Comparison of Accuracy Properties for Confidence Intervals of the Cross-Product Ratio of Binomial Proportions under Direct-Direct Sampling Scheme

[1]CHANAKAN SUNGBOONCHOO, [2]WARARIT PANICHKITKOSOLKUL, [3]ANDREI VOLODIN

[1,2]Department of Mathematics and Statistics, Thammasat University, Pathum Thani, THAILAND (e-mail: chanakano@hotmail.com).
[3]Department of Mathematics and Statistics, University of Regina, Saskatchewan, CANADA

Abstract: We consider a general problem of the confidence interval for a cross-product ratio $\rho = \frac{p_1(1-p_2)}{p_2(1-p_1)}$ according to data from two independent samples. Each sample may be obtained in the framework of direct Binomial sampling scheme. Asymptotic confidence intervals are constructed in accordance with direct Binomial sampling scheme, with parameter estimators demonstrating exponentially decreasing bias. Our goal is to investigate the cases when the normal approximations (which are relatively simple) for estimators of the cross-product ratio are reliable for the construction of confidence intervals. We use the closeness of the confidence coefficient to the nominal confidence level as our main evaluation criterion, and use the Monte-Carlo method to investigate the key probability characteristics of intervals corresponding to direct Binomial sampling schemes. We present estimations of the coverage probability, expectation and standard deviation of interval widths in tables and provide some recommendations for applying each obtained interval.

Keywords: Asymptotic confidence limits, Cross-product ratio, Delta method, Direct Binomial sampling scheme

Received: September 3, 2020. Revised: February 8, 2021. Accepted: March 10, 2021. Published: March 21, 2021.

1. Introduction

The problem of comparing the success probabilities of Bernoulli trials arises in biological and medical investigations.

In this article, we investigate the accuracy properties for linear and logarithmic asymptotic confidence intervals of the cross-product ratio of Binomial proportions under direct-direct Binomial sample schemes.

In article Ngamkham and Volodins (2016) [1] and PhD thesis Ngamkham (2018) [2], the problem of confidence estimation for the ratio of Binomial proportions was considered. The cross-product ratio statistic is more frequently applied to real data, especially in medical and biological research for analyzing 2×2 contingency tables. This can be explained by its importance for analyzing 2×2 contingency tables; see Lehmann (1959) [3], Section 4.6 on page 143.

Because of that, it is very interesting to investigate statistical inference for the cross-product ratio. There are many authors studied about the cross-product ratio under direct-direct sampling scheme.

Goodman (1964) [4] developed the simple methods of obtaining confidence limits for the cross-product ratio in a 2×2 table, and extended these methods to obtain simultaneous confidence intervals for the $r(r-1)c(c-1)/4$ cross-product ratios in an $r\times c$ table and, likewise, for the relative differences between the corresponding cross-product ratios in K different $r\times c$ tables. Goodman's research indicates an improvement of a method for 2×2 table introduced by Gart (1962) [5], and extends the improved method the $r\times c$ tables. These methods are easier to apply than those given by Cornfield (1956) [6].

Anděl (1973) [7] suggests a method based on logarithmic interactions for comparing the association in k fourfold tables (k independent samples).

Lee (1981) [8] presented the empirical Bayes modification of the cross-product ratio for studying the trend and degree of relationship between two cross-classified factors in a 2×2 contingency table. The Independent Poisson, Product Multinomial, and Multinomial are the three sampling schemes used for determining the cell frequencies in contingency tables. These procedures were studied and compared with the classical procedures, the results indicated that the empirical

Bayes estimation procedures had a lower average squared error than the classical procedures.

Albert and Gupta (1983) [9] investigated the Bayesian approach to the estimation of the cell probabilities for 2×2 and $I\times 2$ tables. In the 2×2 table in which the prior information was declared in terms of the cross-product ratio coefficient. For the $I\times 2$ table they used estimators based on a two-stage prior for the I binomial probabilities, where the first stage was the conjugate beta distribution and the second stage was discrete uniform.

Holland and Wang (1987) [10] used the local dependence function that measures the margin-free dependence to order bivariate distributions.

Wang (1987) [11] applied the characterization of a bivariate normal distribution to generate a table of probability integrals via the iterative proportional fitting algorithm.

McCann and Tebbs (2009) [12] constructed the simultaneous logit-based confidence intervals for odds ratios in the analysis of classification tables with a fixed reference level. They examined six procedures to control the familywise error rate and consider the simultaneous coverage probability and mean interval width, which can be used to construct simultaneous confidence intervals.

Baxter and Marchant (2010) [13] described that the non-randomized trials can provide bias in the effectiveness of any intrusion. This study showed a process to estimate the bias in such trials under the bivariate log-normal and gamma distributions, and the size of the bias under two different bivariate models.

Xu (2012) [14] demonstrated the odds ratio or the cross-product ratio is greater than or equal to one under the generalized proportional hazards model. The author used this property to improve a process of testing when the generalized proportional hazards model is not ideal to use for a data set.

Schaarschmidt et al. (2017) [15] proposed an asymptotic method for computation of simultaneous confidence intervals for user-defined sets of pairwise, between-treatment comparisons and user-defined sets of odds ratios based on the assumption of several independent multinomial samples. An improvement of this method by taking the correlation into account and application of Dirichlet posteriors with vague Dirichlet prior is also considered.

