ESTIMATION PROCEDURES IN GROUP TESTING

BY:

NG'ANG'A LUCY GATHONI

REG NO. I56/75990/2014

This Project is submitted in partial fulfilment for the Degree of Master of Science in Mathematical Statistics in the School of Mathematics

Declaration

This project is my original work and has not been presented for a degree in any other University.

Signature.....

This Project has been submitted for examination with my approval as the University Supervisor.

Signature.....

Dedication

To my mother,

Magret Njoki Ng'ang'a

and my sister

Lydia Wambui Ng'ang'a.

Acknowledgment

I thank Prof. M. M. Manene for his guidance, patience and availability.

I thank Dr. Jared Ongaro for introducing me to maple with which all the computations in this work were done.

I am grateful to Mr. Arthur Muchela for providing me with a work space.

Abstract

Estimation of proportion of defective items using group testing has been proven to be more precise (reduced MSE) and economical than individual testing. Seven point estimators have been explored in this project with their applicability and emphasis laid on the bias corrected and empirical Bayes estimators. Using an example, these two class of estimators have been shown to be superior to the maximum likelihood and classical Bayesian estimators. Four methods of obtaining group size have been examined with an in-depth description of group sizes obtained from optimizing the MSE and the method of adjusting the group size from one testing phase to another which are recommended. Finally, six interval estimates are explored with special consideration laid on the variance stabilized interval (VSI) a Wald interval generated with a stabilized variance (free of the proportion estimate). Using an example this Interval has been shown to overcome the problem of negative endpoints.

Contents

Chapter 1

INTRODUCTION

Group testing occurs when units from a population are pooled and tested as a group for the presence of a particular attribute such a disease, the measurement is usually taken to be dichotomous. If the test is positive it is assumed that at least one of the units in the group is positive ,otherwise it is assumed that all the units are negative.

Group testing can be performed in aim of achieving two objectives , either to identify the positive units in the groups tested or to estimate the proportion of positives in a wider population. This thesis is concerning the estimation of the proportion.

An interesting feature about group testing is the variety of application in different fields in which it is applied.Besides in epidemiology (entomology, ,public health etc.), where group testing is used in either classifying the infected organism from a population or estimate the proportion of the infected organism, group testing has a wider application in industrial manufacturing and engineering in identifying the defectives in a population of products, for instance in making a leak test on large number of gas filled electrical devices. Moreover, group testing is very useful in drug discovery process where screening large of compounds to identify the active ones. Other fields of application include experimental designs, genetics etc...

Group testing has also appeared under other names as "batch sampling" and "pooled testing ". This technique has been shown to reduce the cost of classifying all members of a population according to where or not they possess an attribute and/or estimate the

proportion of the members possessing the attribute of question when the incidence rate is fairly low.

The following assumptions will be made about group testing in this paper.

1. The probability to show the trait of interest independent and identically distributed for each unit in the population.

2. The testing is conducted without error i.e. there are no false negative or false positive results.

3. The units are randomly assigned to the groups.

1.1 Literature Review

Group testing first appeared in statistical literature in 1943 in the context of blood testing. Dorfman suggested using group testing with polled blood samples followed by one at a time retesting for individuals in any group that had tested positive. Dorfman's goal was to to classfie each of the individuals in the population as infected with syphilis or not while reducing the expected number of tests.Further analysis Dorfman revealed that for the economic application of the group testing the prevalence rate must be sufficiently small to make it worth while.

Sterrett (1957) introduced a variation of Dorfman's procedure by replacing one at a time testing with testing one by one until the defective one is found then the untested units are tested as a group. The efficiency of this methods were discussed by Sterrett(1957) and Sobel and Groll (1959).

Watson (1961) (see also Gurnow 1965) applied group testing in screening factors in experimental designs in which several factors considered may have effect or not on the response variable. Factors are grouped and the group is treated as a single factor if there is no effect, each factor is assumed to have no effect.

Gibbs and Gower (1960) were the first to explore the estimation using group testing in use of use of multiple-transfer in plant virus transmission studies using the maximum like-

lihood estimation to estimate the proportion and realized that the estimate is positively biased.

Keith Thompson (1962) used group testing in estimation the proportion of vectors in a natural population of insects this was the first time group testing was used to estimate a proportion by assuming a binomial model. He also proposed a formula to calculate group size k given the prior proportion by optimising the asymptotic variance. This method is discussed in detail and compared with others in this thesis. In addition Keith proposed Wald confidence intervals for the proportion.

Sobel and Elashoff (1975) explored the effects of their halving procedure to the MLE of the proportions. Under halving procedure a defective set is divided into equal subsets one of which chosen at random is retested. If this subset is defective it may be halved again. They discovered that this reduces the MSE of the estimate where p (true proportion) is $< 2/3$

Walter (1980) explored the properties of the maximum likelihood estimates when using the groups of different sizes.

Mundel (1984) and Hwang (1984) developed multi-stage group testing procedures in which the groups that fail the group test are subdivided successfully into subgroups for retest. The traditional approach of estimating P has been the maximum likelihood estimation and there are variation of MLE depending on the action taken after a group tests positive.

Swallow 1985 provided an in depth analysis of the point estimate properties of the $\mathbb{MLE}$ by exploring the bias and $\mathbb{MSE}$ of the $\mathbb{MLE}$ to arrive to a conclusion that using the optimal group size k, group testing can be more precise that individual testing, followed by Swallow (1987) in which he proposed using a group size less than the optimal which favours both the mean squared error and the cost per information.

Burrows (1987) proposed an improved MLE with superior bias and mean square error properties. This estimator improved the efficiency of group testing and extended the range of conditions where group testing is more efficient than individual testing. He also derived a simple formulation of optimal group sizes for situations where the number of

groups is fixed by resource limitations.

Hughes-Oliver and Swallow (1992) investigated the sensitivity in choice of group size in realizing the potential benefits of group testing (reduced cost and MSE) which depend on using an appropriate group size. Followed by an adaptive group testing MLE (1994) which resulted from adjusting the group size from time to time throughout the testing phase using all accumulated data to obtain the adaptive estimate of the proportion based on all the data collected. The performance of this adaptive estimator was evaluated parallel to normal MLE and was found to be superior.

Hepworth and Watson (2009) investigated the bias of the MLE when testing groups of different sizes using fixed and sequential procedures. They proposed numerical methods methods of for correcting the bias which produces almost an unbiased estimator.

Lew and Levy 1989 developed a Bayesian estimate of the proportion that was unbiased and adjusted for sensitivity and specificity of the screening test.

Gastwirth with Johnson (1991) used a Bayesian methods, which utilize prior results and update them as new data become available, to estimate the prevalence of AIDs in 1980's using USA and Canada blood donors.

Chaubey and Li (1995) developed a bayes estimator for the proportion for sample of fixed sizes by considering the prior of p(true proportion) or prior of p(estimated proportion) the two estimators performance are evaluated in comparison with MLE. It was observed that the Bayesian estimators were superior compared to the MLE.

Tebbs, Bilder and Moser (2003) developed an parametric empirical Bayesian procedure ,which estimates the model hyper-parameter with a maximum likelihood procedure instead of specifying a prior to avoid poor choices of the prior of the hyper-parameter, to estimate the proportion using a beta type prior distribution and a squared error loss function. They revealed that the empirical Bayes estimator using the squared loss function and found that estimator is superior over the usual $\mathbb{MLE}$ for small group sizes and small proportion. Also, they proposed a interval estimator of the proportion using the $100(1-\alpha)\%$ empirical credible interval. This interval estimator of P will be discussed in length in this thesis.

Tebbs and Bilder (2004) Proposed and thoroughly compared new interval estimators of the proportion in terms of coverage and mean length. He also proposed an new variance of P' that does not involve P to develop a variance stabilized interval with the $\mathbb{MLE}$ as the estimator and found that although the interval estimate is computationally complicated the interval overcame the problem of the negative endpoints.

Xiang, Walter and Shunpu (2007) proposed two additional empirical estimators using two scaled loss functions which improve the Bayesian approach to estimating p in terms of minimizing the $\mathbb{MSE}$ of the Bayes estimator of p small p. They showed that the new estimators are preferred over the estimator from the usual squared-error loss function and the MLE small p. In this thesis we thoroughly review some of the most used estimators (both point and interval). We also compare the Chaubey Li and Burrows estimators that were found to de-bias the MSE and further investigate the robustness of the group size to the errors of prior proportion.

1.2 Notation and definitions

MLE: Maximum likelihood estimator

MSE: Mean squared error

CLT: Central limit theorem

1.3 Summary

This thesis reviews both point and interval estimates of group testing. Chapter 2 contains the derivation of the maximum likelihood estimator and its properties further discussed. First the biased property of the $\mathbb{MLE}$ discussed and two bias corrected estimators are described. Secondly, the proof of consistency followed by the mention of its asymptotic normality property. Following the realization that the properties depend largely on the group size, the chapter also contains a lengthy discussion in detail of the four ways that

can be used in choosing the group size.

On the other hand the chapter contains the Bayesian approach of estimation. This include the classical Bayesian estimation approach using the squared error loss function. Additionally, three parametric empirical Bayesian estimators are also discussed in detail. These are developed using 1: squared error loss function and two scaled loss function.

Chapter 3 is generally on interval estimation of group testing. It kicks off with three Wald intervals. First, the wald interval proposed by Thomspson using the exact variance of the estimator is discussed. Secondly,we have a Wald interval estimate using the asymptotic variance of the estimate. The final Wald interval is based on a stabilized variance, not to depend on the unknown proportion, proposed by Tebbs and Bilder. Another exact interval estimate is discussed on this chapter to try and overcome the limitation of the Wald intervals. Finally, Bayesian interval estimates are discussed in detail, this is both classic and empirical approaches. These interval were obtained from both credible interval and highest posterior distribution means.

Chapter 2

Point Estimation

Group testing prevalence estimates can be categorized into two types:

1. Maximum Likelihood estimation

2. Bayesian estimation

2.1 Maximum likelihood estimation

Group testing starts with experimental units whose responses are assumed to be independent and identically distributed Bernoulli random variable with the probability of possessing the trait as P. The maximum likelihood estimator $\mathbb{MLE}$ of P means that of all the values of that P could assume $\hat{p}$ as the one that maximises the probability or the likelihood of the observed data. These random variable are grouped in to groups(n) of size k.

Let $p =$ the probability of selecting at random an individual with the attribute,

$1 - p =$ the probability of selecting at random an individual without the attribute,

$(1 - p)^k =$ the probability of obtaining by random selection a unit in a group of size k who are without the attribute.

$\hat{p} = 1 - (1-p)^k$ = the probability of obtaining by random selection a group of size k with at least one experimental unit with the attribute.

Under these assumptions the number of groups possessing the attribute X has a binomial distribution with parameters n and $\hat{p} = 1 - (1-p)^k$.

Thus

$$f(x_i, p) = \binom{n}{x}(1 - (1-p)^k)^x((1-p)^k)^{n-x} \tag{2.1}$$

Therefore, the MLE can be derived as follows

$$l(x; p) = \sum_{i=0}^{n} \binom{n}{x}(1 - (1-p)^k)^x((1-p)^k)^{n-x} \tag{2.2}$$

$$\log l(x; p) = \prod_{i=0}^{n} \binom{n}{x} + x \log 1 - (1-p)^k + (n-x)\log(1-p)^k \tag{2.3}$$

$$\frac{d \log l(x; p)}{dx} = \frac{x}{1 - (1-p)^k} + \frac{n-x}{(1-p)^k} \tag{2.4}$$

$$0 = \frac{x}{1 - (1-p)^k} + \frac{n-x}{(1-p)^k} \tag{2.5}$$

$$-\frac{x}{1 - (1-p)^k} = \frac{n-x}{(1-p)^k} \tag{2.6}$$

$$\hat{p} = 1 - (1 - \frac{x}{n})^{\frac{1}{k}} \tag{2.7}$$

With mean

$$E(\hat{p}) = \sum_{x}^{n} \hat{p} L(x, p) \tag{2.8}$$

$$E(\hat{p}) = 1 - \sum_{x}^{n} (1 - \frac{x}{n})^{\frac{1}{k}} * \binom{n}{x}(1 - (1-p)^k)^{x_i}((1-p)^k)^{n-x_i} \tag{2.9}$$

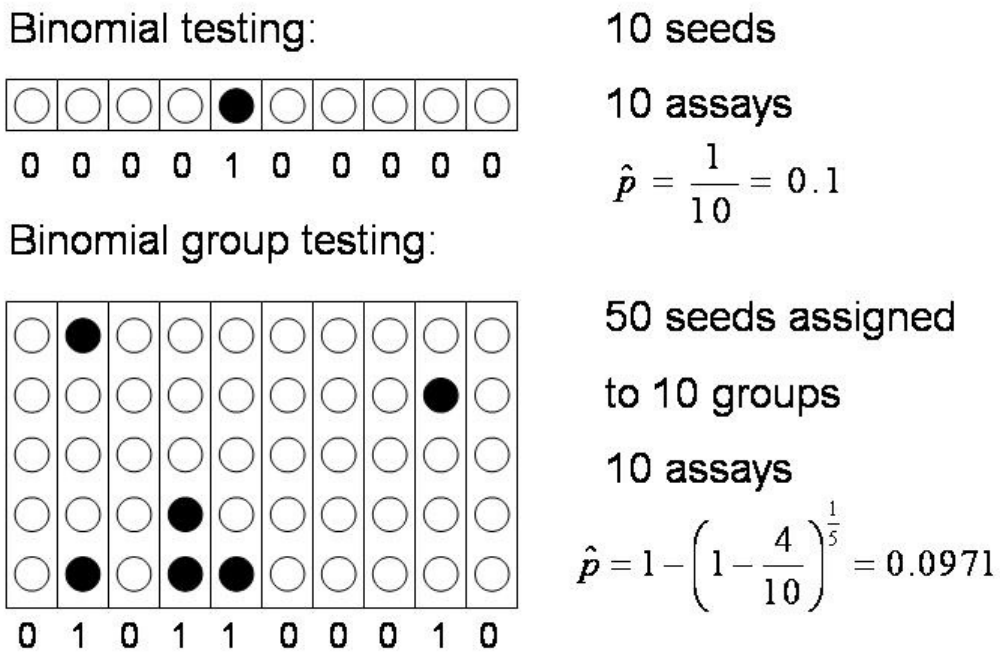

Figure 2.1: An example illustrating group testing estimate

and variance

$$Var(\hat{p}) = 1 - \sum_{x}^{n}(1 - \frac{x}{n})^{\frac{2}{k}} * \binom{n}{x}(1 - (1-p)^k)^{x_i}((1-p)^k)^{n-x_i} - [1 - E(p)]^2 \quad (2.10)$$

The estimate depends on k (group size). The group size k is the one that adjusts the group proportion to individual prevalence.

For instance figure 2.1 shows an example the group proportion was 4/10 while the actual proportion was $3/25 = 0.12$ but the estimated proportion was 0.0971. Group size k influences the estimated proportion. Figure 2.2 shows the relationship of the group proportion in relation to the individual proportion.

