The Korean Communications in Statistics
Vol. 13 No. 3, 2006, pp.787-794

Estimation in Group Testing when a Dilution Effect exists[1]

Sehyug Kwon[2]

Abstract

In group testing, the test unit consists of a group of individuals and each group is tested to classify units from a population as infected or non-infected or estimate the infection rate. If the test group is infected, one or more individuals in the group are presumed to be infected. It is assumed in group testing that classifi -cation of group as positive or negative is without error. But, the possibility of false negatives as a result of dilution effects happens often in practice, specially in many clinical researches. In this paper, dilution effect models in group testing are discussed and estimation methods of infection rate are proposed when a dilution effect exists.

Keywords : Group testing; dilution effect model; estimation.

1. Introduction

In group testing, $k > 1$ individuals from a population are combined into a test unit and a single test is run on the pooled sample. If the pooled sample tests non-infected, all k individuals are considered to be non-infected. When the pooled sample tests infected, at least one individual in the sample is presumed to be infected. When the infection rate is small, group testing has been shown to be more efficient than one-at-a-time testing $(k=1)$ not only in classifying individuals as infected or not in the sense of the expected number of tests to identify all units (Dorfman, 1943) but also in estimating the population infection rate in the sense of minimizing the mean squared error of the maximum likelihood estimator (Thompson, 1962; Swallow, 1985).

In the most of approaches to estimate the infection rate, one of the underlying assumptions is that there is no false positive and/or negative in testing the pooled group. False positive makes us to classify incorrectly a non-infected group as infective, which is a rare event and less of interest in practice. False negative (dilution effect) leads us to fail to identify infected individuals even if there is

[1] The paper was supported by the research grant of Hannam University.

[2] Professor, Department of Statistics, Hannam University, Daejon, 306-791, Korea.
E-mail : wolfpack@hannam.ac.kr

some infected individuals in the pooled group. When a dilution effect exists, the MLE is computed by the smaller value and all individuals in the group will be free to go from further classification even if some are infected by a serious disease.

The studies to estimate prevalence of HIV-1 seropositivity have widely used group testing (Gastwirth and Hammick, 1989; Kline et al., 1989; Litvak et al., 1994; Tu et al., 1995). For screening HIV infected individuals, it had been shown that false positives or false negatives are negligible when a group size of up to 15 is used (Emmanuel et al., 1988; Cahoon-Young et al., 1989; Kline et al., 1989; Monzon et al., 1992). But, the estimation problem has been rarely studied when dilution effects exist, which happens often in practice. Chen and Swallow (1995) showed that group testing is still efficient even in the presence of a dilution effect when the infection rate is small. Hung and Swallow (1999) proposed two dilution effect models and discussed them by examining the behavior of MSE (Mean Squared Error) and bias, but did not cover an estimation procedure.

Retesting individuals from the infected groups is not worthwhile in estimating the infection rate, for it would only negligibly reduce the mean squared error of the estimator (Chen and Swallow, 1990). But, when a dilution effect exist, a retesting scheme is needed to estimate two parameters, the infection rate and the dilution effect. Chen and Swallow (1990) also showed that halving procedures are nearly optimal and convenient.

In this paper, dilution effect models are discussed and approaches to estimate the infection rate in the presence of dilution effects are proposed. Section 2 describes the usual model of group testing and summarizes two dilution effect models proposed by Hung and Swallow (1999). In section 3, modified dilution effect models for estimations is suggested and estimation methods are proposed and discussed. Conclusions and further discussions are summarized in section 4.

2. Usual Group Testing Model and Dilution Effect Models

Suppose that p is the population infection rate, n is the number of testing groups, and k is the group size. Let $D=\sum_{i=1}^{n} x_i$ be the total number of infected groups from n testing groups, where x_i is the test result of ith group with x_i being 1 (infected) or 0 (non-infected). In the usual group testing model, D is distributed as Binomial $(n, 1-(1-p)^k)$ with the following assumptions: (1)The individuals in the population are independently and identically distributed as Bernoulli (p), (2)The same group size k is used for each n groups and total number of individuals in group testing is equal to nk, and (3)There is no false negative and false positive. The MLE of the infection rate is obtained as

$$\hat{p} = 1 - (1 - D/n)^{1/k}. \tag{1}$$

By Jensen's inequality, $E(\hat{p}) > p$ for $k > 1$. Therefore, the MLE in the usual group testing model overestimates the population infection rate p.

