The Statistician (1998)
47, *Part* 1, *pp.* 149–158

Estimating the prevalence of a rare disease: adjusted maximum likelihood

Elham Rahme
McGill University, Quebec, Canada

and Lawrence Joseph†
Montreal General Hospital and McGill University, Montreal, Canada

[Received January 1997. Revised May 1997]

Summary. A common problem in medical research is to estimate the prevalence of a disease, i.e. to determine the proportion of individuals with the disease in a given population at a particular point in time. This can be accomplished by applying a diagnostic test to a sample of subjects from the target population. When an error-free test is not available, one must take into account the potential for misclassification errors to avoid misleading results. In this paper we suggest a new adjustment to the standard maximum likelihood estimator (MLE) of the prevalence, useful in the common situation when the MLE equals 0. Results of simulations are presented to compare the new estimator with the standard MLE. We also comment on confidence intervals and sample size determination for these situations.

Keywords: Diagnostic test; Maximum likelihood estimation; Prevalence; Rare disease; Sample size; Sensitivity; Specificity

1. Introduction

In public health, it is often important to estimate the proportion of individuals with a given disease in a given population at a particular point in time, known as the disease prevalence. One way to estimate disease prevalence is to obtain a random sample from the target population, and to test each individual in the sample for the disease. If the test used is error free, often referred to as a gold standard test, then the number of diseased individuals in the sample is the same as the number of positive test results, and estimating the prevalence is the classical problem of estimating a binomial proportion. Gold standard tests rarely if ever exist, however, since even a theoretically perfect test can be rendered less perfect by human, laboratory or other errors. Even when they exist, gold standard tests may be difficult to perform, highly invasive, very costly or time consuming, so that alternative tests are often considered. In developing alternative tests, their performance must be evaluated. In particular, the sensitivity of a test is the probability that a truly diseased individual will correctly register a positive test, whereas the specificity of a test is the probability of a negative test in a truly disease-free individual. When the sensitivity and the specificity of a diagnostic test are known, many researchers including Rogan and Gladen (1987) and Taragin *et al.* (1993) have proposed the use of a maximum likelihood estimator (MLE) to estimate the prevalence. See Walter and Irwig (1988) for a comprehensive review of methods related to this problem.

†*Address for correspondence*: Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada.
E-mail: joseph@binky.ri.mgh.mcgill.ca

0039–0526/98/47149

The MLE performs well under most circumstances. When the prevalence of the disease is low, however, as for many diseases, the MLE is quite often 0, even when the unobserved number of truly diseased subjects in the sample may not be 0. For example, consider the case where 16 positive results are observed in 100 tests. With a perfect test, the obvious point estimate of the prevalence is 16%. However, if the specificity of the test is 80%, then at least 20 positive tests would be expected, even if the prevalence is 0. To correct for this, Lew and Levy (1989) considered a Bayesian approach. They proposed the use of the posterior mean from a uniform prior distribution as an estimator of disease prevalence. The choice of a non-informative prior distribution, however, can have a substantial effect on the point estimate of the prevalence when the disease is rare. In particular, point estimates arising from a uniform prior density may differ from the point estimate suggested by other reasonable 'non-informative' choices, such as the standard Jeffreys prior density (Gelman *et al.*, 1995). In addition, calculating the posterior mean involves numerical integration and can therefore be difficult to calculate quickly. Here we shall present a simple adjustment to the MLE that is useful for rare diseases. We also provide formulae to calculate confidence intervals, and we discuss the sample sizes required for these confidence intervals to be smaller than a given width.

2. Maximum likelihood estimation

Suppose that the sensitivity and the specificity of a diagnostic test are known and equal to $s < 1$ and $c < 1$ respectively. Since the accuracy of diagnostic tests that have the sum of their sensitivity and specificity below 1 can be improved by reversing what is considered to be a positive test, without loss of generality we shall assume that $s + c > 1$. Consider a random sample of size n from the population under study, and let p denote the probability of testing positive, which includes both true and false positive results. Denote by X the number of individuals from the sample who test positively, and let θ denote the true prevalence of the disease in that population. We have

