Sample Size Determination for Estimating Ratio, Subject to Misclassification

A. Zeinal: Guilan University
S.M. Taheri: Isfahan University of Technology

Abstract

We investigate the problem of estimating the ratio of subjects that share a particular characteristic in a population, but some degree of misclassification is possible. The misclassification probabilities are considered as the random variables, and we study a Bayesian approach to this problem. We propose a method to calculate the exact posterior density of the ratio, based on some prior distributions. We, then, study a method to determine the sample size, using average coverage criterion. We also investigate the effect of different prior distributions on the sample size.

1. Introduction and Preliminaries

Suppose we investigate the existence of a particular characteristic in a population to estimate the ratio of subjects that share it. If the distinguishing method is error free, then the well-known sample size formula, based on the normal approximation to the binomial distribution, can be used. This gives

$$n = \left(\frac{2Z_{\alpha/2}}{w} \right)^2 \theta(1-\theta)$$

where w is the confidence interval width.

Now, suppose that for classifying the subjects, we have some misclassification errors. We may distinguish a subject having the characteristic as a subject that dosen't have that characteristic and vice versa. This problem is common in some researches, for example in the estimate of the prevalence of a disease based on some medical tests.

Keywords: Average Coverage Criterion; Bayes Estimator; Binomial Distribution; Misclassification; Sample Size

It should be mentioned that, in each method of classification, two principal criteria, sensitivity and specificity, are important. Sensitivity is the probability of distinguishing a subject as a positive subject truly. Specificity is the probability of distinguishing a subject as a negative subject truly.

Let θ be the actual ratio of the positive subjects and p be the ratio of the positive subjects distinguished by the method. If sensitivity and specificity are respectively shown as s and c, we have

$$p = \theta\ s + (1-\theta)\ (1-c)\ . \tag{1}$$

Supposing s and c as fixed constants, Rahme and Joseph [5] conveyed the following formula for the adjusted sample size

$$n_{adj} = \left(\frac{2Z_{\alpha/2}}{w(s+c-1)} \right)^2 p(1-p)\ , \tag{2}$$

where p is calculated in equation (1) based on a value of θ. In practice, of course, θ is unknown and therefore the researcher must estimate the value of p based on a primary sampling or some other information, and then calculate the necessary sample size. Equation (2) also demonstrates that both of the sensitivity and specificity have a very large influence on sample size. As expected, when $s = c = 1$, the method is error free, $p = \theta$ and equation (2) reduces to the standard binomial sample size formula.

The above problem is specially challenging when the degree to which misclassification occurs is not exactly known. So, the problem of determining the sample size for estimating the ratio has also been considered from the Bayesian point of view. This approach, first, has been studied without considering misclassification in a series of researches such as those of Adcock [1,2] and Gould [3]; and has recently been studied by Rahme et al. [6] subject to misclassification.

In Bayesian approach, the posterior density of θ and then the sample size are calculated based on some prior distributions of θ, s, and c. In this respect, Joseph et al. [4] used the Gibbs sampling to estimate the posterior density of

θ. Recently, Rahme et al. [6] used the Monte Carlo approximation to obtain the posterior density of θ.

In this paper, we use the Bayesian approach to determine the sample size for estimating the ratio, subject to misclassification when the sensitivity and specificity are unknown. In this respect, following the investigation of Rahme et al. [6], we calculate the exact posterior density of θ in the second section of this paper. We also show that, using the Beta densities as the priors, the posterior density of θ will be a convex linear combination of Beta probability density functions.

In Section 3 of this paper, using average coverage criterion and symmetric intervals around the posterior mean, we will propose a formula for determining the sample size, based on the given posterior density in Section 2. Then we will compare the results with those of Rahme et al.'s work [6].

In Section 4, the influence of the different prior distributions on the sample size will be examined, numerically.

2. Bayes Estimator for Ratio when Sensitivity and Specificity are Unknown

In this section, we apply a Bayesian approach to estimate ratio, θ, when sensitivity and specificity are unknown. Considering the prior information about sensitivity, specificity, and θ, we first calculate the posterior density of θ and then we obtain a Bayes estimator of θ under squared error loss function.

