Seed Science Research (2015) **25**, 12–28 doi:10.1017/S096025851400035X

Optimal sample sizes for group testing in two-stage sampling

Osval Antonio Montesinos-López[1]*, Kent Eskridge[2], Abelardo Montesinos-López[3] and José Crossa[4]

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

(Received 21 March 2014; accepted after revision 17 October 2014; first published online 28 November 2014)

Abstract

Optimal sample sizes under a budget constraint for estimating a proportion in a two-stage sampling process have been derived using individual testing. However, when group testing is used, these optimal sample sizes are not appropriate. In this study, optimal sample sizes at the cluster and individual levels are derived for group testing. First, optimal allocations of clusters and individuals are obtained under the assumption of equal cluster sizes. Second, we obtain the relative efficiency (RE) of unequal versus equal cluster sizes when estimating the average population proportion, $\tilde{\pi}$. By multiplying the sample of clusters obtained assuming equal cluster size by the inverse of the RE, we adjust the sample size required in the context of unequal cluster sizes. We also show the adjustments that need to be made to allocate clusters and individuals correctly in order to estimate the required budget and achieve a certain power or precision.

Keywords: group testing, optimal power, precision, relative efficiency, sample size

Introduction

Group testing is becoming increasingly popular because it can substantially reduce the number of required diagnostic tests compared to individual testing. Dorfman (1943) proposed the original group testing method in which g pools of size s are randomly formed from a sample of n individuals selected from the population using simple random sampling (SRS). Dorfman's method has been extended in many ways. For example, there are group testing regression models for fixed effects, for mixed effects, for multiple-disease group testing data, with imperfect diagnostic tests [with sensitivity (S_e), specificity $(S_p) < 1$, or with dilution effect], and non-parametric group testing methods, among others (Yamamura and Hino, 2007; Hernández-Suárez *et al.*, 2008; Chen *et al.*, 2009; Zhang *et al.*, 2013).

Group testing methods have been used to detect diseases in potential donors (Dodd *et al.*, 2002); to detect drugs (Remlinger *et al.*, 2006); to estimate and detect the prevalence of human (Verstraeten *et al.*, 1998), plant (Tebbs and Bilder, 2004) and animal (Peck, 2006) diseases; to detect and estimate the presence of transgenic plants (Yamamura and Hino, 2007; Hernández-Suárez *et al.*, 2008); and to solve problems in information theory (Wolf, 1985) and even in science fiction (Bilder, 2009). When individuals are not nested within clusters, the issue of the number of pools the sample should have to achieve a certain power or precision for estimating the proportion of interest $\tilde{\pi}$ has been solved (Yamamura and Hino, 2007; Hernández-Suárez *et al.*, 2008; Montesinos-López *et al.*, 2010, 2011). In practice, however, populations often have a multilevel structure, with individuals nested within clusters that may themselves be nested within higher-order clusters. For example, in the detection of transgenic corn in Mexico, sample plants are nested in fields, which are nested in geographical areas. For such surveys, at least two stages may arise, and outcomes within the same cluster tend to be more alike than outcomes from different clusters. To account for such correlated outcomes, more clusters are needed to achieve the same precision as SRS which generates outcomes that are independent (Moerbeek, 2006).

Multistage surveys are often justified because it is difficult or impossible to obtain a sampling frame or list of individuals, or it may be too expensive to take an SRS. For example, it would not be possible to take an SRS of corn plants in Mexico due to travel costs

*Correspondence
Email: oamontes1@ucol.mx

between fields. Instead of using SRS, multistage or cluster sampling methods would typically be employed in this situation. Sampling units of two or more sizes are used and larger units, called clusters or primary sampling units (PSUs), are selected using a probability sampling design. Then some or all of the smaller units (called secondary sampling units or SSUs) are selected from each PSU in the sample. In the example of sampling for transgenic corn, PSU = field and SSU = plant. This design would be less expensive to implement than an SRS of individuals, due to the reduction in travel costs. Also, cluster sampling does not require a list of households or persons in the entire country. Instead, a list is constructed for the PSUs selected to be in the sample (Lohr, 2008).

In a non-group testing context, optimal sample size gives the most precise estimate of the proportion of interest and the largest test power or precision given a fixed sampling budget (Van Breukelen *et al.*, 2007). It can also be defined as the cheapest sample size that gives a certain power or precision of the estimate of interest (Van Breukelen *et al.*, 2007). It is less costly to sample a few clusters with many individuals per cluster than many clusters with just a few individuals per cluster because sampling in an already selected cluster may be less expensive than sampling in a new cluster (Moerbeek *et al.*, 2000). However, simulation studies in a non-group testing context indicate that it is more important to have a larger number of clusters than a larger number of individuals per cluster (Maas and Hox, 2004). In a group testing context, no work has been published on the optimal sample size in two-stage sampling, given a specified sampling budget. Thus new methods are needed to determine the required number of clusters and pools per cluster, given a certain budget, for obtaining a desired precision for estimating the proportion of interest using group testing.

Often optimal sample size calculations for multistage sampling completely assume equal cluster sizes (equal number of individuals per cluster). However, in practice, there are large discrepancies in cluster sizes, and ignoring this imbalance in cluster size could have a major impact on the power and precision required for the parameter estimates. For this reason, sample size formulas have to be adjusted for varying cluster sizes. One approach used to compensate for this loss of efficiency is to develop correction factors to convert the variance of equal cluster size into the variance of the unequal cluster size (Moerbeek *et al.*, 2001a; Van Breukelen *et al.*, 2007, 2008; Candel and Van Breukelen 2010). This correction factor is normally constructed as the inverse of the relative efficiency (RE), which is calculated as the ratio of the variances of the parameter of interest of equal versus unequal cluster sizes. This RE concept has been used in mixed-effects models for continuous and binary data to study loss of efficiency due to varying cluster sizes in a non-group testing context for the estimation of fixed parameters and for variance components (Van Breukelen *et al.*, 2007, 2008; Candel *et al.*, 2008). In the group testing framework, the RE concept has not been used to adjust optimal sample sizes under the assumption of equal cluster sizes.

In this study, we obtain optimal sample sizes in two stages in a group testing context using a multilevel logistic group testing model where we assume that clusters are randomly sampled from a large number of clusters. First, under the assumption that cluster sizes do not vary, we derive analytical expressions for the optimal allocation of clusters and individuals under a budget constraint. These analytical expressions were derived by linearization using a first-order marginal quasi-likelihood to approach the multilevel logistic group testing model. Although equal sample sizes per cluster are generally optimal for parameter estimation, they are rarely feasible. For this reason, we derived an approximate formula for the relative efficiency of unequal versus equal cluster sizes for adjusting the required sample sizes for estimating the proportion in a group testing context. The approximate RE obtained is a function of the mean, the variance of cluster size and the intraclass correlation. The proposed expressions are also useful for estimating the budget required to achieve a certain power or precision when the goal is to achieve a confidence interval of a certain width or to obtain a pre-specified power for a given hypothesis.

Materials and methods

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, and assuming that the responses y_{ij} are independent and Bernoulli distributed with probabilities π_i and that these probabilities are not related to any covariable (Moerbeek *et al.*, 2001a). Then the linear predictor using a logit link is equal to

$$\eta_i = \text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + b_i \qquad (1)$$

where η_i is the linear predictor that is formed from a fixed part (β_0) and a random part (b_i), which is Gaussian iid with mean zero and variance σ_b^2. Therefore, equation (1) can be written in terms of the probability of a positive individual as:

$$\pi_i = \pi_i(\beta_0, \sigma_b) = [1 + \exp\{-(\beta_0 + b_i)\}]^{-1}. \qquad (2)$$

The mixed logit model for binary responses can be written as the probability π_i plus a level 1 residual,

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

denoted e_{ij}:

$$y_{ij} = \pi_i + e_{ij}$$

where e_{ij} has zero mean and variance $(y_{ij}|b_i) = \pi_i(1-\pi_i)$ (Goldstein, 1991, 2003; Rodríguez and Goldman, 1995; Candy, 2000; Moerbeek *et al.*, 2001b; Skrondal and Rabe-Hesketh, 2007; Candel and Van Breukelen, 2010). This model is widely used for estimating optimal sample sizes when the variance components are assumed known (Goldstein, 1991, 2003; Rodríguez and Goldman, 1995; Candy, 2000; Moerbeek *et al.*, 2001a).