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

since each positive test either arises as a true positive, with probability θs, or as a false positive, with probability $(1 - \theta)(1 - c)$. Since θ, s and c must all lie in the interval [0, 1], equation (1) implies that p must lie in the interval $[1 - c, s]$. One common estimator of p is its MLE. As discussed in Rohatgi (1984),

$$\mathrm{MLE}(p) = \begin{cases} X/n, & \text{if } 1 - c < X/n < s, \\ 1 - c, & \text{if } X/n \leqslant 1 - c, \\ s, & \text{if } X/n \geqslant s. \end{cases}$$

Using equation (1) and the invariance property of MLEs (see Casella and Berger (1990), p. 294), the MLE of θ is

$$\mathrm{MLE} = \begin{cases} \dfrac{X/n - (1 - c)}{s + c - 1}, & \text{if } 1 - c < X/n < s, \\ 0, & \text{if } X/n \leqslant 1 - c, \\ 1, & \text{if } X/n \geqslant s. \end{cases} \tag{2}$$

The MLE performs reasonably well for most values of θ. When θ is small, however, the MLE is quite often 0, even when the unobserved number of truly diseased subjects in the sample, Y, is not 0. In general, $P(Y = 0) = (1 - \theta)^n$, whereas $P(\mathrm{MLE} = 0) = P(X/n \leqslant 1 - c)$, and the latter can be much larger than the former.

Table 1. Probability of no positive subjects in a sample of size n, $P(Y = 0)$, *versus* the probability that the MLE of the prevalence, θ, is 0†

θ	n	$P(Y = 0)$	$P(MLE) = 0$
0.050	100	0.006	0.205
0.040	100	0.017	0.252
0.030	100	0.048	0.306
0.020	100	0.133	0.366
0.010	100	0.366	0.431
0.010	500	0.007	0.350
0.005	500	0.082	0.423

†The calculations shown are for the case when the sensitivity is 0.9, and the specificity is 0.8.

Table 1 illustrates this for various values of θ and n, when $s = 0.9$ and $c = 0.8$. For Table 1 we used the normal approximation to the binomial distribution to calculate $P(X/n \leqslant 1 - c)$. Since

$$P(X/n \leqslant 1 - c) = P\left[\frac{X/n - p}{\surd\{p(1 - p)/n\}} \leqslant \frac{1 - c - p}{\surd\{p(1 - p)/n\}}\right],$$

we have

$$P(\text{MLE} = 0) \approx \Phi\left[\frac{1 - c - p}{\surd\{p(1 - p)/n\}}\right],$$

where $\Phi(t)$ denotes the standard normal cumulative distribution function evaluated at t, and where p is given by equation (1).

Table 1 shows that $P(\text{MLE}) = 0$ can occur more than 50 times as often as $P(Y = 0)$. Although here the sensitivity and specificity were set to 0.9 and 0.8 respectively, similar results occur for other sensitivity and specificity values. When it is likely that there is one or more positive subjects in the sample, it is obviously preferable not to use 0 as the point estimate of the prevalence. In the next section, we present an adjustment to the standard MLE that produces a positive estimate.

3. Adjustment to the maximum likelihood estimator

The numerator of equation (2) is $X/n - (1 - c)$ when $1 - c < X/n < s$, which produces a negative estimate when $X/n \leqslant 1 - c$. We shall develop an adjusted estimator that will subtract a quantity less than $1 - c$ when $X/n \leqslant 1 - c$, resulting in an estimate that remains greater than 0 even in this case.

Suppose that we have a sample of size n. Let Z be the unobserved latent data representing the number of truly positive subjects out of X positively testing subjects, and let Y be the unobserved total number of truly positive subjects in the sample. See Table 2.

By definition, $E(X/n) = p$, $E(Y/n) = \theta$, $E(Z/Y) = s$ and $E\{(X - Z)/(n - Y)\} = 1 - c$, so that, for example, the relationship $p = \theta s + (1 - \theta)(1 - c)$ is equivalent to

$$E(X/n) = E(Y/n)\,E(Z/Y) + \{1 - E(Y/n)\}\,E\{(X - Z)/(n - Y)\}. \tag{3}$$

Let $X = x$ denote the observed number of positive tests in a given study. To motivate the definition of an adjusted maximum likelihood estimator (AMLE) when the $\text{MLE} = 0$, i.e. when $x/n \leqslant 1 - c$, assume that equation (3) remains true when x is given. Then $(x - Z)/(n - Y)$ would be the point estimate of $1 - c$ from the sample, but this is not directly observable. We