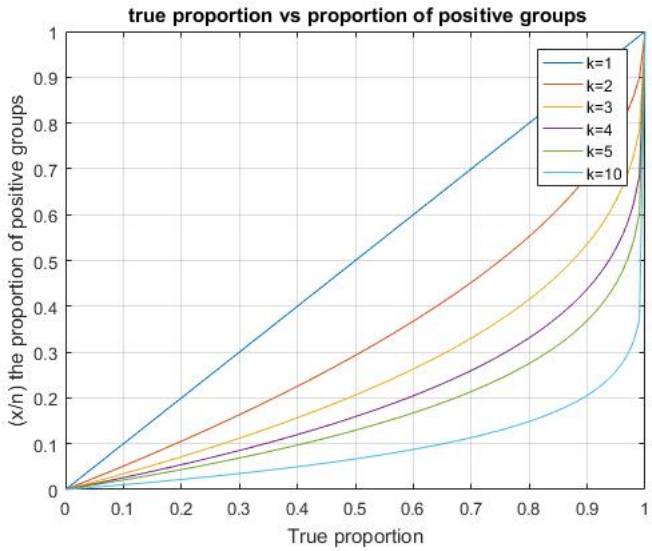

Figure 2.2: The graph above shows the relationship between the true proportion and the proportion of the positive groups

2.1.1 Properties of the estimate

We first explore the properties and the asymptotic properties of the estimate.

The estimator is biased

Let X_i be drawn i.i.d from $f(\theta, X)$. An estimator $\hat{\theta}$ is said to be biased if $E(\hat{\theta}) \neq \theta$.

Although $(1 - \frac{x}{n})$ the proportion of the groups that test negative to the attribute is an unbiased estimator of $(1-p)^k$, a bias is introduced in taking the k-th root of $(1 - \frac{x}{n})$. Thus the estimator is biased.

Proof

This can be proved using the Jensen's inequality.

Given that

$$\hat{p} = 1 - (1 - \frac{x}{n})^{\frac{1}{k}} \tag{2.11}$$

$$E(\hat{p}) = 1 - E(1 - \frac{x}{n})^{\frac{1}{k}} \geq 1 - (1 - E(\frac{x}{n}))^{\frac{1}{k}} \tag{2.12}$$

$$E(\hat{p}) \geq 1 - (1 - (1-P)^k)^{\frac{1}{k}} \tag{2.13}$$

$$E(\hat{p}) \geq P \tag{2.14}$$

Thus the MLE overestimates the true proportion and its is therefore inaccurate.

$$E(\hat{p}) = \sum_{x}^{n} \hat{p} L(x, p) \tag{2.15}$$

$$E(\hat{p}) = 1 - \sum_{x}^{n} (1 - \frac{x}{n})^{\frac{1}{k}} * \binom{n}{x} (1 - (1-p)^k)^x ((1-p)^k)^{n-x} \tag{2.16}$$

$$Bias(\hat{p}) = E(\hat{p} - p) \tag{2.17}$$

$$Bias(\hat{p}) = \sum_{x}^{n} [(1-p) - (1 - \frac{x}{n})^{\frac{1}{k}}] \times \binom{n}{x} (1 - (1-p)^k)^x ((1-p)^k)^{n-x} \tag{2.18}$$

From table 2.1 the general trend of bias is evident that for fixed group size and p, proportion, the bias decreases as n increases this is expected since the $\mathbb{MLE}$ is asymptotically unbiased. Its is also notable that for small k and p the bias is negligible even for small n. This is evident for all k and n when p=0.01 the bias is always less than 0.0006 which would be negligible for most cases. Thus if the researcher has some prior information that the true proportion is less than or equal to or less than 0.01, he/she has the freedom of choosing any k and n depending on the resources available without worry of the estimator being biased.

However, for all n, $p \geq 0.10$ and large k, the bias is relatively large to be ignored. For example, when p= 0.15 and k=15 the bias from table 2.1 is 0.33 which would increase the estimated proportion to 0.48 which would be exaggerated more than three times. This insists on the caution of choice of k.

Figure 2.3 shows that for a fixed n the bias $[E(\hat{p}) - p]$ is positive, increases as k increases thus strongly dependent on the group size. for instance when k=4 the bias is never great and its negligible as long as p is less than about 0.4. When k=22 the bias is negligible as long as p is less than 0.06.

Alternative MLE estimators

Several authors have attempted to come up with modified estimators that reduce the bias.

P	K	n					
		10	20	30	50	100	200
0.01	2	0.000265	0.000129	0.000085	0.000051	0.000025	0.000012
	5	0.000436	0.000211	0.000139	0.000083	0.000041	0.000020
	10	0.000508	0.000244	0.000160	0.000095	0.000047	0.000024
	15	0.000544	0.000261	0.000171	0.000101	0.000050	0.000025
0.02	2	0.000533	0.000259	0.000171	0.000102	0.000051	0.000025
	5	0.000895	0.000431	0.000284	0.000169	0.000084	0.000042
	10	0.001077	0.000515	0.000338	0.000200	0.000099	0.000049
	15	0.001198	0.000566	0.000371	0.000219	0.000108	0.000054
0.03	2	0.000805	0.000391	0.000258	0.000154	0.000076	0.000038
	5	0.001377	0.000662	0.000486	0.000259	0.000129	0.000064
	10	0.001720	0.000815	0.000534	0.000317	0.000157	0.000078
	15	0.002023	0.000927	0.000625	0.000357	0.000176	0.000088
0.04	2	0.001080	0.000525	0.000347	0.000206	0.000103	0.000051
	5	0.001886	0.000904	0.000595	0.000353	0.000175	0.000087
	10	0.002460	0.001149	0.000751	0.000445	0.000220	0.000109
	15	0.003283	0.001355	0.000880	0.000518	0.000256	0.000127
0.05	2	0.001359	0.000660	0.000435	0.000259	0.000129	0.000064
	5	0.002425	0.001158	0.000761	0.000451	0.000224	0.000111
	10	0.003357	0.001523	0.000993	0.000587	0.000289	0.000144
	15	0.005721	0.001871	0.001207	0.000707	0.000347	0.000127
0.07	2	0.001928	0.000935	0.000617	0.000367	0.000183	0.000091
	5	0.003602	0.001735	0.001118	0.000662	0.000328	0.000163
	10	0.006271	0.002419	0.0001565	0.000919	0.000453	0.000225
	15	0.020179	0.003464	0.002067	0.001194	0.000583	0.000288
0.10	2	0.002815	0.001361	0.000899	0.000534	0.000265	0.000132
	5	0.005728	0.002643	0.001727	0.001020	0.000505	0.000251
	10	0.019048	0.004403	0.002723	0.001580	0.000773	0.000382
	15	0.090470	0.014182	0.004927	0.002356	0.001122	0.000550
0.15	2	0.004393	0.002112	0.001392	0.000827	0.000410	0.000204
	5	0.011530	0.004627	0.002992	0.001756	0.000865	0.000429
	10	0.094557	0.017750	0.006950	0.003409	0.001623	0.000795
	15	0.326017	0.134159	0.058109	0.014140	0.003145	0.001448
0.20	2	0.006128	0.002926	0.001924	0.001142	0.000566	0.000282
	5	0.025096	0.007549	0.004716	0.002739	0.001339	0.000663
	10	0.240669	0.083522	0.032951	0.009154	0.003245	0.001552
	15	0.539802	0.374276	0.261921	0.131038	0.027258	0.004423
0.25	2	0.008079	0.003814	0.002503	0.001484	0.000735	0.000366
	5	0.054146	0.013264	0.007296	0.004099	0.001984	0.000978
	10	0.393960	0.220857	0.127287	0.046265	0.008375	0.003055
	15	0.641724	0.555621	0.482812	0.366772	0.188461	0.053939

Table 2.1: The bias of MLE

1. Chaubey and Li (1995)

Lovision, 1994 expanded the MLE into a Taylor series around the true parameter

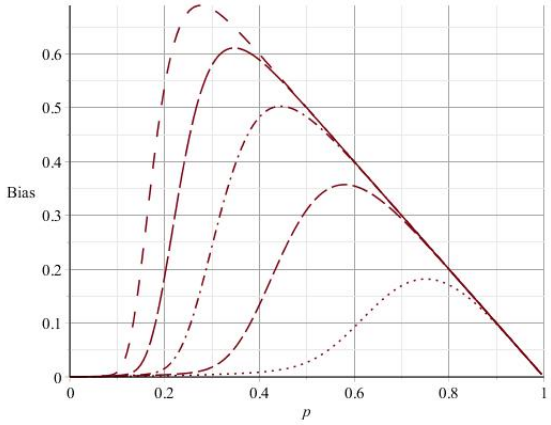

Figure 2.3: The plot of bias($\hat{p}$) for n=50 and k=4,7,11,16,22

P so as to eliminate the leading term of the bias.

$$\hat{p} = p + (\hat{p} - p)\frac{d\hat{p}}{dp} + (\hat{p} - p)^2 \times \frac{1}{2} \times \frac{d^2\hat{p}}{dP^2} + ... \tag{2.19}$$

Taking term by term expectation of the Taylor series expansion then

$$E(\hat{p}_1) = P + P(1-P)\lceil\frac{k-1}{2nk^2}\rceil[\frac{1}{n(1-p)} + \frac{(1-2P)(2k-1)}{3kn^2(1-p)^2} + ...] \tag{2.20}$$

$$E(\hat{p}_1) \approx \hat{p} + \frac{k-1}{2nk^2}[\frac{1-(1-P)^k}{(1-P)^{k-1}}] \tag{2.21}$$

where

$$\frac{k-1}{2nk^2}[\frac{1-(1-p)^k}{(1-p)^{k-1}}]$$

is the bias approximated using the Taylor series approximation.

The comparison of approximate bias and the exact bias as shown in the table 2.2 shows that the approximation is adequate when P is small. Since the MLE of $(1-p)^k = [1 - \frac{x}{n}]$

then the MLE of $1-(1-P)^k = \frac{x}{n}$

and the MLE of $(1-P)^{k-1} = (1 - \frac{x}{n})^{\frac{(k-1)}{k}}$

P	K	n					
		10		20		30	
		exact	approx.	exact	approx.	exact	approx.
0.005	2	0.00013	0.00013	0.00006	0.00006	0.00004	0.00004
	5	0.00022	0.00020	0.00010	0.00010	0.00007	0.00007
	10	0.00025	0.00023	0.00012	0.00012	0.00008	0.00008
	15	0.00026	0.00024	0.00013	0.00012	0.00008	0.00008
0.01	2	0.0003	0.0003	0.0013	0.0013	0.0009	0.0008
	5	0.0004	0.0004	0.0021	0.0020	0.00014	0.0014
	10	0.0005	0.0005	0.0024	0.0024	0.0016	0.0016
	15	0.0005	0.0005	0.0026	0.0025	0.0017	0.0017
0.05	2	0.0014	0.0013	0.0007	0.0006	0.0004	0.0004
	5	0.0024	0.0022	0.0012	0.0011	0.0008	0.0007
	10	0.0036	0.0029	0.0015	0.0014	0.0010	0.0010
	15	0.0057	0.0034	0.0019	0.0017	0.0012	0.0011
0.10	2	0.0028	0.0026	0.0014	0.0013	0.0009	0.0009
	5	0.0057	0.0050	0.0026	0.0025	0.0017	0.0017
	10	0.0190	0.0076	0.0044	0.0038	0.0027	0.0025
	15	0.0905	0.0108	0.0142	0.0054	0.0049	0.0036
0.25	2	0.0081	0.0073	0.0038	0.0036	0.0025	0.0024
	5	0.0541	0.0193	0.0133	0.0096	0.0073	0.0064
	10	0.3940	0.0566	0.2209	0.0283	0.1273	0.0189
	15	0.6417	0.1723	0.5556	0.0861	0.4828	0.0574

Table 2.2: The exact bias and approximate bias of MLE

Thus a bias corrected estimator is proposed $\hat{p}_1$

$$\hat{p}_1 = \hat{p} - \frac{k-1}{2nk^2}[\frac{x/n}{1 - x/n^{(k-1)/k}}] \tag{2.22}$$

Where $\hat{p} = 1 - (1-p)^{\frac{1}{k}}$

The ratio of the bias of corrected MLE to that of uncorrected MLE, we find that the bias correction is effective for small k as well as for the larger k. Substantial reduction in bias is observed for small values of P.

2. Burrows (1987)

The MLE can be written as $\hat{p} = 1 - (y/n)^{1/k}$ where $y = n - x$ is the number of negative groups. If y/n is replaced by $(y + a)/(n + b)$ and the expression is expanded as a power series of n^{-1}, it is found that the bias term is eliminated when

P	K	n				
		10	20	30	50	100
0.005	2	-0.0006	-0.0001	-0.0001	0.0000	0.0000
	5	-0.0222	-0.0105	-0.0069	-0.0041	-0.0020
	10	-0.0300	-0.0143	-0.0094	-0.0056	-0.0027
	15	-0.0331	-0.0157	-0.0103	-0.0061	-0.0030
0.01	2	-0.0006	-0.0001	-0.00003	0.0000	0.0000
	5	-0.0225	-0.0107	-0.0070	-0.0041	-0.0021
	10	-0.0310	-0.0147	-0.0096	-0.0057	-0.0028
	15	-0.0349	-0.0164	-0.0107	-0.0068	-0.0031
0.05	2	0.000008	0.0002	0.0002	0.0001	0.0001
	5	-0.0252	-0.0116	-0.0076	-0.0045	-0.0022
	10	-0.0200	0.0190	-0.0122	-0.0071	-0.0035
	15	0.2334	-0.0245	-0.0162	0.0093	-0.0045
0.10	2	0.0009	0.0006	0.0004	0.0003	0.0001
	5	-0.0191	-0.0135	-0.0086	-0.0050	-0.0025
	10	0.4870	-0.0021	-0.0182	-0.0104	-0.0049
	15	0.8729	0.5116	0.1177	0.0165	-0.0086
0.25	2	0.0098	0.0024	0.0016	0.0010	0.0005
	5	0.6048	0.1241	-0.0012	-0.0092	-0.0042
	10	0.9688	0.9264	0.8714	0.7130	0.1809
	15	0.9954	0.9904	0.9851	0.9735	0.9371

Table 2.3: The ratio of Chaubey's Estimators bias to that of MLE

$a = b = \frac{1}{2}(k-1)/k$. The substitution then leads to

$$\hat{p}_2 = 1 - [\frac{2k(n-x)+k-1}{2kn+k-1}]^{1/k} \tag{2.23}$$

The 2.23 estimator does so well in reducing bias as illustrated by table 2.4. Table 2.4 shows the percentage reduction of bias calculated as

$$[1 - [\frac{Bias(\hat{p}_2)}{Bias(\hat{p})}]] \times 100$$

$$Bias(\hat{p}) = E(\hat{p}) - p$$

For the cases where the percentage reduction is less than 100% then $Bias(\hat{p}_2)$ is positive and less than $Bias(\hat{p})$. With the cases where the percentage is greater than 100% then $Bias(\hat{p}_2)$ is negative.

