Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 69(1) 2015 27-47

Three-stage Optimal Sampling Plans for Group Testing Data

Osval A. Montesinos-López[1,2], Kent Eskridge[2], Abelardo Montesinos-López[3] and Jose Crossa[4]

[1]Facultad de Telemática, Universidad de Colima, Avenida Universidad 333, Col. Las Víboras, C.P. 28040 Colima, México.
[2]University of Nebraska, Statistics Department, Lincoln, Nebraska, USA.
[3]Departamento de Estadística, Centro de Investigación en Matemáticas (CIMAT), Guanajuato, México.
[4]Biometrics and StatisticsUnit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, Mexico, D.F., Mexico.

Received 19 August 2014; Revised 23 February 2015; Accepted 05 March 2015

SUMMARY

In surveys, sample size planning is important for achieving precise estimates at a low cost. However, this issue is not adequately addressed for group testing data obtained from a three-stage sampling process. In this study, we obtained the optimal allocation of localities (l), fields (m) and pools per field (g) in a three-stage group testing survey for a given pool size (s). These optimal values were obtained under the assumption of equal locality and field sizes. To handle the unequal sample size case, we derived the relative efficiency (RE) of unequal *versus* equal locality and field sizes to estimate the proportion. By multiplying the sample of localities and fields obtained assuming equal cluster size by the inverse of the corresponding REs, we adjusted the sample size required in the context of unequal localities and field sizes. We also show the adjustments needed for correctly allocating localities and fields in order to estimate the required budget and achieve a certain power or precision.

Keywords: Three-stage, Group testing, Optimal sample size, Relative efficiency, Power, Precision.

1. INTRODUCTION

Group testing, attributable to Dorfman (1943), was developed to make possible to perform the very large number of tests for venereal diseases required by the United States Defense Forces during World War II (Federer 1994). This technique is useful whenever a large population of individuals has to be subjected to the same test. The idea is to group (pool) objects and test pools instead of individuals. If a pool tests negative, then all subjects in the pool are declared free of infection. A pool tests positive if at least one individual within the pool is positive. Group testing is useful to estimate the proportion of a population that possesses a rare characteristic (Schliep *et al.* 2003).

The successful application of group testing for classification and for estimating prevalence has been documented. In classification, group testing has been used in drug detection, blood donation, detection of rare diseases in plants, animals and humans, science fiction, information theory, and identification of clones from a genomic library for a particular gene (Wolf 1985, Dodd *et al.* 2002, Remlinger *et al.* 2006, Verstraeten *et al.* 1998, Bilder 2009 and Schliep *et al.* 2003). For estimating prevalence, it has been used to estimate the proportion of transgenic maize and the prevalence of certain diseases in animals and humans (West Nile virus, foot and mouth disease, HIV, syphilis, chlamydia and gonorrhea, among others) (Tebbs and Bilder 2004, Peck 2006, Hernández-Suárez *et al.* 2008, Yamamura

Corresponding author: Jose Crossa
E-mail address: j.crossa@cgiar.org

and Hino 2007, Montesinos-López *et al.* 2010, 2011). To determine sample size when estimating prevalence (π), most research assumes that a simple random sample (SRS) of individuals was taken, even though it is a well-known fact that SRS is often not a cost-efficient strategy. However, surveys conducted to estimate the proportion of transgenic maize in Mexico have had a multilevel structure, where: (1) researchers take a sample of localities (primary sampling units=PSU) of the frame of localities; (2) in each locality they take a sample of fields from the frame of fields (secondary sampling unit=SSU); (3) in each field they take a sample of plants (elementary sampling units); (4) from the sample of plants per field, they form pools of size (s); and (5) a diagnostic test is performed on each pool (Piñeyro-Nelson *et al.* 2009).

This data-collecting approach (multistage sampling) is common in large national databases because it is the most cost-efficient sampling design when the population of interest consists of subpopulations, also called clusters, that are used for selection. In practice, multistage samples are preferred because the interviewing or testing costs are greatly reduced if the individuals are geographically or organizationally grouped. Such sample designs reflect the organization of the natural and social worlds. Also, multistage surveys do not require a list of all elementary units, since the sample is selected in stages, often taking into account the hierarchical (nested) structure of the population. However, multistage sampling design leads to dependent observations, and failing to deal with this properly in the statistical design and analysis may lead to erroneous inferences.

In practice, unequal locality and field sizes are the rule, and it is not possible to specify correctly the distribution of locality and field sizes at the beginning of a study. Therefore, first we obtained optimal sample sizes for a three-stage group testing survey assuming equal locality and field sizes, *i.e.*, we used average locality and field sizes. Optimal sample sizes were derived using a mixed logistic group testing model and a first-order marginal quasi-likelihood (MQL) approach, where we assumed clusters were randomly sampled from a large number of clusters. To compensate for the loss of efficiency due to varying locality and field sizes, we derived the relative efficiency (RE) to adjust the optimal sample sizes. The required sample size for varying locality and field sizes can be obtained by multiplying the required sample size for an average cluster (locality or field) size by the inverse of the RE (Ahn *et al.* 2012).

In this article, section 2 presents the random logistic model used for individual testing. Section 3 describes the random logistic model used for group testing. Section 4 provides an approximate marginal variance of the proportion. In section 5, we derive sample sizes without constraints. Section 6 gives the optimal samples (l, m, g) given a pool size (s) under constraints and assuming equal cluster size, while section 7 provides tables for sample size determination assuming equal cluster sizes. Section 8 gives adjustments for unequal locality and field sizes. Section 9 gives an example for estimating the proportion of transgenic plants, and the discussion and conclusions are presented in section 10**.**

2. RANDOM LOGISTIC MODEL FOR INDIVIDUAL TESTING

In the context of individual testing, the standard random logistic model is obtained by conditioning on all fixed and random effects. The responses y_{ijk} are independent and Bernoulli distributed with probabilities π_{ij}, assuming that these probabilities are not related to any covariable (Moerbeek *et al.* 2001a). Thus the linear predictor using a logit link is equal to

$$\begin{aligned}\eta_{ij} &= logit\ (\pi_{ij}) = \ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)\\ &= \beta_0 + a_i + b_{ij} \qquad (1)\end{aligned}$$

where η_{ij} is the linear predictor that is formed from a fixed part (β_0) and two random parts (a_i and b_{ij}), where a_i~$N(0,\ \sigma_a^2)$ and $b_{ij} \sim N(0,\ \sigma_b^2)$. We also assume that the random components are mutually independent. Therefore, Equation (1) can be written in terms of the probability of a positive individual as:

$$\begin{aligned}\pi_{ij} &= \pi_{ij}(\beta_0,\ \sigma_a,\ \sigma_b)\\ &= [1 + \exp\{-(\beta_0 + a_i + b_{ij})\}]^{-1} \qquad (2)\end{aligned}$$

The mixed logit model for binary responses can be written as the probability π_{ij} plus a level 1 residual denoted e_{ijk}:

$$y_{ijk} = \pi_{ij} + e_{ijk} \qquad (3)$$

where e_{ijk} has zero mean and variance $V(y_{ijk}|a_i, b_{ij}) = \pi_{ij}(1-\pi_{ij})$ (Goldstein 1991-2003, Goldstein and Rabash 1996, Breslow and Clayton 1993, Rodriguez and Goldman 1995, Candy 2000, Moerbeek and Van Breukelen 2001a, Candel and Van Breukelen 2010, Skrondal and Rebe-Hesketh 2007). This model is widely used for estimating optimal sample sizes since the variance components are assumed to be known (Goldstein 1991-2003, Rodriguez and Goldman, 1995, Candy 2000, Moerbeek and Van Breukelen 2001a, b).

3. RANDOM LOGISTIC MODEL FOR GROUP TESTING

Suppose that, within field j, each plant in the *ith* locality is randomly assigned to one of the g_{ij} pools. Let $y_{ijkr} = 1$ if the *kth* plant in the *rth* pool in the *jth* field in locality i is positive and $y_{ijkr} = 0$ otherwise, for $i = 1, 2, ..., l$, $j = 1, 2, ..., m_i$, $r = 1, 2, ..., g_{ij}$ and $k = 1, 2, ..., s_{ijr}$ as the pool size. Note that y_{ijkr} is not observed, except when the pool size is 1. Let $Z_{ijr} = I_{(\sum_{k=1}^{s_{ijr}} y_{ijkr}>0)}$ be the indicator variable, whether the *rth* pool inside the field j in locality i is positive ($Z_{ijr} = 1$) or negative ($Z_{ijr} = 0$). Thus we only observe the random binary variable Z_{ijr} that takes the value of $Z_{ijr} = 1$ if the *rth* pool in field j and locality i tests positive, and $Z_{ijr} = 0$ otherwise. Conditional on the random effects [a_i and b_{ij}], pools within field j and locality i are independent. Therefore, the probability that the *rth* pool in field j and locality i is positive is given as

$$\pi_{ij}^p = P(Z_{ijr} = 1 | a_i, b_{ij}) = S_e + (1 - S_e - S_p)\prod_{k=1}^{s_{ijr}}(1-\pi_{ijk})$$

where S_e and S_p denote the sensitivity and specificity of the diagnostic test, respectively. S_e and S_p are assumed constant and close to 1 (Chen *et al.* 2009). For simplicity, in planning the required sample size we will assume an equal pool size, s, in all fields. This is a reasonable assumption for sample size determination, and π_{ij}^p is reduced to:

$$\pi_{ij}^p = P(Z_{ijr} = 1 | a_i, b_{ij}) = S_e + \varphi(1-\pi_{ij})^s \qquad (4)$$

where $\varphi = (1 - S_e - S_p)$. The mixed group testing logit model for binary responses can be written as the probability π_{ij}^p, plus a level 1 residual denoted e_{ijr}^p:

$$Z_{ijr} = \pi_{ij}^p + e_{ijr}^p \qquad (5)$$

where π_{ij}^p, as given in Equation (4) and e_{ijl}^p, has zero mean and variance $V(Z_{ijr}|a_i, b_{ij}) = \pi_{ij}^p(1-\pi_{ij}^p)$. Now let $\theta = (\beta_0, \sigma_a, \sigma_b)$ denote the vector of all estimable parameters. The multilevel likelihood is calculated for each level of nesting. First, the full conditional likelihood ignoring the constant for pool r in field j and locality i, is given by:

$$L_{ijr}(\theta | a_i, b_{ij}) = [\pi_{ij}^p]^{Z_{ijr}}[1-\pi_{ij}^p]^{1-Z_{ijr}} \qquad (6)$$

By multiplying the conditional likelihood (Equation 6) by the density of a_i and b_{ij}, and integrating out the random effects, we get the marginal (unconditional) likelihood.

$$L(\theta | y) = \prod_{i=1}^{l}\{\int[\prod_{j=1}^{m_i}\int(\prod_{r=1}^{g_{ij}} L_{ijr}(\theta | a_i, b_{ij})$$

$$f(b_{ij})db_{ij})f(a_i)da_i]\},$$

where $f(a_i)$ is the density function of a_i and $d(b_{ij})$ is the density function of b_{ij}. Unfortunately, this unconditional likelihood is intractable. There are various ways of approximating the marginal likelihood function. Two of them are: (1) to use integral approximations such as the Gaussian quadrature; and (2) to linearize the non-linear part using Taylor series expansion (TSE) (Moerbeek *et al.* 2001a, Breslow and Clayton 1993). The marginal form of the generalized linear mixed model (GLMM) is of interest here, since it expresses the variance as a function of the marginal mean.