Table 2. Observed and latent data when a diagnostic test is given to a sample of n individuals†

		Test result		
		+	−	
True disease status	+	Z	$Y-Z$	Y
	−	$X-Z$	$n-X-Y+Z$	$n-Y$
		X	$n-X$	n

†The variable X represents the number of subjects observed to test positively, and Y represents the unobserved number of truly diseased subjects. The number of correctly identified positive subjects, Z, is also not observed.

suggest the expected value of $(x-Z)/(n-Y)$, given x and n, as an estimate of $1-c$. When the number of diseased individuals in the sample is small with respect to n, $(x-Z)/(n-Y)$ will be approximately normally distributed with mean $1-c$ and variance $c(1-c)/n$. Letting $H=(x-Z)/(n-Y)$, we then need to calculate $E(H|x)$. According to Table 2,

$$X/n=(Y/n)(Z/Y)+(1-Y/n)(X-Z)/(n-Y),$$

so that

$$(X-Z)/(n-Y)\leqslant X/n\leqslant Z/Y.$$

This follows, since we assume that $s>1-c$, and p is a convex combination of s and $1-c$ according to equation (1).

If we can calculate $E(H|x)$, the following approximation to the term $E(Y/n|x)$, which can be used as an estimator of θ given data x, can then be derived (see Appendix A):

$$E(Y/n|x)\approx\frac{x/n-E(H|x)}{s-E(H|x)},$$

where, as indicated above, $H|x$ follows a truncated normal density with mean $1-c$ and variance $c(1-c)/n$, but with the constraint that $H\leqslant x/n$. A detailed derivation of $E(H|x)$ is given in Appendix A, where it is shown that

$$E(H|x)=1-c-\sqrt{\left\{\frac{c(1-c)}{2\pi n}\right\}}\exp\left[-\frac{\{x/n-(1-c)\}^2}{2c(1-c)/n}\right]\bigg/\Phi\left[\frac{x/n-(1-c)}{\surd\{c(1-c)/n\}}\right].$$

An AMLE of θ can then be defined as

$$\text{AMLE}=\begin{cases}\dfrac{x/n-(1-c)}{s+c-1}, & \text{if } 1-c<x/n<s,\\[2ex] \dfrac{x/n-E(H|x)}{s-E(H|x)}, & \text{if } x/n\leqslant 1-c,\\[2ex] 1, & \text{if } x/n\geqslant s.\end{cases}$$

The AMLE is equivalent to the MLE given by equations (2) except when $x/n\leqslant 1-c$, when the latter produces a point estimate of 0. For example, if $s=0.9$, $c=0.8$ and 16 positive results are observed in 100 tests, AMLE $=0.028$, whereas from equations (2) the MLE is 0. This estimate is easier to calculate than the Bayesian posterior mean as suggested by Lew and Levy (1989), since

Table 3. Examples comparing the AMLE with Bayesian posterior mean estimates of disease prevalence†

Example	n	x	s	c	*AMLE*	*Uniform*	*Jeffreys*
1	100	16	0.9	0.8	0.028	0.032	0.018
2	483619	704	0.9930	0.9982	1×10^{-5}	5×10^{-5}	5×10^{-6}
3	773	279	0.55	0.63	0.055	0.063	0.036
4	96	8	0.89	0.74	0.0125	0.0176	0.0089

†A sample size of n subjects results in x positive test results. The test is assumed to have sensitivity s and prevalence c. The column labelled AMLE provides the adjusted maximum likelihood estimate of the prevalence, whereas the last two columns provide the Bayesian posterior means from uniform and Jeffreys prior densities respectively.

we only need a table of the standard normal distribution along with a hand calculator with square root and exponential functions. Nevertheless, using numerical integration, the posterior mean for θ based on a uniform prior density is 0.032. In contrast, the standard Jeffreys non-informative prior density for θ, which in this case is a beta density with shape and scale parameters both equal to 0.5 (Gelman *et al.*, 1995), gives a posterior mean of 0.018, almost half the size of the estimate based on the uniform prior density. The AMLE estimate is in this case located between the two Bayesian point estimates.

