

ESTIMATION PROBLEM IN GROUP SCREENING DESIGNS "

By

Muhua George Odweso

This thesis is submitted in fulfilment for the degree of Doctor of Philosophy in
Mathematical Statistics in the School of Mathematics

UNIVERSITY OF NAIROBI

©2009, Muhua George Odweso

All Rights Reserved





Declaration

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

Signature:  Date: 03/09/2010
Muhua George Odweso

This thesis has been submitted for examination with our approval as University supervisors.

Signature:  Date: 07-09-2010
Prof. J.A.M. Ottieno

Signature:  Date: 07-09-2010
 Prof. J.O. Owino

Dedication

This work is dedicated to my late mother and my family.

Acknowledgements

I am greatly indebted to my first supervisor Prof. J.A.M Ottieno who has supported me along my journey from my undergraduate studies and who, for countless hours throughout my research was willing to guide me in the core research activity. I am especially grateful for his endless effort, tireless and valuable guidance and dedication and inspired ideas, without which this thesis would not have been written.

I owe the courage to finish this work to all my colleagues at the School of Mathematics. This includes not only my colleague lecturers and professors but also support staff such as secretaries, computer technologists and administrators. I would specifically like to express my appreciation and gratitude to my second supervisor and former Director of the School of Mathematics, Prof. John Okoth Owino, for his support, guidance and encouragement, and for enabling me win a scholarship to pursue this degree. In this regard, I would like to thank DAAD-German Academic Exchange Service for their fellowship which has seen me through the financial aspects of my research.

I must not forget to thank my wife **Gloria** and my children for their love, patience, understanding and perseverance of putting up with many days when I was away from home in pursuit of this degree. Without their support, this thesis would not have been completed.

Finally, I would like to pay special tribute to my late mother, **Mrs Domtila Muhua** without whose effort and sacrifice, I would not have reached this far. May the Almighty God rest her soul in eternal peace.

Abstract

In many situations units of some sort are to be tested perhaps to label them as defective (non-satisfactory) or non-defective (satisfactory). Often testing the units one by one is inefficient especially when they are cheap to obtain relative to the cost of the test. In such cases it is often preferable to form groups of units and test all units in a group simultaneously. This is called "group testing" or "group screening" in statistical literature.

This thesis considers estimation of the prevalence rate in a group screening design using two methods: maximum likelihood method and Bayesian method.

Most of the work in group screening have concentrated on designs without errors. In this work we have undertaken a review of these works and extending them to the case with errors, which involves taking the accuracy of the screening or testing equipment into consideration.

It is shown that for small group sizes and large values of p and k , the bias is considerable while for low values of p and k , there is relatively little bias, and that the best design for a given experiment depends on the prevalence rate, tolerable mean square error, the sampling cost of an individual sample and the cost of performing a single test.

We have also shown how Bayes methods can be incorporated in the group screening problem. For low prevalence rate, we have shown that Bayes estimator outperforms the maximum likelihood estimator in terms

of bias and mean square error. Also we have argued that interval estimates based on Bayes method may be more appropriate, as to avoid intervals extending outside the parameter space.

By taking measurement errors into consideration, we have shown that by group testing we not only achieve a cost saving, but also an increase in the estimation accuracy.

For estimating infection rate in a population of organisms, when sample pools of unequal sizes are analyzed, we have suggested an iterative method of determining successive estimates of the infection rate, resulting in an estimator which can easily be evaluated and upgraded, using the average size of positive pools.

Contents

Declaration	i
Dedication	ii
Acknowledgements	iii
Abstract	iv
List of variables	x
1 General Introduction	1
1.1 The Concept of Group Screening Designs	1
1.2 Terminologies and Notations	2
1.3 Applications	4
1.3.1 <i>Insect-Vector Problem</i>	4
1.3.2 <i>Rodent-Bacterium Problem</i>	5
1.3.3 <i>Blood Testing Problem</i>	6
1.3.4 <i>Plant Testing Problem</i>	6
1.3.5 <i>Other Applications</i>	7
1.4 Literature Review	8

1.5	Statement of the Problem	13
1.6	Objectives of the Study	14
1.7	Methodology	14
2	Maximum Likelihood Estimation With Equal Probabilities but Without Errors in Decisions	16
2.1	Introduction	16
2.2	Binomial Model	17
2.2.1	Maximum Likelihood Estimator of p	18
2.3	Properties of the Estimator, \hat{p}_k	18
2.3.1	The Mean of the Estimator and its Biasedness . . .	18
2.3.2	The Variance of the Estimator	27
2.3.3	Asymptotic Variance of the Estimator	28
2.3.4	The Behavior of the Asymptotic Variance	30
2.3.5	Mean Squared Error	32
2.3.6	The Cost Function	36
2.3.7	Efficiency	38
3	Maximum Likelihood Estimation With Equal Probabilities and With Errors in Decisions	39
3.1	Introduction	39
3.2	Sensitivity-Specificity Approach	40
3.2.1	Maximum Likelihood Estimator of p	42
3.2.2	The Mean of the Estimator and its Biasedness . .	44
3.2.3	The Variance of the Estimator	48

3.2.4	Asymptotic Variance of the Estimator	50
3.3	Relative Efficiency	52
3.4	The Test of Hypothesis Approach	53
3.4.1	Special Case	59
3.4.2	Alternative Approach	60
4	Bayesian Estimation With and Without Errors in Decision	63
4.1	Introduction	63
4.2	Estimation Without Errors in Decisions	64
4.3	Estimation of α and β	67
4.4	Special case: A prior on p with $\alpha = 1$	68
4.5	Prior on p^*	71
4.6	Derivation of Moments	74
4.7	Comparison of Estimators	76
4.7.1	Point Estimate Characteristics	76
4.7.2	Asymptotic Distribution and Interval Estimation	83
4.8	Estimation With Errors in Decisions	84
4.8.1	Derivation of a Bayesian Estimator	85
4.8.2	Alternative Approach	86
4.8.3	Approximation of the Variance	88
5	Group Screening Design With Equal Probabilities but With Unequal Group Sizes	89
5.1	Introduction	89
5.2	Estimation without Errors in Decision	90

5.2.1	Notations and Method	90
5.2.2	Poisson Model Approximation to Binomial Model	94
5.2.3	Asymptotic Variance of \hat{p}	95
5.3	Special Cases	97
5.3.1	Chiang-Reeves model	97
5.3.2	Bhattacharya model	98
5.3.3	Griffiths Approach	99
5.4	Estimation With Errors in Decision	100
6	Summary and Conclusions	102
6.1	Summary	102
6.2	Conclusion	105
6.3	Recommendation for Further Research	106
	Bibliography	107

List of Variables

- f : the population size
- g : the number of groups or group-factors
- k : group size
- r : number of defective group-factors
- p_1 : proportion of defectives when $k = 1$
- p_k : proportion of defectives when $k > 1$
- p^* : proportion of defectives in a group
- C_1 : the total cost for one at a time test
- C_k : the total cost for group test
- C_S : the sampling cost of one individual sample
- C_A : the cost of performing one test
- η : the sensitivity of the screening test
- θ : the specificity of the screening test
- π_1^* : the probability that a group-factor is declared defective
- α_1 : the probability of declaring a non-defective group-factor defective
- β_1 : the probability of declaring a defective group-factor non-defective
- $1 - \alpha_1$: the probability of declaring a non-defective group-factor non-defective
- $1 - \beta_1$: the probability of declaring a defective group-factor defective
- $\pi_1(s\phi_1, \alpha_1)$: the power of the test

Chapter 1

General Introduction

1.1 The Concept of Group Screening Designs

In many practical situations, units of some sort are to be tested perhaps to label them as defective (non-satisfactory) or non-defective (satisfactory). The units might be blood samples, insects or electrical components, for which defective could mean testing positive for a disease, carrying a disease causing agent, or being faulty respectively. Often testing the units one by one is inefficient, especially when very few of them are defective and they are cheap to obtain relative to the cost of the test. In such cases, it is often preferable to form groups of units and test all units in a group simultaneously. In statistical literature, this is referred to as "group testing" or "group screening". Usually the outcome of the group test is dichotomous. When the outcome is satisfactory, one concludes that all units in the group are satisfactory; if it is not, one concludes that at least one unit is defective but does not know which one(s) or how many. If the aim is to classify each individual as positive or negative then individuals in positive groups will

have to be retested in smaller groups, perhaps individually until all positive ones have been identified . This is known as **classification problem**. In many applications the aim is to estimate only the proportion of positives in the population, not specifying which particular individuals are positive. This is called **estimation problem**.

The main aim of group screening is to reduce the expected number of tests in the classification problem by eliminating a large number of non-defective factors in a batch, thereby effecting substantial saving in the cost of the experiment. For the estimation problem, the aim is to determine the optimal group size that improves the precision of the estimator of p , the proportion of positives in the population. Thus, group screening is a cost effective method of detecting defective individuals and estimating the prevalence rate of attribute in these individuals in a given population.

1.2 Terminologies and Notations

Terms used in group screening designs have been specific to particular areas of application. Owing to this, terminologies describing the same concept are diversified and sometimes confusing. Group screening designs have also been called *group testing, pooling, probing, composite sampling, multi-vector transfer, pooling organisms*, by researchers such as Sobel and Groll (1959), Gibbs and Gower (1960), Swallow (1985,1987), Chao and Swallow (1990 and so on. Groups have been termed *group-factors, batches, composite samples, test plants, pools, pooled sera, samples*, and so on. Defective and non-defective have been called *significant and insignificant, unsatisfactory*

and satisfactory, bad and good, non-conforming and conforming, contaminated and non-contaminated, active and inactive, positive and negative, infected and non-infected, diseased and healthy, non-reactive and reactive, and so on.

Terms such as *repeated trials* (Graff and Roellfs(1972)), *stepwise* (Manene(1985)), *curtailed and hierarchial* (Johnson, Kotz and Wu(1992)) have been used to describe special types of group testing.

Various notations have been used for the population size, group size, total number of groups or pools formed, number of groups found defective or non-defective, the probability of a factor being defective or non-defective and the corresponding probability of a group factor being defective or non-defective.

Some notations by various researchers are shown below.

Table 1 Table of Notations in group-screening designs

Description	Watson	Bhattacharrya	Thompson	Chiang
Population size	f	N	m	N
Group size	k	m	k	m
No. of groups	g	n	n	n
No.of defective groups	r	n-R	$n - \sum x_i$	x
No.of non-defective groups	g-r	R	$\sum x_i$	n-x
Prob of defective factor	p	θ	p	p
Prob of non-defective factor	1-p	$1 - \theta$	1-p	1-p
Prob of defective group-factor	p^*	1-p	1-h(p)	π
Prob of non-defective group-factor	$1 - p^*$	p	h(p)	$1 - \pi$

In this work we shall adopt Watson's(1961) notations.

1.3 Applications

Various researchers have used group screening to estimate the proportion of defective or infected individuals in a given population. Applications include studying the spread of insect borne diseases, quantifying resistance factors in plants and estimating proportions of infected or infective individuals in a population among others.

In epidemiologic investigation of disease agents transmitted by arthropod vectors, it is often necessary to estimate the infection rate in the vector population. Because the number of specimens in the sample in most cases is quite high, it is practically impossible to assay each specimen individually. Instead the specimens are randomly divided into a number of pools and each pool is tested as a unit. If a pool includes at least one infected specimen the test shows positive, whereas a negative test is obtained when no infected individual is present in the pool. The ratio of the number of positive pools to the total number of specimens in all pools, is called the *infection rate* which bears a direct relationship to the prevalence or risk of the diseases in the human and other populations. The areas of application that are used in our study and the work done by some researchers in these areas is mentioned in the following sub sections.

1.3.1 *Insect-Vector Problem*

Thompson(1962) used group testing to estimate the proportion of vectors capable of transmitting aster-yellows virus in a natural population of the six spotted leaf hoppers. A group of insects were randomly chosen

and caged with each plant in a given time interval. Only the test plant result from a group of insects was used in estimating the proportion of viruliferous insects in the population from which the random sample was drawn.

Chiang and Reeves(1962) suggested simple and accurate methods of estimating the infection rate in a mosquito population, using group screening design. They considered a pooling device where mosquitoes collected from the field were divided into pools, each pool containing nearly constant number of mosquitoes which was then tested for the presence of virus.

Walter *et al*(1980) used pools of variable sizes to determine if certain mosquitoes could transovarially transmit yellow fever virus. A yellow fever infected adult population of *Aedes aegypt* produced a progeny population which was hatched, reared to adults, separated by sex and grouped in pools for virus assay. Two strains of yellow fever virus were used. Strain A isolated from a dead monkey and strain H isolated from a pool of Haemagogus mosquitoes. Temporal variation occurred during the laval development period, requiring from 6 to 20 days for all mosquitoes to pupate. This biological process resulted in pools of variable sizes.

1.3.2 *Rodent-Bacterium Problem*

Sobel and Elashoff(1975) considered a case where rodents are collected from the harbour of a large city and after being killed and dissected their liver was carefully examined under a microscope, for the presence or absence of a specific type of bacterium. The goal was to study the pro-

portion of rodents that carry this type of bacterium, using an economical experimental design. In this application the cost of obtaining the animals was assumed negligible compared to the cost of the test.

1.3.3 *Blood Testing Problem*

The standard procedure for screening individuals for antibody to the HIV is by Enzyme Linked Immunosorbent Assay (ELISA). Other kits such as Western Blot(WB) are also used, though more expensive but more reliable. Gastwirth and Hammick(1989) estimated prevalence, rate by firstly combining blood samples from k individuals into a single batched sample using ELISA followed by Western Blot for confirmation.

Kline *et al*(1989) undertook a study to assess whether the testing of pooled sera was technically feasible, cost effective and an accurate method of determining HIV seroprevalence in large population based on surveys. A series of experiments were performed to estimate the reliability of ELISA method using pooled sera.

Davies, Grizzle, and Bryan(1973) estimated the probability of post-transfusion hepatitis, when patients received several blood products by generalizing to pools for which each unit does not have the same probability of transmitting the disease.

1.3.4 *Plant Testing Problem*

Gibbs and Gower(1960) used the multiple vector transfer method to determine the frequency with which a virus disease is transmitted to

another by taking samples from a population and testing more than one sample on each test plant. The number of test plants that became infected were then used to determine the proportion of diseased plants in the population.

1.3.5 *Other Applications*

Other areas of application that have recently been considered include quality control application by Xiang Fang *et al* (2007), industrial experimentation by Vine *et al*(2008), breast cancer research tumor growth study by Herald-Weeden-Fekjar *et al*(2008)and statistical analysis of $E(f_{nod})$ -optimal mixed level supersaturated designs by Koukorinas and Mylona (2009).

1.4 Literature Review

Group screening was first introduced into the literature by Dorfman (1943) as an economical way of testing blood samples of army inductees in order to detect the presence of syphilis infection. He proposed that rather than test each blood sample individually, portions of each of the samples could be pooled together and the pooled sample tested first. If the pooled sample was non-infected, all the inductees would be passed with no further test. Otherwise, each portion of the blood samples would be tested individually. The aim here was to reduce the expected total number of tests and the expected total cost of inspection, with the assumption that the prevalence rate was low. He plotted graphs of prevalence rate p versus group sizes. Optimal group sizes occurred at stationary points on the curves but he did not give a general expression for obtaining the optimal group sizes.

Various researchers have extended this concept to estimate the proportion of defectives under different conditions. These conditions include: (i) group screening with equal group sizes, equal probability and with no errors in decision by Gibbs and Gower(1960), Thompson(1962), Chiang and Reeves(1962),Kerr(1971), Bhattacharyya *et al*(1979) and Swallow(1985,1987) (ii) group screening with equal group sizes, equal probability but with errors in decision by Gastwirth and Hammick(1989) (iii) group screening with unequal group sizes, equal probability and with no errors in decision by Griffith(1972),Walter et al(1980) and Le(1981).

Based on these conditions, a theoretical framework for studying esti-

mation problem in Group screening designs is suggested, as shown in figure 1.1 below.

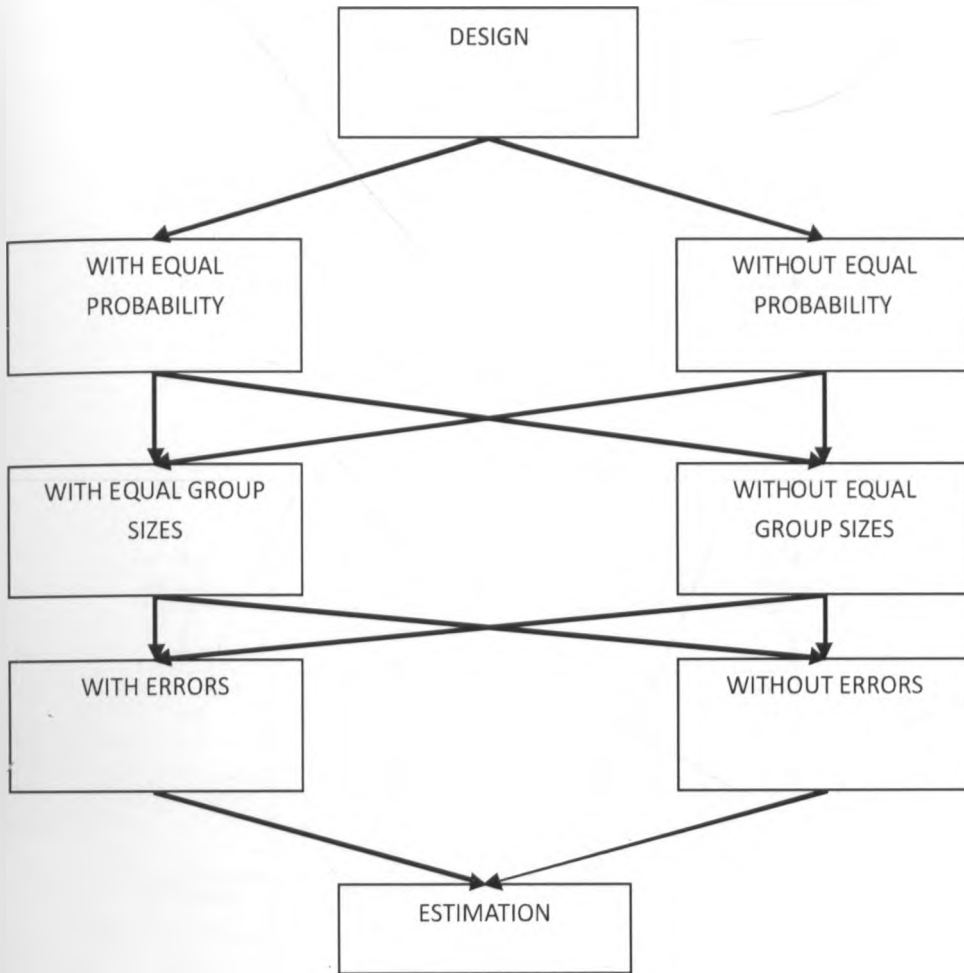


Figure 1.1: Theoretical Framework for Estimation in Group Screening Designs

From the framework more conditions are identified and therefore, the figure is split into various frames as shown below. These frames were used to identify the gaps in the research conducted. The study concentrated on the frames 1(a), 1(b) and 1(c).

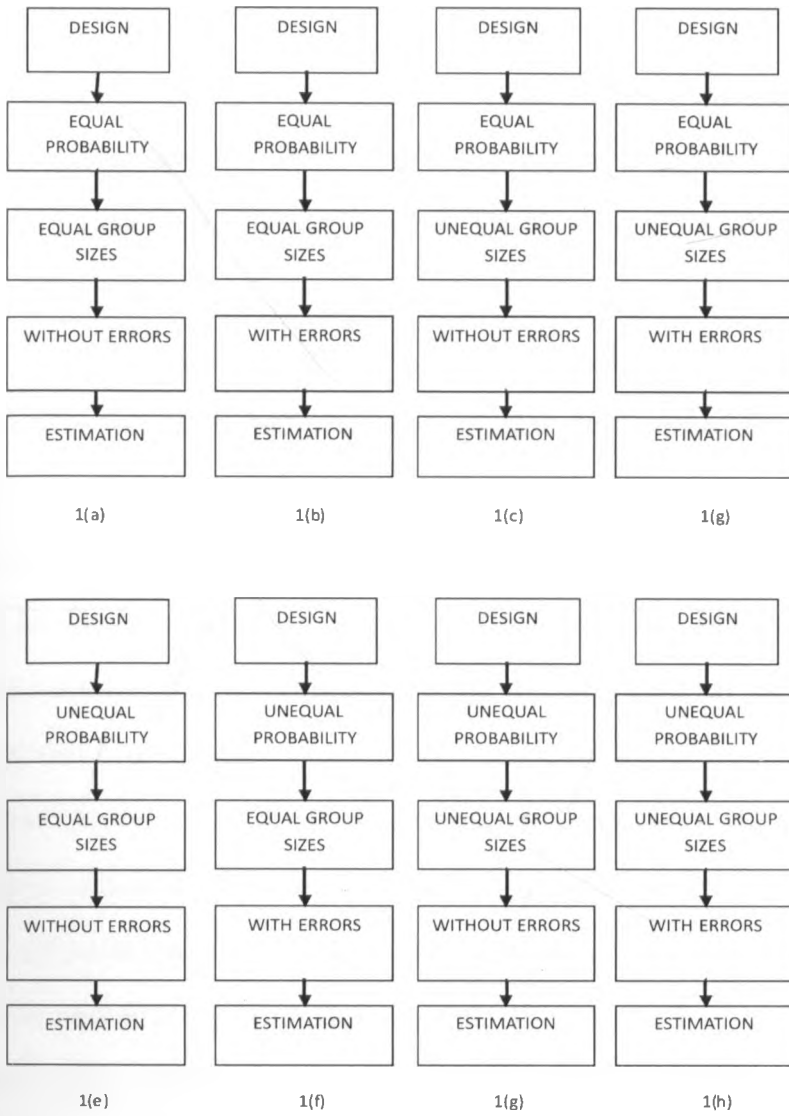


Figure 1.2: Approaches to the study of Estimation in Group Screening Designs

Under the first condition given in diagram 1(a), Thompson(1962) used a Binomial model and the maximum likelihood estimation method to estimate the infection rate. Some properties of the estimator of p such as biasedness and Best Asymptotic Normal (BAN) were investigated and the optimum number of insects per plant is determined by finding the value of k that minimizes the asymptotic variance using differential calculus. This work can be reviewed and the optimal group size determined by minimizing the mean squared error.

