STATISTICAL RESEARCH REPORT
Institute of Mathematics
University of Oslo

No 2
1975

INFERENCE THEORY IN CONTINGENCY TABLES

by

Jan F. Bjørnstad

ABSTRACT.

This paper is divided into two parts. The first part gives a review on the measures of association which have been suggested in the literature. The aim of this review has been to guide an investigator in his choice of a measure in a given situation. It is strongly emphasized that one should only choose between measures which can be given a probabilistic interpretation.

The second part deals with testing of independence in a two-way table when the number of observations is large. The hypothesis "exact independence" will then nearly always be rejected. It is consequently a need for defining a notion "almost independence" and develop tests for this hypothesis. This is done by first considering testing of approximately exact hypotheses in the general multinomical case. Secondly we treat the problem of choosing an "almost independence" hypothesis by using an appropriate measure of association as a basis. Thirdly the theory for the general multinomial case is applied to such measures.

Key words: measure of association, almost independence.

CONTENTS

PART 2 : TESTING ALMOST INDEPENDENCE

PART ONE:

A REVIEW ON MEASURES OF ASSOCIATION IN CONTINGENCY TABLES

I. INTRODUCTION.

I.1. Some situations where measures of association are used.

The problem of choosing a measure of association appears when one wishes to examine the association between two factors or attributes in one or several contingency tables. There are especially two situations where measures of association can be of interest. One is comparison of dependence in several tables, and the other is testing for independence in one table.

It has become apparent that when testing for independence in a large datasample the exact independence hypothesis will nearly always be rejected, even in situations where the dependence evidently is very little. What one really wants to do is to accept independence between two factors even when there exist a slight degree of association. We then say that the factors are almost independent. So, instead of testing exact independence we want to test almost independence. When determinating the almost independence hypothesis, we have the problem of choosing a measure of association.

Part one is meant as a guidance as to which measure one ought to choose. In part two we consider the problem of testing almost independence, For comparison of tables we refer to [1].

I.2. An introductory discussion on measures of association.

The conception of association between two attributes will often be vague and not precise. Usually there are, however, special features of the association which we want to measure in a given situation. These relevant features of association can some times be specified as a part of the purpose of an investigation. A measure of association should in consequence be constructed from a relevant model for the particular case, so that it renders as much information as possible about the interesting features of association. So the desire is, that for a given situation the measure of association measures the features which are interesting to that particular situation. I.e., we sharpen the definition of association when constructing relevant, suitable measures.

If several measures are constructed for the same situation, one ought to choose the measure one believes gives the most evident expression for the relevant features of association. In addition it is required that the measures can be given a simple operational (probabilistic) interpretation such that, for one thing, values of a measure for different tables can be compared.

It seems natural when looking at measures of association, to seperate between the following five cases:

1) Ordered case.

There exists for each factor an underlying ordering between the categories .

One example can be : A: level of education and B: incomelevel.

2) Unordered symmetrical case.

There is no natural or relevant ordering. Moreover the factors appear symmetrically; there is no reason to give one factor precedence to the other.

3) Unordered asymmetrical case.

This situation occurs when one of the factors, say B, is of primary interest and there is no ordering in the two factors. This can happen if the factor A "precedes" B chronologically or causally. One example can be A: occupation and B: attitude to a certain problem.

4) Reliability-case.

This situation appears when $v = w$, and A and B assume the same categories but refer to two different methods. Let for instance A and B be two psychological tests both of which classify deranged individuals as to the type of mental disorder from which they suffer.

5) Mixed case.

The categories of one of the factors possess a natural, relevant ordering, the other do not. One example of this situation can be A: level of income and B: geographical classification.

Outside of these five cases we treat the 2×2 - situation separately. Most of the measures which is considered in the cases 1) - 4) can be found in [5] and [6]. The ordered situation is treated with special thoroughness, since it seems to occur quite frequently. The measures discussed there will all vary in the interval $[-1,1]$. As a measure for the degree of association in the ordered case, one can use the square of these measures.

II. THE INDEPENDENCE-SITUATION IN A TWO-WAY CONTINGENCY TABLE.

The following situation is considered. Two factors (later, also called attributes), A and B, can naturally be divided in respectively v and w categories $A_1,\dots,A_v$ and $B_1,\dots,B_w$. At each trial one and only one of the categories A_i & B_j will occur. Let Y, Z be two random variables defined by:

$$\begin{aligned} Y &= i \quad \text{if } A_i \text{ occur for } i = 1,\dots,v \\ Z &= j \quad \text{if } B_j \text{ occur for } j = 1,\dots,w \end{aligned} \tag{1}$$

The number of trials being executed is n. The outcome of each of the n trials is stochastically independent of the outcome from the other trials. At each trial the probability for occurence of A_i & B_j is p_{ij}. The probability for A_i then becomes $p_{i.} = \sum_{j=1}^{w} p_{ij}$, and the probability for B_j becomes $p_{.j} = \sum_{i=1}^{v} p_{ij}$. I.e. $p_{ij} = P(Y=i \cap Z=j)$, $p_{i.} = P(Y=i)$ and $p_{.j} = P(Z=j)$, for $i = 1, \ldots v$ and $j = 1, \ldots w$.

The factors A and B are said to be exact independent if Y and Z are stochastically independent. I.e. the hypothesis of exact independence between A and B can be expressed as

$$H: p_{ij} = p_{i.}\, p_{.j} \quad \text{for } i = 1, \ldots v \text{ and } j = 1, \ldots w. \qquad (2)$$

Let X_{ij} be the observed frequency in class A_i & B_j during the n trials, and let $q_{ij} = X_{ij}/n$. Let further $q_{i.} = \sum_j q_{ij}$ and $q_{.j} = \sum_i q_{ij}$. The statistical data can be arranged in a two-way contingency table:

A \ B	B_1	B_2	....	B_w	Sum
A_1	X_{11}	X_{12}	....	X_{1w}	$X_{1.}$
A_2	X_{21}	X_{22}	....	X_{2w}	$X_{2.}$
⋮	⋮				⋮
A_v	X_{v1}	X_{v2}	....	X_{vw}	$X_{v.}$
Sum	$X_{.1}$	$X_{.2}$	....	$X_{.w}$	n

Here is $X_{i.} = \sum_{j=1}^{v} X_{ij}$ and $X_{.j} = \sum_{i=1}^{r} X_{ij}$, that is $X_{i.}$ is the number of occurences of A_i and $X_{.j}$ is the number of occurences of B_j.

III. Ordered case.

III.1. Three ordinally invariant measures.

The situation under consideration is a relevant ordering between the categories within both factors. Let us first hive a definition of an ordinally invariant measure.

Definition 1 A measure g is said to be ordinally invariant if it is unchanged under similar types of monotone transformations of Y and Z, and if the sigh of g switches under unlike types of transformations. This means that $g(Y,Z) = g(f(Y),h(Z))$ if f and h are both strictly increasing or both strictly decreasing functions, and $g(Y,Z) = -g(f(Y),h(Z))$ if one of the functions is strictly decreasing, and the other is strictly increasing.

Since Y and Z are measured on an ordinal scale, such that the succession of the their possible values, but not the distance between them, has meaning, we require that a measure for this situation is ordinally invariant. In addition the measure g ought to satisfy two demands:

(i) $-1 \leq g \leq 1$

(ii) A, B exact independent $\Rightarrow$ $g = 0$.

If the range of g is bounded and g is symmetric in origo (i) can always be fulfilled by norming the measure.

We will describe three ordinally invariant measures, satisfying (i) and (ii), which are all modifications of a fundamental quantity. They are denoted by

1) γ , proposed by Goodman & Kruskal ([5],p.748).

2) τ_b , Kendall's rank-correlation coefficient modified to contingency tables.

3) τ_c , suggested by Stuart, [17].

The measure γ is also discussed in [11]. Later we will see that there are reasons for prefering γ to the others.
All the measures, but especially γ, have a simple probabilistic interpretation. We consider this measure first.

III.2. Construction of a natural measure, γ.

Let (Y_1,Z_1) and (Y_2,Z_2) be two independent random variables with the same distribution as (Y,Z).
γ is defined by

$$\gamma = P\{(Y_1-Y_2)(Z_1-Z_2) > 0 | Y_1 \neq Y_2 \cap Z_1 \neq Z_2\}$$
$$- P\{(Y_1-Y_2)(Z_1-Z_2) < 0 | Y_1 \neq Y_2 \cap Z_1 \neq Z_2\}.$$

It is immediately seen that γ is ordinally invariant.

Let $\pi_t = P(Y_1 = Y_2 \cup Z_1 = Z_2)$.
$\pi_s = P\{(Y_1-Y_2)(Z_1-Z_2) > 0\}$.
$\pi_d = P\{(Y_1-Y_2)(Z_1-Z_2) < 0\}$.

That is, π_s is the probability that the variables are concordant, and πd is the probability that they are discordant.
In this case we find it natural to extend the definition of exact independence to:

<u>Definition 2</u> <u>Two factors A and B are said to be ordering-independent (o.i.) if $\pi_s = \pi_d$.</u>

γ can be expressed as follows:

$$\gamma = \frac{\pi_s - \pi_d}{1 - \pi_t} \qquad (3)$$

Besides, since $\pi_t + \pi_s + \pi_d = 1$: $\gamma = \pi_s - \pi_d / \pi_s + \pi_d$.
Hence it is seen that $\gamma \in [-1,1]$, such that (i) is satisfied.
One finds that

$$\pi_t = \sum_{i=1}^{v} p_{i.}^2 + \sum_{j=1}^{w} p_{.j}^2 - \sum_{i=1}^{v}\sum_{j=1}^{w} p_{ij}^2 .$$

$$\pi_s = 2 \sum_{i=1}^{v-1}\sum_{j=1}^{w-1} p_{ij} \{ \sum_{i'>i}\sum_{j'>j} p_{i'j'} \} .$$

$$\pi_d = 2 \sum_{i=1}^{v-1}\sum_{j=2}^{w} p_{ij} \{ \sum_{i'>i}\sum_{j'<j} p_{i'j'} \} .$$

Further it can be shown that A, B exact independent implies that $\pi_s = \pi_d$, such that definition 2 actually is an extension of exact independence.

Hence γ satisfies (i) and (ii). In addition (see [5]), γ has the following properties:

(iii) A, B exact independent $\Rightarrow \gamma = 0$, but the converse need not hold except in the 2×2 - case.

(iv) γ is well-defined provided not all of the positive all-probabilities are concentrated in one single row or column.

In the 2×2 - table the measure reduces to

$$\gamma = \frac{p_{11}p_{22}-p_{12}p_{21}}{p_{11}p_{22}+p_{12}p_{21}} = \frac{\Delta-1}{\Delta+1} \qquad (4)$$

where $\Delta = p_{11}p_{22}/p_{12}p_{21}$ is the cross-product ratio.

Measures of association in the 2×2 - table will be discussed later in VIII.

III. 3. Two alternative measures, τ_b and τ_c.

Let us first consider the following situation.

Let U,V be continuous random variables. Kendall's rankcorrelation coefficient τ for (U,V) is defined by:

$$\tau = P\{(U_1-U_2)(V_1-V_2) > 0\} - P\{(U_1-U_2)(V_1-V_2) < 0\} \qquad (5)$$

where (U_1,V_1) and (U_2,V_2) are two independent random variables distributed as (U,V) ([11], p.822).

τ can be considered as the correlation coefficient between the signs of $U_1 - U_2$ and $V_1 - V_2$.

Let $(u_1,v_1),\ldots,(u_n,v_n)$ be n observations of (U,V) . We say that there are no ties if $u_i \neq u_j$ and $v_i \neq v_j$ for $i \neq j$ and $i = 1,\ldots,n,\ j = 1,\ldots,n$. In a contingency table there will occur ties if at least two observations fall in the same row or column, something that always will happen if $n > \min(v,w)$.

In the event of no ties γ is reduced to τ . In other words γ is a modification of τ to the situation with ties. We will now consider two other modifications to the situation with ties.

Let the situation be as in III.2.

Dendall's rank correlation coeffisient for contingency tables is defined by (our definition):

$$\tau_b = \frac{\pi_s - \pi_d}{\sqrt{P(Y_1 \neq Y_2)P(Z_1 \neq Z_2)}} \tag{6}$$

(Notice that $\gamma = (\pi_s - \pi_d)/P(Y_1 \neq Y_2 \cap Z_1 \neq Z_2)$.)

Let $\pi_y = P(Y_1 \neq Y_2)$ and $\pi_z = P(Z_1 \neq Z_2)$.

$$\pi_y = 1 - \sum_{i=1}^{v} p_{i.}^2 \ .$$

$$\pi_z = 1 - \sum_{j=1}^{w} p_{.j}^2 \ .$$

In the 2×2 - case we have

$$\tau_b = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1.}p_{2.}p_{.1}p_{.2}}} \tag{7}$$

τ_b satisfies (i) and (ii) in III.1., since $\tau_b = 0 \Leftrightarrow \pi_s = \pi_d$.

In addition τ_b has the following properties:

(iii) τ_b is well-defined provided not all positive cellprobabilities are concentrated in one single row or columm.

(iv) τ_b is ordinally invariant.

With regard to (i) it should be mentioned that the limits ± 1 are never attained except in a $v \times v$ - table where $\sum_{i=1}^{v} p_{ii} = 1$.
It is also worth noticing that τ_b^2 can be considered as a generalisation of $\beta = (p_{11}p_{22} - p_{12}p_{21})^2 / p_{1.}p_{2.}p_{.1}p_{.2}$ to a $v \times w$ ordered situation, while the traditional chi-square measure

$$\sigma^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} \frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

is a generalisation of β to the situation with no relevant ordering. For other traditional measures in that situation we refer to IV. 3.

The third measure τ_c is defined by:

$$\tau_c = \frac{\pi_s - \pi_d}{(m-1/m)} \tag{8}$$

where $m = \min(v,w)$.
The norming factor $m/m-1$ is a consequence of lemma 1.

<u>LEMMA 1. $-\frac{m-1}{m} \leq \pi_s - \pi_d \leq \frac{m-1}{m}$. The limits are attained in the case all the cellprobabilities are equal to 0 outside a longest diagonal of the table, and equal to $^1/m$ in the diagonal.</u>

<u>Proof</u>. The number of cells in a longest diagonal is equal to m. Assume first that $v = m$. Then $\max(\pi_s - \pi_d)$ occur when $\sum_{i=1}^{m} p_{i,i+k} = 1$ for a given k, $0 \leq k \leq w - m$ and $p_{i,i+k} = \frac{1}{m}$ for $i = 1,\ldots,m$.
Hence:

$$\max(\pi_s - \pi_d) = \frac{2}{m^2} \sum_{i=1}^{m-1} \sum_{i'>i} 1 = \frac{m-1}{m}.$$

Correspondingly, $\min(\pi_s - \pi_d)$ occur when

$$\sum_{i=o}^{m-1} p_{m-i,k+i} = 1 \quad \text{for a given } k\,, \quad 1 \leq k \leq w-m+1$$

and $p_{m-i,k+i} = \frac{1}{m}$ for $i = 0,1,\ldots,m-1$.

This gives that

$$\min(\pi_s - \pi_d) = -2\sum_{i=1}^{m-1} \sum_{i'>m-i} 1/m^2 = -\frac{m-1}{m} \,.$$

In case $w = \min(v,w)$, the proof is completely analoguous. The difference is only that $\max(\pi_s - \pi_d)$ occur when $0 \leq k \leq v-m$, $\sum_{j=1}^{m} p_{j+k,j} = 1$ and $p_{j+k,j} = 1/m$, and $\min(\pi_s - \pi_d)$ occur when $\sum_{j=1}^{m} p_{v-k-j,j} = 1$ for $-1 \leq k \leq v-m-1$ and $p_{v-k-j,j} = \frac{1}{m}$.

Q.E.D.

Lemma 1 gives that (i) is fulfilled, where now the limits -1 and $+1$ can be attained also when $v \neq w$. Condition (ii) also holds, and moreover, τ_c is ordinally invariant and always well defined. In the 2×2 - case $\tau_c = 4(p_{11}p_{22} - p_{12}p_{21})$.

Let now $\hat{\tau}_b$ be the estimator for τ_b obtained when substituting the relative frequencies q_{ij} instead of p_{ij} in the expression for τ_b . I.e.

$$\hat{\tau}_b = \frac{P_s - P_d}{\sqrt{P_y \cdot P_z}} \tag{9}$$

where $P_y = \pi_y(q)$, $P_z = \pi_z(q)$, $P_s = \pi_s(q)$ and $P_d = \pi_d(q)$.

(Stuart [17] shows a similar result for $P_s - P_d$ as we have done for $\pi_s - \pi_d$).

$\hat{\tau}_b$ can be considered as a special case of a generalized empirical correlation coeffisient (see [10], p.19). We give a short review of it. $(y_1,z_1),\ldots,(y_n,z_n)$ are the n independent observations that is executed. To every pair $\{(y_i,z_i),(y_j,z_j)\}$ a Y - score a_{ij} and a Z - score b_{ij} are assigned such that $a_{ij} = -a_{ji}$ and

$b_{ij} = -b_{ji}$ ($\Rightarrow a_{ii} = b_{ii} = 0$). The general empirical correlation coefficient is defined as:

$$\Gamma = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2}} \tag{10}$$

For example, the usual empirical (product) correlation coefficient

$$\frac{\sum_i (y_i-\bar{y})(z_i-\bar{z})}{\sqrt{\sum_i (y_i-\bar{y})^2 \cdot \sum_i (z_i-\bar{z})^2}}$$

is obtained by putting $a_{ij} = y_j - y_i$ and $b_{ij} = z_j - z_i$.

The following result shows how $\hat{\tau}_b$ appear as a special case of Γ.

LEMMA 2. <u>Let the Y-scores a_{ij} and Z-scores b_{ij} be given by</u>

$$a_{ij} = \begin{cases} +1 & \text{if } y_i < y_j \\ 0 & \text{if } y_i = y_j \\ -1 & \text{if } y_i > y_j \end{cases}, \quad b_{ij} = \begin{cases} +1 & \text{if } z_i < z_j \\ 0 & \text{if } z_i = z_j \\ -1 & \text{if } z_i > z_j \end{cases}.$$

<u>Then $\Gamma = \hat{\tau}_b$.</u>

<u>Proof.</u>

The number of ordered pairs among K elements is $K(K-1)$.
This implies that the number of ordered pairs among the total $n(n-1)$ ordered pairs for which $a_{ij} = 0$ is equal to $\sum_{i=1}^{v} X_{i.}(X_{i.}-1)$, such that $\sum_{i,j} a_{ij}^2 = n(n-1) - \sum_i X_{i.}(X_{i.}-1) = n^2 - \sum_i X_{i.}^2$.
Similarily: $\sum_{i,j} b_{ij}^2 = n^2 - \sum_j X_{.j}^2$.
This implies that the denominator in Γ can be expressed as

$$\{\sum_{i\neq j} a_{ij}^2 \cdot \sum_{i\neq j} b_{ij}^2\}^{\frac{1}{2}} = n^2\sqrt{(1-\sum_i q_{i.}^2)(1-\Sigma q_{i.}^2)(1-\sum_j q_{.j}^2)} = n^2\sqrt{P_y P_z}.$$

Let $U = \sum_{i,j} a_{ij}b_{ij} = 2\sum_{i<j} a_{ij}b_{ij}$, since $a_{ij}b_{ij} = a_{ji}b_{ji}$.

($\sum_{i<j} a_{ij}b_{ij}$ is called the total score S in Kendall, [10].)

After some calculation one finds that

$$\sum_{i<j} a_{ij}b_{ij} = \sum_{r=1}^{v-1}\sum_{k=1}^{w-1} X_{rk}\Big(\sum_{i>r}\sum_{j>k} X_{ij}\Big) - \sum_{r=1}^{v-1}\sum_{k=2}^{w} X_{rk}\Big(\sum_{i>r}\sum_{j<k} X_{ij}\Big),$$

which gives $U = n^2(P_s - P_d)$, and hence

$$\Gamma = \frac{P_s - P_d}{\sqrt{P_y \cdot P_z}} .$$

Q.E.D.

In the case of no ties we always have a_{ij}, b_{ij} equal to $+1$ or -1, for $i \neq j$, and therefore completely analogous to lemma 2, we see that $\Gamma = \hat{\tau}$ where $\hat{\tau}$ is the sample-statistic of τ, definǎd by (5). That is, $\hat{\tau}_b$ is the natural modification of $\hat{\tau}$, based on Γ.

III.4. A valuation of the measures γ, τ_b and τ_c.

The first thing to notice is that the three measures are all modifications of the difference $\pi_s - \pi_d$ to the situation with ties. The most natural modification is obviously γ, where one looks at the conditional probabilities given no ties. Both τ_b and τ_c seems to be somewhat artificial as modifications of $\pi_s - \pi_d$. Especially τ_c, which is only a norming of $\pi_s - \pi_d$.

Another thing one should take note of (regarding τ_b) is that originally it was the empirical rank correlation coeffisient $\hat{\tau}$ that was modified to $\hat{\tau}_b$, with starting-point at the generalized emperical correlation coeffisient Γ given by (10) (see [10] and [17]). The definition (6) is a result of substituting the probabilities p_{ij} instead of q_{ij} in $\hat{\tau}_b$. (τ_b is not mentioned in any of the articles that we give references to.) Hence, we have that while γ is the natural modification of τ based upon $\pi_s - \pi_d$, $\hat{\tau}_b$ is the natural modification of $\hat{\tau}$ based upon Γ. It is the parameter

that interests us. The correct thing to do must therefore be to modify the parameter, and thereafter look at the estimation problem, not to go the other way as Kendall did with $\hat{\tau}$.
The conclusion must therefore be that γ is the most natural and suitable measure in the ordered case.

None of the suggested measures in the ordered situation are invariant by permutations of rows or columns (of cell-probabilities) in the table, naturally. In the next situation to be considered the measures will remain unchanged under such permutations.

IV. UNORDERED SYMMETRICAL CASE.

IV. 1. A symmetrical prediction model.

Two measures of association, γ and η, suggested by Goodman & Kruskal, [5] and [6], will be discussed. In addition we mention some of the traditional measures of association which, however, cannot be given any operational interpretation.
The measures λ and η will be simple functions of error-probabilities within a certain model of prediction to be described. To give the model of prediction meaning it will be assumed that the cell-probabilities p_{ij} are known when constructing the measures λ and η. The two measures are the same function of probabilities for false predictions, based on two different methods of prediction. The symmetrical prediction model the measures are constructed from is as follows (see [5], p.743):

In a given trial one predict with probability 0.5 the B-class and with probability 0.5 the A-class. (Either A or B's class is predicted, each factor having probability equal to 0.5 for being drawn out for prediction.) If B's class is to be guessed, prediction is made on the basis of

(1) no further information, and

(2) given the A-category.

Similar, if A shall be predicted.

IV. 2. The measures λ and η based on respectively optimal and proportional prediction.

Goodman & Kruskal suggests two alternative methods of prediction.

a) Optimal prediction.

If B is drawn out: Predict in case (1) the class B_j with $p_{\cdot j} = \max_{j'} p_{\cdot j'}$, and in case (2), given A_i: Predict the class B_j with $p_{ij} = \max_{j'} p_{ij'}$. Same method is used if A is drawn out. Let

$$Q_1 = P \text{ (correct optimal prediction in case (1))}$$

and

$$Q_2 = P \text{ (correct optimal prediction in case (2))}.$$

b) Proportional prediction.

If B is drawn out: Predict in case (1) B_j with probability $p_{\cdot j}$, for $j = 1,\dots,w$ and in case (2), given A_i: Predict B_j with probability $p_{ij}/p_{i.}$ for $j = 1,\dots,w$. Similar if A is drawn out.

Let

$$P_1 = P \text{ (correct proportional prediction in case (1))}$$

$$P_2 = P \text{ (correct proportional prediction in case (2))}.$$

The measures λ and η are now defined as

$$\lambda = \frac{(1-Q_1)-(1-Q_2)}{Q-Q_1} = \frac{Q_2-Q_1}{1-Q_1} \tag{11}$$

$$\eta = \frac{(1-P_1)-(1-P_2)}{1-P_1} = \frac{P_2-P_1}{1-P_1} \tag{12}$$

One notice that λ and η both are relative decrease in probability of error in prediction from unknown to known characteristic for the factor which is not predicted.

