ESTIMATION AND TESTING OF ODDS RATIO : A REVIEW

MANISHA G. ABHYANKAR

A DISSERTATION SUBMITTED IN PARTIAL
FULFILMENT OF THE REQUIREMENT FOR THE
DEGREE OF MASTER OF PHILSOPHY

DEPARTMENT OF STATISTICS
UNIVERSITY OF POONA
PUNE - 411007

1 9 9 5

ACKNOWLEDGEMENT

It is of great pleasure to express my deep sense of gratitude towards Dr A. V. Kharshikar for his valuable guidance, unfailing patience and constant encouragement throughout this work.

I am grateful to Professor J. V. Deshpande, former Head, and Professor S. R. Paranjape, Head of the Department of Statistics, University of Poona for the research and library facilities provided to me.

My thanks are due to Professor B. K. Kale, Dr G. B. Marathe, Dr S. Kunte and Professor J. V. Deshpande for the excellent teaching of the course work.

I also wish to thank all my friends in the Department for their co-operation.

It is my duty to thank C.S.I.R., New Delhi, for awarding me the Junior Research Fellowship.

(Manisha G. Abhyankar)

Department of Statistics,
University of Poona,
Pune - 411007

INDEX

Chapter 1

Introduction

CHAPTER 1

INTRODUCTION

1.1 : Prelimineries

Contingency table :

A contingency table is a table of observed frequencies classified according to different categories. In this dissertation, we concentrate on two-way and three-way contingency tables.

When the population is cross classified with respect to two classifications or polytomies, we get a two-way continency table. For example, a group of individuals can be classified according to two attributes viz. sex and smoking habit. The attribute sex has two levels (i) male and (ii) female, while a person can be classified as smoker or non-smoker according to smoking habit. Thus each individual is classified in one of the four cells. This classification is known as 2x2 contingency table, which is the simplest form of two-way contingency tables. In general, suppose we have two attributes A and B where A can be classified in 'r' levels, A_1, A_2, ..., A_r; and B can be classified in 's' levels B_1, B_2, ..., B_s and we get a two-way contingency table with 'rs' cells.

A three-way contingency table is obtained if the population is clasified according to three attributes. For example, suppose, we are interested in the relationship between smoking and cancer. Then we may classify the individuals according to disease status (present, absent) and smoking habit (smoker, non-smoker). A three-way contingency table is then obtained if each individual is further classified according to third criterion e.g. sex (male, female).

The variables with respect to which the population is cross-classified may be classified as ordinal or nominal variables. A categorical variable is referred to as "ordinal" when there is clear ordering of the categories but the absolute distances among them are unknown. For example, the variable "education" is ordinal when measured with categories grammar school, high school, college, post-graduate etc. Political philosophy is ordinal because it can be classified as liberal, moderate and conservative. Here a person classified as moderate is more liberal than a person classified as conservative, but there is no obvious way to quantify numerically how much more liberal that person is. A categorical variable is referred to as "nominal" when ordering of categories is unimportant. Examples of nominal variables are sex, race, country of residence, marital status etc.

An ordinal variable is quantitative in the sense that each level on its scale can be compared in terms of whether it corresponds to a greater or smaller magnitude of a certain characteristic than another level. A nominal variable is qualitative and distinct levels of such variables differ in quality, not in quantity.

Association :

The concept of association refers to 'dependence' which may or may not be causal between two or more variables.

While analysing contingency tables, it is natural to test whether the factors are independent as a first step. One cannot stop at testing for independence when there is ground to believe that there is association between the factors, and when it is only a formality to test for independence.

A statistical explication of an association is made in terms of various measures of association, the correlation coefficient being a familiar example. The odds ratio and the relative risk are two others. In 2x2 contingency table, χ^2/N can be looked upon as product moment correlation coefficient between the two characteristics.

Goodman and Kruskal (1954) reviewed a large amount of literature that dates from the late 19th century where many statisticians and non-statisticians have contributed by proposing measures of association in consideration to problem at hand and these have been proposed for use in as varied fields as sociology, psychology, meterology, linguistics etc.

In this dissertation, we concentrate on odds ratio or cross-product ratio which is one of the most important measures of association.

Odds ratio in a 2x2 contingency table :

Consider general 2x2 population cross-classification in the following table.

Row \ Column	1	2	
1	p_{11}	p_{12}	$p_{1.}$
2	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

Within row 1, the odds that the response is in column 1 instead of column 2 is defined to be

$$O_1 = p_{11}/p_{12}$$

Within row 2, the corresponding odds equals

$$O_2 = p_{21}/p_{22} .$$

The ratio of these odds $\psi = \frac{O_1}{O_2} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$ is referred to as the 'odds ratio'. An alternative name for it is the cross-product ratio since ψ equals the ratio of products $p_{11}p_{22}$ and $p_{12}p_{21}$ of proportions from cells that are diagonally opposite.

Properties of odds ratio :

(i) Odds ratio ψ takes values from 0 to ∞.

(ii) Note that each odds O_i can be expressed as

$$O_i = p_{1(i)}/p_{2(i)}$$

where $p_{j(i)}$ represent conditional probability of (i,j)-th cell given the i-th row. Thus,

$$\psi = \frac{p_{1(1)}/p_{2(1)}}{p_{1(2)}/p_{2(2)}} .$$

The two conditional distributions $(p_{1(1)}, p_{2(1)})$ and $(p_{1(2)}, p_{2(2)})$ are identical and hence the two variables are independent if and only if $O_1 = O_2$. In this case the odds ratio $\psi = 1$.

(iii) If $1 < \psi < \infty$; individuals in row 1 are more likely to make the first response than are individuals in row 2.

(iv) If $0 \le \psi < 1$; individuals in row 1 are less likely to make the first response than are individuals in row 2.

(v) ψ is invariant under change in both rows and columns i.e. if both rows and columns are interchanged, ψ remains the same.

(vi) If the order of rows or columns is reversed, the new value of ψ is simply the inverse of the original value.

(vii) ψ is invariant under the transformation $p_{ij} \rightarrow t_i s_j p_{ij}$ for any sets of positive numbers $\{t_i\}$, $\{s_j\}$ that preservs $\sum_i \sum_j p_{ij} = 1$.

Odds ratios in rxs contingency table :

For the general rxs contingency table, odds ratios can be formed using each of the $\binom{r}{2} = \frac{r(r-1)}{2}$ pairs of rows in combination with each of the $\binom{s}{2} = \frac{s(s-1)}{2}$ pairs of columns. For rows i and i' and columns j and j' the odds ratio $\frac{p_{ij} p_{i'j'}}{p_{ij'} p_{i'j}}$ uses four cells occuring in a rectangular pattern and there are $\binom{r}{2}\binom{s}{2}$ odds ratios of this type. The independence of the two variables is equivalent to condition that all these population odds ratios equal one.

There is much redundancy of information when the entire set of these odds ratios is used to characterize the association in a rxs contingency table. For example, consider the set of (r-1) (s-1) odds ratios

$$\psi_{ij} = \frac{p_{ij}\, p_{rs}}{p_{is}\, p_{rj}} \quad , \; i = 1,2,\ldots, r-1; \; j = 1,2,\ldots,s-1 \tag{1.1.1}$$

Each odds ratio is formed using the rectangular array of cells determined by rows i and r and columns j and s. These (r-1)(s-1) odds ratios determine all $\binom{r}{2}$ $\binom{s}{2}$ odds ratios that can be formed from pairs of rows and pairs of columns. Independence of the two variables is therefore also equivalent to the condition that the odds ratios in the set (1.1.1) are identically equal to one.

The construction (1.1.1) for froming a minimal set of odds ratios that determine the entire set is not unique. For example another basic set of (r-1) (s-1) odds ratios is

$$\psi^{*}_{ij} = \frac{p_{ij} \quad p_{i+1\ j+1}}{p_{i\ j+1}\ p_{i+1\ j}} \text{ , } i = 1,2,\ldots, r-1;\ j = 1,2,\ldots,s-1 \qquad (1.1.2)$$

In this dissertation, we consider basic set of the odds ratios in (1.1.1).

1.2 : Outline of the Chapters

Two major areas of statistical inference are estimation of parameters and testing of hypothesis. As mentioned earlier, in this dissertation, we are concerned with the analysis of two-way and three-way contingency tables through odds ratios. Chapter II and III deal with two-way contingency tables while Chapter IV deals with three-way contingency tables.

The most simple form of two-way contingency table viz. 2x2 contingency table is discussed in chapter II. For 2x2 contingency table, the parameter of interest is the odds ratio or cross-product ratio (ψ). Point and interval estimation for ψ based on Fisher's exact distribution is discussed along with various asymptotic methods. Last section of chapter II reviews literature regarding the procedures for testing the hypothesis of independence of two varibales.

The most general case of two-way contingency tables is a rxs contingency table (r, s > 2). For simplicity of notations, we restrict to a 2x3 contingency table in Chapter III. The method of interval estimation based on exact distribution for 2x2 case does not readily extend to the 2x3 table. We have considered various asymptotic methods for interval as well as point estimation for the two odds ratios in a 2x3 contingency table. Test procedures to check independence are considered in last section.

In chapter IV, we consider analysis of the simplest form of three-way contingency table, viz., 2x2x2 contingency table. Assuming that the odds ratio remains constant across the two 2x2 tables, point and interval estimation procedures for this common odds ratio ψ are discussed along with the procedures for testing the hypothesis $H_o : \psi = 1$. It may be noted that the procedures described for two 2x2 tables readily extend to t (t > 2) 2x2 tables.

Various paradoxes that may occur while amalgamating the contingency tables are discussed in chapter V.

Material presented in all these chapters is a review of the results regarding odds ratio in 2x2, 2x3 and 2x2x2 contingency tables. We have given illustrative examples throughout, with a view to explain the theory discussed, as it is used in practice.

Chapter 2

2 x 2 Contingency Tables : One Odds Ratio

CHAPTER 2

2 X 2 CONTINGENCY TABLES : ONE ODDS RATIO

2.1 : Introduction

Odds ratio as a measure of association in a 2x2 contingency table :

In 2x2 contingency tables only one degree of freedom is available to measure association. Due to this most standard measures are said to reduce to functions of cross-product ratio or chi-square statistic (Bishop, Fienberg and Holland 1975). Here we study the cross-product ratio or odds ratio which is often used as an index of association in a 2x2 contingency table.

Sampling schemes :

The extensive literature on 2x2 tables reflects the range of different sampling schemes that may underlie such tables. Barnard (1947a) was the first to observe that there were at least three distinct sampling schemes leading to a 2x2 table. Kudo and Tamuri (1978) have observed that other sampling schemes, based on negative binomial distribution may arise. Here we consider three distinct schemes described by Barnard (1947a) corresponding to tables having zero, one or two margins fixed.

Origin I data (no fixed margins/double dichotomy) :

Consider that we have a four-fold universe in which individuals can be classified according to two attributes A and B each at two levels. This doubly dichotomous population can be represented by

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

In this population p_{ij} ($i = 1,2$; $j = 1,2$) represent the proportions falling into i-th level of A and j-th level of B.

p_{11}/p_{12} represents the odds of being in the first level of B for the first level of A while p_{21}/p_{22} denote the odds of being in the first level of B for the second level of A and odds ratio, ψ, is defined by

$$\psi = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}} .$$

If we draw N individuals at random (independently) from this population, our sample outcome may take the form

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

Thus, in this case the sample $(x_{11}, x_{12}, x_{21}, x_{22})$ is treated as realization of $(X_{11}, X_{12}, X_{21}, X_{22}) \sim$ Multinomial $(N, p_{11}, p_{12}, p_{21}, p_{22})$. Note that in this case none of the margins is fixed in advance; only total sample size, N, is fixed.

Origin II data (one set of margins fixed) :

Origin II data is that we have two binomial populations A and $\bar{A}$. We have a random sample of size n_1 from population A and a random sample of size n_2 from population $\bar{A}$ (row margins are fixed). We then observe the members of B and $\bar{B}$.

If p_1 and p_2 denote proportion of B in populations A and $\bar{A}$ respectively, our population situation would be

	B	$\bar{B}$
A	p_1	$(1-p_1)$
$\bar{A}$	p_2	$(1-p_2)$

Odds of B in population A are defined as $p_1/(1-p_1)$ and that in population $\bar{A}$ are defined as $p_2/(1-p_2)$ so that the odds ratio is defined as

$$\psi = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

The sample outcome may be represented by

	B	$\bar{B}$	
A	x	n_1-x	n_1
$\bar{A}$	y	n_2-y	n_2
	m	N−m	N

Here x is realization of $X \sim Bi(n_1, p_1)$ while y is realization of $Y \sim Bi\ (n_2, p_2)$.

Origin III data (both the sets of margins fixed) :

Origin III is that we have conducted a comparative trial. We had for use N individuals or experimental units. We allocate n_1 of these at random to receive treatment A and remaining n_2 to treatment $\bar{A}$. This determines n_1, n_2 margin. Moreover, if none of the N individuals was assigned treatment A, a given number m (unknown to experimenter) would be fated to be in B category. And if there is no treatment effect this number will not be changed by the experiment. Subject to this condition, therefore other set of margins is also determined. Our sample outcome may be represented by

	B	$\bar{B}$
A	x	n_1-x
$\bar{A}$	m-x	n_2-m+x

The classic example of the two fixed margins case is Fisher's (1935) tea-tasting lady. However, situations in which both margins are unarguably fixed in advance are rare. We consider the first two sampling schemes in the following.

Statistical inference :

There are two major areas of statistical inference - the estimation of parameters and testing of hypothesis.

For 2x2 contingency table, the odds ratio is the parameter of interest. Point and interval estimation for the odds ratio is discussed in sections 2.2 and 2.3 respectivley. Section 2.4 discusses the procedures for testing the hypothesis of independence.

2.2 : Point estimation for the odds ratio

The odds ratio is often used as an index of association in 2x2 contingency tables. There is a considerable literature on 'confidence interval' for the 'odds ratio' in 2x2 tables which we review later. However, it is usually a point estimator that is used as best single statistic to report study results.

We obtain unconditional and conditional maximum likelihood estimators for the odds ratio.

Theorem 2.2.1 : Unconditional maximum likelihood estimator (UMLE) of the population odds ratio is given by the observed cross-product ratio. We denote it by $\hat{\psi}_{un}$.

Proof : Case I

When none of the margins is fixed; $(x_{11}, x_{12}, x_{21}, x_{22})$ is realization of multinomial distribution $(N, p_{11}, p_{12}, p_{21}, p_{22})$. The unconditional likelihood can be written as

$$L = L(x_{11}, x_{12}, x_{21}, x_{22} \mid p_{11}, p_{12}, p_{21}, p_{22})$$

$$= \frac{N!}{x_{11}!\ x_{12}!\ x_{21}!\ x_{22}!}\ p_{11}^{x_{11}}\ p_{12}^{x_{12}}\ p_{21}^{x_{21}}\ p_{22}^{x_{22}} \ .$$

Maximizing L or ln L with respect to p_{ij} (i = 1,2,; j = 1,2) with the restriction that $\sum_i \sum_j p_{ij} = 1$ gives MLE's of p_{ij} (i=1,2;

j=1,2) as x_{ij}/N (i = 1,2; j = 1,2). Hence the UMLE of the odds ratio is given by

$$\hat{\psi}_{Un} = \frac{x_{11}\, x_{22}}{x_{12}\, x_{21}}$$

Note that $\hat{\psi}_{Un}$ is not defined if $x_{12}x_{21} = 0$

<u>Case II</u>

When one set of margins (n_1 and n_2) is fixed; x and y can be considered as realizations of $X \sim Bi(n_1, p_1)$ and $Y \sim Bi(n_2, p_2)$ respectively. The unconditional likelihood can be written as

$$L = L(x, y \mid p_1, p_2)$$
$$= \binom{n_1}{x} p_1^x (1-p_1)^{n_1 - x} \binom{n_2}{y} p_2^y (1-p_2)^{n_2 - y}$$

MLE's of p_i (i = 1,2) are given by x/n and y/n respectively by maximizing L or ln L w.r.t. p_i (i = 1,2). Hence the UMLE of the population odds ratio is given by

$$\hat{\psi}_{Un} = \frac{x(n_2 - y)}{y(n_1 - x)}.$$

Note that $\hat{\psi}_{Un}$ is not defined if $y(n_1 - x) = 0$.

□

Theorem 2.2.2 : If we condition on both the margins of a 2x2 contingency table; the conditional likelihood is given by

$$g(x \mid m, \psi) = \frac{\binom{n_1}{x}\binom{n_2}{m-x}\psi^x}{\sum_{j=r_1}^{r_2}\binom{n_1}{j}\binom{n_2}{m-j}\psi^j} \qquad (2.2.1)$$

where $r_1 = \max(0, m-n_2)$ and $r_2 = \min(m, n_1)$.

Proof : Since we have to condition on both the sets of the margins; we begin with conditioning on one set of margins. When only one set of margins (n_1, n_2) is fixed; the likelihood is given by

$$L = L(x,y|p_1,p_2) = \binom{n_1}{x} p_1^x (1-p_1)^{n_1-x} \binom{n_2}{y} p_2^y (1-p_2)^{n_2-y} .$$

In terms of the odds ratio, the likelihood can be written as

$$L = \binom{n_1}{x} \binom{n_2}{y} (1-p_1)^{n_1} (1-p_2)^{n_2} \left[\frac{p_1(1-p_2)}{p_2(1-p_1)} \right]^x \left[\frac{p_2}{1-p_2} \right]^{x+y}$$

Thus

$$L = L(x, y |\psi, \nu) = u(x,y) v(\psi,\nu) \psi^x \nu^{x+y} \qquad (2.2.2)$$

where $\psi = \dfrac{p_1(1-p_2)}{p_2(1-p_1)}$ and $\nu = \dfrac{p_2}{1-p_2}$.

We observe that (2.2.2) belongs to two parameter exponential family and X and X+Y are sufficient statistic.

Now we obtain the conditional distribution of X given X+Y=m. Since n_1, n_2 and m are fixed; the value of X is limited to lie inclusively between $\max(0, m-n_2)$ and $\min(m,n_1)$. Let $r_1 = \max(0, m-n_2)$ and $r_2 = \min(m,n_1)$.

$$P(X = x \mid X+Y = m) = \frac{P(X=x , Y = m-x)}{P(X+Y = m)} . \qquad (2.2.3)$$

From (2.2.2)

$$P(X = x, Y = m-x) = u(x, m-x) v(\psi, \nu) \psi^x \nu^m$$

and

$$P(X + Y = m) = \sum_{x=r_1}^{r_2} u(x,\ m-x)\ v(\psi,\nu)\ \psi^x\ \nu^m$$

Hence, from (2.2.3)

$$P(X = x \mid X+Y = m) = \frac{\binom{n_1}{x}\binom{n_2}{m-x}\psi^x}{\sum_{j=r_1}^{r_2}\binom{n_1}{j}\binom{n_2}{m-j}\psi^j}$$

$$= K(m,\ \psi)\ \binom{n_1}{x}\ \binom{n_2}{m-x}\ \psi^x$$

where $K(m,\ \psi) = \dfrac{1}{\sum_{j=r_1}^{r_2}\binom{n_1}{j}\binom{n_2}{m-j}\psi^j}$

Thus, the conditional likelihood is given by

$$g(x \mid m,\ \psi) = \frac{\binom{n_1}{x}\binom{n_2}{m-x}\psi^x}{\sum_{j=r_1}^{r_2}\binom{n_1}{j}\binom{n_2}{m-j}\psi^j}$$

Note that the conditional likelihood depends on ψ only.

□

Theorem 2.2.3 : The conditional maximum likelihood estimator (CMLE) of the population odds ratio ψ is obtained by maximizing the conditional likelihood

$$g(x \mid m, \psi) = \frac{\binom{n_1}{x}\binom{n_2}{m-x}\psi^x}{\sum_{j=r_1}^{r_2}\binom{n_1}{j}\binom{n_2}{m-j}\psi^j}$$

w.r.t. ψ. We denote it by $\hat{\psi}_{Cn}$.

Proof : To justify the argument of maximizing $g(x|m,\psi)$ to get CMLE; we discuss the definition of ancillarity in the presence of a nuisance parameter (Godambe 1980).

Let the abstract sample space be $\mathcal{X} = \{x\}$ and the abstract parameter space be $\Omega = \{\theta\}$. The density function w.r.t. some measure μ on $\mathcal{X}$ is $p(x,\theta)$. Further $\theta = (\theta_1, \theta_2)$ where $\theta_1 \in \Omega_1$ being a real interval and $\theta_2 \in \Omega_2$ where Ω_2 is any specified set such that $\Omega = \Omega_1 \times \Omega_2$. Let θ_1 be the parameter of interest and θ_2 be the nuisance parameter.

It is assumed that there exists a statistic T having the following two properties.

(i) The conditional density f_t of x given T = t depends on θ only through θ_1, i.e.

$$p(x, \theta) = f_t(x, \theta_1) \,.\, h(t, \theta)$$

where h is the marginal density of T.

(ii) The class of distributions of T corresponding to $\theta_2 \in \Omega_2$ is complete for each fixed $\theta_1 \in \Omega_1$.

Definition 2.2.1 : Any statistic T satisfying conditions (i) and (ii) above is said to be an ancillary statistic w.r.t. θ_1; the marginal distribution of T is said to contain no information about θ_1 ignoring θ_2.

For 2x2 contingency table, from (2.2.2) the likelihood can be written as

$$L(x, y \mid \psi, \nu) = L = \binom{n_1}{x} \binom{n_2}{y} (1-p_1)^{n_1} (1-p_2)^{n_2} \psi^x \nu^{x+y}$$

Here ψ is the parameter of interest with $\Omega_1 = [0, \infty)$ while ν is the nuisance parameter with $\Omega_2 = [0, \infty)$. We observe that $g(x \mid m, \psi)$, the conditional distribution of X given X+Y = m depends on ψ only. Further the marginal density of X+Y

$$P(X+Y = m) \propto \frac{1}{K(m, \psi)} \nu^m, \quad m = 0, 1, \ldots, N \qquad (2.2.4)$$

Thus, we can write

$$P(X = x, Y = y \mid \psi, \nu) = P(X = x \mid X+Y = m, \psi) . P(X+Y = m \mid \psi, \nu).$$

The class of distributions in (2.2.4) is complete for $\nu \in \Omega_2$ and fixed $\psi \in \Omega_1$ because for fixed $\psi \in \Omega_1$, $K(m, \psi) > 0 \;\forall\; m$. Thus the definition 2.2.1 is applicable and we conclude that the statistic X+Y is ancillary w.r.t. ψ and the marginal density of X+Y is said to contain no information about ψ ignoring ν. Hence inference on ψ can be based on the conditional distribution of X given X+Y. The CMLE is then obtained by maximizing $g(x \mid m, \psi)$.

□

Example 2.2.1

Consider all possible 2x2 tables with n_1 = 15, n_2 = 10 and m = 9. We present below UMLE and CMLE of the population odds ratio.

Table 2.2.1 : UMLE and CMLE for the population odds ratio

Sr. No.	Table		$\hat{\psi}_{Un}$	$\hat{\psi}_{Cn}$
1	0	15	0.0000	0
	9	1		
2	1	14	0.0178	0.0237
	8	2		
3	2	13	0.0659	0.0764
	7	3		
4	3	12	0.1667	0.1812
	6	4		
5	4	11	0.3636	0.3796
	5	5		
6	5	10	0.7500	0.7588
	4	6		
7	6	9	1.5555	1.5285
	3	7		
8	7	8	3.5000	3.3301
	2	8		
9	8	7	10.2857	9.3845
	1	9		
10	9	6	unbounded	unbounded
	0	10		

Discussion :

While evaluating a point estimate of the population odds ratio, problem of zero cell arises. In general if we denote the observed table by

a	b	n_1
c	d	n_2
m	N-m	N

then UMLE of ψ is given by $\hat{\psi}_{Un}$ = ad/bc while CMLE of ψ is obtained by equating one of the observed cell frequency (say the "a" cell) to its expectation under the noncentral hypergeometric distribution with noncentrality parameter ψ. Thus $\hat{\psi}_{Cn}$ is obtained by solving

$$a = E(a \mid n_1, n_2, a+c, b+d, \hat{\psi}_{Cn}). \qquad (2.2.5)$$

Note that $\hat{\psi}_{Un}$ is undefined if bc = 0 and $\hat{\psi}_{Cn}$ is undefined if bc=0, a+c = 0 or b+d = 0.

Caution is obviously required when reporting an actual data set with a zero cell (Mantel 1986, Hauck 1986), although many examples of such report exist in literature.

In evaluating estimators of the odds ratio, one faces a dilemma. If zero-celled tables are excluded from the admissible sample space, estimators are always defined, but the evaluation is over only part of the sample space. It is suggested (Mantel 1992) that when a table with zero cell occurs, instead of point estimate one should look for interval estimate. On the other hand if zero-celled tables are included in the admissible sample space, some estimators are not defined and must be modified if they are to be evaluated. We discuss below some modifications for estimation of ψ or equivalently $\beta = \ln \psi$.

Note that UMLE of β is given by $\hat{\beta}_{Un} = \ln(\hat{\psi}_{Un})$ and CMLE of β is given by $\hat{\beta}_{Cn} = \ln(\hat{\psi}_{Cn})$. $\hat{\beta}_{Un}$ and $\hat{\beta}_{Cn}$ are undefined if abcd =0. To allow estimation of $\hat{\beta}_{Un}$ under the condition that abcd = 0; Haldane (1955) suggested adding a correction term $\varepsilon = 1/2$ to all the four cells, to modify the estimator proposed earlier by Woolf (1955). Thus, Haldane's modified UMLE of β is given by

$$\hat{\beta}_{Un} = \ln((a + 1/2)(d + 1/2)/(b + 1/2)(c+1/2)) .$$

Adding a positive constant only to zero counts was proposed by Grizzle, Starmer, Koch (1969), but Cox (1970), Goodman (1970) and Walter and Cook (1991) recommended always adding a constant. More generally, Gart, Pettigrew and Thomas (1985) have shown that the optimal ε correction for 2xk table depends on k. Adding ε = 1/2 in all tables gives better bias and mse properties for $\hat{\beta}$ than if it is added only as necessary when a zero cell occurs (Walter 1985). Although, not quite as clear cut the same preference also applies to unconditional estimation of ψ (Walter 1987).

Jewell (1986) proposed the following estimator for the estimation of the population odds ratio

$$\hat{\psi}_J = \frac{(ad)}{(b+1)(c+1)}$$

Its correction of $\varepsilon = 1$ to b and c cells is intended to reduce positive bias of UMLE and also make it defined for all possible tables. However, the corresponding estimator of β is still undefined if ad = 0. Walter and Cook (1991) have suggested modified Jewell type β estimator

$$\hat{\beta}_J = \ln((a + 1/2)(d + 1/2) / (b+1)(c+1)) .$$

Jewell (1986) has shown that $\hat{\psi}_J$ performs well on the basis of bias and mse.

Modifications suggested for conditional m.l.e. are as follows (Walter and Cook 1991).

First, if abcd ≠ 0, $\hat{\psi}_{Cn}$ is defined directly by (2.2.5). Second, if a = 0 or d = 0 only; $\hat{\psi}_{Cn}$ is calculated using the solution to

$$a + 1/2 = E(a \mid n_1, n_2, a+c, b+d, \hat{\psi}_{Cn})$$

Similarly, if b = 0 or c = 0; $\hat{\psi}_{Cn}$ is defined as solution to

$$a - 1/2 = E(a \mid n_1, n_2, a+c, b+d, \hat{\psi}_{Cn}) .$$

Finally, if a+c = 0 or b+d = 0 $\hat{\psi}_{Cn}$ is defined to be equal to 1.

Walter and Cook (1991) studied modified UMLE, CMLE and Jewell's estimator for ψ and ln ψ. They recommended that $\beta = \ln\psi$ to be estimated rather than ψ be estimated and modified UMLE should be used for this purpose. There is a considerable instability and occasionally large bias associated with most estimators on the arithmatic scale, particularly for small N. Jewell's estimator $\hat{\psi}_J$ has relatively good bias and mse properties, but it is not invariant under table orientation. Therefore if symmetry w.r.t. table orientation is an important criterion, then modified UMLE should be used. (Walter and Cook, 1991).

Note that the various modifications of UMLE of the odds ratio shows the prior belief of the experimenter. Consider that only table total is fixed at N, and x_{11}, x_{12}, x_{21}, x_{22} are realizations of a multinomial (N, p_{11}, p_{12}, p_{21}, p_{22}). Now if (p_{11}, p_{12}, p_{21}, p_{22}) follow Dirichlet distribution with parameters (1/2, 1/2, 1/2, 1/2) then posterior distribution of (p_{11}, p_{12}, p_{21}, p_{22}) given (x_{11}, x_{12}, x_{21}, x_{22}) is also Dirichlet with parameters (x_{11} + 1/2, x_{12} + 1/2, x_{21} + 1/2, x_{22} + 1/2) so that $\hat{p}_{ij} = \frac{x_{ij} + 1/2}{N+2}$ i = 1,2 ; j = 1,2 and hence

$$\hat{\psi}_{un} = \frac{(x_{11} + 1/2)\,(x_{22} + 1/2)}{(x_{12} + 1/2)\,(x_{21} + 1/2)} .$$

If the prior belief is changed to Dirichlet distribution with parameters (1/2, 1, 1, 1/2) then the odds ratio estimate is given by

$$\hat{\psi}_{un} = \frac{(x_{11} + 1/2)\,(x_{22} + 1/2)}{(x_{12} + 1)\quad(x_{21} + 1)}\,.$$

Example 2.2.2 :

In the following, we study bias of various estimators discussed earlier for tables with

(a) $n_1 = 15$, $n_2 = 10$, $m = 9$ and

(b) $n_1 = 17$, $n_2 = 18$, $m = 15$.