Kerr(1971) used the binomial model and the maximum likelihood estimation method to estimate the infection rate. The biasedness of the estimate was studied using graphs. But he did not give an expression for determining the biasedness.

Bhattacharyya *et al*(1979) considered two models of the population: the finite population model where the whole population was sampled and the infinite population model where sampling was done from super population. For the finite population model, the method of moments was used while for the infinite model the MLE method was used based on the binomial distribution to estimate the infection rate. They studied the biasedness of the estimator. They did not consider the other properties of the estimator, and the optimal value of k for both finite and infinite models.

Kline *et al*(1989) model results was based on the groups that were identified positive in estimating p , but did not take into consideration the fact that positive pools can be misclassified as negative. The estimated value of p was actually lower than than the true value.

Tebbs and Bilder(2003) and Lew and Levy (1989) studied empirical Bayes approach in group screening designs. They considered a beta distribution with $\alpha = 1$ as the conjugate of the binomial distribution and determined the posterior mean \hat{p}_{eb} , assuming an estimate $\hat{\beta}$ of β . Comparison between MLE and Bayesian estimates was made using the relative bias, showing that EB estimate outperforms MLE. They also discussed interval estimation, where credible interval or the highest posterior density (HPD) construction method was used. This work can be extended to the general case where α and β are unknown and the case $\alpha = 1$ treated as a special case. Griffiths(1973)also studied MLE for the beta-binomial distribution and its application to the households distribution of total number of cases of a disease.

Under condition 1(b), Gastwirth and Hammick(1989) obtained two estimators \hat{p}_1 and \hat{p}_2 based on the maximum likelihood method and the method of moments. Two cost functions were formulated, the first based on individual testing and the second on group screening. They were mainly interested in estimating p but not in determining the optimal group size that reduces the relative cost. They also made the assumption that sensitivity and specificity of the test kits were constants. In practise this is not always true as the results of the tests are influenced by several factors.

Xin, Litvak and Pagano(1994,1995) applied the model to HIV screening where they showed that pooling sera samples not only achieves cost savings but increases the estimation accuracy. They also showed that pooled testing increases the probability of estimating prevalence substan-

tially compared to non pooled testing. This work can be extended to the case where, misclassification which leads to high false prevalence rates is taken into consideration.

Under the third set of conditions given in diagram 1(c), Chiang and Reeves(1962) considered two pool sizes where the pool size in one group was markedly different from the other. They used the joint probability function of the number of positive pools in the two groups and the maximum likelihood method to estimate the infection rate. They studied the biasedness of the estimator and its asymptotic behavior.

Walter *et al*(1980) used the binomial model and the maximum likelihood method to estimate the infection rate. They then used the Newton-Raphson iteration method to determine the relationship between successive estimates. They studied the asymptotic behavior of the estimate.

Le(1981) reviewed the work of Walter *et al*(1980) and came up with a new estimator which could be solved non iteratively using the Poisson model and the method of moments. Bayesian estimation method can be applied and other properties such as mean square error and efficiency studied.

1.5 Statement of the Problem

Dorfman's original problem was to identify individuals with rare attribute. In this case the objective was to optimize the problem cost effectively. The assumption was that the probability p of defective was known.

What if p is not known. Once the individuals have been identified

then p can be calculated. However, if the individuals are not identified for confidentiality purposes, how do we estimate p ? This was the statement of the problem that we have tackled.

1.6 Objectives of the Study

The main objective of this study was to estimate the prevalence of defectives in group-screening designs for the identified routes in the conceptual framework. The methods of Maximum Likelihood Estimation and the Bayesian approach were used.

Specifically we were interested in

- (i) Identifying variables used in the design and using the variables to find the conditional probabilities appropriate for estimation problem.
- (ii) Reviewing the works by other researchers using the conditional probabilities.
- (iii) Deriving estimators (MLE and Bayes) and studying their properties such as Unbiasedness, Consistency and Efficiency.
- (iv) comparing the estimators using for example Relative Bias, Mean Square Error and Relative Efficiency.

1.7 Methodology

The goal of group screening procedure is to study the efficiency of the design, which is determined by optimizing a cost function. Some of the methods that were used in achieving this goal include:-

- (i) Differential and integral calculus used in determining the optimum group size
- (ii) Numerical approximations used in cases where iterations are involved
- (iii) Statistical methods such as the Cramer-Rao lower bound and the Delta method were used in studying the asymptotic properties of the estimator.
- (iv) Computer Simulation applied when comparing the efficiency of the group screening design with that of the traditional method of one at a time testing.

Chapter 2

Maximum Likelihood Estimation With Equal Probabilities but Without Errors in Decisions

2.1 Introduction

In this chapter, we review the estimator of the prevalence rate p under the condition of equal probability, equal group sizes and without errors in decision, using the maximum likelihood method of estimation. In section 2.2, we study the binomial model and its maximum likelihood estimator. The properties of the estimator such as biasedness, asymptotic variance, mean squared error and efficiency are discussed in section 2.3. The following three general assumptions are made in this chapter; (1) the status of individuals are taken to be independently and identically distributed Bernoulli random variables, (2) testing errors are negligible, and (3) test cost dominates the cost of sampling individuals and is independent of group size.

2.2 Binomial Model

If s is the number of defective factors in a group-factor of size k and p is the probability that a factor is defective, then s follows a binomial distribution with parameters p and k . The probability that a group-factor is defective is given by

$$\begin{aligned} p^* &= \sum_{s=1}^k \binom{k}{s} p^s q^{k-s} \\ &= 1 - q^k, \text{ where } q = 1 - p. \end{aligned} \quad (2.1)$$

If r is the number of defective group-factors then its probability function is

$$f(r) = \begin{cases} \binom{g}{r} p^{*r} (1 - p^*)^{g-r} & r = 0, 1, 2, \dots, g \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

Thus, the mean of r is

$$\begin{aligned} E(r) &= gp^* \\ &= g(1 - q^k), \end{aligned} \quad (2.3)$$

and the variance of r is

$$\begin{aligned} \text{var}(r) &= gp^*(1 - p^*) \\ &= g(1 - q^k)q^k. \end{aligned} \quad (2.4)$$

We now consider the maximum likelihood estimation of the binomial model.

2.2.1 Maximum Likelihood Estimator of p

Using the binomial model, the likelihood function of r is given by

$$L = \binom{g}{r} p^{*r} (1 - p^*)^{g-r}.$$

Taking logarithms, we have

$$\ln L = \ln \binom{g}{r} + r \ln p^* + (g - r) \ln(1 - p^*) \quad (2.5)$$

Differentiating with respect to p^* and equating to zero, we have

$$\hat{p}^* = \frac{r}{g}. \quad (2.6)$$

But since $\hat{p}^* = 1 - \hat{q}^k$, eqn (2.6) can be simplified to give

$$\hat{p}_k = 1 - \left(1 - \frac{r}{g}\right)^{\frac{1}{k}}. \quad (2.7)$$

If $k = 1$, we have the traditional estimator

$$\hat{p}_1 = \frac{r}{g}. \quad (2.8)$$

2.3 Properties of the Estimator, \hat{p}_k

In this section, we consider the properties of the estimator such as biasedness, asymptotic variance, mean squared error and efficiency.

2.3.1 The Mean of the Estimator and its Biasedness

Since \hat{p}_k is a function of a binomial variable r , we have

$$E(\hat{p}_k) = \sum_{r=0}^g \left(1 - \left(1 - \frac{r}{g}\right)^{\frac{1}{k}}\right) \binom{g}{r} p^{*r} (1 - p^*)^{g-r}$$

$$\begin{aligned}
&= 1 - \sum_{r=0}^g \left(\frac{g-r}{g}\right)^{\frac{1}{k}} \binom{g}{g-r} p^{*r} (1-p^*)^{g-r} \\
&= 1 - \sum_{i=0}^g \left(\frac{i}{g}\right)^{\frac{1}{k}} \binom{g}{i} ((1-p)^k)^i (1 - (1-p)^k)^{g-i}, \quad (2.9)
\end{aligned}$$

where $i = g - r$. Rewriting (2.9), the mean of \hat{p}_k is expressed as

$$\begin{aligned}
E(\hat{p}_k) &= 1 - \left(\frac{1}{g}\right)^{\frac{1}{k}} \sum_{i=0}^g i^{\frac{1}{k}} \binom{g}{i} ((1-p)^k)^i (1 - (1-p)^k)^{g-i} \\
&= 1 - \left(\frac{1}{g}\right)^{\frac{1}{k}} \alpha_1, \quad (2.10)
\end{aligned}$$

where

$$\alpha_1 = \sum_{i=0}^g i^{\frac{1}{k}} \binom{g}{i} ((1-p)^k)^i (1 - (1-p)^k)^{g-i}. \quad (2.11)$$

When $k = 1$, (2.10) becomes

$$\begin{aligned}
E(\hat{p}_1) &= 1 - \frac{1}{g} \sum_{i=0}^g i \binom{g}{i} (1-p)^i p^{g-i} \\
&= p. \quad (2.12)
\end{aligned}$$

Thus, for $k = 1$, \hat{p}_k is an unbiased estimator of p .

Proposition 2.3.1

The value of the estimator \hat{p}_k given in equation (2.7) overestimates the prevalence rate of defectives when k is greater than one ($k > 1$).

Proof:

To prove this proposition, we use Jensen's inequality, which states that, if $f(x)$ is a convex function then $E(f(x)) \geq f(E(x))$ and if $f(x)$ is concave then $E(f(x)) \leq f(E(x))$, provided the expectations exist and are finite.

Approach 1:

Assume that

$$\hat{p}_k = 1 - \left(1 - \frac{r}{g}\right)^{\frac{1}{k}}$$

is continuous. Then we have

$$\frac{\partial \hat{p}_k}{\partial r} = \frac{1}{gk} \left(1 - \frac{r}{g}\right)^{\frac{1}{k}-1}$$

and

$$\frac{\partial^2 \hat{p}_k}{\partial r^2} = \frac{k-1}{(gk)^2} \left(1 - \frac{r}{g}\right)^{\frac{1}{k}-2} \geq 0.$$

Therefore, \hat{p} is convex.

Since \hat{p}_k is a function of r , say $\phi(r)$, then by Jensen's inequality,

$$E[\phi(r)] \geq \phi[E(r)].$$

Thus,

$$\begin{aligned} E(\hat{p}_k) &\geq 1 - \left(1 - \frac{E(r)}{g}\right)^{\frac{1}{k}} \\ &= 1 - \left(1 - \frac{gp^*}{g}\right)^{\frac{1}{k}} \\ &= p. \end{aligned}$$

Approach 2:

Let

$$x = \left(1 - \frac{r}{g}\right)^{\frac{1}{k}}$$

then

$$\frac{\partial x}{\partial r} = \frac{-1}{gk} \left(1 - \frac{r}{g}\right)^{\frac{1}{k}-1}$$

and

$$\frac{\partial^2 x}{\partial r^2} = \frac{1-k}{(gk)^2} \left(1 - \frac{r}{g}\right)^{\frac{1}{k}-2} \leq 0 \quad \text{for } k > 1$$

Therefore, x is concave.

Thus,

$$E\left(1 - \frac{r}{g}\right)^{\frac{1}{k}} \leq \left(1 - \frac{E(r)}{g}\right)^{\frac{1}{k}}$$

implying that

$$1 - E\left(1 - \frac{r}{g}\right)^{\frac{1}{k}} \geq 1 - \left(1 - \frac{E(r)}{g}\right)^{\frac{1}{k}}$$

and that

$$E\left(1 - \frac{r}{g}\right)^{\frac{1}{k}} \geq \left(1 - \frac{E(r)}{g}\right)^{\frac{1}{k}}$$

Hence

$$\begin{aligned} E(\hat{p}_k) &\geq 1 - \left(1 - \frac{gp^*}{g}\right)^{\frac{1}{k}} \\ &= 1 - \left(1 - \frac{gp^*}{g}\right)^{\frac{1}{k}} \\ &= p. \end{aligned}$$

Approach 3:

From formula (2.8),

$$E(\hat{p}_k) = 1 - \frac{1}{g^{\frac{1}{k}}} E(g - r)^{\frac{1}{k}}.$$

Let

$$y = (g - r)^{\frac{1}{k}},$$

then,

$$\frac{\partial y}{\partial r} = \frac{-1}{k} (g - r)^{\frac{1}{k} - 1}$$

and

$$\frac{\partial^2 y}{\partial r^2} = \frac{1 - k}{k^2} (g - r)^{\frac{1}{k} - 2} \leq 0.$$

Therefore,

$$E(g - r)^{\frac{1}{k}} \leq (g - E(r))^{\frac{1}{k}}$$

or

$$E\left(1 - \frac{1}{g^{\frac{1}{k}}}(g - r)^{\frac{1}{k}}\right) \geq 1 - \frac{1}{g^{\frac{1}{k}}}(g - E(r))^{\frac{1}{k}}$$

and hence

$$\begin{aligned} E(\hat{p}_k) &\geq 1 - \frac{1}{g^{\frac{1}{k}}}(g - gp^*)^{\frac{1}{k}} \\ &= p. \end{aligned}$$

Thus, for $k > 1$, \hat{p}_k overestimates p with exact bias given by

$$\vec{B}_{exact} = E(\hat{p}_k) - p. \quad (2.13)$$

The table below shows the values of $E(\hat{p}_k)$ for selected p , k and g .

Table 2. $E(\hat{p}_k)$ for selected values of p, g and k .

Number of groups, g

$p=0.01$

k	10	20	30	40	70	100
1	1.000E-02	1.000E-02	1.000E-02	1.000E-02	1.000E-02	1.000E-02
5	1.044E-02	1.021E-02	1.014E-02	1.010E-02	1.006E-02	1.004E-02
10	1.051E-02	1.024E-02	1.016E-02	1.012E-02	1.007E-02	1.005E-02
15	1.055E-02	1.026E-02	1.017E-02	1.013E-02	1.007E-02	1.005E-02
20	1.057E-02	1.027E-02	1.018E-02	1.013E-02	1.008E-02	1.005E-02

$p=0.05$

k	10	20	30	40	70	100
1	5.000E-02	5.000E-02	5.000E-02	5.000E-02	5.000E-02	5.000E-02
5	5.243E-02	5.116E-02	5.076E-02	5.057E-02	5.032E-02	5.022E-02
10	5.336E-02	5.152E-02	5.099E-02	5.074E-02	5.042E-02	5.029E-02
15	5.572E-02	5.187E-02	5.121E-02	5.089E-02	5.050E-02	5.035E-02
20	6.498E-02	5.240E-02	5.146E-02	5.107E-02	5.060E-02	5.041E-02

$p=0.10$

k

1	0.1000	0.1000	0.1000	0.1000	0.1000	0.1000
5	0.1057	0.1026	0.1017	0.1013	0.1007	0.1005
10	0.1190	0.1044	0.1027	0.1020	0.1011	0.1008
15	0.1905	0.1142	0.1049	0.1031	0.1016	0.1011
20	0.3391	0.1695	0.1227	0.1091	0.1026	0.1017

$p=0.20$

k

1	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000
5	0.2251	0.2075	0.2047	0.2035	0.2019	0.2013
10	0.4407	0.2835	0.2330	0.2156	0.2050	0.2032
15	0.7398	0.5743	0.4619	0.3846	0.2676	0.2273
20	0.9022	0.8211	0.7506	0.6889	0.5440	0.4435

We note from the table that for small group sizes and large values of p and k , the bias is considerably large. However, for low values of p and k , even fairly small group sizes suffice to yield estimates of p , with relatively small bias.

Because the MLE can result in serious bias, we consider the possibility of bias correction. Using standard methods given in Rao (1973), the following approximation is suggested.

$$B_{crctd} = \frac{k-1}{2gk^2} \left[\frac{1 - (1-p)^k}{(1-p)^{k-1}} \right]. \quad (2.14)$$

The table below shows the corrected bias for selected values of p , k and g .

Table 3. Bias Corrected $E(\hat{p}_k)$ for selected values of p, g and k .

Number of groups, g					
$p=0.005$					
k	5	10	15	20	30
1	0.005000	0.005000	0.005000	0.005000	0.005000
2	0.005250	0.005130	0.005080	0.005063	0.005042
5	0.005400	0.005200	0.005130	0.005101	0.005067
10	0.005460	0.005230	0.005150	0.005115	0.005077
20	0.005500	0.005250	0.005170	0.005125	0.005083
$p=0.01$					
k	5	10	15	20	30
1	0.01000	0.01000	0.01000	0.01000	0.01000
2	0.01050	0.01025	0.01017	0.01013	0.01008
5	0.01082	0.01041	0.01027	0.01020	0.01014
10	0.01094	0.01047	0.01031	0.01024	0.01016
20	0.01105	0.01052	0.01035	0.01026	0.01017

$p=0.05$

k

1	0.05000	0.05000	0.05000	0.05000	0.05000
2	0.05257	0.05128	0.05086	0.05064	0.05043
5	0.05444	0.05222	0.05148	0.05111	0.05074
10	0.05573	0.05287	0.05191	0.05143	0.05096
20	0.05808	0.05404	0.05269	0.05202	0.05135

$p=0.10$

k

1	0.10000	0.10000	0.10000	0.10000	0.10000
2	0.10528	0.10264	0.10176	0.10132	0.10088
5	0.10999	0.10499	0.10333	0.10250	0.10166
10	0.11513	0.10757	0.10504	0.10378	0.10252
20	0.13089	0.11544	0.11030	0.10772	0.10515

$p=0.25$

k

1	0.25000	0.25000	0.25000	0.25000	0.25000
2	0.26458	0.25729	0.25486	0.25365	0.25243
5	0.28857	0.26928	0.26286	0.25964	0.25643
10	0.36311	0.30656	0.28770	0.27828	0.26885
20	0.88983	0.80991	0.62328	0.52996	0.43664

We note that there is substantial reduction in bias for small values of p , even for large values of k .

2.3.2 The Variance of the Estimator

The variance of the estimator is given as,

$$\begin{aligned}\text{var}(\hat{p}_k) &= E[\hat{p}_k - E(\hat{p}_k)]^2 \\ &= E[\hat{p}_k - 1 + 1 - E(\hat{p}_k)]^2 \\ &= E(1 - \hat{p}_k)^2 - (1 - E(\hat{p}_k))^2.\end{aligned}$$

But

$$\begin{aligned}E(1 - \hat{p}_k)^2 &= \sum_{r=0}^g \left(1 - \frac{r}{g}\right)^{\frac{2}{k}} \binom{g}{r} p^{*r} (1 - p^*)^{g-r} \\ &= \sum_{i=0}^g \left(\frac{i}{g}\right)^{\frac{2}{k}} \binom{g}{i} p^{*g-i} (1 - p^*)^i.\end{aligned}$$

Therefore,

$$\text{var}(\hat{p}_k) = \sum_{i=0}^g \left(\frac{i}{g}\right)^{\frac{2}{k}} \binom{g}{i} p^{*g-i} (1 - p^*)^i - (1 - E(\hat{p}_k))^2. \quad (2.15)$$

Alternatively, let

$$\alpha_2 = \sum_{i=0}^g i^{\frac{2}{k}} \binom{g}{i} p^{*g-i} (1 - p^*)^i.$$

But from (2.10)

$$E(\hat{p}_k) = 1 - \frac{\alpha_1}{g^{\frac{1}{k}}}.$$

Hence,

$$\begin{aligned}\text{var}(\hat{p}_k) &= \frac{\alpha_2}{g^{\frac{2}{k}}} - \left(\frac{\alpha_1}{g^{\frac{1}{k}}}\right)^2 \\ &= \frac{\alpha_2 - \alpha_1^2}{g^{\frac{2}{k}}}.\end{aligned} \quad (2.16)$$

When $k = 1$

$$\begin{aligned}
 \text{var}(\hat{p}_k) &= \frac{1}{g^2} E(i^2) - (1-p)^2 \\
 &= \frac{1}{g^2} (\text{var}(i) + (E(i))^2) - (1-p)^2 \\
 &= \frac{1}{g^2} \{g(1-p)p + (g(1-p))^2 - g^2(1-p)^2\} \\
 &= \frac{p(1-p)}{g}.
 \end{aligned} \tag{2.17}$$

2.3.3 Asymptotic Variance of the Estimator

Here, we show that as the number of groups becomes large, the variance of the estimator is best asymptotically normally distributed. We approach this using two different methods.

