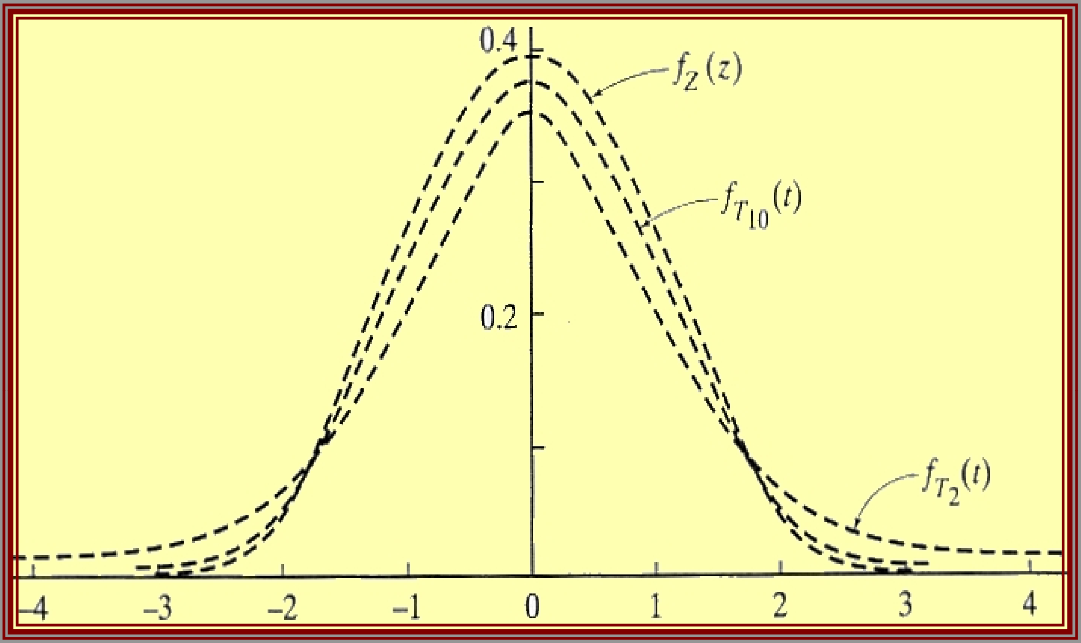
ISSN 1726-3328

JPSS

A comprehensive journal of probability and statistics
for theorists, methodologists, practitioners, teachers, and others



JOURNAL OF PROBABILITY AND STATISTICAL SCIENCE

Volume 18 Number 2 August 2020

ISSN 1726-3328

JPSS

Journal of Probability and Statistical Science

Volume 18 Number 2 August 2020

Table of Contents

Theory and Methods

Teaching and Applications

Appendix

Published by: Susan Rivers' Cultural Institute, Hsinchu City, Taiwan, ROC.

Journal of Probability and Statistical Science
18(2), 121-128, Aug. 2020

A Note on the Confidence Estimation for the Ratio of Binomial Proportions with the Inverse-Direct Sampling Scheme

Parichart Pattarapanitchai Kamon Budsaba Tannen Acoose Andrei Volodin
Thammasat University *University of Regina*

ABSTRACT We continue to investigate the estimation of the binomial proportions ratio. In this article we focus on estimating confidence, by considering the question of confidence interval construction for a ratio of two proportions using data from two independent Bernoulli samples. We focus on the case when using the Inverse Binomial sampling schemeto obtainthe first sample and the Direct Binomial sampling scheme to obtain the second. Unfortunately in so-called First Special case of the Inverse-Direct sampling scheme, the confidence interval construction for the ratio of binomial proportion has a serious disadvantage: the upper bound of this interval may be negative. This creates a problem because the value of the parameter under estimation is always positive. In this article we investigate the proportion of negative upper bound for different number of trials.

Keywords Confidence estimators; Direct binomial sampling; Inverse binomial sampling; Ratio of binomial proportions.

1. Introduction

In prospective studies, biological experiments, and the comparison of manufacturing processes for quality control in industry, the ratio of two binomial proportions occurs. Statistical procedures for the ratio of binomial proportions (often called the relative risk) are also quite common in clinical trials, epidemiological studies, and the pharmaceutical setting. In epidemiological problems, such as cohort studies in two groups, the risk ratio or odds ratio, is related to vaccine efficiency and attributable risk. A clinical trial is designed to test the effectiveness of the new drug to reduce mortality.

