Here we develop a statistical model for analyzing the data generated by a novel protocol for estimating the viability of algal cysts. First we will present an overview of the method. Then we will present a probabilistic model of the method, and subsequent statistical methods used to infer confidence intervals of germination frequency from the type of data generated by this method. Finally we will use this mathematical model to inform experimental design under various conditions.

1 Overview of the Method

The method consists of the following steps: 1) estimate the concentration (cysts/ml) of algal cysts in a mud sample using microscope counts, 2) pipette a small volume of mud into multiple wells in a multi-well plate along with seawater or growth media, 3) wait an appropriate amount of time for germination, and then count the number of wells with viable cells.

This method is much faster and less labor-intensive than traditional cyst isolation experiments, but the data it generates are less straight-forward to analyze due to the possibility of having less than or greater than 1 cyst per well (also the uncertainty on the concentration of cysts in the mud, but we'll get to that later). Because of this, a mathematical model needs to be developed to allow the valid interpretation of the type of data this method produces.

2 Forward Model

First we develop an appropriate "forward" model describing the probabilistic mechanisms at play.

In our idealized framework, each of n wells receive a volume V of mud containing a concentration C of cysts. On average, each well will receive CV cysts, but in reality, the number of cysts in a given well, x, will be a Poisson distributed random variable with $\lambda = CV$.

$$P(x) = \frac{\lambda^x}{x!}e^{-\lambda} \tag{1}$$
$$\lambda = CV$$

Every cyst then has a probability g of germinating. Thus the probability that no cells germinate in a well given the number of cysts in it, $P(0|x)$, is:

$$P(0|x) = (1-g)^x \tag{2}$$

We can combine this information and solve for the likelihood that a well will have no cells germinate in it, $P(0)$.

$$P(0) = \sum_{x=0}^{\infty} P(0|x)P(x) = \sum_{x=0}^{\infty} (1-g)^x \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda g} \tag{3}$$

The total number of wells with no germinated cells, k, will be distributed as a binomial distribution.

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4}$$

$$p = e^{-\lambda g} \tag{5}$$

3 Backward Model

In the last section we showed how the number of wells with no germinated cells, k, is binomially distributed with $p = e^{-\lambda g}$. In this section we'll use this information to build estimates of g based on k.

3.1 Simple Method

Since k is binomially distributed, we might try using well-known methods for estimating confidence intervals on p for a binomial distribution, and then transforming these estimates of p into estimates of g using $p = e^{-\lambda g}$. If you observe k successes in n trials, the most likely estimate, $\hat{p}$ is given by:

$$\hat{p} = \frac{k}{n} \tag{6}$$

while the $\alpha\%$ confidence intervals are usually given as:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{7}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution. Transforming these estimates using $p = e^{-\lambda g}$, we come up with:

$$\hat{g} = -\frac{\ln(k/n)}{\lambda} \tag{8}$$

$$g \text{ C.I.} = -\frac{1}{\lambda} \ln\left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \tag{9}$$

This solution will not work for all cases. Since $0 < k < n$, for small k, it is possible to have $\hat{g} > 1$, which is impossible. However, for large k and λ, it will be approximately correct.

3.2 Bayesian Estimation

We can use Bayesian Estimation to turn our previous expression for the probability of seeing k given g (and λ and n) into an expression for the probability of g given that we saw k. From our previous derivation we have:

$$P(k|g) = \binom{n}{k}(e^{-\lambda g})^k(1-e^{-\lambda g})^{n-k} \tag{10}$$

Next we need to assume prior distributions for g and k. Here we will assume that the prior distributions of both g and k are uniform, with g uniformly distributed between $0 \leq g \leq 1$, and k uniformly distributed between, $0 \leq k \leq n$, reflecting the possible ranges of each variable. Using Bayes' Theorem and these assumed priors:

$$P(g|k) \propto (e^{-\lambda g})^k(1-e^{-\lambda g})^{n-k} \tag{11}$$

To convert from a proportional relation to a proper PDF we need to solve for a normalization factor.

$$\begin{gathered}\int_0^1 (e^{-\lambda g})^k(1-e^{-\lambda g})^{n-k}dg = -\frac{1}{\lambda}\left[\beta_{e^{-\lambda}}(a,b) - \beta_1(a,b)\right],\ k \neq 0 \qquad (12)\\ a = k \\ b = n-k+1\end{gathered}$$

where $\beta_x(a,b)$ is the incomplete beta function.

$$\beta_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt \tag{13}$$

Note that this expression is only valid for $k \neq 0$. While it is possible to derive an exact solution for the normalization factor (and the CDF for that matter) in the case where $k = 0$, in practice it is more efficient (and accurate) to numerically integrate in this special case.

