STATISTICAL RESEARCH REPORT No 5
Institute of Mathematics august 1975
University of Oslo

COMPARISON OF CONTINGENCY TABLES.
II: GENERAL CASE.

by

Jan F. Bjørnstad

ABSTRACT

A multiple testprocedure for comparison of any set of two-way contingency tables is proposed. The comparison-method is a generalization of a method for independent tables presented earlier by the author in [2].

Key words: Contingency table, measure of association, multiple comparison procedure.

CONTENTS

1. INTRODUCTION

In [2] the author proposed several methods for comparing independent two-way contingency tables by use of measures of association. In this paper we consider comparison of two-way tables generally, allowing dependence, and generalize a method given in [2] to this case. Further we define precisely the notion of two independent contingency tables, and show that this definition is consistent with the one formulated in [2] and [3]. At last a very simple proof of theorem 3 in [2] for general linear functions is presented, and we state some more properties of that method. Before we consider the general situation with several independent of dependent contingency tables, we first look at a general model for two contingency tables and present the main theorem.

2. A MULTINOMIAL MODEL FOR TWO CONTINGENCY TABLES. THE MAIN THEOREM.

The situation with two tables can be described as a multinomial model with two dependent sequences as follows. In sequence j, r_j events can occur with probabilities

$$p_{1j},\dots,p_{r_j,j}$$

for $j=1,2$. $\sum_{i=1}^{r_j} p_{ij} = 1$. Let $r=r_1+r_2$.

We assume all p_{ij} positive. Let k_n be the total number of independent trials, and let n_j be the total number of trials in sequence j, for $j=1,2$. Let $n=n_1+n_2$. It is assumed that $n \geq k_n$. I.e. some of the trials may give observations in both sequences.

Let I denote this set of trials and $m=\#(I)$. Then $k_n = m+(n_1-m)+(n_2-m) = n-m$. For the trials in I we let μ_{ij} be the probability of class i in sequence 1 and class j in sequence 2 , for $i=1,\dots,r_1$ and $j=1,\dots,r_2$. N_{ij} is the number of observations in cell i of sequence j , for $j=1,2$ and $i=1,\dots,r_j$. Then $n_j = \sum_{i=1}^{r_j} N_{ij}$. The relative frequences are denoted by $q_{ij}= N_{ij}/n_j$. Let $\pi = m/n$, $\pi_1 = n_1/n$ and $\pi_2 = n_2/n$. π,π_1,π_2 are considered as constants as n tends to infinity, and $\pi \geq 0$, $\pi_1>0$ and $\pi_2>0$.

We use the following notations:

$$p_1 = (p_{11},\dots,p_{r_1 1})'$$
$$p_2 = (p_{12},\dots,p_{r_2 2})'$$
$$q_1 = (q_{11},\dots,q_{r_1 1})'$$
$$q_2 = (q_{12},\dots,q_{r_2 2})'$$

$$q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}, \qquad p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} .$$

Let $\Sigma_1 = \{\sigma_{ij}\}$ be the covariance matrix of $\sqrt{n_1}q_1$ and let $\Sigma_2 = \{\tau_{ij}\}$ be the covariance matrix of $\sqrt{n_2}q_2$. Then

$$\sigma_{ij} = \begin{cases} p_{i1}(1-p_{i1}) & \text{for } i=j \\ -p_{i1}p_{j1} & \text{for } i\neq j \end{cases}$$

$$\tau_{ij} = \begin{cases} p_{i2}(1-p_{i2}) & \text{for } i=j \\ -p_{i2}p_{j2} & \text{for } i\neq j \end{cases}$$

Further we let $\Lambda = \{\rho_{ij}\}$ where

$$\rho_{ij} = \mu_{ij} - p_{i1}p_{j2} \quad \text{for} \quad i=1,\dots,r_1 \; ; \quad j=1,\dots,r_2 \; .$$

We see that

$$\text{cov}\,(q_{i1},q_{j2}) = \frac{\pi}{\pi_1\pi_2}\,\rho_{ij} \; .$$

The first result concerns the simultaneous asymptotic distribution of

$$\sqrt{n}\begin{bmatrix}\pi_1 q_1 - \pi_1 p_1\\ \pi_2 q_2 - \pi_2 p_2\end{bmatrix}$$

LEMMA 1.

$$\sqrt{n}\begin{bmatrix}\pi_1(q_1-p_1)\\ \pi_2(q_2-p_2)\end{bmatrix} \xrightarrow{\mathcal{D}} N_r(0,\Sigma) \tag{1}$$

where

$$\Sigma = \begin{bmatrix}\pi_1\Sigma_1 & \pi\Lambda\\ \pi\,\Lambda' & \pi_2\Sigma_2\end{bmatrix} .$$

$N_r(0,\Sigma)$ denotes the r-dimensional normal distribution with mean zero and covariance matrix Σ.

Proof.

Let $N_i = (N_{1i},\dots,N_{r_i i})'$ for $i=1,2$.

Let us first consider the trials from the set I, and define X_{ij}, Y_{ij} as follows:

$$X_{ij} = \begin{cases}1 & \text{if event no.}i \text{ in sequence 1 occur in trial no.}j\\ 0 & \text{otherwise}\end{cases}$$

$$Y_{ij} = \begin{cases}1 & \text{if event no.}i \text{ in sequence 2 occur in trial no.}j\\ 0 & \text{otherwise}\end{cases}$$

The m observations in I can be formulated as

$$U_j = (X_{1j},\dots,X_{r_1 j},\ Y_{1j},\dots,Y_{r_2 j})' \quad \text{for } j=1,\dots,m .$$

Let $M_{i1} = \sum_{j=1}^{m} X_{ij}$ and $M_{i2} = \sum_{j=1}^{m} Y_{ij}$, $M_1=(M_{11},\dots,M_{r_1 1})'$, $M_2 = (M_{12},\dots,M_{r_2 2})'$. $U_1,\dots,U_m$ are independent and identically distributed with mean p and covariance matrix

$$\Gamma = \begin{bmatrix} \Sigma_1 & \Lambda \\ \Lambda' & \Sigma_2 \end{bmatrix}$$

From the multivariate central limit theorem we then have that

$$\sqrt{m}\begin{bmatrix} M_1/m - p_1 \\ M_2/m - p_2 \end{bmatrix} \xrightarrow{\mathcal{D}} N_r(0,\Gamma) \quad \text{as } m\to\infty . \qquad (2)$$

For the rest of the trials in sequence 1 we let
L_{i1} be the number of observations in cell i .
For the rest of the trials in sequence 2 we let L_{i2} be number of observations in cell i . Let

$$L_1 = (L_{11},\dots,L_{r_1,1})$$

$$L_2 = (L_{12},\dots,L_{r_2,2})$$

$$n_1' = n_1-m \quad , \quad n_2' = n_2-m$$

Assume now that $\pi_i > \pi$ for $i=1,2$ such that n_1' , $n_2' \to \infty$ as $n \to \infty$.