(a)

Table 2.2.2 : Table of estimates

a c	b d	$\hat{\psi}_1$ Unconditional estimate	$\hat{\psi}_2$ (modified unconditional) adding 1/2 to zero count only	$\hat{\psi}_3$ (modified unconditional) adding 1/2 to all the cells	$\hat{\psi}_4$ Jewell's estimate	$\hat{\psi}_5$ Conditional estimate	$\hat{\psi}_6$ Modified conditional
0	15	0.0000	0.0037	0.0051	0.0000	0.0000	0.0093
9	1						
1	14	0.0178	0.0178	0.0304	0.0148	0.0237	0.0237
8	2						
2	13	0.0659	0.0659	0.0864	0.0537	0.0764	0.0764
7	3						
3	12	0.1667	0.1667	0.1938	0.1319	0.1812	0.1812
6	4						
4	11	0.3636	0.3636	0.3913	0.2778	0.3796	0.3796
5	5						
5	10	0.7500	0.7500	0.7566	0.5454	0.7588	0.7588
4	6						
6	9	1.5555	1.5555	1.4662	1.0500	1.5285	1.5285
3	7						
7	8	3.5000	3.5000	3.0000	2.0741	3.3301	3.3301
2	8						
8	7	10.2857	10.2857	7.1778	4.5000	9.3845	9.3845
1	9						
9	6	∞	0.0333	30.6923	12.8571	∞	22.0485
0	10						

Table 2.2.3 : Table of expectations

	$E(\hat{\psi}_1)$	$E(\hat{\psi}_2)$	$E(\hat{\psi}_3)$	$E(\hat{\psi}_4)$	$E(\hat{\psi}_5)$	$E(\hat{\psi}_6)$
ψ=0.2	0.2679	0.2667	0.2865	0.1999	0.2797	0.2785
ψ=0.7	1.0635	1.0628	1.0029	0.6995	1.0435	1.0565
ψ=1.0	1.5822	1.5784	1.4865	0.9975	1.5278	1.5780
ψ=2.0	3.1755	3.1090	3.1561	1.9576	2.9906	3.3941
ψ=4.0	5.2892	4.7774	6.5982	3.6104	4.9055	6.5748

(b)

Table 2.2.4 : Table of estimates

a c	b d	$\hat{\psi}_1$ Unconditional estimate	$\hat{\psi}_2$ (modified unconditional) adding 1/2 to zero count only	$\hat{\psi}_3$ (modified unconditional) adding 1/2 to all the cells	$\hat{\psi}_4$ Jewell's estimate	$\hat{\psi}_5$ Conditional estimate	$\hat{\psi}_6$ Modified conditional
0	17	0.0000	0.0059	0.0064	0.0000	0.0000	0.0091
15	3						
1	16	0.0178	0.0178	0.0282	0.0157	0.0212	0.0212
14	4						
2	15	0.0513	0.0513	0.0657	0.0446	0.0573	0.0573
13	5						
3	14	0.1071	0.1071	0.1255	0.0923	0.1156	0.1156
12	6						

4	13	0.1958	0.1958	0.2174	0.1667	0.2061	0.2061
11	7						
5	12	0.3333	0.3333	0.3562	0.2797	0.3444	0.3444
10	8						
6	11	0.5454	0.5454	0.5652	0.4500	0.5551	0.5551
9	9						
7	10	0.8750	0.8750	0.8823	0.7071	0.8783	0.8783
8	10						
8	9	1.3968	1.3968	1.3719	1.1000	1.3835	1.3835
7	11						
9	8	2.2500	2.2500	2.1493	1.7143	2.1971	2.1971
6	12						
10	7	3.7143	3.7143	3.4364	2.7083	3.5681	3.5681
5	13						
11	6	6.4167	6.4167	5.7008	4.4000	6.0386	6.0386
4	14						
12	5	12.0000	12.0000	10.0649	7.5000	10.9747	10.9747
3	15						
13	4	26.0000	26.0000	19.8000	13.8667	22.7302	22.7302
2	16						
14	3	79.3333	79.3333	48.3333	29.7500	63.7444	63.7444
1	17						
15	2	∞	270.0000	229.4000	90.0000	∞	151.6293
0	18						

Table 2.2.5 : Table of expectations

	$E(\hat{\psi}_1)$	$E(\hat{\psi}_2)$	$E(\hat{\psi}_3)$	$E(\hat{\psi}_4)$	$E(\hat{\psi}_5)$	$E(\hat{\psi}_6)$
ψ=0.2	0.2393	0.2388	0.2579	0.1999	0.2483	0.2472
ψ=0.7	0.8841	0.8841	0.8787	0.6999	0.8806	0.8806
ψ=1.0	1.2942	1.2957	1.2574	1.0000	1.2736	1.2751
ψ=2.0	2.7825	2.7839	2.5576	1.9999	2.7825	2.6641
ψ=4.0	6.2234	6.2844	5.3372	3.9990	5.7322	5.7659

2.3 : Interval estimation for the odds ratio

2.3.1 Confidence limits based on Fisher's "exact" distribution :

Here we consider the confidence limits for the odds ratio based on Fisher's exact distribution. As shown earlier, the conditional probability of the observations for the subset of samples in which all the marginal totals are fixed by the condition

$$x + y = m$$

is given by

$$g(x|m, \psi) = \frac{\binom{n_1}{x}\binom{n_2}{m-x}\psi^x}{\sum_{j=0}^{m}\binom{n_1}{j}\binom{n_2}{m-j}\psi^j} \qquad (2.3.1)$$

Here and henceforward we assume that $r_1 = \max(0, m-n_2) = 0$ and $r_2 = \max(n_1, m) = m$. The conditional probability of the sample observation given the marginal totals, thus depends only on ψ, the unknown parameter.

An exact 100(1-α)% confidence interval for ψ is given by $(\psi_L(x),\ \psi_U(x))$, where $\psi_L(x)$ is such that

$$\sum_{j=x}^{m} g(j\,|\,m,\ \psi_L(x)) = \alpha/2 \qquad (2.3.2)$$

and $\psi_U(x)$ is such that

$$\sum_{j=0}^{x} g(j\,|\,m,\ \psi_U(x)) = \alpha/2 \qquad (2.3.3)$$

The probability that this interval fails to contain ψ is $P(\psi_U(X)<\psi) + P(\psi_L(X) > \psi)$. We now prove that each of the above exclusion probability can not exceed $\alpha/2$.

Let

$$F(x\,|\,\psi) = \sum_{j=0}^{x} g(j\,|\,m,\ \psi) \qquad (2.3.4)$$

be the c.d.f. of X.

We now prove the following lemma.

Lemma 2.3.1 : For any fixed x, $0 \le x < m$; $F(x|\psi)$ is a decreasing function of ψ.

Proof :

$$F(x\,|\,\psi) = \sum_{j=0}^{x} g(j\,|\,m,\ \psi)$$

$$= \frac{\sum_{j=0}^{x} \binom{n_1}{j}\binom{n_2}{m-j}\psi^j}{\sum_{j=0}^{m} \binom{n_1}{j}\binom{n_2}{m-j}\psi^j}$$

$$= \left[1 + \frac{\sum_{j=x+1}^{m} \binom{n_1}{j} \binom{n_2}{m-j} \psi^{j-x}}{\sum_{j=0}^{x} \binom{n_1}{j} \binom{n_2}{m-j} \psi^{j-x}} \right]^{-1}$$

$$= \left[1 + \frac{P(x|\psi)}{Q(x|\psi)} \right]^{-1}$$

where $P(x|\psi) = \sum_{j=x+1}^{m} \binom{n_1}{j} \binom{n_2}{m-j} \psi^{j-x}$

and $Q(x|\psi) = \sum_{j=0}^{m} \binom{n_1}{j} \binom{n_2}{m-j} \psi^{j-x}$

Since $\psi \geq 0$; $P(x|\psi)$ is an increasing function of ψ and $Q(x|\psi)$ is a decreasing function of ψ. Thus, $F(x|\psi)$ is a decreasing function of ψ.

□

Now,

$$\begin{aligned} P(\psi_U(X) < \psi) &= P[F(X|\psi) < F(X \mid \psi_U(X))] \\ &= P[F(X|\psi) < \alpha/2] \\ &\leq P[F(X|\psi) \leq \alpha/2] \\ &\leq \alpha/2 . \end{aligned}$$

By a symmetric argument, $P(\psi_L(x) > \psi) \leq \alpha/2$. This proves that $(\psi_L(x), \psi_U(x))$ is a conservative $100(1-\alpha)\%$ confidence interval. Due to discreteness of the distribution of X, there are generally no values of ψ_L and ψ_U which satisfy the equations (2.3.2) and (2.3.3) exactly. The value of ψ_L is taken to be the largest such that the expression in (2.3.2) is $\leq \alpha/2$, whereas the value of $\psi_U > \psi_L$ is taken to be the smallest such that the expression in (2.3.3) is $\leq \alpha/2$. When any entry in the table is zero, only one sided intervals may be found.

The term "exact confidence limits" is often applied to this procedure. This is somewhat confusing because "exact" does not refer to the confidence coefficient being exactly $(1-\alpha)$, but rather to the fact that the limits, which only approximate a $(1-\alpha)$ confidence interval, are based on the exact conditional distribution of X.

Example 2.3.1

Let $n_1 = 15$, $n_2 = 10$ and $m = 9$. In the following we give 95% exact limits for the odds ratio in each table with above marginal totals.

Table 2.3.1 : Limits based on exact distribution

Sr. No.	a c	b d	95% Exact Limits Lower limit ψ_L	 Upper Limit ψ_U
1	0 9	15 1	0.0000	0.2250
2	1 8	14 2	0.00038	0.2968
3	2 7	13 3	0.00513	0.6528
4	3 6	12 4	0.01900	1.3218
5	4 5	11 5	0.04900	2.6371
6	5 4	10 6	0.1077	5.4812
7	6 3	9 7	0.2191	12.9482
8	7 2	8 8	0.43653	42.6595
9	8 1	7 9	0.8947	504.9055
10	9 0	6 10	2.8585	∞

Gart's approximation to exact confidence limits :

For small numbers, Gart (1962) has suggested approximation to exact confidence limits. The suggested approximations have been found to yield good results whenever $\frac{x(m-x)}{n_1+n_2} \le 1$, although for values of this quantity close to unity, the interval tend to be too narrow.

Consider $\sum_{i=0}^{x} P(i)$ where

$$P(i) = \frac{\binom{n_1}{i}\binom{n_2}{m-i}}{\binom{n_1+n_2}{m}}$$

Wise (1954) has given an approximation to above sum by

$$\sum_{i=0}^{x} P(i) \simeq \sum_{i=0}^{x} \binom{m}{i} c^i (1-c)^{m-i}$$

where

$$c = \frac{2n_1 - k + 1}{2(n_1+n_2) - m+1} \qquad (2.3.5)$$

k being number of terms in the sum in question, in this case (x+1).

We have the exact confidence limits given by

$$\frac{\sum_{i=0}^{x} \binom{n_1}{i}\binom{n_2}{m-i}\psi_U^i}{\sum_{i=0}^{m} \binom{n_1}{i}\binom{n_2}{m-i}\psi_U^i} = \frac{\alpha}{2} \qquad (2.3.6)$$

and

$$\frac{\sum_{j=x}^{m} \binom{n_1}{i} \binom{n_2}{m-i} \psi_L^i}{\sum_{i=0}^{m} \binom{n_1}{i}\binom{n_2}{m-i} \psi_L^i} = \frac{\alpha}{2} \qquad (2.3.7)$$

Since these equations have a certain symmetry, we shall explicitly consider only (2.3.8). Dividing the numerator and denominator of (2.3.8) by $\binom{n_1+n_2}{m}$ we find that coefficients of ψ_U^i are the terms of hypergeometric distribution. Applying the approximation mentioned earlier we obtain

$$\frac{\sum_{i=0}^{x} \binom{m}{i} (c_2\psi_U)^i (1-c_2)^{m-i}}{\sum_{i=0}^{m} \binom{m}{i} (c_2\psi_U)^i (1-c_2)^{m-i}} \simeq \frac{\alpha}{2}$$

where we have c_2 for the moment unspecified. Summation of the denominator yields

$$\sum_{i=0}^{x} \binom{m}{i} \left[\frac{c_2\psi_U}{(1-c_2)+c_2\psi_U} \right]^i \left[\frac{1 - c_2}{(1-c_2)+c_2\psi_U} \right]^{m-i} \simeq \frac{\alpha}{2}$$

$$\Leftrightarrow \sum_{i=x+1}^{m} \binom{m}{i} (c')^i (1-c')^{m-i} \simeq 1 - \frac{\alpha}{2} \qquad (2.3.8)$$

where $c' = \dfrac{c_2\psi_U}{(1-c_2)+c_2\psi_U}$.

By using the relation between binomial sum and incomplete beta, we have

$$I_{c'}(x+1,\ m-x) \simeq 1 - \frac{\alpha}{2}\ .$$

Since the lower $100\nu\%$ point of F distribution with ν_1, ν_2 degrees of freedom (denoted by $F_{1-\nu}(\nu_1,\nu_2)$) satisfies the equation

$$I_{\frac{\nu_1 F}{\nu_2+\nu_1 F}}(\nu_1/2,\ \nu_2/2) = \nu$$

We can write (2.3.8) as

$$I_{\frac{\nu_1 F}{\nu_2+\nu_1 F}}(\nu_1/2,\ \nu_2/2) \simeq 1 - \alpha/2$$

where $\nu_1 = 2(x+1)$, $\nu_2 = 2(m-x)$ and F is lower $100(1-\frac{\alpha}{2})\%$ point of F distribution with ν_1 and ν_2 d.f.

Now $$c' = \frac{\nu_1 F}{\nu_2+\nu_1 F}$$

$$\Leftrightarrow \quad \frac{c_2\psi_U}{(1-c_2)+c_2\psi_U} = \frac{\nu_1 F}{\nu_2+\nu_1 F}$$

$$\Leftrightarrow \quad \psi_U = \frac{\nu_1}{\nu_2} \times \frac{1-c_2}{c_2} F_{\alpha/2}(\nu_1,\ \nu_2)$$

$$= \frac{x+1}{m-x} \times \frac{1-c_2}{c_2} F_{\alpha/2}(2x+2,\ 2m-2x)\ .$$

A similar argument beginning with (2.3.7) leads to

$$\psi_L = \frac{x}{(m-x+1)} \times \frac{1-c_1}{c_1} \times \frac{1}{F_{\alpha/2}[2(m-x+1),\ 2x]}\ .$$

It would seem reasonable here to use Wise's formula (2.3.5) in evaluating c_1 and c_2, however this leads to difficulties. Since (2.3.5) involves number of terms in the sum and this number of terms is different in the numerator and denominator of (2.3.6) and (2.3.7), its direct application would unduly complicate the final result. It was found that (Gart 1962) for the upper limit letting

$$k_2 = \frac{(x+1)^2 + (m+1)^2}{x + m + 2} \qquad (2.3.9)$$

in (2.3.5) yielded the most accurate limits over a large number of cases. Here of course, k_2 is the weighted mean of number of terms in the numerator and denominator, the weights being proportional to number of terms.

Analogously, for the lower limit, we let

$$k_1 = \frac{(m-x+1)^2 + (m+1)^2}{2m - x + 2} \qquad (2.3.10)$$

in evaluating c_1 from (2.3.5).

Thus we have the following approximate $(1-\alpha)100\%$ confidence limits for the odds ratio

$$\psi_L = \frac{x(2n_2 - k_1 + m)}{(2n_1 - k_1 + 1)(m-x+1)} \times \frac{1}{F_{\alpha/2}(2m-2x+2,\ 2x)} \qquad (2.3.11)$$

and

$$\psi_U = \frac{(2n_2 - m + k_2)(x+1)}{(2n_1 - k_2 + 1)(m-x)} \times F_{\alpha/2}(2x+2,\ 2m-2x) \qquad (2.3.12)$$

Example 2.3.2

Here we calculate the limits for the odds ratio using Gart's procedure for 2x2 tables considered in example 2.3.1.

Table 2.3.2 : Limits based on Gart's approximation

Sr. No.	a c	b d	95% limits Lower limit ψ_L	 Upper Limit ψ_U
1	0 9	15 1	0.0000	0.4342
2	1 8	14 2	0.0025	0.8211
3	2 7	13 3	0.0263	1.2857
4	3 6	12 4	0.0735	2.9830
5	4 5	11 5	0.1446	3.1729
6	5 4	10 6	0.2451	5.4600
7	6 3	9 7	0.3897	10.9994
8	7 2	8 8	0.6077	31.7531
9	8 1	7 9	1.4565	339.3172
10	9 0	6 10	2.3028	∞

2.3.2 Limits based on approximate distribution :

Finding confidence limits based on exact conditional distribution is a process which Fisher (1962) described as too lengthy to be recommended. An attempt to avoid such lengthy calculations for moderately large sample sizes leads to the investigation of asymptotic approximations given by Cornfield (1956) and Fisher (1962). Though the formulae given by Fisher and Cornfield were somewhat different, essentially both the methods lead to the same limits. We discuss both the methods.

Approximate limits proposed by Fisher :

Let the observed table be represented as

x	n_1-x	n_1
y	n_2-y	n_2
m	$N-m$	N

First determine the value of z which provides solution to the equation

$$z^2 \left\{ \frac{1}{x-z} + \frac{1}{n_1 - x + z} + \frac{1}{y+z} + \frac{1}{n_2-y-z} \right\} = \chi^2_\alpha \qquad (2.3.13)$$

where χ^2_α is upper 100α percentile of chisquare distribution. Then assuming that equation (2.3.13) has real solutions z_1 and z_2 $(z_1 > z_2)$ take

$$\psi_L = \frac{(x-z_1)\,(n_2-y-z_1)}{(n_1-x+z_1)\,(y+z_1)} \qquad (2.3.14)$$

and

$$\psi_U = \frac{(x-z_2)\ (n_2-y-z_2)}{(n_1-x+z_2)\ (y+z_2)} \qquad (2.3.15)$$

as two sided approximate confidence limits for the true odds ratio. These limits are at approximate probability level α.

Fisher gave the following justification for the use of the equation (2.3.13).

If the expectation in the four cells of 2x2 table were

$x-z$	n_1-x+z
$y+z$	n_2-y-z

their cross product ratio will be

$$\frac{(x-z)\ (n_2-y-z)}{(y+z)\ (n_1-x+z)}$$

and the chisquare statistic for the observations would be

$$X^2 = z^2 \left\{ \frac{1}{x-z} + \frac{1}{n_1-x+z} + \frac{1}{y+z} + \frac{1}{n_2-y-z} \right\} \qquad (2.3.16)$$

with one degree of freedom.

Applying Yates' correction for continuity Fisher used

$$x_c^2 = (\ |z| - \frac{1}{2}\ )^2 \left\{ \frac{1}{x-z} + \frac{1}{n_1-x+z} + \frac{1}{y+z} + \frac{1}{n_2-y-z} \right\}$$

is his numerical calculations.

Approximate limits proposed by Cornfield (1956) :

The mathematical derivation suggested by Cornfield (1956) for obtaining approximate confidence limits for the odds ratio is as follows.

In obtaining the limiting distribution for (2.3.1), which we now denote by $f(x)$; we are faced with an initial difficulty arising from inability to evaluate the constant term of the denominator. This difficulty is avoided by seeking instead the limiting distribution of the ratio

$$\frac{f(x)}{f(\tilde{x})}$$

where $\tilde{x}$ is the mode of the distribution (2.3.1) and is defined by the inequality

$$\frac{(\tilde{x}+1)(n_2-m+\tilde{x}+1)}{(n_1-\tilde{x})(m-\tilde{x})} \geq \psi \geq \frac{\tilde{x}(n_2 - m + \tilde{x})}{(n_1-\tilde{x}+1)\ (m-\tilde{x}+1)} \qquad (2.3.17)$$

For large samples, it is sufficient to write

$$\psi = \frac{\tilde{x}(n_2 - m + \tilde{x})}{(n_1-\tilde{x})\ (m-\tilde{x})} \qquad (2.3.18)$$

Before obtaining the limiting distribution $f(x)$, let us prove the following lemma.

Lemma 2.3.2 : If we use Stirling's formula for factorial n and expand all the terms of the form $\log(1+x)$ to the quadratic and set terms of the form

$$\frac{\tilde{x}+1/2}{\tilde{x}}, \quad \frac{n_1-\tilde{x}+1/2}{n_1-\tilde{x}}, \quad \frac{m-\tilde{x}+1/2}{m-\tilde{x}}, \quad \text{and} \quad \frac{n_2-m+\tilde{x}+1/2}{n_2-m+\tilde{x}}$$

equal to unity, we have as a limiting expression

$$-2\log\left\{\frac{f(x)}{f(\tilde{x})}\right\} = (x - \tilde{x})^2 \left\{\frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}}\right\}$$

Proof :

$$- 2\log\left\{ \frac{f(x)}{f(\tilde{x})} \right\} = - 2\log \left\{ \frac{\binom{n_1}{x} \binom{n_2}{m-x} \psi^{x}}{\binom{n_1}{\tilde{x}} \binom{n_2}{m-\tilde{x}} \psi^{\tilde{x}}} \right\}$$

$$= - 2\log\left\{ \left[\frac{\tilde{x}!\ (n_1-\tilde{x})!\ (n_2-m+\tilde{x})!\ (m-\tilde{x})!}{x!\ (n_1-x)!\ (n_2-m+x)!\ (m-x)!} \right]\left[\frac{\tilde{x}(n_2-m+\tilde{x})}{(n_1-\tilde{x})(m-\tilde{x})} \right]^{x-\tilde{x}} \right\} .$$

Now using Stirling's approximation $n! = \sqrt{2\pi}\ n^{n+1/2}\ e^{-n}$ we have

$$- 2\log\left\{ \frac{f(x)}{f(\tilde{x})} \right\}$$

$$= - 2\log\left\{ \frac{\tilde{x}^{\tilde{x}+ \frac{1}{2}} (n_1-\tilde{x})^{n_1-\tilde{x}+ \frac{1}{2}} (m-\tilde{x})^{m-\tilde{x}+ \frac{1}{2}} (n_2-m+\tilde{x})^{n_2-m+\tilde{x}+ \frac{1}{2}}}{x^{x+ \frac{1}{2}} (n_1-x)^{n_1-x+ \frac{1}{2}} (m-x)^{m-x+ \frac{1}{2}} (n_2-m+x)^{n_2-m+x+ \frac{1}{2}}} \right.$$

$$\left. \left[\frac{\tilde{x}(n_2-m+\tilde{x})}{(n_1-\tilde{x})(m-\tilde{x})} \right]^{x-\tilde{x}} \right\}$$

$$= 2\log \left\{\left(\frac{x}{\tilde{x}} \right)^{x+ \frac{1}{2}}\left(\frac{n_1-x}{n_1-\tilde{x}} \right)^{n_1-x + \frac{1}{2}}\left(\frac{m-x}{m-\tilde{x}} \right)^{m-x+ \frac{1}{2}}\left(\frac{n_2-m+x}{n_2-m+\tilde{x}}\right)^{n_2-m+x+ \frac{1}{2}}\right\}$$

$$= 2(x + \tfrac{1}{2}) \log\left\{1 + \frac{x-\tilde{x}}{\tilde{x}} \right\} + 2(n_1-x + \tfrac{1}{2}) \log \left\{1 + \frac{\tilde{x}-x}{n_1-\tilde{x}} \right\}$$

$$+ 2(m - x + \tfrac{1}{2}) \log \left\{1 + \frac{\tilde{x}-x}{m -\tilde{x}} \right\}$$

$$+ 2(n_2-m+x + 1/2) \log\left\{1 + \frac{x - \tilde{x}}{n_2-m+\tilde{x}} \right\}$$

$$= 2(x + \tfrac{1}{2}) \left\{ \frac{x-\tilde{x}}{\tilde{x}} - \frac{1}{2} \frac{(x-\tilde{x})^2}{\tilde{x}^2} \right\}$$

$$+ \ 2(n_1 - x + 1/2) \left\{ \frac{\tilde{x}-x}{n_1-\tilde{x}} - \frac{1}{2} \frac{(\tilde{x}-x)^2}{(n_1-\tilde{x})^2} \right\}$$

$$+ \ 2(m-x + 1/2) \left\{ \frac{\tilde{x}-x}{m-\tilde{x}} - \frac{1}{2} \frac{(\tilde{x}-x)^2}{(m-\tilde{x})^2} \right\}$$

$$+ \ 2(n_2-m+x + \tfrac{1}{2}) \left\{ \frac{x-\tilde{x}}{n_2-m+\tilde{x}} - \frac{1}{2} \frac{(x-\tilde{x})^2}{(n_2-m+\tilde{x})^2} \right\}$$

$$= \frac{2(x+1/2)}{\tilde{x}} \left\{ (x - \tilde{x}) - \frac{(x-\tilde{x})^2}{2\tilde{x}} \right\}$$

$$+ \ \frac{2(n_1-x+1/2)}{(n_1 - \tilde{x})} \left\{ (\tilde{x} - x) - \frac{(\tilde{x} - x)^2}{2(n_1-\tilde{x})} \right\}$$

$$+ \ \frac{2(m-x+1/2)}{(m-\tilde{x})} \left\{ (\tilde{x} - x) - \frac{(\tilde{x} - x)^2}{2(m - \tilde{x})} \right\}$$

$$+ \ \frac{2(n_2-m+x+1/2)}{(n_2-m+\tilde{x})} \left\{ (x - \tilde{x}) - \frac{(x - \tilde{x})^2}{2(n_2 - m + \tilde{x})} \right\}$$

$$= \frac{2(\tilde{x} + 1/2 + x-\tilde{x})}{\tilde{x}} \left\{ (x - \tilde{x}) - \frac{1}{2} \frac{(x - \tilde{x})^2}{\tilde{x}} \right\}$$

$$+ \ \frac{2(n_1 - \tilde{x} + 1/2 + \tilde{x}-x)}{(n_1 - \tilde{x})} \left\{ (\tilde{x} - x) - \frac{1}{2} \frac{(\tilde{x} - x)^2}{n_1 - \tilde{x}} \right\}$$

$$+ \frac{2(m - \tilde{x} + 1/2 + \tilde{x} - x)}{(m - \tilde{x})} \left\{ (\tilde{x} - x) - \frac{1}{2} \frac{(\tilde{x} - x)^2}{(m - \tilde{x})} \right\}$$

$$+ \frac{2(n_2 - m + \tilde{x} + 1/2 + x - \tilde{x})}{(n_2 - m + \tilde{x})} \left\{ (x - \tilde{x}) - \frac{1}{2} \frac{(\tilde{x} - x)^2}{(n_2 - m + \tilde{x})} \right\}$$

$$= 2\left\{ 1 + \frac{x - \tilde{x}}{\tilde{x}} \right\} \left\{ (x - \tilde{x}) - \frac{1}{2} \frac{(x - \tilde{x})^2}{\tilde{x}} \right\}$$

$$+ 2\left\{ 1 + \frac{\tilde{x} - x}{n_1 - \tilde{x}} \right\} \left\{ (\tilde{x} - x) - \frac{(\tilde{x} - x)^2}{2(n_1 - \tilde{x})} \right\}$$

$$+ 2\left\{ 1 + \frac{\tilde{x} - x}{m - \tilde{x}} \right\} \left\{ (\tilde{x} - x) - \frac{(\tilde{x} - x)^2}{2(m - \tilde{x})} \right\}$$

$$+ 2\left\{ 1 + \frac{x - \tilde{x}}{n_2 - m + \tilde{x}} \right\} \left\{ (x - \tilde{x}) - \frac{(x - \tilde{x})^2}{2(n_2 - m + \tilde{x})} \right\}$$

$$= 2\left\{ (x - \tilde{x}) + \frac{1}{2} \frac{(x - \tilde{x})^2}{\tilde{x}} - \frac{1}{2} \frac{(x - \tilde{x})^3}{\tilde{x}^2} \right\}$$

$$+ 2\left\{ (\tilde{x} - x) + \frac{1}{2} \frac{(\tilde{x} - x)^2}{(n_1 - \tilde{x})} - \frac{1}{2} \frac{(\tilde{x} - x)^3}{(n_1 - \tilde{x})^2} \right\}$$

$$+ 2\left\{ (\tilde{x} - x) + \frac{1}{2} \frac{(\tilde{x} - x)^2}{(m - \tilde{x})} - \frac{1}{2} \frac{(\tilde{x} - x)^3}{(m - \tilde{x})^2} \right\}$$

$$+ 2\left\{ (x - \tilde{x}) + \frac{1}{2} \frac{(x - \tilde{x})^2}{(n_2 - m + \tilde{x})} - \frac{1}{2} \frac{(x - \tilde{x})^3}{(n_2 - m + \tilde{x})^2} \right\}$$

$$\simeq (X - \tilde{X})^2 \left\{ \frac{1}{\tilde{x}} + \frac{1}{n_1 - \tilde{x}} + \frac{1}{m - \tilde{x}} + \frac{1}{n_2 - m + \tilde{x}} \right\}$$

□

We now obtain the limiting distribution f(x).

Theorem 2.3.1 Limiting distribution for (2.3.1) is normal with mean $\tilde{x}$ defined by (2.3.18 and variance given by

$$\left\{ \frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}} \right\}^{-1}.$$

Proof :

To obtain the value of maximum ordinate, we use (2.3.1) to write

$$\frac{1}{f(\tilde{x})} = \sum_{x=0}^{m} f(x)/f(\tilde{x})$$

Approximating the summation with an integration from $-\infty$ to ∞; we have

$$\frac{1}{f(\tilde{x})} = \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2} (x-\tilde{x})^2 \left\{ \frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}} \right\}\right\} dx$$

$$\therefore \frac{1}{f(\tilde{x})} = \sqrt{2\pi} \times \frac{1}{\sqrt{\frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}}}}$$

so that

$$f(\tilde{x}) = \frac{1}{\sqrt{2\pi}} \left\{ \frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}} \right\}^{1/2}$$

Hence the limiting distribution f(x) is normal with mean $\tilde{x}$ and variance

$$\left[\frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}} \right]^{-1}$$

Hence the theorem.

