Statistical Inference for the Rate Ratio in a Two-way Contingency Table with an Empty Cell

الاستدلال الإحصائي حول المعدل النسبي في جداول التوافق الثنائية المحتوية على خلايا خالية

Mahmoud Okasha

Department of Applied Statistics. Faculty of Economics. Al-Azhar University. Gaza. Palestine

Email: m.okasha@palnet.com.

Received: (25/1/2005), Accepted: (28/8/2005)

Abstract:

In many statistical data analyses, the problem of analyzing contingency tables that contain empty cells is commonly encountered in all fields of research. The odds ratio and risk ratio are not applicable in such a case because of an unidentifiablity problem. In the present paper the rate ratio between the second negative response given an initial negative response and the initial negative response in a two-way contingency table with a zero-count in one of the off-diagonal cells is utilized. Hypothesis testing and confidence interval construction for the rate ratio based on the Wald's test statistic and its logarithmic transformation will be reviewed. Inference based on large sample theory and small-sample on the rate ratio of this case is discussed. The asymptotic performance of the Wald's test statistic and its logarithmic transformation is examined. By adopting these statistics, *full unconditional exact* small-sample procedures that have been proposed by Tang and Tang (2002) are discussed. The procedures are modified by utilizing the maximum likelihood estimator of the rate ratio and the conditional likelihood functions. The accuracy of all the methods is empirically assessed. We show that our *modified conditional* procedures are more reliable than both exact and asymptotic procedures in terms of coverage probability and expected interval width. The methodology is applied to data on family planning in the Gaza Strip and other examples from the literature.

Key Words: rate ratio, two-way contingency tables, empty cells, Wald's test, logarithmic transformation, coverage probability.

المقدمة:

يواجه الكثير من الباحثين أثناء التحليل الإحصائي للبيانات الوصفية في مجالات العلوم المختلفة حالات وجود خلايا ذات تكرارات صفرية حقيقية في جداول توافق ثنائية، في مثل هذه الحالات لن يتمكن الباحث من حساب أي من النسب والمعدلات المعروفة مثل نسب الخلاف أو نسب المخاطرة بسبب ظهور مشكلة "عدم التعريف" نتيجة لتساوي أحد التكرارات بالصفر، ولعلاج هذه المشكلة تقوم هذه الدراسة على توظيف المعدل النسبي بين الاستجابة السلبية للظاهرة الثانية بمعلومية وقوع استجابة سلبية للظاهرة الأولى مقارنة بالاستجابة السلبية للظاهرة الأولى وإجراء استدلال إحصائي حول هذا المعدل وذلك في جدول توافق ثنائي يحتوي على خلية خالية في أحد خلايا قطره الثانوي، وقد تم تقدير المعدل النسبي وبناء فترة ثقة له واشتقاق اختبارات للفرضيات المتعلقة به مع توظيف كلاً من إحصاء والد المعروف والإحصاء المتعلق بتحويلته اللوغاريتمية، وقد تم مراجعة التوزيع التقاربي لهذين الإحصاءين وكذلك نتائج استخداميهما في حالة العينات الكبيرة، كذلك فقد تم دراسة واستخدام التوزيع الاحتمالي الدقيق لكل من هذين الإحصاءين، وبناء عليه فقد تم دراسة الطرق الدقيقة الغير مشروطة التي قام بتطويرها Tang and Tang (2002) لحالات العينات الصغيرة باستخدام هذين الإحصاءين، ولقد تم تطوير هذه الطرق في هذه الدراسة لحساب الاحتمالات ومستويات الثقة عن طريق استخدام مقدر الأرجحية العظمى للمعدل النسبي وتوظيف دوال الأرجحية العظمى الكاملة للمعدل النسبي غير المشروطة والمشروطة وذلك لتطبيقها في حالات العينات الصغيرة، وقد استخدمت أيضاً طرق المحاكاة للتحقق من دقة جميع الطرق التي تم التعرض لها، وسوف يتبين من ذلك أن

هذه الطرق سواء كانت الغير مشروطة أو المشروطة التي قمنا بتطويرها تعطي نتائج أدق ويمكن الاعتماد عليها بشكل أفضل من الطرق السابقة في تقدير فترات الثقة واختبارات الفرضيات للمعدل النسبي في جداول التوافق الثنائية المحتوية على تكرار صفري حقيقي في أحد خلاياها وذلك باستخدام معياري احتمال التغطية لفترات الثقة وطول الفترة، وفي النهاية تم تطبيق جميع الطرق التي تم التعرض إليها في هذه الدراسة على بيانات حقيقية منها المنشور في المجلات العلمية ومنها الميدانية عن بحث يتعلق بتنظيم الأسرة في قطاع غزة.

1. Introduction :

The subject of this paper has been stimulated from the analysis of a family planning survey where we met many sparse two-way contingency tables that contain empty cells. In family planning surveys we have samples of men and women classified by their knowledge and attitudes towards family planning and their practicing status of family planning. We expect that all persons who do not know about or have negative attitudes towards family planning do not practice it. The problem of interest in this paper is to make inference on the rate of applying family planning methods among those who know and have positive attitudes towards family planning compared to those who know about family planning methods.

In many cases of 2x2 contingency tables it is expected that one cell in the table could contain a zero-count. For such an empty cell even though the cell has a zero-count, its true probability may be greater than zero. This means that, in such a case, it could be theoretically possible to have observations in the cell if the sample size was sufficiently large. However, we are interested in contingency tables which have empty cells for which observations are theoretically impossible. Such a cell has a true probability equals to zero and the cell count is zero regardless of the sample size. Contingency tables containing empty cells are often called *incomplete* tables and the empty cell is called a *structural zero* if its true probability equals to zero (Agresti, 1996). In such a case, the zero-count is not an observation and is not part of the data. On the other hand, empty cells with true probability greater than zero are common and intensively discussed in many articles in the literature including Agresti (1990 and 1996) and Bishop et. al. (1975). However, little discussion was found in the literature on contingency tables that contain empty cells with true zero probability (Tang and Tang, 2002). In this paper, we discuss in detail the problems of estimation, hypotheses testing and confidence interval construction for the rate ratio between the second negative response given an initial negative response and the initial negative response in two-way contingency tables that contain empty cells with true zero probability.

2. Rate Ratios for 2x2 Contingency Tables with Empty Cells:

We assume that *X* and *Y* are two binary response variables each having – the generic terms - *negative* and *positive* responses and were cross-classified into a 2x2 contingency table. We introduce an empty cell in the off-diagonal cell that corresponds with a positive response to the initial variable *X* and a negative response to the second variable *Y* in the following summary table.

Initial Response (X)	Second Response (Y)		Total
	Negative (Y=0)	*Positive* (Y=1)	
Negative (X=0)	$a(\pi_{11})$	$b\ (\pi_{12})$	$a+b\ (\pi_{1+})$
Positive (X=1)	0	$c\ (\pi_{22})$	$c\ (\pi_{22})$
Total	$a(\pi_{11})$	$b+c\ (\pi_{+2})$	n (1)

Suppose that we have a sample of *n* randomly selected observations from a population of interest, classified on the two response variables, *X* and *Y*. Let $\pi_{ij} = P(X = i, Y = j)$ denote the probability that (*X,Y*) falls in the cell in row *i* and column *j*. The probabilities $\{\pi_{ij}\}$ then form the joint distribution of *X* and *Y*. They satisfy the conditions $0 < \pi_{\mathrm{ij}} < 1;\ (\mathrm{i},\mathrm{j}) = (1,1), (1,2), (2,2)$, $\pi_{21} = 0$ and

$\sum_{i,j} \pi_{ij} = 1$. The marginal distributions are the row and column totals of the joint probabilities. These are denoted by $\{\pi_{i+}\}$for the row variable and $\{\pi_{+j}\}$ for the column variable. The cell counts are denoted by $\{a,b,c\}$ with $a+b+c=n$ denoting the total sample size.