Let $f(x,\theta)$ be the joint density of X and θ, and $g(x)$ be the marginal probability density function of X, where X is the number of subjects have been diagnosed as virus defected in a sample of size n of a population. Then the posterior function of θ will be

$$f(\theta|x)=\frac{f(x,\theta)}{g(x)} \tag{3}$$

where $$f(x,\theta)=\int_0^1\int_0^1 l(x|\theta,s,c)\, f(\theta,s,c)\, dsdc\ ,$$

$$g(x)=\int_0^1 f(x,\theta)\,d\theta$$

and $f(\theta,c,s)$ is the joint prior density function of $s,c,$ and θ. In addition, the likelihood function $l(x|\theta,s,c)$ is as follows

$$l(x|\theta,s,c)=\binom{n}{x}\{\theta\ s+(1-\theta)(1-c)\}^x\{\theta(1-s)+(1-\theta)\ c\}^{n-x}.$$

Suppose that $\theta,s,$ and c are independent Beta random variables (Remarks 1 and 2 below) with parameters as $(\alpha_\theta,\beta_\theta)$, (α_s,β_s) *and* (α_c,β_c), respectively. Then

$$f(x,\theta)=A\int_0^1\int_0^1\{\theta s+(1-\theta)(1-c)\}^x\{\theta(1-s)+(1-\theta)c\}^{n-x}$$
$$\theta^{\alpha_\theta-1}(1-\theta)^{\beta_\theta-1}s^{\alpha_s-1}(1-s)^{\beta_s-1}c^{\alpha_c-1}(1-c)^{\beta c-1}\,dsdc \qquad (4)$$

where

$$A=\frac{\binom{n}{x}}{B(\alpha_\theta,\beta_\theta)B(\alpha_s,\beta_s)B(\alpha_c,\beta_c)}.$$

Rahme et al. [6] has estimated the expression (4) using the Monte Carlo approximation, whereas we propose an exact rule for the posterior density using binomial expansion.

Theorem 1. Considering the above assumptions, the posterior density of θ is

$$f(\theta|x)=\frac{\sum_{k=0}^{x}\sum_{l=0}^{n-x}\binom{x}{k}\binom{n-x}{l}B1B2\theta^{\alpha_\theta+l+k-1}(1-\theta)^{\beta_\theta+n-l-k-1}}{\sum_{k=0}^{x}\sum_{l=0}^{n-x}\binom{x}{k}\binom{n-x}{l}B1B2B3} \qquad (5)$$

where

$$\begin{aligned}B1&=B(\alpha_s+k,\beta_s+l),\\ B2&=B(n-x-l+\alpha_c,x-k+\beta_c),\\ B3&=B(\alpha_\theta+l+k,\beta_\theta+n-l-k).\end{aligned} \qquad (6)$$

Proof. Using binomial expansion of $\{\theta\ s+(1-\theta)(1-c)\}^x$ and $\{\theta(1-s)+(1-\theta)c\}^{n-x}$, in (4), the following formula is induced

$$f(x,\theta)=A\theta^{\alpha_\theta-1}(1-\theta)^{\beta_\theta-1}$$

$$\int_0^1\int_0^1\left[\sum_{k=0}^{x}\binom{x}{k}(\theta s)^k((1-\theta)(1-c))^{x-k}\right]$$

$$\left[\sum_{l=0}^{n-x}\binom{n-x}{l}(\theta(1-s))^l((1-\theta)c)^{n-x-l}\right]c^{\alpha_c-1}(1-c)^{\beta_c-1}s^{\alpha_s-1}(1-s)^{\beta_s-1}dsdc$$

$$=A\theta^{\alpha_\theta-1}(1-\theta)^{\beta_\theta+n-1}$$

$$\int_0^1\int_0^1\left[\sum_{k=0}^{x}\sum_{l=0}^{n-x}\binom{x}{k}\binom{n-x}{l}\theta^{l+k}(1-\theta)^{-l-k}s^{\alpha_s+k-1}(1-s)^{\beta_s+l-1}c^{n-x-l+\alpha_c-1}(1-c)^{x-k+\beta_c-1}\right]dsdc$$

