The Korean Communications
in Statistics Vol. 4, No. 1, 1997
pp. 317-325

Upper Bounds for the Infection Rate in Group Testing[1)]

Sehyug Kwon[2)]

Abstract

Group testing is an efficient method to classify units from a population as infected or non-infected and useful in estimating the infection rate when the population infection rate is small. Upper bounds are the focus of interest in group testing, but has not been studied extensively. In this paper, the upper bound derived from the uniformly most powerful test is proposed and compared with the classical approaches, Thompson's and Bhattacharyya et al.'s methods.

1. Introduction

When few of the units are infected, one-by-one test is not likely to be efficient to classify units whose outcome is dichotomous. Dichotomous (or binary) means that the test result of units can take only one of two possible values, infected or non-infected. It is preferable to form groups of units and simultaneously test units in a group, instead of testing each unit separately, which is called group testing. If the test outcome of a group is not infected, we conclude that there is no infected unit in the group. When infected, we say that at least one unit is infected. Unfortunately, we do not know which one is infected or how many are infected. Therefore, some retesting scheme will be needed to identify the infected units in the population under group testing.

Dorfman (1943) suggested the idea of grouping to identify efficiently all syphilitic men in a population by Wasserman-type blood testing. It is called the classification problem, one of major interests in group testing. The other is the estimation problem. He recommended that, after individual blood samples had been drawn, portions of each be pooled in groups of equal size k and the groups or pools tested. For infected groups, he suggested retesting individuals one-by-one to identify the infected individuals. He showed that when the probability of an individual being infected is smaller than 0.3, his suggested procedure was better than one-at-a-time testing in the sense of the expected number of tests to classify all units as infected or not. Ungar (1960) showed that whenever the population infection rate (p) is in the

1) The paper was supported by the research grant of Hannam University.
2) Assistant Professor, Department of Applied Statistics, Hannam University, Taejon 300-791, Korea

interval $[0, \frac{(3-5)^{1/2}}{2}]$, group testing scheme is more efficient than one-at-a-time testing in the sense that the expected number of tests required to classify all of the units as infected or not is smaller, which has been the usual goal in the classification problem.

According to retesting scheme, classification procedures in group testing may be divided into non-Dorfman procedures and Dorfman procedures. Dorfman procedures test the units of an infected group one-by-one. Non-Dorfman procedures divide infected groups into subgroups for retesting. This step is repeated until all units in the population are classified. Strategies for subgrouping have been studied extensively. Sobel and Groll (1959) considered a procedure which describes a mode of action for any given value of (1-p) in a binomial sample. Finucan (1964) considered dividing defective groups into more than two subgroups and testing them separately. Chen and Swallow (1990) showed that the procedure that repeatedly subdivide unclassified groups into two equal-size groups are nearly optimal and very convenient.

The estimation problem was first considered by Thompson (1962) who used the idea of group testing to estimate the proportion of six-spotted leafhoppers capable of transmitting aster-yellows virus using the maximum likelihood estimator (mle). Bhattacharyya, Karandinos and DeFoliart (1979) derived the mle when the population is finite. Davis, Grizzle and Bryan (1975) explored the multivariate case to estimate the probability of contracting hepatitis. Sobel and Elashooff (1975) derived the mle when units can be retested. Chen and Swallow (1990) showed that retesting, even if feasible, can increase the precision of the mle only very slightly in sense of mean squared error.

Because group testing is most attractive when p is small, the upper bounds for p should be the focus of interest, but have not been studied widely. Thompson(1962) gave an approximate confidence interval for p and Bhattacharyya, Karandinos and DeFoliart (1979) developed a method for both hypothesis testing and confidence interval estimation. Both approaches (classical approaches) are based on the asymptotic normality of the mle of p and derived two sided confidence intervals, but did not mention upper bounds which are more useful in group testing. In this paper, upper bounds for p are derived based on the uniformly most powerful test and shown to have a number of desirable properties.