Broadly speaking, in a 2x2 contingency table, the *odds ratio* is defined using the joint probabilities as:

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \qquad (1)$$

The *risk ratio* is however, defined as the ratio of "risks" for the two groups. In the case of a 2x2 contingency table with an empty cell where $\pi_{21} = 0$, the odds ratio and the risk ratio are unidentifiable. The rate ratio however, is defined as the ratio of the "negative" responses for the two groups. Let π_{11} denote the probability of the negative initial negative second response, π_{12} denote the probability of the negative initial positive second response and π_{1+} denote the probability of negative initial response. The rate ratio between the initial negative response and the second negative response is defined as π_{11}/π_{1+}. For intensive discussion and use of the rate ratio, odds ratio and risk ratio see Chan *et. al.* (2003).

However, inference will be made here on the rate ratio between the second negative response given an initial negative response and the initial negative response.

Using the above terms, the probability of the second negative response given an initial negative response is $P(\mathrm{Y}=0 \mid \mathrm{X}=0) \; = P(\mathrm{Y}=0,\mathrm{X}=0)\,/\,\mathrm{P}(\mathrm{X}=0) \; = \pi_{11}/\,\pi_{1+}$, and the probability of the initial negative response is $P(\mathrm{X}=0) \, = \, \pi_{1+}$. Therefore the rate ratio δ can be expressed as the proportion of the two probabilities as follows:

$$\delta = \frac{P(Y=0 \mid X=0)}{P(X=0)} = \frac{\pi_{11}}{\pi_{1+}^{2}} \qquad (2)$$

In this paper, our discussion will be limited to the rate ratio δ given in (2). The issues of estimating the rate ratio δ, using the observed cell counts, hypothesis testing and confidence interval construction procedures will be discussed. Accuracy of the test statistics and the confidence intervals will be assessed using simulation methods; and the results will be illustrated using real data sets.

3. Estimation of the Rate Ratio:

In the underlying case and using the notations of the above summary table, the likelihood function of the numbers of occurrences *a* and *b* among *n* cases can be written as:

$$f(a,b\,|\,\pi_{11},\pi_{12}) = \frac{\mathrm{n!}}{\mathrm{a!b!(n - a - b)!}}\,\pi_{11}^{\mathrm{a}}\,\pi_{12}^{\mathrm{b}}\,(1-\pi_{11}-\pi_{12})^{n-a-b} \qquad (3)$$

Now, from equation (2) we can easily see the equalities $\pi_{12} = \sqrt{\frac{\pi_{11}}{\delta}} - \pi_{11}$ and $\pi_{22} = 1 - \sqrt{\frac{\pi_{11}}{\delta}}$ hence we observe that $0 < \pi_{11} < min(\delta, 1/\delta)$. Consequently, the likelihood function of the numbers of occurrences *a* and *b* among *n* cases can be written as:

$$L(\pi_{11},\delta) = \frac{\mathrm{n!}}{\mathrm{a!b!(n - a - b)!}}\,\pi_{11}^{\mathrm{a}}\left(\sqrt{\frac{\pi_{11}}{\delta}} - \pi_{11}\right)^{\mathrm{b}}\left(1 - \sqrt{\frac{\pi_{11}}{\delta}}\right)^{n-a-b} \qquad (4)$$

The first issue in this paper is the estimation of the rate ratio δ. Using the maximum likelihood principles we could estimate both parameters, δ and π_{11}, over the entire parameter space, by maximizing the log likelihood function as follows:

$$log\ L(\pi_{11},\delta) = log(\frac{n!}{a!b!(n-a-b)!}) + a.log(\pi_{11}) + b.log\left(\sqrt{\frac{\pi_{11}}{\delta}} - \pi_{11}\right) + (n-a-b)log\left(1 - \sqrt{\frac{\pi_{11}}{\delta}}\right)$$

(5)

$$\frac{\partial\, log\ L(\pi_{11},\delta)}{\partial\pi_{11}} = \frac{a}{\pi_{11}} + \frac{b\left(\frac{1}{2\sqrt{\pi_{11}\delta}} - 1\right)}{\sqrt{\frac{\pi_{11}}{\delta} - \pi_{11}}} - \frac{\frac{(n-a-b)}{2\sqrt{\pi_{11}\delta}}}{1 - \sqrt{\frac{\pi_{11}}{\delta}}} = 0$$

(6)

$$\frac{\partial\, log\ L(\pi_{11},\delta)}{\partial\delta} = -\frac{b\left(\frac{\sqrt{\pi_{11}}}{2\delta\sqrt{\delta}} - 1\right)}{\sqrt{\frac{\pi_{11}}{\delta} - \pi_{11}}} - \frac{\frac{(n-a-b)\sqrt{\pi_{11}}}{2\delta\sqrt{\delta}}}{1 - \sqrt{\frac{\pi_{11}}{\delta}}} = 0$$

(7)

Assuming that $\delta \neq \pi_{11}$ and $\delta \neq 1/\pi_{11}$ the equations yield:

$$\hat{\delta} = \frac{(n-a)^2\,\hat{\pi}_{11}}{\left((n-a-b)\hat{\pi}_{11} + b\right)^2}$$ and

(8)

$$(n+a+b)\sqrt{\hat{\delta}}\hat{\pi}_{11} - (2a\hat{\delta} + 2b\hat{\delta} + a + n)\sqrt{\hat{\pi}_{11}} + (2a + b)\sqrt{\hat{\delta}} = 0$$

(9)

Formula (9) can be rewritten as:

$$A\hat{\pi}_{11}^{2} + \mathrm{B}\,\hat{\pi}_{11} \;+\; C \;=\; 0 \qquad \text{where}$$

$$\mathrm{A} = \;2\mathrm{bn}(\mathrm{n} - \mathrm{a} - \mathrm{b}) \;\;;$$

$$\mathrm{B} \;=\; -\,2\mathrm{b}(\mathrm{n} + \mathrm{a})(\mathrm{n} - \mathrm{a} - \mathrm{b}) \;\;;$$

$$\mathrm{C} \;=\; 2\mathrm{ab}(\mathrm{n} - \mathrm{a} - \mathrm{b})$$

(10)

The solution of the above quadratic equation yields the following MLE estimator for the probability π_{11} and the rate ratio δ:

$$\hat{\pi}_{11} \;=\; \frac{a}{n} \qquad \text{and} \quad \hat{\delta} \;=\; \frac{an}{(a+b)^{2}}$$

(11)

The above estimators of π_{11} and δ can also be obtained through estimating π_{11} and π_{12} using equation (2) and the likelihood function in (3) (Okasha & Al-Krunz , 2000).