4. APPROXIMATE MARGINAL VARIANCE OF THE PROPORTION

The marginal model can be fitted by integrating the random effects out of the log-likelihood and maximizing the resulting marginal log-likelihood or, alternatively, by using an approximate method based on TSE (Breslow and Clayton 1993). Next, π_{ij}^p is approximated using a first-order TSE around $a_i = 0$ and $b_{ij} = 0$, as

$$\pi_{ij}^p \approx \pi_{ij}^p|_{a_i=b_{ij}=0} + \left.\frac{\partial\pi_{ij}^p}{a_i}\right|_{a_i=b_{ij}=0}(a_i-0) + \left.\frac{\partial\pi_{ij}^p}{b_{ij}}\right|_{a_i=b_{ij}=0}(b_{ij}-0)$$

$$(7)$$

$$\pi_{ij}^p \approx \pi_{ij}^p \,|_{a_i=b_{ij}=0} + s\varphi(1-\pi_{ij})^{s-1}\pi_{ij}(1-\pi_{ij})\,|_{a_i=b_{ij}=0}\,(a_i) + s\varphi(1-\pi_{ij})^{s-1}\pi_{ij}(1-\pi_{ij})\,|_{a_i=b_{ij}=0}\,(b_{ij})$$

$$\pi_{ij}^p \approx \tilde{\pi}^p + s\varphi(1-\tilde{\pi})^{s-1}\tilde{\pi}(1-\tilde{\pi})a_i + s\varphi(1-\tilde{\pi})^{s-1}\tilde{\pi}(1-\tilde{\pi})b_{ij} \qquad (7)$$

where $\tilde{\pi}^p \approx \pi_{ij}^p\,|_{a_i=b_{ij}=0} = Se + \varphi(1-[1+\exp(-\beta_0)]^{-1})^s$ and $\tilde{\pi} \approx \pi_{ij}\,|_{a_i=b_{ij}=0} = [1+\exp(-\beta_0)]^{-1}$, since a_i and b_{ij} are independent and identically distributed (iid) and we use the fact that

$$\frac{\partial \pi_{ij}^p}{a_i} = \frac{\partial \pi_{ij}^p}{\pi_{ij}}\frac{\partial \pi_{ij}}{\partial a_i}, \frac{\partial \pi_{ij}^p}{b_{ij}} = \frac{\partial \pi_{ij}^p}{\pi_{ij}}\frac{\partial \pi_{ij}}{\partial b_{ij}},$$

$$\frac{\partial \pi_{ij}}{\partial a_i} = \frac{\partial \pi_{ij}}{\partial b_{ij}} = \frac{\partial \pi_{ij}}{\partial \eta_{ij}} = \pi_{ij}\,(1-\pi_{ij}) \text{ and}$$

$$\frac{\partial \pi_{ij}^p}{\pi_{ij}} = s\varphi(1-\pi_{ij})^{s-1}$$

Now, by substituting Equation (7) in Equation (5) we can approximate Equation (5) by

$$Z_{ijr} \approx \tilde{\pi}^p + s\varphi(1-\tilde{\pi})^{s-1}\tilde{\pi}(1-\tilde{\pi})a_i + s\varphi(1-\tilde{\pi})^{s-1}\tilde{\pi}(1-\tilde{\pi})b_{ij} + e_{ijr}^p \qquad (8)$$

Therefore, the approximate marginal variance based on a first-order TSE of the responses of a pool is equal to:

$$Var(Z_{ijr}) \approx \{s\varphi(1-\tilde{\pi})^{s-1}\}^2\{\tilde{\pi}(1-\tilde{\pi})\}^2[\sigma_a^2 + \sigma_b^2] + \tilde{\pi}^p(1-\tilde{\pi}^p)$$

where the variance of e_{ijr}^p was approximated by $\tilde{\pi}^p(1-\tilde{\pi}^p)$. Note that $\bar{Z} = \frac{\sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{r}^{g} Z_{ijr}}{lmg}$ is a moment estimator of $E(\pi_{ij}^p)$ and its variance is equal to:

$$Var(\bar{Z}) \approx \frac{\{s\varphi(1-\tilde{\pi})^{s-1}\}^2\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l} + \frac{\{s\varphi(1-\tilde{\pi})^{s-1}\}^2\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm} + \frac{\tilde{\pi}^p(1-\tilde{\pi}^p)}{lmg} \qquad (9)$$

Recall that we will select a sample of *l* localities and *m* fields, assuming that the same number of pools per field will be obtained, *i.e.*, $g = \bar{g}$. Since the probability of success is not a constant over trials but varies systematically from field to field, the parameter π_{ij} is a random variable with a probability distribution. Therefore, it is reasonable to work with the expected value of π_{ij} across localities and fields to determine sample size. To approximate $E(\pi_{ij})$, we take advantage of the relationship between $\bar{Z}$ and $E(\pi_{ij}^p)$:

$$\bar{Z} = E(\pi_{ij}^p) = E(S_e + \varphi(1-\pi_{ij})^s) = E(S_e) + E(\varphi(1-\pi_{ij})^s) = S_e + \varphi E(K) \qquad (10)$$

where $K = (1-\pi_{ij})^s$. Using a first-order TSE around $a_i = 0$ and $b_{ij} = 0$, we can approximate K as

$$K \approx K\,|_{a_i=b_{ij}=0} + \left.\frac{\partial K}{a_i}\right|_{a_i=b_{ij}=0}(a_i - 0) + \left.\frac{\partial K}{b_{ij}}\right|_{a_i=b_{ij}=0}(b_{ij} - 0)$$

$$K \approx \tilde{K} + s(1-\tilde{\pi})^{s-1}\tilde{\pi}(1-\tilde{\pi})a_i + s(1-\tilde{\pi})^{s-1}\tilde{\pi}(1-\tilde{\pi})b_{ij} \qquad (11)$$

where $\tilde{K} \approx K\,|_{a_i=b_{ij}=0} = (1-[1+\exp(-\beta_0)]^{-1})^s = (1-\tilde{\pi})^s$ and we use the fact that

$$\frac{\partial K}{a_i} = \frac{\partial K}{\pi_{ij}}\frac{\partial \pi_{ij}}{\partial a_i}, \frac{\partial K}{b_{ij}} = \frac{\partial K}{\pi_{ij}}\frac{\partial \pi_{ij}}{\partial b_{ij}},$$

$$\frac{\partial \pi_{ij}}{a_i} = \frac{\partial \pi_{ij}}{\partial b_{ij}} = \frac{\partial \pi_{ij}}{\partial \eta_{ij}} = \pi_{ij}(1-\pi_{ij}) \text{ and}$$

$$\frac{\partial K}{\pi_{ij}} = s(1-\pi_{ij})^{s-1}$$

Then

$$E(K) \approx \tilde{K}$$

But doing TSE of first order we obtain that $(1 - E(\pi_{ij}))^s \approx (1-\tilde{\pi})^s = \tilde{K}$ and so

$$E(K) \approx (1 - E(\pi_{ij}))^s$$

That is, we approximate $E(K) = E[(1-\pi_{ij})^s]$ by $[1 - E(\pi_{ij})]^s$. This implies that $E(\pi_{ij}^p) \approx S_e + \varphi(1-E(\pi_{ij}))^s$, and since $\bar{Z}$ is an estimator for $E(\pi_{ij}^p)$, then an estimator for $E(\pi_{ij})$ can be obtained from

$$S_e + \varphi(1 - E(\pi_{ij}))^s \approx \bar{Z}$$

Therefore, an estimator for $E(\pi_{ij})$ is

$$E\widehat{(\pi_{ij})} \approx 1 - \left(\frac{S_e - E\widehat{(\pi_{ij}^p)}}{\varphi}\right)^{1/s} = 1 - \left(\frac{S_e - \bar{Z}}{\varphi}\right)^{\frac{1}{s}}$$

The variance of this estimator, $E\widehat{(\pi_{ij})}$, can be approximated from the variance of $\bar{Z}$ (Equation 9) with a first-order TSE around $E(\pi_{ij}^p)$ of the function $g(z) = \left(\frac{s_e - z}{\varphi}\right)^{\frac{1}{s}}$. After some algebra we get:

$$V(E(\widehat{\pi_{ij}})) \approx \left(\frac{\partial g(z)}{\partial z}\bigg|_{z=E(\pi_{ij}^{p})} \right)^2 Var(\bar{Z})$$

where $\frac{\partial g(z)}{\partial z} = \frac{1}{s}\left(\frac{s_e - z}{\varphi}\right)^{\frac{1}{s}-1} \frac{1}{\varphi} = \frac{1}{s\varphi(1-\tilde{\pi})^{s-1}}$. However, since $E(\pi_{ij}^{p})$ doesn't have a close exact form, we replace this with $\tilde{\pi}^{p}$ and obtain

$$V(E(\widehat{\pi_{ij}})) = V(\hat{\pi}) \approx \frac{\sigma_a^{2*}}{l} + \frac{\sigma_b^{2*}}{lm} + \frac{V(\delta)}{lmg} \qquad (12)$$

where σ_a^{2*} = $\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2$, $\sigma_b^{2*} = \{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$,

$V(\delta) = \frac{(s_e - \tilde{\pi}^{p})^{\frac{2}{s}-2}\tilde{\pi}^{p}(1-\tilde{\pi}^{p})}{s^2(\varphi)^{2/s}}$ and $\tilde{\pi}^{p} = S_e + \varphi(1-\tilde{\pi})^{s}$.