Having solved for the normalization factor, we can right a proper probability density function for g given n, k, and λ.

$$\mathrm{PDF}(g|k,\lambda,n) = \frac{-\lambda(e^{-\lambda g})^k(1-e^{-\lambda g})^{n-k}}{\beta_{e^{-\lambda}}(a,b) - \beta_1(a,b)},\ k \neq 0 \tag{14}$$

Integrating the PDF yields the CDF:

$$\mathrm{CDF}(g|k,\lambda,n) = \frac{\beta_{e^{-\lambda g}}(a,b) - \beta_1(a,b)}{\beta_{e^{-\lambda}}(a,b) - \beta_1(a,b)} = \frac{\beta\mathrm{CDF}(e^{-\lambda g},a,b) - 1}{\beta\mathrm{CDF}(e^{-\lambda},a,b) - 1},\ k \neq 0 \tag{15}$$

where βCDF is the cumulative density function of the beta distribution with shape parameters a and b. We can make the substitution because $\beta_x(a,b) = \beta\mathrm{CDF}(x,a,b) \cdot \beta_1(a,b)$.

We can also solve for the inverse cumulative distribution function, or quantile distribution function (QDF), which for a given probability, p, returns the g for which $\mathrm{CDF}(g) = p$.

$$p = \frac{\beta_{e^{-\lambda g}}(a,b) - \beta_1(a,b)}{\beta_{e^{-\lambda}}(a,b) - \beta_1(a,b)} = \frac{\beta\mathrm{CDF}(e^{-\lambda g},a,b) - 1}{\beta\mathrm{CDF}(e^{-\lambda},a,b) - 1} \tag{16}$$

Solving for an analytic expression

$$\mathrm{QDF}_g(p|k,\lambda,n) = -\frac{1}{\lambda}\ln\left(\beta\mathrm{QDF}\left(c,a,b\right)\right) \tag{17}$$

$$c = p \cdot \beta\mathrm{CDF}(e^{-\lambda},a,b) - p + 1 \tag{18}$$

Due to the technical limits of floating point computation, for small k and λ it is easier to solve for the QDF directly using a computational method like a bisection search, rather than using the analytic expression in Equation 17.

However we solve for it, the Bayesian QDF derived above provides a more robust method of estimating confidence intervals on g than the earlier simple method. The middle of the posterior distribution is given by $\mathrm{QDF}_g(0.5)$, while the 95% confidence interval will be defined by $\mathrm{QDF}_g(0.025)$ and $\mathrm{QDF}_g(0.975)$.

3.3 Incorporating Uncertainty on λ

We have been treating $\lambda = CV$ as a fixed and known quantity. In reality there is uncertainty on λ, both because the concentration of cysts is estimated using microscope counts, and hence has some uncertainty associated with it, and because the actual volume of mud placed in each well will also have some variance depending on the type of pipette used. In most cases the uncertainty on C will be the dominant source of error, so it's the only one I'll consider here. However, I'll add that the approach I use to account for the uncertainty on C could easily incorporate an estimate of the uncertainty on V if this quantity was known and significant.

We estimate the concentration of cysts C, by counting N_{mud} cysts in V_{mud} volume. The posterior PDF of C given this knowledge is:

$$\mathrm{PDF}(C|N_{mud},V_{mud}) = \Gamma(C \cdot V_{mud}, \alpha = N_{mud}, \beta = 1) \tag{19}$$

where $\Gamma(x, \alpha, \beta)$ is the PDF of the Gamma distribution with shape parameter α and inverse scale parameter β.

