{-u}\,du$, et sa fonction caractéristique est :

$$\int_{-\infty}^{+\infty} e^{itu-u}\,du = \frac{1}{1 - it}$$

En faisant de même pour toutes les cellules et en sommant $- \Sigma \log p = U$, U a pour fonction caractéristique $(1 - it)^{-k}$. On en déduit que $2U$ suit sous H_0 une loi du χ^2 à $2k$ degrés de liberté, d'où le test.

La démonstration suppose que la loi de probabilité de p est continue. Ce n'est approximativement vrai que si les effectifs dans chaque cellule sont assez grands. Dans le cas contraire on se reportera à [3] ou [4].

B - Somme des χ^2.

Ce test sera également utilisé quand on ignore H_1.

Sous H_0 chacune des quantités $\sum \frac{(\text{observés} - \text{attendus})^2}{\text{attendus}}$ calculée sur une cellule est un χ^2 à 1 degré de liberté. La somme de tous ces X^2 indépendants est sous H_0 un χ^2 avec k degrés de liberté.

C - Somme des χ comptés avec leur signe.

Les tests précédents ne tiennent pas compte du signe des différences entre les pourcentages. Si on suppose que toutes ces différences sont de même sens, le test qui suit est préférable :

La quantité $\frac{p_i^1 - p_i^2}{\sqrt{\text{var}\,(p_i^1 - p_i^2)}} = \frac{d_i}{\sqrt{\text{var}\,d_i}}$ suit approximativement une loi normale réduite. Or cette quantité vaut $\varepsilon_i\sqrt{\chi_i^2}$ ($\varepsilon_i = \pm 1$). On en déduit que $\Sigma\,\varepsilon_i\sqrt{\chi^2}$ suit une loi normale de moyenne nulle et d'écart type $\sqrt{k}$. D'où le test.

Remarque concernant les 3 tests qui précèdent.

Ils présentent l'inconvénient suivant : toutes les cellules ont le même poids, malgré les différences qui peuvent exister entre les N_i. Or la puissance des tests effectués sur une cellule où N est petit est faible, et cette faible puissance se répercutera sur le test final. Il faut donc se méfier de ces 3 méthodes quand les effectifs des divers tableaux 2 × 2 sont par trop différents.

Méthodes d'ajustement.

D - Première méthode.

L'hypothèse H_1 correspondante est la suivante : toutes les différences des pourcentages dans les diverses cellules $d_i = p_i^1 - p_i^2$ sont en espérance mathématique indépendantes de i et égales à une valeur δ.

On voit que l'on est alors naturellement amené à :

1/ trouver une estimation d de δ (la meilleure),

2/ comparer d à 0 (hypothèse nulle) et, si d = 0,

3/ chiffrer l'association par d. Le problème posé sera alors entièrement résolu.

Reprenons ces différents points :

1/ δ est comme d'habitude estimé par d, combinaison linéaire des d_i

$d = \dfrac{\Sigma\, w_i\, d_i}{\Sigma w_i}$ de sorte que $E(d) = \delta$.

Sa variance est $\text{var } d = \dfrac{\Sigma\, w_i^2\, \text{var } d_i}{(\Sigma\, w_i)^2}$.

Les w_i sont tels que d ait la meilleure précision (var d minimum) ce qui conduit à :

$$w_i = \frac{1}{\text{var } d_i}$$

2/ Le test de comparaison de d à 0 s'effectue au moyen du critère

$$\chi^2 = \frac{d^2}{\text{var } d} = \frac{(\Sigma w_i\, d_i)^2}{(\Sigma w_i)^2 \times \dfrac{\Sigma\, w_i^2\, \text{var } d_i}{(\Sigma\, w_i)^2}} = \frac{(\Sigma\, w_i\, d_i)^2}{\Sigma w_i}$$

On a besoin d'estimer var d_i : on le fait par $v_i = \dfrac{p_i^1\, q_i^1}{n_i^1} + \dfrac{p_i^2\, q_i^2}{n_i^2}$

Cette méthode est analogue à celle qu'utilisent les démographes pour comparer les taux de mortalité de 2 populations P_1 et P_2 de structure d'âge différente.