Method 1: Based on Cramer-Rao Bound

The Cramer-Rao bound states that for an estimator \hat{p}_k ,

$$\lim_{g \rightarrow \infty} \text{var}(\hat{p}_k) = \frac{1}{-E\left(\frac{\partial^2 \ln L}{\partial p^2}\right)}.$$

But from eqn. (2.5),

$$\ln L = \ell = \ln \binom{g}{r} + r \ln p^* + (g-r) \ln(1-p^*),$$

implying that

$$\frac{\partial \ell}{\partial p} = \left(\frac{r}{p^*} - \frac{g-r}{1-p^*}\right) \frac{\partial p^*}{\partial p}$$

and

$$\frac{\partial^2 \ell}{\partial p^2} = \left(-\frac{r}{p^{*2}} - \frac{g-r}{(1-p^*)^2}\right) \left(\frac{\partial p^*}{\partial p}\right)^2 + \left(\frac{r}{p^*} - \frac{g-r}{1-p^*}\right) \frac{\partial^2 p^*}{\partial p^2}.$$

Therefore,

$$\begin{aligned}
 -E\left(\frac{\partial^2 \ell}{\partial p^2}\right) &= \left(\frac{g}{p^*} + \frac{g}{1-p^*}\right)\left(\frac{\partial p^*}{\partial p}\right)^2 + (g-g)p^* \frac{\partial^2 p^*}{\partial p^2} \\
 &= \left(\frac{g}{p^*} + \frac{g}{1-p^*}\right)\left(\frac{\partial p^*}{\partial p}\right)^2 \\
 &= \frac{g\left(\frac{\partial}{\partial p}(1-(1-p)^k)\right)^2}{(1-(1-p)^k)(1-p)^k} \\
 &= \frac{gk^2(1-p)^{k-2}}{1-(1-p)^k}
 \end{aligned}$$

and hence,

$$\lim_{g \rightarrow \infty} \text{var}(\hat{p}_k) = \frac{1-(1-p)^k}{gk^2(1-p)^{k-2}}. \quad (2.18)$$

Method 2: Based on the Delta Method

Here, we use the fact that if y is a random variable and a function of another random variable x ie $y = f(x)$, then

$$\text{var}(y) \simeq \text{var}(x)(f'(x))^2.$$

In this case, $p^* = 1 - (1-p)^k \Rightarrow \hat{p}_k = 1 - (1 - \hat{p}^*)^{\frac{1}{k}}$.

Therefore,

$$\begin{aligned}
 \text{var}(\hat{p}_k) &\approx \text{var}(\hat{p}^*)\left(\frac{\partial \hat{p}_k}{\partial p^*}\right)^2 \\
 &= \frac{p^*(1-p^*)}{g} \left(\frac{\partial}{\partial p^*}(1-(1-p^*)^{\frac{1}{k}})\right)^2 \\
 &= \frac{p^*(1-p^*)}{g} \left(\frac{1}{k}(1-p^*)^{\frac{1}{k}-1}\right)^2
 \end{aligned}$$

$$\begin{aligned}
&= \frac{p^*(1-p^*)^{\frac{2}{k}-1}}{gk^2} \\
&= \frac{(1-(1-p)^k)(1-p)^{2-k}}{gk^2} \\
&= \frac{1-(1-p)^k}{gk^2(1-p)^{k-2}}. \tag{2.19}
\end{aligned}$$

Thus, the two methods give the same asymptotic variance for p and shows that the estimator is Best Asymptotically Normal for large g . That is, for fixed k and $g \rightarrow \infty$

$$\sqrt{g}(\hat{p}_k - p) \rightarrow Normal(0, \frac{1-(1-p)^k}{gk^2(1-p)^{k-2}}). \tag{2.20}$$

In the next section, we are interested in how the asymptotic variance behaves for fixed group-factor size and fixed total number of group-factors.

2.3.4 The Behavior of the Asymptotic Variance

(a) For fixed group factor size, k

$$\begin{aligned}
\text{var}(\hat{p}_k) &= \frac{(1-p)^2}{gk^2} ((1-p)^{-k} - 1) \\
&= \frac{\text{constant}}{g},
\end{aligned}$$

which implies that

$$\frac{\partial}{\partial g} \text{var}(\hat{p}_k) = \frac{-\text{constant}}{g^2} < 0, \quad \text{for } 0 < p < 1.$$

Thus, for fixed k , the variance of the estimator is inversely proportional to the number of group-factors g . This implies that the larger the number of group-factors, the smaller the variance and hence the more reliable will be

the estimate.

(b) For fixed total number of factors, $f = gk$,

$$\text{var}(\hat{p}_k) = \frac{(1-p)^2 \left((1-p)^{-k} - 1 \right)}{f k}$$

We check whether the variance is monotonic increasing or decreasing.

Consider the two functions

$$\frac{(1-p)^2 \left((1-p)^{-k} - 1 \right)}{f k}$$

and

$$\frac{(1-p)^2 \left((1-p)^{-(k+1)} - 1 \right)}{f (k+1)}$$

By binomial expansion

$$\begin{aligned} \frac{(1-p)^{-k} - 1}{k} &= \frac{\left(1 + kp + \frac{k(k+1)p^2}{2!} + \frac{k(k+1)(k+2)p^3}{3!} + \dots \right) - 1}{k} \\ &= p \left(1 + \frac{(k+1)p}{2!} + \frac{(k+1)(k+2)p^2}{3!} + \dots \right) \quad (2.21) \end{aligned}$$

and

$$\begin{aligned} \frac{(1-p)^{-(k+1)} - 1}{k+1} &= \frac{\left(1 + (k+1)p + \frac{(k+1)(k+2)p^2}{2!} + \frac{(k+1)(k+2)(k+3)p^3}{3!} + \dots \right) - 1}{k+1} \\ &= p \left(1 + \frac{(k+2)p}{2!} + \frac{(k+2)(k+3)p^2}{3!} + \dots \right). \quad (2.22) \end{aligned}$$

Since eqn (2.22) is greater than eqn (2.21) for all $k \geq 1$, it implies that the function is monotonic increasing and hence the variance of the estimate increases as the group-factor size increases.

In the next section, we (1) find the optimal value of k , ie, the group-factor

size which minimizes MSE and total cost, for a given level of p and fixed number of groups g ; (2) compare the MSE of \hat{p}_k when $k = 1$ and when $k > 1$; and (3) compare the cost of achieving tolerable level of MSE with group and individual sampling plans.

2.3.5 Mean Squared Error

The mean squared error of an estimator is the average squared deviation from the true value of the parameter, which incorporates measures of both accuracy (bias) and precision (variance) of the estimator. In this case,

$$\begin{aligned}
 MSE(\hat{p}_k) &= E[\hat{p}_k - E(\hat{p}_k)]^2 + [E(\hat{p}_k) - p]^2 \\
 &= E[\hat{p}_k^2 + [E(\hat{p}_k)]^2 - 2\hat{p}_k E(\hat{p}_k)] + [[E(\hat{p}_k)]^2 + p^2 - 2pE(\hat{p}_k)] \\
 &= E(\hat{p}_k^2) + p^2 - 2pE(\hat{p}_k) \\
 &= E(\hat{p}_k - p)^2 \\
 &= E\left[1 - \left(1 - \frac{r}{g}\right) - p\right]^2 \\
 &= E\left[(1 - p) - \left(1 - \frac{r}{g}\right)^{\frac{1}{k}}\right]^2 \\
 &= (1 - p)^2 + E\left[1 - \frac{r}{g}\right]^{\frac{2}{k}} - 2(1 - p)E\left(1 - \frac{r}{g}\right)^{\frac{1}{k}} \\
 &= (1 - p)^2 + \sum_{i=0}^g \left(\frac{i}{g}\right)^{\frac{2}{k}} \binom{g}{i} [(1 - p)^k]^i [1 - (1 - p)^k]^{g-i} \\
 &\quad - 2(1 - p) \sum_{i=0}^g \left(\frac{i}{g}\right)^{\frac{1}{k}} \binom{g}{i} [(1 - p)^k]^i [1 - (1 - p)^k]^{g-i} \\
 &= (1 - p)^2 + \sum_{i=0}^g \left(\frac{i}{g}\right)^{\frac{1}{k}} \left[\left(\frac{i}{g}\right)^{\frac{1}{k}} - 2(1 - p)\right] \binom{g}{i} \delta_k^i (1 - \delta_k)^{g-i}, \quad (2.23)
 \end{aligned}$$

where $\delta_k = (1 - p)^k$.

The table below gives the MSE for selected values of p , g and k . The un-

derlined values represent minimal MSE, which correspond to optimal k .

Table 4. MSE for selected values of p, g and k .

Number of group-factors, g

p=0.20						
k	10	20	30	40	70	100
1	1.600E-02	8.000E-03	5.333E-03	4.000E-03	2.286E-03	1.600E-03
2	9.802E-03	4.681E-03	3.078E-03	2.293E-03	1.300E-03	9.068E-04
3	<u>8.356E-03</u>	3.663E-03	2.375E-03	1.396E-03	9.884E-04	6.876E-04
4	1.005E-02	<u>3.284E-03</u>	2.081E-03	1.207E-03	8.506E-04	5.895E-04
5	1.768E-02	3.365E-03	<u>1.975E-03</u>	1.123E-03	7.861E-04	5.424E-04
6	3.492E-02	4.547E-03	2.040E-03	<u>1.416E-03</u>	<u>7.630E-04</u>	5.232E-04
7	6.391E-02	8.713E-03	2.577E-03	1.508E-03	7.691E-04	<u>5.230E-04</u>
8	0.104167023	1.886E-02	4.653E-03	1.950E-03	8.025E-04	5.384E-04
9	0.153026644	3.797E-02	1.047E-02	3.614E-03	8.854E-04	5.693E-04
10	0.206815486	6.763E-02	2.297E-02	8.410E-03	1.166E-03	6.240E-04

p=0.10

k	10	20	30	40	70	100
1	9.000E-03	4.500E-03	3.000E-03	2.250E-03	1.286E-03	9.000E-04
2	5.068E-03	2.450E-03	1.616E-03	1.206E-03	6.845E-04	4.779E-04
3	3.723E-03	1.759E-03	1.153E-03	8.571E-04	4.846E-04	3.378E-04
4	3.088E-03	1.420E-03	9.243E-04	6.854E-04	3.861E-04	2.688E-04
5	<u>2.807E-03</u>	1.224E-03	7.917E-04	4.642E-04	3.285E-04	2.283E-04
6	2.876E-03	1.103E-03	7.076E-04	5.213E-04	2.914E-04	2.023E-04
7	3.494E-03	1.027E-03	6.521E-04	4.785E-04	2.663E-04	1.845E-04
8	5.031E-03	<u>9.870E-04</u>	6.151E-04	3.541E-04	2.488E-04	1.721E-04
9	7.984E-03	9.931E-04	5.915E-04	4.296E-04	2.365E-04	1.632E-04
10	1.291E-02	1.087E-03	<u>5.794E-04</u>	3.266E-04	2.280E-04	1.570E-04
11			5.816E-04	4.102E-04	2.224E-04	1.527E-04
12				<u>4.093E-04</u>	2.192E-04	1.500E-04
13				4.165E-04	<u>2.180E-04</u>	1.487E-04
14					2.186E-04	<u>1.485E-04</u>
15						1.494E-04

$p=0.01$

k	10	20	30	40	70	100
1	9.900E-04	4.950E-04	3.300E-04	2.475E-04	1.414E-04	9.900E-05
2	5.246E-04	2.553E-04	1.687E-04	1.260E-04	7.159E-05	5.001E-05
5	2.212E-04	1.056E-04	6.933E-05	5.162E-05	2.922E-05	2.038E-05
10	1.159E-04	5.470E-05	3.580E-05	2.661E-05	1.503E-05	1.047E-05
15	8.058E-05	3.768E-05	2.460E-05	1.826E-05	1.030E-05	7.160E-06
20	6.302E-05	2.919E-05	1.901E-05	1.410E-05	7.940E-06	5.520E-06
25	5.282E-05	2.414E-05	1.568E-05	1.161E-05	6.530E-06	4.540E-06
30	4.710E-05	2.079E-05	1.347E-05	9.960E-06	5.590E-06	3.890E-06
35	<u>4.610E-05</u>	1.844E-05	1.191E-05	6.980E-06	4.930E-06	3.420E-06
40	5.296E-05	1.670E-05	1.076E-05	7.940E-06	4.440E-06	3.080E-06
45	7.469E-05	1.537E-05	9.000E-06	7.270E-06	4.060E-06	2.820E-06
50	1.235E-04	1.435E-05	9.100E-06	6.751E-06	3.760E-06	2.610E-06
60	3.853E-04	<u>1.300E-05</u>	8.100E-06	4.730E-06	3.330E-06	2.310E-06
70	1.088E-03	1.307E-05	<u>7.500E-06</u>	4.320E-06	3.030E-06	2.090E-06
100	1.030E-02	1.186E-04	7.700E-06	4.740E-06	2.570E-06	1.770E-06

$p=0.05$

k	10	20	30	40	70	100
1	4.750E-03	2.375E-03	1.583E-03	1.188E-03	6.786E-04	4.750E-04
2	2.582E-03	1.253E-03	8.277E-04	6.179E-04	3.510E-04	2.451E-04
5	1.196E-03	5.598E-04	3.657E-04	2.716E-04	1.533E-04	1.067E-04
8	8.745E-04	3.888E-04	2.518E-04	1.862E-04	1.045E-04	7.268E-05
10	<u>8.509E-04</u>	3.346E-04	2.152E-04	1.257E-04	8.877E-05	6.163E-05
15	2.404E-03	<u>2.750E-04</u>	1.709E-04	9.836E-05	6.909E-05	4.777E-05
20		3.789E-04	<u>1.569E-04</u>	1.120E-04	6.104E-05	4.198E-05
25			2.068E-04	<u>1.111E-04</u>	<u>5.816E-05</u>	3.971E-05
30				1.702E-04	5.853E-05	<u>3.951E-05</u>
35						4.096E-05

In general, the MSE decreases to a minimum as the optimal value of k is attained, then increases. Lower MSE values occur with larger number of group-factors, g . After the minimum MSE is reached, the rate of increase in MSE with k decreases as the number of group-factors increases.

2.3.6 The Cost Function

Define the cost of obtaining the estimate, when $k = 1$ as

$$C_1 = f \times (C_S + C_A), \quad (2.24)$$

and the cost of the group testing as

$$C_k = f \times \left(C_S + \frac{C_A}{k} \right), \quad (2.25)$$

where C_1 and C_k are total costs for one at a time tests and group tests respectively, C_S is the sampling cost of one individual sample and C_A is the cost of performing one test. The reduction in cost is emphasized when C_A is substantially greater than C_S . Suppose C_A is five times C_S from either an individual test or group test. The table above may be useful in comparing alternative experiments. For example, if the objective is to achieve a tolerable MSE of say 0.001 when $p = 0.05$. When 10 groups are tested, a design in which $k = 10$ yields an MSE of 0.0008509. When 30 groups are tested, a design in which $k = 2$ yields an MSE of 0.0008277. The traditional plan ($k = 1$) would require approximately 70 groups to achieve the desired MSE of 0.0006786. The costs of these three alternative experiments are estimated as follows:

$$\begin{aligned}
 C_k &= g \times k \times C_S + g \times C_A \\
 C_{k=10} &= 10 \times 10 \times 1 + 10 \times 5 \\
 &= 150 \text{ cost units} \\
 C_{k=2} &= 30 \times 2 \times 1 + 30 \times 5 \\
 &= 210 \text{ cost units} \\
 C_{k=1} &= 70 \times 1 \times 1 + 70 \times 5 \\
 &= 420 \text{ cost units.}
 \end{aligned}$$

The first alternative therefore is the most cost-effective. It requires only 35 percent of the cost of the traditional plan to achieve an MSE of 0.001 in this example. Clearly, the best design for a given experiment depends on p , the tolerable MSE, C_A and C_S . A researcher can use the techniques to

select the least costly design that achieves the experimental objectives.

2.3.7 Efficiency

We define Efficiency as the ratio of the MSE of the estimator at optimal k to the MSE when $k = 1$. That is,

$$RE = \frac{MSE(OPTIMUM)}{MSE(TRADITIONAL)}$$

The table below gives the ratio of MSE when optimal k is used to the MSE when the traditional plan $k = 1$ is used.

Table 5. Relative Efficiency for the values of p and g as given in Table 4 above.

p	Number of group-factors, g					
	10	20	30	40	70	100
0.2	0.522245	0.410449	0.370228	0.354037	0.333805	0.326887
0.1	0.311921	0.219328	0.193137	0.181892	0.169565	0.165007
0.05	0.179135	0.11579	0.099078	0.093547	0.08571	0.083187
0.01	0.046562	0.026262	0.022727			

At low values of p , the MSE for optimal k is always less than the MSE for the traditional plan for a given number of group-factors. Increasing the number of group-factors results in (1) reducing MSE (2) increasing optimal k and (3) reducing the ratio of optimal MSE to traditional MSE.

Chapter 3

Maximum Likelihood Estimation With Equal Probabilities and With Errors in Decisions

3.1 Introduction

In this chapter, we consider group screening design with equal probability, equal group sizes and with errors in decision. The presence of measurement errors resulting from the limited precision of tests makes estimation using traditional methods, impossible in some screening situations. Ignoring measurement errors leads to severe bias, and inference about the prevalence becomes unsatisfactory. The probability of a group factor being declared defective is determined using two approaches, the sensitivity-specificity approach discussed in section 3.2 and test of hypothesis approach considered in section 3.4. Relative efficiency of the group estimate against the traditional estimate is considered in section 3.3.

3.2 Sensitivity-Specificity Approach

Sensitivity is defined as the probability of correctly identifying a defective item, while specificity is the probability of correctly classifying a non-defective item. Define the following indicator variables:

$$T_i = \begin{cases} 1, & \text{if the } i\text{th batch screening test is positive} \\ 0, & \text{otherwise.} \end{cases}$$

and

$$D_i = \begin{cases} 1, & \text{if at least one individual in the } i\text{th batch is positive} \\ 0, & \text{otherwise.} \end{cases}$$

Further, let

$$\begin{aligned} \eta &= \text{Prob}(T_i = 1/D_i = 1) \\ &= \text{sensitivity of the screening test} \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} \theta &= \text{Prob}(T_i = 0/D_i = 0) \\ &= \text{specificity of the screening test.} \end{aligned} \tag{3.2}$$

Note from eqn (2.1) that

$$\text{Prob}(D_i = 1) = 1 - q^k$$

and

$$\text{Prob}(D_i = 0) = q^k.$$

Watson's (1961) result on group-factors declared defective can be proved using the indicator variables, as given in the following theorem.

Theorem 3.2.1

When screening with errors in decisions and using the sensitivity-specificity approach, the probability that a group-factor of size k is declared defective is given as

$$\pi_1^* = [1 - (1 - p)^k]\eta + (1 - p)^k(1 - \theta).$$

where p , k , η and θ are as defined before.

Proof

By definition

$$\begin{aligned} \pi_1^* &= \text{Prob}(T_i = 1) \\ &= \text{Prob}(T_i = 1, D_i = 1) + \text{Prob}(T_i = 1, D_i = 0) \\ &= \text{Prob}(T_i = 1/D_i = 1)\text{Prob}(D_i = 1) + \text{Prob}(T_i = 1/D_i = 0)\text{Prob}(D_i = 0) \\ &= \eta(1 - q^k) + (1 - \theta)q^k \\ &= \eta - (\eta + \theta - 1)q^k \\ &= [1 - (1 - p)^k]\eta + (1 - p)^k(1 - \theta). \end{aligned} \tag{3.3}$$

This completes the proof.

Note that as $p \rightarrow 1$, $[1 - (1 - p)^k]\eta \rightarrow \eta$, hence π_1^* is a monotonic increasing function of p . That is, the higher the prevalence the more likely it is that a group-factor will test positive. This leads to the condition that $1 - \theta \leq \pi_1^* \leq \eta$, with equality if and only if $p = 0$ or $p = 1$ respectively.

3.2.1 Maximum Likelihood Estimator of p

Using the binomial model, the likelihood function is given by

$$L = \binom{g}{r} \pi_1^{*r} (1 - \pi_1^*)^{g-r},$$

where r is the number of group-factors declared defective.

Taking logarithms we have

$$\ln L = \ell = \ln \binom{g}{r} + r \ln \pi_1^* + (g - r) \ln(1 - \pi_1^*).$$

Differentiating partially with respect π_1^* and equating to zero gives

$$\hat{\pi}_1^* = \frac{r}{g}. \quad (3.4)$$

That is,

$$\eta - (\eta + \theta - 1) \hat{q}^k = \frac{r}{g},$$

which implies that

$$\hat{q}^k = \left(\frac{1}{\eta + \theta - 1} \right) \left(\eta - \frac{r}{g} \right),$$

so that assuming that η and θ are known,

$$\hat{p} = 1 - \left[\left(\frac{1}{\eta + \theta - 1} \right) \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k}}$$

or

$$\hat{p} = 1 - \left[\frac{\eta - \hat{\pi}_1^*}{\eta + \theta - 1} \right]^{\frac{1}{k}}. \quad (3.5)$$

If $\eta = \theta = 1$ and $\pi_1^* = \frac{r}{g}$, we have

$$\hat{p} = 1 - \left[1 - \frac{r}{g} \right]^{\frac{1}{k}},$$

which is the same as the result in equation (2.7), for the case of no errors. Since both sensitivity and specificity for testing kits are always $\gg 0.5$, $\eta + \theta - 1 > 0$. Thus \hat{p} is well defined if $\hat{\pi}_1^* \leq \eta$. By monotonicity of π_1^* as a function of p , it follows that

$$1 - \theta \leq \hat{\pi}_1^* = \frac{r}{g} \leq \eta. \quad (3.6)$$

This is the condition for the prevalence to be estimable given the observed ratio $\frac{r}{g}$ and the sensitivity η and specificity θ . Thus, the probabilities of estimating prevalence are given by

$$\begin{aligned} \Pr[\hat{\pi}_1^* \geq 1 - \theta] &= \Pr[r \geq g(1 - \theta)] \\ &= \sum_{r=[g(1-\theta)]}^g \binom{g}{r} \pi_1^{*r} (1 - \pi_1^*)^{g-r} \\ &= \sum_{i=0}^{g\theta} \binom{g}{i} (1 - \pi_1^*)^i \pi_1^{*g-i}, \end{aligned} \quad (3.7)$$

and

$$\begin{aligned} \Pr[\hat{\pi}_1^* \leq \eta] &= 1 - \Pr[r \geq g\eta] \\ &= 1 - \sum_{r=g\eta}^g \binom{g}{r} \pi_1^{*r} (1 - \pi_1^*)^{g-r} \\ &= 1 - \sum_{i=0}^{g(1-\eta)} \binom{g}{i} (1 - \pi_1^*)^i \pi_1^{*g-i}. \end{aligned} \quad (3.8)$$

These probabilities may be evaluated by summing up the terms of the binomial distribution, and may be used to study whether group screening

improves the probability of estimating prevalence for various values of η and θ for fixed number of groups g . Thus, in the case of a rare trait, specificity plays a major role in investigating the effect of pooling and estimating the prevalence based on test results than sensitivity. Thus, the inequality $\pi_1^* \geq 1 - \theta$ is more useful in determining the probability of prevalence. In the case of high prevalence, equation (3.8) indicates that sensitivity plays a major role in the estimation than specificity.

