International Journal of Advanced Scientific Research
ISSN: 2456-0421; Impact Factor: RJIF 5.32
www.allscientificjournal.com
Volume 2; Issue 4; July 2017; Page No. 70-80

Towards a cost effective method for estimating disease infection rates in plants

John L Sirengo

Masinde Muliro University of Science and Technology, Kenya

Abstract
Pool-testing involves investigating more than one unit with a single test for cost effectiveness. Rather than testing an individual plant for the presence of a disease, plants are put into groups of more than one unit and subjected to a single test to determine the presence of a disease. Therefore, if the disease is absent a single test will be sufficient; otherwise, each individual plant is tested. The study used this procedure to estimate the infection rate of a disease. The method of maximum likelihood estimator for equal and unequal group sizes is discussed. For comparison with one-at-time testing procedure, properties of the constructed estimators have also been discussed. Empirical evidence via simulation of the optimal group sizes and asymptotic relative efficiency have been achieved by use of a statistical package. From the simulated results, it emerged that the proposed procedure has smaller asymptotic variance than one-at-a-time testing procedure. Therefore, the proposed procedure is more efficient than the usual one-at-a-time testing.

Keywords: cost effective method, estimating disease infection rates, plants

Introduction
Plant diseases are responsible for major losses in agricultural production throughout the world (Tebbs & Swallow, 2003; Otim-Nape *et al*., 2000) [13, 9]. Plants are used for food, fat, fibre, clothing, industrial products and medicine. The quality and quantity of these products are adversely affected by plant diseases. Although some people argue that the problem of plant disease is insignificant and has been over emphasized, many of these diseases are causing impediments to economic growth and development in many African societies (Thresh *et al*., 1998) [15].

Infection of Plant Diseases
Rational and most effective control of plant diseases is possible only if the disease is correctly diagnosed by identifying the nature of infection of the disease and the life cycle stages of the involved parasite, i.e., its mode of reproduction active structures produced under the favourable condition for rapid and wide dispersal and the structures produced to overcome adverse conditions. One must identify the general modes of reproduction of different plant parasites, viz. fungi, bacteria, viruses (not true parasites), and nematodes and the structures which enable these parasites to survive under most unfavourable conditions, like the off season of suitable host crop(s) and/or exceptional high temperature and dry weather conditions prevailing in the dry seasons in Kenya.
All parasitic as well as viral diseases are transmissible. The parasites transmit infectious pathogens to suitable host plants with ability to spread from host to host and from one area to another. The microscopic parasites or sub-microscopic infectious agents, namely viruses causing plant diseases, are technically referred to as 'pathogens'.
The transfer of pathogens from one host or place to another is referred to as 'dissemination' or 'dispersal'. Dissemination of plant diseases is analysed in relation to different stages of diseases as follows:

- Primary infection: Contact of a pathogen with a suitable host plant and initiation of the disease first time in the season of a crop/plant. Often, a few or several plants in the population are likely to get primarily infected.
- Secondary spread: When a plant or few plants are primarily infected, rapid multiplication of the pathogen sets in under favourable climatic conditions, which helps 'secondary spread' of the disease.

Although in majority of plant diseases the above two phases occur sequentially, primary infections occur only once in the season and secondary spread often recurs several times in the same season thus causing rapid infection over wide areas. However, in some cases there is no secondary spread in the same season, i.e. the pathogen multiplies only once during the crop period and the spread of disease is observed only during the following crop season by way of increased primary infection. Mildews, leaf spots, blasts, blights etc. are common examples of primary infection while some smut diseases in cereal or grain crops, where black powder is formed in place of grains, fall in the secondary infection. Control measures, therefore, need to be directed to avoid the primary infection and also to check the secondary spread.

Continuous and Discontinuous Infection
Infection of disease is termed as 'continuous' when it occurs naturally by way of growth, multiplication and spread of the pathogen in an area where the disease is established. At times, however, in an area where a particular disease has never occurred, it may get introduced through human agency carrying diseased material to a new locality or to a distant place for the purpose of introduction of new plants, crop varieties etc. Such infection, obviously, is unnatural and is regarded as 'discontinuous' infection. Different control measures can be enforced to check such discontinuous infection.

Direct and Indirect Infection
Based on the methods of disease infection and related suitable methods of disease control, infection can be classified into direct and indirect infection.

Direct Infection
This refers to disease infection where the pathogen is carried externally or internally on the seed or planting material like cuttings, sets, tubers, bulbs etc. False smut diseases as well as Helminthosporin Blight disease of wheat are common examples of fungal diseases carried internally through apparently healthy seeds. Ring rot and Brown rot of potato caused by bacteria are carried internally through the tubers. The well-known whip smut and red rot of sugarcane are fungal diseases carried internally in the planting sets. Mosaic and leaf rot of potato which are viral diseases are also carried inside the infected tubers.
On the other hand, in external mode of infection the pathogen is carried externally over the surface of a seed or vegetatively propagated plant parts like sets, tubers, bulbs etc. or may even be carried as a physical mixture of fungal structures with the seed. The common grain smut is an example of the former type while the fungal structures called 'sclerotia' having the size of a grain or slightly bigger in case of the Ergot disease are often likely to be transmitted in the form of physical mixture with the seed.

