Proceedings of The 6th Sino-International Symposium on Probability, Statistics, and Quantitative Management
Conference held on May 30, 2009 at Fo Guang Univ., Taiwan, ROC

Date published: **October 31, 2009**
pp. 2-15

Confidence Intervals for a Ratio of Binomial Proportions Based on Unbiased Estimators

Kamon Budsaba / *Thammasat University*
Thuntida Ngamkham / *Thammasat University*
Andrei Volodin / *University of Western Australia*
Igor Volodin / *Kazan State University*

ABSTRACT A general statement of the interval estimation problem of two proportions ratio according to data from two independent samples is considered. Each sample may be obtained in the framework of direct or inverse binomial sampling. Five asymptotic confidence intervals are constructed in accordance with different types of sampling schemes. Main probability characteristics of intervals are investigated by the Monte Carlo method: coverage probability, median, expectation and standard deviation of intervals length. The results of the simulations are presented in tables and some recommendations for an application of each of the intervals obtained is presented. Sufficiently complete review of the literature for the problem is also presented.

Keywords Confidence limits; Ration of binomial proportions; Inverse binomial sampling; Asymptotic confidence limits.

1. Introduction

Generally speaking, the problem we are solving can be formulated in the following way. Let $X_1, X_2, \cdots$ and $Y_1, Y_2, \cdots$ be two independent Bernoulli sequences with probabilities of success p_1 and p_2 respectively. Observations are done in sequential schemes of samplings with Markov's stopping times ν_1 and ν_2. According to the results of the observations

□ Received May 2009, presented as an Invited Speech on May 30, 2009, revised September 2009.
□ Kamon Budsaba and Thuntida Ngamkham are affiliated to the Department of Mathematics and Statistics at Thammasat University, Rangsit Center, Pathumthani 12121, Thailand. Andrei Volodin (corresponding author) is a Professor in the School of Mathematics and Statistics at the University of Western Australia, Crawley, Perth, WA 6009 Australia; email: andrei@maths.uwa.edu.au. Igor Volodin is affiliated to the Department of Mathematical Statistics at Kazan State University, Kazan 420008, Russia.
□ AMS 2000 subject classifications: Primary 62F25; Secondary 62F12, 62L12.

$X^{(\nu_1)} = (X_1, \ldots, X_{\nu_1})$ and $Y^{(\nu_2)} = (Y_1, \ldots, Y_{\nu_2})$ it is required to construct a confidence interval for the parametric function $\theta = p_1/p_2$.

Up to our knowledge, this statistical problem has been solved only for schemes of sampling with a fixed number of observations $\nu_i = n_i,\ i = 1, 2$, while an unbiased estimation of odds ratio with inverse binomial sampling was used in Roberts (1993).

A complexity of the problem stated can be explained by two reasons. The first is the absence of uniformly most powerful test for hypothesis testing $\theta = \theta_0$ with one-sided or two-sided alternative in the case of an arbitrary hypothetical value θ_0. As it is known (see, for example Lehmann (1984), Section 4.5), the uniformly most powerful unbiased test exists for the values of the cross-product ratio $\rho = p_1(1 - p_2)/p_2(1 - p_1)$, but this is not what we need. Hence, it seems to be impossible to use the standard method of uniformly most accurate confidence boundaries construction based on an acceptance region of the corresponding test, and other tests should be applied or the method of pivot functions with an additional estimation of the nuisance parameter should be used.

The second, but not less important difficulty that precludes from the pivot functions with good accuracy properties construction, is the absence of an unbiased estimation for the parametric function $1/p$ for Bernoulli trials with the fixed sample size n (see Lehmann (1998), Chapter 2, Section 1, Example 1.2; general theory of unbiased estimation is presented in the monograph Voinov and Nikulin (1993)). But if the inverse, not direct binomial sampling method is used, then such unbiased estimation exists and this is the starting point for our investigation on the confidence limits construction for a ratio of probabilities of success.

Now we provide a brief discussion of the literature pertaining to this subject in order to compare our results with already known.

The first easily computed methods of confidence estimation θ have been suggested Noether (1957) and Guttman (1958). A review of these early methods may be found in Sheps (1959). Methods of confidence estimation of the ratio of proportions as a diagnostic test that can detect a disease, are used in McNeil et al. (1975).

Next, some methods based on the corresponding tests for significance have been developed. For example, Thomas et al. (1977) suggests to apply the method based on fixed marginals in the two-by-two tables for confidence interval construction. Santher et al. (1980) develops and generalizes this method and suggested three related exact methods for finding such intervals.

Katz et al. (1978) suggests three methods of lower confidence limit for θ, and the limits are defined as solutions of some equations. Numerical comparison shows that the method in which the logarithmic transformation is applied to the ratio of estimates of probabilities is preferential. Some modifications of these methods, that take their origin in Fiellers method, are discussed in Bailey (1987).

Santher et al. (1980) derives exact intervals for the risk ratio from Cornfield's (1956) confidence interval for the odds ratio.

Koopman (1984), and Miettinen and Nurminen (1985) proposed methods based on asymptotic likelihood for hypothesis $\theta = \theta_0$ testing with the alternative $\theta \neq \theta_0$. In Koopman (1984) this method compared with the one recommended by Katz et all (1978).

