Aplikace matematiky

Tomáš Havránek

On simultaneous inference in multidimensional contingency tables

Aplikace matematiky, Vol. 23 (1978), No. 1, 31–38

Persistent URL: http://dml.cz/dmlcz/103728

ON SIMULTANEOUS INFERENCE IN MULTIDIMENSIONAL CONTINGENCY TABLES

TOMÁŠ HAVRÁNEK

(Received October 15, 1976)

Investigating the GUHA-methods (cf. [4], [6] and [7]) we meet an interesting method of treatment of multidimensional contingency tables (an up-to-date detailed description of these methods can be found in [7]). In fact, in such methods the computer processes a set of 2×2 tables derived from the original $2 \times 2 \times \ldots \times 2$ table and on each such 2×2 table the hypothesis of independence is tested. It is clear that we face here the problem of simultaneous inference. As was shown by Anděl [1] the interaction test has good properties in such situations. In the present note we prove a theorem showing the properties of this test for the set of tables derived in a particular manner. These derived tables cannot be obtained via interactions or generalized interactions, hence our Theorem 1.8 covers situations distinct from those of Anděl's Theorem 3.

I. DERIVED CONTINGENCY TABLES

1.1. Consider the following examples: We have a $2 \times 2 \times 2$ contingency table with the frequencies

$$n_{111}, n_{101}, \quad n_{110}, n_{100}$$
$$n_{011}, n_{001}, \quad n_{010}, n_{000};$$

such a table refers to three properties (n_{101} is the frequency of objects possessing the first and the third property, but not the second). We can ask whether the first property P_1 is associated with the *conjunction* of the second and the third property (P_2 and P_3). Then we must consider the following derived 2×2 table:

$$n_{111}, \quad n_{101} + n_{110} + n_{100}$$
$$n_{011}, \quad n_{001} + n_{010} + n_{000}$$

and test an appropriate hypothesis of dependence.

Such a way of deriving contingency tables is characteristic for the present GUHA-methods; even rather complicated cases described by means of propositional calculus can occur (cf. also [9], with applications in sociology).

1.2. Remark. The usage of logical means is not incidental, it has a rather deep reason:

If we consider a model theory of finite dichotomously valued (i.e. finite classical) monadic models of a finite type t and we restrict ourselves to facts invariant with respect to the isomorphisms of models, we see that all such models can be represented by finite dimensional (t-dimensional) contingency tables. If M is such a model, it can be represented by a table (or vector)

$$\boldsymbol{n}(\boldsymbol{M}) = \langle n_{0,\ldots,0}, \ldots, n_{1,\ldots,1} \rangle ;$$

$n_{i_1,\ldots,i_t}$ is the frequency of objects having the value $\langle i_1, \ldots, i_t \rangle$. Thus the logical apparatus of the model theory can be (and is) a good formal means for investigating contingency tables.

1.3. Suppose we observe some random properties (i.e. alternative random variables) on a set of objects M; let $P_1, \ldots, P_t$ be the names of the properties. Random properties observed on objects form a sequence of t-dimensional random variables. We shall assume that this is a sequence of independently and identically distributed variables. Hence the distributional properties are described by a vector of probabilities

$$\boldsymbol{p} = \langle p_{0,\ldots,0}, p_{0,\ldots,0,1}, \ldots, p_{1,\ldots,1} \rangle ,$$

where $p_{i_1,\ldots,i_t}$ is the probability that the first property assumes the value i_1, the value i_2 etc. We shall suppose here $p_{i_1,\ldots,i_t} > 0$ for each $\langle i_1, \ldots, i_t \rangle \in \{0, 1\}^t$. The vector of the observed frequencies $\boldsymbol{n}$ is multinomially distributed with a parameter $\boldsymbol{p}$.