$$=A\theta^{\alpha_\theta-1}(1-\theta)^{\beta_\theta+n-1}$$

$$\left[\sum_{k=0}^{x}\sum_{l=0}^{n-x}\binom{x}{k}\binom{n-x}{l}\theta^{l+k}(1-\theta)^{-l-k}B(\alpha_s+k,\beta_s+l)B(n-x-l+\alpha_c,x-k+\beta_c)\right]$$

$$=\sum_{k=0}^{x}\sum_{l=0}^{n-x}Z_{x,n,k,l}\theta^{\alpha_\theta+l+k-1}(1-\theta)^{\beta_\theta+n-l-k-1} \tag{7}$$

where

$$Z_{x,n,k,l}=\frac{\binom{x}{k}\binom{n-x}{l}\binom{n}{x}B(\alpha_s+k,\beta_s+l)B(n-x-l+\alpha_c,x-k+\beta_c)}{B(\alpha_\theta,\beta_\theta)B(\alpha_s,\beta_s)B(\alpha_c,\beta_c)}.$$

Therefore the marginal probability function of $g(x)$ is resulted as follows

$$g(x)=\sum_{k=0}^{x}\sum_{l=0}^{n-x}Z_{x,n,k,l}B(\alpha_\theta+l+k,\beta_\theta+n-l-k). \tag{8}$$

By substituting the expressions (7) and (8) in formula (3), we will gain (5).

Remark 1. In this work, as other similar works, prior information in the form of a Beta density will be assumed. This family of distributions was selected since its region of positive density from 0 to 1, matches the range of all parameters of interest, and because it is a flexible family, in that a wide variety of density shapes can be derived by selecting different choices of α and β. It also has the advantage of being the conjugate prior distribution for the binomial likelihood, a property that simplifies the derivation of the posterior distributions.

The particular Beta prior density for each test parameter could be selected by matching the center of the range with the mean of the Beta distribution, given by $\alpha/(\alpha+\beta)$; and matching the standard deviation of the Beta distribution, given by $\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$ with one-quarter of the total range [4].

Remark 2. It will often be reasonable that $\theta, s,$ and c are a priori independent, given that the test methodology (e.g. the cut-off values for continuous tests) remains fixed. This is because the performance of the test within positive and negative subgroups of patients may not be affected by the prevalence of the disease in the population, and prior knowledge about the sensitivity and specificity given any fixed cut-off usually is gained by independently applying the test to known positive and negative subjects (see also[6]).

Proposition 1. Regarding the assumptions in Theorem 1, $f(\theta|x)$, the posterior density function of θ, is a convex linear combination of Beta probability density functions.

Proof. Let

$$R_{x,n,i} = \frac{\sum_{k=0}^{x}\sum_{l=0}^{n-x}{}_{k+l=i}\binom{x}{k}\binom{n-x}{l}B1B2B3}{\sum_{k=0}^{x}\sum_{l=0}^{n-x}\binom{x}{k}\binom{n-x}{l}B1B2B3},$$

$$T_i(\theta) = \frac{\theta^{\alpha_\theta+i-1}(1-\theta)^{\beta_\theta+n-i-1}}{B(\alpha_\theta+i, \beta_\theta+n-i)}.$$

Then

$$f(\theta|x) = \sum_{i=0}^{n} R_{x,n,i}T_i(\theta),$$

that is the posterior density function of θ is a convex linear combination of Beta probability density functions.

Proposition 2. The Bayes estimator of θ under squared error loss function is

$$\hat{\theta}(x,n)=\frac{\sum_{k=0}^{x}\sum_{l=0}^{n-x}\binom{x}{k}\binom{n-x}{l}B1B2B4}{\sum_{k=0}^{x}\sum_{l=0}^{n-x}\binom{x}{k}\binom{n-x}{l}B1B2B3} \tag{9}$$

where $$B4=B(\alpha_\theta+l+k+1,\beta_\theta+n-l-k). \tag{10}$$

Proof. It is known that, Bayes estimator under squared error loss function is equal to $E(\theta|x)$, i.e., the expectation of the posterior density. So, the relation (9) is calculated easily.