Now $Q_i = \frac{1}{2} \cdot \{$ P (correct optimal prediction of B's characteristic in case (i)) + P(correct optimal prediction of A's characteristic in case (i))$\}$.

and similar for P_i. We find the following expressions for Q_i and P_i, λ and η.

$$Q_1 = \tfrac{1}{2}(p_{.m} + p_{m.})$$

$$Q_2 = \tfrac{1}{2}\left(\sum_{i=1}^{v} p_{im} + \sum_{j=1}^{w} p_{mj}\right) ,$$

where $p_{.m} = \max_j p_{.j}$, $p_{m.} = \max_i p_{i.}$, $p_{im} = \max_{j'} p_{ij'}$ and $p_{mj} = \max_{i'} p_{i'j}$.

$$\lambda = \frac{\sum_{i=1}^{v} p_{im} + \sum_{j=1}^{w} p_{mj} - p_{.m} - p_{m.}}{2 - p_{.m} - p_{m.}} \qquad (13)$$

$$P_1 = \tfrac{1}{2}\left(\sum_{i=1}^{v} p_{i.}^2 + \sum_{j=1}^{w} p_{.j}^2\right) .$$

$$P_2 = \tfrac{1}{2} \sum_{i=1}^{v} \sum_{j=1}^{w} p_{ij}^2\left(\frac{1}{p_{i.}} + \frac{1}{p_{.j}}\right) .$$

It is easily seen that η can be formulated as follows:

$$\eta = \frac{\sum_{i=1}^{v} \sum_{j=1}^{w} (p_{ij} - p_{i.}p_{.j})^2\left(\frac{1}{p_{i.}} + \frac{1}{p_{.j}}\right)}{2 - \sum_{i=1}^{v} p_{i.}^2 - \sum_{j=1}^{w} p_{.j}^2} \qquad (14)$$

Some properties of λ:

(i) λ is welldefined, except when one $p_{ij} = 1$

(ii) $0 \leq \lambda \leq 1$

(iii) A,B exact independent $\Rightarrow \lambda = 0$

(iv) λ is unchanged by permutations of rows and columns (of cell-probabilities) in the contingency table.

Some properties of η:

(i) η is well-defined, except when one $p_{ij} = 1$

(ii) $0 \leq \eta \leq 1$

(iii) A,B exact independent <=> $\eta = 0$

(iv) η is unchanged by permutations of rows and columns

In the 2×2-case η equals β. That is, η equals the chi-sequare measure φ^2, and τ_b^2, in the 2×2-table.

Which of the measures that is best suited for a given situation will depend on the method of prediction that is relevant for the situation. Usually it is perhaps most interesting to guess the most likely Y or Z-value, that is optimal prediction. One should, however, notice that λ is a somewhat "coarser" measure than η. By that we mean that if the association between A and B changes slightly, then λ will not necessarily reveal it.

IV. 3. Traditional mesures of association.

The most usual traditional measures of association are based on the chi-square measure, already mentioned:

a) $$\varphi^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} \frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \sum_{i=1}^{v} \sum_{j=1}^{w} \frac{p_{ij}^2}{p_{i.}p_{.j}} - 1 ,$$

(also called the mean square contingency in the literature.)

Three variations of this measure are mentioned in [5], p.739-740.

b) $K = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$ (suggested by K. Pearson).

c) $T = \sqrt{\frac{\varphi^2}{\sqrt{(v-1)(w-1)}}}$ (suggested by Tschuprow).

d) $C = \varphi^2/\min(v-1,w-1)$ (suggested by Cramér).

It is readily seen that $K,T,C \in [0,1]$ and that:
A,B exact independent <=> $\varphi^2 = K = T = C = 0$. It is difficult to give a probabilistic interpretation of these measures. Measures based on φ^2 are in other words not particularily meaningful . Goodman & Kruskal, [5], give a wider account of such measures without interpretation.

A measure not based on φ^2 was suggested by J.F. Steffensen in 1933. (See [6], p.140.)

$$\text{e)} \quad \psi^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} p_{ij} \frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}(1-p_{.i})p_{.j}(1-p_{.j})} .$$

Some properties:

(a) $\psi^2 = 0$ <=> A and B exact independent, and

(b) $0 \leq \psi^2 \leq 1$.

ψ^2 is a weighed average (with p_{ij} as weights) of all 2×2 mean square contingencies formed from each of the vw cells and its complement.

Teo simple measures are

$$\text{f)} \quad \varkappa_1 = \max_{i,j} | p_{ij} - p_{i.}p_{.j} | \qquad \text{and}$$

$$\text{g)} \quad \varkappa_2 = \max_{i,j} \left| \frac{p_{ij}}{p_{i.}} - p_{.j} \right|$$

It seems that $\varkappa_2$ is a more elucidating measure than $\varkappa_2$ (because, for one thing one usually set up tables with $q_{ij}/q_{i.}$ and consider the difference $q_{ij}/q_{i.} - q_{.j}$ when valuating the association in the table).

Let us now consider the case where one factor is of primary interest.

V. UNORDERED ASYMMETRICAL CASE.

V.1. An asymmetrical prediction model.

Let us assume that the factor B is of primary interest. Two measures, λ_b and η_b, suggested by Goodman & Kruskal, [5], are to be considered. The measures λ_b and η_b corresponds to λ and η in IV, with the difference that they are constructed in an asymmetrical model of prediction. For the model to have meaning we will assume, as in IV.1., that the p_{ij}'s are known when we construct the measures λ_b and η_b. The asymmetrical model, given in [5], p.741, is as follows:

In a given trial the B-class is to be predicted, on the basis of

1) No further information, and
2) Given the A-category.

Now, since B is the vital factor the relevant features of association are essentially of the type: "The difference" between correct B-prediction given A and correct B-prediction given no information. Accordingly the asymmetrical prediction model described above is a relevant model for constructing measures of association.

V.2. The measures λ_b and η_b based on respectively optimal and proportional prediction.

Optimal and proportional prediction for B are completely analogous to the definitions a) and b) in IV.2. That is

a) Optimal prediction means that one predict the most probable B-class in case(1), given no information, and (2) given A_i.

b) Proportional prediction means that one in case (1) predict B_j with probability $p_{.j}$ for $j = 1,\dots,w$, and in case (2),

given A_i, predict B_j with probability $p_{ij}|p_{i.}$ for $j = 1,\dots,w$.
The definition of λ_b and η_b are the same as the definition of λ and η in (11) and (12).

Let $Q_i^b = P$ (correct optimal prediction of B in case (i)) for $i = 1,2$, and $P_i^b = P$ (correct proportional prediction of B in case (i)) for $i = 1,2$.

Then:

$$\lambda_b = \frac{(1-Q_1)^b - (1-Q_2)^b}{1-Q_1^b} = \frac{Q_2^b - Q_1^b}{1-Q_1^b} \tag{15}$$

$$\eta_b = \frac{(1-P_1)^b - (1-P_2)^b}{1-P_1^b} = \frac{P_2^b - P_1^b}{1-P_1^b} \tag{16}$$

Both λ_b and η_b are relative decrease in probability of error in prediction from unknown to known A. The measures can be expressed in the following form:

$$\lambda_b = \frac{\sum_{i=1}^{v} p_{im} - p_{.m}}{1-p_{.m}} \tag{17}$$

$$\eta_b = \frac{\sum_{i=1}^{v}\sum_{j=1}^{w} p_{ij}^2/p_{i.} - \sum_{j=1}^{w} p_{.j}^2}{1-\sum_{j=1}^{w} p_{.j}^2} = \frac{\sum_{i=1}^{v}\sum_{j=1}^{w} \frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}}}{1-\sum_{j=1}^{w} p_{.j}^2}. \tag{18}$$

Some properties of λ_b:

(i) λ_b is indeterminate if and only if one $p_{.j} = 1$

(ii) $0 \leq \lambda_b \leq 1$

(iii) A,B exact independent $\Rightarrow \lambda_b = 0$

(iv) λ_b is invariant under permutation of rows and columns.

The properties (i), (ii), and (iv) are valid also for η_b. In addition we have:

(iii)': A,B exact independent <=> $\eta_b = 0$.

If A is the primary factor the measures will be completely corresponding:

$$\lambda_a = \frac{\sum_{j=1}^{w} p_{mj} - p_{m.}}{1 - p_{m.}} \tag{19}$$

$$\eta_a = \frac{\sum_{i=1}^{v} \sum_{j=1}^{w} p_{ij}^2 / p_{.j} - \sum_{i=1}^{v} p_{i.}^2}{1 - \sum_{i=1}^{v} p_{i.}^2} . \tag{20}$$

As to which of the measures λ_b or η_b that are most suitable in a given situation we refer to the discussion in IV.2. about λ and η.

VI. RELIABILITY - CASE.

VI.1. The unordered symmetrical case.

The situation is described in I.1. (see also [5], p.756). The characteristical thing in this case is that $A_i = B_i$ for $i = 1,\dots,v$. In this situation one is often interested in the degree of agreement between the two methods which A and B generally refer to. For the case where the categories does not hold a relevant ordering, Goodman & Kruskal, [5], construct a measure based on the symmetrical model of prediction given in IV.1. The prediction method is as follows:

In case (1) predict that B_i with $p_{i.} + p_{.i} = p_{M.} + p_{.M} = \max_{i'}(p_{i'.} + p_{.i'})$. Similar if A is drawn out.

In case (2), given A_i, predict B_i. Correspondingly if A is to be predicted.

Let Λ_i = P (correct prediction in case (i)) for $i = 1,2$.

The proposed measure is defined analogous to λ, η, λ_b and η_b :

$$\lambda_r = \frac{(1-\Lambda_1)-(1-\Lambda_2)}{1-\Lambda_1} = \frac{\Lambda_2-\Lambda_1}{1-\Lambda_1} \tag{21}$$

One finds that:

$$\Lambda_1 = \tfrac{1}{2}(p_{M.}+p_{.M})$$

$$\Lambda_2 = \sum_{i=1}^{v} p_{ii}$$

such that

$$\lambda_r = \frac{\sum_i p_{ii} - \frac{1}{2}(p_{M.}+p_{.M})}{1-\frac{1}{2}(p_{M.}+p_{.M})} \tag{22}$$

Some properties: i) $-1 \leq \lambda_r \leq 1$

ii) λ_r assumes no particular value in case A and B are exact independent, but as Goodman & Kruskal argue a measure as λ_r would only be used where there is known to be dependence between the methods A and B, so this undesirable quality is not so important.

VI.2. The ordered case.

In this situation it has been customary to use measures of the type

$$\pi_k = \sum_{|i-j|\leq k} p_{ij} \quad \text{for a chosen } k .$$

For example, π_0 $\left(= \sum_{i=1}^{v} p_{ii}\right)$ is the probability that the methods "agree" (that is give the same result).

VII. MIXED SITUATION.

A case which has not been discussed in any of the articles which we refer to is the situation where we have a nominal level for one of the variables (Y,Z) and an ordinal level for the other. We shall here try to forward some suggestions for measures of association in this case. Let us for the sake of simplicity suppose that Y holds an ordinal level. The kind of measure one ought to choose will depend on the features of dependence one is mainly interested in. It seems natural to separate between the following three situations.

a) Asymmetrical situation. B is of primary interest.

b) Asymmetrical situation. A is of primary interest.

c) Symmetrical situation.

a) B has primary interest.

There is no interesting ordering in B's classes, so it seems reasonable that an asymmetrical prediction model as in V.1 is relevant here. Consequently the measure should be constructed from that model. λ_b and η_b are therefore suitable measures.

b) A has primary interest.

Since the classes for the primary factor hold a relevant ordering it would be reasonable to require that the measure in any case is not invariant under permutation of rows in the contingency tables. This implies that all measures in the unordered case are not eligible. A suitable measure then seems to be a measure constructed for the ordered case, which means γ since this measure was found to be the most natural of three measures valuated in III.

c) Symmetrical situation.

As mentioned earlier, this situation appears when there is no reason to give one factor priority in preference to the other. Intuitively it seems natural that a measure of association in this case is a function of two measures D_1, D_2, where D_1 is a measure for the ordered situation ($-1 \leq D_1 \leq 1$), and D_2 is a measure constructed for the unordered situation ($0 \leq D_2 \leq 1$).

Such a function $h(D_1,D_2)$ should idealistically have the following properties:

1) Invariant under permutation of columns
2) Not invariant under permutation of rows.

It seems however, that this is a much too ambitious assumption. A more unprecise condition is:

h should utilize the information from D_1 and D_2 to "the same amount".

Besides it can be desirable that

$$h(D_1,D_2) = 0 \iff D_1 = D_2 = 0 \qquad (23)$$

Examples of such measures are:

i) $h(D_1,D_2) = a(|D_1| + D_2)$
ii) $h(D_1,D_2) = b(D_1^2 + D_2)$; a and b are constants.

The measures i) and ii) will be non-negative. If it is believed that the condition (23) is immaterial other measures of the form $c(D_1 + D_2)$ and $d(D_1 \cdot D_2)$ can be used, where c and d are constants.

At last we will consider the 2×2-case.

VIII. THE 2×2 - TABLE.

VIII. 1. Deducement of a measure of association.

The 2×2 - contingency table can be described in the following manner:

	B	$\bar{B}$
A	p_{11}	p_{12}
$\bar{A}$	p_{21}	p_{22}

(24)

We want to measure the association between the two attributes A and B . $\bar{A}$ and $\bar{B}$ are their negations (complements).

It is readily seen that the cell-probabilities can be expressed in the following way:

$$p_{11} = p_{1.}p_{.1} + (\Delta-1)p_{12}p_{21}$$
$$p_{12} = p_{1.}p_{.2} - (\Delta-1)p_{12}p_{21}$$
$$p_{21} = p_{2.}p_{.1} - (\Delta-1)p_{12}p_{21}$$
$$p_{22} = p_{2.}p_{.2} + (\Delta-1)p_{12}p_{21}$$

Here $\Delta = p_{11}p_{22}/p_{12}p_{21}$ is the cross-product ratio.
The exact independence hypothesis can be formulated as

$$H: \ p_{11}p_{22} = p_{12}p_{21} \quad (\Leftrightarrow \Delta = 1). \tag{25}$$

There are certain reasonable requirements a measure of association for (A,B) should satisfy in the 2×2-table (see [2] and [13],p.4). In most cases the following three demands are reasonable:

1) The measure must be a function of the conditional probability of B given A, $p_{11}/p_{11}+p_{12}$, and the conditional probability of B given $\bar{A}$, $p_{21}/p_{21}+p_{22}$, or, alternatively, of the conditional probability of A given B, $p_{11}/p_{11}+p_{21}$, and the conditional probability of A given $\bar{B}$, $p_{12}/p_{12}+p_{22}$.

2) The alternative measures in 1) must be equal.

3) The measure must change monotonically, for a given set of marginals $p_{1.}$ and $p_{.1}$, as the association becomes stronger.

The demands 1), 2) and 3) implies that the measure of association must be a one-to-one function H of the cross-product ratio Δ. (From Edwards, [2].)

$H(\Delta)$ is invariant under multiplication of rows and/or columns. That is, $H(\Delta)$ gives the same value to table (24) and the table:

	B	$\bar{B}$
A	$r_1c_1p_{11}$	$r_1c_2p_{12}$
$\bar{A}$	$r_2c_1p_{21}$	$r_2c_2p_{22}$

(26)

for all non-negative r_1, r_2, c_1, c_2 such that

$$r_1c_1p_{11} + r_1c_2p_{12} + r_2c_1p_{21} + r_2c_2p_{22} = 1 .$$

It is with this shown that the natural choice of a measure of association in the 2×2-table essentially is the cross-product ratio Δ .

We now mention four measures which are one-to-one functions of Δ .

Yule's coefficient of association:

$$d_1 = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\Delta - 1}{\Delta + 1} = 1 - \frac{2}{\Delta + 1} \qquad (27)$$

(d_1 is the ordinal measure γ in the 2×2-case).

Yule's coefficient of colligation:

$$d_2 = \frac{\sqrt{p_{11}p_{22}} - \sqrt{p_{12}p_{21}}}{\sqrt{p_{11}p_{22}} + \sqrt{p_{12}p_{21}}} = \frac{\sqrt{\Delta} - 1}{\sqrt{\Delta} + 1} = 1 - \frac{2}{\sqrt{\Delta} + 1} \qquad (28)$$

$$\rho = \ln \Delta \qquad (29)$$

and of course Δ itself.

Yule's two measures are strictly increasing when Δ increases.

Let us now define what we mean by positive and negative association between A and B in the table (24).

Definition 3. If $\Delta > 1$ $(p_{11}p_{22} > p_{12}p_{21})$ we say there are positive association (p.a.) between A and B . If $\Delta < 1$ A and B are negative associated (n.a.).

Some properties on Yule's two measures:

(i) $-1 \leq d_i \leq 1$, and $d_i > 0$ if p.a., $d_i < 0$ if n.a.; for $i = 1,2$.

(ii) $d_i = 0 \iff$ Exact independence

(iii) d_i assumes the value -1 when $p_{11} = 0$ or $p_{22} = 0$, for $i = 1,2$.
d_i assumes the value $+1$ when $p_{12} = 0$ or $p_{21} = 0$, for $i = 1,2$.

If we are not interested in the direction of dependence, but only in the degree of association, we can use one of the measures d_1^2, d_2^2 or ρ^2 .

VIII. 2. An alternative measure of association.

It can of course occur situations where other measures than those based on Δ can be applicable. Here we mention one:

Kendall's rank correlation coefficient: $\tau_b = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1.}p_{2.}p_{.1}p_{.2}}}$

or τ_b^2 if we are only interested in the degree of association. (For other measures see [2] and [6].)

IX. CONCLUDING COMMENTS.

As we have seen, most of the measures constructed from a given model have the property of being zero if there is no association relatively to the relevant features of association the measure is constructed for, even if other types of association possibly are present. This is what we had to expect, since we sharpen the "definition" of association in the different cases. Notice that for all situations, except VI, exact independence will imply that the measure is zero.

Finally we will again, as in I.2., strongly emphasize that when determining a measure of association for a given contingency table, one should choose that measure which gives the best information about the interesting features of association.

PART TWO:

TESTING ALMOST INDEPENDENCE

I. INTRODUCTION.

I.1. Some practical problems in larger investigations.

On testing for independence in a two-way contingency table it has been customary to use a chi-square test on the hypothesis of exact independence. As mentioned in part one, it is well known that when the number of observations is large the power of the chi-square test is so high that the hypothesis describing exact independence nearly always will be rejected. At larger investigations, say in the Central Bureau of Statistics of Norway in Oslo, the purpose of using tests for independence can be to decide which tables that are to be published from the investigation. If two factors are not associated, the value of the corresponding two-way table is too little to be published, because one can then be content with the marginal distribution for each factor's classification. Because of the high power of the chi-square test it is then not suitable as an assitance for setting up the tables in a large investigation.

As mentioned in part one, one should instead accept independence even if the factors are only almost independent, by which we mean that the degree of dependence is not materially signigicant with respect to the subject investigated.

This problem can be solved by extending the exact hypothesis to include cases where the degree of association is less than a certain limit, and thereafter develop tests for the extended independence-hypothesis.

Let us first give an ecample to show how the classical test can be less suitable for the purpose described above, when there are many observations.

Example 1, from [15].

We shall examine if there was significant association between participation and occupation at the Storting elections in Norway in 1969. The number of persons being interviewed was 2702. There are eight occupational groups. With participation we mean whether the interview-object has voted or not (according to the object's own statement). The result is given in the table below.

Table 1.

Occupational group / participation	1	2	3	4	5	6	7	8	Total
Voters	169	141	429	618	45	268	753	16	2439
Non-voters	19	16	43	56	14	36	75	4	263
Total	188	157	472	674	59	304	828	20	2702

Source: [15], table 17 and 19

For each occupational group we can calculate the relative frequency that voted/not voted. It gives the following table:

Table 2.

Occupational group	Voters	Non-Voters	Number of respondents
1	0.90	0.10	188
2	0.90	0.10	157
3	0.91	0.09	472
4	0.92	0.08	674
5	0.76	0.24	59
6	0.88	0.12	304
7	0.91	0.09	828
8	0.80	0.20	20
All occupations	0.90	0.10	2702

It seems likely to believe that the dependence between occupation and participation is very little. (There are few observations from the occupational groups, 5 and 8, which shows significant departure.) However, the usual chi-square test rejects the exact independence-hypothesis for significance levels greater than 0.005. The test gives therefore a result in contradiction to what we find reasonable by inspection of table 2 above. (We shall later see that the new tests proposed in this paper will lead us to accept the independence hypothesis.)

I.2. The multinomial situation.

The situation given below covers several of the cases where a chi-square test usually has been used, for instance

a) Testing of goodness of fit for a specified distribution to certain variables.

b) Testing of independence between two factors.

As mentioned earlier, we are particularly interested in case b). We consider the following situation: A sequence of n independent trials is executed. At each trial one and only one of r characteristics

$$A_1, A_2, \dots, A_r$$

can appear with probabilities $p_1, p_2, \dots, p_r$.

where $\sum_{i=1}^{r} p_i = 1$

Let X_i be the number of appearences of A_i in the sequence, and let $q_{in} = X_i/n$, for $i = 1, \dots, r$.

A priori we assume that the probabilities $p_1, \dots, p_r$ are unknown, and $p_i > 0$ for $i = 1, \dots, r$.

The general hypothesis to be tested is

$$H_o : p_i = \varphi_i(\theta) \quad \text{for } i = 1, \dots, r \tag{1}$$

where $\theta = (\theta_1, \ldots \theta_m) \in \Omega$ and Ω includes a non-degenerate interval of a m-dimensional real space R^m.
Each function φ_i is assumed to have continuous partial derivatives.
The number of observations n is assumed to be large.
Further the matrix $M = \{\varphi_i^{-\frac{1}{2}} \frac{\partial \varphi_i}{\partial \theta_s}\}$ of order $r \times m$ is of rank m at the true value of θ. Let $\hat{\theta} = (\hat{\theta}_1, \ldots \hat{\theta}_m)$ be an efficient estimator of θ in the sense of Rao ([16] p.285), and let

$$Z = \sum_{i=1}^{r} \frac{(X_i - n\varphi_i(\hat{\theta}))^2}{n\varphi_i(\hat{\theta})} \tag{2}$$

The asymptotic distribution of Z is χ^2 with $r-1-m$ degrees of freedom (Rao [16] p.325).

Let us in addition to the above conditions assume that for $\delta > 0$ there exists $\varepsilon > 0$ such that

$$\inf_{|\theta-\theta^o|>\delta} \sum_{i=1}^{k} p_i(\theta^o) \ln \frac{p_i(\theta^o)}{p_i(\theta)} \geq \varepsilon \tag{3}$$

where θ^o is the true value of θ and $|\theta-\theta^o|$ is the distance between θ and θ^o.
We then have that the maximum likelihood (m.l.) estimator $\tilde{\theta}$ is efficient and can be used in (2). For this result we refer to Rao, ([16] p.296). $\tilde{\theta}$ is the value of θ which maximizes
$[\prod_{i=1}^{r} \varphi_i(\theta)^{q_{in}}]^n$.
With approximate level ε we now reject H_o when

$$Z > z(r-1-m, \varepsilon) \tag{4}$$

where $z(r-1-m,\varepsilon)$ is $(1-\varepsilon)$-fractile in the chi-square distribution with $r-1-m$ degrees of freedom. This test is called the chi-square test for goodness of fit. The approximation to the chi-square distribution is usually applicable when $n\varphi_i(\hat{\theta}) \geq 5$ for $i = 1,2,\ldots r$.

Let us consider case b). The situation is described in part 1, ch.II, and we have a multinomial sequence of trials with $v \cdot w$ categories. The exact independence hypothesis is

$$H: p_{ij} = p_{i.}\, p_{.j} \quad \text{for } i = 1, \ldots v \text{ and } j = 1, \ldots w \tag{5}$$

We see that the hypothesis (5) has the same form as (1), with $\theta = (p_{1.}, \ldots p_{v-1.}, p_{.1}, \ldots, p_{.w-1})$ and $m = v + w - 2$. In this situation the conditions above are satisfied and the m.l. estimators for θ are efficient and equal to $X_{i.}/n$ and $X_{.j}/n$ for $p_{i.}$ and $p_{.j}$ respectively.