5. SAMPLE SIZES WITHOUT CONSTRAINTS

5.1 How to obtain a Certain Width of the Confidence Interval

Assume the researcher is interested in choosing the number of localities, given the number of fields (*m*), pools per fields (*g*) and pool size (*s*), to obtain a specified width (ω) of the confidence interval (CI) of the proportion. Assuming that the distribution of $\hat{\pi}$ is approximately normal with a mean $\hat{\pi}$ and a fixed variance $V(\hat{\pi})$, then the (1 – α)100% Wald confidence interval of $\hat{\pi}$ is given by $\hat{\pi} \mp Z_{1-\alpha/2}\sqrt{V(\hat{\pi})}$, where $Z_{1-\alpha/2}$ is the quantile 1 – α/2 of the standard normal distribution. Therefore, the observed width of the CI is equal to $W = 2Z_{1-\alpha/2}\sqrt{V(\hat{\pi})}$. The quantity $2Z_{1-\alpha/2}\sqrt{V(\hat{\pi})}$ (added and subtracted from the observed proportion, $\hat{\pi}$) in the CI is defined as *W*/2 (where *W* is the full width of the CI; *W* or *W/2* can be set *a priori* by the researcher depending on the desired precision). Therefore, if the researcher wants a CI width of ω for the full width, we can obtain the required number of localities, *l*, by making $2Z_{1-\alpha/2}\sqrt{\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm} + \frac{V(\delta)}{lmg}}$ = ω and solving for *l*. The required number of localities, *l*, is equal to:

$$l = \frac{4Z_{1-\alpha/2}^2}{\omega^2}[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m} + \frac{V(\delta)}{mg}] \qquad (13)$$

Recall that Equation (13) is useful for obtaining the required number of localities given a number of fields, pools per field and a pool size. However, this sample size (Equation 13) is not optimal.

5.2 How to obtain a Certain Power

Assume we are interested in testing $H_0 : \tilde{\pi} = \tilde{\pi}_0 vs\ H_1 : \tilde{\pi} > \tilde{\pi}_0$. For example, the European Union (Anonymous 2003) requires that the proportion of genetically modified (GM) seed impurities in a seed lot be lower than 0.005. Here, given a number of fields (*m*), pools per field (*g*) and pool size (*s*), we want to reach a power of (1 – γ) with significance level α, when $\delta = |\tilde{\pi}_1 - \tilde{\pi}_0|$. To perform a test with a type I error rate α and a type II error rate γ, the following must hold:

$$Z_{1-\alpha} = (\hat{\pi} - \tilde{\pi}_0)/\sqrt{V(\tilde{\pi}_0)} \text{ and } Z_{1-\gamma} = (\hat{\pi} - \tilde{\pi}_1)/\sqrt{V(\tilde{\pi}_0)}$$

Here $V(\tilde{\pi}_0)$ is the variance of $\hat{\pi}$ but using the value of the null hypothesis. Both $Z_{1-\alpha}$ and $Z_{1-\gamma}$ have a standard normal distribution since the variance components are assumed known. According to Cochran (1977) and Moerbeek *et al.* (2000), this results in the relation:

$$V2 = \frac{|\delta|^2}{(Z_{1-\alpha} + Z_{1-\gamma})^2}, \qquad (14)$$

If we change the alternative hypothesis to $H_1 : \tilde{\pi} < \tilde{\pi}_0$, Equation (14) is still valid, but not if the alternative is $H_1 : \tilde{\pi} \neq \tilde{\pi}_0$, in which case $Z_{1-\alpha}$ needs to be replaced by $Z_{1-\alpha/2}$ in Equation (14).

Then, given a certain number of fields, pools per field and pool size, which is the required number of localities, *l*, needed to achieve a power level (1 – γ) for a desired δ? To obtain the required *l*, we need to solve for *l* from $\left[\frac{\{\tilde{\pi}_0(1-\tilde{\pi}_0)\}^2\sigma_a^2}{l} + \frac{\{\tilde{\pi}_0(1-\tilde{\pi}_0)\}^2\sigma_b^2}{lm} + \frac{V(\delta_0)}{lmg}\right] = \frac{|\delta|^2}{(Z_{1-\alpha} + Z_{1-\gamma})^2}$. Therefore, by solving for *l*, the required number of localities (*l*) is equal to:

$$l = \frac{(Z_{1-\alpha} + Z_{1-\gamma})^2}{|\delta|^2} \times \left[\{\tilde{\pi}_0(1-\tilde{\pi}_0)\}^2\sigma_a^2 + \frac{\{\tilde{\pi}_0(1-\tilde{\pi}_0)]^2\sigma_b^2}{m} + \frac{V(\delta_0)}{mg}\right] (15)$$

Equation (15) gives the required number of localities given the number of fields (*m*), number of pools per field (*g*) and pool size (*s*), but this is not optimal.

6. OPTIMAL SAMPLE SIZES UNDER CONSTRAINTS FOR A GIVEN POOL SIZE

6.1 Minimizing Variance Subject to a Budget Constraint

Assume that we have a fixed sampling budget for estimating the average population proportion $\tilde{\pi}$. The question of interest is: what is the optimal allocation of localities (*l*), fields (*m*) and pools per field (*g*), given a pool size (*s*), for estimating the proportion with minimum variance, subject to the following budget constraint?

$$C = lmgsc_1 + lmgc_2 + lmc_3 + lc_4 \quad (c_u > 0,\ l, m, g, s \geq 2,\ u = 1, 2, 3, 4) \qquad (16)$$

where C is the total sampling budget available, c_1 is the cost of sampling and measuring a plant in an already sampled field, c_2 is the cost of testing a pool of size (*s*), c_3 is the cost of sampling and measuring a field, and c_4 is the cost of sampling and measuring a locality. The values of c_3 and c_4 are average values, since at the time the survey is being planned, it is not known which localities and fields per locality will be sampled and the travel times are not the same for all localities and fields per locality. The budget C and costs c_1, c_2, c_3 and c_4 are given in dollars but can be changed to any other currency. Optimal allocation of the units can be performed using Lagrange multipliers. By combining Equations (12) and (16), we obtain the Lagrangean

$$L(l, m, g, \lambda) = L = V(\hat{\pi}) + \lambda[C - (lmgsc_1 + lmgc_2 + lmc_3 + lc_4)] \qquad (17)$$

where $V(\hat{\pi})$ given by Equation (12) is the objective function that will be minimized with respect to *l*, *m* and *g*, given a pool size (*s*) subject to the constraint given in Equation (16), and λ is the Lagrange multiplier. The partial derivatives of Equation (17) with respect to λ, *l*, *m* and *g* are

$$\frac{\partial L}{\partial \lambda} = C - (lmgsc_1 + lmgc_2 + lmc_3 + lc_4) = 0;$$

$$\text{then } l = \frac{C}{mgsc_1 + mgc_2 + mc_3 + c_4}$$

$$\frac{\partial L}{\partial g} = -\frac{V(\delta)}{g^2ml} - \lambda lm(sc_1 + c_2) = 0;$$

$$\text{then } \lambda = -\frac{V(\delta)}{g^2m^2l^2(sc_1 + c_2)}$$

$$\frac{\partial L}{\partial m} = -\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm^2} - \frac{V(\delta)}{m^2gl} - \lambda[lgsc_1 + lgc_2 + lc_3] = 0$$

$$\frac{\partial L}{\partial l} = -\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l^2} - \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{ml^2} - \frac{V(\delta)}{l^2mg}$$

$$-\lambda[mgsc_1 + mgc_2 + mc_3 + c_4] = 0$$

By solving these equations, we obtain the optimal values for *l*, *m* and *g* (see Appendix A):

$$l = \frac{C}{mgsc_1 + mgc_2 + mc_3 + c_4}, \text{ where } m = \sqrt{\frac{c_4}{c_3}\frac{\sigma_b}{\sigma a}},$$

$$g = \sqrt{\frac{c_3}{(sc_1 + c_2)}}\frac{\sqrt{V(\delta)}}{\{\tilde{\pi}(1-\tilde{\pi})\}\sigma_b} \qquad (18)$$

First, we calculate the number of pools per field, *g*, and then we can calculate the number of fields per locality, *m*. Using these values, we calculate the number of localities to be sampled, *l*. To maintain the hierarchical structure, *g*, *m* and *l* should be rounded out to the nearest integer. They are assumed equal to 2 if *g*, *m* and *l* are less than 2. Note that Equation (18) is a generalization of the optimal sample size for continuous data in three-level sampling given by Cochran (1977).

The best practice to calculate the optimal values of *g*, *m*, and *l* is to provide reasonable estimates of the costs (C, c_1, c_2, c_3 and c_4) pool size (s), $\tilde{\pi}, S_e$, S_p, variance components (σ_a^2, σ_b^2). However, we know that this is not always easy thus for this reason we encourage the researchers to perform a pilot study with the same structure (three stages with pooled data) to be able to compute reasonable estimates. Also, we encourage to perform a literature review and examine similar studies so that getting appropriate values of the required parameters to be able to estimate the optimal values of *g*, *m*, and *l*. For choosing the cost values (c_1, c_2, c_3 and c_4) we believe that when planning a study with this structure, the researchers need to investigate carefully these costs since they can change drastically from one region to another depending of the socio economic and geographical aspects of the region. Also, if the researcher does not have experience in this type of studies she/he need to consult experts that can help her/him to define the cost values to choose. In summary, a combination of performing a pilot study, with an appropriate literature review and consulting with experts are the factors to consider for choosing the most reasonable parameters for calculating the optimal values of *g*, *m*, and *l*.

6.2 Minimizing the Budget to obtain a Certain Width of the Confidence Interval

So far, the allocation of units minimizing $V(\hat{\pi})$ has been derived under the condition that budget sampling and measuring is fixed to a certain value. However, many times the researcher wants the optimal allocation of units to minimize the sampling and measuring budget in order to obtain a specified width (ω) of the confidence interval (CI) of the proportion. The solution to this optimization problem is the same as minimizing the total budget subject to a variance constraint. The variance constraint is obtained from the (1 – α)100% Wald confidence interval of $\hat{\pi}(\hat{\pi} \mp Z_{1-\frac{\alpha}{2}}\sqrt{V(\hat{\pi})})$ (given in section 5.1). Since the width of the CI is equal to $W = 2Z_{1-\alpha/2}\sqrt{V(\hat{\pi})},$ and since we specified the required width of the CI to be ω, this implies that $V(\hat{\pi}) = \omega^2 / 4Z^2_{1-\alpha/2}$. Therefore, the optimization problem is to minimize the sampling budget as given in Equation (16) under the condition that $V(\hat{\pi}) = \omega^2 / 4Z^2_{1-\alpha/2}$ is fixed. That is, we want to minimize $C = lmgsc_1 + lmgc_2 + lmc_3 + lc_4$ subject to $V(\hat{\pi}) = V_0$. Again, using Lagrange multipliers, the corresponding Lagrangean is: $L(l, m, g, \lambda) = L = lmgsc_1 + lmgc_2 + lmc_3 + lc_4 + \lambda[V(\hat{\pi}) - V_0]$. Now the partial derivatives of L with respect to λ, l, m and g are:

$$\frac{\partial L}{\partial \lambda} = \frac{[\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm} + \frac{V(\delta)}{lmg} - V_0 = 0;$$

then $l = [\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m} + \frac{V(\delta)}{mg}]/V_0$

$$\frac{\partial L}{\partial g} = lm(sc_1 + c_2) - \lambda\frac{V(\delta)}{lmg^2} = 0; \text{ then}$$

$$\lambda = \frac{l^2 g^2 m^2 (sc_1 + c_2)}{V(\delta)}$$

$$\frac{\partial L}{\partial m} = lg(sc_1 + c_2) + lc_3 - \frac{\lambda}{lm^2}[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}] = 0$$

$$\frac{\partial L}{\partial l} = mg(sc_1 + c_2) + mc_3 + c_4 - \lambda$$

$$\times \left[\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l^2} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{ml^2} + \frac{V(\delta)}{l^2 mg}\right] = 0$$

By solving these equations, we have that the optimal values are (see Appendix B):

$$l = [\{\tilde{\pi}(1-\tilde{\pi})\}]^2\sigma_a^2 + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m} + \frac{V(\delta)}{mg}]/V_0, \text{ where}$$

$$m = \sqrt{\frac{c_4}{c_3}\frac{\sigma_b}{\sigma_a}},\ g = \sqrt{\frac{c_3}{(sc_1 + c_2)}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b} \tag{19}$$

Note that the number of fields per locality, *m*, and the number of pools per field, *g*, required when we minimize the budget subject to $V(\hat{\pi}) = V_0$ (Equation 19), are the same as when minimizing $V(\hat{\pi})$ subject to a budget constraint (Equation 18). However, the expression for obtaining the required number of localities, *l*, is different. In this case, the value of $V_0 = \omega^2 / 4Z^2_{1-\alpha/2}$ is substituted into Equation (19) and the expression for the required number of localities is

$$l = \frac{4Z^2_{1-\alpha/2}}{\omega^2}\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m} + \frac{V(\delta)}{mg}\right].$$

Another way of obtaining the same solution to this problem is given in Appendix C.