En effet ils ramènent les 2 populations à la même structure d'âge, celle d'une population P_3 prise comme référence, puis ils calculent par des règles de trois ce que seraient les taux de mortalité de P_1 et P_2 si elles avaient la structure d'âge de P_3. Ils obtiennent ainsi 2 taux dits "taux standardisés (par âge)" qu'il faut comparer. On calcule généralement leur différence d, qui est, comme ci-dessus, de la forme $\Sigma\, a_i\, d_i$.

Dans le choix de P_3, donc des a_i, les démographes sont guidés par le souci de commodité : on peut prendre par exemple la population d'un pays déterminé à un instant donné, ou une population moyenne (de plusieurs pays), ou même une population théorique, la population "rectangulaire", qui a l'intérêt de se prêter à des calculs particulièrement simples.

Cette façon de procéder n'est correcte que sous l'hypothèse que nous avons faite, à savoir que les espérances mathématiques des différences des taux de mortalité de P_1 et P_2 sont indépendantes de la tranche d'âge.

En effet on estime d par :

$$d = \frac{\Sigma c_i d_i}{\Sigma c_i}$$

(où c_i est l'effectif de la ième tranche de la population de référence P), dont l'espérance mathématique vaut :

$$E(d) = \frac{\Sigma c_i E(\delta_i)}{\Sigma c_i}$$

Si $E(\delta_i)$ n'est pas constant, E(d) peut prendre, selon le choix des c_i, n'importe quelle valeur entre le plus grand et le plus petit des $E(\delta_i)$. En particulier si les $E(\delta_i)$ ne sont pas tous de même signe, E(d) pourra être positive, négative ou nulle.

Si l'hypothèse est vérifiée, n'importe quelle fonction $\frac{\Sigma c_i d_i}{\Sigma c_i}$ estime δ.

Cependant la pondération que nous avons donnée $c_i = w_i = \frac{1}{\text{var } d_i}$ conduit à la meilleure précision : elle correspond à une population de référence P_3' intermédiaire entre P_1 et P_2. On peut vérifier que plus la population choisie diffère de P_3', moins la précision est bonne : c'est ce qui a lieu en particulier avec la population rectangulaire, très loin de P_1 et P_2 donc de P_3'.

L'hypothèse faite sera peut-être vérifiée si les pourcentages varient peu d'une cellule à l'autre ; elle ne le sera sans doute pas si les p_i varient beaucoup. Ceci est tout-à-fait trivial, on ne peut envisager comme équivalentes les deux différences 10 % - 5 % = 50 % - 45 % = 5 %.

En pratique, pour ne pas appliquer à tort un test inadéquat, on testera l'homogénéité des différents d_i par la formule habituelle :

$$\chi^2 = \sum \frac{1}{\text{var } d_i} (d_i - d)^2 \text{ avec } k - 1 \text{ degrés de liberté.}$$

Finalement la méthode est la suivante :

a) tester l'hétérogénéité des d_i ;

b) si non significatif, tester l'association et éventuellement la chiffrer par d.

Remarque - On a la décomposition classique des χ^2 :

$$\sum \frac{d_i^2}{\text{var } d_i} = \sum \frac{1}{\text{var } d_i} (d_i - d)^2 + \frac{[\Sigma (d_i/\text{var } d_i)]^2}{\Sigma 1/\text{var } d_i}$$

$$\chi^2_K = \chi^2_{K-1} + \chi^2_1 ,$$

où le premier terme est la somme des k χ^2 calculés sur chaque cellule

$$\left(\text{avec var } d_i = \frac{p_i^1 q_i^1}{n_i^1} + \frac{p_i^2 q_i^2}{n_i^2}\right)$$

le deuxième teste l'hétérogénéité entre cellules ;

le troisième teste l'association.