Table 3 summarizes the above example and three additional published examples. Example 2 is from Johnson and Gastwirth (1991), which discussed tests for the detection of the human immunodeficiency virus that have very high sensitivity and specificity. In contrast, example 3, from Centor (1992), discussed the use of serum creatine kinase for the diagnosis of myocardial infarction, a test which has relatively poor sensitivity and specificity. For illustration, we used a specificity of 0.63, rather than the 0.65 value suggested by Centor (1992), since with $c = 0.65$ the AMLE equals the usual MLE. The results are similar, however, whichever value of c is used. Finally, example 4 comes from Lew and Levy (1989), where chest radiographs were used for the diagnosis of pulmonary hypertension. In all four examples, and over the broad range of sensitivity and specificity values encountered, the point estimate from the AMLE fell between the two Bayesian estimates. Although for rare diseases the posterior mean from the uniform prior density will always give higher estimates than the posterior mean from the Jeffreys prior density, it seems difficult to prove that the AMLE will always fall between the two Bayesian estimates. Therefore, it may be possible to find a counter-example, although we did not find any in these or other examples that we investigated.

4. Comparison of the adjusted maximum likelihood estimator with the maximum likelihood estimator

Although the above example suggests that the AMLE may improve on the usual MLE when the latter is 0, we performed a simulation to quantify the improvement. We considered several prototypic situations, where the specificity is 0.8, the sensitivity is either 0.9 or 0.8 and the prevalence runs from 1% to 5%, in increments of 1%. For each of these 10 situations, we ran 10000 simulations with sample sizes of $n = 100$ and $n = 500$, for 20 different cases.

For each simulation, we calculated the number of times that the MLE was 0, and the mean AMLE conditional on the MLE being 0. The mean-squared error (MSE) is defined as the average of the squared deviations between the estimator and the true parameter value in each simulation. We calculated the square root of the MSE of the MLE, denoted by RMSE(MLE), so that

$$\mathrm{RMSE(MLE)} = \sqrt{\left\{ \sum_{i=1}^{10000} (\mathrm{MLE} - \theta)^2 \Big/ 10000 \right\}}.$$

We used the square root so that it would be on the same scale as the estimators. Similarly,

$$\mathrm{RMSE(AMLE)} = \sqrt{\left\{ \sum_{i=1}^{10000} (\mathrm{AMLE} - \theta)^2 \Big/ 10000 \right\}}.$$

Since we are interested in the cases where the MLE and the AMLE differ, we defined the conditional RMSE to be the RMSE conditional on MLE = 0. We denote the square root of the conditional MSE of the AMLE by CRMSE(AMLE), i.e.

$$\mathrm{CRMSE(AMLE)} = \sqrt{\left\{ \sum_{\mathrm{MLE}=0} (\mathrm{AMLE} - \theta)^2 \Big/ k \right\}}.$$

CRMSE(AMLE) = θ, since

$$\mathrm{CRMSE(MLE)} = \sqrt{\left\{ \sum_{\mathrm{MLE}=0} (\mathrm{MLE} - \theta)^2 \Big/ k \right\}} = \sqrt{\left\{ \sum_{\mathrm{MLE}=0} (0 - \theta)^2 \Big/ k \right\}} = \theta.$$

Table 4 contains the results of these simulations. Although both estimators are biased, in most of the cases the RMSE of the AMLE is smaller than that of the MLE. For example, for $\theta = 0.05$,