□Received February 2020, revised April 2020, in final form May 2020.

□Parichart Pattarapanitchai and Kamon Budsaba are affiliated to the Department of Mathematics and Statistics at Thammasat University, Rangsit Campus, Pathum Thani, 12120 Thailand. Tannen Acoose and Andrei Volodin (corresponding author) are affiliated to the Department of Mathematics and Statistics at University of Regina, Regina, SK, S4S 0A2 Canada; email address of Andrei Volodin: andrei@uregina.ca.

JPSS: ISSN 1726-3328

People are interested in knowing whether or not certain air pollutants increase the chances of a disease by twofold in various public health applications. One of the main objectives of clinical trials is to test whether a new drug performs better than acurrent drug (or placebo) to cure a specific disease. The problem can be identified as investigating whether the new drug has a success rate of 1.5 times the current drug (or placebo). Conversely, if we evaluate the incidence rate of the disease (more applicable in vaccine studies), then we are interested in testing whether or not the new treatment reduces the chances of the disease occurring by a certain amount (for example, 50 percent).

In this article we continue our investigation into the estimation of the ratio of Binomial proportions started in Ngamkham *et al.* [1] and Ngamkham [2]. The literature on the estimation of the ratio of Binomial proportions is very extensive and this problem has attracted the attention of statisticians for more than 70 years. We refer readers to the extensive list of references given in Ngamkham *et al.* [1].

The general problem of estimating the ratio of Binomial proportions 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. The observations are done according to the sequential sampling schemes with Markov stopping times ν_1 and ν_2. From the results of observations $X^{(\nu_1)} = (X_1, \cdots, X_{\nu_1})$ and $Y^{(\nu_2)} = (Y_1, \cdots, Y_{\nu_2})$, it is necessary to identify the sampling scheme and corresponding statistic for the estimator of the ratio $\theta = p_1 / p_2$.

We now remind readers of the definitions of the Direct and Inverse Binomial sample schemes. To fix the notation, we present it here and make the article more self-contained. For details we refer to Ngamkham *et al.* [1] and Ngamkham [2].

Direct Binomial sampling: A random vector $X^{(n)} = (X_1, \cdots, X_n)$ with Bernoulli components and a fixed number of observations n is observed. In the case of direct Binomial sampling, there is no unbiased estimate for the parametric function p^{-1}. An estimate of p^{-1} with an exponentially decreasing rate of bias is

$$\frac{n+1}{n\bar{X}_n + 1},$$

where

$$\bar{X}_n = \frac{T}{n} \qquad \text{and} \qquad T = \sum_{k=1}^{n} X_k .$$

Note that $\bar{X}_n$ is asymptotically normal with the mean p and variance $p(1-p)/n$.

Inverse Binomial sampling: A Bernoulli sequence $Y^{(\nu)} = (Y_1, \cdots, Y_\nu)$ is observed with a stopping time

$$\nu = \min\left\{ n : \sum_{k=1}^{n} Y_k \geq m \right\}.$$

That is, the components of the sequence $Y_1, Y_2, \cdots$ are observed until the given number m of successes appears. In the case of Inverse Binomial sampling the statistic $\overline{Y}_m = \nu / m$ is asymptotically normal with mean $1/p$ and variance $(1-p)/(mp^2)$. The statistic $(m-1)/(\nu-1)$ is an unbiased estimate of *p*. In the following, keep the notation $X_1, X_2, \cdots$ for a Bernoulli sequence obtained by the direct sampling scheme and $Y_1, Y_2, \cdots$ for a Bernoulli sequence obtained by the inverse sampling scheme.

In this article, we consider the Inverse-Direct sampling scheme, where the first sample is obtained by the Inverse Binomial sampling scheme with the probability of success p_1 and a stopping time that is defined by the fixed number of success *m*. The second sample is obtained by the Direct Binomial sampling scheme with the probability of success p_2 and a fixed sample size *n*.

