

# Sequential Risk-Efficient Estimation for the Ratio of Two Binomial Proportions

Hokwon Cho

Department of Mathematical Sciences  
University of Nevada-Las Vegas  
Las Vegas, NV 89154-4020, USA

## Abstract

A risk-efficient sequential point estimator is considered for the ratio of two independent binomial proportions based on maximum likelihood estimation under squared error loss and cost proportional to the observations. It is assumed that the cost per observation is constant. First-order asymptotic expansions are obtained for large-sample properties of the proposed procedure. Performance of the procedure is studied through the criteria of risk efficiency and regret analysis. Monte Carlo simulation is carried out to obtain the expected sample size that minimizes the risk and to examine its finite sample behavior. An example is provided to illustrate its use.

*Key words:* Risk-efficient sequential point estimator; ratio of two binomial proportions; first order asymptotics; risk efficiency; regret analysis.

## 1 Introduction

Let  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  be two sequences of independent Bernoulli random variables with non-zero probabilities  $p_0$  and  $p_1$ , respectively. Moreover, let  $\theta = p_1/p_0$ . The problem addressed in this paper is to estimate the true ratio of the two binomial proportions when the loss incurred is of the form

$$L_n = (\hat{\theta}_n - \theta)^2 + cn, \quad (1)$$

where  $\hat{\theta}_n$  is the maximum likelihood estimator (MLE) and  $c (> 0)$  is the known cost per unit of observations  $(X, Y)$ . Then, the risk is

$$\begin{aligned} R_n(c) &= E(\hat{\theta}_n - \theta)^2 + cn \\ &= \text{Var}(\hat{\theta}_n) + B^2 + cn, \end{aligned} \quad (2)$$

where  $B$  represents the bias which is defined by  $B = E(\hat{\theta}_n) - \theta$ .

The ratio of two binomial proportions arises in prospective studies, biological experiments or comparison of manufacturing processes for quality control in industry. It has been an important tool for measuring the risk ratio (Katz et al., 1978, Fleiss, 1981, and Bailey, 1987) or the relative risk (Gart, 1985). In epidemiological problems, such as cohort studies in two groups, the risk ratio or odds ratio is related to vaccine efficacy and attributable risk (Walter, 1976). However, all of these methods have dealt with approximate interval estimators based on the logarithmic method (Katz et al., 1978), chi-square method (Koopman, 1984), power divergence family of statistics (Bedrick, 1987) or likelihood score method (Gart and Nam, 1988). By contrast, Aitchison and Bacon-Shone (1981) obtained exact credible intervals using a Bayesian approach. For sequential point estimation, see Ghosh, Mukhopadhyay and Sen (1997).

For estimating the ratio  $\theta$ , it would be desirable to take  $n$  as large as possible so that the risk can be made adequately small. This may not, however, be practical, since drawing (or measuring) observations involve an expense, and therefore drawing a large sample would naturally increase costs. Thus, it seems quite appropriate to incorporate a cost function  $c(n)$  into the loss function. For instance, suppose we incur a sampling or manufacturing cost  $c$  per unit and collect the data  $(\mathbf{x}, \mathbf{y})$ . After measuring some observations, we must decide whether to terminate sampling and estimate the ratio  $\theta$  or to continue sampling. It will be shown in Section 3 that it is not possible to specify a sample size in advance to solve the problem stated above.

We now propose a risk-efficient sequential point estimator of the ratio of two binomial proportions that has not been considered before, which provides an optimal sample size under squared error loss incorporating with the cost of observation.

## 2 Properties of Ratios: $\hat{\theta}_n$ and $\tilde{\theta}_n$

Let  $R$  and  $S$  be two independent binomial random variables with parameters  $(n, p_0)$  and  $(n, p_1)$ , respectively. Let the observed proportions be denoted by  $\hat{p}_0 = r/n$  and  $\hat{p}_1 = s/n$ , when  $n$  pairs  $(X_i, Y_i)$ ,  $i = 1, 2, \dots, n$  are observed. Define the ratio  $\hat{\theta}_n = S/R$  and the modified ratio by

$$\tilde{\theta}_n = \frac{S}{R + \varepsilon}$$

where  $\varepsilon$  ( $0 < \varepsilon < 1$ ) is an auxiliary constant to avoid the case of undefined  $\hat{\theta}_n$  when  $R = 0$ . One can take  $\varepsilon = 1/2$  commonly for practical purpose. (See also Bailey (1987), Gart and Nam (1988)).

### 2.1 Expectations and Bias

Now, we consider the expectation of the estimator,  $\tilde{\theta}_n$  and its bias.

$$E\left(\tilde{\theta}_n\right) = E\left(\frac{S}{R + 1/2}\right) = E(S) E\left(\frac{1}{R + 1/2}\right) = np_1 E\left(\frac{1}{R + 1/2}\right). \quad (3)$$