□

Now denote by $\tilde{x}_2$ the largest root of quartic in $\tilde{x}$.

$$(\tilde{x} - x - \tfrac{1}{2})^2 \left\{ \frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}} \right\} = \chi^2_\alpha$$

and by $\tilde{x}_1$ the smallest real root of

$$(\tilde{x} - x + \tfrac{1}{2})^2 \left\{ \frac{1}{\tilde{x}} + \frac{1}{n_1-\tilde{x}} + \frac{1}{m-\tilde{x}} + \frac{1}{n_2-m+\tilde{x}} \right\} = \chi^2_\alpha$$

where χ^2_α is upper 100α% point of chisquare distribution. Denote by ψ_U and ψ_L the value of ψ obtained by substituting $\tilde{x}_2$ and $\tilde{x}_1$ in (2.3.18) respectively. Thus,

$$\psi_L = \frac{\tilde{x}_1(n_2-m+\tilde{x}_1)}{(n_1-\tilde{x}_1)\ (m-\tilde{x}_1)} \qquad (2.3.19)$$

and

$$\psi_U = \frac{\tilde{x}_2(n_2-m+\tilde{x}_2)}{(n_1-\tilde{x}_2)\ (m-\tilde{x}_2)} \qquad (2.3.20)$$

Then $P(\tilde{x}_1 \le x \le \tilde{x}_2) = 1-\alpha$ asymptotically and since ψ is a monotone function of x, the asymptotic probability that the statement $\psi_L \le \psi \le \psi_U$ is correct is also equal to $(1-\alpha)$.

Example 2.3.3 :

Consider the following subset of data resulted from a case-control study of menopausal estrogen use and endometrial cancer. (Schlesselman, 1982).

Table 2.3.3 : Use of Oral Conjugated Estrogens (OCE) for Cases of Endometrial Cancer and controls

OCE	Endometrial cancer	control	Total
Yes	55	19	74
No	128	164	292
	183	183	366

The approximate 95% confidence limits for the true odds ratio using Cornfield's method are given by

$$\psi_L = 2.0271 \quad \text{and} \quad \psi_U = 6.8392.$$

2.3.3 Some more approximate methods :

We now consider two more approximate methods viz.

(1) Goodman's modification to Fisher's approximate method and (2) Gart's (1962) approximate method (for large numbers).

1. Goodman's modification to Fisher's approximate method :

Goodman (1964) has suggested a method for obtaining confidence limits for the odds ratio which is based on Fisher's approximate method.

Let $\hat{p}_{11} = x/N$, $\hat{p}_{12} = (n_1-x)/N$, $\hat{p}_{21} = y/N$, $\hat{p}_{22} = (n_2-y)/N$.

The chisquare statistic (2.3.16) can be rewritten as

$$X^2 = N\varepsilon^2\left\{ \frac{1}{p_{11}-\varepsilon} + \frac{1}{p_{12}+\varepsilon} + \frac{1}{p_{21}+\varepsilon} + \frac{1}{p_{22}-\varepsilon} \right\} \qquad (2.3.21)$$

where $\epsilon = z/N$.

When the marginal proportions in 2x2 contingency table are fixed and $N \to \infty$, then ε will converge in probability to zero and the statistic (2.3.21) will be asymptotically equivalent to

$$W^2 = z^2(x^{-1} + (n_1-x)^{-1} + (m-x)^{-1} + (n_2-m+x)^{-1}) \qquad (2.3.22)$$

Instead of solving (2.3.13), the quartic equation given by Fisher (1962) which can be rewritten as

$$X^2 = \chi^2_\alpha \qquad (2.3.23)$$

Goodman has recommended to replace X^2 in (2.3.16) by W^2, thus obtaining

$$W^2 = \chi^2_\alpha$$

$$\text{i.e.} \quad z^2S^2 = \chi^2_\alpha \qquad (2.3.24)$$

where

$$S^2 = \left[\frac{1}{x} + \frac{1}{n_1-x} + \frac{1}{y} + \frac{1}{n_2-y} \right] \qquad (2.3.25)$$

Equation (2.3.24) has two real solutions viz.

$$z_1^* = \sqrt{x^2_\alpha} / S \quad \text{and} \quad z_2^* = - z_1^*$$

Replacing z_1 and z_2 in (2.3.14) and (2.3.15) by z_1^* and z_2^* respectively, we obtain the approximate two sided confidence limits

$$\psi_L^* = \frac{(x-z_1^*)(n_2-y-z_1^*)}{(y+z_1^*)(n_1-x+z_1^*)} \qquad (2.3.26)$$

and

$$\psi_U^* = \frac{(x+z_1^*)\ (n_2-y+z_1^*)}{(y-z_1^*)\ (n_1-x-z_1^*)} \qquad (2.3.27)$$

Denoting $z_1^*/N = \varepsilon^*$, we can write ψ_L^* and ψ_U^* as

$$\psi_L^* = \frac{(\hat{p}_{11} - \varepsilon^*)\ (\hat{p}_{22} - \varepsilon^*)}{(\hat{p}_{21} + \varepsilon^*)\ (\hat{p}_{12} + \varepsilon^*)} \qquad (2.3.28)$$

$$\psi_U^* = \frac{(\hat{p}_{11} + \varepsilon^*)\ (\hat{p}_{22} + \varepsilon^*)}{(\hat{p}_{21} - \varepsilon^*)\ (\hat{p}_{12} - \varepsilon^*)} \qquad (2.3.29)$$

When marginal proportions are held fixed, and $N \to \infty$, (2.3.28) and (2.3.29) will be asymptotically equivalent to

$$\psi_L^* = \frac{\hat{P}_{11}\hat{P}_{22}}{\hat{P}_{12}\hat{P}_{21}} \left[1 - \varepsilon^* \left(\frac{1}{P_{11}} + \frac{1}{P_{12}} + \frac{1}{P_{21}} + \frac{1}{P_{22}} \right) \right] \qquad (2.3.30)$$

$$\psi_U^* = \frac{\hat{P}_{11}\hat{P}_{22}}{\hat{P}_{12}\hat{P}_{21}} \left[1 + \varepsilon^* \left(\frac{1}{P_{11}} + \frac{1}{P_{12}} + \frac{1}{P_{21}} + \frac{1}{P_{22}} \right) \right] \qquad (2.3.31)$$

respectively.

We thus have confidence limits given by

$$\psi_L^* = \hat{\psi}_{Un}[1 - \sqrt{\chi_\alpha^2}\ S] \text{ and } \psi_U^* = \hat{\psi}_{Un}[1 + \sqrt{\chi_\alpha^2}\ S] \qquad (2.3.32)$$

where $\hat{\psi}_{Un}$ denote observed cross product ratio.

The above limits are easy to calculate and they will serve as approximations to Fisher's approximate limits in the case where $N \to \infty$ and both the sets of marginal totals are fixed.

We now prove that the limits obtained in (2.3.32) will also serve as confidence limits for the odds ratio when one set of margins or none of the margins is fixed.

For the case of one set of margins fixed; we obtain asymptotic variance for $\hat{\psi}_{Un}$.

Lemma 2.3.2 : When one set of margins is fixed; $\hat{\psi}_{Un}$ is CAN for ψ with asymptotic variance given by

$$\psi^2 \left\{ \frac{1}{n_1 p_1 (1-p_1)} + \frac{1}{n_2 p_2 (1-p_2)} \right\}$$

Proof : When one set of margins is fixed, likelihood can be written as

$$L = P(X = x,\ Y = y \mid p_1,\ p_2)$$

$$= \binom{n_1}{x} \binom{n_2}{y} p_1^x (1 - p_1)^{n_1 - x}\ p_2^y (1-p_2)^{n_2 - y}$$

Note that it is a two parameter exponential family with parameter (p_1, p_2). M.L.E. of p_i $(i = 1,2)$ is given by $\hat{p}_1 = x/n_1$ and $\hat{p}_2 = y/n_2$ respectively and $(\hat{p}_1,\ \hat{p}_2)$ is consistent for $(p_1,\ p_2)$. Further $\begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} \sim AN \left\{ \begin{pmatrix} p_1 \\ p_2 \end{pmatrix},\ I^{-1}(p_1,\ p_2) \right\}$ where $I(p_1, p_2)$ denote information matrix and is given by

$$I(p_1,\ p_2) = \begin{bmatrix} E(-\dfrac{\partial^2 \ln L}{\partial p_1^2}) & E(-\dfrac{\partial^2 \ln L}{\partial p_1 \partial p_2}) \\ E(-\dfrac{\partial^2 \ln L}{\partial p_2\ \partial p_1}) & E(-\dfrac{\partial^2 \ln L}{\partial p_2^2}) \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{n_1}{p_1(1-p_1)} & 0 \\ 0 & \dfrac{n_2}{p_2(1-p_2)} \end{bmatrix}$$

Now consider the transformation $\psi = \dfrac{p_1(1-p_2)}{p_2(1-p_1)}$. Using the invariance property of CAN estimators $\hat{\psi}_{Un}$ is CAN for ψ with asymptotic variance given by $JI^{-1}(p_1,p_2)J'$ where

$$J = \left(\frac{\partial\psi}{\partial p_1} \quad \frac{\partial\psi}{\partial p_2} \right)$$

$$= \left(\frac{p_1(1-p_2)}{p_2(1-p_1)^2} \qquad \frac{-\,p_1}{p_2^2(1-p_1)} \right).$$

Hence asymptotic variance

$$\text{Var}\,(\hat{\psi}_{Un}) = \frac{p_1(1-p_2)^2}{n_1p_2^2(1-p_1)^3} + \frac{p_1^2(1-p_2)}{n_2p_2^3(1-p_1)^2}$$

$$= \frac{p_1^2(1-p_2)^2}{p_2^2(1-p_1)^2}\left[\frac{1}{n_1p_1(1-p_1)} + \frac{1}{n_2p_2(1-p_2)} \right]$$

$$= \psi^2\left[\frac{1}{n_1p_1(1-p_1)} + \frac{1}{n_2p_2(1-p_2)} \right]$$

□

Similarly, when none of the margins is fixed, asymptotic variance is given by $\text{Var}(\hat{\psi}_{Un}) = \frac{\psi^2}{N}\left\{ \frac{1}{P_{11}} + \frac{1}{P_{12}} + \frac{1}{P_{21}} + \frac{1}{P_{22}} \right\}$.

Thus in both the cases estimate of asymptotic variance of $\hat{\psi}_{Un}$, $\hat{\text{Var}}(\hat{\psi}_{Un})$, is given by $\hat{\psi}^2_{Un}S^2$. Hence in both the cases ψ^*_L and ψ^*_U given by (2.3.32) are approximate limits.

2. We now describe the method proposed by Gart (1962) for large numbers.

The chisquare statistic for testing independence in a 2x2 contingency table is given by

$$X^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\bar{p}\,\bar{q}\,(\frac{1}{n_1} + \frac{1}{n_2})} \qquad (2.3.33)$$

Here unknown p $(= p_1 = p_2)$ is estimated by

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \quad \text{and} \quad \bar{q} = 1 - \bar{p}$$

Confidence limits on $(p_1 - p_2)$ with approximate confidence coefficient $(1-\alpha)$ are found by modifying this statistic (2.3.33) and solving the equation

$$\frac{(\hat{p}_1 - \hat{p}_2 - (p_1 - p_2))^2}{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \chi^2_\alpha \qquad (2.3.34)$$

In an analogous way, confidence limits for the odds ratio may be found. The chisquare statistic (2.3.33) may be alternatively written as

$$X^2 = \frac{(\hat{p}_1\hat{q}_2 - \hat{p}_2\hat{q}_1)}{\bar{p}\,\bar{q}\,(\frac{1}{n_1} + \frac{1}{n_2})} \qquad (2.3.35)$$

This expression being algebraically identical to (2.3.33). The modification of (2.3.35) proceeds as follows. In the numerator, substitute $Z^2 = (\hat{p}_1\hat{q}_2 - \psi\hat{p}_2\hat{q}_1)^2$ and in the denominator substitute estimated variance of Z, denoted by $\hat{V}(Z)$. Now for all ψ, $E(Z)=0$ and Z is a well behaved function of sums of independent random variables. Thus, $Z^2/\hat{V}(Z)$ has limiting chisquare distribution.

We derive expression for variance of Z in the following lemma.

Lemma 2.3.3 : Exact expression for variance of Z is given by

$$v(Z) = \frac{p_2 q_2}{n_2}(p_1 + \psi q_1)^2 + \frac{p_1 q_1}{n_1}(\psi p_2 + q_2)^2 + \frac{p_1 q_1 p_2 q_2}{n_1 n_2}(1-\psi)^2$$

(2.3.36)

Proof :

$$Z = \hat{p}_1 \hat{q}_2 - \psi \hat{p}_2 \hat{q}_1$$

$$= \frac{X}{n_1} \frac{(n_2 - Y)}{n_2} - \psi \frac{Y}{n_1} \frac{(n_1 - X)}{n_2} .$$

Now

$$v(Z) = v(E(Z|Y)) + E(v(Z|Y)) \quad (2.3.37)$$

$$E(Z|Y) = E\left\{ \frac{X(n_2 - Y)}{n_1 n_2} - \frac{Y(n_1 - X)\psi}{n_1 n_2} \mid Y \right\}$$

$$= \frac{(n_2 - Y)}{n_1 n_2} \times n_1 p_1 - \frac{Y \psi}{n_1 n_2}(n_1 - n_1 p_1)$$

$$= p_1 - \left(\frac{p_1 + \psi q_1}{n_2} \right) Y$$

$$\therefore \quad v(E(Z|Y)) = \frac{p_2 q_2}{n_2}(p_1 + \psi q_1)^2 \quad (2.3.38)$$

Now

$$v(Z|Y) = \frac{(n_2 - Y)^2}{n_1^2 n_2^2} n_1 p_1 q_1 + \frac{Y^2 \psi^2 n\ p_1 q_1}{n_1^2 n_2^2} + \frac{2(n_2 - Y) Y \psi\ n_1 p_1 q_1}{n_1^2 n_2^2}$$

$$E(v(Z|Y)) = \frac{p_1 q_1}{n_1} [1 - 2p_2 + \frac{p_2 q_2}{n_2} + p_2^2]$$

$$+ \frac{p_1 q_1}{n_1} \psi^2 [\frac{p_2 q_2}{n_2} + p_2^2]$$

$$+ \quad 2\psi \frac{p_1 q_1}{n_1} [p_2 - \frac{p_2 q_2}{n_2} - p_2^2]$$

$$= \frac{p_1 q_1}{n_1} (\psi p_2 + q_2)^2 + \frac{p_1 q_1 p_2 q_2}{n_1 n_2} (1 - \psi)^2 \qquad (2.3.39)$$

From (2.3.37), (2.3.38) and (2.3.39)

$$V(Z) = \frac{p_2 q_2}{n_2} (p_1 + \psi q_1)^2 + \frac{p_1 q_1}{n_1} (p_2 \psi + q_2)^2 + \frac{p_1 q_1 p_2 q_2}{n_1 n_2} (1 - \psi)^2$$

□

Now, an unbiased estimator of var(z), except for terms of $O(n_1^{-1} n_2^{-1})$, $O(n_1^{-2} n_2^{-1})$, $O(n_1^{-1} n_2^{-2})$, $O(n_1^{-2} n_2^{-2})$ is

$$\hat{v}(Z) = \frac{\hat{p}_2 \hat{q}_2}{(n_2 - 1)} (\hat{p}_1 + \psi \hat{q}_1)^2 + \frac{\hat{p}_1 \hat{q}_1}{(n_1 - 1)} (\psi \hat{p}_2 + \hat{q}_2)^2 + \frac{\hat{p}_1 \hat{q}_1 \hat{p}_2 \hat{q}_2}{n_1 n_2} (1 - \psi)^2 \qquad (2.3.40)$$

Upon the correction for continuity, we are led to estimation equation

$$(Z \pm \frac{1}{2} (\frac{1}{n_1} + \frac{1}{n_2}))^2 = \chi_\alpha^2 \hat{v}(Z) \qquad (2.3.41)$$

where the 'plus' sign is associated with upper limit and the 'negative' sign with lower limit.

Equation (2.3.41) is a quadratic in ψ and roots are limits of an interval with confidence coefficient approximately equal to $(1-\alpha)$.

Hence by solving the equation (2.3.41) we have the lower limit

$$\psi_L = \frac{\hat{\psi}_{Un}\left\{1 + (\frac{1}{n_1} + \frac{1}{n_2})(\chi^2_\alpha - \frac{1}{2\hat{p}_1\hat{q}_2}) - \sqrt{\chi^2_\alpha}\sqrt{\left[\frac{1}{n_1\hat{p}_1\hat{q}_1} + \frac{1}{n_2\hat{p}_2\hat{q}_2} + \frac{A}{n_1 n_2} + B(\frac{1}{n_1} + \frac{1}{n_2})\right]}\right\}}{1 - \chi^2_\alpha\left[\frac{\hat{p}_1}{\hat{q}_1 n_1} + \frac{\hat{q}_2}{\hat{p}_2 n_2}\right]}$$

(2.3.42)

and upper limit

$$\psi_U = \frac{\hat{\psi}_{Un}\left\{1 + (\frac{1}{n_1} + \frac{1}{n_2})(\chi^2_\alpha - \frac{1}{2\hat{p}_1\hat{q}_2}) + \sqrt{\chi^2_\alpha}\sqrt{\left[\frac{1}{n_1\hat{p}_1\hat{q}_1} + \frac{1}{n_2\hat{p}_2\hat{q}_2} + \frac{A}{n_1 n_2} + B(\frac{1}{n_1} + \frac{1}{n_2})\right]}\right\}}{1 - \chi^2_\alpha\left[\frac{\hat{p}_1}{\hat{q}_1 n_1} + \frac{\hat{q}_2}{\hat{p}_2 n_2}\right]}$$

(2.3.43)

where

$$A = \frac{1}{\hat{\psi}_{Un}}\left[(1 - \hat{\psi}_{Un})^2 - \chi^2_\alpha\left(\frac{\hat{p}_1}{\hat{q}_1} - \frac{\hat{q}_2}{\hat{p}_2}\right)^2\right]$$

and

$$B = \frac{1}{\hat{p}_1\hat{q}_1\hat{p}_2\hat{q}_2}\left(\frac{\hat{q}_1}{n_2} + \frac{\hat{p}_2}{n_1}\right).$$

In the derivation of (2.3.42) and (2.3.43) terms of $O(n_1^{-1}n_2^{-1/2})$ $O(n_1^{-1/2}\ n_2^{-1})$, $O(n_1^{-1}n_2^{-1})$, $O(n_1^{-1/2}n_2^{-3/2})$, and $O(n_1^{-3/2}n_2^{-1/2})$ have been dropped.

It is suggested (Gart 1962) that (2.3.42) and (2.3.43) be used for tables in which

$$\frac{x(m-x)}{n_1+n_2} > 1$$

although when the quantity is close to 1 the limits tend to be too wide.

If the quantity greatly exceeds one, to be specific when

$$\frac{x(m-x)}{n_1+n_2} \geq 5$$

the following somewhat simpler formulae may be used.

$$\psi_L = \hat{\psi}_{Un}\left\{1 - \sqrt{\chi^2_\alpha}\ \sqrt{\frac{1}{n_1\hat{p}_1\hat{q}_1} + \frac{1}{n_2\hat{p}_2\hat{q}_2}}\ \right\} \qquad (2.3.44)$$

$$\psi_U = \hat{\psi}_{Un}\left\{1 + \sqrt{\chi^2_\alpha}\ \sqrt{\frac{1}{n_1\hat{p}_1\hat{q}_1} + \frac{1}{n_2\hat{p}_2\hat{q}_2}}\ \right\} \qquad (2.3.45)$$

Equations (2.3.44) and (2.3.45) are obtained from (2.3.42) and (2.3.43) by dropping terms which are of $O(n_1^{-1})$, $O(n_2^{-1})$ and $O(n_1^{-1/2}n_2^{-1/2})$. Note that limits given by (2.3.44) and (2.3.45) are same as given by Goodman's modification to Fisher's method.

Example 2.3.4 : Consider again the 2x2 table in example 2.3.3.

Note that $\dfrac{x(m-x)}{n_1+n_2} > 5$.

Hence we use (2.3.44) and (2.3.45) to calculate 95% confidence limits for the odds ratio.

$$\psi_L = 1.5931$$

and $$\psi_U = 4.7451$$

Note that the same limits are obtained using Goodman's modification to Fisher's approximate method.

2.3.4 Logit method :

Another approach to setting approximate confidence limits on ψ is based on a transformation of approximate limits for $\beta = \ln \psi$.

In case where one set of margins is fixed or where none is fixed; the statistic $\ln \psi_{Un}$ is m.l.e. of β. Haldane (1955) and Anscombe (1956) have shown that an approximately unbiased estimate of β can be derived by adding 1/2 to each of the cells of 2x2 table and using

$$\hat{\beta} = \ln((x + 1/2)(n_2-y+1/2) / (y+1/2)(n_1 - x+1/2)) .$$

The estimated asymptotic variance of $\hat{\beta}$ is given by

$$\hat{Var}(\hat{\beta}) = \left\{ \frac{1}{x+1/2} + \frac{1}{n_1-x+1/2} + \frac{1}{y+1/2} + \frac{1}{n_2-y+1/2} \right\} .$$

Since $\hat{\beta}$ in large samples has a normal distribution, an approximate $(1-\alpha)100\%$ confidence limits for β is

$$\hat{\beta} \pm z_{\alpha/2} \sqrt{\frac{1}{x+1/2} + \frac{1}{n_1-x+1/2} + \frac{1}{y+1/2} + \frac{1}{n_2-y+1/2}}$$

where $z_{\alpha/2}$ is the point on the unit normal distribution that is exceeded with probability $\alpha/2$.

Approximate lower and upper confidence limits for ψ are found by taking antilogs of the limits for β. Thus,

$$\psi_L = \hat{\psi}_{Un} \exp\left\{ - z_{\alpha/2} \sqrt{\widehat{var}(\hat{\beta})} \right\} \tag{2.3.46}$$

and

$$\psi_U = \hat{\psi}_{Un} \exp\left\{ + z_{\alpha/2} \sqrt{\widehat{var}(\hat{\beta})} \right\} \tag{2.3.47}$$

These limits will also serve as approximations to Fisher's confidence limits where both sets of margins are fixed.

Example 2.3.5 : Consider again the 2x2 table in example 2.3.1. 95% confidence limits for the odds ratio using equations (2.3.46) and (2.3.47) are given by

$$\psi_L = 2.1075$$

$$\psi_U = 6.5270$$

In the following table, we summarise limits given by various approximate procedures.

Table 2.3.4 : 95% limits for the odds ratio given by various approximate procedures

a b	limits using Cornfield's method		limits using Gart's method		limits using logit method	
c d	ψ_L	ψ_U	ψ_L	ψ_U	ψ_L	ψ_U
53 19						
128 164	2.0271	6.8392	1.5931	4.7450	2.1075	6.5270

2.4 : Testing the hypothesis of independnce

Various test procedures are suggested for testing the hypothesis of no association in a 2x2 contingency table. We discuss the following tests.

(I) Fisher's exact test

(II) Gart's approximation to Fisher's exact test.

(III) Uncorrected chisquare test.

(IV) Continuity corrected chisquare test.

(I) Fisher's exact test :

We first consider the one sided tests only. An exact one sided test of significance of null hypothesis of no association i.e. $H_0 : \psi = 1$ Vs $H_1 : \psi < 1$ can be obtained from distribution of X conditional on all the margins, being fixed.

p-value in this case is given by

$$P_E = \sum_{i=0}^{x} g(i \mid m, 1) \quad (2.4.1)$$

where

$$g(i \mid m, 1) = \frac{\binom{n_1}{i}\binom{n_2}{m-i}}{\binom{n_1+n_2}{m}} \quad (2.4.2)$$

Similarly, p-value corresponding to the hypothesis $H_0 : \psi=1$ Vs $H_1 : \psi > 1$ is calculated as

$$P_E = \sum_{i=x}^{m} g(i \mid m, 1) \quad (2.4.3)$$

How to calculate two-tailed probability is a matter of dispute among statisticians which we discuss later.

(II) Gart's approximation to Fisher's exact test :

Gart (1962) has proposed the following approximation to Fisher's exact test.

Consider the hypothesis $H_o : \psi = 1$ Vs $H_1 : \psi > 1$. Then the p-value calculated from the exact test is given by (2.4.1). Wise (1954) has shown that a good first approximation to p is found by approximating the hypergeometric terms in (2.4.1) by binomial so that

$$p \simeq \sum_{i=o}^{x} \binom{m}{i} c^{i} (1-c)^{m-i} \qquad (2.4.4)$$

where

$$c = \frac{2n_1 - k + 1}{2(n_1+n_2) - m + 1}$$

k being the number of terms in the sum in question, (x+1) in this case. Using the well-known relationship between binomial sum and incomplete beta function, we find

$$p \simeq I_{1-c}\,(m-x,\ x+1)$$

$$\therefore\ p \simeq 1 - I_c(x+1,\ m-x)$$

$$\therefore\ I_c(x+1,\ m-x) \simeq 1 - p \qquad (2.4.5)$$

Since the lower $100\nu\%$ point of F distribution with ν_1 and ν_2 degrees of freedom $(F_{1-\nu}(\nu_1,\nu_2))$ satisfies the equation

$$I_{\frac{\nu_1 F}{\nu_1 F + \nu_2}}(\nu_1/2, \nu_2/2) = \nu \qquad (2.4.8)$$

we have from (2.4.5)

$$\nu_1 = 2(x+1), \quad \nu_2 = 2(m-x)$$

$$c = \frac{\nu_1 F}{\nu_1 F + \nu_2}$$

$$\leftrightarrow \frac{1}{c} = 1 + \frac{\nu_2}{\nu_1 F}$$

$$\leftrightarrow \frac{1-c}{c} = \frac{\nu_2}{\nu_1 F}$$

$$\leftrightarrow F_{1-(1-p)}(\nu_1, \nu_2) = \frac{1-c}{c} \times \frac{\nu_2}{\nu_1}$$

$$= \frac{2n_1 - x}{2n_2 - m + x + 1} \times \frac{m-x}{x+1}$$

$$\therefore \; P\left\{F(2x+2,\ 2m-2x) \geq \frac{m-x}{x+1} \times \frac{2n_1 - x}{2n_2 - m + x + 1}\right\} = p\,.$$

Thus, we may perform an approximate test by computing

$$F = \frac{(m-x)\,(2n_1 - x)}{(x+1)\,(2n_2 - m + x + 1)} \qquad (2.4.6)$$

and comparing this statistic with upper 100α% point of F distribution with 2(x+1) and 2(m-x) degrees of freedom; α being predetermined level of significance. This approximate test has been found to be accurate for tables with small numbers, a good rule is to use this test whenever $\frac{x(m-x)}{n_1+n_2} \leq 1$.

Now consider the other one-sided test.

$H_0 : \psi = 1$ Vs $H_1 : \psi < 1$. Exact p-value in this case is given by (2.4.3). Again using the approximation given by Wise

$$p \simeq \sum_{i=x}^{m} \binom{m}{i} (c')^{i} (1 - c')^{m-i} \qquad (2.4.7)$$

where $c' = \dfrac{2n_1 - m + x}{2(n_1+n_2) - m+1}$.

Again using the relationship between binomial sum and incomplete beta function,

$$I_{c'}(x,\ m-x+1) \simeq p \qquad (2.4.8)$$

Using relation between F-distribution and incomplete beta function we have

$$I_{\frac{\nu_1 F}{\nu_1 F + \nu_2}} (\nu_1/2,\ \nu_2/2) = p$$

so that $\nu_1 = 2x$, $\nu_2 = 2(m-x+1)$

$$c' = \frac{\nu_1 F}{\nu_1 F + \nu_2}$$

$$\therefore \quad F = \frac{\nu_2}{\nu_1} \times \frac{c'}{1-c'}$$

$$= \frac{2(m-x+1)}{2x} \times \frac{2n_1-m+x}{2n_2-x}$$

$$\therefore \quad P\left\{F(2x,\ 2(m-x+1) \le \frac{2(m-x+1)}{2x} \times \frac{2n_1-m+x}{2n_2-x}\right\} \simeq p .$$

Thus, we perform an approximate test by computing

$$F = \frac{(m-x+1)}{x} \times \frac{2n_1-m+x}{2n_2-x} \qquad (2.4.8)$$

and comparing this statistic with lower 100α% point of F distribution with 2x and 2(m-x+1) degrees of freedom, α being predetermined level of significance.

(III) Chisquare (χ^2) test :

In the case where data is generated either by sampling from two binomial distributions or by sampling from single multinomial distribution, the usual χ^2 statistic for testing independence is a 2x2 contingency table is given by

$$X^2 = \frac{N[x(n_2-y) - y(n_1-x)]^2}{n_1 n_2 m(N - m)} \qquad (2.4.9)$$

Under the null hypothesis of no association, the statistic X^2 has an approximate chisquare distribution with one degree of freedom when the total sample size is large.

A chisquare test is essentially a two-sided test. For testing the one sided hypothesis, we use the approximate unit normal deviate $z = \pm \sqrt{\chi^2_\alpha}$.

When the data is generated by sampling from two binomial distributions, equivalently, the one sided hypothesis $H_0 : p_1 = p_2$ Vs $H_1 : p_1 > p_2$ can be tested using the statistic

$$T_N = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\,(1/n_1 + 1/n_2)}} \qquad (2.4.10)$$

where

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \qquad \bar{q} = 1 - \bar{p}$$

$$\hat{p}_1 = x/n_1 \qquad \hat{p}_2 = y/n_2 \ .$$

The significance p-value for the observations is given by

$$P_N = p(T_N \geq t_n)$$

$$= \int_{t_n}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(- \frac{1}{2} z^2\right) dz$$

where t_n is the value of T_N given by the observed table.

If $\hat{p}_1 < \hat{p}_2$; the test function is not calculated. If $\hat{p}_1 \geq \hat{p}_2$; then the test function is calculated.

Apart from T_N, the statistic

$$T_{N'} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \qquad (2.4.11)$$

is also used to test the hypothesis. Sathe (1982) has suggested the following statistic

$$T_{N''} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_2} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_1}}} \qquad (2.4.12)$$

Similarly, p-value for the other one sided hypothesis $H_0 : p_1 = p_2$ Vs $H_1 : p_1 < p_2$ can be calculated by

$$p_N = p(T_N \le t'_n)$$

$$= \int_{-\infty}^{t'_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2} dz$$

where t'_n is the value of T_N for the observed table. Here test function is calculated if $\hat{p}_1 < \hat{p}_2$. The hypothesis can also be tested using the test statistics $T_{N'}$, and $T_{N''}$.