We know that

$$\sqrt{n_1'}\left(\frac{L_1}{n_1'} - p_1\right) \to N_{r_1}(0,\Sigma_1) \quad \text{as} \quad n_1' \to \infty \tag{3}$$

and
$$\sqrt{n_2'}\left(\frac{L_2}{n_2'} - p_2\right) \to N_{r_2}(0,\Sigma_2) \quad \text{as} \quad n_2' \to \infty \tag{4}$$

We see that

$$\sqrt{n}(\pi_i q_i - \pi_i p_i) = \sqrt{n}\,\pi_i\left(\frac{N_i}{n_i} - p_i\right) = \sqrt{n}\,\frac{\pi_i \cdot n_i'}{n_i}\left(\frac{L_i}{n_i'} - p_i\right) + \sqrt{n}\,\frac{\pi_i \cdot m}{n_i}\left(\frac{M_i}{m} - p_i\right).$$

for $i=1,2$.

Let $X_i^n = \sqrt{n}\,\dfrac{\pi_i \cdot n_i'}{n_i}\left(\dfrac{L_i}{n_i'} - p_i\right)$ for $i=1,2$

and $Y_i^n = \sqrt{n}\,\dfrac{\pi_i \cdot m}{n_i}\left(\dfrac{M_i}{m} - p_i\right)$ for $i=1,2$.

$$X_i^n = \sqrt{\frac{n\pi_i}{n_i}} \cdot \sqrt{\frac{n_i'}{n_i \cdot \left(\frac{\pi_i - \pi}{\pi_i}\right)}} \cdot \sqrt{\pi_i - \pi}\sqrt{n_i'}\left(\frac{L_i}{n_i'} - p_i\right) \xrightarrow{\mathfrak{D}} N_{r_i}(0, (\pi_i - \pi)\,\Sigma_i)$$

from (3) and (4) .

Let $Z_1^n = \begin{pmatrix} X_1^n \\ X_2^n \end{pmatrix}$ and $Z_2^n = \begin{pmatrix} Y_1^n \\ Y_2^n \end{pmatrix}$. Then

$$\sqrt{n}\begin{pmatrix} \pi_1 q_1 - \pi_1 p_1 \\ \pi_2 q_2 - \pi_2 p_2 \end{pmatrix} = Z_1^n + Z_2^n \;;$$
Z_1^n and Z_2^n are independent for all n .

Let $a_n = \left(\dfrac{n\pi_1}{n_1} \cdot \dfrac{m\pi_1}{n_1\pi}\right)^{\frac{1}{2}}$ $b_n = \left(\dfrac{n\pi_2}{n_2} \cdot \dfrac{m\pi_2}{n_2\pi}\right)^{\frac{1}{2}}$. Then $a_n \to 1$ and $b_n \to 1$

and

$$Z_2^n = \sqrt{n}\begin{bmatrix} a_n & 0 \\ 0 & b_n \end{bmatrix}\sqrt{m}\begin{bmatrix} M_1/m - p_1 \\ M_2/m - p_2 \end{bmatrix}$$

Hence $Z_2^n \xrightarrow{\mathfrak{D}} N_r(0, \pi\,\Gamma)$.

Let now $\lambda = \binom{\lambda_1}{\lambda_2}$ be a fixed $r \times 1$ vector λ_1 is $r_1 \times 1$ and λ_2 is $r_2 \times 1$. Then

$$\lambda'(Z_1^n + Z_2^n) = \lambda' Z_1^n + \lambda' Z_2^n = \lambda_1' X_1^n + \lambda_2' X_2^n + \lambda' Z_2^n$$

Let $V_i^n = \lambda_i' X_i^n$ and $W^n = \lambda' Z_2^n$. Then V_1^n, V_2^n, W^n are independent and

$$V_1^n \xrightarrow{\mathcal{D}} N(0, \lambda_1'(\pi_1 - \pi)\Sigma_1 \lambda_1)$$

$$V_2^n \xrightarrow{\mathcal{D}} N(0, \lambda_2'(\pi_2 - \pi)\Sigma_2 \lambda_2)$$

$$W^n \xrightarrow{\mathcal{D}} N(0, \lambda' \pi \Gamma \lambda) .$$

Hence

$$\lambda'(Z_1^n + Z_2^n) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

where $\sigma^2 = \lambda_1'(\pi_1 - \pi)\Sigma_1\lambda_1 + \lambda_2'(\pi_2 - \pi)\Sigma_2\lambda_2 + \lambda'\pi\Gamma\lambda$.

We see that $\sigma^2 = \lambda_1'\pi_1\Sigma_1\lambda_1 + \lambda_2'\pi_2\Sigma_2\lambda_2 + \lambda_1'\pi\Lambda\lambda_2 + \lambda_2'\pi\Lambda'\lambda_1$

$$= \lambda'\Sigma\lambda .$$

This gives $\quad Z_1^n + Z_2^n \xrightarrow{\mathcal{D}} N_r(0, \Sigma)$.

We have now proved (1) when $\pi_i > \pi$ for i=1,2.
If one π_i or both are equal to π, we can put one or both of (X_1^n, X_2^n) equal to zero and the result follows.

Q.E.D.

Let M_{ij} be the number of observations from I that falls in cell i of sequence 1 and cell j of sequence 2, and let $m_{ij} = M_{ij}/m$. Further we assume that f is a function in r variables with continuous partial derivatives.

Let

$$f_i(p) = \frac{\partial f}{\partial x_i}\Big|_{x=p}$$

$$\hat{f}_i = f_i(q)$$

$$\bar{f}p_1 = \sum_{i=1}^{r_1} p_{i1} f_i(p)$$

$$\bar{f}p_2 = \sum_{i=1}^{r_2} p_{i2} f_{i+r_1}(p)$$

$$\bar{f}q_1 = \sum_{i=1}^{r_1} q_{i1} \hat{f}_i$$

$$\bar{f}q_2 = \sum_{i=1}^{r_2} q_{i2} \hat{f}_{i+r_1} .$$

Further we define

$$\sigma_f^2 = \frac{1}{\pi_1} \sum_{i=1}^{r_1} p_{i1}(f_i(p)-\bar{f}p_1)^2 + \frac{1}{\pi_2} \sum_{i=1}^{r_2} p_{i2}(f_{i+r_1}(p)-\bar{f}p_2)^2$$

$$+\frac{2\pi}{\pi_1\pi_2} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \rho_{ij} f_i(p) f_{r_1+j}(p) . \qquad (5)$$

We will from now on use the notations

$$p_i = (p_{1i},\ldots,p_{r_i i}) \quad i=1,2$$

$$q_i = (q_{1i},\ldots,q_{r_i i}) \quad i=1,2$$

$$p = (p_1,p_2) \text{ and } q = (q_1,q_2)$$

Lemma 1 states that

$$\sqrt{n}\left[(\pi_1 q_1,\pi_2 q_2) - (\pi_1 p_1,\pi_2 p_2)\right] \xrightarrow{\mathfrak{D}} N_r(0,\Sigma) \qquad (6)$$

We have the following fundamental result.