Niebuhr and Trabs (2019) [16] examined the impact of weighted data for the estimation of a discrete probability distribution for one-dimensional distributions. The weighting of observations usually increase estimation variances. In the two-dimensional discrete distribution, this research assumes that one marginal distribution is known. This additional information in one category of a contingency table allows for adjusting the estimation of another marginal if there is some degree of association between the two categories. For the marginals can not be assume that the marginals are independent, the authors presented to use the adjusted estimators in applications.

Martín Andrés et al. (2020) [17] considered the two-tailed asymptotic inferences about the odds ratio in cross-sectional studies (under the multinomial sampling). The research investigated 15 different methods, 5 of which were new and 10 were classic. They proposed new methods and compared them with other procedures.

A common practice in statistics is to take the log transformation of highly skewed data and construct confidence intervals for the population average on the basis of transformed data. To support our theoretical findings, we apply the Monte-Carlo method to investigate the key probability characteristics of the linear and logarithmic confidence intervals. We use $N = 10^5$ for the number of replications in our Monte-Carlo simulations.

A mathematical statement of the problem is as follows. Let $X^{(n_1)} = \left(X_{11},...,X_{1n_1}\right)$ and $X^{(n_2)} = \left(X_{21},...,X_{2n_2}\right)$ be two independent sequences of Bernoulli random variables with success probabilities p_1 and p_2, respectively. The observations are done according to the sequential sampling schemes with Markov stopping times ν_1 and ν_2. From the results of observations $X^{(\nu_1)} = (X_1,...,X_{\nu_1})$ and $X^{(\nu_2)} = (X_1,...,X_{\nu_2})$, it is necessary to identify the most accurate method of estimating the cross-product ratio.

Each sample may be obtained in the framework of direct binomial sampling scheme.

Direct binomial sampling. In this scheme, a random vector $X^{(n)} = (X_1,...,X_n)$ with Bernoulli components and a fixed number of observations n is observed. Note that $T = \sum_{i=1}^{n} X_i$ has the Binomial distribution $B(n,p)$, which has two parameters n and p, where n is a natural number and $0 < p < 1$. If a random variable T has Binomial distribution, then its probability mass function is

$$P\{T = t\} = \binom{n}{t} p^t (1-p)^{n-t},\ t = 0,1,...,n.$$

The random variable $\overline{X}_n = \dfrac{T}{n}$ is asymptotically normal with a mean $\mu_X = p$ and variance $\sigma_X^2 = \dfrac{p(1-p)}{n}$.

In the following, we will keep the notation $X_1, X_2,...$ for a Bernoulli sequence obtained by the direct sampling scheme.

2. Estimation of Proportion p and Its Reciprocal p^{-1}

First, we consider the problem of estimating parameter p (success probability) and parametric function $\dfrac{1}{p}$ for the Bernoulli trials. It seems difficult to estimate $\dfrac{1}{q}$, where $q = 1-p$, so we avoid this expression in our further

derivations by expressing it in terms of $\frac{p}{q}$ and $\frac{1}{p}$; see Section III. In this section we discuss how to estimate p and $\frac{1}{p}$. The following formulae is derived in Ngamkham (2018) [2] and Ngamkham and Volodins (2016) [1];
we present them in Table I.

Table I. Estimators for the proportion p and the reciprocal p^{-1} for direct-direct sampling scheme

	Proportion p	Reciprocal p^{-1}
Direct Sampling Scheme	$\hat{p}_n = \overline{X}_n$	$\widehat{p^{-1}}_n = \frac{n+1}{n\overline{X}_n+1}$ $\approx \widetilde{p^{-1}}_n = \frac{1}{\overline{X}_n}$

In the case of direct-direct sampling scheme, the estimate $p^{-1}{}_n$ is biased. Ngamkham (2018) [2] and Ngamkham and Volodins (2016) [1] proved that $\text{Bias}\left(p^{-1}{}_n\right) = \frac{1}{p} - E\left(p^{-1}{}_n\right) = \frac{1}{p}(1-p)^{n+1}$ is decreasing with an exponential rate as $n \to \infty$. The estimate p_n is unbiased estimator.

3. Estimating the Parametric Functions $\frac{p}{q}$ and $\frac{q}{p}$

To solve the problems stated in the Introduction, it is necessary to construct estimates, preferably with exponentially decreasing bias, for the parametric functions $\frac{p}{q}$ and $\frac{q}{p}$, where $q = 1 - p$ for direct-direct Binomial sampling scheme of Bernoulli trials.

From the point of view of estimation, the simplest case is an estimation of the parametric function $\frac{q}{p}$. Really,

$$\frac{q}{p} = \frac{1-p}{p} = \frac{1}{p} - 1$$

and we already know how to estimate $\frac{1}{p}$ for direct Binomial sampling scheme of Bernoulli trials from section II.

In the case of direct binomial sampling, we use statistics $p^{-1}{}_n = \frac{n+1}{n\overline{X}_n+1}$ as an estimator of p^{-1} with an exponentially decreasing bias.

We proceed with estimation of the parametric function $\frac{p}{q}$.

Proposition I. In the case of direct binomial sampling the statistics

$$p/q_n = \frac{n\overline{X}_n}{n+1-n\overline{X}_n}$$

Estimate the parametric functions $\frac{p}{q}$ with an exponentially decreasing bias.