Note. Considering $\alpha_\theta,\beta_\theta$ as positive integers, the posterior density function of θ is a polynomial of $\alpha_\theta+\beta_\theta+n-2$ order.

Example1. According to some available information about θ:"the prevalence of a particular virus in a population", the prior distribution of $B(1,3)$ has been considered for θ. Based on the previous experiences on accuracy and inaccuracy of the test results, for a virus diagnostic test, the prior distributions of $B(60,0.1)$ and $B(30,0.1)$ have, respectively, been considered for s and c. Suppose that, in a random sample with the size of $n=6$ of the above population, two subjects have been diagnosed as virus defected. In this case, using the expressions (6) and (10), we have

$$\begin{aligned}B1&=B(60+k,0.1+l)\\B2&=B(34-l,2.1-k)\\B3&=B(1+l+k,9-l-k)\\B4&=B(2+l+k,9-l-k).\end{aligned}$$

Therefore, the posterior probability density function of θ is calculated as follows

$$\begin{aligned}f(\theta|x)&=0.00001(1-\theta)^2\theta^6+0.00091(1-\theta)^3\theta^5+0.04161(1-\theta)^4\theta^4\\&+1.58637(1-\theta)^5\theta^3+245.50751(1-\theta)^6\theta^2+1.44239(1-\theta)^7\theta+0.02264(1-\theta)^8\\&=242.54\theta^8-1455.36\theta^7+3637.34\theta^6-4845.24\theta^5\\&+3625.82\theta^4-1442.44\theta^3+236.04\theta^2+1.26\theta+0.23\end{aligned}$$

Furthermore, on the basis of (9), the Bayes estimator of θ under squared error loss function becomes

$$\hat{\theta} = \frac{\sum_{k=0}^{2}\sum_{l=0}^{4}\binom{2}{k}\binom{4}{l}B1B2B4}{\sum_{k=0}^{2}\sum_{l=0}^{4}\binom{2}{k}\binom{4}{l}B1B2B3} = 0.29782.$$

3. Sample Size Determination

Suppose that in a sample of size n of a population, the number of sample subjects distinguished as positive subjects is x. Then

$$\left[\hat{\theta}(x,n)-\frac{w}{2},\hat{\theta}(x,n)+\frac{w}{2}\right]$$

is a confidence interval with the width of w for θ. The coverage probability of this interval depends on x and n. This gives

$$\text{coverage}(x,n)=\int_{\hat{\theta}(x,n)-\frac{w}{2}}^{\hat{\theta}(x,n)+\frac{w}{2}} f(\theta|x)\,d\theta\,.$$

Although x is unknown, its probability function, i.e., $g(x)$ is available. Therefore, the expectation of confidence interval (in other words, average coverage) is initially known and calculated as follows

$$\sum_{x=0}^{n} \text{coverage}(x,n)\; g(x).$$

Thus, to have a confidence interval with the minimum average coverage of $1-\alpha$, the sample size n must be chosen in such a way that

$$\sum_{x=0}^{n}\int_{\hat{\theta}(x,n)-\frac{w}{2}}^{\hat{\theta}(x,n)+\frac{w}{2}} f(x,\theta)\,d\theta \geq 1-\alpha\,. \qquad (11)$$

The substitution of expression (7) in (11), results in the determination of the smallest value of n so that

$$\sum_{x=0}^{n}\sum_{k=0}^{x}\sum_{l=0}^{n-x} Z_{x,n,k,l}\int_{\hat{\theta}(x,n)-\frac{w}{2}}^{\hat{\theta}(x,n)+\frac{w}{2}} \theta^{\alpha_\theta+l+k-1}(1-\theta)^{\beta_\theta+n-l-k-1}\,d\theta \geq 1-\alpha\,.$$

Example 2. Rahme et al. [6] have considered the following parameters for the prior distributions of $\theta, s,$ and c in a numerical example

$$(\alpha_\theta, \beta_\theta) = (6,14)$$

$$(\alpha_c, \beta_c) = (44.1,0.1)$$

$$(\alpha_s, \beta_s) = (130.1,6.1).$$

Using Monte Carlo approximation, they obtained 348 for the minimum value of $n,$ so that the average coverage confidence interval of width $w = 0.1$ is at least 0.95.