Ngamkham [3] states that, relative to other sampling schemes the Inverse-Direct sample scheme, estimator performs the worst. It appears that the Mean Squared Error (MSE) for the Inverse-Direct sample scheme estimator is approximately ten times larger than the MSE for the Special Case Direct-Inverse estimator. However, there are two Special Cases of the Inverse-Direct sampling schemes that were not considered in Ngamkham [3]. In Ngamkham *et al.* [1] and Ngamkham [2], both Inverse-Direct Special Cases were considered, albeit briefly, and without any actual derivation of formulae. There is also a concern with the Special Case of Inverse-Direct sampling scheme shown in Section 2, which is not mentioned in Ngamkham *et al.* [1] and Ngamkham [2]. In the current article, all these issues are discussed.

In this article, we concentrate most on the Special Cases of the Inverse-Direct sampling scheme where the first sample is obtained by the Inverse sampling scheme and the second sample is obtained by the Direct sampling scheme where the number of trials n is the same as the number of observations in the first experiment. Article Ngamkham (2020) [3] fails to mention a situation of the First Inverse-Direct Special Case, and there is a typographical error in the formula for the variance in Ngamkham [2].

Now we remind readers of point estimators for the ratio of Binomial proportions for the Special Case of the Inverse-Direct sampling scheme. There are two such estimators, and in this article we are interested in the so-called First Special Case of the Inverse-Direct Sampling scheme. It appears in the following framework.

Special cases of the Inverse-Direct Sampling Scheme: The first sample is obtained by the Inverse sampling method with parameters (m, p_1). A proposed sampling plan for the second sample follows. Let ν be the (random) sample size for the first sample: the value achieved after m successes. This value, ν, from the first sample is used in the planning of the second sample. For the second sample, the number of trials, n, is the same as the number of observations in the first experiment; set $n = \nu$. Denote

$$T_\nu = \sum\nolimits_{k=1}^{\nu} X_k .$$

The random variable ν does not depend on $X_1, X_2, \cdots$; therefore, it is possible to calculate the mean value and variance of T_ν and its distribution. Since there is a typographical error in the formulae for the variance in Ngamkham [2] for these calculations, and no actual calculations are presented in Ngamkham *et al.* [1], we present the correct derivation.

Note that $T_\nu = \sum_{i=1}^{\nu} X_i$ is the sum of n independent identically distributed Bernoulli random variables $X_1, X_2, \cdots$ with parameters p_2 and ν, which has the Pascal (Negative Binomial) distribution with parameters m and p_1. Hence,

$$E(X_1) = p_2, \quad Var(X_1) = p_2(1-p_2), \quad E(\nu) = \frac{m}{p_1}, \text{ and } Var(\nu) = \frac{m(1-p_1)}{p_1^2}$$

(see, for example, Section 2.1 in Ngamkham [2]). By Theorem III.6.2 of Gut [4],

$$E(T_\nu) = E(\nu)E(X_1) = \frac{mp_2}{p_1} = \frac{m}{\theta}$$

and

$$Var(T_\nu) = E(\nu)Var(X_1) + [E(X_1)]^2 Var(\nu) = \frac{m}{p_1} p_2(1-p_2) + p_2^2 \frac{m(1-p_1)}{p_1^2}$$

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

Note that T_ν / m is an unbiased estimate for the ratio $p_2 / p_1 = 1/\theta$. As we mentioned above, $(m-1)/(\nu-1)$ is an unbiased estimator of p_1. Hence, in the formula for the variance of T_ν / m we substitute θ^{-1} and p_1 by these estimators:

$$\overset{\square}{Var}\left(\frac{T_\nu}{m}\right) = \frac{1}{m}\left[\frac{T_\nu}{m} + \frac{T_\nu^2}{m^2}\left(1 - 2\frac{m-1}{\nu-1}\right)\right].$$

Thus, the main idea of the First Special Case of the Inverse-Direct sampling scheme is to estimate $1/\theta$ by the estimator T_ν / m. The asymptotically $(1-\alpha)$-confidence interval for θ^{-1} is

$$\frac{T_\nu}{m} \mp z_{\alpha/2}\sqrt{\overset{\square}{Var}\left(\frac{T_\nu}{m}\right)},$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution.

