Original Research Paper Statistics

Maximum Likelihood Approach in Estimating the Prevalence of Tuberculosis

Neha Gupta	Department of Statistics, University of Jammu, Jammu
Rahul Gupta	Department of Statistics, University of Jammu, Jammu
J P Singh Joorel	Department of Statistics, University of Jammu, Jammu

KEYWORDS

INTRODUCTION AND BACKGROUND

Tuberculosis (TB) is caused by a bacterium called *Mycobacterium tuberculosis*. The bacteria usually attack the lungs, but TB bacteria can attack any part of the body such as the kidney, spine, and brain. If not treated properly, TB disease can be fatal. Tuberculosis (TB) is one of the most common infections in the world. Though India is the second-most populous country in the world, India has more new TB cases annually than any other country. In 2013, out of the estimated global annual incidence of 9.4 million TB cases, 2 million were estimated to have occurred in India, thus contributing to a fifth of the global burden of TB. The incidence of TB in India is estimated based on findings of the nationwide annual risk of tuberculosis infection (ARTI) study conducted in 2000-2003. The prevalence of TB has been estimated at 3.8 million bacillary cases for the year 2000, by an expert group of Govt. of India. However the recent estimate by WHO gives a Prevalence of 3 million.

Prevalence is a dimensionless, unit-free value ranging from zero to one (zero to 100 , if expressed as a percentage). Depending on the context, an investigator might be interested in prevalence of infection, infectious animals or disease.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 positive test results, and estimating the prevalence is the classical problem of estimating a binomial proportion. Gold standard test rarely if ever exist, however, since even a theoretically perfect test can be rendered less perfect by human, laboratory or other error. Even when they exist, gold standard tests may be difficult to perform, highly invasive, very costly of 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. Walter and Irwig (1988) have given a comprehensive review of methods related to this problem. 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' choice, such as the standard Jeffreys prior density (Gleman 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 prevalence estimate in the line of Rahne and Joseph(1998) and apply it for estimating Prevalence of Tuberculosis, for Jammu Division. 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.

Maximum Likelihood estimation

Suppose that the sensitivity and 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 include both true and false positive results. Denote by X the number of individuals from the sample who test positively, and let denote the true prevalence of the disease in that population. We have

p = s+ (1-) (1-c) 1)

since each positive test either arises as a true positive, with probability or as a false positive, with the probability (1-)(1-c). Since , s and c all must 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),

$$\text{MLE (p)} = \begin{cases} \frac{X}{n} & if\ 1-c < \frac{X}{n} < s \\ 1-c, & if \frac{X}{n} \leq 1-c \\ s, & if \frac{X}{n} \geq s \end{cases}$$

Using equation (1) and the Invariance property of MLEs, the MLE of is

$$\text{MLE} = \begin{cases} \frac{\left[\frac{X}{n}-(1-c)\right]}{s+c-1} & if\ \ 1-c < \frac{X}{n} < s \\ 0 & if \frac{X}{n} \leq 1-c \\ 1 & if \frac{X}{n} \geq s \end{cases}$$

The MLE performs reasonably well for most values of . When 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) = , whereas

P(MLE=0)=P(X/n ≤ 1-c),and the later can be much larger than the former.

Table 1 illustrates this for various values of and n, when s=0.9 and c=0.8

We use the normal approximation to the binomial distribution to calculate P(X/n). Since

$$\mathrm{P(X/n \le 1\text{-}c) = P}\left[\frac{\frac{x}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{1-c-p}{\sqrt{\frac{p(1-p)}{n}}}\right],$$

we have

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

P(MLE=0) ≈ Φ,

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

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

DISTRICT		n	P(Y=0)	P(MLE=0)
JAMMU	0.001354	1000	0.2579	0.4721
RAJOURI	0.001297	1000	0.2731	0.4721
DODA	0.005552	1000	0.0038	0.3821
POONCH	0.001173	5000	0.0028	0.4443
UDHAM-PUR	0.001329	5000	0.0013	0.4721
KATHUA	0.007897	5000	0.0004	0.4522

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

ADJUSTMENT TO THE 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 ≤ 1-c. We shall develop an adjusted estimator that will subtract a quantity less than 1-c when X/n ≤ 1-c, resulting in an estimate that remains greater than zero even in this case.

Suppose that we have a sample of size n. Let Z be an 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) = , E (Z/Y) = s and E {(X-Z)/ (n-Y)} = 1-c, so that, for example, the relationship p = s+ (1-) (1-c) is equivalent to

E(X/n) = E(Y/n)E(Z/Y)+{1-E(Y/n)}E{(X-Z)/(n-Y)} (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 MLE = 0, i.e. when x/n ≤ 1-c, assume the 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 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).

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

		TEST RESULT		
		+	-	
TRUE DISEASE	+	Z	Y-Z	Y
STATUS	-	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.