2. Assumptions and the mle of p

The following assumptions are made here:

(1) The infected units are randomly distributed as Bernoulli with the true p.

(2) The number of groups (n) is fixed in advance and costs are not considered.

(3) The group size (k) for each of the n groups is the same.

(4) Units are not to be retested.

(5) Classification (testing) of groups is without error.

Assumption (1) is fundamental in group testing. Assumption (2) and (3) are made for

simplicity. Group size can usually be controlled in practice. For the fixed n case, Swallow (1985) constructed a table of optimal group sizes for various combinations of (n, p) by min mizing the mean squared error of the maximum likelihood estimator (mle) of p.

Assumption (4) is made on the grounds that retesting would only negligibly reduce the mean squared error of the estimator as mentioned in Chen and Swallow (1990). Assumption (5) implies that there are no false negative and false positive effects. False negative effects mean that even though one or more infected units in a group to be tested are present, the test fails to detect their presence. False positive effects make us to classify incorrectly a non-infected group as infective.

Suppose that the population infection rate is p, the optimal group size is k and the number of group of size k to be tested is n. Let X_i be the test outcome for i-th group. The outcome of X_i is either infected, the value of 1 with p or non-infected, the value of 0 with (1-p). The probability distribution function of X_i is

$$f(x_i)=[1-(1-p)^k]^{x_i}[(1-p)^k]^{1-x_i} \quad \text{where } x_i=0,\ 1.$$

Therefore, the mle of p is $\hat{p}=1-(1-\sum_{i=1}^{n} X_i/n)^{1/k}$.

Since $E(\hat{p})=1-E[(1-\sum_{i=1}^{n} X_i/n)^{1/k}] \geq 1-[E(1-\sum_{i=1}^{n} X_i/n)]^{1/k} = p$, for the group size $k > 1$, the mle is not an unbiased estimator, but overestimates p. However, Thompson noted that the bias will be small as long as p and k are small where group testing is more efficient than one-at-a-time testing.

For computing the mle ($\hat{p}$), the group size (k) should be obtained in advance. The optimal group size has been taken from Swallow (1985)'s table and others that have given tables of optimal group sizes for combinations of the (n, p). Because it is often assumed that when the number of groups (n) is fixed, choosing the optimal value for k depends on knowing the value of the parameter (p), but it is unrealistic. In practice, although p is unknown, there sometimes is at least some prior information about the value of p from preliminary data or other considerations. At least an upper bound on p may be available and, if so, it has been suggested that it should be used in choosing k (see, for example, Swallow, 1985). Taking the lower bound of p (i.e., underestimating p) in choosing k causes us to use too large a value of k, increasing the bias in $\hat{p}$ and inflating its mean squared error, which was shown by Swallow (1985).

3. Upper bounds for p

Thompson showed that the mle of p is distributed asymptotically normally and converges in probability to p as n approaches infinity. He suggested the following approximate confidence interval for p:

$$\hat{p} \pm t_{\frac{\alpha}{2},\, n-1}\left[\sum_{i=0}^{n}\left\{(1-\hat{p})-\left(\frac{i}{n}\right)^{1/k}\right\}^2\binom{n}{i}\{(1-\hat{p})^k\}^i\{1-(1-\hat{p})^k\}^{n-i}\right]^{1/2}.$$

Bhattacharyya et al. (1979) developed a confidence interval for p using the central limit theorem as follows:

$$\hat{p} \pm z_{\frac{\alpha}{2}}\left[\frac{1}{k}\left\{\frac{r}{n(n-r)}\right\}^{1/2}\left(\frac{n-r}{n}\right)^{1/k}\right].$$

Both approaches give symmetric confidence intervals for the true p despite the fact that the small-sample distribution of the mle of p is asymmetric, especially so when p is either very small or close to 1. Thus, their confidence intervals may lie out of the parameter space. Moreover, both approaches did not mention upper bounds, and upper bounds in classical approaches have been obtained by only replacing $t_{\alpha/2}$ and $z_{\alpha/2}$ with t_α and z_α and using the right (upper) side.

