The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 619-629

Bayes Estimators in Group Testing

Sehyug Kwon[1)]

Abstract

Binomial group testing or composite sampling is often used to estimate the proportion, p, of positive(infects, defectives) in a population when that proportion is known to be small; the potential benefits of group testing over one-at-a-time testing are well documented. The literature has focused on maximum likelihood estimation. We provide two Bayes estimators and compare them with the MLE. The first of our Bayes estimators uses an uninformative Uniform(0, 1) prior on p; the properties of this estimator are poor. Our second Bayes estimator uses a much more informative prior that recognizes and takes into account key aspects of the group testing context. This estimator compares very favorably with the MSE, having substantially lower mean squared errors in all of the wide range of cases we considered. The priors uses a Beta distribution, $Beta(\alpha, \beta)$, and some advice is provided for choosing the parameter α and β for that distribution.

Keywords : Group testing, Bayes, MLE, Beta prior, Uniform prior

1. Introduction

Binomial random variables are of interest in many applications: plants may or may not be virus-infected, individuals may or may not be HIV seropositive, or a product may or may not be defective (see, e.g., Sobel and Groll, 1959; Gibbs and Gower, 1960; Thompson, 1962; Emmanual et al., 1988; Calhoon-Young et al., 1989; Litvak, Tu, and Pagano, 1994; Tu, Litvak, and Pagano, 1994, 1995). The proportion p of positives (infecteds, defectives, etc.) in the population or, equivalently the probability that an individual is positive, is often of specific interest.

When p is small, group testing (or composite sampling) has been shown to be much more efficient than one-at-a-time testing of samples in minimizing the mean squared error (MSE) of $\hat{p}$, the maximum likelihood estimator (MLE) of p (Gibbs and Gower, 1960; Thompson, 1962;

1) Professor, Department of Statistics, HANNAM University, Daejon 305-761, KOREA,
E-mail: wolfpack@hannam.ac.kr

Sobel and Elashoff, 1975; Swallow, 1985). With group testing, k individual samples are combined, and a single test is done on the pooled sample. If the pooled sample tests negative, then all of the k individual samples are taken to be negative; if the pooled sample tests positive, then at least one of the individual samples is presumed to be positive, but we do not know which one(s) or how many. If identification of the positive individual(s) is required, then one must implement some scheme for retesting individuals in positive groups. However, Chen and Swallow (1990) have shown that when one's interest is confined to estimating the overall p, retesting is of little, if any, value; accordingly, retesting schemes are not considered further in this paper. In some applications, retesting is not feasible anyway.

In using group testing, it is critical that an appropriate group size is used. Unfortunately, the optimal group size depends on the (unknown) proportion one is trying to estimate, as well as on the number n of groups to be tested. Using a group size that is much larger than the optimum size can lead to large bias and, thereby, inflated MSE of $\hat{p}$, MLE. It has been recommended that the appropriate group size would be for $p=p_u$, where p_u is a value believed to be an upper bound for p. This strategy will lead to using a smaller-than-optimal group size for the true p, but the user will realize most of the benefits of group testing (Swallow, 1985). Swallow (1985) and others give tables of optimal group size for various combinations of the true proportion and the number of groups to be tested.

There may be additional constraints on how large the group size k can be in particular contexts. For example, in HIV screening with enzyme-linked immunosorbent assay (ELISA) tests, one may want to limit k to, say, 15 or fewer to allay concerns about false negatives arising as a result of dilution effects (Tu et al., 1995).

Virtually all of the group-testing literature related to estimation have focused on the maximum likelihood estimator (MLE) of p. Although Kumar and Sobel (1975) considered Bayesian approaches to minimizing the expected number of tests required to classify every individual as positive or negative using group testing, Bayesian estimation of p remains to be explored.

In this paper, we consider Bayesian estimation of p under the squared error loss function, and compare two Bayes estimators with the MLE by computing their MSE's. We discuss the underlying model and the MLE in Section 2, and develop our Bayes estimators in Section 3. In Section 4, we compare the MLE and Bayes estimators over a range of conditions, i.e., for a number combinations of n, the number of test groups, p and k through Monte Carlo simulation. And in Section 5, we provide a discussion and conclusion.