Formally incorporating the uncertainty on λ requires us to perform a convolution which is mathematically difficult. A more tractable method for approximately incorporating this information is to draw m random λ from the distribution of λ given N_{mud} and V_{mud} and then treat this sample as the true distribution of λ. Performing the convolution yields:

$$P(g|k, n, \lambda_1, \ldots, \lambda_m) \propto \sum_{i=1}^{m} (e^{-\lambda_i g})^k (1 - e^{-\lambda_i g})^{n-k} \tag{20}$$

Solving for the normalization factor:

$$\int \sum_{i=1}^{m} (e^{-\lambda_i g})^k (1 - e^{-\lambda_i g})^{n-k} dg = \sum_{i=1}^{m} -\frac{1}{\lambda_i} \beta_{e^{-\lambda_i g}}(a, b) \tag{21}$$

$$\text{PDF}(g|k, n, \lambda_1, \ldots, \lambda_m) = \frac{\sum_{i=1}^{m} (e^{-\lambda_i g})^k (1 - e^{-\lambda_i g})^{n-k}}{\sum_{i=1}^{m} \frac{1}{\lambda_i} \left[\beta_1\left(a, b\right) - \beta_{e^{-\lambda_i}}\left(a, b\right)\right]} \tag{22}$$

$$\text{CDF}(g|k, n, \lambda_1, \ldots, \lambda_m) = \frac{\sum_{i=1}^{m} \frac{1}{\lambda_i} [\beta_1(a, b) - \beta_{e^{-\lambda_i g}}(a, b)]}{\sum_{i=1}^{m} \frac{1}{\lambda_i} [\beta_1(a, b) - \beta_{e^{-\lambda_i}}(a, b)]} \tag{23}$$

As before, Equation 23 can be rewritten in terms of the CDF of the beta distribution.

3.4 Experiments Run Under Different Conditions

Suppose we run one series of wells using λ_1 which yields k_1 wells with no germinated cells, and another using λ_2 which yields k_2 wells with no germinated cells. It would be good if we had a way to combine the information from both of these experiments in order to provide a more precise estimate of g than either one can provide by itself.

Treating each series of wells as an independent trial:

$$P(k_1, k_2|g, \lambda_1, \lambda_2) \propto \left(e^{-\lambda_1 g}\right)^{k_1} \left(1 - e^{-\lambda_1 g}\right)^{n_1 - k_1} \left(e^{-\lambda_2 g}\right)^{k_2} \left(1 - e^{-\lambda_2 g}\right)^{n_2 - k_2} \tag{24}$$

As before, we need to integrate to find a normalization factor, but unfortunately an analytical solution is not easily found, so we leave the equations as integrals to be computed numerically.

$$\text{PDF}(g|k_1, \lambda_1, n_1, k_2, \lambda_2, n_2) = \frac{(e^{-\lambda_1 g})^{k_1} (1 - e^{\lambda_1 g})^{n_1 - k_1} (e^{-\lambda_2 g})^{k_2} (1 - e^{\lambda_2 g})^{n_2 - k_2}}{\int_0^1 (e^{-\lambda_1 g})^{k_1} (1 - e^{\lambda_1 g})^{n_1 - k_1} (e^{-\lambda_2 g})^{k_2} (1 - e^{\lambda_2 g})^{n_2 - k_2} dg} \tag{25}$$

More generally for t experiments:

$$\mathrm{PDF}(g|k_1,\lambda_1,n_1,\ldots,k_t,\lambda_t,n_t)=\frac{\prod_{i=1}^{t}(e^{-\lambda_i g})^{k_i}(1-e^{-\lambda_i g})^{n_i-k_i}}{\int_0^1\prod_{i=1}^{t}(e^{-\lambda_i g})^{k_i}(1-e^{-\lambda_i g})^{n_i-k_i}dg} \tag{26}$$

$$\mathrm{CDF}(g|k_1,\lambda_1,n_1,\ldots,k_t,\lambda_t,n_t)=\frac{\int_0^g\prod_{i=1}^{t}(e^{-\lambda_i g})^{k_i}(1-e^{-\lambda_i g})^{n_i-k_i}dg}{\int_0^1\prod_{i=1}^{t}(e^{-\lambda_i g})^{k_i}(1-e^{-\lambda_i g})^{n_i-k_i}dg} \tag{27}$$