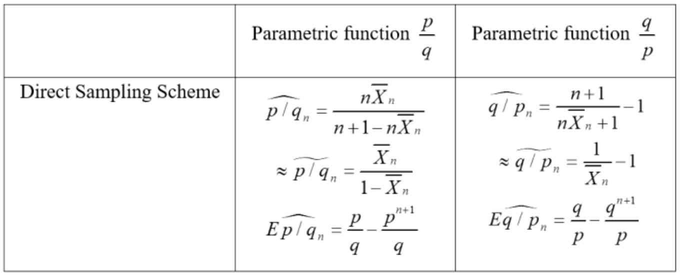
Proof: As we know, the statistic $T = n\overline{X}_n$ has the Binomial distribution $B(n, p)$. Therefore

$$\begin{aligned} Ep/q_n &= E\frac{n\overline{X}_n}{n+1-n\overline{X}_n} \\ &= E\frac{T}{n+1-T} = \sum_{k=0}^{n} \frac{k}{n+1-k}\binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=1}^{n} \frac{k}{n+1-k}\frac{n!}{k!(n-k)!} p^k q^{n-k} \text{ for } k = 0 \text{ we have zero term} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n+1-k)!} p^k q^{n-k} \\ &= \sum_{j=0}^{n-1} \frac{n!}{j!(n-j)!} p^{j+1} q^{n-j-1} \text{ make a substitution } j = k-1 \\ &= \frac{p}{q}\left[\sum_{j=0}^{n} \frac{n!}{j!(n-j)!} p^j q^{n-j} - \frac{n!}{n!0!} p^n q^0\right] \\ &= \frac{p}{q}\left[(p+q)^n - p^n\right] = \frac{p}{q}(1-p^n). \end{aligned}$$

Therefore, $\text{Bias}(p/q_n) = \frac{p}{q} - Ep/q_n = \frac{p^{n+1}}{q}$ is decreasing with an exponential rate as $n \to \infty$.

To summarize, we present Table II. for the estimation of $\frac{p}{q}$ and its reciprocal. In Table II, n is fixed numbers, $\{X_1, ...\}$ is sequences of independent Bernoulli random variables with the parameter $p, T = \sum_{k=1}^{n} X_k, \overline{X}_n = \frac{T}{n}$.

Table II. Estimators for the parametric functions $\frac{p}{q}$ and $\frac{q}{p}$ for direct sampling scheme



	Parametric function $\frac{p}{q}$	Parametric function $\frac{q}{p}$
Direct Sampling Scheme	$\widehat{p/q}_n = \frac{n\overline{X}_n}{n+1-n\overline{X}_n}$ $\approx \widetilde{p/q}_n = \frac{\overline{X}_n}{1-\overline{X}_n}$ $E\widehat{p/q}_n = \frac{p}{q} - \frac{p^{n+1}}{q}$	$\widehat{q/p}_n = \frac{n+1}{n\overline{X}_n+1} - 1$ $\approx \widetilde{q/p}_n = \frac{1}{\overline{X}_n} - 1$ $E\widehat{q/p}_n = \frac{q}{p} - \frac{q^{n+1}}{p}$

4. Point Estimator for the Cross-product Ratio

In Table III, we present estimator of the cross-product ratio of two proportions $\rho = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{p_1}{q_1} \times \frac{q_2}{p_2} = p/q_{n_1} \times q/p_{n_2}$ for direct-direct binomial sampling scheme. We compute the estimator of the cross-product ratio from two independent samples from direct-direct sampling scheme; from Table II. we used the parametric functions $\frac{p}{q}$ and $\frac{q}{p}$ substituted by $\frac{p_1}{q_1} = p/q_{n_1}$ and $\frac{q_2}{p_2} = q/p_{n_2}$, respectively.

Table III. Estimator of the cross-product ratio of two proportions $\rho = \frac{p_1(1-p_2)}{p_2(1-p_1)}$ and its approximation for direct-direct sampling scheme.

Sampling Scheme	Second Sample Direct
First Sample Direct	$\hat{\rho}_{n_1,n_2} = \frac{n_1\overline{X}_{n_1}}{n_1+1-n_1\overline{X}_{n_1}}\left(\frac{n_2+1}{n_2\overline{X}_{n_2}+1}-1\right)$ $\approx \tilde{\rho}_{n_1,n_2} = \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}\left(\frac{1}{\overline{X}_{n_2}}-1\right)$

Estimate of the cross-product ratio ρ is continuous functions of statistics $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ with finite second moments; therefore the estimate is asymptotically normal. $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ are independent. Now we find the asymptotic of the mean and variance of this estimate using the standard Delta method.

5. Delta-Method

Delta method can be explained briefly in the following way; for details we refer interested reader; see Lehmann (2004) [18], Section 2.5 on page 85.

Let $g(v_1, v_2)$ be a differentiable scalar function of two variables. Consider an estimator $T = g(V_1, V_2)$, which is a function of two other basic statistics V_1 and V_2. Usually statistics V_1 and V_2 have a simple form and are jointly asymptotically normal. The asymptotic distribution of an estimator T is found with the help of delta-method, which is a procedure of stochastic representation of T with the accuracy $O_P\left(\frac{1}{\sqrt{n}}\right)$, where n is the sample size.

By the Delta-method, we expand function g into a Taylor series at the point $\mu_1 = EV_1$ and $\mu_2 = EV_2$:

$$g(V_1, V_2) = g(\mu_1, \mu_2) + \sum_{i=1}^{2} \frac{\partial g(\mu_1, \mu_2)}{\partial v_i}(V_i - \mu_i) + \text{Remainder}.$$

It is possible to prove that the remainder term of the expansion converges in probability to zero with the rate $O_P\left(1/\sqrt{\min\{n_1, n_2\}}\right)$ as sample sizes n_1 and n_2 tends to infinity. We have that $g(V_1, V_2) - g(\mu_1, \mu_2)$ is asymptotically normal with a mean of zero and variance

$$E\left[\sum_{i=1}^{2} \frac{\partial g(\mu_1, \mu_2)}{\partial v_i}(V_i - \mu_i)\right]^2.$$

Therefore, the test statistics T is asymptotically normal with mean $g(\mu_1, \mu_2)$ and the variance of the form that is expressed through the elements of the covariance matrix of basic statistics V_1, V_2 and the coefficients $\frac{\partial g(\mu_1, \mu_2)}{\partial v_i}$.