The optimal group size has to be known before the group testing experiment. The choice of the group size to minimize the MSE of $\hat{p}$ has been mostly recommended and widely used (Thompson 1962; Swallow, 1985). Unfortunately, since the optimal group size is a function of the unknown p, the prior guess on p is still needed to choose the optimal group size. If a bound for the true value of p from preliminary data or other considerations is available, which is realistic in practice, the upper bound of it has been suggested in choosing the optimal group size. Taking the lower bound causes us to use too large a value of k, which increases the bias of $\hat{p}$ and inflate its MSE (Swallow, 1985).

When dilution effects (false negatives) exist, the test fails to detect the infection of a test group even though one or more infected individuals in the group are present. False positives make us to classify incorrectly a non-infected group as infective, which is less common in practice than false negatives. False negatives make us to underestimate the true infection rate p, which may offset the overestimation property of group testing somewhat. In the classification, all individuals in the group by dilution effects is free to go from further classification even if it is infected by a serious disease. Classification problems in a dilution effect have been studied somewhat in the sense of obtaining the group size which makes the dilution effect negligible, but estimation problems in a dilution effect have not been studied widely, specially for estimating the infection rate.

Let m_i be the number of infected individuals in the ith group. Then when a dilution effect exists, the probability of x_i being infected can be written

$$\begin{aligned} P(x_i = 1) &= P(x_i = 1, m_i \geqq 0) \\ &= \sum_{j=1}^{k} P(x_i = 1 | m_i = j) P(m_i = j). \end{aligned} \tag{2}$$

Hung and Swallow (1999) proposed the following two dilution effect models:

$$P(x_i = 1 | m_i = j) = \frac{j}{(k-j)f + j} \tag{3}$$

and

$$P(x_i = 1 | m_j = j) = I_{(\frac{j}{k} \geq \frac{1}{f_1})} + \frac{j}{(e^{k - f_1 j} - 1) f_2 + d} I_{(\frac{j}{k} < \frac{1}{f_1})}. \tag{4}$$

In model (3), each non-infected individual contributes a specific dilution factor, $f \leq 1$. For $f = 0$, there is no false negative effect. When $f > 0$, the probability of a false negative in the ith group increases with the number of non-infected units $(k-j)$. They considered a threshold in model (4), where I represents the indicator function. For $(j/k) \geq (1/f_1)$, no dilution effect exists. When the proportion of infected individuals in the ith group j/k is less than $1/f_1$, the distribution of the probability of the infected group being diluted exponentially decays against j/k with coefficient f_2. They discussed the proposed dilution effect models in the sense of MSEs with the various p, but did not discussed an estimation approach to estimate the infection rate. In this paper, two modified dilution effect models are proposed based on Hung and Swallow's models (1999) and estimation methods of the population infection rate are discussed.

3. Modified dilution effect models and Estimations

As described in equation (2), since the probability of the test result of ith group being infected is $P(x_i=1)=\sum_{j=1}^{k} P(x_i=1|m_i=j)P(m_i=j)$, the distribution of the number of infected groups from n test groups is

$$D \sim Binomial(n, \sum_{j=1}^{k} P(x_i=1|m_i=j)P(m_i=j)). \qquad (5)$$

When no dilution effect exists, $P(x_i=1|m_i=j)=1$ for $j \geq 1$ and the proportion in equation (5) reduced to $1-(1-p)^k$ which is the same as in the usual group testing model. When a dilution effect exists, the proportion is a function of the infection rate p and a dilution effect f.

With model (3), the MLE of the infection rate can be obtained by solving

$$\sum_{j=1}^{k} \frac{j}{(k-j)f+j}\binom{k}{j}p^j(1-p)^{k-j} = \frac{D}{n}. \qquad (6)$$

Equation (6) seems too tedious to get solutions for (p,f) by mathematical expansions. Therefore, for $j \geq 1$, a modified dilution effect model $1-(k-j)f$ is proposed in this paper without a loss of the Hung and Swallow's (1999) basic concept in their first dilution effect model, where the value of f differs and $f \leq 1/(k-j)$ satisfying the probability property of probability being greater. Model (3) is a logarithmically increasing function of j, while the proposed dilution model in this paper is a linear increasing function of j. Now, Equation (6) can be written as

$$1-k\hat{f}(1-\hat{p})-(1-k\hat{f})(1-\hat{p})^k = \frac{D}{n}. \qquad (7)$$

No dilution effect, $f=0$ makes equation (7) the usual group testing equation and gives the MLE as equation (1). When a dilution effect exists, two parameters