Finally, if for t experiments one wishes to perform the approximate convolution drawing m λ from the posterior distribution of each λ_i, the PDF and CDF are:

$$\mathrm{PDF}(g|k_1,\vec{\lambda}_1,n_1,\ldots,k_t,\vec{\lambda}_t,n_t)=\frac{\sum_{j=1}^{m}\prod_{i=1}^{t}(e^{-\lambda_i^j g})^{k_i}(1-e^{-\lambda_i^j g})^{n_i-k_i}}{\int_0^1\sum_{j=1}^{m}\prod_{i=1}^{t}(e^{-\lambda_i^j g})^{k_i}(1-e^{-\lambda_i^j g})^{n_i-k_i}dg} \tag{28}$$

$$\mathrm{CDF}(g|k_1,\vec{\lambda}_1,n_1,\ldots,k_t,\vec{\lambda}_t,n_t)=\frac{\int_0^g\sum_{j=1}^{m}\prod_{i=1}^{t}(e^{-\lambda_i^j g})^{k_i}(1-e^{-\lambda_i^j g})^{n_i-k_i}dg}{\int_0^1\sum_{j=1}^{m}\prod_{i=1}^{t}(e^{-\lambda_i^j g})^{k_i}(1-e^{-\lambda_i^j g})^{n_i-k_i}dg} \tag{29}$$

4 Experimental Design

4.1 Optimal Efficiency

We can use this statistical model to inform the experimental design and increase the efficiency of this method. In order to be as efficient as possible, we should maximize dk/dg, that is we should design our experiment so that the expected number of wells with no germinated cells, k, changes rapidly with g. Since the expected number of wells with no germination scales with p, this is equivalent to maximizing dp/dg.

$$\frac{dp}{dg}=-\lambda e^{-g\lambda} \tag{30}$$

Maximizing:

$$\begin{aligned}\frac{d}{d\lambda}\left(\frac{dp}{dg}\right)&=e^{-g\lambda}(g\lambda-1)=0\\ g\lambda&=1\\ \lambda&=\frac{1}{g}\end{aligned} \tag{31}$$

So, for maximum efficiency one should setup the experiment such that $\lambda=1/g$.

4.2 Experimental Constraints and Precision

It may not be possible in all cases to follow the guideline that $\lambda = 1/g$. For instance, if the concentration of cysts or the germination probability are low, then following $\lambda = 1/g$ may require a larger volume of mud than can fit in an individual well. We may also be constrained by the total volume of mud we have available (for instance if analyzing slices from a sediment core). In this case what guidelines should we follow in order to maximize the precision of our method?

Let the total volume of mud distributed amongst the germination wells be given by V_{germ}.

$$V_{germ} = n \cdot V_{well} \tag{32}$$

Because $\lambda = V_{well} {\cdot} C$, the number of wells we run, n, and the number of expected germinations per well, λ, will be related as:

$$\lambda = \frac{V_{germ} C}{n} \tag{33}$$

The average number of wells with no germinated cells, $\hat{k}$, will be given as:

$$\hat{k} = n \cdot p = n e^{-\lambda g} \tag{34}$$

Plugging in $\hat{k}$ for k allows us to evaluate the most likely 95% confidence intervals for a hypothetical experiment.

Figure 1 shows the results of a series of hypothetical experiments where a fixed volume of mud is divided amongst different numbers of wells. While increasing n initially results in a more precise estimate of g, once $\lambda < 1/g$ the effect is negligible. After this point the only way to increase the precision is to run more mud. Therefore, under ideal conditions, the precision of the method is limited by the total number of (expected) cysts in all wells.