For large values of n_1 and n_2, the estimator of the cross-product ratio ρ is functions of statistics $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ with finite second moments; therefore, the estimate is asymptotically normal. Our immediate task is to find the asymptotic of the mean and variance of this estimator, for which we explore the standard Delta method describe above. In our case, the method is based on a Taylor series expansion in the neighborhoods of the mean value of the statistic $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$. It is possible to calculate variance because statistics $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ are independent.

We consider direct-direct sampling scheme:

Direct-direct: Fix two natural numbers n_1 and n_2. Let $X^{(n_1)} = \left(X_{11}, ..., X_{1n_1}\right)$ and $X^{(n_2)} = \left(X_{21}, ..., X_{2n_2}\right)$ be two independent sequences of Bernoulli random variables. We know that the sample means for both samples $V_1 = \overline{X}_{n_1}$ and $V_2 = \overline{X}_{n_2}$ are asymptotically normal and jointly approximately normal, because the samples are independent. The form of the function $g(V_1, V_2)$ will be presented in section VI. The accuracy is $O_P\left(1/\sqrt{\min\{n_1, n_2\}}\right)$.

6. Asymptotic Distribution of Estimators for the Cross-product Ratio

For large values of n_1 and n_2, the estimator of the cross-product ratio ρ is functions of statistics $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ with finite second moments; therefore, the estimate is asymptotically normal. Our immediate task is to find the asymptotic of the mean and variance of this estimator, for which we explore the standard Delta method describe in Section V.

Remember that (see, for example Proposition 3 and 4, Ngamkham (2018) [2]): statistic $\overline{X}_n$ has a mean p and variance $\frac{pq}{n}$, and is asymptotically normal with these parameters.

If we use formulae for ρ, then our expressions for asymptotic variance are quite cumbersome. Hence we use the approximate estimators ρ in Delta method derivations.

In the following we will see that the normal approximation for estimator ρ for direct-direct sampling scheme has the structure of mean and variance:

$$\text{Asymptotic Mean } = \rho \text{ and}$$

$$\text{Asymptotic Variance } = \rho^2 s^2(p_1, p_2).$$

In the following, we will call $s^2(p_1, p_2)$ a variance component.

Direct-direct Sampling Scheme:

From Table III, in this case the statistic of interest is

$$\rho_{n_1,n_2} = \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}\left(\frac{1}{\overline{X}_{n_2}}-1\right) = g_{dd}(V_1, V_2) = \frac{V_1}{1-V_1}\left(\frac{1}{V_2}-1\right), \text{ where}$$

$V_1 = \overline{X}_{n_1}$ and $V_2 = \overline{X}_{n_2}$. In this particular case the function

$g_{dd}(v_1, v_2) = \frac{v_1}{1-v_1}\left(\frac{1}{v_2}-1\right)$. Note that

$EV_i = p_i, VarV_i = \frac{p_i q_i}{n_i}, i = 1, 2$ and

$$g_{dd}(p_1, p_2) = \frac{p_1}{1-p_1}\left(\frac{1}{p_2}-1\right) = \rho.$$

Partial derivatives are:

$$\frac{\partial g_{dd}(v_1, v_2)}{\partial v_1} = \frac{1}{(1-v_1)^2}\left(\frac{1}{v_2}-1\right) \text{ and } \frac{\partial g_{dd}(v_1, v_2)}{\partial v_2} = -\frac{v_1}{(1-v_1)v_2^2},$$

and hence

$$\frac{\partial g_{dd}(p_1, p_2)}{\partial v_1} = \frac{q_2}{q_1^2 p_2} \text{ and } \frac{\partial g_{dd}(p_1, p_2)}{\partial v_2} = -\frac{p_1}{q_1 p_2^2}.$$

A linear term Taylor expansion in the neighborhoods of the mean values of the statistics takes the form

$$\rho_{n_1,n_2} = g_{dd}(V_1, V_2) \approx \rho + \frac{q_2}{q_1^2 p_2}\left(\overline{X}_{n_1} - p_1\right) - \frac{p_1}{q_1 p_2^2}\left(\overline{X}_{n_2} - p_2\right).$$

From this, the estimator ρ_{n_1,n_2} is approximately normal with Mean=ρ and (remembering that $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ are independent)

$$\begin{aligned}\text{Variance} &= \frac{q_2^2}{q_1^4 p_2^2}\frac{p_1 q_1}{n_1} + \frac{p_1^2}{q_1^2 p_2^4}\frac{p_2 q_2}{n_2} \\ &= \rho^2\left[\left(\frac{p_1}{q_1}\right)(p_1^{-1})^2/n_1 + \left(\frac{p_2}{q_2}\right)(p_2^{-1})^2/n_2\right].\end{aligned}$$

In this case, the variance component

$$s^2(p_1, p_2) = \left(\frac{p_1}{q_1}\right)(p_1^{-1})^2/n_1 + \left(\frac{p_2}{q_2}\right)(p_2^{-1})^2/n_2.$$

From Table III, we estimate ρ as

$$\rho = \frac{n_1\overline{X}_{n_1}}{n_1+1-n_1\overline{X}_{n_1}}\left(\frac{n_2+1}{n_2\overline{X}_{n_2}+1}-1\right).$$