Hence, from (2) we have that

$$Z = \sum_{i=1}^{v} \sum_{j=1}^{w} \frac{\left(X_{ij} - n\frac{X_{i.}}{n}\cdot\frac{X_{.j}}{n}\right)^2}{n \cdot \frac{X_{i.}}{n} \cdot \frac{X_{.j}}{n}} = n\left(\sum_{i=1}^{v} \sum_{j=1}^{w} \frac{X_{ij}^2}{X_{i.} X_{.j}} - 1\right) \tag{6}$$

is approximately chi-square distributed with $(r-1)(w-1)$ degrees of freedom. The chi-square test for H now becomes

$$\text{Reject } H \text{ when } Z > z((v-1)(w-1), \epsilon) \tag{7}$$

I.3. Statistical hypotheses as idealized theory of "reality".

As mentioned earlier it seems that, where the data sample is extensive the chi-square test nearly always reject the exact independence-hypothesis.

We will now look further into this matter.

There are many situations where the mull-hypothesis only can be expected to be approximately true. In such situations one can say the statistical hypothesis is an "idealizing of reality", and will therefore be called an idealized hypothesis. An idealized hypothesis is then a hypothesis that cannot be expected to be exact true. Such a situation occur, for instance, usually when we test whether some gariables are normally distributed. Often it also seems reasonable to believe that two factors can be almost independent, but not exactly independent. If this is the case it can explain, to a certain extent, why the usual independence test rejects the exact hypothesis for large n.

Before looking closer at this, let us consider the following general situation.

Let X_n be a random variable with distribution depqnding on n and on a parameter θ which apriori lies in a set Ω. Let $\omega_0 \subset \Omega$ represent an idealized hypothesis as described above. δ is a test for

$$H_0 : \theta \in \omega_0 \qquad (8)$$

with critical region τ_n.

Definition 3. <u>δ is called consistent if $P_\theta(X_n \in \tau_n) \to 1$ as $n \to \infty$ for all $\theta \in \Omega - \omega_0$.</u>

It is readily shown that the chi-square test (4) for the hypothesis (1) is consistent. Especially the chi-square test for the exact independence hypothesis is consistent. This leads to the fact that in a very large sample, small and unimportant departures from the hypothesis (1) are almost certain to be detected. If then the hypothesis is an idealized hypothesis, the chi-square test will reject it nearly always when there are many observations. This of course is not a particular feature of the chi-square tests, but will apply to any consistent test for an idealized hypothesis.

When testing an idealized hypothesis we are generally interested in rejecting the hypothesis only when it is considerably wrong. This is, as mentioned in I.1., the case when we test for independence. The usual test for independence will however as we have given an example of, reject the exact hypothesis even in cases of almost independence.

One way of avoiding this difficulty, suggested by Hodges jr. & Lehmann, [9], is to extend the region of hypothesis to include situations close enough to the hypothesis so that the difference is not materially significant with regard to the specific problem we are investigating.

Let us in this connection turn back to the general situation with the idealized hypothesis (8). The extended region of hypothesis is represented by the set $\omega_1 \supset \omega_0$. If we know $\theta \in \omega_1$ we will still accept the idealized hypothesis H_0 . Let δ' be a test for the extended hypothesis

$$H_1 : \theta \in \omega_1$$

The significance level of the test will be $\max_{\theta \in \omega_1} \beta(\theta)$ where $\beta(\theta)$ is the power function of δ' .

What we are doing is to keep the power under a level α in situations unsignificantly different from H_0 which means that $\max_{\theta \in \omega_1} \beta(\theta) \leq \alpha$. For consistent tests (for H_0) the power will converge to 1 as $n \to \infty$ in the set $\omega_1 - \omega_0$.

Following the idea of Hodges jr. & Lehmann [9], one way to extend the region of hypothesis is to introduce into the parameterspace a measure, say $\Delta(\theta)$, of the "distance" of θ from H_0 reflecting at least roughly the materiality of departures from H_0 . H_1 is then defined as the set of θ for which $\Delta(\theta)$ does not exceed a specified value Δ_0 . The choice of Δ_0 will present problems similar to those encountered in choosing the alternative at which specified power is to be obtained.

I. 4. The independence problem.

We will treat the exact hypothesis of independence as described in I.4., that is, we intend to enlarge the exact hypothesis to situations indicating almost independence (abbreviated a.i.). As Hodges jr. & Lehmann suggest we are going to do this by means of a measure for the "distance" to the true parameter point from the exact hypothesis of independence. This will then be a measure for degree of association. The extended hypothesis is then defined to be the set of parameters for which this measure does not exceed a specified value c. The first problem to handle is to choose a measure of association. This is done in part one. We are thus left with two problems to be considered in this part.

(a) Extension of the region of hypothesis

This extension will of course depend on the measure of association that is chosen for the actual situation.

(b) Development of tests for the extended hypothesis satisfying at least approximately a given level α.

Besides proposing tests for almost independence we shall in chapter III develop confidence intervals for the various measures of association mentioned in [5]. Also in chapter III we discuss natural extensions to a.i. for the most important measures. First, however, we consider in chapter II the problem (b) for general extended hypotheses in the multinomial case. The conditions given in I.2. are assumed to hold in II. The theory developed in chapter II will be applied to testing and interval-estimation for measures of association. A three-decision procedure for the problem is also considered.

II. TESTS FOR EXTENDED HYPOTHESES.

II. 1. General case.

II.1. (i) The main theorem.

Consider a multinomial sequence of n trials with r classes described in I.2. The following notations will be used.

$X_n \xrightarrow{D} X$: X_n converges in distribution to X.

$X_n \xrightarrow{P} X$: X_n converges in probability to X.

$x(p)$: upper p-fractile in $N(0,1)$.

$z(k,p)$: upper p-fractile in $\chi^2(k)$ where $\chi^2(k)$ denotes the chi-square distribution with k degrees of freedom.

$\Phi(x)$: the distribution function for $N(0,1)$.

Let now d be a function in r variables admitting continuous partial derivatives of the first order. Let further

$$\sigma_d^2 = \sum_{i=1}^{r} p_i(a_i - \bar{a})^2 \tag{9}$$

where

$$a_i(p) = \frac{\partial d}{\partial p_i} \quad \text{for } i = 1,\dots,r ,$$

and

$$\bar{a}(p) = \sum_{i=1}^{r} a_i p_i .$$

Consistent estimators (called C-estimators) for $d(p)$ and σ_d^2 are given by respectively

$$\hat{d}_n = d(q_n) \quad \text{and} \quad S_d^2 = \sum_{i=1}^{r} q_{in} (\hat{a}_i - \hat{a})^2 \tag{10}$$

where

$$q_{in} = X_i/n \quad \text{and} \quad q_n = (q_{1n},\dots,q_{rn}) ,$$

$$\hat{a}_i = a_i(q_n) \quad \text{and} \quad \hat{\bar{a}} = \bar{a}(q_n) \; .$$

The main result for our problem can now be stated as follows.

THEOREM 1. <u>Assume $\sigma_d > 0$, i.e. there exists an i such that</u>

$$a_i(p) \neq \bar{a}(p) \tag{11}$$

<u>Then</u>

$$1) \; \frac{\sqrt{n}(\hat{d}_n - d)}{\sigma_d} \xrightarrow{D} N(0,1) \quad \text{and} \quad 2) \; \frac{\sqrt{n}(\hat{d}_n - d)}{S_d} \xrightarrow{D} N(0,1) \; .$$

To be able to prove the theorem we need a result which follows from Rao ([16], p.321).

LEMMA 3. Let T_n be a k-dimensional statistic $(T_{1n},\ldots,T_{kn})$ such that

$$\sqrt{n}(T_n - \theta) = \{\sqrt{n}(T_{1n} - \theta_1),\ldots,\sqrt{n}(T_{kn} - \theta_k)\} \xrightarrow{D} N_k(0,\Sigma)$$

where Σ is a covariance-matrix with elements $\sigma_{ij}(\theta)$. Let further g be a function of k variables with continuous partial derivatives of the first order. Then

1) $$\sqrt{n}\, V_n = \sqrt{n}\, [g(T_n) - g(\theta)] \xrightarrow{D} N(0,\sqrt{\nu(\theta)})$$

provided $\nu(\theta) \neq 0$ where

$$\nu(\theta) = \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma_{ij}(\theta) \frac{\partial g}{\partial \theta_i} \cdot \frac{\partial g}{\partial \theta_j} \; .$$

2) If σ_{ij} is a continuous function of θ and $\nu(\theta) \neq 0$ then

$$\frac{\sqrt{n}\, V_n}{\sqrt{\nu(T_n)}} \xrightarrow{D} N(0,1) \; .$$

Proof of theorem 1.

Let $p = (p_1,\dots,p_r)$

We define the following r-dimensional random variable:

$U_n = (U_{1n},\dots,U_{rn})$ where

$$U_{in} = \begin{cases} 1 & \text{if i'th characteristic appear in trial number } n. \\ 0 & \text{otherwise.} \end{cases}$$

$E(U_{in}) = p_i$, $var(U_{in}) = \sigma_{ii}(p) = p_i(1-p_i)$ and $cov(U_{in},U_{jn}) = \sigma_{ij}(p) = -p_i . p_j$.

That is, $U_1, U_2, \dots$ are independent identically distributed (i.i.d.) random variables with expection p and covariancematrix $\Sigma = \{\sigma_{ij}(p)\}$.

Further we see that $\bar{U}_{in} = \frac{1}{n} \sum_{k=1}^{n} U_{ik} = \frac{X_i}{n} = q_{in}$,

and with it $\bar{U}_n = (\bar{U}_{1n},\dots,\bar{U}_{rn}) = (q_{1n},\dots,q_{rn}) = q_n$.

From the multivariate central limit theorem for i.i.d. random variables (see for example Rao [16], 2c.) it follows that

$$\sqrt{n}(\bar{U}_n - p) = \sqrt{n}(q_n - p) = (\sqrt{n}(q_{1n}-p_1),\dots,\sqrt{n}(q_{rn}-p_r)) \xrightarrow{D} N_r(0,\Sigma).$$

The conditions in lemma 3 are now fulfilled with $T_n = q_n$, $\theta = p$ and $g = d$. In addition we find that

$$\nu(p) = \sum_{i=1}^{r} \sum_{j=1}^{r} \sigma_{ij}(p) \frac{\partial d}{\partial p_i} \cdot \frac{\partial d}{\partial p_j} = \sum_{i=1}^{r} p_i(1-p_i) a_i^2 - \sum_{i \neq j} p_i p_j a_i a_j$$

Simple calculation gives

$$\nu(p) = \sum_{i=1}^{r} p_i (a_i - \bar{a})^2 = \sigma_d^2$$

Since $\sigma_d > 0$, we have $\nu(p) > 0$.

Then, from lemma 3:

1) $\sqrt{n}(\hat{d}_n - d)/\sigma_d \xrightarrow{D} N(0,1)$.

In addition σ_{ij} is a continuous function of p so that lemma 3-2) can be applied, giving

2) $\sqrt{n}(\hat{d}_n - d)/\sqrt{\nu(q_n)} = \sqrt{n}(\hat{d}_n - d)/S_d \xrightarrow{D} N(0,1)$.

Q.E.D.

Let us assume that $d(x) \in [M_1, M_2]$ for $x \in S = \{(p_1, \dots, p_r) | p_i > 0, \Sigma p_i = 1\}$.

Then it is seen that condition (11) in theorem 1 implies $d(x) \in \langle M_1, M_2 \rangle$, $x \in S$. Especially if d is non-negativ, (11) implies that $d(p) > 0$.

II.1. (ii). The N-test for an extended hypothesis.

We return to the problem of extended, approximately idealized hypotheses. Since $\sum_i p_i = 1$, the point $p = (p_1, \dots, p_r)$ lies on a hyperplane in the r-dimensional euclidean space. The standard hypothesis can be formulated as follows (see [9]).

H: p lies on a specified surface ζ .

(Usually this will be an idealized hypothesis as defined in I.3.). Instead of testing H, we are interested in testing an extended hypothesis that p lies in a region close enough to ζ such that ζ "almost is true".

Let now d be a non-negativ function of p, considered as a measure of the distance to p from ζ .

In the contingency table, d is a measure of association. A natural assumption should then be: $d(p) = 0 \iff p \in \zeta$. Unfortunately as shown in part 1 this is not true for a number of measures of association. On the other hand we will always have $p \in \zeta \implies d(p) = 0$, where ζ now denotes the exact hypothesis of independence.

We must assume in order to use theorem 1, that d possesses continuous partial derivatives of the first order. The extended hypothesis can now be formulated as follows.

$$H^* : \ d(p) \leq c \qquad (13)$$

Here c is chosen so that "H is almost true" under H^*. We propose the following test, called the normal-test (N-test), for H^*: Reject H^* when

$$\sqrt{n}(\hat{d}_n - c)/S_d > x(\alpha) \qquad (14)$$

The powerfunction $\beta_n(p)$ for the N-test has the following asymptotical property:

THEOREM 2. <u>Assume $\sigma_d > 0$. Then</u>

$$\lim_{n\to\infty} \beta_n(p) = \begin{cases} 0 & \text{if } d(p) < c \\ \alpha & \text{if } d(p) = c \\ 1 & \text{if } d(p) > c \end{cases}$$

<u>Proof</u>.

<u>a) $d(p) = c$</u>.

$$\lim \beta_n(p) = \lim_{n\to\infty} P_p\left(\frac{\sqrt{n}(\hat{d}_n - c)}{S_d} > x(\alpha)\right) = 1 - \Phi(x(\alpha)) = \alpha$$

from theorem 1.

<u>b) $d(p) < c$</u>.

d is continuous giving $\hat{d}_n \xrightarrow{P} d(p)$ which is equivalent with

$$\lim_{n\to\infty} P(\hat{d}_n - d > x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x > 0 \end{cases}$$

This gives

$$0 \leq \lim_{n\to\infty} \beta_n(p) \leq \lim_{n\to\infty} P_p(\hat{d}_n - c > 0) = \lim_{n\to\infty} P_p(\hat{d}_n - d > c - d) = 0 .$$

<u>c) $d(p) > c$</u>.

S_d is a continuous function in q_n implying $S_d \xrightarrow{P} \sigma_d > 0$.

We therefore have:

$$V_n = \frac{\hat{d}_n - c}{S_d} \xrightarrow{P} \frac{d-c}{\sigma_d} = a > 0 .$$

Let $Y_n = \sqrt{n}\,V_n$. $V_n - a = (Y_n - \sqrt{n}\,a)/\sqrt{n} \xrightarrow{D} 0$.

Let now $0 < \varepsilon < a$, and $b = a - \varepsilon$.

$$(Y_n - \sqrt{n}\,a)/\sqrt{n} > -\varepsilon \iff Y_n > \sqrt{n}\,b \quad \text{which imply}$$

$$\lim_{n\to\infty} P(Y_n > \sqrt{n}\,b) = 1 .$$

Since $b > 0$ the result follows.

Q.E.D.

By applying theorem 1 we can construct confidence intervals for a chosen measure $d(p)$ with confidence level equal to $1-\alpha$ asymptotically. Assume that $M_1 < d(p) < M_2$ and that $\sigma_d > 0$. From theorem 1

$$\lim_{n\to\infty} P(\hat{d}_n - \frac{S_d}{\sqrt{n}}\, x(\tfrac{\alpha}{2}) < d < \hat{d}_n + \frac{S_d}{\sqrt{n}}\, x(\tfrac{\alpha}{2})) = 1-\alpha .$$

A confidence interval for $d(p)$ with asymptotic confidence level equal to $1-\alpha$ is hereby given:

$$d(p) \in \langle \max\,(M_1, \hat{d}_n - \frac{S_d}{\sqrt{n}}\, x(\tfrac{\alpha}{2}))\,,\ \min\,(M_2, \hat{d}_n + \frac{S_d}{\sqrt{n}}\, x(\tfrac{\alpha}{2}))\rangle . \qquad (15)$$

In the next shapter we shall consider two-way contingency tables. Usually then M_1 is equal to 0, but situations where it is natural to separate between directions of association will occur frequently. In such cases the measures can take negative values. They will vary from -1 to 1. Onesided confidence intervals for positive $d(p)$ is deduced from the following equality:

$$\lim_{n\to\infty} P(\sqrt{n}\ \frac{(\hat{d}_n - d)}{S_d} < x(\alpha)) = 1-\alpha ,$$

which gives a $(1-\alpha)$ confidence interval of the form

$$d(p) \in \langle \max\left(0, \hat{d}_n - \frac{S_d}{\sqrt{n}} x(\alpha)\right), M\rangle \tag{16}$$

if $d(p) < M$.

Let now $k(q_n) = \hat{d}_n - S_d\, x(\alpha)/\sqrt{n}$ and assume $k(q_n) > 0$.

For $c < k(q_n)$, the hypothesis

$$H^* : d(p) \leq c$$

will be rejected by N-test at level α, since

$$c < k(q_n) \iff \sqrt{n}\,(\hat{d}_n - c)/S_d > x(\alpha) .$$

In other words, the set $\{d \leq k(q_n)\}$, is the maximum extended region of hypothesis that will be rejected when observing q_n.

If then $k(q_n) \leq 0$ then all hypotheses $H : d(p) \leq c$, $c > 0$ is accepted (at level α).

II. 2. A special case.

Under certain conditions we can apply the theory from Neyman ([14], ch.4) on the hypothesis

$$H' : d(p) = c .$$

Let us assume that the distance measure d is given. To use the theory of Neyman it is sufficient (it seems likely to believe that in view of the theory in II.1. it is not necessary), to find $\theta_1, \ldots, \theta_{r-2}$ and functions $f_1, \ldots, f_r$ such that

$$p_i = f_i(d, \theta_1, \ldots, \theta_{r-2}) \quad \text{for } i = 1, \ldots, r \tag{17}$$

In addition the functions f_i must have continuous partial derivatives of second order.

This is the case that Hodges jr. & Lehmann, [9], consider, though it seems that they have not been aware of the problem of finding such functions $f_1, \ldots, f_r$. For testing independence in a two-way contingency table we have not succeeded, with our choice of measures of association, in finding $\theta_1,\dots,\theta_{r-2}$ so that (17) holds. Therefore, the theory in this section will not be used when testing almost independence. One of the situations where (17) is satisfied is the case where ζ consists of a single point p^o and $d^{\frac{1}{2}}$ is the euclidean distance from p^o to p.

LEMMA 4. Let $d(p) = \sum_{i=1}^{r}(p_i - p_i^o)^2$. Then there exists (polar coordinates) $\theta_1,\dots,\theta_{r-2}$ and functions $f_1,\dots,f_r$ so that (17) is fulfilled.

Proof.

Let $\mu_i = p_i - p_i^o$ for $i = 1,\dots,r$. Then $d = \sum_{i=1}^{r} \mu_i^2$ and $\sum_{i=1}^{r}\mu_i = 0$. There exists polar coordinates $\theta,\theta_1,\dots,\theta_{r-2}$ so that we have:

$$\begin{aligned}
\mu_1 &= \sqrt{d}\,\sin\theta_o \\
\mu_2 &= \sqrt{d}\,\cos\theta_o\,\sin\theta_1 \\
&\vdots \\
\mu_{r-1} &= \sqrt{d}\,\cos\theta_o\,\cos\theta_1 \dots \cos\theta_{r-3}\,\sin\theta_{r-2} \\
\mu_r &= \sqrt{d}\,\cos\theta_o\,\cos\theta_1 \dots \cos\theta_{r-3}\,\cos\theta_{r-2}
\end{aligned} \qquad (*)$$

where $-\frac{\pi}{2} \le \theta_o \le \frac{\pi}{2}$ so $\cos\theta_o \ge 0$.

Define $a(\theta_1,\dots\theta_{r-2}) = \sin\theta_1 + \cos\theta_1\sin\theta_2 + \dots + \cos\theta_1 \dots \cos\theta_{r-3}\sin_{r-2} + \cos\theta_1 \dots \theta_{r-2}$ and let $\theta = (\theta_1,\dots,\theta_{r-2})$. Assume first that $a(\theta) \neq 0$ and $d > 0$.

Now using the fact $\sum_{i=1}^{r}\mu_i = 0$ and $\cos\theta_o \ge 0$ we see that

$$\begin{aligned}
\cos\theta_o &= 1/\sqrt{1+a^2(\theta)} \\
\sin\theta_o &= a(\theta)/\sqrt{1+a^2(\theta)}
\end{aligned} \qquad (**)$$

This holds trivially when $a(\theta) = 0$, so (**) is valid for all $a(\theta)$ and $d > 0$.

Let now $f_1,\dots,f_r$ be functions of (d,θ) given by

$$f_1(d,\theta_1,\dots,\theta_{r-2}) = p_1^o - \frac{\sqrt{d}\ a(\theta)}{\sqrt{1+a^2(\theta)}}$$

$$f_2(d,\theta_1,\dots,\theta_{r-2}) = p_2^o + \sqrt{\frac{d}{1+a^2(\theta)}}\sin\theta_1$$

$$\vdots$$

$$f_{r-1}(d,\theta_1,\dots,\theta_{r-2}) = p_{r-1}^o + \sqrt{\frac{d}{1+a^2(\theta)}}\cos\theta_1\dots\cos\theta_{r-2}$$

$$f_r(d,\theta_1,\dots,\theta_{r-2}) = 1 - \sum_{i=1}^{r-1} f_i(d,\theta_1,\dots,\theta_{r-2})$$

We see that $p_i = p_i^o + \mu_i = f_i(d,\theta)$ for $i = 1,\dots,r$.

If $d = 0$ then $p_i = p_i^o = f_i$ so we have

$$p_i = f_i(d,\theta)\ ,\quad i = 1,\dots,r \quad \text{for all values of } \theta \text{ and } d \geq 0 .$$

Q.E.D.

In this chapter we will assume that (17) is true and that d has continuous partial derivatives of second order.

The next section gives a short review of Neyman's BAN-estimators under the condition of (17) and $H' : d = c$.

II. 2 (i). BAN-estimators.

Neyman introduced the term BAN-estimator, where BAN is an abbreviation of "best asymptotically normal".

Definition 4. A function $\hat{\theta}_k$ of q_n not depending directly on n is called a BAN-estimator of the parameter θ_k if it satisfies the following four conditions:

(i) $\hat{\theta}_k \xrightarrow{P} \theta_k$

(ii) $\sqrt{n}\,(\hat{\theta}_k - \theta_k) \xrightarrow{D} N(0,\sigma_k^2)$

(iii) Let ν be another function satisfying (i) and (ii) with σ^2 equal to the asymptotical variance of $\sqrt{n}\,\nu$. Then $\sigma_k \leq \sigma$

(iv) $\hat{\theta}_k$ possess continuous partial derivatives with respect to each q_{in}, $i = 1,\ldots,r$

Neyman shows that the following three types of estimators are BAN.

A) ML-estimator $\hat{\theta}_k$. Let $\hat{\theta} = (\hat{\theta}_1,\ldots,\hat{\theta}_{r-2})$. Then $\hat{p}_i = f_i(c,\hat{\theta})$.

B) Minimum chi-square estimator $\bar{\theta}_k \cdot \bar{\theta} = (\bar{\theta}_1,\ldots,\bar{\theta}_{r-2})$. $\bar{p}_i = f_i(c,\bar{\theta})$.

C) Modified minimum chi-square estimator $\theta_k^* \cdot \theta^* = (\theta_1^*,\ldots,\theta_{r-2}^*)$. $p_i^* = f_i(c,\theta^*)$.

