Jiří Anděl

The most significant interaction in a contingency table

Aplikace matematiky, Vol. 19 (1974), No. 4, 246–252

Persistent URL: http://dml.cz/dmlcz/103538

THE MOST SIGNIFICANT INTERACTION IN A CONTINGENCY TABLE

JIŘÍ ANDĚL

(Received July 6, 1973)

1. INTRODUCTION

Let us consider a $r \times c$ contingency table (n_{ij}), which is a sample of the size $n = \sum\sum n_{ij}$ from the multinomial distribution with the probabilities (π_{ij}). Suppose $\pi_{ij} > 0$ for all i, j. Let $\boldsymbol{\alpha} = (\alpha_{ij})$ be a matrix with real elements such that

(i) all the row totals are zeros,

(ii) all the column totals are zeros,

(iii) $\boldsymbol{\alpha} \neq \boldsymbol{0}$.

The set of such matrices will be denoted by M.

We shall suppose that $n_{ij} > 0$ for all i, j. (If some $n_{ij} = 0$ then it has been proposed to insert $\frac{1}{2}$ instead of 0. See [3].) Goodman [4] defined the interaction

$$\delta(\boldsymbol{\alpha}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \alpha_{ij} \ln \pi_{ij} \tag{1}$$

and showed that

$$d(\boldsymbol{\alpha}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \alpha_{ij} \ln n_{ij} \tag{2}$$

is the maximum likelihood estimate for $\delta(\boldsymbol{\alpha})$ and that

$$S(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{r} \sum_{j=1}^{c} \alpha_{ij}^2 n_{ij}^{-1}\right)^{1/2} \tag{3}$$

is a consistent estimate for the standard deviation of $d(\boldsymbol{\alpha})$.

In the case of independence

$$\pi_{ij} = p_i q_j \quad \text{for all } i, j\,, \quad \text{where} \quad \sum_{i=1}^{r} p_i = \sum_{j=1}^{c} q_j = 1 \tag{4}$$

and thus $\delta(\boldsymbol{\alpha}) = 0$ for any $\boldsymbol{\alpha} \in M$. Formula (44) in [4] implies that the random variable

$$W^2 = \sup_{\boldsymbol{\alpha} \in M} [d^2(\boldsymbol{\alpha})/S^2(\boldsymbol{\alpha})] \tag{5}$$

has for $n \to \infty$ asymptotically χ^2-distribution with $(r - 1)(c - 1)$ degrees of freedom. Obviously, W^2 is the same as that given in [4], formula (32). We shall find that $\boldsymbol{\alpha} \in M$ which actually gives the supremum in (5).

2. THE MOST SIGNIFICANT INTERACTION

Lemma 1. *Denote E_m the m-dimensional Euclidean space. Let $\mathbf{u} \in E_m$ be a given vector, $\mathbf{u} \neq \mathbf{0}$. Let $\mathbf{A}$ be a given positive definite matrix of the type $m \times m$. Then*

$$\sup_{\mathbf{x} \in E_m} [(\mathbf{u}'\mathbf{x})^2/(\mathbf{x}'\mathbf{A}\mathbf{x})] = \mathbf{u}'\mathbf{A}^{-1}\mathbf{u} \tag{6}$$

and the supremum is reached for

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{u} \,. \tag{7}$$

Proof. See [5], § 1 f. 1.

Note, that $k\mathbf{x}$, $k \neq 0$, gives the same value of (6) as $\mathbf{x}$.

Lemma 2. *Let*

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} \dots \mathbf{B}_{1m} \\ \\ \mathbf{B}_{m1} \dots \mathbf{B}_{mm} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \dots \\ \mathbf{U}_m \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_1 \\ \dots \\ \mathbf{V}_m \end{pmatrix}$$

where $\mathbf{B}_{ij}$, $\mathbf{U}_i$ and $\mathbf{V}_i$ are blocks of the type $h \times h$. Denote $\mathbf{I}$ the unit matrix. Let the matrices $\mathbf{B}$ and $\mathbf{Q} = \mathbf{I} + \mathbf{V}'\mathbf{B}^{-1}\mathbf{U}$ be regular. Then the matrix $\mathbf{A} = \mathbf{B} + \mathbf{U}\mathbf{V}'$ is regular and

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{U}\mathbf{Q}^{-1}\mathbf{V}'\mathbf{B}^{-1} \,. \tag{8}$$

Proof follows from the fact that the product

$$(\mathbf{B} + \mathbf{U}\mathbf{V}')(\mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{U}\mathbf{Q}^{-1}\mathbf{V}'\mathbf{B}^{-1})$$

gives the unit matrix.