To obtain the plug-in estimator of the variance component, we substitute estimations for $\frac{p}{q}$ and p^{-1} (see Tables I and II),

namely $p_1/q_1 = \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}, p_2/q_2 = \frac{\overline{X}_{n_2}}{1-\overline{X}_{n_2}}$, $p_1^{-1} = \frac{1}{\overline{X}_{n_1}}$, and

$p_2^{-1} = \frac{1}{\overline{X}_{n_2}}$ and obtain

$$\begin{aligned}\hat{s}^2 &= \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}\left(\frac{1}{\overline{X}_{n_1}}\right)^2/n_1 + \frac{\overline{X}_{n_2}}{1-\overline{X}_{n_2}}\left(\frac{1}{\overline{X}_{n_2}}\right)^2/n_2 \\ &= \frac{1}{n_1(1-\overline{X}_{n_1})\overline{X}_{n_1}} + \frac{1}{n_2(1-\overline{X}_{n_2})\overline{X}_{n_2}}.\end{aligned}$$

Asymptotic distribution of logarithm of estimate for the cross-product ratio:

In the following we will see that the normal approximation for estimator $\log\ \rho$ for direct-direct sampling scheme has the structure of mean and variance:

$$\text{Asymptotic Mean of } \log\ \rho = \log\ \rho \text{ and}$$

$$\text{Asymptotic Variance of } \log\ \rho = s^2(p_1, p_2).$$

If we use formulae for ρ, then our expressions for asymptotic variance are quite cumbersome. Hence we use the approximate estimators ρ in Delta method derivations.

Direct-direct Sampling Scheme:

From Table III, in this case the statistic of interest is

$$\log\ \rho_{n_1,n_2} = \log\left(\frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}\left(\frac{1}{\overline{X}_{n_2}}-1\right)\right) = gl_{dd}(V_1,V_2)$$
$$= \log(V_1)-\log(1-V_1)+\log(1-V_2)-\log(V_2),$$

where $V_1 = \overline{X}_{n_1}$ and $V_2 = \overline{X}_{n_2}$. In this particular case the function

$$gl_{dd}(v_1,v_2) = \log(v_1)-\log(1-v_1)+\log(1-v_2)-\log(v_2).$$

Note that $EV_i = p_i, VarV_i = \frac{p_i q_i}{n_i}, i = 1,2$ and

$$gl_{dd}(p_1,p_2) = \log(p_1)-\log(1-p_1)+\log(1-p_2)-\log(p_2) = \log\rho.$$

Partial derivatives are:

$$\frac{\partial gl_{dd}(v_1,v_2)}{\partial v_1} = \frac{1}{v_1}+\frac{1}{1-v_1} \text{ and } \frac{\partial gl_{dd}(v_1,v_2)}{\partial v_2} = -\frac{1}{v_2}-\frac{1}{1-v_2},$$

and hence

$$\frac{\partial gl_{dd}(p_1,p_2)}{\partial v_1} = \frac{1}{p_1}+\frac{1}{q_1} = \frac{1}{p_1 q_1} \text{ and}$$

$$\frac{\partial gl_{dd}(p_1,p_2)}{\partial v_2} = -\frac{1}{p_2}-\frac{1}{q_2} = -\frac{1}{p_2 q_2}.$$

Linear term Taylor expansion in the neighborhoods of the mean values of the statistics takes the form:

$$\log\ \rho_{n_1,n_2} = gl_{dd}(V_1,V_2) \approx \log\rho + \frac{1}{p_1 q_1}\left(\overline{X}_{n_1}-p_1\right) - \frac{1}{p_2 q_2}\left(\overline{X}_{n_2}-p_2\right).$$

From this, the estimator $\log\ \rho_{n_1,n_2}$ is approximately normal with Mean $= \log\rho$ and (remind that $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ are independent)

$$\text{Variance} = s^2 = \frac{1}{p_1^2 q_1^2}\frac{p_1 q_1}{n_1} + \frac{1}{p_2^2 q_2^2}\frac{p_2 q_2}{n_2}$$
$$= \frac{p_1}{q_1}(p_1^{-1})^2/n_1 + \frac{p_2}{q_2}(p_2^{-1})^2/n_2.$$

To obtain the plug-in estimator of the variance, we substitute estimations for $\frac{p}{q}$ and p^{-1} (see Tables I and II), namely

$$p_1/q_1 = \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}},\ p_2/q_2 = \frac{\overline{X}_{n_2}}{1-\overline{X}_{n_2}}, p_1^{-1} = \frac{1}{\overline{X}_{n_1}}, \text{ and}$$

$p_2^{-1} = \frac{1}{\overline{X}_{n_2}}$ and obtain that

$$\hat{s}^2 = \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}\left(\frac{1}{\overline{X}_{n_1}}\right)^2/n_1 + \frac{\overline{X}_{n_2}}{1-\overline{X}_{n_2}}\left(\frac{1}{\overline{X}_{n_2}}\right)^2/n_2$$
$$= \frac{1}{n_1\overline{X}_{n_1}(1-\overline{X}_{n_1})} + \frac{1}{n_2\overline{X}_{n_2}(1-\overline{X}_{n_2})}.$$

7. Confidence Limits

As mentioned, the asymptotic for mean and variance of estimator ρ for the cross-product ratio ρ for direct-direct sampling scheme has the structure: $E\rho = \rho$ and $Var\rho = \rho^2 s^2(p_1,p_2)$, where $s^2(p_1,p_2)$ is the variance component.