However,

$$\begin{aligned}
E\left(\frac{1}{R+1/2}\right) &= E\left(\frac{1}{np_0 + R - np_0 + 1/2}\right) \\
&= \frac{1}{np_0} E\left[\left(1 + \frac{R - np_0 + 1/2}{np_0}\right)^{-1}\right] \\
&= \frac{1}{np_0} E\left[1 - \left(\frac{R - np_0 + 1/2}{np_0}\right) + \left(\frac{R - np_0 + 1/2}{np_0}\right)^2 + \dots\right] \\
&= \frac{1}{np_0} \left[1 - \frac{1}{2np_0} + \frac{np_0(1-p_0)}{(np_0)^2} + \frac{1}{4(np_0)^2} + \dots\right] \\
&= \frac{1}{np_0} - \frac{1}{2(np_0)^2} + \frac{np_0(1-p_0)}{(np_0)^3} + \frac{1}{4(np_0)^3} + \dots. \tag{4}
\end{aligned}$$

Hence, it follows from Eqs. (3) and (4) that

$$E(\tilde{\theta}_n) = np_1 \left[ \frac{1}{np_0} - \frac{1}{2(np_0)^2} + \frac{np_0(1-p_0)}{(np_0)^3} + \frac{1}{4(np_0)^3} + \dots \right]. \tag{5}$$

Therefore, the bias  $\tilde{B}$  becomes

$$\begin{aligned}
\tilde{B} = E(\tilde{\theta}_n) - \theta &= -\frac{p_1}{2np_0^2} + \frac{p_1(1-p_0)}{np_0^2} + O(n^{-2}) \\
&= \frac{p_1(-p_0 + 1/2)}{np_0^2} + O(n^{-2}). \tag{6}
\end{aligned}$$

Thus,  $\tilde{B}^2 = O(n^{-2})$  and can be neglected in the expansion for the risk.

For sufficiently large  $n$ , it follows from Eq. (5) that

$$E(\tilde{\theta}_n) \simeq \frac{np_1}{np_0} = \theta.$$

That is,  $\tilde{\theta}_n$  is an asymptotically unbiased estimator of  $\theta$ .

## 2.2 Asymptotic Variances

To get the variance of  $\hat{\theta}_n$ , we consider the maximum likelihood estimates of  $\theta$  and  $p_0$ , and their information matrix. From the observed sample of  $n$  pairs of  $(X_i, Y_i)$ ,  $i = 1, 2, \dots, n$ , the likelihood function is

$$\begin{aligned}
L(\theta, p_0) &\propto p_0^r (1-p_0)^{n-r} (p_0\theta)^s (1-p_0\theta)^{n-s} \\
&= p_0^{r+s} (1-p_0)^{n-r} \theta^s (1-p_0\theta)^{n-s}, \tag{7}
\end{aligned}$$

where  $p_1 = p_0\theta$ . The log-likelihood function of Eq. (7) is then

$$\begin{aligned}
l(\theta, p_0) &\propto (r+s) \log p_0 + (n-r) \log(1-p_0) \\
&\quad + s \log \theta + (n-s) \log(1-p_0\theta). \tag{8}
\end{aligned}$$

The maximum likelihood estimate of  $\theta$ ,  $\hat{\theta}_n$  is obtained by setting  $\partial l(\theta, p_0) / \partial \theta = 0$ , which gives

$$\hat{\theta}_n = s/np_0,$$

and letting  $\partial l(\theta, p_0) / \partial p_0 = 0$  gives the MLE of  $p_0$ ,

$$\hat{p}_0 = r/n.$$

Then, from Eq. (8) Fisher's information matrix about  $(\theta, p_0)$  is given by

$$\begin{aligned} \mathbf{I}(\theta, p_0) &= \begin{bmatrix} E\left(-\frac{\partial^2 l(\theta, p_0)}{\partial \theta^2}\right) & E\left(-\frac{\partial^2 l(\theta, p_0)}{\partial \theta \partial p_0}\right) \\ E\left(-\frac{\partial^2 l(\theta, p_0)}{\partial \theta \partial p_0}\right) & E\left(-\frac{\partial^2 l(\theta, p_0)}{\partial p_0^2}\right) \end{bmatrix} \\ &= n \begin{bmatrix} \frac{p_0^2}{p_1(1-p_1)} & \frac{1}{1-p_1} \\ \frac{1}{1-p_1} & \frac{1}{p_0} \left(\frac{1}{1-p_0} + \frac{\theta}{1-p_1}\right) \end{bmatrix}. \end{aligned} \quad (9)$$

So,

$$\mathbf{I}^{-1}(\theta, p_0) = \frac{\theta(1-p_0)(1-p_1)}{n} \begin{bmatrix} \frac{1}{p_0} \left(\frac{1}{1-p_0} + \frac{\theta}{1-p_1}\right) & -\frac{1}{1-p_1} \\ -\frac{1}{1-p_1} & \frac{p_0^2}{p_1(1-p_1)} \end{bmatrix}. \quad (10)$$

Therefore, from Eq. (10) the asymptotic variance of  $\hat{\theta}_n$  is

$$\begin{aligned} \text{Var}(\hat{\theta}_n) &= \frac{\theta(1-p_0)(1-p_1)}{n} \left[ \frac{1}{p_0} \left(\frac{1}{1-p_0} + \frac{\theta}{1-p_1}\right) \right] \\ &= \frac{\theta(1+\theta-2\theta p_0)}{np_0}. \end{aligned} \quad (11)$$

Furthermore, using Slutsky's theorem and the asymptotic normality of  $\hat{\theta}_n$ , it follows that

$$\sqrt{n}(\hat{\theta}_n - \theta) \stackrel{d}{\simeq} N(0, \sigma^2), \quad (12)$$

where  $\sigma^2 = \theta(1+\theta-2\theta p_0)/p_0$ .

