ANNALES DE LA FACULTÉ DES SCIENCES DE TOULOUSE

HENRI CAUSSINUS

Contribution à l'analyse statistique des tableaux de corrélation

Annales de la faculté des sciences de Toulouse 4^e série, tome 29 (1965), p. 77-183

<http://www.numdam.org/item?id=AFST_1965_4_29__77_0>

Contribution à l'analyse statistique des tableaux de corrélation

par Henri CAUSSINUS

INTRODUCTION

L'étude des tables de contingence, ou tableaux de corrélation, est l'une des premières qui ont préoccupé les statisticiens. Cependant, jusqu'à une période récente, n'avaient été envisagés, à quelques rares exceptions près, que des tableaux à deux dimensions et, liées à ceux-ci, deux questions : test de l'hypothèse d'indépendance entre les deux classifications et mesures globales d'association. Plus récemment, l'on a assisté à une recrudescence de l'intérêt porté aux tableaux de corrélation; de nouveaux problèmes étaient examinés; c'étaient d'un côté, sous l'impulsion de M. FRÉCHET, les problèmes d'existence et de construction de tableaux à marges fixées ayant certaines propriétés, d'un autre côté, la comparaison de plusieurs tables de contingences et l'étude voisine des tables à trois dimensions ou plus; les publications sur ce dernier sujet sont essentiellement d'origine anglo-saxonne et orientées vers les problèmes d'estimation et de tests.

Toutefois, pour tout ce qui concerne l'étude de la structure d'une table de contingence, peu de travaux avaient été entrepris. Il nous a semblé utile de rechercher des méthodes statistiques qui permettraient l'analyse du mode de corrélation des variables qualitatives. Dans ce but, un premier pas pouvait consister, lorsque l'hypothèse d'indépendance était inadmissible, à éprouver des hypothèses plus faibles que cette dernière mais de physionomie voisine. Ceci nous conduisit à envisager des tables de contingence « volontairement » tronquées; mais cette étude pouvait avoir un intérêt au-delà de cette question, puisqu'elle était applicable aux tables de contingence tronquées par la force des choses, soit par accident, soit à cause de la nature même du problème envisagé. Vu sous ce nouvel aspect, le problème paraissait d'actualité, il venait de susciter plusieurs publications. Celles-ci ne traitaient cependant que des cas particuliers, importants mais assez simples, concernant surtout l'estimation des paramètres d'une distribution multinomiale par adjonction de plusieurs échantillons incomplets. Nous avons essayé d'abord de poser les bases d'une étude très générale de ces questions (Chapitre I), puis de trouver des méthodes d'estimation aussi simples que possible pour tous les cas de troncature (Chapitre II) ; nous avons constaté

que l'on pouvait se ramener pour cela à la construction d'un certain tableau à marges fixées et notre étude utilise quelques théorèmes relatifs à cette question, établissant ainsi une jonction entre les préoccupations de l'école française et de l'école anglo-saxonne.

Après cela, nous exposons quelques tests pour les tables de contingence tronquées. Ceux-ci peuvent être utiles à des fins variées, d'abord pour les tables tronquées proprement dites, ensuite pour les tables de contingence usuelles dont ils peuvent permettre d'analyser plus avant la structure. Toutes ces questions ont été traitées au Chapitre III. Nous comparons aussi nos techniques à de nouvelles méthodes que l'on peut adapter d'une publication récente de GOODMAN sur la recherche d'intervalles de confiance pour les interactions du premier ordre, publication qui a paru alors que le présent travail était déjà entrepris.

Dans le Chapitre IV, nous étudions quelques généralisations. Nous montrons d'abord comment l'on peut étudier quelques nouvelles hypothèses sur la structure d'un tableau de corrélation à deux dimensions. Nous envisageons ensuite les tableaux à trois dimensions.

Essayant enfin de poursuivre notre étude sur l'analyse d'une table de contingence (à deux dimensions) nous avons étudié plus particulièrement au Chapitre V, des modèles d'association pour des tableaux carrés. Outre les propriétés mathématiques de ces modèles, nous examinons leur signification et leur portée pratique. Comme on le verra, celle-ci semble assez vaste, en particulier pour les tableaux qui décrivent certains processus de transition (au sens le plus large) de groupes d'individus.

Pour terminer, nous exposons au Chapitre VI quelques applications des méthodes que nous avons introduites, tant pour illustrer les techniques mathématiques impliquées que les possibilités pratiques de ces méthodes.

M. le Professeur Léopold ESCANDE, Membre de l'Institut, a bien voulu présenter à l'Académie des Sciences plusieurs notes contenant les principaux résultats de ce mémoire; qu'il veuille bien trouver ici l'expression de ma vive et respectueuse gratitude.

M. le Professeur Jean COMBES, qui fut l'un de mes premiers maîtres à la Faculté des Sciences de Toulouse, m'a fait l'honneur d'accepter la présidence du Jury; qu'il soit assuré de ma très sincère reconnaissance.

M. le Professeur Roger HURON, qui m'a accueilli dans son Laboratoire de Statistique, est à l'origine de ce travail; les conseils et les encouragements qu'il m'a prodigués ont été nécessaires à son aboutissement; je le prie de croire à ma profonde gratitude et à mon fidèle attachement.

M. le Professeur Henri MASCART a bien voulu me proposer le sujet de seconde thèse et m'a aidé à sa préparation avec une bienveillante sollicitude, je l'en remercie très vivement.

M. le Professeur Jean MÉRIC a bien voulu se joindre au Jury, me témoignant une nouvelle fois sa sympathie; je lui adresse ici mes remerciements chaleureux.

Mes remerciements vont aussi à M^lle Élisabeth LAMBERT, Collaboratrice technique, qui m'a aidé pour la mise en œuvre des méthodes numériques, et à M^me Irène CORACIN qui a assuré avec dévouement et compétence la mise en pages de ce travail.

Je ne peux citer ici tous les autres chercheurs et techniciens du Laboratoire de Statistique, pourtant leur amitié dévouée m'est une aide constante : qu'ils soient assurés de toute ma sympathie.

CHAPITRE PREMIER

ÉTUDE DES RELATIONS DU TYPE H_R

A. — THÉORÈMES GÉNÉRAUX PRÉLIMINAIRES

Nous avons rassemblé dans cette partie les énoncés et démonstrations de quelques théorèmes purement algébriques, dont nous nous servirons dans l'étude statistique qui suit.

1. Notations et définitions.

On considère un ensemble de r indices que l'on notera 1,2, ..., r et un ensemble de s indices notés 1,2, ..., s. Soit C l'ensemble produit. R étant une partie de C, on pose :

$$R_{i.} = \{j : (i, j) \in R\} \qquad R_{.j} = \{i : (i, j) \in R\}$$

$$R^* = \bigcup_i R_{i.} \qquad \bar{R} = \bigcup_j R_{.j}$$

$$\tilde{R} = \bar{R} \times R^* = \{(i, j) : i \in \bar{R} \text{ et } j \in R^*\}$$

Les propriétés suivantes se déduisent immédiatement de ces définitions :

$$\text{si } R \subset C \text{ et } S \subset C$$

$$(R \cup S)_{.j} = R_{.j} \cup S_{.j}\ , \quad (R \cup S)_{i.} = R_{i.} \cup S_{i.}$$

$$\overline{(R \cup S)} = \bar{R} \cup \bar{S}\ , \quad (R \cup S)^* = R^* \cup S^*$$

Dimensions de R : Si $R \subset C$ est telle que $Card\,(\bar{R}) = l$, $Card\,(R^*) = c$ on dira que R est (l, c).

Connexité : On dira que $R \subset C$ est connexe si

$$\forall\, R_1 \subset R\ (R_1 \neq \phi \text{ et } R_1 \neq R)$$

$$\bar{R}_1 \cap \overline{(R - R_1)} \neq \phi \text{ ou } R_1^* \cap (R - R_1)^* \neq \phi$$

Relation H_R : Soit G un groupe dont nous noterons la loi de composition multiplicativement (l'opposé de $x \in G$ sera désigné par x^{-1}). Sauf mention du contraire, les lettres indicées (g_{ij}, p_i, q_j, etc...) seront des éléments de G.

A tout (i, j) appartenant à C on associe un élément de G, soit g_{ij}. H_R sera définie par :

$$\forall i \in \bar{R},\ \exists p_i \in G \text{ et } \forall j \in R^*,\ \exists q_j \in G$$

$$\text{tels que } \forall\ (i, j) \in R,\ g_{ij} = p_i q_j.$$

H_R vraie, signifie donc que les g_{ij} sont tels que le système $g_{ij} = p_i q_j$ pour tout $(i, j) \in R$ admet une solution que l'on notera simplement (p, q). Cette

solution n'est pas unique, en effet, quel que soit $k \in G$, si $p_i' = p_i k$ $(i \in \bar{R})$ et $q'_j = k^{-1} q_j$ $(j \in R^*)$ l'ensemble (p', q') répond aussi à la question. Cependant, l'on conviendra de dire que la solution (p, q) est unique, si pour toute autre solution (p', q') il existe $k \in G$ tel que $p_i' = p_i k$ pour tout $i \in \bar{R}$, ce qui implique $q_j' = k^{-1} q_j$ pour tout $j \in R^*$.

II. Propriétés des parties connexes.

Nous démontrons ici quelques propriétés simples qui nous seront utiles par la suite; dans cette étude R désigne toujours une partie de C et nous supposerons (ce qui ne restreint pas la généralité) $\bar{R} = \{1, 2, .., l\}$ et $R^* = \{1, 2, .., c\}$ lorsque ceci pourra simplifier les notations.

Lemme I. — Si R est connexe, quelle que soit $R' \subset R$ telle que $R'^* \neq R^*$ et $R'_{.j} = R_{.j}$ pour tout j appartenant à R'^*, il existe j appartenant à $R^* - R'^*$ tel que $R_{.j} \cap \bar{R}' \neq \phi$.

En effet, l'on a ici $R^* - R'^* = (R - R')^*$ et donc :

$$R'^* \cap (R - R')^* = R'^* \cap (R^* - R'^*) = \phi$$

Si le lemme ci-dessus était faux, l'on aurait en outre $R_{.j} \cap \bar{R}' = \phi$ pour tout $j \in (R - R')^*$ et donc $\left[\bigcup_{j \in (R-R')^*} R_{.j} \right] \cap \bar{R}' = \phi$, soit $\overline{(R - R')} \cap \bar{R}' = \phi$, ce qui contredirait l'hypothèse de connexité de R.

Bien entendu, l'on démontrera de même la propriété duale : Si R est connexe, quelle que soit $R' \subset R$ telle que $\bar{R}' \neq R$ et $R'_{i.} = R_{i.}$ pour tout $i \in \bar{R}'$, il existe $i \in \bar{R} - \bar{R}'$ tel que $R_{i.} \cap R'^* \neq \phi$.

Lemmes II. — Soit R connexe (l, c) et $r = \{(i, j)\} \subset C$, alors :

II a) Si $r \subset \tilde{R}$, $R \cup r$ est connexe (l, c).

II b) Si $i \notin \bar{R}$ et $j \in R^*$, $R \cup r$ est connexe $(l + 1, c)$

II c) Si $i \in \bar{R}$ et $j \notin R^*$, $R \cup r$ est connexe $(l, c + 1)$

Pour montrer la proposition IIa) supposons $R \cup r$ non connexe; dans ces conditions, l'on devrait avoir, soit $\bar{r} \cap \bar{R} = r^* \cap R^* = \phi$, ce qui est contraire à l'hypothèse, soit pour une partie $R' \subset R$, $\overline{(R' \cup r)} \cap \overline{(R - R')} = (R' \cup r)^* \cap (R - R')^* = \phi$; ceci implique :

$$\overline{(R' \cup r)} \cap \overline{(R - R')} = \bar{R}' \cap \overline{(R - R')} \cup \bar{r} \cap \overline{(R - R')} = \phi$$

donc nécessairement $\bar{R} \cap \overline{(R - R')} = \phi$

et de même $R^* \cap (R - R')^* = \phi$

ce qui est contraire à l'hypothèse.

Les démonstrations de II b et II c sont identiques; montrons par exemple IIb. Il est évident d'abord que $R \cup r$ est $(l + 1, c)$. Montrons que $R \cup r$ est connexe : si l'on considère une partition de $R \cup r$ en deux compo-

santes, elle est nécessairement du type $(R' \cup r,\ R - R')$; dans le cas où R' est vide, l'on a $r^* \cap R^* \neq \phi$ par hypothèse et si $R' \neq \phi$:

$$(R' \cup r)^* \cap (R - R')^* = [R'^* \cap (R - R')^*] \cup [r^* \cap (R - R')^*]$$

$$\text{et } \overline{(R' \cup r)} \cap \overline{(R - R')} = [\bar{R}' \cap \overline{(R - R')}] \cup [\bar{r} \cap \overline{(R - R')}]$$

Utilisant la connexité de R, l'une des deux intersections ci-dessus est non vide. C. q. f. d.

En appliquant autant de fois qu'il sera nécessaire les trois résultats ci-dessus on obtient :

Lemme III. — Si R est connexe (l,c) et la partie $S \subset C$ est telle que $S^* \subset R^*$ (resp. $\bar{S} \subset \bar{R}$) et $\bar{S} - \bar{R}$ a l_1 éléments (resp. $S^* - R^*$ a c_1 éléments), $R \cup S$ est connexe $(l + l_1,\ c)$ [resp. $l,\ c + c_1$)].

Nous montrerons maintenant le :

Lemme IV. — a) Si R est connexe et telle que, pour un i donné, $R_{i.}$ a un seul élément j, la partie $S = R - \{(i, j)\}$ est connexe $(l - 1,\ c)$.

b) Si R est connexe et telle que, pour un j donné, $R_{.j}$ a un seul élément i, la partie $S = R - \{(i, j)\}$ est connexe $(l,\ c - 1)$.

Raisonnons sur le cas a) ; il est évident que S est $(l - 1,\ c)$; montrons qu'elle est connexe. Notant $r = \{(i, j)\}$, on a $R = S \cup r$ et par hypothèse $\bar{S} \cap \bar{r} = \phi$ (1) ce qui implique $S^* \cap r^* \neq \phi$ puisque R est connexe.

Supposons S non connexe; ceci entraîne l'existence de $S' \subset S$ ($S' \neq S$ et $S' \neq \phi$) telle que :

$$\bar{S}' \cap \overline{(S - S')} = \phi \qquad (2)$$

$$S'^* \cap (S - S')^* = \phi \qquad (3)$$

On peut toujours supposer que l'unique élément de r^* qui est aussi un élément de S^* appartient à S'^*; dans ces conditions il vient par (3) : $r^* \cap (S - S')^* = \phi$ (4).

$S' \cup r$ et $S - S'$ constituent une partition de R telle que :

$$\overline{(S' \cup r)} \cap \overline{(S - S')} = (\bar{S}' \cup \bar{r}) \cap \overline{(S - S')}$$

$$= [\bar{S}' \cap \overline{(S - S')}] \cup [\bar{r} \cap \overline{(S - S')}] = \phi$$

en utilisant (1) et (2), et :

$$(S' \cup r)^* \cap (S - S')^* = (S'^* \cup r^*) \cap (S - S')^*$$

$$= [S'^* \cap (S - S')^*] \cup [r^* \cap (S - S')^*] = \phi$$

en utilisant (3) et (4).

Dans ces conditions R ne serait pas connexe, ce qui est contraire à l'hypothèse. Donc S est connexe.

Théorème I. — Si R est connexe $(l,\ c)$ alors Card $(R) \geqslant l + c - 1$. Posons $t = \text{Card } (R)$. On admet $t \leqslant l + c - 1$ (1). Supposons $l \geqslant c$. Dans ces conditions, il existe $i \in \bar{R}$ tel que $R_{i.}$ est réduit à un seul élément, soit j (en effet, dans le cas contraire, l'on aurait $t \geqslant 2\,l \geqslant l + c > l + c - 1$).

Soit alors r la partie réduite à l'élément (i, j) ; R-r a t-1 éléments, et, d'après le lemme IV-a, R — r est connexe $(l-1, c)$; sur R — r on pourra raisonner comme ci-dessus, l'inégalité (1) gardant toujours le même sens puisque les deux nombres sont diminués d'une unité.

En appliquant $l - c + 1$ fois ce processus de formation, on arrivera à une partie connexe $(c-1, c)$ à $t - (l - c + 1)$ éléments que l'on traitera de façon analogue en permutant seulement les rôles des indices, etc.... On pourra ainsi, en $l + c - 2$ pas, construire une partie $(1, 1)$ à $t - (l + c - 2)$ éléments. Or cette partie a nécessairement un élément, donc $t = l + c - 1$, et t ne peut être strictement inférieur à $l + c - 1$.

On remarquera que, quels que soient l et c, il est possible de construire une partie $R \subset C$ connexe de dimension (l, c), à $l + c - 1$ éléments : par exemple la partie formée des éléments $(i, 1)$ pour $i = 1, 2, .., l$ et $(1, j)$ pour $j = 1, 2, .., c$ (partir de $\{(1, 1)\}$ qui est évidemment connexe et appliquer les lemmes II-b et II-c).

D'autres exemples sont donnés par le théorème suivant.

Théorème II. — De toute partie R connexe (l, c) à t éléments $(t > l + c - 1)$ il est possible d'extraire une partie connexe (l, c) à $l + c - 1$ éléments.

La démonstration ci-dessous donne, en outre, une méthode de construction d'une telle partie.

On partira par exemple de l'ensemble R_1 des couples $(1, j)$ où $j \in R_{1.}$. Si $R_{1.}$ a c_1 éléments, on peut supposer qu'il s'agit de 1, 2, .., c_1. R_1 est connexe $(c_1, 1)$, a c_1 éléments, est inclue dans R. Pour $R_1 = R$ le théorème est démontré; sinon considérons tous les indices i différents de 1 tels que $R_{1.} \cap R_{i.} \neq \phi$ (il en existe en vertu du Lemme I) : numérotons 2, 3, .., l_1 ces indices; la partie R_2 sera constituée de la réunion de R_1 et de $l_1 - 1$ couples $(i, j) \in R$ où i vaut successivement 2, 3, .., l_1 et où, à chaque i, est associé un j et un seul (j peut être 1, 2, .., ou c_1), ce qui est possible puisque $R_{1.} \cap R_{i.} \neq \phi$ pour tout $i = 1, 2, .., l_1$. R_2 est une partie (l_1, c_1) à $l_1 + c_1 - 1$ éléments, contenue dans R et connexe d'après le lemme III.

Si $l_1 = l$ et $c_1 = c$ la démonstration est terminée, sinon, appliquons, le Lemme I avec $R' = \widetilde{R}_2 \cap R$ et considérons les indices j tels que $R_{.j} \cap \overline{R'} \neq \phi$, soient $c_1 + 1, c_1 + 2, \ldots, c_2$ ces indices; R_3 sera formée de la réunion de R_2 et de $c_2 - c_1$ éléments (i, j) où $j = c_1 + 1, c_1 + 2, \ldots, c_2$ successivement, et à chaque j est associé un indice $i \in \overline{R}_2$. R_3 est (l_1, c_2) à $l_1 + c_2 - 1$ éléments, contenue dans R et connexe (Lemme III).

Si $l_1 \neq l$ et $c_2 \neq c$, on continue le processus jusqu'à obtention d'une partie T à $l' + c' - 1$ éléments telle que $l' = l$ ou $c' = c$. Si, par exemple, $l' = l$, on adjoindra à T pour chaque j $(j = c' + 1, \ldots, c)$ un élément $(i, j) \in R$.

La partie $S \subset R$ ainsi obtenue est bien connexe (l, c), avec $l + c - 1$ éléments. Evidemment, S n'est pas en général unique.

III. Propriétés de la relation H_R.

Le premier problème qui se pose est la recherche des parties R de C pour lesquelles le système (p, q) lié à H_R est unique dans le sens défini au paragraphe 1.

Théorème III. — (Unicité).

Lorsque H_R est vraie, le système (p, q) lié à H_R est unique, si et seulement si R est connexe.

1° Supposons R non connexe; il existe donc $R' \subset R (R' \neq \phi$ et $R' \neq R)$ telle que (avec $R'' = R - R'$) :

$$\bar{R}' \cap \bar{\bar{R}}'' = R'^* \cap R''^* = \phi$$

Supposons : $\forall (i, j) \in R, g_{ij} = p_i q_j$

Posons $\quad p'_i = p_i k$ pour $i \in \bar{R}'$

$\quad p'_i = p_i h$ pour $i \in \bar{R}''$

avec $k \in G$, $h \in G$, $k \neq h$, ce qui est possible puisque $\bar{R}' \cap \bar{R}''$ est vide, et

$$q'_j = k^{-1} q_j \text{ pour } j \in R'^*$$
$$q'_j = h^{-1} q_j \text{ pour } j \in R''^*$$

ce qui est possible puisque $R'^* \cap R''^*$ est vide.

On a bien pour tout $(i, j) \in R'$ comme pour tout $(i, j) \in R''$, donc quel que soit $(i, j) \in R$: $g_{ij} = p_i q_j = p'_i q'_j$, et l'unicité est en défaut puisque $k \neq h$.

2° Supposons R connexe (l, c) et $p_{ij} = p_i q_j = p'_i q'_j$ pour tout $(i, j) \in R$, d'où pour tout $j \in R^*$ et pour tout $i \in R_{.j}$: $p_i^{-1} p'_i = k_j$. Le théorème sera démontré, si l'on montre que pour tout $j \in R^*$, $k_j = k$ (k indépendant de j).

Partant, par exemple, de $j = 1$, l'on a $p_i^{-1} p'_i = k_1$ pour tout $i \in R_{.1}$. D'après le lemme I, il existe des indices j $(j \neq 1)$ tels que $R_{.1} \cap R_{.j} \neq \phi$: considérons les tous, on peut supposer qu'il s'agit des indices 2, 3, .., j_1. Ainsi pour tout j inférieur ou égal à j_1, il existera au moins un $i \in R_{.1} \cap R_{.j}$ et donc tel que $p_i^{-1} p'_i = k_1 = k_j$ d'où : $k_1 = k_2 = \ldots = k_{j_1}$.

Si $j_1 = c$ le théorème est démontré.

Si $j_1 \neq c$, considérons le nouvel ensemble $\bigcup_{k=1}^{k=j_1} R_{.k}$; d'après le lemme I, il existe des indices j, différents de 1, 2, .. j_1, tels que :

$$\bigcup_{h=1}^{h=j_1} R_{.h} \cap R_{.j} \neq \varnothing$$

On peut supposer que ce sont les indices $j_1 + 1$, $j_1 + 2$, .., j_2; il vient en

raisonnant comme plus haut : $k_1 = k_2 = \ldots = k_{j_2}$. Après m opérations analogues on aura montré :
$k_1 = k_2 = \ldots = k_{j_m}$ avec $j_m > j_{m-1}$; donc, après un nombre fini de pas, l'on obtiendra bien $k_1 = k_2 = \ldots = k_c = k$.

Théorème IV. — Si $S \subset R$, $H_R \Longrightarrow H_S$. (On aura donc aussi : non $H_S \Longrightarrow$ non H_R).

La démonstration est immédiate : si il existe p_i et q_j tels que, pour tout $(i, j) \in R$, $g_{ij} = p_i q_j$, cette relation sera vraie en particulier pour $(i, j) \in S$.

Théorème V. — Si $S = R \cup \{(h, k)\}$ avec $h \in \overline{R}$ et $k \notin R^*$ (ou bien $h \notin \overline{R}$ et $k \in R^*$), $H_R \Longleftrightarrow H_S$.

On a d'une part, puisque R est contenue dans S : $H_S \Longrightarrow H_R$. D'autre part, H_R s'écrit :

$$\forall i \in \overline{R},\ \exists p_i \in G \text{ et } \forall j \in R^*,\ \exists q_j \in G$$
$$\text{tels que } \forall (i, j) \in R,\ g_{ij} = p_i q_j.$$

Si l'on se place dans le cas où $h \in \overline{R}$ et $k \notin R^*$, on a $\overline{S} = \overline{R}$ et $S^* = R^* \cup \{k\}$; en posant $q_k = p_h^{-1} g_{hk}$ l'on vérifie immédiatement que H_S est vraie quel que soit g_{hk}. Donc $H_R \Rightarrow H_S$, ce qui achève la démonstration. Ce théorème est encore vrai pour $h \notin \overline{R}$ et $k \notin R^*$ (mais, dans ce cas, S n'est pas connexe).

Théorème VI. — Si R est connexe (l, c) à $l + c - 1$ éléments, H_R est vraie quels que soient les g_{ij} associés aux éléments de R.

Il suffit de reprendre la décomposition du théorème I (dans le cas $t = l + c - 1$). On pourra obtenir R en partant d'une partie (1, 1) soit r, à laquelle on ajoutera $l + c - 2$ fois un élément dans les conditions d'application du théorème précédent. D'où $H_R \Longleftrightarrow H_r$, or H_r est évidemment toujours vraie.

Remarque. — Si R n'est pas connexe, même ayant un nombre d'éléments inférieur ou égal à $l + c - 1$, il est possible que H_R ne soit pas vraie.

Théorème VII. — Si R et S sont telles que $\overline{R} \cap \overline{S} = R^* \cap S^* = \phi$, ($H_R$ et H_S) $\Longleftrightarrow H_{(RUS)}$

D'une part, puisque $R \subset R \cup S$ et $S \subset R \cup S$, l'on a

$$H_{(RUS)} \Longrightarrow (H_R \text{ et } H_S) \qquad \text{(Théorème IV)}.$$

D'autre part, H_R et H_S sont définies par :

$$\forall i \in \overline{R},\ \exists p_i \in G \text{ et } \forall j \in R^*,\ \exists q_i \in G$$
$$\text{tels que } \forall (i, j) \in R,\ g_{ij} = p_i q_j$$
$$\text{et } \forall i \in \overline{S},\ \exists p_i \in G \text{ et } \forall j \in S^*,\ \exists q_j \in G$$
$$\text{tels que } \forall (i, j) \in S,\ g_{ij} = p_i q_j.$$

Puisque $\overline{R} \cap \overline{S}$ et $R^* \cap S^*$ sont vides et en utilisant $\overline{R} \cup \overline{S} = \overline{R \cup S}$ et $R^* \cup S^* = (R \cup S)^*$, la conjonction de ces deux assertions est équivalente à :

$\forall\, i \in \overline{R \cup S}$, $\exists\, p_i \in G$ et $\forall\, j \in (R \cup S)^*$, $\exists\, q_j \in G$

tels que $\forall\ (i, j) \in R \cup S$, $g_{ij} = p_i q_j$.

Donc H_R et $H_S \Longrightarrow H_{(R \cup S)}$

Remarque. — On a utilisé le fait que $\overline{R} \cap \overline{S}$ et $R^* \cap S^*$ sont vides, et ces conditions sont bien suffisantes pour impliquer la conclusion du théorème VII, mais elles ne sont pas nécessaires et il est possible de trouver d'autres conditions sur R et S qui entraînent : (H_R et H_S) $\Longrightarrow H_{(R \cup S)}$ (en dehors aussi de la condition triviale R = S).

Le théorème VII est surtout important pour le corollaire qui suit. Notons d'abord que, si R n'est pas connexe, il sera possible de la décomposer en composantes connexes R_i telles que :

$$\bigcup_i R_i = R \text{ et } \overline{R}_i \cap \overline{R}_j = R_i^* \cap R_j^* = \phi \text{ pour tout } (i, j) \text{ où } i \neq j.$$

L'on a alors immédiatement le :

Corollaire. — Si R admet k composantes $R_1, R_2, \ldots, R_k$,

$$H_R \Longleftrightarrow (H_{R_1},\ H_{R_2},\ \ldots \text{ et } H_{R_k}).$$

Théorème VIII. — Si R est connexe (l, c) à t éléments $(t > l + c - 1)$, H_R fixe les valeurs de $v = t - l - c + 1$ des g_{ij} en foncton de $l + c - 1$ d'entre eux.

D'après le théorème II, il est possible d'extraire de R une partie connexe (l, c) à $l + c - 1$ éléments, soit S. H_S est vérifiée quels que soient les g_{ij} associés aux éléments de S, mais puisque S est connexe le système (p, q) obtenu est unique et tel que $p_i q_j$ est bien déterminé pour tout $(i, j) \in \overline{R}$ et en particulier pour tout $(i, j) \in R - S$. Donc, les g_{ij} associés aux $t - l - c + 1$ éléments de R — S ne peuvent pas être quelconques, mais sont fixés en fonction de (p, q), c'est-à-dire en fonction des g_{ij} associés aux éléments de S.

Corollaire. — Si R est (l, c) à t éléments et admet k composantes connexes, H_R fixe les valeurs de $v = t - l - c + k$ des g_{ij} en fonction de $l + c - k$ d'entre eux.

Supposons que la i^{me} composante connexe R_i soit (l_i, c_i) avec t_i éléments; H_{R_i} fixe alors $v_i = t_i - l_i - c_i + 1$ des g_{ij} en fonction de $l_i + c_i - 1$ d'entre eux. $H_{R_1}, H_{R_2} \ldots$ et H_{R_k} fixent donc $v = \sum_{i=1}^{k} v_i$ des g_{ij} en fonction de $\sum_{i=1}^{k} (l_i + c_i - 1)$ d'entre eux, et, puisque $\sum_{i=1}^{k} t_i = t$, $\sum_{i=1}^{k} l_i = l$ (car $\overline{R}_i \cap \overline{R}_j = \phi$ pour $i \neq j$), $\sum_{i=1}^{k} c_i = c$, la proposition est démontrée.

IV. Relations avec la théorie des graphes.

Considérons le graphe Γ non orienté dont les sommets sont les éléments de R, et dont l'ensemble $\bar{\Gamma}$ des arêtes est défini par :

$$\forall\ (i,\ j) \in \mathrm{R} \text{ et } \forall\ (h,\ k) \in \mathrm{R},\ [(i,\ j) \times (h,\ k)] \in \bar{\Gamma} \text{ si}$$

$$h = i,\ k \neq j \text{ et non } \exists (i,\ s) \in \mathrm{R} \text{ tel que } s \in]j,\ k[$$

$$\text{ou } k = j,\ k \neq i \text{ et non } \exists (s,\ j) \in \mathrm{R} \text{ tel que } s \in]i,\ h[$$

($s \in]j,\ k[$ signifiant : s strictement supérieur au plus petit et inférieur au plus grand des deux indices j et k, étant entendu que les deux ensembles *finis* d'indices peuvent toujours être ordonnés).

Avec cette définition, on voit d'abord que la connexité de R, telle qu'elle est définie plus haut, est équivalente à la connexité du graphe Γ.

On pourra montrer que, si R est $(l,\ c)$ à t éléments le nombre d'arêtes de Γ est $2t - l - c$.

Si R admet k composantes connexes, le nombre cyclomatique du graphe sera $\nu = t - l - c + k$, c'est-à-dire le nombre des g_{ij} fixés par la relation $\mathrm{H_R}$. Or, on sait que ν est le nombre de cycles linéairement indépendants du graphe, et, en fait, tout cycle de Γ correspond bien à une relation entre les g_{ij} imposée par $\mathrm{H_R}$. Si ces considérations permettent peut-être de rendre plus parlantes certaines propriétés, nous ne pensons pas cependant qu'elles auraient pu nous dispenser des démonstrations qui précèdent, ni même les abréger.

V. Étude d'une extension.

Si l'on considère maintenant $g_{ij} \in \mathrm{G_0} = \mathrm{G} \cup \{0\}$, où l'élément 0 est tel que, $\forall\ x \in \mathrm{G}$, $0x = x0 = 0$, on pourra toujours définir $\mathrm{H_R}$ de la même manière que précédemment, en remplaçant simplement G par $\mathrm{G_0}$. Si $\mathrm{H_R}$ est vraie, et si $g_{hk} = 0$, l'on aura :

$$p_h q_k = 0 \text{ ce qui entraîne } p_h = 0 \text{ ou } q_k = 0 \text{ d'où :}$$

$$\forall j \in \mathrm{R}_{h.},\ g_{hj} = 0 \text{ ou bien } \forall\ i \in \mathrm{R}_{.k},\ g_{ik} = 0.$$

Supposons par exemple la deuxième condition remplie et notons R′ la partie déduite de R en lui retranchant ses éléments de la forme $(i,\ k)$, k fixé, $i \in \mathrm{R}_{.k}$.

On aura encore ici, puisque $\mathrm{R}' \subset \mathrm{R}$: $\mathrm{H_R} \Longrightarrow \mathrm{H_{R'}}$

Mais on a aussi :

$$(\mathrm{H_{R'}} \text{ et } \forall\ i \in \mathrm{R}_{.k}\ g_{ik} = 0) \Longrightarrow \mathrm{H_R}$$

en effet, $\mathrm{H_{R'}}$ s'écrit :

$$\forall\ i \in \bar{\mathrm{R}}',\ \exists\ p_i \in \mathrm{G_0} \text{ et } \forall\ j \in \mathrm{R}'^{*},\ \exists\ q_j \in \mathrm{G_0}$$

$$\text{tels que } \forall\ (i,\ j) \in \mathrm{R}',\ g_{ij} = p_i q_j$$

et, avec $q_k = O$, on aura bien :

$$\forall i \in \bar{R},\ \exists\ p_i \in G_0 \text{ et } \forall j \in R^*,\ \exists\ q_j \in G_0$$
$$\text{tels que } \forall\ (i, j) \in R,\ g_{ij} = p_i q_j$$

c'est-à-dire H_R.

En recommençant un raisonnement analogue s'il existe $(i, j) \in R'$ pour lequel $g_{ij} = 0$, l'on voit que l'on peut ramener l'étude de H_R à l'étude précédente pour une partie R' telle que : $\forall\ (i, j) \in R'$, $g_{ij} \in G$.

Cependant, même si R est connexe, R' ne l'est pas nécessairement. Dans ces conditions, le théorème III précédent n'est pas nécessairement vérifié dans le cas présent, c'est-à-dire en remplaçant G par G_0; il en sera de même des théorèmes V - VI et VIII. Par contre, les théorèmes IV et VII restent valides.

VI. Propriétés d'une nouvelle relation.

1. Définition.

On considère toujours une partie R de C; on associe à chaque élément de R un élément g_{ij} de G.

La relation H_R^- est définie par :

$$\forall\ i \in \bar{R} \cup R^*,\ \exists\ f_i \in G \text{ tels que } \forall\ (i, j) \in R\ g_{ij} = f_i f_j^{-1}$$

(on notera f l'ensemble des f_i ainsi définis.)

Remarque. — Alors que H_R est invariante sous une renumérotation quelconque des indices appartenant à $\bar{R}$ et de ceux appartenant à R*, il n'en est pas ainsi de H_R^- : cette opération peut être faite sur les éléments de $\bar{R} - R^*$ ou ceux de $R^* - \bar{R}$, mais si l'on change le nom des indices de $\bar{R} \cap R^*$, ce doit être de la même façon pour ceux de $\bar{R}$ et ceux de R*.

2. Propriétés.

a) On a évidemment: $H_R^- \Longrightarrow H_R$ (mais la réciproque est fausse si $R^* \cap \bar{R}$ n'est pas vide).

b) *Unicité.* — En utilisant la propriété analogue de H_R, on montre facilement que si R est connexe l'ensemble f lié à H_R^- est unique en ce sens que pour toute autre solution f', il existe $k \in G$ tel que $f_i' = f_i k$ pour tout $i \in \bar{R} \cup R^*$.

c) Si R et S sont connexes (l, c) et $S \subset R$, alors :

$$(H_R \text{ et } H_S^-) \iff H_R^-$$

En effet, $H_R^- \Longrightarrow H_S^-$ est évident, et $H_R^- \Longrightarrow H_R$ a été noté plus haut. Réciproquement, pour tout $(i, j) \in R$ on a $g_{ij} = p_i q_j$ en vertu de H_R, et pour tout

$(i, j) \in S$, $g_{ij} = f_i f_j^{-1}$ en vertu de H_S^- ; de là pour tout $(i, j) \in S$ on a

$$g_{ij} = p_i q_j = f_i f_j^{-1}$$

et en utilisant le théorème d'unicité il vient :

$$\forall i \in \bar{R},\ p_i = f_i k \text{ et } \forall j \in R^*,\ q_i = k^{-1} f_j^{-1}$$

$$\text{donc } \forall\ (i, j) \in R,\ g_{ij} = p_i q_j = f_i f_j^{-1}$$

H_R^- est ainsi vérifiée.

d) Si R est connnexe (l, c) à t éléments, le nombre des relations entre les g_{ij} imposées par H_R est

$$v' = t - l - c + 1 + d \text{ où } d = Card\,(\bar{R} \cap R^*).$$

Démonstration. — On peut extraire de R une partie S connexe (l, c) à $l + c - 1$ éléments (Théorème II). Pour $(i, j) \in S$ on peut écrire $g_{ij} = f_i q_j$ (de façon unique au sens utilisé jusqu'ici). D'après la propriété précédente H_R^- est équivalente à $H_R \cap H_S^-$; or, on peut toujours poser $q_j = f_j^{-1}$ pour $j \in R^* - \bar{R}$, mais H_S^- impose entre les g_{ij}, où (i, j) appartient à S, les d relations $q_j = f_j^{-1}$ pour $j \in \bar{R} \cap R^*$, d'autre part, H_R fixe en fonction de ces g_{ij} où $(i, j) \in S$, les $t - l - c + 1$ g_{ij} où $(i, j) \in R - S$.