Random logistic model for group testing

Suppose that, within the ith field, each plant is randomly assigned to one of the g_i pools; let $y_{ijk} = 0$ if the kth plant in the jth pool in field i is negative, or $y_{ijk} = 1$ otherwise for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, g_i$ and $k = 1, 2, \ldots, s_{ij}$ as the pool size. Note that y_{ijk} is not observed, except when the pool size is 1. Define the random binary variable Z_{ij} that takes the value of $Z_{ij} = 1$ if the jth pool in field i tests positive and $Z_{ij} = 0$ otherwise. Therefore, the two-level generalized linear mixed model (Breslow and Clayton, 1993; Rabe-Hesketh and Skrondal, 2006) for the response Z_{ij} is exactly the same as that given for individual testing in equation (1). Conditional on the random effect $[b_i]$, the statuses of pools within field i are independent, and assuming that the statuses of pools from different fields are independent, the probability that the jth pool in field i is given as

$$\begin{aligned} P(Z_{ij} = 1|b_i) &= \pi_i^p \\ &= S_e + (1 - S_e - S_p)\prod_{k=1}^{s_{ij}}(1 - \pi_{ijk}) \end{aligned} \tag{3}$$

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, we will assume an equal pool size, s, in all clusters, and under this assumption equation (3) reduces to:

$$P(Z_{ij} = 1|b_i) = \pi_i^p = S_e + \varphi\,(1 - \pi_i)^s \tag{4}$$

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

$$Z_{ij} = \pi_i^p + e_{ij}^p \tag{5}$$

where π_i^p is as given in equation (4) and e_{ij}^p has zero mean and variance $V(Z_{ij}|b_i) = \pi_i^p(1 - \pi_i^p)$. Now let $\boldsymbol{\theta} = (\beta_0, \sigma_b)$ denote the vector of all estimable parameters. The multilevel likelihood is calculated for each level of nesting. First, the conditional likelihood for pool j in field i is given by:

$$L_{ij}(\boldsymbol{\theta}\Big|b_i) = \left[\pi_i^p\right]^{Z_{ij}}[1 - \pi_i^p]^{1-Z_{ij}}. \tag{6}$$

By multiplying the conditional likelihood (equation 6) by the density of b_i and integrating out the random effects, we get the marginal (unconditional) overall likelihood:

$$L(\boldsymbol{\theta}|y) = \prod_{i=1}^{m}\left\{\int\prod_{j=1}^{g_i}L_{ij}(\boldsymbol{\theta}\Big|b_i)f(b_i)db_i\right\},$$

where $f(b_i)$ is the density function of b_i. 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 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.

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_i^p is approximated using a first-order TSE around $b_i = 0$, as

$$\begin{aligned} \pi_i^p &\approx \pi_i^p\big|_{b_i=0} + \left.\frac{\partial\,\pi_i^p}{b_i}\right|_{b_i=0}(b_i - 0) \\ \pi_i^p &\approx \pi_i^p\big|_{b_i=0} + s\varphi\,(1 - \pi_i)^{s-1}\pi_i(1 - \pi_i)\big|_{b_i=0}(b_i) \\ \pi_i^p &\approx \tilde{\pi}^p + s\varphi\,(1 - \tilde{\pi})^{s-1}\tilde{\pi}(1 - \tilde{\pi})b_i \end{aligned} \tag{7}$$

where $\tilde{\pi}^p = \pi_i^p\big|_{b_i=0} = Se + \varphi\left(1 - \left[1 + \exp\left(-\beta_0\right)\right]^{-1}\right)^s$ and $\tilde{\pi} = \pi_i|_{b_i=0} = \left[1 + \exp\left(-\beta_0\right)\right]^{-1}$, since b_i are independent and identically distributed (iid), and we use the fact that

$$\frac{\partial\,\pi_i^p}{b_i} = \frac{\partial\,\pi_i^p}{\pi_i}\frac{\partial\,\pi_i}{\partial\,b_i}, \quad \frac{\partial\,\pi_i}{\partial\,b_i} = \frac{\partial\,\pi_i}{\partial\,\eta_i} = \pi_i(1 - \pi_i) \text{ and}$$

$$\frac{\partial\,\pi_i^p}{\pi_i} = s\varphi\,(1 - \pi_i)^{s-1}$$

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

$$Z_{ij} \approx \tilde{\pi}^p + s\varphi\,(1 - \tilde{\pi})^{s-1}\tilde{\pi}(1 - \tilde{\pi})b_i + e_{ij}^p. \tag{8}$$

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

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

$$\mathrm{Var}(Z_{ij}) \approx \{s\varphi(1-\tilde{\pi})^{s-1}\}^2\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \tilde{\pi}^p(1-\tilde{\pi}^p)$$

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

$$Var(\bar{Z}) \approx \frac{\{s\varphi(1-\tilde{\pi})^{s-1}\}^2\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m} + \frac{\tilde{\pi}^p(1-\tilde{\pi}^p)}{mg} \quad (9)$$

Recall that we will select a sample of 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_i is a random variable with a probability distribution. Therefore, it is reasonable to work with the expected value of π_i across fields to determine sample size. To approximate $E(\pi_i)$, we take advantage of the relationship between $\bar{Z}$ and $E(\pi_i^p)$:

$$\begin{aligned}\bar{Z} = E(\pi_i^p) &= E(S_e + \varphi(1-\pi_i)^s) \\ &= E(S_e) + E(\varphi(1-\pi_i)^s) = S_e + \varphi E(K) \quad (10)\end{aligned}$$

where $K = (1-\pi_i)^s$. Using a first-order TSE around $b_i = 0$, we can approximate K as

$$K \approx K|_{b_i=0} + \left.\frac{\partial K}{b_i}\right|_{b_i=0}(b_i - 0)$$

$$K \approx \tilde{K} + s(1-\tilde{\pi})^{s-1}\tilde{\pi}(1-\tilde{\pi})b_i \quad (11)$$

where $\tilde{K} = K|_{b_i=0} = \left(1 - \left[1 + \exp(-\beta_0)\right]^{-1}\right)^s = (1-\tilde{\pi})^s$ and we use the fact that

$$\frac{\partial K}{b_i} = \frac{\partial K}{\pi_i}\frac{\partial \pi_i}{\partial b_i}, \quad \frac{\partial \pi_i}{\partial b_i} = \frac{\partial \pi_i}{\partial \eta_i} = \pi_i(1-\pi_i) \text{ and}$$

$$\frac{\partial K}{\pi_i} = s(1-\pi_i)^{s-1}.$$

Then

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

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

$$E(K) \approx (1-E(\pi_i))^s.$$

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

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

Therefore, an estimator for $E(\pi_i)$ is

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

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

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

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

$$\begin{aligned}V\left(\widehat{E(\pi_i)}\right) = V(\hat{\pi}) &\approx \frac{\sigma_b^{2*}}{m} + \frac{V(\delta)}{mg} \\ &= \frac{(\sigma_b^{2*} + V(\delta))[(\bar{g}-1)\rho + 1]}{m\bar{g}} \quad (12)\end{aligned}$$

where $\sigma_b^{2*} = \{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$, $V(\delta) = \frac{(Se - \tilde{\pi}^p)^{\frac{2}{s}-2}\tilde{\pi}^p(1-\tilde{\pi}^p)}{s^2(\varphi)^{2/s}}$, $\tilde{\pi}^p = S_e + \varphi(1-\tilde{\pi})^s$ and $\rho = \sigma_b^{2*}/[\sigma_b^{2*} + V(\delta))]$ is the intraclass correlation coefficient that measures the amount of variance between clusters (fields).

Results and discussion

Optimal sample size assuming equal cluster size

Minimizing variance subject to a budget constraint

Now assume we have a fixed sampling budget for estimating the population proportion, π. The question of interest is how to allocate clusters (m) and pools per cluster (g) to estimate the proportion $\tilde{\pi}$ with minimum variance, subject to the budget constraint:

$$\mathrm{C} = mgc_1 + mc_2 \quad (c_l > 0,\ m, g \geq 2,\ l = 1, 2) \quad (13)$$

where C is the total sampling budget available, c_1 is the cost of obtaining a pool of s plants from a field, and c_2 is the cost of obtaining a cluster. The optimal allocation of units can be obtained using Lagrange multipliers. By combining equations (12) and (13), we obtain the Lagrangean

$$L(m, g, \lambda) = L = V(\hat{\pi}) + \lambda[\mathrm{C} - (mgc_1 + mc_2)] \quad (14)$$

where $V(\hat{\pi})$, given by equation (12), is the objective function that will be minimized with respect to m and g,

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

subject to the constraint given in equation (13), and λ is the Lagrange multiplier. The partial derivatives of equation (14) with respect to λ, m and g are:

$$\frac{\partial L}{\partial \lambda} = 0 = \mathrm{C} - (mgc_1 + mc_2);\ \text{then}\ m = \frac{C}{c_2 + gc_1}$$

$$\frac{\partial L}{\partial g} = 0 = -\frac{V(\delta)}{g^2 m} - \lambda m c_1;\ \text{then}\ \lambda = -\frac{V(\delta)}{g^2 m^2 c_1}$$

$$\frac{\partial L}{\partial m} = 0 = -\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2 \sigma_b^2}{m^2} - \frac{V(\delta)}{m^2 g} - \lambda[gc_1 + c_2].$$

Solving these equations results in the optimal values for m and g (see Appendix A):

$$m = \frac{C}{c_2 + gc_1},\ \text{where}\ g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}. \qquad (15)$$

First, we calculate the number of pools per field, g, rounded to the nearest integer. Using this value, we calculate the number of fields to sample, m, rounded to the nearest integer. Note that equation (15) is a generalization of the optimal sample sizes for continuous data for two–level sampling given by Brooks (1955) and Cochran (1977).