E - Méthode de Cochran [5].

L'hypothèse alternative la plus complète est que chaque d_i a pour espérance mathématique un δ_i donné variable avec i. Il s'agit de chercher le meilleur test relativement à cette hypothèse. La quantité :

$$\frac{(\Sigma w_i d_i)^2}{\Sigma w_i^2 \frac{P_i Q_i}{N_i}} \quad \text{où} \quad \frac{1}{N_i} = \frac{1}{n_i^1} + \frac{1}{n_i^2},$$

et où P_i est calculé en mélangeant les 2 populations, est sous H_0 un χ^2 à 1 d.d.l., puisque sous cette même hypothèse

$$V_i = \text{var } d_i = P_i Q_i \left(\frac{1}{n_i^1} + \frac{1}{n_i^2}\right).$$

Reste à calculer les w . Cochran montre que le test sera le plus puissant à conditions de choisir :

$$w_i = \frac{\delta_i}{\frac{p_i'^1 q_i'^1}{n_i^1} + \frac{p_i'^2 q_i'^2}{n_i^2}}$$

où $p_i'^1$ et $p_i'^2$ désignent les vrais pourcentages inconnus dans chacune des 2 populations.

δ_i est la plupart du temps inconnu, mais sans doute faible : on peut donc supposer que $p_i'^1 q_i'^1 = p_i'^2 q_i'^2 = p_i' q_i'$ (p_i' désignant la moyenne de $p_i'^1$ et $p_i'^2$).

Finalement $w_i = \frac{\delta_i N_i}{p_i' q_i'}$ où δ_i, p_i', q_i' sont encore inconnus. Cependant si δ_i est constant en échelle probit, on peut vérifier que $\frac{\delta_i}{p_i' q_i'}$ est à peu près indépendant de p et q, d'où le test :

$$\chi^2 = \frac{(\Sigma N_i d_i)^2}{\Sigma N_i P_i Q_i}$$

L'hypothèse finalement admise - δ_i constant en échelle probit - est plus raisonnable que l'hypothèse δ_i constant admise en (D), qui conduisait à considérer comme équivalentes les différences 10 % - 5 % et 50 % - 45 %.

Imaginons que la variable Y soit en fait quantitative, de distribution normale $(\mu, 1)$; le classement en 2 catégories $\bar{y}$ et y correspondant à un seuil y_o, ($\bar{y}$ pour $Y < y_o$). Soient μ_i^1 et μ_i^2 les moyennes pour les populations $\bar{x}$ et x dans la cellule i Les valeurs théoriques de p_i^1 et p_i^2 seront alors :

$$\bar{\omega}_i^1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_o} e^{-\frac{(u-\mu_i^1)^2}{2}} du = \Pi (y_o - \mu_i^1)$$

$$\bar{\omega}_i^2 = \Pi (y_o - \mu_i^2)$$

d'où : probit $\bar{\omega}_i^1$ - probit $\bar{\omega}_i^2$ = $\mu_i^2 - \mu_i^1$.

Dans un tel cas, l'hypothèse faite revient ainsi à supposer que, pour la variable quantitative Y, la différence est constante d'une cellule à l'autre, c'est-à-dire qu'il n'y a pas d'interaction entre X et Z.

Remarque - Dans le cas extrême où, pour chaque niveau z , il n'y a qu'un seul sujet de la population $\bar{x}$ et un seul de la population x, qui peuvent présenter ou non le caractère y, on tombe dans la méthode des couples pour les caractères qualitatifs. On a alors des couples de sujets, l'un $\bar{x}$ et l'autre x, qui peuvent selon la réponse y être du type $\bar{y}\bar{y}$, yy, $y\bar{y}$, yy. Si on désigne par b et c les effectifs de ces 2 derniers types, rappelons que le test classiquement utilisé [6] est :

$$\chi^2 = \frac{(b - c)^2}{b + c} \quad \text{avec 1 degré de liberté.}$$

On peut montrer très simplement que ce résultat est celui que donnerait la méthode de Cochran dans ces conditions particulières.