All the results until the end of 80s were summarized in Gart and Nam (1988). In this paper they provide a comprehensive survey of various approximation methods of confidence limits constructions for the ratio of probabilities based on the properties of goodness of fit with Pearson's chi-square test, invariance, universality of an application for all observations and computational simplicity. Also, asymptotic methods were improved by taking into account the asymptotic asymmetry of statistics (see also Gart and Nam (1990)). The results obtained are extended for the case of estimating the common ratio in a series of two-by-two tables, which was considered before in Gart (1985). Extensive numerical illustrations are provided, which allow to compare accuracy properties of the methods of interval estimation of probabilities ratio. Instead of iterative algorithms for calculating the approximate confidence intervals that have been provided by Koopman (1984), Gart and Nam (1988a), Nam (1995) gives the analytical solutions for upper and lower confidence limits in closed form.

For an interval estimator construction, Bedrick (1987) used the special power divergent family of statistics. Intervals based on inverting the Pearson, likelihood-ratio, and Freeman-Tukey statistics are included in this family. Asymptotic efficiency, coverage probability, and expected interval length are investigated. Comparisons of methods are provided by numerical examples.

The bootstrap method of a confidence interval construction for θ is suggested in Kinsella (1987).

Coe and Tamhane (1993) provided a method for small sample confidence intervals construction for the difference of probabilities, based on an extension of known Sterne's method for constructing small sample confidence intervals for a single success probability. Modifications of the algorithm for ratio probabilities are also indicated.

Nam and Blackwelder (2002) developed a superior alternative to the Walds interval and gave corresponding sample size formulas. Bonett and Price (2006) proposed alternatives to the Nam-Blackwelder confidence interval based on combining two Wilson score intervals. Two sample size formulas are derived to approximate the sample size required to achieve an interval estimate with desired confidence level and length.

Extensive numerical illustrations for comparison of exact and asymptotic methods for θ confidence intervals constructions are presented in the thesis by Mukhopadhyay (2003).

We construct asymptotic confidence intervals for a few schemes of direct and inverse sampling and illustrate their characteristics by the results of statistical modeling (see Tables). In each cell of tables the following characteristics are presented: actual confidence level (the nominal confidence level is chosen to be 0.095), median, expectation, and standard deviation of the length of a corresponding confidence interval. For each interval 10 000 random numbers with

Bernoulli and/or Negative Binomial distributions were generated with parameters (probabilities of success) $p_2 \geq p_1$ $(= 0.1\,(0.1)\,(0.9))$. The tables presented contain only a part of the results for the probability values 0.1 (0.2) (0.9) and not for all values of sample sizes n and m, but the conclusions (see the last section) are made according to all obtained results of statistical modeling.

2. Confidence Limits with Using Direct and Inverse Binomial Sampling Methods

Sufficiently simple method of asymptotic confidence limits construction for the ratio of probabilities $\theta = p_1/p_2$ exists in the case when the stopping moment for observations from the Bernoulli sequence with success probability p_1 are priory fixed ($\nu_1 = n$), that is, the observations are done in the framework of a direct binomial sampling, while observations from the sequence with success probability p_2 are done as an inverse binomial sampling, that is, the stopping time ν is defined by the number of the observation that results in achieving $m\ (\geq 1)$ successes.

The likelihood function of the random samples $\left(X^{(n)},\, Y^{(\nu)}\right)$ depends on the components of these samples only through the values of complete sufficient statistics $\left(\sum_1^n X_k,\, \nu\right)$. The distribution of the statistics $T = \sum_{k=1}^n X_k$ follows binomial law $B(n,\, p_1)$, and the distribution of ν follows Pascal law $P(m,\, p_2)$. It is well known, the statistic $\overline{X}_n = T/n$ has the expected value $\mu_1 = p_1$, variance $\sigma_1^2 = p_1(1-p_1)/n$, and it is asymptotically ($n \to \infty$) normal with parameters $(\mu_1,\, \sigma_1^2)$. Statistic $\overline{Y}_m = \nu/m$ has the expected value $\mu_2 = 1/p_2$, variance $\sigma_2^2 = (1-p_2)/mp_2^2$, and it is asymptotically ($m \to \infty$) normal with parameters $(\mu_2,\, \sigma_2^2)$.

Hence (see Lehmann (1998), Chapter 2, Section 1), $\hat{\theta}_{n,m} = \overline{X}_n\overline{Y}_m$ is an unbiased estimation of probabilities ratio θ such that uniformly by all values of $p_1,\ p_2$ it minimizes any risk function with convex loss function and it is asymptotically ($n, m \to \infty$) normal with mean $\mu = \theta$ and variance

$$\sigma^2 = \frac{p_1(1-p_1)}{p_2^2 n} + \frac{p_1^2(1-p_2)}{p_2^2 m} = \theta\left[\frac{p_2^{-1}-\theta}{n} + \frac{\theta - p_1}{m}\right]. \tag{1}$$

The last statement immediately follows from the following easy to prove lemma.

Lemma 1 *Let X_n be asymptotically ($n \to \infty$) normal $(\mu_1,\, \sigma_1^2/n)$ and Y_m be asymptotically ($m \to \infty$) normal $(\mu_2,\, \sigma_2^2/m,)$, then $X_n \cdot Y_m$ is asymptotically ($n,\, m \to \infty$) normal with parameters $\mu = \mu_1\mu_2$ and $\sigma^2 = \mu_2^2\sigma_1^2/n + \mu_1^2\sigma_2^2/m$.*

Proof. Introduce a normalized random variable

$$Z_{n,m} = \frac{X_n - \mu_1}{\sigma_1}\sqrt{n} \cdot \frac{Y_m - \mu_2}{\sigma_2}\sqrt{m},$$

which under simultaneous limits n and m to infinity has a nondegenerate distribution. Then

$$X_n \cdot Y_m = Z_{n,m} \cdot \frac{\sigma_1 \sigma_2}{\sqrt{nm}} + X_n\,\mu_2 + Y_m\,\mu_1 - \mu_1\,\mu_2.$$

Hence, by Slutsky's theorem, the asymptotic distribution of $X_n \cdot Y_m$ coincides with asymptotic distribution of $X_n\,\mu_2 + Y_m\,\mu_1 - \mu_1\,\mu_2$. □

The results obtained, allows us to the state the following theorem.