Indirect Infection
In this type of infection, the pathogen spreads itself by way of its persistent growth or certain structures of the pathogen carried independently by natural agencies like wind, water, animals, insects, mites, nematodes, birds etc. In indirect infection, plant diseases can be transmitted through any of the following modes:

Autonomous Infection
It takes place by continuous and persistent growth of the threads or 'hyphae' of the causal fungi in soil, characteristic of several wood rotting fungi attacking forest trees and some fruit plants. Some root rotting fungi infecting certain seasonal crops also are transmitted by this method. The autonomous dispersal of such soil fungi may range from a few centimetres to several meters in a single season. Some plant parasitic nematodes also exhibit active but limited mobility in the soil.

Wind Dispersal
Fungal spores produced externally on host surfaces are most easily carried by wind currents and this is the most dangerous mode of infection of plant pathogenic fungi like those causing powdery and downy mildews, leaf spots, blasts, blights and rust diseases. Extensive and severe epidemics of plant diseases are mostly the results of wind dispersal of the pathogens. Wind dissemination involves four stages relating to the spores viz. production of countless spores, their liberation in the wind currents, dispersal along with the wind and deposition on new susceptible host surfaces where they cause infection under favourable climatic conditions. Apart from spores, bits of fungal threads and nematode cysts are also amenable to wind infection in certain cases.

Water Dissemination
Disease infection through the agency of water is relatively less important compared to the wind infection. Splashing rain drops mostly transmit the foliar diseases from leaf to leaf, from shoot to shoot and even from plant to plant in case of closely spaced crops such as maize. Such infection is usually accompanied by wind dispersal as well. Plant pathogens requiring high humidity conditions like the fungi causing downy mildew diseases or bacteria causing canker of citrus are well adapted to this kind of short distance water dispersal. Certain soil inhabiting pathogenic fungi and bacteria causing root and collar rots, wilts, foot rots, etc. are likely to be transmitted to much longer distances through the agencies like irrigation water, streams and rivers.

Animals, Insects and Birds
Farm animals serve as disease transmitting agents in some cases. They are likely to carry the pathogen externally on their body surface, particularly on legs and hoofs, or internally through their intestinal tract. Commonly, the soil inhabiting fungi causing rots and wilts are carried externally while certain smut fungi causing diseases to grain crops are transmitted through the intestinal tract.
Most of the viral diseases of plants are transmitted through the agency of different insects such as white flies, mealy bugs, etc. Both types of insects viz. sucking and chewing or/biting are capable of transmitting viral diseases. The infection may be simply 'mechanical' or it may be 'biological'. In the latter case the specific insect and the specific viral pathogen have some kind of association or relationship between the two. Insects in such cases are called the 'vectors' for the particular viral pathogen. In case of mechanical infection the pathogen is simply carried externally or internally by the insect. Certain bacterial and several fungal pathogens are also known to be carried by insects.
Although birds play a very minor role in disease infection, in cases of dispersal of seeds of higher flowering parasite. Birds transmit loranthus parasites both externally and internally in certain trees such as mango trees.

Implements and Tools
Farm implements used for cultivation of soil are often likely to transmit plant pathogens from one place to another. The pathogens in this case are usually carried in the form of bits of plant disease debris lying in the soil. Similarly tools used for carrying out operations like cutting, pruning, budding, grafting and thinning, also help in the infection of certain diseases from plant to plant. Several viral diseases are disseminated through the budding and grafting operations.

Nematodes
Nematodes have been observed to transmit viral, bacterial and fungal plant diseases. Nematodes feed externally on the roots of a host plant causing injuries to roots which become the avenues for entrance of fungal and bacterial pathogens infecting plant roots. The fan-leaf virus of grapevine is a well-known example of infection through a species of nematodes.

Human Dispersal
Humans are often responsible for infection of plant diseases in two ways viz. The first is workers handling seedlings, other planting material or fruits are likely to get personally in contact with plant pathogens like fungi or bacteria. While handling the diseased material, a worker can unknowingly and indirectly transmit the pathogens to healthy seedlings or plant parts through contaminated hands.
The second mode of infection for which humans are responsible is the most efficient and equally dangerous phenomenon of infection of plant diseases between distant geographical areas often separated by physical barriers like lakes, valleys, mountains or deserts, etc. Such long distances infection of a disease to an area hitherto free from the disease is usually accomplished by the transport of infected seed, nursery stock or timber, etc. Therefore, infection happens through propagating material.
In the history of plant pathology this second mode of infection has often resulted in the transmission of some of the worst plant disease epidemics to new areas. Often, the local host plant stock, which is yet to adapt to the pathogen, remains highly susceptible. Fungal diseases such as the late blight of potatoes and downy mildew of grapes, bunchy top of bananas, which is a viral disease, and bacterial blight of paddy, are examples of some diseases introduced to Kenya from other countries. Practically, all the countries of the world have suffered from such introduction of new plant diseases through human agency.