The upper bound for p based on the uniformly most powerful test is derived in this paper. The p.d.f of x_i can be written as

$$f(x_i) = \mathrm{Exp}\left\{\mathrm{Ln}[\,(1-p)^k] + x_i \times \mathrm{Ln}\left[\frac{1-(1-p)^k}{(1-p)^k}\right]\right\}$$

It is a one-parameter exponential family distribution, and $\sum X_i$ (=the number of infected groups to be tested) is the complete sufficient statistic for p. Moreover, since the following facts are obvious, the uniformly most powerful test is obtained:

(1) p is real,

(2) $\mathrm{Ln}\left[\frac{1-(1-p)^k}{(1-p)^k}\right]$ is strictly increasing in p, and

(3) the p.d.f. of X_i is the exponential family.

Let p^* be $1-(1-p)^t$. $\sum X_i$ is the optimal test statistic for testing $H_0 : p^* \leq p_0$ vs $H_a : p^* > p_0$ through choosing t and γ in the following test function to satisfy $E_{p_0}[\Phi(\sum X_i)] = \alpha$, where α is the size of the test. The test function is

$$\phi(x) = \begin{cases} 1 & \text{when } \sum x_i > t \\ \gamma & \text{when } \sum x_i = t \\ 0 & \text{when } \sum x_i < t \end{cases}$$

Computing $\sum X_i$ is discrete, exact size α tests will require randomization. Inversion of the test to obtain confidence intervals is not easy sometimes. Fortunately, exact confidence intervals can still obtained by determining the endpoints numerically (see Wilks, 1962), and the resulting intervals will possess the properties one would, under nonrandomization tests, attain by inverting a test. Since there is a one-to-one correspondence between the test of hypothesis and the confidence interval, the upper bound for p can be obtained as follows: Given a sample, the set of all values of p that would not be rejected in testing the null hypothesis

$H_0: p^* \leq p_0$ at the significance level, a, would be the $100(1-a)\%$ upper bound for p^*. Therefore, The upper bounds for p can be obtained by getting the region of p_0's such that the null hypothesis can not be rejected and computing $\{1-(1-p_0)^{1/k}\}$

4. Comparisons

As mentioned, group testing is mostly used when p is small, and upper bounds is of greatest interest. Hence, the combinations of (n, p) used are the following for the comparison: n=10, 20, 30, and 50 with p=0.01, 0.02, 0.05, and 0.1. The optimal group sizes for combinations of (n, p) are obtained from Swallow's (1985) table.

For the optimal group sizes, Figure 1 shows the mle and 95% upper bounds when n=20 with p=0.01, 0.02, 0.05, and 0.1, and Figure 2 displays them when p=0.05 with n=10, 20, 30, 50. The horizontal lines mark the true p and the vertical line indicate the values of r having maximum probability at fixed (n, p, k). The value shown as r is the number of defective groups found among the n tested.