B. — INTRODUCTION A L'ÉTUDE STATISTIQUE DES TABLES DE CONTINGENCE TRONQUÉES

I. Notations et définitions.

On considère maintenant une population dont les éléments présentent chacun un caractère A sous une forme α_i et un caractère B sous une forme β_j. L'association des caractères α_i et β_j sera notée $\alpha_i \beta_j$ $[(i, j) \in C]$.

La probabilité qu'un individu tiré au hasard de cette population ait le caractère $\alpha_i \beta_j$ est :

$$Pr\,[\alpha_i \beta_j] = p_{ij} \qquad \Big(\sum_{(i,j) \in C} p_{ij} = 1,\ p_{ij} \geqslant 0\Big) \qquad (1)$$

Sachant que (i, j) appartient à un sous-ensemble donné R de C, cette probabilité est (en supposant $\sum_{(i,j) \in R} p_{ij} \neq 0$) :

$$Pr[\alpha_i \beta_j / (i, j) \in R] = p_{ij} \Big(\sum_{(i,j) \in R} p_{ij}\Big)^{-1} = P_{ij} \qquad (2)$$

II. L'hypothèse H_R ; ses différentes formes.

R étant toujours une partie de C, supposons d'abord que, pour tout $(i, j) \in R$, p_{ij} est strictement positif. Prenant pour G le groupe multiplicatif des nombres réels strictement positifs, on peut associer à tout (i, j) de R, la probabilité $p_{ij} = Pr[\alpha_i \beta_j]$; la relation H_R définie plus haut comme étant

une propriété de ces nombres, sera considérée maintenant comme une hypothèse statistique.

En associant aux éléments de R les nombres P_{ij} au lieu des p_{ij}, on pourra définir toujours de la même façon une relation H_R' entre ces nombres.

Il est immédiat que H_R' et H_R sont équivalentes puisque, pour tout (i, j) de R, le rapport P_{ij}/p_{ij} est constant.

Enfin H_R et H_R' sont équivalentes à H_R'' et H_R''' définies respectivement par :

$$H_R'' : \forall\ i \in \bar{R},\ \exists\ p_i \text{ tels que}$$

$$\forall\ (i, j) \in R,\ Pr\ [\alpha_i/\beta_j,\ (i, j) \in R] = p_i \left(\sum_{h \in R._j} p_h \right)^{-1}$$

$$H_R''' : \forall i \in R^*,\ \exists\ q_j \text{ tels que}$$

$$\forall (i, j) \in R,\ Pr\ [\beta_j/\alpha_i,\ (i, j) \in R] = q_j \left(\sum_{h \in R_{i.}} q_h \right)^{-1}$$

Montrons par exemple : $H_R' \Longleftrightarrow H_R''$

$$\text{On a } Pr\ [\alpha_i/\beta_j,\ (i, j) \in R] = \frac{Pr\ [\alpha_i \beta_j/(i, j) \in R]}{Pr\ [\beta_j/(i, j) \in R]} = \frac{P_{ij}}{\sum_{h \in R._j} P_{hj}}$$

donc, si H_R' est vraie, il existe des p_i et q_j tels que :

$$Pr\ [\alpha_i/\beta_j,\ (i, j) \in R] = \frac{p_i q_j}{\sum_{h \in R._j} p_h q_j} = \frac{p_i}{\sum_{h \in R._j} p_h}$$

d'où : $H_R' \Longrightarrow H_R''$.

Réciproquement, si H_R'' est vraie, il existe des p_i tels que :

$$Pr\ [\alpha_i/\beta_j,\ (i, j) \in R] = \frac{p_i}{\sum_{h \in R._j} p_h}$$

et $Pr\ [\alpha_i \beta_j/(i, j) \in R] = Pr\ [\beta_j/(i, j) \in R]\ Pr\ [\alpha_i/\beta_j,\ (i, j) \in R]$

nous donne : $P_{ij} = \sum_{h \in R._j} P_{hj} \cdot \frac{p_i}{\sum_{h \in R._j} p_h}$

On peut alors poser, puisque $\sum_{h \in R._j} P_{hj}$ et $\sum_{h \in R._j} p_h$ ne dépendent que de j

$$q_j = \left(\sum_{h \in R._j} P_{hj} \right) \left(\sum_{h \in R._j} p_h \right)^{-1} \text{ pour tout } j \in R^*$$

d'où $H_R'' \Longrightarrow H_R'$.

On montrera de même l'équivalence de H_R' et H_R'''. Finalement :

$$H_R \Longleftrightarrow H_R' \Longleftrightarrow H_R'' \Longleftrightarrow H_R'''.$$

Remarque. — H_C ou $H_{\overline{R}}$ sont des hypothèses d'indépendance classiques.

Discussion. — Si certains des p_{ij}, $(i, j) \in R$, sont nuls, on utilisera les résultats du A — V. Dans la pratique, il suffira de voir si les p_{ij} nuls sont compatibles avec H_R ou non, et, dans le premier cas, d'étudier au lieu de H_R l'hypothèse $H_{R'}$, où $R' \subset R$ est telle que p_{ij} est nul pour $(i, j) \in R - R'$, alors que, pour tout $(i, j) \in R'$, p_{ij} est strictement positif. On pourra donc dans l'étude de H_R se ramener toujours à des parties R telles que p_{ij} est différent de zéro pour tout (i, j) de R.

III. Hypothèse H_R et interactions.

Supposons $p_{ij} > 0$ pour tout $(i, j) \in C$. A la suite de GOODMAN (1964 a), nous désignerons sous le nom *d'interactions du premier ordre* dans une table de contingence $r \times s$, les quantités :

$$\delta(u, p) = \sum_{(i.j) \in C} u_{ij} \operatorname{Log} p_{ij}$$

où u est un vecteur de $\mathcal{R}^{rs}$ à rs composantes u_{ij}, $(i, j) \in C$, telles que :

$$\forall\, i \in \overline{R} \sum_{j=1}^{s} u_{ij} = 0 \qquad \text{et} \qquad \forall\, j \in R^* \sum_{i=1}^{r} u_{ij} = 0$$

($\mathcal{R}$ désigne l'ensemble des nombres réels).

Sous l'hypothèse H_R, l'on aura :

$$\delta(u, p) = \sum_{(i,j) \in C} u_{ij} \operatorname{Log} p_{ij} = 0$$

pour tout u tel que :

$$\forall (i, j) \in C - R, \qquad u_{ij} = 0 \tag{1}$$

$$\forall\, i \in \overline{R}, \sum_{j \in R_{i.}} u_{ij} = 0 \qquad \text{et} \qquad \forall\, j \in R^*, \sum_{i \in R_{.j}} u_{ij} = 0 \tag{2}$$

En effet, si u vérifie (1) et (2) et si H_R est vraie, l'on a :

$$\delta(u, p) = \sum_{(i,j) \in R} u_{ij} (\operatorname{Log} p_i + \operatorname{Log} q_j)$$

$$= \sum_{i \in \overline{R}} \operatorname{Log} p_i \sum_{j \in R_{i.}} u_{ij} + \sum_{j \in R^*} \operatorname{Log} q_j \sum_{i \in R_{.j}} u_{ij} = 0$$

Nous allons montrer que la réciproque est vraie. On désignera par u à partir de maintenant un vecteur de $\mathcal{R}^t$ ($t = \operatorname{Card}(R)$) de composantes u_{ij}, $(i, j) \in R$. On notera I (R) l'ensemble des vecteurs de $\mathcal{R}^t$ tels que (2) est vérifié : I (R) est un sous-espace vectoriel de $\mathcal{R}^t$. Pour tout $i \in \overline{R}$,

notons $e^{(i)}$ le vecteur de $\mathcal{R}^t$ de composantes $e_{hk}^{(i)} = \delta_{ih}$ (δ_{ih} est le symbole de KRONECKER). De même, $f^{(j)}$ est le vecteur de $\mathcal{R}^t$ de composantes $f_{hk}^{(j)} = \delta_{jk}$.

On a pour tout $i \in \overline{R}$:

$$\sum_{(h,k)\in R} e_{hk}^{(i)} u_{hk} = \sum_{(h,k)\in R} \delta_{ih} u_{hk} = \sum_{k\in R_i.} u_{ik}$$

et de même pour tout $j \in R^*$:

$$\sum_{(h,k)\in R} f_{hk}^{(j)} u_{hk} = \sum_{h\in R._j} u_{hj}$$

Donc la condition (2) est équivalente à l'orthogonalité de u et des vecteurs $e^{(i)}$ et $f^{(j)}$.

Soit maintenant E le sous-espace vectoriel de $\mathcal{R}^t$ engendré par $e^{(i)}$, $i \in \overline{R}$, et F le sous-espace engendré par $f^{(j)}$, $j \in R^*$. D'après ce que l'on vient de voir, I (R) est l'espace orthogonal complémentaire de E + F.

Supposons :

$$\forall\, u \in I\,(R) \qquad \delta\,(u,\, p) = \sum_{(i,j)\in R} u_{ij} \operatorname{Log} p_{ij} = 0$$

ceci entraîne que le vecteur de composantes $\operatorname{Log} p_{ij}$ appartient à E + F. Or, tout vecteur g de E + F peut s'écrire :

$$g = \sum_{i\in\overline{R}} \lambda_i\, e^{(i)} + \sum_{j\in R^*} \mu_j f^{(j)}$$

donc :

$$\forall\ (h,\, k) \in R \qquad g_{hk} = \sum_{i\in\overline{R}} \lambda_i\, \delta_{ih} + \sum_{j\in R^*} \mu_j\, \delta_{jk} = \lambda_h + \mu_k$$

$\operatorname{Log} p_{hk}$ s'écrira donc $\operatorname{Log} p_{hk} = \lambda_h + \mu_k$ et par suite p_{hk} pourra s'écrire sous la forme $p_{hk} = p_h\, q_k$ pour tout $(h,\ k) \in R$. Dans ces conditions H_R est vérifiée. On a donc le :

Théorème IX. — H_R est équivalente à l'ensemble des relations entre les interactions du premier ordre :

$$\forall\, u \in I\,(R), \qquad \delta\,(u,\ p) = 0$$

Remarque. — En vertu du théorème d'unicité III, on peut montrer que, si R est connexe, E + F est de rang $l + c - 1$. En effet, en écrivant que la combinaison linéaire :

$\sum_{i\in\overline{R}} \lambda_i\, e^{(i)} + \sum_{j\in R^*} \mu_j\, f^{(j)}$ est nulle, on obtient :

$$\forall\,(h,\ k) \in R,\ \lambda_h + \mu_k = 0$$

Cette condition est réalisée pour

$$\lambda_h = \lambda \ (\forall h \in R) \text{ et } \mu_k = -\lambda \ (\forall k \in R^*);$$

mais cette solution est unique en vertu du théorème III puisque R est connexe (associer à tout $(h, k) \in R$ l'élément zéro, G étant le groupe additif des nombres réels).

Donc I (R) est de rang $t - l - c + 1$ et l'ensemble des relations $\delta(u, p) = 0$, $u \in I(R)$, contient $v = t - l - c + 1$ relations indépendantes (on retrouve un résultat précédent). En notant $u^{(k)}$ $(k = 1, 2, \ldots, v)$ v vecteurs indépendants de I(R), il est possible d'extraire des relations $\delta(u, p) = 0$, v relations indépendantes $\delta(u^{(k)}, p) = 0$ qui sont équivalentes à H_R.

IV. Problèmes statistiques concernant l'hypothèse H_R.

Notons d'abord qu'il sera possible de se restreindre à l'étude de parties connexes, en utilisant le théorème VII et son corollaire.

Dans ces conditions, si l'on écrit, sous l'hypothèse H_R, $P_{ij} = p_i q_j$, l'ensemble (p, q) est unique (théorème III) à une proportionnalité près, et bien déterminé si l'on fixe $\sum_{i \in \overline{R}} p_i$ (ou $\sum_{j \in R^*} q_j$); si bien que, sous sous l'hypothèse H_R, ces nombre p_i et q_j peuvent être considérés comme les paramètres de la distribution conditionnelle des caractères $\alpha_i \beta_j$ sachant que $(i, j) \in R$.

On pourra se proposer, *à partir d'un échantillon tiré au hasard de la population parente :*

a) D'estimer les paramètres p_i et q_j.

b) De réaliser un test de l'hypothèse H_R.

Selon la méthode d'échantillonnage, les problèmes posés diffèrent, mais nous verrons qu'ils sont très voisins dans la pratique. On étudiera les cas où l'échantillon est extrait sans restriction de la population complète, puis où l'échantillon obtenu est extrait uniquement du sous-ensemble des éléments de la population parente ayant les caractères $\alpha_i \beta_j$ où $(i, j) \in R$, la taille de l'échantillon étant seule fixée, ou bien aussi certaines « marges », par exemple le nombre d'individus ayant le caractère β_j, $j \in R^*$ (étude de plusieurs distributions multinomiales, supposées homogènes, dont certaines peuvent être « tronquées »).

Un échantillon est caractérisé par la fréquence des diverses catégories; l'ensemble des fréquences relatives aux caractères $\alpha_i \beta_j$ où $(i, j) \in R$ (R strictement contenue dans C) sera appelé tableau de corrélation (ou table de contingence) tronqué (*e*) ou incomplet (*e*). R sera désignée comme la partie de C associée à ce tableau.

Les problèmes posés par les tables de contingence tronquées ne sont pas limités à ceux qui concernent une hypothèse du type H_R; nous commencerons par l'étude de ces derniers, mais nous en aborderons de nouveaux par la suite.

V. Domaines d'application.

L'étude statistique des tables de contingence tronquées et de H_R en particulier peut être utile dans les cas suivants :

a) Certaines données sont manquantes par accident dans un échantillon. Si elles concernent les caractères α_i β_j où $(i, j) \in C - R$, on pourra étudier H_R faute de pouvoir étudier H_C.

b) Il n'est possible de sélectionner de la population parente que des individus ayant les caractères α_i β_j où $(i, j) \in R$.

c) En présence d'une table de contingence ordinaire (échantillon extrait au hasard de la population parente complète), l'étude de H_R peut présenter un intérêt particulier dans le cas où H_C est fausse. On peut ainsi envisager d'analyser la forme de la corrélation.

Notons enfin que les résultats concernant l'hypothèse H_R peuvent avoir quelque intérêt dans l'étude de problèmes apparemment assez différents : nous en verrons un exemple au Chapitre V.

VI. Travaux antérieurs.

Si aucune étude tout à fait systématique ne semble avoir été faite, divers auteurs ont étudié des problèmes particuliers relatifs aux tables de contingence tronquées. Un des premiers en date doit être WAITE (1915) ; HARRIS (1927, 1929) envisage deux questions dont la première est identique à celle de WAITE.

Le problème que se posent ces deux auteurs est différent de ceux indiqués plus haut puisqu'ils désirent essentiellement calculer un « coefficient de contingence ». Ils sont cependant amenés à chercher des « fréquences théoriques » sous une hypothèse H_R où R est une partie « triangulaire » simple $[(i, j) \in R$ si $i \leqslant j]$; la solution de HARRIS est critiquée par PEARSON (1930) qui avait suggéré sur des bases intuitives la solution de Waite; cette dernière s'accorde avec les résultats que l'on obtiendra par la suite par la méthode du maximum de vraisemblance.

Le second problème de Harris concerne un cas où R n'est pas connexe : un traitement par séparation des deux parties connexes est préconisé par PEARSON (1930).

Les études plus récentes rentrent toutes dans le cadre fixé plus haut :

WATSON (1956) envisage le problème (a) des données manquantes pour un cas particulier et donne quelques indications sur le cas général.

KASTENBAUM (1958) traite un autre cas particulier.

BATSCHELET (1960 a), puis GEPPERT (1961) ont étudié les tables triangulaires lorsque les totaux par colonne (ou par ligne) sont fixés. BATSCHELET (1960 b) donne une application à un intéressant problème pratique du type (b) concernant des distributions multimoniales tronquées.

ASANO (1965) généralise sur quelques points les études de ces derniers auteurs sur la réunion de distributions multinomiales tronquées.

CHAPITRE II

PROBLÈMES D'ESTIMATION POUR UNE TABLE DE CONTINGENCE TRONQUÉE

Une population parente étant donnée, pour laquelle H_R est supposée vérifiée, on étudie dans cette partie les problèmes d'estimation des paramètres inconnus attachés à cette population.

I. Distribution d'un échantillon — Statistiques exhaustives.

Pour les raisons données au chapitre précédent, R sera (sauf mention du contraire) connexe.

Un échantillon est constitué d'éléments tirés au hasard, de façon non exhaustive, de la population parente. On notera x_{ij} la fréquence observée de l'élément $\alpha_i \beta_j$.

H_R peut être étudiée pour différents types d'échantillons :

1er cas. — Echantillon de taille fixée extrait de la population réduite aux éléments $\alpha_i \beta_j$ où $(i, j) \in R$. Ici l'hypothèse H_R s'écrit : $P_{ij} = Pr\,[\alpha_i \beta_j / (i, j) \in R] = p_i q_j$.

La vraisemblance de l'échantillon est donc, en appelant p (resp. q) l'ensemble des p_i pour $i \in \bar{R}$ (resp. l'ensemble des q_j pour $j \in R^*$) et en désignant par n la taille de l'échantillon :

$$L_1(p, q) = \frac{n!}{\prod_{(i,j) \in R} (x_{ij})!} \prod_{(i,j) \in R} (p_i q_j)^{x_{ij}}$$

$$= \frac{n!}{\prod_{(i,j) \in R} (x_{ij})!} \prod_{i \in \bar{R}} p_i^{a_i} \prod_{j \in R^*} q_j^{b_j}$$

avec $a_i = x_{i.} = \sum_{j \in R_{i.}} x_{ij}$, $b_j = x_{.j} = \sum_{i \in \bar{R}_{.j}} x_{ij}$

et la relation : $\sum_{(i,j) \in R} p_i q_j = 1$ (1)

Puisque R est supposée connexe, les paramètres p_i et q_j sont parfaitement déterminés si l'on pose par exemple :

$$\sum_{i \in \bar{R}} p_i = 1 \qquad (2)$$

$$\text{ou bien} \quad \sum_{j \in R^*} q_j = 1 \qquad (2')$$

On notera a (resp. b) l'ensemble des a_i pour $i \in \overline{R}$ (resp. l'ensemble des b_j pour $j \in R^*$). D'après la forme de L_1 la statistique (a, b) est exhaustive pour les paramètres (p, q) (cf : critère de NEYMAN-FISHER).

L'existence d'une telle statistique exhaustive, nous incite à chercher des estimateurs des paramètres par la méthode du maximum de vraisemblance, puisque cette méthode conduira certainement à des estimateurs fonctions de ces seules statistiques exhaustives. Cette remarque est surtout valable pour de petits échantillons; pour de grands échantillons, il sera peut-être aussi intéressant de connaître n'importe quel autre estimateur R.B.A.N. (1) ayant d'aussi bonnes propriétés *asymptotiques*.

2e cas. — On dispose d'un échantillon extrait de la population parente complète.

Sa vraisemblance sous H_R est :

$$L_2 = \frac{\left(\sum_{(i,j) \in C} x_{ij}\right)!}{\prod_{(i.j) \in C} (x_{ij})!} \prod_{(i.j) \in C-R} p_{ij}^{x_{ij}} \prod_{i \in \overline{R}} p_i^{a_i} \prod_{j \in R^*} q_j^{b_j} \left[1 - \sum_{(i,j) \in C-R} p_{ij}\right]^n$$

avec $p_{ij} = Pr\,[\alpha_i \beta_j]$ pour $(i, j) \in C - R$

$p_i q_j = Pr\,[\alpha_i \beta_j / (i, j) \in R]$ pour $(i, j) \in R$

$n = \sum_{(i,j) \in R} x_{ij}$ (ici n est donc aléatoire).

La statistique (a, b) est toujours exhaustive pour (p, q), donc, a fortiori, l'ensemble des x_{ij} pour (i, j) appartenant à R. Dans ces conditions, il suffira, pour estimer (p, q), de considérer la distribution des x_{ij} tels que $(i, j) \in R$ et nous sommes ramenés au cas précédent.

Remarque. — On obtiendra un résultat analogue si l'on étudie H_R à partir d'un échantillon extrait de la population réduite aux éléments $\alpha_i \beta_j$ où $(i, j) \in S$ si $R' \subset S \subset C$.

3e cas. — Les totaux marginaux b sont fixés.

Nous écrivons H_R sous la forme équivalente H_R'', on a :

$$Pr\,[\alpha_i / \beta_j\ ,\ (i, j) \in R] = \frac{p_i}{\sum_{h \in R_{.j}} p_h}$$

1. Regular Best Asymptotically Normal : estimateur asymptotiquement normal et à variance minimum, régulier en ce sens qu'il est fonction des fréquences observées continûment dérivable par rapport à chacune d'elles (cf. NEYMAN, 1949).

et la vraisemblance de l'échantillon est :

$$L_3(p) = \frac{\prod_{j \in R^*} b_j !}{\prod_{(i,j) \in R} (x_{ij}) !} \prod_{i \in \bar{R}} p_i^{a_i} \prod_{j \in R^*} \left(\sum_{h \in R._j} p_h \right)^{-b_j}$$

La statistique a est exhaustive pour p.

D'autre part, l'on notera que, dans ce cas, la condition $\sum_{i \in \bar{R}} p_i = 1$ s'introduit naturellement si l'on veut $p_i = \Pr[\alpha_i]$, ce qui est possible : il suffit de remarquer que, si H_R est vraie, $\Pr[\alpha_i/\beta_j, (i, j) \in R]$ s'exprime de la même façon en fonction des p_i et en fonction des $\Pr[\alpha_i]$, et d'invoquer l'unicité de p (R est connexe). C'est pour cette raison que nous avons introduit la condition (2) alors que d'autres sont équivalentes (par exemple, fixer l'un des p_i).

II. Équations de vraisemblance.

Considérons le premier cas; ces équations s'obtiennent en rendant L_1 extremum compte tenu des relations (1) et (2) On démontre que les estimateurs $\hat{p}_i$ et $\hat{q}_j$ doivent vérifier le système :

$$(I) \quad \begin{cases} \hat{p}_i \sum_{j \in R_i.} \hat{q}_j = \frac{a_i}{n} & \text{pour tout } i \in \bar{R} \qquad (3) \\ \hat{q}_j \sum_{i \in R._j} \hat{p}_i = \frac{b_j}{n} & \text{pour tout } j \in R^* \qquad (3') \end{cases}$$

Dans ce système, l'on pourra remplacer soit (3) soit (3') respectivement par :

$$\hat{p}_i \sum_{j \in R_i.} \frac{b_j}{\sum_{h \in R._j} \hat{p}_h} = a_i \text{ pour tout } i \in \bar{R} \qquad (4)$$

ou :

$$\hat{q}_j \sum_{i \in R._j} \frac{a_i}{\sum_{h \in R_i.} \hat{q}_h} = b_j \text{ pour tout } j \in R^* \qquad (4')$$

On doit adjoindre à ces équations l'une des deux relations :

$$\sum_{i \in \bar{R}} \hat{p}_i = 1 \tag{5}$$

ou

$$\sum_{j \in R^*} \hat{q}_j = 1 \tag{5'}$$

Écriture matricielle.

On pourra supposer sans perdre de généralité :

$$\bar{R} = \{1, 2, .., l\} \qquad R^* = \{1, 2, .., c\}$$

On notera $\hat{p}$ la matrice colonne $[\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_l]'$

— $\hat{q}$ » » $[\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_c]'$

— a » » $[a_1, a_2, \ldots, a_l]'$

— b » » $[b_1, b_2, \ldots, b_c]'$

La matrice à l lignes et c colonnes $T = \| t_{ij} \|$ que l'on pourra appeler, par analogie avec la théorie des blocs incomplets, matrice d'incidence [cf. Dugué 1958, p. 274], sera définie par : $t_{ij} = 1 \qquad \forall (i, j) \in R$

$$t_{ij} = 0 \qquad \forall (i, j) \in \bar{R} - R$$

Étant donné un vecteur ligne (ou colonne) $x = [x_1 x_2 .. x_k]$ on notera x_D la matrice carrée diagonale d'ordre k dont les termes diagonaux sont

$$x_1, x_2, \ldots, x_k.$$

Dans ces conditions le système (1) s'écrit :

$$\hat{p}_D T \hat{q} = \frac{1}{n} a \tag{6}$$

$$\hat{q}_D T' \hat{p} = \frac{1}{n} b \tag{6'}$$

Si l'on possède un échantillon extrait de la population complète les $\hat{p}_i$ et $\hat{q}_j$ vérifient, sous l'hypothèse H_R, les mêmes équations que ci-dessus. L'on a, en outre, $\hat{p}_{ij} = \frac{x_{ij}}{N}$ pour $(i, j) \in C - R$, $N = \sum_{(i,j) \in C} x_{ij}$ étant la taille de l'échantillon.

Si les b_j sont fixés, L_3 sera extremum pour $\hat{p}$ vérifiant les équations (4) et (5).

On pourra ici encore, introduire des quantités $\hat{q}_j$ définies par (3') pour retrouver exactement les mêmes équations que dans les cas précédents. (Nous verrons plus loin que cette remarque peut s'avérer utile.)

Il y a donc finalement absolue identité entre les problèmes d'estimation posés par les différents cas d'échantillonnage.

« *Fréquences théoriques* ».

On appellera f_{ij} l'espérance mathématique sous l'hypothèse H_R de la v. a. x_{ij} [2], soit $f_{ij} = E(x_{ij}/H_R)$; les « fréquences théoriques » sont les quantités $\hat{f}_{ij}$ obtenues en remplaçant dans l'expression des f_{ij} les paramètres inconnus par leurs estimations. En reprenant les trois cas d'échantillonnage, on a :

1er cas : $\hat{f}_{ij} = n \hat{p}_i \hat{q}_j$ pour tout $(i, j) \in R$.

2e cas : $\hat{f}_{ij} = x_{ij}$ pour tout $(i, j) \in C - R$.

$$\hat{f}_{ij} = N \hat{p}_i \hat{q}_j \left(1 - \sum_{(h,k) \in C-R} \hat{p}_{hk}\right) = n \hat{p}_i \hat{q}_j \text{ pour tout } (i, j) \in R.$$

3e cas : $$\hat{f}_{ij} = \frac{b_j \hat{p}_i}{\sum_{h \in R_{\cdot j}} \hat{p}_h} \text{ pour tout } (i, j) \in R.$$

ce qui redonne la formule du premier cas en introduisant $\hat{q}_j$ de la façon indiquée plus haut.

Ainsi, l'on a toujours sous H_R, pour tout (i, j) appartenant à R :

$$\hat{f}_{ij} = n \hat{p}_i \hat{q}_j = b_j \frac{\hat{p}_i}{\sum_{h \in R_{\cdot j}} \hat{p}_h} = a_i \frac{\hat{q}_j}{\sum_{h \in R_{i \cdot}} \hat{q}_h} \qquad (7)$$

En notation matricielle, la matrice $F = n \hat{p}_D T \hat{q}_D$ renferme les fréquences théoriques $\hat{f}_{ij}$, ses éléments étant $\hat{f}_{ij}$ pour $(i, j) \in R$ et zéro dans le cas contraire.

III. Étude des équations de vraisemblance.

Avant d'étudier la question de la résolution proprement dite des équations de vraisemblance, nous envisagerons d'abord les diverses façons d'aborder ce problème, comment on peut essayer de le simplifier, si même une solution existe et, dans ce cas, si elle est unique.

---

2. Tant qu'aucune confusion ne sera à craindre, nous noterons de la même façon une variable aléatoire (v. a.) et sa valeur observée.

Rappelons que les paramètres p_i et q_j sont des nombres réels positifs; nous imposerons donc aux estimations $\hat{p}_i$ et $\hat{q}_j$ la même condition. Notant $\mathcal{R}_+^k$ le sous-ensemble de $\mathcal{R}^k$ dont les points ont toutes les coordonnées positives, on appellera dès lors solution des équations de vraisemblance, uniquement une solution p, q telle que $p \in \mathcal{R}_+^l$, $q \in \mathcal{R}_+^c$. Par la suite, et tant qu'aucune confusion ne sera à craindre, nous remplacerons $\hat{p}$, $\hat{q}$, $\hat{f}$ par p, q, f.

1. Généralités.

— Dans la pratique, il faudra résoudre *le plus simple* des deux systèmes (4) et (4′); si c'est par exemple (4), nous obtiendrons d'abord p, et de là q très simplement par (3′) et F par (7). Cette remarque banale montre toutefois combien il peut être important dans la pratique d'introduire q même dans le cas où les marges b sont fixées, et quel avantage on peut tirer d'une étude simultanée des trois cas d'échantillonnage.

— Si l'on connaît F, on peut en déduire une solution p, q. Ayant $\frac{f_{ij}}{n} = p_i q_j$ pour tout $(i, j) \in R$ et R connexe, l'application $F \to (p, q)$ est univoque, du moins si $f_{ij} \neq 0$ pour tout $(i, j) \in R$. Le calcul pratique ne présente aucune difficulté.

— On pourra réduire la dimension du système (4) chaque fois que, pour deux indices i_1 et i_2, $R_{i_1 \cdot}$ sera égal à $R_{i_2 \cdot}$ (de même pour (5) si $R_{\cdot j_1} = R_{\cdot j_2}$). Les propositions montrées ci-dessous signifient dans la pratique, que l'on peut grouper dans une table de contingence incomplète toutes les lignes présentant des « troncatures identiques »; les équations obtenues pour la table avec groupement donnent des estimations p_i' qu'il suffit ensuite de partager en parties proportionnelles aux totaux marginaux des lignes de la table initiale qui ont été groupées en la i^{me} ligne de la table « condensée ». (même chose pour les colonnes.)

On pourra dès lors, si l'on veut, restreindre notre étude à des cas où les $R_{i\cdot}$ sont tous différents, ainsi que les $R_{\cdot j}$.

Précisons tout ceci :
si l'on suppose $R_{(l-1)\cdot} = R_{l\cdot}$ et si l'on note R′ la partie de C obtenue en retranchant à R les éléments de la forme (l, j) où $j \in R_{l\cdot}$, on aura :

$$\bar{R}' = \bar{R} - \{l\},\ R_{i\cdot} = R'_{i\cdot} \text{ pour tout } i \in \bar{R}',$$
$$R^* = R'^*,\ R'_{\cdot j} = R_{\cdot j} - \{l\} \text{ pour tout } j \in R^*$$

Les équations (4) et (5) pourront s'écrire, respectivement pour la table initiale et pour la table condensée :

$$\begin{cases} \text{(E)} \quad p_i \sum_{j \in R_{i\cdot}} \dfrac{b_j}{\sum_{h \in R_{\cdot j}} p_h} = a_i \quad \text{pour tout } i \in \bar{R} \\ (e) \quad \sum_{i \in \bar{R}} p_i = 1 \end{cases}$$

et

$$\begin{cases} \text{(E')} \quad p'_i \sum_{j \in R_{i\cdot}} \dfrac{b_j}{\sum_{h \in R'_{\cdot j}} p'_h} = a'_i \quad \text{pour tout } i \in \bar{R}' = \bar{R} - \{l\} \\ (e') \quad \sum_{i \in \bar{R}'} p'_i = 1 \end{cases}$$

avec $a_i = a'_i$ pour tout $i \in \bar{R}'$

$a_{l-1} + a_l = a'_{l-1}$

Proposition I. — Si $p \in \mathcal{R}_+^l$ est solution de (E) et si l'on définit p' par :

$p_i' = p_i \quad \forall\ i \in \bar{R}' - \{l-1\}.$

$p'_{l-1} = p_{l-1} + p_l$

p' est solution de (E')

En effet, sous nos hypothèses : $l \in R_{\cdot j} \Leftrightarrow l-1 \in R_{\cdot j} \Leftrightarrow l-1 \in R'_{\cdot j}$,

De là : $$\sum_{h \in R_{\cdot j}} p_h = \sum_{h \in R'_{\cdot j}} p'_h \qquad \text{pour tout } j \in R^*$$

Ainsi les $(l-2)$ relations (E) obtenues pour $i \in \bar{R}' - \{l-1\}$ impliquent les $(l-2)$ relations de (E') correspondantes. Ajoutant membre à membre les deux relations restantes de (E), en tenant compte de $R_{(l-1)\cdot} = R_{l\cdot}$, on voit que p' vérifie la dernière relation de (E'). D'autre part, l'on a : $p' \in \mathcal{R}_+^{l-1}$.

Proposition II. — Si $p' \in \mathcal{R}_+^{l-1}$ est une solution de (E') et si l'on définit p par : $p_i = p'_i$ pour tout $i \in \bar{R}' - \{l-1\}$

$$p_{l-1} = \frac{a_{l-1}}{a_{l-1} + a_l}\ p'_{l-1}$$

$$p_l = \frac{a_l}{a_{l-1} + a_l}\ p'_{l-1}$$

p est solution de (E).

La démonstration est analogue à la précédente pour $(l-2)$ équations de (E).

Pour $i = l-1$, le premier membre de l'équation de (E) est :

$$\frac{a_{l-1}}{a_{l-1}+a_l}\, p'_{l-1} \sum_{j \in R_{(l-1)\cdot}} \frac{b_j}{\sum_{h \in R'_{\cdot j}} p'_h}$$

qui est égal, puisque p' est solution de (E'), à

$$\frac{a_{l-1}}{a_{l-1}+a_l}\, a'_{l-1} = a_{l-1} \qquad \text{(id. pour } i = l\text{)}.$$

Remarques. — 1) Les solutions p et p' de (E) et (E') envisagées ci-dessus sont telles que, si p vérifie (e), p' vérifie (e') et réciproquement.

2) Si (E', e') admet deux solutions distinctes p' et p'^* (appartenant à $\mathcal{R}_+^{l-1}$) les solutions p et p^* correspondantes de (E, e) appartiennent à $\mathcal{R}_+^{l}$ et sont distinctes. En effet, si p' et p'^* diffèrent, il existe au moins deux indices i appartenant à $\overline{R'}$ tels que $p' \neq p'_i{}^*$; l'un d'eux (soit h) est donc différent de $l-1$ et ainsi : $p_h \neq p_h{}^*$.

On en déduit :

Proposition III. — Si (E, e) admet une solution unique $p \in \mathcal{R}_+^{l}$, (E', e') admet aussi une solution unique $p' \in \mathcal{R}_+^{l-1}$.

En effet, (E', e') admet une solution déduite de p (proposition I) et l'on vient de voir que si (E', e') admettait deux solutions distinctes, il en serait de même de (E, e) ce qui est contraire à l'hypothèse.

2. Existence et unicité des solutions.

A ce sujet, il est possible d'utiliser dans notre cas une étude de THIONET (1964) [3], envisagée par cet auteur essentiellement en rapport avec le problème du remplissage d'un tableau projectif à double entrée. En définitive, il s'agit bien de cela ici : « remplir » le tableau à l lignes et c colonnes F par des zéros pour $(i, j) \notin R$ et des nombres f_{ij} produits d'une fonction de i seul par une fonction de j seul pour tout $(i, j) \in R$, les marges de F étant fixées pour satisfaire les équations (3) et (3').

D'après cette étude, le système (4) (5) admet une solution et une seule p telle que $p \in \mathcal{R}_+^{l}$, à condition qu'il existe un tableau à double entrée $\| y_{ij} \|$ à l lignes et c colonnes dont les marges soient a et b et tel que

$$\forall (i, j) \in \widetilde{R} - R \qquad y_{ij} = 0.$$

3. On pourrait partir aussi de certain résultats de BIRCH (1963).

Or, dans notre cas, au moins un tel tableau existe, le tableau des observations :

$$y_{ij} = x_{ij} \text{ si } (i, j) \in \mathrm{R}$$

$$y_{ij} = 0 \quad \text{si } (i, j) \in \widetilde{\mathrm{R}} - \mathrm{R}.$$

La propriété invoquée n'est toutefois étudiée par son auteur que pour le cas d'un tableau carré : $l = c$. Mais elle peut se généraliser à tous les cas; en effet, tout tableau rectangulaire peut être obtenu (si $l < c$) par regroupement de lignes présentant la même troncature d'un certain tableau carré à c lignes et c colonnes pour lequel le système des équations de vraisemblance admet une solution unique $p \in \mathcal{R}_+^c$, d'où, par applications successives de la proportion III, l'existence et l'unicité d'une solution $p \in \mathcal{R}_+^l$ du système initialement envisagé.

Discussion. — *a*) la connexité de R est cependant nécessaire; si R n'est pas connexe et admet m composantes connexes, $(\mathrm{R}_1, \ldots, \mathrm{R}_m)$, le système (4) se scinde en m systèmes indépendants dont la solution est unique à condition de fixer les m sommes des p_i pour $i \in \overline{\mathrm{R}}_k$ $(k = 1, 2, \ldots, m)$.

b) *Cas où certains totaux marginaux sont nuls.*

Partons du tableau F des f_{ij} et des équations (3) et (3') :

$$a_i = 0 \Longrightarrow x_{ij} = f_{ij} = 0 \qquad \forall\, j \in \mathrm{R}_{i.}$$

$$b_j = 0 \Longrightarrow x_{ij} = f_{ij} = 0 \qquad \forall\, i \in \mathrm{R}_{.j}$$

En supprimant du tableau initial les lignes et les colonnes dont les marges sont nulles, on obtient un tableau réduit auquel est associée la partie R' de C. En posant $p_i = 0$ (resp. : $q_j = 0$) pour i tel que $a_i = 0$ (resp. : j tel que $b_j = 0$), les équations de (3) et (3') dont le second membre est nul seront satisfaites, les équations restantes sont les équations de vraisemblance relatives au tableau réduit : elles donneront une solution (p, q) unique si mais *seulement si R' est connexe* (ce qui n'est pas nécessaire, même si R est connexe).

c) les conditions d'existence d'un tableau à marges données et à zéros fixés, se présentent sous forme d'inégalités [4]. Dans le cas limite où l'une de ces inégalités devient une égalité les équations de vraisemblance n'ont plus de solution (ou si l'on veut admettent une solution dégénérée).