Formula (8) is well-known for $h = 1$.

Let $\mathbf{C}$ be a matrix of the type $s \times t$, $\mathbf{C} = (c_{ij})_{i=1}^{s}\,_{j=1}^{t}$. If

$$\mathbf{c} = (c_{11}, \dots, c_{1t}, c_{21}, \dots, c_{2t}, \dots, c_{s1}, \dots, c_{st})' \,,$$

then we write

$$(c_{ij})_{i=1}^{s}\,_{j=1}^{t} \leftrightarrow \mathbf{c} \,.$$

Put $v_{ij} = \ln n_{ij}$, $u_{ij} = \ln(n_{ij}n_{rc}/n_{ic}n_{rj})$ and introduce vectors $\boldsymbol{a}$, $\boldsymbol{x}$, $\boldsymbol{v}$, $\boldsymbol{u}$ by

$$(\alpha_{ij})_{i=1}^{r}{}_{j=1}^{c} \leftrightarrow \boldsymbol{a}\,, \quad (\alpha_{ij})_{i=1}^{r-1}{}_{j=1}^{c-1} \leftrightarrow \boldsymbol{x}\,,$$
$$(v_{ij})_{i=1}^{r}{}_{j=1}^{c} \leftrightarrow \boldsymbol{v}\,, \quad (u_{ij})_{i=1}^{r-1}{}_{j=1}^{c-1} \leftrightarrow \boldsymbol{u}\,.$$

Denote

$$\boldsymbol{D} = \operatorname{diag}\{n_{11}^{-1}, \ldots, n_{1c}^{-1}, n_{21}^{-1}, \ldots, n_{2c}^{-1}, \ldots, n_{r1}^{-1}, \ldots, n_{rc}^{-1}\}\,,$$

$$\boldsymbol{B}_k = \begin{pmatrix} n_{k1}^{-1} + n_{kc}^{-1} & n_{kc}^{-1} & \ldots & n_{kc}^{-1} \\ n_{kc}^{-1} & n_{k2}^{-1} + n_{kc}^{-1} & \ldots & n_{kc}^{-1} \\ \ldots & \ldots & \ldots & \ldots \\ n_{kc}^{-1} & n_{kc}^{-1} & \ldots & n_{k,c-1}^{-1} + n_{kc}^{-1} \end{pmatrix}, \quad k = 1, 2, \ldots, r\,,$$

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B}_1 + \boldsymbol{B}_r & \boldsymbol{B}_r & \ldots & \boldsymbol{B}_r \\ \boldsymbol{B}_r & \boldsymbol{B}_2 + \boldsymbol{B}_r & \ldots & \boldsymbol{B}_r \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{B}_r & \boldsymbol{B}_r & \ldots & \boldsymbol{B}_{r-1} + \boldsymbol{B}_r \end{pmatrix}.$$

Theorem 3. *We have*

(9) $$W^2 = \boldsymbol{u}'\boldsymbol{A}^{-1}\boldsymbol{u}\,.$$

W^2 corresponds to the matrix $\boldsymbol{\alpha}$ the elements of which are given by

(10) $$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{u}$$

and

(11) $$\alpha_{ic} = -\sum_{i=1}^{c-1} \alpha_{ij} \quad (i \neq r)\,, \qquad \alpha_{rj} = -\sum_{i=1}^{r-1} \alpha_{ij} \quad (j \neq c)\,,$$
$$\alpha_{rc} = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \alpha_{ij}\,.$$