P	K	n					
		10	20	30	50	100	200
0.005	2	97.55	98.76	99.18	99.53	99.86	100.33
	5	98.07	99.01	99.34	99.61	99.88	100.22
	10	98.22	99.09	99.39	99.65	99.88	100.18
	15	98.25	99.11	99.41	99.68	99.99	100.63
0.01	2	97.53	98.75	99.15	99.46	99.59	99.27
	5	98.04	98.99	99.32	99.58	99.75	99.72
	10	98.16	99.06	99.37	99.63	99.87	100.15
	15	98.17	99.06	99.36	99.61	99.77	99.77
0.05	2	97.41	98.69	99.12	99.47	99.73	99.87
	5	97.76	98.87	99.24	99.54	99.77	99.87
	10	97.56	98.77	99.18	99.51	99.76	99.89
	15	97.83	98.55	99.04	99.43	99.71	99.83
0.10	2	97.21	98.61	99.07	99.44	99.72	99.89
	5	97.24	98.64	99.09	99.45	99.72	99.86
	10	98.34	98.11	98.76	99.27	99.64	99.82
	15	100.39	98.78	98.29	98.87	99.46	99.74
0.20	2	96.68	98.34	98.92	99.35	99.68	99.84
	5	97.48	97.78	98.56	99.16	99.59	99.80
	10	103.13	100.69	99.38	98.27	98.95	99.51
	15	106.24	103.75	102.54	101.23	99.58	98.58
0.25	2	96.28	98.22	98.82	99.29	99.65	99.82
	5	99.23	97.40	98.05	98.88	99.45	99.73
	10	106.39	103.25	101.72	99.98	98.35	99.04
	15	111.08	107.67	105.94	104.06	101.98	100.29

Table 2.4: The percentage reduction of bias of Burrows estimator

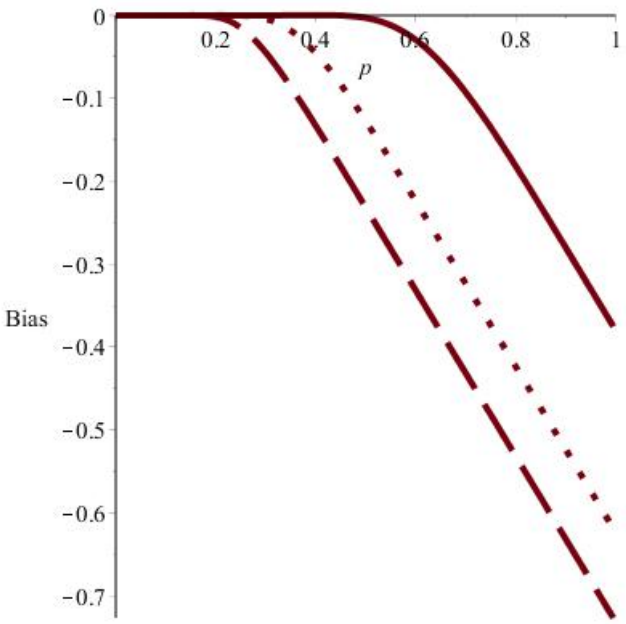

Figure 2.4: The burrows bias plot

Although this estimator have almost eliminated the bias, for larger k and p or large n and small p then p_2 produces a negative bias (this is the situation where p is greater than $E(\hat{p})$).

Consistent

A Consistent estimator is such that as the sample size (data points) increases the estimator approaches the parameter it is actually estimating.

Mathematically, An estimator of θ say $T_n$is consistent if it converges in probability to θ.

$$\lim_{n\to\infty} P(|T_n - \theta| \geq \varepsilon) = 0 forall \varepsilon > 0 \tag{2.24}$$

Proof Using the Chebychev's inequality

$$P(|T_n - \theta| \geq \varepsilon) \leq E(\frac{(T_n - \theta)^2}{\varepsilon^2}) \tag{2.25}$$

$$\lim_{n\to\infty} P(|T_n - \theta| \geq \varepsilon) \leq lim_{n\to\infty} E(\frac{(T_n - \theta)^2}{\varepsilon^2}) \tag{2.26}$$

Where $E((T_n - \theta)^2)$ is the MSE of P'

But $\lim_{n\to\infty} MSE = 0$ Thus

$$\lim_{n\to\infty} P(|T_n - \theta| \geq \varepsilon) = 0 \tag{2.27}$$

$forall\varepsilon > 0$

Asymptotically normally distributed

This simply means that as we get more data, averages of random variables behave like normally distributed random variables.This property is based on the Taylor series theorem.

$$\sqrt{N(\hat{p} - P)} \approx Normal(0, \frac{1-(1-p)^k}{k^2(1-p)^{k-2}}) \tag{2.28}$$

2.2 Experimental design

In the usual group testing scheme, it is assumed that n the number of groups (tests) and k the size of each group are determined before data collection is begins.In choosing n one usually considers constraints related to cost(the budget) and feasibility of implementation (available resources). The choice of group size, k,may also be constrained. For instance there might be a practical limitation on the size of the group, no more than 30 individuals.

2.2.1 Choice of group size

There has been a degree of discomfort associated with group testing because of the sensitivity of the procedures to the size used for the groups in obtaining the data. Whereas great gains can be achieved by testing in groups (rather than one at a time) when near optimal group sizes are used, the losses can be overwhelming when highly inappropriate group sizes are used. The possibility of such of such occurrences limits the appeal of group testing for potential users who prefer a procedure for which is assured a minimal level of performance. Thus the efficiency of group testing largely depends on the group

size k, the number units in a group.

If k is too large the estimated proportion is close to 1 and all groups are likely to test positive, this makes the experiment very uninformative. However, if k is too small the estimated proportion is closer to zero than necessary which makes the experiment very expensive. Therefore it is desirable to have an optimal k that will balance the trade-off. In this thesis examines the four methods for determining the group size when estimating a proportion: (1) choose the group size that makes the probability that a group shows the trait is equal to that a group does not show the trait; (2) choose the group size that minimizes the the asymptotic variance of the estimate of the proportion; (3) choose the group that minimises the exact mean squared error of the estimate of the proportion; (4) adapting the group size from one stage to another throughout the testing phase.

In application, each of the methods mentioned above requires the user to specify an initial value of the proportion (prior proportion), based on whatever the prior information the user may process. The group size obtained through each of the following methods depend on the initial value.

1. **Chiang and Reeves (1962)** Chiang and Reeves considered the idea of equalizing the probability of obtain a group that shows the trait and the group that does not show the trait.

   Let p = the probability of selecting at random an individual with the trait,

   $1 - p$ = the probability of selecting at random an individual without the trait,

   $(1 - p)^k$ = the probability of obtaining by random selection a unit in a group of size k who are without the trait.

   Thus $1 - (1 - p)^k$ = the probability of obtaining by random selection a unit in a group of size k with the trait.

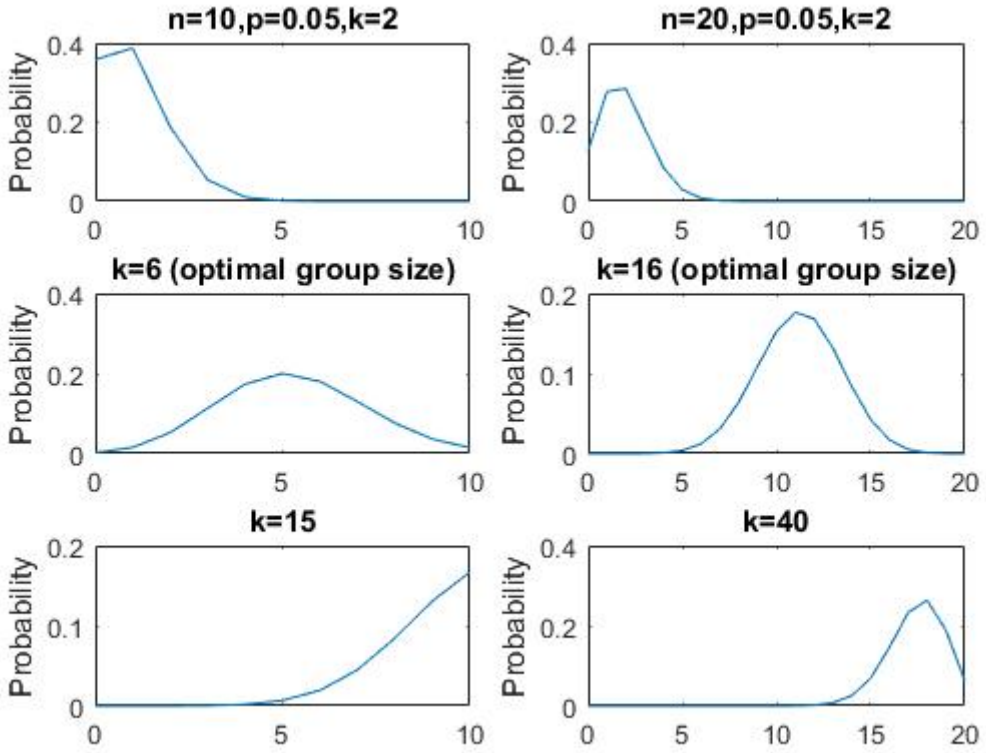

Figure 2.5: The effect of group size to the estimated proportion

That is to choose k such that

$$1 - (1-p)^k = (1-p)^k = \frac{1}{2} \tag{2.29}$$

$$\log(1 - (1-p)^k) = k \log(1-p)$$

But $1 - (1-p)^k = \frac{1}{2}$

$$\log \frac{1}{2} = k \log(1-p)$$

Therefore

$$k_1 = \frac{\log(1/2)}{\log(1-p)} \tag{2.30}$$

With this choice of k, one would then obtain estimate $\hat{p}$ using Eqt. 2.7. The general

P	0.001	0.005	0.01	0.02	0.05	0.1	0.15	0.20	0.25	0.3
K	693	138	69	34	14	7	4	4	2	2

Table 2.5: Group sizes for different proportion as proposed by Chiang and Reeves (1962)

trend is as the proportion increases the group sizes decreases.

2. **Thompson 1960**

Thompson suggested the choice k to be based on the minimisation of the asymptotic variance of the estimated proportion.

$$\lim_{n\to\infty} V(P') = \frac{1}{nE[\frac{dlogf(x,p)}{dp}]^2} \tag{2.31}$$

Alternatively,

$$\lim_{n\to\infty} V(P') = \frac{1}{-nE[\frac{d^2logf(x,p)}{dp^2}]} \tag{2.32}$$

$$f(x,p) = \binom{n}{x}(1-(1-p)^k)^x((1-p)^k)^{n-x} \tag{2.33}$$

$$\log f(x,p) = log\binom{n}{x} + xlog(1-(1-p)^k) + k(n-x)log(1-p) \tag{2.34}$$

$$\frac{d\log f(x,p)}{dp} = \frac{d}{dp}[x\log(1-(1-p)^k)] + \frac{d}{dp}[k(n-x)\log(1-p)] \tag{2.35}$$

let$m = 1-(1-p)^k$ then $dm = k(1-p)^{k-1}dp$

$$\frac{d}{dp}[x\log(1-(1-p)^k)] = \frac{x}{(1-(1-p)^k)}[k(1-p)^{k-1}] \tag{2.36}$$

$$\frac{d}{dp}[k(n-x)\log(1-p)] = \frac{k(n-x)}{(1-p)} \tag{2.37}$$

Thus

$$\frac{d\log f(x,p)}{dp} = \frac{x}{(1-(1-p)^k)}[k(1-p)^{k-1}] + \frac{k(n-x)}{(1-p)} \tag{2.38}$$

$$\frac{d^2\log f(x,p)}{dp^2} = \frac{d}{dp}[\frac{xk(1-p)^{k-1}}{(1-(1-p)^k)}] + \frac{d}{dp}[\frac{k(n-x)}{(1-p)}] \tag{2.39}$$

First, let $u = xk(1-p)^{k-1}$ then $du = -xk(k-1)(1-p)^{k-2}dp$

and $v = 1-(1-p)^k$ then $\frac{d}{dp}[\frac{k(n-x)}{(1-p)}]dv = k(1-p)^{k-1}$

consequently $\frac{d}{dp}[\frac{u}{v}] = \frac{u'v-v'u}{v^2}$

$$\frac{d}{dp}[\frac{xk(1-p)^{k-1}}{(1-(1-p)^k)}] = [\frac{-xk(k-1)(1-p)^{k-2}\times 1-(1-p)^k] - [k(1-p)^{k-1}\times xk(1-p)^{k-1}}{(1-(1-p)^k)^2}] \tag{2.40}$$

$$\begin{aligned}\frac{d}{dp}[\frac{xk(1-p)^{k-1}}{(1-(1-p)^k)}] &= \frac{-xk(k-1)(1-p)^{k-2}\times 1-(1-p)^k - xk^2(1-p)^{2(k-1)}}{(1-(1-p)^k)^2} \\ &= \frac{xk(1-p)^{k-1}}{(1-(1-p)^k)^2}[(k-1)(1-p)\times(1-(1-p)^k) - k(1-p)^{k-1}]\end{aligned} \tag{2.41}$$

On the other hand

$$\frac{d}{dp}[\frac{k(n-x)}{(1-p)}] = \frac{k(n-x)}{(1-p)^2} \tag{2.42}$$

Thus

$$\frac{d^2 logf(x,p)}{dp^2} = \frac{-xk(1-p)^{k-1}}{(1-(1-p)^k)^2}[(k-1)(1-p)\times(1-(1-p)^k) - k(1-p)^{k-1}] - \frac{k(n-x)}{(1-p)^2} \tag{2.43}$$

Since the $x \sim bin(n,(1-(1-p)^k))$ then $E(x) = n(1-(1-p)^k)$

$$\begin{aligned}E(\frac{d^2 logf(x,p)}{dp^2}) = \frac{-nk(1-p)^{k-1}}{(1-(1-p)^k)}[&(k-1)(1-p) \\ &\times(1-(1-p)^k) - k(1-p)^{k-1}] - kn(1-p)^{k-2}\end{aligned} \tag{2.44}$$

$$E(\frac{d^2 logf(x,p)}{dp^2}) = \frac{-nk(1-p)^{k-2}}{(1-(1-p)^k)}[(k-1)(1-(1-p)^k) + k(1-p)^k + (1-(1-p)^k)] \tag{2.45}$$

$$E(\frac{d^2 logf(x,p)}{dp^2}) = \frac{-nk^2(1-p)^{k-2}}{(1-(1-p)^k)} \tag{2.46}$$

$$\lim_{n\to\infty} V(P') = [\frac{-nk^2(1-p)^{k-2}}{(1-(1-p)^k)}]^{-1} \tag{2.47}$$

$$\lim_{n\to\infty} V(P') = \frac{1-(1-p)^k}{nk^2(1-p)k-2} \tag{2.48}$$

To minimise the asymptotic variance of $\hat{p}$.