II - CAS GENERAL -

$$\begin{cases} X \text{ qualitatif à r états } x_1\ x_2\ \ldots\ x_r\ ; \\ Y \text{ qualitatif à s états } y_1\ y_2\ \ldots\ y_s\ ; \\ Z \text{ qualitatif à k états } z_1\ z_2\ \ldots\ z_k\ . \end{cases}$$

Les données de base se présentent alors sous la forme de k tableaux s × r, le $i^{ème}$ étant le suivant :

X / Y	x_1	x_2		x_r	Total
y_1	a_i^{11}	a_i^{12}		a_i^{1r}	m_i^1
y_2	a_i^{21}	a_i^{22}		a_i^{2r}	m_i^2
y_s	a_i^{s1}	a_i^{s2}		a_i^{sr}	m_i^s
Total	n_i^1	n_i^2		n_i^r	N_i

Seules parmi les méthodes précédentes s'appliquent :

- la combinaison des tests ;
- la somme des χ^2 ,

dont on a déjà dit qu'elles étaient peu puissantes. Mais il en existe d'autres :

F - Extension du χ^2.

Dans chacune des cellules on calcule les effectifs attendus. Puis on regroupe toutes les cellules en une seule où les effectifs observés sont la somme des effectifs observés correspondants dans les diverses cellules. On fait de même pour les effectifs attendus. Et on applique le critérium habituel :

$$\chi^2 = \sum \frac{(\text{observés} - \text{attendus})^2}{\text{attendus}} \quad \text{avec } (r - 1)\ (s - 1) \text{ degrés de liberté.}$$

Cette méthode, si elle ne semble qu'approximativement exacte même asymptotiquement, a le mérite d'être très simple (un exemple particulièrement intéressant est traité dans [7]).

G - Autre méthode.

Elle est décrite en [8].

Exposons-en le principe sur l'exemple suivant : X à 2 niveaux (x_1, x_2), Y à 3 niveaux (y_1, y_2, y_3).

Comme dans la méthode ci-dessus on calcule dans chaque cellule les déviations entre effectifs observés et effectifs attendus, qu'on regroupe en une seule table.

Y \ X	x_1	x_2	Total
y_1	D_1	D'_1	O
y_2	D_2	D'_2	O
y_3	D_3	D'_3	O
Total	O	O	

Parmi les six D du tableau, deux seulement permettent de calculer tous les autres (puisque leur somme par lignes et par colonnes est toujours nulle), par exemple D_1 et D_2.

On sait qu'asymptotiquement $\frac{(D_1)^2}{\text{var } D_1}$ suit sous H_0 une loi du χ^2 à 1 degré de liberté, et, d'après la théorie de la régression linéaire, que la quantité :

$$D_{2,1} = D_2 - \frac{\text{cov } (D_1 D_2)}{\text{var } D_1} D_1$$

est indépendante en probabilité de D_1 ; donc, toujours sous H_0,

$$\frac{(D_1)^2}{\text{var } D_1} + \frac{(D_{2,1})^2}{\text{var } D_{2,1}} = \chi^2 \quad \text{avec 2 d.d.l.}$$

Dans le cas plus général où Y a s classes, on est amené à calculer des quantités de la forme :

$$\frac{(D_1)^2}{\text{var } D_1} + \frac{(D_{2,1})^2}{\text{var } D_{2,1}} + \frac{(D_{3,12})^2}{\text{var } D_{3,12}} + \ldots = \chi^2 \quad \text{avec } (s-1) \text{ degrés de liberté.}$$

où on a :

$$D_{p,12\ldots(p-1)} = D_p - \frac{\text{cov } D_p D_1}{\text{var } D_1} D_1 - \frac{\text{cov } D_p D_{2,1}}{\text{var } D_{2,1}} D_{2,1} - \ldots - \frac{\text{cov } D_p D_{p-1,12\ldots(p-2)}}{\text{var } D_{p-1,12\ldots(p-2)}} D_{p-1,12\ldots p-2}$$