The matrix $\boldsymbol{A}^{-1}$ is given by the formula

(12) $$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{B}_1^{-1} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_2^{-1} & \ldots & \boldsymbol{0} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{B}_{r-1}^{-1} \end{pmatrix} - \begin{pmatrix} \boldsymbol{B}_1^{-1} \\ \boldsymbol{B}_2^{-1} \\ \ldots \\ \boldsymbol{B}_{r-1}^{-1} \end{pmatrix} \Big(\sum_{k=1}^{r} \boldsymbol{B}_k^{-1}\Big)^{-1} (\boldsymbol{B}_1^{-1}, \boldsymbol{B}_2^{-1}, \ldots, \boldsymbol{B}_{r-1}^{-1})\,,$$

where

(13) $$\boldsymbol{B}_k^{-1} = \begin{pmatrix} n_{k1} & 0 & \ldots & 0 \\ 0 & n_{k2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & n_{k,c-1} \end{pmatrix} - n_{k.}^{-1} \begin{pmatrix} n_{k1} \\ n_{k2} \\ \ldots \\ n_{k,c-1} \end{pmatrix} (n_{k1}, n_{k2}, \ldots, n_{k,c-1})\,,$$

$$n_{k.} = \sum_{j=1}^{c} n_{kj}\,, \quad k = 1, 2, \ldots, r\,.$$

Proof. We see that $d^2(\boldsymbol{\alpha})/S^2(\boldsymbol{\alpha}) = (\boldsymbol{a}'\boldsymbol{v})^2/(\boldsymbol{a}'\boldsymbol{D}\boldsymbol{a})$. Inserting for α_{ic}, α_{rj} according to (11) we get $d^2(\boldsymbol{\alpha})/S^2(\boldsymbol{\alpha}) = (\boldsymbol{u}'\boldsymbol{x})^2/(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})$. Formulas (9) and (10) follow from Lemma 1, formulas (12) and (13) from Lemma 2.

Our method for expressing the inverse of $\boldsymbol{A}$ is identical with the methods presented in [1] and [2]. The main reason of our previous analysis is that, when the statistic W^2 is significant, we can refer to the related value of $\boldsymbol{\alpha}$ which contributed most to the significance.

In view of numerical evaluation it is appropriate to construct the contingency table in such a way that $c \leqq r$, because the matrix $\boldsymbol{B}_k^{-1}$ is of the type $(c-1) \times (c-1)$ and its inverse is then easier to compute.

3. EQUIVALENCE WITH THE χ^2-TEST

Let us have two sequences of random variables $\{X_n\}$ and $\{Y_n\}$. The symbol $\xrightarrow{D}$ will denote the convergence in distribution, and $\xrightarrow{P}$ the convergence in probability. We say that sequences $\{X_n\}$ and $\{Y_n\}$ are asymptotically equivalent, if $X_n - Y_n \xrightarrow{P} 0$. This equivalence will be denoted by $X_n \sim Y_n$.

Lemma 4. (*a*) *If* $X_n \xrightarrow{P} X$, *then* $X_n \xrightarrow{D} X$.

(*b*) *If* $X_n \xrightarrow{D} X$, $Y_n \xrightarrow{P} 0$, *then* $X_n Y_n \xrightarrow{P} 0$.

(*c*) *If* g *is a continuous function and* $X_n \xrightarrow{P} X$, *then* $g(X_n) \xrightarrow{P} g(X)$.

The assertions (a) and (c) hold for random vectors, too.

Proof. See [5], § 2c.4.

Denote

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(n_{ij} - \frac{n_{i.}n_{.j}}{n} \right)^2 \bigg/ \frac{n_{i.}n_{.j}}{n},$$

where

$$n_{i.} = \sum_{j=1}^{c} n_{ij}, \quad n_{.j} = \sum_{i=1}^{r} n_{ij}.$$

Both of the variables W^2 and χ^2 depend on n, but this dependence will not be denoted explicitly. In the papers [1] and [2] it was noted that W^2 and χ^2 are asymptotically equivalent under the hypothesis of independence. We give a direct proof of this assertion.