1.4. Consider composite properties named by *elementary conjunctions*, i.e. consistent conjunctions of literals. (Literals are atomic formulas P_j and negated atomic formulas $\neg P_j$; a conjunction of literals is consistent if each predicate P_j occurs in it at most once. Examples of elementary conjunctions: $P_2 \,\&\, \neg P_4$, P_1, $\neg P_1 \,\&\, P_2 \,\&\, \neg P_4$; example of an inconsistent conjunction of literals: $P_1 \,\&\, \neg P_1 \,\&\, P_4$.) The meaning of a statement "an object possesses the property named by an elementary conjunction" is obvious (e.g. an object possesses $P_2 \,\&\, \neg P_4$ iff it possesses P_2 and does not possess P_4). If we have now two disjoint conjunctions φ_1, φ_2 (i.e. φ_1, φ_2 have no common predicates) we can ask whether they are associated (not independent).

1.5. Example. In the example from 1.1 we speak about the association of P_1 and $P_2 \,\&\, P_3$. If we consider the association of P_1 and $P_2 \,\&\, \neg P_3$, we have to use the table

$$\begin{array}{ll} n_{110}\,, & n_{100} + n_{111} + n_{101} \\ n_{010}\,, & n_{000} + n_{011} + n_{001}\,. \end{array}$$

Similarly for four properties P_1, P_2, P_3, P_4 (say sex, treatment, symptom A, symptom B) one can construct a table corresponding to the relation between P_1 & P_2 and P_3 & $\neg P_4$.

The above presented tables are constructed using the following rule: Suppose we want to construct the table corresponding to elementary conjunctions φ_1, φ_2. The boolean functions corresponding to φ_1 and φ_2 are used. Denote them by $(\varphi_1)^*$, $(\varphi_2)^*$ respectively. In each cell of the 2×2 table we place the sum of frequencies with indices for which the boolean functions give values in the corresponding cell of the following pattern:

$$\begin{matrix} 1,1\,, & 1,0 \\ 0,1\,, & 0,0\,. \end{matrix}$$

In our example:

indices i_1, i_2, i_3	boolean $(P_1)^*$	functions (or $(P_2 \,\&\, P_3)^*$	$(P_1)^*$	$(P_2 \,\&\, \neg P_3)^*$)
0 0 0	0	0	(0	0)
0 0 1	0	0	(0	0)
0 1 0	0	0	(0	1)
0 1 1	0	1	(0	0)
1 0 0	1	0	(1	0)
1 0 1	1	0	(1	0)
1 1 0	1	0	(1	1)
1 1 1	1	1	(1	0)\,.

Now the reader can easily construct a table for the relation between P_1 & P_2 and P_3 & $\neg P_4$ (here 1110 corresponds to 1, 1, etc.) Note that we could use other forms of open formulae in the same way.

1.6. If $\boldsymbol{n}$ is a t-dimensional table and φ, ψ two disjoint composite properties (without any common atomic property), then the following derived table is relevant for the association of φ and ψ:

$$T(\varphi, \psi, \boldsymbol{n}) = \begin{matrix} \sum_{11}(\varphi, \psi, \boldsymbol{n})\,, & \sum_{10}(\varphi, \psi, \boldsymbol{n}) \\ \sum_{01}(\varphi, \psi, \boldsymbol{n})\,, & \sum_{00}(\varphi, \psi, \boldsymbol{n})\,. \end{matrix}$$

Here

$$\sum_{ij}(\varphi, \psi, \boldsymbol{n}) = \sum_{\{\langle i_1,\ldots,i_t\rangle;(\varphi)^*(i_1,\ldots,i_t)=i,(\psi)^*(i_1,\ldots,i_t)=j\}} n_{i_1,\ldots,i_t}\,.$$

Clearly, we can apply a test of independence in 2×2 contingency tables to our derived table $T(\varphi, \psi, \boldsymbol{n})$.

1.7. If we consider a set S of pairs of disjoint composite properties, we face a simultaneous inference problem. Particularly, we can ask what is the probability that, when testing the independence of pairs from S in a tables $\boldsymbol{n}$, one or more er-

roneous inferences occur under the assumption of the total independence; i.e. the probability of the global error of the first kind.