2. The Assumed Model and the MLE for p

The binomial group testing model makes the following assumptions:

(i)A fraction p of the population is positive for the trait of interest, and the positives are randomly distributed throughout the population. Individual samples can be viewed as independent, identically distributed (iid) Bernoulli random variables; when groups are formed, a binomial model applies. In applications where positives are encountered in clusters, for example, this assumption would be violated.

(ii)The number n of groups to be tested is fixed in advance. The optimal choice of group size k depends in part on n.

(iii)The same group size k is used for each of the n groups. Group size is usually easily controlled in practice. Walter, Hildreth, and Beaty (1980) and Le (1981) investigated group testing with unequal group size, and, of course, the problem is then far messier.

(iv)Classification of group as positive or negative is without error. In group testing applications, the possibility of false negatives as a result of dilution effects must always be considered. For more on dilution effects and misclassification errors, see, e.g., Hwang (1984), Chen and Swallow (1990, 1995), and Hung and Swallow (1999).

(v)Members of positive groups will not be retested.

Let $D=\sum_{i=1}^{n} d_i$ be the number of positive (infected or defective) groups in the observed data, where d_i is the test result for the ith group with d_i=1 (infected) or 0 (non-infected) for $i = 1, 2, \ldots, n$. Then D is distributed Binomial $(n, 1-(1-p)^k)$. The MLE for p is

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

By Jensen's inequality, $E(\hat{p}) \geq p$. So, for $k > 1$ the MLE is not an unbiased estimator, but instead overestimates p. Thompson (1962) noted that the bias will be small as long as p and k are small, and group testing is most advantageous when p is small.

3. Bayes Estimators

Suppose an observation vector $\underline{d} = (d_1, d_2, \ldots, d_n)$ is a random sample from $f(x|p)$ for $p \in \Theta$. In group testing, p is a single unknown parameter and the parameter space Θ is [0,1]. We can specify a prior distribution $\pi(p)$ for p to get a Bayes estimator. The posterior distribution of p is then

$$\pi(p|\underline{d}) \propto \pi(p) l(p|\underline{d})$$

where $l(p \mid \underline{d})$ is the likelihood function. Under the squared error loss function, the posterior

mean $E(p \mid d_1, d_2, \ldots, d_n)$ is the Bayes estimator (Casella and Berger, 1990). Moreover, $E(g(p)|d_1, d_2, \ldots, d_n)$ is the Bayes estimator for $g(p)$, a function of p (Bickel and Doksum, 1977). In group testing, we can put the prior on p before choosing the group size k or on $p_0 = 1-(1-p)^k$ after choosing it. For a particular experiment, prior information (a reasonable upper bound) on p is useful in choosing the group size anyway. Therefore, to put the prior on p rather than on p_0 may seem more natural. With the prior distribution on p, the distribution of p_0 can be obtained by variable transformation.

3.1 Bayes Estimator 1

The uniform distribution on the parameter space $[0, 1]$ can be considered as an non-informative prior distribution $\pi(p)$ for p since the infection rate must be in $[0, 1]$. Then, the prior distribution of p_0 is

$$\pi(p_0) = \frac{1}{k}(1-p_0)^{1/k-1} \text{ for } 0 \le p_0 \le 1$$

Therefore, the joint distribution of $(d_1, d_2, \ldots, d_n; p)$ is

$$f(d_1, d_2, \ldots, d_n; p) = \frac{1}{k}(1-p_0)^{1/k-1}(p_0)^D(1-p_0)^{n-D}$$

and the joint distribution of $(d_1, d_2, \ldots, d_n)$ is

$$f(d_1, d_2, \ldots, d_n) = \frac{1}{k}Beta(D+1, n+\frac{1}{k}-D)$$

where $Beta(a;b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ with $B(.,.)$ and $\Gamma(.)$ denoting beta and gamma functions, respectively. Then, the posterior distribution $\pi(p_0 \mid d_1, d_2, \ldots, d_n)$ is

$$\frac{p_0^D(1-p_0)^{1/k+n-D-1}}{B(D+1; n+1/k-D)}.$$

For the squared error loss function criteria, the Bayes estimator for p is then

$$1 - \frac{\Gamma(2/k+n-D)\Gamma(1/k+n+1)}{\Gamma(2/k+n+1)\Gamma(1/k+n-D)}. \tag{2}$$