Testing of Plant Disease
At the Kenya Agricultural Research Institute (KARI) laboratories, plant pathologists routinely provide scientific diagnostic results to resolve plant health issues. Preventative measures are often less expensive than curative approaches. Preventive measures impede disease development in plants as well as reducing or eliminating losses in production quality and quantity. Plant analyses encompass a wide variety of greenhouse vegetables, fruits, field crops

Plant Disease Diagnosis
Plant disease diagnosis depends on the nature of the causing agent. There are three major disease causing agents in plants, namely bacteria, fungi and viruses.

Bacteria
Bacteria enter plants through wounds or natural openings. Once inside, bacteria continue to multiply. Bacteria can, for instance, liquefy a potato tuber in just a few days. They can be detected by use of PCR methods.

Fungi
Fungi and fungi-like organisms that cause diseases on plants are a diverse. Because of their large numbers and diversity, identification of these pathogens is difficult, expensive and dependent on the expertise and experience of a plant pathologist. Plant Diagnostic Scan for fungal plant diseases is a quantitative method that indicates the level of fungal infection. With this technique, microorganisms are detected not only from plant tissues but also from water and soil samples, including turf grasses. The accurate and sensitive detection of pathogens obtained by the Diagnostic Scan allows growers to monitor disease-causing organ-isms before symptoms develop and yields become affected.

Viruses and Virus-like Agents
The Pest Diagnostic laboratory provides detection of viruses and virus-like agents in plants by ELISA (immunoassay) or PCR detection methods.
The process of testing and detecting disease pathogens in plants encompasses the collection of samples of plant tissues, leaves or tubers. The collected samples are then sent to laboratories for testing.

Statement of the Problem
To estimate disease infection rate in plants, samples of plants tissues are collected and send to laboratories for inspection. Inspection of plant tissues, one-at-a-time, is costly and requires enormous resources as the population size increase, i.e. $N \rightarrow \infty$ where N is plant population. This study proposes a cost effective method that addresses the issues raised.

Literature Review
Group Testing Procedure
The idea of sampling plant tissues together here in between is referred to as a group in order to fit into the pooling algorithm. This idea of group testing originated with Dorfman (1943) [3] during World War II as an economical method of testing blood specimen of army inductees in order to detect the presence of infection. The basic idea was to put the population into groups each of size k and perform tests on each group. Dorfman (1943) [3] has given the group size k depending on the known infection rate, which maximizes the expected number of items classified per test. The main benefit of group testing is that it reduces the expense and effort incurred compared to one-at-a-time testing. Dorfman (1943) [3] showed that if the prevalence rate of a disease is small then

group testing can lead to worthwhile savings, i.e. reduce the number of tests by about 80%. Dorfman (1943) [3] assumed that if H is the infection rate then,
$(1-H)$ = the probability of selecting at random an individual free from infection,
$(1-H)^k$ = the probability of obtaining by random selection of group size k individuals all of whom are free from infection,
$1-(1-H)^k$ = the probability of obtaining by random selection a group of k individuals that contains at least one individual infected,

$$\frac{N}{k} = the\ number\ of\ groups\ of\ size\ k\ constructed\ from\ a\ population\ of\ size\ N$$

The above formulation will be vital in our model derivation in subsequent work. Dorfman (1943) [3] showed that the expected number of tests denoted by E(G) obtained by grouping procedure is

$$E(X) = \frac{N}{K} + NH \qquad (1)$$

That is, the population size divided by group size when the infection rate is small, then group

testing can lead to worthwhile saving. Chiang and Reeves (1962) [2] recommended $k = \frac{\log\left(\frac{1}{2}\right)}{log(1-H)}$, which gives an equal chance of positive or negative results. Thompson (1962) [14] and Kerr (1971) [7] proposes a group size k as $k = \frac{1.6}{H}$ by minimizing the asymptotic variance of the maximum likelihood estimator H, but suggested a smaller k than this because of the potential bias in the estimate when there are not many groups when estimating the infection of the disease in a population. Hwang (1975) [3] describes Dorfman's procedure as a partition of units into any number of disjoint groups such that a group test is performed on each of them. Hwang (1975) [5] also prefers group testing when a population consist of *k* stochastically independent units where unit *i* has a probability of p_i of being defective, which is called a generalized binomial group test (GBGT) problem when $p_i = p$ for all units, then the generalized binomial group test problem reduces to a binary group testing problem considered under Dorfman procedure. He gives an efficient dynamic programming algorithm for obtaining an optimal Dorfman procedure for the generalized binomial problem with finite *k*. An optimal group testing procedure in this case implies a procedure which minimizes the expected number of tests. There is an upper bound on the size of group test which is incorporated into Dorfman procedure that can in fact reduce the amount of computation (Hwang, 1975) [5]. Hwang (1976) [6] calculates the expected cost of identifying each group size *k* which minimizes this cost.
Burrow (1986) [1] has come up with an improved estimation of disease transmission rates by group testing. He calculates an alternative estimate $\hat{p}^*$ to $\hat{p}$ given as

$$\hat{p}^* = 1 - \left(\frac{2kR + k - 1}{2nk + k - 1}\right)^{\frac{1}{k}} \qquad (2)$$