We know that the interaction test has good simultaneous test properties (cf. [1]). Hence we apply this test here.

Consider a 2×2 table $\begin{pmatrix} a, & b \\ c, & d \end{pmatrix}$. By the interaction test we reject the hypothesis of independence if

$$\log(ad/bc)\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)^{-1/2} \geqq \mathcal{N}_{\alpha/2}$$

where $\mathcal{N}_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the normalized normal distribution. We can apply this test to the set S of pairs of composite properties. For $S = \{\langle\varphi_1, \psi_1\rangle, \ldots, \langle\varphi_k, \psi_k\rangle\}$ we use some significance levels $\alpha_1, \ldots, \alpha_k$ and apply the interaction test to the i-th pair as follows: put $\begin{pmatrix} a, & b \\ c, & d \end{pmatrix} = T(\varphi_i, \psi_i, \boldsymbol{n})$ and $\alpha = \alpha_i$ and use the above described decision rule. Then we can prove the following theorem:

1.8. Theorem Let n be a contingency table and let $S = \{\langle\varphi_1, \psi_1\rangle, \ldots, \langle\varphi_k, \psi_k\rangle\}$ be a set of disjoint pairs of composite properties. Put $\alpha_S = 1 - \prod_{i=1}^{k}(1-\alpha_i)$. Then the probability of the global error of the first kind is asymptotically (in the cardinality of samples) less than or equal to α_S.

II. PROOFS

2.1. We have now to introduce some further notation. Having a pair $\langle\varphi_i, \psi_i\rangle$ we put

$$g_i(\boldsymbol{n}) = \log \sum_{11}(\varphi_i, \psi_i, \boldsymbol{n}) - \log \sum_{01}(\varphi_i, \psi_i, \boldsymbol{n}) - \log \sum_{10}(\varphi_i, \psi_i, \boldsymbol{n}) + \log \sum_{00}(\varphi_i, \psi_i, \boldsymbol{n}) \tag{1}$$

and

$$s_i(\boldsymbol{n}) = \left(\frac{1}{\sum_{11}(\varphi_i, \psi_i, \boldsymbol{n})} + \frac{1}{\sum_{01}(\varphi_i, \psi_i, \boldsymbol{n})} + \frac{1}{\sum_{10}(\varphi_i, \psi_i, \boldsymbol{n})} + \frac{1}{\sum_{00}(\varphi_i, \psi_i, \boldsymbol{n})}\right)^{1/2}.$$

Hence our test rejects the independence hypothesis on the derived table $T(\varphi_i, \psi_i, \boldsymbol{n})$ iff

$$\frac{|g_i(\boldsymbol{n})|}{s_i(\boldsymbol{n})} \geqq \mathcal{N}_{\alpha_i/2}.$$

2.2. Lemma. *Let $\boldsymbol{X}_n$ be a sequence of k-dimensional random vectors having asymptotically the $\mathcal{N}(0, A)$ distribution (normal distribution with zero means and regular covariance matrix A). Let $\sigma_i^2(\boldsymbol{X}_n)$ be consistent non-zero estimates of a_{ii}*

$(i = 1, \ldots, k)$. *Then* $\boldsymbol{X}_n^*$, *where* $X_{in}^* = X_{in}/\sigma_i(\boldsymbol{X})$ *for* $i = 1, \ldots, k$, *has asymptotically the* $\mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^*)$ *distribution, where* $\boldsymbol{A}^*$ *is a matrix with diagonal elements equal to* 1.

(The lemma follows from 2.c.4.(×) in [10].)

