ISSN 1995-0802, Lobachevskii Journal of Mathematics, 2021, Vol. 42, No. 2, pp. 435–450. © Pleiades Publishing, Ltd., 2021.

Confidence Estimation of the Cross-Product Ratio of Binomial Proportions under Different Sampling Schemes

Chanakan Sungboonchoo[1*], Thuntida Ngamkham[2], Wararit Panichkitkosolkul[1***], and Andrei Volodin[3****]**

(Submitted by A. M. Elizarov)

[1]Department of Mathematics and Statistics, Thammasat University, Phathum Thani, Thailand
[2]Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta T2N 1N4, Canada
[3]Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan S4S 0A2, Canada

Received May 18, 2020; revised May 31, 2020; accepted June 07, 2020

Abstract—We consider a general problem of the interval estimation 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 or inverse Binomial sampling schemes. Asymptotic confidence intervals are constructed in accordance with different types of sampling schemes, 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 all possible combinations of 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.

DOI: 10.1134/S1995080221020190

Keywords and phrases: *Cross-product ratio, Binomial proportions, Direct Binomial sampling scheme, Inverse Binomial sampling scheme, Confidence estimation, Confidence limits, Asymptotic confidence limits, Delta method, Normal approximation, Monte-Carlo simulation*

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 confidence estimation of the cross-product ratio of Binomial proportions for different 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. This can be explained by its importance for analyzing 2×2 contingency tables; see Lehmann (1997) [3], Section 4.6.

Because of that, it is very interesting to investigate statistical inference for the cross-product ratio.

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

*E-mail: chanakano@hotmail.com
**E-mail: thuntida.ngamkham@ucalgary.ca
***E-mail: wararit@mathstat.sci.tu.ac.th
****E-mail: andrei@uregina.ca

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 to the $r \times c$ table. 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 mathematical statement of the problem is as follows. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ 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 $X^{(\nu_1)} = (X_1, \ldots, X_{\nu_1})$ and $Y^{(\nu_2)} = (Y_1, \ldots, Y_{\nu_2})$, it is necessary to identify the most accurate method of estimating the cross-product ratio $\rho = \frac{p_1(1-p_2)}{p_2(1-p_1)}$.

Each sample may be obtained in the framework of direct or inverse binomial sampling schemes.

Direct binomial sampling. In this scheme, a random vector $X^{(n)} = (X_1, \ldots, X_n)$ 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

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

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

Inverse binomial sampling. In this scheme, a Bernoulli sequence $Y^{(\nu)} = (Y_1, \ldots, Y_\nu)$ is observed with a stopping time $\nu = \min\{n : \sum_{k=1}^{n} Y_k \geq m\}$, where m is a fixed number of successes. From Ngamkham [2] and Ngamkham and Volodins [1] we know that ν has the Pascal distribution $P(m, p)$, which has two parameters m and p, where m is a natural number and $0 < p < 1$. If a random variable ν has Pascal distribution, then its probability mass function is

$$P\{\nu = k\} = \binom{k-1}{m-1} p^m (1-p)^{k-m}, \quad k = m, m+1, m+2, \ldots$$

The random variable $\overline{Y}_m = \nu/m$ is asymptotically normal with a mean $\mu_Y = 1/p$ and variance $\sigma_Y^2 = (1-p)/mp^2$.

In the following, we will keep the notation $X_1, X_2, \ldots$ for a Bernoulli sequence obtained by the direct sampling scheme and $Y_1, Y_2, \ldots$ for a Bernoulli sequence obtained by the inverse sampling scheme.

2. ESTIMATION OF PROPORTION p AND ITS RECIPROCAL p^{-1}

First we consider the problem of estimation of parameter p (success probability) and parametric function $\frac{1}{p}$ for the Bernoulli trials. It seems difficult to estimate $\frac{1}{q}$, where $q = 1 - p$, so we will try to avoid this expression in our further derivations and express it in terms of $\frac{p}{q}$ and $\frac{1}{p}$; see Section 3. In this section we discuss how to estimate p and $\frac{1}{p}$. The following formulae is derived in Nagamkham [2] and Ngamkham and Volodins [1]; we present them in Table 1.

In the case of direct binomial sampling, the estimate $\widehat{p^{-1}}_n$ is biased. Nagamkham [2] and Ngamkham and Volodins [1] proved that $\text{Bias}(\widehat{p^{-1}}_n) = \frac{1}{p} - E\widehat{p^{-1}}_n = \frac{1}{p}(1-p)^{n+1}$ is decreasing with an exponential rate as $n \to \infty$. All other cases $\widehat{p}_n, \widehat{p}_m$, and $\widehat{p^{-1}}_m$ provide unbiased estimators.

3. ESTIMATION OF 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 two schemes 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 both schemes of Bernoulli trials from Section 2.

In the case of direct binomial sampling, we use statistics $\widehat{p_n^{-1}} = \frac{n+1}{n\overline{X}_n+1}$ as an estimator of p^{-1} with an exponentially decreasing bias. In the case of inverse binomial sampling, we use statistics $\widehat{p_m^{-1}} = \overline{Y}_m = \nu/m$ as an unbiased estimator of p^{-1}.

Now we proceed with an estimation of the parametric function $\frac{p}{q}$.

Proposition 3.1. *In the case of direct binomial sampling, the statistics*

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

estimates 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

$$E\widehat{p/q}_n = E\frac{n\overline{X}_n}{n+1-n\overline{X}_n} = E\frac{T}{n+1-T} = \sum_{k=0}^{n} \frac{k}{n+1-k}\binom{n}{k}p^k q^{n-k}$$

$$= \sum_{k=1}^{n} \frac{k}{n+1-k}\frac{n!}{k!(n-k)!}p^k q^{n-k} \text{ for } k=0 \text{ we have zero term}$$

$$= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n+1-k)!}p^k q^{n-k}$$

$$= \sum_{j=0}^{n-1} \frac{n!}{j!(n-j)!}p^{j+1} q^{n-j-1} \text{ make a substitution } j=k-1$$

$$= \frac{p}{q}\left[\sum_{j=0}^{n} \frac{n!}{j!(n-j)!}p^j q^{n-j} - \frac{n!}{n!0!}p^n q^0\right] = \frac{p}{q}[(p+q)^n - p^n] = \frac{p}{q}(1-p^n).$$