Theorem 5. *The variables W^2 and χ^2 are asymptotically equivalent under the hypothesis of independence.*

Proof. Define

$$y_{ij} = n^{-1/2}(n_{ij} - np_i q_j)\,, \quad i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, c\,,$$

where p_i and q_j are mentioned in (4). Obviously

$$\sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij} = 0\,. \tag{14}$$

It is well-known that y_{ij} have an asymptotically simultaneous normal distribution (so that each y_{ij} converges in distribution) and that $n^{-1/2} y_{ij} \xrightarrow{P} 0$. Using Lemma 4 we obtain

$$\chi^2 \sim \sum_{i=1}^{r} \sum_{j=1}^{c} (y_{ij} - q_j \sum_{m=1}^{c} y_{im} - p_i \sum_{k=1}^{r} y_{kj})^2 / p_i q_j =$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} \left(2 + \frac{1 - p_i - q_j}{p_i q_j}\right) y_{ij}^2 + \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{\substack{m=1 \\ j \neq m}}^{c} \left(2 - \frac{1}{p_i}\right) y_{ij} y_{im} +$$

$$+ \sum_{\substack{i=1 \\ i \neq k}}^{r} \sum_{k=1}^{r} \sum_{j=1}^{c} \left(2 - \frac{1}{q_j}\right) y_{ij} y_{kj} + 2 \sum_{\substack{i=1 \\ i \neq k}}^{r} \sum_{k=1}^{r} \sum_{j=1}^{c} \sum_{\substack{m=1 \\ j \neq m}}^{c} y_{ij} y_{km} = S\,.$$

It follows from the Taylor formula that

$$n^{1/2} u_{ij} = n^{1/2} \ln (n_{ij} n_{rc} / n_{ic} n_{rj}) \sim \tag{15}$$

$$\sim y_{ic}/p_i q_c + y_{rj}/p_r q_j - y_{rc}/p_r q_c - y_{ij}/p_i q_j = w_{ij}\,.$$

Further we get

$$\frac{1}{n} \mathbf{B}_k^{-1} \xrightarrow{P} p_k \begin{pmatrix} q_1 & 0 & \ldots & 0 \\ 0 & q_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & q_{c-1} \end{pmatrix} - p_k \begin{pmatrix} q_1 \\ q_2 \\ \ldots \\ q_{c-1} \end{pmatrix} (q_1, q_2, \ldots, q_{c-1}) \tag{16}$$

and using (8)

$$n(\sum_{k=1}^{r} \mathbf{B}_k^{-1})^{-1} \xrightarrow{P} \begin{pmatrix} q_1^{-1} & 0 & \ldots & 0 \\ 0 & q_2^{-1} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & q_{c-1}^{-1} \end{pmatrix} + q_c^{-1} \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix} (1, 1, \ldots, 1)\,. \tag{17}$$

In view of (9), (12), (15), (16) and (17) we obtain after some computation

$$W^2 \sim \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} p_i q_j w_{ij}^2 - \sum_{i=1}^{r-1} p_i (\sum_{j=1}^{c-1} q_j w_{ij})^2 -$$

$$- \sum_{j=1}^{c-1} q_j (\sum_{i=1}^{r-1} p_i w_{ij})^2 + (\sum_{i=1}^{r-1} \sum_{j=1}^{c-1} p_i q_j w_{ij})^2 = R\,.$$

Inserting for w_{ij} from (16) we obtain after an elementary but rather long computation that

$$R = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(1 + \frac{1 - p_i - q_j}{p_i q_j}\right) y_{ij}^2 + \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{\substack{m=1 \\ j \neq m}}^{c} \left(1 - \frac{1}{p_i}\right) y_{ij} y_{im} +$$

$$+ \sum_{\substack{i=1 \\ i \neq k}}^{r} \sum_{k=1}^{r} \sum_{j=1}^{c} \left(1 - \frac{1}{q_j}\right) y_{ij} y_{kj} + \sum_{\substack{i=1 \\ i \neq k}}^{r} \sum_{k=1}^{r} \sum_{\substack{j=1 \\ j \neq m}}^{c} \sum_{m=1}^{c} y_{ij} y_{km} .$$

We see that the difference between S and R is

$$\left(\sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij}\right)^2$$

which is 0 in view of (14). Thus we have $\chi^2 \sim S = R \sim W^2$, which implies $\chi^2 \sim W^2$.