R refers to non infected groups and $\hat{p}$ is the estimator for estimating the infection rate where;

$$\hat{p} = 1 - \left(1 - \frac{x}{n}\right)^{\frac{1}{k}} \qquad (3)$$

Burrow (1986) [1] shows that bias and mean square error properties of Equation (2) are superior to those of Equation (3), except when $k = 1$ where both estimates are identical to the minimum variance unbiased estimate. He concludes that the use of $\hat{p}^*$ rather than $\hat{p}$ increases the efficiency of group testing and changes the choices of *k* in relation to infection when *N* is fixed.
Swallow (1987) [11] has obtained the less than optimal group size *k* used in group testing by using the relative mean square error in estimating the rate of disease infection in plants. He shows that if k^* less than optimal group size and k the optimal group size is then,

$$RMSE(K) = \frac{MSE\left(\widehat{P,K}\right)}{\left(\widehat{P,K^*}\right)} X100 \qquad (4)$$

Swallow (1987) [11] concludes that using a group size smaller than optimal is even more cost effective than one-at-a-time testing especially when the infection rate is low.
Tebbs *et al.*(2003) [13] emphasizes on group testing and shows how it can be done in an increasing order of probabilities in order to reduce the bias and mean square as compared to individual testing. The maximum likelihood estimator (MLE) of the proportion of infected units in a population using pools is upwardly biased estimator of the population proportion. Hepworth and Watson (2009) have investigated this bias of the MLE when testing groups of different sizes using fixed and sequential procedures. They observe that the possibility of obtaining all positive groups contributes substantially to the bias and by using

analytical method, i.e. the simple iterative technique. Hepworth and Watson (2009) have been able to correct the bias for fixed procedures satisfactorily but for the sequential procedures with their uneven bias pattern a numerical method which produces an almost unbiased estimator was proposed. Nyongesa (2012) has proposed hierarchical estimation that improves efficiency of the estimators.

Materials and Methods

Problem Formulation

Suppose we have a large number of plant tissues, say *N*, where N → ∞, we are interested in estimating the infection rate of a disease in these plants. For simplicity we denote infection rate by *H*. The obvious method is to test each of the *N* plant tissues and establish the infected ones. Since N → ∞, one-at-a-time testing is time consuming, costly, tedious and may result into poor estimation of *H*. A cost effective method would be to classify the *N* plants into groups and these groups might be of equal or unequal sizes. We shall investigate the problem for both equal and unequal sizes.

The idea here is to split *N* into *n* groups each of size *k* (equal sizes). Subject each group to testing. If the group is free of infection then it is dropped from further investigation. Suppose X groups are infected out of *n* constructed groups then for simplicity we shall assume that *X* is binomially distributed.

$$f(X|H) = \begin{cases} \binom{n}{x}(1-(1-H)^k)^x((1-H)^k)^{n-x} \\ 0, otherwise \end{cases} \quad ,x = 1,2,3,\dots\dots\dots,n \qquad (5)$$

Where *H* denotes disease infection rate, *n* the number of constructed groups and *x* number of infected groups. In the next section we demonstrate how Equation (5) is derived.

Model Derivation

To derive Equation (5) we shall require the following indicator function:

$$D_i = \begin{cases} 1, if\ the\ i^{th} group\ is\ infected, i = 0,1,2,\dots,n \\ 0\ otherwise \end{cases}$$

The probability that an i^{th} group is not infected is

$$pr(D=0) = \pi\pi\dots\pi = \pi^k \qquad (6)$$

by independence assumption and the probability that it is infected is

$$Pr(D=1) = 1 - Pr(D=0) = 1 - \pi^k \qquad (7)$$

by applying Equation (6). If X groups test positive, then X ~ B(n, $1-\pi^k$), as established in Equation (5).

Estimation of Infection Rate with Equal Group Size

The model that can be used to estimate the infection rate with equal group sizes is given by Equation (5).To obtain the maximum likelihood estimator of the infection rate we take the likelihood function of Equation (5) to obtain

$$L(H|X) = \binom{n}{x}(1-(1-H)^k)^x((1-H)^k)^{n-x}$$

The $\widehat{H}$ that maximizes $\ln l(H\backslash X)$ is obtained by solving

$$\frac{\partial}{\partial H}\log L(H|X) = \frac{\partial}{\partial H}(\log\binom{n}{x} + x\log(1-(1-H)^k + (n-x) \qquad (8)$$

by letting

$$\pi = 1 - H \qquad (9)$$

Substituting Equation (9) into Equation (8),we have

$$\frac{\partial}{\partial H}\log L(H|X) = \frac{\partial}{\partial \pi}\left(\log\binom{n}{x} + x\log(1-\pi^k\right) + (n-x)\log\pi^k) \qquad (10)$$

applying Equation (9) and simplifying we have

$$\widehat{H} = 1 - \left(1 - \frac{x}{n}\right)^{\frac{1}{k}} \quad (11)$$

Equation (11) yields the estimator of infection rate for a fixed group of size k. It is customary in inferential statistics to investigate the properties of the estimator.

Asymptotic Variance of the Estimator

For large number of plants, i.e. nk → ∞, Tebbs and Swallow (2003) [13] have shown that the estimator of infection rate is consistent and approximately normally distributed with asymptotic variance as

$$V(\widehat{H}) = E\left[-\frac{\partial^2}{\partial H^2} log(H \backslash X)\right]^{-1} \quad (12)$$

Applying Equation (12) on the model Equation (5) upon taking logs of the likelihood function and differentiating twice with respect to H, we have

$$Var(\widehat{H}) = \frac{1 - (1-H)^k}{nk^2(1-H)^{k-2}} \quad (13)$$

To demonstrate our computation we simulate the asymptotic variance of H versus k as provided in Tables 1, 2 and 3.