3.2 Bayes Estimator 2

Because group testing is most useful when p is small and should be considered only in such cases, the distribution of p seen in group testing applications is strongly skewed to the right, with p almost certainly < 0.3 and usually < 0.05. This suggests using a prior distribution that reflects this skewness. There is a one-to-one relationship between p and p_0, with the distribution of p_0 also being skewed to the right. For mathematical simplicity and the conjugate property, we consider using a Beta distribution with parameters (α, β), Beta (α, β), as a prior; beta distributions are also appealing in that they can be made to assume a great variety of shapes by the choice of the two parameters, α and β. Although the beta prior can be placed on either p or p_0, it is mathematically easier in this case to put the prior on p_0.

When $\pi(p_0)$ is $Beta(\alpha, \beta)$ distribution, the posterior distribution function $\pi(p_0 \mid d_1, d_{2,\ldots,}d_n)$ is

$$\frac{p_0^{\alpha+D-1}(1-p_0)^{\beta+n-D-1}}{B(\alpha+D;\beta+n-D)}.$$

For the squared error loss function criterion, the Bayes estimator for p is then

$$1-\frac{\Gamma(\alpha+\beta+n)\Gamma(\beta+1/k+n-D)}{\Gamma(\beta+n-D)\Gamma(\alpha+\beta+1/k+n)}. \tag{3}$$

How does one choose (α, β) for the prior distribution, $\pi(p_0)$? When $\alpha = \beta$ and both are greater than 1, the pdf of $Beta(\alpha, \beta)$ is symmetric. The pdf becomes more skewed to the right as α increases and is more skewed to the left as β increases. Since p_0 is a function of p and k, the shape of the distribution function of p_0 depends on the distribution of p and k. The skewness of the prior distribution of p to reflect the practical parameter space of [0, 0.3] makes the distribution of p_0 also skewed. As k becomes larger, the skewness of p_0 goes from right-skewed to left-skewed. Thus, we may choose (α,β) as follows: $\beta > \alpha > 1$ for small k and $\alpha > \beta > 1$ for large k. The distribution of p_0 depends primarily on the distribution of the true p, and the group size k is chosen with the value of (n, p) in mind. Therefore, (n, p) should be simultaneously considered in selecting reasonable (α, β).

4. Comparison of the MLE and Two Bayes Estimators

For comparing the MLE and our two Bayes estimators, we proceed as follows:

(i)We consider all combination of $n = 10, 20,$ and 30 with p=0.01, 0.02, 0.05, 0.1, 0.2. For each combination (n, p), the group size k is chosen by Swallow's(1985) Table1.

(ii)Generate $(n \times k)$ individuals that are *iid* distributed Bernoulli(p) and form n test groups.

(iii)Calculate the MLE and Bayes estimators by formulaee, (1), (2), and (3). The Bayes estimators of formula (2) and (3) are called Bayes 1 and Bayes 2, respectively.

(iv)Repeat steps (ii) and (iii) $5,000$ times for each (n,p) and by the bootstrap get $E(\hat{p})$, Bias$(\hat{p})$ and MSE$(\hat{p})$.

In the simulations, we use the Beta(2,3) distribution as the prior distribution $\pi(p_0)$ in computing Bayes 2. This prior is slightly skewed to the right; the distribution of p_0 is then more severely skewed to the right. We exclude the case $D=n$ in computing, because, in that case, the MLE of p is equal to 1. The simulation results are summarized in Table 1.