4. Testing of Hypothesis Concerning the Rate Ratio :

For testing the null hypothesis $H_0 : \delta = \delta_0$ versus different alternatives, where δ is the rate ratio that takes the form defined in (2) above, relevant tests which had been intensively studied by many authors including Lui (1998) are the Wald's test which takes the form:

$$T_1 = T_1(a,b,\delta_0) = \frac{na - \delta_0(a+b)^2}{\sqrt{na(n-a)}}$$

(12)

and the logarithmic transformation test which takes the form :

$$T_2 = T_2(a,b,\delta_0) = \frac{\sqrt{n}\left(log(na) - 2\,log(a+b) - log(\delta_0)\right)}{\sqrt{(n-a)/a}}$$

(13)

The above test statistics are the core of the present paper and will be used in making inference on the rate ratio. For the purpose of testing the null hypothesis $H_0 : \delta = \delta_0$, we will define $\tilde{t}_j = \tilde{t}_j(a,b,\delta_0)$, $j = 1,2$ as the observed values of each of the above test statistics T_1 and T_2 computed using the observations (a,b). In the subsequent sections, different possible procedures for estimating the *p*-value for testing the null hypothesis H_0 versus different possible alternatives and confidence interval estimation based also on the same test statistics T_1 and T_2 will be discussed and modifications will be proposed.

In the hypothesis testing problem, assuming that the sample size (*n)* is sufficiently large, it is well known that under the null hypothesis H_0 both Wald's test statistic T_1 and the logarithmic transformation test statistic T_2 asymptotically follow the standard normal distribution. The asymptotic *p*-values which will be referred as $p_j^{As}(a,b)$ can then be given as follows:

$$p_j^{As}(a,b)=\begin{cases} 2\times min(\Phi(\tilde{t}_j),1-\Phi(\tilde{t}_j)) & \text{if} \quad \mathrm{H_A} : \delta \neq \delta_0 \\ \Phi(\tilde{t}_j) & \text{if} \quad \mathrm{H_A} : \delta < \delta_0 \\ 1\text{-}\Phi(\tilde{t}_j) & \text{if} \quad \mathrm{H_A} : \delta > \delta_0 \end{cases} \tag{14}$$

for $j=1,2$. The term Φ(.) refers to the standard normal distribution function and $\tilde{t}_j=\tilde{t}_j(a,b,\delta_0)$ are the estimated test statistics using the observed frequencies (a,b) and the null value of the rate ratio δ_0 .

The asymptotic method is computationally simple and performs well for large sample size (*n*) as the distributions of t_1 and t_2 are close to the standard normal for $n \geq 50$. An empirical study showed that the asymptotic *p*-values are very often close to the simulated p-values given π_{11} and δ_0 and the normal probability plots of the simulated cases showed that there is no evidence against the normality of the above tests. If however, the sample size is too small or the data have a sparse structure, the asymptotic tests will not be suitable and the true *p*-value will always be greater than the pre-specified nominal level. In the case of too small sample size and under the null hypothesis $H_0 : \delta = \delta_0$, following Suissa and Shuster (1985) Tang and Tang (2002) proposed *exact* methods where the nuisance parameter π_{11} could be eliminated by maximizing the null likelihood over the complete domain of π_{11}. It should be noted here that the monotonicity property of the convexity assumption described in Hsueh *et. al.* (2001), Okasha & Al-Krunz ((2000) and Bindslev (1997); is not preserved by either T_1 or T_2. This could be demonstrated through examples by trying different values for n, a, b and δ_0. For the two one-sided alternative hypotheses, the maximization is conducted on the entire nuisance parameter space under the null hypothesis H_0. Detailed discussion on the maximum likelihood estimators for inverse problems with nuisance parameters can be found in Bindslev (1997).

Thus, Tang and Tang (2002) defined the exact p-value $P_j^{Ex}(a,b)$ as follows:

$$P_j^{Ex}(a,b) = \begin{cases} \sup\limits_{0 \le \pi_{11} < min(\delta_{0},1/\delta_0)} \left\{ P(\left|T_j\right| \ge \left|\tilde{t}_j\right| \mid \delta_0, \pi_{11}) \right\} & \text{if} \quad \mathrm{H_A} : \delta \ne \delta_0 \\ \sup\limits_{\delta \ge \delta_0} \left\{ \sup\limits_{0 \le \pi_{11} \le min(\delta,1/\delta)} \left\{ P(T_j \le \tilde{t}_j \mid \delta_0, \pi_{11}) \right\} \right\} & \text{if} \quad \mathrm{H_A} : \delta < \delta_0 \\ \sup\limits_{\delta < \delta_0} \left\{ \sup\limits_{0 \le \pi_{11} \le min(\delta,1/\delta)} \left\{ P(T_j \ge \tilde{t}_j \mid \delta_0, \pi_{11}) \right\} \right\} & \text{if} \quad \mathrm{H_A} : \delta > \delta_0 \end{cases} \tag{15}$$

where:

$$P(T_j \le \tilde{t}_j \mid \delta_0, \pi_{11}) = \sum_{\substack{(a,b) \in \Omega \\ \text{such that } \mathrm{T_j} \le \tilde{t}_j}} \frac{n!}{a!\,b!\,(n-a-b)!} \pi_{11}^{a} \left(\sqrt{\frac{\pi_{11}}{\delta_0}} - \pi_{11} \right)^{b} \left(1 - \sqrt{\frac{\pi_{11}}{\delta_0}} \right)^{n-a-b} \tag{16}$$

and $\Omega = \left\{ (a,b);\ \ 0 \le a,b \le n \ \text{ and } \ 0 \le \mathrm{a} + \mathrm{b} \le \mathrm{n} \right\}$ for $\mathrm{j} = 1,2$.

The term *exact* here refers to the utilization of an exact distribution in calculating the *p*-value of an observation. Tang and Tang (2002) suggested an approximate method to eliminate the nuisance parameter π_{11} through estimating its value at its corresponding maximum likelihood estimate under the null hypothesis $H_0 : \delta = \delta_0$. Assuming that $\tilde{\pi}_{11}$ is the value that maximizes the log null likelihood function in (5), they concluded that, if $B^2 - 4AC \geq 0$, then $\sqrt{\tilde{\pi}_{11}}$ is the smaller root of $Ax^2 - Bx + C = 0$ and if $B^2 - 4AC < 0$, then the log likelihood is an increasing function of $\sqrt{\pi_{11}}$ and hence $\tilde{\pi}_{11} = min\ (\delta_0, 1/\delta_0)$. The approximate p-value $P_j^{Ap}(\tilde{a}, \tilde{b})$ is then defined as:

$$P_j^{Ap}(a,b) = \begin{cases} P(|T_j| \geq |\tilde{t}_j| \,|\, \delta_0,\ \pi_{11} = \tilde{\pi}_{11}) & \text{if} \quad \mathrm{H_A} : \delta \neq \delta_0 \\ P(T_j \leq \tilde{t}_j |\ \delta_0,\ \pi_{11} = \tilde{\pi}_{11}) & \text{if} \quad \mathrm{H_A} : \delta < \delta_0 \\ P(T_j \geq \tilde{t}_j |\ \delta_0,\ \pi_{11} = \tilde{\pi}_{11}) & \text{if} \quad \mathrm{H_A} : \delta > \delta_0 \end{cases} \tag{17}$$

It should be observed here that, while type I error rates of the asymptotic and exact tests are always less than or equal to the pre-specified nominal level, approximate methods may sometimes have error rates greater than the pre-specified nominal level. It is known that, for any given test statistic, particularly the two test statistics given in (12) and (13) denoted by $T_j(j = 1,2)$ and for any method of estimating the p-value, we reject H_0 at nominal level α if $P_j^i(a,b) \leq \alpha$ with i = *As, Ex, Ap*; where *As*, *Ex* and *Ap* refer to the above asymptotic, exact and approximate tests respectively.