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

Proposition 3.2. *In the case of inverse binomial sampling, the statistics*

$$\widehat{p/q}_m = \frac{m-1}{m\overline{Y}_m - m + 1}$$

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

Proof. As we know, the statistic $\nu = m\overline{Y}_m$ has the Pascal distribution $P(m,p)$. Therefore,

$$E\widehat{p/q}_m = E\frac{m-1}{m\overline{Y}_m - m+1} = E\frac{m-1}{\nu - m + 1} = (m-1)\sum_{k=m}^{\infty}\frac{1}{k-m+1}\binom{k-1}{m-1}p^m q^{k-m}$$

$$= (m-1)p^m \sum_{i=0}^{\infty}\frac{1}{i+1}\binom{i+m-1}{m-1}q^i \text{ make a substitution } i=k-m$$

$$= \frac{(m-1)p^m}{q}\sum_{i=0}^{\infty}\binom{i+m-1}{m-1}\frac{q^{i+1}}{i+1} = \frac{(m-1)p^m}{q}\sum_{i=0}^{\infty}\binom{i+m-1}{m-1}\int_0^q t^i dt$$

We can interchange the sum and integral signs by Fubini—Tornelli theorem

$$= \frac{(m-1)p^m}{q}\int_0^q \sum_{i=0}^{\infty}\binom{i+m-1}{m-1}t^i dt = \frac{(m-1)p^m}{q}\int_0^q (1-t)^{-m}dt$$

because for any $|t| < 1$ and natural number $m \geq 1 : (1-t)^{-m} = \sum_{i=0}^{\infty}\binom{i+m-1}{m-1}t^i$,

see, for example Lemma 1, Ngamkham (2018) [2]

$$= \frac{(m-1)p^m}{q}\left[\frac{p^{1-m}}{m-1} - \frac{1}{m-1}\right] = \frac{p}{q} - \frac{p^m}{q}$$

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

	Proportion p	Reciprocal p^{-1}
Direct Sampling Scheme	$\widehat{p}_n = \overline{X}_n$	$\widehat{p_n^{-1}} = \frac{n+1}{n\overline{X}_n+1} \approx \widetilde{p^{-1}}_n = \frac{1}{\overline{X}_n}$
Inverse Sampling Scheme	$\widehat{p}_m = \frac{m-1}{m\overline{Y}_m-1} \approx \widetilde{p}_m = \frac{1}{\overline{Y}_m}$	$\widehat{p^{-1}}_m = \overline{Y}_m$

Table 2. Estimators for the parametric functions $\frac{p}{q}$ and $\frac{q}{p}$ for direct and inverse sampling schemes

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

Table 3. Estimators of the cross-product ratio of two proportions $\rho = \frac{p_1(1-p_2)}{p_2(1-p_1)}$ and their approximations for all possible combinations of direct and inverse sampling schemes

Sampling Schemes	Second Sample Direct	Second Sample Inverse
First Sample Direct	$\widehat{\rho}_{n_1,n_2} = \frac{n_1\overline{X}_{n_1}}{n_1+1-n_1\overline{X}_{n_1}}\left(\frac{n_2+1}{n_2\overline{X}_{n_2}+1} - 1\right)$ $\approx \widetilde{\rho}_{n_1,n_2} = \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}\left(\frac{1}{\overline{X}_{n_2}} - 1\right)$	$\widehat{\rho}_{n,m} = \frac{n\overline{X}_n}{n+1-n\overline{X}_n}(\overline{Y}_m - 1)$ $\approx \widetilde{\rho}_{n,m} = \frac{\overline{X}_n(\overline{Y}_m-1)}{1-\overline{X}_n}$
First Sample Inverse	$\widehat{\rho}_{m,n} = \left(\frac{m-1}{m\overline{Y}_m-m+1}\right)\left(\frac{n+1}{n\overline{X}_n+1} - 1\right)$ $\approx \widetilde{\rho}_{m,n} = \frac{1}{\overline{Y}_m-1}\left(\frac{1}{\overline{X}_n} - 1\right)$	$\widehat{\rho}_{m_1,m_2} = \left(\frac{m_1-1}{m_1\overline{Y}_{m_1}-m_1+1}\right)(\overline{Y}_{m_2} - 1)$ $\approx \widetilde{\rho}_{m_1,m_2} = \frac{\overline{Y}_{m_2}-1}{\overline{Y}_{m_1}-1}$

Therefore, $\mathrm{Bias}(\widehat{p/q}_m) = \frac{p}{q} - E\widehat{p/q}_m = \frac{p^m}{q}$ is decreasing with an exponential rate as $m \to \infty$.□

To summarize, we present Table 2 for the estimation of p/q and its reciprocal. In Table 2, m and n are fixed numbers, $\{X_1, \ldots\}$ and $\{Y_1, \ldots\}$ are sequences of independent Bernoulli random variables with the parameter $p, T = \sum_{k=1}^{n} X_k$, $\overline{X}_n = T/n$, $\nu = \min\{n : \sum_{k=1}^{n} Y_k \geq m\}$, and $\overline{Y}_m = \nu/m$.

4. POINT ESTIMATOR FOR THE CROSS-PRODUCT RATIO

In Table 3, we present all estimators of the cross-product ratio of two proportions $\rho = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{p_1}{q_1} \times \frac{q_2}{p_2}$ for these four possible sampling schemes.

All four estimates of the cross-product ratio ρ are continuous functions of statistics $\overline{X}_n$ and $\overline{Y}_m$ with finite second moments; therefore the estimates are asymptotically normal. Now we find the asymptotic of the mean and variance of these estimates, for which we explore the standard Delta method.

5. DELTA-METHOD

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 $\mathcal{O}_P(1/\sqrt{n})$, where n is the sample size.

By the Delta-method, we expand function g into a Taylor series at the point $\mu_1 = \mathrm{E}V_1$ and $\mu_2 = \mathrm{E}V_2$:

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

It is possible to prove that the remainder term of the expansion converges in probability to zero with the rate $\mathcal{O}\left([\min\{m, n\}]^{-1/2}\right)$ as sample sizes m and n 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 m and n, all four estimators of the cross-product ratio ρ are differentiable functions of statistics $\overline{X}_n$ and $\overline{Y}_m$ with finite second moments; therefore, the estimates are asymptotically normal. Our immediate task is to find the asymptotic of the mean and variance of these estimates, for which we explore the standard Delta method described above. In our case, the method is based on a Taylor series expansion in the neighborhoods of the mean values of the statistics $\overline{X}_n$ and $\overline{Y}_m$. It is possible to calculate variances in all four cases because statistics $\overline{X}_n$ and $\overline{Y}_m$ are independent.