3.2.2 The Mean of the Estimator and its Biasedness

The mean of the estimator \hat{p} is given by

$$\begin{aligned} E(\hat{p}) &= \sum_{r=0}^g \left\{ 1 - \left[\left(\frac{1}{\eta + \theta - 1} \right) \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k}} \right\} \binom{g}{r} \pi_1^{*r} (1 - \pi_1^*)^{g-r} \\ &= 1 - \sum_{r=0}^g \left\{ \left[\left(\frac{1}{\eta + \theta - 1} \right) \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k}} \right\} \binom{g}{r} \pi_1^{*r} (1 - \pi_1^*)^{g-r}. \end{aligned} \quad (3.9)$$

Letting $i = g - r$, we have

$$\begin{aligned} E(\hat{p}) &= 1 - \sum_{i=0}^g \left[\frac{1}{(\eta + \theta - 1)^{\frac{1}{k}}} \left(\eta - 1 + \frac{i}{g} \right)^{\frac{1}{k}} \binom{g}{i} \pi_1^{*g-i} (1 - \pi_1^*)^i \right] \\ &= 1 - \sum_{i=0}^g \left[\frac{1}{(g(\eta + \theta - 1))^{\frac{1}{k}}} (g(\eta - 1) + i)^{\frac{1}{k}} \binom{g}{i} \pi_1^{*g-i} (1 - \pi_1^*)^i \right] \\ &= 1 - \frac{\alpha_1^*}{(g(\eta + \theta - 1))^{\frac{1}{k}}}, \end{aligned} \quad (3.10)$$

where

$$\alpha_1^* = \sum_{i=0}^g (g(\eta - 1) + i)^{\frac{1}{k}} \binom{g}{i} \pi_1^{*g-i} (1 - \pi_1^*)^i. \quad (3.11)$$

Equation(2.10) is a special case of equation (3.10), by letting $\eta = \theta = 1$ and $\pi_1^* = [1 - (1 - p)^k] + (1 - p)^k$. When $k = 1$,

$$\begin{aligned}
 E(\hat{p}) &= 1 - \frac{1}{g(\eta + \theta - 1)}(g(\eta - 1) + E(i)) \\
 &= 1 - \frac{1}{g(\eta + \theta - 1)}(g(\eta - 1) + g(1 - \pi_1^*)) \\
 &= 1 - \frac{1}{\eta + \theta - 1}(\eta - \{\eta - (\eta + \theta - 1)q^k\}) \\
 &= 1 - q^k \\
 &= p.
 \end{aligned}$$

Thus, for $k = 1$, \hat{p} is an unbiased estimator of p . However, for $k > 1$ \hat{p} is biased and overestimates p . This claim can be proved using Jensen's inequality.

In this study, we assume that $\hat{p} = 1 - [\frac{1}{\eta + \theta - 1}(\eta - \frac{r}{g})]^{\frac{1}{k}}$ is continuous. Then,

$$\frac{\partial \hat{p}}{\partial r} = \frac{1}{gk} \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k} - 1}$$

and

$$\frac{\partial^2 \hat{p}}{\partial r^2} = \frac{k - 1}{(gk)^2} \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k} - 2} \geq 0,$$

so long as $\eta \geq \frac{r}{g}$.

Therefore, under this condition

$$\begin{aligned}
 E(\hat{p}) &= E \left\{ 1 - \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k}} \right\} \\
 &\geq 1 - \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{E(r)}{g} \right) \right]^{\frac{1}{k}} \\
 &= 1 - \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{g\pi_1^*}{g} \right) \right]^{\frac{1}{k}} \\
 &= 1 - \left[\frac{1}{\eta + \theta - 1} (\eta - \pi_1^*) \right]^{\frac{1}{k}}
 \end{aligned}$$

$$\begin{aligned}
&= 1 - \left[\frac{1}{\eta + \theta - 1} (\eta - (\eta - (\eta + \theta - 1)q^k)) \right]^{\frac{1}{k}} \\
&= 1 - \left[\frac{1}{\eta + \theta - 1} (\eta + \theta - 1)q^k \right]^{\frac{1}{k}} \\
&= 1 - q \\
&= p.
\end{aligned}$$

That is, $E(\hat{p}) \geq p$.

Alternatively, let

$$x = \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k}},$$

so that

$$\frac{\partial x}{\partial r} = \frac{-1}{gk} \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k} - 1}$$

and

$$\frac{\partial^2 x}{\partial r^2} = \frac{1 - k}{(gk)^2} \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k} - 2} \leq 0,$$

for $k > 1$ and $\eta \geq \frac{r}{g}$.

Therefore,

$$\begin{aligned}
E \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k}} &\leq \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{E(r)}{g} \right) \right]^{\frac{1}{k}} \\
&= \left[\frac{1}{\eta + \theta - 1} (\eta - \pi_1^*) \right]^{\frac{1}{k}} \\
&= \left[\frac{1}{\eta + \theta - 1} (\eta - (\eta - (\eta + \theta - 1)q^k)) \right]^{\frac{1}{k}} \\
&= q.
\end{aligned}$$

Therefore,

$$E \left[\frac{1}{\eta + \theta - 1} \left(\eta - \frac{r}{g} \right) \right]^{\frac{1}{k}} \leq q,$$

which implies that

$$-E\left[\frac{1}{\eta + \theta - 1}\left(\eta - \frac{r}{g}\right)\right]^{\frac{1}{k}} \geq -q.$$

That is,

$$1 - E\left[\frac{1}{\eta + \theta - 1}\left(\eta - \frac{r}{g}\right)\right]^{\frac{1}{k}} \geq 1 - q.$$

Hence,

$$E(\hat{p}) \geq p.$$

Thus, \hat{p} overestimates p .

Also, suppose we express $E(\hat{p})$ as

$$E(\hat{p}) = 1 - \frac{1}{(g(\eta + \theta - 1))^{\frac{1}{k}}} E(\eta g - r)^{\frac{1}{k}}.$$

Let $y = (\eta g - r)^{\frac{1}{k}}$. Then

$$\frac{\partial y}{\partial r} = -\frac{1}{k}(\eta g - r)^{\frac{1}{k}-1}$$

and

$$\frac{\partial^2 y}{\partial r^2} = \frac{1-k}{k^2}(\eta g - r)^{\frac{1}{k}-2} \leq 0 \quad \text{for } k > 1.$$

Therefore,

$$\begin{aligned} E(\eta g - r)^{\frac{1}{k}} &\leq (\eta g - E(r))^{\frac{1}{k}} \\ &= (\eta g - g\pi_1^*)^{\frac{1}{k}} \\ &= g^{\frac{1}{k}}(\eta - \pi_1^*)^{\frac{1}{k}} \\ &= g^{\frac{1}{k}}((\eta + \theta - 1)q)^{\frac{1}{k}} \\ &= (g(\eta + \theta - 1))^{\frac{1}{k}}q. \end{aligned}$$

Thus,

$$\frac{E(\eta g - r)^{\frac{1}{k}}}{(g(\eta + \theta - 1)q^k)^{\frac{1}{k}}} \leq q,$$

which implies that

$$1 - \frac{E(\eta g - r)^{\frac{1}{k}}}{(g(\eta + \theta - 1)q^k)^{\frac{1}{k}}} \geq 1 - q$$

and hence for $k > 1$ and $\eta \geq \frac{r}{g}, \hat{p}$ overestimates p .

3.2.3 The Variance of the Estimator

The variance of the estimator is given by

$$\begin{aligned} \text{var}(\hat{p}) &= E[\hat{p} - E(\hat{p})]^2 \\ &= E(1 - \hat{p})^2 - (1 - E(\hat{p}))^2. \end{aligned}$$

But

$$\begin{aligned} E(1 - \hat{p})^2 &= \frac{1}{(\eta + \theta - 1)^{\frac{2}{k}}} E\left(\eta - \frac{r}{g}\right)^{\frac{2}{k}} \\ &= \frac{1}{(\eta + \theta - 1)^{\frac{2}{k}}} \sum_{r=0}^g \left(\eta - \frac{r}{g}\right)^{\frac{2}{k}} \binom{g}{r} \pi_1^{*r} (1 - \pi_1^*)^{g-r} \\ &= \frac{1}{(\eta + \theta - 1)^{\frac{2}{k}}} \sum_{i=0}^g \left(\eta - 1 + \frac{i}{g}\right)^{\frac{2}{k}} \binom{g}{i} \pi_1^{*g-i} (1 - \pi_1^*)^i \\ &= \frac{1}{(g(\eta + \theta - 1))^{\frac{2}{k}}} \sum_{i=0}^g (g(\eta - 1) + i)^{\frac{2}{k}} \binom{g}{i} \pi_1^{*g-i} (1 - \pi_1^*)^i \\ &= \frac{\alpha_2^*}{(g(\eta + \theta - 1))^{\frac{2}{k}}}, \end{aligned} \tag{3.12}$$

where $i = g - r$ and

$$\alpha_2^* = \sum_{i=0}^g (g(\eta - 1) + i)^{\frac{2}{k}} \binom{g}{i} \pi_1^{*g-i} (1 - \pi_1^*)^i. \quad (3.13)$$

Therefore,

$$\text{var}(\hat{p}) = \frac{\alpha_2^*}{(g(\eta + \theta - 1))^{\frac{2}{k}}} - (1 - E(\hat{p}))^2. \quad (3.14)$$

Suppose $\eta = \theta = 1$, then

$$\begin{aligned} \text{var}(\hat{p}) &= \frac{1}{g^{\frac{2}{k}}} \sum_{i=0}^g i^{\frac{2}{k}} \binom{g}{i} \pi_1^{*g-i} (1 - \pi_1^*)^i - (1 - E(\hat{p}))^2 \\ &= \frac{E(i^{\frac{2}{k}})}{g^{\frac{2}{k}}} - (1 - E(\hat{p}))^2. \end{aligned} \quad (3.15)$$

Thus, equation (2.16) is a special case of equation (3.15). When $k = 1$ equation (3.14) becomes,

$$\begin{aligned} \frac{E(i^2)}{g^2} - (1 - p)^2 &= \frac{1}{g^2} (\text{var}i + (E(i))^2) - (1 - p)^2 \\ &= \frac{1}{g^2} (g(1 - p)p + g^2(1 - p)^2) - (1 - p)^2 \\ &= \frac{(1 - p)p}{g}. \end{aligned} \quad (3.16)$$

3.2.4 Asymptotic Variance of the Estimator

The asymptotic variance of the estimator in this case is explained by the following result given Xin, Litvak and Pagano(1995).

Theorem 3.2.2

When screening with errors in decisions and using the sensitivity-specificity approach, the asymptotic variance of the estimator \hat{p} of p , the prevalence rate given in equation (3.5), is expressed as

$$\text{var}(\hat{p}) = \frac{(1-p)^2}{f} \frac{(1-\pi_1^*)}{(\eta + \theta - 1)^2} \frac{(1-p)^{-2k} \pi_1^*}{k}$$

for $0 < p < 1$, where k , π_1^* , η , and θ are as defined earlier.

Proof

We use two alternative methods to prove this result.

Method 1: Based on Cramer-Rao Bound

(i) Using the formula

$$\lim_{g \rightarrow \infty} \text{var}(\hat{p}) = \frac{1}{-E\left(\frac{\partial^2 \ln L}{\partial p^2}\right)},$$

the binomial model given in equation (3.4) implies that

$$\frac{\partial \ell}{\partial p} = \left(\frac{r}{\pi_1^*} - \frac{g-r}{1-\pi_1^*}\right) \frac{\partial \pi_1^*}{\partial p},$$

and

$$\frac{\partial^2 \ell}{\partial p^2} = \left(-\frac{r}{\pi_1^{*2}} - \frac{g-r}{(1-\pi_1^*)^2}\right) \left(\frac{\partial \pi_1^*}{\partial p}\right)^2 + \left(\frac{r}{\pi_1^*} - \frac{g-r}{1-\pi_1^*}\right) \frac{\partial^2 \pi_1^*}{\partial p^2}.$$

Therefore,

$$\begin{aligned} -E\left(\frac{\partial^2 \ell}{\partial p^2}\right) &= \left[\frac{g\pi_1^*}{\pi_1^{*2}} + \frac{g-g\pi_1^*}{(1-\pi_1^*)^2}\right] \left(\frac{\partial \pi_1^*}{\partial p}\right)^2 - \left[\frac{g\pi_1^*}{\pi_1^*} - \frac{g-g\pi_1^*}{1-\pi_1^*}\right] \frac{\partial^2 \pi_1^*}{\partial p^2} \\ &= g \left[\frac{1}{\pi_1^*} + \frac{1}{(1-\pi_1^*)}\right] \left(\frac{\partial \pi_1^*}{\partial p}\right)^2. \end{aligned}$$

Hence,

$$\lim_{g \rightarrow \infty} \text{var}(\hat{p}) = \frac{\pi_1^*(1 - \pi_1^*)}{g \left(\frac{\partial \pi_1^*}{\partial p} \right)^2}.$$

(ii) Using the formula

$$\lim_{g \rightarrow \infty} \text{var}(\hat{p}) = \frac{1}{E \left(\frac{\partial \ell}{\partial p} \right)^2},$$

$$\begin{aligned} \left(\frac{\partial \ell}{\partial p} \right)^2 &= \left[\frac{r}{\pi_1^*} \frac{\partial \pi_1^*}{\partial p} - \frac{g-r}{1-\pi_1^*} \frac{\partial \pi_1^*}{\partial p} \right]^2 \\ &= \left[\frac{r}{\pi_1^*} - \frac{g-r}{1-\pi_1^*} \right]^2 \left[\frac{\partial \pi_1^*}{\partial p} \right]^2, \end{aligned}$$

$$\begin{aligned} E \left(\frac{\partial \ell}{\partial p} \right)^2 &= E \left[\frac{r}{\pi_1^*} - \frac{g-r}{1-\pi_1^*} \right]^2 \left[\frac{\partial \pi_1^*}{\partial p} \right]^2 \\ &= \left[\frac{E(r^2) - 2g\pi_1^*E(r) + g^2\pi_1^{*2}}{\pi_1^{*2}(1-\pi_1^*)^2} \right] \left[\frac{\partial \pi_1^*}{\partial p} \right]^2 \\ &= \left[\frac{\text{var}(r) + (E(r))^2 - 2g\pi_1^*E(r) + g^2\pi_1^{*2}}{\pi_1^{*2}(1-\pi_1^*)^2} \right] \left[\frac{\partial \pi_1^*}{\partial p} \right]^2 \\ &= \left[\frac{g\pi_1^*(1-\pi_1^*)}{\pi_1^{*2}(1-\pi_1^*)^2} \right] \left[\frac{\partial \pi_1^*}{\partial p} \right]^2 \\ &= \left[\frac{g}{\pi_1^*(1-\pi_1^*)} \right] \left[\frac{\partial \pi_1^*}{\partial p} \right]^2. \end{aligned}$$

Thus,

$$\lim_{g \rightarrow \infty} \text{var}(\hat{p}) = \frac{\pi_1^*(1 - \pi_1^*)}{g \left(\frac{\partial \pi_1^*}{\partial p} \right)^2},$$

where

$$\left(\frac{\partial \pi_1^*}{\partial p} \right)^2 = k^2 [(\eta + \theta - 1)^2 q^{2k-2}].$$

Method 2: Based on the Delta Method

Here, let $x = \hat{p}$ and $y = \pi_1^*$ as given in equation (4.3). Then,

$$(1-p)^k = \frac{\eta - \pi_1^*}{\eta + \theta - 1}$$

giving

$$\hat{p} = 1 - \left[\frac{\eta - \pi_1^*}{\eta + \theta - 1} \right]^{\frac{1}{k}}$$

Therefore,

$$\begin{aligned} \text{var}(\hat{p}) &\simeq \text{var}(\pi_1^*) \left[\frac{\partial \hat{p}}{\partial \pi_1^*} \right]^2 \\ &= \frac{g\pi_1^*(1-\pi_1^*)}{g^2} \left[\frac{\partial}{\partial \pi_1^*} \left\{ 1 - \left[\frac{\eta - \pi_1^*}{\eta + \theta - 1} \right]^{\frac{1}{k}} \right\} \right]^2 \\ &= \frac{\pi_1^*(1-\pi_1^*)}{g} \left[\frac{(\eta - \pi_1^*)^{\frac{2}{k}-2}}{k^2(\eta + \theta - 1)^{\frac{2}{k}}} \right] \\ &= \frac{\pi_1^*(1-\pi_1^*)}{gk^2(\eta + \theta - 1)^{\frac{2}{k}}} (\eta - \pi_1^*)^{\frac{2}{k}-2} \\ &= \frac{\pi_1^*(1-\pi_1^*)}{gk^2(\eta + \theta - 1)^{\frac{2}{k}}} [(\eta + \theta - 1)q^k]^{\frac{2}{k}-2} \\ &= \frac{\pi_1^*(1-\pi_1^*)}{gk^2} [(\eta + \theta - 1)^{-2} q^{2-2k}] \\ &= \frac{(1-p)^2}{f} \frac{(1-\pi_1^*)}{(\eta + \theta - 1)^2} \frac{(1-p)^{-2k} \pi_1^*}{k} \end{aligned}$$

where $f = gk$.

3.3 Relative Efficiency

Although group screening increases the probability of obtaining a positive estimate of the prevalence, such a screening strategy would not be attractive, if it yielded a less accurate estimate. This would seem to be the case

since fewer tests are required. But this is not true as shown in the following result.

Theorem 3.3.1

When screening with errors in decisions for a fixed sample size f and for small p , the asymptotic variance of the estimator \hat{p} of p is a monotonic decreasing function of k , where \hat{p} is as given in equation (3.5).

Proof

$$\text{var}(\hat{p}) = \frac{(1-p)^2}{f} \frac{(1-\pi_1^*)}{(\eta+\theta-1)^2} \frac{(1-p)^{-2k}\pi_1^*}{k}.$$

Since π_1^* increases in k , $1-\pi_1^*$ is a decreasing function of k . For small p ,

$$\begin{aligned} \frac{(1-p)^{-2k}\pi_1^*}{k} &= \frac{[\{(1-p)^{-2k} - (1-p)^{-k}\}\eta + (1-p)^{-k}(1-\theta)]}{k} \\ &\simeq \frac{[(1+2pk-1-pk)\eta + (1+pk)(1-\theta)]}{k} \\ &= p(1-\theta+\eta) + \frac{1-\theta}{k}, \end{aligned}$$

which is also a decreasing function in k , indicating that the variance decreases with an increase in group size k . Thus, group screening gives rise to a more accurate estimator of prevalence with fewer tests.

Efficiency can also be defined as a ratio of asymptotic variances $\text{var}(\hat{p}_{k>1})/\text{var}(\hat{p}_{k=1})$, for a fixed sample size and various values of η , θ and p .

3.4 The Test of Hypothesis Approach

Consider the indicator variables T_i and D_i defined in section(3.2) above. Further, let S_i be the number of defective factors in the i th group-factor.

We then determine the conditional probabilities $Prob(T_i/D_i)$ and $Prob(T_i/S_i)$.

To do so, we develop a test of hypothesis as follows:

$$H_0 : S_i = 0$$

against

$$H_1 : S_i = s.$$

Watson(1961) defined the power of this test as

$$\begin{aligned} \pi_1(s\phi_1, \alpha_1) &= \text{the probability of declaring a group with } s \text{ defective factors defective} \\ &= Prob(T_i = 1/S_i = s), \end{aligned}$$

where

$$\begin{aligned} \alpha_1 &= \text{the level of significance in the first stage} \\ &= \text{probability of declaring a non-defective group-factor defective} \\ &= Prob(T_i = 1/D_i = 0) \\ &= \pi_1(0, \alpha_1), \end{aligned}$$

and

$$\phi_1 = \frac{\Delta}{\sqrt{\frac{\sigma^2}{g+h}}}, \quad h = 1, 2, 3, 4,$$

with Δ as the effect of a group and σ^2 as the variance. Thus $\pi_1(s\phi_1, \alpha_1)$ is a function of k .

With these notations, Watson(1961) obtained the following result.