In this case we have that:

1) $\hat{\theta}$ maximizes $[\prod_{i=1}^{r} f_i(c,\theta,\ldots,\theta_{r-2})]^{q_{in}n}$

2) $\bar{p}$ minimizes $n\,\Sigma\,(q_{in}-p_i)^2/p_i$ under the condition $d(p) = c$

3) p^* minimizes $n\,\Sigma\,(q_{in}-p_i)^2/q_{in}$ under the condition $d(p) = c$

A fourth type of BAN-estimator is also given by Neyman (see [14], theorem 5 and 6). Let $p' = (p_1,\ldots,p_{r-1})$, and let us assume there are μ restrictions on $p_1,\ldots,p_{r-1}$ (in this case $\mu = 1$).

$$F_t(p') = 0 \quad \text{for } t = 1,2,\ldots,\mu \quad (\mu \leq r-1) \tag{18}$$

$$\text{Let } Q = \sum_{i=1}^{r} (X_i - np_i)^2/X_i \tag{19}$$

F_t is assumed to have continuous partial derivatives of second order. Let now

$$F_t^*(p',q_n') = F_t(q_n') + \sum_{i=1}^{r-1} b_{t,i}(p_i - q_{in}) \tag{20}$$

where $q_n' = (q_{in},\ldots,q_{r-1,n})$ and $b_{t,i} = ({}^{\partial F}t/\partial p_i)|p' = q_n'$.
Neyman shows that minimizing of Q under the linear restrictions

$$F_t^*(p',q_n') = 0 \tag{21}$$

leads to BAN-estimators $\hat{p}_i$ of p_i when p' satisfy (18).
Let $\hat{p}_r = 1 - \sum_{i=1}^{r-1} \hat{p}_i$. The fourth type, D , of BAN-estimators for p is now equal to $\hat{p} = (\hat{p}_1,\ldots,\hat{p}_r)$.
Let $\tilde{p} = (\tilde{p}_1,\ldots,\tilde{p}_r)$ be a BAN-estimator of p under H' of type A, B, C or D , and assume that d has continuous partial derivatives of second order. Then we have from ([14], lemma 12, p.268):

LEMMA 5 Under H' : $d(p) = c$:

$$\chi_b^2 = n \sum_{i=1}^{r} \frac{(q_{in}-\tilde{p}_i)^2}{\tilde{p}_i} \xrightarrow{D} \chi^2(1)$$

and

$$\chi_b^2 = n \sum_{i=1}^{r} \frac{(q_{in}-\tilde{p}_i)^2}{q_{in}} \xrightarrow{D} \chi^2(1) .$$

II.2 (ii). Asymptotically equivalent tests for an extended hypothesis.

Hodges jr. & Lehmann ([9], p.267) suggests the following tests for the extended hypothesis H* , (which now can be applied since (17) is assumed to hold):

TEST I :

Reject H* if

$$\hat{d}_n > c \tag{21a}$$

and

$$n \sum_{i=1}^{r} \frac{(q_{in}-\tilde{p}_i)^2}{\tilde{p}_i} > z(1,2\alpha) . \tag{21b}$$

TEST II :

Reject H* if

$$\hat{d}_n > c \tag{22a}$$

and $$n \sum_{i=1}^{r} \frac{(q_{in}-\tilde{p})^2}{q_{in}} > z(1,2\alpha) \ . \qquad (22b)$$

Let now $\hat{\varphi}_n = d(q_{in},\dots,q_{r-1,n}, 1-\sum_{i=1}^{r-1} q_{in})$, and let $b_i(q_n') = \frac{\partial \hat{\varphi}_n}{\partial q_{in}}$, for $i = 1,\dots,r-1$. BAN-estimators under H' of type D are obtained by minimizing Q given by (19) under the restrictions

$$\begin{aligned} &\hat{\varphi}_n + \sum_{i=1}^{r-1} b_i(q_n)(p_i-q_{in}) - c = 0 \\ &p_r = 1 - \sum_{i=1}^{r-1} p_i \end{aligned} \qquad (23)$$

We must have $q_{in} > 0$ for $i = 1,\dots,r$, otherwise Q is undefined. This means that q_n' is a inner point in the set S given in II.1(i) and hence $b_i(q_n') = \hat{a}_i - \hat{a}_r$ so that the restrictions (23) are equivalent with

$$\begin{aligned} &\hat{d}_n + \sum_{i=1}^{r} \hat{a}_i(p_i-q_{in}) - c = 0 \\ &p_r = 1 - \sum_{i=1}^{r-1} p_i \end{aligned} \qquad (24)$$

The following interesting result is now true.

LEMMA 6 <u>Let $\hat{p} = (\hat{p}_1,\dots,\hat{p}_r)$ be BAN-estimator of p under H' of type D</u>.

<u>Then</u>

$$Z_1 = \min_{(23)} Q = n \sum_{i=1}^{r} \frac{(q_{in}-\hat{p}_i)^2}{q_{in}} = n \frac{(\hat{d}_n-c)^2}{S_d^2} \ .$$

<u>Note</u>. As a result of lemma 6 we have, even if (17) is not explicitly assumed, that under H' and (11)

$Z_1 \xrightarrow{D} \chi^2(1)$, when d has continuous partial derivatives of second order, which is true for most of the measures of association considered in part 1.

Proof.

$Q = n \sum_{i=1}^{r} (q_{in}-p_i)^2/q_{in}$ is to be minimized under the restrictions (24). We apply the method of Lagrange and form the Lagrange function

$$F(p) = Q + \lambda_1(\sum_{i=1}^{r} p_i - 1) + \lambda_2(\hat{d}_n + \sum_{i=1}^{r} \hat{a}_i(p_i - q_{in}) - c)$$

The first order conditions are:

$$\frac{\partial F}{\partial p_i} = -\frac{2n}{q_{in}}(q_{in}-p_i) + \lambda_1 + \lambda_2\hat{a}_i = 0 \quad \text{for } i = 1,\dots,r . \qquad (*)$$

Now for each $p \neq p^o$ we see that

$$F(p) - F(p^o) > \sum_{i=1}^{r} \frac{\partial F}{\partial p_i}\Big|_{p=p^o} (p_i - p_i^o)$$

so that F is a (strictly) convex function (see [19], p.231) Then we know ([19], p.265) that a value $\hat{p}$ of p satisfying (*) and (24) will minimize Q under (24). From (*):

$$\hat{p}_i = q_{in}(1-\frac{\lambda_1}{2n}) - \frac{\lambda_2\hat{a}_i q_{in}}{2n} , \quad \text{for } i = 1,\dots,r .$$

We determine λ_1 and λ_2 so that $\hat{p}$ satisfies (24):

$$\Sigma\, q_{in} \frac{\lambda_1}{2n} = -\frac{\lambda_2}{2n}\hat{\bar{a}} \Rightarrow \lambda_1 = -\lambda_2\,\bar{a} \qquad (**)$$

and

$$\frac{1}{2n}\sum_{i=1}^{r}(\hat{\bar{a}}_i q_{in}\lambda_1 + \hat{a}_i^2\, q_{in}\lambda_2) = d_n - c$$

$$\Rightarrow \lambda_1\hat{\bar{a}} + \lambda_2 \sum_{i=1}^{r} q_{in}\hat{a}_i^2 = 2n(\hat{d}_n - c) \qquad (***)$$

(**) and (***) give

$$\lambda_2 = \frac{2n(\hat{d}_n - c)}{S_d^2} \quad \text{and} \quad \lambda_1 = -\frac{2n\hat{\bar{a}}}{S_d^2}[\hat{d}_n - c] .$$

This implies the following expression of $q_{in} - \hat{p}_i$:

$$q_{in} - \hat{p}_i = \frac{(\hat{d}_n - c)}{S_d^2}(-q_{in}\hat{\bar{a}} + q_{in}\hat{a}_i) = \frac{q_{in}(\hat{a}_i - \hat{\bar{a}})(\hat{d}_n - c)}{S_d^2}$$

Hence

$$Z_1 = \min_{(23)} Q = n \sum_{i=1}^{r} \frac{q_{in}^2(\hat{a}_i-\hat{\bar{a}})^2(\hat{d}_n-c)^2}{q_{in}\, S_d^4} = \frac{n(\hat{d}_n-c)^2}{S_d^4} \sum_{i=1}^{r} q_{in}(\hat{a}_i-\hat{\bar{a}})^2 = n(\hat{d}_n-c)^2/S_d^2$$

Q.E.D.

It will be shown that asymptotically the eight tests I and II are in a certain sense equivalent. To define precisely the notion of asymptotically equivalent tests, consider the general situation where X_n is a random variable with distribution depending on n and of a parameter $\theta \in \Omega$. The hypothesis to be tested is

$$H: \theta \in \omega_o \quad \text{against} \quad \theta \in \Omega - \omega_o$$

Let φ_1^n, φ_2^n be two non-randomized tests for H. $\varphi_i^n(x)$ is the probability of rejecting the hypothesis having observed $X_n = x$.

<u>Definition 5. φ_1^n and φ_2^n are called asymptotically equivalent (a.e.) tests if for all $\theta \in \Omega$:</u>

$$\lim_{n\to\infty} P_\theta(\varphi_1^n \neq \varphi_2^n) = \lim_{n\to\infty} P_\theta(\varphi_1^n=1 \cap \varphi_2^n=0) + \lim_{n\to\infty} P_\theta(\varphi_1^n=0 \cap \varphi_2^n=1) = 0$$

The following result will be used to show the equivalence of tests I and II.

LEMMA 7. <u>The 2α-level tests in (21b) and (22b) for $H': d = c$ are (pairwise) asymptotically equivalent.</u>

This result follows directly from ([14], theorem 7) and from the fact that if φ_1^n, φ_3^n are a.e. and φ_2^n, φ_3^n the same then φ_1^n, φ_2^n are a.e.

LEMMA 8. <u>The tests I and II for H^* are asymptotically equivalent.</u>

<u>Proof</u>.

Let φ_1^n and φ_2^n be two of the tests in (21b) and (22b), arbitrarily chosen. The corresponding tests for H^* are denoted by ψ_1^n and ψ_2^n :

$$\psi_i^n = 1 \text{ if and only if } \hat{d}_n > c \text{ and } \varphi_i^n = 1 .$$

Hence: $\lim_{n\to\infty} P\,(\psi_1^n \neq \psi_2^n/_p)$

$$= \lim_{n\to\infty} P(\hat{d}_n > c \cap \varphi_1^n = 1 \cap \varphi_2^n = 0/p) + \lim_{n\to\infty} P(\hat{d}_n > c \cap \varphi_1^n = 0 \cap \varphi_2^n = 1/p)$$

$$\leq \lim_{n\to\infty} P(\varphi_1^n = 1 \cap \varphi_2^n = 0/p) + \lim_{n\to\infty} P(\varphi_1^n = 0 \cap \varphi_2^n = 1/p) = 0$$

for all p

Q.E.D.

Now, $z(1,2\alpha) = x^2(\alpha)$, so the N-test is the same as test II when $\tilde{p}$ is a BAN-estimator of type D. Hence the N-test is a.e. with the seven other tests in I and II under the assumption of (17). Let now $\beta_{k,n}(p)$, for $k = 1,\ldots,8$ be the power functions of the eight tests I and II. They have the same asymptotical property as the power function for the N-test as shown in the following result.

THEOREM 3. Assume $\sigma d > 0$. Then for $k = 1,\ldots,8$:

$$\lim_{n\to\infty} \beta_{k,n}(p) = \begin{cases} 0 & \text{for } d(p) < c \\ \alpha & \text{for } d(p) = c \\ 1 & \text{for } d(p) > c \end{cases}$$

Proof.

Let the power functions for the tests I and II be denoted by respectively β_n^1 and β_n^2. That is,

$$\beta_n^1(p) = P_p(Z_{1n} > z_o \cap \hat{d}_n > c)$$

$$\beta_n^2(p) = P_p(Z_{2n} > z_o \cap \hat{d}_n > c)$$

where $z_o = z(1,2\alpha)$ and

$$Z_{1n} = n \sum_{i=1}^{r} \frac{(q_{in}-\tilde{p}_i)^2}{\tilde{p}_i} \quad \text{and} \quad Z_{2n} = n \sum_{i=1}^{r} \frac{(q_{in}-\tilde{p}_i)^2}{q_{in}}$$

a) $\underline{d(p) < c}$

$\hat{d}_n - d \overset{D}{\to} 0$ and $c - d > 0$ which implies

$$0 \leq \lim_n \beta_n^i(p) \leq \lim_n P_p(\hat{d}_n > c) = \lim_n P_p(\hat{d}_n - d > c - d) = 0 .$$

b) $\underline{d(p) > c}$

From Neyman ([14], lemma 14) we have that $\lim_n P(Z_{in} > z_o) = 1$ for $d(p) \neq c$. In this case we also have that $\lim_n P(\hat{d}_n > c) = 1$ since $c - d < 0$. Hence $\lim \beta_n^i(p) = 1$ for $i = 1,2$.

c) $\underline{d(p) = c}$

Let us return to the notations $\beta_{n,k}$ and let $\beta_{8,n}$ be the power function of test II, type D. $\beta_{8,n}$ is hence, from lemma 6 the power function of the N-test. Therefore $\beta_{n,8}(p) \underset{n\to\infty}{\to} \alpha$. Let now $Z_n^* = n(\hat{d}_n - c)^2/S_d^2$ and let Z_n be anyone of the other seven quantities in (21b) and (22b) with β_n as the corresponding power function of the test for H^*, i.e. β_n is anyone of $\beta_{i,n}$ for $i = 1,\ldots,7$. Now by using the asymptotical equivalence with the N-test we find that

$$\lim_{n\to\infty} \beta_n(p) = \lim_{n\to\infty} P_p(Z_n > z_o \cap \hat{d}_n > c) = \lim_{n\to\infty} P_p(Z_n > z_o \cap \hat{d}_n > c \cap Z_n^* > z_o).$$

and

$$\alpha = \lim_{n\to\infty} \beta_{8,n}(p) = \lim P_p(Z_n^* > z_o \cap \hat{d}_n > c) = \lim_n P_p(Z_n^* > z_o \cap \hat{d}_n > c \cap Z_n > z_o).$$

Hence $\lim_n \beta_n(p) = \lim_{n\to\infty} \beta_{8,n}(p) = \alpha$.

Q.E.D.

II. 2(iii) Comments and an example.

As mentioned earlier we have not been able to show that (17) is satisfied for our choice of measures of association. We will therefore use the N-test in that situation. It is also worth noticing that for applying Neyman's theory one of the requirements is that the distance measure d has continuous partial derivatives of second order, while it is sufficient that d only possesses continuous partial derivatives of first order to apply the N-test. (For testing almost independence, however, this is no problem with the measures in part 1.)

Let us give an example, (from [9]) of choice of d for a completely specified hypothesis and application of the N-test.

Let the idealized hypothesis be

$$H : p_1 = \dots = p_r = \frac{1}{r}$$

and choose $d = \sum_{i=1}^{r} (p_i - \frac{1}{r})^2$. This is a special case of lemma 4, so with this choice (17) is satisfied. The extended hypothesis is:

$$H^* : \sum_{i=1}^{r} (p_i - \frac{1}{r})^2 \leq c \, .$$

We find $\hat{a}_i = 2(q_{in} - \frac{1}{r})$, $\hat{\hat{a}} = 2 \sum_{i=1}^{r} q_{in}^2 - 2 \cdot \frac{1}{r}$

so the N-test is to reject H^* when

$$\sqrt{n} \, \frac{\sum_{i=1}^{r} (q_{in} - \frac{1}{r})^2 - c}{2\{\sum_{i=1}^{r} q_{in}(q_{in} - \sum_{i=1}^{r} q_{in}^2)^2\}^{\frac{1}{2}}} > x(\alpha) \, .$$

II. 3. A test procedure for a three-decision problem.

Sometimes one can be interested in taking one of three decisions of the type:

1) Assert $d < c_1$ or 2) assert $d > c_2$ $(c_2 > c_1)$ or
3) make no inference

A test procedure for this problem is proposed:

1) Assert $d < c_1$ if $\sqrt{n}(\hat{d}_n - c_1)/S_d < -x(\alpha)$ (25)

2) Assert $d > c_2$ if $\sqrt{n}(\hat{d}_n - c_2)/S_d > x(\alpha)$ (26)

3) If neither (25) nor (26) is valid no inference is made.

We call this procedure the N_3-method.

The N_3-method has the following asymptotical property.

THEOREM 4. Assume that $\sigma_d > 0$. Then

$$\lim_{n\to\infty} P\text{ (At least one false assertion)} = \begin{cases} \alpha & \text{when } d = c_1 \text{ or } d = c_2 \\ 0 & \text{otherwise} \end{cases}$$

Proof. Let P_a denote the probability of at least one false assertion and let $U_n = \sqrt{n}(\hat{d}_n - c_1)/S_d$ and $V_n = \sqrt{n}(\hat{d}_n - c_2)/S_d$.

(i) $d = c_1$

$$\lim_n P_a = \lim_n P(U_n < -x(\alpha)) + \lim_n P(V_n > x(\alpha)) = \alpha + \lim_n P(V_n > x(\alpha))$$

Now, since $\hat{d}_n \xrightarrow{D} d$: $0 \leq \lim_n P(V_n > x(\alpha)) \leq \lim P(\hat{d}_n > c_2) = 0$.

(ii) $d = c_2$ Completely similar to (i) we get:

$$\lim_n P_a = \lim_n P(V_n > x(\alpha)) = \alpha .$$

(iii) $c_1 < d < c_2$

$$\lim_n P_a \leq \lim_n P(\hat{d}_n - c_1 < 0) + \lim_n P(\hat{d}_n - c_2 < 0) = 0 .$$

(iv) $d < c_1$

$$\lim_n P_a = \lim_n P(V_n > x(\alpha)) \leq \lim_n P(\hat{d}_n > c_2) = 0 .$$

(v) $d > c_2$

$$\lim_n P_a = \lim_n P(U_n < -x(\alpha)) \leq \lim_n P(\hat{d}_n < c_1) = 0 .$$

Q.E.D.

One application of this is to the independence-problem where one can choose c_1 and c_2 such that $d < c_1$ indicates a.i. and $d > c_2$ indicates strong association.
The usual a.i. hypothesis, $d \leq c_1$ is suitable mainly when one is interested in, if possible, to establish whether there is an association in the table. In cases where the interest lies in stating either a.i. or strong association a three-desision procedure like N_3 will be suitable.

III. TESTS FOR ALMOST INDEPENDENCE.

III. 1. Assumptions and notations.

The problems to be considered refer to the situation in a two-way contingency table, described in part 1, ch.II. Let $q_{ij} = X_{ij}/n$, $q_{i.} = X_{i.}/n$, $q_{.j} = X_{.j}/n$ and $q = (q_{11},\dots,q_{vw})$. Let further $p = (p_{11},\dots,p_{vw})$. We will assume that no p_{ij} is equal to zero. For every measure of association, d, it is in addition assumed that the following conditions are satisfied: *)

(a) d has continuous partial derivatives as function of p (27)

(b) There exists (r,s) such that

$$\frac{\partial d}{\partial p_{rs}} \neq \sum_{i=1}^{v} \sum_{j=1}^{w} p_{ij} \frac{\partial d}{\partial p_{ij}} .$$

The following notations for a particular measure d are used (if nothing else is said):

*) Three measures suggested by Goodman & Kruskal, λ, λ_b, λ_r (see part 1), do not fulfill a). There is however developed a similar theory for these measures in [7].

$$\sigma_d^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} p_{ij}(d_{ij}-d^*)^2 \tag{28}$$

where $d_{ij} = \frac{\partial d}{\partial p_{ij}}$ for $i = 1,\ldots,v$ and $j = 1,\ldots,w$

and

$$d^* = \sum_{i=1}^{v} \sum_{j=1}^{w} d_{ij}p_{ij} .$$

Further

$\hat{d} = d(q)$, $\hat{d}$ is the C-estimator of d .

The C-estimator for σ_d^2 is given by:

$$S_d^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}(\hat{d}_{ij}-\hat{d}^*)^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}\hat{d}_{ij}^2 - \hat{d}^{*2} \tag{29}$$

where

$$\hat{d}_{ij} = \frac{\partial d}{\partial p_{ij}}\Big|_{p=q}$$

$$\hat{d}^* = \sum_{i=1}^{v} \sum_{j=1}^{w} \hat{d}_{ij}q_{ij}$$

From theorem 1 we see that $\sqrt{n}(\hat{d}-d) \overset{D}{\rightarrow} N(0,\sigma_d)$. σ_d^2 is therefore called the asymptotic variance of $\sqrt{n}\hat{d}$.

At this point we like to mention that the results in Goodman & Kruskal, [8], for multinomial sampling over the entire two-way table is a special case of formula (28) for σ_d^2 . It should also be said that the author did not have any knowledge of the work in [8], while working on this theory for measures of association.

Theorem 1 also gives

$$\sqrt{n}\,\frac{(\hat{d}-d)}{S_d} \overset{D}{\rightarrow} N(0,1) \tag{30}$$

This asymptotical property will be applied for testing and interval estimation of the measures of association given in part 1. We should emphasize that the major value of theorem 1 lies in the

fact that it finds the asymptotic variance σ_d^2 .
As in part 1 we will deal with measures applicable to the following situations: 1) Ordered case - 2) Unordered symmetrical case - 3) Unordered asymmetrical case - 4) Reliability-case and 5) The 2×2 - table.
In every case, except for 4) and 5) it is shown how the a.i. hypothesis can be determined based on the different suggested measures. For each choice the estimator S_d^2 is found. The N-test for the a.i. hypothesis follows then from II. 1.(ii), and by applying (15) and (16) twosided and onesided intervals can be given. The application of the N-s-procedure in II. 3 also follows directly when S_d^2 is known.

III. 2. Ordered case and the measures γ, τ_b, τ_c.

The three measures considered in part 1 for this situation was

$$\gamma = \frac{\pi_s - \pi_d}{\pi_s + \pi_d}$$

$$\tau_b = \frac{\pi_s - \pi_d}{\sqrt{\pi_y \cdot \pi_z}}$$

$$\tau_c = \frac{\pi_s - \pi_d}{(m-1)/m}$$

Here is $m = \min(v,w)$, and π_s, π_d, π_y and π_z are given in part 1 , ch. III. 2 and III. 3.
In part 1 , chapter III.(iv) we discussed the three measures and found that γ is the most natural and suitable measure. Testing for a.i. should consequently be based on γ^2 . An a.i. hypothesis will be determined later. (As mentioned earlier since γ can be negative, we apply γ^2 as a measure for degree of association.)
We recall that in III. 1. it is assumed that $p_{ij} > 0$ for every

i and j . This implies as shown in II that the measures vary in the open interval $\langle -1,1\rangle$.

III. 2.(i) The asymptotic variance of $\sqrt{n}\,\hat{\gamma}$.

We will now as in part 1 follow the notation of Goodman & Kruskal ([7], p.322) by letting $P_s = \pi_s(q)$, $P_d = \pi_d(q)$ and $P_t = \pi_t(q)$. Let now $P_s^* = n^2 \cdot P_s$, $P_d^* = n^2 P_d$ and $P_t^* = n^2 \cdot P_t$.
Let (Y_1,Z_1), (Y_2,Z_2) and (Y_3,Z_3) be three independent random variables with the same distribution as (Y,Z) (see part 1 ,ch.II).
Define the following probabilites:

$$\pi_{ss} = P\{(Y_1-Y_2)(Z_1-Z_2) > 0 \cap (Y_1-Y_3)(Z_1-Z_3) > 0\}$$

$$\pi_{sd} = P\{(Y_1-Y_2)(Z_1-Z_2) > 0 \cap (Y_1-Y_3)(Z_1-Z_3) < 0\}$$

$$\pi_{dd} = P\{(Y_1-Y_2)(Z_1-Z_2) < 0 \cap (Y_1-Y_3)(Z_1-Z_3) < 0\}$$

Let P_{ss}, P_{sd}, P_{dd} be the following consistent estimators for these probabilites:

$$P_{ss} = \pi_{ss}(q) = \frac{1}{n^3}\sum_i\sum_j X_{ij}\{\sum_{i'>i}\sum_{j'>j} X_{i'j'} + \sum_{i'<i}\sum_{j'<j} X_{i'j'}\}^2$$

$$P_{sd} = \pi_{sd}(q) = \frac{1}{n^3}\sum_i\sum_j X_{ij}\{\sum_{i'>i}\sum_{j'>j} X_{i'j'} + \sum_{i'<i}\sum_{j'<j} X_{i'j'}\} x$$
$$\{\sum_{i'>i}\sum_{j'<j} X_{i'j'} + \sum_{i'<i}\sum_{j'>j} X_{i'j'}\}$$

$$P_{dd} = \pi_{dd}(q) = \frac{1}{n^3}\sum_i\sum_j X_{ij}\{\sum_{i'>i}\sum_{j'<j} X_{i'j'} + \sum_{i'<i}\sum_{j'>j} X_{i'j'}\}^2$$

Let further $P_{ss}^* = n^3 P_{ss}$, $P_{sd}^* = n^3 P_{sd}$ and $P_{dd}^* = n^3 P_{dd}$. The C-estimator $\hat{\gamma}$ for γ now is

$$\hat{\gamma} = \frac{P_s - P_d}{P_s + P_d} = \frac{P_s^* - P_d^*}{P_s^* + P_d^*} \qquad (31)$$

The asymptotic variance σ_γ^2 and its C-estimator S_γ^2 are given in the following result.