Theorem 1 *If $n, m \to \infty$, then an asymptotic $(1-\alpha)$-confidence region (interval) for the parametric function θ as defined by the inequality*

$$\left|\theta - \hat{\theta}_{n,m}\right| \le \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\theta\left(\frac{\overline{Y}_m - \theta}{n} + \frac{\theta - \overline{X}_n}{m}\right)}. \tag{2}$$

The interval with bounds

$$\hat{\theta}_{n,m} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\overline{X}_n \overline{Y}_m \left(\frac{\overline{Y}_m(1 - \overline{X}_n)}{n} + \frac{\overline{X}_n(\overline{Y}_m - 1)}{m}\right)}, \tag{3}$$

is an asymptotically $(1-\alpha)$-confidence interval for θ.

Proof. The statements follow from the asymptotic normality of the estimate$\hat{\theta}_{n,m}$. If in the right hand side of formula (1) for the asymptotic variance of the estimate we change p_1 and p_2^{-1} on their consistent estimates $\overline{X}_n$ and $\overline{Y}_m$ respectively, then we obtain the asymptotically confident region (2). If additionally in (1) we change θ on its estimate $\hat{\theta}_{n,m}$, then we obtain the confident interval (3). □

Note that the left and right bounds of interval (3) are the asymptotically lower and upper $(1-\alpha/2)$-confidence bounds for the parametric function θ.

The important part of the suggested plan realization of the estimate of θ is the choice of the number m. The (random) sample size for the second sample depends on this number. If a statistician could obtain the same size of sample n which she had in the first sample and moreover has some prior knowledge of the type $p_2 > p_1$, then the following sampling plan for the second stage of the statistical experiment can be suggested. Repeat observations until the same number of successes as in the first experiment, that is set $m = T$. Of course, we consider only the case when the value of T is greater than zero. Then for the estimate of $1/p_2$ it is natural to consider the statistics $\overline{Y}_T = \nu/T$, where the conditional distribution of ν is the Pascal distribution $P(T,\, p_2)$ and the unconditional distribution is obtained by taking the expectation of this distribution by the truncated at zero Binomial distribution T. The estimate of the parameter θ is $\hat{\theta}_n = \nu/n$.

Table 1 (conf. int. (3))

n=50, m=50

	0,9	0,7	0,5	0,3	0,1	P2	
0,1	0,901 0,192 0,191 0,039	0,94 0,279 0,277 0,061	0,955 0,445 0,452 0,116	0,966 0,954 0,997 0,451	0,97 4,71 5,074 2,407		P1
0,3	0,939 0,294 0,291 0,021	0,953 0,414 0,414 0,04	0,963 0,663 0,669 0,093	0,965 1,421 1,45 0,316		0,526 0,077 0,076 0,031	0,9
0,5	0,932 0,31 0,309 0,011	0,908 0,418 0,42 0,033	0,897 0,66 0,665 0,086		0,764 0,222 0,216 0,036	0,891 0,188 0,186 0,014	0,7
0,7	0,89 0,268 0,262 0,03	0,759 0,317 0,303 0,072		0,902 0,472 0,475 0,042	0,914 0,297 0,297 0,016	0,93 0,218 0,219 0,005	0,5
0,9	0,628 0,125 0,126 0,051		0,976 1,026 1,035 0,157	0,965 0,475 0,477 0,047	0,953 0,292 0,292 0,019	0,951 0,207 0,206 0,009	0,3
P1		0,991 3,49 3,622 1,216	0,986 0,688 0,695 0,139	0,973 0,321 0,322 0,054	0,955 0,197 0,196 0,029	0,935 0,136 0,135 0,019	0,1
	P2	0,1	0,3	0,5	0,7	0,9	

n=100, m=100

n=50, m=100

	0,9	0,7	0,5	0,3	0,1	P2	
0,1	0,839 0,136 0,133 0,026	0,859 0,194 0,193 0,039	0,89 0,317 0,315 0,071	0,939 0,678 0,689 0,237	0,97 3,396 3,556 1,325		P1
0,3	0,829 0,203 0,2 0,011	0,832 0,276 0,274 0,017	0,869 0,441 0,441 0,04	0,918 0,966 0,974 0,145		0,8 0,152 0,146 0,042	0,9
0,5	0,789 0,208 0,205 0,008	0,74 0,256 0,249 0,031	0,73 0,385 0,374 0,07		0,893 0,379 0,379 0,038	0,965 0,283 0,282 0,017	0,7
0,7	0,702 0,161 0,155 0,033	0,41 0,149 0,148 0,063		0,961 0,722 0,728 0,096	0,975 0,45 0,453 0,037	0,988 0,317 0,319 0,012	0,5
0,9	0,661 0,109 0,11 0,049		0,984 1,463 1,496 0,327	0,988 0,69 0,695 0,099	0,992 0,426 0,428 0,042	0,991 0,295 0,295 0,018	0,3
P1		0,987 4,77 5,153 2,344	0,994 0,968 0,992 0,259	0,99 0,455 0,46 0,09	0,987 0,281 0,281 0,046	0,981 0,195 0,193 0,028	0,1
	P2	0,1	0,3	0,5	0,7	0,9	

n=100, m=50

Lemma 2 *If $n \to \infty$, then the estimate $\hat{\theta}_n$ is asymptotically normal with the mean $\mu = \theta$ and variance $\sigma^2 = \theta\left(2p_2^{-1} - \theta - 1\right)/n$.*