Table 1: Simulated asymptotic variance for various group sizes, k_i, and infection rate H = (0.01, 0.02, 0.03, 0.04) and n = 20

H			K		
	1	5	10	15	20
0.01	0.000495	0.000101	0.000052	0.000035	0.000027
0.02	0.000980	0.000204	0.000108	0.000076	0.000060
0.03	0.001455	0.000310	0.000168	0.000121	0.000099
0.04	0.001920	0.000417	0.000232	0.000173	0.000145

Table 2: Simulated asymptotic variance for various group sizes, k_i, and infection rate H = (0.01, 0.02, 0.03, 0.04) and n = 30

H			K		
	1	5	10	15	20
0.01	0.000330	0.000067	0.000035	0.000024	0.000018
0.02	0.000653	0.000136	0.000072	0.000050	0.000040
0.03	0.000970	0.000206	0.000112	0.000081	0.000066
0.04	0.001280	0.000278	0.000155	0.000115	0.000097

Table 3: Simulated asymptotic variance for various group sizes, k_i, and infection rate H = (0.01, 0.02, 0.03, 0.04) and n = 40

H			K		
	1	5	10	15	20
0.01	0.000248	0.000051	0.000026	0.000018	0.000014
0.02	0.000490	0.000102	0.000054	0.000038	0.000030
0.03	0.000728	0.000155	0.000084	0.000061	0.000049
0.04	0.000960	0.000209	0.000116	0.000086	0.000073

From the tables we observe that as *H* increases the asymptotic variance increases and, in fact, for n = 40, k = 5 and H = 0.01 the asymptotic variance is 0.000051 and when H = 0.02 the asymptotic variance is 0.000102.This implies with an increase in infection rate there are high chances of being detected. We also observe that as*k* increases the asymptotic variances decreases implying an increase in group size leads to increased efficiency of the estimator. This concurs with the findings of Swallow (1987) [11] who proposes group testing procedure when the infection rate is very small with a group size that minimizes the MSE.

Bias and Mean Square Error (MSE) of the Estimator

The bias of the estimator H measures the accuracy of the estimator and is defined as

$$Bias(\widehat{H}) = E(\widehat{H}) - H \quad (14)$$

To compute bias, we obtain the expected value of H.

$$E(\hat{H}) = 1 - E\left(1 - \frac{x}{n}\right)^{\frac{1}{k}}$$

Suppose k = 1

$$E(\hat{H}) = 1 - E\left(1 - \frac{x}{n}\right) = 1 - 1 + E\left(\frac{x}{n}\right) = \frac{nH}{H} = H \qquad (15)$$

Substituting Equation (15) into Equation (14), we obtain the bias as zero implying that for k = 1, H is unbiased estimator of H. We now derive the bias when k > 1, i.e.

$$E(\hat{H}) = E\left[1 - (1 - \frac{x}{n})^{\frac{1}{k}}\right] = 1 - E(1 - \frac{x}{n})^{\frac{1}{k}} \qquad (16)$$

expanding $(1 - \frac{x}{n})^{\frac{1}{k}}$ we get

$$1 + \left(\frac{1}{k}\right)\left(-\frac{x}{n}\right) + \frac{\left(\frac{1}{k}\right)\left(\frac{1}{k} - 1\right)}{2!}(-\frac{x}{n})^2 + \frac{\left(\frac{1}{k}\right)\left(\frac{1}{k} - 1\right)\left(\frac{1}{k} - 2\right)}{3!}(-\frac{x}{n})^3 + \cdots$$

$$= 1 - \frac{x}{nk} + \frac{1}{2n^2}\left(\frac{1}{k}\right)\left(\frac{1}{k} - 1\right)x^2 - \frac{1}{6n^3}\left(\frac{1}{k}\right)\left(\frac{1}{k} - 1\right)\left(\frac{1}{k} - 2\right)x^3 + \cdots \qquad (17)$$

substituting Equation (17) into Equation (16), we get

$$E(\hat{H}) = 1 - E\left[1 - \frac{x}{nk} + \frac{1}{2n^2}\left(\frac{1}{k}\right)\left(\frac{1}{k} - 1\right)x^2 - \frac{1}{6n^3}\left(\frac{1}{k}\right)\left(\frac{1}{k} - 1\right)\left(\frac{1}{k} - 2\right)x^3 + ..\right]$$

$$= 1 - 1 + \frac{1}{nk}E(x) + \frac{1}{2n^2}\left(\frac{1}{k}\right)\left(\frac{1}{k} - 1\right)E(x^2) + O(E(X^3))$$

Using the fact that $E(X) = n\pi$ and $E(X^2) = n(n-1)\pi^2 + n\pi$, we have

$$E(\hat{H}) = \frac{nH}{nk} + \frac{1}{2n^2}\left(\frac{1}{k}\right)\left(\frac{1}{k} - 1\right)(n(n-1)H^2 + nH) + O(E(X^3))$$

$$\frac{H}{k} - \left(\frac{H(1 - H + nH)}{2nk^2} - \frac{H(1 - H + nH)}{2nk}\right) + O(E(X^3))$$

$$Bias(\hat{H}) = \left(\frac{H}{k} - \frac{H(1 - H + nH)}{2nk^2} - \frac{H(1 - H + nH)}{2nk}\right) - H + O(E(X^3)) \qquad (18)$$

Suppose k = 1, Equation (18) gives zero as shown in the substitution of Equation (16) into Equation (15).