La méthode, qui devient rapidement assez compliquée, nécessite la connaissance des quantités var D_α et cov (D_α, D_β). (α et β variant de 1 à s)

$$\text{var } D_\alpha = \sum_i \text{var } D_{\alpha i}$$

(somme des variances dans chaque cellule) ;

$$\text{cov } (D_\alpha, D_\beta) = \sum_i \text{cov } (D_{\alpha i}, D_{\beta i})$$

(somme des covariances dans chaque cellule).

On se sert, pour les calculer, des formules donnant les moments de la loi hypergéométrique.

Enfin la méthode peut être généralisée au cas du tableau à s lignes et r colonnes.

III - UN EXEMPLE TRAITE PAR LES DIFFERENTES METHODES -

Dans une enquête portant sur l'étiologie du cancer des bronches, on a comparé le pourcentage de sujets éthyliques chez les cancéreux du poumon et chez un groupe témoin. Les chiffres obtenus étaient les suivants (tableau I).

Tableau I

	Témoins	Cancer des bronches
Non éthyliques	462	279
Ethyliques	320	266
Total	782	545
% d'éthyliques	41	49

Le χ^2 à 1 degré de liberté vaut 7,9 et semble indiquer une association entre alcool et cancer des bronches.

Cependant les 2 populations étudiées diffèrent beaucoup quant à l'usage du tabac (tableau II, $\chi^2 = 141,8$ avec 2 d.d.l.).

Tableau II

	Témoins	Cancer des bronches
Non fumeurs	148	23
Petits et moyens fumeurs	420	210
Grands fumeurs	214	312
Total	782	545

D'autre part il y a association entre alcool et tabac. C'est ce qui apparaît sur le tableau III ($\chi^2 = 6,1$ avec 2 d.d.l.).

Tableau III

Population témoin

	Non fumeurs	Petits et moyens fumeurs	Grands fumeurs	Total
Non éthyliques	95	255	112	462
Ethyliques	53	165	102	320
Total	148	420	214	782
% de sujets éthyliques	36	39	48	

Une comparaison sur l'éthylisme doit donc tenir compte de la quantité fumée.

Les données de base se présentent sous la forme des 3 tableaux 2 × 2 suivants :

Tableaux IV

	Témoins	Cancer des bronches
Non éthyliques	95	13
Ethyliques	53	10
Total	148	23
% de sujets éthyliques	36	43

Non fumeurs

	Témoins	Cancer des bronches
Non éthyliques	255	128
Ethyliques	165	82
Total	420	210
% de sujets éthyliques	39	39

Petits et moyens fumeurs

	Témoins	Cancer des bronches
Non éthyliques	112	138
Ethyliques	102	174
Total	214	312
% de sujets éthyliques	48	56

Grands fumeurs

Les χ^2 calculés sur chaque tableau valent respectivement 0,503 (p = 0,48); 0,003 (p = 0,96) et 3,345 (p = 0,07) et ne sont pas significatifs.