THEOREM 1.

If $\sigma_f > 0$ then

1)

$$\frac{\sqrt{n}(f(q) - f(p))}{\sigma_f} \xrightarrow{\mathcal{D}} N(0,1) \tag{7}$$

2)

$$\frac{\sqrt{n}(f(q) - f(p))}{\hat{\sigma}_f} \xrightarrow{\mathcal{D}} N(0,1) \tag{8}$$

where $$\hat{\sigma}_f^2 = \frac{1}{\pi_1} \sum_{i=1}^{r_1} q_{i1}(\hat{f}_i - \bar{f}q_1)^2 + \frac{1}{\pi_2} \sum_{i=1}^{r_2} q_{i2}(f_{i+r_1}(q) - \bar{f}q_2)^2$$

$$+ \frac{2\pi}{\pi_1\pi_2} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} (m_{ij} - q_{i1}q_{j2})\hat{f}_i \hat{f}_{j+r_1} \,. \tag{9}$$

Proof.

Let g be a function in r variables defined by

$$g(x_1,\dots,x_r) = f\left(\frac{x_1}{\pi_1},\dots,\frac{x_{r_1}}{\pi_1}, \frac{x_{r_1+1}}{\pi_2},\dots,\frac{x_r}{\pi_2}\right)$$

Then from lemma 1 and Rao,[5],p.321 we have that

$$\sqrt{n}\,(g(\pi_1 q_1,\pi_2 q_2) - g(\pi_1 p_1,\pi_2 p_2)) \xrightarrow{\mathcal{D}} N(0,\sigma_g^2) \quad ,$$

provided $\sigma_g^2 > 0$.

Let $\theta = (\pi_1 p_1, \pi_2 p_2)$. Then

$$\sigma_g^2 = \sum_{i=1}^{r} \sum_{j=1}^{r} \Sigma_{ij} \frac{\partial g}{\partial x_i}\bigg|_{x=\theta} \cdot \frac{\partial g}{\partial x_j}\bigg|_{x=\theta}$$

where $\Sigma = \{\Sigma_{ij}\}$.

Now from the definition of g,

$$g(\pi_1 q_1, \pi_2 q_2) = f(q)$$

and

$$g(\pi_1 p_1, \pi_2 p_2) = f(p) .$$

Hence

$$\sqrt{n}\,(f(q)-f(p)) \xrightarrow{\mathcal{D}} N(0,\sigma_g^2)$$

$\frac{\partial g}{\partial x_i} = \frac{\partial f}{\partial y_i} \cdot \frac{1}{\pi_1}$ for $i=1,\dots,r_1$ and $\frac{\partial g}{\partial x_i} = \frac{\partial f}{\partial y_i} \cdot \frac{1}{\pi_2}$ for $i=r_1+1,\dots,r$.

This gives

$$\frac{\partial g}{\partial x_i}\Big|_{x=\theta} = f_i(p)\frac{1}{\pi_1} \quad \text{for} \quad i \leq r_1,$$

and

$$\frac{\partial g}{\partial x_i}\Big|_{x=\theta} = f_i(p)\frac{1}{\pi_2} \quad \text{for} \quad r_1 < i \leq r .$$

Hence

$$\sigma_g^2 = \sum_{i=1}^{r_1} \sum_{j=1}^{r_1} \pi_1 \sigma_{ij} \cdot f_i(p) f_j(p) \left(\frac{1}{\pi_1}\right)^2$$

$$+ \sum_{i=r_1+1}^{r} \sum_{j=r_1+1}^{r} \pi_2 \tau_{i-r_1, j-r_1} f_i(p) f_j(p) \left(\frac{1}{\pi_2}\right)^2$$

$$+ 2 \sum_{i=1}^{r_1} \sum_{j=r_1+1}^{r} \pi \rho_{i, j-r_1} \frac{1}{\pi_1 \pi_2} f_i(p) f_j(p)$$

$$= \frac{1}{\pi_1}\left[\sum_{i=1}^{r_1} p_{i1} f_i^2(p) - (\bar{f}p_1)^2\right] + \frac{1}{\pi_2}\left[\sum_{i=1}^{r_2} p_{i2} f_{i+r_1}^2(p) - (\bar{f}p_2)^2\right]$$

$$+ \frac{2\pi}{\pi_1 \pi_2} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \rho_{ij} f_i(p) f_{j+r_1}(p) .$$

Hence $\sigma_g^2 = \sigma_f^2$ defined by (5) and 1) is proved.

2) follows from the fact that $\hat{\sigma}_f^2$ is a consistent estimator of σ_f^2 .

Q.E.D.

The next chapter presents first the general situation with k contingency tables. Then the comparison of two tables is considered, and we apply theorem 1 to comparison of measures of association. At last comparison of k tables is discussed, and a method generalizing the multiple normal-tests in [2] is presented.

3. MULTIPLE GN-TESTS FOR DIFFERENCES IN MEASURES OF ASSOCIATION.

3a) Assumptions and notations.

k two-way contingency tables are considered. The number of row- and column-classes in table no. i are respectively v_i and w_i, for $i=1,\dots,k$. Let $r_i=v_i \cdot w_i$. Let p_{ijh} denote the cell-probabilities in table h with $p_{ijh} > 0$ and

$$\sum_{i=1}^{v_h} \sum_{j=1}^{v_h} p_{ijh} = 1 \quad \text{for} \quad h=1,\dots,k .$$

q_{ijh} is the relative frequency in cell (i,j) of table h. Let n_h be the number of observations in table h. We let $n = \sum_{h=1}^{k} n_h$ and $\pi_h = n_h/n$. For each pair (i,j) of tables we let I_{ij} be the set of trials that gives observations in both table i and table j, and let $n_{ij} = \#(I_{ij})$ and $\pi_{ij} = n_{ij}/n$. For the trials in I_{rt}, μ_{ijhl}^{rt} is the probability of falling in cell (i,j) of table r and cell (h,l) of table t. All π_{ij} and π_h are considered as constants as n tends to infinity. $\pi_{ij} \geq 0$ and $\pi_h > 0$.

Let M_{ijhl}^{rt} be the absolute frequency from the set I_{rt} that falls in cell (i,j) of table r and cell (h,l) of table t. The relative frequencies are denoted by

$m^{rt}_{ijhl} = M^{rt}_{ijhl}/n_{rt}$.