Minimizing the budget to obtain a certain width of the confidence interval

Often a researcher is interested in choosing the number of clusters and pools per cluster to minimize the total budget, C, to obtain a specified width (ω) of the confidence interval (CI) of the proportion of interest. Assuming that the distribution of $\hat{\pi}$ is approximately normal with a mean $\tilde{\pi}$ and a fixed variance $\mathrm{Var}(\hat{\pi})$, then the $(1-\alpha)100\%$ Wald confidence interval of $\tilde{\pi}$ is given by $\hat{\pi} \mp Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\pi})}$, where $Z_{1-\alpha/2}$ is the quantile $1-\alpha/2$ of the standard normal distribution. Therefore, the observed width of the CI is equal to $W = 2Z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\pi})}$, and since we specified the required width of the CI to be ω, this implies that $V(\hat{\pi}) = \omega^2/4Z^2_{1-\alpha/2}$. Here the optimization problem is to minimize the sampling budget as given in equation (13) under the condition that $V(\hat{\pi})$ (equation 12) is fixed. That is, we want to minimize $\mathrm{C} = mgc_1 + mc_2$ subject to $V(\hat{\pi}) = V_0$. Again, using Lagrange multipliers, the corresponding Lagrangean is $L(m, \mathrm{g}, \lambda) = L = mgc_1 + mc_2 + \lambda[V(\hat{\pi}) - V_0]$. Now the partial derivatives of L with respect to λ, m and g are

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

$$m = \left[\{\tilde{\pi}\,(1-\tilde{\pi})\}^2 \sigma_b^2 + \frac{V(\delta)}{g}\right]/V_0$$

$$\frac{\partial L}{\partial g} = 0 = mc_1 - \lambda\frac{V(\delta)}{g^2 m} - \lambda m c_1;\ \text{then}\ \lambda = \frac{g^2 m^2 c_1}{V(\delta)}$$

$$\frac{\partial L}{\partial m} = 0 = gc_1 + c_2 - \frac{\lambda}{m^2}\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2 \sigma_b^2 + \frac{V(\delta)}{g}\right].$$

Solving these equations for the optimal values gives (see Appendix B):

$$m = \left[\{\tilde{\pi}(1-\tilde{\pi})\}^2 \sigma_b^2 + \frac{V(\delta)}{g}\right]/V_0,\ \text{where} \qquad (16)$$

$$g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}$$

Note that the number of pools per cluster, g, required when we minimize the cost subject to $V(\hat{\pi}) = V_0$ is the same as when minimizing $V(\hat{\pi})$ (equation 14) subject to a budget constraint. However, the expression for obtaining the required number of clusters, m, is different. In this case, the value of $V_0 = \omega^2/4Z^2_{1-\alpha/2}$ is substituted into equation (16) and the expression for the required number of clusters is $m = \frac{4Z^2_{1-\alpha/2}}{\omega^2}\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2 \sigma_b^2 + \frac{V(\delta)}{g}\right]$. Another way of obtaining the same solution to this problem is given in Appendix C.

It is useful to consider the problem without a budget constraint. For a fixed number of pools per cluster (g), with a CI width of ω, we can get the required number of clusters, m, by making $2Z_{1-\alpha/2}\sqrt{\frac{g\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{gm} + \frac{V(\delta)}{mg}} = \omega$ and solving for m. The required number, m, is equal to:

$$m = \frac{4Z^2_{1-\alpha/2}}{\omega^2}\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2 \sigma_b^2 + \frac{V(\delta)}{g}\right] \qquad (17)$$

Equation (17) is the same expression as derived in equation (16) for the required number of clusters for minimizing the total budget subject to a variance constraint. However, equation (17) produces optimal allocation of clusters, m, only when we replace the values of $g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}$ in equation (17).

Minimizing the budget to obtain a certain power

Assume a threshold is defined *a priori*, and our main interest is to test $H_0 : \tilde{\pi} = \tilde{\pi}_0$ versus $H_1 : \tilde{\pi} > \tilde{\pi}_0$. 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 the issue of interest is to determine a sampling plan (i.e. m and g) budget required for this test to have a specified power $(1-\gamma)$ and significance level α when $\delta = |\tilde{\pi}_1 - \tilde{\pi}_0|$. For performing a test with a type I error rate of α and a type II error rate of γ when $\tilde{\pi} = \tilde{\pi}_1$

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

under H_1, the following must hold:

$$Z_{1-\alpha} = (\hat{\pi} - \tilde{\pi}_0)/\sqrt{\text{var}(\widehat{\pi_0})} \text{ and}$$

$$Z_{1-\gamma} = (\hat{\pi} - \tilde{\pi}_1)/\sqrt{\text{var}(\widehat{\pi_0})}.$$

Here $\text{var}(\widehat{\pi_0})$ is the variance of $\hat{\pi}$ but under the value of the null hypothesis. Both $Z_{1-\alpha}$ and $Z_{1-\gamma}$ have a standard normal distribution because the variance components are assumed known. According to Cochran (1977) and Moerbeek *et al.* (2000), these equations result in the relation:

$$V_2 = \frac{|\delta|^2}{(Z_{1-\alpha} + Z_{1-\gamma})^2}. \quad (18)$$

If we change the alternative hypothesis to $H_1 : \tilde{\pi} < \tilde{\pi}_0$, equation (18) is still valid, but if we change to a two-sided test $H_1 : \tilde{\pi} \neq \tilde{\pi}_0$, $Z_{1-\alpha}$ in equation (18) is replaced by $Z_{1-\alpha/2}$. This is because we want the required budget for this test to have the specified power $(1 - \gamma)$ and significance level α when $\delta = |\tilde{\pi}_1 - \tilde{\pi}_0|$.

Similarly, we are interested in minimizing the total budget to obtain a specified power $(1 - \gamma)$. This implies that $V(\widehat{\pi_0})$ is a fixed quantity and equal to equation (18). Therefore, the problem is exactly the same as minimizing the budget to obtain a certain width of the confidence interval, but with a value of V_0 equal to equation (18), since we want to minimize $\min(C = mgc_1 + mc_2)$ subject to $V(\hat{\pi}) = V_2$. Thus the optimal allocation of clusters and pools per cluster is also given in equation (16) but using equation (18) in place of V_0, $V(\delta_0) = \frac{(Se - \tilde{\pi}_0^p)^{\frac{2}{s}-2}\tilde{\pi}_0^p(1-\tilde{\pi}_0^p)}{s^2(Se+Sp-1)^{2/s}}$ in place of $V(\delta)$, and $\tilde{\pi}_0$ in place of $\tilde{\pi}$; therefore, $\tilde{\pi}_0^p = S_e + (1 - S_e - S_p)(1 - \tilde{\pi}_0)^s$ since these values need to be calculated under the null hypothesis. This implies that $m = \frac{(Z_{1-\alpha}+Z_{1-\gamma})^2}{|\delta|^2}\left[\{\tilde{\pi}_0(1 - \tilde{\pi}_0)\}^2\sigma_b^2 + \frac{V(\delta_0)}{g}\right]$ and $g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta_0)}}{\tilde{\pi}_0(1-\tilde{\pi}_0)\sigma_b}$.

Again, assuming no budget constraint and a given number of pools per cluster, g, we can solve for the required number of clusters, m, to achieve a power level $(1 - \gamma)$ for a desired δ. To get the required m we need to make $\text{var}(\widehat{\pi_0}) = \frac{|\delta|^2}{(Z_{1-\alpha}+Z_{1-\gamma})^2}$ and solve for m. Therefore, solving for m from equation (18) indicates that the required number of clusters (m) is equal to:

$$m = \frac{(Z_{1-\alpha} + Z_{1-\gamma})^2}{|\delta|^2}\left[\{\tilde{\pi}_0(1 - \tilde{\pi}_0)\}^2\sigma_b^2 + \frac{V(\delta_0)}{g}\right]. \quad (19)$$

Here, also, equation (19) is the same as that obtained for m from equation (16) but with $V_0 = \frac{|\delta|^2}{(Z_{1-\alpha}+Z_{1-\gamma})^2}$. For this reason, equation (19) produces optimal values if we use $g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta_0)}}{\tilde{\pi}_0(1-\tilde{\pi}_0)\sigma_b}$.

Behaviour of the optimal sample size for equal cluster sizes

Figure 1a presents several graphs that demonstrate the behaviour of the optimal sample size for equal cluster sizes and values of $\sigma_b^2 = 0.25$. Most of the time the optimal sample size requires fewer clusters (m) than pools per cluster (g) since the ratio (m/g) is usually less than 1. However, for values of $\sigma_b^2 \geq 0.65$ and $\pi > 0.04$, $m/g > 1$, and more clusters (m) than pools per cluster (g) are required. Figure 1a illustrates that when the variability between clusters, σ_b^2, is greater than the variability within clusters, $V(\delta)$, more clusters than pools per cluster are needed when the remaining parameters are fixed.

Figure 1b illustrates the behaviour of the ratio (m/g) as a function of the cost of enrolling clusters in the study c_2. As c_2 increases, the ratio (m/g) decreases, which is expected since the cost of including a cluster increases relative to the cost of enrolling pools, which does not change. Figure 1c shows that the number of clusters, m, decreases as the expected width of the CI increases (ω), which makes sense, since a narrow expected width (ω) of the CI implies that the estimation process is more precise, and vice versa. In Fig. 1d, we can see that the required number of clusters, m, increases when a larger power is required.

Correction factor for unequal cluster sizes

Although equal cluster sizes are optimal for estimating the proportion of interest, they are rarely encountered in practice. Variation in the actual size of the clusters (fields, localities, hospital, schools, etc.), non-response and dropout of individuals (among others) generate unequal cluster sizes in a study (Van Breukelen *et al.*, 2007). Cluster size variation increases bias and causes 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 for adjusting the optimal sample size under the assumption of equal cluster sizes. The relative efficiency of equal versus unequal cluster sizes for the estimator of the proportion of interest, $RE(\hat{\pi})$, is defined as:

$$RE(\hat{\pi}) = \frac{\text{Var}(\hat{\pi}|\mathbf{s}_{\text{equal}})}{\text{Var}(\hat{\pi}|\mathbf{s}_{\text{unequal}})} \quad (20)$$

where $\text{Var}(\hat{\pi}|\mathbf{s}_{\text{equal}})$ denotes the variance of the proportion estimator given a design with equal cluster sizes, $\text{Var}(\hat{\pi}|\mathbf{s}_{\text{unequal}})$ denotes a similar value for an unequal cluster size design, but with the same number of clusters m and the same total number of pools ($N = \sum_{i=1}^{m} g_i$) as in the equal cluster size design.