Table 4. Square root of the MSEs for the MLE and AMLE†

	θ	n	s	*P(MLE = 0)*	*mean(AMLE)*	*RMSE(MLE)*	*RMSE(AMLE)*	*CRMSE(AMLE)*
1	0.05	100	0.9	0.2286	0.035	0.051	0.046	0.016
2	0.05	100	0.8	0.2495	0.040	0.059	0.054	0.012
3	0.04	100	0.9	0.2843	0.035	0.050	0.045	0.009
4	0.04	100	0.8	0.3129	0.040	0.056	0.051	0.009
5	0.03	100	0.9	0.3328	0.034	0.047	0.044	0.009
6	0.03	100	0.8	0.3581	0.039	0.054	0.051	0.013
7	0.02	100	0.9	0.3636	0.027	0.037	0.035	0.008
8	0.02	100	0.8	0.4130	0.039	0.051	0.051	0.021
9	0.01	100	0.9	0.4719	0.033	0.043	0.045	0.024
10	0.01	100	0.8	0.4774	0.038	0.050	0.054	0.030
11	0.05	500	0.9	0.0315	0.017	0.026	0.025	0.033
12	0.05	500	0.8	0.0573	0.019	0.030	0.028	0.030
13	0.04	500	0.9	0.0706	0.017	0.025	0.024	0.023
14	0.04	500	0.8	0.1058	0.019	0.029	0.027	0.021
15	0.03	500	0.9	0.1299	0.017	0.024	0.022	0.014
16	0.03	500	0.8	0.1780	0.019	0.027	0.024	0.012
17	0.02	500	0.9	0.2319	0.016	0.022	0.020	0.005
18	0.02	500	0.8	0.2685	0.018	0.024	0.022	0.004
19	0.01	500	0.9	0.3598	0.016	0.020	0.019	0.006
20	0.01	500	0.8	0.3834	0.017	0.023	0.023	0.008

†The sensitivity of the test is denoted by s, while the specificity is held constant at $c = 0.8$. The column labelled *P*(MLE = 0) presents the proportion of times that MLE = 0 in 10000 simulations, and mean(AMLE) is the average of the AMLE for these cases. The sample size is denoted by n. RMSE(MLE) is the square root of the MSE of the MLE, RMSE(AMLE) is the square root of the MSE of the AMLE and CRMSE(AMLE) is the square root of the MSE of the AMLE conditional on MLE = 0. The square root of the MSE of the MLE when it equals 0 is equal to θ (see text).

$n = 100$, $s = 0.9$ and $c = 0.8$, the number of times that the MLE was 0 out of 10000 simulations was 2286, or about 23% of the time. The conditional RMSE of the AMLE is 0.016, whereas the conditional RMSE of the MLE is much larger, 0.05. If we increase the sample size to 500, the number of times that the MLE is 0 out of 10000 simulations drops to 315. This is because, when the sample size is increased, x/n becomes closer to the true value $p > 1 - c$, so the probability of observing $x/n \leqslant 1 - c$ decreases. We note that the performance of the AMLE is poorer than that of the MLE only in rows 8–10 of Table 4, when both the sample size and the prevalence are relatively small. When the prevalence is near 0, we would usually wish to have-a sample size that is much larger than 100, since small values of θ are usually paired with small values of d (see equation (4) in Section 6). For example, to estimate a prevalence of $\theta = 0.02$ to an accuracy of $\pm d = 0.01$ with a 95% confidence interval requires a sample size of $n = 753$, even with a perfect diagnostic test ($s = c = 1$). Hence, situations like those in rows 8–10 of Table 4 should not frequently arise in practice. With appropriate sample sizes, the RMSE of the AMLE appears always to be smaller than the RMSE of the MLE, and in particular the conditional RMSE of the AMLE is substantially lower than that of the usual MLE.

5. Confidence intervals

Using the normal approximation to the binomial distribution, an approximate confidence interval for p is given by the intersection of the interval

$$(X/n - Z_{\alpha/2}\surd\{p(1-p)/n\},\ X/n + Z_{\alpha/2}\surd\{p(1-p)/n\})$$

with the interval $[1 - c, s]$, where $Z_{\alpha/2}$ is the usual standard normal upper $100(1 - \alpha/2)$% quantile. Since p is unknown, it is usually approximated by x/n. In the current context, however, p is restricted to the interval $[1 - c, s]$. Therefore, we approximate p by x/n only when $1 - c < x/n < s$ and by $1 - c$ when $x/n \leqslant 1 - c$. Straightforward algebra then shows that the interval

$$\left(\frac{X/n + c - 1 - Z_{\alpha/2}\surd\{p(1-p)/n\}}{s + c - 1},\quad \frac{X/n + c - 1 + Z_{\alpha/2}\surd\{p(1-p)/n\}}{s + c - 1}\right)$$

intersected with the interval $[0, 1]$ is an approximate $100(1 - \alpha)$% confidence interval for the prevalence θ. The interval contains both the MLE and the AMLE.