Exemple. — Considérons le tableau (2,3) ci-après où la troncature porte sur la case (1,3). Nous avons la condition d'existence $a_1 + b_3 \leqslant n$ soit $b_3 \leqslant a_2$. Supposons en fait : $b_3 = a_2$, alors la case (2,3) contient nécessairement a_2, et les cases (2,1) et (2,2) contiennent chacune zéro. Il est évidemment impossible de construire le tableau des f_{ij} avec les marges données, tel que f_{ij} soit le produit d'une fonction de i seul par

4. Voir à ce sujet les travaux de M. FRÉCHET dont une liste est donnée dans l'article que nous mentionnons (1960), et plus particulièrement une note de 1959.

une fonction de j seul. En quelque sorte, on peut considérer que ceci est le cas limite où p_1 est infiniment grand devant p_2, q_1 et q_2 infiniment petits devant q_3.

b_1	b_2		a_1
0	0	a_2	a_2
b_1	b_2	b_3	n

En fin de compte, l'on voit que les équations de vraisemblance peuvent ne pas avoir de solution dans ce dernier cas (c), mais peuvent aussi avoir une solution indéterminée (en fait, une infinité de solutions) dans le cas précédent (b). Nous reviendrons plus loin sur ces points.

3. Remarque sur « l'extrêmum » de la vraisemblance.

Dans l'application de la méthode du maximum de vraisemblance, l'on cherche en général un « extrêmum » de la vraisemblance par annulation des dérivées partielles. C'est cette méthode qui est utilisée par tous les auteurs ayant étudié les tables de contingence tronquées. Nous indiquons ici comment l'on peut éviter les dérivations, et surtout montrer que l'extrêmum obtenu est bien un *maximum*.

L'on vient de voir qu'il existe en général un tableau de corrélation f_{ij} et un seul, admettant les marges (a, b) et dont les éléments sont nuls pour $(i, j) \notin \mathrm{R}$ et produits d'une fonction de i seul par une fonction de j seul pour $(i, j) \in \mathrm{R}$. Or, considérant, par exemple, le premier cas d'échantillonnage, rendre $\mathrm{L}_1(p, q)$ maximum est équivalent à rendre maximum la quantité :

$$\mathrm{V}(p, q) = \sum_{(i,j)\in \mathrm{R}} x_{ij} \operatorname{Log}(n\, p_i\, q_j)$$

compte tenu de la relation $\sum_{(i,j)\in \mathrm{R}} n\, p_i q_j = n$

Or, on a :

$$\mathrm{V}(p, q) = \sum_{(i,j)\in \mathrm{R}} (x_{ij} - f_{ij}) \operatorname{Log}(n\, p_i\, q_j) + \sum_{(i,j)\in \mathrm{R}} f_{ij} \operatorname{Log} n\, p_i\, q_j$$

En écrivant $\operatorname{Log} n\, p_i q_j = \operatorname{Log} n + \operatorname{Log} p_i + \operatorname{Log} q_j$ et en utilisant le fait que les tableaux $\| x_{ij} \|$ et $\| f_{ij} \|$ ont les mêmes marges, on montrera : $\mathrm{V}(p, q) = \sum_{(i,j)\in \mathrm{R}} f_{ij} \operatorname{Log} n\, p_i q_j$.

En vertu d'un résultat classique, compte tenu de la relation

$$\sum_{(i,j)\in \mathrm{R}} f_{ij} = \sum_{(i,j)\in \mathrm{R}} \operatorname{Log}\, n\, p_i q_j = n,$$

V sera *maximum* pour $n\,p_i q_j = f_{ij}$ pour tout $(i, j) \in R$, ce qui est possible puisque les f_{ij} satisfont H_R.

Un argument semblable peut être utilisé pour les autres cas d'échantillonnage.

IV. Résolution des équations de vraisemblance.

Utilisant les résultats précédents, nous pourrons supposer toutes les marges différentes de zéro, et aussi, si l'on veut, tous les $R_{i.}$ différents (ainsi que les $R_{.j}$).

1. Possibilités de solutions explicites.

Le système (4) peut se résoudre de façon explicite dans certains cas :

a) $l = 2$.

Tenant compte de (5) : $p_1 + p_2 = 1$, (4) se réduit à une seule équation linéaire à une inconnue, à savoir :

$$\sum_{j \in R_{1.} - R_{2.}} b_j + \sum_{j \in R_{1.} \cap R_{2.}} b_j\, p_1 = a_1$$

C'est un cas de ce type qui est considéré par KASTENBAUM (1958) et aussi par WATSON (1956) si l'on tient compte de la possibilité de regroupement des lignes semblables.

Remarque. — Si $\sum_{j \in R_{1.} \cap R_{2.}} b_j = 0$, p_1 et p_2 sont indéterminés : nous nous trouvons dans un des cas discutés au III-2-b pour lesquels R′ n'est pas connexe.

b) *Table triangulaire.*

Supposons $R_{i.} = \{1, 2, \ldots, j_i\}$ avec j_i fonction non décroissante de i. On a donc : $R_{.j} = \{i_j, i_j + 1, \ldots, l\}$ où i_j est fonction non décroissante de j.

Alors, le système (4) est récurrent et l'on pourra expliciter sa solution vérifiant aussi (5).

Ce cas a déjà été étudié par BATSCHELET (1960 a) et GEPPERT (1961). Nous indiquons cependant une autre méthode de calcul qui est, peut-être, plus rapide.

On peut démontrer simplement par récurrence sur i :

$$c_i = \frac{p_i}{1 - p_1 - p_2 \cdots - p_{i-1}} = \frac{a_i}{\sum_{j \in R_{i.}} b_j - \sum_{h=1}^{i-1} a_h}$$

On explicite de là les p_i à partir de :

$$p_i = c_i \prod_{k=1}^{i-1} (1 - c_k)$$

d'où :

$$p_i = a_i \frac{\prod_{k=1}^{i-1} \left(\sum_{j \in R_k.} b_j - \sum_{h=1}^{k} a_h \right)}{\prod_{k=1}^{i} \left(\sum_{j \in R_k.} b_j - \sum_{h=1}^{k-1} a_h \right)}$$

On remarquera que, si $R_{k.} = R_{(k+1).}$ il y a des simplifications immédiates; ce sont ces formes simplifiées que donne GEPPERT. Cependant, nous avons vu plus haut que de telles simplifications peuvent s'introduire de façon plus générale.

Des permutations de lignes ou colonnes permettront de ramener à celui-ci d'autres cas semblables, parmi lesquels celui qu'Asano (1965) traite sous le nom de « Nested Case ».

Remarques : 1° Le cas où certains a_i ou b_j sont nuls ne pose ici aucune difficulté puisque en retranchant d'une table triangulaire une ligne ou une colonne, l'on obtient une nouvelle table triangulaire.

2° Supposons maintenant les a_i et b_j tous différents de zéro, le cas limite (c) est envisagé par GEPPERT (1961) pages 62-63. D'après cet auteur, l'on peut alors conserver l'expression ci-dessus des p_i qui est toujours définie; ceci permet d'associer à tout échantillon une estimation de p bien déterminée (mais il n'en va pas de même pour q). Cette façon de faire est d'autant plus recommandable que les estimateurs ainsi obtenus sont sans biais [5] comme le montre GEPPERT dans le cas où les b_j sont fixés, et de là, dans le cas où les b_j sont aléatoires puisque (en reprenant la notation $\hat{p}_i$ pour l'estimateur) :

$$E(\hat{p}_i) = E[E(\hat{p}_i/b)] = E(p_i) = p_i.$$

c) *Cas où* R *est connexe* (l, c) à $l + c - 1$ *éléments.*

Plutôt que de reprendre les systèmes d'équations, il suffira de remarquer que la méthode d'estimation utilisée consiste à rendre minima une certaine « distance » Δ (cf. NEYMAN, 1949) entre les points de $\mathcal{R}_+^t$ ($t = l + c - 1$)

5. Et donc « meilleurs estimateurs sans biais » puisque fonctions des seules statistiques exhaustives.

de coordonnées x_{ij} et f_{ij} $[(i, j) \in R]$, avec la condition que f_{ij} soit le produit d'une fonction de i par une fonction de j; or, ici, cette condition est réalisée quels que soient les f_{ij} (cf. Chapitre I. Théorème VI), à condition toutefois que f_{ij} soit différent de zéro pour tout $(i, j) \in R$; si $x_{ij} \neq 0$ pour tout (i, j) de R, il est donc possible de rendre Δ nul (et ce sera bien là son minimum) en posant :

$$f_{ij} = n\, p_i q_j = x_{ij} \qquad \forall (i, j) \in R.$$

De là, on aura facilement, si besoin est, les estimateurs p et q.

Notons que le cas traité ici est beaucoup moins restreint qu'il ne peut paraître, étant données les possibilités de regroupement de lignes ou colonnes vues plus haut. Se ramènent à ce cas, tous les tableaux à lignes ou colonnes « enchaînées » (Chained case d'ASANO, 1965) ce qui rend les calculs très rapides en évitant le maniement de formules assez lourdes.

(On peut utiliser cette méthode, en particulier, pour tout tableau à deux lignes ou deux colonnes.)

Discussion. — 1° Le cas où il existe $(i, j) \in R$ tel que $x_{ij} = 0$ rentre dans le cas limite (c) étudié plus haut; on peut encore, si l'on veut, définir des estimateurs de p, certains p_i pouvant être nuls (mais alors certains q_j sont infinis).

2°) Si pour i donné a_i est nul, l'on aura $p_i = 0$, puis en supprimant la i^{me} ligne, on obtiendra un tableau réduit auquel sera associée une partie R′ non connexe : p ne sera pas défini. Contrairement à ce qui se passe pour une table triangulaire, il paraît donc difficile ici d'associer à tout échantillon une estimation bien définie de p; on peut se demander dans ces conditions s'il est possible d'envisager certaines propriétés statistiques des estimateurs, telle l'absence de biais avancée par ASANO (1965).

d) On peut trouver des solutions explicites dans d'autres cas, quoique de façon moins simple et peut-être moins utile. Par exemple, dans certaines tables tronquées où $l = 3$ (c quelconque) on obtiendra une équation du 2e degré en p_1 dont on pourra déterminer simplement la racine à considérer.

2. Méthodes itératives.

a) *Utilisation de l'algorithme R.A.S.* (6)

Cet algorithme permet de construire un tableau de corrélation $\| t_{ij} \|$ de marges données et « ressemblant » à un autre tableau $\| t_{ij}^{(0)} \|$ en ce sens que t_{ij} est de la forme $\mu_i \nu_j t_{ij}^{(0)}$ (cf. THIONET, 1963 et 1964). Si les totaux marginaux désirés sont

$$a = [a_1, a_2, \ldots, a_l]' \text{ et } b = [b_1, b_2, \ldots, b_c]'$$

6. Nous reprenons la terminologie utilisée par THIONET dans les articles cités. On trouvera en Appendice, à la fin de ce mémoire, quelques indications bibliographiques sur cette méthode.

on construit la suite de matrices $\| t_{ij}^{(n)} \|$ où :

$$\begin{cases} t_{ij}^{(2m)} = t_{ij}^{(2m-1)} \dfrac{b_j}{\sum_{i=1}^{l} t_{ij}^{(2m-1)}} \\ t_{ij}^{(2m+1)} = t_{ij}^{(2m)} \dfrac{a_i}{\sum_{j=1}^{c} t_{ij}^{(2m)}} \end{cases}$$

Ici, il faudra prendre

$\| t_{ij}^{(0)} \| = \mathrm{T}$ (matrice d'incidence définie plus haut).

Si la suite $t_{ij}^{(n)}$ converge, sa limite est nécessairement la matrice F et nous donnera donc les f_{ij} cherchés (et aussi p et q par un calcul immédiat).

Une condition nécessaire de convergence est *l'existence* d'un tableau ayant les mêmes zéros que T et les marges a et b imposées, et, dans notre problème, cette condition est bien réalisée (tableau des observations). Cette condition est-elle suffisante? D'après THIONET (1964), ceci est vraisemblable quoique non prouvé.

Pour notre part, nous avons pu montrer qu'il s'agissait bien d'une condition suffisante pour quelques cas particuliers relativement à la position des zéros de T. Nous exposons la démonstration en Appendice.

Remarque. — Si $\mathrm{R}_{i.}$ ne contient qu'un seul élément j, on a $f_{ij} = x_{ij}$; il suffit de construire un tableau comportant une ligne de moins.

b) *Autres méthodes.*

— WATSON (1956) suggère une méthode d'estimation consistant à introduire dans le tableau initial des quantités y_{ij} pour $(i, j) \notin \mathrm{R}$, telles que :

$$y_{ij} = \frac{\left(\sum_{j \notin \mathrm{R}_{i.}} y_{ij} + a_i \right) \left(\sum_{i \notin \mathrm{R}_{.j}} y_{ij} + b_j \right)}{n + \sum_{(i,j) \notin \mathrm{R}} y_{ij}} \qquad (8)$$

(en notant ici $(i, j) \notin \mathrm{R}$ la relation $(i, j) \in \bar{\mathrm{R}} - \mathrm{R}$, etc...).

Si l'on pose alors

$$\mathrm{N} = n + \sum_{(i,j) \notin \mathrm{R}} y_{ij}$$

$$\forall i \in \bar{\mathrm{R}},\ A_i = a_i + \sum_{j \notin \mathrm{R}_{i.}} y_{ij} \text{ et } \forall j \in \mathrm{R}^*,\ B_j = b_j + \sum_{i \notin \mathrm{R}_{.j}} y_{ij}$$

il vient :

$$\forall i \in \overline{R},\ p_i = A_i/N; \qquad \forall j \in R^*,\ q_j = B_j/N;\ \forall (i,j) \in R,\ f_{ij} = \frac{A_i B_j}{N}$$

(avec par construction, $\frac{A_i B_j}{N} = y_{ij}$ pour $(i, j) \notin R$). Le problème est ainsi ramené à la résolution de (8), système d'équations à t' inconnues avec $t' = \text{Card}\,(\widetilde{R} - R)$.

On peut utiliser le processus itératif suivant :

$$y_{ij}^{(k+1)} = \frac{\left(\sum_{j \notin R_{i.}} y_{ij}^{(k)} + a_i\right)\left(\sum_{i \notin R_{.j}} y_{ij}^{(k)} + b_j\right)}{n + \sum_{(i,j) \notin R} y_{ij}^{(k)}} \tag{9}$$

mais la convergence reste problématique, les valeurs de départ $y_{ij}^{(o)}$, a priori arbitraires, peuvent d'ailleurs jouer un rôle.

Remarques. — 1) Cette technique cherchant à estimer les fréquences manquantes est analogue à celle utilisée dans les tout premiers essais (cf. WAITE, 1915) ; à celle-ci se rattache aussi la technique de BATSCHELET et GEPPERT pour les tables triangulaires les plus générales.

2) Pour une table (l, c) $(l < c)$ dans laquelle on a regroupé lignes et colonnes présentant la même troncature, on a toujours $t' \geqslant l - 1$; or, si l'on tient compte de (5), le système (4) comprend $l - 1$ équations indépendantes seulement, c'est-à-dire jamais davantage que le système (8).

— On peut utiliser un processus du même type que (9) directement avec le système (4) ; là aussi, le problème des valeurs de départ se pose (quoiqu'il soit possible de partir d'estimateurs convergents mais non efficients des p_i) et la convergence n'est pas assurée.

— Notons enfin que l'on peut toujours utiliser la traditionnelle méthode de NEWTON-RAPHSON (comme indiqué par ASANO, 1965), mais elle est assez lourde; cependant l'on sait que la suite des valeurs itérées converge en probabilité si n tend vers l'infini lorsque les valeurs de départ sont des estimateurs convergents de p (KALE, 1962). En fait, par une transformation de cette méthode, l'on peut obtenir de nouveaux estimateurs efficients par un calcul fini comme il sera vu plus loin.

c) *Étude d'un cas particulier important.*

Nous envisageons ici le cas où R est telle que :

$$\begin{cases} l = c \qquad (\overline{R} = 1, 2, \ldots, l; \qquad R^* = 1, 2, \ldots, l). \\ (i, j) \in R \iff i \neq j. \end{cases}$$

Nous indiquons une méthode itérative simple et rapide pour obtenir la solution des équations de vraisemblance dans ce cas.

On a ici, en tenant compte de (5) : $\sum_{h \in R_j} p_h = 1 - p_j \quad (\forall j \in R^*)$

et, en posant

$$x = \sum_{i=1}^{l} \frac{b_i}{1-p_i} \tag{10}$$

(4) devient : $$x p_i^2 - (x + a_i - b_i) p_i + a_i = 0 \tag{11}$$

soit :

$$p_i = \frac{x + a_i - b_i + \varepsilon_i \sqrt{\Delta_i}}{2x} \tag{12}$$

avec $\varepsilon_i = \pm 1$ et $\Delta_i = (x + a_i - b_i)^2 - 4 a_i x = (x - a_i - b_i)^2 - 4 a_i b_i$.

L'équation (5) implique :

$$(l - 2) x + \sum_{i=1}^{l} \varepsilon_i \sqrt{\Delta_i} = 0 \tag{13}$$

Quels que soient les ε_i, si (13) admet une solution x, $x \geqslant n$, (12) donnera des p_i correspondants dont on peut montrer qu'ils sont tous positifs et fournissent donc une solution de notre problème. L'unicité de la solution étant assurée, il suffira d'extraire de (13) par choix des ε_i une équation admettant une solution x réelle supérieure à n.

Montrons qu'il suffit de considérer une équation du type :

$$f(x) = (l-2) x - \sum_{i=1}^{l} \sqrt{\Delta_i} = 0 \tag{13'}$$

ou $$g_j(x) = (l - 2) x - \sum_{i \neq j} \sqrt{\Delta_i} + \sqrt{\Delta_j} = 0 \tag{13''}$$

On notera $x_m = \sup (\sqrt{a_i} + \sqrt{b_i})^2$; pour $x \geqslant x_m$, $f(x)$ est définie et aussi $g_j(x)$ quel que soit j :

1° *Si* $f(x_m) \geqslant 0$, $f(x) = 0$ admet une racine et une seule $x \geqslant x_m$; en effet, l'on montre facilement que $f(x)$ est décroissante pour x supérieur à x_m et $f(x) \to -\infty$ si $x \to +\infty$; en outre $f(n)$ est positif si n est supérieur à x_m, donc la racine x est toujours supérieure à n.

Pour la recherche de cette racine, la méthode de NEWTON est certainement convergente en partant d'une valeur x_0 voisine de x_m; on peut améliorer cette valeur de départ, par exemple en prenant $x_0 = n$ si $n > x_m$; à ce sujet, on pourra remarquer aussi que x définie par (10) peut s'écrire :

$x = n(1 - \sum_{i=1}^{l} p_i q_i)^{-1}$.

2° Si $f(x_m) < 0$, soit j l'indice de la plus grande des quantités $(\sqrt{a_i} + \sqrt{b_i})$, (13′) n'a pas de solution, mais $g_j(x) = 0$ en admet une puisque $g_j(x_m) = f(x_m) < 0$ et $g_j(x) \to 2(n - a_j - b_j) > 0$ [7] si $x \to +\infty$. Elle est évidemment supérieure à n, puisque dans ces cas n est inférieur à x_m.

De même que précédemment, cette racine pourra être cherchée par la méthode de NEWTON.

La valeur de x annulant $f(x)$ ou $g_j(x)$ sera ensuite portée dans (12) pour obtenir p, avec :

$\varepsilon_i = -1$ pour tout i si x annule $f(x)$

$\varepsilon_i = -1$ pour $i \neq j$ et $\varepsilon_j = 1$ si x annule $g_j(x)$.

Remarques. — 1° Dans la pratique, la méthode indiquée ci-dessus est plus rapide que les méthodes générales adaptées à ce cas particulier; le fait de pouvoir utiliser un processus itératif à une seule variable permet d'effectuer facilement les calculs sans l'aide d'ordinateur d'autant plus que la convergence est en général très rapide (par exemple, dans plusieurs cas traités, avec l variant de 4 à 8, et dans un cas où $l = 14$, n variant de 100 à 1 000, 4 ou 5 itérations suffisent pour obtenir sur x une précision de 8 chiffres significatifs).

D'autre part, on peut augmenter la rapidité des calculs par une variante de cette méthode, comme il est indiqué plus loin (V - 3).

2° Il est possible d'adapter très simplement la méthode ci-dessus à de nouveaux modèles de troncature; à titre d'exemple, si R est $(l - 1, l)$ et telle que $(i, j) \in R \iff i \neq j$, il suffira d'ajouter une ligne de zéros au tableau des données. Si la table tronquée est obtenue par suppression de certaines cases d'une diagonale (peut-être pas toutes comme ci-dessus) l'on pourra encore adopter une méthode itérative analogue.

(Rappelons qu'il est toujours possible d'utiliser les résultats du III - 1 pour étendre encore le domaine d'application.)

V. Étude des grands échantillons — Nouveaux estimateurs R.B.A.N.

1. Introduction.

D'après les considérations du Chapitre I (B-2) on peut supposer P_{ij} différent de zéro pour tout $(i, j) \in R$. Supposons en outre que n tend vers l'infini (ou, si les totaux marginaux b_j sont certains, tous ceux-ci tendent vers l'infini).

Dans ces conditions,

$$\forall (i, j) \in R, \quad Pr[x_{ij} = 0] \to 0$$

il en résulte en particulier

$$\forall i \in \overline{R}, \quad Pr[a_i = 0] \to 0$$

$$\forall j \in R^*, \quad Pr[b_j = 0] \to 0$$

7. En écartant le cas limite $a_j + b_j = n$ pour lequel, comme on l'a déjà vu (III-2), le système d'équations envisagé n'est pas résoluble.

Ainsi, les divers cas limite envisagés précédemment qui pouvaient conduire à des difficultés dans la résolution des équations de vraisemblance, se présentent, pour n grand, avec une probabilité qui tend vers zéro; donc, avec une probabilité tendant vers un, si n tend vers l'infini, les estimateurs $(\hat{p}, \hat{q})$ seront bien déterminés.

D'autre part, les théorèmes généraux sur la méthode d'estimation employée assurent la *convergence* et *l'efficacité asymptotique* de $(\hat{p}, \hat{q})$.

En définitive, quel que soit R, l'estimateur $(\hat{p}, \hat{q})$ est R.B.A.N., mais dans certains cas son calcul peut être long, nécessitant l'utilisation de méthodes itératives. Nous donnons ci-après des techniques permettant de déterminer d'autres estimateurs R.B.A.N. par un *calcul fini.*

L'étude suivante est avant tout valable pour n grand; dans ces conditions, nous négligerons les cas limite qui, nous l'avons vu plus haut, ne peuvent se produire qu'avec une probabilité voisine de zéro.

2. Recherche d'estimateurs convergents.

Nous introduisons ici deux familles d'estimateurs *convergents* mais non R.B.A.N.

a) Considérant toujours R connexe (l, c), il est possible d'extraire de R une partie S connexe (l, c) à $l + c - 1$ éléments (Ch. I-A. Théorème II). L'échantillon réduit aux observations $\alpha_i \beta_j$ où $(i, j) \in S$ pourra nous fournir facilement des estimateurs de (p, q) (cf. IV-1-c), estimateurs convergents d'après ce qui précède (puisque si n tend vers l'infini, la taille de l'échantillon réduit tend presque sûrement vers l'infini).

Evidemment, dans la mesure où la partie S définie ci-dessus n'est pas unique, l'on pourra déterminer de cette façon plusieurs estimateurs différents. Cette remarque (jointe au fait que, par la méthode précédente, une partie de l'information utilisable est ignorée) peut justifier la recherche d'une deuxième famille d'estimateurs qui seraient fonctions de toutes les fréquences observées et dans la détermination desquels ne rentrerait aucun choix peut-être subjectif.

b) Nous proposons ici une méthode de « moindres-carrés » qui conduit à des estimateurs pour p moyennant la résolution d'un système d'équations *linéaires.*

On notera $y_{ij} = \frac{x_{ij}}{n}$ les fréquences relatives.

Soit : $$\Phi^2(y, p) = \sum_{(i,j) \in R} y_{ij} \left(\frac{p_i}{y_{ii}} - \frac{\sum_{k \in R_{\cdot j}} p_k}{\sum_{k \in R_{\cdot j}} y_{kj}} \right)^2,$$

On choisira comme estimateurs des p_i les quantités $\tilde{p}_i$ qui minimisent Φ^2 avec la condition

$$\sum_{i=1}^{l} \tilde{p}_i = 1 \tag{14}$$

Introduisant le multiplicateur de Lagrange λ, on est conduit à un système linéaire :

$$\sum_{j \in R_{m.}} \left(\frac{\tilde{p}_m}{y_{mj}} - \frac{\sum_{k \in R_{.j}} \tilde{p}_k}{\sum_{k \in R_{.j}} y_{kj}} \right) = \lambda \tag{15}$$

pour $m = 1, 2, \ldots, l$

qui peut aussi s'écrire :

$$\tilde{p}_m \sum_{j \in R_{m.}} y_{mj}^{-1} - \sum_{k=1}^{l} \left[\sum_{j \in R_{m.} \cap R_{k.}} n\, b_j^{-1} \right] \tilde{p}_k = \lambda \tag{16}$$

$(m = 1, 2, \ldots, l)$

λ est à déterminer par la condition (14). Dans la pratique on prendra λ ($\neq 0$) quelconque; (16) admet alors comme solution des $\tilde{p}_i'$ tels que :

$$\tilde{p}_i = \tilde{p}'_i \left(\sum_{h=1}^{l} p'_h \right)^{-1}$$

Matriciellement, (16) peut s'écrire sous la forme :

$$\mathrm{M}\,\hat{p} = \lambda [11 \ldots 1]'$$

et en notant Δ_i la somme des éléments de la ième ligne de M^{-1} on a :

$$\tilde{p}_i = \Delta_i \left(\sum_{i=1}^{l} \Delta_i \right)^{-1}$$

On peut ainsi ramener le calcul des $\hat{p}_i$ au calcul des seuls mineurs d'ordre $(l-1)$ de M.

Propriétés des estimateurs $\tilde{p}_i$.

1. Les $\tilde{p}_i$ sont définis (et définis de façon unique) avec une probabilité qui tend vers un lorsque n tend vers l'infini.

On notera p_i° et q_j° les vraies valeurs des paramètres p_i et q_j. Rappelons que l'on a supposé $p_i^\circ \neq 0 \ \forall i \in R$; $q_j^\circ \neq 0 \ \forall j \in R^*$. Si $t = \mathrm{Card}\ (R)$, notons

y le point de $\mathcal{R}_+^l$ de coordonnées y_{ij}, et $p^\circ\, q^\circ$ le point de coordonnées $p_i^\circ\, q_j^\circ$. La fonction de p :

$$\Phi^2(p_\circ\, q_\circ, p) = \sum_{(i,j)\in R} \frac{p_i''}{q_j^\circ} \left(\frac{p_i}{p_i^\circ} - \frac{\sum_{k\in R.j} p_k}{\sum_{k\in R.j} p_k^\circ} \right)^2$$

admet évidemment un minimum de zéro pour $p = p^\circ$; puisque R est connexe, ce minimum est unique par application du théorème III (Ch. I).

Le système (16) est donc régulier au point $y = p^\circ\, q^\circ$; d'autre part, les éléments de la matrice M sont continus dans un voisinage de $p^\circ\, q^\circ$ (puisque, $\forall (i,j) \in R$, $p^\circ_i\, q^\circ_j \neq 0$) et donc il existe un voisinage V du point $p^\circ\, q^\circ$ tel que (16) est régulier dès que y appartient à V. Or, lorsque n tend vers l'infini, y converge presque sûrement vers $p^\circ\, q^\circ$, donc, quel que soit V, $Pr[y \in V]$ tend vers un. La propriété annoncée est ainsi démontrée.

2. La distribution asymptotique conjointe des l *v.a.* $\sqrt{n}(\tilde{p}_i - p_i^\circ)$ est normale.

$\tilde{p}_i$ est une fonction de y, soit $\tilde{p}_i = f_i(y)$.
On a vu que $\Phi^2(p^\circ\, q^\circ, p)$ avait un minimum unique pour $p = p^\circ$; donc $f_i(p^\circ\, q^\circ) = p_i^\circ$

On a alors, pour $k = 1, 2, \ldots, l$

$$\tilde{p}_k - p_k^\circ = f_k(y) - f_k(p^\circ\, q^\circ)$$

$$= \sum_{(i,j)\in R} (y_{ij} - p''_i\, q^\circ_j) \left(\frac{\partial f_k}{\partial y_{ij}} \right)_{y = p^\circ\, q^\circ} + \varepsilon_k \qquad (17)$$

où $\sqrt{n}\rho_k$ tend vers zéro en probabilité si n tend vers l'infini.

La distribution asymptotique conjointe des t v.a. $\sqrt{n}(y_{ij} - p^\circ_i\, q^\circ_j)$ est normale; il en sera de même de la distribution des formes linéaires

$$\sum_{(i,j)\in R} \sqrt{n}(y_{ij} - p''_i\, q^\circ_j) \left(\frac{\partial f_k}{\partial y_{ij}} \right)_{y = p^\circ\, q^\circ}$$

(la démonstration est immédiate en utilisant les propriétés des fonctions caractéristiques); si l'on ajoute à cette dernière v.a. la v.a. $\sqrt{n}\rho_k$ qui tend vers zéro en probabilité, la distribution limite sera encore la même (cf. CRAMER 1946, p. 254), ce qui démontre la proposition annoncée.

3. Les estimateurs p_i ne sont pas en général asymptotiquement efficaces.

Il suffit de montrer ceci sur un cas particulier : prenons par exemple $R = \tilde{R}$, on obtient :

$$\tilde{p}_k = \left[\sum_l \frac{1}{y_{kj}} \cdot \sum_h \frac{1}{\sum_j \frac{1}{y_{hj}}} \right]^{-1}$$

et

$$\left(\frac{\partial f_k}{\partial y_{ij}}\right)_{p^\bullet q^\bullet} = \left[(q^\bullet_j)^2 \sum_h \frac{1}{q^\circ_h} \right]^{-1} (\delta_{ik} - p^\circ_k)$$

où δ_{ik} est le symbole de KRONECKER.

L'utilisation de (17) avec l'expression ci-dessus de $\left(\frac{\partial f_k}{\partial y_{ij}}\right) p^\circ q^\circ$ et le résultat classique sur la distribution multinomiale :

$$E\left[(y_{ij} - p^\circ_i q^\circ_j)(y_{i'j'} - p^\circ_{i'} q^\bullet_{j'})\right] = \frac{1}{n}(\delta_{ii'}\delta_{jj'} p^\circ_i q^\bullet_j - p^\circ_i q^\circ_j p^\circ_{i'} q^\circ_{j'})$$

permet d'obtenir, pour la variance asymptotique v_k de la v.a. $\sqrt{n}(\tilde{p}_k - p^\circ_k)$, l'expression :

$$v_k = p^\circ_k (1 - p^\circ_k) \frac{\sum_j \left(\frac{1}{q^\circ_j}\right)^3}{\left(\sum_h \frac{1}{q^\circ_h}\right)^2}$$

Or, dans le cas simple où l'on se trouve, l'on sait que $\hat{p}_k = \sum_j y_{kj}$ est un estimateur efficace, et la variance de $\sqrt{n}(\hat{p}_k - p^\circ_k)$ est $p^\circ_k (1 - p^\circ_k)$.

Le carré de l'efficacité relative de $\tilde{p}_k$ est donc :

$$E^2 = \frac{\left(\sum_h \frac{1}{q^\circ_h}\right)^2}{\sum_j \left(\frac{1}{q^\circ_j}\right)^2}$$

par application de l'inégalité de CAUCHY-SCHWARTZ et en remarquant que dans le cas étudié ici $\sum p^\circ_i = 1$ entraîne $\sum q^\circ_j = 1$, on montrera $E^2 \leqslant 1$.

L'égalité a lieu seulement si tous les q°_j sont égaux; ce cas est non seulement très particulier, mais encore indiscernable dans la pratique, puisque les q°_j sont inconnus.

3. Une nouvelle classe d'estimateurs R.B.A.N.

Posons : $$\psi_i(p) = p_i \sum_{j \in R_{i\cdot}} \frac{b_j}{\sum_{h \in R_{\cdot j}} p_h} - a_i$$

On aura

$$\frac{\partial \psi_i}{\partial p_k} = \delta_{ik} \sum_{j \in R_{i\cdot}} \frac{b_j}{\sum_{h \in R_{\cdot j}} p_h} - p_i \sum_{j \in R_{i\cdot} \cap R_{k\cdot}} \frac{b_j}{\left(\sum_{h \in R_{\cdot j}} p_h\right)^2}$$

Revenons aux équations de vraisemblance (4), elles s'écrivent :

$$\psi_i(\hat{p}) = 0 \text{ pour tout } i \in \bar{R}$$

On définit δp_i ($i = 1, 2, \ldots, l$) comme solutions du système linéaire :

$$\sum_{k=1}^{l} \delta p_k \left(\frac{\partial \psi_i}{\partial p_k}\right)_{\tilde{p}} = \psi_i(\tilde{p}) \qquad \text{pour } i = 1, 2, \ldots, l.$$

$$\sum_{k=1}^{l} \delta p_k = 0$$

On a là un système à l équations et l inconnues, en général régulier, en choisissant $l-1$ des l premières équations dont on montrera facilement qu'elles ne sont pas indépendantes. Ce système est celui associé à la résolution de (4) par la méthode itérative de Newton-Raphson mais l'on peut montrer qu'une seule itération suffit pour obtenir des estimateurs R.B.A.N. si $\tilde{p}$ est un estimateur de p $0(\sqrt{n})$-convergent (Ferguson, 1958), en d'autres termes :

$$\ddot{p}_i = \tilde{p}_i + \delta p_i \text{ pour } i = 1, 2, \ldots, l$$

sont des estimateurs R.B.A.N. des paramètres p_i.

Les estimateurs convergents obtenus au paragraphe précédent sont bien $0(\sqrt{n})$-convergents, si bien que l'on saura toujours trouver un estimateur R.B.A.N. de p par la résolution de un ou de deux systèmes *linéaires* à l inconnues. Vu la simplicité de l'une des équations (qui exprime $\Sigma p_i = 1$) on peut considérer que l'on a, en fait, un système linéaire de $(l-1)$ équations à autant d'inconnues.

A partir de $\ddot{p}$ il est immédiat de déterminer un estimateur R.B.A.N. de q. En effet (3') donne $\hat{q}$ sous forme d'une fonction de $\hat{p}$ soit $\hat{q}_j = \varphi_j(\hat{p})$ pour tout $j \in R^*$.

Il est à peu près évident que $\overset{\cdot\cdot}{q}_j = \varphi_j(\overset{\cdot\cdot}{p})$ est un estimateur R.B.A.N. de q. En effet, on peut écrire :

$$\sqrt{n}\,(\hat{q}_j - \overset{\cdot\cdot}{q}_j) = \sqrt{n}\,[\varphi_j(\hat{p}) - \varphi_j(\overset{\cdot\cdot}{p})] = \sum_i \sqrt{n}\,(\hat{p}_i - \overset{\cdot\cdot}{p}_i)\left(\frac{\partial \varphi_j}{\partial p_i}\right)_{\hat{p}} + \rho\sqrt{n}$$

où $\rho\sqrt{n} \xrightarrow{\text{Prob}} 0$ si $n \to +\infty$.

Or $\overset{\cdot\cdot}{p}$ est construit de telle sorte que $\sqrt{n}\,(\hat{p}_i - \overset{\cdot\cdot}{p}_i)$ tende vers zéro en probabilité lorsque n tend vers l'infini, et d'autre part, $\left(\frac{\partial \varphi_j}{\partial p_i}\right)_{\hat{p}}$ tend en probabilité vers une limite certaine finie.

Donc $\sqrt{n}\,(\hat{q}_j - \overset{\cdot\cdot}{q}_j) \xrightarrow{\text{Prob}} 0$ pour tout $j \in \mathrm{R}^*$. Ce qui entraîne bien que $\overset{\cdot\cdot}{q}$ est R.B.A.N. puisque $\hat{q}$ l'est.

Remarque. — Au IV-2-c nous étudions un processus itératif pour chercher $\hat{p}$ dans un cas particulier. On a, à partir de (10) et de (3′) :

$$\frac{x}{n} = \sum_{i=1}^{l} \frac{b_i / n}{1 - \hat{p}_i} = \sum_{i=1}^{l} \hat{q}_i$$

et x/n est donc un estimateur R.B.A.N. de $\sum_{i=1}^{l} q_i$. Cet estimateur se cherche en résolvant (13) comme indiqué; mais si l'on utilise alors la méthode de Newton en partant d'une valeur x_0 telle que x_0/n soit un estimateur $0(\sqrt{n})$-convergent de $\sum_{i=1}^{l} q_i$ (et l'on sait trouver un tel estimateur) une seule itération nous donnera une valeur x_1, telle que x_1/n est estimateur R.B.A.N. de $\sum_{i=1}^{l} q_i$ (car le processus itératif utilisé est du deuxième ordre).

En reportant cette valeur x_1 dans (12), l'on obtiendra $\overset{\wedge\wedge}{p}$, estimateur R.B.A.N. de p; la démonstration est identique à la démonstration ci-dessus concernant $\overset{\cdot\cdot}{q}$.