We consider the following four possible scenarios:

1. Direct-direct. Fix two natural numbers n_1 and n_2. Let $X^{(n_1)} = (X_{11}, \ldots, X_{1n_1})$ and $X^{(n_2)} = (X_{21}, \ldots, X_{2n_2})$ 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 6. The accuracy is $\mathcal{O}_P(1/\sqrt{\min\{n_1, n_2\}})$.

2. Direct-inverse. Fix two natural numbers n and m. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_\nu$ be two independent sequences of Bernoulli random variables, where $\nu = \min\{n : \sum_{k=1}^{n} Y_k \geq m\}$. We know that the sample mean for the first samples $V_1 = \overline{X}_n$ and statistic $V_2 = \overline{Y}_m = \nu/m$ 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 6. The accuracy of the Delta method in this case is $\mathcal{O}_P(1/\sqrt{\min\{n, m\}})$.

3. Inverse-direct. Fix two natural numbers n and m. Let $Y_1, \ldots, Y_\nu$ and $X_1, \ldots, X_n$ be two independent sequences of Bernoulli random variables, where $\nu = \min\{n : \sum_{k=1}^{n} Y_k \geq m\}$. We know that the statistic $V_1 = \overline{Y}_m = \nu/m$ and the sample mean for the second samples $V_2 = \overline{X}_n$ 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 6. The accuracy of the Delta method in this case is $\mathcal{O}_P(1/\sqrt{\min\{n, m\}})$.

4. Inverse-inverse. Fix two natural numbers m_1 and m_2. Let $Y_{11}, \ldots, Y_{1\nu_1}$ and $Y_{21}, \ldots, Y_{2\nu_2}$ be two independent sequences of Bernoulli random variables, where $\nu_1 = \min\{n : \sum_{k=1}^{n} Y_{1k} \geq m_1\}$ and $\nu_2 = \min\{n : \sum_{k=1}^{n} Y_{2k} \geq m_2\}$. We know that the statistics $V_1 = \overline{Y}_{1m} = \nu_1/m_1$ and $V_2 = \overline{Y}_{2m} = \nu_2/m_2$ 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 6. The accuracy of the Delta method in this case is $\mathcal{O}_P(1/\sqrt{\min\{m_1, m_2\}})$.

6. ASYMPTOTIC DISTRIBUTION OF ESTIMATORS FOR THE CROSS-PRODUCT RATIO

For large values of n and m, all four estimators of the cross-product ratio ρ are differentiable functions of statistics $\overline{X}_n$ and $\overline{Y}_m$ with finite second moments; therefore the estimators are asymptotically normal. Our immediate task is to find the asymptotic of the mean and variance of these estimators. For this, we explore the standard Delta method described in Section 5.

Remember that (see, for example Propositions 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,

statistic $\overline{Y}_m$ has a mean $1/p$ and variance $\frac{q}{mp^2}$, and is asymptotically normal with these parameters.

If we use formulae for $\widehat{\rho}$, then our expressions for asymptotic variance are quite cumbersome. Hence we use the approximate estimators $\tilde{\rho}$ in Delta method derivations.

In the following we will see that the normal approximation for estimators $\tilde{\rho}$ for all sampling schemes have the same structure of means and variances:

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

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

6.1. Direct-direct Sampling Scheme

From Table 3, in this case the statistic of interest is $\widetilde{\rho}_{n_1,n_2} = \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}\left(\frac{1}{\overline{X}_{n_2}} - 1\right) = g_{dd}(V_1, V_2) = \frac{V_1}{1-V_1}\left(\frac{1}{V_2} - 1\right)$, 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 $EV_i = p_i$, $\text{Var}V_i = p_i q_i/n_i, i = 1, 2$ and $g_{dd}(p_1, p_2) = \frac{p_1}{1-p_1}\left(\frac{1}{p_2} - 1\right) = \rho$.

Partial derivatives are:

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

and hence

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

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

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

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

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

In this case, the variance component

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

From Table 3, we estimate ρ as

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

To obtain the plug-in estimator of the variance component, we substitute estimations for $\frac{p}{q}$ and p^{-1} (see Tables 1 and 2), namely $\widetilde{p_1/q_1} = \frac{\overline{X}_{n_1}}{1-\overline{X}_{n_1}}$, $\widetilde{p_2/q_2} = \frac{\overline{X}_{n_2}}{1-\overline{X}_{n_2}}$, $\widetilde{p_1^{-1}} = \frac{1}{\overline{X}_{n_1}}$, and $\widetilde{p_2^{-1}} = \frac{1}{\overline{X}_{n_2}}$ and obtain

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

6.2. Direct-inverse Sampling Scheme

From Table 3, the statistic of interest is $\widetilde{\rho_{n,m}} = \frac{\overline{X}_n(\overline{Y}_m-1)}{1-\overline{X}_n} = g_{di}(V_1, V_2) = \frac{V_1(V_2-1)}{1-V_1}$, where $V_1 = \overline{X}_n$ and $V_2 = \overline{Y}_m$. In this particular case the function $g_{di}(v_1, v_2) = \frac{v_1(v_2-1)}{1-v_1}$.

Note that $EV_1 = p_1, EV_2 = \frac{1}{p_2}, \mathrm{Var}V_1 = p_1q_1/n, \mathrm{Var}V_2 = \frac{q_2}{mp_2^2}$ and $g_{di}(p_1, \frac{1}{p_2}) = \frac{p_1\left(\frac{1}{p_2}-1\right)}{1-p_1} = \rho$.