6.3 Minimizing the Budget to obtain a Certain Power

Assume a threshold is defined *a priori*, and our main interest is to test $H_0: \tilde{\pi} = \tilde{\pi}_0 \ vs \ H_1: \tilde{\pi} > \tilde{\pi}_0$. We want to determine a sampling plan (*i.e.*, *l*, *m* and *g* given a pool size) for minimizing the budget required for this test to have a specified power (1 – γ) and significance level α, when $\delta = |\tilde{\pi}_1 - \tilde{\pi}_0|$. Again, $V(\widehat{\pi_0})$ is a fixed quantity and equal to Equation (14), since we want to minimize the total budget to obtain a specified power (1 – γ). Therefore, we want to minimize $C = lmgsc_1 + lmgc_2 + lmc_2 + lc_4$ subject to $V(\hat{\pi}) = V_2$. The optimal allocation of fields and pools per field is also given in Equation (19) but using

Equation (14) to get V_0. Here, $V(\delta_0)$ $= \frac{(Se-\tilde{\pi}_0^p)^{\frac{2}{s}-2}\tilde{\pi}_0^p(1-\tilde{\pi}_0^p)}{s^2(Se+Sp-1)^{2/s}}$ is used in place of $V(\delta)$, and $\tilde{\pi}_0$ in place of $\tilde{\pi}$. This implies that

$$l = \frac{(Z_{1-\alpha}+Z_{1-\gamma})^2}{|\delta|^2}$$

$$\left[\{\tilde{\pi}_0(1-\tilde{\pi}_0)\}^2\sigma_a^2 + \frac{\{\tilde{\pi}_0(1-\tilde{\pi}_0)\}^2\sigma_b^2}{m} + \frac{V(\delta_0)}{mg}\right],$$

where $m = \sqrt{\frac{c_4}{c_3}\frac{\sigma_b}{\sigma_a}}$, $g = \sqrt{\frac{c_3}{sc_1+c_2}}\frac{\sqrt{V(\delta_0)}}{\tilde{\pi}_0(1-\tilde{\pi}_0)\sigma_b}$.

7. TABLES FOR DETERMINING THE OPTIMAL SAMPLE SIZE ASSUMING EQUAL LOCALITY AND FIELD SIZES

This section contains Tables 1 and 2, which help to calculate the optimal sample size assuming equal locality and field sizes given a pool size (s). These two tables can be used to minimize the variance given a budget constraint or to minimize the budget given a variance constraint. For example, assume that we want optimal values of l, m and g to minimize the $V(\hat{\pi})$ given a pool size (s = 20) and a budget constraint equal to C = 20000. Also assume that after a literature review, we estimate c_1 = 10, c_2 = 35, c_3 = 400, c_4 = 1200, $\sigma_a^2 = 0.25$, $\sigma_b^2 = 0.15$, S_e = S_p = 0.95, and $\tilde{\pi} = 0.01$.

This implies that c_2/c_1 = 3.5, c_3/c_1 = 40, $\frac{c_4}{c_1} = 120$. Then, looking at Table 1 in the intersection between the value of $\tilde{\pi} = 0.01$ (column four) and the value of $\sigma_b^2 = 0.15$ (second column) for S_e = S_p = 0.95 and s = 20, we get the required optimal values of fields per locality (m = 2, column 3) and pools per field (g = 9, column 4). Finally, using the optimal values of m and g and the cost, we can calculate the optimal number of localities as:

$$l = \frac{C}{mgsc_1 + mgc_2 + mc_3 + c_4}$$

$$= \frac{20000}{(2)(9)(20)(10)+(2)(9)35+(2)(400)+1200}$$

$$= 3.21 \approx 4$$

Table 1. Optimal sample sizes (m, g) given a pool size (s = 10, 20) for group testing in three stages for five values of σ_b^2.

			$\tilde{\pi}$									
	σ_b^2	m	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
			S_e = S_p = 0.90									
	0.05	2	41	25	20	17	16	15	14	14	14	13
	0.1	2	29	18	14	12	11	11	10	10	10	9
s =10	0.15	2	23	15	12	10	9	9	8	8	8	8
	0.2	2	20	13	10	9	8	7	7	7	7	7
	0.25	2	18	11	9	8	7	7	6	6	6	6
	0.05	2	19	13	11	10	10	10	10	10	11	12
	0.1	2	14	9	8	7	7	7	7	7	8	8
s = 20	0.15	2	11	8	6	6	6	6	6	6	6	7
	0.2	2	10	7	6	5	5	5	5	5	6	6
	0.25	2	9	6	5	5	4	4	5	5	5	5
			S_e = S_p = 0.95									
	0.05	2	32	21	17	15	14	13	13	12	12	12
	0.1	2	23	15	12	11	10	9	9	9	8	8
s = 10	0.15	2	19	12	10	9	8	8	7	7	7	7
	0.2	2	16	11	9	8	7	7	6	6	6	6
	0.25	2	15	10	8	7	6	6	6	5	5	5
	0.05	2	16	12	10	9	9	9	9	9	9	10
	0.1	2	11	8	7	6	6	6	6	6	7	7
s = 20	0.15	2	9	7	6	5	5	5	5	5	5	6
	0.2	2	8	6	5	5	4	4	4	4	5	5
	0.25	2	7	5	4	4	4	4	4	4	4	4
			S_e = S_p = 0.98									
	0.05	2	28	19	16	14	13	12	12	11	11	11
	0.1	2	20	14	11	10	9	9	8	8	8	8
s = 10	0.15	2	16	11	9	8	8	7	7	7	6	6
	0.2	2	14	10	8	7	7	6	6	6	6	6
	0.25	2	12	9	7	6	6	6	5	5	5	5
	0.05	2	15	11	9	9	8	8	8	8	8	9
	0.1	2	10	8	7	6	6	6	6	6	6	6
s = 20	0.15	2	8	6	5	5	5	5	5	5	5	5
	0.2	2	7	5	5	4	4	4	4	4	4	4
	0.25	2	7	5	4	4	4	4	4	4	4	4

Ten values of the proportion $(\tilde{\pi})$ and two values of $\sigma_a^2 = 0.25$, 0.5, $\frac{c_2}{c_1}$ = 3.5, $\frac{c_3}{c_1}$ = 40, $\frac{c_4}{c_1}$ = 120 and three combinations of S_e and S_p.

Table 1 also can be used to calculate the optimal allocations of l, m and g given a pool size (s) to minimize the budget, C, subject to a variance constraint

Table 2. Optimal sample sizes (*m, g*) given a pool size (s = 10, 20*)* for group testing in three stages for five values of $\frac{c_3}{c_1}$.

			$\tilde{\pi}$									
	c_3/c_1	m	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
			$S_e = S_p = 0.90$									
	15	2	12	8	6	5	5	5	4	4	4	4
	25	2	16	10	8	7	6	6	6	5	5	5
s = 10	35	2	19	12	9	8	7	7	7	6	6	6
	45	2	22	13	11	9	8	8	8	7	7	7
	55	2	24	15	12	10	9	9	8	8	8	8
	15	2	6	4	3	3	3	3	3	3	3	4
	25	2	8	5	4	4	4	4	4	4	4	5
s = 20	35	2	9	6	5	5	5	5	5	5	5	6
	45	2	10	7	6	5	5	5	5	6	6	6
	55	2	11	8	7	6	6	6	6	6	7	7
			$S_e = S_p = 0.95$									
	15	2	10	7	5	5	4	4	4	4	4	4
	25	2	13	8	7	6	6	5	5	5	5	5
s = 10	35	2	15	10	8	7	7	6	6	6	6	6
	45	2	17	11	9	8	7	7	7	6	6	6
	55	2	19	13	10	9	8	8	7	7	7	7
	15	2	5	4	3	3	3	3	3	3	3	3
	25	2	6	5	4	4	3	3	3	4	4	4
s = 20	35	2	8	5	5	4	4	4	4	4	4	5
	45	2	9	6	5	5	5	5	5	5	5	5
	55	2	10	7	6	5	5	5	5	5	5	6
			$S_e = S_p = 0.98$									
	15	2	9	6	5	4	4	4	4	4	3	3
	25	2	11	8	6	6	5	5	5	5	4	4
s = 10	35	2	13	9	8	7	6	6	6	5	5	5
	45	2	15	10	9	8	7	7	6	6	6	6
	55	2	16	11	9	8	8	7	7	7	7	6
	15	2	5	3	3	3	2	2	2	2	3	3
	25	2	6	4	4	3	3	3	3	3	3	3
s = 20	35	2	7	5	4	4	4	4	4	4	4	4
	45	2	8	6	5	5	4	4	4	4	4	5
	55	2	9	6	5	5	5	5	5	5	5	5

Ten values of the proportion $(\tilde{\pi})$. $\sigma_b^2 = 0.2$, $\sigma_a^2 = 0.5$, three combinations of S_e and S_p, $\frac{c_2}{c_1} = 3.5$, and $\frac{c_4}{c_1} = 25, 55, 85, 115, 145$.

equal to $V_0 = \omega^2 / 4Z^2_{1-\alpha/2}$. To get a CI width of $\omega = 0.015$ with 95%, then $Z_{1-\alpha/2} = 1.96$, which implies that $V_0 = \frac{(0.015)^2}{4(1.96^2)} = 0.00001464$. Assuming the same values of c_1, c_2, c_3, c_4, σ_a^2, σ_b^2, S_e, S_p, $\tilde{\pi}$ and $s = 20$, then the optimal values of fields per locality and pools per field are 2 and 9, respectively (using Table 1 exactly as above). However, now the optimal number of localities is calculated as:

$$l = \left[\{\tilde{\pi}(1-\tilde{\pi})\}^2 \sigma_a^2 + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2 \sigma_b^2}{m} + \frac{0.0007603}{mg} \right] / V_0$$

$$= \left[\{0.01(1-0.01)\}^2 (0.25) + \frac{\{0.01(1-0.01)\}^2 (0.15)}{2} \right.$$

$$\left. + \frac{V(\delta)}{(2)(9)} \right] / 0.00001464 = 6.23 \approx 7$$

If we wish to calculate the optimal values for a given power, we can also use Table 1 and the last formula to calculate the required number of localities but using the V_0 calculated with Equation (14). Table 2 should be used exactly as Table 1, assuming a given pool size; the only difference is that now there are five options for the ratio $\frac{c_3}{c_1}, \frac{c_4}{c_1}$ and only one value of $\sigma_b^2 = 0.20$.