LEMMA 9

$$\sigma_\gamma^2 = \frac{16}{(1-\pi_t)^4}\{\pi_s^2\,\pi_{dd} - 2\pi_s\pi_d\pi_{sd} + \pi_d^2\pi_{ss}\}$$

The C-estimator can be expressed in two alternative ways:

1) $$S_\gamma^2 = \frac{16}{(1-P_t)^4}\{P_s^2P_{dd} - 2P_sP_dP_{sd} + P_d^2P_{ss}\}$$

2) $$S_\gamma^2 = \frac{n\cdot 16}{(n^2-P_t^*)^4}\{P_s^{*2}P_{dd}^* - 2P_s^*P_d^*P_{sd}^* + P_d^{*2}P_{ss}^*\}$$

Proof. The expression 1) for S_γ^2 follows immediately from σ_γ^2. The expression 2) follows from 1) since $1-P_t = n^{-2}(n^2-P_t^*)$. It is left to show the expression for σ_γ^2.

From III. 1.

$$\sigma_\gamma^2 = \sum_{i=1}^{v}\sum_{j=1}^{w} p_{ij}(\gamma_{ij}-\gamma^*)^2 \quad \text{where} \quad \gamma_{ij} = \frac{\partial\gamma}{\partial p_{ij}} \quad \text{and}$$

$$\gamma^* = \sum_{i=1}^{v}\sum_{j=1}^{w}\gamma_{ij}p_{ij}.$$

We find

$$\gamma_{ij} = \frac{4}{(1-\pi_t)^2}\{\pi_d\cdot\alpha_{ij} - \pi_s\beta_{ij}\}.$$

where $$\alpha_{ij} = \sum_{i'>i}\sum_{j'>j}p_{i'j'} + \sum_{i'<i}\sum_{j'<j}p_{i'j'} \quad \text{and}$$

$$\beta_{ij} = \sum_{i'>i}\sum_{j'<j}p_{i'j'} + \sum_{i'<i}\sum_{j'>j}p_{i'j'}$$

(see also [7], p.362).

$$\gamma^* = \frac{4}{(1-\pi_t)^2}\{\pi_d\sum_{i=1}^{v}\sum_{j=1}^{w}p_{ij}\alpha_{ij} - \pi_s\sum_{i=1}^{v}\sum_{j=1}^{w}p_{ij}\beta_{ij}\} = 0$$

Hence

$$\sigma_\gamma^2 = \sum_i \sum_j p_{ij}\gamma_{ij}^2 = \frac{16}{(1-\pi_t)^4} \{\pi_d^2 \sum_{ij} p_{ij}\alpha_{ij}^2 + \pi_s^2 \sum_{i,j} p_{ij}\beta_{ij}^2 - 2\pi_s\pi_d \sum_{ij} p_{ij}\alpha_{ij}\beta_{ij}\}$$

$$= \frac{16}{(1-\pi_t)^4} \{\pi_s^2\pi_{dd} + \pi_d^2\pi_{ss} - 2\pi_s\pi_d\pi_{sd}\}$$

Q.E.D.

Lemma 9 is a simplification of the proof in [7]. In ([8], p.416) Goodman & Kruskal apply the same simplification.

A confidence interval for γ with asymptotic confidence level equal to $1-\alpha$ is now given by

$$\gamma \in \langle \max(-1, \hat{\gamma} - \frac{S_\gamma}{\sqrt{n}} x(\tfrac{\alpha}{2})),\ \min(1, \hat{\gamma} + \frac{S_\gamma}{\sqrt{n}} x(\tfrac{\alpha}{2}))\rangle \ . \qquad (32)$$

If $\hat{\gamma} = 1$ or -1 then $S_\gamma^2 = 0$. Goodman & Kruskal suggests then the degenerated interval $\gamma = 1$ (-1 if $\hat{\gamma} = -1$) when n is large. Since $\gamma \in \langle -1, 1\rangle$, $\lim_{n\to\infty} P(\hat{\gamma} = \pm 1) = 0$. Hence the probability of getting a degenerated interval will be very small for large n. For a more thorough discussion we refer to [7], p.324.

III. 2(ii). Determination of a.i. hypothesis based on γ. Estimation of γ^2.

When testing for a.i., the hypothesis will be that the degree of association is less than or equal to a certain upper bound. This means that the "direction" of the association is immaterial.

We will, as mentioned earlier, use γ^2 as a measure for degree of association. A criterion for a.i. is given by:

$$-\epsilon \leq \gamma \leq \epsilon \qquad (33)$$

Choice of ϵ must necessarily be somewhat arbitrary, since the notion almost independence hardly can be given a realistic precise definition. However, we know from part 1 that γ is a difference between two (conditional) probabilities. A reasonable choice of ϵ will therefore be of size 0.01 - 0.10 . It should also be observed that the ϵ-choice can rest on the given situation at hand. If one from experience know that two factors always possess a certain degree of association, then one possibly ought to choose ϵ somewhat larger than if one apriori knows the factors can be approximately independent.

The hypothesis for a.i. can now be formulated as:

$$H^* : \gamma^2 \leq c$$

where c is of size 0.0001 - 0.01 .

As a matter of course one finds the asymptotic variance of $\sqrt{n}\,\hat{\gamma}^2$ equal to $4\gamma^2\sigma_\gamma^2$ and its C-estimator is equal to $4\hat{\gamma}^2 S_\gamma^2$.

Hence:

$$\frac{\sqrt{n}(\hat{\gamma}^2-\gamma^2)}{2|\hat{\gamma}|S_\gamma} \xrightarrow{D} N(0,1) \quad \left(\text{or: } \frac{\sqrt{n}(\hat{\gamma}+\gamma)(\hat{\gamma}-\gamma)}{2|\hat{\gamma}|S_\gamma} \xrightarrow{D} N(0,1)\right).$$

The N-test for H^* : Reject when

$$\frac{\sqrt{n}(\hat{\gamma}^2-\epsilon^2)}{2|\hat{\gamma}|S_\gamma} > x(\alpha) \quad \left(\text{or: } \frac{\sqrt{n}(\hat{\gamma}+\epsilon)(\hat{\gamma}-\epsilon)}{2|\hat{\gamma}|S_\gamma} > x(\alpha)\right). \tag{34}$$

A twosided confidence interval for γ^2 (from (15)):

$$\gamma^2 \in \left\langle \max\left(0, \hat{\gamma}^2 - \frac{2|\hat{\gamma}|S_\gamma}{\sqrt{n}}\, x(\tfrac{\alpha}{2})\right), \min\left(1, \hat{\gamma}^2 + \frac{2|\hat{\gamma}|S_\gamma}{\sqrt{n}}\, x(\tfrac{\alpha}{2})\right)\right\rangle \tag{35}$$

From (16) we get a onesided interval:

$$\gamma^2 \in \left\langle \max\left(0, \hat{\gamma}^2 - \frac{2|\hat{\gamma}|S_\gamma}{\sqrt{n}}\, x(\alpha)\right), 1 \right\rangle . \tag{36}$$

As mentioned in II, $\hat{\gamma}^2 - 2|\hat{\gamma}|S\, x(\alpha)n^{-\frac{1}{2}}$ is the maximum c such that the hypothesis: $\gamma^2 \leq c$ is rejected. The interval (36) tells us therefore more about the strength of the degree of association than the result of the N-test.

In [7] it is shown that

$$\sigma_\gamma^2 \leq 2(1-\gamma^2)/(1-\pi_t) \tag{37}$$

By using the estimator for the upper bound in (37), $2(1-\hat{\gamma}^2)/(1-P_t)$, instead of S_γ^2 in (34) one gets a simple computation of the test-statistic. In return the test becomes more conservative, that is the asymptotical level will be $\leq \alpha$. By using $2(1-\hat{\gamma}^2)/(1-P_t)$ instead of S_γ^2 in the two given confidence intervals the corresponding asymptotical confidence level becomes $\geq 1-\alpha$.

Goodman & Kruskal, [7], treats more thoroughly the use of (37) to construct confidence interval for γ. (Notice that if we use (32) as starting-point, we can construct another interval for γ^2. If the limits in (32) have opposite signs, this interval will be larger that (35).)

III.2. (iii) The asymptotic variances of $\sqrt{n}\,\hat{\tau}_b$ and $\sqrt{n}\,\hat{\tau}_c$.

$\hat{\tau}_b$, $\hat{\tau}_c$ are the C-estimators of τ_b, τ_c.

The C-estimators S_b^2 and S_c^2 for the asymptotic variances σ_b^2 and σ_c^2 of respectively $\sqrt{n}\,\hat{\tau}_b$ and $\sqrt{n}\,\hat{\tau}_c$ are given in the following result.

LEMMA 10

$$S_b^2 = \frac{1}{(P_z P_y)^3} \{4P_y^2 P_z^2 (P_{ss}+P_{dd}-2P_{sd}) + (P_s-P_d)^2 (P_z^2 \sum_{i=1}^{v} q_{i.}^3 +$$

$$+ P_y^2 \sum_{j=1}^{w} q_{.j}^3 + 2P_y P_z \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij} q_{i.} q_{.j}) +$$

$$+ 4P_y P_z (P_s-P_d)(P_z \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij} q_{i.} (\hat{\alpha}_{ij}-\hat{\beta}_{ij}) +$$

$$+P_y \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}q_{.j}(\hat{\alpha}_{ij}-\hat{\beta}_{ij})) - (P_s-P_d)^2(P_y+P_z)^2\} .$$

where

$$\hat{\alpha}_{ij} = \sum_{i'>i} \sum_{j'>j} q_{i'j'} + \sum_{i'<i} \sum_{j'<j} q_{i'j'} .$$

$$\hat{\beta}_{ij} = \sum_{i'>i} \sum_{j'<j} q_{i'j'} + \sum_{i'<i} \sum_{j'>j} q_{i'j'}$$

$$S_c^2 = \frac{4m^2}{(m-1)^2} \{P_{ss}+P_{dd}-2P_{sd}-(P_s-P_d)^2\} .$$

Proof.

From III. 1. $S_b^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}(\hat{\tau}_{b,i,j}-\hat{\tau}_b^*)^2 = \sum_{i=1}^{v} q_{ij}\hat{\tau}_{b,i,j}^2-\hat{\tau}_b^{*2}$,

where $\hat{\tau}_{b,i,j} = \frac{\partial\hat{\tau}_b}{\partial q_{ij}}$ and $\hat{\tau}_b^* = \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}\ \hat{\tau}_{b,i,j}.$; $\hat{\tau}_b = \frac{P_s-P_d}{\sqrt{P_yP_z}}$,

$P_y = \pi_y(q)$, $P_z = \pi_z(q)$.

After some calculation we find that

$$\hat{\tau}_{b,i,j} = \frac{1}{(P_yP_z)^{3/2}} \{2(\hat{\alpha}_{ij}-\hat{\beta}_{ij})P_yP_z + (P_s-P_d)(q_{i.}P_z+q_{.j}P_y)\} .$$

This leads to

$$\hat{\tau}_b^* = (P_yP_z)^{-3/2}\{(P_s-P_d)(P_y+P_z)\} .$$

The result for S_b^2 now follows easily. Now $\hat{\tau}_c = \frac{P_s-P_d}{(m-1)/m}$

so $S_c^2 = \sum_{i,j} q_{ij}\ \hat{\tau}_{c,i,j} - \hat{\tau}_c^*$, where

$\hat{\tau}_{c,i,j} = \partial\hat{\tau}_c/\partial q_{ij} = \frac{m}{m-1}\ 2(\hat{\alpha}_{ij}-\hat{\beta}_{ij})$ and $\hat{\tau}_c^* = \frac{2m}{m-1}(P_s-P_d)$.

Hence

$$S_c^2 = \frac{4m^2}{(m-1)^2} \{P_{ss}+P_{dd}-2P_{sd}-(P_s-P_d)^2\}$$

Q.E.D.

Confidence intervals with level $1-\alpha$ are given by

$$\tau_b \in \langle \max(-1, \hat{\tau}_b - \frac{S_b}{\sqrt{n}} x(\frac{\alpha}{2})), \min(1, \hat{\tau}_b + \frac{S_b}{\sqrt{n}} x(\frac{\alpha}{2}))\rangle .$$

$$\tau_c \in \langle \max(-1, \hat{\tau}_c - \frac{S_c}{\sqrt{n}} x(\frac{\alpha}{2})), \min(1, \hat{\tau}_c + \frac{S_c}{\sqrt{n}} x(\frac{\alpha}{2}))\rangle .$$

III. 3. Unordered symmetrical case and the measures λ, η.

Two suitable measures were proposed in this case (part 1, ch. IV):

$$\lambda = \frac{\sum_{i=1}^{v} p_{im} + \sum_{j=1}^{w} p_{mj} - p_{.m} - p_{m.}}{2 - p_{.m} - p_{m.}}$$

where $p_{.m} = \max_j p_{.j}$, $p_{im} = \max_j p_{ij}$ and similar for $p_{m.}$ and p_{mj}.

$$\eta = \frac{\sum_{i=1}^{v} \sum_{j=1}^{w} (p_{ij} - p_{i.}p_{.j})^2 (p_{i.}^{-1} + p_{.j}^{-1})}{2 - \sum_i p_{i.}^2 - \sum_j p_{.j}^2}$$

We shall in this chapter also consider the traditional measures in part 1, ch. IV. 3 and the simple measures

$$\max_{i,j} |p_{ij} - p_{i.}p_{.j}| \quad \text{and} \quad \max_{i,j} \left|\frac{p_{ij}}{p_{i.}} - p_{.j}\right|$$

A disadvantage with λ compared to η is that $\lambda = 0$ does not necessarily imply exact independence. Especially when choosing an a.i. hypothesis this is an unfortunate property. It therefore looks like one ought to choose η as a basis for an a.i. hypothesis.

III. 3(i) The asymptotic variances of $\sqrt{n}\,\hat{\lambda}$ and $\sqrt{n}\,\hat{\eta}$.

The C-estimators $\hat{\lambda}$ and $\hat{\eta}$ are given by:

$$\hat{\lambda} = \frac{\sum_{i=1}^{v} q_{im} + \sum_{j=1}^{w} q_{mj} - q_{.m} - q_{m.}}{2 - q_{.m} - q_{m.}} = \frac{\sum_{i=1}^{v} X_{im} + \sum_{j=1}^{w} X_{mj} - X_{.m} - X_{m.}}{2n - X_{.m} - X_{m.}}$$

where $q_{im} = \max_{j'} q_{ij'}$, $q_{mj} = \max_{i'} q_{i'j}$, $q_{.m} = \max_{j} q_{.j}$, $q_{m.} = \max_{i} q_{i.}$.
$X_{im} = nq_{im}$, $X_{mj} = nq_{mj}$, $X_{.m} = nq_{.m}$, $X_{m.} = nq_{m.}$.

$$\hat{\eta} = \frac{\hat{P}_2 - \hat{P}_1}{1-\hat{P}_1}$$

where $\hat{P}_1 = \frac{1}{2}(\sum_{i=1}^{v} q_{i.}^2 + \sum_{j=1}^{w} q_{.j}^2)$ and $\hat{P}_2 = \frac{1}{2} \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}^2 (\frac{1}{q_{i.}} + \frac{1}{q_{.j}})$

Since, by assumption, $p_{ij} > 0$ λ and η are well-defined. However λ does not have continuous partial derivatives as a function of the p_{ij}'s . On the other hand λ has of course continuous partial derivatives as a function of p_{im} , p_{mj} , $p_{.m}$ and $p_{m.}$ for $i = 1,\dots,v$ and $j = 1,\dots,w$. This can be utilized in a similar way. This was done by Goodman & Kruskal, [7]. From [7] we have that if p_{mj} , p_{im} , $p_{m.}$ and $p_{.m}$ are uniquely defined and $\lambda \in \langle 0,1 \rangle$ then

$$\frac{\sqrt{n}\,(\hat{\lambda} - \lambda)}{S_\lambda} \xrightarrow{D} N(0,1)$$

Here the estimator S_λ^2 for the asymptotic variance of $\sqrt{n}\,\hat{\lambda}$ is given by:

$$S_\lambda^2 = \frac{1}{(2-U_o)^2}\{(2-U_o)(2-U_\Sigma)(U_o+U_\Sigma+4-2U_*)-2(2-U_o)^2(1-\Sigma^* q_{im}) \\ -2(2-U_\Sigma)^2(1-q_{**})\}$$

where

$$U_o = q_{.m} + q_{m.}$$

$$U_\Sigma = \sum_{i=1}^{v} q_{im} + \sum_{j=1}^{w} q_{mj}$$

$\Sigma^* q_{im} = \sum_i \sum_j q_{ij}$ over all (i,j) such that $q_{ij} = q_{im} = q_{mj}$

q_{**} = that q_{ij} where $q_{i.} = q_{m.}$ and $q_{.j} = q_{.m}$

$$U_* = \sum_i^r q_{im} + \sum_j^c q_{mj} + q_{*m} + q_{m*} .$$

$\Sigma^r_i q_{im}$ denotes the sum of the q_{im}'s over those values of i for which q_{im} is in the same column as $q_{.m}$. $\Sigma^c_j q_{mj}$ is the sum over those q_{mj} such that q_{mj} is in the same row as $q_{m.}$ q_{*m} is that q_{im} with $q_{i.} = q_{m.}$ and q_{m*} is that q_{mj} with $q_{.j} = q_{.m}$.

Now, by using theorem 1 a corresponding result is true for $\hat{\eta}$.

LEMMA 11

The C-estimator for the asymptotic variance of $\sqrt{n}\,\hat{\eta}$ is

$$S^2_\eta = \frac{1}{4(1-\hat{P}_1)^4}\{\sum_{i=1}^{v}\sum_{j=1}^{w} q_{ij}[2q_{ij}(\frac{1}{q_{.j}}+\frac{1}{q_{.j}})(1-\hat{P}_1) - \hat{\gamma}_{ij}(1-\hat{P}_1) - 2(q_{i.}+q_{.j})(1-\hat{P}_2)]^2 - 4[\hat{P}_2-2\hat{P}_1+\hat{P}_1\hat{P}_2]^2\}$$

Here is $\hat{\gamma}_{ij} = \sum_{s=1}^{w}\frac{q^2_{is}}{q^2_{i.}} + \sum_{r=1}^{v}\frac{q^2_{rj}}{q^2_{.j}}$.

Proof.

From theorem 1:

$$S^2_\eta = \sum_{i=1}^{v}\sum_{j=1}^{w} q_{ij}\hat{\eta}^2_{ij} - \eta^{*2}, \quad \hat{\eta}_{ij} = \frac{\partial\hat{\eta}}{\partial q_{ij}} \text{ and } \hat{\eta}^* = \sum_{i,j}\hat{\eta}_{ij}q_{ij}.$$

We find:

$$\frac{\partial\hat{P}_2}{\partial q_{ij}} = q_{ij}(\frac{1}{q_{i.}}+\frac{1}{q_{.j}}) - \frac{1}{2}\hat{\gamma}_{ij}, \quad \frac{\partial\hat{P}_1}{\partial q_{ij}} = (q_{i.}+q_{.j}).$$

Hereby

$$\hat{\eta}_{ij} = \frac{1}{2(1-\hat{P}_1)^2}\{2q_{ij}(\frac{1}{q_{i.}}+\frac{1}{q_{.j}})(1-\hat{P}_1)-\hat{\gamma}_{ij}(1-\hat{P}_1)-2(q_{i.}+q_{.j})(1-\hat{P}_2)\}.$$

$$\hat{\eta}^* = \frac{2}{2(1-\hat{P}_1)^2}\{\hat{P}_2-2\hat{P}_1+\hat{P}_1\hat{P}_2\}$$

The result follows. Q.E.D.

Twosided confidence intervals for λ and η are given by

$$\lambda \in \langle \max\,(0,\hat{\lambda}-\frac{S_\lambda}{\sqrt{n}}x\,(\frac{\alpha}{2})),\ \min\,(1,\hat{\lambda}+\frac{S_\lambda}{\sqrt{n}}x\,(\frac{\alpha}{2}))\rangle$$

$$\eta \in \langle \max\,(0,\hat{\eta}-\frac{S_\eta}{\sqrt{n}}x(\frac{\alpha}{2})),\ \min\,(1,\hat{\eta}+\frac{S_\eta}{\sqrt{n}}x\,(\frac{\alpha}{2}))\rangle\ .$$

Onesided confidence intervals:

$$\lambda \in \langle \max\,(0,\hat{\lambda}-\frac{S_\lambda}{\sqrt{n}}x\,(\alpha)),1\rangle\ .$$

$$\eta \in \langle \max\,(0,\hat{\eta}-\frac{S_\eta}{\sqrt{n}}x\,(\alpha)),1\rangle\ .$$

III. 3(ii) Determination of a.i. hypothesis based on η.

Consider the hypotheses

$$H_1^* : \lambda \leq c_1 \tag{38}$$

$$H_2^* : \eta \leq c_2 \tag{39}$$

From III.3(ii) it follows that we shall reject H_1^*

reject H_1^* when $\sqrt{n}\,\frac{(\hat{\lambda}-c_1)}{S_\lambda} > x(\alpha)$.

and reject H_2^* when $\sqrt{n}\,\frac{(\hat{\eta}-c_2)}{S_\eta} > x(\alpha)$.

Let us assume that we have chosen η as a measure for degree of association. We wish to determine c_2 in (39) such that H_2^* becomes an a.i. hypothesis. In addition it is desirable that c_2 does not depend on the dimension $v \times w$ of the table. We can therefore choose c_2 in the 2×2-table. Then we have that η equals τ_b^2. A criterion for a.i. based on τ_b is given by

$$-\delta \leq \tau_b \leq \delta \tag{40}$$

From this we determine $c_2 = \delta^2$

Regarding the choice of δ the same problems as in III. 2.(ii) arise, but similar to ϵ in (33), it seems natural to choose a value of δ of size 0.01 to 0.10.

III. 3.(iii). Confidence intervals for traditional measures.

We shall give confidence intervals of the four measures φ^2, K, T and C, listed in part 1, ch.IV.3.

Let

$$\hat{\varphi}^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} \frac{(q_{ij}-q_{i.}q_{.j})^2}{q_{i.}q_{.j}} = \sum_{i=1}^{v} \sum_{j=1}^{w} \frac{X_{ij}^2}{X_{i.}X_{.j}} - 1 ,$$

and let further σ_φ^2 be the asymptotic variance of $\sqrt{n}\,\hat{\varphi}^2$. S_φ^2 is the C-estimator of σ_φ^2. Then we have:

LEMMA 12.

$$S_\varphi^2 = \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}(2\hat{\alpha}_{ij}-\hat{\mu}_i-\hat{\beta}_j)^2 - \{\sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}(2\hat{\alpha}_{ij}-\hat{\mu}_i-\hat{\beta}_j)\}^2$$

where

$$\hat{\alpha}_{ij} = \frac{1}{q_{i.}q_{.j}}(q_{ij}-q_{i.}q_{.j})$$

$$\hat{\mu}_i = \frac{1}{q_{i.}^2} \sum_{j=1}^{w} \frac{(q_{ij}-q_{i.}q_{.j})}{q_{.j}}$$

$$\hat{\beta}_j = \frac{1}{q_{.j}^2} \sum_{i=1}^{v} \frac{(q_{ij}-q_{i.}q_{.j})}{q_{i.}} .$$

Proof.

$$\hat{\varphi}^2 = \sum_{\substack{i=1\\ i\neq r}}^{v} \sum_{\substack{j=1\\ j\neq s}}^{w} \frac{(q_{ij}-q_{i.}q_{.j})^2}{q_{i.}q_{.j}} + \sum_{\substack{i=1\\ i\neq r}}^{v} \frac{(q_{is}-q_{.i}q_{.s})^2}{q_{i.}q_{.s}} + \sum_{\substack{j=1\\ j\neq s}}^{w} \frac{(q_{rj}-q_{r.}q_{.j})^2}{q_{r.}q_{.j}} + \frac{(q_{rs}-q_{r.}q_{.s})^2}{q_{r.}q_{.s}} .$$

This gives that

$$\hat{\varphi}_{rs} = \frac{\partial\hat{\varphi}^2}{\partial q_{rs}} = 0 + \sum_{i \neq r} \frac{(q_{is}-q_{i.}q_{.s})}{(q_{i.}q_{.s})^2}\{-2q_{i.}^2 q_{.s}-q_{i.}\} +$$

$$\sum_{j \neq s} \frac{(q_{rj}-q_{r.}q_{.j})}{q_{r.}q_{.j}}(-2q_{r.}q_{.j}^2-q_{.j}) + \frac{(q_{rs}-q_{r.}q_{.s})}{(q_{r.}q_{.s})^2}\{2q_{r.}q_{.s}-2q_{r.}^2 q_{.s}-$$

$$2q_{r.}q_{.s}^2-q_{r.}-q_{.s}\} .$$

Hence we have

$$\hat{\varphi}_{rs} = 2\hat{\alpha}_{rs}-\hat{\beta}_s-\hat{\mu}_r \qquad \text{and the result follows.}$$

Q.E.D.