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

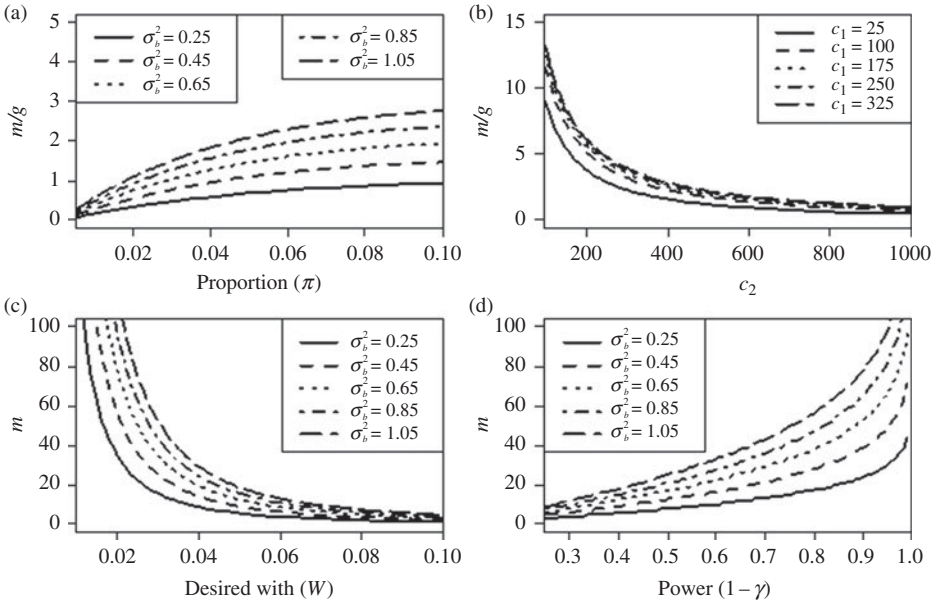

Figure 1. Ratio of the number of clusters and number of pools per cluster. (a) Ratio of the number of clusters and number of pools per cluster m/g as a function of the proportion ($\tilde{\pi}$), for $C = 10,000$, $c_1 = 250, c_2 = 800, s = 10, S_e = 0.98, S_p = 0.96$ and five different values of σ_b^2. (b) Ratio of m/g as a function of c_2, for $C = 10,000$, $\sigma_b^2 = 0.5$, $\tilde{\pi} = 0.05$, $s = 10, S_e = 0.98, S_p = 0.96$ and several values of c_1. (c) Required number of clusters, m, as a function of the desired confidence interval width (ω), for $c_1 = 50, c_2 = 1600$, $\tilde{\pi} = 0.05, s = 10, S_e = 0.98, S_p = 0.96$, and five different values of σ_b^2. (d) Required number of clusters, m, as a function of the desired power $(1 - \gamma)$, for $c_1 = 50, c_2 = 1600$, $\tilde{\pi} = 0.04, d = 0.015, s = 10, S_e = 0.98, \mathrm{S}_p = 0.96, \alpha = 0.05$ and five different values of σ_b^2.

Thus $RE(\hat{\pi})$ is equal to:

$$RE(\hat{\pi}) = \frac{(\sigma_b^{2*} + V(\delta)/\bar{g})/m}{\sum_{i=1}^{m}[(\sigma_b^{2*} + V(\delta)/g_i]/m^2} = \frac{\bar{g}+\alpha}{\bar{g}}\frac{1}{m}\sum_{i=1}^{m}\left[\frac{g_i}{g_i+\alpha}\right] \quad (21)$$

where $\sigma_b^{2*} = \{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$ and $\alpha = V(\delta)/\sigma_b^{2*}$. Note that equation (21) is equal to that derived for the RE of equal versus unequal cluster sizes in cluster randomized and multicentre trials given by Van Breukelen *et al.* (2007) to recover 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. Since our RE was expressed as that derived by Van Breukelen *et al.* (2007), we use their approach to obtain a Taylor series approximation of equation (21), expressing the RE as a function of the intraclass correlation ρ, and the mean and standard deviation of cluster size. It is important to point out that equation (21) is expressed in terms of pools instead of individuals, as in the formula of Van Breukelen *et al.* (2007). Therefore, we assumed that the cluster sizes $g_i (i = 1, 2, \ldots, m)$ are realizations of a random variable U having mean μ_g and standard deviation σ_g. According to Van Breukelen *et al.* (2007), equation (21) can be considered a moment estimator of

$$RE(\hat{\pi}) = \frac{\bar{g}+\alpha}{\bar{g}}E\left(\frac{U}{U+\alpha}\right). \quad (22)$$

If we define $\lambda = (\mu_g/(\mu_g + \alpha))$, and the coefficient of variation of the random variable U by $CV = \sigma_g/\mu_g$, then by using derivations similar to those reported by Van Breukelen *et al.* (2007, pp. 2601–2602; see Appendix D), we obtain the following second-order Taylor approximation of the expectation part of equation (22) $E(\frac{U}{U+\alpha}) \approx \lambda\{1 - CV^2\lambda(1-\lambda)\}$. The second-order Taylor approximation of equation (21) is:

$$RE(\hat{\pi})_t \approx \left\{1 - CV^2\lambda(1-\lambda)\right\}. \quad (23)$$

It is evident that $RE(\hat{\pi})_t$ does not depend on the number of clusters m, but rather on the distribution of cluster sizes (mean and variance) and intraclass correlations. When $\sigma_b^{2*} \to 0$ (and thus $\rho \to 0$) or $\sigma_b^{2*} \to \infty$ (and thus $\rho \to 1$), we have $RE \to 1$. For $0 < \sigma_b^{2*} < \infty$ (and thus $0 < \rho < 1$), we can see that $RE < 1$, implying that equal cluster sizes are optimal. For practical purposes, we will denote $RE(\hat{\pi})_t = RE_t$. To correct for the loss of efficiency due to the assumption of equal cluster sizes, one simply divides the number of clusters (m) given in equation (15) or (16) by the expected RE resulting from equation (23). Also, it is evident that the number of clusters will increase the budget to

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

Table 1. Cluster size distribution used for calculating relative efficiency

Distribution	Cluster size			Cluster frequencies			CV
	g_a	g_b	g_c	f_a	f_b	f_c	
$m = 18, \bar{g} = 22, s = 10$							
Uniform	4	22	40	6	6	6	0.668
Unimodal	4	22	40	2	14	2	0.386
Bimodal	4	22	40	8	2	8	0.771
Positively skewed	8	26	44	8	6	4	0.643
$m = 48, \bar{g} = 20, s = 10$							
Uniform	5	20	35	16	16	16	0.612
Unimodal	5	20	35	8	32	8	0.433
Bimodal	5	20	35	22	4	22	0.718
Positively skewed	10	24	42	24	16	8	0.583

f_a number of clusters of size g_a (small), f_b number of clusters of size g_b (medium), f_c number of clusters of size g_c (large); CV = coefficient of variation. Two numbers of clusters were studied: $m = 18$ with average pools per cluster $\bar{g} = 22$, and $m = 48$ with average pools per cluster $\bar{g} = 20$. In both cases, the pool size was $s = 10$.

$C^* = C\left(\frac{1}{RE_t}\right)$, whereas the optimal number of pools per cluster (g) does not change.

Comparison of the relative efficiency and its Taylor approximation

To compare the RE of equation (21), its Taylor approximation (equation 23) was performed for four cluster size distributions: uniform, unimodal, bimodal and positively skewed. Three different cluster sizes, g_a, g_b, g_c, with frequencies f_a, f_b, f_c, were evaluated (see Table 1). For each of the four distributions, both REs [asymptotic (equation 21) and Taylor approximation (equation 23)] were computed and plotted as a function of the intraclass correlation (the values used were from 0.0 to 0.3).

Figure 2 shows that for the four distributions (uniform, unimodal, bimodal and positively skewed), the RE drops from 1 at $\rho = 0$ to minimum at ρ somewhere between $\rho = 0.05$ and 0.1, and then increases, returning to 1 for $\rho = 1$. Lower RE values are observed when there is more cluster size variation (as in the case of bimodal distribution with larger values of $CV > 0.70$). For this reason, by comparing the four distributions, we can see in Fig. 2 that the positively skewed distribution gives the highest RE, followed by the unimodal, uniform and bimodal distributions. These results are in line with results reported by Van Breukelen *et al.* (2007, 2008) and Candel and Van Breukelen (2010) for studies of cluster randomized trials for normal data and binary results in a non-group testing context.

Figure 2 also shows that the Taylor approximation (equation 23, denoted as RE_t) of the RE given in equation (21) is acceptable in most cases. However, it is clearly affected by the distribution of the cluster sizes, the number of clusters, the number of pools per cluster and the value of the intraclass correlation.