Partial derivatives are:

$$\frac{\partial g_{di}(v_1, v_2)}{\partial v_1} = \frac{v_2-1}{(1-v_1)^2} \quad \text{and} \quad \frac{\partial g_{di}(v_1, v_2)}{\partial v_2} = \frac{v_1}{1-v_1},$$

and hence

$$\frac{\partial g_{di}(p_1, \frac{1}{p_2})}{\partial v_1} = \frac{q_2}{q_1^2 p_2} \quad \text{and} \quad \frac{\partial g_{di}(p_1, \frac{1}{p_2})}{\partial v_2} = \frac{p_1}{q_1}.$$

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

$$\widetilde{\rho}_{n,m} = g_{di}(V_1, V_2) \approx \rho + \frac{q_2}{q_1^2 p_2}(\overline{X}_n - p_1) + \frac{p_1}{q_1}\left(\overline{Y}_m - \frac{1}{p_2}\right).$$

From this, the estimator $\widetilde{\rho}_{n,m}$ is approximately normal with Mean $= \rho$ and (remembering that $\overline{X}_n$ and $\overline{Y}_m$ are independent)

$$\text{Variance} = \frac{q_2^2}{q_1^4 p_2^2}\frac{p_1q_1}{n} + \frac{p_1^2}{q_1^2}\frac{q_2}{mp_2^2} = \rho^2\left[\left(\frac{p_1}{q_1}\right)(p_1^{-1})^2/n + \left(\frac{p_2}{q_2}\right)p_2^{-1}/m\right].$$

In this case, the variance component

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

From Table 3, we estimate ρ as

$$\widehat{\rho} = \frac{n\overline{X}_n}{n+1-n\overline{X}_n}(\overline{Y}_m - 1).$$

To obtain the plug-in estimator of the variance component, we substitute estimations for $\frac{p}{q}$ and p^{-1} (see Tables 1 and 2), namely $\widetilde{p_1/q_1} = \frac{\overline{X}_n}{1-\overline{X}_n}, \widetilde{p_2/q_2} = \frac{1}{\overline{Y}_m-1}, \widetilde{p_1^{-1}} = \frac{1}{\overline{X}_n}$ and $\widetilde{p_2^{-1}} = \overline{Y}_m$ and obtain

$$\widehat{s}^2 = \frac{\overline{X}_n}{1-\overline{X}_n}\left(\frac{1}{\overline{X}_n}\right)^2/n + \frac{1}{\overline{Y}_m-1}\overline{Y}_m/m = \frac{1}{\overline{X}_n(1-\overline{X}_n)}/n + \frac{\overline{Y}_m}{\overline{Y}_m-1}/m.$$

6.3. Inverse-direct Sampling Scheme

From Table 3, the statistic of interest is $\widetilde{\rho_{m,n}} = \frac{1}{\overline{Y}_m-1}\left(\frac{1}{\overline{X}_n} - 1\right) = g_{id}(V_1, V_2) = \frac{1}{V_1-1}\left(\frac{1}{V_2} - 1\right)$, where $V_1 = \overline{Y}_m$ and $V_2 = \overline{X}_n$. In this particular case the function $g_{id}(v_1, v_2) = \frac{1}{v_1-1}\left(\frac{1}{v_2} - 1\right)$. Note that $EV_1 = \frac{1}{p_1}, EV_2 = p_2, \mathrm{Var}V_1 = \frac{q_1}{mp_1^2}, \mathrm{Var}V_2 = p_2q_2/n$ and $g_{id}(\frac{1}{p_1}, p_2) = \frac{1}{\frac{1}{p_1}-1}\left(\frac{1}{p_2} - 1\right) = \rho$.

Partial derivatives are:

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

and hence

$$\frac{\partial g_{id}(\frac{1}{p_1}, p_2)}{\partial v_1} = -\left(\frac{p_1}{q_1}\right)^2\frac{q_2}{p_2} \quad \text{and} \quad \frac{\partial g_{id}(\frac{1}{p_1}, p_2)}{\partial v_2} = -\frac{p_1}{q_1}\left(\frac{1}{p_2}\right)^2.$$

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

$$\widetilde{\rho}_{m,n} = g_{id}(V_1, V_2) \approx \rho - \left(\frac{p_1}{q_1}\right)^2 \frac{q_2}{p_2}\left(\overline{Y}_m - \frac{1}{p_1}\right) - \frac{p_1}{q_1}\left(\frac{1}{p_2}\right)^2 (\overline{X}_n - p_2).$$

From this, the estimator $\widetilde{\rho}_{m,n}$ is approximately normal with Mean $= \rho$ and (remembering that $\overline{Y}_m$ and $\overline{X}_n$ are independent)

$$\text{Variance} = \left(\frac{p_1}{q_1}\right)^4 \left(\frac{q_2}{p_2}\right)^2 \frac{q_1}{mp_1^2} + \left(\frac{p_1}{q_1}\right)^2 \left(\frac{1}{p_2}\right)^4 \frac{p_2 q_2}{n} = \rho^2 \left[\frac{p_1}{q_1} p_1^{-1}/m + \frac{p_2}{q_2}\left(p_2^{-1}\right)^2/n\right].$$

In this case, the variance component

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

From Table 3, we estimate ρ as

$$\widehat{\rho} = \left(\frac{m-1}{m\overline{Y}_m - m + 1}\right)\left(\frac{n+1}{n\overline{X}_n + 1} - 1\right) \quad \text{or} \quad \widetilde{\rho} = \frac{1}{\overline{Y}_m - 1}\left(\frac{1}{\overline{X}_n} - 1\right).$$

To obtain the plug-in estimator of the variance component, we substitute estimations for $\frac{p}{q}$ and p^{-1} (see Tables 1 and 2), namely $\widetilde{p_1/q_1} = \frac{1}{\overline{Y}_m - 1}, \widetilde{p_1^{-1}} = \overline{Y}_m$, and $\widetilde{p_2^{-1}} = \frac{1}{\overline{X}_n}$ and $\widetilde{p_2/q_2} = \frac{\overline{X}_n}{1-\overline{X}_n}$ and obtain

$$\widehat{s}^2 = \frac{1}{\overline{Y}_m - 1}\overline{Y}_m/m + \frac{\overline{X}_n}{1 - \overline{X}_n}\left(\frac{1}{\overline{X}_n}\right)^2/n = \frac{\overline{Y}_m}{m(\overline{Y}_m - 1)} + \frac{1}{n\overline{X}_n(1 - \overline{X}_n)}.$$

6.4. Inverse-inverse Sampling Scheme

From Table 3, the statistic of interest is $\widetilde{\rho}_{m_1,m_2} = \frac{\overline{Y}_{m_2} - 1}{\overline{Y}_{m_1} - 1} = g_{ii}(V_1, V_2) = \frac{V_2 - 1}{V_1 - 1}$, where $V_1 = \overline{Y}_{m_1}$ and $V_2 = \overline{Y}_{m_2}$. In this particular case the function $g_{ii}(v_1, v_2) = \frac{v_2 - 1}{v_1 - 1}$. Note that $EV_i = \frac{1}{p_i}, \text{Var}V_i = \frac{q_i}{m_i p_i^2}, i = 1, 2$ and $g_{ii}\left(\frac{1}{p_1}, \frac{1}{p_2}\right) = \frac{\frac{1}{p_2} - 1}{\frac{1}{p_1} - 1} = \rho$.