Because all three of these estimators are biased, they are compared through their mean squared errors (MSE's). In terms of MSE, the Bayes 2 estimator is by far the best in all cases and the Bayes 1 is worst in all cases. Because the $Uniform(0,1)$ prior for Bayes 1 fails to reflect the fact that p will be small in group testing, it is to be expected that Bayes 1 will do poorly. In terms of the bias, $E(\hat{p}-p)$, Bayes 1 is also consistently the worst. The MLE always overestimates p, reflecting its known positive bias. More often than not, the bias in the MLE is smaller than in the Bayes 2 estimator, but their biases don't differ greatly. This is specially true for small values of p, which are those that are likely to prevail in applications. When p is small, lower bounds for the MLE can be negative where is out of parameter space. It makes the MLE better than Bayes 2 in some small p. Moreover, Bayes 2 depends on the prior selected for $\pi(p_0)$. Bayes 2 as shown in Table 1 took $\pi(p_0) \sim B(2,3)$. Other choices of α and β could improve the performance of Bayes 2 in particular cases, i.e., under particular values of (n,p,k), and an experimenter typically has some advance information on (n,p,k). Usually the experimenter has a ballpark guess or a prior upper bound for p in mind; n may well be known approximate or, in some cases, set by the experimenter; an appropriate choice of k depends on n and p, as illustrated in Swallow's Table 1 (1985).

Figures 1-3 illustrate how changes in α and β affect Bayes 2. Each figure shows all three estimators, MLE Bayes 1 and Bayes 2, which we computed by formulae (1), (2) and (3) for $D=0,1,\ldots,(n-1)$. These are exact, not simulated results. Values of k from Swallow's (1985) Table 1 were used. The values of d are on the horizontal axis, and $\hat{p}$ on the vertical axis. The horizontal line in each plot is at the true value of p. In each of Figures 1 and 2, two sub-figures show the effect of increasing the value of β in $Beta(\alpha,\beta)$ from $\beta=2$ to $\beta=5$. The MLE and Bayes 1 depend on (n,p,k), but do not depend on α and β, which addresses the skewness of the prior distribution of p. In both Figure 1 and 2, increasing the value of β from 2 to 5 is shown to reduce the bias in Bayes 2, substantially reducing it for larger values of d(shown on the horizontal axis). In figure 3, the sub-figures illustrate the

effect of increasing the value of α in $Beta(\alpha, \beta)$ from 2 to 3 in a particular case. Again, the bias of Bayes 2 is reduced, although only slightly. Section 5 provides some discussion on selecting values of α and β that may be more helpful to the practitioner.

Table 1: Simulation Results

		MLE		Bayes1		Bayes2	
n	p	$E(\hat{p})$	MSE$(\hat{p})$	$E(\hat{p})$	MSE $(\hat{p})$	$E(\hat{p})$	MSE $(\hat{p})$
10	0.01	0.0106	0.00004	0.0140	0.00006	0.0122	0.00002
	0.02	0.0214	0.00015	0.0277	0.00022	0.0237	0.00008
	0.05	0.0527	0.00269	0.0654	0.01539	0.0548	0.00059
	0.10	0.1064	0.00273	0.1264	0.00347	0.1054	0.00121
	0.20	0.2045	0.00796	0.2290	0.00821	0.1906	0.00323
20	0.01	0.0104	0.00002	0.0118	0.00002	0.0106	0.00001
	0.02	0.0210	0.00005	0.0232	0.00006	0.0200	0.00003
	0.05	0.0523	0.00026	0.0570	0.00033	0.0485	0.00015
	0.10	0.1031	0.00092	0.1112	0.00110	0.0955	0.00053
	0.20	0.2059	0.00326	0.2174	0.00353	0.1891	0.00192
30	0.01	0.0105	0.000010	0.0114	0.000012	0.0106	0.000008
	0.02	0.0204	0.000023	0.0216	0.000027	0.0192	0.000016
	0.05	0.0517	0.000161	0.0546	0.000194	0.0478	0.000103
	0.10	0.1029	0.000557	0.1080	0.000658	0.0949	0.000361
	0.20	0.2062	0.001925	0.2137	0.002126	0.1892	0.001274

