Journal of Mathematical Extension
Vol. 5, No. 1 (2010), 87-95

Large Sample Inference on the Ratio of Two Independent Binomial Proportions

H. Haghbin *
Shiraz University

M. R. Mahmoudi
Shiraz University

Z. Shishebor
Shiraz University

Abstract. The asymptotic distribution for the ratio of sample proportions in two independent bernoulli populations is introduced. The presented method can be used to derive the asymptotic confidence interval and hypothesis testing for the ratio of population proportions. The performance of the new interval is comparable with similar confidence intervals in the large sample cases. Then the simulation study is provided to compare our confidence interval with some other methods. The proposed confidence set has a good coverage probability with a shorter length.

AMS Subject Classification: 62F03; 62F05; 62F12.
Keywords and Phrases: Binomial distribution, Cramer's theorem, ratio of proportions, Slutsky's theorem.

1. Introduction

It is of interest to make inference about the ratio of the proportions two independent binomials. This parameter is more applicable than the difference of proportions in some applications. The advantage of using ratio instead of difference lies in the fact that the difference of two small

Received October 2010; Final Revised December 2010
*Corresponding author

proportions is also small and has no meaningful description (see for instance [1]). Many researchers have presented some methods to inference about the ratio of proportions. This can be found in [2-7]. ([8]) Compared several Wald-type intervals, as the special cases of a non-iterative approximation to a Bayesian interval. ([9]) proposed an exact unconditional joint confidence set for two binomial parameters estimated from independent samples. In the present work, the asymptotic distribution for the ratio of sample proportions is presented. It will be applied to construct asymptotic confidence interval and perform test statistics. This method is the most efficient way in comparison with other methods where sample size is large.

2. Large Sample Inference

Let $X_1, ..., X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from two Bernoulli distributions with the parameters p_1 and p_2, respectively. We are interested in making inference about the parameter $R = \frac{p_1}{p_2}$. Since $\bar{p}_1 = \frac{\sum_{i=1}^{m} X_i}{m}$ and $\bar{p}_2 = \frac{\sum_{i=1}^{n} Y_i}{n}$are consistent estimators for p_1 and p_2, it seems reasonable to estimate R by $\bar{R} = \frac{\bar{p}_1}{\bar{p}_2}$. Note that by strong large number theorem, $\bar{p}_2 \underset{\rightarrow}{a.s}\ p_2 \neq 0$ and therefore, $\bar{R}$ is well-defined in large sample theory. There is no loss in assuming m=n . In the following theorem, we will give the asymptotic distribution of $\bar{R}$.

Theorem 2.1. *Under the above assumptions,*

$$\sqrt{n}\left(\bar{R} - R\right) \underset{\rightarrow}{\mathcal{L}} N\left(0, \sigma^2\right) \quad as \quad n \to \infty$$

where $\sigma^2 = \frac{R}{p_2}\left(R\left(1 - p_2\right) + 1 - p_1\right).$

Proof. By the central limit theorem, $\sqrt{n}\left(\bar{p}_i - p_i\right)$converges in law to $N\left(0, p_i\left(1 - p_i\right)\right)$ as $n \to \infty$ for $i = 1, 2$. By independence of $\bar{p}_1$and $\bar{p}_2$and Slutsky's theorem we have

$$\sqrt{n}\left(\begin{pmatrix}\bar{p}_1\\ \bar{p}_2\end{pmatrix} - \begin{pmatrix}p_1\\ p_2\end{pmatrix}\right) \underset{\rightarrow}{\mathcal{L}} N\left(0, \Sigma\right) \quad \text{as} \quad n \to \infty$$

where

$$\Sigma = \begin{bmatrix} p_1(1-p_1) & 0 \\ 0 & p_2(1-p_2) \end{bmatrix}$$

Let's define $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ as $g(x_1, x_2) = \frac{x_1}{x_2}$. Then the gradient function with respect to g is $\bigtriangledown g(x_1, x_2) = \left(\frac{1}{x2}, -\frac{x_1}{x2}\right)$. Also $\bigtriangledown g(p_1, p_2)\Sigma \bigtriangledown g(p_1, p_2)^T = \sigma^2$. Since $\bigtriangledown g$ is continuous in neighborhood of (p_1, p_2), therefore, by Cramer's rule we have

$$\sqrt{n}\left(g(\bar{p}_1, \bar{p}_2) - g(p_1, p_2)\right) = \sqrt{n}\left(\bar{R} - R\right) \underset{\rightarrow}{\mathcal{L}} N\left(0, \sigma^2\right) \ as \ n \to \infty$$

By the theorem we have just proved,

$$T_n = \sqrt{n}\left(\frac{\bar{R} - R}{\sigma}\right) \underset{\rightarrow}{\mathcal{L}} N(0,1) \quad \text{as} \quad n \to \infty \tag{1}$$

This result can be used to construct an asymptotic confidence interval and hypothesis testing as follows:

2.1 Asymptotic Confidence Interval

Since the parameter σ in T_n depends on the unknown parameter R, it cannot be used as a pivotal quantity for the parameter R. In the following theorem, we try to estimate the parameter σ.