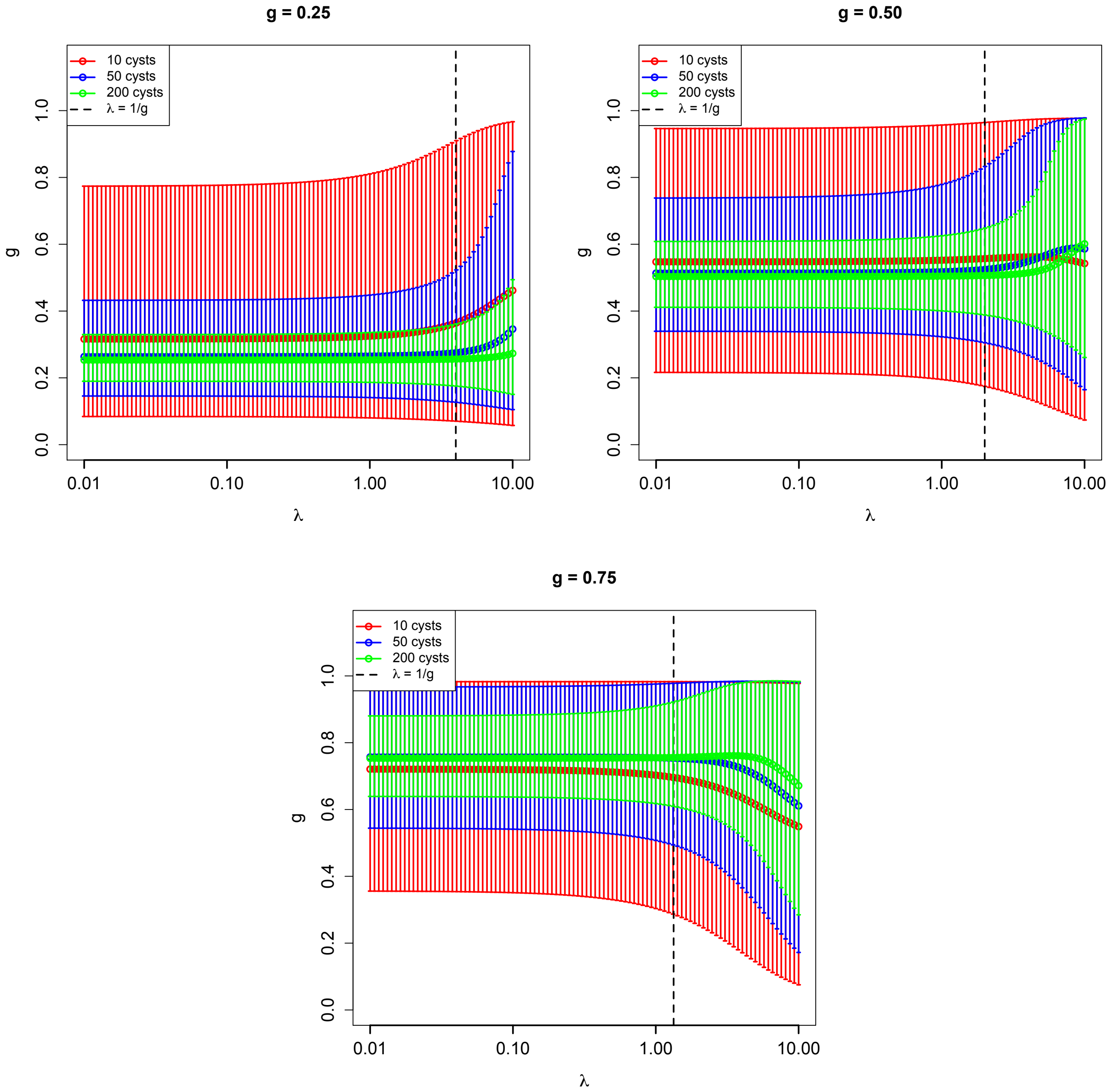

Figure 1: Hypothetical 95% confidence intervals on g where a fixed volume of mud is divided amongst n wells. λ is related to n by $\lambda = V_{germ}C/n$. Confidence intervals grow more precise as n increases and λ decreases, but the effect asymptotes on a scale proportional to $\lambda = 1/g$. For a fixed volume of mud, increasing n stops having an effect once $\lambda < 1/g$. After this point, in order to increase the precision one must increase the expected total number of cysts analyzed, i.e. run more mud.

