Estimation of the Prevalence Under Misclassification

Instr. Meric Colak
Baskent University,
Dept. of Health Management
06530, Ankara, Turkey
meric@baskent.edu.tr

Mehtap Akcil Temel, PhD
Baskent University,
Dept. of Statistics
06530,Ankara, Turkey
matemel@baskent.edu.tr

Ergun Karaagaoglu, PhD
Hacettepe University,
Dept. of Biostatistics
Ankara, Turkey
ekaraaga@hacettepe.edu.tr

1. Introduction And Methodology

It is often necessary to estimate the prevalence of a disease in the underlying population on the basis of a screening test that cannot discriminate diseased and non-diseased subjects with 100% certainty. When such imperfect gold standard screening tests are administered to a sample of individuals, the proportion of subjects with a positive test result therefore cannot be used as an estimate of the true prevalence in the population. In this study, two different algorithms for prevalence estimation are given under misclassification. The first one is Maximum Likelihood Estimate (MLE) of prevalence using the Relative Likelihood Function under misclassification. The second one is a method, which requires the evaluation of a logistic function.

2.Approximate Bayesian Estimate of the True Population Prevalence

When the Bayesian estimate and proportion of positives in a test with known sensitivity and specificity are plotted for different sample sizes, points fit very well to a logistic function in the form of, $y=1/(1+e^{-(b0+b1x)})$, where y is the Bayesian estimate and x is the proportion of subjects with a positive test result. SPSS program is used to estimate constants b_0 and b_1 for different sample sizes and for different sensitivity and specificity values, as a total of 1225 functions. Approximate Bayesian estimate of y, Π_a, is obtained. For practical purposes, only the coefficients for some selected sensitivities, specificities and sample sizes are tabulated.

3.Construction of Likelihood and Relative Likelihood

Under suppositions of independence and of constant probabilities governing the misclassification process, the distribution of the number of cases showing the disease is given as convolution of two binomials compounded by hyper geometric by Johnson and Kotz. If ‘A’ among the ‘N’ in the population have the disease then the probability of ‘a’ having it among the ‘n’ in the sample is hyper geometric, while the chance that, ‘x’, which is the total number of persons either correctly or incorrectly classified as diseased after the inspection of the sample, of the ‘n’ will show the disease; given that ‘a’ have it can be formally expressed as;

$$\sum_{a=a_1}^{a=a_2}\sum_{k=k_1}^{k=k_2}\binom{a}{k}h^{k}(1-h)^{a-k}\binom{n-a}{x-k}(1-q)^{x-k}q^{n-a-x+k}\Big/\binom{N}{n}$$

$k_1=\max[0,(a-x+n)]$ and $k_2=\min[x,a]$ $a_1=\max[0,(n-N+A)]$ and $a_2=\min[A,n]$

And h is the sensitivity, q is the specificity values.

After dividing the Likelihood (L) by its maximum, Maximum Likelihood Estimate of a parameter, Relative Likelihood (RL) are obtained. Simulation study is introduced to observe such relations and to see how good of estimators we obtained for the prevalence. A finite population of

100 is obtained and classified as diseased or non-diseased by the true device, where no classification error exists, and prevalence of disease is recorded as 0.77. At each iteration, a sample of n=50 person is drawn from the population. They are classified by the true device and then by the fallible device with the given error probabilities and the number of diseased persons in no error case, a, is obtained as 39 in the sample and that number in error case, x, is obtained for each misclassified sample. At each iteration, likelihood probabilities are calculated under error and no error case. MLE of prevalence is obtained under misclassification for different values of sensitivity and specificity, after the computation of all iterations the average value of MLE values under error case are obtained as estimate value for prevalence.

4. Discussion

If the sensitivity and specificity of a test that is used for screening individuals for a particular disease are less than unity, the proportion of subjects screened positive by that test cannot be used as an estimate of the true population prevalence. The estimation procedures proposed with logistic function approximation to the Bayesian estimate require simple calculations. As well as being simple, it produces results that always lie between 0 and 1, and are very close to those obtained by Bayesian techniques. Users, who are not familiar with complicated mathematical calculations, can use this method and obtain the required estimates to a high degree of accuracy.

According to the results of simulation study in which Likelihood and Relative Likelihood probabilities are calculated under error and no error case, increases in the sensitivity and specificity result in increases in the estimated values. However, MLE starts with 0.768 and approaches to 0.77 but the estimate value under error case starts with 0.717 and approaches to 0.767. While sensitivity and specificity changed from 0.9 to 0.99. In all cases, MLE of prevalence is found almost equal to 0.77, actual value. When sensitivity is fixed to 0.9 and specificity is increased from 0.9 to 0.99 estimate value under error case are decreased from 0.717 to 0.698, while MLE value are approximately remained at 0.77. When specificity is fixed to 0.9 and sensitivity is increased from 0.9 to 0.99, estimate value under error case is increased from 0.717 to 0.789, while MLE value are almost equal to 0.77 in all cases. As a result, MLE of prevalence values under misclassification give better results than the estimated values where the error parameters are taken as known. If we compare the estimate values of prevalence calculated with two methods, we can say that, they are likely each other.

REFERENCE

Colak, M. (1994). An Enumerative Inferential Procedure In the Presence of Measurement Error. Master Desertion, Ankara, Turkey.

Karaagaoglu, E.(1999). Estimation of Prevalence of a Disease from Screening Tests, Turkish journal of Medical Sciences, 29, 425-430.

Johnson, N.L and Kotz, S. (1982). Errors in Inspection and Grading: Distributional Aspects of Screening and Hierarchical Screening, Comm. Statist., 11(18), 1997-2016.

RESUME

L'estimation de prévalant qui appartient à la population, la condition d' employer les contrôles par tests qui ne sont pas "or standard", la relative vraisemblance fonction trouvera par en employant logistique fonction.