According to Table 2,

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

So that

(X-Z)/(n-Y) ≤ X/n ≤ 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, the following approximation to the term E, which can be used as an estimator of given data x, can then be derived as

E then approximates E

Substituting these approximation into equation (3) considering x as fixed gives

$$\mathrm{x/n} \approx \mathrm{E}\left(\frac{Y}{n}\middle|x\right)\mathrm{s} + \left\{1 - E\left(\frac{Y}{n}\middle|x\right)\right\}\mathrm{E}(H|x)$$

$$\mathrm{x/n} \approx \mathrm{E}\left(\frac{Y}{n}\middle|x\right)\mathrm{s} + \mathrm{E}(H|x) - \mathrm{E}(H|x)\,\mathrm{E}\left(\frac{Y}{n}\middle|x\right)$$

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

$$\mathrm{x/n} - \mathrm{E}(H|x) \approx \mathrm{E}\left(\frac{Y}{n}\middle|x\right)[s - \mathrm{E}(H|x)]$$

$$\left[\frac{\left(\frac{x}{n} - E(H|x)\right)}{s - E(H|x)}\right] \approx \mathrm{E}\left(\frac{Y}{n}\middle|x\right)$$

follows a truncated normal density with mean (1-c) and variance c(1-c)/n , but with the constraint that H ≤ x/n. A detailed derivation of E shows that

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

An AMLE of can be defined as:

$$\mathrm{AMLE} = \begin{cases} \dfrac{\frac{x}{n}-(1-c)}{s+c-1} & if\ 1 - c < \frac{x}{n} < s, \\ \dfrac{\frac{x}{n}-E(H|x)}{s-E(H|x)} & if\ \frac{x}{n} \le 1 - c, \\ \quad 1 & if\ \frac{x}{n} \ge s \end{cases}$$

The AMLE is equivalent to the MLE given by equation (2) except when x/n ≤ 1-c, when the latter produces a point estimate of 0. For example if s = 0.9 and c = 0.8 and 16 positive results are observed in 100 tests, AMLE = 0.018, whereas from equation (2) the MLE is 0. This estimate is easier to calculate than the Bayes posterior mean as suggested by Lew and Levy (1989), since we only need a table of standard normal distribution along with a hand calculator with square root and exponential functions.

Below Table 3 summarizes the above example and two additional published example.

TABLE 3.

Example	n	x	s	c	AMLE
1	100	16	0.9	0.8	0.018
2	773	279	0.55	0.63	0.054
3	96	8	0.89	0.74	0.134
4	545886	4974	0.9	0.8	0.286

A sample size of n subjects results in x positive test results. AMLE provides the adjusted maximum likelihood estimate of the prevalence.

Example 2 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 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. Example 3 comes from Lew and Levy (1989), where chest radiographs were used for the diagnosis of pulmonary hypertension. Example 4 is the collection of data from Jammu Division on Tuberculosis.

CONFIDENCE INTERVALS

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

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

With the interval [1-c, s] where is the usual standard normal upper 100(1-α/2)% quantile.

Using Data

X/n = 0.00091

Then C.I ≈ (0, 0.025)

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 1-4/3c < x/n < 3/4s and by 1-4/3c when x/n ≤ 1-4/3c.

Straightforward algebra then shows that the interval

$$\left(\frac{\frac{x}{n}+\frac{4}{3}c-1-Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}}{\frac{3}{4}s+\frac{4}{3}c-1},\frac{\frac{x}{n}+\frac{4}{3}c-1+Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}}{\frac{3}{4}s+\frac{4}{3}c-1}\right)$$

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

Sample Size for Estimating the Prevalence

In planning a sampled survey, an investigator may wish to determine the sample size that is needed to estimate the prevalence to within an accuracy of using a 100(1-□)% confidence interval. Again using the normal approximation to the binomial distribution, it can be easily shown that the sample size required is

$$n=\frac{Z^\wedge 2\,\alpha/_2}{d^2(s+c-1)^\wedge 2} \qquad (4)$$

where s and c are sensitivity and specificity of the test respectively and p is given by equation (1), based on given value of . In practice is unknown, so one may wish to select a final sample size after examining the size suggested by a range of values. Equation (4) demonstrates that the sum of the sensitivity and specificity has a very large influence on sample size requirements. As expected, when s=c=1, the test is error free, p= 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.

References:

1. Casella,G. and Berger, R.L.(1990) Statistical Inference. California: Wadsworth and Brooks.
2. Centor, R.M. (1992) Estimating confidence intervals of likelihood ratios. Med. Decsn Makng,12, 229-233
3. Gelman, A., Carlin J.B., Stern, H.s. and Rubin, D.B.(1995) Bayesian Data Analysis: Chapman and Hall.
4. Johnson, W.O. and Gastwirth, J.L.(1991) Bayesian inference for medical screening tests: approximation useful for the analysis of acquired immune deficiency syndrome. J.R. Statist. Soc.B,53,427-439.
5. Joseph, L, Gyorkos, T. and Coupal, L. (1995) Bayesian estimate of disease prevalence and the parameters of diagnostic test in the absence of a gold standard. Am. J. Epidm., 141,263-272
6. Lew, R.A. Levy, P.S. (1989) Estimation of prevalence on the basis of screening tests.Statist. Med., 8, 1225-1230
7. Rahme, E. and Joseph, L.(1998) Estimating the prevalence of a rare disease:-Adjusted maximum Likelihood. The Statistician., 47, Part 1, 149-158.
8. Rogan, W.J. and Gladen, B. (1987) Estimating prevalence from the result of a screening tests. Am. J. Epidem., 107, 71-76.
9. Rohatgi, V. K. (1984) Statistical Inference. New York: Wiley.
10. Taragin, M.L., Wildman, D. and Trout, R. (1993) Assessing disease prevalence from inaccurate tests results: teaching an old dog new tricks. Med. Decsn Makng, 14, 269-273.
11. 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.