It is known that χ^2 has asymptotically chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom. From $\chi^2 \sim W^2$ it follows immediately that W^2 has the same asymptotic distribution as χ^2. Thus we obtained another proof of this fact.

References

[1] *Goodman L. A.:* On Plackett's test for contingency table interactions. J. Roy. Statist. Soc. Ser. B 25 (1963), 179–188.

[2] *Goodman L. A.:* Simple methods for analyzing three — factor interaction in contingency tables. J. Amer. Statist. Assoc. 59 (1964), 319–352.

[3] *Goodman L. A.:* Interactions in multidimensional contingency tables. Ann. Math. Statist. 35 (1964), 632–646.

[4] *Goodman L. A.:* Simultaneous confidence limits for cross — product ratios in contingency tables. J. Roy. Statist. Soc. Ser. B 26 (1964), 86–102.

[5] *Rao C. R.:* Linear statistical inference and its applications. Wiley, New York 1965.

Souhrn

NEJVÝZNAMNĚJŠÍ INTERAKCE V KONTIGENČNÍ TABULCE

Jiří Anděl

Budiž (n_{ij}) kontingenční tabulka typu $r \times c$, kterou lze pokládat za výběr o rozsahu $n = \sum\sum n_{ij}$ z multinomického rozdělení s kladnými pravděpodobnostmi (π_{ij}). Nechť $\boldsymbol{\alpha} = (\alpha_{ij})$ je nenulová matice typu $r \times c$, jejíž všechny řádkové i sloupcové součty jsou rovny nule. Množinu takových matic označíme M. Interakcí, která odpovídá matici $\boldsymbol{\alpha}$, se nazývá veličina

$$\delta(\boldsymbol{\alpha}) = \sum\sum \alpha_{ij} \ln \pi_{ij} .$$

Jejím maximálně věrohodným odhadem je

$$d(\boldsymbol{\alpha}) = \sum\sum \alpha_{ij} \ln n_{ij} .$$

Konsistentním odhadem pro var $d(\boldsymbol{\alpha})$ je

$$S^2(\boldsymbol{\alpha}) = \sum\sum \alpha_{ij}^2 n_{ij}^{-1} .$$

V případě nezávislosti v kontingenční tabulce (n_{ij}) platí $\delta(\boldsymbol{\alpha}) = 0$ pro každou matici $\boldsymbol{\alpha} \in M$. Goodman [4] dokázal, že pak pro každou $\boldsymbol{\alpha} \in M$ má náhodná veličina $d^2(\boldsymbol{\alpha})/S^2(\boldsymbol{\alpha})$ asymptoticky χ^2-rozdělení s jedním stupněm volnosti. Z jeho výsledků dále vyplývá, že při zmíněné nezávislosti má náhodná veličina

$$W^2 = \sup_{\boldsymbol{\alpha} \in M} \left[d^2(\boldsymbol{\alpha})/S^2(\boldsymbol{\alpha})\right]$$

asymptoticky χ^2-rozdělení s $(r - 1)(c - 1)$ stupni volnosti.

V článku je vypočtena matice $\boldsymbol{\alpha} \in M$, pro kterou podíl $d^2(\boldsymbol{\alpha})/S^2(\boldsymbol{\alpha})$ dosahuje své maximální hodnoty W^2. Výsledek je uveden ve větě 3. Znalost této matice $\boldsymbol{\alpha}$, která odpovídá nejvýznamnější interakci v kontingenční tabulce (n_{ij}), může pomoci při hledání zdroje případné závislosti a při interpretaci výsledku.

V závěru článku je pak přímou metodou dokázáno, že v případě nezávislosti je veličina W^2 asymptoticky ekvivalentní veličině χ^2, která se běžně užívá při vyhodnocování kontingenčních tabulek. Této ekvivalence si všiml již Goodman v pracích [1] a [2]. Naznačil tam důkaz opírající se o dosti složitou Waldovu teorii. Důkaz uvedený v článku má tu výhodu, že je proveden pouze pomocí elementárních prostředků teorie pravděpodobnosti.

Author's address: RNDr. *Jiří Anděl*, CSc., Matematicko-fyzikální fakulta Karlovy university, Sokolovská 83, 186 00 Praha 8 - Karlín.