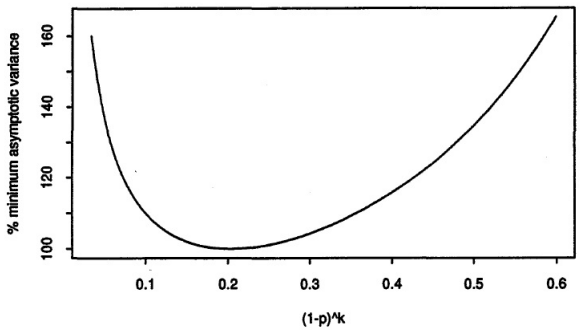

Figure 2.6: The asymptotic variance of $\hat{p}$ as a percentage of its minimum value for different values of $(1-p)^k$

$$\frac{d}{dk}[\frac{1-(1-p)^k}{nk^2(1-p)k-2}] = \frac{1}{nk^3(1-p)^{k-2}}[2(1-p)^k - klog(1-p) - 2] \tag{2.49}$$

which has a unique solution in k for all $0 < p < 1$ when equated to zero.

$$0 = \frac{1}{nk^3(1-p)^{k-2}}[2(1-p)^k - klog(1-p) - 2] \tag{2.50}$$

This solution maybe represented by the equation

$$2(1-p)^k - klog(1-p) = 2 \tag{2.51}$$

Thus the approximaton

$$k_2 \approx \frac{-1.5936}{ln(1-p)} \tag{2.52}$$

which is obtained by solving $(1-p)^k \approx 0.2032$ for k is sufficient to minimise the asymptotic variance of p. k_2 is the solution to 2.52

P	0.001	0.005	0.01	0.02	0.05	0.1	0.15	0.20	0.25	0.3
k_2	1593	318	159	79	31	15	10	7	6	3

Table 2.6: Group sizes for different proportion as proposed by Thompson(1962)

Figure 2.6 plots the asymptotic variance of p as a percentage of its minimum value versus different values of $(1-p)^k$.We see that this function is indeed minimised at

$(1-p)^k = 0.2032$, but we also see that there is a range of values of $(1-p)^k$ that are close to the minimum value. For instance, if $(1-p)^k \epsilon [0.1641, 0.2477]$ then there is no more than 1% increase in asymptotic variance, so that any k such that $(1-p)^k$ is in this region is acceptable.

For example if p=0.05, then this region is $27 \leq k \geq 35$ instead of the single value that k=31. Figure refasy vs grp also shows that outside this interval the rate of increase of of the asymptotic variance is greater for values of k such that $(1-p)^k < 0.2032$ than for the values of of k such that $(1-p)^k > 0.2032$. since larger values of of k will result smaller values of $(1-p)^k$. That is, the percentage increase in the asymptotic variance is greater when k is larger than optimal than when is smaller than optimal.

Hence, if one cannot use the optimal value of k or a value such that $(1-p)^k \epsilon [0.1641, 0.2477]$ then it is better to use a smaller that optimal values rather that a larger than optimal value.

3. **Swallow 1985**

He proposed the choice of group size (k) to be chosen so as to give the smallest mean squared error(MSE).This is due to the fact that the estimator is biased, the MSE is the best measure of precision because it takes into account both the variance(precision) and the bias(accuracy) of the estimator.

The MSE of the estimator is given as

$$MSE(\hat{p}) = E(\hat{p} - p)^2 \tag{2.53}$$

$$MSE(\hat{p}) = \sum_{x=0}^{n} (\hat{p} - p)^2 \binom{n}{x} (1 - (1-p)^k)^{x_i} ((1-p)^k)^{n-x_i} \tag{2.54}$$

$$MSE(\hat{p}) = \sum_{x=0}^{n} (1 - (1 - \frac{x}{n})^{\frac{1}{k}} - p)^2 \binom{n}{x} (1 - (1-p)^k)^{x_i} ((1-p)^k)^{n-x_i} \tag{2.55}$$

$$MSE(\hat{p}) = (1-p)^2 + \sum_{x=0}^{n} (1 - \frac{x}{n})^{\frac{1}{k}} [(1 - \frac{x}{n})^{\frac{1}{k}} - 2(1-P)] \binom{n}{x} (1-(1-p)^k)^x ((1-p)^k)^{n-x} \tag{2.56}$$

Clearly the optimal value of group size depends on the unknown value of the true proportion. In practise some prior value for the true proportion(P) say p_0 can be used to determine the group size that minimizes the MSE of P' evaluated when $P = p_0$. That is

$$k = k(n) = argmin_k MSE(\hat{p}; x, p_0, n) \tag{2.57}$$

$$k = k(n) = argmin_k (1-p_0)^2 + \sum_{x=0}^{n} (1-\frac{x}{n})^{\frac{1}{k}} [(1-\frac{x}{n})^{\frac{1}{k}} - 2(1-p_0)] \binom{n}{x} (1-(1-p_0)^k)^x ((1-p_0)^k)^{n-x} \tag{2.58}$$

Equating the first derivative of MSE with respect to k to 0 and solving for k leads to non-existence k.This is due to the fact that the MSE is typically a convex function and strictly increasing, thus it is quite reasonable to assume that the function has a derivative that vanishes at most once.
Using non-derivative optimisation methods one can be able to obtain the optimal group sizes(k) for given prior proportions and the number of groups(n).

The table 2.7 gives the following range of combination of p and n:k the value of k is optimal in the sense that it minimises $\mathbb{MSE}(p)$; the $\mathbb{MSE}$ of $\hat{p}$ when k=k* $[\mathbb{MSE}(\hat{p}; k^*)]$ and for comparison the $\mathbb{MSE}$ of $\hat{p}$ when k=1 $[\mathbb{MSE}(\hat{p}; 1)]$ For example when p=0.04 and n=20 table 2.7 indicates that the optimal group size is 19 for which the $\mathbb{MSE}(\hat{p}) = 0.000180$, if k=1 had been used instead , $\hat{p}$ would have had $\mathbb{MSE}(\hat{p}) = 0.001920$,which is 10.67 times the minimum value realized with optimal k. Thus using the optimal group size group testing can be more precise than individual testing. Also, a general observation can be made from table 2.7, that the group sizes decrease as p increases.

From the above two tables, table 2.6 and 2.7, it can observed that, minimal k due to MSE depends on n (the number of groups) and as n increases the group size are always less than the value of optimal k obtained from the asymptotic variance.

4. **Hughes-Oliver and Swallow 1994**

P		n							
		10	20	30	50	60	80	100	
0.01	k	34	65	88	119	121	134	143	
	MSE(p,k)	0.000046	0.000013	0.000007	0.000004	0.000003	0.000002	0.0000016	0.0
	MSE(p,1)	0.000990	0.000495	0.000330	0.000198	0.000165	0.000124	0.000099	0.
0.02	k	19	36	47	62	64	69	71	
	MSE(p,k)	0.000162	0.000048	0.000027	0.000014	0.000011	0.000008	0.0000065	0.0
	MSE(p,1)	0.001960	0.000980	0.000653	0.000391	0.000327	0.000245	0.000196	0.
0.03	k	14	25	30	40	43	45	48	
	MSE(p,k)	0.000337	0.000104	0.000059	0.000031	0.000025	0.000018	0.000015	0.
	MSE(p,1)	0.002910	0.001455	0.000970	0.000582	0.000485	0.000364	0.000291	0.
0.04	k	11	19	25	30	43	35	35	
	MSE(p,k)	0.000565	0.000180	0.000102	0.000055	0.000045	0.000032	0.000026	0.
	MSE(p,1)	0.003840	0.001920	0.0001280	0.000768	0.000640	0.000480	0.000384	0.
0.05	k	9	16	20	25	25	25	30	
	MSE(p,k)	0.000842	0.000274	0.000157	0.000084	0.000069	0.000050	0.000040	0.
	MSE(p,1)	0.004750	0.002375	0.001583	0.000950	0.0000792	0.000594	0.000475	0.
0.06	k	8	13	17	21	22	23	23	
	MSE(p,k)	0.001158	0.000385	0.000222	0.000120	0.000098	0.000071	0.000056	0.
	MSE(p,1)	0.005640	0.002820	0.001880	0.001128	0.000940	0.000705	0.000564	0.
0.08	k	6	10	13	16	16	17	17	
	MSE(p,k)	0.001922	0.000656	0.000382	0.000208	0.000170	0.000124	0.000097	0.
	MSE(p,1)	0.007360	0.003680	0.002453	0.001472	0.001227	0.000920	0.000736	0.
0.10	k	5	8	10	12	13	13	14	
	MSE(p,k)	0.002807	0.000987	0.000579	0.000317	0.000258	0.000189	0.000149	0.
	MSE(p,1)	0.009000	0.004500	0.003000	0.001800	0.001500	0.001125	0.000900	0.
0.15	k	4	6	7	8	8	9	9	
	MSE(p,k)	0.005409	0.002014	0.001202	0.000665	0.000544	0.000398	0.000314	0.
	MSE(p,1)	0.012750	0.006375	0.004250	0.002550	0.002125	0.001594	0.001275	0.
0.20	k	3	4	5	6	6	6	7	
	MSE(p,k)	0.008356	0.003284	0.001975	0.001100	0.000901	0.000662	0.000523	0.
	MSE(p,1)	0.016000	0.008000	0.005333	0.003200	0.002667	0.002000	0.001600	0.
0.25	k	3	3	4	5	5	55	5	
	MSE(p,k)	0.012089	0.004735	0.002831	0.001597	0.001306	0.000959	0.000757	0.
	MSE(p,1)	0.018750	0.009375	0.006250	0.003750	0.003125	0.002344	0.001875	0.
0.3	k	2	3	3	4	4	4	4	
	MSE(p,k)	0.014516	0.006000	0.003769	0.002114	0.001733	0.001276	0.001009	0.
	MSE(p,1)	0.021000	0.010500	0.007000	0.004200	0.003500	0.002625	0.002100	0.

Table 2.7: The optimal group sizes

Having noted in their previous works that the benefits of group testing both economical and reduction of MSE are sensitive to the choice of group size and the assumed prior proportion . Hughes and swallow proposed adapting the group size from time to time throughout the testing phase using all the accumulated data and obtaining a final estimate P (proportion) based on all the data collected.

The adaptive estimator is obtained by testing groups in stages and updating the group size from one stage to the next. This general scheme allows the number of groups N for each stage to be arbitrary but known before the experiment begins while group sizes k_i are determined sequentially during the experiment. The challenge comes in deciding how one should update the group size in a manner that will lead to using a group size as close to the optimal group size as possible.

They presented an update based on the MLE of P (the proportion) obtained using the data from the previous stage. The group size selected is the one that minimises the MSE of the estimate of P that would be obtained if only the data from the next stage were to be used. With P replaced in the MSE formula by the most recent MLE of P. The group size of the for the first stage is still based on a prior value p_0 of P.

They limited the description and the discussion to two stages partly fro simplicity and partly because asymptotic results suggest that two stages yield optimal group size as possible.

To enable direct comparison with the non adaptive estimator a total of N test will be performed in two stages such that $N_1 = \lambda N$ tests are performed in the first stage and $N_2 = 1 - \lambda N$ tests are performed in the second stage, where λ is assumed to be known before the experiment begins.

The group size for the first stage is based on some prior value p_0 for the true P and the number of tests to be performed in the first stage of size k_1 calculated as

$$k_1 = argmin[\frac{1-(1-p_0)^k}{k^2(1-p_0)^{k-2}}]. \tag{2.59}$$

And then proceed to test λN groups each of size k_1 for the trait of interest. The

number of groups with the trait X_1 has a binomial distribution with parameters λN and $1-(1-p)^{k_1}$. Thus the intermediate MLE of P is determined as

$$\hat{p_1} = p_1\hat{(x_1)} = 1 - [1 - \frac{X_1}{\lambda N}] \tag{2.60}$$

The group of the second is then obtained as

$$k_2 = k_2(x_1) = argmin[\frac{1-(1-\hat{p_1})^k}{k^2(1-\hat{p_1})^{k-2}}] \tag{2.61}$$

That is, for each realization of $X_1 = x_1$ k_2 minimizes the MSE of an estimate of P, where the true MSE in the MSE formula replaced by $p_1\hat{(x_1)}$. Hence k_2 is a random variable that derives its randomness from X_1. stage two proceeds by testing $(1-\lambda)N$ groups, each of size $k_2(x_1)$. Let X_2 is the number of these groups showing the trait. Hence conditioned on X_1, X_2 has a binomial distribution. Specifically, $X_2|X_1 = x_1$ has a binomial distribution with parameters $(1-\lambda)N$ and $1-(1-p)^{k_2}$.

$$f(X_2|X_1) = \binom{(1-\lambda)N}{x_2}(1-(1-p)^{k_2})^{x_2}|((1-p)^{k_2})^{n_2-x_2}$$

The final two stage adaptive estimator of P, $\hat{p}_A$ is the MLE based on the joint distribution of X_1 and X_2.

From the bivariate distribution

$$P(X = x, Y = y) = P(X = x|Y = y) \times P(Y = y)$$

Thus

$$f(X_1, X_2) = f(X_2|X_1) * f(X_1) \tag{2.62}$$

Letting $\lambda N = n_1$ and $(1-\lambda)N = n_2$

$$f(X_1, X_2|P) = \binom{n_1}{x_1}(1-(1-p)^{k_1})^{x_1}|((1-p)^{k1})^{n_1-x_1}\times \binom{n_2}{x_2}(1-(1-p)^{k_2})^{x_2}|((1-p)^{k_2})^{n_2-x_2} \quad (2.63)$$

The log likelihood is therefore

$$lnL(X_1, X_2|P) = ln\binom{n_1}{x_1} + ln\binom{n_2}{x_2} + x_1 ln\{\frac{1-(1-p)^{k_1}}{(1-p)^{k_1}}\} + x_2 ln\{\frac{1-(1-p)^{k_2}}{(1-p)^{k_2}}\} + [n_1k_1 + n_2k_2]ln(1-p) \quad (2.64)$$

$$\frac{dlnL(X_1, X_2|P)}{dp} = \frac{x_1k_1}{1-(1-p)^{k_1}} + \frac{x_1k_1}{1-(1-p)^{k_1}} - \frac{n_1k_1 - n_2k_2}{(1-p)} \quad (2.65)$$

Equating the derivative with zero and replacing back the $\lambda N = n_1$ and $(1-\lambda)N = n_2$

Thus the MLE is the solution to

$$\frac{k_1x_1}{1-(1-p)^{k_1}} + \frac{k_2x_2}{1-(1-p)^{k_1}} = \lambda N k_1 + (1-\lambda)Nk_2 \quad (2.66)$$

And the MSE can be calculated as

$$MSE(\hat{p}_A) = E_{X1}\{E_{X2}[(\hat{p}_A - P)^2]|X_1\} \quad (2.67)$$

To assess the asymptotic behaviour of $\hat{p}_A$ one must first assess the asymptotic behaviour of the non random $k_1(N)$ and the random $k_2(X_1N)$. The second stage group size approaches the asymptotically optimal group size regardless of the value of the initial p_0 if the proportion estimates are consistent.