4.3 Counting Cysts vs. Running Wells

The accuracy and precision of our estimate of g rely to a great extent on the accuracy of our knowledge of C, the concentration of cysts in the mud. There is therefore a tradeoff between allocating mud and labor towards cyst counts vs. running germination plates. Getting a good estimate of g relies on finding the proper balance between these two tasks.

Imagine we have a fixed volume of mud, V_{total}, which we can allocate between counting and germination plates such that:

$$V_{total} = V_{germ} + V_{count} \tag{35}$$

The expected number of cysts we will count, $\hat{N}_{mud}$, will be:

$$\hat{N}_{mud} = C \cdot V_{count} \tag{36}$$

and as before the expected number of wells with no germination, $\hat{k}$, will be:

$$\hat{k} = ne^{-\lambda g} \tag{37}$$

So, as in the previous section, we can plug these most likely values into our formulae, and see how the confidence intervals on g will be influenced by the proportion of mud we count vs. germinate.

Figure 2 shows how the tradeoff between these two activities influences the potential confidence intervals on g. While spending a disproportionate amount of time and mud on either one can result in an inefficient mismatch (uncertainty on one will dominate the total), using about 60% of the available mud on germination is often ideal, though the difference from 50-70% is usually minimal. The sweet spot, from an efficiency standpoint, is to keep the ratio of mud used for counting to mud used for germination plates in the neighborhood of 1:1 or 1:2