4. Utilisation d'une méthode de Neyman.

Les techniques exposées ci-dessus permettent de trouver des estimateurs R.B.A.N. par résolution d'un système linéaire à $(l-1)$ inconnues, quelle que soit la partie R.

Pour ramener cette rercherche à un système linéaire, NEYMAN (1949) donne une méthode très générale basée sur la « linéarisation » des liaisons

existant entre les probabilités des diverses catégories, et la minimisation d'une quantité χ_1^2 sous ces nouvelles conditions.

L'adaptation de cette méthode à l'étude des tables de contingence tronquées est immédiate. Elle donne des fréquences théoriques f^*_{ij} (estimateurs R.B.A.N. des f_{ij}) par la résolution d'un système linéaire à ν inconnues ($\nu = t - l - c + 1$), de la façon suivante :

sous l'hypothèse H_R (R connexe) les P_{ij} sont soumis à ν liaisons (Ch. I-théorème VIII) que l'on peut écrire, en suivant les notations de Neyman :

$$F_k(P) = 0 \qquad k = 1, 2, \ldots, \nu.$$

On pose, pour $k, h = 1, 2, \ldots, \nu$; $(i, j) \in R$

$$\left(\frac{\partial F_k}{\partial p_{ij}}\right)_{P=y} = b_{kij}$$

$$\bar{b}_k = \sum_{(ij)\in R} y_{ij}\, b_{kij}$$

$$v_{hk} = \frac{1}{n} \sum_{(i,j)\in R} y_{ij}\, b_{hij}\, b_{kij}$$

Dans ces conditions, l'on aura :

$$f^*_{ij} = x_{ij}\left[1 + \frac{1}{n} \sum_{k=1}^{\nu} (b_{kij} - \bar{b}_k)\,\mu_k\right] \text{ pour tout } (i, j) \in R.$$

où les μ_k sont obtenus en résolvant le système linéaire :

$$\sum_{k=1}^{\nu}\left(v_{hk} - \frac{1}{n}\, b_h\, b_k\right)\mu_k = F_h(y) \qquad (h = 1, 2 \ldots, \nu).$$

Mais ceci peut être simplifié. En effet, si l'on exprime les liaisons sous forme d'interactions du premier ordre (ce qui est toujours possible; cf : Ch. I-B-III) on aura, pour $k = 1, 2, \ldots, \nu$:

$$F_k(P) = \sum_{(i,j)\in R} u_{ij}^{(k)} \operatorname{Log} P_{ij} = \sum_{(i,j)\in R} u_{ij}^{(k)} \operatorname{Log} p_{ij} = \delta\,(u^{(k)}, p)$$

avec :

$$\sum_{i\in R_j} u_{ij}^{(k)} = 0 \qquad \text{pour } j \in R^*, \quad \sum_{j\in R_i} u_{ij}^{(k)} = 0 \text{ pour } i \in \bar{R}$$

$$b_{kij} = n\, u_{ij}^{(k)} / x_{ij}$$

$$F_k(y) = \sum_{(i,j)\in R} u_{ij}^{(k)} \operatorname{Log} x_{ij} = \delta\,(u^{(k)}, x) = d_k$$

$\bar{b}_k = 0$

$v_{hk} = \sum_{(i,j)\in R} \frac{u_{ij}^{(k)} u^{(h)}{}_{ij}}{x_{ij}}$ pour $h, k = 1, 2, \ldots, \nu$

et $\forall (i, j) \in R \quad f^*_{ij} = x_{ij} + \sum_{k=1}^{\nu} u_{ij}^{(k)} \mu_k$

avec $\mu = V^{-1} d$

en posant

$$\mu = [\mu_1, \mu_2, \ldots, \mu_\nu]', \quad d = [d_1, d_2, \ldots, d_\nu]' \quad V = ||v_{hk}||$$

Remarques. — 1. La matrice V est un estimateur de la matrice des variances et covariances du vecteur aléatoire d. La nullité de $\bar{b}_k$ est l'analogue de ce fait utilisé par PLACKETT (1962) puis GOODMAN (1963 a, 1964 a) que la matrice des variances et covariances de constrastes de v.a. du type Log x_{ij} s'obtient de la même façon que si ces variables étaient indépendantes.

2. Les estimateurs $f^*{}_{ij}$ sont invariants si l'on remplace les liaisons indépendantes $\delta(u^{(k)}, p) = 0$ par ν autres liaisons indépendantes.

En effet, soit $\dot{u}^{(1)}, \dot{u}^{(2)}, \ldots, \dot{u}^{(\nu)}$, ν vecteurs indépendants de I(R). On notera U la matrice (ν, t) dont les lignes sont les vecteurs $u^{(h)}$ et les quantités déduites des précédentes en remplaçant $u^{(h)}$ *par* $\dot{u}^{(h)}$ seront distinguées par un point. Il existe une matrice M carrée régulière d'ordre ν telle que $\dot{U} = M U$.

Si l'on désigne par $l(x)$ le vecteur de $\mathcal{R}^t$ de composantes Log x_{ij} et par Z la matrice diagonale $(t \times t)$ dont les termes diagonaux sont $1/x_{ij}$, $(i, j) \in R$, on aura :

$$d = U\, l(x)$$
$$V = U Z U'$$

d'où $\dot{d} = \dot{U}\, l(x) = M U\, l(x) = M d$

$V = \dot{U} Z \dot{U}' = M U Z U' M' = M V M'$ donc $\dot{V}^{-1} = M'^{-1} V^{-1} M^{-1}$

enfin $\mu = V^{-1} d$

et $\dot{\mu} = \dot{V}^{-1} \dot{d} = M'^{-1} V^{-1} M^{-1} M d = M'^{-1} V^{-1} d = M'^{-1} \mu$

$f^*{}_{ij} = x_{ij} + \sum_{k=1}^{\nu} u_{ij}^{(k)} \mu_k$ et $\sum_{k=1}^{\nu} u_{ij}^{(k)} \mu_k$ est le terme de la colonne d'indice ij du vecteur $U'\mu$; or, en vertu de ce qui précède,

$$\dot{U}'\dot{\mu} = U' M' M'^{-1} \mu = U'\mu,$$

d'où l'invariance annoncée.

VI. Conclusions sur les méthodes d'estimation.

Le nombre des méthodes qui s'offrent à nous, nous oblige à effectuer un choix. On sera guidé pour cela par les indications données au cours de leur présentation; nous les résumons et les complétons quelque peu.

Le fait que l'estimateur $(\hat{p}, \hat{q})$ soit exhaustif quelle que soit la taille de l'échantillon, confère à la méthode du maximum de vraisemblance un intérêt tout particulier. Lorsque $\hat{p}$ peut être explicité, il doit être utilisé. Dans le cas particulier envisagé au IV-2-c, la méthode préconisée là (modifiée le cas échéant à partir des remarques de V-3) ne demande pas en général de calculs prohibitifs (par contre l'algorithme décrit au IV-2-b s'est avéré en pratique d'une convergence très lente). Pour la troncature la plus générale, l'algorithme R.A.S. a été programmé et utilisé sur plusieurs exemples au moyen de l'ordinateur I.B.M. 7044 qui équipe l'Institut de Calcul Numérique de la Faculté des Sciences de Toulouse; la programmation est très simple, peut servir à des fins variées, et la convergence s'est toujours avérée sur les exemples considérés assez rapide (une dizaine d'itérations pour une précision de 7 ou 8 chiffres). Pour de « petites tables » on peut envisager d'utiliser cette méthode sans l'aide d'un ordinateur; quelle que soit la précision obtenue, il est toujours possible d'écrire $\hat{f}_{ij} = \hat{p}_i \hat{q}_j$ pour tout (i, j) appartenant à R et de déduire ainsi des valeurs approchées de $\hat{p}$ et $\hat{q}$.

Pour les grands échantillons, on a vu comment le processus itératif pouvait être évité. Si l n'est pas très grand la méthode exposée au V-3 sera retenue, mais elle n'est guère plus recommandable que la méthode itérative lorsque l et c sont grands (par exemple supérieurs à 10) et que le recours à l'ordinateur est pratiquement nécessaire. Enfin, la dernière méthode indiquée est évidemment intéressante lorsque ν est petit; mais les fréquences théoriques f^*_{ij} obtenues ne peuvent pas se mettre en général sous la forme $p^*_i q^*_j$; ceci est très gênant pour les problèmes d'estimation proprement dits (mais non pour le calcul éventuel d'une statistique de test). Signalons enfin, que lorsque des estimateurs non efficients peuvent être suffisants, il est toujours possible d'en obtenir de façon simple (V-2).

CHAPITRE III

TESTS DE H_R APPLICATION A L'ANALYSE STATISTIQUE D'UN TABLEAU DE CORRÉLATION

Nous indiquons dans cette partie divers tests de l'hypothèse H_R dans une table de contingence tronquée associée à R. Les tests envisagés sont essentiellement de la famille du χ^2; toutes les propriétés importantes sont asymptotiques; dans la pratique, pour des échantillons de taille nécessairement finie, les résultats obtenus doivent être considérés comme des approximations; toutefois, lorsqu'aucun malentendu ne peut en découler, on notera souvent comme une égalité, sans préciser la taille d'un échantillon, ce qui n'est vrai qu'asymptotiquement : par exemple, l'on parlera de « test de seuil α ». A titre de complément, nous étudions en outre un test « exact » pour un cas simple. Nous signalons les possibilités d'utilisation de H_R à l'analyse statistique de la structure d'un tableau de corrélation ordinaire.

Enfin, nous discutons l'emploi, comme alternative ou complément aux méthodes présentées, de tests basés sur des intervalles de confiance pour les interactions du premier ordre.

I. Tests de H_R — Cas des grands échantillons.

Si la v. a. Y suit une loi de χ^2 à ν degrés de liberté (ddl) on notera $l(\alpha, \nu)$ la valeur telle que :

$$Pr\,[Y > l(\alpha, \nu)] = \alpha$$

On notera $\bar{f}_{ij}$ les fréquences théoriques obtenues sous l'hypothèse H_R par une méthode quelconque conduisant à des estimateurs R.B.A.N.

Si l'on tire au hasard de la population réduite aux éléments $\alpha_i \beta_j$ où $(i, j) \in R$, un échantillon de taille fixée n, la v. a.

$$\chi^2_R = \sum_{(i,j)\in R} \frac{(x_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}$$

suit asymptotiquement une loi de χ^2 à $\nu = t - l - c + 1$ ddl, si H_R est vraie. On lui associera la région critique $\chi^2 > l(\alpha, \nu)$, pour obtenir un test de seuil α.

Ce résultat est encore valide si l'on remplace $\hat{f}_{ij}$ par n'importe quelle autre quantité $\bar{f}_{ij}$.

Remarques. — 1. Ceci est vrai si R est connexe, et seulement dans ce cas. Si R admet k composantes connexes R_i $(i = 1, 2, \ldots, k)$, le nombre de

degrés de liberté de χ^2_R est $t - l - c + k$ (théorème VIII et corollaire) ; on pourra dans ce dernier cas étudier H_{R1}, H_{R2}, ..., H_{Rk} (cf. théorème VII et corollaire) et retenir H_R si l'ensemble des hypothèses H_{R_i} est retenu. On notera que, si l'on combine les k tests en ajoutant les χ^2 correspondants, on retrouve :

$$\chi^2_R = \sum_{i=1}^{k} \chi^2_{R_i}$$

2. Le test décrit plus haut est valable pour les trois cas d'échantillonnage envisagés, en effet :

Si l'échantillon dont on dispose est extrait de la population complète (2e cas), la contribution au χ^2 des fréquences x_{ij}, $(i, j) \in C - R$, est nulle.

Si des totaux marginaux (par exemple b, mais ce peut être seulement une partie des b_j, $j \in R^*$) sont fixés, l'on a vu que l'expression des fréquences $\hat{f}_{ij}$ ne changeait pas, le même test est toujours utilisable (NEYMAN 1949). (Nous avons aussi donné (1962 a) une démonstration de ce fait, en utilisant les méthodes de CRAMER 1946).

3. On a vu au Chapitre I que, si les parties S et R sont telles que $S = R \cup \{h, k\}$ avec $(h, k) \notin \bar{R}$, alors H_R et H_S sont équivalentes (Théorème V) ; ainsi, l'on peut se contenter d'éprouver H_R à la place de H_S. D'ailleurs, il y a une absolue identité entre les deux tests car $\hat{f}_{hk} = x_{hk}$ et les $\hat{f}_{ij}$, $(i, j) \in R$, sont les mêmes sous les hypothèses H_R et H_S (cf. Chapitre II-IV-2. a. Remarque). De là, $\chi^2_S = \chi^2_R$. L'équivalence de H_R et H_S implique ensuite l'égalité des d.d.l. respectifs de χ^2_S et χ^2_R.

Ainsi, à titre d'exemple, le test de H_R dans la table tronquée étudiée par KASTENBAUM (1958), se réduit à un test d'indépendance dans une table 2×2.

Le test du χ^2, tel qu'il est décrit ci-dessus, peut être remplacé par des tests asymptotiquement équivalents : rapport de vraisemblance, χ^2_1 (cf. NEYMAN, 1949). Pour tous ces tests, les calculs numériques présentent à peu près les mêmes difficultés, puisque la partie la plus longue consiste en général pour chacun à obtenir des fréquences théoriques $\hat{f}_{ij}$ sous l'hypothèse H_R (du moins lorsque R est telle que les $\hat{f}_{ij}$ peuvent être difficilement explicités). Si l'on ne désire pas la valeur des fréquences théoriques, on peut remarquer que le χ^2_1 de NEYMAN se simplifie un peu lorsque l'on pose : $\bar{f}_{ij} = f^*_{ij}$. L'on aura, en conservant les notations du Chapitre II (V-4) :

$$\chi^2_{1R} = W^2 = \sum_{(i,j) \in R} \frac{(f^*_{ij} - x_{ij})^2}{x_{ij}} = \mu' V \mu = \mu' d = d' V^{-1} d$$

W^2 ne dépend pas du choix des ν interactoins indépendantes qui composent le vecteur d (cf. Chapitre II-V-4. Remarque 2).

Récemment, BHAPKAR (1966) a montré que cette statistique est toujours identique à celle de WALD (1943), ce qui est bien vérifié pour le cas actuel.

La comparaison des difficultés d'emploi des divers tests est parallèle à la comparaison des méthodes d'estimation et pourra donc être négligée ici. Nous n'avons pas cherché à préciser la validité de ces tests pour des valeurs modérées ou petites de n : les difficultés matérielles d'un traitement satisfaisant sont énormes, et vraisemblablement sans commune mesure avec l'intérêt des résultats que l'on pourrait obtenir; il semble que l'on doive s'en tenir aux indications générales données par COCHRAN (1952, 1954) pour les tests d'indépendance bien que des études récentes par la méthode de Monte-Carlo, sur certains cas particuliers, puissent être plus encourageantes (voir par exemple LEWONTIN et FALSENSTEIN, 1965).

II. Application à l'analyse d'un tableau de corrélation.

1. Introduction.

Les hypothèses envisagées dans ce travail sont paramétriques, en ce sens qu'elles peuvent s'exprimer comme l'appartenance d'un paramètre θ à un certain ensemble. Nous considérons que θ peut varier *a priori* dans un espace Ω : l'hypothèse $\mathcal{H}$ sera définie par $\theta \in \Omega$, c'est l'hypothèse la moins restrictive, dans la pratique celle dont on est *sûr* qu'elle est vraie. Une hypothèse *simple* est un élément de $\mathcal{H}$: elle est équivalente à $\theta = \theta_0 \in \Omega$, θ_0 fixé; une hypothèse *composée* est une partie de $\mathcal{H}$: elle est équivalente à $\theta \in \omega \subset \Omega$.

Soit $H_1 <=> \theta \in \omega_1$ et $H_2 <=> \theta \in \omega_2$, $H_1 => H_2$ est équivalent à $\omega_1 \subset \omega_2$; on écrira parfois $H_1 \subset H_2$ pour $H_1 => H_2$. Si une suite d'hypothèses composées H_i est telle que $H_1 => H_2 => \ldots => \mathcal{H}$ ($H_1 \subset H_2 \subset \ldots\ldots \subset \mathcal{H}$), on dira qu'il s'agit d'hypothèses *emboîtées* ou *gigognes* (en anglais : nested hypotheses).

Lorsque l'on parle de test d'une hypothèse H_0 sans plus de précision, l'alternative envisagée est $\mathcal{H} - H_0$. Soit $H_0 \subset H_1 \subset \mathcal{H}$; si l'on envisage un test de H_0 contre les alternatives appartenant à H_1 (c'est-à-dire contre $H_1 - H_0$), l'on dira qu'il s'agit d'un *test restreint* de H_0 à l'intérieur de H_1.

Revenant à notre problème plus particulier, nous étudions par la suite ce que peut apporter l'introduction de H_R à l'étude d'une table de contingence complète (à laquelle est associée la partie C); dans cette dernière, H_0 est l'hypothèse d'indépendance complète entre les deux classifications. On notera ν_R le nombre de liaisons imposées par H_R entre les probabilités p_{ij}.

Les statistiques notées χ^2 peuvent être l'une quelconque des statistiques de tests asymptotiquement équivalentes citées plus haut; les indices se

réfèrent toujours au type d'hypothèse sous laquelle est calculée cette statistique (et jamais à quelque modification dans la façon de la calculer).

2. Test restreint de H_C.

Puisque $R \subset C$, on a $H_C \Rightarrow H_R$. Si l'on désire effectuer un test de H_C contre une alternative bien définie H_R, on peut calculer :

$$\chi^2_{H_C/H_R} = \chi^2_C - \chi^2_R$$

qui suit asymptotiquement une loi de χ^2 à $\nu_C - \nu_R$ ddl si H_C est vraie (NEYMAN, 1949) ; une région critique de seuil α sera donc :

$$\chi^2_{H_C/H_R} > l(\alpha, \nu_C - \nu_R)$$

Adopter cette région critique au lieu de la région critique habituelle $\chi^2_C > l(\alpha, \nu_C)$, pourra augmenter la puissance du test contre les alternatives contenues dans H_R (cf. FIX, HODGES et LEHMANN 1959).

3. Étude d'hypothèses gigognes.

a) *Première méthode.*

Les parties $R_1, R_2, \ldots, R_m$ sont telles que :

$$R_1 \subset R_2 \subset \ldots \subset R_m \subset C$$

ou même, plus généralement, telles que :

$$H_C \Rightarrow H_{Rm} \Rightarrow H_{Rm-1} \Rightarrow \ldots \Rightarrow H_{R2} \Rightarrow H_{R1}$$

Un test à décisions multiples peut être basé sur le calcul des χ^2 successifs. On sait, par extension immédiate de l'étude de NEYMAN (1949), que les v. a. :

$$\chi^2_1 = \chi^2_{H_{R_1}/\mathcal{H}_0} = \chi^2_{R_1}$$

$$\chi^2_2 = \chi^2_{H_{R_2}/H_{R_1}} = \chi^2_{R_2} - \chi^2_{R_1}$$

$$\chi^2_i = \chi^2_{H_{Ri}/H_{Ri-1}} = \chi^2_{R_i} - \chi^2_{R_{i-1}}$$

etc...

suivent asymptotiquement des lois de χ^2 *indépendantes,* centrées si H_C est vraie, les d.d.l. respectifs étant

$$\nu_1 = \nu_{R1}, \nu_2 = \nu_{R2} - \nu_{R1}, \ldots\ldots, \nu_i = \nu_{Ri} - \nu_{Ri-1}, \ldots.$$

Les décisions peuvent être prises ainsi :

Accepter H_{R1} si et seulement si $\chi^2_1 < l(\alpha_1, \nu_1)$

Accepter H_{R2} si et seulement si H_{R1} est acceptée et $\chi^2_2 < l(\alpha_2, \nu_2)$

.....

Accepter H_{R_i} si et seulement si $H_{R_{i-1}}$ est acceptée et $\chi^2_i < l(\alpha_i, \nu_i)$

. . . . . .

Une telle méthode semble logique dans la mesure où elle procède des hypothèses les moins restrictives aux plus restrictives; si H_{R_j} est rejetée, il est normal de rejeter H_{R_i} pour $i > j$ car $H_{R_i} \Rightarrow H_{R_j}$; d'autre part, il est naturel, ayant accepté $H_{R_{i-1}}$ d'éprouver H_{R_i} « à l'intérieur de $H_{R_{i-1}}$ » par un test restreint; enfin, l'indépendance stochastique des diverses statistiques de tests est un attrait supplémentaire.

Malgré cette indépendance, il faut souligner que l'épreuve de H_{R_i} $(i \neq 1)$ est *conditionnée* par le résultat des épreuves de H_{R_j} pour $j < i$. Ainsi, il conviendrait d'étudier en détail les caractéristiques (seuil, puissance) réelles des tests ainsi définis.

Soit α^*_i le seuil de signification du test de H_{R_i} $(i \neq 1)$; on a :

$$1 - \alpha^*_i = Pr\,[\text{accepter } H_{R_i}/H_{R_i}]$$

$$= Pr\,[\chi^2_1 < l(\alpha_1, \nu_1) \text{ et } \chi^2_2 < l(\alpha_2, \nu_2) \text{ et } \ldots . \text{ et } \chi^2_i < l(\alpha_i, \nu_i)/H_{R_i}]$$

Puisque les v. a. χ^2_j $(j = 1, 2, .., i)$ sont indépendantes d'une part, et que, d'autre part, toutes suivent des lois de χ^2 centrées sous l'hypothèse H_{R_i} ($H_{R_i} \Rightarrow H_{R_j}$ pour $j < i$), il vient :

$$1 - \alpha^*_i = \prod_{j=1}^{i} (1 - \alpha_j)^2 \quad \text{soit} \quad \alpha^*_i = 1 - \prod_{j=1}^{i} (1 - \alpha_j)^2$$

Le seuil de signification n'est pas α_i, mais α^*_i qui lui est *supérieur*.

Etudions maintenant l'erreur de deuxième espèce, relative à H_{R_i}, c'est-à-dire la probabilité d'accepter H_{R_i} alors qu'elle est fausse : c'est une fonction de l'alternative effectivement vérifiée; supposons celle-ci telle que la distribution de χ^2_i (qui est alors voisine d'un χ^2 non centré à ν_i d.d.l.) admette un paramètre de non centralité λ.

Supposons d'abord $H_{R_{i-1}}$ vraie; toujours avec les mêmes approximations sur la distribution des statistiques en cause, l'on a :

— Si l'on réalise un test $\mathcal{C}_1$ de H_{R_i} comme il est indiqué ci-dessus, l'erreur de deuxième espèce est :

$$\beta_1(\alpha_1, \alpha_2, \ldots, \alpha_i; \lambda) = \prod_{j=1}^{i} (1 - \alpha_j) \cdot \beta(\alpha_i, \lambda)$$

en posant $\beta(\alpha_i, \lambda) = Pr[\chi^2_i < l(\alpha_i, \nu_i)]$.

— Si l'on réalise un test restreint unique ($\mathcal{C}_2$) de H_{R_i} à l'intérieur de $H_{R_{i-1}}$, *au seuil* α^*_i, l'erreur de deuxième espèce est :

$$\beta_2(\alpha^*_i, \lambda) = Pr[\chi^2_1 < l(\alpha^*_i, \nu_i)] = \beta(\alpha^*_i, \lambda)$$

Pour comparer les « performances » de ces deux tests (ce qui a un sens puisqu'ils sont de même seuil), il faut comparer β_1 à β_2.

Nous n'essayerons pas ici de faire une comparaison numérique précise pour toutes les valeurs des paramètres, mais simplement de montrer, par un argument simple, quel sera le signe de la différence $\beta_1 - \beta_2$ dans les cas les plus intéressants.

Comme ceci est toujours vérifié dans la pratique, nous supposerons les α_j $(j = 1, 2, \ldots, i)$ petits, et surtout, petits devant i pour que soit suffisamment valable l'approximation $\prod_{j=1}^{i} (1 - \alpha_j) \simeq 1 - \sum_{j=1}^{i} \alpha_j$

En second lieu, nous supposerons que, pour des valeurs de la variable comprises entre $l(\alpha_i, \nu_i)$ et $l(\alpha^*_i, \nu_i)$ la densité du χ^2 centré (à ν_i d.d.l.) est inférieur à celle du χ^2 non centré (à ν_i d.d.l.) de paramètre de non centralité λ. Ceci est toujours le cas, sauf si λ est trop grand (alors, l'erreur de deuxième espèce est voisine de zéro pour les deux types de tests).

Dans ces conditions, l'on a :

$$\beta(\alpha_i, \lambda) - \beta(\alpha^*_i, \lambda) > \alpha^*_i - \alpha_i$$

$$\text{or } \alpha^*_i \simeq \sum_{j=1}^{i} \alpha_j \qquad \text{soit} \qquad \alpha^*_i - \alpha_i \simeq \sum_{j=1}^{i-1} \alpha_i$$

d'où, l'inégalité :

$$\beta(\alpha_i, \lambda) - \beta(\alpha^*_i, \lambda) > \sum_{j=1}^{i-1} \alpha_j \tag{1}$$

(Il est évident que la différence $\beta(\alpha_i, \lambda) - \beta(\alpha^*_i, \lambda)$ peut être précisée en utilisant des tables de la distribution non centrale de χ^2).

L'on a ensuite :

$$\beta_1 = \beta(\alpha_i, \lambda) \prod_{j=1}^{i-1} (1 - \alpha_j) \simeq \beta(\alpha_i, \lambda) - \beta(\alpha_i, \lambda) \sum_{j=1}^{i-1} \alpha_j \tag{2}$$

Finalement, $\beta_1 - \beta_2 = \beta_1 - \beta(\alpha_i, \lambda) + \beta(\alpha_i, \lambda) - \beta(\alpha^*_i, \lambda)$

nous donne en vertu de (1) et (2) :

$$\beta_1 - \beta_2 > [1 - \beta(\alpha_i, \lambda)] \sum_{j=1}^{i-1} \alpha_j > 0$$

Donc, si l'on suppose H_{Ri-1} vraie, l'erreur de deuxième espèce est moins importante en utilisant un test restreint unique $\mathcal{C}_2$ de H_{Ri} contre $H_{R_{i-1}}$ qu'en réalisant un test de H_{Ri} lors d'une suite de tests d'hypothèses gigognes.

Cependant, si H_{Ri-1} est fausse, les valeurs des deux tests peuvent être renversées, puisque, dans β_1, apparaît à la place de $(1 - \alpha_{i-1})$ un facteur $Pr[\chi^2_{i-1} < l(\alpha_{i-1}, \nu_{i-1})]$ qui lui est *inférieur* quelle que soit l'actuelle alternative remplaçant H_{Ri-1}.

Il serait important de comparer en outre $\mathcal{C}_1$ et $\mathcal{C}_2$ au test $\mathcal{C}_3$ de $\mathbf{H}_{\mathbf{R}_i}$ basé simplement sur la région critique $\chi^2_{R_i} > l(\alpha^*_i, \nu_{Ri})$ ($\mathcal{C}_3$ est encore de seuil α^*_i comme $\mathcal{C}_1$ et $\mathcal{C}_2$). On sait que la puissance de $\mathcal{C}_3$ est inférieure à celle de $\mathcal{C}_2$ si H_{Ri-1} est vraie, mais peut être supérieure quand H_{Ri-1} est fausse (FIX, HODGES et LEHMANN, 1959); dans ces conditions, il est tentant de penser que $\mathcal{C}_1$ et $\mathcal{C}_3$ pourraient avoir dans tous les cas des performances voisines, et il serait intéressant d'entreprendre une vérification. C'est une étude que nous espérons aborder ultérieurement en nous plaçant dans un cadre plus général que celui qui est envisagé ici; d'ailleurs, la discussion entamée plus haut est valable pour toute épreuve d'hypothèses gigognes effectuée par une suite de tests successifs analogues au test du χ^2.

b) *Deuxième méthode.*

Dans la pratique, il arrive souvent que le praticien commence par éprouver une hypothèse assez restrictive, ici ce sera H_C, puis, l'ayant trouvée inacceptable, cherche quelque structure de remplacement. Supposons, que lorsque le test de H_C a été trouvé significatif, l'on cherche à éprouver H_R; R doit être fixée à l'avance, (avant l'expérience), ou, du moins, choisie pour des raisons que l'on peut considérer comme « extérieures » à la vue de l'échantillon. Alors, les décisions sont prises ainsi :

Accepter H_C (et en particulier H_R) si $\chi^2_C < l(\alpha_1, \nu_C)$
Rejeter H_C mais accepter H_R si $\chi^2_C > l(\alpha_1, \nu_C)$ et $\chi^2_R < l(\alpha_2, \nu_R)$
Rejeter H_R si $\chi^2_C > l(\alpha_1, \nu_C)$ et $\chi^2_R > l(\alpha_2, \nu_R)$

Les caractéristiques du test de H_R ne sont pas les mêmes que s'il était effectué seul. Nous nous proposons de faire ultérieurement une étude plus détaillée dans un cadre plus général de tests à décisions multiples; mais, nous noterons déjà ici quelques faits simples :

$$Pr[\text{Rejet de } H_R] = Pr[\chi^2_C > l(\alpha_1, \nu_C) \text{ et } \chi^2_R < l(\alpha_2, \nu_R)]$$
$$= Pr[\chi^2_R > l(\alpha_2, \nu_R)] \cdot Pr[\chi^2_C > l(\alpha_1, \nu_C) / \chi^2_R > l(\alpha_2, \nu_R)]$$
Donc : $Pr[\text{Rejet de } H_R] \leqslant Pr[\chi^2_R > l(\alpha_2, \nu_R)]$

Cette inégalité est vraie quelle que soit la distribution des v.a. considérées, c'est-à-dire que H_R soit vraie ou non. Mais l'égalité peut avoir lieu. En effet, on a toujours $\chi^2_C > \chi^2_R$ ou au moins $Pr[\chi^2_C - \chi^2_R < 0] \to 0$ si la taille de l'échantillon tend vers l'infini, puisque $\chi^2_C - \chi^2_R$ suit une loi de χ^2 (centré ou non) asymptotiquement.

D'où : $\chi^2_R > l(\alpha_2, \nu_R) \implies \chi^2_C > l(\alpha_2, \nu_R)$
Si $l(\alpha_2, \nu_R) \geqslant l(\alpha_1, \nu_C)$ (3)
on a $\chi^2_R > l(\alpha_2, \nu_R) \implies \chi^2_C > l(\alpha_1, \nu_C)$
soit $Pr[\chi^2_C > l(\alpha_1, \nu_C) / \chi^2_R > l(\alpha_2, \nu_R)] = 1$
Donc si (3) est vérifié, l'on a :

$$Pr[\text{Rejet de } H_R] = Pr[\chi^2_R > l(\alpha_2, \nu_R)]$$

Les caractéristiques (seuil, puissance) du test de H_R réalisé à la suite du test de H_C, sont alors les *mêmes* que les caractéristiques du test « isolé » de H_R.

On remarquera que la condition (3) est impossible avec $\alpha_1 = \alpha_2$ car $\nu_C > \nu_R$; mais elle peut être réalisée par un choix convenable de α_1 et α_2 ($\alpha_2 < \alpha_1$).

Remarque. — Au sujet d'un test de H_R suivant un test significatif de H_C, il faut noter que, le plus souvent, R risque d'être choisie plus ou moins en raison de l'échantillon observé. D'un point de vue pratique, cette démarche nous paraît naturelle, à condition de ne pas appliquer alors sans discernement des méthodes mathématiques reposant sur des bases qui sont, de toute évidence, violées. A titre d'exemple, on peut être tenté de choisir R parce qu'un certain nombre de fréquences x_{hk}, $(h, k) \in C - R$, paraissent apporter à la valeur élevée de χ^2_C la contribution décisive. On cherche à « mettre en évidence » une partie de la table complète telle que χ^2_R n'est pas significatif. Dans ce cas, il faut considérer χ^2_R comme le plus petit des χ^2 obtenus en tronquant la table initiale.

Si la région critique adoptée est $\chi^2_R > l(\alpha, \nu_R)$, le *choix* de R abaisse le seuil réel de signification mais aussi la puissance. Si $\chi^2_R > l(\alpha, \nu_R)$, on pourra rejeter R à un seuil inférieur à α; mais si $\chi^2_R < l(\alpha, \nu_R)$ une conclusion ne devra être tirée que prudemment; ceci peut fournir toutefois une indication importante quand une expérience nouvelle peut être envisagée : le choix de R sera fait alors a priori.

III. Intervalles de confiance pour des interactions du premier ordre et leur utilisation pour le test de H_R.

Une méthode simple a été introduite par GOODMAN (1964 a), qui permet de trouver, pour un nombre quelconque d'interactions du premier ordre, des intervalles de confiance simultanés de coefficient de sécurité au moins égal à $1 - \alpha$. Goodman montre que la probabilité de l'événement

$$\forall u \in I(C),\ |\delta(u, p) - \delta(u, x)| \leqslant \sqrt{l(\alpha, \nu_C)\ v(u, x)} \qquad (4)$$

est égale à $(1 - \alpha)$; dans l'expression ci-dessus $v(u, x)$ est l'estimation de la variance de $\delta(u, x)$ c'est-à-dire

$$v(u, x) = \sum_{(i,j) \in C} u^2_{ij} / x_{ij}$$

On pourra montrer en raisonnant de façon analogue :
La probabilité que soient vérifiées les inégalités

$$\forall u \in I(R),\quad |\delta(u, p) - \delta(u, x)| \leqslant \sqrt{l(\alpha, \nu_R)\ v(u, x)} \qquad (5)$$

est égale à $(1 - \alpha)$.

D'où l'on obtient, pour l'ensemble des interactions $\delta(u, p)$ où $u \in I(R)$, les intervalles de confiance simultanés de coefficient de sécurité $(1 - \alpha)$:

$$\delta(u, x) \pm \sqrt{l(\alpha, \nu_R)\, v(u, x)} \qquad (6)$$

Ces intervalles sont plus courts que les intervalles

$\delta(u, x) \pm \sqrt{l(\alpha, \nu_C)\, v(u, x)}$ que l'on peut déduire de (4).

On peut montrer que, dès que l'un de ces intervalles ne contient pas zéro, alors W^2 (tel qu'il est défini au paragraphe I de ce Chapitre) est supérieur à $l(\alpha, \nu_R)$.

Ceci permet d'utiliser, pour éprouver H_R, la méthode suivante (suggérée dans un contexte peu différent par Goodman) :

Chercher parmi les interactions paraissant les plus « suspectes » s'il en existe une telle que l'intervalle $\delta(u, x) \pm \sqrt{l(\alpha, \nu_R)\, v(u, x)}$ ne contienne pas zéro; c'est-à-dire, chercher s'il existe $u \in I(R)$ tel que $[\delta(u, x)]^2 > l(\alpha, \nu_R)\ v(u, x)$.

— S'il en existe un, on a nécessairement $W^2 > l(\alpha, \nu_R)$ et le test du χ^2 est significatif : on rejette H_R. Sinon, on calculera effectivement W^2 ou quelqu'autre statistique équivalente.

La première opération étant la plus simple on peut éviter, par cette démarche, beaucoup de calculs.

L'on peut se contenter de choisir ν interactions, c'est-à-dire ν vecteurs indépendants $u^{(k)}$ appartenant à I (R), de calculer les intervalles de confiance correspondants et de rejeter H_R si et seulement si au moins l'un d'entre eux contient zéro. Le test ainsi réalisé est uniformément moins puissant que le test basé sur W^2 (ou quelque autre χ^2 équivalent). Cependant, il est facile de construire pour ces interactions des intervalles de confiance simultanés, de coefficient de sécurité au moins égal à $(1 - \alpha)$ plus courts que les précédents (cf. Goodman, 1964 a), à savoir les ν intervalles :

$$d_k \pm c\left(\frac{\alpha}{\nu}\right) \sqrt{v(u^{(k)}, x)}$$

($v(u^{(k)}, x) = v_{kk}$ avec les notations introduites précédemment); $c(\alpha)$ est défini par : $Pr[|Z| > c(\alpha)] = \alpha$ si Z est une v.a. normale réduite (d'où la relation $c(\alpha) = \sqrt{l(\alpha, 1)}$).

Utilisant ces nouvaux intervalles de confiance, on peut rejeter H_R à un seuil de signification moindre que α, si l'un d'eux ne contient pas zéro. Ce test est uniformément plus puissant que celui basé sur les intervalles de confiance plus longs (6); l'efficacité relative par rapport au χ^2 dépend de l'alternative (pour une discussion similitaire, voir Scheffé 1959, p. 82-83). Il nous semble cependant que son emploi, dans le contexte considéré, est moins recommandable que l'emploi du χ^2, simplifié le cas échéant par l'utilisation des intervalles (6). En effet, ce dernier test présenté n'est pas indépendant du choix des ν interactions

utilisées; or, il sera bien rare dans notre problème que ce choix puisse être guidé par quelque considération objective.

IV. Quelques remarques supplémentaires et conclusion sur les tests relatifs aux grands échantillons.

1. Test « isolé » de H_R.

Nous résumons d'abord ce qui précède au sujet du test « isolé » de H_R. Ces remarques sont valables que H_R soit liée à une table de contingence complète ou tronquée. L'on a étudié différentes formes du test du χ^2, la plupart des calculs passent par la résolution de problèmes d'estimation ou en sont très voisins; pour certaines formes de R ces calculs sont assez simples; pour d'autres ils sont plus longs, la principale difficulté est alors soit la mise en œuvre d'un processus itératif, soit l'inversion d'une matrice dont l'ordre peut ne pas dépasser cependant $\sup(l-1,\ c-1,\ \nu)$.

