ANNALES DE LA FACULTÉ DES SCIENCES DE TOULOUSE

LEO A. GOODMAN

Contributions to the statistical analysis of contingency tables : notes on quasi-symmetry, quasi-independence, log-linear models, log-bilinear models, and correspondence analysis models

Annales de la faculté des sciences de Toulouse 6^e série, tome 11, nº 4 (2002), p. 525-540

<http://www.numdam.org/item?id=AFST_2002_6_11_4_525_0>

© Université Paul Sabatier, 2002, tous droits réservés.

L'accès aux archives de la revue « Annales de la faculté des sciences de Toulouse » (http://picard.ups-tlse.fr/~annales/) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

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

Annales de la Faculté des Sciences de Toulouse Vol. XI, n° 4, 2002
pp. 525–540

Contributions to the statistical analysis of contingency tables: Notes on quasi-symmetry, quasi-independence, log-linear models, log-bilinear models, and correspondence analysis models [*]

LEO A. GOODMAN [1]

RÉSUMÉ. — Cet article commence par comparer le concept de quasi-symétrie dû à Henri Caussinus avec un concept parent mais différent que j'appellerai la quasi-symétrie de Karl Pearson ; et nous trouverons que le concept de Caussinus est préférable et plus utile que celui de Pearson. J'introduirais alors un ensemble de modèles log-bilinéaires quasi symétriques qui sont plus parcimonieux que le modèle de quasi-symétrie de Caussinus, et un ensemble de modèles d'analyse des correspondances quasi symétriques qui sont plus parcimonieux que le modèle correspondant de Pearson ; et nous établirons que les modèles log-bilinéaires quasi symétriques sont préférables et plus utiles que leurs vis-à-vis, les modèles d'analyse des correspondances quasi symétriques. Notre attention est concentrée dans cet article sur le modèle de quasi-symétrie de Caussinus et sur les modèles qui lui sont apparentés. En addition aux modèles mentionnés ci-dessus, nous commenterons brièvement les modèles de quasi-indépendance, et verrons comment les modèles de quasi-indépendance et de quasi-symétrie peuvent être considérés comme des précurseurs directs des modèles log-linéaires et log-bilinéaires. Finalement nous étudierons la relation entre quasi-indépendance et quasi-symétrie ; et je citerai ici, pour les lecteurs qui pourraient être intéressés, une série particulière d'articles concernant la quasi-indépendance, publiés pendant une période de trente trois ans commençant en 1961.

ABSTRACT. — This paper begins by comparing Henri Caussinus' concept of quasi-symmetry with a related but different concept that I shall call Karl Pearson's quasi-symmetry; and we shall find that Caussinus' concept is preferable to and more useful than Pearson's concept. I shall then introduce a set of quasi-symmetric log-bilinear models that are more

[*] Reçu le 18 septembre 2001, accepté le 18 septembre 2002

[1] Departments of Statistics and Sociology, University of California at Berkeley, Berkeley, California 94720, U.S.A. Email: lgoodman@socrates.berkeley.edu

parsimonious than Caussinus' quasi-symmetry model, and a set of quasi-symmetric correspondence analysis models that are more parsimonious than Pearson's quasi-symmetry model; and we shall find that the quasi-symmetric log-bilinear models are preferable to and more useful than the corresponding quasi-symmetric correspondence analysis models.

Our main focus of attention in this paper is on Caussinus' quasi-symmetry model and on other models related to it. In addition to the other models referred to above, we shall also comment briefly in this paper on the quasi-independence model, and on how the quasi-independence and quasi-symmetry models can be viewed as direct precursers of the log-linear models and the log-bilinear models. Finally, we shall comment on the relationship between quasi-independence and quasi-symmetry; and I shall also include here, for those readers who may be interested, citations to a particular series of articles pertaining to quasi-independence, published over a thirty-three year period beginning in 1961.

1. Introductory comments

The concept of quasi-symmetry introduced by Henri Caussinus has been and continues to be a major contribution to the statistical analysis of square contingency tables (in which there is a one-to-one correspondence between the row and column categories). This is well known. But I wonder whether those who are familiar with Caussinus' concept are aware of the fact that it is, in a certain sense, "infinitely better" than an alternative concept of "quasi-symmetry". The alternative concept can be viewed as an expression for "quasi-symmetry" obtained with the correspondence analysis approach and/or with an approach based on Karl Pearson's perspective. The present paper will explain why Caussinus' concept is infinitely better than the alternative.

I shall also introduce here a set of models that can be viewed as special cases of Caussinus' quasi-symmetry model and that are more parsimonious than his quasi-symmetry model. In addition, I shall introduce a set of models that can be viewed as special cases of the alternative model of "quasi-symmetry" and that are more parsimonious than this alternative model. We shall also find here that each model in the set of models that are special cases of Caussinus' quasi-symmetry model are, in a certain sense, infinitely better than the corresponding model in the set of models that are special cases of the alternative model of quasi-symmetry.