Now, for the above hypothesis testing problem, we propose using the same method of Tang and Tang (2002) with the utilization of the maximum likelihood estimators $\hat{\pi}_{11}$ and $\hat{\delta}$, that we derived and in result (11) above, in computing the *p*-values for the Wald's test T_1 and its logarithmic transformation T_2. Let $\tilde{t}_j = \tilde{t}_j(a,b,\delta_0);\ j = 1,2$ be the observed values of the test statistics T_1 and T_2 estimated using the cell frequencies *a* and *b* as well as the null rate ratio δ_0. The modified *p*-values $P_j^M(a,b)$ for the two tests can then be computed as:

$$P_j^M(a,b) = \begin{cases} P(\,|T_j| \geq |\tilde{t}_j(a,b,\delta_0)|\,\big|\,\delta = \hat{\delta}, \pi_{11} = \hat{\pi}_{11}); & \text{if} \quad \mathrm{H_A}: \delta \neq \delta_0 \\ P(T_j \leq \tilde{t}_j(a,b,\delta_0)\,|\,\delta = \hat{\delta}, \pi_{11} = \hat{\pi}_{11}); & \text{if} \quad \mathrm{H_A}: \delta < \delta_0 \\ P(T_j \geq \tilde{t}_j(a,b,\delta_0)\,|\,\delta = \hat{\delta}, \pi_{11} = \hat{\pi}_{11}); & \text{if} \quad \mathrm{H_A}: \delta > \delta_0 \end{cases}$$

(18)

where the above probabilities can be computed using the exact distribution of the cell frequencies *a* and *b*, given the estimated probability of the first cell $\hat{\pi}_{11}$ and the estimated rate ratio $\hat{\delta}$, using the form below:

$$P\left(T_j \le \tilde{t}_j(a,b,\delta_0) \,\middle|\, \delta = \hat{\delta}, \pi = \hat{\pi}_{11}\right) = \sum_{\substack{(k,l)\in\Omega \text{ such that} \\ \mathrm{T_j} \le \tilde{t}_j(a,b,\delta_0)}} \frac{n!}{k!\,l!\,(n-k-l)!}\pi_{11}^{k}\left(\sqrt{\frac{\hat{\pi}_{11}}{\hat{\delta}}}-\hat{\pi}_{11}\right)^{l}\left(1-\sqrt{\frac{\hat{\pi}_{11}}{\hat{\delta}}}\right)^{n-k-l} \tag{19}$$

for $\mathrm{j} = 1,2$; where $\Omega = \{(k,l);\ 0 \le k,l \le n \ \ \& \ \ 0 \le k+l \le \mathrm{n}\}$. Empirical results showed that the proposed method works well for small sample size and produces roughly similar results as well as being computationally easier than Tang's exact methods which require very heavy and sophisticated computing. Moreover, for large sample size with n>50 the asymptotic method may be applied using formula (14) instead of formulae (18) and (19) for estimating the p-values where π_{11} and δ are estimated using the maximum likelihood principles that as in (11).

The test procedures described above can be modified further by utilizing the conditional distribution of the observations a and b given the marginal probability at $\pi_{1+} = \hat{\pi}_{1+}$. In practice this probability is very often known since we usually know the marginal totals and percentages. Mehta & Patel (1997) followed a similar approach in computing exact p-values for some nonparametric tests in contingency tables. The result of this produces the conditional *p*-value $P_j^C(a,b)$ which can then be defined as:

$$P_j^C(a,b) = \begin{cases} P(\,|T_j| \geq |\tilde{t}_j(a,b,\delta_0)|\,\big|\,\delta = \hat{\delta}, \pi_{1+} = \hat{\pi}_{1+}); & \text{if} \quad \mathrm{H_A}: \delta \neq \delta_0 \\ P(T_j \leq \tilde{t}_j(a,b,\delta_0)\,|\,\delta = \hat{\delta}, \pi_{1+} = \hat{\pi}_{1+}); & \text{if} \quad \mathrm{H_A}: \delta < \delta_0 \\ P(T_j \geq \tilde{t}_j(a,b,\delta_0)\,|\,\delta = \hat{\delta}, \pi_{1+} = \hat{\pi}_{1+}); & \text{if} \quad \mathrm{H_A}: \delta > \delta_0 \end{cases}$$

(20)

where the maximum likelihood principles are used for estimating π_{1+} and δ as $\hat{\pi}_{1+} = (a+b)/n$ and $\hat{\delta} = na/(a+b)^2$. Hence it can be easily observed that $\hat{\pi}_{1/1} = \hat{\delta}\hat{\pi}_{1+} = \dfrac{a}{a+b}$. Thus the *p*-values may be computed using the following cumulative conditional distribution:

$$P(T_j \leq t_j \,\big|\, \delta = \hat{\delta}, \pi_{1+} = \hat{\pi}_{1+})$$

$$= \sum_{\substack{(k,l)\in\Omega \text{ such that} \\ \mathrm{k+l=m} \,\&\, \mathrm{T_j} \leq t_j}} \frac{P(T_j = t_j, \delta = \hat{\delta} \,|\, \pi_{1+} = \hat{\pi}_{1+})}{P(\delta = \hat{\delta} \,|\, \pi_{1+} = \hat{\pi}_{1+})}$$

$$= \sum_{\substack{(k,l)\in\Omega \text{ such that} \\ \mathrm{k+l=m} \,\&\, \mathrm{T_j} \leq t_j}} \left\{ \frac{n!}{k!\,l!\,(n-k-l)!} \left(\hat{\delta}\hat{\pi}_{1+}^2\right)^k \left(\hat{\pi}_{1+} - \hat{\delta}\hat{\pi}_{1+}^2\right)^l \left(1 - \hat{\pi}_{1+}\right)^{n-k-l} \Big/ \right.$$

$$\left. \frac{n!}{(k+l)!\,(n-k-l)!} \left(\hat{\pi}_{1+}\right)^{k+l} \left(1 - \hat{\pi}_{1+}\right)^{n-k-l} \right\}$$

$$= \sum_{\substack{(k,l)\in\Omega \text{ such that} \\ \text{k+l=m \& } \mathrm{T_j} \le t_j}} \left\{ \frac{(k+l)!}{k!\,l!} \left(\hat{\delta}\hat{\pi}_{1+}\right)^k \left(1-\hat{\delta}\hat{\pi}_{1+}\right)^l \right\}$$

$$= \sum_{\substack{k\in\Omega \text{ such that} \\ \mathrm{T_j} \le t_j}} \binom{m}{k} \left(\hat{\pi}_{1/1}\right)^k \left(1-\hat{\pi}_{1/1}\right)^{m-k}$$

(21)

for $\mathrm{j}=1,2$; where $\Omega = \{(k,l);\ 0 \le k,l \le n \ \ \& \ \ 0 \le k+l \le \mathrm{n}\}$.

Here $\hat{\pi}_{1/1}$ refers to the maximum likelihood estimator of the conditional probability of the second negative response given the initial negative response where $\hat{\pi}_{1/1} = a/(a+b)$ as above.