Now, we consider the asymptotic variance of  $\tilde{\theta}_n = S/(R+1/2)$ .

$$\begin{aligned} \text{Var}(\tilde{\theta}_n) &= E\left[\left(\frac{S}{R+1/2}\right)^2\right] - \left[E\left(\frac{S}{R+1/2}\right)\right]^2 \\ &= E(S^2) \cdot E\left(\frac{1}{R+1/2}\right)^2 - \left[E(S) \cdot E\left(\frac{1}{R+1/2}\right)\right]^2. \end{aligned} \quad (13)$$

Noting that

$$\begin{aligned}
E\left(\frac{1}{R+1/2}\right)^2 &= E\left[\frac{1}{np_0}\left(1+\frac{R-np_0+1/2}{np_0}\right)^{-1}\right]^2 \\
&= \frac{1}{(np_0)^2}E\left[1-2\left(\frac{R-np_0+1/2}{np_0}\right)+3\left(\frac{R-np_0+1/2}{np_0}\right)^2+\dots\right] \\
&= \frac{1}{(np_0)^2}\left[1-\frac{1}{np_0}+\frac{3np_0(1-p_0)}{(np_0)^2}+\frac{3}{4(np_0)^2}+\dots\right].
\end{aligned}$$

Using this and the previous result in Eq. (4), it follows from Eq. (13),

$$\begin{aligned}
Var(\tilde{\theta}_n) &= [np_1(1-p_1)+(np_1)^2]\left[\frac{1}{(np_0)^2}\left\{1-\frac{1}{np_0}+\frac{3np_0(1-p_0)}{(np_0)^2}+\frac{3}{4(np_0)^2}+\dots\right\}\right] \\
&\quad -\left[np_1\left\{\frac{1}{np_0}-\frac{1}{2(np_0)^2}+\frac{(1-p_0)}{(np_0)^2}+\frac{1}{4(np_0)^3}+\dots\right\}\right]^2 \\
&= \left[\frac{p_1(1-p_1)}{np_0^2}+\theta^2\right]\left[1+\frac{2-3p_0}{np_0}+\frac{3}{4(np_0)^2}+\dots\right] \\
&\quad - (np_1)^2\left[\frac{1}{(np_0)^2}-\frac{1}{(np_0)^3}+\frac{2(1-p_0)}{(np_0)^3}+\frac{1}{4(np_0)^4}+\frac{(1-p_0)^2}{(np_0)^4}+\frac{(1-p_0)}{2(np_0)^4}+\dots\right]
\end{aligned}$$

After simplification it gives us

$$\begin{aligned}
Var(\tilde{\theta}_n) &= \left[\frac{\theta(1-p_1)}{np_0}+\theta^2\right]\left[1+\frac{2-3p_0}{np_0}+\frac{3}{4(np_0)^2}+\dots\right] \\
&\quad - (np_1)^2\left[\frac{1}{(np_0)^2}\left\{1+\frac{1-2p_0}{np_0}+\frac{1}{4(np_0)^2}+\frac{(1-p_0)^2}{(np_0)^2}+\dots\right\}\right] \\
&= \left[\frac{\theta(1-p_1)}{np_0}+\theta^2\right]\left[1+\frac{2-3p_0}{np_0}+O(n^{-2})\right]-\theta^2\left[1+\frac{1-2p_0}{np_0}+O(n^{-2})\right] \\
&\approx \frac{\theta(1-p_1)}{np_0}\left(1+\frac{2-3p_0}{np_0}\right)+\theta^2\left[1+\frac{2-3p_0}{np_0}\right]-\theta^2\left[1+\frac{1-2p_0}{np_0}\right] \\
&= \frac{\theta(1-p_1)}{np_0}\left(1+\frac{2-3p_0}{np_0}\right)+\theta^2\left(\frac{1-p_0}{np_0}\right) \\
&= \frac{\theta(1-p_1)}{np_0}+\frac{\theta^2(1-p_0)}{np_0}+O(n^{-2}) \\
&\approx \frac{\theta(1+\theta-2\theta p_0)}{np_0}. \tag{14}
\end{aligned}$$

From the results of Eqs. (11) and (14), we see that two ratios have the asymptotically same variance. Hence, we conclude that two estimators,  $\hat{\theta}_n = S/R$  and  $\tilde{\theta}_n = S/(R+1/2)$  are asymptotically equivalent for large  $n$ .

**Remark 2.1** One can say that  $n \left( \hat{\theta}_n - \tilde{\theta}_n \right)$  converges to  $\theta/2p_0$  in probability as  $n$  tends to  $\infty$ .

### 3 Stopping Rule $N$ and Its Properties

In this section, we derive the risk-efficient stopping rule for the proposed sequential procedure and study the asymptotic properties of the rule.