Theorem 3.4.1

When screening with errors in decisions, the probability of declaring a defective group-factor defective is given by

$$\pi'_1 = \frac{\sum_{s=1}^k \pi_1(s\phi_1, \alpha_1) \binom{k}{s} p^s q^{k-s}}{1 - q^k}.$$

where the parameters are as defined earlier.

Proof

Suppose the group test is positive and the group factor is actually defective, then we have,

$$\begin{aligned} \pi'_1 &= \text{Prob}(T_i = 1 / S_i \geq 1) \\ &= \frac{\text{Prob}(T_i = 1, S_i \geq 1)}{\text{Prob}(S_i \geq 1)} \\ &= \frac{\sum_{s=1}^k \text{Prob}(T_i = 1, S_i = s)}{\sum_{s=1}^k \text{Prob}(S_i = s)} \\ &= \frac{\sum_{s=1}^k \text{Prob}(T_i = 1 / S_i = s) \text{Prob}(S_i = s)}{\sum_{s=1}^k \text{Prob}(S_i = s)} \\ &= \frac{\sum_{s=1}^k \pi_1(s\phi_1, \alpha_1) \binom{k}{s} p^s q^{k-s}}{\sum_{s=1}^k \binom{k}{s} p^s q^{k-s}} \\ &= \frac{\sum_{s=1}^k \pi_1(s\phi_1, \alpha_1) \binom{k}{s} p^s q^{k-s}}{1 - q^k}. \end{aligned} \tag{3.17}$$

This completes the proof.

Based on this theorem, we obtain the following result.

Theorem 3.4.2

When screening with errors in decisions, the probability that a group-factor is declared defective is given by

$$\pi_1^* = \pi_1' - (\pi_1' - \alpha_1)q^k.$$

where the π_1^* and α_1 are as in theorem (3.4.1) and the other parameters as earlier defined.

Proof

$$\begin{aligned} \pi_1^* &= \text{Prob}(T_i = 1) \\ &= \text{Prob}(T_i = 1, D_i = 1) + \text{Prob}(T_i = 1, D_i = 0) \\ &= \text{Prob}(T_i = 1/D_i = 1)\text{Prob}(D_i = 1) + \text{Prob}(T_i = 1, D_i = 0)\text{Prob}(D_i = 0) \\ &= \pi_1'(1 - q^k) + q^k \\ &= \pi_1' - (\pi_1' - \alpha_1)q^k. \end{aligned}$$

This completes the proof.

The maximum likelihood estimator of π_1^* is given in equation (3.4), as

$\hat{\pi}_1^* = \frac{r}{g}$ with $\text{var}(\hat{\pi}_1^*) = \frac{\pi_1^*(1-\pi_1^*)}{g}$. But

$$\pi_1^* = \pi_1' - (\pi_1' - \alpha_1)q^k.$$

Therefore,

$$(\pi_1' - \alpha_1)q^k = \pi_1' - \pi_1^*,$$

giving

$$q^k = \frac{\pi'_1 - \pi_1^*}{\pi'_1 - \alpha_1}.$$

Thus,

$$(1 - p)^k = \frac{\pi'_1 - \pi_1^*}{\pi'_1 - \alpha_1},$$

implying that

$$p = 1 - \left[\frac{\pi'_1 - \pi_1^*}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \equiv f(\pi_1^*), \text{ say.}$$

Therefore,

$$\hat{p} = 1 - \left[\frac{\pi'_1 - \frac{r}{g}}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \equiv \phi(r), \text{ say.}$$

Using the delta method, if $p = f(\pi_1^*)$, then

$$\begin{aligned} \text{var}(\hat{p}) &= \left[\frac{df}{d\pi_1^*} \right]^2 \text{var}(\hat{\pi}_1^*) \\ &= \left[\frac{d}{d\pi_1^*} \{ \pi'_1 - (\pi'_1 - \alpha_1)(1 - p)^k \} \right]^2 \left[\frac{\pi_1^*(1 - \pi_1^*)}{g} \right] \\ &= \left[\frac{d}{d\pi_1^*} \left\{ 1 - \left[\frac{\pi'_1 - \pi_1^*}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \right\} \right]^2 \left[\frac{\pi_1^*(1 - \pi_1^*)}{g} \right] \\ &= \left[\frac{1}{k} \left\{ \frac{\pi'_1 - \pi_1^*}{\pi'_1 - \alpha_1} \right\}^{\frac{1}{k}-1} \right]^2 \left[\frac{\pi_1^*(1 - \pi_1^*)}{g} \right] \\ &= \frac{1}{k^2} \left\{ \frac{\pi'_1 - \pi_1^*}{\pi'_1 - \alpha_1} \right\}^{\frac{2}{k}-2} \left[\frac{\pi_1^*(1 - \pi_1^*)}{g} \right] \\ &= \frac{1}{gk^2} \pi_1^*(1 - \pi_1^*) \left\{ \frac{\pi'_1 - \pi_1^*}{\pi'_1 - \alpha_1} \right\}^{\frac{2}{k}-2}. \end{aligned}$$

Using Jensen's inequality, if $\phi(r)$ is convex, then

$$E[\phi(r)] \geq \phi[E(r)].$$

Condition for $\hat{p} = \phi(r)$ to be convex is derived as follows:

$$\begin{aligned} \frac{\partial \hat{p}}{\partial r} &= \frac{\partial}{\partial r} \left\{ 1 - \left[\frac{\pi'_1 - \frac{r}{g}}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \right\} \\ &= \frac{1}{gk} \left[\frac{\pi'_1 - \frac{r}{g}}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}-1}, \end{aligned}$$

implying that

$$\frac{\partial^2 \hat{p}}{\partial r^2} = \frac{k-1}{(gk)^2} \left[\frac{\pi'_1 - \frac{r}{g}}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}-2}.$$

Note that

$$\frac{\partial^2 \hat{p}}{\partial r^2} \geq 0 \text{ if } k \geq 1 \text{ and } \pi'_1 \geq \frac{r}{g}.$$

Under these conditions, then $\phi(r)$ is convex. Hence,

$$E[\phi(r)] \geq \phi[E(r)].$$

That is,

$$\begin{aligned} E \left\{ 1 - \left[\frac{\pi'_1 - \frac{r}{g}}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \right\} &\geq 1 - \left[\frac{\pi'_1 - E\left(\frac{r}{g}\right)}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \\ &= 1 - \left[\frac{\pi'_1 - \frac{g\pi'_1}{g}}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \\ &= 1 - \left[\frac{\pi'_1 - \pi'_1}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \\ &= 1 - \left[\frac{\pi'_1 - [\pi'_1 - (\pi'_1 - \alpha_1)q^k]}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \end{aligned}$$

$$\begin{aligned}
&= 1 - \left[\frac{(\pi'_1 - \alpha_1)q^k}{\pi'_1 - \alpha_1} \right]^{\frac{1}{k}} \\
&= 1 - (q^k)^{\frac{1}{k}} \\
&= 1 - q \\
&= p.
\end{aligned}$$

Thus, $E(\hat{p}) \geq p$ showing that \hat{p} overestimates p .

3.4.1 Special Case

The result obtained by Thompson(1962) can be considered as a special case, where $\pi'_1 = 1$ and $\alpha_1 = 0$. For instance,

$$\pi_1^* = \pi'_1 - (\pi'_1 - \alpha_1)q^k = 1 - q^k = \hat{p}^*.$$

Then $\pi'_1 = \hat{p}^* = \frac{r}{g}$,

$$\text{var}(\pi_1^*) = \text{var}(\hat{p}^*) = \frac{p^*(1-p^*)}{g} = \frac{q^k(1-q^k)}{g}.$$

Thus,

$$\begin{aligned}
\text{var}(\hat{p}) &= \frac{p^*(1-p^*)^{\frac{1}{k}-1}}{gk^2} \\
&= \frac{(q^k)^{\frac{1}{k}-1}(1-q^k)}{gk^2} \\
&= \frac{1-q^k}{gk^2q^{k-2}}
\end{aligned}$$

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

3.4.2 Alternative Approach

Let

$$\begin{aligned}\beta_1 &= \text{probability of declaring a defective group-factor non-defective} \\ &= \text{Prob}(T_i = 0/D_i = 1).\end{aligned}$$

Therefore,

$$\begin{aligned}1 - \alpha_1 &= \text{probability of declaring a non-defective group-factor non-defective} \\ &= \text{Prob}(T_i = 0/D_i = 0) \\ &= \text{specificity of the test } \theta,\end{aligned}$$

and

$$\begin{aligned}1 - \beta_1 &= \text{probability of declaring a defective group-factor defective} \\ &= \text{Prob}(T_i = 1/D_i = 1) \\ &= \text{sensitivity of the test } \eta.\end{aligned}$$

Thus, π_1^* can be defined in terms of α_1 and β_1 as follows.

Theorem 3.4.3

When screening with errors in decisions and using the test of hypothesis approach, the probability that a group-factor is declared defective is given by

$$\pi_1^* = (1 - \beta_1) - (1 - \alpha_1 - \beta_1)q^k.$$

where β_1 is the probability of declaring a defective group factor non-defective and the other parameters are as defined earlier.

Proof

$$\begin{aligned}\pi_1^* &= Prob(T_i = 1) \\ &= Prob(T_i = 1, D_i = 1) + Prob(T_i = 1, D_i = 0) \\ &= Prob(T_i = 1/D_i = 1)Prob(D_i = 1) + Prob(T_i = 1, D_i = 0)Prob(D_i = 0) \\ &= \pi_1'(1 - q^k) + \alpha_1 q^k \\ &= (1 - \beta_1)(1 - q^k) + \alpha_1 q^k \\ &= 1 - \beta_1 - q^k + \alpha_1 q^k + \beta_1 q^k \\ &= (\alpha_1 + \beta_1 - 1)q^k + (1 - \beta_1) \\ &= (1 - \beta_1) - (1 - \alpha_1 - \beta_1)q^k.\end{aligned}\tag{3.18}$$

This completes the proof.

Thus, equations (3.3) and (3.18) give similar results for the probability of declaring a group-factor defective.

The maximum likelihood estimator of p using π_1^* in equation (3.18) is therefore,

$$\begin{aligned}\hat{p} &= 1 - \left[\frac{1 - \beta_1 - \frac{r}{g}}{1 - \beta_1 - \alpha_1} \right]^{\frac{1}{k}} \\ &= 1 - \left[\frac{1 - \beta_1 - \pi_1^*}{1 - \beta_1 - \alpha_1} \right]^{\frac{1}{k}}.\end{aligned}\tag{3.19}$$

Since $0 \leq \hat{p} \leq 1$, it follows that

$$0 \leq 1 - \left[\frac{1 - \beta_1 - \pi_1^*}{1 - \beta_1 - \alpha_1} \right]^{\frac{1}{k}} \leq 1.$$

Thus,

$$\frac{1 - \beta_1 - \pi_1^*}{1 - \beta_1 - \alpha_1} \leq 1,$$

which implies that $1 - \beta_1 - \pi_1^* \leq 1 - \beta_1 - \alpha_1$ and hence

$$\alpha_1 \leq \pi_1^*.\tag{3.20}$$

Similarly,

$$\frac{1 - \beta_1 - \pi_1^*}{1 - \beta_1 - \alpha_1} \geq 0,$$

which implies that

$$\pi_1^* \geq 1 - \beta_1.\tag{3.21}$$

Expressions (3.20) and (3.21) can thus be used to determine the probability of prevalence p as is the case with expressions (3.7) and (3.8), for known values of α_1 and β_1 and since π_1^* is a function of r which has a binomial distribution.

Chapter 4

Bayesian Estimation With and Without Errors in Decision

4.1 Introduction

In this chapter, we consider the Bayesian method of estimation applied to group-screening design with equal prior probability, equal group sizes with and without errors in decision. The assumption here is that there is prior knowledge of the infection rate which is random with known probability distribution. The rest of the chapter is organized as follows. The Bayesian estimator without errors is considered in section 4.2. Section 4.3 gives an alternative approach to estimating the parameters of the beta distribution, α and β . The case where $\alpha = 1$ is discussed as a special case in section 4.4. The prior on p^* , the probability that a group-factor is declared defective is discussed in section 4.5. Section 4.6 gives the comparison between the MLE and the Bayesian estimator based on biasedness, mean square error and confidence interval. Estimation with errors in decisions is

considered in section 4.7.

4.2 Estimation Without Errors in Decisions

Since the probability of a factor being defective is small, we use a family of prior distribution appropriate for rare traits. Thus, we use the beta distribution since it is a conjugate of the binomial distribution.

If r is the number of defective groups out of g groups formed, then

$$f(r/p) = \begin{cases} \binom{g}{r} [1 - (1-p)^k]^r (1-p)^{k(g-r)}, & r = 0, 1, 2, \dots, g \\ 0, & \text{otherwise.} \end{cases} \quad (4.1)$$

Thus, the joint distribution of r and p is given by

$$\begin{aligned} f(r, p) &= f(r/p) \cdot f(p) \\ &= \binom{g}{r} [1 - (1-p)^k]^r (1-p)^{k(g-r)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \end{aligned} \quad (4.2)$$

Let

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}. \quad (4.3)$$

Then

$$f(r, p) = \binom{g}{r} \{B(\alpha, \beta)\}^{-1} p^{\alpha-1} (1-p)^{kg-kr+\beta-1} [1 - (1-p)^k]^r. \quad (4.4)$$

With this joint distribution, we determine the marginal probability density function as given in the following theorem .

Theorem 4.2.1

The marginal probability density $f(r)$ is given by

$$f(r) = \left(\frac{\beta}{\alpha + \beta}\right)^{k(g-r)} \binom{g}{r} \sum_{j=0}^r \binom{r}{j} (-1)^j \left(\frac{\beta}{\alpha + \beta}\right)^{kj}.$$

Proof

$$\begin{aligned} f(r) &= \int_0^1 \binom{g}{r} \{B(\alpha, \beta)\}^{-1} p^{\alpha-1} (1-p)^{kg-kr+\beta-1} [1 - (1-p)^k]^r dp \\ &= \int_0^1 \binom{g}{r} r \{B(\alpha, \beta)\}^{-1} p^{\alpha-1} (1-p)^{kg-kr+\beta-1} \left(\sum_{j=0}^r \binom{r}{j} (-1)^j (1-p)^{kj}\right) dp \\ &= \binom{g}{r} \{B(\alpha, \beta)\}^{-1} \int_0^1 p^{\alpha-1} (1-p)^{kj+kg-kr+\beta-1} \sum_{j=0}^r \binom{r}{j} (-1)^j dp \\ &= \binom{g}{r} \{B(\alpha, \beta)\}^{-1} \left(\sum_{j=0}^r \binom{r}{j} (-1)^j B(\alpha, kj + kg - kr + \beta)\right) \\ &= \binom{g}{r} \sum_{j=0}^r \binom{r}{j} (-1)^j \frac{B(\alpha, kj + kg - kr + \beta)}{B(\alpha, \beta)} \\ &\simeq \binom{g}{r} \sum_{j=0}^r \binom{r}{j} (-1)^j \left(\frac{\beta}{\alpha + \beta}\right)^{kj+kg-kr} \\ &= \left(\frac{\beta}{\alpha + \beta}\right)^{k(g-r)} \binom{g}{r} \sum_{j=0}^r \binom{r}{j} (-1)^j \left(\frac{\beta}{\alpha + \beta}\right)^{kj}. \end{aligned} \tag{4.5}$$

The approximation in the above equation can be given the following justification. For large N , Abramowitz and Stegun (1960) showed that

$$\frac{\Gamma(N+a)}{\Gamma(N+b)} \simeq N^{a-b}. \tag{4.6}$$

To determine the estimates of the hyperparameters α and β , we differentiate expression (4.5) partially with respect to α and β . Suppose the estimates are $\hat{\alpha}$ and $\hat{\beta}$, then the posterior distribution of p is given as

$$\begin{aligned} f(p/r) &= f(r, p)/f(r) \\ &= \frac{p^{\hat{\alpha}-1}(1-p)^{kg-kr+\hat{\beta}-1}[1-(1-p)^k]^r}{\sum_{j=0}^r \binom{r}{j} (-1)^j B(\hat{\alpha}, kj + kg - kr + \hat{\beta})}, \end{aligned} \quad (4.7)$$

leading to the following result.

Theorem 4.2.2

With expression (4.7), the Bayes estimator of p , based on squared error loss is given by

$$\hat{p} = \frac{\sum_{j=0}^r \binom{r}{j} (-1)^j \frac{\hat{\alpha} \hat{\beta}^{k(j+g-r)}}{(\hat{\alpha} + \hat{\beta})^{k(j+g-r)}}}{\sum_{j=0}^r \binom{r}{j} (-1)^j \left(\frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}}\right)^{k(j+g-r)+1}}.$$

Proof

$$\begin{aligned} \hat{p} &= \frac{\int_0^1 p \cdot p^{\hat{\alpha}-1} (1-p)^{kg-kr+\hat{\beta}-1} [1-(1-p)^k]^r dp}{\sum_{j=0}^r \binom{r}{j} (-1)^j B(\hat{\alpha}, kj + kg - kr + \hat{\beta})} \\ &= \frac{\int_0^1 p^{\hat{\alpha}} (1-p)^{k(g-r)+\hat{\beta}-1} [1-(1-p)^k]^r dp}{\sum_{j=0}^r \binom{r}{j} (-1)^j B(\hat{\alpha}, k(j+g-r) + \hat{\beta})} \\ &= \frac{\sum_{j=0}^r \binom{r}{j} (-1)^j B(\hat{\alpha} + 1, k(j+g-r) + \hat{\beta})}{\sum_{j=0}^r \binom{r}{j} (-1)^j B(\hat{\alpha}, k(j+g-r) + \hat{\beta})} \\ &= \frac{\sum_{j=0}^r \binom{r}{j} (-1)^j \frac{B(\hat{\alpha}+1, k(j+g-r)+\hat{\beta})}{B(\hat{\alpha}, \hat{\beta})}}{\sum_{j=0}^r \binom{r}{j} (-1)^j \frac{B(\hat{\alpha}, k(j+g-r)+\hat{\beta})}{B(\hat{\alpha}, \hat{\beta})}} \end{aligned}$$

$$= \frac{\sum_{j=0}^r \binom{r}{j} (-1)^j \frac{\hat{\alpha} \beta^{k(j+g-r)}}{(\hat{\alpha} + \beta)^{k(j+g-r)}}}{\sum_{j=0}^r \binom{r}{j} (-1)^j \left(\frac{\hat{\alpha}}{\hat{\alpha} + \beta}\right)^{k(j+g-r)+1}}. \quad (4.8)$$

Hence, the proof.

4.3 Estimation of α and β

From the marginal probability density function of r above,

$$f(r) = \int_0^1 \binom{g}{r} \{B(\alpha, \beta)\}^{-1} p^{\alpha-1} (1-p)^{k(g-r)+\beta-1} [1 - (1-p)^k]^r dp. \quad (4.9)$$

For small p , we propose the following theorem.

Theorem 4.3.1

The marginal density function of r is given by

$$f(r) = k^r \binom{g}{r} \frac{\alpha^r \beta^{k(g-r)}}{(\alpha + \beta)^{r+k(g-r)}}.$$

Proof

For small p , $(1-p)^k \simeq 1 - kp$. Therefore, we have

$$\begin{aligned} f(r) &\simeq \int_0^1 \binom{g}{r} \{B(\alpha, \beta)\}^{-1} p^{\alpha-1} (1-p)^{k(g-r)+\beta-1} (kp)^r dp \\ &= k^r \binom{g}{r} \{B(\alpha, \beta)\}^{-1} \int_0^1 p^{\alpha+r-1} (1-p)^{k(g-r)+\beta-1} dp \\ &= k^r \binom{g}{r} \{B(\alpha, \beta)\}^{-1} B(\alpha + r, k(g-r) + \beta) \\ &= k^r \binom{g}{r} \frac{\Gamma(\alpha + \beta) \Gamma(\alpha + r) \Gamma(k(g-r) + \beta)}{\Gamma \alpha \Gamma \beta \Gamma(\alpha + r + k(g-r) + \beta)} \end{aligned}$$

$$= k^r \binom{g}{r} \frac{\alpha^r \beta^{k(g-r)}}{(\alpha + \beta)^{r+k(g-r)}}, \quad (4.10)$$

a more simplified expression of the marginal density function of r .

The hyperparameters α and β can be estimated by maximizing the likelihood function leading to a digamma function, or for fixed α , we find $\hat{\beta}$ by differentiating $f(r)$ partially with respect to β and equating to zero. Thus,

$$\frac{\partial f(r)}{\partial \beta} = \frac{k^r \binom{g}{r} [(\alpha + \beta)^{r+k(g-r)} r \alpha^r k(g-r) \beta^{k(g-r)-1} - \alpha^r \beta^{k(g-r)} (r+k(g-r)) (\alpha + \beta)^{r+k(g-r)-1}]}{(\alpha + \beta)^{2(r+k(g-r))}} = 0.$$

Solving this equation, we get

$$\hat{\beta} = \frac{\alpha k(g-r)}{r}. \quad (4.11)$$

This shows that $\hat{\beta}$ can be expressed as a function of number of positive pools and the pool size.