If the sample sizes for direct-direct sampling scheme tend to infinity, then

$$P\left(\left|\rho-\rho\right| \le z_{\alpha/2}\rho s(p_1,p_2)\right) \sim 1-\alpha,$$

where $z_{\alpha/2}$ is $(1-\alpha/2)$-quantile of the standard normal distribution. Since $s^2(p_1,p_2)$ is a continuous function of its arguments, replacing $s^2(p_1,p_2)$ by plug-in estimator $\hat{s}^2$ from Section VI.

Therefore, if the sample sizes in direct-direct sampling scheme tend to infinity, then the interval with the following end-points,

$$\rho\left(1\mp z_{\alpha/2}\hat{s}\right) \qquad (1)$$

is the asymptotic $(1-\alpha)$-confidence sets for the cross-product ratio ρ. We will call it the linear confidence interval to distinguish it from logarithmic confidence interval.

Direct-direct Sampling Scheme:

When both samples are obtained by the direct sampling scheme with sample sizes n_1 and n_2, then, according to Table III and Section VI:

$$\rho_{n_1,n_2} = \frac{n_1\overline{X}_{n_1}}{n_1+1-n_1\overline{X}_{n_1}}\left(\frac{n_2+1}{n_2\overline{X}_{n_2}+1}-1\right) \text{ and}$$

$$\hat{s}^2 = \frac{1}{n_1(1-\overline{X}_{n_1})\overline{X}_{n_1}} + \frac{1}{n_2(1-\overline{X}_{n_2})\overline{X}_{n_2}}.$$

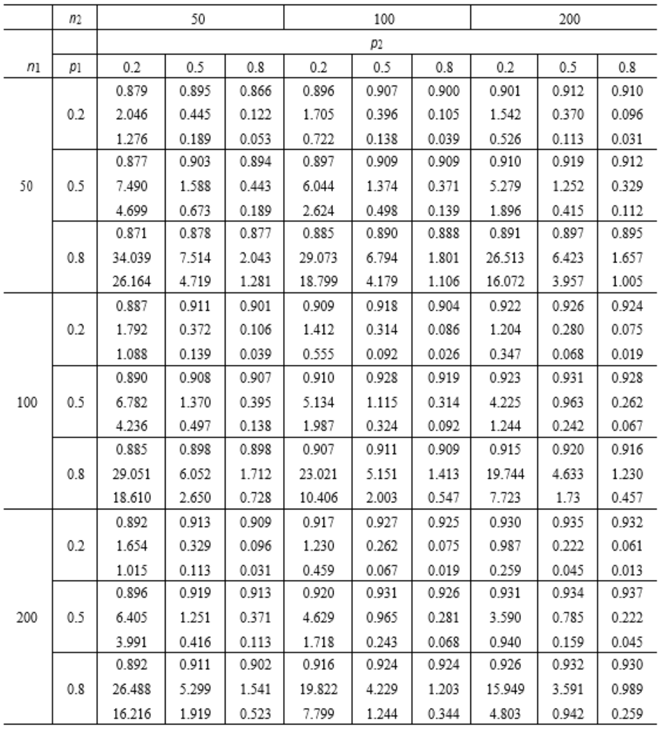
Hence, the asymptotic $n_1, n_2 \to \infty$ confidence interval (1) based on the relative frequencies $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ of successes (sample means) in each sample and can be written as

$$\frac{n_1\overline{X}_{n_1}}{n_1+1-n_1\overline{X}_{n_1}}\left(\frac{n_2+1}{n_2\overline{X}_{n_2}+1}-1\right)\left(1\mp z_{\alpha/2}\sqrt{\frac{1}{n_1(1-\overline{X}_{n_1})\overline{X}_{n_1}}+\frac{1}{n_2(1-\overline{X}_{n_2})\overline{X}_{n_2}}}\right). \quad (2)$$

Below, we provide the results of statistical modeling in Table IV. For each pair (n_1, n_2) of sample sizes and values (p_1, p_2) of success probabilities, we present the Monte-Carlo estimations of the coverage probability, mean width, and standard deviation of the width for the confidence interval (2). The nominal level is assumed to be 0.95.

The results of Table IV show that the interval (2) has a confidence level lower than nominal and an error not larger than 0.02 only for $n_1, n_2 = 200$ and $p_1, p_2 \geq 0.2$.

Table IV. Coverage probability, width, and standard deviation for confidence interval (2)