#### 3.1 Risk-Efficient Stopping Rule

Taking Eq. (6) into account, the risk given in Eq. (2) can be rewritten as

$$\begin{aligned} R_n(c) &\simeq \text{Var}(\tilde{\theta}_n) + cn \\ &= \frac{\theta(1 + \theta - 2\theta p_0)}{np_0} + cn = \frac{\sigma^2}{n} + cn. \end{aligned} \quad (15)$$

Then, analytically, the risk in Eq. (15) is minimized by solving

$$g(n) = \frac{\partial}{\partial n} [R_n(c)|\theta] = \frac{-\theta(1 + \theta - 2\theta p_0)}{n^2 p_0} + c = 0,$$

which yields

$$n = c^{-1/2} \sigma \quad (16)$$

where  $\sigma = [\theta(1 + \theta - 2\theta p_0)/p_0]^{1/2}$ . Hence, we take the optimal fixed-sample size  $n^*$  as the integer such that  $n \leq n^* \leq n+1$ , for estimating  $\theta$  when everything is known, namely

$$n^* = \lceil c^{-1/2} \sigma \rceil + 1 \quad (17)$$

where  $\lceil \cdot \rceil$  indicates the greatest integer function. The minimum risk associated with the optimal fixed-sample size  $n^*$  is

$$R_{n^*}(c) = E \left( \tilde{\theta}_{n^*} - \theta \right)^2 + cn^*.$$

Since both  $\theta$  and  $p_0$  are unknown, there is no fixed-sample size procedure that will attain the risk in Eq. (15). So, the following adaptive sequential rule for determining the sample size is proposed: stop sampling at  $N$  when

$$N = \inf \left\{ n \geq m : n \geq c^{-1/2} \hat{\sigma}_n \right\} \quad (18)$$

where  $m (\geq 2)$  is the initial sample size, and since when  $\hat{p}_{0n} = R/n$  is zero (i.e.,  $R = 0$ )  $\hat{\sigma}_n$  is undefined, replacing  $\hat{p}_{0n}$  with  $(R + 1/2)/n$  we set

$$\hat{\sigma}_n = \begin{cases} \left[ \tilde{\theta}_n \left( 1 + \tilde{\theta}_n - 2\hat{p}_{1n} \right) 2n \right]^{1/2}, & \text{if } R = 0 \\ \left[ \tilde{\theta}_n \left( 1 + \tilde{\theta}_n - 2\hat{p}_{1n} \right) / \hat{p}_{0n} \right]^{1/2}, & \text{if } R \geq 1 \end{cases}.$$

Then, the risk function associated with the stopping time  $N$  is given by

$$R_N(c) = E \left[ (\tilde{\theta}_N - \theta)^2 + cN \right]. \quad (19)$$

### 3.2 Finite Sure Termination

The following result establishes the fundamental property that the proposed stopping rule terminates finitely almost surely. Proofs of this theorem and those to follow are given in the Appendix.

**Theorem 3.1** *Let  $N$  denote the stopping time associated with the proposed sequential procedure. Then  $P\{N = \infty\} = 0$ .*

### 3.3 First Order Asymptotics

Justification of the proposed procedure rests primarily upon its good asymptotic behavior for sufficiently small  $c$ . Theoretically, we are not able to investigate the small sample behavior of the proposed procedure, but it is possible to study the asymptotic behavior of the procedure when sample size is sufficiently large. Therefore, since the random stopping time  $N$  is a function of  $c$ , one can get large enough  $n$  by letting  $c$  get small.

The stopping time  $N$  given by Eq. (18) can be rewritten as

$$N = \inf \left\{ n \geq m : \frac{n}{(c\hat{p}_{0n})^{1/2}} \geq \left( \frac{\tilde{\theta}_n}{\theta} \right)^{1/2} \left[ \frac{\tilde{\theta}_n (1 + \tilde{\theta}_n - 2\tilde{\theta}_n \hat{p}_{0n})}{c\hat{p}_{0n}} \right]^{1/2} \right\}. \quad (20)$$

Then, the rule Eq. (20) takes the form

$$N = N(t) = \min \{ n \geq m : W_n \leq f(n)/t \},$$

where

$$W_n = \left( \frac{\tilde{\theta}_n}{\theta} \right)^{1/2} \left[ \frac{\tilde{\theta}_n (1 + \tilde{\theta}_n - 2\tilde{\theta}_n \hat{p}_{0n})}{c\hat{p}_{0n}} \right]^{1/2},$$

$$f(n) = n,$$

and

$$t = (c\hat{p}_{0n})^{1/2}.$$

Then clearly,  $W_n$  is a sequence of random variables such that  $W_n > 0$ ,  $\lim_{n \rightarrow \infty} W_n = 1$  almost surely (a.s.) since  $\hat{p}_{0n} \rightarrow p_0$  and  $\tilde{\theta}_n/\theta \rightarrow 1$  as  $n \rightarrow \infty$ , respectively. Moreover,  $\lim_{n \rightarrow \infty} f(n) = \infty$  and  $\lim_{n \rightarrow \infty} f(n)/f(n-1) = 1$ .

Since the stopping rule  $N$  is well-defined and non-decreasing as a function of  $t$ , by invoking the results of Chow and Robbins (1965), the first order asymptotics for properties of the proposed sequential procedure are obtained as follows:

**Theorem 3.2** *We have*

- (i)  $\lim_{c \rightarrow 0} N = \infty$  a.s.,  $\lim_{c \rightarrow 0} E(N) = \infty$ ,
- (ii)  $\lim_{c \rightarrow 0} N/n^* = 1$  a.s.,
- (iii)  $\lim_{c \rightarrow 0} E(N)/n^* = 1$ .