Estimating the proportion of transgenic plants – An example

Next we illustrate how to achieve the optimal allocation of fields and pools for minimizing the variance (using equation 15), and for estimating the required budget for a desired CI width and the budget required to obtain a certain power (using equation 16). Carreón-Herrera (2011) collected corn grain in 14 localities of the Sierra Nororiental and 22 localities in the Mixteca Baja, in the State of Puebla, Mexico. She collected a total of 58 kg of grain. Forty-seven samples were obtained from farmers and 11 from DICONSA stores. Of the 58 samples, 36 were white grain, 10 yellow, 8 blue and 4 red. The researchers used the polymerase chain reaction (PCR) to detect the promoter of cauliflower mosaic virus (CaMV-35S), which indicates the presence of transgenic corn. They reported the percentage of the CaMV-35S promoter in each sample. The standard 0.01% was used as the lower limit of reference for the detection of CaMV-35S. The percentages of the CaMV-35S promoter reported varied between 0.01% and 0.25%. However, in a study conducted in the neighbouring state of Oaxaca, Landavazo Gamboa *et al.* (2006) reported a lower value (0.000012% median for the five fields studied) for the percentage of the CaMV-35S promoter. Assuming that we wish to conduct another study in this region of Puebla, we can assume that the expected proportion of transgenes is equal to $\tilde{\pi} = \frac{0.0025 - 0.00000012}{2} = 0.0013$, while the variance between clusters $\sigma^2 = (\frac{\text{range}}{6})^2$. For binomial data, the range relevant to six-sigma approximation is the difference between the maximum and minimum plausible logit (Stroup, 2012). Since we know the lowest ($\tilde{\pi}_L = 0.00000012$) and highest ($\tilde{\pi}_H = 0.0025$) plausible probabilities, we can calculate the logit $l_L = \log\left[\frac{\tilde{\pi}_L}{1-\tilde{\pi}_L}\right] = \log\left[\frac{0.00000012}{1-0.00000012}\right] = -6.9208$ and $l_H = \log\left[\frac{\tilde{\pi}_H}{1-\tilde{\pi}_H}\right] = \log\left[\frac{0.0025}{1-0.0025}\right] = -2.60\,097$; then the range is equal to range $= -2.6009 + 6.9208 = 4.3196$. Therefore, $\sigma_b^2 \cong (4.3196/6)^2 = 0.5184$. Based on a literature review, we decided to use a pool size of 10 plants per pool, $S_e = 0.999$, $S_p = 0.997$, $C = 20,000$ total budget for the study, $c_2 = 850$ cost of enrolling fields in the study, and $c_1 = 70$ cost of enrolling pools composed of $s = 10$ plants in the study. Next we obtained the required sample sizes for minimizing the variance, for

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

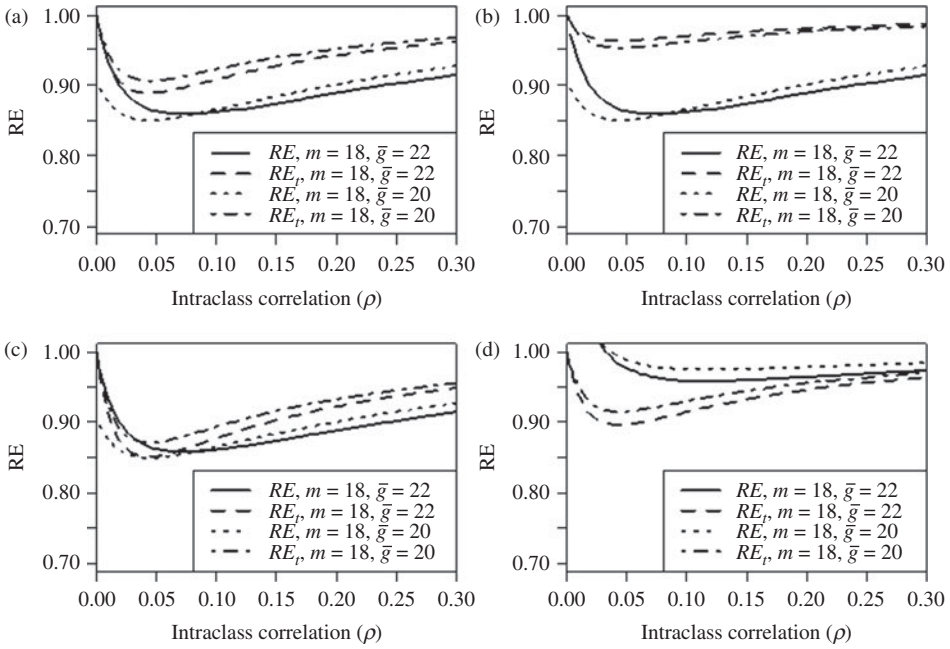

Figure 2. Relative efficiency of unequal versus equal cluster sizes as a function of the intraclass correlation (ρ) for four distributions of cluster size: (a) uniform, (b) unimodal, (c) bimodal and (d) positively skewed distribution.

achieving a certain width of the CI and for obtaining a certain power.

Minimizing the variance

Computing $\tilde{\pi}^p = 0.999 + (1 - 0.999 - 0.997)(1 - 0.0013)^{10} = 0.01587$ and $V(\delta) = \frac{(S_e - \tilde{\pi}^p)^{\frac{2}{s}-2}\tilde{\pi}^p(1-\tilde{\pi}^p)}{s^2(S_e+S_p-1)^{2/s}} = \frac{(0.999-0.01587)^{\frac{2}{10}-2}(0.01587)(1-0.01587)}{10^2(0.999+0.997-1)^{2/10}} = 0.000161$ results in

$$g = \sqrt{\frac{850}{70}}\frac{\sqrt{0.000161}}{0.0013(1-0.0013)\sqrt{0.5184}} = 47.32856$$

$$\approx 47$$

$$m = \frac{C}{c_2 + gc_1} = \frac{20000}{850 + (51)(70)} = 4.8042 \approx 5.$$

This means that we need to select five fields at random from the population of fields, with 47 pools in each field. Thus the total number of plants to select from each field is $g \times s = 47 \times 10 = 470$ plants, which will be allocated at random to form the 47 pools.

Now, if the cluster sizes are unequal, how do we compensate for the loss of efficiency due to varying cluster sizes? Assuming that the mean and standard deviation of cluster sizes are $\mu_g = 177$ and $\sigma_g = 81.5$, respectively, then $CV = \frac{81.5}{177} = 0.4605$, $\alpha = \frac{V(\delta)}{\sigma_b^{2*}} = \frac{0.000161}{\{0.0013(1-0.0013)\}^2 0.5184} = 5$, so $\lambda = (177/(177+5) = 0.9602$. Therefore, $RE_t = \{1 - (0.4605^2)(0.9725)(1 - 0.9725)\} = 0.9943$ and, for practical purposes, adjustment for unequal cluster sizes is not needed. However, to illustrate the method, full efficiency can be restored by taking $m = \frac{4.8042}{0.9943} = 4.8316 \approx 5$ clusters with $g = 47$ pools, and the new total budget will increase to $C^* = \frac{20000}{0.9943} = 20114.65$.

Specified 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 $W = (\tilde{\pi}_U - \tilde{\pi}_L) \leq \omega = 0.0025$. Therefore, $Z_{1-0.05/2} = 1.96$ and $V_0 = \omega^2/4Z^2_{1-\alpha/2} = \frac{0.0025^2}{4*1.96^2} = 0.000000401$. Using the same values of $s, S_e, S_p, \sigma_b^2, \tilde{\pi}, c_2$ and c_1 as given for minimizing the variance, equation (16) gives $g = \sqrt{\frac{850}{70}}\frac{\sqrt{0.000161}}{0.0013(1-0.0013)\sqrt{0.5184}} = 47$, while the number of clusters is equal to:

$$m = \frac{\left[\{0.0013(1 - 0.0013)\}^2(0.5184) + \frac{0.000161}{47}\right]}{0.000000401}$$

$$= 10.5802 \approx 11.$$

Since the value of g does not change, we need 470 plants per field, but now we need 11 fields to reach the required width of a 95% CI. However, this sample size is valid only for equal cluster sizes. If needed, adjustment for unequal cluster sizes is carried out by $m^* = \frac{m}{RE_t}$.

Therefore the budget has to be equal to $C = (47)(11)(70) + 11(850) = 45{,}540$. This implies that the required total budget for obtaining a 95% CI for estimating the proportion ($\tilde{\pi}$) with a desired width of 0.0025 is 2.264 times larger than the previous budget (20,114.65).

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

Now we determine the required number of clusters when there is no budget constraint, and assuming $g = 10$ (pools per cluster). Using equation (17) and assuming the same values of $\omega, s, \alpha, S_e, S_p, \sigma_b^2, \tilde{\pi}$ as were given for minimizing the variance, we have

$$m = \frac{4(1.96)^2}{0.0025^2} \times \left[\{0.0013(1-0.0013)\}^2(0.5184) + \frac{0.000161}{10}\right] = 41.7783 \approx 42.$$

This implies that we need a sample of 42 clusters, each containing 10 pools, assuming equal cluster size. Using unequal cluster sizes and assuming the same mean and standard deviation of cluster sizes, we need $m^* = \frac{41.7783}{0.9943} = 42.0178 \approx 43$ clusters. Of course, in this case, the total budget will be higher than the previously specified budget.

Specified power

Now suppose that we need to know the budget and sample size required for testing $H_0 : \tilde{\pi}_0 = 0.0013$ versus $H_1 : \tilde{\pi}_0 > 0.0013$ at an $\alpha = 0.05$ significance level with a power $(1-\gamma) = 0.9$ (90%) for detecting $\delta \geq 0.002$ and using the same parameter values ($s, S_e, S_p, \sigma_b^2, c_2$ and c_1) as before. Then, $V_0 = V_2 = \frac{0.002^2}{(1.645+1.282)^2} = 0.0000004671$. Since $V(\delta_0) = V(\delta) = 0.000161$, $\tilde{\pi} = \tilde{\pi}_0$, then $g = \sqrt{\frac{850}{70}} \frac{\sqrt{0.000161}}{0.0013(1-0.0013)\sqrt{0.5184}} \approx 47$, and the required number of clusters is equal to:

$$m = \frac{\left[\{0.0013(1-0.0013)\}^2(0.5184) + \frac{0.000161}{47}\right]}{0.000105} = 9.2136 \approx 10.$$

Here, too, we need 470 plants per field, but now we need 10 fields to reach the required power of 90%. To compensate for the unequal cluster sizes and assuming the same mean and standard deviation of the cluster sizes ($\mu_g = 177$ and $\sigma_g = 81.5$), we multiply $m = 9.2136$ by the correction factor (1/0.9943), which gives us $m^* = \frac{9.2136}{0.9943} = 9.2664 \approx 10$ clusters. Here the number of clusters remains the same due to rounding, but this is not always the case.