Proof. The characteristic function of Pascal's distribution $P(m,\ p_2)$ (the distribution of ν given $T = m$) is $\varphi_m(t) = \lambda^m(t)$, where

$$\lambda(t) = \frac{p_2\, \mathrm{e}^{\mathrm{i}t}}{1 - (1 - p_2)\, \mathrm{e}^{\mathrm{i}t}}.$$

Under the assumption that T has truncated at zero binomial distribution, the characteristic function of the unconditional distribution of ν takes the form

$$\begin{aligned}\varphi(t) &= \frac{1}{1-(1-p_1)^n} \cdot \sum_{i=1}^{n} \binom{n}{i} [\, p_1\, \lambda(t)\,]^i (1-p_1)^{n-i} \\ &= \frac{[\, p_1\, \lambda(t) + (1-p_1)\,]^n - (1-p_1)^n}{1-(1-p_1)^n}.\end{aligned}$$

□

The statement of the lemma follows now from the Taylor expansion of the function $\varphi(t)$. The lemma immediately implies the following result.

Theorem 2 *If $n \to \infty$, the asymptotic $(1-\alpha)$-confidence interval for the parametric function θ is defined by the inequality*

$$\left|\theta - \hat{\theta}_n\right| \le \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\theta}{n}\, \left(2\overline{Y}_T - \theta - 1\right)}. \tag{4}$$

The interval bounded by the points

$$\hat{\theta}_n \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\hat{\theta}_n}{n}\left(2\overline{Y}_T - \hat{\theta}_n - 1\right)}, \tag{5}$$

Is the asymptotically $(1-\alpha)$*-confidence interval for* θ.

Characteristics of this interval are presented in Table 2.

Table 2 (Confidence interval (5))

n=30

	0,9	0,7	0,5	0,3	0,1	P2	
0,1	0,826 0,214 0,246 0,088	0,855 0,323 0,347 0,167	0,834 0,5 0,532 0,293	0,816 0,868 0,969 0,583	0,783 2,63 2,882 1,896		
0,3	0,939 0,382 0,381 0,073	0,905 0,513 0,551 0,151	0,897 0,82 0,848 0,263	0,896 1,485 1,523 0,506		0,953 0,255 0,257 0,046	P1 0,9
0,5	0,934 0,405 0,427 0,063	0,93 0,628 0,637 0,134	0,917 0,977 1 0,238		0,94 0,504 0,51 0,075	0,939 0,319 0,322 0,038	0,7
0,7	0,936 0,407 0,413 0,064	0,93 0,644 0,656 0,125		0,929 0,772 0,78 0,143	0,94 0,49 0,496 0,079	0,945 0,327 0,334 0,037	0,5
0,9	0,943 0,324 0,328 0,077		0,917 1,168 1,188 0,302	0,919 0,648 0,656 0,156	0,932 0,422 0,429 0,089	0,945 0,292 0,297 0,043	0,3
P1		0,848 2,158 2,285 1,139	0,857 0,687 0,723 0,337	0,874 0,389 0,404 0,175	0,883 0,263 0,267 0,103	0,903 0,189 0,19 0,055	0,1
	P2	0,1	0,3	0,5	0,7	0,9	

n=50

n=70

	0,9	0,7	0,5	0,3	0,1	P2	
0,1	0,934 0,157 0,161 0,039	0,9 0,228 0,225 0,073	0,879 0,326 0,342 0,125	0,875 0,585 0,61 0,242	0,867 1,878 1,948 0,814		
0,3	0,94 0,251 0,252 0,031	0,932 0,36 0,363 0,063	0,93 0,551 0,558 0,112	0,931 0,996 1,007 0,213		0,944 0,183 0,183 0,022	P1 0,9
0,5	0,944 0,282 0,283 0,026	0,942 0,419 0,421 0,057	0,935 0,653 0,658 0,101		0,946 0,359 0,361 0,038	0,945 0,228 0,229 0,019	0,7
0,7	0,946 0,272 0,273 0,027	0,943 0,428 0,432 0,054		0,941 0,549 0,552 0,071	0,941 0,35 0,352 0,039	0,946 0,235 0,237 0,018	0,5
0,9	0,948 0,217 0,218 0,032		0,935 0,836 0,842 0,151	0,93 0,463 0,467 0,079	0,939 0,303 0,305 0,044	0,942 0,211 0,211 0,021	0,3
P1		0,888 1,594 1,634 0,555	0,895 0,499 0,513 0,168	0,902 0,283 0,289 0,087	0,909 0,188 0,191 0,05	0,918 0,135 0,135 0,027	0,1
	P2	0,1	0,3	0,5	0,7	0,9	

n=100

3. Confidence Limits with Using Only Direct Binomial Sampling

Consider now the standard situation when a statistician has in his hands only the numbers of success

$$n\overline{X}_n = \sum_1^n X_i, \qquad m\overline{Y}_m = \sum_1^m X_i$$

for two binomial experiments $B(n,\ p_1)$ and $B(m,\ p_2)$ with priory fixed sample sizes n and m. Initially for such type of data, asymptotic confidence intervals were constructed on the bases of the statistic if sample means ratio $\overline{X}_n/\overline{Y}_m$, that is, for $1/p_2$, a biased estimation was explored. Moreover, a problem with its irregular behaviour under the absence of successes in trials $B(m,\ p_2)$ appear. As it has been mentioned in introduction, in this case there is no unbiased estimation of the parametric function $1/p_2$. But it is possible to construct an estimation of this function that has exponentially small for $m \to \infty$ value of a bias.