8. ADJUSTING FOR UNEQUAL LOCALITY AND FIELD SIZES

In practice, unequal cluster sizes (localities and fields) are the rule. Cluster size variation increases bias and produces a considerable loss of power and precision in the parameter estimates. For this reason, we will calculate the relative efficiency of unequal *versus* equal cluster sizes (localities and fields) for adjusting the optimal sample size derived under the assumption of equal localities and field sizes. The definition of the relative efficiency of equal *versus* unequal cluster sizes is:

$$RE(\hat{\pi}) = \frac{Var(\hat{\pi} \mid \varsigma \text{ equal})}{Var(\hat{\pi} \mid \varsigma \text{ unequal})} \tag{20}$$

where $Var(\hat{\pi} \mid \varsigma \text{ equal})$ denotes the variance of the proportion estimator given a design with equal cluster sizes, and $Var(\hat{\pi} \mid \varsigma \text{ unequal})$ denotes a similar value for an unequal cluster size design, but with the same number of localities (*l*), fields (*m*) and the same total number of pools $\left(N=\sum_{i=1}^{l}\sum_{j=1}^{m_i} g_{ij}\right)$ as in the equal cluster size design. Given that it is possible to have variability in locality and field sizes, one $RE(\hat{\pi})$ will be calculated for localities $(RE(\hat{\pi})_l)$ and another for fields $(RE(\hat{\pi})_f)$ to incorporate both variabilities. To derive the $RE(\hat{\pi})_l$, we assume that only localities have variations in size, whereas to derive the $RE(\hat{\pi})_f$, we assume that only fields have size variation. If we wish to calculate the total relative efficiency, it should be equal to $RE(\hat{\pi})=RE(\hat{\pi})_l \times RE(\hat{\pi})_f$

Assuming only localities have different sizes, and using Equation (20), $RE(\hat{\pi})_l$ is equal to:

$$RE(\hat{\pi})_l = \frac{\sigma_a^{2*}+\dfrac{\sigma_b^{2*}}{\bar{m}}+\dfrac{V(\delta)}{\bar{m}\bar{g}}/l}{\sum_{i=1}^{l}\left[\sigma_a^{2*}+\dfrac{\sigma_b^{2*}}{m_i}+\dfrac{V(\delta)}{m_i\bar{g}}\right]/l^2}$$

$$=\frac{\left(\sigma_a^{2*}+\dfrac{\sigma_c^{2*}}{\bar{m}}\right)}{\sum_{i=1}^{l}\left[\sigma_a^{2*}+\dfrac{\sigma_c^{2*}}{m_i}\right]/l} = \frac{\bar{m}+\alpha_l}{\bar{m}}\frac{1}{l}\sum_{i=1}^{l}\left[\frac{m_i}{m_i+\alpha_l}\right] \quad (21)$$

where, $\sigma_a^{2*}=\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2$, $\sigma_b^{2*}=\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$, σ_c^{2*} $=\sigma_b^{2*}+\dfrac{V(\delta)}{\bar{g}}$ and $\alpha_l=\sigma_c^{2*}/\sigma_a^{2*}$.

Now, assuming unequal field sizes in locality *i*, the relative efficiency of locality *i* (RE_i) is:

$$RE_i = \frac{\left(\sigma_b^{2*}+\dfrac{V(\delta)}{\bar{g}}\right)/m_i}{\sum_{j=1}^{m_i}\left[\sigma_b^{2*}+\dfrac{V(\delta)}{g_j}\right]/m_i^2} = \frac{\left(\sigma_b^{2*}+\dfrac{V(\delta)}{\bar{g}}\right)}{\sum_{j=1}^{m_i}\left[\sigma_b^{2*}+\dfrac{V(\delta)}{g_j}\right]/m_i}$$

$$=\frac{\bar{g}+\alpha_f}{\bar{g}}\frac{1}{m_i}\sum_{j=1}^{m_i}\left[\frac{g_j}{g_j+\alpha_f}\right]$$

where

$$\alpha_f = V(\delta)/\sigma_b^{2*}.$$

Therefore, the average relative efficiency of the *l* localities, assuming unequal field sizes, is equal to

$$RE(\hat{\pi})_f = \frac{\bar{g}+\alpha_f}{\bar{g}}\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\left[\frac{g_j}{g_j+\alpha_f}\right] \quad (22)$$

Note that Equations (21) and (22) can be considered moment estimators of $\dfrac{\bar{m}+\alpha_l}{\bar{m}}E\left[\dfrac{U_l}{U_l+\alpha_l}\right]$ and $\dfrac{\bar{g}+\alpha_f}{\bar{g}}E\left[\dfrac{U_f}{g_f+\alpha_f}\right]$, respectively, assuming that locality and field sizes (m_i, i = 1, 2, ..., l; g_j, j = 1, 2, ... $\bar{m}$) are realizations of two random variables U_l (with mean μ_l and standard deviation σ_l) and U_f (with mean μ_f and standard deviation σ_f), respectively. Equations (21) and (22) are therefore equal to the equations derived for obtaining the RE of equal *versus* unequal cluster sizes in cluster randomized and multicenter trials given by Van Breukelen *et al.* (2007) for recovering the loss of power when estimating treatment effects using a linear model. Here we use RE to repair the loss of power or precision when estimating the proportion using a random logistic model for group testing. Defining $\lambda_k = (\mu_k/(\mu_k+\alpha_k))$ and the coefficient of variation of the random variable U_k by CV_K, = σ_k/μ_k, with k = *l*, *f*, then the $RE(\hat{\pi})_l$ and $RE(\hat{\pi})_f$ were expressed as the relative efficiency derived by Van Breukelen *et al.* (2007, pp. 2601-2602; see Appendix D), and a second-order Taylor series approximation of Equations (21) and (22) can be obtained. For localities, this is equal to

$$RE(\hat{\pi})_{lt} \approx \{1-CV_l^2\lambda_l(1-\lambda_l)\} \quad (23)$$

And for fields, it is equal to:

$$RE(\hat{\pi})_{ft} \approx \{1-CV_f^2\lambda_f(1-\lambda_f)\} \quad (24)$$

This is possible because the expectation part of the moment estimators of Equations (21) and (22) is equal to $E\left(\dfrac{U_k}{U_k+\alpha_k}\right)\approx\lambda_k\{1-CV_k^2\lambda_k(1-\lambda_k)\}$ with k = *l*, *f*. $RE(\hat{\pi})_{lt}$ and $RE(\hat{\pi})_{ft}$ do not depend on the number of localities and fields, respectively, but rather on the distribution of cluster sizes (mean and variance of localities and fields) and intraclass correlations. Note that CV_f = 0 means that the fields are equal and $RE(\hat{\pi})_{ft}=1$ means that only the sample of localities should be adjusted. Also, to correct for the loss of efficiency due to the assumption of equal locality sizes, one simply divides the number of localities (*l*) by the expected RE resulting from Equation (23). The adjustment for unequal field sizes is the same, but using Equation (24). For practical purposes, we will denote $RE(\hat{\pi})_{lt} = RE_{lt}$ and $RE(\hat{\pi})_{ft} = RE_{ft}$.

9. A NUMERICAL EXAMPLE FOR ESTIMATING THE PRESENCE OF TRANSGENIC MAIZE

In 2004, a study was conducted to detect the presence of genetically modified maize plants in farmers' fields in two localities (7 and 11) in the Sierra Juárez region of the Mexican state of Oaxaca (see Table 3). Thirty fields were sampled in each locality, and 300 leaves were collected from plants randomly chosen throughout each field. Six pools of 50 leaves each were formed from the 300 leaves. DNA was extracted from each pool sample and the presence of CaMV-35S sequences was determined by polymerase chain reaction (PCR) (see Table 3) (Pineyro-Nelson *et al.*, 2009).

Table 3. Number of pools comprised of leaf samples from two localities in Oaxaca, Mexico (2004). With a positive 35S PCR band based on 30 fields per locality and 6 pools per field, each composed of 50 maize leaves.

Locality	Field	Positive pools	Locality	Field	Positive pools
7	6	1	11	7	1
	8	1		17	1
	11	1		19	1
	15	1			
	17	2			
	25	1			
	27	1			
	30	3			

Assuming that we wish to conduct another study in this region of Oaxaca, we can estimate the parameters $(\tilde{\pi}, \sigma_a^2, \sigma_b^2)$ required for calculating the optimal sample size using the information given in Table 3, since the authors only reported the total number of positive fields and pools per locality. Assuming a non-informative sampling process, we estimated the parameters $(\tilde{\pi}, \sigma_a^2, \sigma_b^2)$ by fitting model (1) to these data but taking into account the pooled data (see in Appendix E the Glimmix code used in the estimation process). The resulting estimates were $\tilde{\pi} = 0.0024$, $\sigma_a^2 = 0.57$ and $\sigma_b^2 = 0.77$. After performing a literature review, we decided to use $S_e = 0.999$, $S_p = 0.997$ and $C = 20{,}000$ (total budget for the study); $c_1 = 10$ is the cost of enrolling the plants in the study, $c_2 = 35$ is the cost of each diagnostic test, $c_3 = 300$ cost of enrolling a field in the study, and $c_4 = 500$ the cost of enrolling a locality in the study. Next we will illustrate how we obtained the optimal sample sizes.