(p, f) have to be estimated with one equation. For having one more equation, it is assumed that retesting is feasible, but it would be done only once. When retesting runs, the only individuals of non-infected groups should be retested and the smaller group size have to used than the original group size k to detect the dilution effect. According to retesting scheme, classification procedures in group testing may be divided into non-Dorfman procedures which test the units of an infected group one-by-one and Dorfman procedures which divide infected groups into subgroups for retesting. Halving procedures, which repeatedly subdivide unclassified groups into two equal size groups, are shown to be nearly optimal and convenient (Chen and Swallow, 1990). Thus in retesting experiment, the number of test groups is equal to $2(n-D)$ and the group size k_2 is $k/2$. Let D_2 be the number of infected groups in retesting. The following equation can be obtained in the retesting experiment, similar to equation (7):

$$1-k_2\hat{f}(1-\hat{p})-(1-k_2\hat{f})(1-\hat{p})^{k_2}=\frac{D_2}{2(n-D)}. \tag{8}$$

Thus the estimators for (p, f) can be obtained by solving equation (7) and equation (8) simultaneously. From equation (7), the following expression can be obtained:

$$\hat{f}=\frac{(1-D/n)-(1-\hat{p})^k}{k(1-\hat{p})(1-(1-\hat{p})^{k-1})}. \tag{9}$$

Now, the following equation can be obtained to compute the estimator $\hat{p}$:

$$(1-\frac{D_2}{2(n-D)})=(1-\hat{p})^{k_2}+k_2\frac{(1-D/n)-(1-\hat{p})^k}{k(1-\hat{p})(1-(1-\hat{p})^{k-1})}(1-\hat{p})(1-(1-\hat{p})^{k_2-1}). \tag{10}$$

The estimator for the infection rate $\hat{p}$ can be obtained from equation (10) by the numerical substitution method. The following example shows how one can obtain the estimators of $\hat{p}$ when a dilution effect exist and the dilution effect model is assumed to be $1-(k-j)f$. Let the population units be iid as $Bernoulli(p=0.05)$ where p is the infection rate and the number of testing groups n be 30. Then the optimal group size k is 20 from Swallow's table (1985). The number of infected groups was 19 from a simulated data using RANTBL function of SAS with seed=1. Suppose a dilution effect exists and two infected groups out of 19 infected groups tests non-infected. Then D is 17. The 13 non-infected groups would be tested in the retesting experiment, and divided into 26 test groups with the halving procedure. If 9 groups out of 26 groups test infected in the retesting, the estimator $\hat{p}$ is 0.042 from equation (10).

Since parameter f is the power of exponential form in model (4) which was proposed by Hung and Swallow (1999), the mathematical expansion seems very unlikely. Three parameters in model (4) make retesting on non-infected groups of the previous testing run twice. Therefore, model (4) is not a feasible model for estimating p when a dilution effect exists. Instead, a dilution effect model with a threshold is proposed here in simpler form. The existence of dilution effects $P(x_j = 1 | m_i = j)$ depends on the volume of k, j/k, or $(k-j)$. It seems likely in the real fields that we know the maximum value k of which does not make any dilution effect. For example, in the screening for HIV, it has been shown that the maximum value is 15 as mentioned (Emmanuel et al., 1988; Cahoon-Young et al., 1989; Kline et al., 1989; Monzon et al., 1992). Let k_0 be the maximum group size of which does not make any dilution effect. Then a dilution effect model can be suggested as follows, where a dilution effect is a function of k and j:

$$P(x_i = 1 | m_i = j) = I_{(k \le k_0)} + I_{(k > k_0)} g(k, j) \text{ for } j \ge 1. \tag{11}$$

Two testing procedures, the usual group testing (the first stage of testing experiment) and retesting experiment can be used to estimate the infection rate and the dilution effect as before. Notations are the same as the previous. The probability of x_i being 0 (not infected) can be written

$$\begin{aligned} P(x_i = 0) &= P(x_i = 0, m_i = 0) + P(x_i = 0, m_i \geqq 1) \\ &= (1-p)^s + P(x_i = 0) P(m_i \geqq 1 | x_i = 0). \end{aligned} \tag{12}$$

The second term in equation (12), $\Pr(x_i = 0, m_i \geqq 1)$ can be considered a dilution effect if it exists. The existence of dilution effects depends on the volume of (j/k) where j is the number of defective units in the tested group. If the optimal group size at the first stage of testing experiment k is not greater than k_0, there is no dilution effect and the usual group testing can be applied to estimate the infection rate. When k is greater than $2k_0$, the group size at the first stage is recommended to be $2k_0$ instead of the optimal group size from Swallow's table (1985) to assure that there is no dilution effect at the second stage if halving retesting procedure is used. If k is an odd number, $(k-1)$ should be the group size at the first stage to use halving retesting procedure, for the smaller k reduces the MSE of the estimator.