Au paragraphe III on introduit une technique qui évitera dans certains cas le calcul du χ^2. Nous indiquons ci-dessous une nouvelle possibilité analogue à cette dernière, qui peut réduire considérablement les calculs : si $R' \subset R$, on a $\chi^2_{R'} < \chi^2_R$. Il peut arriver que $\chi^2_{R'}$ soit bien plus facile à calculer que χ^2_R, par exemple si les fréquences théoriques $\hat{f}_{ij}$ peuvent être explicitées sous l'hypothèse $H_{R'}$; un cas très simple est celui où $R' = \tilde{R}'$, mais nous avons vu qu'il en existe d'autres. On peut alors calculer $\chi^2_{R'}$, et si $\chi^2_{R'}$ est supérieur à $l(\alpha, \nu_R)$ rejeter H_R puisque a fortiori χ^2_R sera supérieu à $l(\alpha, \nu_R)$; mais si $\chi^2_{R'} < l(\alpha, \nu_R)$ l'on doit obtenir χ^2_R pour conclure (dans ces conditions, les calculs auront été, au contraire, allongés; il en va de même d'ailleurs pour la méthode basée sur les intervalles de confiance).

2. Analyse de la corrélation.

Pour une table de contingence $(r \times s)$, $(r-1)(s-1)$ composantes indépendantes du χ^2 ont été isolées (Irwin, 1949; Lancaster, 1949) ce qui permet une décomposition « de principe » du χ^2 allant vers une analyse de la corrélation. Cependant, si cette décomposition peut avoir parfois une interprétation pratique, il n'en est pas toujours ainsi. L'utilisation de tables tronquées peut avoir alors un intérêt.

Nous pensons que l'étude de H_R peut être souvent intéressante pour R peu différent de C, cas où un petit nombre de configurations de caractères ont des fréquences « anormales » qui pourraient être seules responsables de la fausseté de H_C. Notons qu'en général nous aurons là des parties R pour lesquelles les problèmes d'estimation sont faciles à résoudre, souvent explicitement, d'où la simplicité du test de H_R. Au sujet du cas de ce type le plus simple $(R = C - \{(1,1)\})$ nous avons donné (1962 b) quelques indications supplémentaires.

Mais l'étude de H_R peut être intéressante en bien d'autres circonstances : un cas particulier important sera étudié plus en détail au Chapitre V.

V. Étude d'un test exact.

Lorsque l'hypothèse H_R fixe seulement une liaison entre les paramètres du modèle ($\nu = 1$), il est facile par les méthodes habituelles de construire un test semblable U.M.P. (uniformément le plus puissant) de H_R contre une alternative unilatérale, et un test semblable U.M.P.U. (uniformément le plus puissant parmi les tests sans biais) contre une alternative bilatérale. Nous explicitons ci-dessous cette démarche pour une partie R de dimension (3, 3) telle que : $(i, j) \in R \Longleftrightarrow i, j = 1, 2, 3$ et $i \neq j$. Il s'agit d'une application simple de la théorie générale (cf : TOCHER, 1950; LEHMANN, 1959); d'autres cas de troncature pourront être étudiés de façon analogue.

Nous distinguerons ici une v. a. de sa valeur observée, en désignant la première par une lettre majuscule et la deuxième par la minuscule correspondante. L'ensemble des observations est ici le vecteur

$$x = [x_{12}, x_{13}, x_{21}, x_{23}, x_{31}, x_{32}]$$

en appelant toujours a (a_1, a_2, a_3) et b (b_1, b_2, b_3) les marges du tableau des observations réduit aux cases associées à R, on notera $\mathfrak{X}(a, b)$ l'ensemble des tableaux 3×3 de marges a et b garnis de zéros sur la diagonale principale et de nombres entiers positifs ailleurs (ou plutôt, l'ensemble des vecteurs x représentant ces tableaux).

Dans ces conditions, on a pour tout x appartenant à $\mathfrak{X}(a, b)$:

$$Pr\,[X = x \,/\, X \in \mathfrak{X}(a, b)] = \frac{\prod_{(i,j)\in R} \left[p_{ij}^{x_{ij}} / (x_{ij})\,! \right]}{\sum_{x \in \mathfrak{X}(a,b)} \prod_{(i,j)\in R} \left[p_{ij}^{x_{ij}} / (x_{ij})! \right]}$$

Posons : $\rho = \dfrac{p_{12}\, p_{23}\, p_{31}}{p_{21}\, p_{32}\, p_{13}}$

On peut écrire :

$$p_{13} = p_1 q_3,\ p_{21} = p_2 q_1,\ p_{23} = p_2 q_3,\ p_{31} = p_3 q_1,\ p_{32} = p_3 q_2$$
$$\text{et } p_{12} = \rho\, p_1 q_2$$

H_R est équivalente à $\rho = 1$

Pour tout $x \in \mathfrak{X}(a, b)$, la probabilité de l'événement $X = x$ sera

$$P_\rho\,[X = x / X \in \mathfrak{X}(a, b)] = C(\rho) \quad \frac{\rho^{x_{12}}}{\prod_{(i,j)\in R} (x_{ij})\,!} \tag{7}$$

avec

$$C(\rho) = \left[\sum_{x \in \mathcal{X}(a,b)} \frac{\rho^{x_{12}}}{\prod_{(i,j)\in R} (x_{ij})!} \right]^{-1}$$

Sous l'hypothèse H_R, cette probabilité devient :

$$P_1 [X = x / X \in \mathcal{X}(a,b)]$$

$$= \left[\prod_{(i,j)\in R} (x_{ij})! \cdot \sum_{x \in \mathcal{X}(a,b)} \frac{1}{\prod_{(i,j)\in R} (x_{ij})!} \right]^{-1} \qquad (8)$$

Ceci donne la distribution conditionnelle de X, les totaux marginaux a et b étant fixés; dans ces conditions, de la valeur de X_{12} (par exemple) se déduisent les valeurs des autres v. a. X_{ij}.

Le test cherché est un test conditionnel basé sur cette distribution. Pour obtenir un seuil α en dépit de son caractère discret, nous introduisons une probabilité de rejet de H_R :

$$\phi(x_{12}) = Pr\ [\text{Rejet de } H_R / X_{12} = x_{12}]$$

(on peut évidemment remplacer la statistique de test x_{12} par tout autre x_{ij}; d'autre part, pour simplifier la typographie, nous omettons maintenant les indices et noterons x pour x_{12} et X pour X_{12}).

Un test U.M.P. de seuil α, de l'hypothèse $\rho = 1$ contre l'alternative $\rho < 1$ est donné par (cf : LEHMANN, 1959) :

$$\begin{cases} \phi(x) = 1 & \text{si } x < K \\ \phi(x) = \pi & \text{si } x = K \\ \phi(x) = 0 & \text{si } x > K \end{cases}$$

où les constantes K et π seront déterminées par la condition :

$$E_1 [\phi(X)] = \alpha \qquad (9)$$

(l'indice 1 à l'opérateur E signifie que l'espérance mathématique est calculée sous l'hypothèse $\rho = 1$, c'est-à-dire à partir de la distribution (8)).

De même pour un test de $\rho = 1$ contre $\rho > 1$, on prendra :

$$\begin{cases} \phi(x) = 1 & \text{si } x > K' \\ \phi(x) = \pi' & \text{si } x = K' \\ \phi(x) = 0 & \text{si } x < K' \end{cases}$$

où K' et π' seront encore déterminés à partir de (9).

Finalement, un test U.M.P.U. de l'hypothèse $\rho = 1$ contre l'alternative $\rho \neq 1$ sera donné par

$$\begin{cases} \phi(x) = 1 & \text{si } x < K_1 \text{ ou } x > K_2 \\ \phi(x) = \pi_i & \text{si } x = K_i \ (i = 1,2) \\ \phi(x) = 0 & \text{si } K_1 < x < K_2 \end{cases}$$

Ici, les constantes K_1, K_2, π_1, π_2 sont à déterminer par la condition (9) et la nouvelle condition :

$$E_1 [X. \phi (X)] = \alpha E_1 (X) \qquad (10)$$

Utilisation pratique. — Pour se servir de ces tests, il faut calculer l'expression (8) pour tout $x \in \mathcal{X}(a, b)$; pour cela, il suffit de calculer des nombres proportionnels à $\left[\prod_{(i,j) \in R} (x_{ij}) ! \right]^{-1}$ pour tout $x \in \mathcal{X}(a, b)$ et de normer ensuite pour que la somme des probabilités soit 1. Ce travail peut être très long pour de grands échantillons, mais sera très abordable en de nombreux cas, justement lorsque l'emploi du χ^2 n'est guère valide à cause de trop petites fréquences.

Ceci fait, la détermination de K et de π (resp. K′ π') pour les tests unilatéraux est immédiate. Pour le test bilatéral, on pourra procéder de la façon suivante : prendre des valeurs d'essai de K_1 et K_2; les relations (9) et (10) donnent alors deux équations linéaires en π_1 et π_2; si les solutions sont telles que $0 \leqslant \pi_i \leqslant 1$ $(i = 1, 2)$, le test est déterminé, sinon on essayera de nouvelles valeurs pour K_1 et K_2. Pour ces essais, on prendra K_1 et K_2 tels que $Pr [X < K_1] \simeq Pr [X > K_2]$ ces deux probabilités étant proches de $\alpha/2$ (mais inférieures à $\alpha/2$) (dans la pratique il sera toujours facile de « deviner » K_1 et K_2 au premier essai, à la rigueur au deuxième).

On trouvera au Chapitre VI une application de ce test; la suite des opérations décrites ci-dessus y est explicitée.

CHAPITRE IV

GÉNÉRALISATIONS

I. Étude des hypothèses $H_R(\lambda)$.

1. Définition.

Les probabilités p_{ij} ayant la même définition que précédemment, on considère une partie de C, soit R, de dimension (l, c), à t éléments. Associons à l'élément (i, j) de R : $g_{ij} = p_{ij}/\lambda_{ij}$, si les g_{ij} vérifient H_R, on dira que $H_R(\lambda)$ est vraie (on notera g le vecteur de $\mathcal{R}^t$ de composantes g_{ij}, λ le vecteur de composantes λ_{ij}). Donc, sous l'hypothèse $H_R(\lambda)$, p_{ij} peut s'écrire $p_{ij} = \lambda_{ij}\, p_i\, q_j$ pour tout (i, j) appartenant à R. L'adaptation des résultats essentiels des chapitres précédents sera immédiate.

En supposant $p_{ij} > 0$ pour tout (i, j) de R, et en introduisant les interactions du premier ordre $\delta(u, p)$, on aura $\delta(u, g) = \delta(u, p) - \delta(u, \lambda)$; d'où $H_R(\lambda)$ est équivalente à :

$$\forall\, u \in I(R) \quad \delta(u, g) = 0 \quad \text{soit} \quad \delta(u, p) = \delta(u, \lambda)$$

(donc $\delta(u, p)$ est une constante bien déterminée).

2. Méthodes d'estimation.

On pourra généraliser un grand nombre des méthodes indiquées au Chapitre II; nous indiquons rapidement comment en nous restreignant, pour simplifier, au cas d'un échantillon extrait de la population parente réduite aux éléments $\alpha_i\, \beta_j$ où $(i, j) \in R$. La statistique (a, b) est encore exhaustive pour les paramètres (p, q) et les équations de vraisemblance s'écrivent :

$$\begin{cases} \forall\, i \in \bar{R} \quad \hat{p}_i \sum_{j \in R_{i\cdot}} \lambda_{ij}\, \hat{q}_j = \dfrac{a_i}{n} \\ \forall\, j \in R^* \quad \hat{q}_j \sum_{i \in R_{\cdot j}} \lambda_{ij}\, \hat{p}_i = \dfrac{b_j}{n} \end{cases} \quad \text{soit} \quad \begin{cases} \sum_{j \in R_{i\cdot}} \hat{f}_{ij} = a_i \\ \sum_{i \in R_{\cdot j}} \hat{f}_{ij} = b_j \end{cases}$$

On montrera comme au Chapitre II l'existence et l'unicité de la solution, celle-ci rendant la vraisemblance maximum. Dans le cas général il n'est pas possible de résoudre explicitement ces équations. Comme au Chapitre II, l'algorithme R.A.S. peut être utilisé : il n'y a aucun changement à apporter à la suite des opérations (cf. Ch. II-IV-2-a), simplement la matrice de départ $\| t^{(0)}_{ij} \|$ ne doit pas être la matrice d'incidence T, mais sera choisie telle que :

$$t^{(0)}_{ij} = \lambda_{ij} \qquad \text{si } (i, j) \in R$$
$$t^{(0)}_{ij} = 0 \qquad \text{si } (i, j) \notin R$$

Si l'algorithme converge, il donnera les fréquences théoriques $\hat{f}_{ij}$ (pour les questions de convergence, voir Ch. II et Appendice).

D'autres estimateurs R.B.A.N. peuvent être obtenus comme solutions d'un système linéaire en calquant les méthodes du Chapitre II; en particulier la technique du IV-4 est de généralisation immédiate puisqu'il suffit de remplacer $\delta(u^{(k)}, p) = 0$ par $\delta(u^{(k)}, p) - \delta(u^{(k)}, \lambda) = 0$; la matrice V est ainsi inchangée, d_k doit être remplacé par

$$d_k - \delta(u^{(k)}, \lambda) = \delta(u^{(k)}, x) - \delta(u^{(k)}, \lambda) = \delta\left(u^{(k)}, \frac{x}{\lambda}\right)$$

3. **Test de $H_R(\lambda)$.**

Si l'on dispose d'un échantillon suffisamment important, le test du χ^2 ou les tests équivalents sont utilisables, avec les fréquences théoriques estimées ci-dessus. Le nombre de degrés de liberté sera toujours

$$\nu_R = t - l - c + 1$$

si R est connexe. On pourra, en suivant les indications du Chapitre III (§ III), se servir des intervalles de confiance pour les interactions $\delta(u, p)$ avec une seule modification : il faudra maintenant situer par rapport à ces intervalles, à la place de zéro, les quantités $\delta(u, \lambda)$.

On notera aussi que le test exact exposé au Chapitre III (§ V) peut être utilisé pour éprouver dans ce cas $H_R(\lambda)$, en remarquant que $H_R(\lambda)$ est équivalente, toujours avec les notations de ce paragraphe, à :

$$\rho = \frac{\lambda_{12}\,\lambda_{23}\,\lambda_{31}}{\lambda_{21}\,\lambda_{32}\,\lambda_{13}} \quad (= \rho_0)$$

Dans (9) et (10) on remplacera E_1 par E_{ρ_0}.

4. **Conclusion sur l'utilisation de l'hypothèse $H_R(\lambda)$.**

Il semble que l'étude de ce type d'hypothèse ait été quelque peu délaissée par les statisticiens. A notre connaissance, l'unique publication sur ce sujet est celle de Lokki (1960) qui étudie surtout $H_C(\lambda)$ et donne quelque indication sur $H_R(\lambda)$; cet auteur cite un exemple pour lequel son étude pourrait être appliquée, toutefois ce problème nous semble plutôt relever, dans la pratique, de la comparaison de deux tables de contingence.

Nous indiquons au Chapitre V un domaine où le modèle étudié ci-dessus peut être approprié, et nous traitons au Chapitre VI un exemple faisant intervenir une hypothèse $H_R(\lambda)$ où la partie R est différente de C.

II. Tables de contingence tronquées à trois dimensions.

L'hypothèse H_R généralise directement pour une table tronquée (à deux dimensions) l'hypothèse d'indépendance. Nous étudions maintenant une

généralisation analogue de l'hypothèse de non-interaction du deuxième ordre pour les tables tronquées à trois dimensions.

1. Définition - Propriétés.

a) *Notations.* — On notera T l'ensemble des triplés ordonnés (i, j, k) où

$$i = 1, 2, \ldots, l_0 \; ; \; j = 1, 2, \ldots, c_0 \; ; \; k = 1, 2, \ldots, s_0$$

Si S est une partie de T, on notera :

$$S_{ij.} = \{k : (i, j, k) \in S\} \qquad S_{i.k} = \{j : (i, j, k) \in S\}$$

$$S_{i..} = \{(j, k) : (i, j, k) \in S\} \qquad S_{.jk} = \{i : (i, j, k) \in S\}$$

$$S_{..k} = \{(i, j) : (i, j, k) \in S\} \qquad S_{.j.} = \{(i, k) : (i, j, k) \in S\}$$

$$P_1(S) = \bigcup_k S_{..k} \qquad P_2(S) = \bigcup_i S_{i..} \qquad P_3(S) = \bigcup_j S_{.j.}$$

$$L_1(S) = \bigcup_{(j,k)} S_{.jk} \qquad L_2(S) = \bigcup_{(i,k)} S_{i.k} \qquad L_3(S) = \bigcup_{(i,j)} S_{ij.}$$

$$l = \text{Card } L_1(S) \qquad c = \text{Card } L_2(S) \qquad s = \text{Card } L_3(S)$$

$$t_1 = \text{Card } P_1(S) \qquad t_2 = \text{Card } P_2(S) \qquad t_3 = \text{Card } P_3(S)$$

$$\text{et } t = \text{Card } (S)$$

Les notations utilisées qui ne sont pas définies ci-dessus sont les mêmes que celles qui sont définies aux Chapitres précédents.

b) *Définition.* — A chaque élément de S on associe $g_{ijk} \in G$ (on pourra supposer tout de suite que G est le groupe multiplicatif des nombres réels positifs). L'hypothèse K_S sera définie par :

$$\forall (i, j) \in P_1(S) \quad \exists \theta_{ij} \in G, \; \forall (j, k) \in P_2(S) \quad \exists \varphi_{jk} \in G, \; \forall (i, k) \in P_3(S) \quad \exists \psi_{ik} \in G$$

tels que $\qquad \forall (i, j, k) \in S \qquad g_{ijk} = \theta_{ij} \varphi_{jk} \psi_{ik}$

On notera θ l'ensemble des θ_{ij}, $(i, j) \in P_1(S)$, φ et ψ auront une signification analogue.

On considère maintenant une population parente constituée d'individus possédant chacun trois caractères, le premier pouvant prendre l_0 niveaux, le deuxième c_0 et le troisième s_0; la probabilité qu'un individu possède pour chacun de ces trois caractères respectivement les niveaux i, j et k, sera notée p_{ijk}. On posera maintenant $g_{ijk} = p_{ijk}$ (mais une généralisation analogue à celle du paragraphe I de ce chapitre serait immédiate).

c) *Propriétés.* — On peut introduire les interactions du deuxième ordre

$$\gamma(u, p) = \sum_{(i, j, k) \in T} u_{ijk} \operatorname{Log} p_{ijk}$$

où les u_{ijk} sont tels que :

$$\forall (i, j, k) \in T \quad \sum_{h=1}^{l_0} u_{hjk} = \sum_{h=1}^{c_0} u_{ihk} = \sum_{h=1}^{s_0} u_{ijh} = 0$$

On sait que l'hypothèse K_T est équivalente à $\gamma(u, p) = 0$. Les interactions associées à S seront celles pour lesquelles $u_{ijk} = 0$ si $(i, j, k) \notin S$. Les interactions écrites par la suite sont toutes de ce type. On notera donc u le vecteur de $\mathcal{R}'$ de composantes u_{ijk}, $(i, j, k) \in S$; I (S) sera le sous-espace de $\mathcal{R}'$ dont les éléments u sont tels que :

$$\forall (i, j) \in P_1(S) \quad \sum_{k \in S_{ij\cdot}} u_{ijk} = 0$$

$$\forall (j,k) \in P_2(S) \quad \sum_{i \in S_{\cdot jk}} u_{ijk} = 0$$

$$\forall (i,k) \in P_3(S) \quad \sum_{j \in S_{i\cdot k}} u_{ijk} = 0$$

On a alors le

Théorème X. — L'hypothèse K_S est équivalente à :

$$\forall u \in I(S), \quad \gamma(u, p) = 0$$

La démonstration est analogue à celle du théorème IX; on introduira les $(t_1 + t_2 + t_3)$ vecteurs de $\mathcal{R}'$

$e^{(ij)}$ de composantes $e^{(ij)}_{i'j'k'} = \delta_{ii'}\,\delta_{jj'}$ $\quad (i, j) \in P_1(S)$
$f^{(jk)}$ de composantes $f^{(jk)}_{i'j'k'} = \delta_{jj'}\,\delta_{kk'}$ $\quad (j, k) \in P_2(S)$
$g^{(ik)}$ de composantes $g^{(ik)}_{i'j'k'} = \delta_{ii'}\,\delta_{kk'}$ $\quad (i, k) \in P_3(S)$

et l'on montrera que l'hypothèse K_S est équivalente à l'appartenance du vecteur de composantes Log p_{ijk} au sous espace linéaire E engendré par ces $(t_1 + t_2 + t_3)$ vecteurs.

Le nombre de liaisons ν_{K_S}, imposées aux paramètres p_{ijk} par l'hypothèse K_S, est égal à la dimension de I (S) (qui est l'espace orthogonal complémentaire de E).

Il nous a paru difficile de donner pour ν_{K_S} une formule générale maniable, ne dépendant de S que par les nombres $t, t_1, t_2, t_3, l, c, s$ et par une propriété « simple » de sa structure, formule qui généraliserait le résultat du Chapitre I (Théorème VIII et Corollaire); en effet, en représentant par exemple S comme la superposition de « strates » $S_{\cdot\cdot k}$, ν_{K_S} dépend non seulement des propriétés de connexité des diverses « strates » mais aussi des propriétés de connexité de leurs intersections, $S_{\cdot\cdot h} \cap S_{\cdot\cdot k}$, pour h et k appartenant à $L_3(S)$.

Par contre, on obtiendra facilement ν_{K_S} si l'on précise des propriétés de ces intersections, par exemple, en posant $L_3(S) = 1, 2, \ldots, s$ ce qui ne restreint pas la généralité, il sera possible dans presque tous les cas pratiques de numéroter les « strates » $S_{\cdot\cdot k}$ de sorte que $S_{\cdot\cdot(k+1)} \subset S_{\cdot\cdot k}$ pour

$k = 1, 2 \ldots, s-1$; supposons en outre $S_{..k}$ connexe pour tout k; on notera $t^{(k)}_1 = \text{Card}\ (S_{..k})$ $\quad l^{(k)} = \text{Card}\ (\bar{S}_{..k})$ $\quad c^{(k)} = \text{Card}\ (S^*_{..k})$.

Les relations $p_{ij1} = \theta_{ij}\,\varphi_{j1}\,\psi_{i1}$ et $p_{ij2} = \theta_{ij}\,\varphi_{j2}\,\psi_{i2}$ impliquent l'existence de nombres p_i et q_i tels que :

$$\forall\ (i, j) \in S_{..2}\ \ p_{ij2}/p_{ij1} = p_i\, q_j$$

ceci fixe $t^{(2)}_1 - l^{(2)} - c^{(2)} + 1$ liaisons entre les paramètres p_{ij2} (Théorème VIII) ; en continuant de la même façon, on aura :

$$\nu_{K_s} = \sum_{k=1}^{s} (t^{(k)}_1 - l^{(k)} - c^{(k)} + 1) = t - t_1 - t_2 - t_3 + l + c + s - 1$$

Remarques. — 1. On peut élargir la portée de ce résultat en permutant les rôles des trois indices.

2. Ce résultat est valable en particulier si les $S_{..k}$ sont des parties connexes égales pour tout k, ce qui permet de l'utiliser lorsque l'on compare des tables de contingence à deux dimensions identiquement tronquées.

3. Pour les parties S qui ne satisfont pas les conditions imposées ci-dessus, on pourra essayer de calculer ν_{K_S} dans chaque cas particulier par une méthode analogue à la précédente, en décomposant en strates et en appliquant les théorèmes du Chapitre I.

2. Indications sur les tests d'une hypothèse K_S.

On supposera disponible un échantillon tiré au hasard de la population parente, x_{ijk} représente le nombre d'individus figurant dans l'échantillon qui présentent pour chacune des trois classifications respectivement les niveaux i, j, et k. Ce qui suit est valable pour divers cas d'échantillonnage, en particulier échantillon tiré au hasard de la population parente complète ou réduite; nous nous placerons néanmoins dans ce dernier cas seulement, pour simplifier.

Les problèmes d'estimation paraissent ici secondaires, mais si l'on veut appliquer le test classique du χ^2, il faut chercher des fréquences théoriques sous l'hypothèse K_S. La méthode du maximum de vraisemblance conduit aux équations :

$$\left\{\begin{array}{lll} \forall\ (i,j) \in P_1(S) & \hat{\theta}_{ij} \sum_{k \in S_{ij\cdot}} \hat{\varphi}_{jk}\,\hat{\psi}_{ik} = \frac{x_{ij\cdot}}{n} & \\ \forall\ (j,k) \in P_2(S) & \hat{\varphi}_{jk} \sum_{i \in S_{\cdot jk}} \hat{\theta}_{ij}\,\hat{\psi}_{ik} = \frac{x_{\cdot jk}}{n} & \text{soit} \\ \forall\ (i,k) \in P_3(S) & \hat{\psi}_{ik} \sum_{j \in S_{i\cdot k}} \hat{\theta}_{ij}\,\hat{\varphi}_{jk} = \frac{x_{i\cdot k}}{n} & \end{array}\right. \left\{\begin{array}{l} \sum_{k \in S_{ij\cdot}} \hat{f}_{ijk} = x_{ij\cdot} \\ \sum_{i \in S_{\cdot jk}} \hat{f}_{ijk} = x_{\cdot jk} \\ \sum_{j \in S_{i\cdot k}} \hat{f}_{ijk} = x_{i\cdot k} \end{array}\right.$$

(le point à la place d'un indice a la signification habituelle

$$x_{ij.} = \sum_{k \in S_{ij.}} x_{ijk}, \text{ etc... et } n = x... = \sum_{(i,j,k) \in S} x_{ijk})$$

On arrive encore à un problème de construction d'un tableau de corrélation (à trois dimensions) à marges fixées, et garni de zéros dans certaines cases déterminées. L'analogue de la suite définie par l'algorithme R.A.S. sera ici :

$$\begin{cases} t_{ijk}^{(3m+1)} = t_{ijk}^{(3m)} \dfrac{x_{ij.}}{\sum\limits_{k \in S_{ij.}} t_{ijk}^{(3m)}} \\ t_{ijk}^{(3m+2)} = t_{ijk}^{(3m+1)} \dfrac{x_{.jk}}{\sum\limits_{i \in S_{.jk}} t_{ijk}^{(3m+1)}} \\ t_{ijk}^{(3m+3)} = t_{ijk}^{(3m+2)} \dfrac{x_{i.k}}{\sum\limits_{j \in S_{i.k}} t_{ijk}^{(3m+2)}} \end{cases}$$

On posera :

$$t_{ijk}^{(0)} = 1 \quad \text{si} \quad (i,j,k) \in S$$
$$t_{jk}^{(0)} = 0 \quad \text{si} \quad (i,j,k) \notin S$$

Dans ces conditions, si l'algorithme converge, il donnera, à la limite, la solution $\hat{f}_{ijk}$ cherchée.

A partir de ces fréquences théoriques $\hat{f}_{ijk}$, on pourra réaliser un test du χ^2 pour éprouver K_S; le nombre de degrés de liberté est ν_{K_S}, dont la valeur est discutée plus haut.

A ce stade, il convient de noter que l'algorithme ci-dessus généralise très directement, pour les tables tronquées, celui qui est proposé par DARROCH (1962) pour les tableaux complets. En fait, alors que dans le cas de deux dimensions, l'étude générale de l'hypothèse H_R est nettement différente de l'étude classique de l'hypothèse H_C, il n'en est plus de même pour les tables à trois dimensions. Dans le cas de l'hypothèse K_T d'absence (complète) d'interaction du deuxième ordre, de grandes simplifications ne sont plus ici rencontrées, les méthodes nécessaires pour les tables complètes, peuvent être généralisées immédiatement.

Les tests les plus simples pour l'hypothèse K_T ont été donnés par GOODMAN (1963 a, 1963 b, 1964 a, 1964 b); ils sont basés sur la comparaison des interactions du 1er ordre de s tables à deux dimensions, ou, si l'on veut, sur l'étude de la nullité des interactions du deuxième ordre. Or, on a vu que K_S était équivalente à l'annulation de certaines de ces interactions, les méthodes de GOODMAN seront donc généralisées facilement

en sélectionnant seulement ν_{K_S} interactions indépendantes liées à K_S. On peut noter d'ailleurs, que GOODMAN (1963 b) suggère la possibilité, pour comparer plusieurs tables, de sélectionner quelques mesures présentant un intérêt particulier et de comparer ces mesures : la comparaison de s tables présentant la même forme de troncature, disons associée à R, (tables tronquées volontairement) est une opération de ce type, on compare ainsi s ensembles de ν_R interactions du premier ordre (en effet, si S est composée de la « superposition » de s « strates » $S_{..k} = R$, $\gamma(u, p) = 0$ pour tout $u \in I(S)$ est équivalent à $\delta(u, p'_k)$ indépendant de k pour tout $u \in I(R)$, en désignant par p'_k le vecteur de composantes $p_{ijk} \sum_{(i.j) \in S_{..k}} p_{ijk}$; la démonstration est immédiate et sera omise).

Pour comparer certaines tables, il pourrait être intéressant de comparer successivement des interactions en nombre de plus en plus grand; ces tests d'hypothèses gigognes se feront comme au Chapitre III, on en trouvera un exemple simple au Chapitre VI.

Remarque. — Dans le cas invoqué ci-dessus, comme dans le cas où le choix des interactions liées à S n'est pas très simple, l'algorithme que nous avons signalé a l'avantage d'une grande facilité d'utilisation pour celui qui dispose d'un ordinateur. Dans le cas où l'on veut éprouver des hypothèses emboîtées, seules les valeurs de départ $t^{(0)}_{ij}$ sont à changer d'un test à l'autre; au moyen de l'ordinateur 7044, de l'Institut de Calcul Numérique de la Faculté des Sciences de Toulouse, nous l'avons utilisé à titre d'essai sur l'exemple donné au Chapitre VI (bien que là d'autres méthodes soient certainement plus simples). La sortie des résultats est pratiquement instantanée.

CHAPITRE V

ANALYSE STATISTIQUE DES TABLEAUX DE CORRÉLATION CARRÉS

Dans cette partie, nous portons plus spécialement notre attention sur la structure de certains tableaux de corrélation carrés. Nous retrouvons ici l'utilisation des méthodes précédentes, mais nous les complétons par de nouvelles méthodes plus spécifiques. Cependant les techniques nouvelles envisagées par la suite procèdent des mêmes idées générales que celles des pages précédentes, et leur étude s'appuie en de nombreux points, parfois au prix de quelques artifices, sur des résultats antérieurs.

I. Préliminaires.

1. Généralités.

Les notations étant les mêmes que précédemment, on appellera tableau carré un tableau pour lequel $l = c$. Un tel tableau représente souvent les fréquences observées de caractères $\alpha_i \beta_j$ tels que la correspondance entre α_i et β_i est très étroite : il peut s'agir par exemple du même caractère observé sur des sujets appariés (père et fils, jumeaux,...) ou observé sur un même sujet à des époques différentes, etc... Alors, les configurations $\alpha_i \beta_i$ jouent en général un rôle tout à fait particulier, ou tout au moins, *peuvent le jouer* (1).

C'est en pensant essentiellement aux applications à des tableaux de ce type que nous avons élaboré l'étude qui suit, notre but étant d'introduire des *modèles* simples permettant de décrire la *structure* d'une classe assez vaste de tels tableaux. Dans un autre genre d'idée, GOODMAN et KRUSKAL (1959, p. 125) donnent quelques indications sur l'utilisation, dans ce cas, de divers indices d'association.

Remarque. — L'on peut même observer les manifestations d'un caractère donné sur deux sujets qu'il est impossible de classer de façon naturelle l'un par rapport à l'autre (par exemple : jumeaux placés exactement dans les mêmes conditions). Ceci a déjà été envisagé sous le nom de « tables de contingence intra-classes » par OKAMOTO et ISHII (1961) et surtout ISHII (1960). Nous ne nous occuperons pas spécialement de ce modèle, mais nous montrerons incidemment comment l'on peut traiter une nouvelle hypothèse le concernant.

Tout au long de ce chapitre, nous désignerons par R l'ensemble des couples (i, j) ou $i, j = 1, 2, \ldots, l$ et $i \neq j$. On aura donc :

$$\bar{R} = R^* = \{1, 2, \ldots, l\}.$$

Pour simplifier les notations, on supposera $\tilde{R} = C$.

1. Dans certains domaines d'applications, les fréquences devant figurer dans la diagonale principale sont même ignorées; l'on se trouve alors en présence d'une table tronquée : nous en verrons des exemples.

Dans le sens de ce qui est dit plus haut, une première hypothèse que l'on peut envisager, justement parce qu'elle ne concerne pas la distribution des configurations $\alpha_i \beta_i$ est l'hypothèse H_R : nous l'étudierons plus loin, mais l'on peut noter tout de suite qu'elle est trop restrictive pour constituer un modèle assez général. L'hypothèse de quasi-symétrie que nous introduirons contiendra H_R comme cas particulier.

2. Symétrie et identité des distributions marginales.

a) *Symétrie.*

L'hypothèse de symétrie, que nous noterons S est l'hypothèse :

$$p_{ij} = p_{ji} \text{ pour tout } (i, j) \in R$$

Il est immédiat qu'elle peut s'écrire aussi :

$$\forall\ (i, j) \in C, \quad p_{ij} = p_{ji}$$
$$\text{ou bien } \forall\ (i, j) \in R, \quad P_{ij} = P_{ji}$$

L'introduction de cette hypothèse semble due à Bowker (1948). Son étude est aisée :

On peut estimer p_{ij} par $\hat{p}_{ij} = \hat{p}_{ji} = \dfrac{x_{ij} + x_{ji}}{2N}$ pour tout $(i, j) \in C$

P_{ij} par $\hat{P}_{ij} = \hat{P}_{ji} = \dfrac{x_{ij} + x_{ji}}{2n}$ pour tout $(i, j) \in R$.

(avec les notations : $N = \sum_{(i,j) \in C} x_{ij}$; $n = \sum_{(i,j) \in R} x_{ij}$)

Que les fréquences x_{ii} soient connues ou non, on peut utiliser pour éprouver cette hypothèse la statistique :

$$\chi^2_S = \sum_{i<j}\sum \frac{(x_{ij} - x_{ji})^2}{x_{ij} + x_{ji}}$$

qui, sous l'hypothèse S, est distribuée comme χ^2 avec $\dfrac{l(l-1)}{2}$ degrés de liberté, sous la condition habituelle que les fréquences théoriques soient assez grandes.

b) *Identité des distributions marginales.*

Le point à la place d'un indice a la signification habituelle :

$$p_{i.} = \sum_{j=1}^{l} p_{ij}, \text{ etc...}$$

L'hypothèse de l'identité des distributions marginales, que nous désignerons par I M est l'hypothèse :

$$p_{i.} = p_{.i} \text{ pour tout } i \in \bar{R}$$

L'on a entre S et I M la relation S=>I M, mais la réciproque est fausse, sauf pour $l = 2$.

L'hypothèse I M a été étudiée par STUART (1955) puis BHAPKAR (1966). Si l'on dispose d'un échantillon de taille fixée N, BHAPKAR propose, pour éprouver cette hypothèse, de calculer la statistique :

$$\chi^2_{IM} = d' \, W^{-1} \, d$$

avec $d_i = x_{i.} - x_{.i}$ et $d = [d_1, d_2, \ldots, d_{l-1}]'$

W étant une matrice carrée d'ordre $l - 1$ telle que :

$$W = [\delta_{kk'} (x_{k.} + x_{.k}) - x_{kk'} - x_{k'k} - N^{-1} d_k d_{k'}]$$

Si l'hypothèse I M est vraie, χ^2_{IM} suit asymptotiquement une loi de χ^2 à $l - 1$ degrés de liberté, d'où le test (il s'agit là du test de WALD, qui est reconnu identique, sous certaines conditions au sujet de l'estimation des paramètres inconnus, au test χ^2_1 de NEYMAN).

La statistique suggérée par STUART ne diffère de la précédente que par la suppression dans l'expression de W du terme $N^{-1} d_{kk'}$, qui est en général très petit.

Remarque. — Pour le cas où l'on ne connaît pas les fréquences diagonales x_{ii} (table tronquée), on montre facilement que l'on doit utiliser le même test que plus haut en remplaçant simplement x_{ii} par zéro. D'ailleurs, les fréquences diagonales n'interviennent explicitement dans χ^2_{IM}, que par N figurant au dernier terme des éléments de W. Si l'on connaît ces fréquences diagonales mais si on les ignore volontairement, la statistique obtenue diffère de celle de BHAPKAR seulement par le changement de N en n dans le dernier terme de W : la différence est, dans la pratique, tout à fait négligeable (quant à la statistique de STUART, elle reste inchangée).

c) *Propriété fondamentale de l'hypothèse* I M.

La définition de deux *hypothèses séparables* par rapport à un certain type de test est donnée par AITCHISON (1962). Les tests utilisés seront le test classique du χ^2 ou les tests équivalents (test de WALD, etc... cf : AITCHISON, 1962) ; nous parlerons donc simplement « d'hypothèses séparables » en sous-entendant toujours que c'est par rapport à ces tests.

Si H est une hypothèse quelconque portant uniquement sur la « structure interne » d'un tableau de corrélation, il semble intuitivement que les hypothèses I M et H doivent être séparables. (En fait, des résultats de ce type sont classiques dans certains cas particuliers.)