	n_2	50			100			200		
		p_2								
n_1	p_1	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
50	0.2	0.879 2.046 1.276	0.895 0.445 0.189	0.866 0.122 0.053	0.896 1.705 0.722	0.907 0.396 0.138	0.900 0.105 0.039	0.901 1.542 0.526	0.912 0.370 0.113	0.910 0.096 0.031
	0.5	0.877 7.490 4.699	0.903 1.588 0.673	0.894 0.443 0.189	0.897 6.044 2.624	0.909 1.374 0.498	0.909 0.371 0.139	0.910 5.279 1.896	0.919 1.252 0.415	0.912 0.329 0.112
	0.8	0.871 34.039 26.164	0.878 7.514 4.719	0.877 2.043 1.281	0.885 29.073 18.799	0.890 6.794 4.179	0.888 1.801 1.106	0.891 26.513 16.072	0.897 6.423 3.957	0.895 1.657 1.005
100	0.2	0.887 1.792 1.088	0.911 0.372 0.139	0.901 0.106 0.039	0.909 1.412 0.555	0.918 0.314 0.092	0.904 0.086 0.026	0.922 1.204 0.347	0.926 0.280 0.068	0.924 0.075 0.019
	0.5	0.890 6.782 4.236	0.908 1.370 0.497	0.907 0.395 0.138	0.910 5.134 1.987	0.928 1.115 0.324	0.919 0.314 0.092	0.923 4.225 1.244	0.931 0.963 0.242	0.928 0.262 0.067
	0.8	0.885 29.051 18.610	0.898 6.052 2.650	0.898 1.712 0.728	0.907 23.021 10.406	0.911 5.151 2.003	0.909 1.413 0.547	0.915 19.744 7.723	0.920 4.633 1.73	0.916 1.230 0.457
200	0.2	0.892 1.654 1.015	0.913 0.329 0.113	0.909 0.096 0.031	0.917 1.230 0.459	0.927 0.262 0.067	0.925 0.075 0.019	0.930 0.987 0.259	0.935 0.222 0.045	0.932 0.061 0.013
	0.5	0.896 6.405 3.991	0.919 1.251 0.416	0.913 0.371 0.113	0.920 4.629 1.718	0.931 0.965 0.243	0.926 0.281 0.068	0.931 3.590 0.940	0.934 0.785 0.159	0.937 0.222 0.045
	0.8	0.892 26.488 16.216	0.911 5.299 1.919	0.902 1.541 0.523	0.916 19.822 7.799	0.924 4.229 1.244	0.924 1.203 0.344	0.926 15.949 4.803	0.932 3.591 0.942	0.930 0.989 0.259

Confidence limits for logarithmic interval:

As mentioned, for direct-direct sampling scheme, the normal approximation for estimator $\log\tilde{\rho}$ show that mean and variance has the structure:

$$\text{Mean}=\log\rho \text{ and } \text{Variance} = s^2(p_1, p_2).$$

If the sample sizes in direct-direct sampling scheme tend to infinity, then using the inequity

$$\left|\log\rho - \log\rho\right| \leq z_{\alpha/2}s(p_1, p_2),$$

(where $z_{\alpha/2}$ is $(1-\alpha/2)$-quantile of the standard normal distribution) and replacing $s^2(p_1, p_2)$ by its estimator that correspond to direct-direct sampling scheme, gives us the following end points for an asymptotically $(1-\alpha)$-confidence interval for the cross-product ratio ρ:

$$\rho\exp\{\mp z_{\alpha/2}\hat{s}\}. \quad (3)$$

Direct-direct Sampling Scheme:

When both samples are obtained by direct sampling scheme with sample sizes n_1 and n_2, then according to Table III and Section VI:

$$\rho_{n_1,n_2} = \frac{n_1\overline{X}_{n_1}}{n_1+1-n_1\overline{X}_{n_1}}\left(\frac{n_2+1}{n_2\overline{X}_{n_2}+1}-1\right) \text{ and}$$

$$\hat{s}^2 = \frac{1}{n_1\overline{X}_{n_1}(1-\overline{X}_{n_1})}+\frac{1}{n_2\overline{X}_{n_2}(1-\overline{X}_{n_2})}.$$

Hence the asymptotic $n_1, n_2 \to \infty$ confidence interval (3) based on the relative frequencies $\overline{X}_{n_1}$ and $\overline{X}_{n_2}$ of successes (sample means) in each sample and can be written as

$$\frac{n_1\overline{X}_{n_1}}{n_1+1-n_1\overline{X}_{n_1}}\left(\frac{n_2+1}{n_2\overline{X}_{n_2}+1}-1\right)\exp\left\{\mp z_{\alpha/2}\sqrt{\frac{1}{n_1\overline{X}_{n_1}(1-\overline{X}_{n_1})}+\frac{1}{n_2\overline{X}_{n_2}(1-\overline{X}_{n_2})}}\right\}. \quad (4)$$

Below, we provide simulation results in Table V. For each pair (n_1, n_2) of sample sizes and values (p_1, p_2) of success probabilities, we present the Monte-Carlo estimations of the coverage probability, mean width, and standard deviation of the width for the confidence interval (4). The nominal level is assumed to be 0.95.

The logarithmic interval (Table V.) has good coverage probability with an error less than 0.01 in most of the cases.

Table V: Coverage probability, width, and standard deviation for logarithmic interval (4)

	n_2	50			100			200		
n_1	p_1 \ p_2	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
50	0.2	0.958 2.436 1.683	0.955 0.506 0.209	0.955 0.142 0.059	0.958 1.925 0.805	0.956 0.439 0.147	0.954 0.119 0.042	0.959 1.704 0.560	0.959 0.406 0.118	0.956 0.106 0.033
	0.5	0.952 8.689 6.148	0.954 1.760 0.748	0.954 0.505 0.209	0.950 6.642 2.937	0.947 1.484 0.540	0.950 0.407 0.151	0.952 5.674 2.052	0.955 1.335 0.444	0.952 0.353 0.121
	0.8	0.945 41.288 37.972	0.953 8.718 6.202	0.957 2.433 1.684	0.949 33.480 25.039	0.953 7.695 5.436	0.956 2.066 1.461	0.953 29.868 20.623	0.954 7.191 5.168	0.957 1.866 1.301
100	0.2	0.956 2.055 1.429	0.952 0.407 0.151	0.956 0.119 0.042	0.953 1.533 0.613	0.952 0.335 0.097	0.948 0.093 0.028	0.954 1.278 0.365	0.951 0.295 0.071	0.950 0.079 0.020
	0.5	0.954 7.684 5.622	0.948 1.480 0.539	0.957 0.439 0.147	0.951 5.499 2.175	0.953 1.173 0.342	0.952 0.335 0.097	0.950 4.426 1.312	0.947 1.001 0.251	0.951 0.274 0.070
	0.8	0.949 33.446 24.528	0.952 6.650 2.969	0.957 1.932 0.813	0.950 25.086 11.708	0.949 5.517 2.193	0.954 1.534 0.602	0.952 21.043 8.444	0.951 4.901 1.878	0.951 1.310 0.497
200	0.2	0.955 1.864 1.334	0.952 0.354 0.121	0.956 0.106 0.032	0.952 1.309 0.500	0.950 0.274 0.070	0.951 0.079 0.020	0.952 1.028 0.271	0.952 0.229 0.046	0.953 0.064 0.013
	0.5	0.954 7.173 5.329	0.954 1.335 0.446	0.959 0.407 0.118	0.951 4.896 1.865	0.946 1.002 0.253	0.952 0.296 0.071	0.950 3.712 0.980	0.948 0.805 0.163	0.953 0.229 0.046
	0.8	0.954 29.847 21.014	0.951 5.695 2.077	0.960 1.703 0.556	0.952 21.128 8.536	0.950 4.429 1.312	0.955 1.278 0.363	0.95 16.621 5.068	0.951 3.713 0.983	0.951 1.030 0.271