(IV) Continuity corrected chisquare (x_c^2) test :

Because X^2 depends only on cell counts and because cell counts assume successive integer values, it seems natural to adjust the cell counts by the amount 1/2 to obtain a continuity correction for X^2. This procedure proposed by Yates (1934), results in the use of the statistic

$$X_c^2 = \frac{N(\left|x(n_2 - y) - y(n_1 - x)\right| - \frac{1}{2} N)^2}{n_1 n_2 \; m(N-m)} \qquad (2.4.13)$$

In case of sampling from two binomial distributions the one sided hypothesis $H_0 : p_1 = p_2$ Vs $H_1 : p_1 > p_2$ can be tested in a similar way to that of chi-square test by using the statistic

$$T_c = \frac{\left|\hat{p}_1 - \hat{p}_2\right| - \frac{1}{2} (\frac{1}{n_1} + \frac{1}{n_2})}{\sqrt{\bar{p}(1-\bar{p}) (\frac{1}{n_1} + \frac{1}{n_2})}} \qquad (2.4.14)$$

which is distributed as N(0,1) for large n_1, n_2 .

The p-value is calculated as

$$p_c = p(T_c \ge t_c)$$

where t_c is the value of T_c for the observed table. Similarly, the other one sided hypothesis can be tested.

Discussion :

Here we are concerned with the test of the hypothesis of no association in a 2x2 table. In case of samples from two binomial populations, the problem reduces to the comparison of two observed proportions; a very old statistical problem that has a very literature associated with it. Upton (1982) has compared the performance of 22 different tests for the problem.

We discuss below the historical background as well as various points raised against the tests discussed above from time to time.

Early dispute regarding the degrees of freedom in a contingency table :

In 1900, Karl Pearson introduced the chisquare test for goodness of fit. The test can be used for testing the hypothesis of no association in a rxs contingency table. For the chisquare test, in addition to deducting one degree of freedom for the number in sample, an additional degree of freedom must be deducted for each additional parameter estimated from the data. In testing for association in contingency tables, the expectations of the cell values are estimated from the marginal totals and the number of degrees of freedom for a rxs contingency table is therefore (r-1) (s-1), not rs-1 as Pearson indicated.

This error is particularly serious in 2x2 tables for which chi-square with one degree of freedom must be used, and not the three degrees of freedom.

Udny Yule (1911) introduced a test for association in a 2x2 tables in his textbook 'Introduction to theory of statistics' using the large sample estimate $\sqrt{\hat{pq}/n}$ for standard error of a proportion p.

Under the null hypothesis of equality of proportions, Yule's test was equivalent to Pearson's chi-square test with one degree of freedom.

Yule did not mention the chi-square test in his text-book, but he evidently soon (Greenwood and Yule, 1915) became aware of the discrepancy between his test and the chi-square test with three degrees of freedom. Shortly afterwords, he constructed 350 2x2 tables and 100 4x4 tables by mechanical devices designed to give independent distributions, and compared the χ^2 - distributions so obtained with those given by theory, but did not immediately publish his results.

Pearson's error regarding the degrees of freedom was pointed out by Fisher (1922). Although Yule was not fully satisfied with Fisher's proof; he then simultaneously published his sampling investigations, which, as was to be expected, confirmed Fisher's results. Pearson did not immediately admit to any error and a considerable controversy arose, but the correctness of Fisher's conclusions ultimately came to be generally accepted.

Exact test and continuity corrected chi-square test :

The chi-square test is of course approximate and will not hold exactly when the expectations of the separate cells of a distribution or contingency table are small. In Statistical Methods for Research Workers (1925), Fisher advanced a rule of thumb that the expected number in any one cell should not be less than five. This rule may infact be adequate, indeed conservative, for tests involving more than one degree of freeedom (Yates 1984).

If the exact distribution related to any particular problem is known the accuracy of chi-square test (or any other aproximate test) can be investigated by comparing its performance with that given by the exact distribution over a range of typical examples.

The exact form of the distribution of a 2x2 table with given marginal totals was suggested by Fisher. This form depends on the restriction that only sets of values conforming to both pairs of observed marginal totals are included in evaluating probabilities.

The results of the investigation by Yates (1934) showed that the approximations given by chi-square to both binomial and 2x2 exact probability, particularly when the parent distributions are approximately symmetrical, are greatly improved by deducting 1/2 from observed deviations from expectations when calculating X^2. When parent distribution is markedly assymetric, continuity correction may not work.

Binomial distribution is markedly asymmetric if p differs greatly from 0.5. Concept of asymmetry in case of the exact form of a 2x2 table can be explained as follows :

We represent a 2x2 contingency table as

a	b	n_1
c	d	n_2
m	N-m	N

where 'a' is the cell with smallest expectation (denoted by e) under the hypothesis of independence. With $n_1 = n_2$, there will be pairs of points representing the integral division on the two tails equidistant from the expected value, e. These will have equal hypergeometric probabilities. Hence hypergeometric distribution is said to be symmetric. For example consider the table

a	b	20
c	d	20
7	33	40

with e = 3.5.

If $n_1 \neq n_2$, but 2e is integral there will still be pairs of points equidistant from e, but also some points on the longer tail that are unpaired. For example, consider the table

a	b	20
c	d	60
14	66	80

with e = 3.5.

For this case, the hyptergeometric distribution will be asymmetric and the associated probabilities will be unequal.

If 2e is not integral, there will be no equidistant pairs. This may be termed as mismatch.

Unconditional approach :

An unconditional approach to this problem was proposed first by Barnard (1945). Barnard (1947a) elaborated his proposal into what he termed as C.S.M. test.

Basis of Barnard's C.S.M. test : In case of sampling from two binomial distributions, all combinations of the possible samples of n_1 and n_2 are considered along with their associated probabilities ranked in the order of the values of $\hat{p}_1 - \hat{p}_2$. A condition, S, of symmetry is applied so that the points (a,b) and $(n_1-a,\ n_2-b)$ are equally significant. Barnard then imposed a natural convexity condition, c, which requires that if the observation $(n_1\hat{p}_1,\ n_2\hat{p}_2)$ is significant, then $(n_1\hat{\tilde{p}}_1,\ n_2\hat{\tilde{p}}_2)$ is significant if $\hat{\tilde{p}}_1 - \tilde{p}_2$ is greater than $\hat{p}_1-\hat{p}_2$. The conditions S and C give only a partial ordring of the sample space. The ordering will be complete if the maximum condition M, is imposed.

Conditions S and C alone pick out two distinct points as being most significant. One can then consider what other points always using S and C, should be regarded as next in the significance ordering and so on. If, for instance we have selected the ordering z_1, z_2, ..., z_k, these being points in the sample space, one then consider subject to S and C, as the next condidate point, say, z; one then looks at Prob(z_1, z_2, ..., z_k, z) which is a function of the unknown common value of p under the null hypothesis and chooses that z that makes this probability a maximum over p.

Barnard (1947a) argued that his test was more powerful than Fisher's exact test. But shortly afterwards (1949) he retracted his proposal saying that the reference set considered by Fisher to calculate the probabilities was correct.

The unconditional approach has attracted support from McDonald et al. (1977), Berkson (1978 a,b) and Kempthorne (1979).

Berkson (1978a) considered the problem of comparison of two observed proportions. The hypothesis of interest is $H_0: p_1=p_2$ Vs $H_1 : p_1 > p_2$. He compared the exact test for one sided case with normal test and normal test with continuity correction for the nominal significance levels 0.05 and 0.01. Here the nominal significance level α is the level α at which H_0 is formally rejected.

If we consider P_N, P_C, and P_E (or P_T for the test T) as the test functions and distributions of random variables P_T, specific values of which are denoted by p_T; then with an ideal test $Prob(P_T \leq \alpha | H_0) = \alpha$. With the three tests considered here, this is not generally true. The actual $Prob(P_T \leq \alpha | H_0)$ is referred to as 'effective α' and denoted by α_e while $\beta_e = Prob(P_T \leq \alpha | H_1)$ represents the 'effective power'.

To compute α_e and β_e, each possible sample with given $n_1=n_2$ was generated successively, and if $a \geq c$, its probability and the value of P_N, P_C and P_E were calculated. If $p_T \leq \alpha$, the probability of the sample was cumulated for the pertinent test T. The sum represents α_e or β_e depending on whether $p_1 = p_2$ or $p_1 > p_2$.

The three tests viz. T_N, T_C and T_E (exact test) are compared in terms of α_e and β_e for the sample sizes ranging from $n_1=n_2=5$ to $n_1 = n_2 = 200$. A test T was said to perform better than the test T' if the effective level of the test T is closer to the nominal α than that of the test T' and at the same time effective power for the test T is more than that of the test T'. Using this principle, Dr Berkson concluded that in case when samples are drawn from two binomial distributions, T_N is preferable to T_E and T_E should not be used.

The preference expressed by Berkson was strongly criticized by Barnard (1979), Basu (1979), Yates (1984), L.C.A. Corsten and deKroon (1979).

Conditional vs unconditional methods :

In the context of testing the significance of association in a 2x2 table, there is considerable debate about whether one should condition on both its margins.

It is sometimes (Kempthorne 1979, Upton 1982) represented that although conditioning on margins is justified by necessity for randomization in a comparative trial; this does not apply to samples from two binomial distributions for which more powerful unconditional tests are available.

The reason advanced for also conditioning on second margin in case of samples from two binomials is that the second margin does not contain information on the association within the table. Hence conditioning is said to eliminate the effect of overall success probability which is a nuisance parameter (Godambe 1980, Yates 1984).

Use of nominal level of significance :

A contributory cause of confusion that affects discontinuous data is the use of conventional nominal levels of significance, such as 5 or 1 percent. This was partly engendered by the use of nominal significance probability in the tables of t and normal distribution.

Exact test and its approximation by Yate's continuity correction have been continuously criticized as being conservative. [Berkson 1978, Upton 1982, D'Agostino et al 1988, Storer and Kim, 1990]. The exact test of course does not give the test with predetermined significance level α. In fact, because of discreteness of hypergeometric distribution; the observed significance level or p-value has the property $p \leq \alpha$, where p depends on marginal totals held fixed, and the test then will be always conservative.

If we use accept-reject rule in case of tests involving discontinuous data, a predetermined level of significance is rarely attained and these tests are always conservative. If we want to achieve a predetermined level of significance, a randomized test should be used. For 2x2 table, one such randomized test, which is infact UMPU, is suggested by Tocher (1950). But randomized tests are not useful in practice. Hence the actual significance probability or the p-value should always be mentioned when reporting on discontinuous data.

Uncorrected versus corrected chi-square tests :

Earlier (Pearson 1947, Plackett 1964, Grizzle 1967) chisquare test and continuity corrected chisquare test were compared using the exceedance probability estimates taking exact test as basis for comparison. These comparisons resulted in the criticism of continuity corrected chisquare test for being overly conservative.

Mantel and Greenhouse (1968) have supported the use of continuity corrected chi-square statistic with a two stage argument saying :

1. The proper probability model to use in a 2x2 table is the one with both sets of marginal totals fixed, which yields the hypergeometric distribution of X^2 and

2. The correction improves probability estimates for the hypergeometric distribution except in pathalogical cases such as when the distribution is sufficiently asymmetric.

Controversy arising in calculating two-tailed probability :

For two sided tests both tails of the distribution must be taken into account. Many authors advocate that the one-tailed p-value should be doubled. An alternative approach (Mantel 1974, Haber 1980) is to calculate p as total probability of tables in either tail, which are at least as extreme as that observed in the sense of having probability at least as small.

Thus as shown in the following example, the same observed one tailed probability can give rise to different two-tailed probabilities according to the rule used.

Example 2.4.1

Let $n_1 = 15$; $n_2 = 10$; $m = 9$.

Table 2.4.1 : Single tail and two-tailed probability based on exact distribution

a	Exact P(a)	Single tail probability	Two-tailed probability (I) doubling the single tail	Two-tailed probability (II) Summing the prob of two tails
0	0.000005	0.000005	0.000010	0.000005
1	0.000330	0.000335	0.000670	0.000335
2	0.006167	0.006502	0 013004	0.008952
3	0.046770	0.053273	0.106546	0.087220
.				
.			.	
.				
8	0.031498	0.033948	0.067896	0.040450
9	0.002450	0.002450	0.004900	0.002785

Table 2.4.2 : Single tail and two tailed probability based on continuity corrected statistic

			Two-tailed probability	
a	Continuity corrected estimate of p(a)	Single tail probability	(I) doubling the single tail	(II) Summing the prob. of two tails
0	0.000014	0.000014	0.000028	0.000014
1	0.000435	0.000449	0.000898	0.000449
2	0.006306	0.006755	0.013510	0.008393
3	0.045860	0.052615	0.105230	0.086835
.				
.				
.				
8	0.032582	0.034229	0.068458	0.040975
9	0.001638	0.001638	0.003276	0.002087

Chapter 3

2 x 3 Contingency Tables : Two Odds Ratios

CHAPTER 3

2 X 3 CONTINGENCY TABLES : TWO ODDS RATIOS

3.1: Introduction

In the previous chapter, we have dealt with point and interval estimation of the odds ratio and the test procedures for the hypothesis of no association in a 2x2 contingency table. Here we discuss extension of some of the results for the two odds ratios in a 2x3 contingency table.

As in 2x2 case, we consider the following two ways in which the data can be generated in a 2x3 contingency table and define population odds ratios.

Origin I data : In this case, none of the margins is fixed, only the sample size N is fixed. The population situation may be represented by

	B_1	B_2	B_3
A_1	P_{11}	P_{12}	P_{13}
A_2	P_{21}	P_{22}	P_{23}

with $\sum_i \sum_j P_{ij} = 1$, $i = 1,2$; $j = 1,2,3$.

If we draw N individuals from this population, our sample outcome may take the form

	B_1	B_2	B_3	
A_1	x_{11}	x_{12}	x_{13}	n_1
A_2	x_{21}	x_{22}	x_{23}	n_2
	m_1	m_2	$N-(m_1+m_2)$	N

Here the sample $(x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23})$ can be considered as realization of $(X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{23}) \sim$ Multinomial $(N, P_{11}, P_{12}, P_{13}, P_{21}, P_{22}, P_{23})$.

Origin II data (One set of margins fixed) :

Origin II is that we have two multinomial population A_1 and A_2. We have a random sample of size n_1 from popoulation A_1 and a random sample of size n_2 from population A_2. In this way, the row margins are fixed. We then observe the members of B_1, B_2 and B_3. The popoulation situation may be represented by

	B_1	B_2	B_3
A_1	P_{11}	P_{12}	P_{13}
A_2	P_{21}	P_{22}	P_{23}

with $\sum_{j=1}^{3} P_{ij} = 1$, $i = 1,2$.

The sample outcome can be represented by

	B_1	B_2	B_3	
A_1	x_{11}	x_{12}	$n_1-x_{11}-x_{12}$	n_1
A_2	x_{21}	x_{22}	$n_2-x_{21}-x_{22}$	n_2
	m_1	m_2	$N-(m_1+m_2)$	N

Here $(x_{i1}, x_{i2}, n_i - x_{i1} - x_{i2})$ can be considered as realization of $(X_{i1}, X_{i2}, X_{i3}) \sim$ Multinomial $(n_i, p_{i1}, p_{i2}, p_{i3})$, $i = 1,2$.

Odds ratios for a 2x3 contingency table : For a 2x3 contingencyy table, we can define three odds ratios, viz.

$$\psi_1 = \frac{P_{11}\,P_{23}}{P_{13}\,P_{21}}, \quad \psi_2 = \frac{P_{12}\,P_{23}}{P_{13}\,P_{22}} \quad \text{and} \quad \psi_3 = \frac{P_{11}\,P_{22}}{P_{12}\,P_{21}}$$

Since ψ_3 can be obtained as a function of ψ_1 and ψ_2 we consider inference on these two odds ratios. For simplicity of notation, we denote these odds ratios by ψ_{11} and ψ_{12} i.e. $\psi_{1j} = \frac{P_{1j}\,P_{23}}{P_{13}\,P_{2j}}$, j = 1, 2. Point estimation of the odds ratios is discussed in section 3.2. Section 3.3 discusses simultaneous interval estimation for the two odds ratios. The last section deals with the tests of hypothesis of independence.

3.2 : Point estimation

We discuss below unconditional maximum likelihood estimation for the odds ratios ψ_{11} and ψ_{12} in a 2x3 contingency table.

Unconditional maximum likelihood estimation :

<u>Case I</u> : When only the total sample size, N, is fixed, $(x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23})$ is realization of $(X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{23}) \sim$ Multinomial $(N, P_{11}, P_{12}, P_{13}, P_{21}, P_{22}, P_{23})$ and the unconditional likelihood function can be written as

$$L = \frac{N\,!}{x_{11}!\,x_{12}!\,x_{13}!\,x_{21}!\,x_{22}!\,x_{23}!} \prod_{i=1}^{2} \prod_{j=1}^{3} p_{ij}^{x_{ij}} \qquad (3.2.1)$$

M.l.e.s of p_{ij} $(i = 1,2,;\ j = 1,2,3)$ are obtained by maximizing L w.r.t. p_{ij} (i = 1,2; j = 1,2,3) with the restriction that $\sum_{i=1}^{2} \sum_{j=1}^{3} p_{ij} = 1$. Thus we have $\hat{p}_{ij} = x_{ij}/N$, i = 1,2; j=1,2,3. Hence the unconditional maximum likelihood estimators (UMLE) for ψ_{11} and ψ_{12} (denoted by $\hat{\psi}_{11}$ and $\hat{\psi}_{12}$) are given by

$$\hat{\psi}_{11} = \frac{x_{11}\, x_{23}}{x_{13}\, x_{21}} \qquad \hat{\psi}_{12} = \frac{x_{12}\, x_{23}}{x_{13}\, x_{22}}$$

Case II : When only one set of margins (n_1, n_2) is fixed, unconditional likelihood can be written as

$$L = \frac{n_1!\ n_2!}{x_{11}!x_{12}!x_{13}!x_{21}!x_{22}!x_{23}!} \prod_{i=1}^{2} \prod_{j=1}^{3} p_{ij}^{x_{ij}} \qquad (3.2.2)$$

Maximizing L w.r.t. p_{ij} (i = 1,2; j = 1,2,3) with the restriction that $\sum_{j=1}^{3} p_{ij} = 1$, i = 1,2 gives x_{ij}/n_i as m.l.e. of p_{ij}, i = 1,2; j = 1,2,3. Hence UMLE's of ψ_{11} and ψ_{12} are given by

$$\hat{\psi}_{11} = \frac{x_{11}\, x_{23}}{x_{13}\, x_{21}} \qquad \hat{\psi}_{12} = \frac{x_{12}\, x_{23}}{x_{13}\, x_{22}}$$

□

Conditional maximum likelihood estimation :

Theorem 3.2.1 : If we conndition on both the margins of a 2x3 contingency table; the conditional likelihood is given by

$$g(x_{11}, x_{12} \mid m_1, m_2, \psi_{11}, \psi_{12})$$

$$= \frac{u(x_{11}, x_{12}, m_1 - x_{11}, m_2 - x_{12})\ \psi_{11}^{x_{11}}\ \psi_{12}^{x_{12}}}{\sum_{i=r_1}^{r_2} \sum_{j=\ell(i)}^{s(i)} u(i,j, m_1 - i, m_2 - j)\ \psi_{11}^{i}\ \psi_{12}^{j}} .$$

where

$u(i,\ j,\ m_1-i,\ m_2-j)$

$$= \frac{n_1!}{i!\ j!\ (n_1-i-j)!} \times \frac{n_2!}{(m_1-i)!\ (m_2-j)!} \ \frac{\psi_{11}^{i}\ \psi_{12}^{j}}{(n_2-m_1-m_2+i+j)!}$$

$$\ell(i) = \max(0,\ m_1 + m_2 - n_2 - i)$$

$$s(i) = \min(n_1 - i,\ m_2)$$

$$r_1 = \max(0,\ m_1 - n_2) \quad \text{and} \quad r_2 = \min(m_1,\ n_1)\ .$$

Proof : Since we have to condition on both the sets of margins; we begin with conditioning on one set of margins. When only one set of margins (n_1, n_2) is fixed, the likelihood is given by

$$L = L(x_{11},\ x_{12},\ x_{21},\ x_{22} \mid p_{11},\ p_{12},\ p_{21},\ p_{22})$$

$$= \frac{n_1!}{x_{11}!x_{12}!(n_1-x_{11}-x_{12})!} \quad \frac{n_2!}{x_{21}!x_{22}!(n_2-x_{21}-x_{22})!} \prod_{i=1}^{2}\prod_{j=1}^{2} p_{ij}^{x_{ij}}$$

$$(1 - p_{11} - p_{12})^{n_1-x_{11}-x_{12}}\ (1 - p_{21} - p_{22})^{n_2-x_{21}-x_{22}}$$

$$= u(x_{11},\ x_{12},\ x_{21},\ x_{22})\ \ p_{13}^{n_1}\ \ p_{23}^{n_2}$$

$$\left[\frac{p_{11}\ p_{23}}{p_{13}\ p_{21}}\right]^{x_{11}} \left[\frac{p_{12}\ p_{23}}{p_{13}\ p_{22}}\right]^{x_{12}} \left[\frac{p_{21}}{p_{23}}\right]^{x_{11}+x_{21}} \left[\frac{p_{22}}{p_{23}}\right]^{x_{12}+x_{22}}$$

where $p_{13} = 1 - p_{11} - p_{12}$ and $p_{23} = 1 - p_{21} - p_{22}$.

Thus, we can write

$$L = L(x_{11},\ x_{12},\ x_{21},\ x_{22} \mid \psi_{11},\ \psi_{12},\ \nu_1,\ \nu_2)$$

$$= u(x_{11},x_{12},x_{21},x_{22})\quad v(\psi_{11},\psi_{12},\nu_1,\nu_2)\quad \psi_{11}^{x_{11}}\quad \psi_{12}^{x_{12}}\quad \nu_1^{x_{11}+x_{21}}\nu_2^{x_{11}+x_{22}} \tag{3.2.3}$$

where $\psi_{11} = \dfrac{p_{11}\, p_{23}}{p_{13}\, p_{21}}$, $\psi_{12} = \dfrac{p_{12}\, p_{23}}{p_{13}\, p_{22}}$

$\nu_1 = \dfrac{p_{21}}{p_{23}}$ and $\nu_2 = \dfrac{p_{22}}{p_{23}}$.

We observe that (3.2.3) belongs to four parameter exponential family and $(X_{11}, X_{12}, X_{11} + X_{21}, X_{12} + X_{22})$ is minimal sufficient statistic. Now,

$$P(X_{11} = x_{11}, X_{12} = x_{12} \mid X_{11} + X_{21} = m_1, X_{12} + X_{22} = m_2)$$

$$= \frac{P(X_{11} = x_{11}, X_{21} = m_1 - x_{11}, X_{12} = x_{12}, X_{22} = m_2 - x_{12})}{P(X_{11} + X_{21} = m_1,\ X_{12} + X_{22} = m_2)}$$

(3.2.4)

Since n_1, n_2, m_1 and m_2 are fixed, the range for x_{11} and x_{12} is given by

$$\max(0, m_1 - n_2) \le x_{11} \le \min(m_1, n_1)$$

and

$$\max(0, m_1+m_2-n_1-x_{11}) \le x_{12} \le \min(n_1 - x_{11}, m_2)$$

or

$$\max(0, m_2 - n_2) \le x_{12} \le \min(m_2, n_1)$$

and

$$\max(0, m_1 + m_2 - n_2 - x_{12}) \le x_{11} \le \min(n_1 - x_{12}, m_2).$$

Now, from (3.2.3)

$$P(X_{11} = x_{11}, X_{12} = x_{12}, X_{21} = m_1 - x_{11}, X_{22} = m_2 - x_{12})$$

$$= u(x_{11}, x_{12}, m_1 - x_{11}, m_2\ 0\ x_{12})\ v(\psi_{11}, \psi_{12}, \nu_1, \nu_2)$$

$$\psi_{11}^{x_{11}}\ \psi_{12}^{x_{12}}\ ,\ \nu_1^{m_1}\ \nu_2^{m_2}$$

and $P(X_{11} + X_{21} = m_1 \ , \ X_{12} + X_{22} = m_2)$

$$= \sum_{x_{11}} \sum_{x_{12}} u(x_{11},\ x_{12},\ m_1-x_{11},\ m_2-x_{12})\ v(\psi_{11},\ \psi_{12},\ \nu_1,\ \nu_2)$$

$$\psi_{11}^{x_{11}}\ \psi_{12}^{x_{12}}\ .\ \ \nu_1^{m_1}\ \nu_2^{m_2}\ .$$

Hence, from (3.2.4), the conditional likelihood is given by

$g(x_{11},\ x_{12}\ |\ m_1,\ m_2,\ \psi_{11},\ \psi_{12})$

$= P(X_{11} = x_{11},\ X_{12} = x_{12}\ |\ X_{11} + X_{21} = m_1,\ X_{12} + X_{22} = m_2)$

$$= \frac{u(x_{11},\ x_{12},\ m_1 - x_{11},\ m_2 - x_{12})\ \psi_{11}^{x_{11}}\ \psi_{12}^{x_{12}}}{\sum_{i=r_1}^{r_2} \sum_{j=\ell(i)}^{s(i)} u(i,j,\ m_1 - i,\ m_2 - j)\ \psi_{11}^{i}\ \psi_{12}^{j}} \qquad (3.2.5)$$

where $\ell(i) = \max(0,\ m_1 + m_2 - n_2 - i)$

$s(i) = \min(n_1 - i,\ m_2)$

$r_1 = \max(0,\ m_1 - n_2)$ and $r_2 = \min(m_1,\ n_1)$.

Note that the conditional likelihood depends on ψ_{11} and ψ_{12} only.

□

Theorem 3.2.2 : Conditional maximum likelihood estimator (CMLE) for $(\psi_{11},\ \psi_{12})$ is obtained by maximizing the conditional likelihood $g(x_{11},\ x_{12}\ |\ m_1,\ m_2,\ \psi_{11},\ \psi_{12})$ given by (3.2.5).

Proof : Justification for the argument of maximizing the conditional likelihood $g(x_{11},\ x_{12}\ |\ m_1,\ m_2,\ \psi_{11},\ \psi_{12})$ to get CMLE can be given using the definition of ancillarity in presence of nuisance parameter. (Godambe 1980).

We have discussed definition of ancillarity in presence of nuisance parameter when the parameter of interest is real valued. The extension of definition of ancillarity in presence of nuisance parameter when both the parameter of interest and nuisance parameter are vector valued is straightforward. We discuss it in the following.

Let the abstract sample be $\mathcal{X} = \{x\}$ and the abstract parameter space be $\Omega = \{\theta\}$. The density function w.r.t. some measure μ on $\mathcal{X}$ is $p(x, \theta)$. Further, $\theta = (\theta_1, \theta_2)$ where θ_1 is vector-valued parameter of interest and θ_2 denotes vector valued nuisance parameter. $\theta_1 \in \Omega_1$ and $\theta_2 \in \Omega_2$ such that $\Omega_1 \times \Omega_2 = \Omega$.

Definition 3.2.1 : Any vector-valued statistic T satisfying the following two conditions is said to be ancillary statistic w.r.t.

(i) The conditional density f_t of observations given T=t depends on θ only through θ_1 i.e.

$$p(x, \theta) = f_t(x, \theta_1) \; . \; h(t, \theta)$$

where h is the marginal density of T.

(iii) The class of distribution of T corresponding to $\theta_2 \in \Omega_2$ is complete for each fixed $\theta_1 \in \Omega_1$.

For 2x3 contingency table, from (3.2.3) the likelihood is

$$L = u(x_{11}, x_{12}, x_{21}, x_{22}) v(\psi_{11}, \psi_{12}, \nu_1, \nu_2) \psi_{11}^{x_{11}} \psi_{12}^{x_{12}} \nu_1^{x_{11}+x_{21}} \nu_2^{x_{12}+x_{22}}$$

Here (ψ_{11}, ψ_{12}) is the parameter of interest and (ν_1, ν_2) denote the nuisance parameter. $\Omega_1 = \Omega_2 = [0, \infty) \times (0, \infty)$. We observe that the conditional distribution of observations given $X_{11}+X_{21} = m_1$ and $X_{12} + X_{22} = m_2$ depends on ψ_{11} and ψ_{12} only. Further the marginal density of $(X_{11} + X_{21}, X_{12} + X_{22})$

$$P(X_{11}+X_{21} = m_1, X_{12}+X_{22} = m_2) \propto \frac{1}{K(m_1, m_2, \psi_{11}, \psi_{12})} \nu_1^{m_1} \nu_2^{m_2} \quad \forall\, m_1, m_2 \tag{3.2.6}$$

where

$$K(m_1, m_2, \psi_{11}, \psi_{12}) = \frac{1}{\sum_{i=r_1}^{r_2} \sum_{j=\ell(i)}^{s(i)} u(i, j, m_1 - i, m_2 - j)\psi_{11}^{i}\psi_{12}^{j}} \tag{3.2.7}$$

Thus, we can write likelihood (3.2.3) as

$$L = g(x_{11}, x_{12} | m_1, m_2, \psi_{11}, \psi_{12}) \cdot P(X_{11}+X_{21} = m_1, X_{12}+X_{22} = m_2 \mid \psi_{11}, \psi_{12}, \nu_1, \nu_2).$$

Note that for fixed $(\psi_{11}, \psi_{12}) \in \Omega_1$, (3.2.6) is a two-parameter exponential family and hence complete. Thus the definition 3.2.1 is applicable and we conclude that the statistic $(X_{11}+X_{21}, X_{12}+X_{22})$ is ancillary w.r.t. (ψ_{11}, ψ_{12}) and the marginal density of $(X_{11}+X_{21}, X_{12} + X_{22})$ is said to contain no information about (ψ_{11}, ψ_{12}) ignoring (ν_1, ν_2). Hence the inference on (ψ_{11}, ψ_{12}) can be based on the conditional likelihood. The CMLE is then obtained by maximizing the conditional likelihood.

Note that the conditional likelihood given by (3.2.5) is a power-series distribution and belongs to two parameter exponential family. Hence moment estimators and maximum likelihood estimators are same.