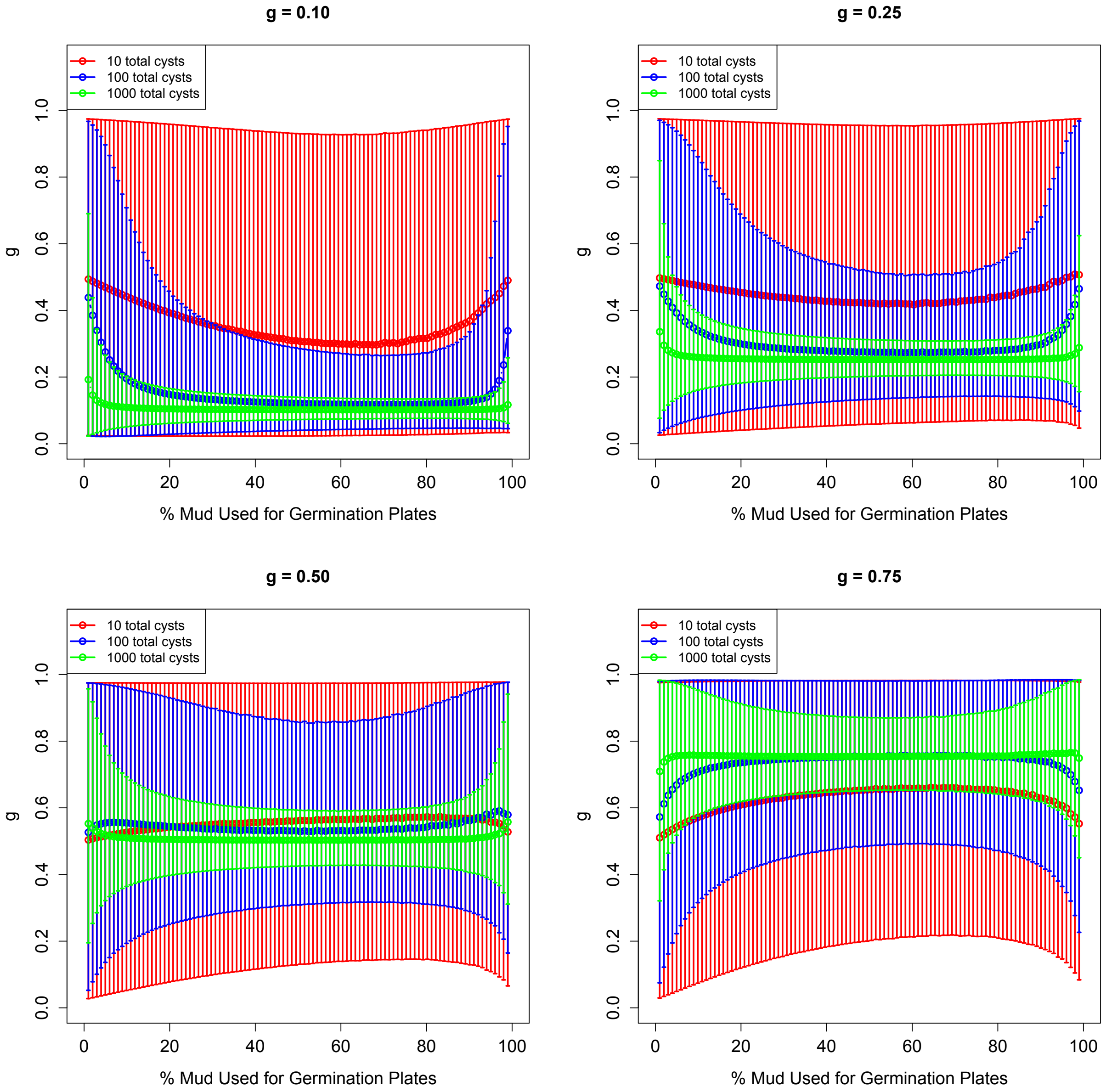

Figure 2: The tradeoff between measuring the concentration of cysts in the mud and performing germination assays. Performing one activity to the extreme detriment of the other can result in less precise estimates of g. In general the most efficient mix of the two is to focus about 60% of the available mud on germination assays, but the difference from 50-70% is usually minimal. Hypothetical confidence intervals were calculated using $\lambda = 0.5/g$, i.e. in a conservative manner to guarantee maximum precision.

4.4 Summary

To summarize the findings of this section:

1. The prime determinant of the precision of an estimate of g is the number of cysts likely included in the analysis ($V_{tot} \cdot C$). If analyzing the same volume of mud, a sample with higher concentrations of cysts will yield a more precise estimate of g than a sample with lower concentrations of cysts. Increasing the volume of mud included in the analysis will always increase the precision of the estimate.

2. There is a critical $\lambda = 1/g$. For the maximum precision per unit of mud used on germination assays set $\lambda \leq 1/g$, or equivalently $n \geq V_{germ}Cg$. In other words, use more wells than the total number of cysts that are likely to germinate.

3. Setting $\lambda = 1/g$ provides the maximum precision per well, since it balances precision per unit of mud with the most mud per well. However, depending on the total supply of mud, and whether or not this volume can fit in an individual well it may not be advantageous or possible to do this.

4. For the maximum precision per unit of mud analyzed, the ratio of the volume of mud used for microscope counts vs. germination assays should be between 1:1 and 1:2.

5 Standard Protocol

Informed by this exercise, we propose the following standard protocol.

1. Set aside 1/3 of the total sample volume for counting cysts and 2/3 for germination plates.

2. Estimate the concentration of cysts in the mud before running germination plates. Count 350 cysts or the total volume set aside for cyst counts, whichever comes first.

3. If the estimated concentration of cysts in the mud from these counts is < 50 cysts/ml, you will pipette $V_{well} = 0.02$ ml into each well of the germination plate. Otherwise set $V_{well} \leq 1/C_{cyst}$, or to the smallest volume that can be accurately pipetted, whichever is larger.

4. Plate $\sim$ 650 cysts or the entire volume of mud set aside for germination plates, whichever comes first.

5. If further accuracy is required and sample remains run additional well plates or count more cysts, trying to keep the ratio of volume used for counts vs. germination plates between 1:1 and 1:2.

6 Mixtures of g

This method implicitly assumes that all cysts in a given sample have the same germination probability g. If instead the cysts in a sample have some mixture of germination probabilities, then the g estimated by this method might not reflect the average germination probabilities of the cysts in the mixture.

We can examine the extent to which this is a problem by considering the extreme case where there is a mixture of two types of cyst with germination probabilities of 0% and 100%. Figure 3 shows how the C.I. on g compares to the population average $\hat{g}$. While there can be divergence between the two, it tends to occur when $\hat{g} > 1/\lambda$, i.e. outside the "safe" bounds established in the laboratory protocol. Considering the extreme nature of this test case, under normal circumstances estimates of g should reflect the population average and be unaffected by mixture effects.