From the above discussion we may conclude that the conditional *p*-value $P_j^C(a,b)$ can be estimated using the conditional probability of the initial negative response given the second negative response, through the cumulative binomial distribution with the parameters $(m,\pi_{1/1})$ where m is the first marginal total and the maximum likelihood principles were used to estimate $\pi_{1/1}$.

5. Confidence Intervals Estimation of the Rate Ratio:

For the problem of confidence intervals estimation, Lui (1998) proposed several test-based confidence intervals for the rate ratio δ . The most important of them are the Wald's test based confidence interval using T_1 in formula (12) and on its logarithmic transformation test using T_2 in formula (13). However, all of the test-based confidence intervals, discussed in Lui (1998), had been established using the large sample theory. Furthermore, Tang and Tang (2002) proposed several other test-based confidence intervals for the rate ratio δ. Agresti (2001) discussed the appropriateness of some exact methods, primarily relating to their conservative nature because of discreteness. In the present section more reliable procedures for constructing test-based confidence intervals for δ based on the estimated probability of the first cell $\hat{\pi}_{11}$ and the estimated rate ratio $\hat{\delta}$ that we derived in (11) are proposed.

To discuss the large-sample asymptotic method, let $\hat{\pi}_{11} = a/n$ and $\hat{\pi}_{1+} = (\mathrm{a}+\mathrm{b})/\mathrm{n}$, then $\hat{\delta} = \hat{\pi}_{11} / \hat{\pi}_{1+}^2 = na/(a+b)^2$, $\hat{\sigma}_1^2 = \hat{\pi}_{11}(1-\hat{\pi}_{11})/\hat{\pi}_{1+}^4$ and $\hat{\sigma}_2^2 = (1-\hat{\pi}_{11})/\hat{\pi}_{11}$ where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the standard errors of $\hat{\delta}$ and $log(\hat{\delta})$ respectively. Lui (1998) proposed the following two 100(1-α)% test based confidence intervals for δ :

The Wald's test based confidence interval (based on T_1) :

$$\left[max\left(\hat{\delta} - z_{\alpha/2}\hat{\sigma}_1 / \sqrt{n}\ ,\ 0\right),\ \hat{\delta} + z_{\alpha/2}\hat{\sigma}_1 / \sqrt{n} \right]$$ and :

the logarithmic transformation based confidence interval (based on T_2) :

$$\left[exp\left\{log(\hat{\delta}) - z_{\alpha/2}\hat{\sigma}_2 / \sqrt{n}\right\},\ exp\left\{log(\hat{\delta}) + z_{\alpha/2}\hat{\sigma}_2 / \sqrt{n}\right\} \right] \qquad (22)$$

where $z_{\alpha/2}$ is the upper $(100*\alpha/2)^{th}$ percentile of the standard normal distribution. For the small-sample case, test-based exact confidence intervals for the rate ratio δ has been established by Tang and Tang (2002) following a method proposed by Chan and Zhang (1999). Following this method, the 100(1-α)% exact confidence intervals, based on the statistics $T_j\ (j = 1{,}2)$ for the rate ratio δ, are given by $\{l_{Exl,j}, l_{Exu,j}\};\ (j = 1{,}2)$, where :

$$l_{Exl,j} = \inf_{\delta}\left\{\delta :\ \sup_{0 \le \pi_{11} \le min(\delta, 1/\delta)} P(T_j \ge t_j \mid \delta, \pi_{11}) > \frac{\alpha}{2}\right\} \quad ;$$

$$l_{Exu,j} = \sup_{\delta}\left\{\delta :\ \sup_{0 \le \pi_{11} \le min(\delta, 1/\delta)} P(T_j \le t_j \mid \delta, \pi_{11}) > \frac{\alpha}{2}\right\} . \qquad (23)$$

A slight modification to the above method has been proposed by Chen (2002) and Agresti and Min (2001) to be used for setting exact confidence intervals for the difference of two independent binomial proportions. Tang and Tang (2002) suggested that 100(1-α)% confidence intervals based on the tests $T_j \ \ (j=1,2)$ for δ may be constructed as $\{l_{Eul,j}, l_{Euu,j}\}$, where :

$$l_{Eul,j} = \inf_{\delta}\left\{\delta : \sup_{0\le\pi_{11}\le min(\delta,1/\delta)} P(|T_j| \ge |t_j| \,|\delta,\pi_{11}) > \alpha\right\} \quad ;$$

$$l_{Euu,j} = \sup_{\delta}\left\{\delta : \sup_{0\le\pi_{11}\le min(\delta,1/\delta)} P(|T_j| \ge |t_j| \,|\delta,\pi_{11}) > \alpha\right\} .$$

(24)

According to Agresti and Min (2001), Chen (2002), and Tang and Tang (2002) the nuisance parameter π_{11} could be again eliminated by evaluating its value at its corresponding maximum likelihood estimate. The maximization was however, conducted on the entire nuisance parameter space under $H_0 : \delta = \delta_0$. The result was the maximum likelihood estimate $\tilde{\pi}_{11}$ of π_{11} as $\tilde{\pi}_{11} = min\ (\delta_0, 1/\delta_0)$. Doing this, a 100(1-α)% approximate confidence interval based on the tests $T_j \ \ (j=1,2)$ for δ is given by $\{l_{Apl,j}, l_{Apu,j}\}$, where :

$$l_{Apl,j} = \inf_{\delta}\{\delta : P(|T_j| \ge |t_j| \,|\delta,\tilde{\pi}_{11}) > \alpha\} \quad ;$$

$$l_{Apu,j} = \sup_{\delta}\left\{\delta : P(|T_j| \geq |t_j| \,|\, \delta, \tilde{\pi}_{11}) > \alpha\right\} . \qquad (25)$$

The above confidence intervals are claimed to perform well with very high coverage probabilities but they include tedious computing and severe conservativeness in the sense that they produce error rates very much higher than the α-level. For n>50 the methods require highly sophisticated computer programming.

Empirical studies on the above procedures for constructing exact confidence intervals showed that, in addition to the fact that the methods are computationally tedious, the resulting confidence intervals are unnecessarily long with confidence intervals that guarantee very much greater than 100(1-α)% coverage probability. This is because of the discrete nature of a and b and hence $\hat{\delta}$ particularly when the sample size is too small. But the main reason for that phenomenon is the method of elimination of the nuisance parameter π_{11} over its parameter space under the null hypothesis $H_0 : \delta = \delta_0$ as $min(\delta_0, 1/\delta_0)$which showed that it is a biased estimate for π_{11}.