A confidence interval for φ^2 is now given by:

$$\varphi^2 \in \langle \max\left(0,\hat{\varphi}^2-\frac{S_\varphi}{\sqrt{n}}x(\tfrac{\alpha}{2})\right),\ \hat{\varphi}^2+\frac{S_\varphi}{\sqrt{n}}x(\tfrac{\alpha}{2})\rangle$$

Let $L_1 = \max\left(0,\hat{\varphi}^2-\frac{S_\varphi}{\sqrt{n}}x(\frac{\alpha}{2})\right)$ and $L_2 = \hat{\varphi}^2+\frac{S_\varphi}{\sqrt{n}}x(\frac{\alpha}{2})$.

Confidence intervals for the measures K, T, C can now be stated as follows

$$K \in \langle \{\frac{L_1}{1+L_1}\}^{\frac{1}{2}} , \quad \{\frac{L_2}{1+L_2}\}^{\frac{1}{2}}\rangle$$

$$T \in \langle \{\frac{L_1}{\sqrt{(v-1)(w-1)}}\}^{\frac{1}{2}} , \quad \min\left(1,\{\frac{L_2}{\sqrt{(v-1)(w-1)}}\}^{\frac{1}{2}}\right)\rangle$$

$$C \in \langle \frac{L_1}{\min(v-1,w-1)} , \quad \min\left(1,\frac{L_2}{\min(v-1,w-1)}\right)\rangle$$

In the next section we consider interval estimation of two quantities which are not especially suitable measures of association (they are much too coarse), but very straight to deal with, since their values are easily interpreted.

III. 3 (iv). Confidence intervals for $\max_{i,j} |p_{ij}-p_{i.}p_{.j}|$ and $\max_{i,j} |\frac{p_{ij}}{p_{i.}} - p_{.j}|$.

$$\varkappa_1 = \max_{i,j} |p_{ij} - p_{i.}p_{.j}| .$$

$\varkappa_1$ is not derivative and we can therefore not use theorem 1 to construct confidence intervals for $\varkappa_1$. Consider instead

$$D(p) = \sum_{i=1}^{v} \sum_{j=1}^{w} (p_{ij}-p_{i.}p_{.j})^2 .$$

D has obviously continuous partial derivatives.

It is assumed that (27b) holds. A confidence interval for D is hence given by:

$$D \in \langle \max\left(0, \hat{D} - \frac{S_D}{\sqrt{n}} x\left(\frac{\alpha}{2}\right)\right), \hat{D} + \frac{S_D}{\sqrt{n}} x\left(\frac{\alpha}{2}\right)\rangle \tag{41}$$

where $\hat{D} = D(q)$ and S_D^2 is the C-estimator for the asymptotic variance of $\sqrt{n}\,\hat{D}$, given below.

LEMMA 13

$$S_D^2 = 4\{\sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}(q_{ij}-q_{i.}q_{.j}-\hat{\xi}_i-\hat{\nu}_j)^2 - \left[\sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}(q_{ij}-q_{i.}q_{.j}-\hat{\xi}_i-\hat{\nu}_j)\right]^2$$

where

$$\hat{\xi}_i = \sum_{k=1}^{w} q_{.k}(q_{ik}-q_{i.}q_{.k}) \quad \text{for } i = 1,\dots,v$$

and

$$\hat{\nu}_j = \sum_{r=1}^{v} q_{r.}(q_{rj}-q_{r.}q_{.j}) \quad \text{for } j = 1,\dots,w .$$

Proof. By using the same splitting on $\hat{D}$ as was used on $\hat{\varphi}^2$ in the proof of lemma 12 we see that

$$\hat{D}_{ij} = \frac{\partial\hat{D}}{\partial q_{ij}} = 2\{(q_{ij}-q_{i.}q_{.j}) - \sum_{k=1}^{w} q_{.k}(q_{ik}-q_{i.}q_{.k}) - \sum_{r=1}^{v} q_{r.}(q_{rj}-q_{r.}q_{.j})\}$$

$$= 2(q_{ij}-q_{i.}q_{.j}-\hat{\xi}_i-\hat{\nu}_j)$$

Hence the result for S_D^2 follows.

Q.E.D.

The confidence interval (41) for D can now be applied to construct an interval for $\varkappa_1$. Let $C_1 = C_1(q)$ and $C_2 = C_2(q)$ be respectively lower and upper limit in (41). Then the following result is valid.

LEMMA 14

a) $v = w = 2$ gives: $C_1 < D < C_2 \iff \frac{1}{2}\sqrt{C_1} < \varkappa_1 < \frac{1}{2}\sqrt{C_2}$

b) $v = 2$, $w > 2$ (or $w = 2$, $v > 2$) gives: $C_1 < D < C_2 \Rightarrow \sqrt{\frac{C_1}{vw}} < \varkappa_1 < \sqrt{\frac{C_2}{2}}$

c) $v > 2$, $w > 2$ gives: $C_1 < D < C_2 \Rightarrow \sqrt{\frac{C_1}{vw}} < \varkappa_1 < \sqrt{C_2}$.

Proof.

a) $|p_{ij} - p_{i.}p_{.j}| = |p_{11}p_{22} - p_{12}p_{21}|$ for $i = 1,2$ and $j = 1,2$ (see part 1, ch.VIII)

Hence: $D = 4\varkappa_1^2$ and the result follows.

b) Assume $v = 2$. For $w = 2$ the procedure is completely analoguous. We have: $|p_{1j} - p_{1.}p_{.j}| = |p_{2j} - p_{2.}p_{.j}|$, such that

$$D = 2\sum_{j=1}^{w}(p_{1j}-p_{1.}p_{.j})^2 \quad \text{and} \quad \varkappa_1 = \max_j |p_{1j}-p_{1.}p_{.j}|$$

This implies: $C_1 < D < C_2 \iff \frac{C_1}{2} < \sum_j (p_{1j}-p_{1.}p_{.j})^2 < \frac{C_2}{2} \Rightarrow$

$$\Rightarrow \frac{C_1}{2} < w\varkappa_1^2 \text{ and } \varkappa_1^2 < \frac{C_2}{2} \Leftrightarrow \frac{C_1}{2w} < \varkappa_1^2 < \frac{C_2}{2} \Leftrightarrow \sqrt{\frac{C_1}{vw}} < \varkappa_1 < \sqrt{\frac{C_2}{2}} .$$

c) $C_1 < D \Rightarrow C_1 < vw\varkappa_1^2 \Leftrightarrow C_1/vw < \varkappa_1^2$,

and

$D < C_2 \Rightarrow \varkappa_1^2 < C_2$, and the result follows.

Q.E.D.

Confidence intervals for $\varkappa_1$ in the three different cases in lemma 14 are now given by (since $\varkappa_1 \in [0,1]$):

a) <u>$v = w = 2$</u>

$$\varkappa_1 \in \langle \{\max(0, \hat{\varkappa}_1^2 - \frac{S_D}{4\sqrt{n}} x(\tfrac{\alpha}{2}))\}^{\frac{1}{2}}, \min(1, \{\hat{\varkappa}_1^2 + \frac{S_D}{4\sqrt{n}} x(\tfrac{\alpha}{2})\}^{\frac{1}{2}}) \rangle$$

where $\hat{\varkappa}_1^2 = (q_{11}q_{22} - q_{12}q_{21})^2$. Asymptotic confidence level is equal to $1-\alpha$.

b) <u>$v = 2, w > 2$, or $v > 2, w = 2$</u>

$$\varkappa_1 \in \langle \{\max(0, \frac{\hat{D}}{vw} - \frac{S_D}{vw\sqrt{n}} x(\tfrac{\alpha}{2}))\}^{\frac{1}{2}}, \min(1, \{\tfrac{1}{2}\hat{D} + \frac{S_D}{2\sqrt{n}} x(\tfrac{\alpha}{2})\}^{\frac{1}{2}}) \rangle .$$

Asymptotic confidence level is $\geq 1-\alpha$.

c) <u>$v > 2, w > 2$</u>

$$\varkappa_1 \in \langle \{\max(0, \frac{\hat{D}}{vw} - \frac{S_D}{vw\sqrt{n}} x(\tfrac{\alpha}{2})\}^{\frac{1}{2}}, \min(1, \{\hat{D} + \frac{S_D}{\sqrt{n}} x(\tfrac{\alpha}{2})\}^{\frac{1}{2}} \rangle .$$

Asymptotic confidence level is $\geq 1-\alpha$.

In all three cases the lower limit equals $(\frac{C_1}{vw})^{\frac{1}{2}}$.

The upper limit becomes, if it is less than 1 , in case a) equal to $(\frac{C_2}{vw})^{\frac{1}{2}}$ in b) equal to $(\frac{C_2}{2})^{\frac{1}{2}}$ and in c) equal to $C_2^{\frac{1}{2}}$.

When trying to construct confidence interval for *)

---

*) $\varkappa_2$ is a measure which really is best suited in the asymmetrical situation (with B as the primary factor).

$$\varkappa_2 = \max_{i,j} \left|\frac{p_{ij}}{p_{i.}} - p_{.j}\right|$$

we run across the same problem as with $\varkappa_1$. We are not able to construct interval for $\varkappa_2$ based directly on $\hat{\varkappa}_2 = \max_{i,j} \left|\frac{q_{ij}}{q_{i.}} - q_{.j}\right|$, on the basis of the theory developed in chapter II. Therefore we will first construct an asymptotical $(1-\alpha)$-confidence interval for

$$E(p) = \sum_{i=1}^{v} \sum_{j=1}^{w} \left(\frac{p_{ij}}{p_{i.}} - p_{.j}\right)^2 .$$

This interval is given by

$$E \in \langle \max\left(0, \hat{E} - \frac{S_E}{\sqrt{n}}\, x(\tfrac{\alpha}{2})\right),\ \hat{E} + \frac{S_E}{\sqrt{n}}\, x(\tfrac{\alpha}{2})\rangle \tag{42}$$

where $\hat{E} = E(q)$ and S_E^2 is the C-estimator of the asymptotic variance for $\sqrt{n}\,\hat{E}$, given in the next lemma.

LEMMA 15

$$S_E^2 = 4\left\{\sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}\left(\frac{q_{ij}-q_{i.}q_{.j}}{q_{i.}^2} - \hat{E}_i - \hat{F}_j\right)^2 - \left[\sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}\left(\frac{q_{ij}-q_{i.}q_{.j}}{q_{i.}^2} - \hat{E}_i - \hat{F}_j\right)\right]^2\right\}$$

<u>where</u>

$$\hat{E}_i = q_{i.}^{-3} \sum_{k=1}^{w} (q_{ik}-q_{i.}q_{.k})q_{ik} \qquad \text{for } i = 1,\dots,v$$

<u>and</u>

$$\hat{F}_j = \sum_{r=1}^{v} \frac{q_{rj}-q_{r.}q_{.j}}{q_{r.}} \qquad \text{for } j = 1,\dots,w .$$

<u>Proof</u>. By using the same splitting of $\hat{E}$ as of $\hat{D}$ in lemma 13 and $\hat{\varphi}^2$ in lemma 12 we see that

$$\hat{E}_{rs} = \frac{\partial\hat{E}}{\partial q_{rs}} = -2\sum_{\substack{i=1\\i\neq r}}^{v}\frac{q_{is}-q_{i.}q_{.s}}{q_{i.}} - 2\sum_{\substack{j=1\\j\neq s}}^{w}\frac{(q_{rj}-q_{r.}q_{.j})q_{rj}}{q_{r.}^{3}} +$$

$$2\,\frac{(q_{rs}-q_{r.}q_{.s})}{q_{r.}^{3}}(q_{r.}-q_{rs}-q_{r.}^{2})$$

$$= 2\{\frac{q_{rs}-q_{r.}q_{.s}}{q_{r.}^{2}} - \hat{E}_{r} - \hat{F}_{s}\} .$$

The result follows immediately.

Q.E.D.

Let the lower and upper bound in (42) be denoted by $K_1 = K_1(q)$ and $K_2 = K_2(q)$.

LEMMA 16.

1) $\underline{w = 2 \text{ gives: } K_1 < E < K_2 \Rightarrow \sqrt{\frac{K_1}{vw}} < \varkappa_2 < \sqrt{\frac{K_2}{2}}}$

2) $\underline{w > 2 \text{ gives: } K_1 < E < K_2 \Rightarrow \sqrt{\frac{K_1}{vw}} < \varkappa_2 < \sqrt{K_2}}$

Proof. 1) Let $\Delta_i = |p_{i1}-p_{i.}p_{.1}|$ for $i = 1,\dots,v$. Since $|p_{i1}-p_{i.}p_{.1}| = |p_{i2}-p_{i.}p_{.2}|$ we have that $|\frac{p_{ij}}{p_{i.}} - p_{.j}| = \frac{\Delta_i}{p_i}$ for $j = 1,2$, i.e.

$$|\frac{p_{i1}}{p_{i.}} - p_{.1}| = |\frac{p_{i2}}{p_{i.}} - p_{.2}| \quad \text{for } i = 1,2,\dots,v .$$

This gives: $E = 2\sum_{i=1}^{v}(\frac{p_{i1}}{p_{i.}} - p_{.1})^2$ and $\varkappa_2 = \max_i |\frac{p_{i1}}{p_{i.}} - p_{.1}|$ such that

$K_1 < E < K_2 \Rightarrow \frac{K_1}{2v} < \varkappa_2^2 < \frac{K_2}{2}$ and the result 1) follows.

2) $w > 2$: $K_1 < E < K_2 \Rightarrow \frac{K_1}{vw} < \varkappa_2^2 < K_2 \Leftrightarrow \sqrt{\frac{K_1}{vw}} < \varkappa_2 < \sqrt{K_2}$.

Q.E.D.

Confidence intervals for $\varkappa_2$ with asymptotic confidence levels $\geq 1-\alpha$ in the two cases now become (since $\varkappa_2 \in [0,1]$),

1) <u>w = 2</u>

$$\varkappa_2 \in \langle \{\max (0, \frac{\hat{E}}{vw} - \frac{S_E}{vw\sqrt{n}} x (\frac{\alpha}{2}))\}^{\frac{1}{2}}, \min (1, \{\frac{\hat{E}}{2} + \frac{S_E}{2\sqrt{n}} x (\frac{\alpha}{2})\}^{\frac{1}{2}}) \rangle$$

2) <u>w > 2</u>

$$\varkappa_2 \in \langle \{\max (0, \frac{\hat{E}}{vw} - \frac{S_E}{vw\sqrt{n}} x (\frac{\alpha}{2}))\}^{\frac{1}{2}}, \min (1, \{\hat{E} + \frac{S_E}{\sqrt{n}} x (\frac{\alpha}{2})\}^{\frac{1}{2}} \rangle$$

Next we consider the measures proposed in part 1 when one factor is of primary interest. We shall assume for simplicity that the primary factor is B .

III. 4. Unordered asymmetrical case and the measures λ_b, η_b.

The measures suggested for this situation was

$$\lambda_b = \frac{\sum_{i=1}^{v} p_{im} - p_{\cdot m}}{1 - p_{\cdot m}}$$

$$\eta_b = \frac{\sum_{i=1}^{v} \sum_{j=1}^{w} \frac{(p_{ij} - p_{i.} p_{.j})^2}{p_{i.}}}{1 - \sum_{j=1}^{w} p_{\cdot j}^2}$$

C-estimators of λ_b and η_b are given by

$$\hat{\lambda}_b = \frac{\sum_{i=1}^{v} q_{im} - q_{\cdot m}}{1 - q_{\cdot m}} = \frac{\sum_{i=1}^{v} X_{im} - X_{\cdot m}}{n - X_{\cdot m}}$$

$$\hat{\eta}_b = \frac{\hat{P}_2^b - \hat{P}_1^b}{1 - \hat{P}_1^b}$$

where $\hat{P}_2^b = \sum_{i=1}^{v} \sum_{j=1}^{w} q_{ij}^2 / q_{i.}$ and $\hat{P}_1^b = \sum_{j=1}^{w} q_{\cdot j}^2$.

III. 4 (i) Asymptotic variances of $\sqrt{n}\,\hat{\lambda}_b$ and $\sqrt{n}\,\hat{\eta}_b$.

All the cell-probabilities are assumed positive so λ_b and η_b are well defined. λ_b does not possess continuous partial derivatives with respect to the p_{ij}'s. However, from [7], we have that if p_{im} and $p_{.m}$ are uniquely defined and $\lambda_b \in \langle 0,1 \rangle$ then

$$\sqrt{n}\,\frac{(\hat{\lambda}_b-\lambda_b)}{S_b} \xrightarrow{D} N(0,1)$$

where $$S_b^2 = \frac{(1-\sum_{i=1}^{v} q_{im})(\sum_{i=1}^{v} q_{im}+q_{.m}-2\Sigma^r q_{im})}{(1-q_{.m})^3}$$

With use of theorem 1 a similar result is obtained for $\hat{\eta}_b$.

LEMMA 17

The C-estimator for the asymptotic variance of $\sqrt{n}\,\hat{\eta}_b$ is given by

$$S_{ob}^2 = \frac{1}{(1-\hat{P}_1^b)^4}\{\sum_{i,j} q_{ij}[2\frac{q_{ij}}{q_{i.}}(1-\hat{P}_1^b) - 2q_{.j}(1-\hat{P}_2^b) - \hat{\rho}_i(1-\hat{P}_1^b)]^2 - [\hat{P}_2^b - 2\hat{P}_1^b + \hat{P}_1^b\hat{P}_2^b]^2\}.$$

$$\hat{\rho}_i = \sum_j (q_{ij}/q_{i.})^2 .$$

(Goodman & Kruskal, [7] p.354, have an error in the result for S_{ob}^2 which they have corrected in [8].)

Proof.

$$S_{ob}^2 = \sum_{i=1}^{v}\sum_{j=1}^{w} q_{ij}(\hat{\eta}_{ij}^o - \hat{\eta}^o)^2 \quad \text{where} \quad \hat{\eta}_{ij}^o = \frac{\partial \hat{\eta}_b}{\partial q_{ij}} \quad \text{and}$$

$$\hat{\eta}^o = \sum_{ij} q_{ij}\hat{\eta}_{ij}^o .$$

It is readily seen that

$$\frac{\partial \hat{P}_1^b}{\partial q_{ij}} = 2q_{.j} \quad \text{and} \quad \frac{\partial \hat{P}_2^b}{\partial q_{ij}} = \frac{2q_{ij}}{q_{i.}} - \sum_{j'=1}^{w} \left(\frac{q_{ij'}}{q_{i.}}\right)^2 .$$

This gives

$$\hat{\eta}_{ij}^o = \frac{1}{(1-\hat{P}_1^b)^2} \{2\frac{q_{ij}}{q_{i.}}(1-\hat{P}_1^b) - 2q_{.j}(1-\hat{P}_2^b) - \hat{p}_i(1-\hat{P}_1^b)\} ,$$

hence

$$\hat{\eta}^o = \frac{1}{(1-\hat{P}_1^b)^2} \{ \hat{P}_2^b - 2\hat{P}_1^b + \hat{P}_1^b\hat{P}_2^b\} .$$

The result follows.

Q.E.D.

We now can state the following confidence intervals

$$\lambda_b \in \langle \max\ (0, \hat{\lambda}_b - \frac{S_b}{\sqrt{n}}\ x(\tfrac{\alpha}{2})),\ \min\ (1, \hat{\lambda}_b + \frac{S_b}{\sqrt{n}}\ x(\tfrac{\alpha}{2}))\rangle .$$

$$\eta_b \in \langle \max\ (0, \hat{\eta}_b - \frac{S_{ob}}{\sqrt{n}}\ x(\tfrac{\alpha}{2})),\ \min\ (1, \hat{\eta}_b + \frac{S_{ob}}{\sqrt{n}}\ x(\tfrac{\alpha}{2}))\rangle .$$

Tests for the hypotheses

$$H_3^* : \lambda_b \leq c_3 \qquad (43)$$

$$H_4^* : \eta_b \leq c_4 \qquad (44)$$

are given by: Reject H_3^* when $\sqrt{n}\,\frac{(\hat{\lambda}_b - c_3)}{S_b} > x(\alpha)$

and reject H_4^* when $\sqrt{n}\,\frac{(\hat{\eta}_b - c_4)}{S_{ob}} > x(\alpha)$.

Choice of c_4 so that (44) becomes an a.i. hypothesis follows from the criterion (40) since $\eta_b = \tau_b^2$ in the 2×2-table, i.e. $c_4 = c_2 = \delta^2$.

III. 5. Reliability-case and the measures λ_r and π_k.

We recall from part 1, ch. VI that the characteristical feature of this situation is that $A_i = B_i$ for $i = 1,\dots,w$.

III. 5.(i) The unordered symmetrical case.

The proposed measure for this situation was

$$\lambda_r = \frac{\sum_{i=1}^{v} p_{ii} - \frac{1}{2}(p_{M.}+p_{.M})}{1-\frac{1}{2}(p_{M.}+p_{.M})}$$

where $\quad p_{M.}+p_{.M} = \max_{i'} (p_{i'.}+p_{.i'})$.

In [7] it is shown that if λ_r is well-defined, different from ± 1 and $p_{M.}+p_{.M}$ is unique, then

$$\sqrt{n}\,\frac{\hat{\lambda}_r-\lambda_r}{S_r} \xrightarrow{D} N(0,1) \tag{45}$$

where $\hat{\lambda}_r = \dfrac{\sum_{i=1}^{v} q_{ij} - \frac{1}{2}(q_{M.}+q_{.M})}{1-\frac{1}{2}(q_{M.}+q_{.M})}$; $q_{M.}+q_{.M} = \max_{i} (q_{i.}+q_{.i})$

and $\quad S_r^2 = [1-\frac{1}{2}(q_{M.}+q_{.M})]^{-2}\{(1-\sum_{i=1}^{v} q_{ii})[\sum_{i=1}^{v} q_{ii}+\frac{1}{4}(q_{M.}+q_{.M}) \times$

$$(1-\sum_{i=1}^{v} q_{ii}-(q_{M.}+q_{.M})) - q_{MM}(\tfrac{3}{2}+\tfrac{1}{2}\sum_{i=1}^{v} q_{ii}-(q_{M.}+q_{.M})]\}$$

q_{MM} is that q_{ii} where $q_{i.}+q_{.i} = q_{M.}+q_{.M}$.

(45) can be used to test hypotheses and construct confidence intervals for λ_r .

III. 5 (ii) The ordered case.

The suggested measures was of type $\pi_k = \sum_{|i-j|\leq k} p_{ij}$.

In this case we can use exact distribution theory. Let $X_k = \sum_{|i-j|\leq k} X_{ij}$. One sees immediately thet X_k have a binomial distribution (π_k, n) i.e.

$$P(X_k = x) = \binom{n}{x}\pi_k^x(1-\pi_k)^{n-x} \quad . \qquad \text{for} \quad x = 0,1,\ldots,n \; .$$

It follows easily how testing and estimation of π_k can be done with optimal methods.

Theorem 1 reduces in this case to the usual central limit theorem for independent, identically distributed random variables.

At last we consider the measures proposed for the 2×2-table.