4.4 Special case: A prior on p with $\alpha = 1$

Here, the prior probability distribution of p is given by,

$$f(p) = \beta(1-p)^{\beta-1} : \quad 0 < p < 1 \quad \text{and} \quad \beta \geq 1.$$

The joint probability distribution function of r and p , conditional on β is thus,

$$\begin{aligned} f(r, p/\beta) &= \binom{g}{r} [1 - (1-p)^k]^r (1-p)^{k(g-r)} \beta (1-p)^{\beta-1} \\ &= \beta \binom{g}{r} [1 - (1-p)^k]^r (1-p)^{k(g-r)+\beta-1}, \end{aligned} \quad (4.12)$$

for $r = 0, 1, \dots, g$ and $0 < p < 1$. The marginal distribution function of r is therefore,

$$f(r/\beta) = \beta \binom{g}{r} \int_0^1 [1 - (1-p)^k]^r (1-p)^{k(g-r)+\beta-1} dp. \quad (4.13)$$

This integral is finite, when $\beta \geq 1$. Using the change of variable technique with $u = (1-p)^k$, it follows that $p = 1 - u^{\frac{1}{k}}$ and that $dp = -(1-p)^{1-k} \frac{du}{k}$. Thus, equation (4.13) becomes

$$\begin{aligned} f(r/\beta) &= \beta k^{-1} \binom{g}{r} \int_0^1 u^{g-r+\frac{\beta}{k}-1} (1-u)^r du \\ &= \frac{\beta \Gamma(g+1) \Gamma(g-r+\frac{\beta}{k})}{k \Gamma(g-r+1) \Gamma(g+\frac{\beta}{k})} \\ &\approx \frac{\beta (g+1)^r}{k (g+\frac{\beta}{k})^{r+1}}, \end{aligned} \quad (4.14)$$

for $r = 0, 1, \dots, g$. Differentiating equation (4.14) with respect to β and equating to zero gives,

$$\frac{1}{k} \left(\frac{(g+\frac{\beta}{k})^{r+1} (g+1)^r - \beta (g+1)^r \frac{(r+1)}{k} (g+\frac{\beta}{k})^r}{\{(g+\frac{\beta}{k})^{r+1}\}^2} \right) = 0,$$

which is simplified as,

$$\hat{\beta} = \frac{gk}{r}. \quad (4.15)$$

The posterior distribution is therefore, given by

$$\begin{aligned} f(p/r) &= \frac{k \Gamma(g+\frac{\hat{\beta}}{k}+1)}{\Gamma(g-r+\frac{\hat{\beta}}{k}) \Gamma(r+1)} (1-p)^{k(g-r)+\hat{\beta}-1} [1 - (1-p)^k]^r \\ &= \frac{k \Gamma(g+\frac{g}{k}+1)}{\Gamma(g-r+\frac{g}{k}) \Gamma(r+1)} (1-p)^{k(g-r)+\frac{gk}{r}-1} [1 - (1-p)^k]^r \end{aligned} \quad (4.16)$$

for $0 < p < 1$. Using the squared error loss function $L(p, a) = (p - a)^2$, the Bayes estimate of p is the mean of the posterior $f(p/r)$ as given in the following theorem by Tebbs and Bilder (2003). The alternative proof of this theorem is given as follows.

Theorem 4.4.1

The mean of the posterior distribution $f(p/r)$ is given by,

$$\hat{p} = 1 - \left(\frac{g - r}{g + 1} \right)^{\frac{1}{k}}.$$

Proof

Using change of variable $u = (1 - p)^k$ again, we have

$$\begin{aligned} \hat{p} &= \int_0^1 p f(p/r, \hat{\beta}) dp \\ &= \int_0^1 \frac{k\Gamma(g + \frac{\hat{\beta}}{k} + 1)}{\Gamma(g - r + \frac{\hat{\beta}}{k})\Gamma(r + 1)} p(1 - p)^{k(g-r)+\hat{\beta}-1} [1 - (1 - p)^k]^r \\ &= \frac{k\Gamma(g + \frac{\hat{\beta}}{k} + 1)}{\Gamma(g - r + \frac{\hat{\beta}}{k})\Gamma(r + 1)} \int_0^1 \frac{1}{k} (1 - u^{\frac{1}{k}}) u^{g-r+\frac{\hat{\beta}}{k}-1} (1 - u)^r du \\ &= \frac{\Gamma(g + \frac{\hat{\beta}}{k} + 1)}{\Gamma(g - r + \frac{\hat{\beta}}{k})\Gamma(r + 1)} \left\{ \int_0^1 u^{g-r+\frac{\hat{\beta}}{k}-1} (1 - u)^r du - \int_0^1 u^{g-r+\frac{\hat{\beta}}{k}+\frac{1}{k}-1} (1 - u)^r du \right\} \\ &= \frac{\Gamma(g + \frac{\hat{\beta}}{k} + 1)}{\Gamma(g - r + \frac{\hat{\beta}}{k})\Gamma(r + 1)} \left\{ \frac{\Gamma(g + \frac{\hat{\beta}}{k} - r)\Gamma(r + 1)}{\Gamma(g + 1 + \frac{\hat{\beta}}{k})} - \frac{\Gamma(g + \frac{\hat{\beta}}{k} + \frac{1}{k} - r)\Gamma(r + 1)}{\Gamma(g + \frac{\hat{\beta}}{k} + \frac{1}{k} + 1)} \right\} \\ &= 1 - \frac{\Gamma(g + \frac{\hat{\beta}}{k} + 1)\Gamma(g - r + \frac{\hat{\beta}}{k} + \frac{1}{k})}{\Gamma(g - r + \frac{\hat{\beta}}{k})\Gamma(g + \frac{\hat{\beta}}{k} + \frac{1}{k} + 1)}, \end{aligned} \tag{4.17}$$

which can be simplified using equation (4.6), to the frequentist estimate

$$\hat{p} = 1 - \left(\frac{g-r}{g+1} \right)^{\frac{1}{k}}. \quad (4.18)$$

4.5 Prior on p^*

In this section, we consider an alternative approach to determining the Bayes estimate based on the relationship between p and p^* , the probability that a group-factor is declared defective. We call the resulting estimate, the *indirect Bayes estimator*, \hat{p}_{Ibys} . The probability of r conditioned on p^* is given by,

$$f(r/p^*) = \binom{g}{r} p^{*r} (1-p^*)^{g-r}.$$

The prior probability of p^* is given by,

$$\begin{aligned} f(p^*) &= \frac{\Gamma(\alpha + \beta)}{\Gamma\alpha\Gamma\beta} p^{*\alpha-1} (1-p^*)^{\beta-1} \\ &= \{B(\alpha, \beta)\}^{-1} p^{*\alpha-1} (1-p^*)^{\beta-1}. \end{aligned} \quad (4.19)$$

The joint probability distribution function of r and p^* is given by,

$$\begin{aligned} f(r, p^*) &= \binom{g}{r} p^{*r} (1-p^*)^{g-r} \{B(\alpha, \beta)\}^{-1} p^{*\alpha-1} \\ &= \{B(\alpha, \beta)\}^{-1} \binom{g}{r} p^{*r+\alpha-1} (1-p^*)^{g-r+\beta-1}. \end{aligned} \quad (4.20)$$

The marginal distribution of r is therefore,

$$f(r) = \int_0^1 \{B(\alpha, \beta)\}^{-1} \binom{g}{r} p^{*r+\alpha-1} (1-p^*)^{g-r+\beta-1} dp^*$$

$$\begin{aligned}
&= \{B(\alpha, \beta)\}^{-1} \binom{g}{r} B(r + \alpha, g - r + \beta) \\
&= \binom{g}{r} \frac{B(r + \alpha, g - r + \beta)}{B(\alpha, \beta)} \\
&\simeq \binom{g}{r} \left(\frac{\alpha}{\beta}\right)^r \left(\frac{\beta}{\alpha + \beta}\right)^g.
\end{aligned} \tag{4.21}$$

Differentiating with respect to β and equating to zero, we get

$$\hat{\beta} = \frac{\hat{\alpha}(g - r)}{r}. \tag{4.22}$$

The posterior distribution of p^* is given by,

$$\begin{aligned}
f(p^*/r) &= \frac{\binom{g}{r} p^{*r} (1 - p^*)^{g-r} \{B(\alpha, \beta)\}^{-1} p^{*\alpha-1}}{\binom{g}{r} \frac{B(r+\alpha, g-r+\beta)}{B(\alpha, \beta)}} \\
&= \frac{p^{*\alpha+r-1} (1 - p^*)^{g-r+\beta-1}}{B(r + \alpha, g - r + \beta)}.
\end{aligned} \tag{4.23}$$

The posterior mean of p^* based on squared error loss is given by,

$$\begin{aligned}
p^* &= \frac{\int_0^1 p^{*\alpha+r} (1 - p^*)^{g-r+\beta-1} dp^*}{B(r + \alpha, g - r + \beta)} \\
&= \frac{B(r + \alpha + 1, g - r + \beta)}{B(r + \alpha, g - r + \beta)} \\
&= \frac{r + \alpha}{g + \alpha + \beta}.
\end{aligned} \tag{4.24}$$

This yields the posterior of p through the transformation $p^* = 1 - (1 - p)^k$, and we compute the Bayes estimator under the square error loss as,

$$p_{Ibys} = 1 - \left(1 - \frac{r + \alpha}{g + \alpha + \beta}\right)^{\frac{1}{k}} \tag{4.25}$$

However, we can directly compute this, without computing the posterior of p as follows.

Theorem 4.5.1

The posterior mean of p is given by,

$$\bar{p}_{Ibys} = 1 - \left(1 - \frac{r + \alpha}{g + \alpha + \beta}\right)^{\frac{1}{k}}.$$

Proof

$$\begin{aligned} \bar{p}_{Ibys} &= \int_0^1 [1 - (1 - p)^{\frac{1}{k}}] f(p^*/r) dp^* \\ &= 1 - \int_0^1 \frac{p^{*\alpha+r-1} (1 - p^*)^{g-r+\beta+\frac{1}{k}-1}}{B(r + \alpha, g - r + \beta)} dp^*. \end{aligned} \quad (4.26)$$

Using the substitution $u = 1 - p^*$, we have

$$\begin{aligned} \bar{p}_{Ibys} &= 1 - \int_0^1 u^{g-r+\beta+\frac{1}{k}-1} (1 - u)^{r+\alpha-1} du \\ &= 1 - \frac{B(g - r + \beta + \frac{1}{k}, r + \alpha)}{B(r + \alpha, g - r + \beta)} \\ &= 1 - \frac{\Gamma(g - r + \beta + \frac{1}{k}) \Gamma(g + \alpha + \beta)}{\Gamma(g - r + \beta) \Gamma(g + \alpha + \beta + \frac{1}{k})} \\ &\approx 1 - \left(\frac{g - r + \beta}{g + \alpha + \beta}\right)^{\frac{1}{k}} \quad \text{by eqn(4.6)} \\ &= 1 - \left(1 - \frac{r + \alpha}{g + \alpha + \beta}\right)^{\frac{1}{k}}, \end{aligned} \quad (4.27)$$

which is easier to compute and therefore may be preferred in practise.

4.6 Derivation of Moments

Given from equation (4.21) that

$$f(r) = \binom{g}{r} \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + r)\Gamma(g - r + \beta)}{\Gamma\alpha\Gamma\beta\Gamma(g + \alpha + \beta)},$$

then

$$f(r-1) = \binom{g}{r-1} \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + (r-1))\Gamma(g - (r-1) + \beta)}{\Gamma\alpha\Gamma\beta\Gamma(g + \alpha + \beta)}.$$

Dividing the two equations, we have

$$\frac{f(r)}{f(r-1)} = \frac{(g-r+1)(r-1+\alpha)}{r(g-r+\beta)}, \quad (4.28)$$

which implies that

$$r(g-r+\beta)f(r) = (g-r+1)(r-1+\alpha)f(r-1). \quad (4.29)$$

Summing equation (4.30) over r gives,

$$\sum_{r=1}^g r(g-r+\beta)f(r) = \sum_{r=1}^g (g-r+1)(r-1+\alpha)f(r-1)$$

That is,

$$(g+\beta) \sum_{r=1}^g rf(r) = g\alpha \sum_{r=1}^g f(r-1) + (g-\alpha) \sum_{r=1}^g (r-1)f(r-1) - \sum_{r=1}^g (r-1)^2 f(r-1).$$

Thus,

$$(g+\beta)M_1 = g\alpha + (g-\alpha)(M_1 - gf(g)) - (M_2 - g^2f(g)),$$

which simplifies to

$$(\alpha + \beta)M_1 = g\alpha$$

Therefore, the first moment

$$E(r) = M_1 = \frac{g\alpha}{(\alpha + \beta)} \quad (4.30)$$

Multiplying equation (4.30) by r and summing over r we have,

$$\sum_{r=1}^g r^2(g - r + \beta)f(r) = \sum_{r=1}^g [(r - 1) + 1][(g - r + 1)][(r - 1 + \alpha)f(r - 1)],$$

which implies that

$$\begin{aligned} (g + \beta) \sum_{r=1}^g r^2 f(r) - \sum_{r=1}^g r^3 f(r) &= g\alpha \left[\sum_{r=1}^g (r - 1) f(r - 1) + \sum_{r=1}^g f(r - 1) \right] \\ &+ (g - \alpha) \left[\sum_{r=1}^g (r - 1)^2 f(r - 1) + \sum_{r=1}^g (r - 1) f(r - 1) \right] \\ &- \sum_{r=1}^g (r - 1)^3 f(r - 1) - \sum_{r=1}^g (r - 1)^2 f(r - 1). \end{aligned}$$

This can then be expressed as,

$$\begin{aligned} (g + \beta)M_2 - M_3 &= g\alpha[M_1 - gf(g) + 1 - f(g)] + (g - \alpha)[M_2 - g^2 + M_1 - gf(g)] \\ &- [M_3 - g^3 f(g)] - [M_2 - g^2 f(g)], \end{aligned}$$

which implies that

$$(\alpha + \beta + 1)M_2 = [g\alpha + g - \alpha]M_1 + g\alpha.$$

Substituting for M_1 from equation (4.31), we have the second moment as

$$M_2 = \frac{g\alpha(g\alpha + g) + g\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)}. \quad (4.31)$$

The moment estimators of α and β are thus given as,

$$\begin{aligned} \hat{\alpha} &= \frac{M_{(1)}M_{(2)} - (g - 1)M_{(1)}^2}{(g - 1)M_{(1)}^2 - gM_{(2)}} \\ \hat{\beta} &= \hat{\alpha} \left(\frac{g - M_{(1)}}{M_{(1)}} \right), \end{aligned} \quad (4.32)$$

same as those derived by Skellan(1948).

4.7 Comparison of Estimators

Here, the point estimate and interval estimate are compared and their characteristics are discussed.

4.7.1 Point Estimate Characteristics

In this section, we compare the Maximum Likelihood Estimate and the Bayesian Estimate through the measures of bias and mean squared error. In a Bayesian framework, the choice of the loss function determines the specific form of the estimator. However, since we are comparing the two estimators on frequentist terms, after the specific form of the estimator is identified, the loss function is no longer used. For p fixed, the bias and mean squared error of \hat{p}_{bys} , respectively are given by,

$$\begin{aligned} Bias(\hat{p}_{bys}) &= E(\hat{p}_{bys} - p) \\ &= \sum_{r=0}^g (\hat{p}_{bys} - p) \binom{g}{r} [1 - (1-p)^k]^r (1-p)^{k(g-r)} \\ &= (1-p) - \sum_{i=0}^g \frac{\Gamma(g + \frac{\hat{\beta}}{k} + 1) \Gamma(i + \frac{\hat{\beta}}{k} + \frac{1}{k})}{\Gamma(i + \frac{\hat{\beta}}{k} + \frac{1}{k} + 1) \Gamma(g + \frac{\hat{\beta}}{k} + \frac{1}{k} + 1)} \\ &\quad \binom{g}{i} [1 - (1-p)^k]^{g-i} (1-p)^{ki} \end{aligned} \quad (4.33)$$

and

$$MSE(\hat{p}_{bys}) = E[(\hat{p}_{bys} - p)^2]$$

$$\begin{aligned}
&= \sum_{r=0}^g [(\hat{p}_{bys} - p)^2] \binom{g}{r} [1 - (1-p)^k]^r (1-p)^{k(g-r)} \\
&= (1-p)^2 + \sum_{i=0}^g \frac{\Gamma(g + \frac{\hat{\beta}}{k} + 1) \Gamma(i + \frac{\hat{\beta}}{k} + \frac{1}{k})}{\Gamma(i + \frac{\hat{\beta}}{k} + \frac{1}{k} + 1) \Gamma(g + \frac{\hat{\beta}}{k} + \frac{1}{k} + 1)} \\
&\quad \left[\frac{\Gamma(g + \frac{\hat{\beta}}{k} + 1) \Gamma(i + \frac{\hat{\beta}}{k} + \frac{1}{k})}{\Gamma(i + \frac{\hat{\beta}}{k} + \frac{1}{k} + 1) \Gamma(g + \frac{\hat{\beta}}{k} + \frac{1}{k} + 1)} - 2(1-p) \right] \\
&\quad \binom{g}{i} [1 - (1-p)^k]^{g-i} (1-p)^{ki}. \tag{4.34}
\end{aligned}$$

In particular, for the indirect Bayes estimator \hat{p}_{Ibys} the bias is given by,

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

and the mean squared error is given by

$$MSE(\hat{p}_{Ibys}) = E[(\hat{p}_{Ibys} - p)^2],$$

where

$$\hat{p}_{Ibys} = 1 - \left(1 - \frac{r + \alpha}{g + \alpha + \beta}\right)^{\frac{1}{k}}.$$

For comparing the Bayes estimator corresponding to a $Beta(\alpha, \beta)$ prior on p , the prior on p^* is chosen to be $Beta(\alpha^*, \beta^*)$, where α^* and β^* are obtained by equating the first two moments. That is,

$$\frac{\beta^*}{\alpha^* + \beta^*} = A \tag{4.35}$$

and

$$\frac{\beta^*(\beta^* + 1)}{(\alpha^* + \beta^* + 1)(\alpha^* + \beta^*)} = B, \tag{4.36}$$

where

$$A = \frac{\Gamma(\alpha + \beta)\Gamma(\beta + k)}{\Gamma\beta\Gamma(\alpha + \beta + k)}$$

and

$$B = \frac{\Gamma(\alpha + \beta)\Gamma(\beta + 2k)}{\Gamma\beta\Gamma(\alpha + \beta + 2k)}$$

Solving for α^* and β^* from equations (4.35) and (4.36), we have

$$\alpha^* = \frac{(1 - A)(B - A)}{A^2 - B}$$

and

$$\beta^* = \frac{A(B - A)}{A^2 - B}$$

For the simple case, where $\alpha = 1$, we find that

$$A = \frac{\Gamma(1 + \beta)\Gamma(\beta + k)}{\Gamma\beta\Gamma(1 + \beta + k)} = \frac{\beta}{\beta + k}$$

and

$$B = \frac{\Gamma(1 + \beta)\Gamma(\beta + 2k)}{\Gamma\beta\Gamma(1 + \beta + 2k)} = \frac{\beta}{\beta + 2k}$$

Thus,

$$\alpha^* = \frac{k\beta[(\beta + k) - (\beta + 2k)]}{\beta[\beta(\beta + 2k) - (\beta + k)^2]} = 1$$

and

$$\beta^* = \frac{\beta^2[(\beta + k) - (\beta + 2k)]}{\beta[\beta(\beta + 2k) - (\beta + k)^2]} = \frac{\beta}{k}$$

Example (4.1):

Liu et al(1997) reported results on 1875 blood donors screened for anti HCV at the Blood Transfusion Service in China. The 1875 serum samples were tested individually ($k = 1$) to examine effectiveness of pooling. With a group size of $k = 5$ and $g = 375$, they got $r = 37$. Using equation (4.15),

$$\hat{\beta} = \frac{1875}{37} = 50.66.$$

Thus, using equation (4.18) the posterior mean is given by,

$$\begin{aligned}\hat{p} &= 1 - \left(\frac{375 - 37}{376} \right)^{\frac{1}{5}} \\ &= 1 - 0.97891 \\ &= 0.021083,\end{aligned}$$

which compares favorably with $\hat{\beta} = 48.13$ and $\hat{p} = 0.020557$, given by Tebbs and Bilder(2003). For the indirect Bayes estimate equation (4.27), we get

$$\begin{aligned}\hat{p}_{Ibys} &= 1 - \left(1 - \frac{37 + 1}{375 + 1 + \frac{50.66}{5}} \right)^{\frac{1}{5}} \\ &= 1 - 0.979494 \\ &= 0.020506.\end{aligned}$$

Comparison can also be made in terms of relative bias and relative efficiency which are respectively defined as,

$$RB(\hat{p}) = \frac{E(\hat{p} - p)}{p}$$

and the relative efficiency to be

$$RE(\hat{p}) = \frac{MSE(\hat{p}_k)}{MSE(\bar{p})},$$

where \hat{p}_k is the estimate for $k > 1$. The tables below give a summary of the relative bias and the relative efficiency for various values of p and k for $g = 10$, based on the maximum likelihood, the direct Bayes and the indirect Bayes estimators.