For minimizing the variance. Once again, we assume a pool size of $s = 50$. Then $\tilde{\pi}^p = 0.999 + (1 - 0.999 - 0.997)(1 - 0.0024)^{50}$ $= 0.115755$ and $V(\delta) = \dfrac{(Se - \tilde{\pi}^p)^{\frac{2}{s}-2}\tilde{\pi}^p(1-\tilde{\pi}^p)}{s^2(S_e + S_p - 1)^{2/s}}$

$$= \frac{(0.999 - 0.115755)^{\frac{2}{50}-2}(0.115755)(1-0.115755)}{50^2(0.999+0.997-1)^{2/50}} =$$

0.0 000522. Therefore

$$g = \sqrt{\frac{c_3}{(sc_1 + c_2)}\frac{\sqrt{V(\delta)}}{\pi(1-\pi)\sigma_b}}$$

$$= \sqrt{\frac{300}{((50)(10)+35)}\frac{\sqrt{0.0000522}}{0.0024(1-0.0024)(0.8775)}}$$

$$= 2.576 \approx 3$$

$$m = \sqrt{\frac{c_4}{c_3}\frac{\sigma_b}{\sigma_a}} = \sqrt{\frac{500}{300}\frac{0.8775}{0.755}} = 1.50 \approx 2,$$

$$l = \frac{C}{mgsc_1 + mgc_2 + mc_3 + c_4}$$

$$= \frac{20000}{(2)(3)(50)(10) + (2)(3)(35) + (2)(300) + 500}$$

$$= 4.64 \approx 5.$$

This means that we need to select five localities at random from the population of localities, two fields at random from each selected locality, and three pools per selected field. Thus the total number of plants that will be used in the study should be $l \times m \times g \times s = 5 \times 2 \times 3 \times 50 = 1500$ plants, *i.e.*, 150 plants per selected field.

Now, if locality and field sizes are unequal, how do we compensate for the loss of efficiency due to varying cluster sizes? Assume the mean and standard deviation of locality and field sizes are μ_k = 177 and σ_k = 81.5, respectively, with k = l, f. Then $CV_l = \frac{81.5}{177}$ = 0.4605, α_l =

$$\frac{\sigma_b^{2*}+\frac{V(\delta)}{\bar{g}}}{\sigma_a^{2*}} = \frac{\{0.0024(1-0.0024)\}^2(0.77)+\frac{0.0000522}{177}}{\{0.0024(1-0.0024)\}^2 0.57},$$

= 1.44, so λ_l = (177/(177 + 1.44)) = 0.9919. Therefore RE_{lt} = {1 – (0.4605^2)(0.9919)(1 – 0.9919)} = 0.9983. Thus efficiency for localities can be restored by taking $l = \frac{4.64}{0.9983} = 4.6479 \approx 5$ localities. Now, for fields, the $CV_f = \frac{81.5}{177}$ = 0.4605 α_f = $\frac{V(\delta)}{\sigma_b^{2*}}$ =

$\frac{0.0000522}{\{0.0024(1-0.0024)\}^2 0.77} = 11.83$, so λ_f = (177/(177 + 11.83) = 0.9374. Therefore RE_{ft} = {1 – (0.4605^2) (0.9374)(1–0.9374)} = 0.9876. Thus efficiency for fields can be restored by taking m = $\frac{1.5}{0.9876}$ = 1.519 ≈ 2 fields.

For a desired CI width. Now suppose that the researcher requires a 95% confidence interval estimate, with a desired width for the proportion of transgenic plants that is equal to W = $(\tilde{\pi}_U - \tilde{\pi}_L) \le \omega = 0.005$. Therefore, $Z_{1-0.05/2}$ = 1.96 and V_0 = $\omega^2/4Z^2_{1-\alpha/2}$ = $\frac{0.005^2}{4*1.96^2}$ = 0.000001627, assuming the same values of s, S_e, S_p, σ_a^2, σ_b^2, $\tilde{\pi}$, c_4, c_3, c_2 and c_1 given for minimizing the variance. Using Equation (19), we obtain that

$$g = \sqrt{\frac{300}{((50)(10)+35)}\frac{\sqrt{0.0000522}}{0.0024(1-0.0024)(0.8775)}} \approx 3,$$

$$m = \sqrt{\frac{500}{300}\frac{0.8775}{0.755}} = 1.50 \approx 2,$$ while the number of localities is equal to:

$$l = [\{0.0024(1 - 0.0024)\}^2(0.57) + \{0.0024(1-0.0024)\}^2(0.77)/2 + \frac{0.0000522}{(2)(3)}]/0.000001627 = 9.3669 \approx 10$$

Since g and m do not change, we need 150 plants per field, 2 fields per locality and 10 localities to reach the required width (0.005) of the 95% CI. Now the budget is two times larger than that obtained for minimizing the variance given a budget constraint. However, this sample size is only valid for equal cluster sizes. If adjustments need to be made for unequal locality sizes and field sizes, they can be carried out by $l^* = \frac{1}{RE_{lt}}$ and $m^* = \frac{m}{RE_{ft}}$, respectively.

Now assume that we wish to determine the required number of localities without a budget constraint, assuming 2 fields per locality, a pool size of 50 and g = 10 pools per field. Using Equation (13) and assuming the same values for ω, α, S_e, S_p, σ_b^2, and $\tilde{\pi}$ that were given for minimizing the variance, we have

$$l = \frac{4(1.96)^2}{0.005^2}[\{0.0024(1 - 0.0024)\}^2(0.57) + \{0.0024(1 - 0.0024)\}^2(0.77)/2 + \frac{0.0000522}{(2)(10)}] = 5.623 \approx 6$$

This implies that we need a sample of 6 localities, 2 fields per locality and 10 pools of size 50 per field. These values do not change, assuming unequal locality and field sizes with the same mean and standard deviations.

For a desired power. Now suppose that we need to know the budget and sample size required for testing $H_0: \tilde{\pi}_0$ = 0.0024 vs $H_1: \tilde{\pi}_0$ > 0.0024 at a α = 0.05 significance level with a power (1 – γ) = 0.9 (90%) for detecting δ = 0.003 and using the same parameters (s, S_e, S_p, σ_b^2, c_4, c_3, c_2 and c_1) as in the example for minimizing the variance. Then,

$$V_0 = V_2 = \frac{0.003^2}{(1.645+1.282)^2} = 0.000001051.$$

Since $V(\delta_0)$ = $V(\delta)$ = 0.0000522, $\tilde{\pi} = \tilde{\pi}_0$, then

$$g = \sqrt{\frac{300}{((50)(10)+35)}\frac{\sqrt{0.0000522}}{0.0024(1-0.0024)(0.8775)}} \approx 3,$$

$m = \sqrt{\frac{500}{300}\frac{0.8775}{0.755}} = 1.50 \approx 2$, while the number of localities is equal to:

$$l = [\{0.0024(1 - 0.0024)\}^2\,(0.57) + \{0.0024(1 - 0024)\}^2(0.77)/2 + \frac{0.0000522}{(2)(3)}]/0.000001051 = 14.5 \approx 15$$

Here, again, we need 150 plants per field, 2 fields per locality and 15 localities to reach the required power of 90%. This means that the required budget is three times larger than that obtained for minimizing the variance given a budget constraint. To compensate for the unequal locality and field sizes and assuming the same mean and standard deviation of the sizes (μ_k = 177 and σ_k = 81.5, for $k = l, f$), we have to multiply the localities and fields obtained by correction factors $\frac{1}{RE_{lt}}$ and $\frac{1}{RE_{ft}}$, respectively.

Now let's assume that we decided to use 10 pools per field (g), 2 fields per locality and a pool size of 50, without a budget constraint and using the same values of S_e, S_p, $\sigma_a^2, \sigma_b^2, \alpha$, $(1 - \gamma)$, $\delta = \tilde{\pi}_1 - \tilde{\pi}_0$ as above. Then using Equation (15), the required number of localities, l, is equal to:

$$l = \frac{(1.645+1.282)^2}{0.003^2}\left[\{0.0024(1-0.0024)\}^2(0.57) + \frac{\{0.0024(1-0.0024)\}^2(0.77)}{2} + \frac{0.0000522}{(2)(10)}\right]$$

$$= 8.7 \approx 9$$

This means that to perform the study, we need 9 localities, 2 fields per locality and 10 pools per field of size 50. These values do not change, assuming unequal locality and field sizes with the same mean and standard deviation.

10. DISCUSSION AND CONCLUSIONS

In this paper, we derived optimal sample sizes for group testing in a three-stage sampling process under a budget or variance constraint. Given a pool size (s) and using Lagrange multipliers, we derived formulae to produce the optimal allocation of localities (l), fields (m) and pools per field (g). Although these formulae are similar to those derived by Cochran (1977; p. 285), they assume that all localities and fields are of the same size. However, in practice, this assumption is rarely satisfied; for this reason, we derived correction factors (inverse of the relative efficiency) to adjust the optimal sample sizes for unequal locality and field sizes. We also show examples of how to calculate the optimal values of l, m and g using these formulae when we wish to obtain a certain precision (width of the confidence interval) or a specified power.

If sample sizes for precision or power without a budget constraint are required, Equations (13) and (15) can be used for precision and power, respectively. However, these sample sizes are not optimal, since the values of m, g and s are given by the researcher in a non-optimal way.

There are two important aspects that should be taken into account, given that our optimal sample sizes were derived using a first-order TSE approach under the assumption that the variance components are known. First, the optimal sample sizes will be slightly biased, based on Monte Carlo simulations (Goldstein and Rabash 1996, Moerbeek and Van Breukelen 2001a,b, Moerbeek and Maas 2005, Candel and Van Breukelen 2010). Second, we assumed a relatively simple covariance structure for deriving the optimal sample sizes. These approximate sample sizes should be reasonable and can be calculated easily. However, further study of the performance of the proposed optimal sample sizes is required. Finally, it is important to carefully choose the required parameters in order to be able to obtain the optimal values of g, m and l, since if they are smaller than the true values (underestimated) the estimated sample size (g, m, and l) will be smaller as well, whereas if they are larger than the true values (overestimated) we will obtain sample size (g, m, and l) larger than the required in order to fulfill the a priory specified precision and thus researchers will be wasting resources.

REFERENCES

Ahn, C., Hu, F. and Lee, S.C. (2012). Relative efficiency of unequal versus equal cluster sizes for the nonparametric weighted sign test estimators in clustered binary data. *Drug Inform. J.*, **46(4)**, 428-433.

Anonymous (2003). Regulation (EC) 1829 of the European Parliament and the European Council of 22 September 2003 on genetically modified food and feed. *Official J. Europ. Union,* L, 268.

Bilder, C. (2009). Human or Cylon? Group Testing on Battlestar Galactica. *Chance,* **22(3)**, 46-50.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88(421)**, 9-25.

Candel, M.J. and Van Breukelen, G.J. (2010). Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Stat. Medi.*, **29(14)**, 1488.

Candy, S.G. (2000). The application of generalized linear mixed models to multi-level sampling for insect population monitoring. *Environ. Ecol. Statist.,* **7(3)**, 217-238. Chen, P., Tebbs, J. and Bilder, C. (2009). Group Testing Regression Models with Fixed and Random Effects, *Biometrics,* **65(4),** 1270-1278.

Cochran, W.G. (1977). *Sampling Techniques*. New York, Wiley. 3rd ed. Dodd, R.Y., Notari, E. and 4th. Stramer, S.L. (2002). Current prevalence and incidence of infectious disease markers and estimated window period risk in the American Red Cross blood donor population. *Transfusion*, **42(8),** 975-979.