Partial derivatives are:

$$\frac{\partial g_{ii}(v_1, v_2)}{\partial v_1} = -\frac{v_2 - 1}{(v_1 - 1)^2} \quad \text{and} \quad \frac{\partial g_{ii}(v_1, v_2)}{\partial v_2} = \frac{1}{v_1 - 1},$$

and hence

$$\frac{\partial g_{ii}\left(\frac{1}{p_1}, \frac{1}{p_2}\right)}{\partial v_1} = -\left(\frac{p_1}{q_1}\right)^2 \frac{q_2}{p_2} \quad \text{and} \quad \frac{\partial g_{ii}\left(\frac{1}{p_1}, \frac{1}{p_2}\right)}{\partial v_2} = \frac{p_1}{q_1}.$$

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

$$\widetilde{\rho}_{m_1,m_2} = g_{ii}(V_1, V_2) \approx \rho - \left(\frac{p_1}{q_1}\right)^2 \frac{q_2}{p_2}\left(\overline{Y}_{m_1} - \frac{1}{p_1}\right) + \frac{p_1}{q_1}\left(\overline{Y}_{m_2} - \frac{1}{p_2}\right).$$

From this, the estimator $\widetilde{\rho}_{m_1,m_2}$ is approximately normal with Mean $= \rho$ and (remembering that $\overline{Y}_{m_1}$ and $\overline{Y}_{m_2}$ are independent)

$$\text{Variance} = \left(\frac{p_1}{q_1}\right)^4 \left(\frac{q_2}{p_2}\right)^2 \frac{q_1}{m_1 p_1^2} + \left(\frac{p_1}{q_1}\right)^2 \frac{q_2}{m_2 p_2^2} = \rho^2 \left[\frac{p_1}{q_1}\frac{1}{p_1 m_1} + \frac{p_2}{q_2}\frac{1}{p_2 m_2}\right].$$

In this case, the variance component

$$s^2(p_1, p_2) = \frac{p_1}{q_1}\frac{1}{p_1 m_1} + \frac{p_2}{q_2}\frac{1}{p_2 m_2}.$$

From Table 3, we estimate ρ as

$$\widehat{\rho} = \left(\frac{m_1 - 1}{m_1\overline{Y}_{m_1} - m_1 + 1}\right)(\overline{Y}_{m_2} - 1) \quad \text{or} \quad \widetilde{\rho} = \frac{\overline{Y}_{m_2} - 1}{\overline{Y}_{m_1} - 1}.$$

To obtain the plug-in estimator of the variance component, we substitute estimations for $\frac{p}{q}$ and p^{-1} (see Tables 1 and 2), namely $\widetilde{p_1/q_1} = \frac{1}{\overline{Y}_{m_1}-1}, \widetilde{p_2/q_2} = \frac{1}{\overline{Y}_{m_2}-1}, \widetilde{p_1^{-1}} = \overline{Y}_{m_1}$, and $\widetilde{p_2^{-1}} = \overline{Y}_{m_2}$ and obtain

$$\widehat{s}^2 = \frac{1}{\overline{Y}_{m_1} - 1}\overline{Y}_{m_1}/m_1 + \frac{1}{\overline{Y}_{m_2} - 1}\overline{Y}_{m_2}/m_2 = \frac{\overline{Y}_{m_1}}{\overline{Y}_{m_1} - 1}/m_1 + \frac{\overline{Y}_{m_2}}{\overline{Y}_{m_2} - 1}/m_2.$$

For the construction of the linear confidence intervals for the cross-product ratio ρ, the formulae for the variance component plays a crucial role. We collect the variance components in the following table.

Remark 6.1. For the schemes of inverse binomial sampling with parameters (p, m), the mean sample size is $E\nu = m/p$. If the observations are obtained in the scheme of direct sampling with the same probability p of the success and sample size $n = m/p$, then, on average, it is equivalent to the scheme of inverse sampling from the point of view of the cost for the experiment. The variance of the estimator $\tilde{\rho}_{m_1,m_2}$ coincides with the variance of the estimator $\tilde{\rho}_{n_1,n_2}$, if $m_1 = n_1p_1$ and $m_2 = n_2p_2$. Therefore *the schemes direct-direct and inverse-inverse are equivalent in the same sense from the point of view of asymptotic precision of the estimates for the cross-product ratio*. Of course, the same conclusion is true for all pairs of sampling schemes with the corresponding substitution of m by np.

We use Remark 6.4 in our comparisons of estimator accuracy for different sampling schemes.

7. CONFIDENCE LIMITS

As mentioned, the asymptotic for means and variances of estimators $\widehat{\rho}$ for the cross-product ratio ρ for all sampling schemes have the same structure: $E\widehat{\rho} = \rho$ and $\text{Var}\widehat{\rho} = \rho^2 s^2(p_1, p_2)$, where $s^2(p_1, p_2)$ is the variance component. If the sample sizes for both sampling schemes tend to infinity, then

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

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 their plug-in estimators $\hat{s}^2$ presented in Table 4, we obtain the same asymptotic equality. Therefore, if the sample sizes in both sample schemes tend to infinity, then the intervals with the following end-points,

$$\widehat{\rho}(1 \mp z_{\alpha/2}\hat{s}) \tag{1}$$

are 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 that we will consider in our further investigations.

7.1. 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 Tables 3 and 4:

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

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

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

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

Table 4. Plug-in estimators of the variance component of estimators $s^2(p_1, p_2)$ for all possible combinations of direct and inverse sampling schemes

Sampling Schemes	Second Sample Direct	Second Sample Inverse
First Sample Direct	$\frac{1}{n_1(1-\overline{X}_{n_1})\overline{X}_{n_1}} + \frac{1}{n_2(1-\overline{X}_{n_2})\overline{X}_{n_2}}$	$\frac{1}{n\overline{X}_n(1-\overline{X}_n)} + \frac{\overline{Y}_m}{m(\overline{Y}_m-1)}$
First Sample Inverse	$\frac{\overline{Y}_m}{m(\overline{Y}_m-1)} + \frac{1}{n\overline{X}_n(1-\overline{X}_n)}$	$\frac{\overline{Y}_{m_1}}{m_1(\overline{Y}_{m_1}-1)} + \frac{\overline{Y}_{m_2}}{m_2(\overline{Y}_{m_2}-1)}$

Below, we provide the results of statistical modeling in Table 5. For each pair (n_1, n_2) of sample sizes and values (p_1, p_2) of success probabilities, we present 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 5 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 $p_1, p_2 \geq 0.2$.