MLE: Maximum likelihood estimator

Bayes 1: Bayes estimator with $\pi(p)$ $Uniform[0,1]$

Bayes 2: Bayes estimator with $\pi(p_0)$ $Beta(2,3)$

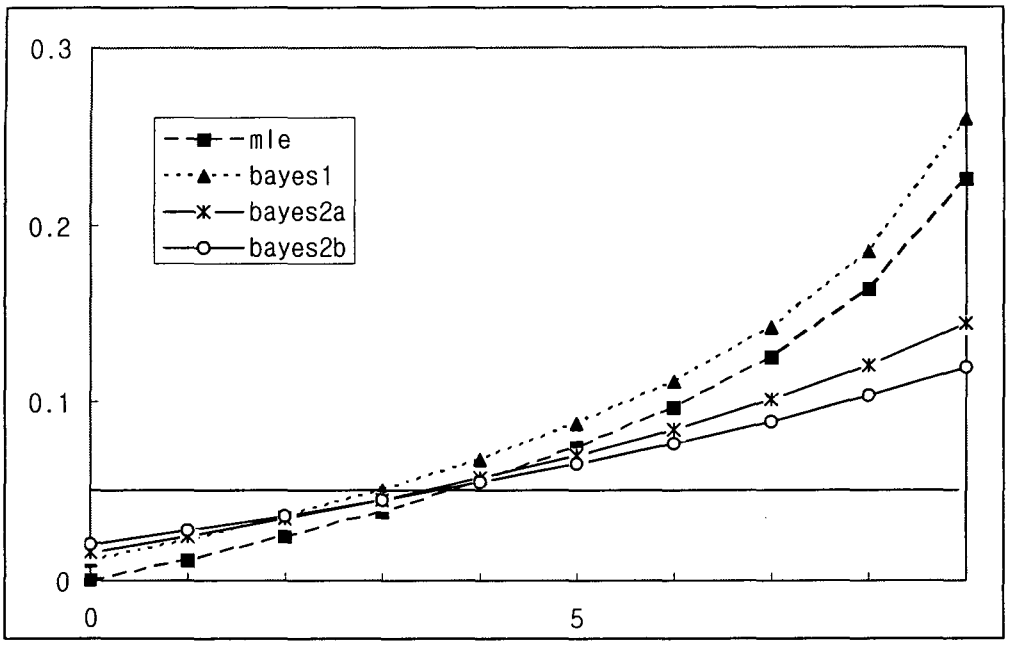

Figure 1. $(n, p, k) = (10, 0.05, 9)$, $Bayes2a_{(\alpha=2,\beta=3)}$, $Bayes2b_{(\alpha=3,\beta=5)}$