Dorfman, R. (1943). The detection of defective members of large populations. *The Ann. Math. Stats.,* **14(4)**, 436-440.

Federer, W. (1994). Pooling and other designs for analysing laboratory samples more efficiently. *Statistician*, **43**, 413-422.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika,* **78(1),** 45-51.

Goldstein, H. (2003). *Multilevel Statistical Models*. Third Edition. Edward Arnold, London.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *J. Roy. Statist. Soc., A* **159(3)**, 505-513.

Hernández-Suárez, C.M., Montesinos-López, O.A., McLaren, G. and Crossa, J. (2008). Probability models for detecting transgenic plants. *Seed Sci. Res.,* **18**, 77-89.

Moerbeek, M. and Maas, C.J. (2005). Optimal experimental designs for multilevel logistic models with two binary predictors. *Comm. Statist. –Theory Methods,* **34(5)**, 1151-1167.

Moerbeek, M., van Breukelen, G.J.P. and Berger, M.P.F. (2001a). Optimal experimental Designs for Multilevel Models with Covariates. *Comm. Statist. –Theory Methods,* **30**, 2683-2697.

Moerbeek, M., van Breukelen, G.J. and Berger, M.P. (2000). Design issues for experiments in multilevel populations. *J. Edu. Behavi. Statist.*, **25(3),** 271-284.

Moerbeek, M., van Breukelen, G.J., and Berger, M.P. (2001a). Optimal experimental designs for multilevel logistic models. *J. Roy. Statist. Soc.*, Series D (The Statistician), **50(1)**, 17-30.

Moerbeek, M., van Breukelen, G.J.P. and Berger, M.P.F. (2001b). Optimal experimental designs for multilevel models with covariates. *Comm. Statist., –Theory Methods*, **30**, 2683-2697.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Eskridge, K. and Hernández-Suárez, C.M. (2010). Sample size for detecting and estimating the proportion of transgenic plants with narrow confidence intervals. *Seed Sci. Res.,* **20**, 123-136.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Eskridge, K., and Sáenz-Casas, R.A. (2011). Optimal sample size for estimating the proportion of transgenic plants using the Dorfman model with a random confidence interval. *Seed Sci. Res.,* **21(3)**, 235-246.

Mood, A.M., Graybill, F.A., and Boes, D.C. (1974). *Introduction to the Theory of Statistics* (3rd Edition). McGraw-Hill.

Peck, C. (2006). Going after BVD. *Beef 2006,* **42**, 34-44.

Piñeyro-Nelson, A., van Heerwaarden, J., Perales, H.R., Serratos-Hernández, J.A., Rangel, A., Hufford, M.B., Gepts, P., Garay Arroyo, A., Rivera Bustamante, R., and Álvarez Buylla, E.R. (2009). Transgenes in Mexican maize: molecular evidence and methodological considerations for GMO detection in landrace populations. *Mol. Ecol.,* **18(4)**, 750-761.

Rabe Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *J. Roy. Statist. Society: Series A* (Statistics in Society), **169(4)**, 805-827.

Remlinger, K., Hughes-Oliver, J., Young, S. and Lam, R. (2006). Statistical design of pools using optimal coverage and minimal collision. *Technometrics,* **48,** 133-143.

Rodríguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *J. Roy. Statist. Soc., Series A***158(1)**, 73-89.

Schliep, A. *et al.* (2003). Proceedings of the Computational Systems Bioinformatics; August 11-14; Stanford, CA. Group testing with DNA chips: generating designs and decoding experiments, 84-91.

Skrondal, A. and Rabe-Hesketh, S. (2007). Redundant over dispersion parameters in multilevel models for categorical responses. *J. Edu. Behavi. Statist.* **32,** 419-430.

Tebbs, J. and Bilder, C. (2004). Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. *J. Agril., Bio., Envir. Statist.,* **9(1)**, 79-90.

Van Breukelen, G.J.P., Candel, M.J.J.M. and Berger, M.P.F. (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicenter trials. *Stat. Medi.,* **26,** 2589-2603.

Verstraeten, T., Farah, B., Duchateau, L. and Matu, R. (1998). Pooling sera to reduce the cost of HIV surveillance: a feasibility study in a rural Kenyan district. *Tropical Medi. Intt. Health,* **3,** 747-750.

Wolf, J. (1985). *Born-again group testing: multi access communications.* IEEE *Transactions on Information Theory,* **31(2),** 185-191.

Yamamura, K. and Hino, A. (2007). Estimation of the proportion of defective units by using group testing under the existence of a threshold of detection.*Comm. Statist. - Simul. Comput.,* **36,** 949-957.

Appendix A. Derivation of the optimal solution for minimizing $V(\hat{\pi})$ subject to $C=lmgsc_1 + lmgc_2 + lmc_3 + lc_4$ ($c_u > 0$, l, m, g, $s \geq 2$, $u = 1, 2, 3, 4$) given the pool size (s)

By combining Eq. (12) and (16), we obtain the Lagrangean

$$L(l, m, g, \lambda) = L = V(\hat{\pi}) + \lambda[\mathrm{C} - (lmgsc_1 + lmgc_2 + lmc_3 + lc_4)] \tag{A1}$$

where $V(\hat{\pi}) = \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm} + \frac{V(\delta)}{lmg} = \frac{\sigma_a^{2*}}{l} + \frac{\sigma_b^{2*}}{lm} + \frac{V(\delta)}{lmg}$. λ is the Lagrange multiplier. The partial derivatives of Eq. (17) with respect to λ, m and g are

$$\frac{\partial L}{\partial \lambda} = C - (lmgsc_1 + lmgc_2 + lmc_3 + lc_4) = 0;\ \text{then } l = \frac{c}{mgsc_1 + mgc_2 + mc_3 + c_4}$$

$$\frac{\partial L}{\partial g} = -\frac{V(\delta)}{g^2ml} - \lambda lm(sc_1 + c_2) = 0;\ = \text{then } \lambda = -\frac{V(\delta)}{g^2m^2l^2(sc_1 + c_2)}$$

$$\frac{\partial L}{\partial m} = -\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm^2} - \frac{V(\delta)}{m^2gl} - \lambda[1g(sc_1 + c_2) + lc_3] = 0 \Leftrightarrow \frac{V(\delta)}{g^2l^2m^2(sc_1 + c_2)}[lg(sc_1 + c_2) + lc_3]$$

$$= \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm^2} + \frac{V(\delta)}{lm^2g},\ \text{since } \lambda = -\frac{V(\delta)}{g^2m^2l^2(sc_1 + c_2)}$$

$$\Leftrightarrow V(\delta)lg\ (sc_1 + c_2) + V(\delta)lc_3 = g^2l(sc_1 + c_2)\left[\{\tilde{\pi}(1-\tilde{\pi})^2\sigma_b^2 + \frac{V(\delta)}{g}\right],$$

$$\Leftrightarrow V(\delta)lc_3 = g^2l(sc_1 + c_2)\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$$

$$\Leftrightarrow g = \sqrt{\frac{c_3}{(sc_1 + c_2)}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}$$

$$\frac{\partial L}{\partial l} = \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l^2} - \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{ml^2} - \frac{V(\delta)}{l^2mg} - \lambda[mg(sc_1 + c_2) + mc_3 + c_4] = 0$$

$$\Leftrightarrow \frac{V(\delta)}{g^2m^2l^2(sc_1 + c_2)}[mg(sc_1 + c_2) + mc_3 + c_4] = \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l^2} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l^2m} + \frac{V(\delta)}{l^2mg},\ \text{since}$$

$$\lambda = -\frac{V(\delta)}{g^2m^2l^2(sc_1 + c_2)}$$

$V(\delta)mg(sc_1 + c_2) + V\,(\delta)mc_3 + V(\delta)c_4 = g^2m^2(sc_1 + c_2)\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + g^2m(sc_1 + c_2)$ $\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + V(\delta)mg(sc_1 + c_2)];$

$$V(\delta)mc_3 + V(\delta)c_4 = g^2m^2(sc_1 + c_2)\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + g^2m(sc_1 + c_2)\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$$

$$\Leftrightarrow V(\delta)mc_3 + V(\delta)c_4 = \frac{c_3V(\delta)}{(sc_1 + c_2)\{\pi(1-\pi)\}^2\sigma_b^2}\,[m^2(sc_1 + c_2)\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + m(sc_1 + c_2)\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2;$$

since $g = \sqrt{\frac{c_3}{(sc_1 + c_2)}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}$.

$$\Leftrightarrow V(\delta)mc_3 + V(\delta)c_4 = \frac{c_3V(\delta)\sigma_a^2m^2}{\sigma_b^2} + mc_3V(\delta)$$

$$\Leftrightarrow m = \sqrt{\frac{c_4\sigma_b}{c_3\sigma_a}}$$

Appendix B. Derivation of the optimal solution for minimizing $C = lmgsc_1 + lmgc_2 + lmc_3 + lc_4$ subject to $V(\hat{\pi}) = V_0$

By combining Eq. (12) and (16), we obtain the Lagrangean

$$L(l, m, g, \lambda] = L = (lmgsc_1 + lmgc_2 + lmc_3 + lc_4) + \lambda[V(\hat{\pi}) - V_0] \tag{B1}$$

where $V(\hat{\pi}) = \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm} + \frac{V(\delta)}{lmg}$. Now the partial derivatives of L with respect to λ, l, m and g are

$$\frac{\partial L}{\partial \lambda} = \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{lm} + \frac{V(\delta)}{lmg} - V_0 = 0;\ \text{then } l = [\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m} + \frac{V(\delta)}{mg}]/V_0$$

$$\frac{\partial L}{\partial g} = lm(sc_1 + c_2) - \lambda\frac{V(\delta)}{lmg^2} = 0;\ \text{then } \lambda = \frac{l^2g^2m^2(sc_1 + c_2)}{V(\delta)}$$

$$\frac{\partial L}{\partial m} = lg(sc_1 + c_2) + lc_3 - \frac{\lambda}{lm^2}[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}] = 0$$

$$\Leftrightarrow lg\ (sc_1 + c_2) + lc_3 \quad = \quad \frac{\lg^2(sc_1 + c_2)}{V(\delta)}[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}],\ \text{since}$$

$$\lambda \quad = \quad \frac{l^2g^2m^2(sc_1 + c_2)}{V(\delta)}$$

$$\Leftrightarrow lg(sc_1 + c_2) + lc_3 = \quad \frac{lg^2(sc_1 + c_2)}{V(\delta)}[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2] + lg(sc_1 + c_2),$$

$$\Leftrightarrow lc_3 = \frac{lg^2(sc_1 + c_2)}{V(\delta)}\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$$

$$\Leftrightarrow g = \sqrt{\frac{c_3}{(sc_1 + c_2)\tilde{\pi}}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}$$

$$\frac{\partial L}{\partial l} = mg(sc_1 + c_2) + mc_3 + c_4 - \lambda\left[\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{l^2} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{ml^2} + \frac{V(\delta)}{l^2mg}\right] = 0$$

$$\Leftrightarrow [mg(sc_1 + c_2) + mc_3 + c_4] = \frac{g^2l^2m^2(sc_1 + c_2)}{V(\delta)}\left[\frac{\{\pi(1-\pi)\}^2\sigma_a^2}{l^2} + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{l^2m} + \frac{V(\delta)}{l^2mg}\right],\ \text{since}$$

$$\lambda = \frac{l^2g^2m^2(sc_1 + c_2)}{V(\delta)}$$

$$\Leftrightarrow V(\delta)mg(sc_1 + c_2) + V(\delta)mc_3 + V(\delta)c_4 = g^2m^2(sc_1 + c_2)\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + g^2m(sc_1 + c_2)\ \{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + V(\delta)mg(sc_1 + c_2)]$$

$$\Leftrightarrow V(\delta)mc_3 + V(\delta)c_4 \quad = \quad \frac{V(\delta)m^2c_3\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2}{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2} + V(\delta)mc_3;$$

$$g \quad = \quad \sqrt{\frac{c_3}{(sc_1 + c_2)}}\frac{\sqrt{V(\delta)}}{\{\tilde{\pi}(1-\tilde{\pi})\}\sigma_b}$$

$$\Leftrightarrow m \quad = \quad \sqrt{\frac{c_4}{c_3}\frac{\sigma_b}{\sigma_a}}$$