Since the above derivation is based on a normal approximation, we performed additional simulations, again of size 10000, to estimate the proportion of times that the 95% confidence intervals capture the true prevalence. Using a sample size of $n = 100$, we let θ vary between 1% and 5% by increments of 1%, whereas the sensitivity s varied between 0.7 and 0.9, and the specificity c was either 0.8 or 0.9. See Table 5. In each case that we considered, the coverage was very close to 95% and was often slightly higher than 95%, perhaps indicating a slight conservative bias. The bias decreases with increasing sample size and most probably arises from the use of the normal approximation to the binomial distribution. Similar results occur for other sensitivity and specificity values.

6. Sample size for estimating the prevalence

In planning a prevalence survey, an investigator may wish to determine the sample size that is needed to estimate the prevalence to within an accuracy of $\pm d$ using a $100(1 - \alpha)$% confidence

Table 5. Proportion of times (prop) out of 10000 that the 95% confidence intervals captured θ†

θ	s	c	*prop*	θ	s	c	*prop*
0.05	0.9	0.8	0.953	0.03	0.8	0.9	0.948
0.05	0.7	0.8	0.959	0.03	0.7	0.9	0.969
0.05	0.8	0.9	0.953	0.02	0.9	0.8	0.960
0.05	0.7	0.9	0.933	0.02	0.7	0.8	0.947
0.04	0.9	0.8	0.958	0.02	0.8	0.9	0.965
0.04	0.7	0.8	0.957	0.02	0.7	0.9	0.952
0.04	0.8	0.9	0.960	0.01	0.9	0.8	0.964
0.04	0.7	0.9	0.958	0.01	0.7	0.8	0.962
0.03	0.9	0.8	0.956	0.01	0.8	0.9	0.976
0.03	0.7	0.8	0.962	0.01	0.7	0.9	0.971

†The sample size is 100 for all simulations, s is the sensitivity and c is the specificity.

interval. Again using the normal approximation to the binomial distribution, it can be easily shown that the sample size required is

$$n = \frac{Z^2_{\alpha/2} p(1-p)}{d^2(s+c-1)^2}, \tag{4}$$

where s and c are the sensitivity and specificity of the test respectively and p is given by equation (1), based on a given value for θ. In practice θ is unknown, so that one may wish to select a final sample size after examining the sizes suggested by a range of θ-values. Equation (4) demonstrates that the sum of the sensitivity and the specificity has a very large influence on sample size requirements. As expected, when $s = c = 1$, the test is error free, $p = \theta$ and equation (4) reduces to the standard binomial sample size formula. At the other extreme, an infinite sample size results if $s + c = 1$, i.e. the test is completely uninformative no matter how large the sample size. Most situations should fall between these extremes.

7. Discussion

We have presented a simple adjustment to the MLE of the prevalence of a disease in a given population, which improves on the standard MLE when it is 0. We have also provided formulae for confidence intervals and sample size determination. These methods should be useful for estimating the prevalence of rare diseases and, unlike Bayesian methods, do not require a choice of a non-informative prior density. Similarly, an adjustment to the MLE can be defined for very common conditions, i.e. when the MLE is 1. This can be done using the above results by reversing both what is considered to be a positive test and what is considered to be the disease state.

Throughout this paper, we have assumed that the sensitivity and specificity of the diagnostic test are exactly known. For well-established diagnostic tests, this may often be a reasonable assumption, and indeed most diagnostic test kits include suggested values for the test properties. In other cases, the test properties are not exactly known, so the sensitivity and specificity must be estimated from the sample along with the prevalence of the disease. With three parameters to estimate, but only 1 degree of freedom in the observed data, maximum likelihood approaches cannot be used without imposing constraints (Walter and Irwig, 1988). Bayesian approaches, however, that provide the joint posterior density of all unknown parameters have been developed. See Joseph *et al.* (1995) for details.