Les méthodes proposées donnent les résultats suivants :

a) Combinaison des tests.

p	$\log_{10} p$
0,48	$\bar{1},68124$
0,96	$\bar{1},98227$
0,07	$\bar{2},84510$
	$\bar{2},50861$

$$2u = -2 \frac{\Sigma \log_{10} p}{\log_{10} e} = 6,87$$

qui suit une loi du χ^2 à 6 degrés de liberté, d'où p voisin de 0,40.

b) Somme des χ^2. $\chi^2 = 0,503 + 0,003 + 3,345 = 3,9$ qui avec 3 degrés de liberté correspond à une probabilité comprise entre 0,20 et 0,30.

c) Somme des χ comptés avec leur signe

$$\frac{0,709 - 0,056 + 1,829}{\sqrt{3}} = 1,4$$

qui correspond à une probabilité p telle que :

$$0,15 < p < 0,16$$

d) Première méthode d'ajustement. Les calculs conduisent au tableau suivant :

p_1	q_1	$\frac{p_1 q_1}{n_1} \times 10^2$	p_2	q_2	$\frac{p_2 q_2}{n_2} \times 10^2$	$w \times 10^2$	$w = \frac{1}{v}$	d	wd
0,3581	0,6419	0,1553	0,4349	0,5651	1,0685	1,2238	81,713	+0,0768	6,2756
0,3929	0,6071	0,0568	0,3905	0,6095	0,1133	0,1701	587,889	-0,0024	- 1,4109
0,4766	0,5234	0,1166	0,5577	0,4423	0,0791	0,1957	510,986	+0,0811	41,4410
							1180,588		46,3057

d'où :

Comparaison de d à 0	d.d.l. = 1	$\chi^2 = (\Sigma d_i w_i)^2 / \Sigma w_i$	1,816
Hétérogénéité entre cellules	d.d.l. = 2	χ^2 = par différence	2,030
Total	d.d.l. = 3	$\chi^2 = \Sigma d_i^2 w_i$	3,846

La conclusion est que les 2 populations ne diffèrent pas significativement entre elles.

$\chi_1^2 = 1,8$ correspond en effet à une probabilité p telle que :

$$0,17 < p < 0,18$$

e) Méthode de Cochran. Les calculs sont résumés dans le tableau suivant :

P	Q	PQ	n_1	n_2	$N = \frac{n_1 n_2}{n_1+n_2}$	d	Nd	Npq
0,3684	0,6316	0,2327	148	23	19,906	+0,0768	1,5288	4,6317
0,3921	0,6079	0,2384	420	210	140,000	-0,0024	- 0,3360	33,3704
0,5247	0,4753	0,2494	214	312	126,935	+0,0811	10,2944	31,6563
							11,4872	69,6584

$$\chi^2 = \frac{(\Sigma Nd)^2}{\Sigma Npq} = 1,9$$

$$0,16 < p < 0,17$$

f) Extension du χ^2. Les tableaux des effectifs attendus correspondant aux tableaux IV sont :

	Témoins	Cancer des bronches
Non éthyliques	93,474	14,526
Ethyliques	54,526	8,474

Non fumeurs

	Témoins	Cancer des bronches
Non éthyliques	255,333	127,667
Ethyliques	164,667	82,333

Petits et moyens fumeurs

	Témoins	Cancer des bronches
Non éthyliques	101,711	148,289
Ethyliques	112,289	163,711

Grands fumeurs

qu'on regroupe en un seul tableau d'effectifs attendus qu'on compare au tableau I

	Témoins	Cancer des bronches
Non éthyliques	450,518	290,482
Ethyliques	331,482	254,518

Le test consiste à calculer :

$$\chi^2 = \sum \frac{(\text{attendus} - \text{observés})^2}{\text{attendus}} = \overline{11,482}^2 \left[\frac{1}{450,518} + \frac{1}{290,482} + \frac{1}{331,482} + \frac{1}{254,518} \right]$$

avec un degré de liberté.

$\chi^2 = 1,7$;

$0,19 < p < 0,20$.

g) Méthode G. La déviation D_1, somme des déviations entre effectifs attendus et effectifs observés dans les 3 tableaux III vaut :

$$1,526 - 0,333 + 10,289 = 11,482$$

Le test consiste à voir si elle est significativement différente de 0. La variance de D_1 est la somme des variances des 3 déviations partielles ; ces variances valent :

$$\frac{m_i^1 m_i^2 n_i^1 n_i^2}{N^2(N-1)}$$

soit respectivement :

$$4,659 \ ; \ 33,422 \text{ et } 31,717$$

d'où var $D_1 = 69,798$.