## 4 Evaluation of the Procedure

The performance of the sequential procedure is usually evaluated by comparing two risks: one is  $R_N(c)$ , the risk involved in sequential estimation of  $\theta$  using the proposed procedure, and the other is  $R_{n^*}(c)$ , the risk associated with the optimal fixed-sample size  $n^*$ . As a measure of closeness, two comparisons are made by studying the ratio and the difference of the two risks defined by

- (1) risk-efficiency:  $e_R(N, n^*) = R_N(c)/R_{n^*}(c)$ ,
- (2) regret functions:  $r_R(N, n^*) = R_N(c) - R_{n^*}(c)$ .

However, in most cases there do not exist sequential procedures that are uniformly risk-efficient or which have uniformly minimum regret. Therefore, we consider the risk-efficiency and the regret as  $c$  tends to zero. To show that the sequential procedure is asymptotically risk-efficient, i.e., the ratio  $R_N(c)/R_{n^*}(c) \rightarrow 1$  or  $R_N(c) \sim R_{n^*}(c)$  as  $c \rightarrow 0$ , we establish the following theorem.

**Theorem 4.1** *Let  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  be two sequences of independent Bernoulli random variables with non-zero probabilities  $p_0$  and  $p_1$ , respectively. Define  $\theta = p_1/p_0$ . For  $c > 0$  and  $0 < \theta < \infty$ , define the stopping time  $N$  by Eq. (18). Then the sequential procedure  $(N, \hat{\theta}_N)$  is asymptotically risk-efficient, i.e.,*

$$\lim_{c \rightarrow 0} [e_R(N, n^*)] = \lim_{c \rightarrow 0} [R_N(c)/R_{n^*}(c)] = 1. \quad (21)$$

**Remark 4.1** Since the risks of the optimal fixed-sample size procedure and the sequential procedure both approach zero as  $c$  does, it follows immediately from Theorem 4.1 that

$$\lim_{c \rightarrow 0} [r_R(N, n^*)] = \lim_{c \rightarrow 0} [R_N(c) - R_{n^*}(c)] = 0. \quad (22)$$

**Remark 4.2** From a practical point of view, one can observe binomial variables based on a fixed number of trials, say  $k$ , instead of observing Bernoulli variables at each stage. In this case, the number of trials does not change from stage to stage. Then, the procedure as well as the entire asymptotic theory goes through provided  $Nk$  tends to be large, where  $N$  is now interpreted as the (random) number of stages.

## 5 Numerical Studies

### 5.1 Monte Carlo Simulation

Monte Carlo experimentation is performed to illustrate the behavior and performance of the stopping rule in the proposed sequential procedure as  $c \rightarrow 0$ . The results of the Monte Carlo simulation, based on the sequential rule in Eq. (18), are summarized in the following tables, which show several choices of the

parameter  $\theta$ , namely  $\theta = 2.0, 4.0$  and  $9.0$  with moderate values of  $(p_0, p_1)$ . For more realistic constellation we also add cases with small values of  $(p_0, p_1)$ . These are presented in Tables 5.4-5.6. In practice (e.g. vaccine trials), frequently comparing two groups with small rates have smaller differences. Since  $X$  and  $Y$  can be interchanged when  $\theta \leq 1$ , we have considered only situations in which  $\theta \geq 1$ .

For each selected value of  $c$ , every value in the table is based on 5,000 independent replications with the initial sample size  $m = 3$ . Each table contains the selected value of  $c$ , the estimate of  $\theta$ , the average of the stopping time,  $E(N)$ , the optimal stopping time  $n^*$ , the average risk associated with the stopping time  $N$ ,  $Risk(N)$ , the risk under the optimal fixed-sample size  $n^*$ ,  $Risk(n^*)$ , the risk-efficiency,  $e_R(N, n^*)$ , and the regret,  $r_R(N, n^*)$ .

Table 5.1: For  $\theta = 2.0$  (when  $p_0 = 0.3, p_1 = 0.6$ )

$c$	$\hat{\theta}$	$E(N)$	$n^*$	$Risk(N)$	$Risk(n^*)$	$e_R(N, n^*)$	$r_R(N, n^*)$
.10	1.9703	9.02	11	1.8743	2.1909	.8555	-.3166
.05	1.9962	12.69	16	1.5341	1.5500	.9897	-.0159
.01	1.9867	31.85	35	.7358	.6929	1.0620	.0430
.005	2.0061	46.87	49	.5068	.4899	1.0346	.0169
.002	1.9987	75.25	78	.3153	.3098	1.0177	.0055
.001	2.0006	107.48	110	.2191	.2191	.9999	.0000

Table 5.2: For  $\theta = 4.0$  (when  $p_0 = 0.2, p_1 = 0.8$ )

$c$	$\hat{\theta}$	$E(N)$	$n^*$	$Risk(N)$	$Risk(n^*)$	$e_R(N, n^*)$	$r_R(N, n^*)$
.10	3.9794	21.07	27	5.5905	5.2185	1.0713	.3720
.05	3.9514	30.48	37	4.3440	3.6878	1.1779	.6502
.01	4.0002	78.54	83	1.7395	1.6493	1.0547	.0902
.005	4.0110	113.52	117	1.2140	1.1662	1.0410	.0478
.002	4.0023	181.20	185	.7471	.7376	1.0129	.0095
.001	3.9940	256.62	261	.5258	.5215	1.0082	.0043