Once again, we can utilize the maximum likelihood estimate of the rate ratio $\hat{\delta}$ and eliminate the nuisance parameter π_{11} using its maximum likelihood estimate $\hat{\pi}_{11}$ as derived in (11). The resulting modified 100(1-α)% confidence intervals based on the tests $T_j(\, j = 1,2\,)$ for δ would then be $\{l_{Ml,j}, l_{Mu,j}\}$, where :

$$l_{Ml,j} = \inf_{\delta}\left\{\delta : P(T_j \geq t_j \,|\, \delta = \hat{\delta}, \pi_{11} = \hat{\pi}_{11}) > \frac{\alpha}{2}\right\} \quad ;$$

$$l_{Mu,j} = \sup_{\delta}\left\{\delta : P(T_j \leq t_j \,|\, \delta = \hat{\delta}, \pi_{11} = \hat{\pi}_{11}) > \frac{\alpha}{2}\right\} \; .$$

(26)

where the probabilities could be computed using the following distribution:

$$P(T_j \leq t_j \,|\, \hat{\delta}, \hat{\pi}_{11}) =$$

$$\sum_{\substack{(k,l)\in\Omega \\ \text{such that } \mathrm{T_j} \leq t_j}} \frac{n!}{k!\,l!\,(n-k-l)!}\hat{\pi}_{11}^{k}\left(\sqrt{\frac{\hat{\pi}_{11}}{\hat{\delta}}} - \hat{\pi}_{11}\right)^{l}\left(1-\sqrt{\frac{\hat{\pi}_{11}}{\hat{\delta}}}\right)^{n-k-l}$$

(27)

for $j = 1,2$; where $\Omega = \{(k,l);\;\; \text{and} \;\; 0 \leq k + l \leq \mathrm{n}\}$.

An empirical study carried out by the author and described in the next section showed that the modified confidence intervals in (26) proved to be shorter than Tang's exact confidence intervals and guarantee at least 100(1-α)% coverage probability particularly when n is moderate to large. The modified confidence intervals performed better than those of the exact ones that were based on utilizing $\tilde{\pi}_{11}$ instead which are proposed by Tang and Tang (2002).

The above principles of modifying confidence intervals could be applied on the conditional distribution of the observations *a* and *b* given the marginal probability at $\pi_{1+} = \hat{\pi}_{1+}$. The result of applying this approach produces the following 100(1-α)% conditional confidence interval estimate $\{l_{Cl,j}, l_{Cu,j}\}$ for the rate ratio δ:

$$l_{Cl,j} = \inf_{\delta} \left\{ \delta : P(T_j \geq t_j \middle| \delta, \hat{\pi}_{1+}) > \frac{\alpha}{2} \right\} \quad ;$$

$$l_{Cu,j} = \sup_{\delta} \left\{ \delta : P(T_j \leq t_j \middle| \delta, \hat{\pi}_{1+}) > \frac{\alpha}{2} \right\}$$

(31)

where π_{1+} and δ are estimated using the maximum likelihood principles as $\hat{\pi}_{1+} = (a+b)/n$ and $\hat{\delta} = na/(a+b)^2$ and the probabilities in the expression are estimated using the cumulative binomial distribution, as derived in (24), and given by:

$$P(T_j \leq t_j \middle| \hat{\delta}, \hat{\pi}_{1+}) = \sum_{\substack{k \in \Omega \text{ such that } k \leq m \\ \& \, T_j(k|\hat{\pi}_{1+}, \hat{\delta}) \leq t_j}} \binom{m}{k} (\hat{\pi}_{1/1})^k (1 - \hat{\pi}_{1/1})^{m-k}$$

(32)

for $j = 1,2$; where $\Omega = \{k;\ 0 \le k \le m \le n\}$ and $m = a + b$.

It is clear that the use of the binomial distribution in the present method of estimating confidence intervals is much easier than the use of the multinomial distribution and requires much less computing time.

6. Assessing the tests and confidence interval construction procedures :

Now, we assess the performance of the various tests and confidence intervals discussed in the previous sections, for the rate ratio δ in a 2x2 contingency table with an empty cell, using simulation methods. For this purpose, 2000 samples of different values of a, b, n and δ_0, using the distribution of the numbers of occurrences a and b among n cases in (4), had been generated in each run. Various tests for the null hypothesis $H_0 : \delta = \delta_0$ versus different possible alternatives, using different values of δ_0 , were examined and various confidence intervals for the rate ratio δ were estimated. Results of all the above methods and of the simulation results are compared. To assess the performance of various confidence intervals for δ two assessment measures are considered which are: the coverage probability measure:

$$\phi_{(a,b,n)} = \sum_{(a,b,n)} I\left(\delta \in [l(a,b,n),u(a,b,n)]\right) f(a,b|\pi_{11},\delta)$$

(33)

and the expected interval width:

$$\psi_{(a,b,n)} = \sum_{(\mathrm{a,b,n})} \left(u(a,b,n) - l(a,b,n)\right) f(a,b|\pi_{11},\delta)$$

(34)

where $[l(a,b,n),u(a,b,n)]$ is the confidence interval based on the observed frequencies *a*, *b* and *n* and $I\left(\delta \in l(a,b,n),u(a,b,n)\right)$ is an indicator function of the event $\{\delta \in [l(a,b,n),u(a,b,n)]\}$. The notations *l(a,b,n)* and *u(a,b,n)* are the lower and upper confidence limits as estimated using a specific method.

All the procedures were implemented using the R code (R, 2004) using a PC. For a moderate sample size, estimating confidence intervals and conducting various tests for a given data set, using the asymptotic and simulation methods, with 2000 samples, could be conducted within a few seconds and so as the modified and the conditional methods. They require reasonably less computing time than Tang's exact procedures that could take several minutes. Some of the assessment results are reported in table (1) for different values of π_{11}, δ and sample size *n*.

Table 1 : Coverage probabilities and expected interval widths for 95% confidence intervals based on different values of *n*, *a* and *b* using different methods and the statistic (T_1).$^{(1)}$

			Simulation Method	Asymptotic method		Tang's Exact method		Modified method		Conditional Method	
n	π_{11}	δ	$\psi^{(2)}$	$\varphi^{(3)}$	$\psi^{(4)}$	$\varphi^{(3)}$	$\psi^{(4)}$	$\varphi^{(3)}$	$\psi^{(4)}$	$\varphi^{(3)}$	$\psi^{(4)}$
10	0.01	0.25	2.5	95.6	1.7	100	39.7	100.0	10.0	98.7	5.0
	0.04	1.0	10.0	93.4	4.3	99.0	46.9	100.0	10.0	94.6	5.0
	0.06	1.5	10.0	91.9	5.2	99.5	51.1	100.0	10.0	91.9	5.0
	0.023	0.25	2.5	95.3	1.2	100	19.8	100.0	10.0	99.0	3.3
	0.09	1.0	10.0	93.5	3.0	97.7	28.5	100.0	10.0	94.3	3.3
	0.135	1.5	10.0	89.8	3.9	99.2	53.4	100.0	10.0	89.8	3.3
	0.063	0.25	2.2	93.0	0.8	99.8	4.2	98.8	2.2	98.2	1.6
	0.25	1.0	3.3	89.1	1.9	98.6	8.7	98.4	3.3	89.1	1.9
	0.375	1.5	4.4	90.8	2.4	99.8	11.2	99.3	5	77.7	1.5
	0.16	0.25	0.8	92.9	0.6	98.3	1.0	98.7	0.8	97.6	0.8
	0.64	1.0	1.0	93.9	0.9	100	1.9	98.6	1.7	82.8	0.6
25	0.01	0.25	2.8	92.8	1.1	100	9.9	99.9	6.3	99.1	4.0
	0.04	1.0	6.3	91.0	2.7	99.3	15.0	99.1	6.3	95.1	3.8
	0.06	1.5	8.3	90.2	3.5	97.8	18.3	98.3	6.3	91.1	3.7
	0.023	0.25	1.6	91.4	0.8	100	2.9	99.2	2.1	98.6	1.9
	0.09	1.0	4.0	92.6	2.1	98.2	5.7	98.1	3.1	92.8	2.2
	0.135	1.5	5.6	89.4	3.1	98.1	7.5	98.0	4.6	85.0	2.4
	0.063	0.25	0.9	94.8	0.6	99.3	1.0	99.4	1.0	99.4	1.0
	0.25	1.0	1.7	92.8	1.3	99.4	2.3	96.4	1.9	93.8	1.5
	0.375	1.5	1.8	93.0	1.5	99.6	2.9	96.1	2.4	88.1	1.1
	0.16	0.25	0.5	94.4	0.4	97.2	0.6	99.0	0.6	98.9	0.6
	0.64	1.0	0.6	94.2	0.6	100	1.0	95.7	1.3	85.6	0.4
50	0.01	0.25	1.9	91.6	0.9	100	2.6	99.8	3.1	99.6	3.0
	0.04	1.0	4.2	92.5	2.2	98.1	4.8	99.2	4.1	97.7	3.4
	0.06	1.5	5.5	94.1	3.0	98.0	6.2	97.9	4.1	94.2	3.3
	0.023	0.25	1.0	94.3	0.7	99.9	1.3	99.8	1.6	99.7	1.6
	0.09	1.0	2.2	92.6	1.6	98.4	2.8	98.4	2.4	96.9	2.2
	0.135	1.5	2.6	93.5	2.0	99.1	3.6	96.2	2.8	91.0	2.0
	0.063	0.25	0.6	94.9	0.5	97.4	0.7	99.8	0.8	99.7	0.8
	0.25	1.0	1.1	93.7	0.9	99.6	1.5	96.6	1.5	92.7	0.9
	0.375	1.5	1.2	94.0	1.0	99.7	1.8	94.6	1.9	81.1	0.6
	0.16	0.25	0.3	95.0	0.3	98.0	0.4	99.5	0.5	98.7	0.4
	0.64	1.0	0.5	94.4	0.4	99.9	0.7	93.9	0.5	81.6	0.2