When the optimal group size k at the first stage is in $[k_0 + 1, 2k_0]$ and even number, halving procedures in retesting can be applied to the groups which are tested as non-infected at the first stage for estimating the dilution effect. In retesting, the non-infected group at the first stage should be divided into two testing groups with the size of $k/2$. The number of tested groups in retesting is $2(n-D)$. When the number of infected groups in resting D_2 is 0, there is no dilution effect and equation (12) reduces to the usual group testing model, which means that we gain nothing by retesting. The last term in equation (12)

$\Pr(m_i \geq 1 | x_i = 0)$ can be considered as a dilution effect and estimated by estimating the infection rate in retesting. Therefore, the estimator for the infection rate can be obtained as

$$\hat{p} = 1 - ((1 - \frac{D}{n})(\frac{D_2}{2(n-D)}))^{1/k}. \tag{13}$$

For example, suppose the number of infected groups at the first stage D is 10 with the same population assumption as the previous example. The smaller value of D is the more resonable outcome for the last dilution effect model (11). If 19 groups out of 40 test groups test infected in the retesting experiment with the halving procedure, the estimator $\hat{p}$ is 0.056 from equation (13).

4. Conclusion and Discussion

The group testing is more efficient than one-by-one testing in estimating problem and classification problem when the test outcomes of units are dichotomous and the probability of being success is small. In classical approaches to estimate the infection rate, no false negative (dilution effect) is assumed, but dilution problem happens often in many practical area. Hung and Swallow (1999) discussed dilution effects and the selection of group size by two proposed dilution effect models, but did not touch how to estimate the infection rate. In this paper, two modified dilution effect models are proposed and estimation methods for the population infection rate are discussed when a dilution effect exists. And examples are given to show how to get the estimator $\hat{p}$ with two modified dilution effect models. Halving retesting procedure is used for the retesting scheme. We can extend the last proposed model to the case that the group size of the first stage is greater two times of the maximum value of defected units in a group that does not make any dilution effect somewhat with more than two retestings.

References

[1] Chen, C.L. and Swallow, W.H (1995). Sensitivity Analysis of variable-size group testing and its related models. *Biometrical Journal,* Vol. 37, 173 -181.

[2] Cahoon-Young, B., Chandler, A., Livermore, T., Caudino, J., and Benjamin, R. (1989). Sensitivity and specificity of pooled versus individual sera in a HIV-antibody prevalence study. *Journal of Clinical Microbiology,* Vol. 27, 1893-1895.

[3] Dorfman, R. (1943). The detection of defective members of large population. *Annals of Mathematical Statistics,* Vol. 14, 436-440.

[4] Gastwirth, J.L. and Hammick, P.A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of AIDS antibodies in blood donors. *Journal of Statistical Planning and Inference,* Vol. 22, 15-27.

[5] Emmanuel, J.C. and Bassett, M.T. (1988). Pooling of sera for HIV testing: An economical method for use in developing countries. *Journal of Clinical Pathology,* Vol. 41, 582-585.

[6] Hung, M. and Sallow, W.H. (1999). Robustness of Group Testing in the Estimation of Proportions. *Biometrics,* Vol. 55, 231-237.

[7] Kline, R.L., Brithers, T.A., Brookmeyer, R., Zegar, S., and Quinn, T.C. (1989). Evaluation of HIV seroprevalence in population surveys using pooled sera. *Journal of Clinical Microbiology,* Vol. 27, 1449-1452.

[8] Litvak, E., Tu, X.M., and Pagano, M. (1994). Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association,* Vol. 89, 424-434.

[9] Monzon, O.T., Paladin, F.J.E., Dimaandal, E., Balis, A.M., Samson, C., and Mitxhell, S. (1992). Relevance of antibody content and test format in HIV testing of pooled sera. *AIDS,* Vol. 6, 43-48.

[10] Swallow, W.H. (1985). Group testing for estimation infection rates and probabilities of diseases transmission. *Phytopathology,* Vol. 75, 1376 -1381.

[11] Thompson, K.H. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics,* Vol. 18, 568-578.

[12] Tu, X.M., Litvak, E., and Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika,* Vol. 82, 287-297.

[Received September 2006, Accepted October 2006]