Let $X_1, \ldots, X_n$ be a sample in Bernoulli scheme with success probability p, and $T = \sum_1^n X_i$. For a construction of an estimate $\hat{\theta}_n$ of the parametric function $\theta = 1/p$, we apply the statistic ν, which equals to the number of the last trial with $X_\nu = 1$. Then, by the analogy with the inverse binomial sampling, it is natural to suggest the statistic $\hat{\theta}_n = \nu/T$ as the estimate of θ. But the value of ν in our case is unknown, so it is better to use the projection $\theta_n^* = \theta_n^*(T) = \mathbf{E}\{\hat{\theta}_n \,|\, T\}$ of this statistic on the sufficient statistic T. As it is known, (see Lehmann (1998), Chapter 2, Section 1), a projection does not cause an increase of the risk if the loss function is convex.

Lemma 3 *The projected estimator has the following representation* $\theta_n^* = (n+1)/(T+1)$ *and its mean value is*

$$\mathbf{E}\,\theta_n^*(T) = \frac{1}{p}\left(1 - (1-p)^{n+1}\right).$$

Proof. The joint distribution of statistics ν and T is defined by the probabilities

$$\mathrm{pr}(\nu = k,\, T = t) = \begin{cases} 0, & \text{if} \quad k = 0,\, t \ge 1, \\ (1-p)^n, & \text{if} \quad k = 0,\, t = 0, \\ \binom{k-1}{t-1} p^t (1-p)^{n-t}, & \text{if} \quad t = 1, \ldots, n,\, k = t, \ldots, n. \end{cases}$$

The marginal distribution of statistic T is

$$\mathrm{pr}(T = t) = \binom{n}{t} p^t (1-p)^{n-t}, \quad t = 0, 1, \ldots, n,$$

then the conditional distribution

$$\mathrm{pr}(\nu = k \,|\, T = t) = \begin{cases} 0, & \text{if} \quad k = 0,\, t \ge 1, \\ 1, & \text{if} \quad k = 0,\, t = 0, \\ \binom{k-1}{t-1}/\binom{n}{t}, & \text{if} \quad t = 1, \ldots, n,\, k = t, \ldots, n. \end{cases}$$

All further calculations for mean values are trivial, if we use the well known combinatorial formula

$$\sum_{k=1}^{N} \binom{n+k}{n} = \binom{n+N+1}{n+1}.$$

□

It follows from the lemma proved above that for an estimate of the parametric function $\theta = p_1/p_2$, it is appropriate to take the statistic

$$\hat{\theta}_{n,m} = \frac{\overline{X}_n\,(m+1)}{m\overline{Y}_m + 1},$$

with mean value

$$\mathbf{E}\,\hat{\theta}_{n,m} = \theta\left(1 - (1-p_2)^{m+1}\right).$$

The next theorem provides two kinds of asymptotic confidence intervals for θ.

Theorem 3 *If $n, m \to \infty$, then an asymptotic $(1-\alpha)$-confident region (interval) for the parametric function θ is defined by the inequality*

$$\left| \theta - \hat{\theta}_{n,m} \right| \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\theta \left(\frac{(m+1)(1-\overline{X}_n)}{n(m\overline{Y}_m + 1)} + \theta \frac{(m+1)(1-\overline{Y}_m)}{m(m\overline{Y}_m + 1)} \right)}. \tag{6}$$

The interval with bounds

$$\hat{\theta}_{n,m} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{\theta}_{n,m} \left(\frac{(m+1)(1-\overline{X}_n)}{n(m\overline{Y}_m + 1)} + \theta \frac{(m+1)(1-\overline{Y}_m)}{m(m\overline{Y}_m + 1)} \right)}, \tag{7}$$

is an asymptotically $(1-\alpha)$-confident interval for θ.

Proof. By the analogy with the proof of Theorem 1, the current proof follows from the following asymptotic representation (standard technique of asymptotic normality parameters' calculations for a ratio of two asymptotically normal estimates is used):

$$\hat{\theta}_{n,m} = \frac{\overline{X}_n}{p_2} - \frac{1}{p_2}\left[XY\sqrt{\frac{p_1 p_2 (1-p_1)(1-p_2)}{nm}} + p_1(\overline{Y}_m - p_2) + O_{p_2}\left(\frac{1}{m^2}\right) \right],$$

where

$$X = \frac{\overline{X}_n - p_1}{p_1(1-p_1)}\sqrt{n}, \quad Y = \frac{\overline{Y}_m - p_2}{p_2(1-p_2)}\sqrt{m}\,. \qquad \square$$

Table 3 (Confidence interval (7))

n=50, m=50 (upper part); n=100, m=100 (lower part)