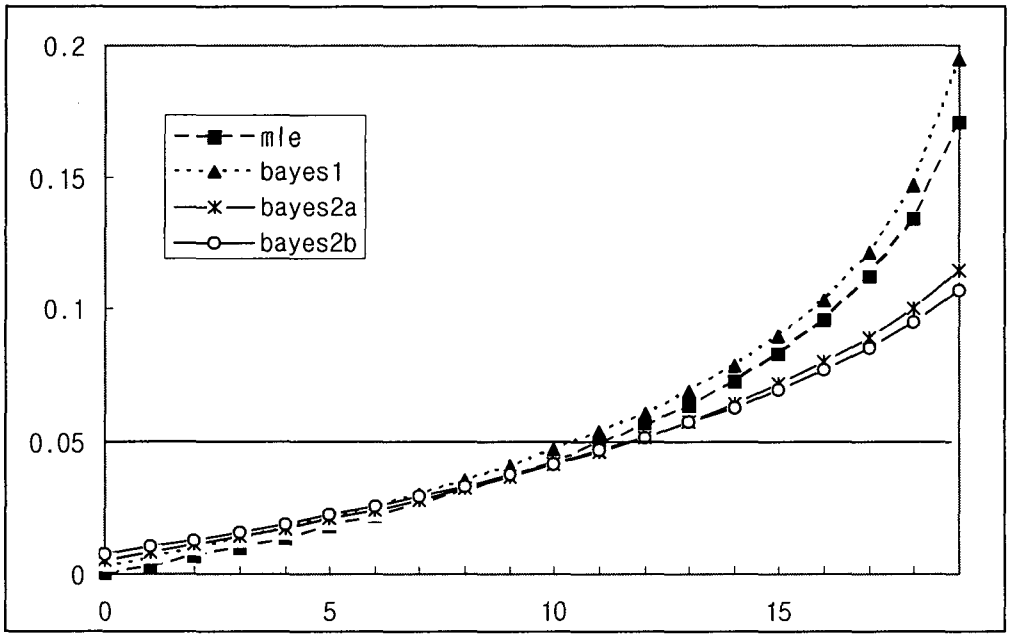

Figure 2. $(n, p, k) = (20, 0.05, 16)$, $Bayes2a_{(\alpha=2,\beta=3)}$, $Bayes2b_{(\alpha=3,\beta=3.8)}$

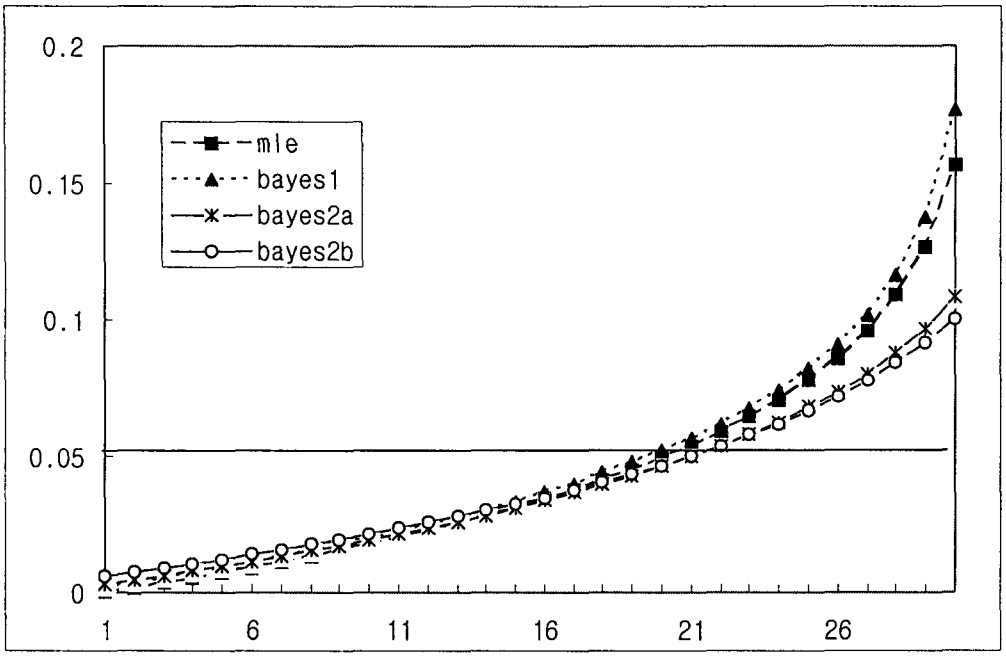

Figure 3. $(n, p, k) = (30, 0.05, 20)$, $Bayes2a_{(\alpha=2,\beta=3)}$, $Bayes2b_{(\alpha=4,\beta=4)}$

5. Discussion and Conclusion

In the binomial group testing (or composite sampling) literature, the focus has been strongly on maximum likelihood approaches to the estimation of p, the proportion of infected or defective individuals in the population. We proposed here two Bayes estimators for p. The prior distribution can be placed on either p or $p_0 = 1-(1-p)^k$, where k is the number of individuals to be pooled in each group. Then, p_0 is the proportion of defective groups, i.e., groups containing 1 or more defective or infected individuals. Whether the prior is placed on p or p_0 is simply a matter of convenience, as there is a one-to-one relationship between p and p_0. Our Bayes 1 estimator places an uninformative $Uniform(0,1)$ prior on p. This estimator is shown to perform poorly. The Bayes 2 estimator places the prior on p_0, and makes better use of our knowledge of the group testing context. In group testing, the value of p will almost surely be less than 0.1 and probably less than 0.05. The prior distribution for p and thus for p_0 should be skewed to reflect this. To accomplish this, we use a beta distribution, $Beta(\alpha, \beta)$, for which the prior mean and variance of p_0 are