Based on this, the following asymptotically $(1-\alpha)$-confident interval of θ for the First Special case of the Inverse-Direct sampling scheme was considered in Ngamkham [1] and Ngamkham [2]:

$$\left(\left(\frac{T_\nu}{m} + z_{\alpha/2}\sqrt{\overset{\square}{Var}\left(\frac{T_\nu}{m}\right)}\right)^{-1}, \left(\frac{T_\nu}{m} - z_{\alpha/2}\sqrt{\overset{\square}{Var}\left(\frac{T_\nu}{m}\right)}\right)^{-1}\right). \qquad (*)$$

2. Negative Upper Bound of the Special Case Inverse-Direct

Unfortunately the $(1-\alpha)$-confidence interval of θ presented above has a serious disadvantage. It can be described as follows. The upper bound of this interval is in the form of subtraction; hence, it becomes negative when

$$\frac{T_\nu}{m} < z_{\alpha/2}\sqrt{\overset{\square}{Var}\left(\frac{T_\nu}{m}\right)}.$$

Obviously this happens for small value of T_ν.

This creates a problem for the confidence interval construction for the First Special Case of the Inverse-Direct sampling scheme. The value of the parameter θ under estimation is always positive, so a negative upper bound for a confidence interval is not acceptable.

In this section, we investigate the proportion of negative upper bounds for different number of trials. For each case, 10^5 simulations of the confidence interval have been performed. The number of occurring negative upper bound was collected for each situation according to the simulation results. In all case we put $p_2 = p_1$. The proportion of the negative upper bound is shown in Figures 1-3.

From Figures 1-3 we see that the higher the p_1, the smaller the proportion of lower negative upper bounds. The large steps tend to decrease if m is increased by one. We consider m to be $50p_1$, $100p_1$, and $200p_1$. The proportion of negative upper bound is presented in Table 1 for each value of m and some values of p_1.