In addition to the various models referred to above, which are related, in one way or another, to Caussinus' quasi-symmetry model, we shall also consider briefly the quasi-independence model; and we shall comment here on how the quasi-independence and quasi-symmetry models can be viewed as direct precursers of the log-linear approach and the log-bilinear approach. We shall also comment on the relationship between quasi-independence and quasi-symmetry. This relationship was also explored by Caussinus [6].

2. Caussinus' quasi-symmetry, symmetric association, Pearson's quasi-symmetry, and symmetric contingency

We shall show in this section how Caussinus' concept of quasi-symmetry is related to a somewhat different concept which I shall call Pearson's quasi-symmetry. But first some necessary notation:

For the square $I \times I$ contingency table, let P_{ij} denote the probability that an observation will fall in the i-th row and j-th column of the table. The Caussinus [6] concept of quasi-symmetry states that P_{ij} can be expressed as

$$P_{ij} = \alpha_i \beta_j \gamma_{ij}, \quad \text{with } \gamma_{ij} = \gamma_{ji}, \tag{1}$$

when there is a one-to-one correspondence between the i-th row category and the i-th column category (for $i = 1, \ldots, I$), and $\alpha_i \geqslant 0$, $\beta_j \geqslant 0$, $\gamma_{ij} \geqslant 0$ (for $i = 1, \ldots, I$; and $j = 1, \ldots, I$). When $\gamma_{ij} = \gamma$ (for $i = 1, \ldots, I$; and $j = 1, \ldots, I$), we can rewrite (1) as

$$P_{ij} = \alpha_i \beta_j \gamma, \tag{2}$$

which states that the row variable (say, variable A) and the column variable (say, variable B) are statistically independent of each other. We can thus view γ_{ij} in (1) as a measure of a particular kind of "nonindependence" or "association"; and the quasi-symmetry model (1) can thus be called a model of "symmetric nonindependence" or "symmetric association" (see Goodman [28]).

Let us now consider a different measure of nonindependence based on Karl Pearson's coefficient of "mean squared contingency":

$$G^2 = \sum_i \sum_j (P_{ij} - P_i^A P_j^B)^2 / (P_i^A P_j^B), \tag{3}$$

where
$$P_i^A = \sum_j P_{ij} = P_{i+}, \quad P_j^B = \sum_i P_{ij} = P_{+j}.$$

From this definition of "mean squared contingency", we can see from (3) that Pearson's measure of "contingency" was

$$C_{ij} = (P_{ij} - P_i^A P_j^B) / (P_i^A P_j^B). \tag{4}$$

(The distribution used in calculating the mean of the squared contingency here is $P_i^A P_j^B$ (for $i = 1, \ldots, I$; and $j = 1, \ldots, I$).) When $C_{ij} = 0$ (for

$i = 1, \ldots, I$; and $j = 1, \ldots, I$), we see again that the row variable A and the column variable B are statistically independent of each other. We can thus view C_{ij} in (4) also as a measure of "nonindependence" or "contingency"; and, analogous to the quasi-symmetry condition that $\gamma_{ij} = \gamma_{ji}$ in (1), I shall call the corresponding condition that $C_{ij} = C_{ji}$ Pearson's condition of quasi-symmetry or Pearson's "symmetric nonindependence" or "symmetric contingency". From (4) we see that the model for Pearson's quasi-symmetry can be expressed as

$$P_{ij} = P_i^A P_j^B D_{ij}, \quad \text{with } D_{ij} = D_{ji}, \tag{5}$$

where $D_{ij} = C_{ij} + 1$ (for $i = 1, \ldots, I$; and $j = 1, \ldots, I$). Note that D_{ij} is simply $P_{ij}/(P_i^A P_j^B)$, which I shall call "Pearson's ratio".

Let us now compare Pearson's quasi-symmetry (5) with Caussinus' quasi-symmetry (1). From (5) we see that the following set of equations will be satisfied:

$$\sum_i D_{ij} P_i^A = \sum_i P_{ij}/P_j^B = 1, \quad \sum_j D_{ij} P_j^B = \sum_j P_{ij}/P_i^A = 1, \tag{6}$$