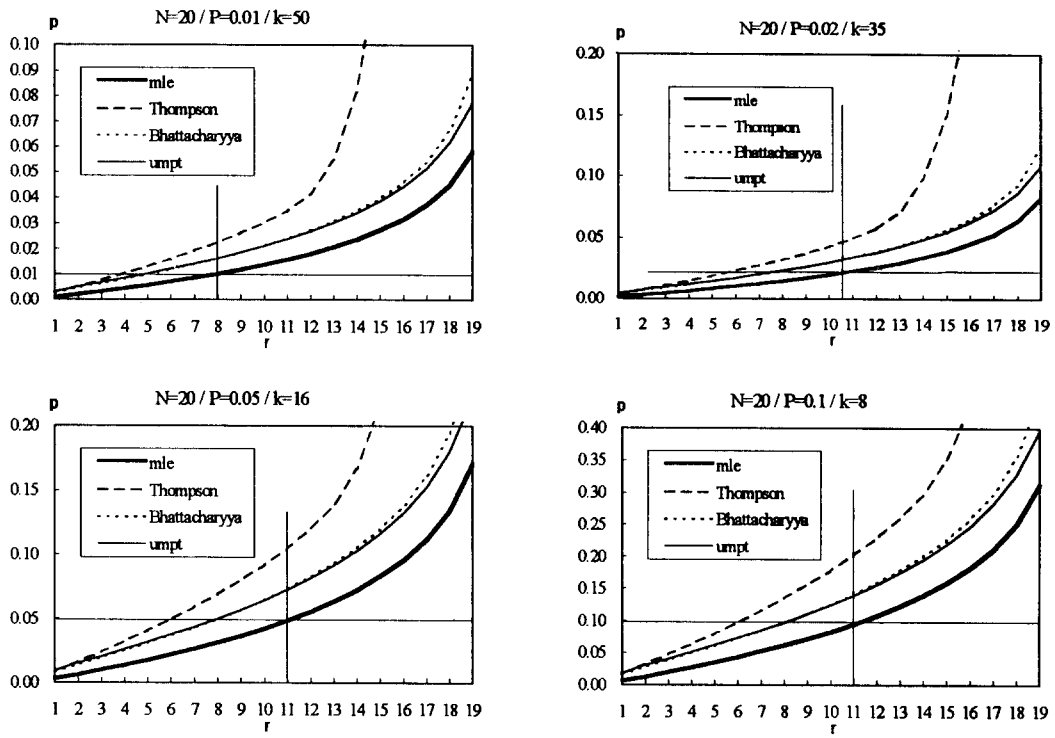

Figure 1. The mle and 95% upper bounds when n=20 with p=0.01, 0.02, 0.05, and 0.1