2.3. Lemma (Šidák [11]). *If variables* $X_1, \ldots, X_k$ *have a k-variate normal distribution with zero means, then*

$$P(|X_1| < c_1, \ldots, |X_k| < c_k) \geqq P(|X_1| < c_1) \ldots P(|X_k| < c_k).$$

2.4. Proof of Theorem 1.8. It is clear that

$$\langle \sqrt{(n)}\,(n_{0,\ldots,0}/n - p_{0,\ldots,0}), \ldots, \sqrt{(n)}\,(n_{1,\ldots,1}/n - p_{1,\ldots,1})\rangle$$

has asymptotically a 2^t-variate normal distribution with zero means and the dispersion matrix

$$\boldsymbol{V} = \begin{pmatrix} p_{0,\ldots,0}(1 - p_{0,\ldots,0}), & -p_{0,\ldots,0}p_{0,\ldots,1}, & \cdots \\ -p_{0,\ldots,1}p_{0,\ldots,0}, & \cdots & \\ & & p_{1,\ldots,1}(1 - p_{1,\ldots,1}) \end{pmatrix}.$$

Consider now functions g_i as defined in (1) and hence the vector variable

$$\boldsymbol{X}_n = \langle \sqrt{(n)}\,(g_1(\boldsymbol{n}/n) - g_1(\boldsymbol{p})), \ldots, \sqrt{(n)}\,(g_k(\boldsymbol{n}/n) - g_k(\boldsymbol{p}))\rangle,$$

where

$$\boldsymbol{n}/n = \langle n_{0,\ldots,0}/n, \ldots, n_{1,\ldots,1}/n\rangle$$

and n is the cardinality of the sample. Thus we obtain, for the diagonal elements of the dispersion matrix of $\boldsymbol{X}_n$,

$$v_i(\boldsymbol{p}) = \sum_{\langle i_1,\ldots,i_t\rangle} \sum_{\langle j_1,\ldots,j_t\rangle} v_{\langle i_1,\ldots,i_t\rangle\langle j_1,\ldots,j_t\rangle} \frac{\partial g_i(\boldsymbol{p})}{\partial p_{i_1,\ldots,i_t}} \frac{\partial g_i(\boldsymbol{p})}{\partial p_{j_1,\ldots,j_t}},$$

where $v_{i_1,\ldots,i_t\rangle\langle j_1,\ldots,j_t\rangle}$ are elements of $\boldsymbol{V}$. For g_i corresponding to $\langle \varphi_i, \psi_i\rangle$ we have $v_i(\boldsymbol{p}) = s_i^2(\boldsymbol{p})$ (to prove it is a principally elementary but tedious exercise). By Lemma 6.a.2. (iii) in [10] we see that $\boldsymbol{X}_n$ has asymptotically a normal distribution with the diagonal elements $v_i(\boldsymbol{p})$. (Moreover, $v_i(\boldsymbol{p}) = s_i^2(\boldsymbol{p}) > \boldsymbol{0}$.)

Note that under the null hypotheses we have $g_i(\boldsymbol{p}) = 0$ for $i = 1, \ldots, k$.

Moreover, we have $s_i^2(\boldsymbol{n}/n) = n\, s_i^2(\boldsymbol{n})$ and $g_i(\boldsymbol{n}/n) = g_i(\boldsymbol{n})$; we see immediately that $s_i^2(\boldsymbol{n}/n) > 0$ for $i = 1, \ldots, k$ and that $s_i^2(\boldsymbol{n}/n)$ is a consistent estimate of $s_i^2(\boldsymbol{p})$. Hence by 2.2

$$\left\langle \frac{\sqrt{(n)}\, g_1(\boldsymbol{n}/n)}{s_1(\boldsymbol{n}/n)}, \ldots \right\rangle = \left\langle \frac{g_1(\boldsymbol{n})}{s_1(\boldsymbol{n})}, \ldots \right\rangle$$

has asymptotically a k-variate normal distribution with zero means and the diagonal elements of the dispersion matrix equal to 1.