	0,9	0,7	0,5	0,3	0,1	P2	
0,1	0,884 0,184 0,181 0,038	0,903 0,239 0,239 0,056	0,9 0,338 0,348 0,103	0,899 0,594 0,647 0,363	0,857 2,085 2,704 2,308		P1
0,3	0,94 0,286 0,286 0,027	0,937 0,388 0,394 0,06	0,938 0,593 0,615 0,14	0,917 1,151 1,244 0,465		0,952 0,183 0,184 0,023	0,9
0,5	0,942 0,32 0,323 0,024	0,95 0,464 0,475 0,072	0,945 0,769 0,801 0,191		0,948 0,359 0,364 0,045	0,952 0,222 0,223 0,016	0,7
0,7	0,952 0,313 0,315 0,032	0,951 0,507 0,518 0,091		0,95 0,549 0,56 0,092	0,948 0,332 0,336 0,036	0,949 0,227 0,228 0,012	0,5
0,9	0,945 0,258 0,259 0,048		0,941 0,832 0,869 0,214	0,946 0,422 0,431 0,067	0,943 0,277 0,279 0,029	0,944 0,202 0,203 0,013	0,3
P1		0,908 1,56 1,806 1,046	0,927 0,426 0,443 0,114	0,929 0,242 0,246 0,047	0,93 0,168 0,17 0,027	0,935 0,13 0,129 0,018	0,1
	P2	0,1	0,3	0,5	0,7	0,9	

n=50, m=100 (upper part); n=100, m=50 (lower part)

	0,9	0,7	0,5	0,3	0,1	P2	
0,1	0,885 0,183 0,18 0,036	0,894 0,236 0,235 0,051	0,909 0,335 0,337 0,082	0,912 0,575 0,595 0,211	0,891 1,924 2,175 1,241		P1
0,3	0,941 0,284 0,283 0,021	0,935 0,375 0,376 0,041	0,943 0,553 0,56 0,086	0,939 1,021 1,057 0,252		0,948 0,224 0,226 0,047	0,9
0,5	0,946 0,312 0,313 0,014	0,94 0,431 0,435 0,039	0,948 0,669 0,681 0,099		0,942 0,438 0,45 0,092	0,955 0,245 0,247 0,032	0,7
0,7	0,945 0,298 0,297 0,021	0,95 0,441 0,444 0,047		0,936 0,664 0,696 0,181	0,949 0,379 0,388 0,07	0,947 0,238 0,24 0,022	0,5
0,9	0,95 0,224 0,223 0,032		0,919 0,997 1,084 0,426	0,945 0,482 0,5 0,121	0,944 0,295 0,302 0,048	0,951 0,206 0,208 0,018	0,3
P1		0,876 1,781 2,363 1,975	0,918 0,468 0,499 0,183	0,927 0,254 0,261 0,064	0,929 0,172 0,174 0,033	0,928 0,131 0,13 0,02	0,1
	P2	0,1	0,3	0,5	0,7	0,9	

Therefore, the confidence interval constructed above is asymptotically equivalent to the interval based on the statistic $\overline{X}_n/\overline{Y}_m$, but the problem that appears when the denominator of the estimate is zero with a positive probability is completely solved, and the estimate with smaller bias is explored. An interested reader may compare with a solution of this problem in the paper Cho (2007); see the beginning of Section 2.)

4. Confidence Limits with Using Only Inverse Binomial Sampling

For the case when both samples are obtained in the schemes $P(m_i, p_i)$, $i = 1, 2$ of the inverse binomial sampling, there exists an unbiased estimate of θ with the uniformly minimal risk for any loss function. Really, for the parametric function $1/p_2$ we have the unbiased estimate ν_2/m_2, and for p_1 under the scheme of inverse sampling for $m_1 \geq 2$ there also exists the unbiased estimate (see Guttman, I. (1958)) $\hat{p}_1 = (m_1 - 1)/(\nu_1 - 1)$. Therefore, the optimal unbiased estimate of $\theta = p_1/p_2$ is

$$\hat{\theta}_{n,m} = \frac{\nu_2 (m_1 - 1)}{(\nu_1 - 1)\, m_2}.$$

We have that $\mathbf{E}\nu_i/m_i = 1/p_i$, $\mathbf{var}(\nu)_i/m_i = (1 - p_i)/m_i\, p_i^2$, $i = 1, 2$, and by the same method of asymptotic analysis for a ratio of two asymptotically normal estimates that we explored in the previous section, we obtain the following theorem.

Theorem 5 *if* $m_i \to \infty$, $i = 1, 2$, *then the interval bounded by the points*

$$\hat{\theta}_{n,m} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{\theta}_{n,m}\left(\frac{\hat{p}_1 (1 - \hat{p}_2)}{m_2} + \hat{\theta}_{n,m}\frac{1 - \hat{p}_1}{m_1}\right)}, \tag{8}$$

where $\hat{p}_i = (m_i - 1)/(\nu_i - 1)$, $i = 1, 2$, *is an asymptotically* $(1 - \alpha)$*-confident interval for* θ.

5. Comparative Analysis of Methods

As it was mentioned in the introduction, we provide more detailed analysis of characteristics of the confidence intervals obtained in the paper. For the direct binomial sampling the sample sizes are chosen to be $n = 30, 50, 100$ for all intervals, except (5), where $n = 30, 50, 70, 100$. For the inverse binomial sampling, in order to obtain a correct comparison of intervals' characteristics, sampling stopped at $m = np_2$ successes, because in this case $E\,\nu = n$. For each fixed values of n, m p_1 and $p_2(\geq p_1)$ a table contains the blocks, in which the following characteristics are provided (from top to bottom): coverage probabilities, length median, expected value, and standard deviation for the confidence interval obtained from the formula with the number presented at the bottom of the table.