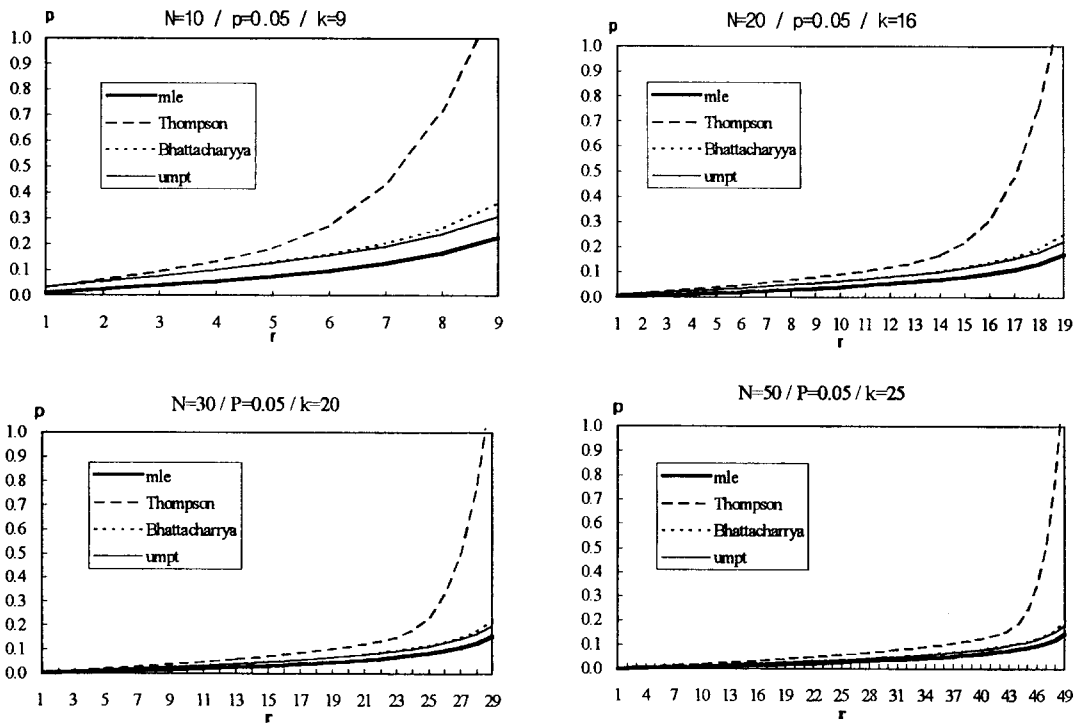

Figure 2. The mle and 95% upper bounds when p=0.05 with n=10, 20, 30, 50

For various combinations of (n, p, k), the similar results are obtained to Figure 1 and Figure 2. Thompson's bound is much larger than the others, and Bhattacharyya et al.'s is somewhat larger than the bound obtained from the uniformly most powerful test (umpt), especially when r is large. Thompson's upper bound is always higher, that is worse, than any other's except when r=1. It does not give the claimed (exact) confidence level when r=1, which will be discussed later. As r increases, the upper bound of Thompson becomes larger and larger, and finally goes up to be greater than 1. Therefore, it gives a completely uninformative confidence bound that spans the entire parameter space.

From Figures 1 and Figure 2, it would appear that neither the Bhattacharyya's nor umpt's upper bound is consistently the lower. For large r, the umpt bound is clearly better, but for small r, it appears that Bhattacharyya's upper bound is slightly lower than umpt's. The contradiction comes from the fact that it cheats a bit in getting symmetric confidence intervals even though the underlying distribution is not symmetric. Therefore, the umpt approach gives an exact $100(1-\alpha)\%$ confidence bound, but Bhattacharrya et al.'s approach does not produce claimed confidence levels. It gives a true level of confidence lower than claimed when r is small. Table 1 shows exact confidence levels of Bhattacharyya et al.'s method when p is 0.01 0.05, and 0.1 with n=20 for small r.

Table 1. The exact confidence levels of Bhattacharyya et al.'s approach

	r \ p	0.01	0.03	0.05	0.1
n=20	1	.9334	.9336	.9340	.9342
	2	.9359	.9361	.9365	.9369
	3	.9384	.9387	.9392	.9397
	4	.9407	.9410	.9416	.9422
	5	.9429	.9430	.9436	.9444
	6	.9447	.9449	.9455	.9464
	7	.9464	.9466	.9472	.9483
	8	.9481	.9483	.9488	-
	9	.9496	.9499	-	-

The exact confidence levels are calculated by $\Pr\{r \le i \mid r \sim \text{Bin}(n, 1-(1-\hat{p_u})^k)\}$, where $\hat{p_u}$ is Bhattacharyya et al.'s upper bound. As seen in table 1, Bhattacharyya et al.'s approach gives smaller confidence levels than actual confidence levels. And as r (or p) get large at the fixed p (or r), the confidence level from Bhattacharrya et al.'s approach goes to the claimed significance level (α =0.05). Similarly, Bhattacharrya et al.'s confidence level goes to the α as n goes lerge, but the group size k has little effect on the exact confidence level of Bhattacharyya's approach.

The values of r are, of course, not equally likely. The expected value of r is $n[1-(1-p)^k]$; values in that neighborhood will be most likely, and those will be found near where the curve for $\hat{p}$ crosses the horizontal line marking the true p. Using the value of r having maximum probability, the most likely lengths of each upper bound with each combination of (n, p) are calculated, and shown in Table 2.

Table 2. The most likely lengths

n→	10			20			30			50		
p↓	Thom.	Bhata.	umpt	Thom.	Bhata.	umpt	Thom.	Bhata.	umpt	Thom.	Bhata.	umpt
0.01	.0250	.0198	.0200	.0224	.0161	.0162	.0219	.0150	.0151	.0216	.0139	.0140
0.02	.0449	.0362	,0362	.0426	.0299	.0298	.0422	.0282	.0280	.0413	.0263	.0262
0.05	.1301	.0998	.0985	.1046	.0729	.0721	.1016	.0677	.0671	.0998	.0638	.0633
0.10	.2236	.1738	.1703	.2022	.1410	.1390	.2135	.1421	.1401	.2012	.1286	.1277

The expected length of each upper bounds is also computed using the following formula for each of a number of combinations of n, p and k.

$$\sum_{i=1}^{n-1}(\text{upper bound when } r=i)\times P[r=i \text{ at fixed } (n,p)]$$

where r, the number of defective groups, takes on values 0 to n. Because upper bounds of Thompson's approach and Bhattacharyya's approach are 0, 1, or not able to be computed, the values r=0 and n are not considered in calculating the expected length. That is, When $\hat{p}$=0 or 1 from r=0 or n, respectively, no informative confidence bound can be obtained. The expected lengths for the combination of (n, p) are computed and summarized in Table 3.