(1) Confidence interval estimation methods using the statistic (T_2) produce approximately the same results as the statistic (T_1). Moreover, other Tang's exact methods produce approximately similar results. Therefore, one table is exhibited here in order to simplify comparisons .

(2) Expected interval width for 95% confidence interval using 2000 simulated samples.

(3) Coverage probability as defined in formula (31).

(4) Expected interval width as defined in formula (32).

The values of π_{11}, δ and n were selected for illustration of the results in all situations in the table that would enable comparison between different methods. They are however constrained by the above relation between π_{11}, π_{1+} and δ. Only results of the Wald's test T_1 are reported in the above table as the logarithmic transformation test produces similar results in most cases. Using a similar approach the power functions for nine different test statistics are studied by Berger (1994).

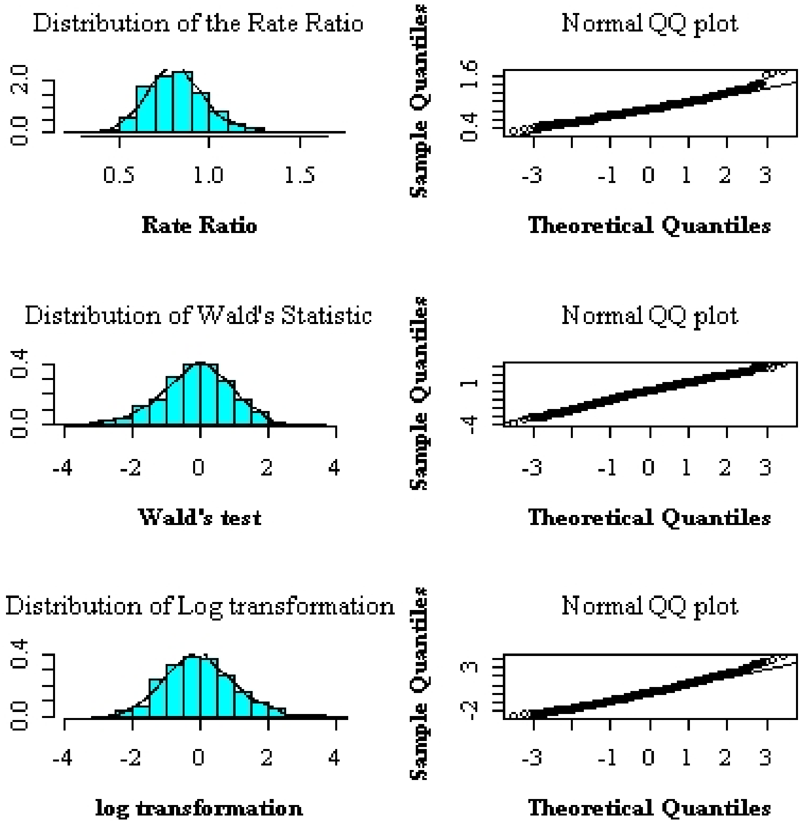
Comparison of the values of the coverage probability φ and the expected interval width ψ for different methods shows that the exact methods proposed by Tang and Tang (2002) are unnecessarily long despite the fact that their coverage probabilities are higher than those of the other methods. The coverage probabilities of the modified methods can be seen clearly slightly below those of the exact methods but their expected interval widths are much smaller than those of the exact methods. The conditional methods (in the last columns in the table) however, produced much shorter confidence intervals than all other methods with coverage probabilities ranging around 95%. An interesting result which can be concluded clearly from the above table is that despite the difference in complications and computing time within different methods, the asymptotic methods for hypotheses testing and confidence intervals construction perform well for moderate and large samples. Conditional confidence intervals indicated that they are reliable in the sense that they are very short and their coverage probabilities are just about the required level.

Moreover, the distributional properties of the estimated rate ratio $\hat{\delta}$ and the test statistics T_1 and T_2 are examined for samples of small and moderate sizes using simulation methods. The results showed that in the majority of cases the distributions of the test statistics T_1 and T_2 are very close to the standard normal distribution. This result supports the above results that the asymptotic methods for hypotheses testing and confidence intervals construction perform well in moderate to large samples and to some extent in small samples. An example showing histograms and the normal probability plots of the estimated rate ratio $\hat{\delta}$ and the test statistics T_1 and T_2 is exhibited in figure (1) for a moderate sample size (n=30).

7. Application of the methods to real data sets :

To illustrate the above methods, we applied them on some real published and unpublished data sets. The first data set is the two-step tuberculosis testing data reported in Toyota et al (1999) and revisited by Tang and Tang (2002). In their study they reported that a=22, b=8, c=4. For this data set, we estimated the rate ratio $\hat{\delta}$ at 0.831. Using the statistic T_1 the 95% confidence interval estimates based on the simulation, asymptotic, Tang's small sample exact, the modified and the conditional methods are given by [0.62, 1.05], [0.62, 1.04], [0.59, 1.18], [0.65, 1.01] and [0.68, 0.98] respectively. As expected, the interval width based on Tang's small sample exact method is relatively wider than those of the other methods. Moreover, based on the above hypothesis testing procedures and using the statistic T_1 for testing the null hypothesis $H_0 : \ \delta = \delta_0$ the p-values are given by 0.066, 0.0543, 0.072, 0.057 and 0.036 respectively. Based on those results, Tang and Tang (2002) concluded that there is no evidence of a booster effect from the data because all their resulting confidence intervals contain the value 1.0 and all their p-values are greater than 0.05. However, using our conditional method we may conclude a different result.

Figure (1): An example showing histograms and normal QQ plots of the simulated rate ratio $\hat{\delta}$ and the test statistics T_1 and T_2 (for n=30).