8. Discussion

The simulation results present in Table IV, the linear asymptotic confidence interval for the cross-product ratio coefficient has a confidence level quite low precision and poor accuracy properties. A common practice in statistics is to take the log transformation of highly skewed data and construct confidence interval for the population average on the basis of transformed data. In this article, we investigate logarithmic confidence interval (the results present in Table V.) that shows better precision and accuracy properties.

9. Conclusion

The linear confidence interval for the cross-product ratio coefficient has a confidence level lower than nominal. In this article, we show that this deficiency may be resolved by considering the logarithmic confidence interval. We recommend the analysis of precision and reliability properties of logarithmic confidence interval for direct-direct sampling scheme. Consideration of accuracy and reliability properties of the point estimator for the cross-product ratio is also an interesting problem.

Acknowledgment

The authors thank the Department of Mathematics and Statistics Faculty of Science and Technology Thammasat University for financial support.

References

[1] T. Ngamkham, A. Volodin, and I. Volodin, "Confidence intervals for a ratio of binomial proportions based on direct and inverse sampling schemes," *Lobachevskii J. Math*, vol. 37, 2016, pp. 466-496.

[2] T. Ngamkham, "Confidence interval estimation for the ratio of binomial proportions and random numbers generation for some statistical models," *Ph. D. Thesis* (Univ. of Regina, 2018).

[3] E. L. Lehmann, *Testing Statistical Hypotheses*. New York: Springer, 1997.

[4] L. A. Goodman, "Simultaneous confidence limits for cross-product ratios in contingency tables," *J. R. Stat. Soc*, Ser. B 26, 1964, pp.86-102.

[5] J. J. Gart, "On the combination of relative risks," *Biometrics*, vol. 18, 1962, pp. 601-610.

[6] J. Cornfield, "A statistical problem arising from retrospective studies," in *Proceedings of the 3rd Berkeley symposium on Mathematical Statistics and Probability,* vol.4, 1956. (Univ. of California Press, Berkeley, CA).

[7] J. Anděl, "On interactions in contingency tables," *Apl. Mat*, vol. 18, 1973, pp. 99-109.

[8] L. F. Lee, "Empirical Bayes estimators for the cross-product ratios for 2 × 2 contingency tables," *Ph. D. Thesis* (Virginia Polytech. Inst. State Univ., 1981).

[9] J. H. Albert and A. K. Gupta, "Estimation in contingency tables using prior information," *J. R. Stat. Soc*, Ser. B 45, 1983, pp. 60-69.

[10] P. W. Holland and Y. J. Wang, "Dependence functions for Continuous bivariate densities," *Commun. Stat. Theory Methods*, vol. 16, 1987, pp. 863-876.

[11] Y. J. Wang, "The probability integrals of bivariate normal distributions: A contingency table approach," *Biometrika*, vol. 74, 1987, pp. 185-190.

[12] M. H. McCann and J. M. Tebbs, "Simultaneous logit-based confidence intervals for odds ratios in 2 × *k* classification tables with a fixed reference level," *Commun. Stat. Simul. Comput*, vol. 38, 2009, pp. 961-975.

[13] P. D. Baxter and P. R. Marchant, "The cross-product ratio in bivariate lognormal and gamma distributions, with an application to non-randomized trials," *J. Appl. Stat*, vol. 37, 2010, pp. 529-536.

[14] J. L. Xu, "A property of the generalized proportional hazards model," *Far East J. Theor. Stat*, vol. 41, 2012, pp. 149-162.

[15] F. Schaarschmidt, D. Gerhard, and C. Vogel, "Simultaneous confidence intervals for comparisons of several multinomial samples," *Comput. Stat. Data Anal*, vol. 106, 2017, pp. 65-76.

[16] T. Niebuhr and M. Trabs, "Profiting from correlations: Adjusted estimators for categorical data," *Appl. Stoch. Models Business Ind*, vol. 35, 2019, pp. 1090-1102.

[17] A. Martín Andrés, J. M. Tapia García, and F. Gayá Moreno, "Two-tailed asymptotic inferences for the odds ratio in cross-sectional studies: Evaluation of fifteen old and new methods of inference," *Research Gate Publ., №340429976*, 2020, pp. 1-33.

[18] E. L. Lehmann, *Elements of Large Sample Theory*. New York: Springer, 2004.