Table 5.3: For  $\theta = 9.0$  (when  $p_0 = 0.1, p_1 = 0.9$ )

$c$	$\hat{\theta}$	$E(N)$	$n^*$	$Risk(N)$	$Risk(n^*)$	$e_R(N, n^*)$	$r_R(N, n^*)$
.10	9.0268	76.28	86	20.0551	17.1814	1.1673	2.8737
.05	8.9908	111.93	122	14.1292	12.1492	1.1630	1.9800
.01	9.0217	266.33	272	5.4885	5.4332	1.0102	.0553
.005	9.0297	380.00	385	3.8465	3.8419	1.0012	.0046
.002	8.9903	600.20	608	2.4777	2.4298	1.0197	.0479
.001	9.0045	853.58	860	1.7587	1.7181	1.0236	.0406

Table 5.4: For  $\theta = 2.0$  (when  $p_0 = 0.05, p_1 = 0.1$ )

$c$	$\tilde{\theta}$	$E(N)$	$n^*$	$Risk(N)$	$Risk(n^*)$	$e_R(N, n^*)$	$r_R(N, n^*)$
.10	2.0397	30.88	34	4.7836	6.6941	.7146	-1.9106
.05	1.9727	39.80	48	3.4848	4.7333	.7362	-1.2486
.01	2.0187	91.11	106	2.0638	2.1166	.9751	-.0528
.005	2.0140	132.20	150	1.5802	1.4967	1.0558	.0835
.002	2.0070	217.15	237	1.0489	.9466	1.1081	.1023
.001	2.0048	316.47	335	.7339	.6693	1.0964	.0645

Table 5.5: For  $\theta = 2.0$  (when  $p_0 = 0.025, p_1 = 0.05$ )

$c$	$\tilde{\theta}$	$E(N)$	$n^*$	$Risk(N)$	$Risk(n^*)$	$e_R(N, n^*)$	$r_R(N, n^*)$
.10	2.0116	51.96	49	6.9244	9.6333	.7187	-2.7103
.05	1.9932	61.94	69	4.7366	6.8188	.6953	-2.0757
.01	2.0123	130.22	153	2.6536	3.0463	.8711	-.3928
.005	2.0138	184.32	216	2.0733	2.1541	.9625	-.0807
.002	2.0070	302.15	341	1.4600	1.3624	1.0717	.0977
.001	2.0023	439.58	482	1.0552	.9633	1.0953	.0918

Table 5.6: For  $\theta = 4.0$  (when  $p_0 = 0.025, p_1 = 0.1$ )

$c$	$\tilde{\theta}$	$E(N)$	$n^*$	$Risk(N)$	$Risk(n^*)$	$e_R(N, n^*)$	$r_R(N, n^*)$
.10	3.9752	68.13	88	13.5284	17.5273	.7718	-3.9989
.05	3.9834	95.36	124	10.9148	12.3935	.8807	-1.4787
.01	3.9877	237.12	278	6.0680	5.5426	1.0948	.5254
.005	4.0077	353.60	392	4.4216	3.9192	1.1282	.5024
.002	4.0155	589.66	620	2.6662	2.4787	1.0756	.1874
.001	3.9800	840.99	877	1.8294	1.7527	1.0437	.0767

From Tables 5.1-5.6, we see that the estimate  $\tilde{\theta}_n$  converges to the corresponding true ratio  $\theta$  as  $c \rightarrow 0$ , and the expected stopping time  $E(N)$  is uniformly smaller than the optimal stopping time  $n^*$ . That is, the suggested procedure requires smaller sample sizes than the fixed-sample procedure. We also observe that as the sampling cost per observation  $c$  becomes smaller, the average random stopping time,  $E(N)$  increases. However, the average risks under both stopping times  $N$  and  $n^*$  decrease as  $c \rightarrow 0$ . Moreover, it appears to be true that as the ratio  $\theta$  increases (by  $p_0$  getting smaller and/or  $p_1$  getting bigger), the expected stopping time  $E(N)$  tends to grow rapidly.

As mentioned earlier, the assumed loss function incorporates both losses from estimation and costs from sampling. In this context, the value of  $c$  plays a role as a sample-inflation factor in the sequential procedure. Analytically, the risk-efficiency approaches one and the regret goes to zero as  $c \rightarrow 0$ . The simulation results provide substantial numerical evidence to conclude that the proposed sequential estimator  $\tilde{\theta}_N$  performs satisfactorily.

**Remark 5.1** Note that some of values of the regret  $r_R(N, n^*)$  in the above tables are shown to be negative; i.e., the risk  $R_N(c)$  associated with the proposed sequential procedure is smaller than the risk  $R_{n^*}(c)$  associated with the

optimal fixed-sample size  $n^*$  in estimating the risk-efficient estimator  $\tilde{\theta}$ . This implies that the proposed sequential procedure outperforms the fixed-sample size procedure, which is considered hypothetically since the parameters are unknown (see Section 3.1).

## 5.2 An Example

A data set consisting of binomial variables based on a fixed number of trials was found instead of a large number of sequential Bernoulli trials. The proposed method is applied to the following data set.