III. 6. The 2×2-table.

Under certain reasonable assumptions on a measure of association in the 2×2-table which we stated in part 1, ch.VIII, the cross-product ratio

$$\Delta = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

or a one-to-one function of Δ was found to be the matural choice of measure. We listed three measures which was one-to-one functions of Δ

$$d_1 = \frac{\Delta-1}{\Delta+1} = 1-\frac{2}{\Delta+1}$$

$$d_2 = \frac{\sqrt{\Delta}-1}{\sqrt{\Delta}+1} = 1-\frac{2}{\sqrt{\Delta}+1} \; .$$

$$\rho = \ln \Delta$$

The exact independence hypothesis can in this case be expressed as

$$H : p_{11}p_{22} = p_{12}p_{21} \; .$$

As a measure for degree of association we can as mentioned in part 1 use any one of the measures d_1^2, d_2^2 or ρ^2. For testing for a.i.

it does not matter which one we choose, since a hypothesis about one measure will be equivalent to hypotheses about the other measures. For testing of d_1^2 we refer to III.2.(ii), since $d_1 = \gamma$ in the 2×2-case. Later it will be shown that there exists a uniformly most powerful unbiased α-level test for a.i. based on ρ^2. Let us therefore show what the a.i. hypothesis based on $d = \rho^2$ is, and state the N-test for that hypothesis.

III. 6.(i) Determination of a.i. hypothesis, and its normal-test.

The a.i. hypothesis based on d can be formulated as:

$$H^* : (\ln \Delta)^2 \leq c \tag{46}$$

As a basis for determinating c we shall use the a.i. criterion (33) which now becomes:

$$-\epsilon \leq \frac{\Delta-1}{\Delta+1} \leq \epsilon \tag{47}$$

(47) is equivalent with

$$\frac{1-\epsilon}{1+\epsilon} \leq \Delta \leq \frac{1+\epsilon}{1-\epsilon}$$

which gives that A and B are a.i. if and only if

$$\frac{1-\epsilon}{1+\epsilon} \leq \Delta \leq \frac{1+\epsilon}{1-\epsilon} \tag{48}$$

It is clear that Δ and Δ^{-1} corresponds to the same degree of association in opposite directions. We see that $\Delta \in \left[\frac{1-\epsilon}{1+\epsilon}, \frac{1+\epsilon}{1-\epsilon}\right]$ if and only if $\Delta^{-1} \in \left[\frac{1-\epsilon}{1+\epsilon}, \frac{1+\epsilon}{1-\epsilon}\right]$, so that (48) is a reasonable criterion for a.i. Further (48) is equivalent with

$$\ln \frac{1-\epsilon}{1+\epsilon} \leq \ln \Delta \leq \ln \frac{1+\epsilon}{1-\epsilon} \iff (\ln \Delta)^2 \leq (\ln \frac{1+\epsilon}{1-\epsilon})^2$$

Hence $c = (\ln \frac{1+\epsilon}{1-\epsilon})^2$, and the a.i. hypothesis (46) becomes equal to

$$H^* : (\ln \Delta)^2 \leq (\ln \frac{1+\epsilon}{1-\epsilon})^2 ,$$

where ϵ is determined in (33). Below we present a table over c-values for some chosen ϵ-values.

ϵ	0.01	0.05	0.10
c	0.0004	0.01	0.04

$$\hat{d} = (\ln \frac{q_{11}q_{22}}{q_{12}q_{21}})^2 = (\ln \frac{X_{11}X_{22}}{X_{12}X_{21}})^2$$

The C-estimator S_d^2 for the asymptotic variance of $\sqrt{n}\,\hat{d}$ are given in the following lemma.

LEMMA 18

$$\underline{S_d^2 = 4\hat{d}nS^2} \qquad (49)$$

$$\underline{\text{where} \quad S^2 = X_{11}^{-1} + X_{22}^{-1} + X_{12}^{-1} + X_{21}^{-1}} \qquad (50)$$

<u>(when using (49) it is assumed that no X_{ij} equals zero).</u>

<u>Proof</u>.

Let $d_{ij} = \frac{\partial d}{\partial p_{ij}}$ and $d^* = \sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij}\,d_{ij}$.

For $i,j = 1,2$ we have:

$$d_{ij} = 2\ln \Delta \cdot \frac{\partial \ln \Delta}{\partial p_{ij}} \quad , \text{ and}$$

$$\frac{\partial \ln \Delta}{\partial p_{ij}} = \begin{cases} p_{ii}^{-1} & \text{for } i = j \\ -p_{ij}^{-1} & \text{for } i \neq j \end{cases}$$

Hence $d_{ii} = \frac{2}{p_{ii}} \ln \Delta$ and $d_{ij} = -\frac{2}{p_{ij}} \ln \Delta$ for $i \neq j$

This gives $d^* = 0$.

Let $\sigma_d^2(p) = \sum_{i,j} p_{ij} d_{ij}^2$. We find: $\sigma_d^2 = 4(\ln \Delta)^2 (\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}})$.

Now $S_d^2 = \sigma_d^2(q)$ and the result follows.

Q.E.D.

From lemma 18 the N-test for H* is now given by:

Reject H* when

$$\frac{\hat{d} - c}{2S\sqrt{\hat{d}}} > x(\alpha) .$$

The condition 27 b) reduces here to assume $\Delta \neq 1$ $(\rho^2 > 0)$.
The theory in the next section, however, is valid even when $\Delta = 1$.

III. 6 (ii) A uniformly most powerful unbiased test for a.i..

Definition 6 Let $H: \theta \in \omega_o$ against $\theta \in \Omega - \omega_o$ be the hypothesis to be tested. Further, let

$$M_\alpha = \{\varphi | \varphi \text{ is an unbiased } \alpha\text{-level test for } H\}$$

The power function of a test φ is called β_φ . Then φ_o is a uniformly most powerful unbiased (UMPU) α-level test for H , if $\varphi_o \in M_\alpha$ and $\beta_{\varphi_o}(\theta) \geq \beta_\varphi(\theta)$ for all $\varphi \in M_\alpha$ and $\theta \in \Omega - \omega_o$.

We shall now find a UMPU α-level test for $H^* : (\ln \Delta)^2 \leq c$, or equivalently for

$$H^{**} : -k \leq \ln \Delta \leq k \qquad (51)$$

where $k = \ln \frac{1+\epsilon}{1-\epsilon}$.

Let us call this test δ_o . Then δ_o will be a UMPU α-level test for an a.i. hypothesis based on every measure which is a one-to-one function of Δ , and where a.i. pr. definition is given by (48). Especially δ_o is a UMPU α-level test for

$$H_o : -\epsilon \leq \frac{\Delta-1}{\Delta+1} \leq \epsilon .$$

Let $X = (X_{11}, X_{12}, X_{21}, X_{22})$. We will use the same notation as in

Sverdrup, [18].

Let P_o be the distribution for X when $p_{11} = p_{12} = p_{21} = p_{22} = \frac{1}{4}$. X has then the following distribution P given by:

$$dP = (4p_{22})^n e^{\tau_1 x_{1.} + \tau_2 x_{.1} + \rho x_{11}} dP_o \quad \text{(see [18], p.40-41)}$$

$$(P(X \in A) = P(A) = \int_A (4p_{22})^n e^{\tau_1 x_{1.} + \tau_2 x_{.1} + \rho x_{11}} dP_o) .$$

Here is $\tau_1 = \ln \frac{p_{12}}{p_{22}}$ and $\tau_2 = \ln \frac{p_{21}}{p_{22}}$, $x_{1.} = x_{11} + x_{12}$ and $x_{.1} = x_{11} + x_{21}$

Let $x = (x_{11}, x_{12}, x_{21}, x_{22})$. From Lehmann ([12], ch.4.4) a UMPU-α-level test δ_o for H^{**} is given by:

$$\delta_o(x) = \begin{cases} 1 & \text{if } x_{11} < C_1(x_{1.}, x_{.1}) \text{ or } x_{11} > C_2(x_{1.}, x_{.1}) \\ \gamma_i(x_{1.}, x_{.1}) & \text{if } x_{11} = C_i(x_{1.}, x_{.1}) \text{ for } i = 1,2 \\ 0 & \text{if } C_1(x_{1.}, x_{.1}) < x_{11} < C_2(x_{1.}, x_{.1}) \end{cases}$$

where C_1, C_2, γ_1, γ_2 are determined by:

$$E_{-k}[\delta_o(X) | X_{1.}, X_{.1}] = E_k[\delta_o(X) | X_{1.}, X_{.1}] = \alpha .$$

To determine C_i, γ_i, $i = 1,2$ we need the conditional distribution for X given the marginals. It is easily seen that

$$P(X = x | X_{1.} = x_1 \cap X_{.1} = y_1) = \begin{cases} P(X_{11} = x_{11} | X_{1.} = x_1 \cap X_{.1} = y_1) & \text{if} \\ & x_{11} + x_{12} = x_1 \text{ and } x_{11} + x_{21} = y_1 \\ 0 \text{ otherwise} & \end{cases}$$

The conditional distribution of X_{11} given the marginals can be expressed as follows:

$$P(X_{11} = x_{11} | X_{1.} = x_1 \cap X_{.1} = y_1) = \frac{\binom{x_1}{x_{11}} \binom{n - x_1}{y_1 - x_{11}} e^{\rho x_{11}}}{\sum_{z=o}^{\min(x_1, y_1)} \binom{x_1}{z} \binom{n - x_1}{y_1 - z} e^{\rho z}} = f_\rho(x_{11} | x_1, y_1)$$

Hence C_1, C_2, γ_1, γ_2 are given by the following two equalities:

$$\sum_{x_{11}=0}^{C_1-1} f_k(x_{11}|x_1,y_1) + \sum_{x_{11}=C_2+1}^{n} f_k(x_{11}|x_1,y_1) + \gamma_1 f_k(C_1|x_1,y_1) + \gamma_2 f_k(C_2|x_1,y_1) = \alpha$$

$$\sum_{x_{11}=0}^{C_1-1} f_{-k}(x_{11}|x_1,y_1) + \sum_{x_{11}=C_2+1}^{n} f_{-k}(x_{11}|x_1,y_1) + \gamma_1 f_{-k}(C_1|x_1,y_1) + \gamma_2 f_{-k}(C_2|x_1,y_1) = \alpha .$$

for all x_1, y_1 such that $0 \leq x_1 + y_1 \leq n$.

III. 6 (iii) Confidence intervals for the cross-product ratio.

Fisher, [3], proposed a method for obtaining confidence interval for Δ which required the solution of a quadratical equation. Goodman, [4], developed a simpler method for constructing confidence interval for Δ with asymptotical confidence level equal to $1-\alpha$. We shall now show that by using theorem 1 we obtain the same confidence interval as the one Goodman gives. Our method is however simpler than the one Goodman proposes. We assume that no X_{ij} equals zero.

Let $\hat{\Delta} = \dfrac{X_{11}X_{22}}{X_{12}X_{21}}$, $\Delta_{ij} = \dfrac{\partial \Delta}{\partial p_{ij}}$ and $\Delta^* = \sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij}\Delta_{ij}$.

Then the following result will be proved:

LEMMA 19

$$S_{\Delta}^2 = n\,\hat{\Delta}^2 S^2$$

where S^2 is defined by (50).

Proof.

$$\Delta_{11} = \frac{p_{22}}{p_{12}p_{21}} , \; \Delta_{22} = \frac{p_{11}}{p_{12}p_{21}} , \; \Delta_{12} = -\Delta/p_{12}, \; \Delta_{21} = -\Delta/p_{21}$$

This implies $\Delta^* = 0$.

Let $\sigma_{\Delta}^2(p) = \sum_{i,j} p_{ij}\Delta_{ij}^2 = \Delta^2(\sum_{i,j} p_{ij}^{-1})$. Hence

$$S_\Delta^2 = \sigma_\Delta^2(q) = n\,\hat{\Delta}^2(\sum_{i,j} X_{ij}^{-1}) = n\,\hat{\Delta}^2 S^2 \,.$$

Q.E.D.

From theorem 1 and lemma 19:

$$\frac{\hat{\Delta}-\Delta}{\hat{\Delta}\cdot S} \overset{D}{\to} N(0,1)$$

A sufficient condition for satisfying a) and b) in III.1 is in this case that $p_{ij} > 0$ for $i=1,2$ and $j=1,2$.
A confidence interval for Δ with approximate confidence level equal to $1-\alpha$ is now given by:

$$\Delta \in \langle \max\,(0,\hat{\Delta}(1-x(\tfrac{\alpha}{2})\cdot S)),\hat{\Delta}(1+x(\tfrac{\alpha}{2})\cdot S)\rangle \tag{52}$$

This is the same interval as Goodman derives ([4], p.90).

From theorem 1 we also have

$$\sqrt{n}\,\frac{\hat{\Delta}-\Delta}{\Delta\sqrt{p_{11}^{-1}+p_{12}^{-1}+p_{21}^{-1}+p_{22}^{-1}}} \overset{D}{\to} N(0,1)$$

and since

$$\sqrt{n}\,S \overset{P}{\to} \sqrt{p_{11}^{-1}+p_{12}^{-1}+p_{21}^{-1}+p_{22}^{-1}}\,:$$

$$\frac{\hat{\Delta}-\Delta}{\Delta\cdot S} \overset{D}{\to} N(0,1)\,.$$

The following relations hold:

$$\{-x(\tfrac{\alpha}{2}) < \frac{\hat{\Delta}-\Delta}{\Delta\cdot S} < x(\tfrac{\alpha}{2})\} \iff \{\Delta(1-x(\tfrac{\alpha}{2})S) < \hat{\Delta} < \Delta(1+x(\tfrac{\alpha}{2})S)\}$$

$$\Rightarrow \{\frac{\hat{\Delta}}{1+x(\tfrac{\alpha}{2})S} < \Delta < \hat{\Delta}I(1-x(\tfrac{\alpha}{2})\cdot S)\} \Rightarrow \{\Delta(1-x(\tfrac{\alpha}{2})S) \leq \hat{\Delta} < \Delta(1+x(\tfrac{\alpha}{2})S)\}\,.$$

The function I is defined by:

$$I(x) = \begin{cases} 1/x & \text{if} \quad x > 0 \\ \infty & \text{if} \quad x \leq 0 \end{cases}$$

The above gives a confidence interval for Δ with asymptotical confidence level equal to $1-\alpha$:

$$\Delta \in \langle \frac{\hat{\Delta}}{1+Sx(\frac{\alpha}{2})}\ ,\ \hat{\Delta}I(1-Sx(\frac{\alpha}{2}))\rangle \qquad (53)$$

It is easily seen that $L(52) < L(53)$, also when $1-Sx(\frac{\alpha}{2}) > 0$, where $L(52)$ is the length of the interval given by (52) and the same for $L(53)$. (52) can be applied to construct confidence intervals for any monotone one-to-one function $H(\Delta)$:

$$\hat{\Delta}(1-x(\tfrac{\alpha}{2})S) < \Delta < \hat{\Delta}(1+x(\tfrac{\alpha}{2})S)$$

$$\Updownarrow$$

$$\begin{cases} H(\hat{\Delta}-\hat{\Delta}x(\frac{\alpha}{2})S) < H(\Delta) < H(\hat{\Delta}+\hat{\Delta}x(\frac{\alpha}{2})S) & \text{if } H \text{ is strictly increasing in } \Delta \\ H(\hat{\Delta}+\hat{\Delta}x(\frac{\alpha}{2})S) < H(\Delta) < H(\hat{\Delta}-\hat{\Delta}x(\frac{\alpha}{2})S) & \text{if } H \text{ is strictly decreasing in } \Delta . \end{cases}$$

For example an interval for $\gamma = \frac{\Delta-1}{\Delta+1}$ based upon (52) would be:

$$\gamma \in \langle \max\left(-1, \frac{\hat{\Delta}(1-x(\frac{\alpha}{2})S)-1}{\hat{\Delta}(1-x(\frac{\alpha}{2})S)+1}\right),\ \frac{\hat{\Delta}(1+x(\frac{\alpha}{2})S)-1}{\hat{\Delta}(1+x(\frac{\alpha}{2})S)+1}\rangle \qquad (54)$$

The interval (54) can also be expressed as follows:

$$\gamma \in \langle \max\left(-1, \frac{X_{11}X_{22}(1-x(\frac{\alpha}{2})S)-X_{12}X_{21}}{X_{11}X_{22}(1-x(\frac{\alpha}{2})S)+X_{12}X_{21}}\right),\ \frac{X_{11}X_{22}(1+x(\frac{\alpha}{2})S)-X_{12}X_{21}}{X_{11}X_{22}(1+x(\frac{\alpha}{2})S)+X_{12}X_{21}}\rangle$$

In the 2×2 - table S_γ^2 can be stated as:

LEMMA 20

$$S_\gamma^2 = \frac{4\hat{\Delta}^2}{(\hat{\Delta}+1)^2} n\, S^2$$

<u>Proof</u>.

$$\gamma = 1 - \frac{2}{\Delta+1} \Rightarrow \frac{\partial\gamma}{\partial p_{ij}} = \frac{2\Delta_{ij}}{(\Delta+1)^2} \quad \text{and hence} \quad \gamma^* = 0$$

$$\Rightarrow \sigma_\gamma^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{4\Delta_{ij}^2}{(\Delta+1)^4} p_{ij} = \frac{4}{(\Delta+1)^4}\sigma_\Delta^2 \quad \text{and hence:}$$

$$S_\gamma^2 = \frac{4}{(\hat{\Delta}+1)^4} S_\Delta^2 = \frac{4\hat{\Delta}^2}{(\hat{\Delta}+1)^4} n\, S^2 .$$

Q.E.D.

The confidence interval (32) for γ now becomes

$$\gamma \in \langle \max\left(-1, \frac{\hat{\Delta}-1}{\hat{\Delta}+1} - \frac{2\hat{\Delta}Sx(\frac{\alpha}{2})}{(\hat{\Delta}+1)^2}\right), \min\left(1, \frac{\hat{\Delta}-1}{\hat{\Delta}+1} + \frac{2\hat{\Delta}Sx(\frac{\alpha}{2})}{(\hat{\Delta}+1)^2}\right)\rangle \qquad (55)$$

If $Sx(\frac{\alpha}{2}) \leq 1$ the interval limits in (55) are equal to respectively

$$\frac{\hat{\Delta}-1}{\hat{\Delta}+1} - \frac{2\hat{\Delta}Sx(\frac{\alpha}{2})}{(\hat{\Delta}+1)^2}, \quad \frac{\hat{\Delta}-1}{\hat{\Delta}+1} + \frac{2\hat{\Delta}Sx(\frac{\alpha}{2})}{(\hat{\Delta}+1)^2}.$$

Now let L_1 be the length of interval (54) and L_2 be the length of (55).

LEMMA 21

a) <u>If $S \cdot x(\frac{\alpha}{2}) \leq 1$, then $L_2 < L_1$.</u>

b) <u>If $S \cdot x(\frac{\alpha}{2}) \geq 1 + \max(\hat{\Delta}^{-1}, \hat{\Delta})$, then $L_2 > L_1$.</u>

c) <u>$P(L_2 < L_1) \underset{n \to \infty}{\to} 1$</u>

<u>Proof</u>. The first thing we notice is that

$$S \cdot x(\tfrac{\alpha}{2}) \leq 1 \iff L_1 = \frac{\hat{\Delta}(1+x(\frac{\alpha}{2}) \cdot S)-1}{\hat{\Delta}(1+x(\frac{\alpha}{2}) \cdot S)+1} - \frac{\hat{\Delta}(1-x(\frac{\alpha}{2}) \cdot S)-1}{\hat{\Delta}(1-x(\frac{\alpha}{2}) \cdot S)+1}$$

$$= \frac{4\hat{\Delta}S \cdot x(\frac{\alpha}{2})}{\hat{\Delta}^2(1-S^2x^2(\frac{\alpha}{2}))+2\hat{\Delta}+1}.$$

In addition: $S \cdot x(\frac{\alpha}{2}) \leq 1 \Rightarrow L_2 = \dfrac{4\hat{\Delta}S \cdot x(\frac{\alpha}{2})}{(\hat{\Delta}+1)^2}$.

Hence: $\dfrac{L_2}{L_1} = \dfrac{\hat{\Delta}^2(1-S^2x^2(\frac{\alpha}{2}))+2\hat{\Delta}+1}{(\hat{\Delta}+1)^2}$ when $S \cdot x(\frac{\alpha}{2}) \leq 1$.

If $Sx(\frac{\alpha}{2}) < 1$, then $L_2/L_1 < 1$ since $\hat{\Delta} > 0$ (no X_{ij} equals 0).

If $Sx(\frac{\alpha}{2}) = 1$, then $L_2/L_1 = \dfrac{2\hat{\Delta}+1}{(\hat{\Delta}+1)^2} < 1$, and a) is proved.

Next, let $Sx(\frac{\alpha}{2}) > 1$. Then $L_1 = \frac{\hat{\Delta}(1+x(\frac{\alpha}{2})S)-1}{\hat{\Delta}(1+x(\frac{\alpha}{2})S)+1} + 1 = \frac{2\hat{\Delta}(1+x(\frac{\alpha}{2})\cdot S}{\hat{\Delta}(1+x(\frac{\alpha}{2})\cdot S)+1}$

The upper bound in (55) equals 1 if and only if

$$Sx(\tfrac{\alpha}{2}) \geq 1 + \hat{\Delta}^{-1} .$$

The lower bound in (55) equals -1 if and only if $S\cdot x(\frac{\alpha}{2}) \geq 1 + \hat{\Delta}$.
Assume now that $S\cdot x(\frac{\alpha}{2}) \geq 1 + \max(\hat{\Delta}^{-1},\hat{\Delta}) = \max(1+\hat{\Delta}^{-1},1+\hat{\Delta})$. Then

$L_2 = 2$ and $L_2/L_1 = \frac{\hat{\Delta}(1+x(\frac{\alpha}{2})\cdot S)+1}{\hat{\Delta}(1+x(\frac{\alpha}{2})\cdot S)} > 1$, and b) is proved.

From a): $\lim_{n\to\infty} P(L_2 < L_1) \geq \lim_n P(S^2 \leq x^{-2}(\frac{\alpha}{2}))$.

Now X_{ij} is binomial (n,p_{ij}) , and by using the fact that

$Y_{ij} = \frac{X_{ij} - n\,p_{ij}}{\sqrt{n\,p_{ij}(1-p_{ij})}}$ is asymptotically normal we find that for

$y > 0$

$$P(X_{ij}^{-1} \leq y) = P(X_{ij} \geq y^{-1}) = 1 - P(Y_{ij} < \frac{y^{-1} - n\,p_{ij}}{\sqrt{n\,p_{ij}(1-p_{ij})}}) \underset{n\to\infty}{\to} 1 .$$

Let now $s > 0$. Then $X_{ij}^{-1} < \frac{s}{4}$ for $i,j = 1,2 \Rightarrow S^2 \leq s$.
This implies that $\lim P(S^2 \leq s) \geq 1 - \lim_n P(\bigcup_{i,j} X_{ij}^{-1} > \frac{s}{4}) = 1$.
This proves c) by putting $s = x^{-2}(\frac{\alpha}{2})$.

Q.E.D.

Asymptotically the interval (55) is therefore better than (54).

A confidence interval for $\rho = \ln\Delta$, based upon (52), is given by:

$$\rho \in \langle \hat{\rho}+\ln(1-Sx(\tfrac{\alpha}{2})),\ \hat{\rho}+\ln(1+Sx(\tfrac{\alpha}{2}))\rangle .$$

Here $\hat{\rho}$ is equal to $\ln\hat{\Delta}$.