in addition to the usual equations

$$\sum_i P_i^A = 1, \quad \sum_j P_j^B = 1. \tag{7}$$

When $D_{ij} = D_{ji}$ in (5), we see from (6) and (7) that the P_i^A and the P_j^B will satisfy the same set of equations; and thus the condition that $D_{ij} = D_{ji}$ in (5) implies that $P_i^A = P_i^B$ (for $i = 1, \ldots, I$) when the contingency table is irreducible. (A contingency table is irreducible if no two rows have the same conditional distributions — i.e., for all pairs of rows, say, rows i and i', with $i \neq i'$, we do not have $P_{ij}/P_i^A = P_{i'j}/P_{i'}^A$, for all $j = 1, \ldots, I$ — *and* if no two columns have the same conditional distributions.) Thus, Pearson's quasi-symmetry in an irreducible contingency table implies that $P_i^A = P_i^B$ (for $i = 1, \ldots, I$); i.e., that the row marginal and the column marginal are homogeneous. Since the row and column marginals are homogeneous in this case, we find that Pearson's quasi-symmetry in this case implies also that the contingency table is symmetric; i.e., that $P_{ij} = P_{ji}$ (for $i = 1, \ldots, I$; $j = 1, \ldots, I$). The fact that Pearson's quasi-symmetry in this case implies symmetry is a serious limitation of this concept of quasi-symmetry. A somewhat related kind of limitation arises also when the contingency table is reducible (i.e., when the table is not irreducible); see Goodman [34]. However, these kinds of limitations do *not* arise when Pearson's quasi-symmetry is replaced by Caussinus' quasi-symmetry.

Caussinus' quasi-symmetry model can hold true in contingency tables in which the row and column marginals are not homogeneous (and also in tables in which the row and column marginals are homogeneous); while Pearson's quasi-symmetry model can not hold true in irreducible contingency tables in which the row and column marginals are not homogeneous. As we noted in the preceding paragraph, if Pearson's quasi-symmetry model holds true in an irreducible contingency table, then the row and columns marginals in the table must be homogeneous and the table must be symmetric (and a somewhat related kind of limitation applies also when the contingency table is reducible). When the usual symmetry model does not hold true and/or the row and column marginals are not homogeneous, then we can examine whether the Caussinus' quasi-symmetry model holds true; but there is no point in examining whether Pearson's quasi-symmetry model holds true. Under these circumstances, Pearson's concept is of no use; and Caussinus' concept is, in this sense, "infinitely better" than Pearson's concept.

3. Caussinus' quasi-symmetry, symmetric association models, Pearson's quasi-symmetry, and symmetric contingency models

Let us now consider the RC association model

$$P_{ij} = \alpha_i \beta_j e^{\phi \mu_i \nu_j}, \tag{8}$$

where the μ_i and ν_j are standardized row scores and standardized column scores, respectively, with

$$\sum_i \mu_i / I = 0, \quad \sum_i \mu_i^2 / I = 1, \quad \sum_j \nu_j / I = 0, \quad \sum_j \nu_j^2 / I = 1, \tag{9}$$

and the parameter ϕ in (8) is called the intrinsic association coefficient; see, e.g., Goodman [28]. Without loss of generality, the sign of the μ_i and the sign of the ν_j in (8) can be chosen so that $\mu_1 < 0$ and $\nu_1 < 0$ (or the signs can be chosen so that, for at least one value of i ($i = 1, \ldots, I$), the $\mu_i < 0$ and the $\nu_i < 0$).

Comparing (1) with (8), we see that

$$\log \gamma_{ij} = \phi \mu_i \nu_j \tag{10}$$

under model (8); and we can refer to model (8) as the RC log-bilinear association model. From (9) and (10), we see that, under model (8), Caussinus' quasi-symmetry condition (namely, that $\gamma_{ij} = \gamma_{ji}$) implies that $\mu_i = \nu_i$ (for $i = 1, \ldots, I$) in this model. Thus, quasi-symmetry here, under model (8), implies that the row scores μ_i and the column scores ν_j are homogeneous in this model. I shall, therefore, call the *RC* association model with homogeneous row and column scores the *RC* symmetric association model.

The *RC* symmetric association model is more parsimonious than Caussinus' quasi-symmetry model (when $I \geqslant 3$). The number of degrees of freedom for testing this quasi-symmetry model is $(I-1)(I-2)/2$; and we find here that the number of degrees of freedom for testing the *RC* symmetric association model is $(I-1)(I-2)$. Thus, there are *twice* as many degrees of freedom for testing the *RC* symmetric association model than there are for testing the quasi-symmetry model; and the description of the *RC* symmetric association model uses $(I-1)(I-2)/2$ fewer parameters than are needed to describe the quasi-symmetry model. (Compare (8)–(9) with (1).) Also, if the *RC* symmetric association model holds true, the quasi-symmetry model will also hold true.

Next let us consider the following model which is somewhat analogous to model (8):

$$P_{ij} = P_i^A P_j^B (1 + \rho x_i y_j), \tag{11}$$

where the x_i and y_j are standardized row scores and standardized column scores, respectively, with

$$\sum_i x_i P_i^A = 0, \quad \sum_i x_i^2 P_i^A = 1, \quad \sum_j y_j P_j^B = 0, \quad \sum_j y_j^2 P_j^B = 1. \tag{12}$$

From (11)–(12) we see that

$$\sum_i \sum_j x_i y_j P_{ij} = \rho; \tag{13}$$

and the parameter ρ in (11) is the correlation coefficient. We shall refer to model (11) as the *RC* contingency model (or the *RC* correlation model). Without loss of generality, the sign of the x_i and the sign of the y_i in (11) can be chosen so that $x_1 < 0$ and $y_1 < 0$ (or the signs can be chosen so that, for at least one value of i ($i = 1, \ldots, I$), the $x_i < 0$ and the $y_i < 0$).