7.2. Direct-inverse Sampling Scheme

When the first sample is obtained by the direct sampling scheme with sample size n and the second sample is obtained by the inverse sampling scheme with the number of successes m, then, according to Tables 3 and 4:

$$\widehat{\rho}_{n,m} = \frac{n\overline{X}_n}{n+1-n\overline{X}_n}(\overline{Y}_m - 1) \quad \text{and} \quad \widehat{s}^2 = \frac{1}{n\overline{X}_n(1-\overline{X}_n)} + \frac{\overline{Y}_m}{m(\overline{Y}_m - 1)}.$$

Hence, the asymptotic $n, m \to \infty$ confidence interval (1) based on $\overline{X}_n$ and $\overline{Y}_m$ can be written as

$$\frac{n\overline{X}_n(\overline{Y}_m - 1)}{n+1-n\overline{X}_n}\left(1 \mp z_{\alpha/2}\sqrt{\frac{1}{n\overline{X}_n(1-\overline{X}_n)} + \frac{\overline{Y}_m}{m(\overline{Y}_m - 1)}}\right). \tag{3}$$

Below, we provide the results of statistical modeling in Table 6. For each pair (n, m) of sample size and number of successes, and values (p_1, p_2) of success probabilities, we present Monte-Carlo estimations of the coverage probability, mean width, and standard deviation of the width for the confidence interval (3). The nominal level is assumed to be 0.95.

The results of Table 6 show that the interval (3) has a confidence level lower than nominal and an error not larger than 0.02 only in the region $p_1, p_2 \geq 0.2$ for $n = 200$, $m = 100$, and in other cases the coverage probability is not influenced by m.

7.3. Inverse-direct Sampling Scheme

When the first sample is obtained by the inverse sampling scheme with the number of successes m, and the second sample is obtained by the direct sampling scheme with a sample size n, then, according to Tables 3 and 4:

$$\widehat{\rho}_{m,n} = \left(\frac{m-1}{m\overline{Y}_m - m + 1}\right)\left(\frac{n+1}{n\overline{X}_n + 1} - 1\right) \quad \text{and} \quad \widehat{s}^2 = \frac{\overline{Y}_m}{m(\overline{Y}_m - 1)} + \frac{1}{n\overline{X}_n(1-\overline{X}_n)}.$$

Hence, the asymptotic $m, n \to \infty$ confidence interval (1) based on $\overline{Y}_m$ and $\overline{X}_n$ can be written as

$$\left(\frac{m-1}{m\overline{Y}_m - m + 1}\right)\left(\frac{n+1}{n\overline{X}_n + 1} - 1\right)\left(1 \mp z_{\alpha/2}\sqrt{\frac{\overline{Y}_m}{m(\overline{Y}_m - 1)} + \frac{1}{n\overline{X}_n(1-\overline{X}_n)}}\right). \tag{4}$$

Below, we provide the results of statistical modeling in Table 7. For each pair (m, n) of number of successes and sample size, and values (p_1, p_2) of success probabilities, we present 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.

Table 5. 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	0.895	0.866	0.896	0.907	0.900	0.901	0.912	0.910
		2.046	0.445	0.122	1.705	0.396	0.105	1.542	0.370	0.096
		1.276	0.189	0.053	0.722	0.138	0.039	0.526	0.113	0.031
	0.5	0.877	0.903	0.894	0.897	0.909	0.909	0.910	0.919	0.912
		7.490	1.588	0.443	6.044	1.374	0.371	5.279	1.252	0.329
		4.699	0.673	0.189	2.624	0.498	0.139	1.896	0.415	0.112
	0.8	0.871	0.878	0.877	0.885	0.890	0.888	0.891	0.897	0.895
		34.039	7.514	2.043	29.073	6.794	1.801	26.513	6.423	1.657
		26.164	4.719	1.281	18.799	4.179	1.106	16.072	3.957	1.005
100	0.2	0.887	0.911	0.901	0.909	0.918	0.904	0.922	0.926	0.924
		1.792	0.372	0.106	1.412	0.314	0.086	1.204	0.280	0.075
		1.088	0.139	0.039	0.555	0.092	0.026	0.347	0.068	0.019
	0.5	0.890	0.908	0.907	0.910	0.928	0.919	0.923	0.931	0.928
		6.782	1.370	0.395	5.134	1.115	0.314	4.225	0.963	0.262
		4.236	0.497	0.138	1.987	0.324	0.092	1.244	0.242	0.067
	0.8	0.885	0.898	0.898	0.907	0.911	0.909	0.915	0.920	0.916
		29.051	6.052	1.712	23.021	5.151	1.413	19.744	4.633	1.230
		18.610	2.650	0.728	10.406	2.003	0.547	7.723	1.73	0.457
200	0.2	0.892	0.913	0.909	0.917	0.927	0.925	0.930	0.935	0.932
		1.654	0.329	0.096	1.230	0.262	0.075	0.987	0.222	0.061
		1.015	0.113	0.031	0.459	0.067	0.019	0.259	0.045	0.013
	0.5	0.896	0.919	0.913	0.920	0.931	0.926	0.931	0.934	0.937
		6.405	1.251	0.371	4.629	0.965	0.281	3.590	0.785	0.222
		3.991	0.416	0.113	1.718	0.243	0.068	0.940	0.159	0.045
	0.8	0.892	0.911	0.902	0.916	0.924	0.924	0.926	0.932	0.930
		26.488	5.299	1.541	19.822	4.229	1.203	15.949	3.591	0.989
		16.216	1.919	0.523	7.799	1.244	0.344	4.803	0.942	0.259

As for the previous direct-inverse sampling scheme, it is possible to conclude that the coverage probability is almost not influenced by m, the number of expected success in the first sample. Moreover, the same remark is equally applicable to the precision properties of the interval. The coverage probability by the interval (4) is significantly lower than nominal (see Table 7). We recommend to apply this interval only if it is known that the success probability in the second sample $p_2 \geq 0.2$, $p_1 = 0.2$ for $n = 200$, $m = 50$ and $p_2 \geq 0.2$, $p_1 = 0.2, 0.5$ for $n = 200$, $m = 100$.