Now we apply Lemma 2.3 to estimate the probability of error. By the lemma we have

$$P(\bigcup_{i=1}^{k} |X_i| \geqq c_i) \leqq 1 - \prod_{i=1}^{k} (1 - P(|X_i| \geqq c_i)). \tag{2}$$

If now $\alpha_1, \ldots, \alpha_k$ are some numbers $(\alpha_i \in (0, 0{\cdot}5))$ and $X_1, \ldots, X_k$ are multinormal variables with zero means and with VAR $(X_i) = 1$ we obtain by (2)

$$P(\bigcup_{i=1}^{k} |X_i| \geqq \mathcal{N}_{\alpha_i/2}) \leqq 1 - \prod_{i=1}^{k} (1 - \alpha_i).$$

Applying the asymptotical properties of $g_i(\boldsymbol{n})/s_i(\boldsymbol{n})$, $i = 1, \ldots, k$, we have

$$\lim_{n \to +\infty} |P(\bigcup_{i=1}^{k} |g_i(\boldsymbol{n})| \geqq \mathcal{N}_{\alpha_i/2}\, s_i(\boldsymbol{n})) - P(\bigcup_{i=1}^{k} |X_i| \geqq \mathcal{N}_{\alpha_i/2})| = 0$$

under null hypotheses, which completes the proof.

III. DISCUSSION

3.1. Our proof depends on the idea of Anděl's proof of Theorem 3 in [1]. But Anděl's proof is not quite complete: it is based on the incorrect Rao's lemma 6.a.2.11. Anděl's proof can be corrected using our Lemma 2.2 to prove that

$$P(|d_1 - \delta_1| < cS_{d_1}, \ldots, |d_w - \delta_w| < cS_{d_w}) -$$
$$- P(|Y_1| < c, \ldots, |Y_w| < c), \quad \text{where} \quad Y_1, \ldots, Y_w$$

are multinormal with zero means and with the diagonal elements of the dispersion matrix equal to 1, converges to zero. (Since $\dfrac{(d_i - \delta_i)}{S_{d_i}}$ $i = 1, \ldots, w$ has asymptotically the desired distribution. In Anděl's proof

$$P(|d_1 - \delta_1| < cS_{d_1}, \ldots, |d_w - \delta_w| < cS_{d_w}) -$$
$$- P(|Y_1| < c\sqrt{(n)}\, S_{d_1}, \ldots, |Y_w| < c\sqrt{(n)'} S_{d_w})$$

need not converge to zero; note that the left expression is incorrect; $\langle S_{d_1}, \ldots, S_{d_w} \rangle$ is a random variable and the probability concerns $\langle Y_1, \ldots, Y_w \rangle$ (see [1], p. 104)).

3.2. Exactly the same error occurs in the proof of Theorem 1 of [1]. The proof of this theorem could be completed using the following easy lemma:

Let $\boldsymbol{X}_n$ be a sequence of k-dimensional random vectors having asymptotically the $\mathcal{N}(\boldsymbol{0}, \boldsymbol{A})$ distribution ($\boldsymbol{A}$ is assumed to be regular). Let $\boldsymbol{A}_n = f(\boldsymbol{X}_n)$ be consistent and regular estimates of $\boldsymbol{A}$. Then $\boldsymbol{X}_n \boldsymbol{A}_n^{-1} \boldsymbol{X}_n'$ has asymptotically the χ^2-distribution with k degrees of freedom. (A proof can be based on 2.c.4 from [10].)

3.3. Both Lemmas 3.2 and 2.2 have the same basis: norming the sample values by estimates of variances (or covariance matrix) based on the same sample values. Note that in all cases (Theorem 1.8 here and Theorems 1 and 3 in [1] and many others) the idea is to establish that $t(\boldsymbol{X}_n)$ *converges in distribution* to $\boldsymbol{X}$, where $\boldsymbol{X}$ has a distribution independent of the sample values.

3.4. All the proofs could be completed using Rao's lemma 6.a.2.11 in its true form, i.e. $\sup |F_n - H_n| \xrightarrow{P} 0$.

However, by the first step we obtain then only the convergence in probability which leads to further complications. The "norming" method seems to be more natural and straightforward.