$$E(P_0) = \frac{\alpha}{\alpha+\beta} \text{ and } V(P_0) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

respectively. Because p_0 is the proportion of positive (infected, defective) group, $E(p_0) = \frac{\alpha}{\alpha+\beta}$ and $(\alpha+\beta)$suggests thinking of α as being related to the proportion of "positives," and thus β as being related to the proportion of "negative". For example, suppose the population proportion of positive individuals p is 0.1 and one planned to test n=10 groups. Swallow's Table 1 (1985) suggests using the optimal group size, $k=5$ for $(n,p)=(10,0.1)$. Then, $p_0 = 1-(1-p)^k$ equals approximately $p_0 = 0.4$. With $n=10$, we might set $\alpha = np_0 = 4$ and $\beta = 6$. For $(\alpha,\beta)=(4,6)$ in Bayes 2, $E(\hat{p}) = 0.1028$ and $MSE(\hat{p}) = 0.00063$, respectively. Both the bias and MSE with $(\alpha,\beta)=(4,6)$ are smaller than found with $(\alpha,\beta)=(2,3)$ in Table 1 where $E(\hat{p}) = 0.1054$ and MSE$(\hat{p})= 0.00121$ for this same case. Further discussion of the assessment of a beta prior distribution can be found by Chaloner and Duncan (1983).

The overall conclusions are that Bayes estimators can compare very favorably with the MSE, provided an informative prior is used. Table 1 compared these biased estimators through their MSE's and showed that our Bayes estimator 2 had the smaller MSE over a wide range of situations using a $Beta(\alpha=2,\beta=3)$ prior for p_0 throughout. We further showed that the Bayes 2 can be improved by selecting (α,β) in $Beta(\alpha,\beta)$ more carefully in a particular case.

References

[1] Bickel, P.L. and Doksum, K.A.(1997). Mathematical Statistics. Holden-Day Inc.,419.

[2] Casella, G. and Berger, R.L.(1990). Statistical Inference. Wadworth and Brooks/Cole, 475.

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

[4] Chaloner, K. M. and Duncan, G. T. (1983). Assessment of a Beta prior distribution: PM elicitation. The Statistician, **32** 174-180.

[5] Chen, C. L. and Swallow, W. H. (1990). Using group testing to estimate a proportion and to test the binomial model. Biometrics **46**, 1035-1046

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

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

[8] Gibbs, A. J. and Gower, J. C. (1960). The use of a multiple-tansfer method in plant virus transmission studies-Some statistical points arising in the analysis of results.

Annals of Applied Biology **48**, 75-83.

[9] Hung, M. and Swallow, W. H. (1999). Robustness of group testing in the estimation of proportions. Biometrics **55**, 231-237.

[10] Hwang, F. K. (1984). Robust group testing. Journal of Qualitative Technology **89**, 189-195.

[11] Kumar, S. and Sobel M. (1975). An asymtotocally optimal Bayes solution for group testing. In A Survey of Statistical Design and Latin Models, J. N. Srivastava (ed.), 367-381. Amsterdam: North Holland.

[12] Le, C. T. (1981). A new estimator for infection rates using pools of variable size. American journal of Epidemiology **114**, 132-136.

[13] 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 **89**, 424-34.

[14] Sobel, M. and Elasfoff, R. M. (1975). Group testing with a new goal, estimation. Biometrika **62**, 1179-1252.

[15] Sobel, M. and Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. The Bell System Technical Journal 38, 1179-1252.

[16] Swallow, W. H. (1985). Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology **75**, 882-889.

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

[18] Tu, X. M., Litvak, E., and Pagano, M. (1994). Screening tests: Can we get more by doing less. Statistics in Medicine **13**, 1905-1919.

[19] _____(1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. Biometrika **82**, 287-297.

[20] Walter, S. D., Hildreth, S. W., and Beaty, B. J. (1980). Estimation of infection rates in populations of organisms using pools of various size. Journal of Epidemiology **112**, 124-128.

[Received July 2004, Accepted October 2004]