In other examples we applied the methods on real data sets from a family planning survey in the Gaza Strip where we have samples of men and women classified by their knowledge and attitudes towards family planning and their practicing status of at least one family planning method. It is expected that all persons who do not know about or have negative attitudes towards family planning would not practice it. Therefore, we have many two-way contingency tables with empty cells. We concentrate on three of those tables with different sample sizes. The overall sample size in the survey is 348 but we apply the methods on the sub-samples of males and females with university level education. The sizes of sub-samples are 35 and 61 for females and males respectively. The three data sets for women, men and the overall sample are (19, 1 and 15), (39, 3 and 19) and (143, 46 and 159) respectively for (positive attitude and applied, positive attitude and never applied and negative attitude and never applied any family planning method). We start by estimating the rate ratio for women with a university level education with a sample size of 35. For this data set, the estimated rate ratio $\hat{\delta}$ is 1.66. Using the statistic T_1 the 95% confidence interval estimates based on the simulation, asymptotic, Tang's small sample exact, the modified and the conditional methods are given be [1.28, 2.33], [1.23, 2.25], [1.16, 2.33], [1.39, 2.33] and [1.49, 1.75] respectively. Based on the above hypothesis testing procedures and using the statistic T_1 for testing the null hypothesis $H_0 : \ \delta = \delta_0$ all the p-values in this case are close to zero. For the males' data set, the estimated rate ratio $\hat{\delta}$ is 1.35 and the 95% confidence interval estimates based on the above methods are given be [1.14, 1.68], [1.09, 1.60], [0.95, 1.63], [0.95, 1.60] and [1.21, 1.45] respectively. All the p-values for tests of the null hypothesis $H_0 : \delta = \delta_0$ using the statistic T_1 in this case are well below 0.05 despite the fact that the small sample exact and the modified confidence intervals do not contain the value 1.0. For the third data set, since the sample size is too big, confidence intervals estimation based on Tang's small sample exact and the modified methods seem to be impossible to compute using a PC. For such a large sample the estimation requires a mainframe computer. The estimated rate ratio $\hat{\delta}$ for this example is 1.39 and the 95% confidence interval estimates, using the statistic T_1 and based on the simulation, asymptotic and the conditional methods are given be [1.24, 1.58], [1.22, 1.57] and [1.32, 1.47] respectively. All the p-values for tests of the null hypothesis $H_0 : \delta = \delta_0$ using the statistic T_1 in this case are close to zero. The conclusion which could be drawn from the above results is that the rate ratios between people who apply at least one family planning method among those who have positive attitudes towards family planning methods in the society and people who have positive attitudes towards family planning are significantly greater than 1. This could be interpreted as the probability of applying at least one family planning method among people who have positive attitude towards family planning is significantly greater than the probability of having positive attitude towards family planning in the society. In other words, this also means that, people who have positive attitude towards family planning are more likely that they will apply family planning methods in the Palestinian society in the Gaza Strip.

8. Conclusions :

In the present article, we showed that the confidence interval estimates based on the asymptotic methods produce the shortest expected interval widths particularly when the sample size is small. However, this advantage can be penalized by their inability to achieve the desired confidence levels. If one needs to guarantee a lower coverage probability, confidence intervals based on Tang's small-sample exact methods or the modified methods provide more reliable performance. The conditional methods however, have been shown to approximately achieve the desired confidence levels and lower bounds on the coverage probability. Moreover, while Tang's small-sample exact methods require so heavy computing that they are hard to estimate using a PC the asymptotic methods and the conditional methods could be estimated easily.

Empirical study showed that, asymptotic methods of hypothesis testing produce actual errors usually greater than the α-level but Tang's small-sample exact methods and the modified methods produce actual errors always less than the α-level. However, the conditional methods of hypothesis testing produce actual error rates very much close to the α-level. Thus, based on our empirical study we would recommend the conditional methods of hypothesis testing for small and moderate samples and the asymptotic methods for large samples.

The Tang's small-sample exact procedures described in this article were implemented using a C program written by Tang and Tang (2002) with their kind permission. All other procedures were implemented using R codes (R Development Core Team; 2004) on a Pentium 4 PC, which are available from the author on request.

9. References :

1. Agresti, A.; “Exact inference for categorical data; recent advances and continuing controversies”; *Statist. Med.* **20**; (2001); 2709-2722.
2. Agresti, A. and Min Y.; “On small sample confidence intervals for parameters in discrete distributions”; *Biometrics*, **57**; (2001); 963-971.

3. Agresti, A.; “An Introduction to Categorical Data Analysis”; Wiley Series in Probability and Statistics; John Wiley & Sons; New York; (1996).

4. Agresti, A.; “Categorical Data Analysis”; Wiley; New York; (1990).

5. Berger, R. L.; “Power Comparison of Exact Unconditional Tests for Comparing Two Binomial Proportions”; Institute of Statistics Mimeo Series No. 2266; North Carolina; (1994).

6. Bindslev, H.; “Maximum likelihood estimators for inverse problems with nuisance parameters”; JET Joint Understanding, JET-R(97)13; Oxfordshire; (1997).

7. Bishop, Y. V. V., Fienberg, S. E. and Holland P. W.; “Discrete Multivariate Analysis”; MIT Press; Cambridge, MA; (1975).

8. Chan , I. S. F., Tang, N. S., Tang, M. L. and Chan, P. S. (2003); “Statistical Analysis of Noninferiority Trials with a Rate Ratio in Small-Sample Matched-Pair Designs”; *Biometrics*; **59**; (1975); pp. 1170.

9. Chan, I. S. F. & Zhang Z.; “Test-based exact confidence intervals for the difference of two binomial parameters”; *Biometrics*; **55**; (1999); 1202-1209.

10. Chen, X; “A guasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases”; *Statistics in Medicine*; **21**; (2002); 943-956.

11. Hsueh, H. M., Liu, J. P., and Chen, J. J.; “Unconditional exact tests for equivalence or noninferiority for paired binary endpoints”; *Biometrics*; **57**; (2001); 478-483.

12. Lui, K. J.; “Interval estimation of the risk ratio between a secondary infection, given a primary infection”; *Biometrics*; **54**; (1998); 706-711.

13. Mehta, C. R. & Patel, N. R.; “Exact Inference for Categorical Data”; Unpublished research paper; Harvard University and Cytel Software Corporation; (1997).

14. Okasha, M. K. and Al-Krunz, S. M.; (In Arabic) “Mathematical Statistics”; Al-Quds Open University; Jerusalem; (2000); (Sec. 4 of Ch. 2).

15. R Development Core Team; “R: A language and environment for statistical computing”; R Foundation for Statistical Computing; Vienna, Austria; (2004).

16. Suissa, S. and Shuster, J. J.; “Exact unconditional sample size for the 2x2 binomial trials”; *Journal of the Royal Statistical Society, Series A*; **148**; (1985); 317-327.

17. Tang, N. and Tang, M.; “Exact Unconditional Inference for Risk Ratio in a correlated 2x2 Table with Structural Zero”; *Biometrics*, **58**; (2002); 972-980.

18. Toyota, M., Kudo, K., Sumiya, M., and Kobori, O.; "High frequency of individuals with strong reactions to tuberculosis among clinical trainees*"; Japanese Journal of Infectious Disease*; **52**; (1999); 128-129.