Considérons les « interactions du premier ordre » $\delta(u, p)$. On aura le :

Lemme V. — Si H_δ est une hypothèse équivalente à :

$$h^{(m)} = \delta(u^{(m)}, p) - C^{(m)} = 0 \quad (\text{pour } m = 1, 2, \ldots, \mu)$$

où, pour tout m, les $C^{(m)}$ sont des constantes, et $u^{(m)}$ appartient à I (C), alors H_δ et I M sont séparables.

La démonstration est une application immédiate du critère donné par AITCHISON (1962). Si B est la matrice d'information, nous avons ici (δ_{ij}

étant le symbole de KRONECKER) $B^{-1} = \| \delta_{ih}\,\delta_{jk} - p_{ij}\,p_{hk} \|$ l'indice de la ligne est l'indice double $(i\,j)$, de même $(h\,k)$ est l'indice de la colonne; $(i\,j)$ prend $l^2 - 1$ valeurs; les v. a. x_{ij}, $(i, j) \in C$, étant liées par une relation certaine, nous en éliminons une, par exemple x_{ll}; on notera $C' = C - \{(l, l)\}$. De même l'on conservera comme paramètres indépendants les p_{ij} tels que (i, j) appartient à C'.

Notons H la matrice à $l^2 - 1$ lignes et μ colonnes dont les éléments sont :

$$\frac{\partial h^{(m)}}{\partial p_{ij}} = \frac{u_{ij}^{(m)}}{p_{ij}} - \frac{u_{ll}^{(m)}}{p_{ll}} \quad (m=1,2,\ldots,\mu\,;\ (i.j) \in C').$$

L'élément de la ligne d'indice $(i\,j)$ et de la colonne d'indice m de la matrice $B^{-1}H$ est :

$$\sum_{(h,\,k)\in C'} (\delta_{ih}\,\delta_{jk}\,p_{hk} - p_{ij}\,p_{hk}) \left(\frac{u_{hk}^{(m)}}{p_{hk}} - \frac{u_{ll}^{(m)}}{p_{ll}}\right)$$

$$= u_{ij}^{(m)} - p_{ij} \sum_{(h,\,k)\,\in\,C} u_{hk}^{(m)}$$

puisque $u^{(m)}$ appartient à I (C), cette expression est simplement $u^{(m)}{}_{ij}$ pour $(i, j) \in C'$, $m = 1, 2, \ldots, \mu$.

L'hypothèse I M fixe entre les paramètres les liaisons :

$$k^{(c)} = p_{c.} - p_{.c} = 0 \qquad c = 1, 2, \ldots, l-1$$

Notons K la matrice à $l^2 - 1$ lignes et $l - 1$ colonnes dont les éléments sont :

$$\frac{\partial\, k^{(c)}}{\partial\, p_{ij}} = \begin{cases} 1 \text{ si } i = c \text{ et } i \neq j \\ -1 \text{ si } i = c \text{ et } i \neq j \\ 0 \text{ dans tous les autres cas.} \end{cases}$$

La matrice $K' B^{-1} H$ a $(l - 1)$ lignes et μ colonnes, l'élément de la i^{me} ligne et m^{me} colonne étant :

$$\sum_{(i,\,j)\in C'} u_{ij}^{(m)} \frac{\partial k^{(c)}}{\partial p_{ij}} = \sum_{j\neq c} u_{cj}^{(m)} - \sum_{i\neq c} u_{ic}^{(m)} = 0$$

ce qui démontre la proposition annoncée.

Remarque. — La démonstration suppose p_{ij} différent de zéro pour (i, j) appartenant à C; c'est la condition pour que $\delta(u, p)$ soit définie. Mais on peut l'étendre à des cas où p_{ij} est différent de zéro seulement pour (i, j) appartenant à R ($R \subset C$) à condition de ne considérer que des hypothèses H_δ définies par $\delta(u^{(m)}, p) = C^{(m)}$ où $u^{(m)}$ appartient à I (R).

II. L'hypothèse de quasi-symétrie.

1. Définition.

La quasi-symétrie (Q S) sera définie par :

$$\text{Q S} \Longleftrightarrow \begin{cases} \forall i \in \bar{\text{R}}\ \exists\ p_i > 0, \quad \forall j \in \bar{\text{R}}\ \exists\ q_j > 0, \quad \forall (i,j) \in \text{R}\ \ \exists\ d_{ij} > 0, \quad d_{ij} = d_{ji} \\ \text{tels que :} \quad \forall\ (i,\ j) \in \text{R} \quad p_{ij} = p_i\, q_j\, d_{ij} \end{cases}$$

Il est équivalent, aux notations près, de remplacer la condition $p_{ij} = p_i\, q_j\, d_{ij}$ par $\text{P}_{ij} = p_i\, q_j\, d_{ij}$ car P_{ij}/p_{ij} est constant pour tout (i, j) appartenant à R.

Les paramètres intervenant dans cette définition ne sont pas déterminés de façon unique. L'on peut d'abord remplacer p_i par kp_i et q_j par $k^{-1}\, q_j$; en outre, si l'on pose $d_{ij} = k\, p''_i\, p''_j\, d'_{ij}$, on aura $d'_{ij} = d'_{ji}$ pour tout $(i, j) \in \text{R}$ et l'on peut écrire : $\forall (i, j) \in \text{R}\ \ p_{ij} = kp_i\, p''_i\, q_j\, p''_j\, d'_{ij} = p'_i\, q'_j\, d'_{ij}$,
Réciproquement, supposons :
$\forall\ (i, j) \in \text{R}\ \ p_{ij} = p_i\, q_j\, d_{ij} = p'_i\, q'_j\, d'_{ij}$ et $d_{ij} = d_{ji}, \quad d'_{ij} = d'_{ji}$
il vient $d'_{ij} = \dfrac{p_i\, q_j}{p'_i\, q'_j}\, d_{ij}$ qui est de la forme

$$d'_{ij} = p''_i\, q''_j\, d_{ij} \text{ avec } p''_i = \frac{p_i}{p'_i} \text{ et } q''_j = \frac{q_j}{q'_j}$$

En écrivant $d_{ij} = d_{ji}$ et $d'_{ij} = d'_{ji}$, on obtient pour tout $(i, j) \in \text{R}$

$$p''_i\, q''_j = p''_j\, q''_i \text{ soit } \frac{p''_i}{q''_i} = \frac{p''_j}{q''_j}$$

Il existe donc k tel que $q''_i = k\, p''_i$ pour tout $i \in \bar{\text{R}}$, d'où nécessairement $d'_{ij} = k\, p''_i\, p''_j\, d_{ij}$.

Remarque. — Si $p_{ij} = p_i\, q_j\, d_{ij} = p'_i\, q'_j\, d'_{ij}$ et $d_{ij} = d_{ji}$, $d'_{ij} = d'_{ji}$, en reprenant l'étude ci-dessus, l'on voit que pour tout $i \in \bar{\text{R}}$ $\dfrac{q_i}{q'_i} = k\, \dfrac{p_i}{p'_i}$.

Si maintenant, partant de $p_{ij} = p_i\, q_j\, d_{ij}$, nous posons $q_i\, q_i\, d_{ij} = c_{ij}$, il vient :

$p_{ij} = \dfrac{p_i}{q_i}\, c_{ij} = \pi_i\, c_{ij}$, avec la condition $c_{ij} = c_{ji}$ pour tout $(i, j) \in \text{R}$. D'où :

Deuxième définition :

La quasi-symétrie peut être définie de façon plus simple (mais sans doute moins naturelle) par :

$$\text{QS} \Longleftrightarrow \begin{cases} \forall i \in \bar{\text{R}}\ \exists\ \pi_i > 0, \quad \forall\ (i, j) \in \text{R} \quad \exists\ c_{ij} > 0, \quad c_{ij} = c_{ji} \\ \text{tels que} \quad \forall (i,\ j) \in \text{R} \quad p_{ij} = \pi_i\, c_{ij} \end{cases}$$

On peut donner des définitions analogues en permutant les rôles des deux indices, ou en remplaçant p_{ij} par les probabilités conditionnelles P_{ij}.

Avec cette deuxième définition les paramètres π_i et c_{ij} sont définis uniquement, à une proportionnalité près. En effet, utilisant la dernière

« Remarque » ci-dessus, si $p_{ij} = \pi_i c_{ij} = \pi'_i c'_{ij}$, l'on aura pour tout $i \in \bar{R}$: $\frac{\pi'_j}{\pi_j} = k$.

Quelques propriétés de la quasi-symétrie.

On a toujours $H_R \Rightarrow QS$.

Ceci est évident en utilisant la première définition de l'hypothèse QS. La réciproque est en général fausse, avec deux exceptions :

Pour $l = 2$: H_R et QS sont équivalentes de façon triviale étant toutes deux vraies quelles que soient les probabilités p_{ij}.

Pour $l = 3$: Si QS est vraie on peut écrire $p_{ij} = \pi_i c_{ij}$ pour $i, j = 1, 2, 3$ et $i \neq j$ ce qui implique

$$\frac{p_{12}\ p_{23}\ p_{31}}{p_{21}\ p_{32}\ p_{13}} = 1$$

Or on a déjà vu que cette condition était équivalente à H_R donc :

$$QS \Rightarrow H_R.$$

Ainsi, dans une table 3×3, on a bien :

$$QS \iff H_R.$$

Notons R′ la partie de C définie par :

$$(i, j) \in R' \iff (i, j) \in R \quad \text{et} \quad i < j.$$

A chaque élément de R′, on associe $g_{ij} = \frac{p_{ij}}{p_{ji}}$

Si QS est vraie l'on a $g_{ij} = \pi_i/\pi_j$

Donc, avec les notations du Chapitre I, G étant ici le groupe multiplicatif des nombres réels positifs, on aura : $QS \Rightarrow H_{R'}^{-}$.

Réciproquement, on peut toujours écrire pour tout $(i, j) \in R$, $p_{ij} = \pi_i c_{ij}$; supposons $H_{R'}^{-}$ vraie, on aura $\frac{\pi_i c_{ij}}{\pi_j c_{ji}} = \frac{\pi_i}{\pi_j}$ d'où $c_{ij} = c_{ji}$ et donc $H_{R'}^{-} \Rightarrow QS$

$$\text{Finalement : } QS \iff H_{R'}^{-}$$

Remarque. — On peut déduire de ceci le nombre ν_{QS} de relations indépendantes imposées aux paramètres p_{ij} par l'hypothèse QS (nombre qui est égal à celui des relations imposées par $H_{R'}^{-}$) par application de résultats du Chapitre I (A-VI). On a ici :

$$\text{Card } (R') = \frac{l\,(l-1)}{2}$$

$$\text{Card } (\bar{R}') = \text{Card } (R'^{*}) = l - 1$$

$$\text{Card } (\bar{R}' \cap R'^{*}) = l - 2 \text{ (en effet } \bar{R}' = \{1, 2, \ldots, l-1\} \text{ et } R'^{*} = \{2, 3, \ldots, l\})$$

D'où :

$$\nu_{QS} = \frac{(l-1)\,(l-2)}{2}$$

2. Étude comparée des hypothèses QS, IM et S.

Quasi-symétrie et interactions du premier ordre :

Si QS est vérifiée, l'on a :

$$\forall\,(i,j)\in \mathrm{R}\quad \frac{p_{ij}}{p_{ji}} = \frac{\pi_i}{\pi_j}$$

Ecrivant cette relation pour les couples (j, k) puis (k, i), on peut éliminer les π_i pour obtenir :

$$\frac{p_{ij}\ p_{jk}\ p_{ki}}{p_{ji}\ p_{ik}\ p_{kj}} = 1$$

En posant $r_{ijk} = p_{ij}\, p_{jk}\, p_{ki}/p_{ji}\, p_{ik}\, p_{kj}$
QS entraîne

$$\text{pour tout } i, j, k \text{ appartenant à } \bar{\mathrm{R}} : r_{ijk} = 1 \qquad (1)$$

(i, j, k différents deux à deux : cependant les relations écrites sont identiquement vérifiées lorsque, par exemple, $j = i$ si $p_{ii} \neq 0$).

Montrons que les relations (1) entraînent QS :

D'après ce que l'on a vu plus haut sur les tables 3×3, on aura d'abord : $r_{123} = 1 \Rightarrow p_{ij} = \pi_i\ c_{ij},\ c_{ij} = c_{ji}$ pour $i, j = 1, 2, 3$ et $i \neq j$.

Pour $h > 3$, posons :

$$c_{1h} = p_{1h}/\pi_1, \qquad c_{2h} = p_{2h}/\pi_2, \qquad \pi_h = p_{h1}/c_{1h}$$

Alors $r(1, 2, h) = 1 \Rightarrow p_{h2} = \pi_h\, c_{2h}$

On a donc $p_{ij} = \pi_i\, c_{ij},\ c_{ij} = c_{ji}$ pour tout couple (i, j) de R où $i = 1, 2\ \ j = 1, 2, \ldots, l$ ou bien $j = 1,2\ \ i = 1, 2, \ldots, l$. Pour tout couple (h, k) de R différent des précédents, les résultats ci-dessus et la relation $r\,(h,k, 1) = 1$ impliquent

$$p_{hk} = \pi_h\ c_{hk} \quad \text{avec } c_{hk} = c_{kh}$$

ce qui achève la démonstration (2).

Ainsi QS est équivalente à l'ensemble des relations (1).

On notera que ces relations ne sont pas toutes indépendantes; mais l'on peut caractériser un ensemble de relations indépendantes impliquant les autres. On vérifie d'abord que la relation $r_{ijk} = 1$ entraîne les cinq autres relations analogues obtenues par une quelconque permutation des trois indices. Pour montrer QS à partir de (1) on a utilisé seulement les relations :

$$r_{1hk} = 1 \text{ pour } h,\ k = 2,\ 3,\ \ldots,\ l\ \ h \neq k \qquad (2)$$

---

2. On aurait pu éviter cette démonstration en faisant appel aux résultats du paragraphe suivant; nous l'avons donnée cependant parce qu'elle permet d'atteindre facilement un ensemble de ν_{QS} liaisons indépendantes.

Elles sont donc suffisantes pour que QS soit vraie, et l'on peut donc en déduire les autres relations (1).

Mais l'on a vu que QS imposait $\nu_{QS} = \frac{(l-1)(l-2)}{2}$ liaisons indépendantes entre les probabilités p_{ij}, or (2) contient bien seulement $C^2_{l-1} = \frac{(l-1)(l-2)}{2}$ relations, qui sont donc indépendantes.

Bien entendu, l'on pourra en général trouver des systèmes différents de ν_{QS} liaisons indépendantes de ce type.

Si l'on introduit les « interactions du premier ordre », QS est équivalente à l'ensemble des relations (qui ne sont pas toutes indépendantes) :

$$\mathrm{Logr}_{ijk} = \mathrm{Logp}_{ij} + \mathrm{Logp}_{jk} + \mathrm{Logp}_{ki} - \mathrm{Logp}_{ji} - \mathrm{Logp}_{ik} - \mathrm{Logp}_{kj} = 0 \quad (3)$$

pour tout i, j, k différents deux à deux et appartenant à $\overline{R}$.

Les quantités Log r_{ijk} sont bien des interactions du premier ordre : en effet, si l'on écrit :

$$\mathrm{Log}\, r_{ijk} = \sum_{(h,h')\in C} u_{hh'} \mathrm{Log}\, p_{hh'}$$

toutes les sommes sur h ou h' des $u_{hh'}$ contiennent soit seulement des zéros, soit seulement deux termes non nuls dont l'un est + 1, l'autre — 1.

De cela, par application du Lemme V, on déduit :

Théorème XI. — Les hypothèses QS et IM sont séparables.

Nous montrerons maintenant le :

Théorème XII. — Pour tout l, a lieu l'équivalence : S <=> QS ∩ IM.

On a déjà noté S=> IM (la réciproque est fausse pour $l > 2$)

De même l'on a S => QS (la réciproque est fausse pour $l > 1$)

On a donc : S => QS ∩ IM.

Supposons réciproquement QS et IM vraies; en tenant compte de QS l'on a

$$p_{i.} = p_{ii} + \sum_{j\neq i} \pi_i c_{ij} \text{ et } p_{.i} = p_{ii} + \sum_{j\neq i} \pi_j c_{ij}$$

En utilisant IM il vient :

$$\forall i \in \overline{R} \quad \sum_{j\neq i} (\pi_i - \pi_j) c_{ij} = 0$$

On a vu que les paramètres c_{ij} et π_i étaient définis à une proportionnalité près : si l'on fixe par exemple $\pi_1 = \pi$, ils sont définis de façon unique. Les équations ci-dessus sont vérifiées si $\pi_i = \pi$ pour tout i appartenant à $\overline{R}$ et c'est là leur unique solution : en effet, on peut les écrire

$$\pi_i = \frac{\sum_{j\neq i} \pi_j c_{ij}}{\sum_{j\neq i} c_{ij}}$$

si bien que chaque π_i est une moyenne pondérée des

autres (avec des poids c_{ij}, tous strictement positifs) ; la contradiction est alors immédiate si les π_i ne sont pas tous égaux.

Ainsi $QS \cap IM \Rightarrow \forall (i,j) \in R,\ p_{ij} = \pi\, c_{ij},\ c_{ij} = c_{ji}$

Dans ces conditions $p_{ij} = p_{ji}$.

Donc : $QS \cap IM \Rightarrow S$.

3. Quasi-symétrie et tables tronquées.

Considérons l'ensemble des couples ordonnés (h, ij) où h peut prendre les valeurs $1, 2, \ldots, l$ (c'est-à-dire $h \in \overline{R}$), et où l'indice double ij est tel que (i,j) appartient à R′ et peut donc prendre $\frac{l(l-1)}{2}$ valeurs.

On notera T le sous-ensemble défini par :

$T = \{(h, ij) : h \in \overline{R},\ (i,j) \in R',\ (i = h \text{ et } j > h) \text{ ou } (j = h \text{ et } i < h)\}$

Toujours avec les notations du Chapitre I, l'on a :

$T_{h.} = \{ij : (i,j) \in R',\ (i = h \text{ et } j > h) \text{ ou } (j = h \text{ et } i < h)\}$

$T_{.ij} = \{h : h = i \text{ ou } h = j\}$

Ainsi, pour ij donné et $(h, ij) \in T$, h peut prendre deux valeurs : i ou j. On a donc Card $(T) = l(l-1)$.

On établit une correspondance biunivoque entre les p_{ij}, $(i,j) \in R$, et les éléments de T en associant à l'élément (h, ij) de T :

$$\begin{cases} p_{ij} & \text{si } h = i \\ p_{ji} & \text{si } h = j \end{cases}$$

Si les p_{ij}, vérifiant l'hypothèse de quasi-symétrie, sont tels que : $\forall (i,j) \in R\ p_{ij} = \pi_i\, c_{ij},\ c_{ij} = c_{ji}$,
on a ainsi associé à $(h, ij) \in T$:

$$\begin{cases} p_{ij} = \pi_i\, c_{ij} & \text{si } h = i \\ p_{ji} = \pi_j\, c_{ji} = \pi_j\, c_{ij} & \text{si } h = j \end{cases}$$

Ainsi dans tous les cas on a associé à (h, ij) la quantité $\pi_h\, c_{ij}$ si bien que H_T est vérifiée. Donc $QS \Rightarrow H_T$.

La réciproque est tout aussi immédiate; finalement, l'on a :

$$QS \Longleftrightarrow H_T.$$

Le nouvel arrangement des p_{ij} est explicité pour $l = 4$ dans le tableau ci-dessous.

$h \backslash ij$	12	13	14	23	24	34
1	$p_{12} = \pi_1\, c_{12}$	$p_{13} = \pi_1\, c_{13}$	$p_{14} = \pi_1\, c_{14}$			
2	$p_{21} = \pi_2\, c_{12}$			$p_{23} = \pi_2\, c_{23}$	$p_{24} = \pi_2\, c_{24}$	
3		$p_{31} = \pi_3\, c_{13}$		$p_{32} = \pi_3\, c_{23}$		$p_{34} = \pi_3\, c_{34}$
4			$p_{41} = \pi_4\, c_{14}$		$p_{42} = \pi_4\, c_{24}$	$p_{43} = \pi_4\, c_{34}$

Remarque importante. — Toutes les propriétés de l'hypothèse QS établies précédemment, sont valables en particulier pour $p_{ii}=0$. Elles sont donc utilisables lorsque la population considérée est privée de ses « éléments diagonaux ».

4. Estimation des paramètres du modèle sous l'hypothèse de quasi-symétrie.

Un échantillon de taille N est extrait au hasard de la population parente : soit x_{ij} la fréquence observée de la catégorie (i, j). Sous l'hypothèse Q S la vraisemblance est :

$$L = \frac{N!}{\prod_{(i,j)\in C} (x_{ij})!} \prod_{(i,j)\in R} (\pi_i c_{ij})^{x_{ij}} \prod_{i\in\bar{R}} p_{ii}^{x_{ii}}$$

Les paramètres inconnus sont $p_{ii}\ (i\in\bar{R})$, $\pi_i\ (i\in\bar{R})$, $c_{ij}\ ((i,j)\in R',\ c_{ji}=c_{ij})$; ils sont identifiables si l'on fixe par exemple la valeur de π_1.

En notant $a_i = \sum_{j\neq i} x_{ij}$

$$s_{ij} = x_{ij} + x_{ji}$$

l'on a :

$$\text{Log } L = C^{te} + \sum_{i\in\bar{R}} a_i \text{ Log } \pi_i + \sum_{(i,j)\in R'} s_{ij} \text{ Log } c_{ij} + \sum_{i\in\bar{R}} x_{ii} \text{ Log } p_{ii}$$

Les statistiques $x_{ii}\ (i\in\bar{R})$, $a_i\ (i\in\bar{R})$, $s_{ij}\ [(i,j)\in R']$ sont donc exhaustives (il y a $\frac{(l+4)(l-1)}{2}$ paramètres indépendants et la statistique exhaustive minimale a cette même dimension).

a) *Méthode du maximum de vraisemblance.*

La méthode du maximum de vraisemblance conduit aux équations :

$$\begin{cases} N\hat{\pi}_i \sum_{j\neq i} \hat{c}_{ij} = a_i & \text{pour tout } i\in\bar{R} \\ N(\hat{\pi}_i + \hat{\pi}_j)\hat{c}_{ij} = s_{ij} & \text{pour tout } (i,j)\in R' \\ N\hat{p}_{ii} = x_{ii} \end{cases}$$

avec $\hat{c}_{ij} = \hat{c}_{ji}$

Soit f_{ij} les fréquences théoriques sous l'hypothèse Q S

$$f_{ii} = N \hat{p}_{ii} \qquad \text{pour tout } i \in \bar{R}$$
$$f_{ij} = N \hat{\pi}_i \hat{c}_{ij} \qquad \text{pour tout } (i, j) \in R$$

On aura $f_{ii} = x_{ii}$ pour tout i; reste à résoudre le système en $\hat{\pi}_i$ et $\hat{c}_{ij}$

$$\left.\begin{array}{ll} \forall\, i \in \bar{R} & N \hat{\pi}_i \sum_{j \neq i} \hat{c}_{ij} = a_i \\ \forall\, (i, j) \in R' & N (\hat{\pi}_i + \hat{\pi}_j)\, \hat{c}_{ij} = s_{ij} \\ & \hat{c}_{ij} = \hat{c}_{ji} \end{array}\right\} \prec\!=\!\succ \left\{\begin{array}{l} \sum_{j \neq i} f_{ij} = a_i \\ f_{ij} + f_{ji} = s_{ij} \end{array}\right. \tag{4}$$

(on peut remarquer que ces équations entraînent $\sum_{i \neq j} f_{ij} = b_j$ avec $b_j = \sum_{i \neq j} x_{ij}$)

Ce sont ces équations (4) que l'on obtiendra aussi (à l'exclusion de la relation $f_{ii} = x_{ii}$) si l'on est en présence d'un échantillon tiré de la population réduite aux sujets ayant les caractères $\alpha_i\, \beta_j$ où $i \neq j$.

Dans les deux cas leur résolution est équivalente à la construction d'un tableau de corrélation carré f_{ij} *tel que*
$f_{ij} = 0$ pour $i = j$
f_{ij} est le produit d'une fonction de i seul par une fonction *symétrique* de i et de j, pour $i \neq j$.
les marges (a_i et b_j) sont fixées
les sommes $f_{ij} + f_{ji} = s_{ij}$ sont fixées.

Il ne semble pas que la solution des équations (4) puisse être explicitée (en dehors du cas trivial $l = 2$).

Considérons la suite matricielle $\| f^{(m)}{}_{ij} \|$ ($m = 1, 2, 3, \ldots$) définie par :

$$\left\{\begin{array}{ll} f_{ij}^{(2m+1)} = f_{ij}^{(2m)} \dfrac{a_i}{\sum_{h \in \bar{R}} f_{ih}^{(2m)}} & \text{pour tout } (i, j) \in R \\ f_{ij}^{(2m+2)} = f_{ij}^{(2m+1)} \dfrac{s_{ij}}{f_{ij}^{(2m+1)} + f_{ji}^{(2m+1)}} & \text{pour tout } (i, j) \in R \end{array}\right.$$

avec $f^{(0)}{}_{ij} = \| 1 - \delta_{ij} \|$

et $f^{(m)}{}_{ii} = 0$ pour tout $i \in \bar{R}$ et pour tout $m = 1, 2, 3, \ldots$

Si cette suite converge, sa limite $\| f_{ij} \|$ est bien un tableau possédant toutes les propriétés cherchées.

Ayant obtenu ces quantités $f_{ij} = N \hat{\pi}_i \hat{c}_{ij}$, il sera facile, si on le désire, d'en déduire les $\hat{\pi}_i$ et $\hat{c}_{ij}$ après avoir introduit une condition qui assure leur identifiabilité.

En fait, si, de façon analogue au paragraphe précédent, les x_{ij} $(i \neq j)$ sont réarrangés dans la table tronquée associée à T, les équations (4) sont les équations de vraisemblance obtenues, à partir de cette nouvelle table, sous l'hypothèse H_T; l'algorithme ci-dessus est l'algorithme R.A.S. envisagé au Chapitre II. On pourra se contenter d'invoquer les résultats de ce chapitre, pour affirmer *l'existence* et *l'unicité* d'une solution des équations (4), L atteignant en ce point son *maximum*.

Remarque. — Revenons à l'algorithme ci-dessus; c'est un cas particulier de l'algorithme R.A.S. pour lequel la convergence n'a pas été prouvée; cependant, dans la pratique, il semble converger et même donner assez rapidement une bonne approximation des fréquences théoriques. Toutefois, sur les exemples numériques considérées, il nous a paru que l'on pouvait toujours obtenir plus rapidement la même approximation, avec le processus à trois stades suivant :

$\|f_{ij}^{(0)}\| = \| 1 - \gamma_{ij} \|$, $f_{ii}^{(m)} = 0$ pour tout $i \in \bar{R}$ et tout m et, pour tout $(i, j) \in R$:

$$\begin{cases} f_{ij}^{(3m+1)} = f_{ij}^{(3m)} \dfrac{a_i}{\sum\limits_{h \in \bar{R}} f_{ih}^{(3m)}} \\ f_{ij}^{(3m+2)} = f_{ij}^{(3m+1)} \dfrac{b_j}{\sum\limits_{h \in \bar{R}} f_{hj}^{(3m+1)}} \\ f_{ij}^{(3m+3)} = f_{ij}^{(3m+2)} \dfrac{s_{ij}}{f_{ij}^{(3m+2)} + f_{ji}^{(3m+2)}} \end{cases}$$

Il est facile de voir que si la suite $\|f_{ij}^{(m)}\|$ converge, sa limite est nécessairement le tableau $\|f_{ij}\|$ cherché (cf. : passage de la première à la deuxième définition de QS).

b) *Autres estimateurs.*

Nous venons de voir que les équations de vraisemblance étaient analogues aux équations obtenues pour une certaine table tronquée; donc, les considérations du Chapitre II sur la résolution de ces équations peuvent être utilisées ici, afin d'obtenir des estimateurs R.B.A.N., intéressants surtout pour des échantillons importants.

En particulier, la méthode du paragraphe (V-4) s'appliquera en écrivant les liaisons entre les p_{ij} sous la forme (3), ou plutôt, en choisissant parmi les relations (3), $\nu_{QS} = \dfrac{(l-1)(l-2)}{2}$ relations indépendantes, par exemple :

$$\text{Log } r_{ihk} = 0 \qquad h, k = 2, 3, \ldots, l \qquad h \neq k.$$

Comme au Chapitre II, on vérifiera que les estimateurs sont invariants si le choix des liaisons est changé.

L'inconvénient de la métohde est toujours le même : on obtient f^*_{ij} qui ne peut se mettre en général sous la forme $\pi^*_i c^*_{ij}$ $(c^*_{ij} = c^*_{ji})$. Les calculs

sont simples, si ce n'est l'inversion d'une matrice d'ordre ν_{QS}; pour $l = 3$ $\nu_{QS} = 1$, pour $l = 4$ $\nu_{QS} = 3$. Pour $l \geqslant 5$, on aura $\nu_{QS} \geqslant l$; or nous avons indiqué une autre méthode qui nécessite seulement l'inversion d'une matrice d'ordre $(l - 1)$: elle pourra être alors préférée, d'autant plus qu'elle donne directement des estimateurs π_i et c_{ij} en même temps que $f_{ij} = \pi_i \pi_j c_{ij}$.

5. Test de l'hypothèse de quasi-symétrie.

Pour éprouver l'hypothèse QS, il est possible d'utiliser le χ^2 ou les divers test équivalents avec des fréquences théoriques obtenues au paragraphe précédent. Ces diverses statistiques seront notées indifféremment χ^2_{QS}; on leur associe la région critique de seuil α :

$$\chi^2_{QS} > l(\alpha, \nu_{QS})$$

La discussion de cet emploi est tout à fait analogue à celle que nous avons faite pour les tests de H_R, puisque QS est équivalente à une hypothèse H_T. Comme pour éprouver H_R, on pourra faire appel aussi à des intervalles de confiance simultanés pour les interactions Log r_{ijk}, c'est-à-dire pour les interactions $\delta(u, p)$, telles que $u \in I(T)$, de la table de contingence à l lignes et $\frac{l(l-1)}{2}$ colonnes associée à $\bar{T}$.

Il est facile en outre d'introduire ici quelques nouvelles possibilités. Nous avons vu : $QS \cap IM \prec\!\!\succ S$. D'autre part, par le théorème XI, QS et IM sont séparables, et si l'on pose $\chi^2_{S/IM} = \chi^2_S - \chi^2_{IM}$ l'on aura :
$\lim pr\ \chi^2_{S|IM} = \lim pr\ \chi^2_{QS}$
(on désigne par *lim pr* la limite en probabilité lorsque la taille de l'échantillon tend vers l'infini).

On pourra donc utiliser, comme étant équivalente aux régions critiques précédentes, la nouvelle région critique de seuil α :

$$\chi^2_{S/IM} > l(\alpha, \nu_{QS})$$

La statistique χ^2_S est simple à calculer, ce dernier test de QS semble donc tout à fait recommandable si l'on désire aussi éprouver l'hypothèse IM, et peut parfois être préféré, même si cette dernière épreuve est de peu d'intérêt.

Inversement, l'on notera la possibilité d'adopter comme région critique de seuil α, pour éprouver IM :

$$\chi^2_{S/QS} = \chi^2_S - \chi^2_{QS} > l(\alpha, \nu_{IM}).$$

D'un autre point de vue, l'on obtient aussi la décomposition de χ^2_S en deux statistiques asymptotiquement indépendantes, permettant les tests respectifs des deux « composantes » IM et QS de l'hypothèse de symétrie S.

6. Remarques sur la signification pratique de l'hypothèse QS et son utilité.

Afin de chercher dans quels types de tableaux l'hypothèse QS peut constituer un modèle satisfaisant, nous essayerons d'abord de donner quelques interprétations pratiques de ce modèle mathématique.

Dire que l'hypothèse QS est vraie pour une table de contingence donnée, c'est dire que les interactions du premier ordre liées à cette table sont égales aux interactions correspondantes d'un certain tableau symétrique (ceci est immédiat à partir de la première définition de QS). En quelque sorte un tableau vérifie QS « s'il est symétrique, ou dissymétrique *seulement* par l'inégalité des distributions marginales » [3] (cf. : S <=> QS ∩ IM).

Supposons qu'un tableau de corrélation carré représente, pour un échantillon donné, les fréquences x_{ij} de sujets ayant un caractère α_i à une première époque déterminée, et le caractère $\beta_j = \alpha_j$ à une deuxième époque (nous noterons plutôt dans ce cas $\alpha_i = \alpha^1{}_i$ et $\beta_i = \alpha^2{}_i$, et le caractère α_i sera aussi appelé un « état »); alternativement, il peut s'agir de couples (ordonnés) d'individus dont le premier possède le caractère $\alpha^1{}_i$, le second le caractère $\alpha^2{}_j$, ou de quelque autre schéma d'association analogue. On peut dire que l'on est en présence d'un processus « migratoire », ce dernier terme devant être pris dans le sens le plus large. Aux lignes du tableau est associée la distribution marginale des sujets dans un premier stade (ou la distribution des sujets n° 1 de chaque couple ordonné), aux colonnes correspond la distribution marginale des sujets dans le second stade (respectivement la distribution des sujets n° 2). La fréquence x_{ij} est la fréquence des « transitions » de type $\alpha^1{}_i\ \alpha^2{}_j$. En terme de probabilité conditionnelle sachant qu'il y a effectivement modification du caractère entre le stade initial et le stade final, l'hypothèse QS s'écrit : $\forall\ (i, j) \in R \qquad P_{ij} = p_i\ q_j\ d_{ij}\ (d_{ij} = d_{ji})$
et peut être interprétée ainsi :

p_i représente la « tendance » d'un sujet à quitter l'état $\alpha^1{}_i$, q_j représente la « tendance » à adopter le nouvel état $\alpha^2{}_j$; d_{ij} apparaît comme un paramètre « d'interaction », paramètre qui caractérise l'intensité des « échanges » entre l'état α_i et l'état α_j, *compte non tenu de leur sens.* Il peut être intéressant de relier cette notion à une notion de « distance » entre ces caractères, cette distance pourrait être mesurée, par exemple, par d_{ij}^{-1} puisqu'il semble naturel de définir deux caractères comme étant d'autant plus « éloignés » qu'il y a peu de sujets qui abandonnent l'un pour acquérir l'autre. Mais, une difficulté majeure naît du fait que ces divers paramètres ne sont pas identifiables. Toutefois, le paramètre $\pi_i = \dfrac{p_i}{q_i}$ est invariant à une proportionnalité près comme on l'a vu au début du paragraphe et l'on peut reparamétrer le modèle sous la forme $P_{ij} = \pi_i\ c_{ij}$. On peut considérer que π_i indique comment a tendance à varier entre le premier et le deuxième stade l'importance relative des effectifs qui possèdent le caractère α_i. Pour interpréter c_{ij}, on remarquera :

$$Pr\ [\alpha^2{}_j\ /\ \alpha^1{}_i\ ,\ \alpha^1{}_i \neq \alpha^2{}_j] = \frac{\pi_i\ c_{ij}}{\pi_i \sum_{h \neq i} c_{ih}}$$

3. D'où la dénomination de « quasi-symétrie » que nous avons adoptée.

donc, la probabilité conditionnelle de α^2_j, α^1_i étant fixé ($i \neq j$), est proportionnelle à c_{ij}.

Pour terminer nous citons quelques faits qui semblent justifier l'introduction et l'étude de cette structure de quasi-sysmétrie :

1° Nous avons trouvé des exemples de natures variées pour lesquels l'ajustement à ce modèle s'est avéré très étroit (voir Chapitre VI). L'étude que nous avons faite peut donc aider à l'interprétation de ces tableaux.

2° Pour l'étude de processus de migration interne (dans le sens démographique le plus restreint) certains spécialistes ont suggéré, pour caractériser l'amplitude des flots migratoires, des modèles mathématiques voisins (voir par exemple : GIL et OMABOE 1963; G. CALOT, discussion sur une communication de R. PRESSAT, 1963). Les modèles avancés par ces auteurs sont plus riches que la quasi-symétrie, en ce sens qu'ils impliquent QS, la réciproque étant fausse : le premier ne laisse que deux paramètres inspécifiés, le second 2 *l*. Mais, pour que de tels modèles soient raisonnables, il faut d'abord que la quasi-symétrie le soit (nous préciserons mieux plus loin, ce que nous entendons par modèle raisonnable : cf. paragraphe IV). Son étude paraît donc utile en permettant de façon simple un premier test du modèle plus strict adopté, et en permettant aussi l'estimation de paramètres à partir des données actuelles, estimation qu'il peut être intéressant de comparer aux appréciations empiriques que l'on a effectuées.

Nous espérons ainsi avoir donné un début de réponse à une question soulevée par M. Calot dans la discussion de l'article de PRESSAT (1963). Quant au modèle précis qui est suggéré là, il peut s'écrire :
$\forall\ (i, j) \in R \quad P_{ij} = p_i\, q_j\, D_{ij}^{-\beta}$ où les D_{ij} sont donnés. Il n'est pas dans notre intention de l'étudier ici en détail, mais nous pensons qu'à partir de notre travail il sera possible de l'aborder de plus près; en fait, il est « intermédiaire » entre la quasi-symétrie et un modèle du type étudié au Chapitre IV : $P_{ij} = p_i\, q_j\, D'_{ij}$ où D'_{ij} est donné. Nous noterons simplement que les résultats du Chapitre IV-I permettent déjà d'éprouver l'hypothèse $\beta = \beta_0$, β_0 fixé.

3° Enfin, la structure définie plus haut et les techniques afférentes, peuvent avoir des applications dans des domaines de la statistique apparemment assez distincts du domaine ordinaire des tableaux de corrélation : nous espérons publier très prochainement une telle application aux *méthodes de comparaison par paires*.