For k > 1

$$\left(\frac{H(2nk - 1 + H - nH + k - Hk + nHk - 2nk^2}{2nk^2}\right) \neq 0,$$

implying that when k>1 the estimator is biased. Therefore the bias component is

$$Bias(\hat{H}) = \frac{H}{k}\left((1 - k) + \frac{1 - H + nH}{2n} + \frac{H - nH - 1}{2nK} + O(E(X^3))\right) \qquad (19)$$

Now that we have obtained the bias, we investigate the mean square error (MSE). In statistical inference MSE incorporates measures of both accuracy (bias) and precision (variance) of the estimator (Swallow, 1987) [11]. Having found the variance and bias of the estimator H, MSE can be stated as

$$MSE(\hat{H}) = Var(\hat{H}) + (Bias(\hat{H}))^2 \qquad (20)$$

To demonstrate our computation we simulate the bias and MSE as Provided in Table 4.

Table 4: Simulated Bias and MSE for various group sizes, k_i, and with infection H = (0.01, 0.02, 0.03, 0.04) and n = 20

H			Bias				MSE			
			K					K		
	1	5	10	15	20	1	5	10	15	20
0.01	0.0000	−0.0080	−0.0090	−0.0094	−0.0095	0.0005	0.0002	0.00013	0.00012	0.0001
0.02	0.0000	−0.0161	−0.0181	−0.0187	−0.0190	0.0010	0.0005	0.00043	0.00043	0.0004
0.03	0.0000	−0.0242	−0.0271	−0.0281	−0.0286	0.0015	0.0009	0.00090	0.00091	0.0009
0.04	0.0000	−0.0323	−0.0362	−0.0374	−0.0381	0.0019	0.0015	0.00154	0.00158	0.0016

In Table 4 we have results of simulated bias and mean square error (MSE) for various group sizes. From the simulated results, we observe that bias of the estimator is zero when k = 1 and absolute bias increases with increase in group size and vice versa for mean square error (MSE) which reduces with increase in group size. We note that group of size 5 provides the minimal absolute asymptotic bias and the observation concurs with that of Swallow (1985) [10] who recommends that a relatively small group size should be used to obtain optimal results.

Estimation of Infection Rate Using Unequal Group Sizes

In most cases it is not easy to construct equal group sizes of plant tissues thus leading to unequal groups for experiments. To obtain maximum likelihood estimator with unequal group sizes, first the model of interest is

$$f(H|X) = \prod_{i=1}^{m} \binom{n_i}{x_i} (1-(1-H)^{k_i})^{x_i} ((1-H)^{k_i})^{n_i - x_i} \qquad (21)$$

Proceeding from the above, we obtain the maximum likelihood estimator from the model. Using Equation (9)and setting m = 1, we have

$$L(H|X) = Log\binom{n_i}{x_i} + x_i Log\pi^{k_i} + (n_i - x_i) Log(\pi)^{k_i} \qquad (22)$$

Differentiating Equation (22) with respect to π and equating to zero, we have

$$(k_i \pi^{k_i - 1}) \left(\frac{x_i \pi^{k_i} + n_i(1-\pi^{k_i}) - x_i(1-\pi^{k_i})}{(1-\pi^{k_i})\pi^{k_i}} \right) = 0,$$

which simplifies to

$$\pi^{2k_{i-1}} = \pi k_{i-1} - \frac{x_i}{n_i} \pi^{k_{i-1}} \qquad (23)$$

from Equation (9), we have

$$\widehat{H} = 1 - \left(1 - \frac{x_i}{n_i}\right)^{\frac{1}{ki}} \qquad (24)$$

Equation (24) provides the estimator of infection rate when unequal group sizes are used. Similarly in the next section we investigate the properties of Equation (24).

Bias and Mean Square Error (MSE) of the Estimator

In this section we derive the bias and MSE of our estimator (Equation 24). We need to compute the expected value of the estimator. Using Equation (24) the expected value is

$$E(\widehat{H}) = \frac{\pi}{k_i} - \left(\frac{\pi(1-\pi+n_i\pi)}{2n_i k_i^2} - \frac{\pi(1-\pi+n_i\pi)}{2n_i k_i} \right) + O(E(X^3)) \qquad (25)$$

Hence

$$Bias(\widehat{H}) = \left(\frac{\pi}{k} - \frac{\pi(1-\pi+n\pi)}{2nk^2} - \frac{\pi(1-\pi+n\pi}{2nk} \right) + O(E(X^3))$$

$$Bias = \frac{\pi}{k_i}\left((1-k_i) + \frac{1-\pi+n_i\pi}{2n_ik_i} + \frac{\pi-n_i\pi-1}{2n_i}\right) + O(E(X^3)) \qquad (26)$$

If we set $k_i = k$ and $n_i = n$, Equation (26) reduces to Equation (13), therefore Equation (26) is a generalization of Equation (13). Proceeding as in equal group size, the mean square error for the estimator when using unequal group sizes is

$$MSE(\hat{H}) = Var(\hat{H}) + Bias(\hat{H}))^2 \qquad (27)$$

In the next Section we investigate the asymptotic variance of Equation (24).