The following notations are used

$$p_h = (p_{11h},\dots,p_{v_h w_h,h}) \quad \text{for} \quad h=1,\dots,k$$

$$q_h = (q_{11h},\dots,q_{v_h w_h,h}) \quad \text{for} \quad h=1,\dots,k \ .$$

$$p = (p_1,\dots,p_k)$$

$$q = (q_1,\dots,q_k) \ .$$

$$m^{rt} = (m^{rt}_{1111},\dots,m^{rt}_{v_r w_r, v_t w_t}) \ .$$

$$m = \{m^{rt}\} \quad \text{for} \quad r=1,\dots,k \quad t=1,\dots,k \ ; \ r<t \ .$$

$$\mu^{rt} = (\mu^{rt}_{1111},\dots,\mu^{rt}_{v_r w_r, v_t w_t})$$

$$\mu = \{\mu^{rt}\} \quad \text{for} \quad r=1,\dots,k \ , \quad t=1,\dots,k \ ; \ r<t \ .$$

Let d be the chosen measure of association with continuous partial derivatives as function of the cell-probabilities. For a presentation of measures of association we refer to the author's review in [1] , part 1 and the original paper [4] by Goodman and Kruskal.

Let d_i be the measure d in table i . Then d_i is a function of r_i variables with continuous partial derivatives. I.e. $d_i = d_i(p_i)$. A consistent estimator of d_i is $\hat{d}_i = d_i(q_i)$.

Let

$$\sigma^2_h = \sum_{i=1}^{v_h} \sum_{j=1}^{w_h} p_{ijh}(d_{ijh} - d^*_h)^2 \quad , \ h=1,\dots,k \ . \tag{10}$$

where

$$d_{ijh} = \frac{\partial d_h}{\partial p_{ijh}} \quad \text{and} \quad d^*_h = \sum_{i=1}^{v_h} \sum_{j=1}^{w_h} d_{ijh} \ p_{ijh}$$

A consistent estimator of σ_h^2 is

$$\hat{\sigma}_h^2 = \sum_{i=1}^{v_h} \sum_{j=1}^{w_h} q_{ijh}(\hat{d}_{ijh} - \hat{d}_h^*)^2 \tag{11}$$

where

$$\hat{d}_{ijh} = d_{ijh}(q_h) \quad \text{and} \quad \hat{d}_h^* = \sum_{i=1}^{v_h} \sum_{j=1}^{w_h} \hat{d}_{ijh} q_{ijh} .$$

Let further

$$\rho_{rt} = \sum_{i=1}^{v_r} \sum_{j=1}^{w_r} \sum_{h=1}^{v_t} \sum_{l=1}^{w_t} \mu_{ijhl}^{rt} d_{ijr} d_{hlt} - d_r^* \cdot d_t^* . \tag{12}$$

It is later seen that $\frac{\pi_{rt}}{\sqrt{\pi_r \pi_t}} \rho_{rt}$ can be considered as the asymptotic covariance of $(\sqrt{n_r}\, \hat{d}_r , \sqrt{n_t}\, \hat{d}_t)$.

A consistent estimator of ρ_{rt} is

$$\hat{\rho}_{rt} = \sum_{i=1}^{v_r} \sum_{j=1}^{w_r} \sum_{h=1}^{v_t} \sum_{l=1}^{w_t} m_{ijhl}^{rt} \hat{d}_{ijr} \hat{d}_{hlt} - \hat{d}_r^* \hat{d}_t^* . \tag{13}$$

We will now first consider the case $k=2$, i.e. comparison of two measures d_1 and d_2 .

3 b). Comparison of two tables.

We simplify our notation for this case, letting $\rho=\rho_{12}$, $\hat{\rho}=\rho_{12}$, $m_{ijhl}= m_{ijhl}^{12}$, $m=m^{12}$, μ_{ijhl}^{12} , $\mu=\mu^{12}$. $I=I_{12}$ and $n_{12}=\#(I)$, $\pi=\pi_{12}$.

We see that the situation is exactly as in section 2.

The result for comparing d_1 and d_2 can now be stated.

THEOREM 2.

Let $\sigma^2 = \frac{1}{\pi_1}\sigma_1^2 + \frac{1}{\pi_2}\sigma_2^2 - \frac{2\pi}{\pi_1\pi_2}\rho$, and assume $\sigma^2 > 0$.

Then

1) $$\frac{\sqrt{n}\,[\hat{d}_1-\hat{d}_2-(d_1-d_2)]}{\sigma} \xrightarrow{\mathfrak{D}} N(0,1) \tag{14}$$

2) $$\frac{\sqrt{n}\,[\hat{d}_1-\hat{d}_2-(d_1-d_2)]}{\hat{\sigma}} \xrightarrow{\mathfrak{D}} N(0,1) \tag{15}$$

where

$$\hat{\sigma}^2 = \frac{1}{\pi_1}\hat{\sigma}_1^2 + \frac{1}{\pi_2}\hat{\sigma}_2^2 - \frac{2\pi}{\pi_1\pi_2}\hat{\rho} \ . \tag{16}$$

Proof.

Theorem 1 is applied by letting

$$f(p) = d_1(p_1) - d_2(p_2) \ .$$

In order to facilitate the notation we replace (i,j) by a single letter i , such that p_{ijh} is replaced by p_{ih} , $i=1,\dots,r_h$ and d_{ijh} is replaced by d_{ih} , $i=1,\dots,r_h$. Similar changes for q_{ijh} and $\hat{d}_{ijh}$, and μ_{ijhl} is replaced by μ_{ij} . We find

$$f_i(p) = \begin{cases} d_{i1}(p_1) & \text{for } i=1,\dots,r_1 \\ -d_{i-r_1}(p_2) & \text{for } i=r_1+1,\dots,r \end{cases} \ .$$

Hence $\overline{f}p_1 = d_1^*$ and $\overline{f}p_2 = -d_2^*$.

In this case we get from (5)

$$\sigma_f^2 = \frac{1}{\pi_1}\sum_{i=1}^{r_1} p_{i1}(d_{i1}-d_1^*)^2 + \frac{1}{\pi_2}\sum_{i=1}^{r_2} p_{i2}(d_{i2}-d_2^*)^2$$

$$+ \frac{2\pi}{\pi_1\pi_2}\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}(\mu_{ij}-p_{i1}p_{i2})d_{i1}(-d_{j2})$$

Hence $\sigma_f^2 = \frac{1}{\pi_1}\sigma_1^2 + \frac{1}{\pi_2}\sigma_2^2 - \frac{2\pi}{\pi_1\pi_2}\left(\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\mu_{i1}d_{i1}d_{j2} - d_1^*d_2^*\right) = \sigma^2$.

Result 1) follows now from theorem 1.

Result 2) is proved by seeing that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

Q.E.D.