Comparing (5) with (11), we see that

$$C_{ij} = \rho x_i y_j, \tag{14}$$

where $C_{ij} = D_{ij} - 1$, under model (11). From (14) we see that, under model (11), Pearson's quasi-symmetry condition (namely, that $C_{ij} = C_{ji}$) implies that the x_i and y_i are proportional (for $i = 1, \ldots, I$) in this model. The *RC* contingency model with proportional row and column scores I shall call the

RC symmetric contingency model. This model is more parsimonious than Pearson's quasi-symmetry model (when $I \geqslant 3$).

We noted earlier herein a serious limitation of Pearson's quasi-symmetry; namely, that Pearson's quasi-symmetry implies symmetry when the contingency table is irreducible, and a somewhat related kind of limitation arises also when the contingency table is reducible. Using the same approach now with the corresponding *RC* symmetric contingency model, we find that this model suffers from the same kind of serious limitations; namely that, if the *RC* symmetric contingency model describes the contingency table, then the usual symmetry model will also describe the contingency table when the table is irreducible, and a somewhat related kind of limitation arises when the contingency table is reducible. However, these kinds of limitations do *not* arise when the *RC* symmetric contingency model is replaced by the *RC* symmetric association model.

4. An example

To illustrate the application of some of the models described in the preceding section, we now consider briefly the analysis of the following 8 × 8 mobility table:

Table 1. — Cross-classification of British male sample according to each subject's occupational status category and his father's occupational status category.

Father's	Subject's status							
status	1	2	3	4	5	6	7	8
1	50	19	26	8	7	11	6	2
2	16	40	34	18	11	20	8	3
3	12	35	65	66	35	88	23	21
4	11	20	58	110	40	183	64	32
5	2	8	12	23	25	46	28	12
6	12	28	102	162	90	554	230	177
7	0	6	19	40	21	158	143	71
8	0	3	14	32	15	126	91	106

These data were studied earlier by Duncan [7], Hauser [37], McCullagh [38], and Goodman [28]. Focusing our attention now only on models pertaining to symmetric association, we present in Table 2 the goodness-of-fit and likelihood-ratio chi-square values obtained when these models are applied to the data in Table 1.

Table 2. — Symmetric association models applied to the data in Table 1 with the main diagonal deleted.

Symmetric association models	Degrees of freedom	Goodness-of-fit chi-square	Likelihood-ratio chi-square
1. Null association	41	555.12	446.84
2. Uniform association	40	55.77	58.44
3. *RC* symmetric association	34	31.21	32.56
4. Quasi-symmetry	21	20.34	22.93

The usual null association model (i.e., the model that states that the row variable and the column variable are statistically independent of each other) and the uniform association can be viewed as special cases of the *RC* symmetric association model; and the *RC* symmetric association model can be viewed as a special case of the quasi-symmetry model. From Table 2 we see that (a) there is a dramatic improvement in fit when the null association model is replaced by any of the three models considered in Table 2 that take into account, in one form or another, the symmetric association; (b) the quasi-symmetry model and the *RC* symmetric association model fit the data well; and (c) there is a dramatic improvement in parsimony when the quasi-symmetry model is replaced by the *RC* symmetric association model.

Because the entries in the eight cells on the main diagonal in Table 1 were deleted in this analysis, the degrees of freedom were reduced by eight for each of the first three models in Table 2 — from 49, 48, and 42 degrees of freedom to 41, 40, and 34, respectively. The number of degrees of freedom is unaffected by the deletion of the main diagonal in the analysis of quasi-symmetry.

5. The $RC(M)$ symmetric association models, the $RC(M)$ symmetric contingency models, and the $RC(M)$ quasi-symmetric correspondence analysis models

Let us now consider the following generalization of the *RC* association model (8):

$$P_{ij} = \alpha_i \beta_j \exp\left[\sum_{k=1}^{M} \phi_k \mu_{ik} \nu_{jk}\right], \tag{15}$$

where $M \leqslant I-1$, and the μ_{ik} and ν_{jk} are standardized row scores and standardized column scores, respectively, with

$$\begin{gathered}\sum_i \mu_{ik}/I = 0, \ \sum_i \mu_{ik}^2/I = 1, \ \sum_j \nu_{jk}/I = 0, \ \sum_j \nu_{jk}^2/I = 1, \\ \sum_i \mu_{ik}\mu_{ik'}/I = 0, \ \sum_j \nu_{jk}\nu_{jk'}/I = 0,\end{gathered} \tag{16}$$

with $k \neq k'$; see, e.g., Goodman [30]. Without loss of generality, the intrinsic association parameters ϕ_k in (15) can be ordered so that $|\phi_1| \geqslant |\phi_2| \geqslant \ldots \geqslant |\phi_M|$; and the sign of the μ_{ik} and the sign of the ν_{jk} in (15) can be chosen so that $\mu_{1k} < 0$ and $\nu_{1k} < 0$, for $k = 1, \ldots, M$ (or the signs can be chosen so that, for each k, there is at least one value of i $(i = 1, \ldots, I)$ with $\mu_{ik} < 0$ and $\nu_{ik} < 0$). Model (15) is called the $RC(M)$ association model; and the $RC(1)$ association model is the same as the RC association model (8).