Table 4 (Confidence interval (8))

n=50, m=50 (upper part); n=100, m=100 (lower part)

	0,9	0,7	0,5	0,3	0,1	P2	
	0,873	0,868	0,855	0,842	0,756		
	0,163	0,209	0,299	0,498	1,435		
0,1	0,183	0,237	0,337	0,569	1,725		
	0,085	0,112	0,167	0,3	1,137		P1
	0,929	0,92	0,903	0,857		0,94	
	0,28	0,374	0,542	0,926		0,179	
0,3	0,284	0,381	0,554	0,956		0,179	0,9
	0,051	0,078	0,127	0,262		0,019	
	0,937	0,921	0,895		0,923	0,94	
	0,319	0,445	0,667		0,334	0,22	
0,5	0,319	0,449	0,673		0,335	0,221	0,7
	0,031	0,063	0,115		0,029	0,013	
	0,938	0,91		0,903	0,933	0,944	
	0,311	0,475		0,474	0,316	0,226	
0,7	0,313	0,477		0,477	0,317	0,227	0,5
	0,027	0,06		0,057	0,031	0,015	
	0,928		0,873	0,917	0,929	0,941	
	0,254		0,669	0,389	0,269	0,201	
0,9	0,253		0,68	0,393	0,271	0,202	0,3
	0,04		0,131	0,063	0,039	0,025	
P1		0,803	0,882	0,902	0,912	0,917	
		1,116	0,378	0,225	0,159	0,124	
		1,221	0,403	0,239	0,168	0,13	0,1
		0,53	0,141	0,077	0,052	0,04	
	P2	0,1	0,3	0,5	0,7	0,9	

n=50, m=100 (upper part); n=100, m=50 (lower part)

	0,9	0,7	0,5	0,3	0,1	P2	
	0,87	0,871	0,864	0,857	0,803		
	0,162	0,21	0,294	0,489	1,462		
0,1	0,182	0,236	0,332	0,552	1,688		
	0,086	0,111	0,159	0,272	0,971		P1
	0,931	0,925	0,92	0,889		0,923	
	0,277	0,364	0,521	0,877		0,219	
0,3	0,281	0,37	0,53	0,898		0,218	0,9
	0,048	0,068	0,107	0,208		0,038	
	0,937	0,929	0,91		0,912	0,94	
	0,312	0,419	0,611		0,398	0,241	
0,5	0,311	0,419	0,613		0,4	0,242	0,7
	0,025	0,044	0,08		0,053	0,027	
	0,934	0,927		0,878	0,921	0,942	
	0,297	0,42		0,548	0,352	0,236	
0,7	0,298	0,421		0,553	0,354	0,238	0,5
	0,015	0,033		0,089	0,048	0,023	
	0,937		0,837	0,902	0,929	0,94	
	0,222		0,737	0,421	0,283	0,205	
0,9	0,22		0,756	0,428	0,287	0,207	0,3
	0,029		0,178	0,082	0,047	0,028	
P1		0,735	0,863	0,892	0,903	0,915	
		1,144	0,389	0,232	0,163	0,124	
		1,279	0,42	0,246	0,172	0,131	0,1
		0,66	0,161	0,084	0,055	0,04	
	P2	0,1	0,3	0,5	0,7	0,9	

In order to estimate the accuracy of a confidence interval characteristics' calculations, each table was reproduced 10 times. For all cases and for all characteristics, we observed a difference only in the third digit after the decimal point. Hence a calculation error should not exceed 0.01.

We start with analysis of the modeling results (see Table 3) for the confidence interval (7), which is a modification of the classical interval with only direct sampling. After we compare characteristics of other intervals with this "classical" case.

Assuming that the coverage probability error in comparison with the nominal should not exceed 0.025, it is possible to make a certain conclusion about strictly low coverage probability for the values $p_1 < 0.2$ for any values of $p_2 \geq p_1$ and $n \leq 50$ (we remind that we have more detailed tables, based on which we make these conclusions). If p_1 becomes bigger than 0.2, then for all values of p_2 and n, the coverage probability does not differ too much from the nominal level (still lower). The smallest value of the coverage probability always corresponds to the equal proportions ($p_1/p_2 = 1$).

The values of median and expectation of the interval length is practically the same in all blocks of the table. This expectations that the length of the interval is symmetrically distributed. Most probably this follows from the application of unbiased estimators for our results. We interpret the phenomenon of the symmetry of the length distribution as an additional fact that these intervals should be used. Median, expectation, and standard deviation of the length are increasing when the ratio p_1/p_2 increases even in the case when probabilities start to take values more than 0.5. For each fixed value of p_1 (p_2, correspondingly), these characteristics have a tendency to become smaller when the value of p_2 is increasing (p_1, correspondingly).

Now it is appropriate to compare the case of only direct binomial sampling with the case of only inverse binomial sampling for both experiments (see Table 4). Here the area of small differences with the nominal level shrinks significantly. Except poor performing values $p_1 < 0.2$ for all $p_2 \geq p_1$, the coverage probability is low for $p_2 \leq 0.5$ and nearly all sample sizes. The behaviour of all characteristics of the confidence interval (8) is the same as for the interval (7), that it the tendency of decreasing for the coverage probability when the ratio of probabilities increasing is preserved, the values of median and expectation for the length of the interval are not significantly different (symmetry of the length distribution), the behaviour of the median, expected length, and standard deviation under the changes of probabilities and their ratio is similar. But this characteristics of the confidence interval (8) length are somehow better than for the interval (7), hence it is recommended to use this method of interval estimation, but only in the region of large values of p_2 (> 0.5) and p_1 (> 0.3).