Theorem 2.1.1. *On the same hypothesis of Theorem 2.1.*

$$T_n^* = \sqrt{n}\left(\frac{\bar{R} - R}{\overset{\wedge}{\sigma}_n}\right) \underset{\rightarrow}{\mathcal{L}} N(0,1) \quad as \quad n \to \infty \tag{2}$$

where $\overset{\wedge}{\sigma}_n = \left(\frac{\bar{R}}{\bar{p}_2}\left(\bar{R}(1-\bar{p}_2) + 1 - \bar{p}_1\right)\right)^{1/2}$.

Proof. By the weak law of large numbers $\bar{p}_i - p_i = o_p(1), i = 1, 2$. From

this and Slutsky's theorem we have

$$\begin{pmatrix} \bar{p}_1 \\ \bar{p}_2 \end{pmatrix} \xrightarrow{P} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \; as \; n \to \infty$$

Let's define $f : (0,1) \times (0,1) \to \mathbb{R}^+$ as

$$f(x,y) = \sqrt{\frac{x(1-x)}{y^2} + \frac{x^2(1-y)}{y^3}}$$

By Slutsky's theorem $\overset{\wedge}{\sigma}_n - \sigma = o_p(1)$.Using Theorem 2.1. completes the proof.
Now T_n^* can be used as a pivotal quantity to construct an asymptotic confidence interval for R,

$$\left(\bar{R} - \frac{\overset{\wedge}{\sigma}_n}{\sqrt{n}} z_{\alpha/2} \; , \; \bar{R} + \frac{\overset{\wedge}{\sigma}_n}{\sqrt{n}} z_{\alpha/2} \right) \tag{3}$$

2.2 Hypothesis Testing

Hypothesis testing about R is important in practice. For instance, the assumption $R = 1$ is equivalent to the assumption $p_1 = p_2$. In general, to test $H_0 : R = R_0$, the test statistic can be

$$T_0 = \sqrt{n} \left(\frac{\bar{R} - R_0}{\sqrt{\frac{R_0}{\bar{p}_2} \left(R_0 \left(1 - \bar{p}_2 \right) + 1 - \bar{p}_1 \right)}} \right) \tag{4}$$

By similar methodology applied in Theorem 2.1.1. it can be shown that under null hypothesis, T_0 has asymptotic standard normal distribution. Note that, in the case $n \neq m$, it is sufficient to replace n by $n^* = \min(m,n)$ in the above results. It is easy to see that the power function of test $H_0 : R = R_0$ on the basis of the test statistic (4), is as follows:

$$\beta(p_1, p_2) = \sum_{i=1}^{m} \sum_{j=1}^{n} \binom{m}{i} \binom{n}{j} p_1^i p_2^j (1-p_1)^{m-i} (1-p_2)^{n-j} I_A(i,j)$$

where

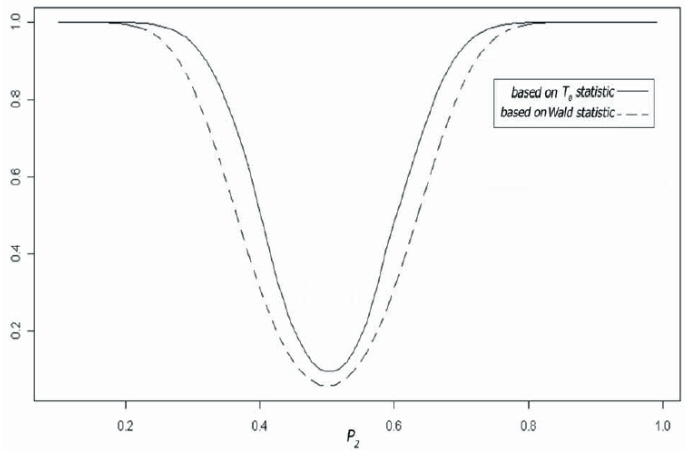
$$A=\left\{(x,y):\sqrt{n^*}\left|\frac{nx}{my}-R_0\right|>z_{\alpha/2}\sqrt{\frac{n}{y}\left(R_0\left(1-\frac{y}{n}\right)+1-\frac{x}{n}\right)}\right\}$$