Asymptotic Variance

The asymptotic variance for the estimator when unequal group sizes are used is obtained by applying the Cramer Rao lower bound on Equation (22).The likelihood function of Equation (22) is

$$L(H|n,x) \propto (1-\pi^{k_i})^{x_i}(\pi^{k_i})^{n_i-x_i}$$
$$Var(\hat{H}) = \frac{1}{\sum_{i=1}^{m}\frac{n_ik_i^2(1-H)^{k_i-2}}{1-(1-H)^{k_i}}} \qquad (28)$$

This equation will be useful in the computation of asymptotic relative efficiency and the optimal group size.

Results and Discussion

In this section we find the optimal group size that reduces the number of tests/costs involved in carrying out investigations as just discussed. First, we consider the problem when equal group sizes are used.

Equal Group Size

As in the previous chapter we shall be interested in both equal and unequal group sizes. To obtain the optimal group size k (the group size that minimizes the variance of Equation (13), we solve for

$$\underset{k}{argmin}\left(\frac{1-(1-H)^k}{nk^2(1-H)^{k-2}}\right) \qquad (29)$$

For fixed n, minimizing (29) with respect to k we have

$$f(k) = Log(1-H)^k - 2(1-H)^k + 2. \qquad (30)$$

Equation (30) cannot be solved analytically, therefore we apply Newton Raphson method to obtain the optimal group size iteratively by

$$k_{i+1} = k_i - \frac{f(k)}{f^I(k)} \qquad (31)$$

where f(k) is given by Equation (30) while f'(k) is the derivative of f(k) with respect to k. The iteration in (31) is stopped when $|k_{i+1} - k_i| \leq \xi$, for an arbitrary ξ. Some computation of optimal k is provided in Table 5 for some selected disease infection rates.

Table 5: Simulated optimal Group Sizes, k_i, for some Selected Infection Rates (H)

H	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**
k	88.84	43.25	28.94	20.42	15.84	12.78	10.58	8.92	7.62
H	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
k	6.57	5.71	4.97	4.34	3.78	3.29	2.83	2.40	1.98

From the table above, we observe that as *H* increases k decreases and in fact when *H* is 0.9 the optimal *k* value is 1.98, when *H* is large prefer one-at-a-time testing. This concurs with the results of Dorfman (1943) who proposes pooling strategies and that the procedure is only visible when the infection rate is small. Some optimal *k* is plotted versus the infection rate *H* as provided in Figure 1 to summarize the observations. The optimal *k* is the value that minimizes the variance and cost.

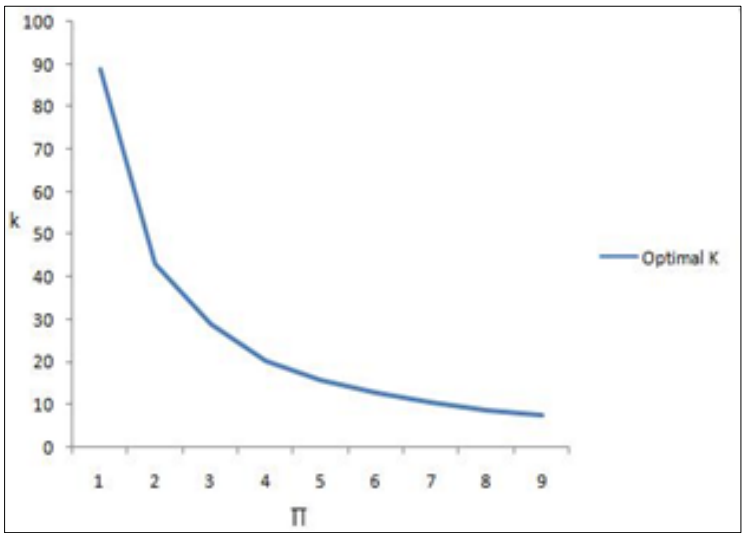

Fig 1: A plot of infection rate versus optimal *k*

From the figure one can observe that there is an inverse relationship between infection rate and the optimal group size *k*. With a low infection the optimal group size will be high as compared to a higher infection rate.

Unequal Group Sizes

To obtain the optimal group size k_i for unequal group sizes, we utilize Equation (23) by solving

$$\binom{argmin}{k}\left(\frac{1}{\sum_{i=1}^{m}\frac{n_i k_i^2(1-H)^{k_i-2}}{1-(1-H)^{k_i}}}\right) \qquad (32)$$

To solve for a k_i that minimizes Equation (32) is achieved via Newton Raphason Method.
By letting

$$f(k_i) = \left(\frac{1}{\sum_{i=1}^{m}\frac{n_i k_i^2(1-H)^{k_i-2}}{1-(1-H)^{k_i}}}\right) \qquad (33)$$

Taking k_i^* to be the initial guess value, k_{i^*+1} is obtained from

$$k_{i+1}^* = k_i^* - \frac{V(\hat{H})}{V'(\hat{H})} \qquad (34)$$

The iteration is stopped if $|k_{i^*+1} - k_i^*| < \varepsilon$. Equation (34) can be easily implemented via statistical software and values of optimal k_i^* generated.