Table 6. Relative Bias for selected values of p, k and $g = 10$.

g=10						
p	k	\hat{p}_{mle}	(α, β)	\hat{p}_{bys}	(α^*, β^*)	\hat{p}_{Ibys}
0.25			(1,3)			
	1	0.00000		0.00000	(1,3.00)	0.00000
	5	0.21661		0.09552	(1,0.60)	0.01252
	10	1.57583		0.29291	(1,0.30)	0.00192
	15	2.56691		0.37393	(1,0.20)	-0.10853
	20	2.88854		0.35010	(1,0.15)	-0.23565
0.10			(1,9)			
	1	0.00000		0.00000	(1,9.00)	0.00000
	5	0.05731		0.03843	(1,1.80)	0.00311
	10	0.19010		0.08031	(1,0.90)	0.00494
	15	0.90474		0.14572	(1,0.60)	0.01453
	20	2.39111		0.23484	(1,0.45)	0.01712
0.05			(1,19)			
	1	0.00000		0.00000	(1,19.0)	0.00000
	5	0.04851		0.02363	(1,3.80)	-0.00740
	10	0.06713		0.04351	(1,1.90)	-0.00431
	15	0.11444		0.06241	(1,1.27)	-0.00053
	20	0.29962		0.08494	(1,0.95)	-0.00039
0.01			(1,99)			
	1	0.00000		0.00000	(1,99.0)	0.00000
	5	0.04360		0.00452	(1,19.8)	-0.00870
	10	0.05082		0.01180	(1,9.90)	-0.01101
	15	0.05451		0.01821	(1,6.60)	-0.01103
	20	0.05734		0.02374	(1,4.95)	-0.01033

Table 7. Relative Efficiency for selected values of p, k and $g = 10$.

g=10						
p	k	\hat{p}_{mle}	(α, β)	\hat{p}_{bys}	(α^*, β^*)	\hat{p}_{Ibys}
0.25			(1,3)			
	1	1.00000		1.96000	(1,3.00)	1.96000
	5	1.03260		4.69672	(1,0.60)	7.04301
	10	0.99481		19.03341	(1,0.30)	80.62304
	15	0.99862		35.90983	(1,0.20)	269.94501
	20	0.99971		59.85294	(1,0.15)	147.47321
0.1			(1,9)			
	1	1.00000		4.00000	(1,9.00)	4.00000
	5	1.17922		1.70664	(1,1.80)	1.91413
	10	1.04641		7.82342	(1,0.90)	10.69631
	15	1.00283		35.66633	(1,0.60)	69.54100
	20	0.99924		71.11520	(1,0.45)	210.27562
0.05			(1,19)			
	1	1.00000		9.00000	(1,19.0)	9.00000
	5	1.12490		2.19461	(1,3.80)	2.35772
	10	1.19852		1.89812	(1,1.90)	2.17101
	15	1.07565		5.68040	(1,1.27)	7.04622
	20	1.01584		24.86763	(1,0.95)	34.74474
0.01			(1,99)			
	1	1.00000		121.00000	(1,99.0)	121.00000
	5	1.09484		9.85463	(1,19.8)	10.12575
	10	1.11742		4.47010	(1,9.90)	4.69181
	15	1.13311		3.14221	(1,6.60)	3.35011
	20	1.14780		2.57100	(1,4.95)	2.77764

We note from the tables, that the biases and the MSE's of the Bayes

estimators are smaller than that of MLE, especially for low prevalence rate. The indirect Bayes estimator in general performs very well for small k and large g , in the sense of having even smaller bias and MSE. This mainly occurs when the experimenter is forced to use smaller group sizes, perhaps due to biological considerations involved in test assays.

4.7.2 Asymptotic Distribution and Interval Estimation

In practise, the the population size and the number of groups are fairly large, so that large sample estimates can be used for inference purposes. The asymptotic distributions of the MLE and that of the Bayesian estimator follow from the general result which obeys the Mann-Wald theorem, given in Rao(1973). For the function h if

$$h(p') = 1 - \left(1 - \frac{p' + b}{1 + c}\right)^a \quad (4.37)$$

then

$$\sqrt{m} \frac{h(\hat{p}') - h(p')}{p'(1 - p')} \sim N\left(0, \left[\frac{a}{1 + c} \left(1 - \frac{p' + b}{1 + c}\right)^{a-1}\right]^2\right). \quad (4.38)$$

For the Bayes estimator $h(\hat{p}') = \hat{p}$, with $b = \frac{\alpha}{g}$, $c = \frac{(\alpha + \beta)}{g}$ and $a = \frac{1}{k}$. Thus,

$$\text{var}(\hat{p}) = \left\{ \frac{g}{k(g + \alpha + \beta)} \left[1 - \frac{gp^* + \alpha}{g + \alpha + \beta}\right]^{\frac{1}{k}-1} \right\}^2 \frac{p^*(1 - p^*)}{g}. \quad (4.39)$$

From the example (4.1) above, $p = 0.0224$ and $\hat{p} = 0.021083$, giving that

$$\begin{aligned} \text{var}(\hat{p}) &= \left\{ \frac{0.2}{1.13101} \left[1 - \frac{0.0224 + 0.00267}{1.13101}\right]^{-0.8} \right\}^2 \frac{0.0224(0.9776)}{375} \\ &= 1.89270 \times 10^{-6}. \end{aligned}$$

The standard error of \hat{p} is therefore 0.0026965, resulting in a 95% confidence interval of (0.0183865, 0.023779). For the maximum likelihood estimator, the confidence interval is given by,

$$\hat{p}_{mle} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\text{var}(\hat{p}_{mle})}{g}},$$

where $\text{var}(\hat{p}_{mle})$ is the asymptotic variance given by $\text{var}\hat{p}_{mle} = k^{-2}[1 - (1 - p)^k](1 - p)^{2-k}$ and $z_{1-\frac{\alpha}{2}}$ denotes the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. Thus, from Example (4.1) above, we get

$$\begin{aligned} \text{var}(\hat{p}_{mle}) &= 5^{-2}[1 - (1 - 0.020557)^5](1 - 0.020557)^{2-5} \\ &= \frac{[1 - (0.979443)^5](0.979443)^{-3}}{25} \\ &= 0.004199. \end{aligned}$$

The standard error is $(0.004199/375)^{0.5} = 0.0033466$ giving a 95% confidence of(0.01905, 0.02575). Thus the indirect Bayesian estimate is more precise than the MLE, since for the same level of confidence it gives a narrower interval.

4.8 Estimation With Errors in Decisions

Estimates based on screening tests can be severely biased, unless adjusted for the sensitivity and specificity of the screening test. One such estimator is the MLE, which can produce negative confidence endpoints. In this section, an attempt is made in studying a Bayesian estimate which always lies between zero and one.

4.8.1 Derivation of a Bayesian Estimator

If r is the number of defective groups out of g groups and π_1^* is the probability that a group-factor is declared defective, then

$$f(r/\pi_1^*) = \binom{g}{r} \pi_1^{*r} (1 - \pi_1^*)^{g-r}. \quad (4.40)$$

Using the conjugate prior for the binomial distribution, the prior distribution of π_1^* is given by,

$$f(\pi_1^*) = \frac{\Gamma(a+b)}{\Gamma a \Gamma b} \pi_1^{*a-1} (1 - \pi_1^*)^{b-1}. \quad (4.41)$$

The joint distribution of r and π_1^* is given by,

$$\begin{aligned} f(r, \pi_1^*) &= \binom{g}{r} \pi_1^{*r} (1 - \pi_1^*)^{g-r} \frac{\Gamma(a+b)}{\Gamma a \Gamma b} \pi_1^{*a-1} (1 - \pi_1^*)^{b-1} \\ &= \binom{g}{r} \{B(a, b)\}^{-1} \pi_1^{*a+r-1} (1 - \pi_1^*)^{g-r+b-1}, \end{aligned}$$

where

$$B(a, b) = \frac{\Gamma a \Gamma b}{\Gamma(a+b)}.$$

The marginal probability density function of r is given by,

$$f(r) = \int_0^1 \binom{g}{r} \{B(a, b)\}^{-1} \pi_1^{*a+r-1} (1 - \pi_1^*)^{g-r+b-1} d\pi_1^*. \quad (4.42)$$

Using the sensitivity-specificity approach, π_1^* in the above equation is replaced by $\eta - (\eta + \theta - 1)q^k$, to give

$$\begin{aligned} f(r) &= \int_0^1 \binom{g}{r} \{B(a, b)\}^{-1} [\eta - (\eta + \theta - 1)q^k]^{a+r-1} [1 - (\eta - (\eta + \theta - 1)q^k)]^{g-r+b-1} \\ &\quad k(\eta + \theta - 1)(1 - p)^{k-1} dp, \end{aligned}$$

which can be solved using numerical integration.

4.8.2 Alternative Approach

In group screening designs, the probability that a group-factor is defective is given by,

$$p^* = \eta p + (1 - \theta)(1 - p),$$

which can be simplified to

$$p^* = (\eta + \theta - 1)p + 1 - \theta,$$

to give

$$\hat{p} = \frac{\theta + p^* - 1}{\eta + \theta - 1}. \quad (4.43)$$

With g groups the number of positive groups r follow a binomial distribution:

$$f(r/p^*) = \binom{g}{r} p^{*r} (1 - p^*)^{g-r}.$$

The expression $p^{*r}(1 - p^*)^{g-r}$ is proportional to the likelihood function $L(p)$. Choosing the uniform distribution as the prior distribution of p , the posterior mean and Bayesian estimator p_{bys} of p is,

$$\hat{p} = \frac{\int_0^1 p L(p) dp}{\int_0^1 L(p) dp}. \quad (4.44)$$

But

$$p = \frac{\theta + p^* - 1}{\eta + \theta - 1}$$

Thus,

$$dp = \frac{dp^*}{\eta + \theta - 1}.$$

Therefore,

$$\begin{aligned}
 p_{bys} &= \frac{\int_{1-\theta}^{\eta} \frac{\theta+p^*-1}{\eta+\theta-1} p^{*r} (1-p^*)^{g-r} \frac{dp}{\eta+\theta-1}}{\int_{1-\theta}^{\eta} p^{*r} (1-p^*)^{g-r} \frac{dp}{\eta+\theta-1}} \\
 &= \frac{1}{\eta+\theta-1} \frac{\int_{1-\theta}^{\eta} (\theta+p^*-1) p^{*r} (1-p^*)^{g-r} dp}{\int_{1-\theta}^{\eta} p^{*r} (1-p^*)^{g-r} dp} \\
 &= \frac{1}{\eta+\theta-1} \left[(\theta-1) + \frac{\int_{1-\theta}^{\eta} p^{*r+1} (1-p^*)^{g-r} dp^*}{\int_{1-\theta}^{\eta} p^{*r} (1-p^*)^{g-r} dp^*} \right] \\
 &= \frac{d+\theta-1}{\eta+\theta-1}, \tag{4.45}
 \end{aligned}$$

where

$$d = \frac{\int_{1-\theta}^{\eta} p^{*r+1} (1-p^*)^{g-r} dp^*}{\int_{1-\theta}^{\eta} p^{*r} (1-p^*)^{g-r} dp^*}.$$

Example (4.2):

A sample survey consisted of the screening of chest radiographs of 96 groups of Black women for the presence of pulmonary hypertension (Lew and Levy (1989)). An enlarged artery was the screen for pulmonary hypertension; 8 groups out of 96 ($p^* = 0.08333$) were positive for enlargement. Assume that the sensitivity and specificity of the procedure are 0.89 and 0.74 respectively.

The Bayesian estimate from eqn (4.38) is,

$$\frac{d + 0.74 - 1}{0.89 + 0.74 - 1}$$

where

$$d = \frac{\int_{0.26}^{0.89} p^{*9} (1-p^*)^{88} dp^*}{\int_{0.26}^{0.89} p^{*8} (1-p^*)^{88} dp^*},$$

$$= 0.2706573363.$$

The estimate is therefore,

$$\frac{0.27 + 0.74 - 1}{0.89 + 0.74 - 1} = 0.0159.$$

The maximum likelihood estimate of p^* from equation (4.36) is,

$$\hat{p} = \frac{(0.0833 + 0.74 - 1)}{(0.89 + 0.74 - 1)} = -0.28.$$

Thus, the Bayesian estimate is non-negative, whereas the MLE is negative.

In this case, the MLE is set to zero.

4.8.3 Approximation of the Variance

Since both the Bayesian and the MLE are consistent estimators, asymptotically there is no significant difference in their variances. Thus, the variance of the Bayesian estimate can be approximated by substituting \hat{p}_{bys}^* in the estimated variance for MLE,

$$\text{var}(\hat{p}_{bys}^*) = \frac{p^*(1 - p^*)}{[g(\eta + \theta - 1)^2]}.$$

In the numerical Example (4.2) above, the variance of \hat{p}_{bys}^* is

$$\text{var}(\hat{p}_{bys}^*) = \frac{0.0159(1 - 0.0159)}{[96(0.89 + 0.74 - 1)^2]} = 0.00041066,$$

and its estimated standard error is $SE = \sqrt{0.00041066} = 0.02026$.

Chapter 5

Group Screening Design With Equal Probabilities but With Unequal Group Sizes

5.1 Introduction

So far we have considered only cases, where the group size is assumed constant throughout the experiment. In this chapter, we consider estimating the prevalence rate in populations with variable group sizes under both the condition of no errors and when errors are taken into consideration, a common situation in practise. We shall assume that, each member of the population is equally likely to be sampled, and that positive and negative members are on average homogeneously distributed across pools. In section 5.2, we consider estimation without errors in decision and use the Maximum Likelihood Estimation Method and Newton-Raphson method of iteration to determine successive estimates. Models studied by various researchers are considered as special cases in section 5.3. Estimation with

errors in decision is briefly discussed in section 5.4.

5.2 Estimation without Errors in Decision

5.2.1 Notations and Method

Let k_i be the number of organisms in a pool, r_i be the number of positive pools of size k_i and g_i be the number of pools each of size k_i . The number of positive pools r_i is binomial with probability distribution,

$$f(r_i) = \binom{g_i}{r_i} (1 - q^{k_i})^{r_i} (q^{k_i})^{g_i - r_i}; r_i = 0, 1, \dots, g_i.$$

The mean of r_i is $E(r_i) = g_i(1 - q^{k_i})$ and the variance is $\text{var}(r_i) = g_i(1 - q^{k_i})q^{k_i}$, where $1 - q^{k_i}$ is the probability that a pool of size k_i is positive.

Conditional on the set of group sizes, the likelihood function is,

$$L = \prod_i \binom{g_i}{r_i} (1 - q^{k_i})^{r_i} (q^{k_i})^{g_i - r_i}, \quad (5.1)$$

where the product is over all the different group sizes used. Taking logarithms on both sides, we have

$$\ln L = \sum_{i=1}^{\ell} \ln \left[\binom{g_i}{r_i} (1 - q^{k_i})^{r_i} (q^{k_i})^{g_i - r_i} \right].$$

Differentiating with respect to q and equating to zero gives,

$$\sum_{i=1}^{\ell} \left[\frac{r_i}{1 - q^{k_i}} (-k_i q^{k_i - 1}) + \frac{k_i(g_i - r_i)}{q} \right] = 0, \quad (5.2)$$

which yields

$$\sum_{i=1}^{\ell} \frac{k_i r_i q^{k_i - 1}}{1 - q^{k_i}} = \sum_{i=1}^{\ell} \frac{k_i(g_i - r_i)}{q}.$$

Therefore,

$$\sum_{i=1}^{\ell} \frac{k_i r_i q^{k_i-1}}{1 - q^{k_i}} = \sum_{i=1}^{\ell} \frac{k_i g_i}{q} - \sum_{i=1}^{\ell} \frac{k_i r_i}{q},$$

implying that,

$$\sum_{i=1}^{\ell} \frac{k_i r_i q^{k_i}}{1 - q^{k_i}} + \sum_{i=1}^{\ell} k_i r_i = \sum_{i=1}^{\ell} k_i g_i$$

and

$$\sum_{i=1}^{\ell} [k_i r_i (\frac{q^{k_i} + 1 - q^{k_i}}{1 - q^{k_i}})] = \sum_{i=1}^{\ell} k_i g_i.$$

Hence,

$$\sum_{i=1}^{\ell} \frac{k_i r_i}{1 - q^{k_i}} = \sum_{i=1}^{\ell} k_i g_i. \quad (5.3)$$

Let the population size f be given by,

$$f = \sum_{i=1}^{\ell} k_i g_i.$$

Then, by equation (5.3)

$$\begin{aligned} f &= \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - q^{k_i}} \\ &= \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - (1 - \hat{p})^{k_i}} \end{aligned} \quad (5.4)$$

and thus,

$$f - \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - q^{k_i}} = 0.$$

Suppose we set the equation as,

$$h(\hat{q}_t) = f - \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - q_t^{k_i}}.$$

Then by differentiating the equation, we have

$$\begin{aligned} h'(\hat{q}_t) = \frac{d}{dq_t} h(\hat{q}_t) &= \sum_{i=1}^{\ell} k_i r_i (1 - q_t^{k_i})^{-2} (-k_i q_t^{k_i-1}) \\ &= \sum_{i=1}^{\ell} \frac{k_i^2 r_i q_t^{k_i-1}}{(1 - q_t^{k_i})^2}. \end{aligned}$$

Using Newton-Raphson method of iteration, we have

$$\begin{aligned} \hat{q}_{t+1} &= \hat{q}_t + \frac{h(\hat{q}_t)}{h'(\hat{q}_t)} \\ &= \hat{q}_t + \frac{f - \lambda_1}{\lambda_2}, \end{aligned} \tag{5.5}$$

where

$$\lambda_1 = \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - q_t^{k_i}}$$

and

$$\lambda_2 = \sum_{i=1}^{\ell} \frac{k_i^2 r_i q_t^{k_i-1}}{(1 - q_t^{k_i})^2}.$$

The initial value in this case is the minimum infection rate (MIR), given by dividing total number of positive pools by total number of specimens used in all the pools and is given by,

$$q_0 = 1 - \sum_{i=1}^{\ell} \left\{ \frac{r_i}{f} \right\}.$$

An alternative iterative method is to solve formula (5.2) for q . That is,

$$\sum_{i=1}^{\ell} \frac{k_i r_i q^{k_i}}{1 - q^{k_i}} = \sum_{i=1}^{\ell} \frac{k_i (g_i - r_i)}{q},$$

giving

$$q = \frac{\sum_{i=1}^{\ell} k_i (g_i - r_i)}{\sum_{i=1}^{\ell} \left\{ \frac{k_i r_i q^{k_i-1}}{1 - q^{k_i}} \right\}}. \tag{5.6}$$

Therefore,

$$q_{t+1} = \frac{\sum_{i=1}^{\ell} k_i (g_i - r_i)}{\lambda_3}, \quad (5.7)$$

where

$$\lambda_3 = \sum_{i=1}^{\ell} \left\{ \frac{k_i r_i q_t^{k_i - 1}}{1 - q_t^{k_i}} \right\},$$

for $t=0,1,2,\dots$

In order to circumvent the iteration involved in equation(5.4), let k_i be replaced with k_+ , which is the average size of positive pools. That is,

$$k_+ = \frac{\sum_{i=1}^{\ell} k_i r_i}{\sum_{i=1}^{\ell} r_i}.$$

Then equation (5.4) becomes,

$$f - \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - q^{k_+}} = 0,$$

which gives,

$$\begin{aligned} f &= \frac{1}{1 - q^{k_+}} \sum_{i=1}^{\ell} k_i r_i \\ &= \frac{1}{1 - q^{k_+}} f_+, \end{aligned}$$

where

$$f_+ = \sum_{i=1}^{\ell} k_i r_i,$$

is the total number of organisms in all positive pools.

Thus,

$$\hat{q} = \left[1 - \frac{f_+}{f} \right]^{\frac{1}{k_+}}. \quad (5.8)$$

Equation (5.6) therefore becomes,

$$q = \left\{ \frac{f - f_+}{\sum_{i=1}^{\ell} k_i r_i} \right\} \frac{1 - q^{k_+}}{q^{k_+ - 1}}.$$

Thus, based on this result, we have

$$q^{k_+} = \left\{ \frac{f - f_+}{f_+} \right\} (1 - q^{k_+}),$$

which implies that

$$\left\{ 1 + \frac{f - f_+}{f_+} \right\} q^{k_+} = \left\{ \frac{f - f_+}{f_+} \right\}$$

and hence,

$$\hat{q} = \left\{ 1 - \frac{f_+}{f} \right\}^{\frac{1}{k_+}}. \quad (5.9)$$

5.2.2 Poisson Model Approximation to Binomial Model

Suppose p is very small and the group size is large, then letting $p = \frac{\pi}{k_+}$ formula (5.4) becomes,

$$\begin{aligned} f &= \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - (1 - p)^{k_i}} \\ &= \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - \left(1 - \frac{\pi}{k_+}\right)^{k_i}} \\ &\approx \sum_{i=1}^{\ell} \frac{k_i r_i}{1 - e^{-\pi}} \\ &= \frac{1}{1 - e^{-\pi}} \sum_{i=1}^{\ell} k_i r_i. \end{aligned} \quad (5.10)$$

That is,

$$f = \frac{1}{1 - e^{-pk_+}} f_+. \quad (5.11)$$

This expression can then be given as,

$$\ln \left[1 - \frac{f_+}{f} \right] = -pk_+,$$

leading to

$$\hat{p} = -\frac{1}{k_+} \ln \left[1 - \frac{f_+}{f} \right]. \quad (5.12)$$

5.2.3 Asymptotic Variance of \hat{p}

From the binomial distribution of r_i , in equation (5.1), variance of r_i is,

$$\text{var}(r_i) = g_i(1 - q^{k_i})q^{k_i}.$$

Asymptotically, for large g_i , $\text{var}(\hat{p})$ can be estimated by $\text{var}(\hat{q})$, as given in the following theorem.