Let λ,N and p_0 be given. If the group size k_1 and k_2 are chosen as described earlier, and if $\hat{p}_A$ is the solution to 2.66 then $\hat{p}_A$ is strongly consistent for P and asymptotically $\hat{p}_A$ has a normal distribution with mean P and variance equal to the reciprocal of the Fishers information using the fact that $E(x_1) = \lambda N(1-(1-p)^{k_1})$

and $E(x_2) = (1-\lambda)N(1-(1-p)^{k_2})$

$$\sqrt{N(\hat{p}_A - P)} \approx Normal(0, \frac{\lambda k_1^2 (1-p)^{k_1-2}}{1-(1-p)^{k_1}} + \frac{(1-\lambda)k_2^2(1-p)^{k_2-2}}{1-(1-p)^{k_2}}) \qquad (2.68)$$

Extending these derivations to more than two stages yields exactly the same asymptotic distribution for $\hat{p}_A$.This is due to the fact that regardless of the number of stages involved, the group size for all but the first stage approaches the optimal k as N approaches infinity. It is in this sense that the two stage adaptive estimator is considered optimal compared to a three stage estimator and any stage estimator.

Robustness of the group size to the choice of prior proportion

Whether the group size is obtained with method 1, 2 or 3 some initial value for the proportion is needed. Since all these methods require specifying a value of p,one way or another in the process of choosing k. Then all these methods would be implemented with p_0 in place of p.

If p_0 is very far from the true P, the group sizes obtained p_0 can be very different from those obtained using the true proportion.

Method 1 and 2 share the questionable property that the recommended group size remains the same , regardless of the number of tests n being performed. Method three recommends different group size for different values of n.

Method 1 is based , namely on equalizing the absolute influence of a group showing the attribute and the influence of the group not showing the attribute, has some appeal but seems to be less practical interest than the criteria behind method 2 and 3. Comparing method 2 and 3 , although the asymptotic variance might be an adequate approximation to the mean squared error when n is large it might be misleading when n is small.

Asymptotically, method 1 and 2 asymptotic distribution is unchanged, this is because the two methods yield group sizes that are independent of the value of n. In contrast, method 3,minimising the exact mean squared error, generally yields a different group size for each different value of n. Thus its behaviour as n tends to infinity is clear of interest. Under reasonable assumptions, the following theorem provided a simple limiting property

of the group sizes yielded from the 3rd method.

Theorem

If the first derivative of $MSE(p_0, k, n)$ *with respect to k equals zero at most once,* $k_2 < \infty$ *and* $k_3(n) \geq 1$ *for all* $n > n_p$ *then* $\lim_{n\to\infty} k_3(n) = k_2$

This result says that under mild conditions, as n tends to infinity the group sizes obtained from minimizing the mean squared error converge to the group sizes obtained from minimising the asymptotic variance. This implication mildly suggests that the mean squared error $\mathbb{MSE}$ converges to the asymptotic variance in probability as n tends to infinity

Convergence in probability

Definition: A sequence of random variables $X_1, X_2, ..., X_n$converges in probability to X if

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

for all $\varepsilon > 0$

Indeed this (that the mean squared error $\mathbb{MSE}$ converges to the asymptotic variance in probability as n tends to infinity)is expected but not immediate result that has the following desirable implication. Although for each different value of n k_3 has the potential to be different, the asymptotic distribution of the estimator $\hat{p}$ has been the same form as in method 2.

Moreover, as n tends to infinity, $k_{3_0}(n)$ does not tend to k_2 rather it tends to k_{2_0}. That is,the effect of the initial p_0 persist even in the limit.

To illustrate the mildness of Theorem above reflabel shows the typical shape of the $\mathbb{MSE}$ as a function of k. reflabel shows the same function plotted against $\log k$ for $1 \leq k \geq 120$. in both figures p=0.01 n=50. The function is not convex, but it is convex in a region containing the minimum. Outside this area the function is strictly increasing.

Having established that Method 2 and 3 are asymptotically equivalent then it follows that one would prefer the two methods over method 1 in an asymptotic sense.However, unless the n is large enough so that $MSE(\hat{p})$ approaches the asymptotic variance fairly closely the value of k that minimises the asymptotic variance maybe highly inefficient for estimating the proportion.

To examine the robustness of the group size yielded from methods 3 above we construct the tables for several choices of p, p_0 and n. For each value of p, the values of p_0 shown are p_0=0.5p, p, 1.5p, 2p. the tables shows the value of k (group size) for these choices as well as th true mean squared error of $\hat{p}$ obtained with these different group sizes.

P	p_0		n						
			10	20	30	50	100	150	200
0.01	0.005	k	58	120	166	230	286	296	301
		MSE(p,k)	0.000307	0.000788	0.00185	0.00521	0.00289	0.000369	0.000473
	0.01	k	34	66	89	119	143	148	150
		MSE(p,k)	0.000046	0.000013	0.000069	0.000036	0.000016	0.000010	0.000008
	0.015	k	24	48	61	81	95	98	100
		MSE(p,k)	0.000054	0.000015	0.000008	0.000004	0.000002	0.000001	0.0000008
	0.02	k	19	36	47	62	71	73	75
		MSE(p,k)	0.000065	0.000018	0.000010	0.000005	0.000002	0.000001	0.0000009
0.02	0.01	k	34	65	88	119	143	148	150
		MSE(p,k)	0.000900	0.00193	0.00390	0.00868	0.00300	0.000391	0.000054
	0.02	k	19	36	47	62	71	73	75
		MSE(p,k)	0.000193	0.000048	0.000027	0.000014	0.000006	0.000004	0.000003
	0.015	k	14	25	32	42	47	49	50
		MSE(p,k)	0.000189	0.000057	0.000031	0.000015	0.000007	0.000005	0.000003
	0.02	k	11	19	25	31	35	36	37
		MSE(p,k)	0.000225	0.000068	0.000036	0.000018	0.000008	0.000005	0.000004
0.05	0.025	k	16	29	38	50	57	59	60
		MSE(p,k)	0.00352	0.00587	0.00969	0.015100	0.003320	0.000486	0.000091
	0.05	k	9	16	20	25	28	29	30
		MSE(p,k)	0.000840	0.000274	0.000157	0.000084	0.000039	0.000025	0.000019
	0.075	k	7	11	14	17	19	19	19
		MSE(p,k)	0.000964	0.000317	0.000178	0.000094	0.000043	0.000028	0.000021
	0.01	k	5	8	10	12	14	14	14
		MSE(p,k)	0.001140	0.000375	0.000210	0.000109	0.000050	0.000033	0.000024
0.10	0.05	k	9	16	20	25	28	29	29
		MSE(p,k)	0.009320	0.012800	0.017900	0.020500	0.004010	0.000727	0.000205
	0.10	k	5	8	10	12	14	14	14
		MSE(p,k)	0.002790	0.000983	0.000578	0.000317	0.000148	0.000097	0.000072
	0.15	k	4	6	7	8	9	9	9
		MSE(p,k)	0.003160	0.001130	0.000653	0.000352	0.000164	0.000107	0.000079
	0.20	k	3	4	5	6	6	7	7
		MSE(p,k)	0.003690	0.001330	0.000771	0.000414	0.000193	0.000125	0.000093

Table 2.8: Group sizes and their associated mean squared error for different prior p.

Taking a closer look on the sensitivity of the group size to the choice of the prior proportion from table 2.8, for example when true proportion p=0.01, n=20 the optimal (minimises the $\mathbb{MSE}$ = 0.000013) group size k=66. If the researcher prior information suggest that the proportion (prior) is $p_0 = 0.005$ then he/she will end up with the optimal group size k=120 with the true MSE=0.000806 which is 60 times the minimal MSE. On the other hand if the researcher prior information suggest that the proportion (prior)is $p_0 = 0.02$, he/she will end up with the optimal group size k=36 with the true MSE=0.000018 which is only 1.38 times the optimal MSE.

Therefore, if $p_0 < p$ we end up choosing a larger group than when $p = p_0$ which exaggerates the true MSE compromising the precision of the estimator. However, if $p_0 > p$ we end up with the a group size that is less than the optimal group size $p = p_0$ which although it increases the true MSE it is still applicable. Hence we can conclude that it is safer to overestimate the prior than to underestimate it. Although, ideally, one would like p_0 as close to the true p as possible , if in-doubt take p_0 too large for a conservative choice of k.

2.3 Bayesian Estimates

Bayesian estimation starts with an assumed probability distribution of the parameters space. With this approach the parameter itself is a random variable and the observations are conditionally independent given the parameters.

The assumed p.d.f of the parameter is called the prior distribution, together the prior and the parametric family give the joint distribution. The joint distribution is used to find the marginal distribution of the observation and by the use of Bayes Theorem the conditional distribution of the parameter given the observation which is called the posterior distribution is computed. Bayes Theorem For Events A and B

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \tag{2.69}$$

Thus with the appropriate densities we can write the bayes formula as

$$f(\theta/x) = \frac{f(x/\theta)f(\theta)}{f(x)} \tag{2.70}$$

2.3.1 Bayes Estimator

In this case, the observations are said to be binomial distributed while the parameter P is assumed to be Beta distributed.

$$f(x/p) = \binom{n}{x}(1-(1-p)^k)^x((1-p)^k)^{n-x} \tag{2.71}$$

while

$$f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \tag{2.72}$$

By use of bayes theorem

$$f(x,p) = \binom{n}{x}B(\alpha,\beta)^{-1}p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x \tag{2.73}$$

The marginal of x can therefore be found by

$$f(x) = \int_0^1 \binom{n}{x} B(\alpha, \beta)^{-1} p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x \mathrm{d}p \tag{2.74}$$

using of the binomial theorem

$[1-(1-p)^k]$ becomes

$\sum_{j=0}^{x} \binom{x}{j}(-1)^j(1-p)^{kj}$

$$f(x) = \binom{n}{x} B(\alpha, \beta)^{-1} \sum_{j=0}^{x} \binom{x}{j}(-1)^j \int_0^1 p^{\alpha-1}(1-p)^{kn+kj-kx+\beta-1}\mathrm{d}p \tag{2.75}$$

$$f(x) = \binom{n}{x} B(\alpha, \beta)^{-1} \sum_{j=0}^{x} \binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta) \tag{2.76}$$

The posterior of p can therefore be evaluated as

$$f(p/x) = \frac{p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x}{\sum_{j=0}^{x} \binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta)} \tag{2.77}$$

Then the Bayes estimator of P is given by

$$\hat{p}_1 = E(p/x) = \int_0^1 p f(p/x)\mathrm{d}p \tag{2.78}$$

$$\hat{p}_1 = \int_0^1 p \frac{p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x}{\sum_{j=0}^{x} \binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta)}\mathrm{d}p \tag{2.79}$$

$$\hat{p}_1 = \int_0^1 \frac{p^{\alpha}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x}{\sum_{j=0}^{x} \binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta)}\mathrm{d}p \tag{2.80}$$

$$\hat{p}_1 = \frac{\sum_{j=0}^{x} \binom{x}{j}(-1)^j B(\alpha+1, kn+kj-kx+\beta)}{\sum_{j=0}^{x} \binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta)} \tag{2.81}$$

However, Beta distribution can have a number of shapes. The prior should correspond to your belief about the proportion.

The figure 2.7 shows the variety of shapes a beta distribution can take. When $\alpha < \beta$ the density has more weight on the lower half and when $\alpha > \beta$ the density has more weight on the upper half. When $\alpha = \beta$ the distribution is symmetric.

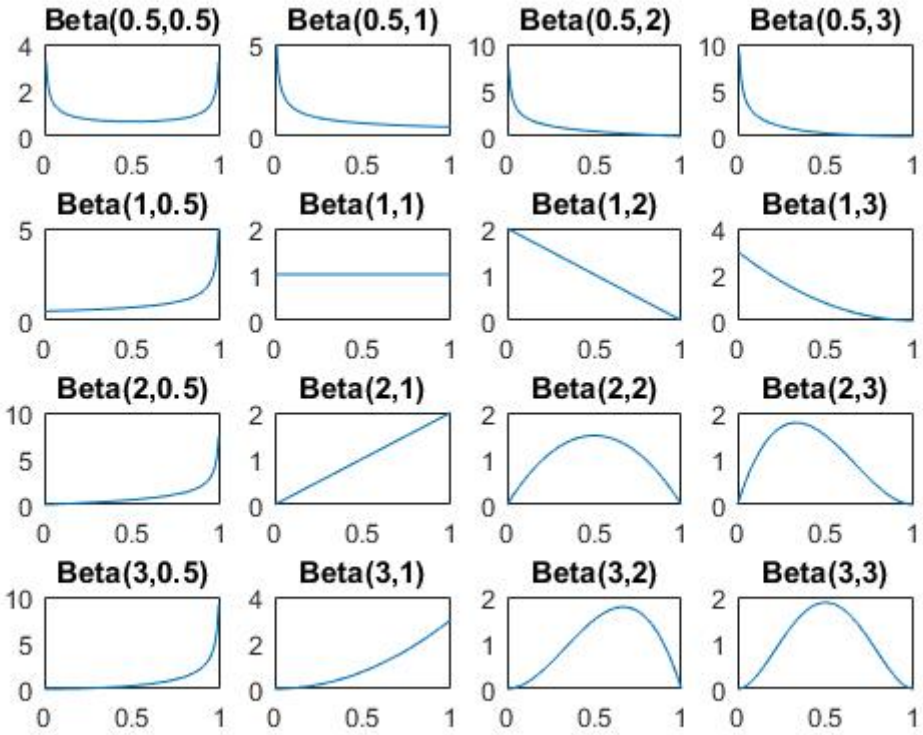

Figure 2.7: Different shapes of beta distribution

The prior beta distribution chosen should correspond to your belief about the proportion. That is a ditribution that matches the location (mean) and scale (standard deviation) of the believed proportion.

For instance if the prior mean of the proportion is P_0' and let σ_0 be the prior standard deviation for the proportion. Then equating these with the mean ($\frac{\alpha}{\alpha+\beta}$)and the standard deviation($\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$) respectively.

We obtain $P_0' = \frac{\alpha}{\alpha+\beta}$ and $\sigma_0 = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$

bearing in mind that $1 - P_0' = \frac{\beta}{\beta+\alpha}$ then $\sigma_0 = \sqrt{\frac{P_0'(1-P_0')}{(\alpha+\beta+1)}}$

solving these two equations for α and β one can obtain the beta parameters of the prior distribution. If we do not have any idea before hand what the proportion P is we might be like to chose a prior that does not favour any one value over the other. In this case w should use the uniform prior that gives equal weight to to all possible values of the proportion.

2.3.2 Empirical Bayes estimator

Empirical Bayesian estimation is a special type of Bayesian estimation which instead of fixing(making an assumption) the prior distribution it is estimated from the data.
Since group testing is only economically viable if the proportion of the defectives is small, we would like to use a family of prior distributions appropriate for small p. the Beta$(1, \beta)$ is such a family. this is because for large values of of β the majority f the probability distribution of the random variable p is close to zero.