Before we look at some important special cases, we will discuss the notion of independence between two contingency tables. For this sake we define the set of variables $X = \{X_{ij}\}$ for $i=1,\ldots,v_1$, $j=1,\ldots,w_1$ and the set $Y = \{Y_{ij}\}$ for $i=1,\ldots,v_2$ and $j=1,\ldots,w_2$ as follows .

$$X_{ij} = \begin{cases} 1 & \text{if observation falls in cell } (i,j) \text{ of table 1} \\ 0 & \text{otherwise} \end{cases} \qquad (17)$$

$$Y_{ij} = \begin{cases} 1 & \text{if observation falls in cell } (i,j) \text{ of table 2} \\ 0 & \text{otherwise} \end{cases} \qquad (18)$$

The situation is that we have n_{12} independent observations (X^k,Y^k) , $k=1,\ldots,n_{12}$ of (X,Y) , then n_1-n_{12} independent observations $(X^k,0)$, $k=n_{12}+1,\ldots,n_2$ of $(X,0)$ and n_2-n_{12} independent observations $(0,Y^k)$, $k=n_1+1,\ldots,n-n_{12}$ of $(0,Y)$. A general formulation of the trials is then $U^k = (U_1^k,U_2^k)$; $k=1,\ldots,k_n$ $(k_n=n-n_{12})$, where

$$U_1^k = \begin{cases} X^k & \text{for } k \le n_1 \\ 0 & \text{for } k > n_1 \end{cases}$$

and

$$U_2^k = \begin{cases} Y^k & \text{for } k \le n_{12} \text{ and } k \ge n_1 \\ 0 & \text{otherwise} \end{cases} .$$

The following natural definition of independent tables is then:

Definition. Table 1 and 2 are said to be independent if U_1^k and U_2^k are independent for $k=1,\dots,k_n$.

Our main goal is to show that this definition is equivalent with q_1 and q_2 stochastically independent. We first need some simple results to prove this.

Since the observations are independent we see that if $n_{12}=0$ then the tables are independent.

First notice that X and Y are independent if and only if (X_{ij},Y_{hl}) are independent for all pairs (i,j) and (h,l) .

Let us now assume $n_{12}>0$. Then we have the following results.

LEMMA 2.

Table 1 and 2 are independent $\iff \mu_{ijhl} = p_{ij1}p_{hl2} \quad \forall(i,j,h,l)$.

Proof.

Since $n_{12}>0$, table 1 and 2 are independent if and only if X and Y are independent which is equivalent with (X_{ij},Y_{hl}) being independent for all (i,j,h,l) , and this again is equivalent with $\mu_{ijhl} = P(X_{ij}=1 \cap Y_{hl}=1) = P(X_{ij}=1)P(Y_{hl}=1) = p_{ij1}\cdot p_{hl2}$.

Q.E.D.

From theorem 2 we see that $\frac{\pi_{\cdot}}{\sqrt{\pi_1\pi_2}}\,\rho$ can be considered as the asymptotic covariance of $(\sqrt{n_1}\hat{d}_1 , \sqrt{n_2}\hat{d}_2)$,

and immediately from lemma 2 we get

$$\text{Table 1 and 2 independent} \Rightarrow \rho = 0 . \qquad (19)$$

<u>LEMMA 3.</u> Let $V_{ij} = \sum_{k=1}^{n_{12}} X_{ij}^k$, $W_{ij} = \sum_{k=1}^{n_{12}} Y_{ij}^k$,

$V = (V_{11}, \dots, V_{v_1 w_1})$

$W = (W_{11}, \dots, W_{v_2 w_2})$

V and W are independent <=> V_{ij} and W_{hl} are independent for all pairs (i,j) and (h,l).

Proof.

=> : Holds generally .

<= : If V_{ij} and W_{hl} are independent then

$0 = \text{cov}\,(V_{ij}\; W_{hl}) = n_{12}(\mu_{ijhl} - p_{ij1}p_{hl2})$, implying

that X_{ij} and Y_{hl} are independent. Hence X and Y are indpendent, giving that X^k and Y^k are independent for $k=1,..,n_{12}$.

$$V = \sum_{k=1}^{n_{12}} X^k, \quad W = \sum_{k=1}^{n_{12}} Y^k \quad .$$

Let $X^o = (X^1, \dots, X^m)$ and $Y^o = (Y^1, \dots, Y^m)$. It is easily shown that X^o and Y^o are independent. Hence, V and W are independent. Q.E.D.

We are now able to prove that definition of independence used in [2] and [3] is consistent with the natural definition given earlier.

<u>THEOREM 3.</u> Table 1 and 2 are independent

⇕

q_1 and q_2 are stochastically independent.

Proof.

1) $n_{12} = 0$. Obvious.

2) $n_{12} > 0$.

$$q_{ij1} = \frac{1}{n_1}(V_{ij} + Z_{ij}) \quad \text{where} \quad Z_{ij} = \sum_{k=n_{12}+1}^{n_1} X_{ij}^k$$

$$q_{ij2} = \frac{1}{n_2}(W_{ij} + T_{ij}) \quad \text{where} \quad T_{ij} = \sum_{k=n_1+1}^{k_n} Y_{ij}^k$$

$$Z = (Z_{11},\dots,Z_{v_1w_1}) \ , \ T = (T_{11},\dots,T_{v_2w_2})$$

<= : All pairs (q_{ij1}, q_{hl2}) are independent. Hence

$$0 = \text{cov}\,(V_{ij} + Z_{ij}, W_{hl} + T_{ij}) = \text{cov}\,(V_{ij}, W_{hl})$$
$$= n_{12}(\mu_{ijhl} - p_{ij1}p_{hl2})$$

From lemma 2 we get that the tables are independent.

=> : X and Y are independent, and therefore (X_{ij}^k, Y_{hl}^k) are independent for all (i,j,h,l) and $k=1,\dots,n_{12}$.

Let $X_o = (X_{ij}^1,\dots,X_{ij}^{n_{12}})$ $Y_o = (Y_{hl}^1,\dots,Y_{hl}^{n_{12}})$. X_o and Y_o are independent and therefore V_{ij} and W_{hl} are independent for all (i,j) and (h,l). From lemma 3 we know then that V and W are independent.

$$q_1 = \frac{1}{n_1}(V + Z) \quad \text{and} \quad q_2 = \frac{1}{n_2}(W + T) .$$

(V,Z) are independent of (W,T) and hence q_1 and q_2 are independent. Q.E.D.

We now like to look into some important special cases, and apply theorem 2 on them.

First we consider the independence case.

LEMMA 4. Table 1 and 2 independent.

$$\Downarrow$$

$$\frac{\sqrt{n}(\hat{d}_1-\hat{d}_2-(d_1-d_2))}{\left(\frac{\hat{\sigma}_1^2}{\pi_1}+\frac{\hat{\sigma}_2^2}{\pi_2}\right)^{\frac{1}{2}}} \xrightarrow{\mathfrak{D}} N(0,1)$$

Proof. $n_{12}=0 \Rightarrow \pi=0 \Rightarrow \hat{\sigma}^2 = \hat{\sigma}_1^2/\pi_1 + \hat{\sigma}_2^2/\pi_2$.

$n_{12}>0 \Rightarrow \rho=0 \Rightarrow \sigma^2=\sigma_1^2/\pi_1+\sigma_2^2/\pi_2$, and hence

$$\frac{\hat{\sigma}_1^2}{\pi_1}+\frac{\hat{\sigma}_2^2}{\pi_2} \xrightarrow{P} \sigma^2 .$$

Q.E.D.

This is the same result as lemma 1 in [2] for two tables. Another case that occur frequently is the situation where all the trials give observations in both tables.