**Example 5.1** Montgomery (1985, pp. 123-127) provided data from a packing process of frozen orange juice concentrate packed in 6-oz cardboard cans. We monitor the fraction of nonconforming cans, where each stage of sampling involves taking 50 observations. For the first three-shift period, 30 stages of samples were collected. For the next three shifts after the adjustment of the process, 24 stages of samples were also collected. Since two sample fractions in stages 15 and 23 in the first period were identified to be anomalous observations, we discarded two pairs of observations from both periods. Let  $X$  be the number of nonconformances in each stage during the first period, and let  $Y$  represent the number of nonconformances during the second period. Then, the 22 pairs of observations  $(x, y)$  are:

Sample No.	1	2	3	4	5	6	7	8	9	10	11
$x$	12	15	8	10	4	7	16	9	14	10	5
$y$	9	6	12	5	6	4	5	3	7	6	2
Sample No.	12	13	14	15	16	17	18	19	20	21	22
$x$	6	17	12	8	10	5	13	11	20	18	15
$y$	4	3	6	4	8	5	6	7	5	6	4

Let  $p_0$  denotes the proportion of nonconforming cans during the first period and let  $p_1$  be the proportion of nonconformances during the second period. Define  $\theta = p_1/p_0$  to be the ratio of two nonconforming proportions  $p_0$  and  $p_1$ . In order to apply our procedure, it is assumed that the sampling cost  $c$  per stage is constant. Under squared error loss, we wish to infer risk-efficient estimates of the ratio  $\theta$  when the sampling cost per stage is  $c$ . The estimate  $\tilde{\theta}_n = \hat{p}_{1n}/\hat{p}_{0n}$  is obtained from sequential observations, where  $n$  denotes the stopping stage. Then, the total sample size after stopping in each period is  $n \times 50$  cans. The initial sample stage  $m = 1$  is taken for all procedures.

1. Firstly, we want to estimate the true ratio  $\theta = p_1/p_0$  under squared error loss when the proportional sampling cost per stage is  $c = 0.01$ . (In fact, in this example, the sampling cost can be regarded as the manufacturing cost.) The proposed sequential procedure stops at sample stage  $n = 3$ ; the estimates are  $\hat{p}_0 = 0.180$ ,  $\hat{p}_1 = 0.233$  and the estimated ratio between the two periods is  $\hat{\theta} = 1.2963$  with the associated risk 0.1178 based on the sample size of  $n = 150$  cardboard cans used for each period.

2. Next, estimate the ratio  $\theta$  of two nonconforming proportions  $p_0$  and  $p_1$  under squared error loss when the sampling unit cost per each stage  $c$  is lowered by a factor of 10, i.e.,  $c = 0.001$ . The suggested procedure terminates at stage  $n = 9$ , which yields the estimates  $\hat{p}_0 = 0.126$ ,  $\hat{p}_1 = 0.210$  and estimated ratio  $\tilde{\theta} = 1.6667$  under the squared error loss function based on 450 cans. The associated risk of the estimated ratio is now 0.0746, which achieves 37% reduction in risk from the previous case (when  $c = 0.01$ ).

Table 5.7 summarizes the results. For each selected value of  $c$ , it shows the stopping stage  $n$ , the sample size  $= n \times 50$ , the estimated ratio  $\tilde{\theta}_n$ , the variance of the estimated ratio,  $Var(\tilde{\theta}_n)$ , and the corresponding risk,  $R_n(c) = Var(\tilde{\theta}_n) + cn$ .

Table 5.7: Risk-Efficient Estimates of  $\theta$  under Squared Error Loss for Orange Juice Packed Data

$c$	Stopping stage	Sample size	$\hat{p}_0$	$\hat{p}_1$	$\tilde{\theta}_n$	$Var(\tilde{\theta}_n)$	$R_n(c)$
.01	3	150	.180	.233	1.2963	.0878	.1178
.008	4	200	.160	.225	1.4063	.0860	.1180
.004	5	250	.152	.196	1.2895	.0644	.0844
.002	6	300	.140	.187	1.3333	.0622	.0742
.0015	7	350	.134	.206	1.5319	.0691	.0796
.0010	9	450	.126	.210	1.6667	.0656	.0746

## 6 Appendix: Proofs of the Theorems

### A.1. Proof of Theorem 3.1

Using the stopping rule in Eq. (18),

$$\begin{aligned}
P\{N = \infty\} &= \lim_{n \rightarrow \infty} P\{N > n\} \\
&\leq \lim_{n \rightarrow \infty} P\left\{n \leq (c^{-1}\hat{\sigma}_n^2)^{1/2}\right\} \\
&= \lim_{n \rightarrow \infty} P\left\{n \leq \left[\tilde{\theta}_n \left(1 + \tilde{\theta}_n - 2\hat{\theta}_n\hat{p}_{0n}\right) / (c\hat{p}_{0n})\right]^{1/2}\right\} \\
&= 0,
\end{aligned} \tag{23}$$

since  $\hat{p}_{0n} \rightarrow p_0$  and  $\hat{p}_{1n} \rightarrow p_1$  almost surely as  $n \rightarrow \infty$ . Therefore,  $\tilde{\theta}_n \rightarrow \theta$ , since  $\tilde{\theta}_n$  is a function of  $\hat{p}_{0n}$  and  $\hat{p}_{1n}$ . Hence, the finite sure termination of the sequential procedure is established.

### A.2. Proof of Theorem 3.2

(i) is easily verified using Eq. (20), and the monotone convergence theorem gives the result.