Table 3. The expected lengths

n→	10			20			30			50		
p↓	Thom.	Bhata.	umpt	Thom.	Bhata.	umpt	Thom.	Bhata.	umpt	Thom.	Bhata.	umpt
0.01	.0369	.0202	.0203	.0233	.0163	.0164	.0220	.0151	.0152	.0215	.0139	.0140
0.02	.0623	.0398	,0397	.0494	.0315	.0313	.0448	.0287	.0285	.0416	.0265	.0264
0.05	.1374	.0958	.0942	.1183	.0773	.0762	.1095	.0711	.0702	.1029	.0656	.0650
0.10	.2591	.1845	.1788	.2235	.1524	.1494	.2127	.1406	.1385	.2034	.1302	.1201

As seen in Table 2 and Table 3, Thompson's method produces the largest expected length and most likely length for all combinations of (n, p, k) as expected. Moreover, as the group size is increasing at fixed (n, p), the ratio of the expected length of Thompson's to the umpt's interval is increasing. Hence, Thompson's approach is clearly the worst of the three.

Bhattacharyya's method does not give an exact confidence level when r is small, especially for small p as discussed earlier. So, the expected length and most likely length of the umpt are smaller than those of Bhattacharyya's upper bounds except for p=0.01 where Bhattacharyya's method does not produce the claimed confidence level as discussed earlier. Moreover, for either the largest expected length and the most likely length, the differences between the Bhattacharyya and umpt approaches increase with p.

For all combinations of (n, p) shown in Table 2 and 3, the umpt approach has shortest expected length and most likely length of upper bound except for very small p (=0.01) where Bhattacharyya et al's procedure does not give the stated confidence levels for some r. Bhattacharyya et al's approach is a close second, and Thompson's is clear not recommendable in obtaining upper bounds for the infection rate.

5. Conclusion

Since the upper confidence bounds are often of interest in group testing, the upper bound for the infection rate p was derived based on the uniformly most powerful test and compared extensively with two previously suggested approaches, Bhattacharyya et al.'s and Thompson's

method. The expected length and most likely length of those three methods were calculated for combinations of (n, p). The umpt has the smallest expected length and most likely length except when r and p are both very small. For r and p both very small, Bhattacharyya et al.'s approach looks best in terms of the lowest upper bound, but this is an artifact of Bhattacharyya et al.'s procedure not giving the stated confidence level. The procedure's actual confidence level is less than claimed. Therefore, the proposed approach is the most attractive of the three approaches for obtaining upper bounds in group testing.

References

[1] Bhattacharyya, G. K. Karandinos, M. G. and DeFoliart G. R. (1979), Point estimates and confidence intervals for infection rates using pooled organisms in epidemiological studies, American Journal of Epistemology 109, 124-131

[2] 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

[3] Davis, C. E., Grizzle, J. E. and Bryan, J. A. (1975), Estimation of the probability of post transfusion hepatitis in hemophilia treatment, Biometrics 29, 386-392

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

[5] Finucan, H. M. (1964), The blood testing problem, Applied Statistics 13, 43-50

[6] 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

[7] Sobel, M. and Elashoff, R. M. (1975), Group testing with a new goal, estimation, Biometrika 62, 181-193

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

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

[10] Ungar, P. (1960),The cutoff point for group testing, Communications on pure and Applied Mathematics 13, 49-54

[11] Wilks, S. D. (1962), Mathematical Statistics. John Wiley and Sons, New York