Here, also, the required budget is C = (10)(47)(70) + 10(850) = 41,400 which implies that the required total budget is 2.058 times larger than the budget for minimizing the variance of the proportion (20,114.65). However, this case guarantees a power of 90% for $\delta \geq 0.002$.

Now consider the problem without a budget constraint with 10 pools per cluster (g); solving for the required number of clusters (m) using the same values of $s, S_e, S_p, \sigma_b^2, \alpha, (1-\gamma)$, $\delta = \tilde{\pi}_1 - \tilde{\pi}_0$ as above, gives

$$m = \frac{(1.645+1.282)^2}{0.002^2} \times \left[\{0.0013(1-0.0013)\}^2(0.5184) + \frac{0.000161}{10}\right] = 44.6386.$$

This means that to perform the study, we need 45 clusters with 10 pools per cluster if the cluster sizes are equal, and $m^* = \frac{44.6386}{0.9943} = 44.8945 \approx 45$ clusters using unequal cluster sizes.

Tables for determining sample size

This section contains tables that help to calculate the optimal sample size. Table 2 gives the optimal allocation of clusters and pools when the goal is to estimate the proportion ($\tilde{\pi}$) with minimum variance using group testing with pool size ($s = 10$). The cost function is C = $mgc_1 + mc_2$ with C = 10,000, with six values of $\sigma_b^2 = 0.15, 0.25, 0.45, 0.65, 0.85, 1.05$; three values of $c_1 = 50, 100, 200$ and $c_2 = 800$. To illustrate how to use Table 2, assume that the proportion of interest is $\tilde{\pi} = 0.035$, and that the variance between clusters is $\sigma_b^2 = 0.25$. Assume the researcher estimates the cost of enrolling clusters in the study as $c_2 = 800$, that the cost of enrolling pools of size s = 10 is $c_1 = 100$ and the total budget for conducting the study is C = 10,000. Since in this case $c_1 = 100$, we will refer to the second subsection of Table 2. We find the value of $\tilde{\pi} = 0.035$ in the first column and the value of $\sigma_b^2 = 0.25$ in columns four and five. The values in the intersection between the value of $\tilde{\pi} = 0.035$ (first column) and the value of $\sigma_b^2 = 0.25$ (columns 4 and 5) are the optimal number of pools per cluster ($g = 11$) and the number of clusters ($m = 6$) required.

Table 3 gives the optimal allocations of clusters (m) and pools per cluster (g) to estimate $\tilde{\pi}$ with a certain width of the confidence interval under the cost function C = $mgc_1 + mc_2$, when $c_1 = 50$ and $c_2 = 800$ and significance level $\alpha = 0.05$. Three values of $\sigma_b^2 = 0.15, 0.25, 0.5$ form the three subsections of Table 3. To illustrate, assume that $\tilde{\pi} = 0.035$, $\sigma_b^2 = 0.25$, $c_1 = 50$, $c_2 = 800$, and the desired width of the confidence interval is equal to $\omega = 0.015$. The optimal m and g are obtained where the value of $\tilde{\pi} = 0.035$ (first column) intersects with the value of $\omega = 0.015$ (columns 6 and 7) in the second subsection corresponding to $\sigma_b^2 = 0.25$. Therefore the optimal numbers of pools per cluster (g) and of clusters (m) are 16 and 4, respectively.

Table 4 should be used when testing a hypothesis, that is, when we want to test $H_0 : \tilde{\pi} = \tilde{\pi}_0$ relative to

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

Table 2. Optimal sample sizes (g and m) for group testing for two stages given a pool size that minimizes the variance of the proportion ($\tilde{\pi}$)

	$\sigma_b^2 = 0.15$		$\sigma_b^2 = 0.25$		$\sigma_b^2 = 0.45$		$\sigma_b^2 = 0.65$		$\sigma_b^2 = 0.85$		$\sigma_b^2 = 1.05$	
$\tilde{\pi}$	g	m	g	m	g	m	g	m	g	m	g	m
					$c_1 = 50$ and $c_2 = 800$							
0.005	65	3	50	4	37	4	31	5	27	5	24	5
0.015	32	5	25	5	19	6	15	7	14	7	12	8
0.025	25	5	19	6	14	7	12	8	10	8	9	8
0.035	21	6	16	7	12	8	10	8	9	9	8	9
0.045	19	6	15	7	11	8	9	8	8	9	7	9
0.055	17	6	14	7	10	8	8	9	7	9	7	9
0.065	17	7	13	7	10	8	8	9	7	9	6	9
0.075	16	7	12	8	9	8	8	9	7	9	6	10
0.085	15	7	12	8	9	9	7	9	6	9	6	10
0.095	15	7	12	8	9	9	7	9	6	9	6	10
					$c_1 = 100$ and $c_2 = 800$							
0.005	46	2	35	3	26	3	22	4	19	4	17	4
0.015	23	4	18	4	13	5	11	6	10	6	9	7
0.025	17	4	13	5	10	6	8	7	7	7	7	7
0.035	15	5	11	6	9	7	7	7	6	8	6	8
0.045	13	5	10	6	8	7	6	7	6	8	5	8
0.055	12	5	10	6	7	7	6	8	5	8	5	8
0.065	12	6	9	6	7	7	6	8	5	8	4	9
0.075	11	6	9	6	6	7	5	8	5	8	4	9
0.085	11	6	8	7	6	7	5	8	5	8	4	9
0.095	11	6	8	7	6	8	5	8	4	9	4	9
					$c_1 = 200$ and $c_2 = 800$							
0.005	32	2	25	2	19	3	16	3	14	3	12	4
0.015	16	3	12	4	9	4	8	5	7	5	6	5
0.025	12	4	9	4	7	5	6	6	5	6	5	6
0.035	10	4	8	5	6	5	5	6	4	6	4	7
0.045	9	4	7	5	5	6	5	6	4	7	4	7
0.055	9	4	7	5	5	6	4	7	4	7	3	7
0.065	8	5	6	5	5	6	4	7	3	7	3	8
0.075	8	5	6	5	5	6	4	7	3	7	3	8
0.085	8	5	6	6	4	6	4	7	3	7	3	8
0.095	8	5	6	6	4	6	4	7	3	7	3	8

With pool size ($s = 10$), cost function $C = mgc_1 + mc_2$ with $C = 10{,}000$ and $c_2 = 800$. For ten values of $\tilde{\pi}$, three values of c_1 and six values of σ_b^2.

$H_1 : \tilde{\pi} > \tilde{\pi}_0$. Using this table, for six values of power (0.70. 0.75, 0.80, 0.85, 0.90, 0.95), significance level ($\alpha = 0.05$), pool size ($s = 10$), $c_1 = 50$ and $c_2 = 800, \sigma_b^2 = 0.5, \delta = 0.01, 0.03, 0.05$, and 10 values of $\tilde{\pi}$ from 0.005, 0.015 to 0.095 with increments of 0.01, we obtain the optimal allocations of clusters (m) and of pools per cluster (g) using the cost function $C = mgc_1 + mc_2$. To illustrate, assume that $\tilde{\pi} = 0.035$, $\sigma_b^2 = 0.5$, $c_1 = 50$, $c_2 = 800$, the desired power is $1 - \gamma = 0.8$, the significance level is $\alpha = 0.05$ and $\delta = 0.03$. Using the second subsection ($\delta = 0.03$), we find the value of $\tilde{\pi} = 0.035$ (first column) and $1 - \gamma = 0.8$ (columns 6 and 7) and at the point where they intersect, we find the required number of pools per cluster ($g = 11$) and the number of clusters ($m = 7$) needed to achieve a power of 80%.

Conclusions

In the present paper, we derived optimal sample sizes for group testing in a two-stage sampling process under a budget constraint. We assumed that the budget for enrolling individuals and clusters in the study is fixed and that we know the variance components. The optimal sample sizes were derived using Lagrange multipliers and produced formulae similar to the methods of Brooks (1955), Cochran (1977, p. 285) and Moerbeek *et al.* (2000) based on minimizing the error variance. This optimal allocation of clusters and pools was derived assuming equal cluster sizes, which are a good approximation when financial resources are scarce. However, since in practice the equality of cluster sizes is rarely satisfied,

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

Table 3. Optimal sample sizes (g and m) for confidence interval estimation using group testing in two stages given a pool size