For (ii), from Eq. (23), since  $N - 1 \leq c^{-1/2} \sigma_{N-1}$ , we have

$$\frac{c^{-1} \sigma_N^2}{c^{-1} \sigma^2} \leq \frac{N}{n^*} \leq \frac{1 + c^{-1} \sigma_{N-1}^2}{c^{-1} \sigma^2}, \quad (24)$$

from which it is easy to see that

$$\sigma_N^2 / \sigma^2 \leq N / n^* \leq c / \sigma^2 + (\sigma_{N-1}^2 / \sigma^2). \quad (25)$$

Taking limits of the inequalities in Eq. (25) as  $c$  goes to zero, we have

$$\lim_{c \rightarrow 0} (\sigma_N^2 / \sigma^2) \leq \lim_{c \rightarrow 0} (N / n^*) \leq \lim_{c \rightarrow 0} (\sigma_{N-1}^2 / \sigma^2).$$

However, the quantities on the extremes of the inequality tends to unity. Hence  $\lim_{c \rightarrow 0} (N / n^*) = 1$ .

(iii) follows from the result of Theorem 2.3 in Cho and Govindarajulu (2006).

### A.3. Proof of Theorem 4.1

Obviously, as  $c \rightarrow 0$ ,  $N \rightarrow \infty$  a.s. To establish the risk efficiency, we must show that as  $c$  tends to zero,

$$\frac{E(\tilde{\theta}_N - \theta)^2 + cE(N)}{E(\tilde{\theta}_{n^*} - \theta)^2 + cn^*} \rightarrow 1. \quad (26)$$

Taking the limit on left-hand side (LHS) of Eq. (26),

$$\begin{aligned} \lim_{c \rightarrow 0} \left[ \frac{E(\tilde{\theta}_N - \theta)^2 + cE(N)}{E(\tilde{\theta}_{n^*} - \theta)^2 + cn^*} \right] &= \lim_{c \rightarrow 0} \left[ \frac{E(\tilde{\theta}_N^2) - 2\theta E(\tilde{\theta}_N) + \theta^2 + cE(N)}{E(\tilde{\theta}_{n^*}^2) - 2\theta E(\tilde{\theta}_{n^*}) + \theta^2 + cn^*} \right] \\ &= \lim_{c \rightarrow 0} \left[ \frac{E(\tilde{\theta}_N^2 / \theta^2) - 2E(\tilde{\theta}_N / \theta) + 1 + cE(N / \theta^2)}{E(\tilde{\theta}_{n^*}^2 / \theta^2) - 2E(\tilde{\theta}_{n^*} / \theta) + 1 + c(n^* / \theta^2)} \right]. \end{aligned}$$

Since  $\tilde{\theta}_N / \theta \rightarrow 1$  a.s., and  $\tilde{\theta}_{n^*} \rightarrow \theta$  a.s. as  $c \rightarrow 0$ , so  $E(\tilde{\theta}_N / \theta) \rightarrow 1$  and  $E(\tilde{\theta}_N^2 / \theta^2) \rightarrow 1$ , by the proof was referenced, not shown, Theorem 3.2 (iii).

Noting that  $\theta \geq 1$  and using Eq. (16), it follows that

$$cn^* / \theta^2 = \sigma c^{1/2} / \theta^2 \leq \sigma c^{1/2}, \quad (27)$$

where  $\sigma = [\theta(1 + \theta - 2\theta p_0) / p_0]^{1/2}$ , while the RHS of Eq. (27) tends to zero as  $c \rightarrow 0$ . Further, we can write

$$cE(N) / \theta^2 = (cn^* / \theta^2) E(N / n^*).$$

The proof is complete upon noting that  $E(N / n^*) \rightarrow 1$  as  $c \rightarrow 0$ .

**Acknowledgement.** The author wishes to thank the Editor and anonymous referees for their careful review of this manuscript. Especially, two referees provided very keen, insightful comments and valuable suggestions which have resulted in a more accurate and readable paper.

## References

- [1] Aitchison, J. and Bacon-Shone, J. (1981). Bayesian risk ratio analysis. *The American Statistician* **35**, 254-257.
- [2] Bailey, B. J. R. (1987). Confidence limits to the risk ratio. *Biometrics* **43**, 201-205.
- [3] Bedrick, E. J. (1987). A family of confidence intervals for the ratio of two binomial proportions. *Biometrics* **43**, 993-998.
- [4] Cho, H. and Govindarajulu, Z. (2006). Sequential confidence limits for the ratio of two binomial proportions. (Submitted for publication.)
- [5] Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics* **36**, 457-462.
- [6] Fleiss, J. (1981). *Statistical Methods for Rates and Proportions*, Wiley, New York.
- [7] Gart, J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of  $2 \times 2$  tables. *Biometrika* **72**, 673-677.
- [8] Gart, J. and Nam, J. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **44**, 323-338.
- [9] Ghosh, M., Mukhopadhyay, N. and Sen, P. (1997). *Sequential Estimation*, Wiley, New York.
- [10] Katz, D., Baptista, J., Azen, S., and Pike, M. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* **34**, 469-474.
- [11] Koopman, P. (1984). Confidence limits for the ratio of two binomial proportions. *Biometrics* **40**, 513-517.
- [12] Walter, S. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics* **32**, 829-849.