III. Étude de l'hypothèse H_R.

L'hypothèse H_R est étudiée de façon générale aux Chapitres II et III; pour le cas qui nous occupe ici, nous indiquons des méthodes (Ch. II IV-2-c et V-3) qui peuvent simplifier cette étude. Nous nous proposons maintenant de discuter la valeur pratique de cette hypothèse et d'examiner l'hypothèse $H_R \cap IM$; on montre comment les méthodes d'estimation et de test qui ont trait à H_R permettent facilement de résoudre les problèmes analogues se rapportant à cette dernière hypothèse.

1. Quelques remarques sur l'hypohèse H_R.

Sous l'hypothèse H_R, on a :

$$\forall (i, j) \in R \qquad P_{ij} = p_i\, q_j$$

ce qui est équivalent à (cf. : Ch. I-B-II) :

$$Pr\,[\beta_j \,/\, \alpha_i\,,\, j \neq i] = \frac{q_j}{\sum\limits_{h \neq i} q_h}$$

ce qui signifie que les sujets ayant le caractère α_i se répartissent suivant les caractères β_j $(j \neq i)$ proportionnellement à des nombres q_j (indépendants de i) attachés à ces caractères.

On posera pour simplifier $\sum\limits_{j=1}^{l} q_j = 1$, ce qui est toujours possible.

Soit P la probabilité de tirer de la population parente un élément $\alpha_i\, \beta_j$ tel que (i, j) appartienne à R; H_R s'écrit

$$p_{ij} = Pr[\alpha_i\, \beta_j] = P\, p_i\, q_j \qquad \text{pour} \quad \text{tout} \quad (i, j) \in R$$

d'où : $$Pr\,[\beta_j \,/\, \alpha_i] = \frac{Pr\,[\alpha_i\, \beta_j]}{Pr\,[\alpha_i]} = \frac{P\, p_i\, q_j}{Pr\,[\alpha_i]} = \phi_i\, q_j \qquad \text{si } i \neq j$$

$$\text{et } Pr[\beta_i/\alpha_i] = 1 - \phi_i \sum_{j \neq i} q_j = 1 - \phi_i + \phi_i\, q_i$$

où les ϕ_i sont des paramètres tels que : $0 \leqslant \phi_i \leqslant \frac{1}{1 - q_j}$. Un modèle voisin a déjà attiré l'attention de quelques auteurs qui étudient la persistance d'un effet dans une chaîne de Markov (Barton, David et Fix, 1962; Goodman, 1964 c). Les différences entre leur étude et la nôtre se situent ansi :

1° Dans le modèle étudié par ces précédents auteurs, on suppose les ϕ_i égaux.

2° Leur problème est traité sous l'angle des chaînes de Markov discrètes du premier ordre : l'expérience porte sur l'observation d'une suite de stades d'une telle chaîne. Nous étudions par contre des tirages indépendants dans une population à deux caractères. On connaît cependant l'analogie entre les méthodes relatives aux tables de contingence et aux chaînes de Markov, et nous comptons prochainement reprendre notre étude pour ce dernier cas. Mais notre modèle convient à l'étude de la persistance d'un effet lorsque l'on possède un échantillon de « transitions » indépendantes pour examiner la succession de plusieurs événements $\alpha_1, \alpha_2, \ldots, \alpha_l$.

L'étude de telles transitions se rattache à celle des phénomènes migratoires (pris au sens le plus large comme au paragraphe précédent). H_R signifie que le seul écart à l'hypothèse d'indépendance H_0 est dans la persistance d'une ou plusieurs situations (ou, au contraire dans une

instabilité exagérée qui serait une persistance négative; il se peut aussi que certains états soient plus persistants que ne le voudrait l'indépendance des deux classifications, et d'autres le soient moins). Dans ces conditions, il peut être intéressant dans la pratique de savoir éprouver le modèle H_R, puis, s'il est conservé, de réaliser un test restreint de H_C à l'intérieur de H_R afin de décider si cette persistance est significative (cf. Ch. III).

2. Étude simultanée des hypothèses H_R et IM.

Puisque H_R est équivalente à la nullité de certaines interactions du premier ordre (Ch. I-B-III) les hypothèses H_R et IM sont séparables (Lemme V) [4].

Théorème XIII. — L'hypothèse $H_R \cap IM$ que l'on notera H_R^+ est équivalente à : $\forall\ i \in \overline{R},\ \exists\ f_i$ tel que $\forall\ (i, j) \in R,\ p_{ij} = f_i f_j$ (5)

(on peut remplacer ci-dessus p_{ij} par la probabilité conditionnelle P_{ij}).

En effet, si $p_{ij} = f_i f_j$, H_R est vraie et d'autre part $p_{ij} = p_{ji}$ d'où S, d'où IM.

Réciproquement, si H_R est vraie, l'on peut poser pour tout $(i, j) \in R$: $p_{ij} = k\, p_i\, q_j$ avec $\sum_{i=1}^{l} p_i = \sum_{i=1}^{l} q_i = 1$; les paramètres p_i et q_j sont définis de façon unique.

On a alors $Pr[\alpha_i] = \sum_{j=1}^{l} p_{ij} = p_{ii} + kp_i\ (1 - q_i)$

$$Pr[\beta_j] = \sum_{i=1}^{l} p_{ij} = p_{jj} + kq_j\ (1 - p_j)$$

IM implique : pour tout $i \in \overline{R}$ $\quad Pr[\alpha_i] = Pr[\beta_i]$ d'où $p_i = q_i$.

Donc $H_R \cap IM$ entraîne $\forall\ (i, j) \in R$, $p_{ij} = kp_i\, p_j$ et de là (5).

H_R fixe $\nu_R = (l-1)(l-2)$ liaisons entre les paramètres p_{ij}; H_R^+ fixe en plus $p_i = q_i$ pour tout $i \in \overline{R}$, soit $l - 1$ liaisons puisque les p_i et q_i sont définis de façon unique et $\sum_{i=1}^{l} p_i = \sum_{i=1}^{l} q_i$. Donc H_R^+ fixe $\nu_R^+ = (l-1)^2$ liaisons entre les paramètres p_{ij}.

Remarques. — 1. Dans une table 3×3, on a $H_R \prec\!\!=\!\!\succ QS$ donc $H_R^+ \prec\!\!=\!\!\succ QS \cap IM$, soit $H_R^+ \prec\!\!=\!\!\succ S$.

2. Pour $l > 3$, on a seulement $H_R \Longrightarrow QS$, soit $H_R \prec\!\!=\!\!\succ H_R \cap QS$; de là H_R^+ est équivalente à $H_R \cap QS \cap IM$, et puisque $QS \cap IM$ est l'hypothèse S on obtient : $H_R^+ \iff H_R \cap S$.

4. Si H_R est remplacée par $H_R(\lambda)$ cette propriété reste vraie.

On notera cependant que H_R et S ne sont pas séparables (par contre, H_R et S sont « séparables à l'intérieur de QS »).

3. Méthodes d'estimation et de test pour l'hypothèse H_R^+

a) *Estimation.* — La vraisemblance d'un échantillon de taille n extrait de la population parente réduite aux éléments $\alpha_i \beta_j$ où $i \neq j$, est sous l'hypothèse H_R^+

$$L = \frac{n!}{\prod_{(i,j) \in R} (x_{ij})!} \prod_{i \in \bar{R}} f_i^{(a_i + b_i)}$$

où $f_i f_j$ est la probabilité conditionnelle de $\alpha_i \beta_j$ sachant que i est différent de j, et $a_i = \sum_{j \in R_{i.}} x_{ij}$, $b_j = \sum_{j \in R_{.j}} x_{ij}$

La méthode du maximum de vraisemblance conduit aux équations

$$\forall i \in \bar{R} \quad \hat{f}_i \sum_{j \in R_{i.}} \hat{f}_j = \frac{a_i + b_i}{2n} = \frac{m_i}{n} \tag{6}$$

Soit, en posant $\hat{P}_{ij} = \hat{f}_i \hat{f}_j$, $\hat{f}_{ij} = n \hat{P}_{ij}$

$$\forall i \in \bar{R} \sum_{j \in R_{i.}} \hat{P}_{ij} = \frac{m_i}{n} \quad \text{ou} \quad \sum_{j \in R_{i.}} \hat{f}_{ij} = m_i$$

La résolution de ces équations est équivalente à la résolution de :

$$\sum_{j \in R_{i.}} \hat{f}_{ij} = n \sum_{j \in R_{i.}} \hat{P}_{ij} = m_i \qquad \forall i \in \bar{R} \tag{7}$$

$$\sum_{i \in R_{.j}} \hat{f}_{ij} = n \sum_{i \in R_{.j}} \hat{P}_{ij} = m_j \qquad \forall j \in \bar{R} \tag{8}$$

$$\hat{P}_{ij} = \hat{p}_i \hat{q}_j \qquad \forall (i, j) \in R \tag{9}$$

en effet, les $\hat{P}_{ij}$ solutions de (7), (8), (9) vérifient à la fois H_R par (9) et IM puisqu'en comparant (7) et (8) on a pour tout $h \in \bar{R}$,

$$\sum_{i \in R_{.h}} \hat{P}_{ih} = \sum_{j \in R_{h.}} \hat{P}_{hj}$$

ainsi les $\hat{P}_{ij}$ peuvent s'écrire $\hat{P}_{ij} = \hat{f}_i \hat{f}_j$.

Or, le système (7), (8), (9) est celui que l'on obtiendrait pour estimer $\hat{p}_i$ et $\hat{q}_j$ sous l'hypothèse H_R, à partir d'un tableau dont les totaux marginaux de la i^{me} ligne et de la i^{me} colonne auraient la valeur com-

mune m_i. Les méthodes de résolution étudiées au Chapitre II sont donc valables ici; il suffira de remplacer les marges a_i et b_i par $m_i = \frac{a_i + b_i}{2}$. En particulier le processus itératif décrit en (Ch. II-IV-2-c) peut être utilisé; dans ce cas, son maniement devient même plus simple.

Remarques. — 1. En utilisant les résultats du Chapitre II, il est immédiat de montrer que les équations (6) admettent une solution et une seule, et que L atteint là son maximum.

2. Si un échantillon de taille fixée N est tiré de la population parente complète, on montrera que les fréquences théoriques $\hat{f}_{ij}$ pour $(i, j) \in R$ sont encore fournies par les équations ci-dessus (où maintenant n est la v.a. $\sum_{(i,j) \in R} x_{ij}$), les fréquences théoriques f_{ii} étant évidemment égales à x_{ii}/N.

b) *Test de* H_R^+

Comme dans les problèmes précédents, on pourra calculer le χ^2 ou les tests équivalents, avec les fréquences théoriques trouvées ci-dessus. Une région critique de seuil α sera :

$$\chi^2_{H_R^+} > l(\alpha, \nu_R^+)$$

D'après le III-2 ci-dessus, on peut utiliser, pour réaliser un test de IM, la nouvelle région critique :

$$\chi^2_{H_R^+} - \chi^2_R = \chi^2_{H_R^+ / H_R} > l(\alpha, \nu_{IM})$$

Dans certains cas, en particulier si l est grand, il peut être avantageux de substituer ce nouveau test de IM à celui qui est décrit au paragraphe II-2-b et qui nécessite alors l'inversion d'une matrice d'ordre élevé.

4. Application aux tables de contingence intraclasses.

Si la population parente est composée de couples *non ordonnés* $\alpha_i \alpha_j$ un échantillon tiré de cette population peut être représenté dans une table de contingence intraclasses. Notons p'_{ij} la probabilité de tirer un couple dont l'un des éléments a le caractère α_i, l'autre le caractère α_j $(i \leqslant j)$ et x_{ij} le nombre de couples ainsi définis figurant dans l'échantillon. L'hypothèse d'indépendance entre les deux classifications peu s'écrire : $p'_{ii} = p_i'^2$ et $p'_{ij} = 2\, p'_i\, p'_j$ pour $i < j$ et a déjà été étudiée (Ishii, 1960). Lorsque cette hypothèse est fausse, ceci risque souvent d'être la conséquence de trop grandes valeurs des p_{ii}. Il peut être intéressant d'étudier alors l'hypothèse moins restrictive (qui correspond à H_R) : $p'_{ij} = p'_i\, p'_j$ pour $i < j$.

Si $P'_{ij} = Pr[\alpha_i \alpha_j / i \neq j]$ on peut écrire cette hypothèse $P'_{ij} = f_i\, f_j$ (pour $i < j$).

Lorsqu'un échantillon de taille n est extrait de la population réduite aux couples $\alpha_i\,\alpha_j$ où $i \neq j$, les estimateurs $\hat{f}_i$ obtenus par le maximum de vraisemblance vérifient les équations :

$$\hat{f}_i \sum_{j \neq i} \hat{f}_j = \frac{m_i}{n} \quad \text{avec} \quad m_i = \sum_{j>i} x_{ij} + \sum_{h<i} x_{hi}$$

Ces équations sont donc analogues à (6) et pourront être résolues de la même façon. On pourra traiter de même le cas d'un échantillon extrait de la population parente complète, et, de façon analogue à ce qui a été fait jusqu'ici, on pourra éprouver l'hypothèse envisagée.

IV. Conclusion sur le choix d'un modèle dans un tableau de corrélation carré.

Dans ce chapitre, nous avons étudié essentiellement deux hypothèses H_R et QS sur la structure interne d'un tableau carré; nous avons tenté d'analyser leur signication pratique afin de mieux situer dans quel domaine ces structures sont réalistes. Nous avons pressenti qu'il pouvait en être ainsi dans certains processus de transition, on peut noter qu'il était tout à fait naturel d'envisager alors parallèlement l'hypothèse d'homogénéité des distributions marginales qui est une hypothèse de stationnarité de ces processus. Nous avons indiqué divers tests relatifs à ces modèles, ainsi que des méthodes d'estimation des paramètres. Les résultats des Chapitres II - III - IV permettent d'analyser quelques nouvelles structures. Cependant, notre travail ne saurait être exhaustif et ces modèles ne peuvent recouvrir toutes les possibilités de structure d'un tableau de corrélation. Nous espérons pourtant qu'ils peuvent s'avérer utiles dans la mesure où ils sont susceptibles d'interprétation simple.

Nous n'avons jamais parlé « d'indice » de corrélation; en effet, notre but était avant tout de caractériser la « forme » d'une association plutôt que son importance, ce dernier problème étant abondamment traité dans la littérature (pour une discussion de ces questions et une importante bibliographie, voir GOODMAN et KRUSKAL, 1954, 1959 et 1963). Mais les deux points de vue sont complémentaires, on en trouvera des exemples au Chapitre VI.

Pour terminer, il convient de signaler une difficulté d'ordre méthodologique qui peut apparaître dans le choix d'un modèle d'association, nous indiquons ensuite une méthode heuristique pour la détourner.

Pour des échantillons de taille relativement faible, si nous éprouvons une hypothèse fausse mais voisine de la réalité, nous conclurons souvent qu' « il n'y a aucune raison de rejeter cette hypothèse »; malgré l'erreur commise, ceci peut être une bonne chose : souvent le praticien désire connaître seulement une structure grossière de la population étudiée; en particulier, il préfère presque toujours un modèle approché simple à un modèle complexe exact (en supposant qu'il soit possible de connaître un « modèle exact »...).

Pour des échantillons de grande taille, tout modèle simplement « approché » sera rejeté par le test d'ajustement approprié, avec une probabilité qui tend vers un. Néanmoins, parmi ces modèles, l'un peut être utile; l'ajustement ne doit pas consister à rejeter un modèle sous prétexte d'inexactitude, mais plutôt à choisir le plus raisonnable. (Il est évident, qu'augmenter le nombre des modèles à notre disposition ne peut pas résoudre le fond du problème).

Définir un modèle « raisonnable » est difficile de façon objective; il conviendrait pour cela de posséder une mesure de l' « écart » entre une distribution observée et les modèles théoriques envisagés; cependant cette mesure ne peut avoir d'une part qu'une valeur relative (mais elle pourrait être suffisante pour *comparer* la validité de plusieurs modèles), d'autre part, il reste possible de juger plus raisonnable qu'un autre, un modèle un peu moins bien adapté mais plus simple.

Comme mesure d'écart à un modèle H, il paraît naturel d'envisager une mesure basée sur $\chi^2_H = \sum \frac{(x_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}$ où les $\hat{f}_{ij}$ sont les fréquences théoriques sous l'hypothèse H; on connaît les critiques dont elle a été l'objet en tant que mesure d'association, mais il s'agit ici d'un problème tout à fait différent et l'on a besoin d'un indicateur de portée générale, calculable de façon analogue sous des hypothèses très diverses. Afin de rendre comparables les mesures obtenues à partir de plusieurs modèles, on peut prendre simplement : $D_H = \frac{\chi^2_H - \nu_H}{N}$

où ν_H est le nombre de ddl associé à χ^2_H et N la taille de l'échantillon disponible.

Les raisons de ce choix sont les suivantes :

Si H est fausse, la v.a. χ^2_H est voisine d'un χ^2 non centré à ν_H ddl et de paramètre de non centralité λ, de la forme $\lambda = N\, d_H$ où d_H est un paramètre inconnu ne dépendant que de H et de la structure vérifiée par la population parente (ceci se déduit de résultats cités dans DIAMOND, MITRA et ROY, 1960). Ce paramètre d_H peut être considéré comme mesure de la distance entre la structure exacte et la structure H. D'autre part, si la v.a. Y suit une loi de χ^2 non centré de paramètres ν (ddl) et λ (non centralité) on a (voir par exemple KENDALL et STUART 1961, page 229) :

$$E(Y) = \nu + \lambda,\ \mathrm{Var}(Y) = 2(\nu + 2\lambda).$$

Dans ces consitions, on peut admettre que $E(D_H)$ est voisin de d_H et $\mathrm{Var}(D_H)$ est très petit si N est grand, d'où le choix suggéré plus haut.

Évidemment, beaucoup de ces remarques sur la recherche d'un modèle mathématique seront valables dans d'autres contextes que celui qui est actuellement le nôtre.

CHAPITRE VI

APPLICATIONS

Afin de montrer comment les suggestions des chapitres antérieurs pourraient compléter l'arsenal des méthodes que les statisticiens ont déjà proposées aux praticiens, nous avons essentiellement choisi des données numériques déjà publiées et étudiées précédemment sous des angles divers.

Ces exemples sont examinés pour illustrer l'utilisation des méthodes statistiques que nous avons étudiées et non préciser tel ou tel point de recherche appliquée; dans ces conditions, si nous sortons de notre domaine pour nous pencher sur la signification pratique d'un résultat numérique, ce sera seulement à titre indicatif, pour tenter de mieux situer l'intérêt de ces méthodes : il est bien évident qu'une interprétation ne peut être entreprise qu'en collaboration avec l'expérimentateur et sort du cadre de ce travail. D'autre part, pour que nos calculs soient valides, il faut connaître parfaitement la technique expérimentale choisie pour s'assurer que les conditions, qui sont à la base des méthodes proposées, ont été respectées; nous supposerons ceci réalisé.

Exemple I.

Référence : TALLIS (1962, p. 349).

		Année 1952		
	Nombre d'agneaux	0	1	2
Année 1953	0	58	52	1
	1	26	58	3
	2	8	12	9

TABLEAU 1

Dans le tableau 1 ci-dessus, 227 brebis ont été classées selon le nombre d'agneaux mis bas lors de deux années consécutives : les lignes correspondent à la première année, les colonnes à la deuxième année.

1. **Test de H_R** (R est définie comme au Chapitre V).

Ici χ^2_R est facile à calculer sous sa forme W^2. Il y a une seule liaison, prenons :

$F_1(P) = \text{Log } P_{12} + \text{Log } P_{23} + \text{Log } P_{31} - \text{Log } P_{21} - \text{Log } P_{13} - \text{Log } P_{32} = O$

On a : $F_1(x) = d_1 = 1{,}386294 \qquad v_{11} = 1{,}599359$

$W^2 = d^2_1/v_{11} = 1,202$

$\nu_R = 1$, $l(0{,}05; 1) = 3{,}841$, le test n'est pas significatif au seuil 0,05.

On a d'autre part, par un calcul classique, $\chi^2_C = 49{,}641$ d'où $\chi^2_{H_C / H_R} = \chi^2_C - \chi^2_R = 48{,}439$ (pour $\nu_C - \nu_R = 3$ ddl) le test est très significatif.

2. Test exact de H_R

Dans cet exemple le test du χ^2 n'est peut-être pas très valide à cause des petites fréquences; on peut appliquer le test exact (cf. Ch. III-V) qui ne demande pas ici des calculs excessifs.

Les marges étant fixées (marges du tableau réduit par la suppression des éléments diagonaux) il existe 5 tableaux (tronqués) ayant ces marges :

	53	0		52	1		51	2		50	3		49	4
25		4	26		3	27		2	28		1	29		0
9	11		8	12		7	13		6	14		5	15	

On peut décrire ces tableaux à partir de la seule v.a. $X = X_{23}$. En utilisant la formule (8) (Ch. III) on obtient les probabiltiés conditionnelles pour les marges données :

$Pr[X = 0] = 0{,}044$ $Pr[X = 1] = 0{,}256$ $Pr[X = 2] = 0{,}423$
$Pr[X = 3] = 0{,}238$ $Pr[X = 4] = 0{,}039$
$E(X) = 1{,}972$

Nous adopterons le seuil $\alpha = 0{,}05$; essayons les valeurs
$K_1 = 0$, $K_2 = 4$
On aura $\alpha\, E(X) = 0{,}0986$

$\phi(X) = 0$ pour $X = 1, 2, 3$, $\phi(0) = \pi_1$ $\phi(4) = \pi_2$

d'où $E[\phi(X)] = \pi_1 . 0{,}044 + \pi_2 . 0{,}039$

et $E[X\phi(X)] = 0{,}039 . 4 . \pi_2 = 0{,}156\, \pi_2$

La relation (9) Ch. III s'écrit :

$0{,}044\, \pi_1 + 0{,}039\, \pi_2 = 0{,}05$

La relation (10) s'écrit :

$0{,}156\, \pi_2 = 0{,}0986$

d'où : $\pi_1 = 0{,}577$ $\pi_2 = 0{,}632$.

Le test est ainsi déterminé par :

$$\begin{cases} \phi(X) = 0 & \text{si } 0 < X < 4 \\ \phi(X) = 0{,}577 & \text{si } X = 0 \\ \phi(X) = 0{,}632 & \text{si } X = 4. \end{cases}$$

Dans le cas actuel $X = 3$ d'où $\phi(X) = 0$ et l'hypothèse H_R est conservée au seuil 0,05.

3. Discussion.

D'après les résultats du paragraphe 1, on doit accepter H_R et rejeter H_0, on peut donc admettre en pratique qu'il y a une association entre les distributions du nombre d'agneaux d'une année à l'autre mais que celle-ci provient de la persistance d'un comportement (par contre, lorsque celui-ci change, il change « au hasard »). Ceci pourrait justifier l'adoption pour indice d'association, d'un simple indice de persistance, par exemple un indice de confiance (reliability) (cf. GOODMAN et KRUSKAL, 1954, p. 758) : on peut utiliser la probabilité d'agrément, estimée ici par $(58 + 58 + 9)/227 = 0,55$. Le calcul de cet indice est très facile; de plus sa distribution simple (binomiale) entraîne un maniement aisé : intervalle de confiance, comparaison de plusieurs indices, etc...

Nous avons donc montré une façon de décrire la structure de ce tableau, permettant d'éviter le recours toujours dangereux à la normalité de quelque distribution latente. Ces résultats peuvent être utiles pour certains types de conclusions; dans d'autres cas, cependant, le coefficient de corrélation de TALLIS peut avoir plus d'intérêt : les deux études se complètent.

4. Test de l'hypothèse IM.

On calcule très rapidement $\chi^2_S = 19,511$

Puisque dans ce cas H_R <==> QS, ayant accepté H_R, on peut éprouver IM par :

$\chi^2_S - \chi^2_R = 19,511 - 1,202 = 18,309$

$\nu = 3 - 1 = 2$ test très significatif.

(le χ^2_{IM} de BHAPKAR est égal à 19,716, celui de STUART à 18,141; ils conduisent à la même conclusion).

Exemple II.

Référence : STUART (1953, p. 109).

1. Le tableau 2.a indique, classées selon quatre degrés, les distances de vision des deux yeux pour 7477 femmes (entre parenthèses figurent les fréquences théoriques sous l'hypothèse QS).

1520 (id)	266 (236.38028)	124 (133.58377)	66 (59.03595)
234 (236.61972)	1512 (id)	432 (418.98576)	78 (88.39452)
117 (107.41623)	362 (375.01424)	1772 (id)	205 (201.56953)
36 (42.96405)	82 (71.60548)	179 (182.43047)	492 (id)

TABLEAU 2.a

STUART (1955), puis BHAPKAR (1966) étudient l'identité des distributions marginales et sont amenés à rejeter cette hypothèse. L'hypothèse QS signifierait que la dissymétrie est due uniquement à la différence de ces distributions marginales : nous avons cherché à éprouver cette hypothèse; la statistique χ^2_{QS} a été calculée de deux façons :

à partir des fréquences théoriques f_{ij} obtenues au moyen de l'algorithme décrit au Chapitre V (et figurant dans le tableau 2.a), on a :

$$\chi^2_{QS} = 7,258$$

la statistique de Wald est ici :

$$W^2_{QS} = 7,224$$

Nous avons $\nu_{QS} = 3$; les décisions à partir de l'une ou de l'autre de ces statistiques coïncident : au seuil 0, 05 on conservera QS [$l\,(0,05;3) = 7,815$] (nous sommes cependant dans le voisinage de la valeur critique).

La quasi-symétrie pourrait donc fournir un modèle assez satisfaisant de la structure de ce tableau de corrélation. On notera que le test de IM peut se réduire au test restreint de S à l'intérieur de QS. On a

$$\chi^2_S = 19,107$$
$$\chi^2_S - \chi^2_{QS} = 11,849$$

(cette valeur est comparable à $\chi^2_{IM} = 11,97$ trouvé par BHAPKAR).

Remarque. — Ici le test de H_R est significatif au seuil 0, 05; il est possible d'en décider très rapidement : on pourra par exemple calculer le χ^2 d'indépendance relatif à la table 2×2 composée des intersections des lignes 1-2 et des colonnes 3-4; on obtient 32, 026, or $\nu_R = 5$ et $l\,(0,05;5) = 11,071$. On a un résultat analogue en calculant pour l'interaction correspondante, l'intervalle de confiance de seuil 0, 05 donné par (6) Chapitre III : on obtient :

$$-1,081 \pm 0,652 \quad \text{soit} \quad [-1,733; -0,429]$$

cet intervalle ne contient pas zéro.

2. Le tableau 2.b produit une classification analogue à la précédente pour 3 242 hommes. On a effectué les mêmes calculs que pour le tableau 2.a, afin d'éprouver les hypothèses gigognes QS, puis S.

821 (id)	112 (116.05471)	85 (80.29707)	35 (35.64821)
116 (111.94524)	494 (id)	145 (148.72085)	27 (27.33390)
72 (76.70291)	151 (147.27912)	583 (id)	87 (86.01796)
43 (42.35180)	34 (33.66611)	106 (106.98209)	331 (id)

TABLEAU 2.b

$\chi^2_{QS} = 1,089$ ($\nu = 3$; QS est acceptée au seuil 0,05)
$\chi^2_S = 4,762$
$\chi^2_S - \chi^2_{QS} = 3,673$ ($\nu = 3$; S est acceptée au seuil 0,05)
(la statistique χ^2_{IM} de Bhapkar vaut 3,678, valeur très voisine de $\chi^2_S - \chi^2_{QS}$).

3. On a comparé enfin les deux tableaux précédents en éprouvant l'identité des interactions du premier ordre :

1° Dans les deux tableaux tronqués (suppression de la diagonale principale qui semble jouer un rôle particulier : on notera S la partie de T associée à cette troncature).

2° Dans les deux tableaux complets (ce dernier test ne devant être effectué que si le précédent n'est pas significatif).

Evidemment on a $K_T \subset K_S$.

$\chi^2_{K_S} = 6,115$

$\nu_{K_S} = 5$, on acceptera K_S au seuil 0,05.

$\chi^2_{K_T} = 29,327$ d'où $\chi^2_{K_T/K_S} = \chi^2_{K_T} - \chi^2_{K_S} = 23,212$
le nombre de degrés de liberté est $\nu_{K_T} - \nu_{K_S} = 4$
On rejettera K_T au seuil 0,05.

En définitive, on admettra que ces deux tableaux ont des structures analogues si la diagonale principale n'est pas prise en compte, mais non lorsqu'elle est considérée; cette conclusion confirme cette idée que les éléments diagonaux jouent un rôle à part.

4. STUART (1953) étudie un indice d'association pour caractériser l'importance de la corrélation entre les deux classifications; il ne trouve pas de différence significative entre les valeurs de cet indice pour chacun des deux tableaux. Nous avons d'abord complété cette étude d'*intensité* de l'association par une étude de sa forme. En comparant ensuite ces deux tableaux, nous avons pu déceler une différence significative précise de leurs structures. Dans ces conditions, une modification du choix de l'indice d'association pourrait très bien aboutir à un changement de conclusion sur la signification de ses valeurs respectives pour l'un et l'autre des deux cas envisagés.

Exemple III.

1954 \ 1962	1	2	3	4	5	6
1	187	13	17	11	3	1
2	4	191	4	9	22	1
3	22	8	182	20	14	3
4	6	6	10	323	7	4
5	1	3	4	2	126	17
6	0	2	2	5	1	153

TABLEAU 3

Le tableau 3 a été construit à partir de données qui nous ont été communiquées par MM. KAYSER (Faculté des Lettres et Sciences humaines de Toulouse) et LEDRUT (C.N.R.S. Sociologie) que nous tenons à remercier. Ces auteurs ont classé un échantillon d'individus d'après leur catégorie socio-professionnelle en 1954 puis 1962 (on a retenu 6 catégories). La forte corrélation entre les deux classements est évidente; pour essayer de préciser le mode de transition, nous avons éprouvé l'hypothèse QS. Nous avons obtenu

$$\chi^2_{QS} = 26,765 \qquad (\nu_{QS} = 10).$$

Donc, ici, l'hypothèse QS sera rejetée au seuil 0,05. Nous avons obtenu d'autre part $\chi^2_S = 51,783$ d'où :

$\chi^2_S - \chi^2_{QS} = 25,018$ $\qquad (\nu_{IM} = 5)$

Donc, l'hypothèse IM est aussi rejetée au seuil 0,05.

Exemple IV.

L'exemple suivant est tiré de données de PEARSON reproduites récemment par GOOD (J. R. Stat. Soc. B, 1956, 18). Il est du même type que le précédent : 775 couples père-fils sont classés selon le métier de l'un et de l'autre (14 métiers sont considérés).

La présence de très petites fréquences, en particulier de nombreux zéros, rend peu recommandable l'emploi du test du χ^2; nous verrons cependant que les conclusions que nous en tirerons sont très « nettes », basées sur un χ^2 éloigné de sa valeur critique; dans ces conditions, elles semblent malgré tout assez sûres.

Etude de QS. — Sous l'hypothèse QS il y a 16 fréquences théoriques nulles; les cases correspondantes ont été écartées pour le calcul de χ^2_{QS}; on a trouvé ainsi :

$$\chi^{*2}_{QS} = 78,679$$

Si l'on se rapporte à la table tronquée associée à QS, on voit que celle-ci a 8 totaux marginaux nuls; il semble alors raisonnable de prendre ici $\nu^*_{QS} = \nu_{QS} - 8$, soit $\nu^*_{QS} = 70$.

Dans cet exemple, le χ^2 observé n'est pas significatif (en fait, il est très près de sa valeur moyenne). L'hypothèse QS est retenue.

On a éprouvé aussi l'hypothèse H_R (contenue dans QS). Malgré les grandes dimensions de ce tableau, l'algorithme décrit au Chapitre II (IV-2-c) est assez maniable et a donné x (avec cinq décimales exactes) en trois itérations (on a pris $x_0 = 550$, on a obtenu $x_1 = 568,58568$ puis $x_2 = x_3 = 568,64697$). Les calculs suivants sont longs à cause des grandes dimensions du tableau, mais élémentaires. On a obtenu : $\chi^2_R = 262,361$ d'où $\chi^2_{R/QS} = 183,682$ (avec $\nu = 85$); le test est très significatif.

On remarquera que, dans ce cas, le calcul de χ^2_R par les autres méthodes que nous avons envisagées, demanderait l'inversion d'une matrice de grandes dimensions, et serait beaucoup plus long que le calcul indiqué ci-dessus.

Exemple V.

	296	242	119	284	101	48
	328	418	35	479	97	46
	295,6696	241,6405	117,2589	286,2235	101,3851	47,8225
	297,8075	242,2693	116,1214	284,3474	101,1132	48,3411
67		138	47	54	19	9
140		260	44	83	32	30
67,3304		138,0359	46,4887	54,9355	18,7012	8,5084
67,6806		140,5902	46,4188	52,8611	18,1241	8,3252
242	608		270	340	105	43
396	588		140	472	124	41
242,3595	607,9641		266,0698	342,9504	105,4231	43,2331
241,6607	617,0699		266,2063	336,9516	103,3736	42,7379
141	248	319		159	53	43
220	309	518		165	69	84
142,7411	248,5113	322,9302		150,7177	53,8354	44,2642
140,7571	247,5847	323,4960		155,8811	51,6374	43,6437
204	171	244	79		221	50
274	122	163	34		398	94
201,7766	170,0645	241,0496	87,2823		219,4639	49,3631
201,4021	164,7488	239,2625	91,0857		222,7292	49,7717
81	65	84	36	246		26
21	12	15	7	162		27
80,6149	65,2988	83,5769	35,1646	247,5361		25,8088
80,6662	63,6227	82,6772	33,9852	250,8685		26,1803
227	177	205	174	332	154	
336	131	202	159	690	374	
227,1776	177,4916	204,7669	172,7358	332,6369	154,1912	
229,8333	174,1664	203,7048	171,1826	334,0904	156,0224	

TABLEAU 4

Nous empruntons à GIL et OMABOE (1963) cet exemple qui concerne le nombre de migrants (en centaines) entre diverses régions du Ghana; pour plus de précision, voir la référence citée. Il s'agit d'une table tronquée, les mouvements à l'intérieur d'une même région n'étant pas considérés.

Dans chaque case du tableau 4 on a écrit (de haut en bas) le nombre correspondant de migrants (en centaines) puis, exprimés avec la même unité, les fréquences théoriques sous les hypothèses suivantes :

1) Hypothèse étudiée par GIL et OMABOE qui peut s'écrire :
$p_{ij} = \alpha\, p_i\, q_j\, (d_{ij}^{-1})^{\beta}$ où toutes les quantités p_i, q_j, d_{ij} sont fixées, α et β sont les seuls paramètres inspécifiés.

2) Hypothèse QS.

3) Hypothèse $p_{ij} = p_i\, q_j\, d_{ij}^{-1}$ où les d_{ij} sont fixés (on les a pris égaux aux distances géographiques entre capitales des diverses régions, telles qu'elles sont utilisées par GIL et OMABOE). C'est une hypothèse du type $H_R(\lambda)$, R étant définie comme au Chapitre V et $\lambda_{ij} = d_{ij}^{-1}$ pour tout $(i, j) \in R$.

En utilisant les fréquences théoriques ci-dessus, nous avons calculé

$$\chi^2_{QS} = 168,303 \qquad (\nu_{QS} = 15)$$
$$\chi^2_{H_R}(\lambda) = 310,810 \qquad (\nu_R = 29)$$

Dans les deux cas le test de l'hypothèse correspondante est très significatif, bien que les fréquences théoriques soient proches des fréquences observées; mais, comme nous l'avons vu plus haut, même si la réalité s'écarte très peu du modèle étudié, il en sera ainsi presque toujours avec de si grands échantillons. (On remarquera par exemple que, si les nombres contenus dans le tableau représentaient des unités (ou même des dizaines) et non des centaines, aucun des deux tests ne serait significatif).

Le calcul des indices D_H donne :

$$D_{H_R}(\lambda) = 0,42 \times 10^{-3}$$
$$D_{QS} = 0,23 \times 10^{-3}$$

Ces deux indices sont du même ordre, même si D_{QS} est un peu plus petit.

On a calculé en outre cet indice pour le modèle de GIL et OMABOE (avec les fréquences théoriques données par ces auteurs), pour trouver :

$$D_{GO} = 0,35 \times 10^{-1}$$

ainsi D_{GO} est environ cent fois plus grand que $D_{H_R(\lambda)}$ et D_{QS}.

Partant de QS ($p_{ij} = p_i\, q_j\, \lambda_{ij}$; p_i, q_j, λ_{ij} inconnus, $\lambda_{ij} = \lambda_{ji}$) le fait de spécifier les paramètres $\lambda_{ij} = d^{-1}{}_{ij}$ fait peu varier l'écart du modèle aux données; par contre, si l'on fixe aussi p_i et q_j (même en laissant arbitraire un paramètre β pour mieux « adapter » les distances) cet écart varie beaucoup. On arrive ainsi à cette conclusion pratique, que le modèle de GIL et OMABOE semble « bon » en ce qui concerne les valeurs imposées aux d_{ij}, mais semble surtout s'écarter de la réalité à cause des valeurs imposées aux p_i et q_j; le praticien peut alors se poser (entre autres) les questions suivantes :

— doit-on conserver le modèle plus riche adopté, malgré l'écart assez important entre fréquences observées et théoriques?

