Journal of Microbiological Methods, 18 (1993) 69–81

MIMET 00582

Efficient sampling designs for microbial processes: a case study

K.R. Johnson[a], R.W. Lundman[a] and M.A. Hamilton[b]

[a]*Center for Interfacial Microbial Process Engineering and* [b]*Department of Mathematical Sciences Montana State University, Bozeman, MT 59717, USA*

(Received 24 August 1992; revision received and accepted 12 March 1993)

Summary

We describe a method for determining optimal sampling designs for experiments involving subsampling in stages. An optimal sampling design either maximizes the precision of the results for a given amount of experimental effort (time, money) or minimizes the effort required to achieve a specified level of precision. The optimization method utilizes information on (i) the cost of obtaining a sample at each stage and (ii) the variability of the data attributable to each sampling stage. To illustrate the method, we use effluent suspended cell data from a packed-bed bioreactor. The sample design involved three stages. First, three primary samples were collected from the reactor. Second, each reactor sample was subsampled in triplicate and a specified volume of each subsample was passed through a filter. Third, the bacteria on the filter were counted in each of ten microscopic fields, each field representing the same proportion of the filter area. The optimization analysis, applied to the mean of the log-transformed bacterial counts, showed that the precision of the sample mean could be substantially improved by expending more sampling effort on the first stage and less on the second and third stages. We also show how the optimization analysis can be applied to other response measurements (e.g., total organic carbon, soluble organic carbon and glucose concentration) and sampling designs involving other numbers of subsampling stages.

Key words: Bioreactor; Experimental design; Random effects model; Subsampling; Variance component; Microbial process

Introduction

Microbiologists often summarize their data with a simple, but meaningful, statistic, e.g., a mean or a geometric mean of the observations. It is important in presenting the results to include an assessment of the uncertainty (variability) of the summary statistic. For example, when the summary statistic is the mean, the standard error of the mean is the usual measure of uncertainty. Experiments that produce highly

Correspondence to: M.A. Hamilton, Center for Interfacial Microbial Process Engineering, Montana State University, Bozeman, MT 59717, USA.

certain summary statistics are more informative.

The following strategy is useful for determining the experimental design that minimizes statistical uncertainty. First, one identifies all the elements of the experimental protocol that contribute to inherent variability in the experimental observations. These considerations lead to a list of important steps in the protocol. Next, one collects data to determine what proportion of the overall variability is contributed by each step. Then one focuses on each step that contributes significant variability and either modifies the experimental methodology to reduce uncertainty, or designs the experiment to take replicates at that step to reduce uncertainty by averaging. In this paper we use a case study of a bioreactor experiment to show how to apply this strategy.

Collecting data from bioreactors frequently involves subsampling in 'stages' or 'levels' [1]. An initial sample is collected directly from the reactor, and then one or more subsamples are obtained from this reactor sample for analysis. There may be any number of stages of subsampling, as subsamples from the reactor samples may themselves be subsampled, etc.

The efficiency of the experiment depends on the allocation of effort among the sampling stages; that is, how many reactor samples, how many subsamples from each reactor sample, etc. Specification of these numbers is called the *sampling design*. We define the *optimal sampling design* as the design that minimizes the statistical variance of the summary statistic for a given amount of experimental effort, as measured in units of time or money.

The purpose of this paper is to show how to apply conventional statistical analyses to determine an optimal sampling design. The optimization approach involves: (i) isolating and quantifying sources of variability in the data collection process; and (ii) estimating the costs associated with each stage of sampling. Mathematical formulas are presented for converting the information from (i) and (ii) into an optimal sampling design. The optimization method can be modified to determine the design that minimizes experimental effort among designs that have a specified small variance of the summary statistic.

We illustrate the optimization procedure using bulk fluid suspended cell data from a pilot run of a packed-bed bioreactor. The illustration is then extended by applying the procedure to three other response variables, each of which could be the target for optimization – total organic carbon (TOC) concentration, soluble organic carbon (SOC) concentration and glucose concentration measurements. The techniques presented are applicable across other experimental variables, reactors and types of experiments.

Materials and Methods

Experiment overview

The ultimate goal of the project was to verify an existing model of biological accumulation and cellular detachment in porous media. The reactor consisted of a glass cylinder (50 mm × 31 mm diameter) packed with 1 mm glass beads. A single species of bacteria, *Pseudomonas aeruginosa*, was grown in the reactor. Glucose, the sole carbon source, was fed to the system at a concentration of 18 mg l^{-1}. The reactor

was operated in a once-through, continuous, upflow manner with a liquid flow rate of 38.5 ml min^{-1}.