In the case $R_0=1$, which is equivalent to $p_1=p_2$, we can also use Wald statistics as follows:

$$W=\frac{\bar{p}_1-\bar{p}_2}{\sqrt{\bar{p}\left(1-\bar{p}\right)\left(\frac{1}{m}+\frac{1}{n}\right)}},$$

where $\bar{p}=\frac{\sum_{i=1}^{m}X_i+\sum_{i=1}^{n}Y_i}{m+n}$. Figure 1. compares the power functions based on the test statistics W and T_0 with $p_1=0.5$. As can be seen, although the probability of the first type error based on W is less than that based on T_0, the power of the test based on $T_0$is greater than that based on W.

Figure 1: Power function to test $H_0: R=1$, when $p_1=0.5$.



Remark 2.3. *This method can be applied to other distributions, such as geometric distribution. By the same method used in the proof of Theorem 1, one can see that if $X_1,...,X_m$ and $Y_1,\ldots,Y_n$ are two independent samples from two geometric distributions, then*

$$\sqrt{n}\left(\bar{R}-R\right)\xrightarrow{\mathcal{L}}N\left(0,\sigma^2\right)\qquad as\quad n\to\infty\quad,$$

where $\sigma^2 = \frac{1}{p_2{}^2}\left(\frac{1-p_1}{p_1{}^2} + R^2\frac{1-p_2}{p_2{}^2}\right)$. *Also, it is easy to show that,*

$$T_n^* = \sqrt{n}\left(\frac{\bar{R}-R}{\hat{\sigma}_n}\right) \xrightarrow{\mathcal{L}} N(0,1) \ \ as\ n \to \infty$$

where $\hat{\sigma}_n = \left(\frac{1}{\bar{p}_2{}^2}\left(\frac{1-\bar{p}_1}{\bar{p}_1{}^2} + \bar{R}^2\frac{1-\bar{p}_2}{\bar{p}_2{}^2}\right)\right)^{1/2}$.

3. Simulation Study

In this section, we provide a simulation study to compare our confidence interval (3) with similar works, in view of the empirical coverage and average length. The best general reference is [8] which compared the following three large sample confidence intervals for R:

Log-limits method:

$$\bar{R}\exp\left(\pm z_{\alpha/2}\sqrt{\frac{1}{\sum_{i=1}^m X_i} + \frac{1}{\sum_{i=1}^n Y_i} - \frac{1}{m} - \frac{1}{n}}\right)$$

$\mathbf{LOG_{0.5}}$ method:

$$\theta_{0.5}\exp\left(\pm z_{\alpha/2}\sqrt{var\left(\log\left(\hat{\theta}_{0.5}\right)\right)}\right),$$

where

$$\log\left(\hat{\theta}_{0.5}\right) = \log\left(\frac{\sum_{i=1}^m X_i + 0.5}{m+0.5}\right) - \log\left(\frac{\sum_{i=1}^n Y_i + 0.5}{n+0.5}\right)$$

and

$$var\left(\log\left(\hat{\theta}_{0.5}\right)\right) = \frac{1}{\sum_{i=1}^m X_i + 0.5} + \frac{1}{\sum_{i=1}^n Y_i + 0.5} - \frac{1}{m+0.5} - \frac{1}{n+0.5}$$

Inverse hyperbolic sine method:

$$\bar{R}\exp\left(\pm 2\sinh^{-1}\left(\frac{z_{\alpha/2}}{2}\sqrt{\frac{1}{\sum_{i=1}^m X_i} + \frac{1}{\sum_{i=1}^n Y_i} - \frac{1}{m} - \frac{1}{n}}\right)\right)$$

We simulate 50000 times of the above confidence intervals with $m = n = 50, 100, 200$ and 500 for different values of p_1 and p_2. The empirical coverage and mean lengths are summarized in Table 1.
We see that, in terms of the empirical probability coverage, our method is weaker than other methods when sample size is small. But in large sample size cases, all the intervals have the same empirical probability coverage. In terms of the length of the interval, our method is the best in all cases. Therefore, the critical region which is constructed by inverting our confidence interval has more power than the critical regions corresponding to the other confidence intervals. This subject can be seen by a simulation study to test $H_0 : R = 1$, the empirical powers of the tests are presented in Table 2.