Recall that

$$f(x/p) = \binom{n}{x}(1-(1-p)^k)^x((1-p)^k)^{n-x} \tag{2.82}$$

Now that

$$f(p) = \beta(1-p)^{\beta-1} \tag{2.83}$$

then

$$f(x,p) = \beta\binom{n}{x}(1-(1-p)^k)^x((1-p)^k)^{n-x}(1-p)^{\beta-1} \tag{2.84}$$

Thus the marginal of x

$$f(x|\beta) = \beta\binom{n}{x}\int_0^1 [1-(1-p)^x](1-p)^{kn-kx+\beta-1}\mathrm{d}p \tag{2.85}$$

Using the change of variable technique

Let $u = (1-p)^k$

which implies that $p = 1 - u^{\frac{1}{k}}$

$\mathrm{d}p = \frac{-1}{k}(1-p)^{1-k}\mathrm{d}u$

thus

$$f(x|\beta) = \beta k^{-1}\binom{n}{x}\int_0^1 u^{n-x+\frac{\beta}{k}-1}(1-u)^x\mathrm{d}u \tag{2.86}$$

$$f(x|\beta) = \beta k^{-1}\binom{n}{x}B(n-x+\frac{\beta}{k}, x+1) \tag{2.87}$$

$$f(x|\beta) = \beta k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\beta}{k}),)}{\Gamma(n+\frac{\beta}{k}+1)} \tag{2.88}$$

Maximizing the above equation with respect to β may be done numerically by solving

$$\frac{\partial}{\partial\beta}(f(x|\beta)) = 0$$

$$\frac{\partial}{\partial\beta}[\beta k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\beta}{k}),)}{\Gamma(n+\frac{\beta}{k}+1)}] \tag{2.89}$$

Taking the logarithms

$$\begin{aligned} 0 =& \frac{\partial}{\partial\beta}\{\log\beta + \log[\Gamma(n-x+\frac{\beta}{k})] - \log[\Gamma(n+\frac{\beta}{k}+1)]\} \\ &= \beta^{-1} + k^{-1}[\psi(n-x+\frac{\beta}{k}) - \psi(n+\frac{\beta}{k}+1)] \end{aligned} \tag{2.90}$$

Where ψ represents the digamma function.

Thus $\hat{\beta}$ is the solution to 3.31. which is the optimal β and therefore β can be replaced by $\hat{\beta}$.

Hence, the posterior of p can be given as

$$f(p/x) = \frac{\hat{\beta}\binom{n}{x}(1-(1-p)^k)^x|((1-p)^k)^{n-x+\hat{\beta}-1}}{\hat{\beta}k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\hat{\beta}}{k}),)}{\Gamma(n+\frac{\hat{\beta}}{k}+1)}} \tag{2.91}$$

$$f(p/x) = k\frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}(1-(1-p)^k)^x|((1-p)^k)^{n-x+\hat{\beta-1}} \tag{2.92}$$

With $f(p/x)$ and a given a loss function , say $L(p,a)$ the empirical estimate of P' with respect to $L(p,a)$ is the value of a that minimizes

$$E[L(p,a)] = \int_0^1 L(p,a)f(p/x)dp \tag{2.93}$$

Considering the general loss function

$$L(p,a) = w(p)[p-a]^2$$

The bayes estimator of p corresponding to this loss function is given by

$$\begin{aligned}\hat{p}_{eb} =& \frac{E[w(p)pf(p|x)]}{E[w(p)f(p|x)]} \\ =& \frac{\int_0^1 w(p)pf(p|x)dp}{\int_0^1 w(p)f(p|x)dp}\end{aligned}$$

Empirical Bayes Estimate using Loss function $L_1(p,a) = (p-a)^2$

Where $L(p,a) = (p-a)^2$ is the squared error loss, thus the mean of the empirical posterior is the empirical estimator.

This is a special case with w(p)= 1. Thus the loss function is

$$L_1(p,a) = (p-a)^2$$

Therefore, the Bayes estimator of p with respect to the squared loss function is the posterior mean E[p/x]

$$\hat{p}_{eb} = \int_0^1 k\frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}p(1-(1-p)^k)^x|((1-p)^k)^{n-x+\hat{\beta-1}}dp \tag{2.94}$$

Using the change of variable

Let $u = (1-p)^k$

which implies that $p = 1-u^{\frac{1}{k}}$

$\mathrm{d}p = \frac{-1}{k}(1-p)^{1-k}\mathrm{d}u$

One obtains a closed form of expression for the posterior mean as

$$\hat{p}_{eb1} = 1 - \frac{\Gamma(n+\frac{\beta}{k}+1)\times\Gamma(n-x+\frac{\beta}{k}+1)}{\Gamma(n-x+\frac{\beta}{k})\times\Gamma(n+\frac{\beta}{k}+1+1/k)} \tag{2.95}$$

Empirical Bayes Estimate Using Loss Function $L_2(p,a) = \dfrac{(p-a)}{p(1-p)}$

Let

$$w(p) = \frac{1}{p(1-p)}$$

Thus the empirical estimator is obtained as follows

$$\hat{p}_{eb2} = \frac{E[\frac{1}{(1-p)}f(p|x)]}{E[\frac{1}{p(1-p)}f(p|x)]} = \frac{\int_0^1[\frac{1}{(1-p)}]f(p|x)dp}{\int_0^1[\frac{1}{p(1-p)}]f(p|x)dp}$$

$$\int_0^1[\frac{1}{(1-p)}]f(p|x)dp = \int_0^1 k\frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}(1-(1-p)^k)^x((1-p)^k)^{n-x+\hat{\beta-}2} \tag{2.96}$$

Let $u=(1-p)^k$. It follows that $p=1-u^{\frac{1}{k}}$ and $dp=-u^{1/k-1}\frac{dk}{k}$

$$= \frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}\int_0^1(1-u)^x u^{n-x+(\hat{\beta}/k)-(1/k)-1}du \tag{2.97}$$

$$= \frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}\frac{\Gamma(x+1)\Gamma(n-x+\frac{\hat{\beta}}{k}-\frac{1}{k})}{\Gamma(n+\frac{\hat{\beta}}{k}-\frac{1}{k}+1)} \tag{2.98}$$

$$= \frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)\Gamma(n-x+\frac{\hat{\beta}}{k}-\frac{1}{k})}{\Gamma(n-x+\frac{\hat{\beta}}{k})\Gamma(n+\frac{\hat{\beta}}{k}-\frac{1}{k}+1)} \tag{2.99}$$

On the other hand

$$\int_0^1[\frac{1}{p(1-p)}]f(p|x)dp = \int_0^1\frac{k\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}p^{-1}(1-(1-p)^k)^x((1-p)^k)^{n-x+\hat{\beta-}2} \tag{2.100}$$

Let $u=(1-p)$. it follows that $p=1-u$ and $dp=-du$

$$= \frac{k\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}\int_0^1\frac{(1-u^k)u^{k(n-x)+\hat{\beta}-2}}{(1-u)}du \tag{2.101}$$

We know that $1-u^k = \sum_{i=0}^{k-1} u^i(1-u)$

$$= \frac{k\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)} \int_0^1 \sum_{i=0}^{k-1} (1-u^k)^{x-1} u^{i+k(n-x)+\hat{\beta}-2} du \tag{2.102}$$

Further, we let $u^k = q$. It follows that $u = q^{\frac{1}{k}}$ and $du = q^{\frac{1}{k}-1}\frac{dq}{k}$

$$= \frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)} \int_0^1 \sum_{i=0}^{k-1} q^{n-x+(i/k)-(1/k)-1}(1-q)^{x-1} dq \tag{2.103}$$

$$= \frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{x\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)} \sum_{i=0}^{k-1} \frac{\Gamma(n-x+i/k+\frac{\hat{\beta}}{k}-1/k)}{\Gamma(n-i/k+\frac{\hat{\beta}}{k}-(1/k)} \tag{2.104}$$

Thus

$$\begin{aligned}
\hat{p}_{eb2} =& \frac{E[\frac{1}{(1-p)}f(p|x)]}{E[\frac{1}{p(1-p)}f(p|x)]} \\
&= \frac{\frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)\Gamma(n-x+\frac{\hat{\beta}}{k}-\frac{1}{k})}{\Gamma(n-x+\frac{\hat{\beta}}{k})\Gamma(n+\frac{\hat{\beta}}{k}-\frac{1}{k}+1)}}{\frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{x\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)} \sum_{i=0}^{k-1} \frac{\Gamma(n-x+i/k+\frac{\hat{\beta}}{k}-1/k)}{\Gamma(n-i/k+\frac{\hat{\beta}}{k}-(1/k)}} \\
&= \frac{x\Gamma(n-x+\frac{\hat{\beta}}{k})-\frac{1}{k}}{\Gamma(n+\frac{\hat{\beta}}{k}-\frac{1}{k}+1)\sum_{i=0}^{k-1} \frac{\Gamma(n-x+i/k+\frac{\hat{\beta}}{k}-1/k)}{\Gamma(n-i/k+\frac{\hat{\beta}}{k}-(1/k)}}
\end{aligned}$$

The difference between loss functions $L_1(p,a) = (p-a)^2$ and $L_2(p,a) = \frac{(p-a)}{p(1-p)}$ is that $w(p) = 1$ for the first one and $w(p) = \frac{1}{p(1-p)}$ for the second one. Therefore, for any $p\epsilon(0,1)$, $L_1(p,a) = (p-a)^2$ has a constant weight 1 and $L_2(p,a) = \frac{1}{p(1-p)}$ increases the weight for the loss $(p-a)^2$ as $p \to 0$ since group testing concerns only small p., it is more appropriate to increase the weight for the loss as p is small.

Empirical Bayes Estimate Using Loss Function $L_3(p,a) = \dfrac{(p-a)^2}{p}$

Thus $w(p) = \frac{1}{p}$. In this case $L_3(p,a) = \dfrac{(p-a)^2}{p}$ increases the weight monotonically as p decreases, so it does note have the problem as $L_2(p,a)$.

Thus the estimator

$$\begin{aligned}\hat{p}_{eb3} =& \frac{E[f(p|x)]}{E[\frac{1}{p}f(p|x)]} \\ =& \frac{\int_0^1 f(p|x)dp}{\int_0^1 \frac{1}{p}f(p|x)dp}\end{aligned}$$

Given that $f(p|x)$ is a probability density function. Then $\int_0^1 f(p|x)dp = 1$

On the other hand

$$\int_0^1 \frac{1}{p}f(p|x)dp = \int_0^1 \frac{k\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}p^{-1}(1-(1-p)^k)^x((1-p)^k)^{n-x+\hat{\beta-}1} \quad (2.105)$$

Let $u = (1-p)$. it follows that $p = 1-u$ and $dp = -du$

$$= \frac{k\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}\int_0^1 \frac{(1-u^k)u^{k(n-x)+\hat{\beta}-1}}{(1-u)}du \quad (2.106)$$

We know that $1-u^k = \sum_{i=0}^{k-1} u^i(1-u)$

$$= \frac{k\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}\int_0^1 \sum_{i=0}^{k-1}(1-u^k)^{x-1}u^{i+k(n-x)+\hat{\beta}-1}du \quad (2.107)$$

Further, we let $u^k = q$. It follows that $u = q^{\frac{1}{k}}$ and $du = q^{\frac{1}{k}-1}\frac{dq}{k}$

$$= \frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}\int_0^1 \sum_{i=0}^{k-1} q^{n-x+(i/k)-(1/k)}(1-q)^{x-1}dq \quad (2.108)$$

$$= \frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{x\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}\sum_{i=0}^{k-1}\frac{\Gamma(n-x+i/k+\frac{\hat{\beta}}{k}-1/k)}{\Gamma(n-i/k+\frac{\hat{\beta}}{k})} \quad (2.109)$$

$$\hat{p}_{eb3} = \frac{E[f(p|x)]}{E[\frac{1}{p}f(p|x)]}$$

$$= \frac{x}{\frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)} \sum_{i=0}^{k-1} \frac{\Gamma(n-x+i/k+\frac{\hat{\beta}}{k}-1/k)}{\Gamma(n-i/k+\frac{\hat{\beta}}{k})}}$$

Chapter 3

Interval Estimation

3.1 Interval Based on the MLE

Interval estimation is the use of sample data to calculate an interval of possible or probable values of an unknown population parameter. In group testing the proportion of the attribute is the unknown population parameter from a binomial distribution with n trials and the probability of success as $1-(1-p)^k$, where k is the group size, compared to the point estimates of group testing there is less literature on interval testing.
Here are some proposed estimation procedure

3.1.1 Wald Intervals

Wald confidence intervals

1. Thompson (1962)

   Thompson proposed a Wald interval of P that is based on the exact variance of P'. Since the actual variance of the parameter P and $(Var(p))$ is not known Thompson used the variance of $\hat{p}(Var_e(p))$ to approximate it. Due to this approximation the students t distribution is used to construct the $100(1-\alpha)\%$ confidence interval.

   The first two moments of P' are given by

$$E(\hat{p})^r = \sum_{x=0}^{n}[1-(1-\frac{x}{n})^{1/k})]^r \binom{n}{x}[1-(1-p)^k]^x(1-p)^{k(n-x)} \tag{3.1}$$

for r=1,2. Thus the exact variance is given by $Var_e(p) = E(\hat{p})^2 - [E(\hat{p})]^2$. Thus the Thompson's proposed interval

$$\hat{p} \pm t_\alpha \sqrt{Var_e(p)} \tag{3.2}$$

Because $(Var_e(\hat{p}))$ is just a consistent estimate of $(Var(P))$, one might not expect this interval to perform well in situations where n is small.

Moreover, in the case where the $\hat{p}$ is small this interval produces a negative lower bound.

2. Bhattacharyya (1979)

   Bhattacharyya (1979) provided an Wald type of interval based on the asymptotic normal distribution of the estimator $\hat{p}$. This is based on the fact that the asymptotic distribution of $\hat{p}$ is normal as mentioned above.

$$\sqrt{n}(\hat{p} - p) \to Normal(0, \frac{1-(1-p)^k}{k^2(1-p)^{k-2}}) \tag{3.3}$$

   Confidence interval can be constructed using the estimator $\hat{p}$ and its variance multiplied by the appropriate quantile of the standard normal distribution.

$$\hat{p} \pm Z_{1-\alpha/2} / \sqrt{\frac{1-(1-p)^k}{k^2(1-p)^{k-2}}} \tag{3.4}$$

   Because of its computationally simplicity, this interval is often used in practise. However, it faces the same shortcoming as the Thompson's interval of producing negative endpoints.

3. Tebbs and Bilder 2004

   Due to the shortcomings of the above methods of estimating the interval, Tebbs and Bilder derived a variance-stablizing interval. Because the $Var(\hat{p})$ is a function of p we consider a transformation whose variance will be free of p.