3.3 : Simultaneous confidence intervals

Confidence set : Confidence sets are generalizations of the familiar notion of confidence intervals. Suppose that $(y_1, y_2, \ldots, y_n)$ are observations whose distribution is completely determined by the unknown values of the parameter $(\theta_1, \theta_2, \ldots, \theta_m)$ and that $(\phi_1, \phi_2, \ldots, \phi_q)$ are specified functions of the parameters. Denote the three points with coordinates $(y_1, y_2, \ldots, y_n)$, $(\theta_1, \theta_2, \ldots, \theta_m)$ and $(\phi_1, \phi_2, \ldots, \phi_q)$ respectively by $\underline{y}$, $\underline{\theta}$ and $\underline{\phi}$ so that $\underline{\phi}$ is a point determined by the value of $\underline{\theta}$ in q-dimensional ϕ-space. Suppose, for every possible $\underline{y}$ in the sample space a region $R(\underline{y})$ in the q-dimensional ϕ-space is determined. Then if the region $R(\underline{y})$ has the property that the probability that it covers the true point $\underline{\phi}$ is a preassigned constant $(1-\alpha)$ no matter what the unknown true parameter point $\underline{\theta}$ is, we say that $R(\underline{y})$ is a confidence set with confidence coefficient $(1-\alpha)$. Note that a confidence interval is a special case when $q = 1$ and $R(\underline{y})$ is an interval in one-dimensional ϕ-space.

Here we are concerned with simultaneous confidence interval estimation for the parameters ψ_{11} and ψ_{12}. We discuss two asymptotic methods viz. (i) Cornfield's method and (ii) Woolf's method.

Cornfield's method : The unconditional likelihood function can be given as

$$L = \prod_{i=1}^{2} n_i! \prod_{j=1}^{3} \frac{p_{ij}^{x_{ij}}}{x_{ij}!}$$

where $\sum_{j=1}^{3} x_{ij} = n_i$ and $\sum_{j=1}^{3} p_{ij} = 1$, $i = 1,2$.

The conditional probability of observations for the subset of samples in which all the marginal totals are fixed by the conditions $\sum_{i=1}^{2} x_{ij} = m_j$, $j = 1,2$ is given by

$$g(x_{11}, x_{12} | m_1, m_2, \psi_{11}, \psi_{12}) = K(m_1, m_2, \psi_{11}, \psi_{12}) \frac{\prod_{i=1}^{2} n_i!}{\prod_{j=1}^{3} x_{ij}!} \prod_{j=1}^{2} \psi_{1j}^{x_{ij}}$$

(3.3.1)

where $K(m_1, m_2, \psi_{11}, \psi_{12})$ is given in (3.2.7).

We find the limiting distribution for (3.3.1) (which we denote by $f(x_{ij})$) exactly as in chapter 2. We denote the values of x_{ij} at the point of maximum density of (3.3.1) by $\tilde{x}_{ij}$, where, in large samples

$$\psi_{ij} = \frac{\tilde{x}_{ij}(n_2 - \sum_{j=1}^{2} m_j + \sum_{j=1}^{2} \tilde{x}_{1j})}{(m_j - \tilde{x}_{1j})(n_1 - \sum_{j=1}^{2} \tilde{x}_{1j})} \tag{3.3.2}$$

Then, by substituting Stirling's formula and making other approximations of chapter 2, we get

$$- 2 \log \left\{ \frac{f(x_{ij})}{f(\tilde{x}_{ij})} \right\} = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(x_{ij} - \tilde{x}_{ij})^2}{\tilde{x}_{ij}} \qquad (3.3.3)$$

Hence, we conclude that the limiting distribution of (3.3.1) is multivariate normal and in consequence the positive definite quadratic form given by right hand side of (3.3.3) is distributed as chi-square with 2 d.f. In that case, the required confidence region in the $\tilde{x}_{ij}$ is defined by

$$\sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(x_{ij} - \tilde{x}_{ij})^2}{\tilde{x}_{ij}} \leq \chi^2_{\alpha,2} \qquad (3.3.4)$$

where x_{ij} are observed values, $\tilde{x}_{ij}$ are the variables of parameter space and $\chi^2_{\alpha,2}$ is upper α percent point of chisquare distribution with 2 d.f. A corresponding region for ψ_{1j} is obtained from (3.3.2) in view of the fact that ψ_{1j} is monotonic in $x_{\ell k}$ $\forall$ j, ℓ, k. We illustrate the procedure in the following examples.

Example 3.3.1 : The data (Cornfield 1956) in the following table shows the distribution of lung cancer and control patients by smoking status. It has been suggested that cigarette smokers have a greater excess risk of developing lung cancer than do pipe and cigar smokers. We propose to consider from the point of view of interval estimation, what evidence the data in the table contain on this point.

	Smoking status			
	Nonsmoker	pipe+cigar smoker	pipe+cigar+cigarette smoker	
Lung cancer	19	15	484	518
Control	56	68	394	518
Total	75	83	878	1036

Let α = 0.05, then the confidence region given by (3.3.4) becomes

$$\frac{(19 - \tilde{x}_{11})^2}{\tilde{x}_{11}} + \frac{(15 - \tilde{x}_{12})^2}{\tilde{x}_{12}} + \frac{[484 - (518 - \tilde{x}_{11} - \tilde{x}_{12})]^2}{(518 - \tilde{x}_{11} - \tilde{x}_{12})}$$

$$+ \frac{[56 - (75 - \tilde{x}_{11})]^2}{75 - \tilde{x}_{11}} + \frac{[68 - (83 - \tilde{x}_{12})]^2}{(83 - \tilde{x}_{12})}$$

$$+ \frac{[394 - (518 - 75 - 83 + \tilde{x}_{11} + \tilde{x}_{12})]^2}{(360 + \tilde{x}_{11} + \tilde{x}_{12})} \le 5.99 \qquad (3.3.5)$$

At the 95 percent level of confidence therefore we shall reject any hypothesis specifying values of $\tilde{x}_{11}$, $\tilde{x}_{12}$ for which the expression set out above exceeds 5.99 and we accept the hypothesis for which it has a lower value. One such set is obtained by setting $\tilde{x}_{12} = x_{12}$ (= 15) in (3.3.5) and solving the quartic

$$(\tilde{x}_{11} - 19)^2 \left\{ \frac{1}{\tilde{x}_{11}} + \frac{1}{503-\tilde{x}_{11}} + \frac{1}{75-\tilde{x}_{11}} + \frac{1}{375+\tilde{x}_{11}} \right\} = 5.99 \qquad (3.3.6)$$

subject to the condition $0 \le \tilde{x}_{11} \le 75$.

The smallest and the largest root of (3.3.6) are 11.52337 and 28.92046 respectively. Hence the confidence limits for ψ_{11} are given by (0.1428, 0.5347). Thus the risk that any pipe + cigar + cigarette smoker will develop lung cancer relative to the risk that a nonsmoker will develop the lung cancer are 1.8702 to 7.0028.

Similarly, we obtain the confidence limits for ψ_{11}. For this let $\tilde{x}_{11} = x_{11}$ (= 19) in (3.3.5) and we solve the quartic

$$(\tilde{x}_{12} - 15)^2 \left\{ \frac{1}{\tilde{x}_{12}} + \frac{1}{499-\tilde{x}_{12}} + \frac{1}{83-\tilde{x}_{12}} + \frac{1}{379+\tilde{x}_{12}} \right\} = 5.99 \quad (3.3.7)$$

subject to the restriction $0 \le \tilde{x}_{12} \le 83$. The smallest and the largest root are 8.392087 and 24.824650 respectively. Hence the confidence limits for ψ_{12} are given by (0.0888, 0.3634). Thus the risk that any pipe + cigar + cigarette smoker will develop the lung cancer relative to the risk that a pipe + cigar smoker will develop the lung cancer are 2.7518 to 11.2613. Note that the chance that there is any error in the two statements made above giving the confidence limits for ψ_{11} and ψ_{12} is less than 0.05.

Example 3.3.2 : The following data refers to a comparison of two different operations for treating duodenal ulcer patients (Agresti, 1984). The operations correspond to removal of various amounts of the stomach. Operation A is drainage and vagotomy while B is 25% resection and vagotomy. The categories of operation have a natural ordering with A being less severe operation. The variable 'dumping severity' describes the extent of a possible undesirable side effect of the operation. The categories of this variable are also ordered, with the response 'none' representing the most desirable result.

	Dumping severity			
Operation	None	Slight	Moderate	
A	61	28	7	96
B	68	23	13	104
	129	51	20	200

Here again, let $\alpha = 0.05$ and the confidence region is given by

$$\frac{(61 - \tilde{x}_{11})^2}{\tilde{x}_{11}} + \frac{(28 - \tilde{x}_{12})^2}{\tilde{x}_{12}} + \frac{[7 - (96 - \tilde{x}_{11} - \tilde{x}_{12})]^2}{96 - \tilde{x}_{11} - \tilde{x}_{12}}$$

$$+ \frac{[68 - (129 - \tilde{x}_{11})]^2}{129 - \tilde{x}_{11}} + \frac{[23 - (51 - \tilde{x}_{12})]^2}{51 - \tilde{x}_{12}}$$

$$+ \frac{[13 - (104 - 129 - 51 + \tilde{x}_{11} + \tilde{x}_{12})]^2}{\tilde{x}_{11} + \tilde{x}_{12} - 76} \leq 5.99 \qquad (3.3.8)$$

To obtain confidence limits for ψ_{11}, we put $\tilde{x}_{12} = 28$ in (3.3.8) and solve the quartic

$$(\tilde{x}_{11} - 61)^2 \left\{ \frac{1}{\tilde{x}_{11}} + \frac{1}{68 - \tilde{x}_{11}} + \frac{1}{129 - \tilde{x}_{11}} + \frac{1}{\tilde{x}_{11} - 48} \right\} = 5.99$$

(3.3.9)

with the restriction that $48 \leq \tilde{x}_{11} \leq 68$. The smallest and the largest root are 56.00276 and 64.83797 respectively. Hence the confidence limits for ψ_{11} are given by (0.5118, 5.3811). Hence the risk that dumping is moderate instead of none for operation B than that for operation A is 0.1858 to 1.9539.

Similarly we obtain the confidence limits for ψ_{12} by putting $\tilde{x}_{11} = 61$ in (3.3.8) and solving the quartic

$$(\tilde{x}_{12} - 28)^2 \left\{ \frac{1}{\tilde{x}_{12}} + \frac{1}{35-\tilde{x}_{12}} + \frac{1}{51-\tilde{x}_{12}} + \frac{1}{\tilde{x}_{12}-15} \right\} = 5.99 \tag{3.3.10}$$

with the restriction that $15 \le \tilde{x}_{12} \le 35$. The smallest and the largest root for (3.3.10) are given by 23.40872 and 31.67342. Hence the confidence limits for ψ_{12} are given by (0.6155, 8.2142). Thus the risk that dumping is moderate instead of slight for operation B than that for operation A is 0.1217 to 1.6247. Here also the chance that there is any error in the statements regarding confidence limits for ψ_{11} and ψ_{12} is less than 0.05.

Woolf's method :

Here we first obtain the $(1-\alpha)100\%$ simultaneous confidence intervals for $\ln \psi_{1j}$, $j = 1,2$. Simultaneous confidence intervals for ψ_{1j} are then obtained by taking the antilogs of the intervals for $\ln \psi_{1j}$.

We first consider the case where a sample of size N is drawn from the population table where no margins are fixed. Let $\hat{\psi}_{1j}$, $j=1,2$ denote the observed cross product ratio i.e.

$$\hat{\psi}_{1j} = \frac{x_{1j}\, x_{23}}{x_{13}\, x_{2j}} \quad , \quad j = 1,2 .$$

Let $\ln \hat{\psi}_{ij} = d_{ij}$, $i = 1$, $j = 1,2$. Denoting the column vector (d_{ij}) by $\underline{d}$, this vector has two elements and its distribution is bivariate normal with mean vector $\underline{\delta} = (\delta_{ij})$ where

$$\delta_{ij} = \ln \psi_{ij} , \quad i = 1, \quad j = 1,2, \ \ldots$$

and the asymptotic variance - covariance matrix of $\underline{d}$ is given by

$$\Sigma = \begin{bmatrix} \frac{1}{Np_{11}} + \frac{1}{Np_{23}} + \frac{1}{Np_{13}} + \frac{1}{Np_{21}} & \frac{1}{Np_{23}} + \frac{1}{Np_{13}} \\ \frac{1}{Np_{23}} + \frac{1}{Np_{13}} & \frac{1}{Np_{12}} + \frac{1}{Np_{23}} + \frac{1}{Np_{13}} + \frac{1}{Np_{22}} \end{bmatrix}$$

The asymptotic variance - covariance structure of the d_{ij} can be determined from the fact that in an analysis of d_{ij} which are contrasts of $\ln x_{ij}$, we can regard $\ln x_{ij}$ as asymptotically uncorrelated with an asymptotic variance $(Np_{ij})^{-1}$ (Goodman 1964).

The dispersion matrix of $\underline{d}$ can be estimated consistently by a matrix V obtained by replacing the $(Np_{ij})^{-1}$ in the dispersion matrix by $(x_{ij})^{-1}$. The asymptotic distribution as $N \to \infty$ of the statistic

$$W^2 = (\underline{d} - \underline{\delta})' \, V^{-1}(\underline{d} - \underline{\delta}) \qquad (3.3.11)$$

is the chi-square distribution with 2 d.f.

The desired confidence set is constructed using (3.3.11). Hence the probability is $(1-\alpha)$ that the chi-square variable in (3.3.11) is $\le \chi^2_{\alpha,2}$ and therefore the set given by

$$\{ \underline{\delta} \mid (\underline{d} - \underline{\delta})' \, V^{-1} \, (\underline{d} - \underline{\delta}) \le \chi^2_{\alpha,2} \} \qquad (3.3.12)$$

is the desired confidence set with confidence coefficient $(1-\alpha)$.

Inequality in (3.3.12) determines an ellipsoid (for the proof, see appendix) in the two - dimensional space with centre at (d_{11}, d_{12}) and the probability that this random ellipsoid covers the true parameter point $(\delta_{11}, \delta_{12})$ is $(1-\alpha)$ no matter what the values of unknown parameter.

Let us now denote by $\underline{x} = (x_1, x_2)$ any point in two - dimensional space of possible values of $(\delta_{11}, \delta_{12})$. Then the above confidence ellipsoid may be formulated as follows. The probability is $(1-\alpha)$ that $(\delta_{11}, \delta_{12})$ lies inside the ellipsoid

$$(\underline{d} - \underline{x})' \, V^{-1} (\underline{d} - \underline{x}) \leq \chi^2_{\alpha,2} \qquad (3.3.13)$$

But $(\delta_{11}, \delta_{12})$ lies inside the ellipsoid (3.3.13) iff it lies between all pairs of parallel planes of support of ellipsoid. If $\underline{h} = (h_1, h_2)$ is any non-zero vector, it can be shown that (see Appendix) the point $(\delta_{11}, \delta_{12})$ lies between two planes of support of ellipsoid (3.3.13) orthogonal to $\underline{h}$ iff

$$| \, \underline{h}' \, \underline{\delta} - \underline{h}'\underline{d} \, | \leq \sqrt{\chi^2_{\alpha,2}} \, (\underline{h}' \, V\underline{h})^{1/2} \qquad (3.3.14)$$

This result can be used to obtain simultaneous confidence intervals for δ_{ij}, i = 1, j = 1,2. In particular, we obtain the following approximate two sided simultaneous confidence intervals for δ_{ij} at a coverage probability at least $(1-\alpha)$

$$d_{ij} \pm \sqrt{\chi^2_{\alpha,2}} \; S_{d_{ij}} \qquad (3.3.15)$$

where

$$S_{d_{ij}} = \sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{23}} + \frac{1}{n_{i3}} + \frac{1}{n_{2j}}} \quad , \quad i = 1, \; j = 1,2.$$

Example 3.3.3 : Consider the data in example 3.3.1. Let α = 0.05. We obtain from (3.3.15) the following simultaneous confidence intervals for the two odds ratios.

$$0.1412 \le \psi_{11} \le 0.5401 \ ; \quad 0.0876 \le \psi_{12} \le 0.3629.$$

Example 3.3.4 : Consider the data in example 3.3.2. Let α = 0.05. We obtain from (3.3.15) the following simultaneous confidence intervals for the two odds ratios.

$$0.4890 \le \psi_{11} \le 5.6760 \ ; \quad 0.5937 \le \psi_{12} \le 8.6188.$$

3.4 : Testing the hypothesis of independence.

We have discussed in chapter 2, the test procedures for testing the hypothesis that row and column effects are independent in a 2x2 contingency table. For 2x3 contingency table, the hypothesis of independence is equivalent to H_o: ψ_{11}=1, ψ_{12} = 1.

We generalize the test procedures for 2x2 table to 2x3 table. In the following, we discuss generalization of Fisher's exact tst, uncorrected chisquare test and continuity corrected chisquare test.

3.4.1 Fisher's exact test :

Let us denote the observed 2x3 contingency table by $\mathcal{X}$. Denote by $\mathcal{J}$ the reference set of all possible 2x3 contingency tables with the same marginal totals of $\mathcal{X}$. Thus

$$\mathcal{J} = \{ \mathcal{Y} : \mathcal{Y} \text{ is } 2\text{x}3 , \sum_{j=1}^{3} y_{ij} = n_i , \sum_{i=1}^{2} y_{ij} = m_j , \ i = 1,2; \ j = 1,2\}.$$

Under the null hypothesis of row and column independence, probability of observing any $\mathbb{Y} \in \mathcal{J}$ can be expressed as product of multinomial coefficients.

$$\frac{m_1!\ m_2!\ (N-m_1-m_2)!\ n_1!\ n_2!}{N!\ \prod_{j=1}^{3} y_{1j}!\ y_{2j}!}$$

The exact significance level or p-value associated with the observed table $\mathbb{X}$ is defined as the sum of the probabilities of all tables in $\mathcal{J}$ that are no more likely than $\mathbb{X}$. Specifically,

$p = \sum_{\mathbb{Y} \in \mathfrak{J}} P(\mathbb{Y})$ where

$$\mathfrak{J} = \{\mathbb{Y} : Y \in \mathcal{J} \text{ and } P(\mathbb{Y})\} \leq P(\mathbb{X})\} .$$

To calculate this p-value, it is required to generate all the 2x3 contingency tables with fixed marginals. The complexity involved in this procedure severely limits the use of Fisher's exact test. But the computer algorithm developed by Mehta and Patel (1983) calculates p-value for a general rxs table in considerably less amount of time.

Fisher's exact test is not the only procedure yielding an exact p-value. There are several alternative methods. In general, we may define a discrepancy measure $d : \mathcal{J} \to \mathbb{R}$ as a function that assigns a real number to each contingency table in the reference set $\mathcal{J}$. If $\mathbb{X}$ is the observed table, an exact test is defined by

$$p = \sum_{\mathbb{Y} \in \mathbb{F}} P(\mathbb{Y})$$

where $\mathbb{F} = \{\mathbb{Y} : \mathbb{Y} \in \mathcal{J} \text{ and } d(\mathbb{Y}) \geq d(\mathbb{X}) \}$.

For Fisher's exact test $d(\mathcal{X}) = 1/P(\mathcal{X})$. Two other commonly used discrepancy measures are Pearson's X^2 statistic and the likelihood ratio statistic.

3.4.2 Chisquare test (uncorrected) :

Let x_{ij} ($i = 1, 2;\ j = 1,2,3$) denote the observed frequencies in a 2x3 contingency table. Under the hypothesis of independence of row and column effects, let e_{ij} denote expected frequencies ($i = 1,2;\ j = 1,2,3$). Then the chi-square statistic is calculated as

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(x_{ij} - e_{ij})^2}{e_{ij}} .$$

When $N \to \infty$, X^2 has an approximate chisquare distribution with 2 d.f.

3.4.3 Continuity corrected chi-square test :

The continuity corrected chisquare statistic is given by

$$X_c^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(|x_{ij} - e_{ij}| - 1/2)^2}{e_{ij}}$$

where x_{ij} and e_{ij} are as before. When $N \to \infty$ X_c^2 has an approximate chisquare distribution with 2 d.f.

APPENDIX

Definition A.1 : A (solid n dimensional) sphere with centre at the point $\underline{a}$ and radius r is the set of all points $\underline{x}$ satisfying

$$\|\underline{x} - \underline{a}\| \le r \quad \text{or} \quad \|\underline{x} - \underline{a}\|^2 \le r^2 \qquad (A.1)$$

Here $\underline{x} = (x_1, x_2, \ldots, x_n)$ and $\underline{a} = (a_1, a_2, \ldots, a_n)$.

The case where $\underline{a} = \underline{0}$ and $r = 1$ viz.

$$\underline{x}' \, \underline{x} \le 1 \qquad (A.2)$$

is called as unit sphere at the origin.

Definition A.2 : An ellipsoid in canonical position is defined to be the result of applying to the unit sphere at the origin a uniform stretch along each axis.

If the unit sphere at the origin, (A.2), is stretched by factors $c_1, c_2, \ldots, c_n$ along the axes, the resulting ellipsoid in canonical position then satisfies

$$\sum_{i=1}^{n} x_i^2 / c_i^2 \le 1 \qquad (A.3)$$

The numbers (c_i) are called semi-axes of the ellpsoid.

The ellipsoid in canonical position is symmetrical in all co-ordinate planes, since replacing x_i by $-x_i$ does not affect (A.3). We may therefore call origin its centre.

Definition A.3 : An ellipsoid is defined as any point set which may be brought by translation and subsequent orthogonal transformation to coincide with an ellipsoid in canonical position.

A translation of a set by a vector $\underline{a}$ consists in displacing each point $\underline{x}$ of the set so that it goes to $\underline{x} + \underline{a}$. This means that a set defined by an equation or an inequality can be translated by the vector $\underline{a}$ by substituting $\underline{x} - \underline{a}$ for $\underline{x}$ or $x_i - a_i$ for x_i.

Geometrical meaning of an orthogonal transformation is a rotation plus possibly some reflections in coordinate planes.

The centre of the ellipsoid just defined is the point into which the centre of the ellipsoid in canonical position goes.

Result A.1 : If M is a symmetric positive definite matrix, the inequality

$$(\underline{x} - \underline{a})' M(\underline{x} - \underline{a}) \leq 1 \qquad (A.4)$$

defines an ellipsoid with centre at $\underline{a}$.

Proof : First, we translate the set defined by (A.4) by the vector $-\underline{a}$, by substituting $\underline{x}+\underline{a}$ for $\underline{x}$, to get

$$\underline{x}' M\underline{x} \leq 1 \qquad (A.5)$$

We know that for the quadratic form $\underline{x}'M\underline{x}$ in n variables, there exists an orthogonal transformation which reduces (A.5) to the form $\sum_{i=1}^{n} \lambda_i^2 x_i^2 \leq 1$, where the $\{\lambda_i\}$are characteristic roots of M. But this is of the form (A.3) which defines an ellipsoid in the canonical position, with semiaxes $\{c_i = \lambda_i^{-1/2}\}$; the λ_i's are positive because we assumed M to be positive definite. Thus (A.4) defines an ellipsoid with centre at $\underline{a}$.

Definition A.4 : For any vector $\underline{h} \neq \underline{0}$, we define the plane through $\underline{0}$ orthogonal to $\underline{h}$ to be the set of points whose location vectors $\underline{x}$ drawn from $\underline{0}$ are orthogonal to $\underline{h}$.

These vector points $\underline{x}$ are on the plane iff $\underline{h}'\underline{x} = 0$. In general, if $\underline{h} \neq \underline{0}$

$$\underline{h}' (\underline{x} - \underline{x}_0) = 0 \tag{A.6}$$

define a plane through point $\underline{x}_0$ and orthogonal to $\underline{h}$.

Under translation by any vector $\underline{a}$, or a linear transform with any nonsingular matrix P; a plane goes into another plane.

The plane through $\underline{x}_0$ orthogonal to $\underline{h} \neq \underline{0}$ divides the n-dimensional space of point $\underline{x}$ into three parts according as the linear function

$$f(\underline{x}) = \underline{h}'(\underline{x} - \underline{x}_0)$$

is 0, > 0 or < 0.

We say that two points $\underline{x}_{(1)}$ and $\underline{x}_{(2)}$ are on the same side of the plane if

$$f(\underline{x}_{(1)})\, f(\underline{x}_{(2)}) \geq 0 \tag{A.7}$$

The relationship of two points being on the same side of a plane is unaltered by a nonsingular linear transformation and translation.

Definition A.5 : We may define a plane of support to the ellipsoid (A.4) as a plane that has at least one point in common with the ellipsoid and such that the ellipsoid is entirely on one side of the plane.

We give the following result without proof (detail proof is given in Scheffe,H. 1959).

Result A.2 : For $\underline{h} \neq \underline{0}$, the equations of planes of support of ellipsoid (A.4) are given by

$$\underline{h}'(\underline{x} - \underline{a}) = \pm (\underline{h}'M\underline{h})^{1/2} \qquad (A.8)$$

We now get the inequality which defines the strip between the two planes (A.8).

We define the set of points between the planes (A.8) to be the set of those points which are on the same side of both planes as the ellipsoid. This is the same as the set of all points $\underline{x}$ which are on the same side of both the planes, as the centre $\underline{a}$ of the ellipsoid. Using the condition (A.7) to determine the latter set of points, we let

$$f_{\pm}(\underline{x}) = \underline{h}'(\underline{x} - \underline{a}) \pm c_h$$

where $c_h = (\underline{h}'M^{-1}\underline{h})^{1/2}$.

For either plane, both $\underline{x}$ and $\underline{a}$ will be on the same side iff

$$f_{\pm}(\underline{x})\ f_{\pm}(\underline{a}) \geq 0$$

$$\text{or} \quad \pm c_h\ \underline{h}'(\underline{x} - \underline{a}) + c_h^2 \geq 0$$

$$\text{or} \quad \pm \underline{h}'(\underline{x} - \underline{a}) \geq - c_h$$

the two conditions (+ and -) will both be satisfied iff

$$- c_h \leq \underline{h}'(\underline{x} - \underline{a}) \leq + c_h$$

$$\text{or} \quad | \underline{h}'(\underline{x} - \underline{a}) | \leq (\underline{h}'\ M^{-1}\underline{h})^{1/2}$$

This is the desired condition defining the strip between two planes of support orthogonal to $\underline{h}$.

Chapter 4

Two 2 x 2 Contingency Tables : Common Odds Ratio

CHAPTER - 4

TWO 2 X 2 CONTINGENCY TABLES : COMMON ODDS RATIO

4.1 : Introduction

The analysis of several 2x2 contingency tables is required in many contexts. The two principal ones are (i) the comparison of binary response random variables for two treatments over a spectrum of different conditions or populations; and (ii) the comparison of the degree of association among two binary random variables over different populations.

Consider the following examples.

Example 4.1.1 : Use of oral contraceptives is said to be related to myocardial infarction (MI), and suppose one is interested in their relationship. An investigator may want to control for the potential confounding effects of age, cigarette smoking, weight, alcohol consumption, diabetes etc. For simplicity, if we consider only one confounding effect, say, age with 5 classes; then we get 5 2x2 tables. Analysis of these tables may throw light on association between use of oral contraceptives and MI adjusted for the effect of age.

Example 4.1.2 : Suppose, in a sociological problem, one is concerned with the relationship of education and sex to attitudes towards the role of women in society, and each respondent is asked if he or she agreed or disagreed with the statement "Women should take care of running their homes and leave running the country upto men". At the same time, education status of the respondent is also noted. Thus the respondents are classified according to sex (male, female), attitude towards women (agreed with the statement, disagreed with the statement) and the education status (say, t levels, $t \geq 2$).

One might have anticipated that those who are more highly educated would generally tend to disagree with the statement, but what is the form of relationship? To what extent do the males and females respond differently? Does any such difference depend on the amount of education received?

We may form t 2x2 tables according to t levels of education. Analysis of these t 2x2 tables may help to answer some of the questions raised above.

In this chapter, we restrict to analysis of two 2x2 contingency tables. We consider two pairs of mutually independent binomial variates X_{1k} and X_{2k} with corresponding parameters p_{1k} and p_{2k} and sample sizes n_{1k} and n_{2k} respectively $(k = 1,2)$. The sample observations can be written as

Table I

x_{11}	$n_{11} - x_{11}$	n_{11}
x_{21}	$n_{21} - x_{21}$	n_{21}
m_{11}	$N_1 - m_{11}$	N_1

Table II

x_{12}	$n_{12} - x_{12}$	n_{12}
x_{22}	$n_{22} - x_{22}$	n_{22}
m_{12}	$N_2 - m_{12}$	N_2

Let $N = \sum_{k=1}^{2} N_k$

Here x_{1k} is realization of $X_{1k} \sim$ Bi (n_{1k}, p_{1k}) and x_{2k} is realization of $X_{2k} \sim$ Bi (n_{2k}, p_{2k}); k = 1,2. Population odds ratio for k-th table is defined as

$$\psi_k = \frac{p_{1k}\, q_{2k}}{p_{2k}\, q_{1k}} \qquad k = 1,2$$

where $q_{ik} = 1 - p_{ik}$, i = 1,2, k = 1,2.

We assume that odds ratio is constant for the two tables i.e. $\psi = \psi_k$, k = 1,2.

In this chapter, we concentrate on inference about this common odds ratio, ψ. Section 4.2 reviews point estimation methods for ψ while section 4.3 discusses interval estimation for ψ as well as testing procedures for the hypothesis $H_0 : \psi = 1$.

4.2 : Point estimation for the common odds ratio

Here we discuss the methods of point estimation for the common odds ratio, ψ. Unconditional maximum likelihood, conditional maximum likelihood and asymptotic method for point estimation of ψ involve iterative calculations. We discuss these iterative methods. We also discuss some noniterative methods for the point estimation of ψ suggested in the literature.