Good properties according to the coverage probability (close to the nominal level) and with characteristics of the length very similar to the interval (7), possesses the confidence interval (5) (see Table 2), where the sample size for the second experiment is defined by the number of successes in the first sample. If we exclude the values $p_1 < 0.2$, then practically acceptable correspondence to the nominal level starts from the size $n = 50$ for the first sample. But even for $n = 30$, the interval is still possible to use when $p_1 > 0.3$ and all values $p_2 \geq p_1$.

Now we discuss the confidence interval (3) (see Table 2). Configuration of the region of acceptable values of coverage probabilities is similar to the region for the interval (8), but the region itself is much wider. As before, we should exclude the values $p_1 < 0.2$, but even for $n = 30$, $m = 30p_2$ the interval may be recommended for all $p_2 > 0.3$. The same recommendations are true for all n and $m < 100p_2$ and only for sample sizes $n = 30, 50$ and $m = 100p_2$, the recommended region of the interval (3) applications is increased to the region $p_2 \geq p_1 > 0.1$. Probability characteristics of the interval (3) length are practically the same as characteristics of interval (7).

Hence, if we order the interval according to the size of the regions of p_1 and p_2 values where we could recommend their application, then the order is the following: (5), (3), (7), (8).

If we compare the coverage probabilities of intervals (5) and (7) with known before, then according to very poor numerical illustrations provided in the previously published papers, they are better than the intervals based on tests (such as Koopman (1984), for example), but obviously are inferior to the confidence intervals of a complex construction Gart–Nam (1988) which were specially designed for small sample sizes.

References

[1] Bailey, B. J. R. (1987). Confidence limits to the risk ratio, *Biometrics*, **43**, 201-205.

[2] Bedrick, E. J. (1987). A family of confidence intervals for the ratio of two binomial proportions, *Biometrics*, **43**, 993-998.

[3] Bonett, D. J. and Price, R. M. (2006). Confidence intervals for a ratio of binomial proportions based on paired data, *Statist. Med.*, **253**, 3039-3047.

[4] Cho, H. (2007). Sequential risk-efficient estimation for the ratio of two binomial proportion, *J. Statist. Plann. and Inf.*, **137**, 2336-2346.

[5] Coe, P. K. and Tamhane, A. C. (1993). Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities, *Commun. Statist. Simul. and Comput.*, **22**(4), 925-938.

[6] Cornfield, J. (1956). A statistical problem arising from retrospective studies, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. IV, J. Neyman (ed.), 135-148. Berkeley: University of California Press.

[7] Gart, J. J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of 2×2 variables, *Biometrika*, **72**, 673-677.

[8] Gart, J. J. and Nam, J. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness, *Biometrics*, **44**, 323-338.

[9] Gart, J. J. and Nam, J. (1990). Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension on multiple tables, *Biometrics*, **46**, 637-643.

[10] Guttman, I. (1958). A note on a series solution of a problem in estimation, *Biometrika*, **45**, 565-567.

[11] Katz, D., Baptista, J., Azen, S. P., and Pike, M. C. (1978). Obtaining confidence intervals for the risk ratio in cohort studies, *Biometrics*, **34**, 469-474.

[12] Kinsella, A. (1987). The exact bootstrap approach to confidence intervals for the risk ratio in cohort studies, *The Statistician*, **36**, 345-347.

[13] Koopman, P. A. R. (1984). Confidence limits to the ratio of two binomial proportions, *Biometrics*, **40**, 513-517.

[14] Lehmann, E. L. (1986). Testing Statistical Hypothesis (2nd Edition). New York: Springer-Verlag.

[15] Lehmann, E. L. (1998). Theory of Point Estimation (2nd Edition). New York: Springer-Verlag.

[16] McNeil, B. J., Keeler, E., and Adelstein, S. J. (1975). Primer on certain elements of medical decision making, *New England J. Med.*, **293**, 211-215.

[17] Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates *Statist. Med.*, **4**, 213-226.

[18] Mukhopadhyay, P. (2003). Exact tests and exact confidence intervals for the ratio of two binomial proportion, *Dissertation, Graduate School at NC State University,*. I-XII, pp. 1-168.

[19] Nam, J. (1995). Confidence limits for the ratio of two binomial proportions based on likelihood scores: Non-iterative method, *Biom. J.*, **37**, 375-379.

[20] Nam, J. and Blackwelder W. C. (2002). Analysis of the ratio of marginal probabilities in a matched-pair setting, *Statist. Med.*, 689-699.

[21] Noether, G. E. (1957). Two confidence intervals for the ratio of two probabilities and some measures of effectiveness, *J. Amer. Statist. Assoc.*, **52**, 36-45.

[22] Roberts, C. (1993). Unbiased estimation of odds ratio by using negative binomial plans, *Biom. J.*, **35**, 581-587.

[23] Santher, T. J. and Snell, M. K. (1980). Small sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables, *J. Amer. Statist. Assoc.*, **75**, 386-394.

[24] Sheps, M. C. (1959). An examination of some methods of comparing several rates of proportions, *Biometrics*, **15**, 87-89.

[25] Thomas, D. G. and Gart, J. J. (1977). A table of exact confidence limits for difference and ratios of two proportions and their odds ratios, *J. Amer. Statist. Assoc.*, **72**, 73-76.

[26] Voinov, V. G. and Nikulin, M. S. (1993). Unbiased estimators and their applications, Vol. 1, Univariate case. Mathematics and its Applications, 263. Kluwer Academic Publishers, Dordrecht.