Comparing (1) with (15), we see that

$$\log \gamma_{ij} = \sum_{k=1}^{M} \phi_k \mu_{ik} \nu_{jk} \tag{17}$$

under model (15). And Caussinus' quasi-symmetry condition (namely, that $\gamma_{ij} = \gamma_{ji}$) implies that

$$\sum_{k=1}^{M} \phi_k \mu_{ik} \nu_{jk} = \sum_{k=1}^{M} \phi_k \mu_{jk} \nu_{ik} \tag{18}$$

(for $i = 1, \ldots, I$; and $j = 1, \ldots, I$) in model (15).

As we did earlier with the RC association model (8), we now introduce the condition that the row scores and the column scores are homogeneous; i.e.,

$$\mu_{ik} = \nu_{ik}, \quad \text{for } i = 1, \ldots, I, \quad \text{and } k = 1, \ldots, M, \tag{19}$$

in model (15). I shall call the $RC(M)$ association model (15), with the row scores and column scores satisfying condition (19), the $RC(M)$ symmetric association model. When condition (19) is satisfied, then condition (18) will also be satisfied, and the association in model (15) is symmetric. In addition, when condition (18) is satisfied, then it is possible to show that condition (19) will also be satisfied.

The $RC(M)$ symmetric association model is more parsimonious than Caussinus' quasi-symmetry model when $M < I - 1$; and the two models are equivalent when $M = I - 1$. We noted earlier that there are $(I-1)(I-2)/2$

degrees of freedom for testing the quasi-symmetry model; and we find here that the number of degrees of freedom for testing the $RC(M)$ symmetric association model is $(I-1)^2 - M(2I-M-1)/2$. (Note that the number of degrees of freedom is $(I-1)(I-2)$ when $M=1$, and the corresponding number of degrees of freedom is $(I-1)(I-2)/2$ when $M = I-1$.)

Next let us consider the following generalization of model (11):

$$P_{ij} = P_i^A P_j^B \left(1 + \sum_{i=1}^{M} \rho_k x_{ik} y_{jk}\right), \tag{20}$$

where $M \leqslant I-1$, and the x_{ik} and y_{jk} are standardized row scores and standardized column scores, respectively, with

$$\begin{gathered} \sum_i x_{ik} P_i^A = 0, \quad \sum_i x_{ik}^2 P_i^A = 1, \quad \sum_j y_{jk} P_j^B = 0, \quad \sum_j y_{jk}^2 P_j^B = 1, \\ \sum_i x_{ik} x_{ik'} P_i^A = 0, \quad \sum_j y_{jk} y_{jk'} P_j^B = 0, \end{gathered} \tag{21}$$

with $k \neq k'$. From (20)–(21), we see that

$$\sum_i \sum_j x_{ik} y_{jk} P_{ij} = \rho_k; \tag{22}$$

and we shall call model (20) the $RC(M)$ contingency model (or the $RC(M)$ correlation model). Without loss of generality, the intrinsic correlation parameters ρ_k can be ordered so that $|\rho_1| \geqslant |\rho_2| \geqslant \ldots \geqslant |\rho_M|$; and the sign of the x_{ik} and the sign of the y_{ik} in (20) can be chosen so that $x_{1k} < 0$ and $y_{1k} < 0$, for $k = 1, \ldots, M$ (or the signs can be chosen so that, for each k, there is at least one value of i ($i = 1, \ldots, I$) with $x_{ik} < 0$ and $y_{ik} < 0$). When $M = 1$, the $RC(M)$ contingency model is the same as the RC contingency model (11); and when $M = I-1$, model (20) can be viewed as equivalent to the basic formula of correspondence analysis (see, e.g., Goodman [30,31]). When $M < I-1$, we have more parsimonious models than the usual correspondence analysis model.

Comparing model (5) with model (20), we see that

$$C_{ij} = \sum_{k=1}^{M} \rho_k x_{ik} y_{jk}, \tag{23}$$

under model (20). From (23) we see that Pearson's quasi-symmetry condition (namely, that $C_{ij} = C_{ji}$) implies that

$$\sum_{k=1}^{M} \rho_k x_{ik} y_{jk} = \sum_{k=1}^{M} \rho_k x_{jk} y_{ik} \tag{24}$$

(for $i = 1, \ldots, I$; and $j = 1, \ldots, I$) in model (20).