4.2.1 Unconditional maximum likelihood estimation :

The unconditional likelihood function, ignoring the binomial coefficients is given by

$$L = \prod_{i=1}^{2} \prod_{k=1}^{2} p_{ik}^{x_{ik}} (1 - p_{ik})^{n_{ik} - x_{ik}} \qquad (4.2.1)$$

Since, we have considered the case of a common odds ratio i.e. $\psi_k = \psi$, k = 1,2; we can write $p_{2k} = \dfrac{p_{1k}}{p_{1k} + \psi\, q_{1k}}$ k = 1,2 and hence the likelihood function (4.2.1) can be written as

$$L = \prod_{k=1}^{2} \psi^{n_{2k} - x_{2k}} \, p_{1k}^{x_{1k}+x_{2k}} \, q_{1k}^{n_{1k}+n_{2k} - (x_{1k}+x_{2k})} \, (p_{1k} + \psi q_{1k})^{-n_{2k}}$$

(4.2.2)

Thus, the likelihood is the function of 3 parameters viz. p_{11}, p_{12} and ψ. The unconditional maximum likelihood estimate (UMLE), denoted by $\hat{\psi}_{Un}$ is then obtained by maximizing (4.2.2) w.r.t. p_{11}, p_{12} and ψ simultaneously. Thus UMLE of the common odds ratio ψ is obtained by solving the system of 3 simultaneous equations which involve iterative calculations.

Example 4.2.1 : As mentioned in example 4.1.1, use of oral contraceptives is said to be related to myocardial infarction (MI). Here, we consider two tables corresponding to two age groups viz. 40-44 and 45-49. The tables are as follows (Schlesselman 1982):

Age group → 40 - 44

OC	MI	control	Total
Yes	6	9	15
No	65	362	427
	71	371	442

Age group → 45 - 49

OC	MI	control	Total
Yes	6	5	11
No	93	301	394
	99	306	405

For the data considered above UMLE of ψ is obtained as $\hat{\psi}_{un}$=3.7897.

4.2.2 Conditional maximum likelihood estimation :

Theorem 4.2.1 : If we condition on both the margins of the two 2x2 tables; the conditional distribution of $X_{11} + X_{12}$ given $X_{11}+X_{21}$ and $X_{12} + X_{22}$ is given by

$$P(X_{11} + X_{12} = s \mid X_{11} + X_{21} = m_{11} , X_{12} + X_{22} = m_{12})$$

$$= \frac{C_s \psi^s}{\sum_{j=r_{1.}}^{r_{2.}} C_j \psi^j} , \quad r_{1.} \le s \le r_{2.}$$

where $C_s = \sum_{x_{11}+x_{12}=s} \prod_{k=1}^{2} \binom{n_{1k}}{x_{1k}} \binom{n_{2k}}{x_{2k}}$

$$r_{1.} = \sum_{k=1}^{2} r_{1k} \quad \text{and} \quad r_{2.} = \sum_{k=1}^{2} r_{2k} .$$

Proof : The unconditional likelihood can be written as

$$L = \prod_{i=1}^{2} \prod_{k=1}^{2} \binom{n_{ik}}{x_{ik}} p_{ik}^{x_{ik}} (1 - p_{ik})^{n_{ik} - x_{ik}}$$

$$= \prod_{k=1}^{2} \binom{n_{1k}}{x_{1k}} p_{1k}^{x_{1k}}(1 - p_{1k})^{n_{1k} - x_{1k}} \binom{n_{2k}}{x_{2k}} p_{2k}^{x_{2k}}(1-p_{2k})^{n_{2k} - x_{2k}}$$

$$= \prod_{k=1}^{2} \binom{n_{1k}}{x_{1k}} \binom{n_{2k}}{x_{2k}}(1 - p_{1k})^{n_{1k}}(1-p_{2k})^{n_{2k}}\left[\frac{p_{1k}(1-p_{2k})}{p_{2k}(1-p_{1k})} \right]^{x_{1k}} \left[\frac{p_{2k}}{1 - p_{2k}} \right]^{x_{1k}+x_{2k}}$$

$$= u(x_{11}, x_{21}, x_{12}, x_{22})\ v(\psi, \nu_1, \nu_2)\psi^{x_{11}+x_{12}} \nu_1^{x_{11}+x_{21}} \nu_2^{x_{12}+x_{22}}$$

(4.2.3)

where $\psi = \psi_k = \dfrac{p_{1k}(1-p_{2k})}{p_{2k}(1-p_{1k})}$, $k = 1, 2$

$$\nu_1 = \frac{p_{21}}{(1 - p_{21})} \quad \text{and} \quad \nu_2 = \frac{p_{22}}{(1 - p_{22})} .$$

We observe that (4.2.3) is a three parameter exponential family and $X_{11}+X_{12}$, $X_{11}+X_{21}$ and $X_{12}+X_{22}$ are minimal sufficient statistic.

First, we obtain the conditional distribution of X_{11}, X_{12} given $X_{11} + X_{21} = m_{11}$, $X_{12} + X_{22} = m_{12}$

$$P(X_{11} = x_{11},\ X_{12} = x_{12} \mid X_{11} + X_{21} = m_{11},\ X_{12} + X_{22} = m_{12})$$

$$= \frac{P(X_{11} = x_{11},\ X_{21} = m_{11} - x_{11},\ X_{12} = x_{12},\ X_{22} = m_{12} - x_{12})}{P(X_{11} + X_{21} = m_{11},\ X_{12} + X_{22} = m_{12})} \qquad (4.2.4)$$

From (4.2.3)

$$P(X_{11} = x_{11},\ X_{21} = m_{11} - x_{11},\ X_{12} = x_{12},\ X_{22} = m_{12} - x_{12})$$

$$= u(x_{11},\ m_{11}-x_{11},\ x_{12},\ m_{12} - x_{12})\ v(\psi,\ \nu_1,\ \nu_2)\psi^{x_{11}+x_{12}} \nu_1^{m_{11}} \nu_2^{m_{12}}$$

(4.2.5)

and

$$P(X_{11} + X_{21} = m_{11},\ X_{12} + X_{22} = m_{12})$$

$$= \sum_{x_{11}=r_{11}}^{r_{21}} \sum_{x_{12}=r_{12}}^{r_{22}}$$

$$u(x_{11},\ m_{11}-x_{11},\ x_{12},\ m_{12} - x_{12})\ v(\psi,\ \nu_1,\ \nu_2)\psi^{x_{11}+x_{12}}\nu_1^{m_{11}}\nu_2^{m_{12}}$$

(4.2.6)

where $r_{1k} = \max(0,\ m_{1k} - n_{2k})$ and $r_{2k} = \min(m_{1k},\ n_{1k})$, k=1,2.

Hence, from (4.2.4), (4.2.5) and (4.2.6)

$$P(X_{11} = x_{11},\ X_{12} = x_{12} \mid X_{11} + X_{21} = m_{11},\ X_{12} + X_{22} = m_{12})$$

$$= \frac{u(x_{11},\ m_{11}-x_{11},\ x_{12},\ m_{12}-x_{12})\ \psi^{x_{11}+x_{12}}}{\sum_{i=r_{11}}^{r_{21}} \sum_{j=r_{12}}^{r_{22}} u(i,\ m_{11}-i,\ j,\ m_{12}-j)\ \psi^{i+j}} .$$

Thus, we can write the conditional likelihood as

$$P(X_{11} = x_{11},\ X_{12} = x_{12} \mid X_{11} + X_{21} = m_{11},\ X_{12} + X_{22} = m_{12})$$

$$= \prod_{k=1}^{2} g(x_{1k} \mid m_{1k}\ ,\ \psi) \qquad (4.2.7)$$

where

$$g(x_{1k} | m_{1k},\ \psi) = \frac{\binom{n_{1k}}{x_{1k}} \binom{n_{2k}}{m_{1k}-x_{1k}} \psi^{x_{1k}}}{\sum_{j=r_{1k}}^{r_{2k}} \binom{n_{1k}}{j} \binom{n_{2k}}{m_{1k}-j} \psi^{j}} \qquad (4.2.8)$$

Now, the conditional distribution of $X_{1.} = X_{11}+X_{12}$ given $X_{11}+X_{21}$, $X_{12}+X_{22}$ is obtained by using (4.2.7)

$$P(X_{1.} = s \mid X_{11} + X_{21} = m_{11},\ X_{12} + X_{22} = m_{12})$$

$$= h(s \mid m_{11}, m_{12}, \psi)$$

$$= \sum_{x_{11}+x_{12}=s} \prod_{k=1}^{2} g(x_{1k} \mid m_{1k}, \psi)$$

$$= \sum_{x_{11}+x_{12}=s} \prod_{k=1}^{2} \frac{\binom{n_{1k}}{x_{1k}} \binom{n_{2k}}{m_{1k}-x_{1k}} \psi^{x_{1k}}}{\sum_{j=r_{1k}}^{r_{2k}} \binom{n_{1k}}{j} \binom{n_{2k}}{m_{1k}-j} \psi^{j}}$$

$$= \frac{C_s\ \psi^s}{\sum_{j=r_{1.}}^{r_{2.}} C_j\ \psi^j} \qquad r_{1.} \le s \le r_{2.} \qquad (4.2.9)$$

where $C_s = \sum_{x_{11}+x_{12}=s} \prod_{k=1}^{2} \binom{n_{1k}}{x_{1k}} \binom{n_{2k}}{x_{2k}}$

$$r_{1.} = \sum_{k=1}^{2} r_{1k}, \qquad r_{2.} = \sum_{k=1}^{2} r_{2k} .$$

Note that (4.2.9) depends only on ψ, the parameter of interest.

□

Theorem 4.2.2 : CMLE for the common odss ratio ψ (denoted by $\hat{\psi}_{Cn}$) is obtained by solving the equation

$$x_{11} + x_{12} = E\{X_{11} + X_{12} \mid X_{11} + X_{21},\ X_{12} + X_{22},\ \psi\}$$

Proof : First, we claim that inference on ψ is based on conditional distribution of $X_{11}+X_{12}$ given $X_{11} + X_{21}$ and $X_{12}+ X_{22}$. There are two lines of argument that lead to above statement. In both the approaches we need to consider methods involving sufficient statistic.

First, we may look for a distribution that depends on the value of ψ, but not on the nuisance parameter (ν_1, ν_2). For any fixed and known value of ψ, a sufficient statistic for remaining parameters is $X_{11} + X_{21}$ and $X_{12} + X_{22}$. Therefore, the distribution of the observations and hence also of $X_{1.} = X_{11} + X_{12}$ given $X_{11} + X_{21}$ and $X_{12} + X_{22}$ depends on ψ only. This is the standard Neyman - Pearson approach to elimination of nuisance parameter (Cox 1970).

The second approach is that if we were given only the vlaues of $X_{11} + X_{21}$ and $X_{12} + X_{22}$, no conclusion could be drawn about ψ, and hence ancillary for ψ ignoring the nuisance parameter (Fisher 1956). This approach can be formalized using the definition of ancillarity in presence of nuisance parameter (Godambe 1980) discussed in chapter 2. Here ψ is the parameter of interest with $\Omega_1 = [0, \infty)$ and (ν_1, ν_2) represent nuisance parameter with $\Omega_2 = [0, \infty) \times [0, \infty)$. Further, we can write the likelihood L, as

$$L = P(X_{11} = x_{11},\ X_{12} = x_{12} \mid X_{11} + X_{21} = m_{11},\ X_{12} + X_{22} = m_{12}, \psi)$$
$$\times\ P(X_{11} + X_{21} = m_{11},\ X_{12} + X_{22} = m_{12},\ \psi,\ \nu_1,\ \nu_2).$$

Thus, conditional distribution of observations given $X_{11} + X_{21}$ and $X_{12} + X_{22}$ depends on ψ, only.

Further, from (4.2.6), we obseve that for fixed ψ, joint distribution of $X_{11} + X_{21}$ and $X_{12} + X_{22}$ belongs to two parameter exponential family with parameters (ν_1, ν_2). Hence it is complete. Thus, the definition 2.2.1 is applicable and we conclude that the statistic $(X_{11} + X_{21}, X_{12} + X_{22})$ is ancillary w.r.t. ψ and marginal density of $(X_{11} + X_{21}, X_{12} + X_{22})$ is said to contain no information about ψ ignoring (ν_1, ν_2). Thus, infernce on ψ in the conditional set up is based on the conditional distribution of $X_{11} + X_{12}$ given $X_{11} + X_{21}$ and $X_{12}+X_{22}$. From (4.2.9)

$$h(s \mid m_{11}, m_{12}, \psi) = \frac{C_s\ \psi^s}{\sum_{j=r_{1.}}^{r_{2.}} C_j\ \psi^j} \qquad r_{1.} \le s \le r_{2.}$$

CMLE of ψ is obtained by maximizing $h(s|m_{11}, m_{12}, \psi)$ w.r.t. ψ i.e. by solving the equation

$$s = \frac{\sum_{j=r_{1.}}^{r_{2.}} j\ C_j \psi^j}{\sum_{j=r_{1.}}^{r_{2.}} C_j\ \psi^j} \qquad (4.2.10)$$

and hence CMLE is obtained by solution to

$$x_{11} + x_{12} = E(X_{11} + X_{12} \mid X_{11} + X_{21}, X_{12} + X_{22}, \psi) .$$

Note that (4.2.9) is a power series distribution. Hence moment estimator and m.l.e. are same.

□

The computation of the coefficients C_j in (4.2.9) is the most difficult task in the application of conditional theory. The computing time utilized by program of Thomas (1975) was large if number of tables are large. But this difficulty has been solved by algorithms developed by Mehta et al. (1985) and Vollset et al. (1991).

Example 4.2.2 : Consider the two tables with marginal totals fixed by

(i) $n_{11} = 15$ $n_{21} = 10$ $m_{11} = 9$

(ii) $n_{12} = 17$ $n_{22} = 18$ $m_{12} = 15$.

Any two tables with the above marginal totals can be represented by

x_{11}	$15 - x_{11}$	15
x_{21}	$10 - x_{21}$	10
9	16	25

x_{12}	$17 - x_{12}$	17
x_{22}	$18 - x_{22}$	18
15	20	35

Note that $X_{1.} = X_{11} + X_{12}$ takes values from 0 to 24. In the following, we give conditional point estimate of the common odds ratio for tables with observed value of $X_{1.}$ equal to 8, 12 and 16.

Table 4.2.1 : Conditional maximum likelihood estimate

Observed value of $X_{1.} = X_{11}+X_{11}$	$\hat{\psi}_{Cn}$
8	0.2667
12	0.8287
16	2.5573

Example 4.2.3 : Consider the data in example 4.2.1. Here

(i) $n_{11} = 15,\ n_{21} = 427,\ m_{11} = 71$

(ii) $n_{12} = 11,\ n_{22} = 394,\ m_{12} = 99.$

Conditional point estimate for the common odds ratio is given by

$$\hat{\psi}_{Cn} = 3.7761.$$

4.2.3 Estimation by asymptotic method :

The exact conditional noncentral distribution of X_{1k}, k=1,2 is given as before by

$$\prod_{k=1}^{2} g(x_{1k} \mid m_{1k}, \psi)$$

where $g(x_{1k} \mid m_{1k}, \psi)$ is given by (4.2.8). In Chapter 2, we have shown that X_{1k} is asymptotically normally distributed with asymptotic mean and variance given by

$$\hat{E}(X_{1k} \mid m_{1k}, \psi) = \tilde{x}_{1k}$$

and $$\hat{V}(X_{1k} \mid m_{1k}, \psi) = \left\{ \frac{1}{\tilde{x}_{1k}} + \frac{1}{n_{1k}-\tilde{x}_{1k}} + \frac{1}{m_{1k}-\tilde{x}_{1k}} + \frac{1}{n_{2k}-m_{1k}+\tilde{x}_{1k}} \right\}^{-1}$$

where $\tilde{x}_{1k}$ is the appropriate solution to the quadratic equation

$$\frac{\tilde{x}_{1k}(n_{2k} - m_{1k} + \tilde{x}_{1k})}{(m_{1k} - \tilde{x}_{1k})(n_{1k} - \tilde{x}_{1k})} = \psi \quad , k = 1, 2 \qquad (4.2.11)$$

The appropriate solution is the one yielding positive numbers for all the factors and divisors on the left hand side of (4.2.11) i.e. all the estimated entries in the 2x2 tables. Hence $X_{1.} = X_{11} + X_{12}$ is asymptotically normally distributed with mean and variance given by

$$\hat{E}(X_{1.} \mid m_{11}, m_{12}, \psi) = \sum_{k=1}^{2} \hat{E}(X_{1k} \mid m_{1k}, \psi) \qquad (4.2.12)$$

and

$$\hat{V}(X_{1.} \mid m_{11}, m_{12}, \psi) = \sum_{k=1}^{2} \hat{V}(X_{1k} \mid m_{1k}, \psi) \qquad (4.2.13)$$

The estimator of ψ which maximizes the asymptotic conditional likelihood is then obtained by (Gart (1970)

$$x_{11} + x_{12} = \hat{E}(X_{1.} \mid m_{11}, m_{12}, \psi) \qquad (4.2.14)$$

which involves trial and error solution of two quadratic equations. We denote this estimator by $\hat{\psi}_a$. Program of Thomas (1975) performs the calculations.

Example 4.2.4 : Consider the two 2x2 tables considered in example 4.2.1, $\hat{\psi}_a$ for this data is given by $\hat{\psi}_a = 3.7866$.

4.2.4 Noniterative Methods :

All the three procedures described above for estimating the common odds ratio require iterative calculations. There are some methods which require no iterative calculations. Here, we discuss two procedures viz. Woolf's procedure and Mantel - Haenszel procedure.

Woolf's Procedure

Woolf (1955) proposed an adjusted estimate of the common odds ratio ψ based on the log odds ratio. Let $\hat{\psi}_k$ denote unconditional point estimate of the odds ratio for the k-th table. An estimate of variance of ln $\hat{\psi}_k$ is given by

$$\hat{v}_k = \frac{1}{x_{1k}} + \frac{1}{n_{1k} - x_{1k}} + \frac{1}{x_{2k}} + \frac{1}{n_{2k} - x_{2k}} \qquad (4.2.15)$$

Taking a weighted average of the log odds ratios with weights being the reciprocals of the estimated variances, $\beta = \ln \psi$ may be estimated by the quantity

$$\hat{\beta}_w = \frac{\sum_{k=1}^{2} (\hat{v}_k)^{-1} \ln \hat{\psi}_k}{\sum_{k=1}^{2} (\hat{v}_k)^{-1}}$$

Thus, the estimate of ψ is given by

$$\hat{\psi}_w = \exp\left\{ \frac{\sum_{k=1}^{2} (\hat{v}_k)^{-1} \ln \hat{\psi}_k}{\sum_{k=1}^{2} (\hat{v}_k)^{-1}} \right\} \qquad (4.2.16)$$

Example 4.2.5 : $\hat{\psi}_w$ for the data considered in example 4.2.1 is given by

$$\hat{\psi}_W = 3.7867 .$$

Mantel - Haenszel Procedure

Mantel and Haenszel (1959) proposed on heuristic grounds a highly efficient method for estimating a summary odds ratio from a series of 2x2 table.

The Mantel - Haenszel estimate (denoted by $\hat{\psi}_{M-H}$) is calculated as

$$\hat{\psi}_{M-H} = \frac{\sum_{k=1}^{2} x_{1k}(n_{2k} - x_{2k})/N_k}{\sum_{k=1}^{2} x_{2k}(n_{1k} - x_{1k})/N_k} \qquad (4.2.17)$$

Note that $\hat{\psi}_{M-H}$ is defined only if $x_{2k} > 0$ and $n_{1k} - x_{1k} > 0 \; \forall \; k$.

The Mantel - Haenszel estimate can be regarded as a weighted average of subgroup (subtable) odds ratio provided none of the values of $n_{1k} - x_{1k}$ and x_{2k} are equal to zero. The odds ratio estimate for the k-th table is given by

$$\hat{\psi}_k = \frac{x_{1k}(n_{2k} - x_{2k})}{x_{2k}(n_{1k} - x_{1k})} .$$

Using the weights $w_k = x_{2k}(n_{1k} - x_{1k})/N_k$, one can write

$$\hat{\psi}_{M-H} = \frac{\sum_{k=1}^{2} w_k \hat{\psi}_k}{\sum_{k=1}^{2} w_k} \qquad (4.2.18)$$

Example 4.2.6 : $\hat{\psi}_{M-H}$ for the data considered in example 4.2.1 is given by

$$\hat{\psi}_{M-H} = 3.7823 .$$

4.2.5 Jackknife estimators :

The estimators that are discussed here are based on the jackknife principle originally proposed by Quenouille (1949). In general, for an estimator $\hat{\theta}$ of a real valued parameter θ, the pseudovalue J_i is based on $\hat{\theta}$ itself and $\hat{\theta}_i$ which is just the estimate obtained if the i-th observation out of N observations, say, is deleted; that is

$$J_i = J_i^N(\theta) = N\hat{\theta} - (N-1)\hat{\theta}_i \ , \quad i = 1, 2, \ldots, N$$

and the jackknife estimator is defined as the arithmetic mean of the N pseudovalues.

$$J = \frac{1}{N} \sum_{i=1}^{N} J_i$$

with the jackknife variance (Tukey 1958) estimate

$$\hat{V}(J) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (J_i - J)^2$$

Here we discuss two jackknife estimators for the common odds ratio, ψ, suggested by Breslow and Liang (1982) and Pigeot (1991).

Breslow and Liang jackknife the logarithm of the Mantel - Haenszel estimator. They calculate jackknife estimator on the basis of pseudovalues that are obtained by omitting one complete table. If there are t tables, the t pseudovalues are

$$J_k = t \log \hat{\psi}_{M-H} - (t-1) \log \hat{\psi}_{M-H,k}, \quad k = 1,2,\ldots, t \qquad (4.2.19)$$

with

$$\hat{\psi}_{M-H,k} = \frac{\sum_{\substack{j=1 \\ j\neq k}}^{t} x_{1j}(n_{2j} - x_{2j})/N_j}{\sum_{\substack{j=1 \\ j\neq k}}^{t} x_{2j}(n_{1j} - x_{1j})/N_j}$$

and the jackknife estimator

$$\hat{\psi}_{JBL} = \frac{1}{t} \sum_{k=1}^{t} J_k \qquad (4.2.20)$$

with jackknife variance estimate

$$\hat{V}(\hat{\psi}_{JBL}) = \frac{1}{t(t-1)} \sum_{k=1}^{t} (J_k - \hat{\psi}_{JBL})^2 \qquad (4.2.21)$$

This method is appropriate, if number of tables, t, is large. When the number $N = \sum_{k=1}^{t} N_k$ of all observations tends to infinity while t is fixed, another approach of omitting each single observation is proposed by Pigeot (1991). We describe this procedure for t = 2.

When each single observation is omitted, we get the following N pseudovalues.

$$J_{\alpha,k} = N \hat{\psi}_{M-H} - (N-1) \hat{\psi}_{\alpha,k} \qquad (4.2.22)$$

with

$$\hat{\psi}_{a,k} = \frac{\dfrac{(x_{1k}-1)(n_{2k}-x_{2k})}{(N_k-1)} + \sum_{\substack{j=1\\j\neq k}}^{2} \dfrac{x_{1j}(n_{2j} - x_{2j})}{N_j}}{\dfrac{x_{2k}(n_{1k}-x_{1k})}{(N_k-1)} + \sum_{\substack{j=1\\j\neq k}}^{2} \dfrac{x_{2j}(n_{1j} - x_{1j})}{N_j}} \qquad k = 1,\ 2$$

(4.2.23)

$J_{b,k}$, $J_{c,k}$, $J_{d,k}$ can be obtained in a similar way (a,b,c,d denote the cells in each contingency table). Each pseudovalue exists exactly x_{1k}, $n_{1k} - x_{1k}$, x_{2k} and $n_{2k} - x_{2k}$ times respectively. Thus, the jackknife estimator is given by

$$\hat{\psi}_{JP} = \frac{1}{N} \sum_{k=1}^{2} \{x_{1k} J_{a,k} + (n_{1k}-x_{1k})J_{b,k} + x_{2k}J_{c,k} + (n_{2k}-x_{2k})J_{d,k}\}$$

$$= N\, \hat{\psi}_{M-H} - \frac{(N-1)}{N} \sum_{k=1}^{2} \{x_{1k}\, \hat{\psi}_{a,k} + (n_{1k}-x_{1k})\hat{\psi}_{b,k} + x_{2k}\hat{\psi}_{c,k} + (n_{2k}-x_{2k})\hat{\psi}_{d,k}\}$$

(4.2.24)

The jackknife variance estimate is given by

$$\hat{V}(\hat{\psi}_{JP}) = \frac{1}{N(N-1)} \sum_{k=1}^{2} \{x_{1k}(\hat{\psi}_{JP} - J_{a,k})^2 + (n_{1k}-x_{1k})(\hat{\psi}_{JP} - J_{b,k})^2$$

$$+ x_{2k}(\hat{\psi}_{JP} - J_{c,k})^2 + (n_{2k} - x_{2k})(\hat{\psi}_{JP} - J_{d,k})^2\}$$

$$= - \frac{1}{(N-1)} \hat{\psi}^2_{JP} + \frac{1}{N(N-1)} \sum_{k=1}^{2} \{x_{1k}\, J^2_{a,k} + (n_{1k}-x_{1k})J^2_{b,k} + x_{2k}\, J^2_{c,k}$$

$$+ (n_{2k} - x_{2k})\, J^2_{d,k}\} \qquad (4.2.25)$$

Now,

$$\frac{1}{N(N-1)}\sum_{k=1}^{2}\{x_{1k}J^2_{a,k} + (n_{1k}-x_{1k})J^2_{b,k} + x_{2k}\ J^2_{c,k} + (n_{2k}-x_{2k})\ J^2_{d,k}\}$$

$$= \frac{1}{N(N-1)}\sum_{k=1}^{2}\{x_{1k}[N\,\hat{\psi}_{M-H} - (N-1)\,\hat{\psi}_{a,k}]^2$$

$$+ (n_{1k} - x_{1k})\,[N\,\hat{\psi}_{M-H} - (N-1)\hat{\psi}_{b,k}]^2$$

$$+ x_{2k}[N\,\hat{\psi}_{M-H} - (N-1)\hat{\psi}_{c,k}]^2$$

$$+ (n_{2k} - x_{2k})[N\,\hat{\psi}_{M-H} - (N-1)\,\hat{\psi}_{d,k}]^2\}$$

$$= \frac{N^2}{(N-1)}\,\hat{\psi}^2_{M-H} + \frac{(N-1)}{N}\sum_{k=1}^{2}\{x_{1k}\,\hat{\psi}^2_{a,k} + (n_{1k}-x_{1k})\hat{\psi}^2_{b,k}$$

$$+ x_{2k}\,\hat{\psi}^2_{c,k} + (n_{2k} - x_{2k})\,\hat{\psi}^2_{d,k}\}$$

$$- 2\hat{\psi}_{M-H}\,\{\frac{N}{N-1}\,(N\,\hat{\psi}_{M-H} - \hat{\psi}_{JP})\}$$

$$= -\frac{N^2}{(N-1)}\,\hat{\psi}^2_{M-H} + \frac{2N}{N-1}\,\hat{\psi}_{M-H}\,\hat{\psi}_{JP}$$

$$+ \frac{(N-1)}{N}\sum_{k=1}^{2}\{x_{1k}\,\hat{\psi}^2_{a,k} + (n_{1k}-x_{1k})\hat{\psi}^2_{b,k}$$

$$+ x_{2k}\,\hat{\psi}^2_{c,k} + (n_{2k} - x_{2k})\,\hat{\psi}^2_{d,k}\} \qquad (4.2.26)$$

Substituting (4.2.26) in (4.2.25), we get

$$\hat{V}(\hat{\psi}_{JP}) = \frac{(N-1)}{N}\sum_{k=1}^{2}\{x_{1k}\ \hat{\psi}^2_{a,k} + (n_{1k}-x_{1k})\hat{\psi}^2_{b,k}$$

$$+ x_{2k}\,\hat{\psi}^2_{c,k} + (n_{2k} - x_{2k})\,\hat{\psi}^2_{d,k}\}$$

$$- \frac{1}{(N-1)}(\,\hat{\psi}_{JP} - N\,\hat{\psi}_{M-H})^2 \qquad (4.2.27)$$

Example 4.2.7 :

Consider again the two 2x2 tables in example 4.2.1. Since t=2, we do not apply method of Breslow and Liang. By applying the method suggested by Pigeot, jackknife estimate of the common odds ratio is given by $\hat{\psi}_{JP}$ = 3.4575 with $\hat{V}(\hat{\psi}_{JP})$ = 2.6000.

4.2.6 Discussion :

The problem of zero cells also arises in multiple 2x2 tables, and we get zero or infinite estimates for the odds ratio or log odds ratio. A possible device for handling infinite estimates is to avoid them by adding some positive constant to cell entry values or at least to the summary table cell entries.

Occurence of zero cells is primarily a small sample problem and it should not be considered when dealing with asymptotic case. In effect, if zeros or infinite estimates are other than a fairly rare problem, that is an indicator that the estimator in question is not appropriate with that sample size. Then forcing application of an estimator by adding constants would be misleading (Hauck 1986, Mantel 1986).

The strategy of adding constant to cell values may be viewed as fine tuning of an estimator (Hauck 1986). It is obsedrved that odds ratio (or log-odds ratio) estimators are biased to some degree in finite samples. Hauck, Anderson and Leathy (1982) noted that adding positive constants to all the cells would shrink the estimators towards one or zero depending on the scale;

thus reducing the tendency of some of these estimators to overestimate the odds ratio and its logarithm.

4.3 : Interval estimation and testing the hypothesis of independence

In this section we describe various methods to test the hypothesis $H_0 : \psi = 1$. We also describe procedures to obtain confidence limits for the common odds ratio, ψ.

First method is the extension of Fisher's exact treatment for single 2x2 table. Cornfield's asymptotic method can also be extended to 2x2x2 table. These two methods involve iterative calculations.

The noniterative procedures to calculate confidence limits for the common odds ratio include Woolf's (1955) method and Mantel - Haenszel procedure (1959).