Appendix C. An alternative way of minimizing the values of *m* and *g* given a pool size and a budget or variance constraint

In section 4.6 we derived optimal values for locality (*l*), fields (*m*) and pools per field (*g*) given a pool size (*s*) using Lagrange multipliers. We found that minimizing the variance given a budget constraint or minimizing the budget given a variance constraint produces the same optimal values for *m* and *g* due to duality. Only the expression for obtaining the optimal allocation of localities is different. For this reason, according to Cochran (1977), we obtain the same solution by minimizing the product of the variance of interest by the budget constraint. This means that the problem of minimization is the same as minimizing the product:

$$C(m, g) = V(\hat{\pi})\,(lmgsc_1 + lmgc_2 + lmc_3 + lc_4)$$

$$C(m, g) = \left[\sigma_a^{2*} + \frac{\sigma_b^{2*}}{m} + \frac{V(\delta)}{m\bar{g}}\right][c_4 + mc_3 + mg\,(sc_1 + c_2)] \tag{C1}$$

The Cauchy-Schwarz inequality (Cochran 1977, p. 77) is

$$(\textstyle\sum_h A_h^2)(\sum_h B_h^2) - (\sum_h A_h B_h)^2 \; = \; \sum_i \sum_{i>j} (A_i B_j - A_j B_i)^2 \geq 0 \tag{C2}$$

Therefore

$$(\textstyle\sum_h A_h^2)(\sum_h B_h^2) \geq (\sum_h A_h B_h)^2 \tag{C3}$$

Making $A_1 = \sqrt{\sigma_a^{2*}},\ A_2 = \sqrt{\dfrac{\sigma_b^{2*}}{m}},\ A_3 = \sqrt{\dfrac{V(\delta)}{m\bar{g}}},\ B_1 = \sqrt{c_4},\ B_2 = \sqrt{mc_3},\ B_3 = \sqrt{mg(sc_1 + c_2)},$ we can express (C1) as (C3):

$$[\sigma_a^{2*} + \frac{\sigma_b^{2*}}{m} + \frac{V(\delta)}{m\bar{g}}][c_4 + mc_3 + mg(sc_1 + c_2)] \geq \left(\sqrt{\sigma_a^{2*} c_4} + \sqrt{\sigma_b^{2*} c_3} + \sqrt{V(\delta)(sc_1 + c_2)}\right)^2 \tag{C4}$$

The product will be minimized, provided that the equality of Equation (C4) holds. Setting the equality of Equation (C4) and expanding both sides, we find that

$$\frac{\sqrt{c_4}}{\sqrt{\sigma_a^{2*}}} = \frac{\sqrt{mc_3}}{\sqrt{\dfrac{\sigma_b^{2*}}{m}}} \geq k \text{ a constant} \tag{C5}$$

and

$$\frac{\sqrt{mc_3}}{\sqrt{\dfrac{\sigma_b^{2*}}{m}}} = \frac{\sqrt{mg(sc_1 + c_2)}}{\sqrt{\dfrac{V(\delta)}{m\bar{g}}}} \geq k \tag{C6}$$

from (C5)

$$\frac{\sqrt{c_4}}{\sqrt{\sigma_a^{2*}}} = \frac{\sqrt{mc_3}}{\sqrt{\dfrac{\sigma_b^{2*}}{m}}} \Rightarrow \frac{\sqrt{c_4}}{\sigma_a^*} = \frac{m\sqrt{c_3}}{\sigma_b^*} \Rightarrow m = \sqrt{\frac{c_4}{c_3}}\,\frac{\sigma_b}{\sigma_a}$$

and also from (C6)

$$\frac{\sqrt{mc_3}}{\sqrt{\frac{\sigma_b^{2*}}{m}}} = \frac{\sqrt{mg(sc_1+c_2)}}{\sqrt{\frac{V(\delta)}{m\bar{g}}}} \Rightarrow \frac{m\sqrt{c_3}}{\sigma_b^{*}} = \frac{mg\sqrt{(sc_1+c_2)}}{\sqrt{V(\delta)}} \Rightarrow g = \sqrt{\frac{c_3}{(sc_1+c_2)}}\,\frac{\sqrt{V(\delta)}}{\{\tilde{\pi}(1-\tilde{\pi})\}\sigma_b}$$

If we wish to minimize the variance given a budget constraint, the optimal allocation of localities can be obtained by solving the budget constraint for *l*, and we get: $l^* = C/(mgsc_1 + mgc_2 + mc_3 + c_4)$. However, if we wish to minimize the budget given a variance constraint (V_0), the optimal value of localities can be obtained by solving the variance constraint for *l*, and we get: $l^* = [\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_a^2 + \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m} + \frac{V(\delta)}{mg}]/V_0$. Thus we get the same solution by using Lagrange multipliers.

Appendix D. Taylor series approximation (Eq. 23) of the RE in (Eq. 21) given by Van Breukelen *et al.* (2007).

Taylor series approximation (23 and 24) is derived from the RE of Equations (21 and 22) in four steps.

Step 1. Let the g_i values be independent realizations of a random variable cluster size U with expectation μ_n and standard deviation σ_n. Equations (21 and 22) are moment estimators of

$$RE(\tilde{\pi}) = \frac{\bar{g}+\alpha}{\bar{g}} E\left(\frac{U}{U+\alpha}\right) \tag{F1}$$

where $\alpha = (1-\rho)/\rho \geq 0$.

Step 2. Define $d = (U - \mu_n)$; then the last term in (F1) can be written as:

$$E\left(\frac{U}{U+\alpha}\right) = E\left\{\left(\frac{\mu_n + d}{\mu_n + \alpha + d}\right) = E\left\{\left(\frac{\mu_n + d}{\mu_n + \alpha}\right)\left(\frac{1}{1+(d/(\mu_n+\alpha))}\right)\right\}\right.$$

The last term is a Taylor series [Mood *et al.* (1974), p. 533, Equation (34)]:

$$\left(\frac{1}{1+(d/\mu_n+\alpha))}\right) = \sum_{j=0}^{\infty}\left(\frac{-d}{\mu_n+\alpha}\right)^j$$

if $-(\mu_n + \alpha) <$ d $< (\mu_n + \alpha)$ to ensure convergence.

Since $d = U - \mu_n$ and $\alpha \geq 0$, this convergence condition will be satisfied, except for a small probability $P(U > 2\mu_n + \alpha)$ for strongly positively skewed cluster size distributions combined with large ρ (= small α). Thus we have:

$$E\left(\frac{U}{U+\alpha}\right) = E\left\{\left(\frac{\mu_n+d}{\mu_n+\alpha}\right)\sum\nolimits_{j=0}^{\infty}\left(\frac{-d}{\mu_n+\alpha}\right)^j\right\} \tag{F2}$$

Step 3. If we ignore all terms d^j with $j > 2$, and rearrange terms in (F2), we will have

$$E\left(\frac{U}{U+\alpha}\right) = \lambda\{1 - CV^2\lambda(1-\lambda)\} \tag{F3}$$

where $\lambda = (\mu_g/\mu_g + \alpha)) \in (0, 1]$, assuming $\bar{g} = \mu_g$ and $CV = \sigma_g/\mu_g$ is the coefficient of variation of the random variable U.

Step 4. Plugging (F3) into (F1) gives:

$$RE(\hat{\pi})_t \approx \{1 - CV^2\lambda(1-\lambda)\} \tag{F4}$$

Remark

Ignoring in (F2) only those d^j terms with $j > 4$ instead of 2, will give

$$RE(\hat{\pi})_t \approx 1 - \{(1-\lambda)[\lambda CV^2 - \lambda CV^3 \text{skew} + \lambda^3 CV^4(\text{kurt}+3)]\} \tag{F5}$$

where skew and kurt are the skewness and kurtosis of the cluster size distribution, that is, skew = the 3rd central moment of the U divided by σ_n^3, and kurt = the 4th moment of U divided by σ_n^4, minus 3 (see, for example, Mood *et al.*, 1974, p.76).

Appendix E. Glimmix code to estimate the proportion and variance components.

```
data corn2004;
 input Locality Field pool yp;
cards;
  11    1    1    0
  11    1    2    0
  11    1    3    0
  11    1    4    0
  11    1    5    0
  11    1    6    0
  11    2    1    0
  11    2    2    0
  11    2    3    0
  11    2    4    0
  11    2    5    0
  11    2    6    0
  .
  .
  .
  11    29   1    0
  11    29   2    0
  11    29   3    0
  11    29   4    0
  11    29   5    0
  11    29   6    0
  11    30   1    0
  11    30   2    0
  11    30   3    0
  11    30   4    0
  11    30   5    0
  11    30   6    0
  7     1    1    0
  7     1    2    0
  7     1    3    0
  7     1    4    0
  7     1    5    0
  7     1    6    0
  .
  .
  .
  7     29   1    0
  7     29   2    0
  7     29   3    0
  7     29   4    0
  7     29   5    0
  7     29   6    0
  7     30   1    0
  7     30   2    1
  7     30   3    1
  7     30   4    1
  7     30   5    0
  7     30   6    0
  ;
```

```
proc glimmix data=corn2004 ;
class Locality field;
model yp(event='1')= /solution dist=binary;
random intercept/ subject=Locality;
random intercept / subject=field(Locality);
prod=1; s=50;
do i = 1 to s;
            p1= exp(_linp_)/(1 + exp(_linp_));
            prod = prod*(1 - p1);
   end;
   _MU_ =1 - prod;
run;
```