   Suppose that g is a real valued differentiable function and consider the statistic

$g(\hat{p})$. A first order Taylor series expansion of $g(\hat{p})$about p is given by

$$g(\hat{p}) = g(p) + g'(p)(\hat{p} - p) \tag{3.5}$$

Because $g(\hat{p})$ is expressed as a linear function of $\hat{p}$ and p converges almost surely to P. It follow Slutky's Theorem that $\sqrt{n}[g(\hat{p}) - g(p)]$ converges to a N(0, g'(p) Var(p'))

Proof We first state the Slutsky's Theorem.

Theorem : If $W_n \to W$ in distribution and $Z_n \to C$ in probability where C is non random constant then

$$W_n Z_n \to CW in dis ribution$$

$$W_n + Z_n \to W + C in distribution$$

Then by **Delta Method** (A generalized central limit theorem)

Let Y_n be a sequence of random variables that satisfy $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ in distribution. For a given function and a specific value of θ , suppose that $g'(\theta)$exists and is not 0. Then,

$\sqrt{n}(g(Y_n) - g(\theta)) \to N(0, \sigma^2 g'(\theta)^2)$

Proof The Taylor expansion of $g'(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + Remainder$$

Where the remainder $\to 0$ as $Y_n \to \theta$. From the assumption that (Y_n) satisfies the standard CLT, We have $(Y_n \to \theta)$ in probability, so it follows that the remainder $\to 0$ in probability as well. Rearranging terms, we have

$$g(Y_n) = g'(\theta)(Y_n - \theta) + Remainder$$

Applying Slutsky's Theorem with $W_n = g'(\theta)(Y_n - \theta)$ and Z_n as the remainder, we

have the right-hand converging to $N(0, \sigma^2 g'(\theta)^2)$

Setting the $g'(p)Var(\hat{p})$ equal to a constant free of P say C_0

$$C_0 = g'(p)Var(\hat{p})$$

where $Var(\hat{p}) = \frac{1-(1-p)^k}{k^2(1-p)k-2}$

$$g'(p) = \sqrt{\frac{C_0 k^2(1-p)k-2}{1-(1-p)^k}} \tag{3.6}$$

Integrating both sides of the above differential equation, we get that

$$g(p) = k\sqrt{C_0} \int \frac{(1-p)^{\frac{k}{2}-1}}{\sqrt{1-(1-p)^k}} dp + C_1 \tag{3.7}$$

where C_1 is a constant free of P. Using a change of variable with $t = 1-(1-p)^k$, it follows that $p = 1-(1-t)^{\frac{1}{k}}$and that $dp = \dfrac{(1-t)^{1/k-1}}{k}$ Thus

$$\begin{aligned} g(P) &= \sqrt{C_0} \int \frac{1}{\sqrt{t(1-t)}} dt + C_1 \\ &= 2\sqrt{C_0}(arcsin\sqrt{t} + C_2) + C_1 \\ &= 2\sqrt{C_0}[arcsin\sqrt{1-(1-p)^k} + C_2] + C_1 \end{aligned} \tag{3.8}$$

where $C_2$is a constant without P.

Taking $C_0 = 0$ and $C_1 = C_2 = 0$, we get that

$$g(P) = 2arcsin\sqrt{1-(1-p)^k} \tag{3.9}$$

Following that $\sqrt{n}(g(Y_n) - g(\theta))$ converges to standard normal distribution

$$g(\hat{p}) \pm Z_{1-\alpha/2}/\sqrt{n} \tag{3.10}$$

serves as approximate $100(1-\alpha)\%$ confidence intervals for g(P). Setting a= g(P)

$$\begin{aligned}
\frac{a}{2} &= arcsin\sqrt{1-(1-p)^k} \\
sin^2(\frac{a}{2}) &= \sqrt{1-(1-p)^k} \\
1-(1-p)^k &= sin^2(\frac{a}{2}) \\
(1-p)^k &= 1-sin^2(\frac{a}{2}) \\
(1-p) &= [1-sin^2(\frac{a}{2})]^{1/k} \\
p &= 1-[1-sin^2(\frac{a}{2})]^{1/k}
\end{aligned}$$

Substituting $g(\hat{p}) \pm Z_{1-\alpha/2}/\sqrt{n}$ for a we arrive at the lower limit and the upper limit respectively of the approximate $100(1-\alpha)\%$ for P as

$$[1-(1-sin^2(a/2))^{1/k}, 1-(1-sin^2(b/2))^{1/k}]$$

Where $a = g(\hat{p}) - Z_{1-\alpha/2}/\sqrt{n}$ and $b = g(\hat{p}) + Z_{1-\alpha/2}/\sqrt{n}$ This interval have immediate advantages over the Wald intervals discussed above

(a) The standard error in the confidence interval calculation is free of estimates of P.

(b) The interval like the underlying finite distribution of $\hat{p}$ is not symmetric.

(c) Negative lower confidence limit are not possible.

This latter point is especially desirable in situations where p is small as is often case in group testing experiments.

3.1.2 Exact confidence intervals

Exact confidence intervals guarantee nominal confidence level. The number of positive groups follow a binomial distribution. But because of the discreteness of the binomial distribution they tend to be conservative. An exact confidence interval is derived from the exact test.

The construction of exact confidence intervals is on group scale, that is, a confidence interval $[\theta_l, \theta_u]$ of group proportion θ is first constructed using the estimator $\hat{\theta} = \frac{x}{n}$. This interval is then transformed in a second step to a confidence interval (P_l, P_u) for the individual probability by applying $\hat{p} = 1 - (1-p)^{1/k}$ on the confidence limits of $[\theta_l, \theta_u]$. The methodology of transferring the confidence intervals from the group scale to the individual scale assumes that the relation between p and θ is monotone for fixed values of k. This assumption is checked using figure 3.1 . Then, it is provided that positive difference in $[\theta_l, \theta_u]$ will result to a positive difference in (P_l, P_u). Because $\theta = x/n$ is

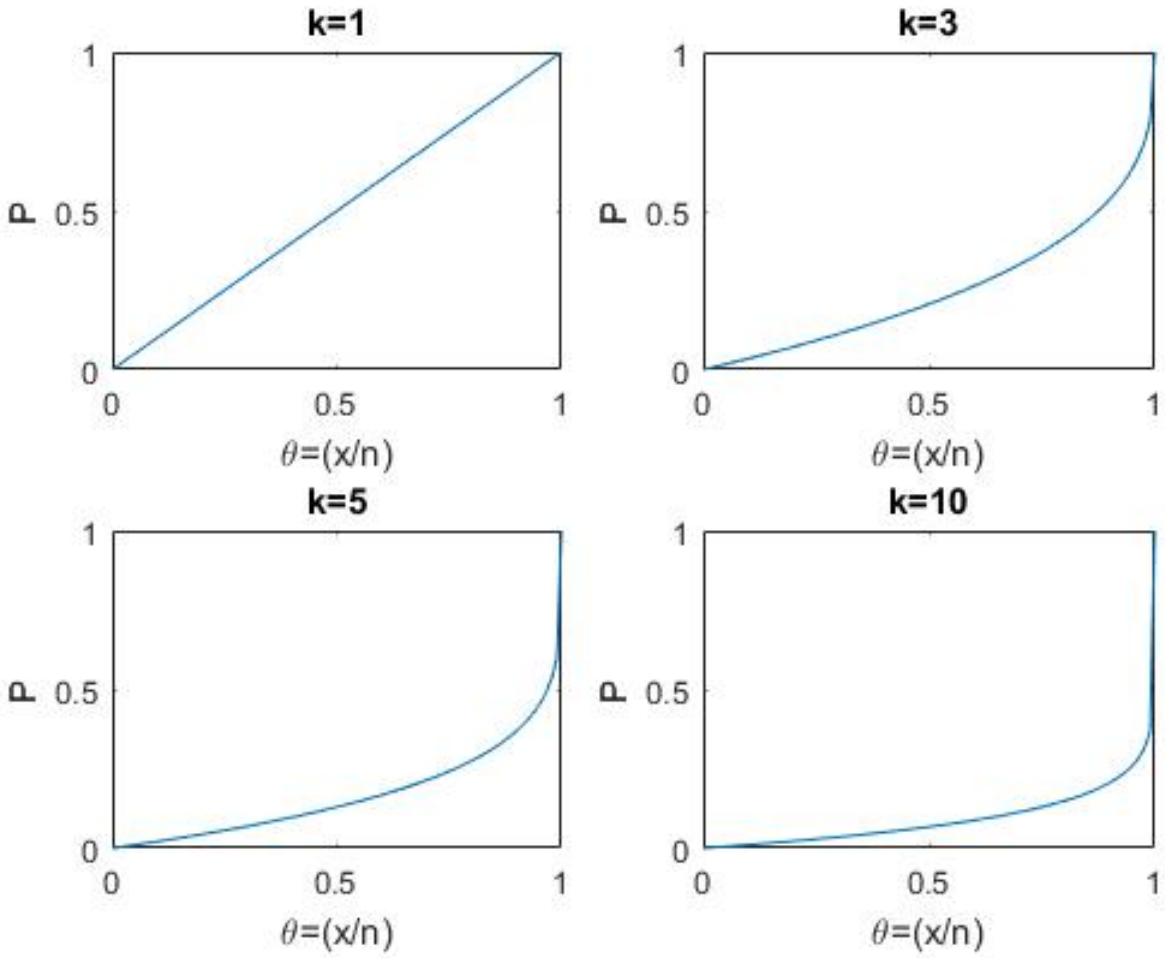

Figure 3.1: Illustrates the monotone relationship of $\hat{p} = 1 - (1-\theta)^{1/k}$ for different group sizes.

the estimator of a simple binomial proportion, the usual methods for construction of confidence limits for the binomial proportion can be applied this way. It follows that the

lower limit of the proportion of positive groups $\theta = (x/n)$ is

$$\sum_{r \leq x} \binom{n}{r} \theta_l^r (1 - \theta_l)^{n-r} = \frac{\alpha}{2} \tag{3.11}$$

And the corresponding upper limit is found

$$\sum_{r \geq x} \binom{n}{r} \theta_u^r (1 - \theta_u)^{n-r} = \frac{\alpha}{2} \tag{3.12}$$

The resulting confidence interval is central with equal probabilities in the tails of the distribution.

Once the limits of θ are found from the above equations, the corresponding limits for p can be found by substituting θ_l and θ_u for θ in $p = 1 - (1 - \theta)^{\frac{1}{k}}$.

3.2 Bayesian Interval Estimation

A Bayesian interval estimate is called a credible interval. Since in Bayesian estimation both the data and the parameter are are random , the credible set is based on the posterior distribution of the parameter.

Given the observation $X = x$ the interval $[P_l, P_u]$ is said to be a $100(1-\alpha)\%$ credible interval for P if the posterior probability of X being in

$$P(P_l \leq x \leq P_u) = 1 - \alpha$$

$$\int_{P_l}^{P_u} f(p|x, \alpha, \beta) = 1 - \alpha$$

In practise P_l and P_u may be determined using an equal tail or the Highest Posterior density (HPD) construction method.

The equal tail method is such that

$$\int_{0}^{P_l} f(p|x, \alpha, \beta) = \frac{\alpha}{2}$$

and

$$\int_{P_u}^{1} f(p|x, \alpha, \beta) = \frac{\alpha}{2}$$

The resulting interval (P_l, P_u) serves as the $100(1-\alpha)\%$ credible interval for P. This equal tail interval is preferred because it is invariant under transformation.

On the other hand, a $100(1-\alpha)\%$ HPD interval is a region that satisfies the following two conditions.

1. The posterior probability of the region is $100(1-\alpha)$.
2. The minimum density of any point within that region is equal to or larger than the density of any point outside that region.

The HPD is an interval in which most of the distribution lies. it is preferred because it is smallest interval.

3.2.1 Bayesian Credible Interval

In the case, the observations are said to be binomial distributed while the parameter P is assumed to be Beta distributed.

$$f(x/p) = \binom{n}{x}(1-(1-p)^k)^x((1-p)^k)^{n-x} \tag{3.13}$$

while

$$f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \tag{3.14}$$

By use of Bayes theorem

$$f(x,p) = \binom{n}{x}B(\alpha,\beta)^{-1}p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x \tag{3.15}$$

The marginal of x can therefore be found by

$$f(x) = \int_0^1 \binom{n}{x}B(\alpha,\beta)^{-1}p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x \mathrm{d}p \tag{3.16}$$

using of the binomial theorem

$[1-(1-p)^k]$ becomes

$\sum_{j=0}^{x}\binom{x}{j}(-1)^j(1-p)^{kj}$

$$f(x) = \binom{n}{x}B(\alpha,\beta)^{-1}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\int_0^1 p^{\alpha-1}(1-p)^{kn+kj-kx+\beta-1}\mathrm{d}p \tag{3.17}$$

$$f(x) = \binom{n}{x}B(\alpha,\beta)^{-1}\sum_{j=0}^{x}\binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta) \tag{3.18}$$

The posterior of p can therefore be evaluated as

$$f(p/x) = \frac{p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x}{\sum_{j=0}^{x}\binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta)} \tag{3.19}$$

Thus, the equal tail $100(1-\alpha)\%$ credible interval $(P_l, P_u)$is given by the solution to

$$\int_0^{P_l} \frac{p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x}{\sum_{j=0}^{x}\binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta)} = \frac{\alpha}{2} \tag{3.20}$$

and

$$\int_{P_u}^{1} \frac{p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x}{\sum_{j=0}^{x}\binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta)} = \frac{\alpha}{2} \tag{3.21}$$

However, the HPD $100(1-\alpha)\%$ is given by

$$\int_{P_l}^{P_u} \frac{p^{\alpha-1}(1-p)^{kn-kx+\beta-1}[1-(1-p)^k]^x}{\sum_{j=0}^{x}\binom{x}{j}(-1)^j B(\alpha, kn+kj-kx+\beta)} = 1-\alpha \tag{3.22}$$

3.2.2 Empirical Bayes credible interval

Group testing is only economically viable if the proportion of the defectives is small, we would like to use a family of prior distributions appropriate for small p. the Beta$(1, \beta)$is such a family. this is because for large values of of β the majority f the probability distribution of the random variable p is close to zero.