	$\omega = 0.01$		$\omega = 0.03$		$\omega = 0.015$		$\omega = 0.07$		$\omega = 0.09$		$\omega = 0.11$	
$\tilde{\pi}$	g	m	g	m	g	m	g	m	g	m	g	m
					$\sigma_b^2 = 0.15$							
0.005	65	3	65	2	65	2	65	2	65	2	65	2
0.015	32	16	32	2	32	2	32	2	32	2	32	2
0.025	25	35	25	4	25	2	25	2	25	2	25	2
0.035	21	61	21	7	21	3	21	2	21	2	21	2
0.045	19	93	19	11	19	4	19	2	19	2	19	2
0.055	17	133	17	15	17	6	17	3	17	2	17	2
0.065	17	171	17	19	17	7	17	4	17	3	17	2
0.075	16	221	16	25	16	9	16	5	16	3	16	2
0.085	15	278	15	31	15	12	15	6	15	4	15	3
0.095	15	333	15	37	15	14	15	7	15	5	15	3
					$\sigma_b^2 = 0.25$							
0.005	50	4	50	2	50	2	50	2	50	2	50	2
0.015	25	22	25	3	25	2	25	2	25	2	25	2
0.025	19	50	19	6	19	2	19	2	19	2	19	2
0.035	16	89	16	10	16	4	16	2	16	2	16	2
0.045	15	134	15	15	15	6	15	3	15	2	15	2
0.055	14	189	14	21	14	8	14	4	14	3	14	2
0.065	13	255	13	29	13	11	13	6	13	4	13	3
0.075	12	331	12	37	12	14	12	7	12	5	12	3
0.085	12	406	12	46	12	17	12	9	12	6	12	4
0.095	12	487	12	55	12	20	12	10	12	7	12	5
					$\sigma_b^2 = 0.5$							
0.005	35	7	35	2	35	2	35	2	35	2	35	2
0.015	18	35	18	4	18	2	18	2	18	2	18	2
0.025	13	86	13	10	13	4	13	2	13	2	13	2
0.035	11	154	11	18	11	7	11	4	11	2	11	2
0.045	10	237	10	27	10	10	10	5	10	3	10	2
0.055	10	327	10	37	10	14	10	7	10	5	10	3
0.065	9	446	9	50	9	18	9	10	9	6	9	4
0.075	9	565	9	63	9	23	9	12	9	7	9	5
0.085	8	725	8	81	8	29	8	15	8	9	8	6
0.095	8	873	8	97	8	35	8	18	8	11	8	8

With pool size ($s = 10$), cost function $C = mgc_1 + mc_2$ subject to $V(\hat{\pi}) = \omega^2/4Z^2_{1-\alpha/2}$ with $c_1 = 50$, $c_2 = 800$ and significance level $\alpha = 0.05$. For ten values of $\tilde{\pi}$, six values of the expected width of the CI (ω), and three values of σ_b^2.

we derived a correction factor (inverse of the relative efficiency) to adjust the optimal sample sizes under equal cluster sizes. It is important to point out that this correction factor does not affect the number of required pools per cluster (g), but only the number of required clusters (m) and the total budget (C).

To determine the optimal sample sizes for equal or unequal cluster sizes, we started by specifying the needed power or precision; we then calculated $V(\hat{\pi})$ as well as the needed budget (C), and later obtained the optimal numbers of clusters (m) and pool per cluster (g) needed. This is extremely important because the researcher will usually plan his/her research in terms of power or precision under a budget constraint. The examples given show how the researcher can estimate the budget needed to reach the desired power or precision for the parameter estimate, equations (17) and (19) can be used for precision and power, respectively. However, the sample sizes given by equations (17) and (19) are not optimal, since the value of g is determined by the researcher according to his beliefs.

It is important to point out that the derived optimal sample sizes are approximate since they were obtained assuming that the proportion ($\tilde{\pi}$) is distributed approximately normal. This produces poor coverage for small sample sizes and also when the proportion ($\tilde{\pi}$) takes extreme values (near 0 and 1). For this reason, under simple random sampling, the exact or Pearson CI or the Wilson CI are preferred (Agresti and Coull, 1998; Agresti and Min, 2001; Brown *et al.*, 2001). Even in group testing it has been demonstrated that the best

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

Table 4. Optimal sample sizes (g and m) for power estimation using group testing in two stages given a pool size

	$1-\gamma=0.70$		$1-\gamma=0.75$		$1-\gamma=0.80$		$1-\gamma=0.85$		$1-\gamma=0.90$		$1-\gamma=0.95$	
$\tilde{\pi}$	g	m	g	m	g	m	g	m	g	m	g	m
						$\delta=0.01$						
0.005	35	2	35	3	35	3	35	3	35	4	35	5
0.015	18	11	18	13	18	15	18	17	18	20	18	25
0.025	13	27	13	30	13	35	13	40	13	48	13	61
0.035	11	47	11	54	11	62	11	72	11	86	11	108
0.045	10	73	10	83	10	96	10	111	10	132	10	167
0.055	10	100	10	115	10	132	10	153	10	182	10	230
0.065	9	137	9	157	9	180	9	209	9	249	9	314
0.075	9	173	9	198	9	228	9	265	9	315	9	398
0.085	8	222	8	254	8	292	8	339	8	404	8	511
0.095	8	268	8	306	8	351	8	409	8	487	8	615
						$\delta=0.03$						
0.005	35	2	35	2	35	2	35	2	35	2	35	2
0.015	18	2	18	2	18	2	18	2	18	3	18	3
0.025	13	3	13	4	13	4	13	5	13	6	13	7
0.035	11	6	11	6	11	7	11	8	11	10	11	12
0.045	10	9	10	10	10	11	10	13	10	15	10	19
0.055	10	12	10	13	10	15	10	17	10	21	10	26
0.065	9	16	9	18	9	20	9	24	9	28	9	35
0.075	9	20	9	22	9	26	9	30	9	35	9	45
0.085	8	25	8	29	8	33	8	38	8	45	8	57
0.095	8	30	8	34	8	39	8	46	8	55	8	69
						$\delta=0.05$						
0.005	35	2	35	2	35	2	35	2	35	2	35	2
0.015	18	2	18	2	18	2	18	2	18	2	18	2
0.025	13	2	13	2	13	2	13	2	13	2	13	3
0.035	11	2	11	3	11	3	11	3	11	4	11	5
0.045	10	3	10	4	10	4	10	5	10	6	10	7
0.055	10	4	10	5	10	6	10	7	10	8	10	10
0.065	9	6	9	7	9	8	9	9	9	10	9	13
0.075	9	7	9	8	9	10	9	11	9	13	9	16
0.085	8	9	8	11	8	12	8	14	8	17	8	21
0.095	8	11	8	13	8	15	8	17	8	20	8	25

With pool size ($s=10$), cost function $C=mgc_1+mc_2$ subject to $V(\hat{\pi})=\frac{|\delta|^2}{(Z_{1-\alpha}+Z_{1-\gamma})^2}$ with $c_1=50$, $c_2=800, \sigma_b^2=0.5$ and significance level $\alpha=0.05$. For ten values of $\tilde{\pi}$, six values of power $(1-\gamma)$ and three values of δ.

options for CI are the Exact and the Wilson CI (Tebbs and Bilder, 2004). For this reason, Montesinos-López, *et al.* (2010) proposed sample sizes for pooled data that guarantee narrow confidence intervals under simple random sampling. However, when the data are clustered it is not appropriate to use these sample size values and it is not possible to obtain exact confidence intervals (as Pearson type). For this reason, the analysis and sample size determination of binary data is usually performed under a generalized linear mixed model framework, which is accepted worldwide since it produces consistent parameter estimates. It is also true that when maximum likelihood is used, the parameter estimates are better than when a Taylor Series Expansion is employed. It is important to point out that since our data are clustered and the response variable is binary under group testing, the variance of the proportion is composed of between and within group variances and both components are affected by the proportion. This is in agreement with the results obtained by Candel and Van Breukelen (2010) in a non-group testing context.

For the reasons above, our optimal sample sizes were derived using a first-order TSE approach under the assumption that the variance components are known. Therefore, it is expected that the optimal sample sizes will be biased, which is supported for several Monte Carlo simulations for estimating fixed and random effects and determining optimal sample size for clustered randomized trials (Goldstein and Rasbash, 1996; Moerbeek *et al.*, 2001b; Candel and Van Breukelen, 2010). Even with the limitations of the proposed method, it is a valuable contribution to the planning of sample size for clustered data under group testing, since it produces an optimal allocation of the required number of clusters and pools given

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

budget constraint. Furthermore, the formulae for sample size determination are easy to use. However, more research is required to study further the proposed optimal sample sizes method.

Conflicts of interest

None.

References

Agresti, A. and Coull, B. (1998) Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician* **52**, 119–126.

Agresti, A. and Min, Y. (2001) On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **57**, 963–971.

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

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

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Brooks, S.H. (1955) The estimation of an optimum subsampling number. *Journal of the American Statisitical Association* **50**, 398–415.

Brown, L., Cai, T. and DasGupta, A. (2001) Interval estimation for a binomial proportion. *Statistical Science* **16**, 101–133.

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. *Statistics in Medicine* **29**, 1488.

Candel, M.J.J.M., Van Breukelen, G.J.P., Kotova, L. and Berger, M.P.F. (2008) Optimality of unequal cluster sizes in multilevel studies with small sample sizes. *Communications in Statistics: Simulation and Computation* **37**, 222–239.

Candy, S.G. (2000) The application of generalized linear mixed models to multi-level sampling for insect population monitoring. *Environmental and Ecological Statistics* **7**, 217–238.

Carreón-Herrera N.I. (2011) Detección de transgenes en variedades nativas de maíz en dos regiones del estado de Puebla. Unpublished doctoral dissertation, Colegio de Postgraduados, Campus Puebla, Puebla, Mexico. Available at: http://www.biblio.colpos.mx:8080/xmlui/bitstream/handle/10521/439/Carreon_Herrera_NI_MC_EDAR_2011.pdf?sequence=1 (accessed 20 February 2014).

Chen, P., Tebbs, J. and Bilder, C. (2009) Group testing regression models with fixed and random effects. *Biometrics* **65**, 1270–1278.

Cochran, W.G. (1977) *Sampling techniques* (3rd edition). New York, Wiley.

Dodd, R.Y., Notari, E. and 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**, 975–979.

Dorfman, R. (1943) The detection of defective members of large populations. *The Annals of Mathematical Statistics* **14**, 436–440.

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

Goldstein, H. (2003) *Multilevel statistical models* (3rd edition). London, Edward Arnold.

Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A* **159**, 505–513.

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

Landavazo Gamboa, D.A., Calvillo Alba, K.G., Espinosa Huerta, E., González Morelos, L., Aragón Cuevas, F., Torres Pacheco, I. and Mora Avilés, M.A. (2006) Caracterización molecular y biológica de genes recombinantes en maíz criollo de Oaxaca. *Agricultura Técnica en México* **32**, 267–279.

Lohr, S.L. (2008) Coverage and sampling. pp. 239–263 *in* de Leeuw, E.D.; Hox, J.J.; Dillman, D.A. (Eds) *International handbook of survey methodology*. New York, Lawrence Erlbaum Associates.

Maas, C.J.M. and Hox, J.J. (2004) Robustness issues in multilevel regression analysis. *Statistica Neerlandica* **58**, 127–137.

Moerbeek, M. (2006) Power and money in cluster randomized trials: when is it worth measuring a covariate? *Statistics in Medicine* **25**, 2607–2617.

Moerbeek, M., van Breukelen, G.J. and Berger, M.P. (2000) Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics* **25**, 271–284.

Moerbeek, M., Van Breukelen, G.J. and Berger, M.P. (2001a) Optimal experimental designs for multilevel logistic models. *Journal of the Royal Statistical Society: Series D (The Statistician)* **50**, 17–30.

Moerbeek, M., van Breukelen, G.J.P. and Berger, M.P.F. (2001b) Optimal experimental designs for multilevel models with covariates. *Communications in Statistics, Theory and 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 Science Research* **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 Science Research* **21**, 235–246.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) *Introduction to the theory of statistics* (3rd edition). New York, McGraw-Hill.

Peck, C. Going after BVD. *Beef* **42**, 34–44.

Rabe-Hesketh, S. and Skrondal, A. (2006) Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**, 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.

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

Rodríguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A* **158**, 73–89.

Skrondal, A. and Rabe-Hesketh, S. (2007) Redundant overdispersion parameters in multilevel models for categorical responses. *Journal of Educational and Behavioral Statistics* **32**, 419–430.

Stroup, W.W. (2012) *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton, Florida, CRC Press.

Tebbs, J. and Bilder, C. (2004) Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. *Journal of Agricultural, Biological, and Environmental Statistics* **9**, 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. *Statistics in Medicine* **26**, 2589–2603.

Van Breukelen, G.J.P., Candel, M.J.J.M. and Berger, M.P.F. (2008) Relative efficiency of unequal cluster sizes for variance component estimation in cluster randomized and multicentre trials. *Statistical Methods in Medical Research* **17**, 439–458.

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 Medicine and International Health* **3**, 747–750.

Wolf, J. (1985) Born-again group testing: multi access communications. *IEEE Transactions on Information Theory* **31**, 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. *Communications in Statistics – Simulation and Computation* **36**, 949–957.

Zhang, B., Bilder, C. and Tebbs, J. (2013) Regression analysis for multiple-disease group testing data. *Statistics in Medicine* **32**, 4954–4966.

Appendix A: Derivation of the optimal solution for minimizing $V(\hat{\pi})$ subject to $C = mgc_1 + mc_2$ ($c_l > 0, m, g \geq 2, l = 1, 2$)

By combining equations (12) and (13), we obtain the Lagrangean

$$L(m, g, \lambda) = L = V(\hat{\pi}) + \lambda[C - (mgc_1 + mc_2)] \quad (14)$$

where $V(\hat{\pi}) = \frac{\tilde{g}\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{\tilde{g}m} + \frac{V(\delta)}{m\tilde{g}}$. λ is the Lagrange multiplier. The partial derivatives of equation (14) with respect to λ, m and g are

$$\frac{\partial L}{\partial \lambda} = 0 = C - (mgc_1 + mc_2); \text{ then } m = \frac{C}{c_2 + gc_1}$$

$$\frac{\partial L}{\partial g} = 0 = -\frac{V(\delta)}{g^2 m} - \lambda m c_1; \text{ then } \lambda = -\frac{V(\delta)}{g^2 m^2 c_1}$$

$$\frac{\partial L}{\partial m} = 0 = -\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m^2} - \frac{V(\delta)}{m^2 g} - \lambda[gc_1 + c_2]$$

$$\Leftrightarrow \frac{V(\delta)}{g^2 m^2 c_1}[gc_1 + c_2] = \frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m^2} + \frac{V(\delta)}{m^2 g},$$

$$\text{since } \lambda = -\frac{V(\delta)}{g^2 m^2 c_1}$$

$$\Leftrightarrow V(\delta)gc_1 + V(\delta)c_2 = g^2 c_1\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}\right],$$

$$\Leftrightarrow V(\delta)c_2 = g^2 c_1\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$$

$$\Leftrightarrow g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}$$

Appendix B: Derivation of the optimal solution for minimizing $C = mgc_1 + mc_2$ subject to $V(\hat{\pi}) = V_0$

By combining equations (12) and (13), we obtain the Lagrangean

$$L(m, g, \lambda) = L = mgc_1 + mc_2 + \lambda[V(\hat{\pi}) - V_0] \quad (14)$$

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

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

$$m = \left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}\right]/V_0$$

$$\frac{\partial L}{\partial g} = 0 = mc_1 - \lambda\frac{V(\delta)}{g^2 m}; \text{ then } \lambda = \frac{g^2 m^2 c_1}{V(\delta)}$$

$$\frac{\partial L}{\partial m} = 0 = gc_1 + c_2 - \frac{\lambda}{m^2}\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}\right]$$

$$\Leftrightarrow gc_1 + c_2 = \frac{g^2 m^2 c_1}{V(\delta)}\left[\frac{\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2}{m^2} + \frac{V(\delta)}{m^2 g}\right],$$

$$\text{since } \lambda = \frac{g^2 m^2 c_1}{V(\delta)}$$

$$\Leftrightarrow gc_1 + c_2 = \frac{g^2 c_1}{V(\delta)}\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}\right],$$

$$\Leftrightarrow c_2 = \frac{g^2 c_1}{V(\delta)}\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2$$

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

$$\Leftrightarrow g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}.$$

Appendix C: Alternative derivation of the optimal solution for minimizing $C = mgc_1 + mc_2$ subject to $V(\hat{\pi}) = V_0$

If the sampling budget is C, the allocation of units as given in equation (15) results in a minimal value of $V(\hat{\pi})$ which in terms of cost is equal to:

$$V(\hat{\pi}) = \frac{c_2 + gc_1}{C}\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}\right] \qquad \text{(C1)}$$

where $g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}$. The solution to this problem can be derived directly since this budget C is also the minimal budget to obtain that particular value of $V(\hat{\pi})$. If there were a smaller budget with other allocations yielding the same $V(\hat{\pi})$, then our allocation (15) given C would not be optimal. This is true because the variable g appears in equation (C1) in the same manner as in equation (12), so that the g of equation (15) is also the value of g which will minimize the cost of the sample if the variance of the estimate of the proportion (p) is fixed. Thus it also minimizes the cost variance product (Brooks, 1955). Thus, if a value of $V(\hat{\pi})$ equal $V_0 = \omega^2/4Z_{1-\alpha/2}^2$ is required, the minimal budget to obtain this $V(\hat{\pi})$ follows by setting $V(\hat{\pi})$ (as given in equation C1) equal to $V_0 = \omega^2/4Z_{1-\alpha/2}^2$. Solving for budget C gives $C = (c_2 + gc_1)\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}\right]/V_0$ and, finally, the corresponding optimal allocation of units follows from equation (15). Since $m = \frac{C}{c_2+gc_1}$ and substituting $C = (c_2 + gc_1)\left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}\right]/V_0$, we obtain $m = \left[\{\tilde{\pi}(1-\tilde{\pi})\}^2\sigma_b^2 + \frac{V(\delta)}{g}\right]/V_0$ and $g = \sqrt{\frac{c_2}{c_1}}\frac{\sqrt{V(\delta)}}{\tilde{\pi}(1-\tilde{\pi})\sigma_b}$.

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

Taylor series approximation (equation 23) is derived from the RE of equation (21) 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 μ_n. Equation (19) is a moment estimator of

$$RE(\hat{\pi}) = \frac{\bar{g}+\alpha}{\bar{g}}E\left(\frac{U}{U+\alpha}\right) \qquad \text{(D1)}$$

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

Step 2

Define $d = (U - \mu_n)$; then the last term in equation (D1) can be written as:

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

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 $\rho(=$ small$\alpha)$. Thus we have:

$$E\left(\frac{U}{U+\alpha}\right) = E\left\{\left(\frac{\mu_n + d}{\mu_n + \alpha}\right)\sum_{j=0}^{\infty}\left(\frac{-d}{\mu_n + \alpha}\right)^j\right\}. \qquad \text{(D2)}$$

Step 3

If we ignore all terms d^j with $j > 2$ and rearrange terms in equation (D2), we will have

$$E\left(\frac{U}{U+\alpha}\right) = \lambda\{1 - CV^2\lambda(1-\lambda)\} \qquad \text{(D3)}$$

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 (D3) into (D1) gives:

$$RE(\hat{\pi})_t \approx \{1 - CV^2\lambda(1-\lambda)\}. \qquad \text{(D4)}$$

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press

Remark

Ignoring in (D2) only those d^j terms with $j > 4$ instead of 2 will give

$$RE(\hat{\pi})_t \approx 1 - \{(1-\lambda)[\lambda CV^2 - \lambda CV^3 \text{skew} + \lambda^3 CV^4(\text{kurt}+3)]\} \quad \text{(D5)}$$

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

https://doi.org/10.1017/S096025851400035X Published online by Cambridge University Press