Table 6. Coverage probability, width, and standard deviation for confidence interval (3)

	m	20			50			100		
		p_2								
n	p_1	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
		0.892	0.879	0.862	0.909	0.904	0.879	0.916	0.912	0.901
	0.2	1.687	0.462	0.148	1.507	0.395	0.115	1.440	0.369	0.102
		0.645	0.193	0.075	0.477	0.135	0.046	0.413	0.112	0.036
		0.900	0.893	0.859	0.915	0.911	0.898	0.919	0.917	0.909
50	0.5	5.955	1.671	0.555	5.127	1.366	0.417	4.818	1.253	0.355
		2.286	0.683	0.266	1.731	0.479	0.163	1.541	0.415	0.126
		0.882	0.878	0.851	0.894	0.890	0.888	0.898	0.896	0.891
	0.8	28.722	7.803	2.442	26.048	6.802	1.947	24.988	6.416	1.741
		18	4.880	1.597	16.073	4.202	1.196	15.238	3.968	1.082
		0.908	0.896	0.876	0.923	0.919	0.902	0.930	0.927	0.918
	0.2	1.384	0.394	0.135	1.155	0.313	0.098	1.068	0.281	0.082
		0.448	0.139	0.059	0.292	0.087	0.033	0.235	0.066	0.023
		0.913	0.905	0.880	0.927	0.923	0.910	0.933	0.931	0.925
100	0.5	5.030	1.467	0.518	4.013	1.110	0.366	3.609	0.964	0.294
		1.575	0.491	0.211	1.028	0.301	0.114	0.840	0.234	0.079
		0.903	0.897	0.872	0.915	0.913	0.904	0.921	0.918	0.911
	0.8	22.598	6.415	2.169	19.008	5.134	1.599	17.679	4.630	1.340
		9.099	2.671	1.022	7.056	1.934	0.637	6.454	1.720	0.502
		0.916	0.909	0.878	0.931	0.926	0.915	0.936	0.932	0.923
	0.2	1.200	0.354	0.127	0.929	0.261	0.088	0.819	0.222	0.070
		0.342	0.110	0.049	0.198	0.061	0.025	0.146	0.043	0.016
		0.918	0.913	0.890	0.933	0.929	0.917	0.937	0.938	0.930
200	0.5	4.508	1.356	0.499	3.331	0.960	0.338	2.829	0.785	0.259
		1.235	0.395	0.183	0.693	0.212	0.088	0.509	0.149	0.056
		0.914	0.907	0.884	0.929	0.925	0.915	0.933	0.931	0.927
	0.8	19.298	5.689	2.042	15.035	4.217	1.421	13.281	3.591	1.124
		6.042	1.865	0.808	3.984	1.152	0.427	3.329	0.913	0.302

7.4. Inverse-inverse Sampling Scheme

When both samples are obtained by the inverse sampling scheme with the number of successes m_1 and m_2, then, according to Tables 3 and 4:

$$\widehat{\rho}_{m_1,m_2} = \left(\frac{m_1 - 1}{m_1\overline{Y}_{m_1} - m_1 + 1}\right)(\overline{Y}_{m_2} - 1) \quad \text{and} \quad \widehat{s}^2 = \frac{\overline{Y}_{m_1}}{m_1(\overline{Y}_{m_1} - 1)} + \frac{\overline{Y}_{m_2}}{m_2(\overline{Y}_{m_2} - 1)}.$$

Table 7. Coverage probability, width, and standard deviation for confidence interval (4)

	n	50			100			200		
		p_2								
m	p_1	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
		0.884	0.900	0.894	0.904	0.914	0.908	0.916	0.917	0.919
	0.2	1.805	0.375	0.106	1.422	0.317	0.087	1.218	0.285	0.076
		1.130	0.156	0.043	0.616	0.113	0.032	0.435	0.094	0.025
		0.875	0.890	0.882	0.888	0.902	0.895	0.900	0.904	0.902
20	0.5	8.010	1.728	0.476	6.662	1.534	0.410	5.981	1.424	0.372
		5.662	0.966	0.265	3.751	0.814	0.214	3.122	0.736	0.189
		0.830	0.831	0.817	0.826	0.842	0.840	0.843	0.844	0.844
	0.8	42.807	9.949	2.614	39.074	9.520	2.441	37.716	9.235	2.340
		40.174	8.831	2.317	34.261	8.337	2.142	33.020	8.027	2.033
		0.893	0.916	0.909	0.917	0.928	0.924	0.930	0.934	0.932
	0.2	1.629	0.321	0.094	1.190	0.251	0.072	0.940	0.209	0.058
		1.004	0.110	0.030	0.448	0.067	0.019	0.259	0.047	0.013
		0.891	0.911	0.901	0.911	0.920	0.921	0.923	0.929	0.924
50	0.5	6.825	1.378	0.397	5.181	1.124	0.316	4.272	0.974	0.264
		4.269	0.520	0.145	2.110	0.360	0.099	1.385	0.283	0.077
		0.870	0.885	0.882	0.889	0.897	0.895	0.900	0.902	0.900
	0.8	32.671	7.044	1.940	27.169	6.295	1.676	24.488	5.865	1.521
		25.044	4.249	1.170	16.670	3.773	1.002	14.730	3.509	0.893
		0.897	0.921	0.915	0.920	0.932	0.929	0.934	0.941	0.936
	0.2	1.563	0.301	0.090	1.105	0.226	0.067	0.831	0.177	0.051
		0.932	0.097	0.026	0.407	0.053	0.015	0.212	0.033	0.009
		0.896	0.917	0.913	0.918	0.929	0.926	0.931	0.938	0.933
100	0.5	6.405	1.253	0.370	4.645	0.968	0.281	3.595	0.788	0.223
		3.943	0.423	0.115	1.741	0.252	0.071	0.975	0.172	0.048
		0.884	0.906	0.898	0.904	0.916	0.918	0.914	0.921	0.925
	0.8	28.117	5.795	1.647	21.860	4.828	1.337	18.451	4.270	1.145
		17.571	2.420	0.661	9.594	1.798	0.488	6.952	1.511	0.405

Hence, the asymptotic $m_1, m_2 \to \infty$ confidence interval (1) based on $\overline{Y}_{m_1}$ and $\overline{Y}_{m_2}$ can be written as

$$\frac{(m_1-1)(\overline{Y}_{m_2}-1)}{m_1\overline{Y}_{m_1}-m_1+1}\left(1\mp z_{\alpha/2}\sqrt{\frac{\overline{Y}_{m_1}}{m_1(\overline{Y}_{m_1}-1)}+\frac{\overline{Y}_{m_2}}{m_2(\overline{Y}_{m_2}-1)}}\right). \qquad (5)$$

Below, we provide the results of statistical modeling in Table 8. For each pair (m_1, m_2) of numbers of successes and values (p_1, p_2) of success probabilities, we present Monte-Carlo estimations of the coverage probability, mean width, and standard deviation of the width for the confidence interval (5). The nominal level is assumed to be 0.95.