Recall that

$$f(x/p) = \binom{n}{x}(1-(1-p)^k)^x((1-p)^k)^{n-x} \tag{3.23}$$

Now that

$$f(p) = \beta(1-p)^{\beta-1} \tag{3.24}$$

then

$$f(x,p) = \beta\binom{n}{x}(1-(1-p)^k)^x((1-p)^k)^{n-x}(1-p)^{\beta-1} \tag{3.25}$$

Thus the marginal of x

$$f(x) = \beta\binom{n}{x}\int_0^1 [1-(1-p)^x](1-p)^{kn-kx+\beta-1}\mathrm{d}p \tag{3.26}$$

Using the change of variable technique

Let $u = (1-p)^k$

which implies that $p = 1-u^{\frac{1}{k}}$

$\mathrm{d}p = \frac{-1}{k}(1-p)^{1-k}\mathrm{d}u$

thus

$$f(x) = \beta k^{-1}\binom{n}{x}\int_0^1 u^{n-x+\frac{\beta}{k}-1}(1-u)^x \mathrm{d}u \tag{3.27}$$

$$f(x) = \beta k^{-1}\binom{n}{x} B(n-x+\frac{\beta}{k}, x+1) \tag{3.28}$$

$$f(x) = \beta k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\beta}{k}),)}{\Gamma(n+\frac{\beta}{k}+1)} \tag{3.29}$$

Maximizing the above equation with respect to β may be done numerically by solving

$$\frac{\partial}{\partial\beta}(f(x|\beta)) = 0$$

$$\frac{\partial}{\partial\beta}[\beta k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\beta}{k}),)}{\Gamma(n+\frac{\beta}{k}+1)}] \tag{3.30}$$

Taking the logarithms

$$\begin{aligned} 0 =& \frac{\partial}{\partial\beta}\{\log\beta + \log[\Gamma(n-x+\frac{\beta}{k})] - \log[\Gamma(n+\frac{\beta}{k}+1)]\} \\ =& \beta^{-1} + k^{-1}[\psi(n-x+\frac{\beta}{k}) - \psi(n+\frac{\beta}{k}+1)] \end{aligned} \tag{3.31}$$

Where ψ represents the digamma function.

Thus $\hat{\beta}$ is the solution to 3.31. which is the optimal β and therefore β can be replaced by $\hat{\beta}$.

Hence, the posterior of p can be given as

$$f(p/x) = \frac{\hat{\beta}\binom{n}{x}(1-(1-p)^k)^x|((1-p)^k)^{n-x+\hat{\beta}-1}}{\hat{\beta}k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\hat{\beta}}{k}),)}{\Gamma(n+\frac{\hat{\beta}}{k}+1)}} \tag{3.32}$$

$$f(p/x) = k\frac{\Gamma(n+\frac{\hat{\beta}}{k}+1)}{\Gamma(n-x+\frac{\hat{\beta}}{k}),\Gamma(x+1)}(1-(1-p)^k)^x|((1-p)^k)^{n-x+\hat{\beta}-1} \tag{3.33}$$

Thus, the equal tail $100(1-\alpha)\%$ credible interval $(P_l, P_u)$is given by the solution to

$$\int_0^{P_l}\frac{\hat{\beta}\binom{n}{x}(1-(1-p)^k)^x|((1-p)^k)^{n-x+\hat{\beta}-1}}{\hat{\beta}k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\hat{\beta}}{k}),)}{\Gamma(n+\frac{\hat{\beta}}{k}+1)}} = \frac{\alpha}{2} \tag{3.34}$$

and

$$\int_{P_u}^{1} \frac{\hat{\beta}\binom{n}{x}(1-(1-p)^k)^x|((1-p)^k)^{n-x+\hat{\beta}-1}}{\hat{\beta}k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\hat{\beta}}{k}),)}{\Gamma(n+\frac{\hat{\beta}}{k}+1)}} = \frac{\alpha}{2} \tag{3.35}$$

However, the HPD $100(1-\alpha)\%$ is given by

$$\int_{P_l}^{P_u} \frac{\hat{\beta}\binom{n}{x}(1-(1-p)^k)^x|((1-p)^k)^{n-x+\hat{\beta}-1}}{\hat{\beta}k^{-1}\frac{\Gamma(n+1)}{\Gamma(n-x+1),}\frac{\Gamma(n-x+\frac{\hat{\beta}}{k}),)}{\Gamma(n+\frac{\hat{\beta}}{k}+1)}} = 1-\alpha \tag{3.36}$$

Chapter 4

Conclusion and recommendation

At a glance this thesis has put together seven point estimators, four criteria of choosing k, the group size, and six interval estimators. The maximum likelihood is the traditional way of estimating the proportion in group testing. With its limitations such as its biased nature, two bias corrected estimators are proposed and the case where prior information is available then Bayesian estimation makes more sense in-order to make use of the information. With Bayesian analysis we get the additional four different point estimators. With the realization that the benefits of group testing are highly dependent on the group size and with ambiguity of how to go about their choice four ways of choosing them are in the literature. For the Interval estimates we have three Wald interval based on different variances, exact interval estimate and the finally two ways of going about interval estimation in Bayesian.

Thus with a researcher considering the use of group testing in his/her estimation quest, he/she might come to a point of confusion due to the large number of estimators available and how specifically t design the experiment with four different ways of choosing from. This project, also puts into consideration of development of the theory.

4.1 Experiment Design

From this review we can conclusively conclude that choosing your group sizes based on reducing $\mathbb{MSE}$ criteria is the best if the researcher has enough prior information of the

phenomenon. This is because this method guarantees precision of the estimator. However, this method requires the researcher to have a prior proportion or at least its upper bound which is hard to be certain of in the most cases of estimation problem. In the case when the researcher lacks this vital information he/she can spit his/her testing procedure into at least two testing stages and adapt/adjust the group size from one testing phase to the next using all the accumulated data to obtain a final group size based on all the data collected. This method guarantees precision since it ends up with an optimal group size no matter the choice of the prior proportion. This adaptive method bears the limitation that it requires a lot of resources and is mathematical computation intensive with the researcher optimization the $\mathbb{MSE}$ at every stage/phase of testing.

For instance, consider a case where the researcher can only afford 30 test (n=30) and the proportion is p=0.05 (unknown to the researcher), he/she chose the prior proportion to be $p_0 = 0.025$ then by optimizing $\mathbb{MSE}$ he/she obtains a group size of 38 (k=38). Perhaps it is also important to not that if the prior proportion is bigger than the true proportion the the group size used will be greater that optimal and poses more threat when than when the prior is less than the true proportion in-terms of precision (MSE). Thus was the bases of at the upper bound above. If x count the number of groups that test positive, then x is distributed as a binomial random variable with parameters 30 and $1-(1-p)^k = 0.876$. Suppose that x=25 is observed the $\mathbb{MLE}$ would be $\hat{p} = 0.046$ with $\mathbb{MSE} = 0.009156$ estimated by 0.003973 with p replaced by $\hat{p}$ everywhere.

On the other hand, if the researcher decides that his/her prior information is not sufficient and decide to adapt the group size k using two stages of testing and thus the half of test (n=15) performed at the first stage with the same conservative prior $p_0 = 0.025$. Then by optimizing the $\mathbb{MSE}$ he/she obtains the first stage group size to be $k_1 = 23$. If x count the number of groups that test positive, then x is distributed as a binomial random variable with parameters 15 and $1-(1-p)^k = 0.6926$. Suppose that x=10 is observed then the intermediate proportion would be $\hat{p}_1 = 0.0466$. Using this estimate as the prior proportion to determine the second testing stage group size k_2 and optimizing the $\mathbb{MSE}$. The second stage group size would be $k_2 = 13$ If x count the number of groups

that test positive, then x is distributed as a binomial random variable with parameters 15 and $1-(1-p)^k = 0.4867$. Suppose that x=7 is observed the adaptive $\mathbb{MLE}$ would be $\hat{p}_A = 0.04682$ with the true $\mathbb{MSE} = 0.000215$ estimated by 0.000182 with $\hat{p}_A$ in place of p.

Thus, it is clear that the adaptive method is superior with not only $\mathbb{MSE}(\hat{p}_A)_e < \hat{\mathbb{MSE}}(\hat{p})$ but also that $\hat{\mathbb{MSE}}(\hat{\mathbb{A}}p)$ is closer to $\mathbb{MSE}(\hat{p}_A)$ than $\hat{\mathbb{MSE}}(\hat{\mathbb{A}}p)$ is to $\mathbb{MSE}(\hat{p})$. However, it is computationally intensive and may not be attractive to non mathematicians.

It is also interesting to note that in this example the advantage of this adaptive estimation increases if the cost of the individuals is high. Whereas in both adaptive and non adaptive method estimates use the same number of test (n=30) the non adaptive uses 1140 (30×38) individuals units, whereas the adaptive estimate uses 540 ($15\times23+15\times13$).

4.2 Point estimates

In the case where the researcher is interested on the point estimate of proportion, he/she will be required to choose on the seven estimates discussed in this work. The most traditional estimate is $\mathbb{MLE}$. in the case where the choice of group size is determined by the technical issues of the testing procedure, then the next thing for the researcher to consider is the accuracy (bias) of the estimate. The $\mathbb{MLE}$ is positively biased. Bias increases with increase of the true proportion, reduces as n increases and is increases as group size increases. Thus the easy logic would to keep the group size small and perform relatively many tests (n). This way the cost are hugely increases losing the primary advantage of group testing. In such a dilemma the researcher would be advised to use one of the bias corrected estimates discussed in this work. The Chaubey Li. biased corrected estimate is however highly dependent on the group size and thus i would recommend its use only if the researcher is confident that he/she is confident that the group size used is optimal, otherwise the bias is hardly corrected and in some cases over corrected as illustrated by table 2.3 with some cases of negative bias. On the other hand if the researcher have doubts that the group sizes may differ from the optimal group, he/she is left to use the Burrows bias corrected $\mathbb{MLE}$ which performs well in $p \leq 0.25$ and $n \leq 200$ as illustrated by table 2.4. Outside this region the burrows bias corrected estimate produces negative bias (over corrects the $\mathbb{MLE}$) as shown in figure 2.4.

The Bayesian estimates are characterized with randomizing the proportion. Thus if the researcher can assign the proportion a probability distribution function he/she would use any Bayes estimates.Traditionally the proportion is believed to be a beta random variable. Thus by the use of classical Bayesian approach we get an additional two hyper-parameters to be estimated as illustrated in this work. However, using the empirical Bayesian approach then distribution is believed to be beta, generally when $\alpha < \beta$ as shown by figure 2.7 but specifically beta with $\alpha = 1$ and large β.. With the posterior mean a squared error loss function has been used. Additional empirical estimate ca be obtained using different loss function. In this work two scaled loss function have been used to obtain their respective empirical Bayesian estimates. To illustrate the choice of

the point estimates we illustrate an experiment done in the use on estimation of Hepatits C estimation in the USA. Hepatits C(HCV) is a viral infection that causes cirrhosis and cancer of the liver. Currently the worldwide sero-prevalence of HCV is estimated to be around 3%. Cost associated with testing HCV are very high, this makes HCV screening experiments excellent candidates for group testing. Neill and Conradie (1992, 1994) proposed the use of group testing to screen for the prevalence of HCV. They determine that commonly used kits could reliably test detect the HCV anti-bodies for group sizes $k \leq 8$. In this application, biological constraint associated with the testing procedure prohibit large group sizes.

Liu (1997) reported 1875 blood donors screened for HCV.using the group size k=5 and in addition, researchers tested 1875 serum individually k=1 so that he could examine the efficiency of group testing.

	Individual test	Pooled tests
Number of pools	1875	375
Positive pools	42	37
Estimate of P	0.0224	$\hat{p}_{mle} = 0.02056$
		$\hat{p}_{eb} = 0.020557$

With n=375 , k=5 and x=37 the empirical Bayes estimate of β is given by

$$\hat{\beta} = argmax_{\beta \geq 1} f(\beta | x = 37) \approx 48.13$$

It is clearly observed from the above table $\hat{p}_{mle}$ and $\hat{p}_{eb} = 0.020557$ provided similar point estimates. Because of a relatively large n one might expect this. However, if the researcher used the bias corrected estimates to correct bias of $\mathbb{MLE}$ then he/she would have ended up with

Chaubey Li	$\hat{p} - \frac{k-1}{2nk^2}[\frac{x/n}{1-x/n^{(k-1)/k}}] = 0.0205391$
Burrows	$1 - [\frac{2k(n-x)+k-1}{2kn+k-1}]^{1/k} = 0.0205368$

The above two estimators seems to perform equally the same this is due to the fact that n is relatively large (n=375) and probably the fact the n used was close to optimal (well designed).

4.3 Interval Estimation.

This thesis reviewed a total of six interval estimates. First we look at he three Wald intervals generated from the asymptotic variance of the estimated proportion, the exact variance of the estimated variance and a stabilized variance (a variance that is free of the estimate of proportion). For this purpose of comparison we name them Wald, Thompson and VSI (Variance Stabilized Interval). To measure their performance we use an example of data obtained from an experiment in Argentina to study the effects of the Mal Rio Cuarto (MRC) virus. The goal of this experiment was to estimate the probability of virus transmission in natural Macropterous plant hoppers populations that are known sources of virus. A total of $n = 24$ plants were used each allocated k=7 plant-hoppers. At the end of the experiment x=3 plant were observed as infected yielding a point estimate $\hat{p} = 1 - (1 - \frac{3}{24})^{1/7} \approx 0.019$. The table 4.3 shows the three interval estimates obtained.

Interval	95%	Length
Wald	(-0.0023, 0.0401)	0.0424
Thompson	(-0.0028, 0.0406)	0.0434
VSI	0.0037, 0.0465	0.0428

From table 4.3 it is clear that VSI overcomes the biggest shortcoming of the other two of producing a negative endpoints. Thus we can confidently conclude that the VSI is a superior interval estimates and recommend its use in place of the Wald and the Thompson interval estimates.

The exact and Bayesian interval estimates can be used depending on the researcher needs.

Bibliography

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

[2] Swallow, W. H. (1985), Group testing for estimating Infection rates and Probability od disease transmission, *Phytopathology,* 75, 882-889.

[3] Hudges-Oliver, J. M.,and Swallow, W. H. (1994), A two stage adaptive group testing procedure for estimating small proportions, *Journal of American Statistical association,* 79, 982-993.

[4] Burrows P. M. (1987), Improved estimation of Pathogen transmission rates by group testing, *Phytopathology,* 77, 363-365.

[5] Chaubey, Y. P. and Li, W. (1993), Comparison between Maximum Likelihood and Bayes Methods for Estimation of Binomial Probability with Sample Compositing, *Journal of official Statistics,* 11, 379-390.

[6] Dorfman, R.(1943), The detection of defective members of a large populations, *Annals of Mathematical statistics,* 14, 436-440.

[7] Tebbs, J. M., Bilder, C. R. and Moser B. K. (2004), Confidence Interval procedures for the probability of disease transmission in multiple vector transfer designs, *Journal of Agricultural, Biological and environmental Statistics,* 9, 75-90.

[8] Tebbs, J. M. and Swallow, W. H. (2003), Estimated ordered binomial proportions with use of group testing, *Biometrica,* 90,471-477.

[9] Sobel, M. and Elashoff, R M. (1975), Group testing with a new goal, estimation, *Biometrica,* 18:568-578.

[10] Hepworth, G. (1996), Exact confidence intervals for proportions estimated by group testing, *Biometrics,* 52, 1134-1146.

[11] Xiang F., Walter W. S. and Shunpu Z.(2007), Improved Empirical Bayes Estimation in Group testing procedure for small proportion, *Communications in Statistics-Theory and Methods,* 36: 2937-2944.

[12] Tebbs, J. M., Bilder, C. R. and Moser B. K.(2003),An Empirical Bayes Group Testing Approach to Estimating Small Proportions, *Communication in Statistics- Theory and Methods,* 32, 983-995.