— doit-on envisager de le modifier, soit en laissant arbitraires les valeurs des p_i et q_j, soit en fixant ces valeurs à partir de considérations pratiques (économiques, démographiques, etc...) qui pourraient être différentes des conditions initialement adoptées?

CONCLUSION

Ce travail débute par une étude des tables de contingence incomplètes; notre intérêt pour cette question vient surtout de ce que l'on peut en tirer quelques méthodes pour analyser la structure d'un tableau de corrélation. Cependant, les possibilités d'analyse ainsi introduites ne sauraient être suffisantes. Les problèmes les plus intéressants, nous ont paru liés à la structure des tableaux carrés; à leur sujet, plusieurs questions avaient été soulevées qui étaient restées sans réponse. Aussi, c'est pour ce cas particulier que nous avons poursuivi la recherche de nouveaux modèles; l'hypothèse de quasi-symétrie, que nous avons introduite, pourrait être d'une utilisation assez vaste; on a d'ailleurs noté que son intérêt dépassait le cadre que nous nous étions fixé et nous espérons en donner des applications nouvelles. Quelques indications sont données sur d'autres modèles; il y a dans cette direction un large champ d'investigations apparemment peu exploré.

Mais, dans la recherche d'un modèle pour la structure d'un tableau de corrélation, comme pour beaucoup de problèmes analogues de statistique, l'essentiel n'est pas l'étude technique d'une structure particulière donnée, bien qu'elle puisse présenter un intérêt en soi, et parfois quelques difficultés. La question cruciale est une question de méthode : c'est le problème des critières de choix entre plusieurs modèles : d'un côté, il rejoint le délicat problème des tests à décisions multiples, d'un autre côté il montre combien la théorie classique des tests n'est pas tout à fait adaptée à cette démarche. Ces questions sont discutées; quelques indications heuristiques sont données, mais elles sont loin d'être absolument satisfaisantes; c'est surtout dans cette direction, sortant du cadre des tableaux de corrélation pour nous placer d'un point de vue plus général, que nous aimerions continuer ultérieurement nos recherches.

APPENDICE

SUR LA CONVERGENCE D'UN ALGORITHME

Nous indiquons ici comment l'on peut montrer, dans des cas particuliers simples, la convergence de l'algorithme R.A.S. Ces démonstrations sont loin de recouvrir le cas général et même les cas les plus intéressants dans la pratique. Cependant, elles apportent un début de preuve à la conjecture de THIONET (1964) selon laquelle les *conditions d'existence* d'un tableau ayant les marges et les zéros donnés *sont suffisantes* pour que converge cet algorithme.

On considère la suite $\|t_{ij}^{(n)}\|$ définie par :

$$t_{ij}^{(p)} = t_{ij}^{(p-1)} \frac{b_j}{\sum_i t_{ij}^{(p-1)}} \tag{1}$$

$$t_{ij}^{(p-1)} = t_{ij}^{(p)} \frac{a_i}{\sum_j t_{ij}^{(p)}} \tag{2}$$

avec

$$i \in \{1, 2, \ldots l\} = \bar{R} \qquad j \in \{1, 2, \ldots, c\} = R^*$$

(l'ensemble de variation de i n'est pas précisé s'il s'agit de $\bar{R}$ celui de j s'il s'agit de R^*)

$$a_i > 0 \text{ pour tout } i \in \bar{R}, \; b_j > 0 \text{ pour tout } j \in R^*.$$

La matrice $\|t_{ij}^{(0)}\|$ est donnée telle que $t_{ij}^{(0)} \geqslant 0$ pour tout $(i, j) \in \bar{R}$.

Nous désignons constamment par p un entier pair strictement positif.

Nous partirons des assertions de THIONET, à savoir :

$$\text{les suites } \quad S_p = \sup_i \frac{a_i}{\sum_j t_{ij}^{(p)}} \quad \text{et } S'_p = \sup_j \frac{\sum_i t_{ij}^{(p+1)}}{b_j}$$

sont décroissantes et convergent vers une limite L

$$\text{les suites } \quad s_p = \inf_i \frac{a_i}{\sum_j t_{ij}^{(p)}} \quad \text{et } s'_p = \inf_j \frac{\sum_i t_{ij}^{(p+1)}}{b_j}$$

sont croissantes et convergent vers une limite $\mathfrak{l}$ $(L \geqslant \mathfrak{l} > 0)$

La convergence de la matrice $\|t_{ij}^{(n)}\|$ sera montrée si l'on prouve $L = \mathfrak{l} = 1$.

1° Supposons : $\forall (i, j) \in \bar{R} \quad t_{ij}^{(0)} > 0$

Il existe $\alpha > 0$ et $A > \alpha$ tels que : $\forall (i, j) \in \bar{R} \qquad \alpha \leqslant t_{ij}^{(\cdot)} \leqslant A.$

Nous montrerons que dans ces cas : il existe toujours a strictement positif indépendant de n tel que $t_{ij}^{(n)} > a$ pour tout n.

Des propriétés des suites s_p et s'_p on déduit : $\forall n$, $\exists d > 0$ tel que :

$$\sum_j t_{ij}^{(n)} \geqslant d \qquad (\forall i \in \overline{R})$$

et (3)

$$\sum_i t_{ij}^{(n)} \geqslant d \qquad (\forall j \in R^*)$$

D'autre part, l'on peut écrire :

$$t_{ij}^{(n)} = t_{ij}^{(0)} \mu_i^{(n)} \nu_j^{(n)} \tag{4}$$

en définissant les suites $\mu_i^{(n)}$, $\nu_j^{(n)}$ par :

$$\mu_i^{(0)} = \nu_j^{(0)} = 1 \text{ pour tout } i \text{ et } j.$$

$$\mu_i^{(p)} = \mu_i^{(p-1)} \quad , \quad \mu_i^{(p+1)} = \frac{a_i}{\sum_j t_{ij}^{(p)}} \mu_i^{(p)}$$

$$\nu_j^{(p)} = \frac{b_j}{\sum_i t_{ij}^{(p-1)}} \nu_j^{(p-1)} \quad , \quad \nu_j^{(p+1)} = \nu_j^{(p)}$$

De (1) l'on déduit :

$$\sum_i t_{ij}^{(p)} = \nu_j^{(p)} \sum_i t_{ij}^{(0)} \mu_i^{(p)} = b_j$$

d'où

$$\nu_j^{(p)} = \frac{b_j}{\sum_i t_{ij}^{(0)} \mu_i^{(p)}}$$

donc :

$$\forall j \in R^*, \qquad \nu_j^{(p)} \geqslant \frac{b_j}{A \sum_i \mu_i^{(p)}} \tag{5}$$

et :

$$\forall j \in R^* \ , \ \nu_j^{(p)} \leqslant \frac{b_j}{\alpha \sum_i \mu_i^{(p)}} \tag{6}$$

et de (4) et (5) l'on déduit :

$$t_{ij}^{(p)} \geqslant \frac{\alpha\, b_j}{A} \frac{\mu_i^{(p)}}{\sum_i \mu_i^{(p)}} \tag{7}$$

Mais l'on a aussi :

$$\sum_j t_{ij}^{(p)} = \mu_i^{(p)} \sum_j t_{ij}^{(0)} \nu_j^{(p)} \leqslant A\, \mu_i^{(p)} \sum_j \nu_j^{(p)} \leqslant A\, \mu_i^{(p)} \frac{\sum_j b_j}{\alpha \sum_i \mu_i^{(p)}}$$

en utilisant (6), et de là :

$$\frac{\mu_i^{(p)}}{\sum_i \mu_i^{(p)}} \geqslant \frac{\alpha \sum_j t_{ii}^{(p)}}{A \sum_j b_j} \geqslant \frac{\alpha\, d}{A \sum_j b_j} \qquad \text{d'après (3)}$$

Portant dans l'inégalité (7) :

$\forall p$, $t_{ij}^{(p)} \geqslant \dfrac{\alpha^2 b_j d}{A^2 \sum\limits_j b_j}$, quantité strictement positive et indépendante de p.

Une démonstration analogue permettra de minorer de la même façon $t_{ij}^{(p+1)}$ (on peut aussi partir de (2) et utiliser le résultat ci-dessus), on aura finalement :

$$\forall n \quad et \quad \forall (i,j) \in \widetilde{R}, \quad \exists\, a > 0, \quad t_{ij}^{(n)} \geqslant a$$

L'on déduira de là la convergence de l'algorithme de la façon suivante.

Posons $\delta = L - l \geqslant 0$.

Partant de (2) et désignant par k l'indice de la plus grande des quantités $\dfrac{a_i}{\sum\limits_j t_{ij}^{(p)}}$, il vient pour tout j :

$$\sum_i t_{ij}^{(p+1)} = \sum_{i \neq k} t_{ij}^{(p)} \frac{a_i}{\sum\limits_j t_{ij}^{(p)}} + t_{kj}^{(p)} \frac{a_k}{\sum\limits_j t_{kj}^{(p)}}$$

d'où :

$$\sum_i t_{ij}^{(p+1)} \geqslant \left(\inf_i \frac{a_i}{\sum\limits_j t_{ij}^{(p)}} \right) \sum_{i \neq k} t_{ij}^{(p)} + t_{kj}^{(p)} \frac{a_k}{\sum\limits_j t_{kj}^{(p)}}$$

Or, des propriétés des suites s_p et S_p on déduit d'une part :

$$\forall\, \varepsilon > 0, \exists\, p_0 \text{ tel que } p > p_0 \quad \Rightarrow \quad \inf_i \frac{a_i}{\sum\limits_j t_{ij}^{(p)}} > l - \varepsilon$$

et d'autre part :

$$\frac{a_k}{\sum\limits_j t_{kj}^{(p)}} \geqslant L = l + \delta$$

d'où, pour tout $j \in R^*$

$\forall\, \varepsilon > 0, \exists\, p_0$ tel que

$$p > p_0 \quad \Rightarrow \sum_i t_{ij}^{(p+1)} > (l-\varepsilon) \sum_{i \neq k} t_j^{(p)} + t_{kj}^{(p)} (l + \delta)$$

et, puisque

$$\sum_i t_{ij}^{(p)} = b_j \quad \text{et} \quad t_{kj}^{(p)} \geqslant a$$

nous aurons :

$$\sum_i t_{ij}^{(p+1)} > l\, b_j - \varepsilon\ b_j + a\, \delta \quad \text{soit} \quad \frac{\sum\limits_i t_{ij}^{(p+1)}}{b_j} > l - \varepsilon + \frac{a\delta}{b_j}$$

avec $\dfrac{a\,\delta}{b_j}$ strictement positif si $\delta \neq 0$, ce qui est incompatible avec le fait que $\inf\limits_j \dfrac{\sum\limits_i t_{ij}^{(p+1)}}{b_j}$ tend vers l si p tend vers l'infini.

D'où :

$$\delta = 0 \quad \text{et} \quad L = l$$

On aura donc :

$$\sum_i t_{ij}^{(p+1)} \to L b_j \quad \text{et} \quad \sum_j \sum_i t_{ij}^{(p+1)} \to L \sum_i b_j$$

or

$$\sum_j \sum_i t_{ij}^{(p+1)} = \sum_i \sum_j t_{ij}^{(p+1)} = \sum_i a_i$$

d'où

$$L = \frac{\sum_i a_i}{\sum_j b_j}$$

Ecrivant alors l'unique condition d'existence pour notre cas :

$$\sum_i a_i = \sum_j b_j \ ,$$

on a bien $L = l = 1$.

2° Supposons $t_{11}^{(0)} = 0$ et $t_{ij}^{(0)} > 0$ pour tout $(i, j) \in R = \tilde{R} - \{(1, 1)\}$ En plus de la condition $\sum_i a_i = \sum_j b_j = N$, il existe alors une deuxième condition d'existence : $a_1 + b_1 \leqslant N$ (nous remplaçons ici n par N, n étant utilisée comme indice des termes d'une suite).

L'étude de l'algorithme dans ce cas sera très proche de l'étude précédente.

On montrera d'abord que les termes $t_{1j}^{(n)}$ et $t_{i1}^{(n)}$ ne peuvent pas devenir très petits, et que, *sous la condition* $a_1 + b_1 < N$, il en est de même des termes $t_{ij}^{(n)}$ quel que soit $(i, j) \in R'$ en notant R' l'ensemble des couples (i, j) où $i \neq 1$ et $j \neq 1$.

Les notations sont les mêmes que plus haut, sauf : $t_{ij}^{(0)} \geqslant \alpha$ seulement pour tout $(i, j) \in R$. La relation (3) est toujours vraie, les relations (5) et (6) deviennent :

$$\nu_j^{(p)} \geqslant \frac{b_j}{A \sum_{i \in R_{\cdot j}} \mu_i^{(p)}} \geqslant \frac{b_j}{A \sum_i \mu_i^{(p)}} \tag{5'}$$

$$\nu_j^{(p)} \leqslant \frac{b_j}{\alpha \sum_{i \in R_{\cdot j}} \mu_i^{(p)}} \tag{6'}$$

La relation (7) est toujours valable pour $(i, j) \in R$; de (6') et du fait que $t_{11}^{(n)}$ est nul quel que soit n, on déduit

$$\sum_j t_{1j}^{(p)} = \mu_1^{(p)} \sum_j t_{1j}^{(0)} \quad \nu_j^{(p)} \leqslant \mu_1^{(p)} \ A \sum_{j>1} \frac{b_j}{\alpha \sum_i \mu_i^{(p)}}$$

(On remarquera que pour $j \geqslant 1$, $R._j = \overline{R}$)
d'où :

$$\sum_j t_{ij}^{(p)} \leqslant \frac{A \sum_j b_j}{\alpha} \frac{\mu_i^{(p)}}{\sum_i \mu_i^{(p)}}$$

soit :

$$\frac{\mu_i^{(p)}}{\sum_i \mu_i^{(p)}} \geqslant \frac{\alpha \sum_j t_{ij}^{(p)}}{A \sum_j b_j} \geqslant \frac{\alpha d}{AN}$$ quantité indépendante de n et strictement positive.

En portant dans (7) pour $i = 1$, on voit que pour tout p et pour $j > 1$ il existe un nombre a' strictement positif tel que $t_{1j}^{(p)}$ reste supérieur à a'.

On a ensuite :

$$t_{1j}^{(p+1)} = t_{1j}^{(p)} \frac{a_i}{\sum_j t_{ij}^{(p)}} \geqslant \frac{a' \inf_i a_i}{N} > 0$$

D'où, $\exists\, a > 0$ tel que $\forall\, n$ et $\forall j > 1$ $\quad t_{1j}^{(n)} \geqslant a$ (8)

On démontrera de même :

$$\exists\, b > 0 \text{ tel que } \forall n \text{ et } \forall i > 1 \quad t_{i1}^{(n)} \geqslant b. \tag{9}$$

Étudions maintenant le comportement des suites $t_{ij}^{(n)}$ pour $(i, j) \in R'$.
On montrera la propriété suivante :
$\exists\, \varepsilon > 0$ tel que

$$(\forall (i,j) \in R',\ t_{ij}^{(n-1)} < \varepsilon \text{ et } t_{ij}^{(n)} < \varepsilon) \Longrightarrow \forall (i,j) \in R',\ t_{ij}^{(n+1)} > t_{ij}^{(n)}$$

On supposera d'abord n pair. $\sum_j t_{ij}^{(p-1)} = a_i$ d'où :

$\forall\, \varepsilon_1 > 0$, $\exists\, \varepsilon > 0$ tel que

$$\forall (i,j) \in R' \quad t_{ij}^{(p-1)} < \varepsilon \tag{10}$$

$$\Downarrow$$

$$\forall\, i > 1 \quad a_i - \varepsilon_1 < t_{i1}^{(p-1)} \leqslant a_i \tag{11}$$

De là : $\forall\, \varepsilon_2 > 0$, $\exists\, \varepsilon_1 > 0$,

$$(11) \Longrightarrow a_2 + \ldots + a_l - \varepsilon_2 < \sum_{i>1} t_{i_1}^{(p-1)} \leqslant a_2 + \ldots + a_l \tag{12}$$

Donc : $\forall \varepsilon_2 > 0$, $\exists\, \varepsilon > 0$, $\quad (10) \Longrightarrow (12)$.

Par ailleurs $t_{i1}^{(p)} = t_{i1}^{(p-1)} \frac{b_i}{\sum_{i>1} t_{i_1}^{(p-1)}}$ et par continuité :

$\forall \varepsilon_3 > 0$, $\exists\, \varepsilon > 0$,

$$(10) \Longrightarrow \frac{b_1 a_i}{a_2 + \ldots + a_l} - \varepsilon_3 < t_{i1}^{(p)} < \frac{b_1 a_i}{a_2 + \ldots + a_l} + \varepsilon_3 \tag{13}$$

et enfin :

$$\frac{\sum_j t_{ij}^{(p)}}{a_i} = \frac{t_{i_1}^{(p)} + \sum_{j>1} t_{ij}^{(p)}}{a_i}$$

donc : $\forall \varepsilon_4 > 0,\ \exists\ \varepsilon > 0,$

$$(\forall (i,j) \in R' \quad t_{ij}^{(p-1)} < \varepsilon \quad \text{et} \quad t_{ij}^{(p)} < \varepsilon)$$

$$\Downarrow$$

$$\forall\, i > 1 \quad , \quad \frac{b_1}{a_2 + \ldots + a_I} - \varepsilon_4 < \frac{\sum_j t_{ij}^{(p)}}{a_i} < \frac{b_1}{a_2 + \ldots + a_I} + \varepsilon_4 \qquad (14)$$

Ecrivons maintenant la *condition d'existence* d'un tableau de marges a_i et b_j possédant un zéro dans la casse (1,1) :

$$a_1 + b_1 < N$$

qui peut s'écrire :

$$\frac{b_1}{a_2 + \ldots + a_I} = 1 - \beta, \quad \beta \text{ fixé strictement positif,}$$

ce qui donne en rapportant dans (14) :

$\forall \varepsilon_4 > 0\ \exists\ \varepsilon > 0$, tel que

$$(\forall\ (i,j) \in R' \quad t_{ij}^{(p-1)} < \varepsilon \quad \text{et} \quad t_{ij}^{(p)} < \varepsilon)$$

$$\Downarrow$$

$$\forall\, i > 1 \qquad 1 - \beta - \varepsilon_4 < \frac{\sum_j t_{ij}^{(p)}}{a_i} < 1 - \beta + \varepsilon_4 \qquad (15)$$

On peut choisir $\varepsilon_4 < \beta$ soit $1 - \beta + \varepsilon_4 < 1$, alors

$$(15) \quad \Longrightarrow \quad \forall\, i > 1 \quad , \quad \frac{a_i}{\sum_j t_{ij}^{(p)}} > 1$$

et donc :

$$t_{ij}^{(p+1)} > t_{ij}^{(p)}$$

pour tout $i > 1$ et tout $j \in R^*$, soit, en particulier, pour tout $(i,j) \in R'$.

Un raisonnement analogue peut être fait en changeant p en $p+1$. Finalement on obtient la propriété annoncée pour n quelconque. On en déduit que les suites $t_{ij}^{(n)}$, $(i,j) \in R'$, ne peuvent pas toutes converger vers zéro : en effet, si elles convergent toutes vers zéro

$$\forall\, \varepsilon > 0\ \exists\ n_\varepsilon \ , \quad n > n_\varepsilon \Longrightarrow \forall (i,j) \in R' \quad t_{ij}^{(n)} < \varepsilon$$

prenant justement pour ε celui qui est défini plus haut, on en déduit un n_ε bien déterminé; si l'on pose $n_0 = n_\varepsilon + 2$, on aura pour tout $n \geqslant n_0$, $t_{ij}^{(n+1)} > t_{ij}^{(n)}$ pour tout $(i,j) \in R'$. Or pour n_0 donné fini $t_{ij}^{(n_0)}$ est différent de zéro si $t_{ij}^{(0)}$ est différent de zéro (c'est le cas pour $(i,j) \in R'$), et les suites $t_{ij}^{(n)}$, minorées à partir de $n = n_0$ par $t_{ij}^{(n_0)}$ ne peuvent converger vers zéro, d'où la contradiction. Ceci entraîne l'existence d'au moins une suite $t_{i_0 j}^{(n)}$, $(i_0, j_0) \in R'$, qui ne converge pas vers zéro, d'où l'existence d'un nombre t strictement positif et d'une suite partielle infinie extraite de

la suite $t^{(n)}_{io\,jo}$, tels que :

$$\forall\, m \in \mathrm{M} \quad . \quad t^{(m)}_{iojo} > t \tag{16}$$

(en désignant par M l'ensemble des indices des termes de la suite partielle définie ci-dessus).

Rassemblons les résultats précédents; en utilisant (4) et les inégalités (8), (9), (16) et en remarquant que les $t^{(n)}_{ij}$ sont tous bornés supérieurement (par exemple par N), il existe des nombres a', a'', b', b'', t', t'' strictement positifs et finis tels que, pour tout m appartenant à M l'on a :

$$a' \leqslant \mu_1^{(m)}\, \nu_j^{(m)} \leqslant a'' \qquad \text{pour tout } j > 1 \tag{17}$$

$$b' \leqslant \mu_i^{(m)}\, \nu_1^{(m)} \leqslant b'' \qquad \text{pour tout } i > 1 \tag{18}$$

$$t' \leqslant \mu_{i_0}^{(m)}\, \nu_{j_0}^{(m)} \leqslant t'' \qquad \text{pour un couple } (i_0, j_0) \in \mathrm{R}' \tag{19}$$

(17) et (19) entraînent l'existence de α' et α'' tels que :

$$0 < \alpha' \leqslant \frac{\mu_{i_0}^{(m)}}{\mu_1^{(m)}} \leqslant \alpha''$$

et (18) entraîne l'existence de nombres β' et β'' tels que :

$$0 < \beta' \leqslant \frac{\mu_i^{(m)}}{\mu_h^{(m)}} \leqslant \beta'' \text{ pour tout } i \text{ et } h \text{ supérieurs à 1,}$$

l'un de ces indices pouvant être en particulier i_0 ; on arrive donc à l'existence de γ' et γ'' tels que :

$$\forall\, i \in \bar{\mathrm{R}} \text{ et } \forall\, h \in \bar{\mathrm{R}}\ ,\ 0 < \gamma' \leqslant \frac{\mu_i^{(m)}}{\mu_h^{(m)}} \leqslant \gamma''$$

de là, l'on conclut à l'existence de η' et η'' tels que :

$$\forall\, i \in \bar{\mathrm{R}} \text{ et } \forall\, m \in \mathrm{M} \qquad 0 < \eta_i' \leqslant \frac{\mu_i^{(m)}}{\sum_i \mu_i^{(m)}} \leqslant \eta_i''$$

et par une méthode analogue :

$$\forall\, j \in \mathrm{R}^* \text{ et } \forall m \in \mathrm{M} \qquad 0 < \eta_1' \leqslant \frac{\nu_j^{(m)}}{\sum_j \nu_j^{(m)}} \leqslant \eta_i''$$

Reprenant alors l'inégalité (7)

$$t_{ij}^{(p)} \geqslant \frac{\alpha\, b_j}{\mathrm{A}} \frac{\mu_i^{(p)}}{\sum_i \mu_i^{(p)}} \quad \text{pour } m \text{ pair}$$

et l'inégalité analogue obtenue en permutant les rôles des lignes et des colonnes :

$$t_{ij}^{(p+1)} \geqslant \frac{\alpha\, a_i}{\mathrm{A}} \frac{\nu_j^{(p+1)}}{\sum_j \nu_j^{(p+1)}} \quad \text{pour } m \text{ impair}$$

on obtiendra :

$$\forall m \in \mathrm{M} \text{ et } \forall (i, j) \in \mathrm{R} \qquad \exists\, a > 0,\ t_{ij}^{(m)} \geqslant a$$

Utilisant cette propriété des suites $t_{ij}^{(n)}$, la démonstration de convergence de l'algorithme R.A.S. sera achevée par un argument voisin de celui que l'on a utilisé dans le premier cas où aucun des $t_{ij}^{(0)}$ n'était nul.

Remarquons tout d'abord que l'inégalité $t_{ij}^{(m)} \geqslant a$ sera vérifiée pour tout $(i, j) \in R$ pour une infinité d'indices m pairs; en effet, ou bien M contient une infinité de tels indices, ou bien M contient une infinité d'entiers impairs; mais dans ce cas, il suffit d'utiliser (1) pour voir que $t_{ij}^{(m)} > a$ pour tout $(i, j) \in R$ entraîne l'existence de $a' > 0$ tel que $t_{ij}^{(m+1)} \geqslant a'$ pour tout $(i, j) \in R$. Soit alors $\| t_{ij}^{(q)} \|$ la suite partielle infinie extraite de $\| t_{ij}^{(n)} \|$ et telle que q est pair et $t_{ij}^{(q)} \geqslant a > 0$ pour tout $(i, j) \in R$; nous appellerons Q la famille des indices de cette sous-suite.

Pour $q \in Q$ fixé, désignons par k, $k \in \overline{R}$, l'indice du plus grand des l nombres $\dfrac{a_i}{\sum\limits_j t_{ij}^{(q)}}$; l'on a comme dans le premier cas, en posant $L - l = \delta$:

$$\forall \varepsilon > 0 \quad \exists q_0 \ , \ q > q_0 \implies \sum_i t_{ij}^{(q+1)} > l b_j - \varepsilon b_j + t_{kj}^{(q)} \cdot \delta$$

d'où

$$s'_q = \inf_j \frac{\sum\limits_i t_{ij}^{(q+1)}}{b_j} > l - \varepsilon + \delta \inf_j t_{kj}^{(q)}$$

Supposant $\delta \neq 0$ et $k \neq 1$ on a $\inf\limits_j t_{kj}^{(q)} \geqslant a$ et on peut choisir ε tel que $a\,\delta - \varepsilon > 0$, ce qui est en contradiction avec le fait que les termes de la suite s'_q restent inférieurs à l : donc, soit $\delta = 0$ et la convergence de l'algorithme est montrée, soit $k = 1$ (car alors $\inf\limits_j t_{1j}^{(q)} = t_{11}^{(q)} = 0$).

L'indice de la plus grande des quantités $\dfrac{a_i}{\sum\limits_j t_{ij}^{(q)}}$ serait donc $i = 1$. Mais par un raisonnement analogue, l'on montre que 1 est aussi l'indice de la plus petite de ces quantités. Dans ces conditions, il existe une suite partielle infinie S_q extraite de S_p et une suite partielle infinie s_q extraite de s_p qui ont même limite; puisque les limites respectives de S_q et s_q sont L et l, on en déduit $L = l$.

Enfin, comme il a été vu plus haut $\sum\limits_i a_i = \sum\limits_j b_j$ entraîne :

$$L = l = 1$$

Remarque. — Pour démontrer que les $t_{ij}^{(m)}$ ne devenaient pas très petits lorsque m était grand, l'on a écrit la condition $a_1 + b_1 < N$, écartant le cas $a_1 + b_1 = N$ où, pourtant, un tableau répondant aux conditions imposées existe; mais alors, il n'existe qu'un seul tableau répondant à la question et tous ses termes sont nuls pour $(i, j) \in R'$; si l'algorithme R.A.S. converge, il est donc nécessaire que $t_{ij}^{(n)}$ tende vers zéro pour $(i, j) \in R'$. C'est ce qui se passe aussi lorsque la condition d'existence est violée, l'algorithme a ten-

dance à « vider » une partie du tableau pour approcher autant que faire se peut les conditions (irréalisables) imposées.

Nous terminerons par une note bibliographique sur l'algorthime que l'on vient d'étudier. Il semble avoir été introduit par DEMING et STEPHAN (1940), de façon assez curieuse, à la suite d'une confusion; celle-ci est notée par STEPHAN (1942) qui étudie surtout un processus « linéaire » assez voisin; STEPHAN affirme en outre avoir prouvé la convergence de l'algorithme étudié plus haut, mais ne donne aucune indication sur cette démonstration. Toutefois, il est très vraisemblable qu'elle ne recouvre pas le cas où certains des $t_{ij}^{(0)}$ sont nuls, puisque celui-ci est toujours écarté dans les articles que nous venons de citer.

BIBLIOGRAPHIE

AITCHISON, J. (1962). Large sample restricted parametric tests. J. R. Stat. Soc. B, 24, p. 234-250.

ASANO, C. (1965). On estimating multinomial probabilities by pooling incomplete samples. Ann. Inst. Stat. Math. 17, p. 1-13.

BARTON, D. E., DAVID, F. N. et FIX, E. (1962). Persistence in a chain of multiples events. Biometrika, 49, p. 351-357.

BATSCHELET, E. (1960 a). Uber eine Kontingenztafel mit fehlenden Daten. Biom. Zeit., 2, p. 236-243.

BATSCHELET, E. (1960 b). Auslesefreie Verteilung des Manifestations alters mit einer Anwendung auf die Respirationsatopien. Biom. Zeit., 2, p. 244-256.

BHAPKAR, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical Data. J. Amer. Stat. Ass., 61, p. 228-235.

BIRCH, M. W. (1963). Maximum likelihood in three-way contingency tables. J. R. Stat. Soc. B, 25, p. 220-233.

BOWKER, A. H. (1948). A test of symetry in contingency tables. J. Amer. Stat. Ass., 43, p. 572-574.

CAUSSINUS, H. (1962 a). Sur certaines généralisations de l'emploi du test du χ^2. C. R. Acad. Sc., t. 254, p. 3306-3308.

CAUSSINUS, H. (1962 b). Sur un problème d'analyse de la corrélation de deux caractères qualitatifs. C. R. Acad. Sc., t. 255, p. 1688-90.

CAUSSINUS, H. (1965). Sur les tables de contingence tronquées. C. R. Acad. Sc., t. 261, p. 5303-5306.

CAUSSINUS, H. (1966 a). Remarques sur les problèmes d'estimation et de tests dans les tables de contingence tronquées. C. R. Acad. Sc., t. 262 A, p. 293-295.

CAUSSINUS, H. (1966 b). Sur l'analyse de certaines tables de contingence. C. R. Acad. Sc., t. 263 A, p. 551-554.

CAUSSINUS, H. (1966 c). Sur la structure des tableaux de corrélation carrés. C. R. Acad. Sc., t. 263 A, p. 795-797.

COCHRAN, W. G. (1952). The χ^2 test of goodness of fit. Ann. Math. Stat. 23, p. 315-345.

COCHRAN, W. G. (1954). Some methods for strengthening the common χ^2 tests. Biometrics, 10, p. 417-451.

CRAMER, H. (1946). Mathematical Methods of statistics, Princeton University Press.

DARROCH, J. N. (1962). Interactions in multi-factor contingency tables. J. R. Stat. Soc. B, 24, p. 251-263.

DELTHEIL, R. et HURON, R. (1959). Statistique mathématique, Collection Armand Colin, Paris.

DEMING, W. E. et STEPHAN, F. F. (1940). On a least square adjustement of a sample frequency table when the expected marginal totals are known. Ann. Math. Stat., 11, p. 427-444.

DIAMOND, E. L., MITRA, S. K. et ROY, S. N. (1960). Asymptotic power and asymptotic independence in the statistical analysis of categorical data. Bull. Inst. Int. Stat., 37, p. 309-329.

DUGUE, D. (1958). Traité de Statistique théorique et appliquée, Masson et C^{ie}, Paris.

FERGUSON, T. S. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. Ann. Math. Stat., 29, p. 1046-1062.

FIX, E., HODGES, J. L. et LEHMANN (1959). The restricted chi-square test, In Probability and Statistics, the Harald Cramer Volume, édité par U. Grenander, Wiley, New-York.

FRECHET, M. (1959). Sur les tableaux de corrélation dont les marges et des cases vides sont données. C. R. Acad. Sc., 249, p. 592.

FRECHET, M. (1960). Sur les tableaux dont les marges et les bornes sont données. Rev. Inst. Int. Stat., 28, p. 10-32.

FRIEDLANDER, D. (1961). Technique for estimating a contingency table given the marginal totals and some supplementary data. J. R. Stat. Soc. A., 124, p. 412-420.

GEPPERT, M. P. (1961). Erwartungstreue plausibelste Schätzer aus dreieckig gestutzten Kontingenztafeln, Biom. Zeit., 3, p. 54-67.

GIL, B. et OMABOE, E. N. (1963). Internal migration differentials from conventionnal census questionnaire items in Ghana. Bull. Inst. Int. Stat. Actes de la 34e Session. Tome XL, 1re livraison, p. 431-446.

GOODMAN, L. A. (1963 a). On Plackett's test for contingency table interactions. J. R. Stat. Soc. B, 25, p. 179-188.

GOODMAN, L. A. (1963 b). On methods for comparing contingency tables. J. R. Stat. Soc. A, 126, p. 94-108.

GOODMAN, L. A. (1964 a). Simultaneous confidence limits for crossproduct ratios in contingency tables. J. R. Stat. Soc. B, 26, p. 86-102.

GOODMAN, L. A. (1964 b). Simple methods for analyzing three-factor interaction in contingency tables. J. Amer. Stat. Ass., 59, p. 319-352.

GOODMAN, L. A. (1964 c). The analysis of persistence in a chain of multiple events. Biometrika, 51, p. 405-411.

GOODMAN, L. A. et KRUSKAL, W. H. (1954). Mesures of association for cross classifications (I). J. Amer. Stat. Ass., 49, p. 732-764.

GOODMAN, L. A. et KRUSKAL, W. H. (1959). Mesures of association for cross classifications. II : Further discussion and references. J. Amer. Stat. Ass., 54, p. 123-163.

GOODMAN, L. A. et KRUSKAL, W. H. (1963). Mesures of association fort cross-classifications. III : approximate sampling theory. J. Amer. Stat. Ass., 58, p. 310-364.

HARRIS, A. J. et TRELOAR, A. E. (1927). On a limitation in the applicability of the contingency coefficient. J. Amer. Stat. Ass., 22, p. 460-472.

HARRIS, A. J. et CHI TU (1929). A second category of limitations in the applicability of the contingency coefficient. J. Amer. Stat. Ass., 24, p. 367-375.

HARRIS, A. J., TRELOAR, A. E. and WILDER, M. (1930). Professor Pearson's note on ours papers on contingency. J. Amer. Stat. Ass., 25, p. 323-327.

IRWIN, J. O. (1949). A note on the subdivision of χ^2 into components. Biometrika, 36, p. 130-134.

ISHII, G. (1960). Intraclass contingency tables. Ann. Inst. Statist. Math. Tokyo, 12, p. 161-207.

KALE (1962). On the solution of likelihood equations by iteration processes. The multiparametric case. Biometrika, 49, p. 479-486.

KASTENBAUM, M. A. (1958). Estimation of relative frequencies of four sperm types in Drosophila Melanogaster. Biometrics, 14, p. 223-228.

KENDALL, M. G. et STUART, A. (1958). The advanced theory of statistics, Volume 1 : distribution theory. Ch. Griffin et Cie, Londres.

KENDALL, M. G. et STUART, A. (1961). The advanced theory of statistics, Volume 2 : inference and relationship. Ch. Griffin et Cie, Londres.

LANCASTER (1949). The derivation and partition of χ^2 in certain discrete distributions. Biometrika, 36, p. 117-129.

LEHMANN, E. L. (1959). Testing statistical hypotheses, Wiley, New-York.

LEWONTIN et FALSENSTEIN (1965). The robustness of homogeneity tests in $2 \times n$ tables. Biometrics, 21, p. 19.

LOKKI, O. (1961). An application of the common chi-squared test. Soc. Sci Fenn., Comm-Physico-Math., 26, 8, p. 1-10.

NEYMAN, J. (1949). Contribution to the theory of χ^2 test. Proceedings of the first Berkeley Symposium, p. 239-273.

OKAMOTO et ISHII (1961). Test of independance in intraclass 2×2 tables. Biometrika, 45, p. 181-190.

PEARSON, K. (1930). On the theory of contingency. I. Note on Professor J. Arthur Harris' paper on the limitation in the applicability of the contingency coefficient. J. Amer. Stat. Ass. 25, p. 320-323.

PLACKETT, R. L. (1962). A note on interactions in contingency tables. J. R. Stat. Soc. B, 24, p. 162-166.

PRESSAT, R. (1963). L'attraction dans les migrations intérieures. Bull. Inst. Intern. Stat. (Actes de la 34e Session), tome XL, 1re livraison, p. 450-460.

ROY, S. N. et KASTENBAUM, M. A. (1956). On the hypothesis of « no interaction » in a multi-way contingency table. Ann. Math. Stat., 27, p. 749-757.

SCHEFFE (1959). The analysis of variance. Wiley, New-York.

STEPHAN, F. F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. Ann. Math. Stat., 13, p. 166-178.

STUART, A. (1953). The estimation and comparison of strengths of assosiation in contingency tables. Biometrika, 40, p. 105-110.

STUART, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. Biometrika, 42, p. 412-416.

TALLIS, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. Biometrics, 18, p. 342-353.

THIONET, P. (1963). Sur certaines variantes des projections du tableau d'échanges interindustriels, Bull. Inst. Inter. Stat., tome XL, 1re livraison, p. 431-446.

THIONET, P. (1964). Note sur le remplissage d'un tableau à double entrée. Journal de la Société de Statistique de Paris, n° 10-11-12, p. 228-247.

TOCHER, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates. Biometrika, 37, p. 130-144.

WAITE, H. (1915). Association of finger-prints. Biometrika, 10, p. 421-478.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Amer. Math. Soc., 54, 426-482. Réimprimé dans : Selected papers in statistics and probability by Abraham Wald.

WATSON, G. S. (1956). Missing and « Mixed-up » frequencies in contingency tables. Biometrics, 12, p. 47-50.

TABLE DES MATIÈRES