LEMMA 5. Assume that $n_1=n_2=n_{12}$. I.e. the set I consists of all the trials.

Then

$$\frac{\sqrt{n_{12}}(\hat{d}_1-\hat{d}_2-(d_1-d_2))}{(\hat{\sigma}_1^2+\hat{\sigma}_2^2-2\hat{\rho})^{\frac{1}{2}}} \xrightarrow{\mathfrak{D}} N(0,1) .$$

Proof.

The estimated asymptotic variance in this case is

$\hat{\sigma}^2 = \frac{1}{2}\hat{\sigma}_1^2 + \frac{1}{2}\hat{\sigma}_2^2 - 4\hat{\rho}$, and the result follows. Q.E.D.

Another common case is considered in the next result.

LEMMA 6.

Assume that $n_1 = n_{12} < n_2$. I.e. all the observations in one table come from the set I . Then

$$\frac{\sqrt{n}\left[(\hat{d}_1-\hat{d}_2) - (d_1-d_2)\right]}{\left(\frac{\hat{\sigma}_1^2}{\pi_1} + \frac{\hat{\sigma}_2^2}{\pi_2} - \frac{2}{\pi_2}\hat{\rho}\right)^{\frac{1}{2}}} \xrightarrow{\mathfrak{D}} N(0,1) .$$

Proof. Obvious.

From theorem 2 we can propose the following test for comparing d_1 and d_2 :

State $d_i > d_j$ if

$$(\hat{d}_i - \hat{d}_j) > x(\tfrac{\alpha}{2})\hat{\sigma}/\sqrt{n} . \tag{20}$$

Here $x(\epsilon)$ is the upper ϵ -fractile in the $N(0,1)$-distribution. It is easily seen that

$$\lim_{n\to\infty} P\text{ (false statement)} = \begin{cases}\alpha & \text{if } d_1=d_2\\ 0 & \text{if } d_1 \neq d_2\end{cases} \tag{21}$$

A confidence interval for the difference d_1-d_2 with asymptotic confidence level equal to $1-\alpha$ is given by

$$d_1-d_2 \in [\hat{d}_1-\hat{d}_2 \pm x(\alpha/2)\,\hat{\sigma}/\sqrt{n}] \tag{22}$$

Let us now consider a case that often will appear, namely that one of the factors in both tables, say the column-factors , are the same. Then the two other factors will be two possible explaining factor to the primary column factor.

σ_1^2 and σ_2^2 will be just as before, but there will be a different expression for ρ , since $\mu_{ijhl} = 0$ for $j \neq l$.

Hence

$$\rho = \sum_{i=1}^{v_1} \sum_{j=1}^{w} \sum_{h=1}^{v_2} \mu_{ijhj} d_{ij1} d_{hj2} - d_1^* d_2^*$$

and $$\hat{\rho} = \sum_{i=1}^{v_1} \sum_{j=1}^{w} \sum_{h=1}^{v_2} m_{ijhj} \hat{d}_{ij1} \cdot \hat{d}_{hj2} - d_1^* d_2^* \quad ; \; w=w_1=w_2 \; .$$

Of course in this case we cannot have independence if $n_{12} > 0$, since $\mu_{ijhl} = 0$ and $p_{ij1} p_{hl2} > 0$ for $j \neq l$

At last in this section we go back to the case where $n_1 = n_2 = n_{12}$. In this case the asymptotic variance of $\sqrt{n_{12}}(\hat{d}_1 - \hat{d}_2)$ was found to be

$$\tau^2 = \sigma_1^2 + \sigma_2^2 - 2\rho \; .$$

After some calculations we find

$$\tau^2 = \sum_{i=1}^{v_1} \sum_{j=1}^{w_1} \sum_{h=1}^{v_2} \sum_{j=1}^{w_2} \mu_{ijhl} (d_{ij1} - d_{hl2})^2 - (d_1^* - d_2^*)^2$$

$$= \sum_i \sum_j \sum_h \sum_l \mu_{ijhl} [(d_{ij1} - d_{hl2}) - (d_1^* - d_2^*)]^2$$

Here all observations in both tables are results of the same trials, so we can consider the two-tables as one four-way (possibly three-way) table. If we let $D = d_1 - d_2$, we see that

$$\tau^2 = \sum_i \sum_j \sum_h \sum_l \mu_{ijhl} \left[\frac{\partial D}{\partial \mu_{ijhl}} - \left(\sum_{i,j} \sum_{h,l} \mu_{ijhl} \frac{\partial D}{\partial \mu_{ijhl}} \right) \right]^2 . \quad (23)$$

In the next section we will consider the general case, comparison of several tables. The multiple N-tests for differences proposed in [2] for independent tables will be generalized to this case.

3 c). Comparison of k tables.

The situation is given in 3 a).

The result we need is a direct consequence of theorem 2.

THEOREM 4.

Let $\sigma_{ij}^2 = \frac{1}{\pi_i}\sigma_i^2 + \frac{1}{\pi_j}\sigma_j^2 - \frac{2\pi_{ij}}{\pi_i\pi_j}\rho_{ij}$, and assume $\sigma_{ij}^2 > 0$.

Then

$$\sqrt{n}\,\frac{[\hat{d}_i-\hat{d}_j-(d_i-d_j)]}{\hat{\sigma}_{ij}} \xrightarrow{\mathcal{D}} N(0,1)$$

where

$$\hat{\sigma}_{ij}^2 = \frac{\hat{\sigma}_i^2}{\pi_i} + \frac{\hat{\sigma}_j^2}{\pi_j} - \frac{2\pi_{ij}}{\pi_i\pi_j}\hat{\rho}_{ij} .$$

Proof.

Let $n' = n_i + n_j$ and $\lambda_1 = \frac{n_i}{n'}$, $\lambda_2 = \frac{n_j}{n'}$ and $\lambda = \frac{n_{ij}}{n'}$.

Then from theorem 2

$$T = \frac{\sqrt{n'}\,(\hat{d}_i-\hat{d}_j-(d_i-d_j))}{(\frac{1}{\lambda_1}\hat{\sigma}_i^2 + \frac{1}{\lambda_2}\hat{\sigma}_j^2 - \frac{2\lambda}{\lambda_1\lambda_2}\hat{\rho}_{ij})^{\frac{1}{2}}} \xrightarrow{\mathcal{D}} N(0,1)$$

Now $\lambda_1 = \frac{\pi_i}{\pi_i+\pi_j}$, $\lambda_2 = \frac{\pi_j}{\pi_i+\pi_j}$ and $\lambda = \frac{\pi_{ij}}{\pi_i+\pi_j}$, hence

$$T = \frac{\sqrt{n'(\pi_i+\pi_j)^{-1}}\,(\hat{d}_i-\hat{d}_j-(d_i-d_j))}{\left(\frac{\hat{\sigma}_i^2}{\pi_i} + \frac{\hat{\sigma}_j^2}{\pi_j} - \frac{2\pi_{ij}}{\pi_i\pi_j}\hat{\rho}_{ij}\right)^{\frac{1}{2}}}$$ and the result follows.