Table 1.CI-1:Confidence interval in(4),CI-2:Log-limits CI-3:LOG0.5 and CI-4: inverse hyperbolic confidence intervals

		p_1=0.3,p_2=0.7		p_1=0.8,p_2=0.7		p_1=0.7,p_2=0.3	
n	CI	coverage	length	coverage	length	coverage	length
50	CI-1	0.93976	0.3965921	0.94702	0.532843	0.94	2.391832
	CI-2	0.95328	0.4114769	0.95194	0.537808	0.95392	2.494745
	CI-3	0.94954	0.4103981	0.95254	0.530169	0.9501	2.376941
	CI-4	0.94932	0.407461	0.94948	0.536542	0.94972	2.466623
100	CI-1	0.9442	0.2797325	0.95006	0.372997	0.94816	1.597835
	CI-2	0.94924	0.2848217	0.9519	0.374674	0.95056	1.628632
	CI-3	0.94918	0.2845568	0.9515	0.372071	0.95056	1.593943
	CI-4	0.94924	0.2835016	0.9519	0.374251	0.95056	1.620625
200	CI-1	0.94748	0.1975773	0.95086	0.262139	0.94892	1.101314
	CI-2	0.95146	0.199349	0.95238	0.262718	0.95138	1.111466
	CI-3	0.9509	0.1992746	0.95214	0.261815	0.95064	1.100122
	CI-4	0.951	0.1988979	0.9523	0.262572	0.95082	1.10888
500	CI-1	0.9492	0.1248825	0.95046	0.165343	0.95068	0.685916
	CI-2	0.95036	0.1253268	0.95118	0.165488	0.95074	0.68838
	CI-3	0.9506	0.1253109	0.95068	0.165262	0.95016	0.685638
	CI-4	0.95014	0.1252149	0.95086	0.165452	0.95048	0.68776

R package version 2.11.1software has been employed for the computations in this simulation.

Table 2. T-1: Empirical powers of the tests based on T_0 in (4),T-2: Test based on Log-limits approach, T-3: Test based on LOG0.5 approach and T-4: Test based on inverse hyperbolic approach

n	C.R	p_1=0.3,p_2=0.7	p_1=0.5,p_2=0.8	p_1=0.2,p_2=0.3	p_1=0.2,p_2=0.5
50	T-1	0.99302	0.92846	0.361	0.95032
	T-2	0.98022	0.8927	0.18808	0.88716
	T3	0.97918	0.892	0.1844	0.88556
	T-4	0.98712	0.8933	0.2044	0.8903
75	T-1	0.99948	0.98586	0.4433	0.99028
	T-2	0.99862	0.97686	0.28258	0.97562
	T-3	0.99862	0.97686	0.2717	0.97492
	T-4	0.99862	0.97748	0.29346	0.97638
100	T-1	1	0.99692	0.51168	0.998
	T-2	0.99994	0.99472	0.35778	0.99486
	T-3	0.99994	0.99472	0.34932	0.9948
	T-4	0.99996	0.99472	0.36698	0.99488
150	T-1	1	0.99992	0.63484	0.99996
	T-2	1	0.99984	0.50654	0.99984
	T-3	1	0.99984	0.50506	0.99984
	T-4	1	0.99984	0.5143	0.99988

Acknowledgement
We would like to thank the editor and referees for their constructive comments.

References

[1] A. Agresti,*Categorical Data Analysis,* Wiley, New York, 2002.

[2] G. E. Noether, Two confidence intervals for the ratio of two probabilities and some measures of effectiveness, *Journal of the American Statistical Association,* 52 (1957), 3645.

[3] D. G. Thomas and J. J. Gart, A table of exact confidence limits for differences and ratios of two proportions and their odds ratio, *Journal of the American Statistical Association,* 72 (1977), 73-76.

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

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

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

[7] R. G. Newcombe, Logit confidence intervals and the inverse sinh transformation, *American Statistician*, 55 (2001), 200-202.

[8] R. M. Price and D. G. Bonett, Confidence intervals for a ratio of two independent binomial proportions, *Statist. Med,.* 27 (2008), 5497-5508.

[9] J. Reiczigel, Z. Abonyi-Tóth a, and J. Singer, An exact confidence set for two binomial proportions and exact unconditional confidence intervals for the difference and ratio of proportions, *Computational Statistics and Data Analysis,* 52 (2008), 5046-5053.

Hossein Haghbin
Department of Statistics
Ph.D Student of Statistics
Shiraz University
Shiraz, Iran.
E-mail: haghbinh@gmail.com

Mohammad Reza Mahmoudi
Department of Statistics
Ph.D Student of Statistics
Shiraz University
Shiraz, Iran.
E-mail: mrmahmuoodi@shirazu.ac.ir

Zohreh Shishebor
Department of Statistics
Associated Professor of Statistics
Shiraz University
Shiraz, Iran.
E-mail: sheshebor@susc.ac.ir