Asymptotic Relative Efficiency (AREs)

To measure the efficiency of our estimators for both equal and unequal group sizes as com-pared to one-at-a-time testing procedure, we compute the AREs.

ARE with Equal Group Sizes

We shall measure the efficiency of our estimator relative to one-at-a-time testing procedure. The variance of one-at-a-time estimator i.e the variance of a Bernoulli distribution is obvious which is given as

$$Var(\hat{H}) = \left(\frac{1}{nk}\right)^2 H(1-H) \qquad (35)$$

With equations (13) and (35) at hand the ARE is given by

$$\frac{Var(\widehat{H_n})}{Var(\widehat{H_1})} = \frac{1-(1-H)^k}{(1-H)^{k-1}H} \qquad (36)$$

Simulation of Equation (36) for various H and group sizes is provided in Table 6.

Table 6: Simulated asymptotic relative efficiency for various equal group sizes, k_i, and infection rate H = (0.01, 0.02, 0.03, 0.04)

H		K		
	5	10	15	20
0.05	5.54	12.73	22.01	34
0.10	6.24	16.81	34.71	65.03
0.15	7.10	23.12	59.20	140.53
0.20	8.21	33.25	109.69	342.94
0.25	9.64	50.27	221.48	938.97
0.30	11.55	80.27	489.24	2923.98

Asymptotic relative efficiency (ARE) is greater than one for all generated values as shown in Table6, increase in group size results into increased efficiency of the estimator. Similar argument in this Section is applicable to unequal group sizes.

Conclusion and Recommendations

In this study we sought to construct the maximum likelihood estimators of disease infection rate in plants both for equal and unequal group sizes. The properties of the maximum likelihood estimator such as bias, mean square error and asymptotic variance are provided in the discussion. To justify the purpose of this study, we also discussed the estimator of unequal group size. We compared the efficiency of group testing and one-at-a-time testing by simulation of the bias and mean square error. It has been shown asymptotically that the maximum likelihood estimator> 1is efficient as compared to one-at-a-time. The constructed maximum likelihood estimator gives an estimator with smaller variance as compared to one-at-a-time testing. A similar conclusion is arrived at in case of unequal group size.

The value k_i (the number of units in a group) that is efficient and minimizes the asymptotic variance for equal group size has been constructed with various values of H. From the results of this study, it has been observed that group testing is viable in situations with low infection rate via Table 5 and Table 6 and this procedure lowers the cost of group testing since only the groups that have been classified as positive are retested. Similar conclusion is arrived at by use of statistical software in case of unequal group size which serves as a generalization for estimation of disease infection rate.

In this study we made several assumptions such as independence of the tests and plant tissues sampled together acting homogeneously. It could be interesting if one investigates the case when the assumptions are dropped.

References

1. Burrow PM. Improved estimation of pathogen transmission rates by group testing. Phytopathology. 1986; 77:363-365.
2. Chiang CL, Reeves WC. Statistical estimation of virus infection ratesin mosquito vector populations. Amer. J. Hygiene. 1962; 75:377-391.
3. Dorfman R. The detection of defective members of large populations. Annals of mathematical statistics. 1943; 14:436-440.
4. Hepworth G, Watson R. Debiased Estimation of Proportions in GroupTesting. Appl. Stats. 2009; 58:105-121.
5. Hwang FK. Generalized Binomial group Testing problem. Journal of the American statistical Association. 1975; 70:923-926.
6. Hwang FK, *Group testing with a dilution effect*. Biometrika. 1976; 63:611-613.
7. Kerr JD. The probability of disease transmission. Biometrics. 1971; 77:219 -222.
8. Nyongesa LK. Hierachical Estimation. Wiley Journal of Biometrical, 2012.
9. Otim-Nape GW, Abua A, Thresh JM, Ssemakula YGN, Acola G, Byabakama B, Martin JCA, Baguma Ogwal S. *The Current Pandemic of cassava mosaic virus disease in East Africa and its control*. NARO, NRI, DFID, Natural Resources Institute, Chatheem, UK. 2000, 100.
10. Swallow WH. Group testing for estimating infection rates and probabilities ofdisease transmission. Phytopathology. 1985; 75(8):568-578.
11. Swallow WH. Relative MSE and cost considerations in choosing group sizefor group testing to estimate infection rates and probability of disease infection. Phytopathology. 1987; 77:1376-1381.
12. Tebbs JM, Swallow WH. Estimating ordered binomial proportions with theuse of group testing. Biometrika. 2003; 90:471-477.
13. Thompson KH. Estimation of the proportion of vectors in a natural populationof insects. Biometrics. 1962; 18:568-578.
14. Thresh JM, Otim-Nape GW, Thankappan M, Muniyappa V. The mosaicdisease of cassava in Africa and India caused by Whitefly borne Gemini viruses. Reviewof plant Pathology. 1998; 77:935-945.