But, using the above prior densities, we have calculated 0.95072, 0.95046, 0.95020, 0.94994 respectively for 348, 347, 346, and 345 as values of $n.$ As a result, the minimum value of n to get the minimum average coverage of 0.95 is equal to 346.

4. The Influence of Priors on the Sample Size

For studying the influence of prior distributions on the sample size, we considered different Beta distributions as priors for θ, s, c and determined the related sample size when $w = 0.1.$

As a result, we obtained the following tables which indicate when the prior distributions of $\theta, s,$ and c change, the sample size essentially changes.

Note that, in terms of sensitivity and specificity, the cases 1, 2, and 4 are similar, but $n_2 < n_4 < n_1$. This is not surprising, because $V_2(\theta) < V_4(\theta) < V_1(\theta)$. In other words, in case 2 we have a more precise prior than in case 4, and so in this case we need fewer samples than in case 4. The same argument is valid in comparing case 4 with case 1.

On the other hand, case 2 and case 3 have the same prior distribution of θ, but $n_2 < n_3$. Since in case 2 we face with a more exact test (in terms of sensitivity and specificity) than in case 3. Note that $E_2(s) > E_3(s)$, $E_2(c) > E_3(c)$, and $V_2(s) < V_3(s)$, $V_2(c) < V_3(c)$.

Table 1(a)
Comparison between sample sizes, based on some different priors

case	s	c	θ	n
1	$B(130.1,6.1)$	$B(44.1,0.1)$	$B(6,14)$	**346**
2	$B(130.1,6.1)$	$B(44.1,0.1)$	$B(1,19)$	**71**
3	$B(65.1,6.1)$	$B(22.1,0.1)$	$B(1,19)$	**101**
4	$B(130.1,6.1)$	$B(44.1,0.1)$	$B(1,9)$	**160**

Table 1(b)
Comparison between sample sizes, in terms of means and variances of priors

case	E(s)	V(s)	E(c)	V(c)	V(θ)
1	0.9552	3.1×10^{-4}	0.9977	4.99×10^{-5}	0.01
2	0.9552	3.1×10^{-4}	0.9977	4.99×10^{-5}	2.26×10^{-3}
3	0.9143	1.1×10^{-3}	0.9950	1.93×10^{-4}	2.26×10^{-3}
4	0.9552	3.1×10^{-4}	0.9977	4.99×10^{-5}	8.18×10^{-3}

Acknowledgement

The authors would like to thank the referees for careful reading of the article and giving valuable comments. The second author is grateful to Isfahan University of Technology Research for the support of this work.

References

1. C.A. Adcock, Bayesian approach to calculating sample sizes for multinomial sampling, Statistician 36 (1988) 155-159.
2. C.A. Adcock, Sample size determination: a review, Statistician 46 (1997) 261-283.
3. A.L. Gould, Sample size for event rate equivalence trials using prior information, Statist. Med. 12 (1993) 2009-2023.
4. L. Joseph, T.W. Gyorkos, and L. Coupal, Bayesian estimation of disease prevalence and parameters of diagnostic tests in the absence of a gold standard, Amer. J. Epidem. 141 (1995) 263-272.
5. E. Rahme, and L. Joseph, Estimating the prevalence of a rare disease: adjusted maximum likelihood, Statistician 47 (1998) 149-158.
6. E. Rahme, L. Joseph, and T.W. Gyorkos, Bayesian sample size determination for estimation binomial parameters from data subject to misclassification, Appl. Statis. 49 (2000) 119-128.