Q.E.D.

From now on we assume that $\sigma_{ij}^2 > 0$ for all $i < j$.

A method for testing all differences d_i-d_j , $i < j$, that is similar to multiple normal-tests presented in [2] will be proposed. Let then

$$T_{ij} = \frac{\sqrt{n}\,(\hat{d}_i-\hat{d}_j)}{\hat{\sigma}_{ij}} \quad . \tag{24}$$

For a fixed sample $(n_1,\ldots,n_k)$ we see that T_{ij} can be expressed as

$$T_{ij} = \frac{(\hat{d}_i-\hat{d}_j)}{\left(\frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\sigma}_j^2}{n_j} - \frac{2n_{ij}}{n_i n_j}\hat{\rho}_{ij}\right)^{\frac{1}{2}}} \tag{25}$$

The following testprodecure for comparing $d_1,\ldots,d_k$ is proposed:

<u>Multiple generalized normal-tests (GN-test).</u>

State $d_i > d_j$ if $T_{ij} > x(\alpha/k(k-1))$. (26)

Let $\underline{d} = (d_1,\ldots,d_k)$ and $\underline{\sigma}_d=(\sigma_1,\ldots,\sigma_k)$. For a set $(\underline{d},\underline{\sigma}_d)$ of values of the parameter we let $\alpha(\underline{d},\underline{\sigma}_d)$ be the probability of at least one false statement "$d_i > d_j$" . We shall consider $\alpha(\underline{d},\underline{\sigma}_d)$ generally and apply the same approach as in [2] .

Let then the index sets V_i for $i=1,\ldots,t$ be as follows:

$V_i \subset \{1,\ldots,k\}$; V_i and V_j are disjoint for $i \neq j$ and

$\bigcup_{i=1}^{t} V_i = \{1,\ldots,k\}$.

Let further v_i be the number of elements in V_i , such that

$$\sum_{i=1}^{t} v_i = k \quad .$$

$\omega(V_1,\ldots,V_t)$ is the parameter set of all $(\underline{d},\underline{\sigma}_d)$ such that $d_i = d_j$ if $i,j \in V_h$ and $d_i \neq d_j$ if (i,j) belongs to different V_h's . The following result is valid.

<u>THEOREM 5.</u>

<u>If $(\underline{d},\underline{\sigma}_d) \in \omega(V_1,\ldots,V_t)$ then</u>

$$\limsup_n \underline{\alpha(\underline{d}\;,\underline{\sigma}_d) \leq (1 - \tfrac{t-1}{k})(1 - \tfrac{t-1}{k-1})\,\alpha} \quad . \tag{27}$$

<u>Proof.</u>

Now, since $\hat{d}_i - \hat{d}_j \xrightarrow{P} d_i - d_j$, it implies that

$$\lim_n P(T_{ij} > x(\alpha/k(k-1)) \mid d_i<d_j) = 0 \text{ , and hence}$$

$$\lim_n P(\bigcup_{g \neq h} \bigcup_{i \in V_g} \bigcup_{j \in V_h} (\text{false statement "}d_i>d_j\text{"})) = 0 \tag{28}$$

The theorem is therefore proved for $t=k$.

Assume $t < k$.

$$\alpha(\underline{d},\underline{\sigma}_d) = P(\bigcup_{h=1}^{t} \bigcup_{\substack{i<j \\ i,j \in V_h}} (\text{state : } d_i \neq d_j)) \quad , \quad \text{from (28) .}$$

Hence

$$\limsup_n \alpha(\underline{d},\underline{\sigma}_d) = \limsup_n P(\bigcup_{h=1}^{t} \bigcup_{\substack{i<j \\ i,j \in V_h}} |T_{ij}| > x(\tfrac{\alpha}{k(k-1)})$$

$$\leq \sum_{h=1}^{t} \sum_{\substack{i<j \\ i,j \in V_h}} \lim_n P(|T_{ij}| > x(\tfrac{\alpha}{k(k-1)}))$$

$$= \sum_{h=1}^{t} \sum_{\substack{i<j \\ i,j\in V_h}} \frac{2\alpha}{k(k-1)} = \frac{2\alpha}{k(k-1)} \sum_{h=1}^{t} \frac{v_h(v_h-1)}{2} = \frac{\alpha}{k(k-1)}\left[\sum_{h=1}^{t} v_h^2 - k\right].$$

In [2] we showed that $\sum_{h=1}^{t} v_h^2 \leq (k-t+1)^2 + (t-1)$. This implies

$$\limsup_n \alpha(\underline{d},\underline{\sigma}_d) \leq \frac{\alpha}{k(k-1)}\left[(k-t+1)^2 - (k-t+1)\right] = (1-\frac{t-1}{k})(1-\frac{t-1}{k-1})\alpha.$$

Q.E.D.

The upper bound in (27) is the same as the one given in [2] for multiple normal-tests for independent tables. Theorem 5 is therefore a generalization of the result in [2] ; it is valid for any set of tables. The upper bound on $\limsup_n \alpha(\underline{d},\underline{\sigma}_d)$ increases as t decreases and has maximum for $t=1$, such that

$$\limsup_n \alpha(\underline{d},\underline{\sigma}_d) \leq \alpha \quad \text{for all } (\underline{d},\underline{\sigma}_d) \tag{29}$$

Simultaneous confidence intervals for all differences d_i-d_j are given by the following relation

$$\limsup_n P(\hat{d}_i-\hat{d}_j-x(\tfrac{\alpha}{k(k-1)})\,\hat{\sigma}_{ij}/\sqrt{n} \leq d_i-d_j \leq \hat{d}_i-\hat{d}_j+x(\tfrac{\alpha}{k(k-1)})\hat{\sigma}_{ij}/\sqrt{n}$$
$$\text{for all } i \neq j) \geq 1-\alpha .$$

The last section in this paper deals with independent contingency tables. A very simple proof of theorem 3 on [2] is given, and we present some properties of the method for linear functions in $d_1,\ldots,d_k$, not given in [2] .

4. Comparison of k independent tables.

Since two tables are independent if and only if q_1 and q_2 are independent we say that k tables are independent if and only if $q_1,\dots,q_k$ are independent, as we did in [2] and [3]. We will now give another, simpler proof of theorem 3 in [2]. First, however, we need an algebraic result.

LEMMA 7. Let y be a $(k\times 1)$-vector. Then

$$y'y \le z \overset{(1)}{\Longleftrightarrow} h'y \le \sqrt{z}\sqrt{h'h} \quad \forall h=(h_1,\dots,h_k)' \overset{(2)}{\Longleftrightarrow} |h'y| \le \sqrt{z}\sqrt{h'h}\ , \quad \forall h = (h_1,\dots,h_k)' .$$

Here $z > 0$.

Proof.