Theorem 5.2.1

For small p , the asymptotic variance of \hat{q} is given by

$$\text{var}(\hat{q}) \approx \frac{\hat{p}}{\sum_{i=1}^l k_i g_i q^{k_i-2}}.$$

Proof

Based on Cramer-Rao method, we have

$$\lim_{g_i \rightarrow \infty} \text{var}(\hat{p}) = \frac{1}{-E \left[\frac{\partial^2 \ln L}{\partial q^2} \right]}.$$

But,

$$\frac{\partial \ln L}{\partial q} = \sum_{i=1}^{\ell} k_i \left[\frac{g_i - r_i}{q} - \frac{r_i q^{k_i-1}}{1 - q^{k_i}} \right]$$

and

$$\frac{\partial^2 \ln L}{\partial q^2} = \sum_{i=1}^{\ell} k_i \frac{\partial}{\partial q} \left[\frac{g_i - r_i}{q} \right] - \sum_{i=1}^{\ell} k_i r_i \frac{\partial}{\partial q} \left[\frac{q^{k_i-1}}{1 - q^{k_i}} \right].$$

Now,

$$\frac{\partial}{\partial q} \left[\frac{q^{k_i-1}}{1 - q^{k_i}} \right] = \frac{(1 - q^{k_i})(k_i - 1)q^{k_i-2}r_i}{(1 - q^{k_i})^2} + \frac{k_i q^{k_i-1} q^{k_i-1} r_i}{(1 - q^{k_i})^2},$$

giving,

$$\begin{aligned} E \left[\frac{\partial^2 \ln L}{\partial q^2} \right] &= \frac{k_i r_i (1 - q^{k_i})}{q^2} - \frac{k_i g_i}{q^2} - \frac{(k_i - 1)(1 - q^{k_i})^2 q^{k_i-2} k_i g_i}{(1 - q^{k_i})^2} - \frac{k_i^2 g_i (1 - q^{k_i}) q^{2k_i-2}}{(1 - q^{k_i})^2} \\ &= -k_i^2 g_i q^{k_i-2} - \frac{k_i^2 g_i q^{2k_i-2}}{1 - q^{k_i}} \\ &= \frac{-k_i^2 g_i q^{k_i-2}}{1 - q^{k_i}}. \end{aligned}$$

Thus,

$$-E \frac{\partial^2 \ln L}{\partial q^2} = \sum_{i=1}^{\ell} \frac{k_i^2 g_i q^{k_i-2}}{1 - q^{k_i}},$$

implying that,

$$\text{var}(\hat{q}) = \frac{1}{\sum_{i=1}^{\ell} \frac{k_i^2 g_i q^{k_i-2}}{1 - q^{k_i}}}.$$

For small p , $1 - q^{k_i} = 1 - (1 - p)^{k_i} \approx k_i p$.

Therefore,

$$f \approx \sum_{i=1}^{\ell} \frac{k_i r_i}{k_i p} = \sum_{i=1}^{\ell} \frac{r_i}{p},$$

leading to,

$$\begin{aligned}\text{var}(\hat{q}) &\approx \frac{1}{\sum_{i=1}^{\ell} \frac{k_i^2 g_i q^{k_i-2}}{k_i p}} \\ &= \frac{\hat{p}}{\sum_{i=1}^{\ell} k_i g_i q^{k_i-2}},\end{aligned}$$

where

$$\hat{p} \approx \sum_{i=1}^{\ell} \frac{r_i}{f}.$$

Hence the proof.

5.3 Special Cases

In this section, models considered by some researchers such as Chiang and Reeves (1962), Bhattacharya et al (1979) and Griffiths (1972) are treated as special cases of the general model presented above.

5.3.1 Chiang-Reeves model

Chiang and Reeves (1962) used two groups with $k_1 = k$ and $k_2 = 2k$. Thus, in this case

$$\begin{aligned}\text{var}(\hat{q}) &= \frac{1}{\frac{g_1 k^2 q^{k-2}}{1-q^k} + \frac{g_2 (2k)^2 q^{2k-2}}{1-q^{2k}}} \\ &= \frac{1-q^{2k}}{k^2 g_1 q^{k-2} (1+q^k) + 4k^2 g_2 q^{2k-2}}\end{aligned}$$

$$\begin{aligned}
&= \frac{1 - q^{2k}}{k^2 q^{k-2} [g_1(1 + q^k) + 4g_2 q^k]} \\
&= \frac{1 - q^{2k}}{k^2 q^{k-2} [g_1 + (g_1 + 4g_2)q^k]}. \tag{5.13}
\end{aligned}$$

5.3.2 Bhattacharya model

In this case, $\ell = 1$ so that

$$f - \frac{k_1 - r_1}{1 - q^{k_1}} = 0,$$

implying that,

$$f = \frac{k_1 - r_1}{1 - q^{k_1}}$$

and hence,

$$f q^{k_1} = f - r_1 k_1.$$

Therefore,

$$\begin{aligned}
\hat{q} &= \left[1 - \frac{k_1 r_1}{f}\right]^{\frac{1}{k_1}} \\
&= \left[1 - \frac{r_1}{f/k_1}\right]^{\frac{1}{k_1}} \\
&= \left[1 - \frac{r_1}{g_1}\right]^{\frac{1}{k_1}}.
\end{aligned}$$

When the suffix 1 is removed, we have

$$\hat{q} = \left[1 - \frac{r}{g}\right]^{\frac{1}{k}}$$

and hence,

$$\begin{aligned}
\text{var}(\hat{q}) &= \frac{1}{g k^2 q^{k-2}} \\
&= \frac{1 - q^k}{g k^2 q^{k-2}}.
\end{aligned}$$

5.3.3 Griffiths Approach

When $g_i = 1 \forall i$, then

$$L = \prod_{i=1}^{\ell} \binom{1}{r_i} (1 - q^{k_i})^{r_i} (q^{k_i})^{g_i - r_i}.$$

In this case, r_i takes the value 1 or 0. For $r_i = 1$, let $i \in M$ and for $r_i = 0$ let $i \notin M$.

Therefore,

$$L = \prod_{i \in M} (1 - q^{k_i}) \prod_{i \notin M} q^{k_i},$$

implying that,

$$\ln L = \sum_{i \in M} \ln(1 - q^{k_i}) + \sum_{i \notin M} \ln q^{k_i}$$

and

$$\frac{\partial}{\partial q} \ln L = \sum_{i \in M} \frac{-k_i q^{k_i-1}}{1 - q^{k_i}} + \sum_{i \notin M} \frac{k_i}{q}.$$

Equating to zero, we have

$$\bar{q} = \frac{\sum_{i \notin M} k_i}{\sum_{i \in M} \left[\frac{k_i q^{k_i-1}}{1 - q^{k_i}} \right]},$$

giving

$$\bar{q}_{t+1} = \frac{\sum_{i \notin M} k_i}{\sum_{i \in M} \left[\frac{k_i q_t^{k_i-1}}{1 - q_t^{k_i}} \right]} \quad \text{for } t = 0, 1, 2, \dots$$

The initial value used was

$$q_0 = \left[\frac{\sum_i (\bar{g}_i - r_i)}{\sum_i g_i} \right]^{\frac{1}{k^*}}$$

$$\begin{aligned}
&= \left[\frac{f - \sum_i r_i}{f} \right]^{\frac{1}{k^*}} \\
&= \left[\frac{\text{Number of groups not infected}}{\text{Total number of groups}} \right]^{\frac{1}{k^*}},
\end{aligned}$$

where k^* is the optimal value of group size.

5.4 Estimation With Errors in Decision

When test error is taken into consideration, the likelihood function is given by,

$$L = \prod_i \binom{g_i}{r_i} \pi_{1i}^{*r_i} (1 - \pi_{1i}^*)^{g_i - r_i}, \quad (5.14)$$

where

$$\pi_{1i}^* = \eta - (\eta + \theta - 1)q^{k_i}.$$

Therefore,

$$\ln L = \sum_{i=1}^{\ell} \ln \left[\binom{g_i}{r_i} \pi_{1i}^{*r_i} (1 - \pi_{1i}^*)^{g_i - r_i} \right]. \quad (5.15)$$

Differentiating equation(5.15) with respect to q and equating to zero, we have

$$\sum_{i=1}^{\ell} \left[\frac{r_i}{\pi_{1i}^*} - \frac{(g_i - r_i)}{1 - \pi_{1i}^*} \right] \frac{\partial \pi_{1i}^*}{\partial q} = 0,$$

which gives,

$$\sum_{i=1}^{\ell} \frac{r_i}{\pi_{1i}^*} \frac{\partial \pi_{1i}^*}{\partial q} = \sum_{i=1}^{\ell} \frac{(g_i - r_i)}{1 - \pi_{1i}^*} \frac{\partial \pi_{1i}^*}{\partial q}. \quad (5.16)$$

But

$$\frac{\partial}{\partial q} \pi_{1i}^* = -k_i(\eta + \theta - 1)q^{k_i - 1}.$$

Thus, equation(5.16) becomes

$$\begin{aligned} \sum_{i=1}^{\ell} \frac{r_i k_i q^{k_i-1}}{\eta - (\eta + \theta - 1)q^{k_i}} &= \sum_{i=1}^{\ell} \frac{k_i (g_i - r_i) q^{k_i-1}}{1 - \eta + (\eta + \theta - 1)q^{k_i-1}} \\ &= \frac{1}{q} \sum_{i=1}^{\ell} \frac{k_i (g_i - r_i) q^{k_i-1}}{1 - \eta + (\eta + \theta - 1)q^{k_i-1}}, \end{aligned}$$

Hence,

$$q = \frac{\mu_1}{\mu_2},$$

where

$$\mu_1 = \sum_{i=1}^{\ell} \frac{k_i (g_i - r_i) q_0^{k_i-1}}{1 - \eta + (\eta + \theta - 1)q_0^{k_i-1}}$$

and

$$\mu_2 = \sum_{i=1}^{\ell} \frac{r_i k_i q_0^{k_i-1}}{\eta - (\eta + \theta - 1)q_0^{k_i}},$$

where q_0 is the initial value. The Newton-Raphson iterative method can then be used to determine successive estimates of the infection rate.

Chapter 6

Summary and Conclusions

6.1 Summary

Our objective in this thesis was to study Estimation Problem in Group Screening Designs with emphasis on the methods of Maximum Likelihood and Bayesian Estimation, under various conditions.

In chapter 1, we reviewed the concept of group screening designs, discussed the two areas of concern; identification of individuals called *classification problem* and estimating proportions called *estimation problem*. Terminologies and notations that have been used by various researchers and are specific to particular areas of application were also discussed. Various areas of application of group screening designs, such as public health, phytopathology, epidemiology and industrial studies were discussed. Based on the areas of application and the literature review, a *theoretical framework* for studying estimation in group screening designs has been developed. It is this framework that formed the basis of our study in the thesis.

In chapter 2, we reviewed group screening without errors in decision

using the maximum likelihood estimation method based on Thompson's (1962) work. The MLE estimate of probability of defective factor was calculated and its properties such as biasedness, BAN, MSE and asymptotic efficiency discussed. For small group sizes and large values of p and k , it was observed that the bias is considerably large, while for small values of p and k , there was relatively low bias. Determination of the optimal group size using MSE was considered and it was observed that the MSE decreases to a minimum, as the optimal value of k is attained, then increases. After the minimum MSE is reached, the rate of increase in MSE with k decreases as the number of group-factors increases. Also discussed was the choice of the most appropriate model based on the cost consideration in terms of the model with the least MSE. It was shown that the best design for a given experiment depends on p , tolerable MSE, the sampling cost of an individual sample and the cost of performing one test.

Maximum likelihood estimation with errors in decisions was discussed in chapter 3. Estimation of prevalence is challenging especially when the prevalence is small. One reason is that the presence of measurement errors resulting from the limited precision of tests makes estimation, using traditional methods, impossible in some screening situations. Measurement error is real, hence ignoring it leads to severe bias, and inference about the prevalence becomes unsatisfactory. Indeed in a low prevalence situation the expected number of false positives is very high, often even higher than the number of true positives. The second reason is that in the low prevalence areas large sample is needed in order to obtain non-zero estimate.

This is usually a very costly and often unrealistic solution. This chapter considered advantages and disadvantages of group testing as alternative solution to this problem. We showed that, by group testing we not only achieve a cost saving, but also an increase in the estimation accuracy. We also discussed the statistical properties of the estimator such as biasedness and asymptotic properties.

In Chapter 4, we discussed a Bayesian procedure to estimate the prevalence rate with equal group sizes, equal probability with and without errors in decisions, using a beta-type prior distribution and a squared-error loss function. Two choices of prior were considered; (i) the prior on p , the population proportion and (ii) the prior on $p^* = 1 - (1 - p)^k$, the group prevalence proportion. The performance of the Bayes estimator was evaluated in terms of bias and relative efficiency in comparison to the Maximum Likelihood Estimator (MLE). It was observed that the Bayes estimator outperforms the MLE, for small group sizes and small p . In addition, we also discussed the asymptotic property of the Bayes estimator and the interval estimation, using the frequentist approach. The methods were illustrated using group-testing data from Liu et al(1997), a prospective hepatitis C virus study conducted in China. In extending the the discussion to Bayesian estimation with errors in decisions, we observed that estimates of prevalence based on screening tests can be severely biased unless adjusted for the sensitivity and specificity of the screening test. One such estimate is the MLE, which can yield an extreme estimate of zero or one that has undesirable characteristic, such as a standard error of zero.

We then showed that, a Bayesian estimator always falls between zero and one.

In chapter 5, we reviewed estimation in a population of organisms when unequal sample pools are analyzed, a common situation in practice, but not one which can be dealt with by existing methodology. An iterative method of determining successive estimates of the infection rate was discussed. An example was given of estimating the infection rate of yellow fever virus in a mosquito population. The minimum infection rate (MIR), which is the ratio of positive pools to the total number of pools was used to estimate the true infection rate (TIR) by considering the MIR as the initial estimate for the iterative procedure of the maximum likelihood method. In order to circumvent the iterations involved, an alternative estimator which can easily be evaluated and upgraded, using the average size of positive pools was suggested. A brief introduction of estimation with errors was given, and this can be considered for further research.

Chapter 6, gives the summary of what has been done chapter by chapter in section 6.1, while section 6.2 gives the conclusions from the study. Suggestions for further research is discussed in section 6.3.

6.2 Conclusion

Group testing is in general economical in the light of reduced average number of units to be tested or samples to be screened, since a group is declared defective as soon as an item is found to be defective. Based on a given total cost per unit to be tested, a group size k may be determined in

advance and fixed. Next, based on the prior knowledge of the population proportion, the group size may be compared to the optimal value of the MLE or that for the Bayes estimator. The direct Bayes method offers a good alternative to the experimenter, as he may have some general ideas about population proportion which can be used to obtain the value of the optimum k , which can be updated in subsequent testing and estimation. Once k is known, a prior on p may be transformed into a prior on p^* and then the alternative Bayes estimator or the indirect Bayes estimator may be used. This estimator is simple to calculate and hence, may be attractive to users. Choice of k may be guided by physical considerations or be chosen by minimizing MSE or be determined from the optimal k for the MLE. For practical use, a fixed k is desirable unless variability of the prior is very high.

6.3 Recommendation for Further Research

Based on the work undertaken in this thesis, the following work for further research is recommended.

- (a) Studying Estimation problem under the condition of unequal group sizes but with errors in decision.
- (b) Applying both the Maximum Likelihood and the Bayesian methods to the case of unequal probabilities .
- (c) Studying the above work using the least squares regression method.
- (d) Using the method of moments to study Estimation problem especially

the case of unequal group sizes, unequal probability and with errors in decision.

(e) Applying Bayesian approach to classification problem could be of interest.

Bibliography

- [1] Abramowitz, M. and Stegun, I.A. (1960): "Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. Washington, D.C.: Bureau of National Standards".

- [2] Bhattacharyya, G.K., Karandinos, M.G. and DeFoliart, G.R., (1979): "Point estimates and Confidence Intervals for Infection Rates Using Pooled Organisms in Epidemiologic Studies" *A.J Epidemiol* **109**: 124-131

- [3] Boswell, M.T. and Patil, G.P. (1987): "A perspective of composite Sampling". *Communication in Statistics-Theory and Methods* **16**: 3069-3093

- [4] Chao, L.C and Swallow, W.H (1990): "Using Group Testing to Estimate a Proportion and to Test the Binomial Model". *Biometrics* **46**:1035-1046

- [5] Chiang, C.L and Reeves, W.C (1962): "Statistical Estimation of virus in infection rates in mosquito vector population". *A. J Hygiene* **76**: 377-391
- [6] Davies, C.E, Grizzle, J.E. and Bryan, J.A. (1973): "Estimation of the Probability of Post Transfusion Hepatitis in Hemophilia Treatment". *Biometrics* **29**: 386-392
- [7] Dorfman, R (1943): "The Detection of Defective Members of a large Population". *Annals of Mathematical Statistics* **14**: 436-440
- [8] Gastwirth, J.L and Hammick, P. (1989):"Estimation of the Prevalence of a Rare Disease, Preserving the Anonymity of the Subjects by Group Testing: Application to Estimating the Prevalence of AIDS Antibodies in Blood Donors." *Journal of Statistical Planning and Inference* **22**: 15-27
- [9] Gibbs, A.J and Gower, J.C (1960):"The Use of a Multiple Transfer Method in Plant Virus Transmission Studies- Some Statistical Methods Arising in the Analysis of Results." *Annals of Applied Biology*, **48**: 75-83

- [10] Graff, L.E and Roelffs, R. (1972):"Group Testing in the Presence of Testing Error: An Extension of Dorfman Procedure." *Technometrics*, **14**: 113-122
- [11] Griffiths, D.A (1972):"A Further Note on the Probability of Disease Transmission." *Biometrics*, **28**: 1133-1139
- [12] Griffiths, D.A (1973):"Maximum Likelihood Estimation for the Beta-Binomial Distribution and An Application to the Household Distribution of the Total Number of a Disease." *Biometrics*, **29**: 637-648
- [13] Johnson, N.L., Kotz, S. and Wu, X.(1992):"Inspection of Errors for Attributes in Quality Control." *Monographs on Statistics and Applied Probability*, **44**, Chapman Hall
- [14] Kerr, J.D (1971):"The Probability of Disease Transmission." *Biometrics*, **27**: 219-222
- [15] Kline, R., Brothers, T., Bookmeyer, R., Zeger, S. and Quinn, T., (1989): "Evaluation of HIV Seroprevalence in Population Surveys Using Pooled Sera." *Journal of Clinical Microbiology* **27**: 1449-1452

- [16] Koukorinas, C. and Mylona, K.(2009): "Group Screening Method for the Statistical Analysis of $E(f_{NOD})$ -Optimal Mixed-Level Supersaturated Designs" *Statistical Methodology* **6**(4): 380-338
- [17] Le, C.T.(1981): "A New estimator for Infection Rates Using Pools of Variable Size" *A.J Epidemiol* **114**: 132-136
- [18] Lew, R.A. and Levy, P. (1989):" Estimation of Prevalence on the Basis of Screening Tests." *Statistics in Medicine* **8**: 1225-1230
- [19] Liu, P., Shi, Z.,Zhang, Y., Xu, Z., Shu, H. and Zhang, X. (1997):" A Prospective Study of a Serum-Pooling Strategy in Screening Blood Donors for Antibody to Hepatitis C Virus." *Transfusion* **37**: 732-736
- [20] Manene, M.M (1985):"Further Investigation of Group Screening Desgns- Stepwise Desings: Ph.D Thesis Submitted to the University of Nairobi."
- [21] Rao, C.R. (1973): "Linear Statistical Inference and Its Applications. 2nd Edition. New York: John Wiley and Sons: p. 426"

- [22] Skellam, J.G. (1948): "A probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials". *Journal of Royal Statistical Society Series B* **10**: 254-261
- [23] Sobel, M. and Groll, P.(1959):"Group-Testing to Eliminate Efficiently all defectives in a Binomial Sample". *The Bell System Journal* **38**: 1179-1557
- [24] Sobel, M. and Elashoff, R.M.(1975):"Group-Testing with a new Goal,Estimation". *Biometrika* **63**: 181-193
- [25] Swallow, H.W. (1985):"Group Testing for Estimating Infection Rates and Probabilities of Disease Transmission." *Am Phytopathological Society* **75**: 882-889
- [26] Swallow, H.W. (1987):"Relative Mean Squared Error and Cost Considerations in Choosing Group Size for Group Testing to Estimate Infection Rates and Probabilities of Disease Transmission".*Phytopathology* **77**: 1376-1381
- [27] Tebbs, J. M. and Bilder, C. R. (2003):"An Empirical Bayes Group Testing Approach To Estimating Small Proportions." *Communications*

- [28] Thompson, K.H.(1962):"Estimation of the proportion of vectors in a natural population of insects". *Biometrics* **18**: 568-578
- [29] Vine, A.E., Lewis, S.M., Dean, A.M. and Brunson, D. (2008):"A Critical Assesment of Two Stage Group Screening through Industrial Experimentation". *Technometrics* **50**: 15-25
- [30] Walter, S.D., Hildreth, S.W. and Beaty, B.J. (1980): "Estimation of Infection Rates in Populations of Organisms Using Pools of Variable Size" *A.J Epidemiol* **112**: 124-128
- [31] Watson, G.S. (1961):"A study of Group-Screening Method". *Technometrics* **3**: 371-388
- [32] Xin, M.T., Litvack, E. and Pagano, M. (1994):"Studies of AIDS and HIV Surveillance, Screening Tests: Can We Get More By Doing Less?" *Statistics in Medicine* **13**: 1905-1919
- [33] Xiang, F., Walter, W.S. and Shunpu, Z. (2007):"Improved Empirical Bayes Estimation in Group Testing Procedure for Small

Proportions." *Communication in Statistics-Theory and Methods* **36**:
2937-2944

- [34] Xin, M.T., Litvack, E. and Pagano, M. (1995): "On the Informativeness and Accuracy of Pooled Testing in Estimating Prevalence of a Rare disease: Application to HIV Screening." *Biometrika* **82**: 287-297