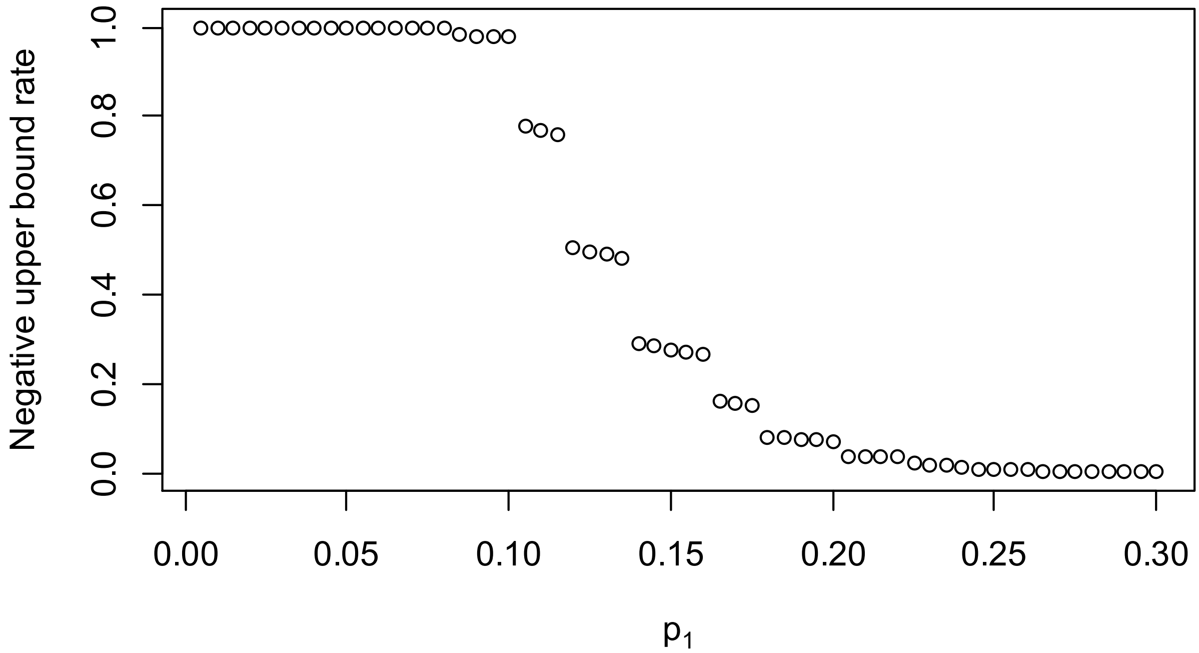


Figure 1 Proportion of negative upper bounds for $m = 50p_1$ and $p_2 = p_1$

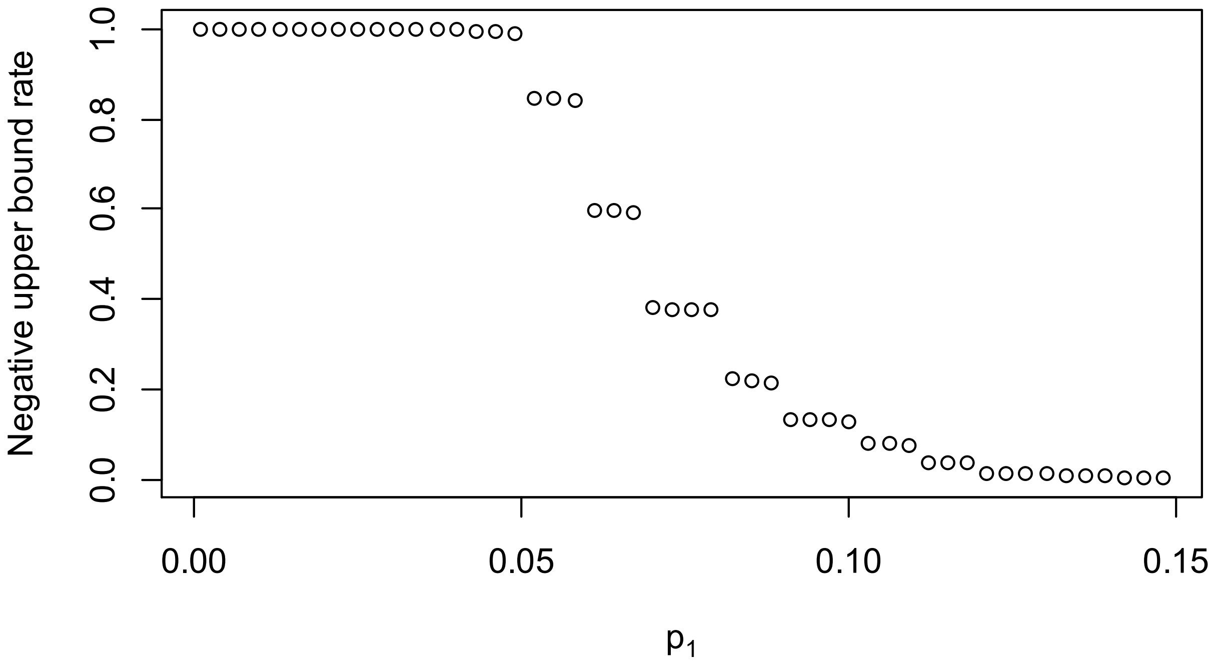


Figure 2 Proportion of negative upper bounds for $m = 100p_1$ and $p_2 = p_1$

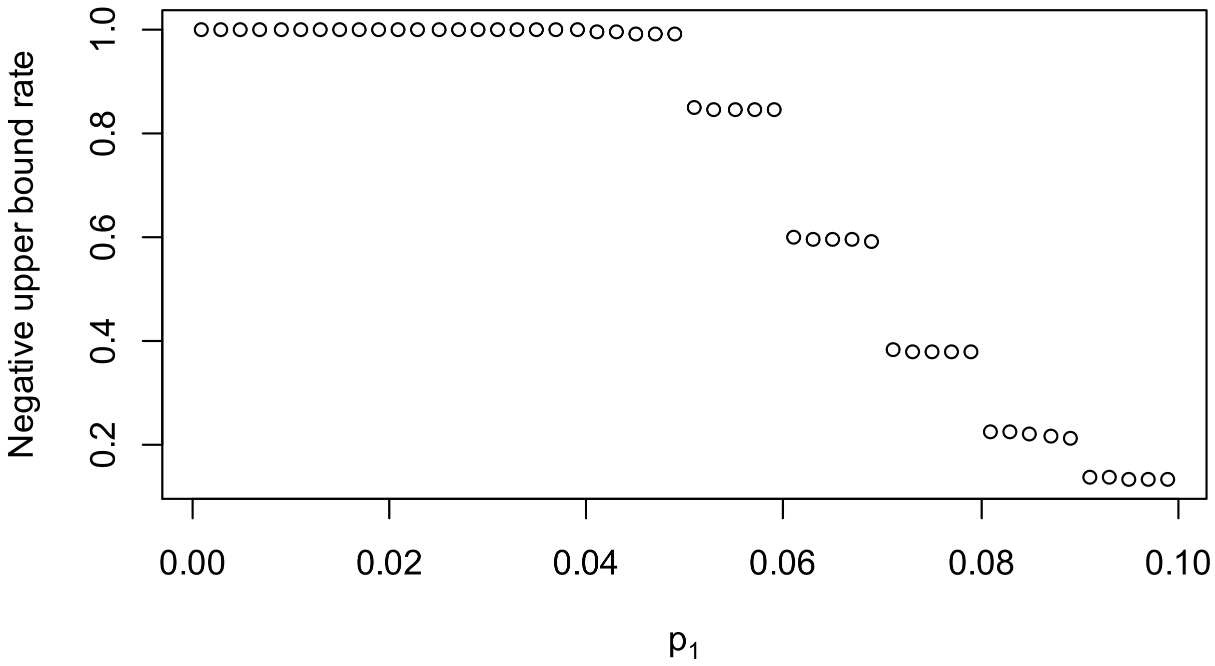


Figure 3 Proportion of negative upper bounds for $m = 200p_1$ and $p_2 = p_1$

Table 1 The negative upper bound rate for each value of m and $p_2 = p_1$

	Proportion of negative upper bounds		
m	$p_1 = m/50$	$p_1 = m/100$	$p_1 = m/200$
1	1.000	1.000	1.000
2	1.000	1.000	1.000
3	1.000	1.000	1.000
4	1.000	1.000	1.000
5	0.981	0.992	0.995
6	0.768	0.844	0.878
7	0.493	0.596	0.650
8	0.278	0.380	0.417
9	0.154	0.219	0.293
10	0.076	0.134	0.143
11	0.036	0.081	0.097
12	0.019	0.039	0.065
13	0.009	0.017	0.039
14	0.004	0.010	0.015
15	0.002	0.005	0.007

There is only a slight difference in proportions in each row of Table 1. This implies that the proportion of negative bounds mainly depends on m.

3. Conclusion and Recommendations

The usage of the confidence interval $(*)$ to estimate θ may cause a problem with the occurrence of negative upper bounds. This obviously provides an incorrect estimation, because the value of θ is always positive. This problem occurs when there is a small T_v value produced by a low m and a small probability of successes.We recommend the number of m to be at least 13 to avoid this problem by more than 95%. In the case $m = 15,$ we can avoid the problem by more than 99% even for small probabilities of success.

Another concern is that, theoretically, it may happen that $T_v = 0$ and formula $(*)$ does not have any sense. This situation can also happen with a low m and small probability of successes p_2. To solve this problem, the recommendation is to consider the estimator conditional on $T_v \neq 0,$ as it is considered in Ngamkham [1], or assume at least one success for the second sample.

Acknowledgments

The work of the last listed author was carried out in the framework of the program of support of the Mathematical Center of the Volga Region Federal District (Project No. 075-02-2020-1478).

References

[1] Ngamkham, T., Volodin, A., and Volodin, I. (2016). Confidence intervals for a ratio of binomial proportions based on direct and inverse sampling schemes, *Lobachevskii Journal of Mathematics*, **37**(4), 466-494.

[2] Ngamkham, T. (2018). Confidence Interval Estimation for the Ratio of Binomial Proportions and Random Numbers Generation for Some Statistical Models (Ph.D. Thesis, University of Regina, 2018).

[3] Ngamkham, T. (2020). Comparison of accuracy properties of point estimators for the ratio of binomial proportions with different sampling schemes, to appear in *Thailand Statistician*.

[4] Gut, A. (2013). Probability: A Graduate Course, Springer Science & Business Media.