Finalement :

$$\chi^2 = \frac{D_1^2}{\text{var } D_1} = 1,9$$

$$0,16 < p < 0,17$$

IV - CONCLUSIONS -

L'exemple traité est démonstratif : alors qu'il semblait y avoir une association entre alcool et cancer des bronches, on constate en fait que cancéreux et témoins ne diffèrent plus quand on prend en considération la consommation de tabac.

Il y a donc, à tenir compte de tels facteurs, un intérêt considérable, qui explique la multiplicité des méthodes proposées par la littérature.

Cependant, en traitant un même exemple numérique par 7 méthodes différentes -avec des résultats souvent voisins- nous ne voudrions pas laisser croire que ces techniques peuvent être utilisées indifféremment.

La similitude des résultats est d'abord affaire de circonstance ; dans l'exemple choisi les proportions p_1 et p_2 approchaient de 50 % dans chaque cellule, de sorte que δ_i peut être à peu près constant à la fois en échelles arithmétique et probit ; ce ne serait évidemment pas le cas pour des proportions plus voisines de 0 % ou 100 %.

Il faut, pour chaque problème particulier, choisir la méthode appropriée.

La combinaison des tests et la somme des χ^2 sont très générales, mais peu puissantes : elles ne sont à utiliser que si l'hypothèse alternative ne peut être énoncée, ce qui correspond à un problème incomplètement formulé.

Une distinction intervient ensuite, celle du nombre de classes pour X et Y :

- S'il est supérieur à 2, pour X ou Y, on n'a guère le choix qu'entre les méthodes F (relativement simple mais pas très exacte) et G (plus exacte, mais compliquée).

- Si X et Y n'ont que 2 classes, cas d'un usage très courant, on n'a pas intérêt à utiliser ces 2 méthodes ; il reste à choisir entre les 3 méthodes C, D, E. Le test C (somme des χ) est évidemment le plus rapide, et pourra être utilisé lorsque les effectifs ne varient pas trop d'une cellule à l'autre. S'il n'en est pas ainsi, on devra employer une des méthodes d'ajustement D ou E. La méthode E correspond à une hypothèse alternative plus raisonnable, et devra être préférée. Cependant, lorsqu'un test préalable aura montré la relative constance des différences d d'une cellule à l'autre, on pourra utiliser le test D, qui est alors plus simple et plus court.

Travail de l'Unité de Recherches Statistiques
de l'Institut National d'Hygiène
(à l'Institut Gustave-Roussy)

BIBLIOGRAPHIE

[1] SCHWARTZ D. - "La méthode statistique en médecine : les enquêtes étiologiques" (Rev. Stat. Appl. 8, n° 3, pp. 5-27, 1960).

[2] KENDALL - "The advanced theory of statistics" (Ed. Griffin & Cie, London, 1946 - vol. II - pp. 132-133).

[3] LANCASTER - "The combination of probabilities arising from data in discrete distributions" (Biometrika 36, pp. 370-382, 1949).

[4] YATES - "A note on the application of the combination of probabilities test to a set of 2 2 tables" (Biometrika 42, pp. 404-411, 1955).

[5] COCHRAN - "Some methods for strengthening the common χ^2 tests" (Biometrics 10, n° 4, pp. 417-451, 1954).

[6] COCHRAN - "The comparison of percentages in matched samples" (Biometrika 37, p. 256, 1950).

[7] BOYD, DOLL - "Gastro-intestinal cancer and the use of liquid paraffin" (The Brit. J. Canc. 8, pp. 231-237, 1954).

[8] MANTEL, HAENSZEL - "Statistical aspects of the analysis of data from retrospective studies of disease" (J. Nat. Canc. Inst. 22, n° 4, pp. 719-748, 1959).