As earlier with the *RC* contingency model, we now introduce the condition that the row scores and the column scores are proportional; i.e.,

$$x_{ik} = y_{ik} c_k, \text{ for } i = 1, \ldots, I, \text{ and } k = 1, \ldots, M, \tag{25}$$

in model (20), with $c_k > 0$. I shall call model (20), with the row scores and column scores satisfying condition (25), the $RC(M)$ symmetric contingency model. When the row scores and the column scores in (20) satisfy condition (25), then condition (24) will also be satisfied, and Pearson's contingency in model (20) is symmetric.

Since model (20) can be viewed as equivalent to the basic formula of correspondence analysis when $M = I - 1$, the $RC(M)$ symmetric contingency model in this case can also be viewed as the $RC(M)$ quasi-symmetric correspondence analysis model. The $RC(M)$ quasi-symmetric correspondence analysis model is equivalent to Pearson's quasi-symmetry model when $M = I - 1$; and the former model is more parsimonious than the latter model when $M < I - 1$.

We noted earlier herein serious limitations of Pearson's quasi-symmetry model and the *RC* symmetric contingency model. Using the same approach now with the $RC(M)$ quasi-symmetric correspondence analysis models, we find that these models also suffer from the same serious limitations; but the corresponding $RC(M)$ symmetric association models do *not.*

Before closing this section, let us return for a moment to the $RC(M)$ symmetric association models. As we noted earlier herein, each of these models can be viewed as a more parsimonious special case of the Caussinus quasi-symmetry model (when $M < I - 1$). Various other kinds of symmetric association models, which are different from the $RC(M)$ symmetric association models, can also be viewed as more parsimonious special cases of the Caussinus quasi-symmetry model. Examples of these other kinds of symmetric association models were introduced in, e.g., Goodman [23,29,32]. These models will not be considered here, as this would go beyond the scope of the present paper.

6. Quasi-independence, quasi-symmetry, multiplicative models, log-linear models, and log-bilinear models

In the earlier sections herein, our attention was focused on the analysis of the square $I \times I$ contingency table. Now we shall consider the analysis of the more general rectangular $I \times J$ contingency table. The quasi-symmetry concept can be applied to the square table (in which there is a one-to-one correspondence between the row and column categories), while the other models to be considered in this section can be applied both to the square table and to the more general rectangular table.

We shall begin now with the quasi-independence model. For the $I \times J$ contingency table, we let P_{ij} denote the probability that an observation will fall in the i-th row and the j-th column of the table (for $i = 1, \ldots, I$; and $j = 1, \ldots, J$). The quasi-independence model pertaining to a given subset S of the $I \times J$ cells in the table states that the P_{ij} can be expressed as

$$P_{ij} = \alpha_i \beta_j \gamma_{ij}, \quad \text{with } \gamma_{ij} = \gamma \quad \text{for } (i,j) \text{ in } S, \tag{26}$$

with $\alpha_i \geqslant 0$, $\beta_j \geqslant 0$, $\gamma_{ij} \geqslant 0$ (for $i = 1, \ldots, I$; and $j = 1, \ldots, J$). Let $\bar{S}$ denote the subset of the $I \times J$ cells that are not in S. For the cells (i,j) in $\bar{S}$, we find that $\gamma_{ij} = P_{ij}/(\alpha_i \beta_j)$ if $(\alpha_i \beta_j) > 0$. Comparing models (1), (2), and (26), we see that each of these models expresses P_{ij} in terms of multiplicative effects; and the γ_{ij} multiplicative effects are subject to somewhat different restrictions in the three models. A model that expresses P_{ij} in terms of multiplicative effects can also be described by expressing $\log P_{ij}$ in terms of the corresponding additive effects (i.e., adding the logarithms of the multiplicative effects). So we can describe these models as “multiplicative models” and/or as “log-additive models” (or log-linear models).

In addition to the multiplicative (or log-linear) models (1), (2), and (26), we can obtain many other multiplicative (or log-linear) models by introducing other restrictions on the γ_{ij} and/or by expressing the γ_{ij} as a product of other multiplicative effects. A more general multiplicative model was introduced in Goodman [23], and models (1), (2), and (26) can be viewed as special cases of this more general model. The iterative methods that can be used to calculate the maximum-likelihood estimate of the P_{ij} under models (1), (2), and (26) can also be generalized in a straightforward way to obtain a more general iterative method that can be used to calculate the maximum-likelihood estimate of the P_{ij} under the more general multiplicative model. (The maximum-likelihood estimate of the P_{ij} under model (2) can be calculated by the iterative method and/or by the usual elementary explicit formula in this case.)

The statistical methods developed for the log-linear analysis of quasi-independence and quasi-symmetry, and the statistical methods developed for the log-linear analysis of three-factor interaction in a three-way contingency table (see, e.g., Goodman [15,16]), could be generalized then in a straightforward way to obtain the corresponding statistical methods developed for the analysis of the more general log-linear models in the multi-way contingency table (see, e.g., Goodman [21,23]); and the statistical methods developed for the log-linear model in the two-way table could be developed further to obtain statistical methods for the analysis of the log-bilinear model (see, e.g., Goodman [28,30]).