4.3.1 Extension of Fisher's exact treatment :

As shown earlier, conditional distribution of $X_{1.} = X_{11} + X_{12}$ given $X_{11} + X_{21}$ and $X_{12} + X_{22}$ is given by

$$h(s \mid m_{11}, m_{12}, \psi) = \frac{C_s \, \psi^s}{\sum_{j=r_{11}}^{r_{2.}} C_j \psi^j} \quad , \quad r_{1.} \le s \le r_{2.} \qquad (4.3.1)$$

where

$$C_s = \sum_{x_{11}+x_{12}=s} \prod_{k=1}^{2} \binom{n_{1k}}{x_{1k}} \binom{n_{2k}}{m_{1k}-x_{1k}}$$

$$r_{1.} = \sum_{k=1}^{2} r_{1k} \quad \text{and} \quad r_{2.} = \sum_{k=1}^{2} r_{2k} \ .$$

Calculation of p-values : The UMP test for $\psi = \psi_o$ is based on the above conditional distribution, it has critical regions of the form $X_{1.} \geq s$ ($X_{1.} \leq s$) for alternatives of the form $\psi > \psi_o$ ($\psi < \psi_o$).

Hence corresponding to an observed value s; the one sided significance level against the alternative $\psi > \psi_o$ is

$$p = \sum_{j=s}^{r_{2.}} h(j \mid m_{11}, m_{12}, \psi_o) \qquad (4.3.2)$$

and against the alternative $\psi < \psi_o$

$$p = \sum_{j=r_{1.}}^{s} h(j \mid m_{11}, m_{12}, \psi_o) \qquad (4.3.3)$$

Note that for particular case of $\psi_o = 1$, $h(j \mid m_{11}, m_{12}, 1)$ is the product of two hypergeometric probabilities.

Example 4.3.1 : Consider the data in example 4.2.1.

Age group → 40 - 44

OC	MI	control	Total
Yes	6	9	15
No	65	362	427
	71	371	442

Age group → 45 - 49

OC	MI	control	Total
Yes	6	5	11
No	93	301	394
	99	306	405

Suppose we want to test the hypothesis $H_0 : \psi = 1$ Vs $H_1 : \psi > 1$. p-value in this case is given by 0.0015.

Confidence limits : As stated above, the UMP tst for $\psi = \psi_o$ is based on $h(s \mid m_{11}, m_{12}, \psi)$ and it has critical region of the form $X_{1.} \geq s$ ($X_{1.} \leq s$) for alternatives of the form $\psi > \psi_o$ ($\psi < \psi_o$). An exact confidence interval may be constructed by inverting this test (Mehta et al. 1985). Specifically, an exact 100(1-α)% confidence interval for ψ is given by $\{\psi_L(s), \psi_U(s)\}$ where $\psi_L(s)$ is such that

$$\psi_L(s) = 0 \qquad \text{if } s = r_{1.}$$

$$\sum_{j=s}^{r_{2.}} h(j \mid m_{11}, m_{12}, \psi_L(s)) = \frac{\alpha}{2} \qquad \text{if } r_{1.} < s \leq r_{2.} \tag{4.3.4}$$

and $\psi_U(s)$ is such that

$$\sum_{j=r_{1.}}^{s} h(j \mid m_{11}, m_{12}, \psi_U(s)) = \frac{\alpha}{2} \qquad \text{if } r_{1.} \leq s < r_{2.}$$

$$\psi_U(s) = \infty \qquad \text{if } s = r_{2.} \tag{4.3.5}$$

The probability that this interval fails to contain ψ is $P(\psi_U(s) < \psi) + P(\psi_L(s) > \psi)$. Each of these exclusion probability can not exceed $\alpha/2$. This can be shown as in a single 2x2 table case. Hence we omit the proof.

Thus $\{\psi_L(s), \psi_U(s)\}$ is a conservative 100(1-α)% confidence interval for the common odds ratio. Due to discreteness of the conditional distribution of $X_{1.}$, one can not guarantee coverage of ψ exactly 100(1-α)% of the time. The above procedure is exact in the sense that it is based on exact distribution theory.

The various computer programs (Thomas 1975, Mehta et al. 1985, Vollset et al. 1991) calculate the confidence limits.

Example 4.3.2 : For the data considered in example 4.2.1, 95% confidence limits are given by

$\psi_L = 1.5514 \qquad \psi_U = 9.0580$.

Example 4.3.3 : For the data considered in example 4.2.1, we give 95% confidence limits in the following.

Table 4.3.1 : 95% exact confidence limits

Observed value of $X_{1.} = X_{11} + X_{12}$	95% exact limits ψ_L	ψ_U
8	0.0755	0.8745
12	0.2597	2.6379
16	0.784	8.9805

4.3.2 Asymptotic Method :

In section 4.2; we have shown that $X_{1.} = X_{11} + X_{12}$ is asymptotically normal with mean $\hat{E}(X_{1.} \mid m_{11}, m_{12}, \psi)$ and variance $\hat{V}(X_{1.} \mid m_{11}, m_{12}, \psi)$ given by (4.2.12) and (4.2.13) respectively. This can be used to test the hypothesis $H_0 : \psi = \psi_0$.

Asymptotic Confidence Limits : $(1 - \alpha)100\%$ confidence limits for the common odds ratio ψ, using the asymptotic distribution are obtained as follows (Gart 1970).

The upper limit ψ_U is obtained by finding the value of ψ for which

$$\frac{(x_{11}+x_{12}) - \hat{E}(X_{1.} \mid m_{11}, m_{12}, \psi) + 1/2}{\sqrt{\hat{V}(X_{1.} \mid m_{11}, m_{12}, \psi)}} = -z_{\alpha/2} \qquad (4.3.6)$$

and the lower limit ψ_L by solving

$$\frac{(x_{11}+x_{12}) - \hat{E}(X_{1.} \mid m_{11}, m_{12}, \psi) - 1/2}{\sqrt{\hat{V}(X_{1.} \mid m_{11}, m_{12}, \psi)}} = z_{\alpha/2} \qquad (4.3.7)$$

where $z_{\alpha/2}$ is upper $\alpha/2$ percentage point of N(0,1).

Example 4.3.4 : Consider the data in example 4.2.1. 95% confidence limits for the common odds ratio using the asymptotic method are given by

$$\psi_L = 2.6343 \qquad \psi_U = 5.3830$$

4.3.3 Noniterative procedures :

Woolf's procedure

As shown in 4.2.4, $\beta = \ln \psi$ is estimated by the quantity

$$\hat{\beta}_w = \frac{\sum_{k=2}^{2} (\hat{v}_k)^{-1} \ln \hat{\psi}_k}{\sum_{k=1}^{2} (\hat{v}_k)^{-1}}$$

where $\hat{v}_k$ is given by (4.2.15).

Tests of significance and confidence interval assume that $\hat{\beta}_w$ is normally distributed.

Let $u = \sum_{k=1}^{2} (\hat{v}_k)^{-1}$,

Then $\hat{var}(\hat{\beta}_w) = 1/u$.

Thus, an approximate chisquare test of the null hypothesis $H_o : \psi = \psi_o$ vs $H_1 : \psi \neq \psi_o$ may be based on the statistic

$$X_v^2 = \frac{(\hat{\beta}_v - \ln \psi_o)^2}{1/u} \tag{4.3.8}$$

Under the null hypothesis, X_v^2 has an approximate chisquare distribution with one degree of freedom. In particular, for $\psi_o=1$, (4.3.8) reduces to

$$X_v^2 = u(\hat{\beta}_v)^2 \tag{4.3.9}$$

Further 100(1-α)% confidence limits are given by

$$\psi_L = \hat{\psi}_v \exp\{- z_{\alpha/2}/\sqrt{u}\} \tag{4.3.10}$$

and

$$\psi_U = \hat{\psi}_v \exp\{+ z_{\alpha/2}/\sqrt{u}\} \tag{4.3.11}$$

Example 4.3.5 : Consider the data in example 4.3.2, 95% confidence limits for the common odds ratio using Woolf's procedure are given by

$\psi_L = 1.7077$ $\psi_U = 8.3967$

Mantel - Haenszel procedure

As discussed in 4.2.4, Mantel - Haenszel (M-H) estimate for ψ is defined as

$$\hat{\psi}_{M-H} = \frac{\sum_{k=1}^{2} x_{1k}(n_{2k} - x_{2k})/(n_{1k} + n_{2k})}{\sum_{k=1}^{2} x_{2k}(n_{1k} - x_{1k})/(n_{1k} + n_{2k})}$$

An approximate test of the hypothesis of no association (H_o: $\psi = 1$) is given as follows. For the k-th subtable, the conditional mean and variance of X_{1k} calculated under H_o, is given by

$$E(X_{1k}) = n_{1k}\ m_{1k}/N_k$$

$$V(X_{1k}) = \frac{n_{1k}\ n_{2k}\ m_{1k}(N_k - m_{1k})}{N_k^2(N_k - 1)} .$$

The Mantel - Haenszel test of H_o : $\psi = 1$ against the two sided alternative H_1:$\psi \neq 1$ is then given by

$$\chi^2_{M-H} = \frac{[\ |\ \sum_{k=1}^{2} x_{1k} - \sum_{k=1}^{2} E(X_{1k})|\ - 1/2]^2}{\sum_{k=1}^{2} V(X_{1k})} \qquad (4.3.12)$$

The 1/2 correction for continuity is used so that the p-value based on χ^2_{M-H} more closely approximates the value based on exact conditional test (Li, Simon and Gart, 1979). The statistic χ^2_{M-H} has an approximate chisquare distribution with one degree of freedom under H_o. For a one sided test, one may use the approximate unit normal deviate $Z = \pm \sqrt{\chi^2_{M-H}}$, the sign being chosen by the direction of alternative hypothesis.

We now consider two asymptotic methods using M-H estimate to obtain confidence limits for ψ. It is assumed that total number of tables and each p_{ik} (i = 1, 2; k = 1,2) remain fixed and n_{ik} = Na_{ik}, i = 1,2; k = 1,2 where a_{ik} are constants such that $0 < a_{ik} < 1$.

I. The first method suggested by Miettinen (1974b, 1976a) is called the test - based method. This method gives approximate $(1-\alpha)$ 100% confidence limits for ψ as

$$\psi_L = \exp\{(1 - z_{\alpha/2}/ \sqrt{X^2_{M-H}}\,)\ln \hat{\psi}_{M-H}\} \qquad (4.3.13)$$

$$\psi_U = \exp\{(1 + z_{\alpha/2}/ \sqrt{X^2_{M-H}}\,)\ln \hat{\psi}_{M-H}\} \qquad (4.3.14)$$

Example 4.3.6 : For the data in example 4.2.1, 95% confidence limits using above method are given by

$$\psi_L = 1.4358 \qquad \psi_U = 10.0164.$$

II. The second method is based on M-H estimate and its variance derived by Hauck (1979).

Hauck (1979) has given an estimate of the variance of $\ln \hat{\psi}_{M-H}$ as

$$\hat{V}(\ln \hat{\psi}_{M-H}) = \frac{\sum_{k=1}^{2} w_k \hat{v}_k}{\sum_{k=1}^{2} w_k}$$

where w_k and $\hat{v}_k$ are as specified earlier.

An approximate $(1-\alpha)$ 100% confidence interval for $\ln \psi$ is thus given by

$$\ln \hat{\psi}_{M-H} \pm z_{\alpha/2} \sqrt{\hat{V}\,(\ln \hat{\psi}_{M-H}\,)} \qquad (4.3.15)$$

Approximate lower and upper confidence limits for ψ are then obtained by taking antilogs of the lower and upper limit for $\ln \psi$. Hence

$$\psi_L = \hat{\psi}_{M-H} \exp\{ - z_{\alpha/2} \sqrt{\hat{V}(\ln \hat{\psi}_{M-H})}\} \qquad (4.3.16)$$

$$\psi_U = \hat{\psi}_{M-H} \exp\{ + z_{\alpha/2} \sqrt{\hat{V}(\ln \hat{\psi}_{M-H})}\} \qquad (4.3.17)$$

Example 4.3.7 : For the data in example 4.2.1, 95% confidence limits using above method are given by

$$\psi_L = 0.9727 \qquad \psi_U = 14.7851.$$

Chapter 5

Paradoxes

CHAPTER 5

PARADOXES

5.1 : Introduction

One of the questions that often arises in the consideration of cross classified discrete data is whether a relatively complex table can be collapsed, or a collection of individual tables pooled in order to yield a simpler table without affecting the conclusions regarding the relationship of interest. For example, instead of examining the effectiveness of a treatment in two subpopulations, one consisting of men and the other consisting of women; one can pool the data across both the sexes and can examine effectiveness of the treatment in the combined population. Thus, pooling of information from subpopulations achieves data compactification. But some characteristics of data are lost in the process. The greatest danger in amalgamating contingency tables is the possibility of a resulting paradox. Many such have been noted since 1903, when Yule first noticed the phenomenon. Section 5.2 defines such paradoxes, viz., Yule's association paradox (YAP), Yule's reversal paradox (YRP) or Simpson's paradox and amalgamation paradox (AMP). Sections 5.3, 5.4 and 5.5 review sufficient conditions for avoidance of these paradoxes.

What would be best is to find a statistical explanation of any paradox when it occurs, namely, a necessary and sufficient condition for the paradox, described in statistical terms. Unfortunately, none seems to have been discovered, so far, for any of the three paradoxes we consider. Necessary condition for occurrence of YRP is discussed in section 5.4.

Notation :

Let $M_i = (a_i, b_i, c_i, d_i)$ i = 0,1,2,..., t-1 denote the 2x2 contingency table corresponding to i-th of t mutually exclusive subpopulations with $a_i\, b_i\, c_i\, d_i \neq 0$. Here (a_i, b_i) and (c_i, d_i) denote the first and second row respectively for the i-th subpopulation.

If the t tables are added together, we obtain a table $M = \Sigma M_i = (A, B; C, D) = (\Sigma_i a_i, \Sigma_i b_i; \Sigma_i c_i, \Sigma d_i)$. The measure of association considered is the odds ratio, so that for M_i, the i-th subpopulation $\psi(M_i) = \frac{a_i d_i}{b_i c_i}$, i = 0, 1, ..., t-1 and for the combined population

$$\psi(M) = \frac{(\Sigma_i a_i)\ (\Sigma_i d_i)}{(\Sigma_i b_i)\ (\Sigma_i c_i)} \quad .$$

In this chapter, we consider two subpopulations and proofs for various theorems are given for the case when t = 2. These proofs easily extend to the case when t > 2. The two subpopulations M_i, i = 0,1 can be explicitly represented by

	M_o				M_1	
	S	$\bar{S}$			S	$\bar{S}$
T	a_o	b_o		T	a_1	b_1
$\bar{T}$	c_o	d_o		$\bar{T}$	c_1	d_1

5.2 : Definitions

Yule (1903) noted that a spurious association between attributes may arise in a population as a result of amalgamation eventhough the attributes are independent in the subpopulations. Mittal (1991) refers to this paradoxical situation as 'Yule's association paradox' (YAP). If the odds ratio is considered as measure of association; the definition of YAP is formalized in the following. .

Definition 5.2.1 : It is possible that $\psi(M_i) = 1$, $i=0,1,\ldots,t-1$; but $\psi(M) \neq 1$ where $M = \sum_i M_i$, denotes the amalgamated table. In other words, it is possible that $a_i d_i = b_i c_i$, $i = 0,1,\ldots,t-1$; but $(\sum_i a_i)(\sum_i d_i) = (\sum_i b_i)(\sum_i c_i)$. Such a paraoxical situation is termed as Yule's association paradox (YAP).

Following example will illustrate YAP (Schlesselman,1982).

Example 5.2.1 : Suppose one wants to investigate a postulated causal relationship between alcohol consumption and myocardial infarction (MI). Since smoking is known to be a cause of MI, subjects are classified into a smoking group and a nonsmoking group as follows.

Smokers (M_0)			Nonsmokers (M_1)		
Alcohol	MI	Control	Alcohol	MI	Control
Yes	63	36	Yes	8	16
No	7	4	No	22	44

Among smokers, odds ratio estimate of MI associated with alcohol consumption is $\psi(M_0) = 1$, with an identical estimate among nonsmokers. If we pool the data by summing the entries across the subgroups, we get the following 2x2 table

	$M\ (= M_0 + M_1)$	
Alcohol	MI	Control
Yes	71	52
No	29	48

For this pooled table, an estimated odds ratio of MI associated with alcohol consumption is $\psi(M) = 2.2599$;thus showing spurious positive association between MI and alcohol consumption.

Definition 5.2.2 : Cohen and Nagel (1934) noticed the paradoxical behavior in the form that is now popularly known as Simpson's paradox, namely,

$$a_i d_i \geq (\leq)\ b_i c_i\ , \quad i = 0,\ 1,\ \ldots,\ t-1$$

$$\text{but} \quad (\sum_i a_i)\ (\sum_i d_i) \leq (\geq)\ (\sum_i b_i)\ (\sum_i c_i)$$

where equality sign holds in only one of the two statements above.

For the two subpopulations M_0 and M_1, the statement of Simpson's paradox can also be given in the following way.

It is possible to have

$$P(S|T) > (<) P(S \mid \bar{T})$$

and have at the same time both

$$P(S|TM_o) < (>) P(S|\bar{T}M_o) \qquad (5.2.1)$$

and $$P(S|TM_1) < (>) P(S|\bar{T}M_1)$$

Note that here one tends to reason intuitively that this is impossible because

$P(S|T)$ = An average of $P(S|TM_o)$ and $P(S|TM_1)$

and $P(S|\bar{T})$ = An average of $P(S|\bar{T}M_o)$ and $P(S|\bar{T}M_1)$

which is easily seen to be true if all the conditioning events have positive probabilities :

$$P(S|T) = P(M_o|T) \; P(S|TM_o) + P(M_1|T) \, P(S|TM_1)$$

$$P(S|\bar{T}) = P(M_o|\bar{T}) \; P(S|\bar{T}M_o) + P(M_1|\bar{T}) \, P(S|\bar{T}M_1)$$

but the reasoning fails because these two averages have different weightings. However, in particular, if M_o and T are independent; then these two weightings coincide and the reasoning correctly shows that (5.2.1) is impossible. The paradox can be said to result from association of T and M_o. We study this in section 5.4.

Since Nagel suspected that he learned the paradox from Yule's work, Mittal (1991) refers to it as Yule's reversal paradox (YRP). Messick and van de Geer (1981) called this as 'reversal paradox'. Following numerical example will illustrate that whenever YRP occurs, it can be misleading.

Example 5.2.2 : Consider the following two tables.

Men (M_o)

	success	failure
Treatment I	60	20
Treatment II	100	50

Women (M_1)

	success	failure
Treatment I	40	80
Treatment II	10	30

For both the subpopulations, odds ratio estimate is 1.5 indicating that treatment I is better than treatment II. If we combine the tables for both the sexes, we have

$M\ (= M_o + M_1)$

	success	failure
Treatment I	100	100
Treatment II	110	80

The odds ratio estimate for the combined table is 0.7273 indicating that treatment II is better than treatment I; thus reversing the direction of association.

Good and Mittal (1987) have defined amalgamation paradox (AMP) as

Definition 5.2.3 : For t subpopulations; we say that amalgamation paradox (AMP) occurs if

$$\max_i \alpha(M_i) < \alpha(M) \quad \text{or} \quad \alpha(M) < \min_i \alpha(M_i) \ , \quad i = 0,1,\ldots,t-1$$

where α is the measure of association and $M = M_0+M_1+\ldots+M_{t-1}$ is the amalgamated table.

If odds ratio is considered as measure of association, then AMP is implied by YRP which in turn is implied by YAP.

Following example illustrates AMP.

Example 5.2.3 : Consider the two tables (15, 12, 5, 8) and (18, 25; 2, 5) with amalgamated table (33, 37; 7, 13). The odds ratios for the two tables and the amalgamated table are 2, 1.8 and 1.6563 respectively, thus resulting in AMP.

5.3 : Yule's association paradox

The main result of this section is the necessary and sufficient condition for avoidance of YAP for two subpopulations (Mittal 1991). Before this, we study some definitions.

Let the t subpopulations be represented by corresponding 2x2 tables $M_i = (a_i, b_i, c_i, d_i)$, $i = 0,1,\ldots,t-1$.

Definition 5.3.1 : The subpopulations are called row-homogeneous if $\max_i r_i \le \min_i s_i$, $i = 0,1,\ldots, t-1$ where (r_i, s_i) is one of the two pairs $(b_i/a_i, d_i/c_i)$; $(d_i/c_i, b_i/a_i)$.

Definition 5.3.2 : The subpopulations are called column - homogeneous if $\max_i r_i \le \min_i s_i$, $i = 0,1,\ldots, t-1$, where (r_i, s_i) is one of the two pairs $(c_i/a_i, d_i/b_i)$; $(d_i/b_i, c_i/a_i)$.

Definition 5.3.3 : The subpopulations are called homogeneous if they are either row or column homogeneous.

Example 5.3.1 : Consider the tables (15, 12; 5, 8) and (18, 25; 2,5).

(a) To check whether the subpopulations are row-homogeneous, let $r_i = b_i/a_i$ and $s_i = d_i/c_i$, i = 1,2. Then $\max_i r_i = 25/18$ and $\min_i s_i = 8/5$. Thus the subpopulations are row-homogeneous as $25/18 < 8/5$.

(b) Further, we note that if (r_i, s_i) corresponds to either of the pairs $(c_i/a_i, d_i/b_i)$; $(d_i/b_i, c_i/a_i)$; $\max_i r_i > \min_i s_i$. Hence the subpopulations are not column - homogeneous.

(c) The subpopulations are homogeneous.

Definition 5.3.4 : Attributes T and S are positively associated viz. $T \sim S$ if $P(S \mid T) > P(S \mid \overline{T})$.

Similarly, we define

Definition 5.3.5 : Attributes T and S are negatively associated if $P(\overline{S} \mid T) > P(\overline{S} \mid \overline{T})$.

In the following lemma we prove reflexive property of the relation ~ defined in the deinition 5.3.4.

Lemma 5.3.1 : The relation ~ in definition 5.3.4 is reflexive.

Proof : Let $T \sim S$. Then we have to prove that $S \sim T$ i.e. $P(T \mid S) > P(T \mid \overline{S})$. Now

$T \sim S$

$\Rightarrow \; P(S|T) > P(S|\overline{T})$

$$\Rightarrow \; \frac{P(ST)}{P(T)} > \frac{P(S\overline{T})}{P(\overline{T})}$$

$$\Rightarrow \quad \frac{P(ST)}{P(S\overline{T})} > \frac{P(T)}{P(\overline{T})}$$

$$\Rightarrow \quad \frac{P(ST)}{P(S\overline{T})} > \frac{P(ST) + P(\overline{S}\,T)}{P(S\overline{T}) + P(\overline{S}\,\overline{T})}$$

Further

$$\frac{P(ST)}{P(S\overline{T})} > \frac{P(ST) + P(\overline{S}\,T)}{P(S\overline{T}) + P(\overline{S}\,\overline{T})} > \frac{P(\overline{S}\,T)}{P(\overline{S}\,\overline{T})} \qquad (5.3.1)$$

The inequality (5.3.1) above follows since the middle term is a convex combination of two end terms. Thus,

$$\frac{P(ST)}{P(S\overline{T})} > \frac{P(\overline{S}\,T)}{P(\overline{S}\,\overline{T})}$$

$$\Rightarrow \quad \frac{P(\overline{S}\,\overline{T})}{P(\overline{S}\,T)} > \frac{P(S\overline{T})}{P(ST)}$$

$$\Rightarrow \quad \frac{P(\overline{S})}{P(\overline{S}T)} > \frac{P(S)}{P(ST)}$$

$$\Rightarrow \quad \frac{P(ST)}{P(S)} > \frac{P(\overline{S}T)}{P(\overline{S})}$$

$$\Rightarrow \quad P(T \mid S) > P(T \mid \overline{S})$$

Hence, the relation ~ is reflexive. □

Now, we give interpretation of homogeneity in terms of conditional probabilities in the following.

Consider first the two subpopulations M_0 and M_1. If T and S are positively associated in one subpopulation and negatively associated in the other subpopulation, then the subpopulations are obviously nonhomogeneous.

Eventhough T and S are similarly associated in both the subpopulations; conditional probabilities may render them nonhomogeneous. This can be shown as follows.

Assume without loss of generality that T is positively associated with S in both the subpopulations M_0 and M_1.

i.e. $P(S|TM_0) > P(S|\bar{T}M_0)$ and $P(S|TM_1) > P(S|\bar{T}M_1)$.

Now

$$P(S|TM_0) > P(S|\bar{T}M_0) \Rightarrow \frac{b_0}{a_0} < \frac{d_0}{c_0} \tag{5.3.2}$$

and

$$P(S|TM_1) > P(S|\bar{T}M_1) \Rightarrow \frac{b_1}{a_1} < \frac{d_1}{c_1} \tag{5.3.3}$$

When $\max\limits_i \frac{b_i}{a_i} = \frac{b_0}{a_0}$, we have three cases satisfying (5.3.2) and (5.3.3), viz.

(i) $\frac{b_1}{a_1} < \frac{d_1}{c_1} < \frac{b_0}{a_0} < \frac{d_0}{c_0}$

(ii) $\frac{b_1}{a_1} < \frac{b_0}{a_0} < \frac{d_1}{c_1} < \frac{d_0}{c_0}$

and (iii) $\frac{b_1}{a_1} < \frac{b_0}{a_0} < \frac{d_0}{c_0} < \frac{d_1}{c_1}$.

In cases (ii) and (iii), we observe that condition of homogeneity is satisfied. But in case (i) condition of homogeneity is not satisfied. Thus, in case (i); we have $\max\left(\frac{b_1}{a_1}, \frac{d_1}{c_1} \right) < \min\left(\frac{b_o}{a_o}, \frac{b_o}{c_o} \right)$ so that both the treatment and nontreatment in population M_1 is better than treatment in M_o indicating that treatment behaves differently in two subpopulations and hence the subpopulations will be nonhomogeneous with regards to the effect of this particular treatment. Note that the case when $\max_i b_i/a_i = b_1/a_1$ can be similarly treated.

Further, if we have t (t > 2) subpooulations that are nonhomogeneous eventhough T and S are similarly associated in all of them; then we must find two subpopulations M_i and M_j, so that both treatment and nontreatment in one subpopulation is more effective than either treatment or nontreatment in the other subpopulation indicating that treatment acts differently in two subpopulations M_i and M_j and hence making the t subpopulations nonhomogeneous.

Now, we discuss the necessary and sufficient condition for avoidance of YAP (Mittal 1991) in the following theorem.

Theorem 5.3.1 : Suppose that the attributes are independent in each of the t subpopulations, viz.,

$$a_i d_i = b_i c_i \quad \text{for } i = 0,1, \ldots, t-1 \tag{5.3.4}$$

If the subpopulations are homogeneous (row or column) then YAP is avoided, viz.,

$$\left(\sum_i a_i \right) \left(\sum_i d_i \right) = \left(\sum_i b_i \right) \left(\sum_i c_i \right) \tag{5.3.5}$$

On the other hand, if YAP is avoided and $t = 2$, then the subpopulations must be row and column homogeneous.

Proof : We first consider the case when $t = 2$. Without loss of generality, suppose that $\max_i b_i/a_i \le \min_i d_i/c_i$, namely the subpopulations are row-homogeneous and that (5.3.4) holds. Then we must have

$$\frac{b_0}{a_0} = \frac{b_1}{a_1} = \frac{d_0}{c_0} = \frac{d_1}{c_1}$$

which easily implies that

$$\frac{\sum_i b_i}{\sum_i a_i} = \frac{\sum_i d_i}{\sum_i c_i} \qquad i = 0, 1$$

so that

$$(\sum_i a_i)(\sum_i d_i) = (\sum_i b_i)(\sum_i c_i)$$

Thus, YAP is avoided.

On the other hand, suppose (5.3.4) and (5.3.5) hold. Then

$$\frac{\sum_i c_i}{\sum_i a_i} = \frac{\sum_i d_i}{\sum_i b_i} \qquad i = 0, 1$$

gives a convex combination of 2 numbers on each side, and the set of these two numbers is the same for each side due to (5.3.4). Further, for 2 subpopulations this will be true iff the corresponding weights on each side are the same, that is

$\dfrac{a_0}{a_0 + a_1} = \dfrac{b_0}{b_0 + b_1}$ or equivalently $\dfrac{a_1}{a_0 + a_1} = \dfrac{b_1}{b_0 + b_1}$ so that

$\frac{a_1}{a_0} = \frac{b_1}{b_0}$ and hence $\frac{d_0}{c_0} = \frac{b_0}{a_0} = \frac{b_1}{a_1} = \frac{d_1}{c_1}$. Thus, the subpopulations are row-homogeneous. Because of symmetry of the argument, they must be column homogeneous.

Now, consider the case when $t > 2$. In this case, the proof for sufficiency of the condition of homogeneity for avoidance of YAP can be given in a similar manner to that of the case when $t = 2$. Following example will illustrate that condition of homogeneity is not necessary for avoidance of YAP when $t > 2$.

Example 5.3.2 : Let $t = 3$ and consider the subpopulations represented by the tables $M_0 = (6, 12; 3, 6)$; $M_1 = (2,3;4,6)$ and $M_2 = (4, 12; 5, 15)$ with amalgamated table $M = (12,27; 12,27)$. We observe that YAP is aovided. Now we check the homogeneity of subpopulations.

(i) Let $r_i = b_i/a_i$ and $s_i = d_i/c_i$ $i = 0,1,2$. Since $r_i = s_i$; $\max_i r_i > \min_i s_i$ and hence the subpopulations are not row homogeneous.

(ii) Now, let $r_i = c_i/a_i$ and $S_i = d_i/b_i$, $i = 0,1,2$.

Since $r_i = s_i$, $i = 0,1,2$; $\max_i r_i > \min_i S_i$ and hence the subpopulations are not column - homogeneous.

Thus, pooling of data from three nonhomogeneous population has not produced YAP. And hence the condition of homogeneity is not necessary for avoidance of YAP when more than two subpopulations are amalgamated.

5.4 : Yule's reversal paradox

In the previous section, we have seen that homogeneity is sufficient for avoidance of YAP. Further it is necessary for avoidance of YAP when only two subpopulations are pooled. In the following theorem, we show that the homogeneity is sufficient for avoidance of YRP (Mittal 1991).

Theorem 5.4.1 : If the subpopulations are homogeneous then YRP is avoided.

Proof : Consider the case when t = 2. We represent a 2x2 contingency table by two vectors (a,b) and (c,d). Definition of homogeneity says that the subpopulations will be homogeneous if $\max_i b_i/a_i \le \min_i d_i/.c_i$, i = 0,1. This is true if maximum slope of (a,b) vectors is less than minimum slope of (c,d) vectors. Adding the two (a,b) vectors, the slope wil still remain smaller than that of the sum of two (c,d) vectors, thus showing that the condition of homogeneity is sufficient to avoid YRP.