(1) : $y'y \le z \Longleftrightarrow \Sigma y_i^2 \le z \Longleftrightarrow \Sigma h_i^2 \Sigma y_i^2 \le z\ \Sigma h_i^2 \quad \forall h = (h_1,\dots,h_k)'$.

From Schwartz inequality we get $\Sigma h_i^2\ \Sigma y_i^2 \ge (\Sigma h_i y_i)^2$

Hence $y'y \le z \Rightarrow (\Sigma h_i y_i)^2 \le z\ \Sigma h_i^2 \Longleftrightarrow (h'y)^2 \le zh'h \Rightarrow h'y \le \sqrt{z}\sqrt{h'h}$

The other way. Let $h=y$ and the result follows.

(2) is obvious. Q.E.D.

If now Y is a $(k\times 1)$ - random variable, it follows from lemma 7 that

$$P(Y'Y \le z) = P(h'Y \le \sqrt{z}\sqrt{h'h}\ ;\ \forall h) = P(|h'Y| \le \sqrt{z}\sqrt{h'h}\ ;\ \forall h) . \tag{30}$$

THEOREM 3 FROM [2].

Simultaneous confidence intervals for all linear functions $\sum_{i=1}^{k} c_i d_i$ are

$$\sum_{i=1}^{k} c_i d_i \in \left[\sum_{i=1}^{k} c_i \hat{d}_i \pm \sqrt{z(k,\alpha)}\ \hat{\sigma}_{c,\hat{d}} \right], \text{ where } \hat{\sigma}^2_{c,\hat{d}} = \sum_{i=1}^{k} \frac{c_i^2 \hat{\sigma}_i^2}{n_i} . \tag{31}$$

Here $z(k,\alpha)$ is the upper α-fractile in the chi-square distribution with k degrees of freedom. Asymptotically the probability is equal to
$(1-\alpha)$ that (31) is true for all $(c_1,\dots,c_k)$.

Proof.

Let $Y_i = \frac{\hat{d}_i - d_i}{\hat{\sigma}_i} \sqrt{n_i}$ and $Y = (Y,\dots,Y_k)'$.

Then $\lim_n P(Y'Y \leq z(k,\alpha)) = 1-\alpha$.

From (30) we see that

$$1-\alpha = \lim_n P(|\Sigma h_i Y_i| \leq \sqrt{z(k,\alpha)} \sqrt{\Sigma h_i^2} \ ; \forall h)$$

$$= \lim_n P(|\Sigma \frac{h_i \sqrt{h_i}}{\hat{\sigma}_i} (\hat{d}_i - d_i)| \leq \sqrt{z(k,\alpha)} \sqrt{\Sigma h_i^2} ; \forall h)$$

Let $\underline{\hat{d}} = (\hat{d}_1,\dots,\hat{d}_k)$ and $\hat{\sigma}_d = (\hat{\sigma}_1,\dots,\hat{\sigma}_k)$.

Let further

$$A = \{\underline{\hat{d}}, \underline{\hat{\sigma}_d} \,||\, \Sigma \frac{h_i \sqrt{n_i}}{\hat{\sigma}_i} (\hat{d}_i - d_i)| \leq \sqrt{z(k,\alpha)} \sqrt{\Sigma h_i^2} \ ; \forall h\}$$

$$B = \{\underline{\hat{d}}, \hat{\sigma}_d \,||\, \Sigma c_i (\hat{d}_i - d_i) \,| \leq \sqrt{z(k,\alpha)}\ \hat{\sigma}_{c,\hat{d}} \quad \forall c\} .$$

Now it is easily seen that $A = B$, and the result follows.

Q.E.D.

From (30) we also get the following result:

$$\lim_n P\left(\Sigma c_i d_i > \Sigma c_i \hat{d}_i - \sqrt{z(k,\alpha)}\, \hat{\sigma}_{c,\hat{d}} \;,\quad \forall c\right) = 1 - \alpha \,. \tag{32}$$

The test for linear functions is then to

state $$\Sigma c_i d_i > 0 \quad \text{if} \quad \Sigma c_i \hat{d}_i > \sqrt{z(k,\alpha)}\, \hat{\sigma}_{c,\hat{d}} \tag{33}$$

Then we have the following result, not shown in [2].

<u>LEMMA 8 .</u>

$$\limsup_n P\,(\text{at least one false statement: } \Sigma c_i d_i > 0) = \begin{cases} \alpha & \text{if } \underline{d} = 0 \\ \leq \alpha & \text{if } \underline{d} \neq 0. \end{cases}$$

<u>Proof.</u>

$$P(\text{no false statement}) = P\left(\bigcap_{\Sigma c_i d_i \leq 0} \Sigma c_i \hat{d}_i \leq \sqrt{z(k,\alpha)}\, \hat{\sigma}_{c,\hat{d}} \right)$$

Assume first $\underline{d} = 0$. Then $\Sigma c_i d_i = 0$ and hence

$$P(\text{no false statement}) = P\left(\bigcap_{\forall c} \Sigma c_i \hat{d}_i \leq \sqrt{z(k,\alpha)}\, \hat{\sigma}_{c,\hat{d}} \right) \underset{n\to\infty}{\longrightarrow} 1 - \alpha \,,$$

from (32) .

If $\underline{d} \neq 0$ then

$$\limsup_n P(\text{no false statement}) \geq \limsup_n P\left(\bigcap_{\Sigma c_i d_i \leq 0} \Sigma c_i \hat{d}_i - \Sigma c_i d_i \leq \sqrt{z(k,\alpha)}\, \hat{\sigma}_{c,\hat{d}} \right)$$

$$\geq \lim_n P\left(\bigcap_{\forall c} \Sigma c_i \hat{d}_i - \Sigma c_i d_i \leq \sqrt{z(k,\alpha)}\, \hat{\sigma}_{c,\hat{d}} \right) = 1 - \alpha \,, \text{ from (32).}$$

Q.E.D.

In [3], section 5 the author presents similar methods as (33) for all linear contrasts, i.e. linear functions $\Sigma c_i d_i$ with $\Sigma c_i = 0$. The difference from (33) is that we substitute $z(k,\alpha)$ with $z(k-1,\alpha)$. The testprocedure for linear contrasts has the same property as the one stated in lemma 8 for the method (33) .

References.

[1] Bjørnstad, J.F. : "Inference theory in contingency tables", Statistical Research Report, No. 2-1975, Univ. of Oslo.

[2] Bjørnstad, J.F.: "Comparison of contingency tables", Statistical Research Report, No. 3-1975, Univ. of Oslo.

[3] Bjørnstad, J.F.: "A general principle for multiple comparison procedure for linear contrasts, and some applications", Statistical Research Report, No. 4-1975, Univ. of Oslo.

[4] Goodman, L.A. & Kruskal, W.H.: "Measures of association for cross classifications", J.Am. Statist.Ass. Vol.49, 732-764, (1954).

[5] Rao, C.R.: "Linear Statistical Inference and Its Applications", John Wiley & Sons, Inc. (1965).