3.5. How numerically good is the improvement of the Bonferroni bounds achieved by Theorem 3 of [1] or Theorem 1.8 of the present note? Consider some usual significance levels (probabilities of the error of the first kind) and a moderate number of tests:

level of one test	number of tests					
	3	5	10	15	20	30
0·001	0·002997	0·004990	0·009955	0·014895	0·019811	0·029569
	0·003	0·005	0·010	0·015	0·020	0·30
0·005	0·014925	0·024751	0·04889	0·07243	0·09539	0·13962
	0·015	0·025	0·05	0·075	0·1	0·15
0·01	0·0297	0·0490	0·0956	0·1399	0·1821	0·2603
	0·03	0·05	0·1	0·15	0·2	0·3
0·02	0·0588	0·0961	0·1829	0·2614	0·3323	0·4545
	0·06	0·1	0·2	0·3	0·4	0·6
0·05	0·143	0·226	0·401	0·537	0·642	0·785
	0·15	0·25	0·5	0·75	1·0	1·0

(the upper numbers are our bounds, the lower are the Bonferroni bounds).

We see immediately that our (and Anděl's) bounds are considerably better than Bonferroni bounds for values of the global probability greater than 0·10, i.e. out of the conventional region of admissible values of probability of an error.

Another comparison of the effectiveness of such bounds for simultaneous inference methods, using the given probability of the global error of the first kind and on this basis computed critical levels and critical values, can be found in [3] (with further references).

References

[1] *Anděl J.* (1973): On interactions in contingency tables. Aplikace matematiky 18, 99—109.

[2] *Anděl J.* (1974): The most significant interaction in contingency table. Aplikace matematiky 19, 246—252.

[3] *Dunn O. J., Massey F. J.* (1965): Estimation of multiple contrasts using *t*-distributions. J. Amer. Statist. Assoc. 60, 573—583.

[4] *Hájek P., Havel I., Chytil M.* (1966): The GUHA method of automatic hypotheses determination. Computing 1, 293—308.
[5] *Hájek P., Havel I., Chytil M.* (1966): GUHA — method of systematic searching for hypotheses. (Czech.) Kybernetika 2, 31—47.
[6] *Hájek, P.* (1968): The problem of a general concept of the GUHA — method. (Czech.) Kybernetika 4, 505—515.
[7] *Hájek P., Havránek T.* (1977): Mechanizing hypothesis formation; mathematical foundations for a general theory. Springer-Verlag, Heidelberg.
[8] *Hájek P., Havránek T.* (1977): On generation of inductive hypotheses. Int. Journ. Mon-Machine Studies 9, 415—438.
[9] *Lazarsfeld P. F.* (1961): The algebra of dichotomous systems, in: Studies in item analysis and prediction, H. Solomon (ed.), Stanford university press.
[10] *Rao C. R.* (1965): Linear statistical inference and its applications, J. Wiley, New York (sec. ed. 1974).
[11] *Šidák Z.* (1967): Rectangular confidence regions for the means of multivariate normal distributions. J. Amer. Statist. Assoc. 62, 626—633.

Souhrn

O SIMULTÁNNÍ INFERENCI V MNOHOROZMĚRNÝCH KONTINGENČNÍCH TABULKÁCH

TOMÁŠ HAVRÁNEK

V článku je studována otázka simultánní inference pro mnohorozměrné kontingenční tabulky typu $2 \times 2 \ldots \times 2$. Je popsána metoda odvozování kontingenčních tabulek 2×2, které odpovídají „složeným" vlastnostem; tyto tabulky nemohou být obdrženy z původní tabulky pomocí obvyklého kolapsování na marginální tabulky. Je dokázána věta asymptoticky omezující hladinu pravděpodobnosti celkové chyby prvního druhu při použití interakčního testu (viz [1]). Je dále diskutována efektivnost takto dosažené hladiny a uvedeny některé doplňky k článku [1].

Author's address: Dr. *Tomáš Havránek,* Matematické středisko biologických ústavů ČSAV Budějovická 1083, 142 20 Praha 4.