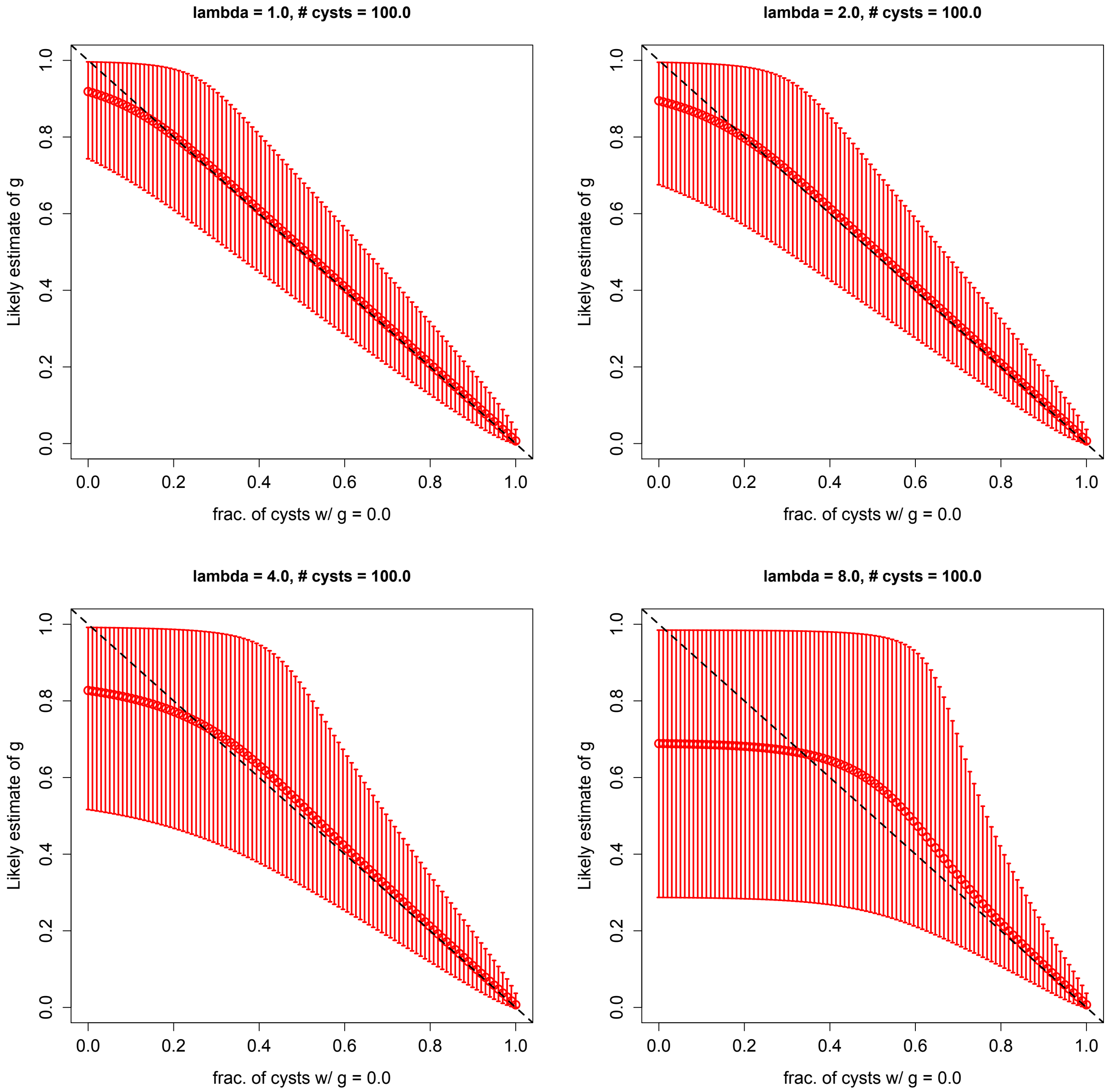

Figure 3: The results of mixed populations of cysts with different germination frequencies. The dotted line shows the average germination frequency of the population, while the red lines show the confidence intervals. While the estimates can deviate from the population average, they do so as $\hat{g} > 1/\lambda$.