Note that the proof easily extends to the case when t > 2.

□

Following example will illustrate that the condition of homogeneity is not necessary for avoidance of YRP. i.e. if YRP is avoided, then the subpopulations need not be homogeneous. In other words, a paradox will not necessarily occur if non homogeneous subpopulations are amalgamated.

Example 5.4.1 : Consider the two tables (13, 17; 8, 10) and (15, 23; 5, 7) with the pooled table (28, 40; 13, 17).

(i) If (r_i, s_i) corresponds to either of the pairs $(b_i/a_i, d_i/c_i)$; $(d_i/c_i, b_i/a_i)$ we observe that $\max_i r_i > \min_i s_i$. Hence, the subpopulations are not row-homogeneous.

(ii) If (r_i, s_i) corresponds to either of the pairs $(c_i/a_i, d_i/b_i)$; $(d_i/b_i, c_i/a_i)$ we observe that $\max_i r_i > \min_i s_i$. Hence, the subpopulations are not column-homogeneous.

But YRP does not occur since the odds ratio estimates are, respectively, 0.9559, 0.9130 and 0.9154.

Thus, if the subpopulations are homogeneous, it is assured that YRP is avoided and if the subpopulations are nonhomogeneous, nothing can be said about the occurence of YRP; but second look at data may reveal some characteristics originally overlooked (Mittal 1991).

Whenever a paradox occurs, one faces difficulty with its interpretation. We now give necessary condition for occurence of YRP (Mittal 1991) which may help in interpretation of YRP.

Theorem 5.4.2 : If $T \sim S$ and YRP occurs while amalgamating $T \times S$ tables over M_0 and M_1 then either (i) $M_0 \sim S$ and $M_0 \sim T$ or (ii) $M_0 \sim \bar{S}$ and $M_0 \sim \bar{T}$.

(Note that when $T \sim \bar{S}$ and YRP occurs, we only need to interchange S and $\bar{S}$ to get the corresponding statement of the theorem).

Proof : The amalgamated table (over M_o and M_1) can be written as

	M	
	S	$\bar{S}$
T	a_o+a_1	b_o+b_1
$\bar{T}$	c_o+c_1	d_o+d_1

Now $T \sim S$, so that $P(S|T) > p(S|\bar{T})$

$$=> \frac{a_o+a_1}{a_o+a_1+b_o+b_1} > \frac{c_o+c_1}{c_o+c_1+d_o+d_1}$$

$$=> \frac{b_o+b_1}{a_o+a_1} < \frac{d_o+d_1}{c_o+c_1}$$

$$=> \frac{a_o+a_1}{c_o+c_1} > \frac{b_o+b_1}{d_o+d_1} \qquad (5.4.1)$$

and since YRP occurs, we have

$$\frac{a_o}{c_o} \le \frac{b_o}{d_o} \quad \text{and} \quad \frac{a_1}{c_1} \le \frac{b_1}{c_1} \qquad (5.4.2)$$

If we write both sides of (5.4.1) as convex combinations of ratios a_i/c_i and b_i/d_i $i = 0,1,$; namely,

$$\frac{a_o+a_1}{c_o+c_1} = \frac{c_o}{c_o+c_1} \times \frac{a_o}{c_o} + \frac{c_1}{c_o+c_1} \times \frac{a_1}{c_1}$$

and

$$\frac{b_o+b_1}{d_o+d_1} = \frac{d_o}{d_o+d_1} \times \frac{b_o}{d_o} + \frac{d_1}{d_o+d_1} \times \frac{b_1}{d_1}$$

then in view of (5.4.2), (5.4.1) is possible only if the interval $[a_1/c_1, b_1/d_1]$ lies entirely to the left or entirely to the right of the interval $[a_o/c_o, b_o/d_o]$, that is

$$\frac{a_1}{c_1} \le \frac{b_1}{d_1} \le \frac{a_o}{c_o} \le \frac{b_o}{d_o} \tag{5.4.3}$$

$$\frac{a_o}{c_o} \le \frac{b_o}{d_o} \le \frac{a_1}{c_1} \le \frac{b_1}{d_1} \tag{5.4.4}$$

Also, in order to achieve (5.4.1); the weight assigned to the smaller of the two ratios a_i/c_i, i = 0,1 must be smaller than the weight assigned to the smaller of the two ratios b_i/d_i i = 0,1.

Now (5.4.3)

$$\Rightarrow \frac{a_1}{c_1} \le \frac{b_1}{d_1} \le \frac{a_o}{c_o} \le \frac{b_o}{d_o}$$

$$\Rightarrow \frac{c_1}{d_1} \le \frac{c_o}{d_o}$$

$$\Rightarrow \frac{c_o}{c_1} \ge \frac{d_o}{d_1}$$

Similarly, (5.4.4) $\Rightarrow \dfrac{c_o}{c_1} \le \dfrac{d_o}{d_1}$.

We note that the assumption of the theorem can also be stated by inequalities such as in (5.4.1) and (5.4.2) with c_i and b_i, i = 0,1 interchanged. This will give

$$\frac{a_1}{b_1} \le \frac{c_1}{d_1} \le \frac{a_o}{b_o} \le \frac{c_o}{d_o} \tag{5.4.5}$$

or

$$\frac{a_o}{b_o} \le \frac{c_o}{d_o} \le \frac{a_1}{b_1} \le \frac{c_1}{d_1} \tag{5.4.6}$$

It is clear that (5.4.5) $\Rightarrow \dfrac{c_o}{c_1} \ge \dfrac{d_o}{d_1}$ while (5.4.6) implies reverse inequality.

Hence, the only possible cases under the assumptions of the theorem are

$$\frac{a_1}{c_1} \le \frac{b_1}{d_1} \le \frac{a_o}{c_o} \le \frac{b_o}{d_o} \quad \text{and} \quad \frac{a_1}{b_1} \le \frac{c_1}{d_1} \le \frac{a_o}{b_o} \le \frac{c_o}{d_o} \qquad (5.4.7)$$

or

$$\frac{a_o}{c_o} \le \frac{b_o}{d_o} \le \frac{a_1}{c_1} \le \frac{b_1}{d_1} \quad \text{and} \quad \frac{a_o}{b_o} \le \frac{c_o}{d_o} \le \frac{a_1}{b_1} \le \frac{c_1}{d_1} \qquad (5.4.8)$$

Now, if $M_o \sim S$, then $P(S|M_o) > P(S_1|M_1)$ so that

$$\frac{a_o+c_o}{b_o+d_o} > \frac{a_1+c_1}{b_1+d_1} \qquad (5.4.9)$$

and if $M_o \sim T$ then $P(T|M_o) > P(T|M_1)$ so that

$$\frac{a_o+b_o}{c_o+d_o} > \frac{a_1+b_1}{c_1+d_1} \qquad (5.4.10)$$

It is easy to see that (5.4.7) implies both (5.4.9) and (5.4.10); while (5.4.8) implies (5.4.9) and (5.4.10) with inequalities reversed (i.e. $M_o \sim \overline{S}$ and $M_o \sim \overline{T}$). This proves the theorem.

□

5.5 : Amalgamation paradox

Following example will show that homogeneity is neither sufficient nor necessary for avoidance of AMP.

Example 5.5.1 : (i) Consider the two subpopulations represented by the tables M_o = (15, 12; 5, 8) and M_1 = (18, 25; 2,5).

In Example 5.3.1, we have shown that these subpopulations are homogeneous. Further example 5.2.3 shows that amalgamation of the tables (over M_o and M_1) results in AMP. Thus, homogeneity is not sufficient to avoid AMP.

(ii) Consider the tables in example 5.4.1, viz. (13, 17; 8, 10) and (15, 23; 5,7). In example 5.4.1, it is shown that the subpopulations are not homogeneous. Further AMP does not occur since the odds ratio estimate for the pooled table is 0.9154 which lies in the interval of odds ratio estimates of the two subpopulations viz. 0.9130, 0.9559.

Good and Mittal (1987) have shown that how AMP can be avoided by suitable designs of the sampling experiments.

Definition 5.5.1 : An experimental design is said to 'row-uniform' or 'row-fair' if for some λ,

$$\frac{a_i + b_i}{c_i + d_i} = \lambda \ , \quad i = 0,\ 1,\ \ldots,\ t-1.$$

Definition 5.5.2 : An experimental design is said to be column uniform or column-fair if for some μ,

$$\frac{a_i + c_i}{b_i + d_i} = \mu \ , \quad i = 0,\ 1,\ 2,\ \ldots,\ t-1.$$

Theorem 5.5.1 : If the design is both row-fair and column-fair then

$$\min_i \psi(M_i) \le \psi(M) \le \max_i \psi(M_i), \ i = 0,\ 1,\ \ldots,\ t-1$$

where $M = M_0 + M_1 + \ldots + M_{t-1}$ represents an amalgamated table.

Proof : We prove the theorem for t = 2. For t > 2 we can first amalgamate two tables and then add further tables one at a time to get the final result.

Let $x_1 = a_o/a_1$, $x_2 = b_o/b_1$, $x_3 = c_o/c_1$ and $x_4 = d_o/d_1$.

For t = 2, we rewrite the conditions of row - uniform and column - uniform design as follows. The design is row-uniform

$$\Rightarrow \frac{a_o+b_o}{c_o+d_o} = \frac{a_1+b_1}{c_1+d_1}$$

$$\Rightarrow \frac{a_o+b_o}{a_1+b_1} = \frac{c_o+d_o}{c_1+d_1}$$

$$\Rightarrow \frac{a_1}{a_1+b_1} \times \frac{a_o}{a_1} + \frac{b_1}{a_1+b_1} \times \frac{b_o}{b_1} = \frac{c_1}{c_1+d_1} \times \frac{c_o}{c_1} + \frac{d_1}{c_1+d_1} \times \frac{d_o}{d_1}$$

$$\Leftrightarrow \quad x_1\delta + x_2(1-\delta) = x_3\delta' + x_4(1 - \delta') \qquad (5.5.1)$$

with $\delta = \dfrac{a_1}{a_1+b_1} > 0$ and $\delta' = \dfrac{c_1}{c_1+d_1} > 0$.

Similarly, the design is column - uniform

$$\Rightarrow \quad x_1\eta + x_3(1 - \eta) = x_2\eta' + x_4(1 - \eta') \qquad (5.5.2)$$

for $\eta = \dfrac{a_1}{a_1+c_1} > 0$ and $\eta' = \dfrac{b_1}{b_1+d_1} > 0$.

Equation (5.5.1) shows that the intervals (x_1, x_2) and (x_3, x_4) must overlap (or as a special case $x_1 = x_2 = x_3 = x_4$) while (5.5.2) shows that the intervals (x_1, x_3) and (x_2, x_4) must overlap. Here, the notation (x_1, x_2) is not intended to imply that $x_1 \le x_2$ etc.

Without loss of generality, we shall assume that

$$\psi(M_o) \le \psi(M_1)$$

$$\Rightarrow \frac{a_o d_o}{b_o c_o} \le \frac{a_1 d_1}{b_1 c_1}$$

$$\Rightarrow \frac{a_o d_o}{a_1 d_1} \le \frac{b_o c_o}{b_1 c_1}$$

$$\Rightarrow x_1 x_4 \le x_2 x_3 \qquad (5.5.3)$$

We have to show that

$$\psi(M_o) \le \psi(M) \le \psi(M_1) \qquad (5.5.4)$$

First, let us assume that $x_1 \le x_4$; then x_2 and x_3 can not both be in the interval $[x_1, x_4]$ since $x_1 \le x_2 \le x_3 \le x_4$ violates (5.5.1) while $x_1 \le x_3 \le x_2 \le x_4$ violates (5.5.2). The only exception is when $x_1 = x_2 = x_3 = x_4$, but the result is then trivially true. For the same reasons, it is not possible that one of x_2 and x_3 is less than x_1 and the other is greater than x_4. If both x_2, x_3 are less than x_1, then (5.5.3) is violated. Thus, x_2 and x_3 both have to exceed x_4. Accordingly, we are left with only two cases,

$$x_1 \le x_4 < x_2 \le x_3 \qquad (5.5.5)$$

and

$$x_1 \le x_4 < x_3 \le x_2 \qquad (5.5.6)$$

Arguing similarly for the case $x_4 \le x_1$, we get two more cases, namely,

$$x_4 \le x_1 < x_2 \le x_3 \qquad (5.5.7)$$

and

$$x_4 \le x_1 < x_3 \le x_2 \qquad (5.5.8)$$

For each one of the four cases above, we will show that (5.5.4) holds.

We note that in all four of the cases (5.5.5), (5.5.6), (5.5.7) and (5.5.8) we have $x_1 \le x_3$ and $x_4 \le x_2$. This implies that $a_0/c_0 \le a_1/c_1$ and $d_0/b_0 \le d_1/b_1$. Therefore

$$\frac{a_0}{c_0} \le \frac{a_0+a_1}{c_0+c_1} \le \frac{a_1}{c_1}$$

as well as

$$\frac{d_0}{b_0} \le \frac{d_0+d_1}{b_0+b_1} \le \frac{d_1}{b_1}$$

and hence (5.5.4) holds readily.

□

The following example shows that just row-uniform or column - uniform design is not sufficient for avoidance of AMP.

Example 5.5.2 : Let M_0 = (3, 1; 1, 9) and M_1 = (889, 203; 381, 2349). Then $\psi(M_0) = \psi(M_1) = 27$. The design is row uniform with $\lambda = 0.4$, but $\psi(M) = 26.9907$. Note that the design is not column-uniform.

It may be noted that the conditions imposed in theorem 5.5.1 are somewhat stringent as in practice it is difficult to obtain a design which is both row and column fair. In practice, we may get a design which is row-uniform and approximately column-uniform or vice versa. Effects of such approximations are discussed by Good and Mittal (1987). Here we do not go into these details.

5.6 : Examples

We need to worry about YAP very rarely in practice, since the precondition $\psi(M_i) = 1$, $i = 0,1,\ldots,t-1$ is unlikely to be satisfied for observational data. Though AMP is more frequent than YRP in data it is YRP that poses critical problems of interpretation and infrence when it occurs.

An actual occurrence of this paradox was observed (Cohen and Nagel, 1934) in a comparison of tuberculosis deaths in New York city and Richmond Virginia, during the year 1910. Although the overall tuberculosis Mortality rate was lower in New York the opposite was observed when the data were separated into two racial categories. Richmond had a lower mortality rate.

Whenever a paradoxical situation arises, what needed is the 'sensible interpretation'. Necessary condition for occurence of YRP states that " If $T \sim S$ and YRP occurs while amalgamating TxS tables over M_o and M_1, then either (i) $M_o \sim S$ and $M_o \sim T$ or (ii) $M_o \sim \bar{S}$ and $M_o \sim \bar{T}$.

Hence apart from the pair of variables of interest; if we investigate associations between remaining pairs of variables, these investigations may help in giving sensible interpretations. The measure of association is the odds ratio. We illustrate this by following example.

Example 5.6.1 : Consider the example constructed by Simpson (1951).

	Male (M_o)			Female (M_1)	
	Alive	Dead		Alive	Dead
Treated	8	5	Treated	12	15
Untreated	4	3	Untreated	2	3

$\psi(M_o) = 24/20$ $\qquad$ $\psi(M_1) = 36/30$

Combined population (M)

	Alive	Dead
Treated	20	20
Untreated	6	6

$\psi(M) = 1$

We observe that there is positive association between treatment and survival in both the sexes, but if we combine the tables, we find that there is no association between treatment and survival in the combined population. To explain such a paradoxical situation; we now consider odds ratio between remaining pairs of variables at each of the two levels of third variable.

	Alive (S)			Dead ($\bar{S}$)	
	Male	Female		Male	Female
Treated	8	12	Treated	5	15
Untreated	4	2	Untreated	3	3

$\psi(S) = 0.33$ $\qquad$ $\psi(\bar{S}) = 0.33$ $\qquad$ (5.6.1)

Odds ratio estimate for the combined table is 0.34. (5.6.2)

Now, consider

	Treated (T)			Untreated ($\bar{T}$)	
	Male	Female		Male	Female
Alive	8	12	Alive	4	2
Dead	5	15	Dead	3	3

$\psi(T) = 2$ $\qquad\qquad$ $\psi(\bar{T}) = 2$ $\qquad\qquad$ (5.6.3)

If we combine the data over T and $\bar{T}$, the odds ratio estimate is 1.93. (5.6.4)

From (5.6.1) and (5.6.2) we observe that there is positive association between females and being treated. In fact data says that proportion of women being treated is three times higher than that of men. (5.6.3) and (5.6.4) suggests that mortality rate for women is twice than that of men regardless of treatment. Hence the sensible interpretation is that "treatment is beneficial".

Example 4.6.2 : Consider the example (Agresti, 1984) which concerns the effect of racial characteristics on the decision regarding whether to impose the death penalty after an individual is convicted for a homicide. The variables considered are `race of defendant' having two categories white and black and `death penalty verdict' having categories `yes' and `no'. The 326 subjects cross classified according to these variables were defendants in homicide indictments in 20 Flourida countries during 1976-77. Following table refers only to indictments for homicides in which defendant and victim were strangers, since death sentences are very rarely imposed when the defendant and the victim had a prior friendship or family relationship.

Death Penalty

Defendant's race	Yes	No
White	19	141
Black	17	149

The odds ratio estimate for the above table is 1.1772. (5.6.5) [Note that for this example, we have calculated odds ratio estimate by adding 0.5 to each entry].

This value of odds ratio means that odds of getting the death penalty were 1.18 times higher for white defendants in the sample than for black defendants. Note that the two - dimensional table presented above is obtained by amalgamating two 2x2 tables; corresponding to two categories of victim's race. We present these tables in the following

Table 5.6.1

Association between defendant's race and death penalty for two categories of victim's race

Victim's race	Defendant's race	Death Penalty	
		Yes	No
White (M_o)	White	19	132
	Black	11	52
		$\psi(M_o) = 0.6719$	(5.6.6)
Black (M_1)	White	0	9
	Black	6	97
		$\psi(M_1) = 0.79$	(5.6.7)

From (5.6.5) to (5.6.7) we observe that association between defendant's race and death penalty is reversed when 'victim's race' is ignored.

To illustrate such a paradoxical situation, we study association between other pairs of variables through odds ratio.

Table 5.6.2

Association between victim's race and defendent's race at each of the two levels of death penalty

Death Penalty	Victim's race	Defendant's race White	Black
Yes (S)	White	19	11
	Black	0	6
		$\psi(S) = 22.0435$	(5.6.8)
No ($\bar{S}$))	White	132	52
	Black	9	97
		$\psi(\bar{S}) = 25.9023$	(5.6.9)

Now consider combined table over S and $\bar{S}$.

Victim's race	Defendant's race White	Black
White	151	63
Black	9	103

For the combined table (over S and $\bar{S}$), the odds ratio estimate is given by 25.9929. (5.6.10)

Now consider the remaiining pair of variables.

Table 5.6.3

Association between death penalty and victim's race at two levels of defendent's race

Defendant's race	Victim's race	Death Penalty	
		Yes	No
White (T))	White	19	132
	Black	0	9
		$\psi(T) = 2.7962$	(5.6.11)
Black ($\bar{T}$)	White	11	52
	Black	6	97
		$\psi(\bar{T}) = 3.2857$	(5.6.12)

Now consider combined table over T and $\bar{T}$

Victim's race	Death Penalty	
	Yes	No
White	30	184
Black	6	106

For the combined table over T and $\bar{T}$, odds ratio estimate is given by 2.7086. (5.6.13)

The odds ratios relating victim's race and defendant's race indicate a very strong association between these two variables. To illustrate, odds of having killed a white are estimated to be 26 times higher for white defendants than for black defendants.

The odds ratios relating death penalty verdict and victim's race indicate that death penalty was more likely to be imposed when the victim was white than when the victim was black.

All these paradoxes suggest an analysis which takes into account interaction between defendant's race and victim's race. One can find details of such an analysis of this problem in Agresti (1984).

REFERENCES

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, John - Wiley and Sons.

Anscombe, F.J. (1956). On estimating binomial response relations, *Biometrika*, 43, 461-464.

Barnard, G.A. (1945). A new test for 2x2 tables, *Nature*, 156, 177.

Barnard, G.A. (1947a) : Significance tests for 2x2 tables, *Biometrika*, 34, 123-138.

Barnard, G.A. (1949). Statistical inference, *Journal of the Royal Statistical Society*, B, 11, 115-139.

Barnard, G.A. (1979). In contradiction to J. Berkson's dispraise - conditional tests can be more efficient. *Journal of the Statistical Planning and Inference*, Vol. 3, No. 3, 181-188.

Basu, D. (1979). Discussion of J. Berkson's paper `In dispraise of exact test', *Journal of the Statistical Planning and Inference*, Vol. 3, No. 3, 189-192.

Berkson, J. (1978a). In dispraise of exact tests, *Journal of the Statistical Planning and Inference*, Vol. 2, No. 1, 27-42.

Berkson, J. (1978b). Do the marginal totals of the 2x2 table contain relevant information respecting the table proportions, *Journal of the Statistical Planning and Inference*, Vol. 2, No. 1, 43-44.

Bishop, Y.M.M., Fienberg S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis - Theory and Practice*, MIT Press, Cambridge Massachusettes and London, England.

Breslow, N.E. and Liang, K.Y. (1982). The variance of the Mantel - Haenscel estimator, Biometrics, 38, 943-952.

Cohen, M. R. and Nagel, E. (1934). An Introduction to Logic and Scientific Method, New York, Harcourt, Brace and Co.

Cornfield, J. (1956). A statistical problem arising from retrospective studies, Proceedings of 3rd Berkley Symposium, 4, 135-148.

Corsten L.C.A. and de- Kroon, J.P.M. (1979). Comment on J. Berkson's paper 'In dispraise of exact test', Journal of the Statistical Planning and Inference, Vol. 3, No.3, 193-198.

Cox, D. R. (1970). The Analysis of Binary Data, London - Chapman and Hall.

D'Agostino, R.B., Chase, W. and Belanger, A. (1988). The appropriateness of some common procedures for testing equality of two independent binomial proportions, The American Statistician, 42, 198-202.

Fisher, R.A. (1922). On the interpretation of χ^2 from contingency tables and the calculation of P, Journal of the Royal Statistical Society, 85, 87-94.

Fisher, R.A. (1934). Statistical Methods for Research Workers, 5th Ed. Edinburgh : Oliver and Boyd.

Fisher, R.A. (1956). Statistical Methods and Scientific Inference, Oliver and Boyd, Edinburgh.

Fisher, R. A. (1962). Confidence limits for a cross product ratio, Australian Journal of Statistics, 4, 41.

Gart, J.J. (1962). Approximate confidence limits for the relative risk, Journal of the Royal Statistical Society, B, 24, 454-463.

Gart, J. J. (1970). Point and interval estimation of the common odds ratio in the combination of 2x2 tables with fixed margins, Biometrika, 57, 471-475.

Gart, J. J.; Pettigrew, H.M. and Thomas, D.G. (1985). The effect of bias, variance estimation, skewness and kurtosis of the empirical logit of weighted least square analysis, Biometrika, 72, 179-190.

Greenwood, M. and Yule, G.U. (1915). The statistics of anti-cholera and anti-typhoid inoculations, and the interpretation of such statistics in gneral, Proceedings of Royal Society Med (Epidemiology), 8, 113-190.

Grizzle, J.E. (1967). Continuity correction in a χ^2 test for 2x2 tables, The American Statistician, 21, 28-32.

Grizzle, J.E.; Starmer, C.F. and Koch, C.G. (1969). Analysis of categorical data by linear models, Biometrics, 25, 489-504.

Godambe, V.P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter, Biometrika, 67, 155-162.

Good, I.J. and Mittal, Y. (1987). The amalgamation and geometry of 2x2 contingency tables, The Annals of Statistics, 15, 694-711.

Goodman, L.A. (1964). Simultaneous confidence limits for cross product ratios in contingency tables, Journal of the Royal Statistical Society, B, 26, 86-102.

Goodman, L. A. (1970). The multivariate analysis of qualitative data - interactions among multiple classification, Journal of the American Statistical Association, 65, 226-256.

Goodman, L.A. and Kruskal, W.H. (1954). Measures of association for cross classifications, Journal of the American Statistical Association, 49, 732-764.

Haber, M. (1980). A comaprison of some continuity corrections for the χ^2 test on 2x2 tables, Journal of the American Statistical Association, 75, 371, 510-515.

Haldane, J.B.S. (1955). The estimation and significance of the logarithm of a ratio of frequencies, Annals of Human Genetics, 20, 309-311.

Hauck, W.W. (1979). The large sample variance of the Mantel - Haenszel estimate of the common odds ratio, Biometrics, 35, 817-819.

Hauck, W.W. (1986). Response to reader reaction - Alternative estimators of the common odds ratio for 2x2 tables by N. Mantel, Biometrics, 42, 201-202.

Hauck, W.W., Anderson, S. and Leathy, F.J. (1982). Finite sample properties of some old and some new estimators of a common odds ratio from multiple 2x2 tables, Journal of the American Statistical Association, 77, 145-152.

Jewell, N.P. (1986). On the bias of commonly used measures of association for 2x2 tables, Biometrics, 42, 351-358.

Kempthorne, O. (1979). In dispraise of exact test - reactions, Journal of the Statistical Planning and Inference, Vol. 3, no. 3, 199-214.

Kudo, A. and Tamuri, T. (1978). 2x2 tables emerging out of different chance mechanisms, Communications in Statistics - Theory and Methods, A7(10), 977-986.

Li, S., Simon, R.M., Gart, J.J. (1979) . Small sample properties of the Mantel - Haenszel test, Biometrika, 66, 181-183.

Mantel, N. (1974). Comment on `Some reasons for not using the Yates continuity correction on 2x2 contingency table' by Conver, W.J. Journal of the American Statistical Association, 69, 378-380.

Mantel, N. (1986). Reader reaction - Alternative estimators of the common odds ratio for 2x2 tables, Biometrics, 42, 199-200.

Mantel, N. (1992). An invalid comparison of several point estimators of the odds ratio in a single 2x2 table, Biometrics, 48, 1289-1295.

Mamel, N. and Greenhouse, S.W. (1968). What is a continuity correction?, The American Statistician, 22, 27-30.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, Journal of National Cancer Institute, 22, 719-748.

McDonald, L.L.; Davis, B.M. and Milliken, G.A. (1977). A nonrandomized unconditional test for comparing two proportions in 2x2 contingency tables, Technometrics, 19, 145-147.

Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in r x c contingency tables, Journal of the American Statistical Association, 78, 427-434.

Mehta, C. R.; Patel, N.R. and Grey, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2x2 contingency tables, Journal of the American Statistical Association, 80, 969-973.

Messick, D. M. and van de Geer, J.P. (1981). A reversal paradox, Psychology Bulletin, 90, 582-593.

Miettinen, O.S. (1974b). Simple interval estimation of risk ratio (abstract), American Journal of Epidemiology, 100, 515-516.

Miettinen, O. C. (1976a). Estimability and estimation in case-referent studies, American Journal of Epidemiology, 104, 609-620.

Mittal, Y. (1991). Homogeneity of subpopulations and Simpson's paradox, Journal of the American Statistical Association, 86, 167-172.

Pearson, E.S. (1947). The choice of statistical test illustrated on the interpretation of data classed in a 2x2 table, Biometrika, 34, 139-167.

Pigeot, I. (1991). A jackknife estimator of a combined odds ratio, Biometrics, 47, 373-381.

Plackett, R.L. (1964). The continuity correction in 2x2 tables. Biometrika, 51, 327-337.

Quenouille, M. H. (1949). Approximate tests of correlation in time series, Journal of the Royal Statistical Society, B, 11, 68-84.

Sathe, Y. S. (1982). Another test for equality of two proportions, Communications in Statistics - Simulations and Computations, 11(3), 373-375.

Scheffe, H. (1959). The Analysis of Variance, John Wiley & Sons, Inc.

Schlesselman, J. J. (1982). Case Control Studies - Design. Conduct Analysis, Oxford University Press.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables, Journal of the Royal Statistical Society, B, 238-241.

Storer, B.E. and Kim, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions, Journal of the American Statistical Association, 85, 146-155.

Thomas, D.G. (1975). Exact and asymptotic methods for combination of 2x2 tables, Computers and Biomedical Research. 8, 423-446.

Tocher, K. D. (1950). Extension of the Neyman - Pearson theory of tests to discontinuous variates, Biometrika, 37, 130-144.

Tukey, J. W. (1958). Bias and confidence is not - quite lare samples (Abstract), Annals of Mathematical Statistics, 29, 614.

Upton, G.J.G. (1982). A comparison of alternative tests for the 2x2 comparative trial, Journal of the Royal Statistical Society, A, 145, 86-105.

Vollset, S.E., Hirji, K.F. and Elashoff, R.M. (1991). Fast computation of exact confidence limits for the common odds ratio in a series of 2x2 tables, Journal of the American Statistical Association, 86, 414, 404-409.

Walter, S.D. (1985). Small sample estimation of log odds ratios from logistic regression and four fold tables, Statistics in Medicine, 4, 437-444.

Walter, S. D. (1987). Point estimation of the odds ratio in sparse contingency tables, In Biostatistics, I. B. MacNeill and G. J. Humphrey (eds.) Dordrecht : D. Reidel.

Walter, S. D. and Cook, R.J. (1991). A comparison of several point estimators of the odds ratio in a single 2x2 contingency table, Biometrics, 47, 795-811.

Wise, M. E. (1954). A quickly converging expansion for cumulative hypergeometric probabilities, direct and inverse, Biometrika, 41, 317-329.

Woolf, B. (1955). On estimating the relation between blood group and disease, Annals of Human Genetics, 19, 251-253.

Yates, F. (1934). Contingency tables involving small numbers and the chisquare test, Journal of the Royal Statistical Society, B, Suppl. 1, no. 2, 217-235.

Yates, F. (1984). Tests of significance for 2x2 contingency tables, Journal of the Royal Statistical Society, A, 147 Part 3, 426-463.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics, Biometrika, 2, 121-134.