7. Quasi-independence and quasi-symmetry, and some references to additional articles pertaining to quasi-independence

We shall now consider how the quasi-independence model (26) differs from the quasi-symmetry model (1).

From (1) we see that Caussinus' quasi-symmetry condition (namely, that $\gamma_{ij} = \gamma_{ji}$ in (1), for $i = 1, \ldots, I$, and $j = 1, \ldots, I$) applies when $i \neq j$; but when $i = j$, the condition is tautological. So the quasi-symmetry model applied to the square $I \times I$ contingency table will yield the same results for the cells that are not on the main diagonal regardless of what the entries in the cells on the main diagonal might be. And, in particular, the model will yield the same results for the cells that are not on the main diagonal even when the entries in the cells on the main diagonal are deleted or when these cells are empty.

From (26) we see that the quasi-independence model can be applied to any given subset S of the $I \times J$ cells in the $I \times J$ contingency table; and the cells that are not in the subset S (i.e., the cells that are in the subset $\bar{S}$, the complement of S) can be viewed as empty cells or as cells in which the entries are deleted. The quasi-independence model can be applied in many different contexts when the contingency table is square and/or when the contingency table is rectangular, when the cells in $\bar{S}$ are the cells on the main diagonal and/or when the cells in $\bar{S}$ are any given subset of the cells in the $I \times J$ table (not necessarily the cells on the main diagonal). In the special case where the quasi-independence model is applied to the square $I \times I$ table *and* where the subset $\bar{S}$ consists of the cells on the main diagonal, then we can see from (26) and (1) that this special case of the quasi-independence model is also a special case of the quasi-symmetry model (when $I > 3$). However, when we consider the more general quasi-independence model applied either to the $I \times I$ table or the $I \times J$ table, where the subset $\bar{S}$ does not consist of the cells on the main diagonal, then the quasi-independence model is not a special case of the quasi-symmetry model. Each of these models can be applied in many different contexts.

Now for some closing comments pertaining to quasi-independence. Just in case some readers of this article may be interested, I shall take the liberty of including here a brief description of the development of my interest in quasi-independence and in the analysis of contingency tables in which some of the cells in the table are deleted or empty or are not of interest. My interest in this subject arose about forty-five years ago when Bill Kruskal and I were working on our second article on measures of association (see Goodman and Kruskal [35]). We considered then measures of association in contingency tables in which the main diagonal is not of interest. The measures of association we proposed for such a table did not require the comparison of the given contingency table (with the main diagonal deleted) with the corresponding table under quasi-independence; and so we did not consider the concept of quasi-independence in that article. Later on I found that there was a need, in many different contexts, for a concept of this kind and for iterative procedures for estimating the P_{ij} when this concept is applied; e.g., in the development of methods for the analysis of the mover-stayer problem, for the analysis of the transaction flows, for the analysis of persistence in a chain of multiple events, for the analysis of mobility tables, for the analysis of status persistence, for the development of scaling methods, for the analysis of triangular contingency tables (see, e.g., Goodman [11–14,17–20,22–27,33]).

The term quasi-perfect mobility was introduced in [17] for the analysis of mobility tables, and the term quasi-independence was introduced in [18]. The citations in the preceding paragraph were limited to my own work on these subjects; and the relevant work of others is referred to in the articles cited above (in the preceding paragraph) and also in, e.g., Agresti [1,2], Bishop *et al.* [5], Fienberg [9], and Haberman [36].

Bibliography

[1] Agresti (A.). — *Categorical Data Analysis*, New York, Wiley, 1990.

[2] Agresti (A.). — *An Introduction to Categorical Data Analysis*, New York, Wiley, 1996.

[3] Becker (M. P.). — Maximum likelihood estimation of the $RC(M)$ association model, *Applied Statistics*, 39 (1990), 152–157.

[4] Bishop (Y. M. M.), Fienberg (S. E.). — Incomplete two-dimensional contingency tables, *Biometrics*, 25 (1969), 119–128.

[5] Bishop (Y. M. M.), Fienberg (S. E.), Holland (P. W.). — *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA, MIT Press, 1975.

[6] CAUSSINUS (H.). — Contribution à l'analyse statistique des tableaux de correlation, *Annales de la Faculté des Sciences de l'Université de Toulouse*, 29 (annèe 1965), (1966), 77–182.

[7] DUNCAN (O. D.). — How destination depends on origin in the occupational mobility table, *American Journal of Sociology*, 84 (1979), 793–803.

[8] FIENBERG (S. E.). — Quasi-independence and maximum likelihood estimation in incomplete contingency tables, *Journal of the American Statistical Association*, 65 (1970), 1610–1616.

[9] FIENBERG (S. E.). — *The Analysis of Cross-Classified Categorical Data*, 2nd ed., Cambridge, MA, MIT Press, 1980.