A confidence interval for ρ^2 (from lemma 16) is given by:

$$\rho^2 \in \langle \max(0,\hat{\rho}^2 - 2|\hat{\rho}|Sx(\tfrac{\alpha}{2})),\ \hat{\rho}^2 + 2|\hat{\rho}|Sx(\tfrac{\alpha}{2})\rangle .$$

III. 6 (iv) Confidence intervals for an alternative measure of association.

A measure not depending on Δ is

$$\tau_b = \frac{p_{11}p_{22} - p_{12}p_{21}}{(p_{1.}p_{2.}p_{.1}p_{.2})^{\frac{1}{2}}}$$

or $\beta = \tau_b^2$ if we are only interested in the degree of association. For testing of a.i. based on β we fefer to III.3 (ii). In the 2×2 - case one finds that S_η^2 (here called S_β^2) can be expressed as follows:

LEMMA 22

$$S_\beta^2 = \hat{\mu}^{-4}\hat{\theta}^2\{q_{11}b_{22}^2+q_{22}b_{11}^2+q_{12}b_{21}^2+q_{21}b_{12}^2-(q_{11}b_{22}+q_{22}b_{11}-q_{12}b_{21}-q_{21}b_{12})^2\}.$$

where

$$\hat{\mu} = q_{1.}q_{2.}q_{.1}q_{.2}$$

$$\hat{\theta} = q_{11}q_{22}-q_{12}q_{21}$$

and

$$b_{ij} = q_{i.}q_{.j}\{q_{i.}q_{.j}(2-q_{i.}-q_{.j}) - q_{ij}(q_{i.}+q_{.j}-2q_{i.}q_{.j})\}.$$

Proof.

$$S_\beta^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} q_{ij}(\hat{\beta}_{ij}-\hat{\beta}^*)^2 \quad \text{where} \quad \hat{\beta}_{ij} = \frac{\partial\hat{\beta}}{\partial q_{ij}} \quad \text{and} \quad \hat{\beta}^* = \sum_{i=1}^{2}\sum_{j=1}^{2}\hat{\beta}_{ij}q_{ij}.$$

Let $M = (q_{11}q_{22}-q_{12}q_{21})^2 = \hat{\theta}^2$.

$$\hat{\beta} = \frac{M}{\hat{\mu}} \qquad \hat{\beta}_{ij} = \frac{1}{\hat{\mu}^2}\{\frac{\partial M}{\partial q_{ij}}\hat{\mu} - M\frac{\partial\hat{\mu}}{\partial q_{ij}}\}.$$

$$\frac{\partial M}{\partial q_{11}} = 2\hat{\theta}\cdot q_{22}, \quad \frac{\partial M}{\partial q_{22}} = 2\hat{\theta}\cdot q_{11}, \quad \frac{\partial M}{\partial q_{12}} = -2\hat{\theta}\cdot q_{21} \quad \text{and} \quad \frac{\partial M}{\partial q_{21}} = -2\hat{\theta}\cdot q_{12}.$$

$$\frac{\partial\hat{\mu}}{\partial q_{11}} = q_{2.}q_{.2}(q_{1.}+q_{.1}),\ \frac{\partial\hat{\mu}}{\partial q_{22}} = q_{1.}q_{.1}(q_{2.}+q_{.2})$$

$$\frac{\partial\hat{\mu}}{\partial q_{12}} = q_{2.}q_{.1}(q_{1.}+q_{.2}),\ \frac{\partial\hat{\mu}}{\partial q_{21}} = q_{1.}q_{.2}(q_{2.}+q_{.1})$$

The above implies:

$$\hat{\beta}_{11} = \hat{\mu}^{-2}\hat{\theta}b_{22}$$

$$\hat{\beta}_{22} = \hat{\mu}^{-2}\hat{\theta}b_{11}$$

$$\hat{\beta}_{12} = -\hat{\mu}^{-2}\hat{\theta}b_{21}$$

and $\hat{\beta}_{21} = -\hat{\mu}^{-2}\hat{\theta}b_{12}$.

The result now follows readily.

Q.E.D.

Confidence interval for β :

$$\beta \in \langle \max\,(0,\hat{\beta} - \frac{S_\beta}{\sqrt{n}}\, x(\tfrac{\alpha}{2})),\ \min\,(1,\hat{\beta} + \frac{S_\beta}{\sqrt{n}}\, x(\tfrac{\alpha}{2}))\rangle\ ,$$

where

$$\hat{\beta} = \frac{\hat{\theta}^2}{\hat{\mu}} = \frac{(X_{11}X_{22}-X_{12}X_{21})^2}{X_{1.}X_{2.}X_{.1}X_{.2}}\ .$$

Let $C_1 = \hat{\beta} - \frac{S_\beta}{\sqrt{n}}\, x(\frac{\alpha}{2})$ and $C_2 = \hat{\beta} + \frac{S_\beta}{\sqrt{n}}\, x(\frac{\alpha}{2})$.

Then an interval for τ_b with confidence level $\geq 1-\alpha$ is given by:

$$\tau_b \in \langle -\min\,(1,\sqrt{C_2}\,)\ ,\ \min\,(1,\sqrt{C_2}\,)\rangle \qquad (56)$$

From III.2 (iii) a confidence interval for τ_b with asymptotic level equal to $1-\alpha$ is given. It is easily seen that

$$S_b^2 = \frac{1}{4\hat{\tau}_b^2}\, S_\beta^2 \qquad (\frac{\partial\beta}{\partial p_{ij}} = 2\tau_b \cdot \frac{\partial\tau_b}{\partial p_{ij}})\ .$$

Hence the interval for τ_b in III.2 (iii) can be expressed as follows:

$$\tau_b \in \langle \max\,(-1,\hat{\tau}_b - \frac{S_\beta}{2|\hat{\tau}_b|\sqrt{n}}\, x(\tfrac{\alpha}{2})),\ \min\,(1,\hat{\tau}_b + \frac{S_\beta}{2|\hat{\tau}_b|\sqrt{n}}\, x(\tfrac{\alpha}{2}))\rangle \qquad (57)$$

Let $K_1 = \hat{\tau}_b - \frac{S_\beta}{2|\hat{\tau}_b|\sqrt{n}}\, x(\frac{\alpha}{2})$ and $K_2 = \hat{\tau}_b + \frac{S_\beta}{2|\hat{\tau}_b|\sqrt{n}}\, x(\frac{\alpha}{2})$.

Let further $L_1^* = 2\sqrt{C_2}$, that is the length of the interval $\langle -\sqrt{C_2}, \sqrt{C_2} \rangle$, and similar $L_2^* = K_2 - K_1$. The following lemma gives us some results about the relation between L_1^* and L_2^* .

LEMMA 23

(a) $C_1 > 0 \Rightarrow L_2^* < \frac{\sqrt{2}}{4} L_1^*$

(b) $\frac{L_1^*}{L_2^*} > 1 \Leftrightarrow \frac{\hat{\tau}_b^2 \sqrt{n}}{S_\beta x(\frac{\alpha}{2})} > \frac{\sqrt{2}-1}{2}$ $(= 0.207)$

(c) $\lim_{n \to \infty} P(L_2^* < L_1^*) = 1$.

Proof.

(a) We find:

$$\left.\begin{aligned} L_1^* &= 2(\hat{\beta} + \frac{S_\beta}{\sqrt{n}}\, x(\frac{\alpha}{2}))^{\frac{1}{2}} \\ L_2^* &= \frac{S_\beta}{|\hat{\tau}_b|\sqrt{n}}\, x(\frac{\alpha}{2}) \end{aligned}\right\} \Rightarrow \left(\frac{L_1^*}{L_2^*}\right)^2 = \frac{4\hat{\tau}_b^2\sqrt{n}}{S_\beta x(\frac{\alpha}{2})}\left[\frac{\hat{\tau}_b^2\sqrt{n}}{S_\beta x(\frac{\alpha}{2})} + 1\right]$$

Now: $C_1 > 0 \Leftrightarrow \hat{\tau}_b^2\sqrt{n} > S_\beta\, x(\frac{\alpha}{2}) \Rightarrow \left(\frac{L_1^*}{L_2^*}\right)^2 > 4(1+1) = 8$

$$\Leftrightarrow \left(\frac{L_1^*}{L_2^*}\right) > 2\sqrt{2} \Leftrightarrow L_2^* < \frac{\sqrt{2}}{4} L_1^* .$$

(b) $\left(\frac{L_1^*}{L_2^*}\right) > 1 \Leftrightarrow 4\hat{\tau}_b^4 n + 4\hat{\tau}_b^2\sqrt{n}\, S_\beta x > S_\beta^2 x^2$, where $x = x(\frac{\alpha}{2})$.

Let $y = \hat{\tau}_b^2\sqrt{n}$ and $b = S_\beta \cdot x$, $y \geq 0$, $b > 0$. It now follows that

$$L_1^*/L_2^* > 1 \Leftrightarrow y^2 + by - b^2 \cdot 4^{-1} > 0 .$$

$$\Leftrightarrow (y - \frac{\sqrt{2}-1}{2} b)(y + \frac{\sqrt{2}+1}{2} b) > 0$$

$$\Leftrightarrow y - \frac{\sqrt{2}-1}{2} b > 0 \Leftrightarrow \frac{y}{b} > \frac{\sqrt{2}-1}{2} .$$

(c) From (b) we now have:

$$\lim_{n\to\infty} P(L_2^* < L_1^*) = \lim_{n\to\infty} P\left(\frac{\hat{\tau}_b\sqrt{n}}{S_\beta} > \frac{\sqrt{2}-1}{2}x\right) .$$

$\tau_b^2 > 0$, so that the interval (56) has meaning only at the assumption of $\tau_b \neq 0$. In that case

$$\frac{\hat{\tau}_b^2}{S_\beta} \xrightarrow{P} \frac{\tau_b^2}{\sigma_\beta} = a > 0 .$$

Let now $Y_n = \sqrt{n}\,\frac{\hat{\tau}_b^2}{S_\beta}$ and let $k > 0$:

$$P(Y_n > k) = P\left(\frac{\hat{\tau}_b^2}{S_\beta} > \frac{k}{\sqrt{n}}\right) \underset{n\to\infty}{\to} 1 \quad \text{since} \quad \frac{\hat{\tau}_b^2}{S_\beta} \xrightarrow{D} a > 0 .$$

Substituting k with $\frac{\sqrt{2}-1}{2}x$ and we have obtained result (c).

Q.E.D.

III. 7. 3 examples.

The three examples to be presented are taken from [15]. In each example we shall use the measures we think are suitable.
Confidence intervals will be stated, and testing of a.i. will be done. As for the numerical calculations these have not been checked, so some reservation must be taken for the results.
For each measure d for degree of association being considered, the a.i. hypothesis in the examples will be chosen equal to

$$H^* : d \leq 0.0025$$

For example, if $d = \gamma^2$, then ϵ is chosen equal to 0.05 in the a.i. criterion (33).

Example 1.

Let us first go back to example 1 from I.1. Let A be occupation.

The eight occupational groups in the table are:

A_1 : Self-employed in agriculture, forestry and fishing.

A_2 : Other self-employed.

A_3 : Wage earners in manufacturing, construction and mining.

A_4 : Other wage earners.

A_5 : Pupils, students.

A_6 : Pensioners.

A_7 : Housewives.

A_8 : Others.

It was shown earlier the chi-square test will assert association between occupation and participation (factor B). When testing a.i. the result will be to accept independence between the factors for those suitable measures that are considered. Usually one will think of the occupational groups $A_1,\dots,A_8$ to hold no relevant ordering, so Y has nominal level. As for the characteristics "voters" and "non-voters" these can be considered both as ordered or not depending on the kind of problem one is trying to elucidate. If, for instance, participation is used as an indicator for the interest of the election, it can be meaningful to say that the characterisics are ordered. We will then have a mixed situation. It seems reasonable to consider B as the primary factor, so from the considerations in part 1 a suitable measure in this case will be γ.

The measure for degree of association is then γ^2, and the a.i. hypothesis becomes:

$$H_1^* : \gamma^2 \leq 0.0025$$

We find that $\hat{\gamma} = 0.0146$ and $\hat{\gamma}^2 = 0.0002$.

Since $\hat{\gamma}^2 < 0.0025$ the a.i. hypothesis is accepted. This means that we accept that the factors are approximately independent. (The

expression "accept" means in this connection: "fail to reject".)
The estimator for the asymptotic variance of $\sqrt{n}\,\hat{\gamma}$ is:

$$S_{\gamma}^{2} = 5.577$$

Hence we have the following confidence interval for γ with asymptotical confidence level equal to 0.95:

$$-0.0744 < \gamma < 0.1036$$

We see that $2|\hat{\gamma}|\frac{S_{\gamma}}{\sqrt{n}} = 0.00133$, so a 95%-interval for γ^2 (35) becomes:

$$0 < \gamma^{2} < 0.0028$$

The onesided confidence interval (36) with $1-\alpha$ equal to 0.95 gets a lower bound equal to zero, which means that all extended hypotheses, $\gamma^2 \leq c$, is accepted at level $\alpha = 0.05$.

Let us consider the situation that arises when there is said to be no relevant ordering between voters and non-voters.
B is still regarded as the factor of primary interest. The situation therefore becomes unordered and asymmetrical. Practicable measures are hence λ_b or η_b. We find $\hat{\lambda}_b = 0$ so all extended hypotheses based on λ_b are accepted. The a.i. hypothesis based on η_b:

$$H_2^* : \eta_b \leq 0.0025$$

Results:

$$\hat{\eta}_b = 0.0072 .$$

Estimated asymptotic variance: $S_{ob}^2 = 0.0525$.
The N-test with level $\alpha = 0.05$. Reject H_2^* if

$$T_\eta = \frac{\sqrt{n}(\hat{\eta}_b - 0.0025)}{S_{ob}} > 1.645$$

We find $T_\eta = 1.07$ which means that we accept H_2^* .
The confidence interval for η_b with $\alpha = 0.05$ is given by

$$0 < \eta_b < 0.0158$$

The onesided interval has lower bound equal to zero, so in fact all extended hypotheses, $\eta_b \leq c$, are accepted at level 0.05 .

Finally we shall state the confidence intervals for
$\varkappa_1 = \max_{i,j} |p_{ij} - p_{i.}p_{.j}|$ and $\varkappa_2 = \max_{i,j} |\frac{p_{ij}}{p_{i.}} - p_{.j}|$, derived in III. 3 (iv).

Results: $\hat{D} = 0.000068$, $S_D^2 = 0.000006$.
95%-interval for $\varkappa_1$: $0 < \varkappa_1 < 0.009$

$$\hat{E} = 0.06179, \quad S_E^2 = 5.9999 .$$

95%-interval for $\varkappa_2$: $0 < \varkappa_2 < 0.2776$.

It is worth noticing that in this example the choice of a.i. hypothesis is no problem because every extended hypothesis is accepted without regard to which of the measures γ , λ_b or η_b that are prefered.

Example 2. In this example the dependence between the factors income and participation will be investigated. We let factor A be (yearly)-income and factor B participation. The number of observations in the Bureau's interview survey was $n = 2702$.
The result of the survey arranged in a two-way contingency table was:

Table 3

Participation / Income	B_1 Voters	B_2 Non-voters	Total
A_1: Less than kr 10.000	400	84	484
A_2: kr 10.000-19.900	517	64	581
A_3: kr 20.000-29.900	785	68	853
A_4: kr 30.000-39.900	398	32	430
A_5: kr 40.000-49.900	194	9	203
A_6: kr 50.000-and more	145	6	151
Total	2439	263	2702

($1 = kr 5.50)

Source: [15], table 11.

For each income group we can compute the proportion of voters/non-voters. This gives the following table:

Table 4

Income group	Voters	Non-voters
A_1	0.83	0.17
A_2	0.89	0.11
A_3	0.92	0.08
A_4	0.93	0.07
A_5	0.96	0.04
A_6	0.96	0.04
All income groups	0.90	0.10

The share of voters increases with increasing income. It seems that there is present a certain degree of association. The question is if the degree of association in the table is significant. The usual chi-square test in I. 2. rejects the exact independence hypothesis for all significance levels > 0.001 .

We will assume a relevant ordering in B . Income establishes obviously an ordering, so the situation is ordered. γ^2 is herewith a suitable measure for degree of association. The a.i. hypothesis is therefore equal to H_1^* in example 1 :

$$H_1^* \ : \ \gamma^2 \leq 0.0025$$

We find:

$$\hat{\gamma} = -0.3080 \,, \quad \hat{\gamma}^2 = 0.0949$$

The hypothesis is rejected with level $\alpha = 0.05$ if

$$T_\gamma = \frac{\sqrt{n}\,(\hat{\gamma}^2 - 0.0025)}{2|\hat{\gamma}|S_\gamma} > 1.645$$

The results are:

$$S_\gamma^2 = 5.329 \quad \text{and} \quad 2|\hat{\gamma}|\frac{S_\gamma}{\sqrt{n}} = 0.0274 \,, \quad \text{which gives}$$

$$T_\gamma = 3.37 \,.$$

Conclusion: Association between income and participation.

95% confidence intervals for γ and γ^2, (32) and (35):

$$-0.3950 < \gamma < -0.2210$$

$$0.0412 < \gamma^2 < 0.1486$$

The onesided confidence interval (36) for γ^2 with $\alpha = 0.05$ becomes:

$$\lambda^2 > 0.0498$$

I.e. that all hypotheses: $\gamma^2 \leq c$, $c < 0.0498$ will be rejected at level 0.05 .

As in example 1 we give 95% confidence intervals for $\varkappa_1$ and $\varkappa_2$:

Results: $\hat{D} = 0.000529 \quad S_D^2 = 0.000095$

$$0.0037 < \varkappa_1 < 0.0212$$

$$\hat{E} = 0.02587 , \quad S_E^2 = 0.1698$$

$$0.0293 < \varkappa_2 < 0.1439$$

Example 3. The association between education and partysympathy shall be examined. Let factor A be level of education and B partysympathy. The starting-point is from [15], table 27 but we have withdrawn those 26 who voted, but did not specify to which party. In addition, the votes for SF and K are added together. The result of the interview survey, expressed by the cell-frequencies X_{ij}, was:

Table 5.

Education \ Party	SF/K	A	V	Sp	Kr.F.	H	Total
A_1: Primary school lower stage	35	748	72	152	107	101	1215
A_2: Primary school upper stage	11	322	71	103	71	171	749
A_3: Secondary school	8	93	44	39	26	97	307
A_4: Post secondary school and University	8	16	21	11	21	65	142
Total	62	1179	208	305	225	434	n=2413

Source: [15], table 27.

In the table the following letters are used for the political parties:

A = Labour Party

H = Conservative Party

Kr.F. = Christian Democrats

K = Communist Party

Sp = Center Party

SF = Socialist People's Party

V = Liberal Party

At each level of education we compute the portion who voted for the different parties. This gives the following table:

Table 6

Level of education	SF/K	A	V	Sp	Kr.F.	H
Primary school, lower stage	0.03	0.62	0.06	0.12	0.09	0.08
Primary school, upper stage	0.01	0.43	0.095	0.14	0.095	0.23
Secondary school	0.03	0.30	0.14	0.13	0.08	0.31
University & postsecondary school	0.06	0.11	0.15	0.08	0.15	0.46
All levels of education	0.03	0.49	0.09	0.13	0.09	0.18

There can be no doubt that level of education and partysympathy are strongly associated. The chi-square test rejects the exact independence hypotheses at any usual level.

It seems to be several alternative ways to interpret the situation in this example. It is quite clear that generally there is a relevant ordering between levels of education. If partysympathy is used to indicate political direction on the scale leftorientated-rightorientated, we might say there is an ordering between the parties in the forgoing meaning. In tables 5 and 6 the parties are ordered (subjectively judged) on such a scale. Relevant measures are ordinal measures, and γ^2 is hence chosen as measure for degree of association. The a.i. hypothesis is:

$H_1^*: \; \gamma^2 \leq 0.0025$

We find: $\hat{\gamma} = 0.3720 \qquad S_\gamma^2 = 1.260$ and hence

$$\sqrt{n}\,\frac{(\hat{\gamma}^2 - 0.0025)}{2|\hat{\gamma}|S_\gamma} = 7.99$$

Conclusion: Education and partysympathy are dependent.

95%- confidence intervals for γ and γ^2, (32) and (35):

$$0.3271 < \gamma < 0.4169$$

$$0.1051 < \gamma^2 < 0.1717$$

The onesided 95%- interval (36) becomes:

$$\gamma^2 > 0.1104$$

I.e. all hypotheses $\gamma^2 \leq c$ where $c < 0.1104$ will be rejected at level 0.05.

The above interpretation of the situation is of course not necessarily always the most relevant one. If the given problem indicates that it is desirable to consider the parties without ordering, but still think of education as ordered, we will have a mixed case.

It then seems reasonable to consider the situation to be asymmetrical with B as primary factor, so that λ_b or η_b will be natural choices. Let us choose η_b as measure for degree of association

The a.i. hypothesis is:

$H_2^*: \; \eta_b \leq 0.0025$

Results: $\hat{\eta}_b = 0.0513$, $S_{ob}^2 = 0.0763$ and hence

$$\frac{\sqrt{n}\,(\hat{\eta}_b - 0.0025)}{S_{ob}} = 8.71$$

Conclusion: We reject H_2^* (at <u>infinitesimal</u> levels < 0.00001)

Further one finds: $\hat{\lambda}_b = 0.0430$.

Estimated asymptotic variance: $S_b^2 = 0.4030$.

Now we get the following twosided 95%- confidence intervals for η_b and λ_b:

$$0.0403 < \eta_b < 0.0623$$

$$0.0177 < \lambda_b < 0.0683$$

Onesided 95%- intervals:

$$\eta_b > 0.0421$$

$$\lambda_b > 0.0218$$

This implies that all hypotheses, $\eta_b \leq c$; $c < 0.0421$ and all hypotheses, $\lambda_b \leq c$; $c < 0.0218$ is rejected.

As a third alternative interpretation the situation is considered as unordered and symmetrical. Suitable measures are then λ and η.

Results:

$$\hat{\eta} = 0.0517$$

$$\hat{\lambda} = 0.0506$$

$$S_\eta^2 = 0.0742$$

$$S_\lambda^2 = 0.2101$$

Onesided 95%- intervals:

$$\eta > 0.0425$$

$$\lambda > 0.0353$$

Twosided 95%- intervals:

$$0.0407 < \eta < 0.0627$$

$$0.0324 < \lambda < 0.0688$$

Finally, we state 95%- intervals for $\varkappa_1$ and $\varkappa_2$:

$$0.0167 < \varkappa_1 < 0.1059$$

$$0.1012 < \varkappa_2 < 0.6236$$

REFERENCES.

[1] Bjørnstad, J.F. (1975): "Comparison of contingency tables", Statistical Research Report, Univ. of Oslo.

[2] Edwards, A.W.F. (1963): "The measure of association in a 2×2-table". J.R.Statist.Soc., A126, 109-114.

[3] Fisher, R.A. (1962): "Confidence limits for a cross-product ratio", Austral.J.Statist., Vol.4, 41.

[4] Goodman, L.A. (1964): "Simultaneous confidence limits for cross-product ratios in contingency tables", J.R.Statist.Soc., A126, 94-108.

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

[6] Goodman, L.A. & Kruskal, W.H. (1959): "Measures of association for cross classifications II. Further discussions and references". J.Am.Statist.Ass., Vol.54, 123-163.

[7] Goodman, L.A. & Kruskal, W.H. (1963): "Measures of association for cross classifications III. Approximate sampling theory". J.Am.Statist.Ass., Vol.58, 310-364.

[8] Goodman, L.A. & Kruskal, W.H. (1972): "Measures of association for cross classifications IV. Simplification of asymptotic variances", J.Am.Statist.Ass., Vol.67, 415-421.

[9] Hodges jr. & Lehmann, E.L. (1954): "Testing the approximate validity of statistical hypotheses", J.R.Statist.Soc., B.16, 261-268.

[10] Kendall, M.G. (1955): "Rank Correlation Methods", Charles Griffin & Co. Lim.

[11] Kruskal, W.H. (1958): "Ordinal measures of association", J.Am. Statist.Ass., Vol 53, 814-861.

[12] Lehmann, E.L. (1959): "Testing Statistical Hypotheses", John Wiley & Sons, Inc.

[13] Mosteller, F. (1968): "Association and estimation in contingency tables", J.Am.Statist.Ass., Vol.63, 1-28.

[14] Neyman, J. (1949): "Contribution to the theory of the χ^2-test", Proc.Berkeley Symp.Math.Statist., 239-273.

[15] NOS A 340: "Storting Elections 1969, Vol.II", Central Bureau of Statistics, Oslo-Norway.

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

[17] Stuart, A. (1953): "The estimation and comparison of strenghts of association in contingency tables", Biometrika, Vol.39, 105-110.

[18] Sverdrup, E. (1971): "Testing statistical hypotheses: The general Neyman-Pearson theory", Statistical Memoirs, University of Oslo.

[19] Sydsæter, K. (1973): "Matematisk analyse, bind II", Universitetsforlaget, Oslo-Bergen-Tromsø.