Table 8. Coverage probability, width, and standard deviation for confidence interval (5)

	m_2	20			50			100		
		p_2								
m_1	p_1	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
20	0.2	0.904	0.896	0.871	0.919	0.915	0.900	0.924	0.920	0.912
		1.397	0.396	0.135	1.171	0.317	0.099	1.084	0.284	0.083
		0.528	0.157	0.063	0.390	0.110	0.037	0.343	0.092	0.029
	0.5	0.891	0.881	0.856	0.904	0.899	0.886	0.906	0.904	0.897
		6.568	1.805	0.584	5.823	1.535	0.452	5.544	1.430	0.396
		3.457	0.974	0.340	2.992	0.813	0.241	2.804	0.738	0.204
	0.8	0.836	0.828	0.816	0.842	0.839	0.829	0.847	0.843	0.837
		39.142	10.159	2.914	37.286	9.465	2.531	36.672	9.217	2.402
		34.443	8.958	2.664	32.416	8.267	2.226	31.872	8.080	2.089
50	0.2	0.916	0.912	0.884	0.930	0.927	0.916	0.936	0.934	0.927
		1.161	0.346	0.126	0.878	0.250	0.086	0.761	0.208	0.067
		0.335	0.106	0.048	0.202	0.060	0.024	0.157	0.045	0.016
	0.5	0.912	0.906	0.878	0.926	0.919	0.911	0.931	0.930	0.921
		5.06	1.474	0.519	4.050	1.118	0.367	3.650	0.973	0.296
		1.695	0.519	0.217	1.188	0.337	0.120	1.029	0.275	0.088
	0.8	0.889	0.893	0.860	0.901	0.897	0.886	0.903	0.902	0.898
		26.820	7.382	2.354	23.908	6.269	1.843	22.739	5.863	1.619
		16	4.480	1.470	14.431	3.736	1.086	13.632	3.531	0.956
100	0.2	0.920	0.915	0.885	0.936	0.933	0.920	0.939	0.939	0.933
		1.072	0.328	0.123	0.760	0.224	0.082	0.620	0.177	0.061
		0.276	0.089	0.043	0.144	0.045	0.020	0.100	0.030	0.012
	0.5	0.919	0.914	0.887	0.932	0.929	0.918	0.936	0.934	0.929
		4.513	1.359	0.499	3.338	0.963	0.338	2.844	0.787	0.259
		1.264	0.401	0.185	0.739	0.222	0.089	0.573	0.163	0.059
	0.8	0.909	0.905	0.874	0.921	0.916	0.909	0.925	0.923	0.919
		21.494	6.167	2.128	17.672	4.809	1.532	16.179	4.269	1.265
		8.215	2.427	0.958	6.259	1.722	0.574	5.660	1.496	0.447

The results of Table 8 show that the interval (5) has the confidence level lower than nominal and an error less than 0.02 when $p_1, p_2 = 0.2, 0.5$ for $m_1 = 50$, $m_2 = 100$ and $p_1 = 0.2, 0.5$, $p_2 \geq 0.2$ for $m_1, m_2 = 100$.

8. CONCLUDING REMARKS AND FURTHER RESEARCH

As we can see from the simulation results presented in the previous sections, in some cases the linear confidence intervals for the cross-product ratio coefficient has the confidence level lower than nominal. We expect that this flow may be overcame by considering the logarithmic confidence intervals, which is the topic of our further research. Consideration of accuracy properties of the point estimators for the cross product ratio is also an interesting problem.

ACKNOWLEDGEMENTS

The third author's research was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, project no. 1.13556.2019/13.1.

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. **37**, 466–496 (2016).
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, *Elements of Large Sample Theory* (Springer, New York, 1997).
4. L. A. Goodman, "Simultaneous confidence limits for cross-product ratios in contingency tables," J. R. Stat. Soc., Ser. B **26**, 86–102 (1964).
5. J. J. Gart, "On the combination of relative risks," Biometrics **18**, 601–610 (1962).
6. J. Cornfield, "A statistical problem arising from retrospective studies," in *Proceedings of the 3rd Berkeley symposium on Mathematical Statistics and Probability* (Univ. of California Press, Berkeley, CA, 1956), Vol. 4.
7. J. Anděl, "On interactions in contingency tables," Apl. Mat. **18**, 99–109 (1973).
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**, 60–69 (1983).
10. P. W. Holland and Y. J. Wang, "Dependence functions for Continuous bivariate densities," Commun. Stat. "— Theory Methods **16**, 863–876 (1987).
11. Y. J. Wang, "The probability integrals of bivariate normal distributions: A contingency table approach," Biometrika **74**, 185–190 (1987).
12. M. H. McCann and J. M. Tebbs, "Simultaneous logit-based confidence intervals for odds ratios in $2 \times k$ classification tables with a fixed reference level," Commun. Stat. "— Simul. Comput. **38**, 961–975 (2009).
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. **37**, 529–536 (2010).
14. J. L. Xu, "A property of the generalized proportional hazards model," Far East J. Theor. Stat. **41**, 149–162 (2012).
15. F. Schaarschmidt, D. Gerhard, and C. Vogel, "Simultaneous confidence intervals for comparisons of several multinomial samples," Comput. Stat. Data Anal. **106**, 65–76 (2017).
16. T. Niebuhr and M. Trabs, "Profiting from correlations: Adjusted estimators for categorical data," Appl. Stoch. Models Business Ind. **35**, 1090–1102 (2019).
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, 1–33 (2020).