Appendix A

A.1. Approximating E(Y/n|x) when x/n ≤ 1 − c

For large sample sizes, $(X - Z)/(n - Y)$ is approximately normally distributed with mean $1 - c$ and variance $c(1 - c)/n$. Therefore, to approximate $E\{(X - Z)/(n - Y)|x\}$, we consider a random variable H that is normally distributed with mean $1 - c$ and variance $c(1 - c)/n$, but with the added constraint that $H \leqslant x/n$. The expected value $E(H|x)$ then approximates $E\{(X - Z)/(n - Y)|x\}$. Substituting these approximations into equation (3) while considering x as fixed gives

$$x/n \approx E(Y/n|x)s + \{1 - E(Y/n|x)\}\, E(H|x),$$

and solving for $E(Y/n|x)$ gives

$$E(Y/n|x) \approx \frac{x/n - E(H|x)}{s - E(H|x)}.$$

A.2. Calculating E(H|x)

$E(H|x)$ is defined by

$$E(H|x) = \int_{-\infty}^{x/n} h\, f_{H|X}(h|x)\, \mathrm{d}h,$$

where $f_{H|X}(h|x)$ denotes the conditional density function of H given $X = x$. Let $f_H(h)$ denote the density function of H, $F_H(h)$ denote the distribution function of H and $F_{H|X}(h|x)$ denote the conditional distribution function of H given $X = x$. We then have

$$\begin{aligned} F_{H|X}(h|x) &= F_{H|X}(h|h \leqslant x/n) \\ &= \begin{cases} \dfrac{F_H(h)}{F_H(x/n)}, & \text{if } h \leqslant x/n, \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

and therefore

$$f_{H|X}(h|x) = \begin{cases} \dfrac{f_H(h)}{F_H(x/n)}, & \text{if } h \leqslant x/n, \\ 0, & \text{otherwise.} \end{cases}$$

Hence

$$E(H|x)\, F_H(x/n) = \int_{-\infty}^{x/n} \frac{1}{\surd\{2\pi c(1 - c)/n\}}\, h \exp\left[-\frac{\{h - (1 - c)\}^2}{2c(1 - c)/n}\right] \mathrm{d}h.$$

Let

$$u = \frac{h - (1 - c)}{\surd\{c(1 - c)/n\}};$$

then

$$\begin{aligned} E(H|x)\, F_H(x/n) &= \int_{-\infty}^{\{x/n - (1-c)\}/\surd\{c(1-c)/n\}} \frac{\surd\{c(1 - c)/n\}u + 1 - c}{\surd(2\pi)} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \\ &= -\sqrt{\left\{\frac{c(1 - c)}{2\pi n}\right\}} \exp\left[-\frac{\{x/n - (1 - c)\}^2}{2c(1 - c)/n}\right] + (1 - c)\, \Phi\left[\frac{x/n - (1 - c)}{\surd\{c(1 - c)/n\}}\right], \end{aligned}$$

so that

$$E(H|x) = 1 - c - \sqrt{\left\{\frac{c(1 - c)}{2\pi n}\right\}} \exp\left[-\frac{\{x/n - (1 - c)\}^2}{2c(1 - c)/n}\right] \Big/ \Phi\left[\frac{x/n - (1 - c)}{\surd\{c(1 - c)/n\}}\right].$$

References

Casella, G. and Berger, R. L. (1990) *Statistical Inference.* California: Wadsworth and Brooks.

Centor, R. M. (1992) Estimating confidence intervals of likelihood ratios. *Med. Decsn Makng*, **12**, 229–233.

Gelman, A., Carlin J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis.* London: Chapman and Hall.

Johnson, W. O. and Gastwirth, J. L. (1991) Bayesian inference for medical screening tests: approximations useful for the analysis of acquired immune deficiency syndrome. *J. R. Statist. Soc.* B, **53**, 427–439.

Joseph, L., Gyorkos, T. and Coupal, L. (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am. J. Epidem.*, **141**, 263–272.

Lew, R. A. and Levy, P. S. (1989) Estimation of prevalence on the basis of screening tests. *Statist. Med.*, **8**, 1225–1230.

Rogan, W. J. and Gladen, B. (1987) Estimating prevalence from the result of a screening test. *Am. J. Epidem.*, **107**, 71–76.

Rohatgi, V. K. (1984) *Statistical Inference*. New York: Wiley.

Taragin, M. I., Wildman, D. and Trout, R. (1993) Assessing disease prevalence from inaccurate test results: teaching an old dog new tricks. *Med. Decsn Makng*, **14**, 269–273.

Walter, S. D. and Irwig, L. M. (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J. Clin. Epidem.*, **41**, 923–937.