[10] GILULA (Z.), HABERMAN (S. J.). — Canonical analysis of two-way contingency tables by maximum likelihood, *Journal of the American Statistical Association*, 81 (1986), 780–788.

[11] GOODMAN (L. A.). — Statistical methods for the mover-stayer model, *Journal of the American Statistical Association*, 56 (1961), 841–868.

[12] GOODMAN (L. A.). — Statistical methods for the preliminary analysis of transaction flows, *Econometrica*, 31 (1963), 197–208.

[13] GOODMAN (L. A.). — A short computer program for the analysis of transaction flows, *Behavioral Science*, 8 (1964), 176–186.

[14] GOODMAN (L. A.). — The analysis of persistence in a chain of multiple events, *Biometrika*, 51 (1964), 405–411.

[15] GOODMAN (L. A.). — Simple methods for analyzing three-factor interaction in contingency tables, *Journal of the American Statistical Association*, 59 (1964), 319–352.

[16] GOODMAN (L. A.). — Interactions in multidimensional contingency tables, *The Annals of Mathematical Statistics*, 35 (1964), 632–646.

[17] GOODMAN (L. A.). — On the statistical analysis of mobility tables, *American Journal of Sociology*, 70 (1965), 564–585.

[18] GOODMAN (L. A.). — The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries, *Journal of the American Statistical Association*, 63 (1968), 1091–1131.

[19] GOODMAN (L. A.). — How to ransack social mobility tables and other kinds of cross-classification tables, *American Journal of Sociology*, 75 (1969), 1–40.

[20] GOODMAN (L. A.). — On the measurement of social mobility: An index of status persistence, *American Sociological Review*, 34 (1969), 831–850.

[21] GOODMAN (L. A.). — The multivariate analysis of qualitative data: Interactions among multiple classifications, *Journal of the American Statistical Association*, 65 (1970), 226–256.

[22] GOODMAN (L. A.). — A simple simultaneous test procedure for quasi-independence in contingency tables, *Applied Statistics*, 20 (1971), 165–177.

[23] GOODMAN (L. A.). — Some multiplicative models for the analysis of cross-classified data, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, (eds. L. M. LeCam *et al.*), Berkeley, University of California Press, 1 (1972), 649–696.

[24] GOODMAN (L. A.). — On the empirical evaluation of certain logical forms of hypotheses using quasi-independence as a standard for comparison, *American Sociological Review*, 39 (1974), 273–277.

[25] GOODMAN (L. A.). — A new model for scaling response patterns: An application of the quasi-independence concept, *Journal of the American Statistical Association*, 70 (1975), 755–768.

[26] GOODMAN (L. A.). — On quasi-independence in triangular contingency tables, *Biometrics*, 35 (1979), 651–655.

[27] GOODMAN (L. A.). — The analysis of qualitative variables using more parsimonious quasi-independence models, scaling models, and latent structures., *Qualitative and Quantitative Social Research: Papers in Honor of Paul F. Lazarsfeld*, (eds. R. K. Merton *et al.*), New York, Free Press (1979), 119–137.

[28] GOODMAN (L. A.). — Simple models for the analysis of association in cross-classifications having ordered categories, *Journal of the American Statistical Association*, 74 (1979), 537–552.

[29] GOODMAN (L. A.). — Multiplicative models for the analysis of occupational mobility tables and other kinds of cross-classification tables, *American Journal of Sociology*, 84 (1979), 804–819.

[30] GOODMAN (L. A.). — The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries, *The Annals of Statistics*, 13 (1985), 10–69.

[31] GOODMAN (L. A.). — Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables, with discussion, *International Statistical Review*, 54 (1986), 243–309.

[32] GOODMAN (L. A.). — Measures, models, and graphical displays in the analysis of cross-classified data, with discussion, *Journal of the American Statistical Association*, 86 (1991), 1085–1138.

[33] GOODMAN (L. A.). — On quasi-independence and quasi-dependence in contingency tables, with special reference to ordinal triangular contingency tables, *Journal of the American Statistical Association*, 89 (1994), 1059–1063.

[34] GOODMAN (L. A.). — A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis, *Journal of the American Statistical Association*, 91 (1996), 408–428.

[35] GOODMAN (L. A.), KRUSKAL (W. H.). — Measures of association for cross-classifications, II: Further discussion and references, *Journal of the American Statistical Association*, 54 (1959), 123–163.

[36] HABERMAN (S. J.). — *Analysis of Qualitative Data*, Vols. 1 and 2 (1978,1979), New York: Academic Press.

[37] HAUSER (R. M.). — Some exploratory methods for modeling mobility tables and other cross-classified data, *Sociological Methodology*, ed. Karl F. Schuessler, San Francisco, Jossey–Bass (1980), 413-458.

[38] MCCULLAGH (P.). — The analysis of matched pairs with qualitative data, *Technical Report No. 75*, Department of Statistics, University of Chicago, 1978.