Sample collection and analysis

For this pilot run, samples were collected at approximately 24 h intervals from both the influent and effluent lines for twelve consecutive days. A 'subsampling' design (Fig. 1) was employed. For bulk fluid suspended cell counts, three stages of sampling were employed. First, three samples were pulled directly from both the influent and effluent lines. (We collected influent suspended cell data for a mass balance analysis of the reactor.) Second, each of these reactor samples was subsampled in triplicate and each subsample was filtered. Third, the cells on the filter were counted in each of ten random microscopic fields, each field of known, specified area. This plan required a total of $3 \times 3 \times 10 = 90$ microscopic field counts per day and location. Two sampling stages were employed for the three dissolved constituents (TOC, SOC and glucose). Three samples were pulled from each reactor line, and two subsamples were drawn from each for the analytic determination (total of $3 \times 2 = 6$ determinations per day and location). Sample volumes were 20 ml for cell counting and glucose concentrations and 10 ml for TOC and SOC concentrations.

The cell counting procedure consisted of double staining the bacteria with 4'-6-diamidino-2-phenylindol and Acridin orange, filtering the bacteria onto a black 0.2 μm polycarbonate filter and enumerating with epifluorescent microscopy [2].

Organic carbon concentrations were measured with a Dohrmann DC-80 Carbon Analyzer. All samples were adjusted to pH 2, using concentrated phosphoric acid, and purged with oxygen to remove inorganic carbon and carbon dioxide. TOC concentrations were measured directly from the reactor samples. SOC concentrations were measured from the supernate of reactor samples after centrifugation for 10

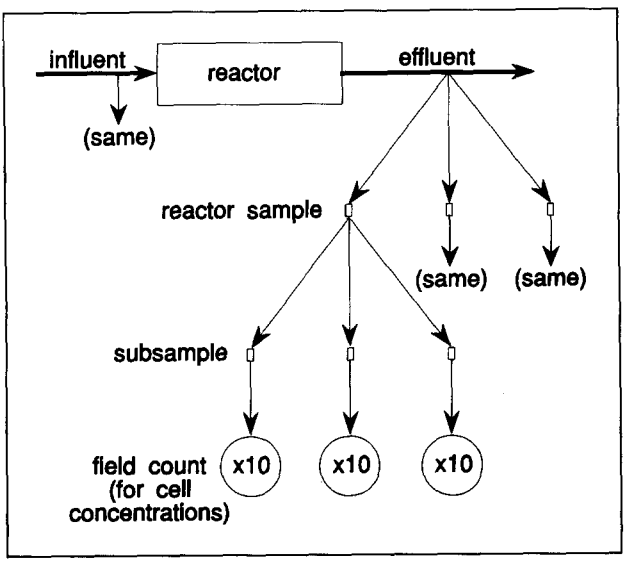

Fig. 1. Schematic of sample collection design.

min (10 000 rpm).

Glucose concentrations of filtered samples were measured colorimetrically using the Sigma Diagnostics Enzymatic Glucose Determination (Procedure 510). Samples were filtered through a 0.2 μm membrane to remove interfering particulates. Enzyme was reacted with the glucose to form oxidized ortho-dianisidine, which is brown in color. Sample absorbencies were measured with a Varian DMS 90 UV-Visible spectrophotometer. Sample glucose concentrations were estimated using a simple linear regression analysis developed from absorbencies of glucose standard solutions.

Statistical model

A conventional 'components of variance' model can be used to represent the data. We decided, for reasons presented below, that the data should be log-transformed to conform to the conventional model. Throughout the rest of this paper, the term 'microscopic field count' represents the $\log_{10}$-transform of the number of cells within the microscopic field of view.

The microscopic field count can be expressed as the true mean plus a random deviation from that mean, where the deviation is due to inherent variation contributed at each stage of sampling. In notation,

$$Y_{ijk} = \mu + D_{ijk}, \text{ where} \tag{1}$$

$Y_{ijk} = k^{th}$ microscopic field count ($\log_{10}$ transformed counts) from subsample j of the i^{th} reactor sample, μ is a constant, the true overall mean, and D_{ijk} is the associated random deviation.

The deviation D_{ijk} can be partitioned to isolate the various contributions to the variability of Y_{ijk} as

$$D_{ijk} = A_i + B_{j(i)} + \varepsilon_{k(ij)} \tag{2}$$

where A_i = deviation of the i^{th} reactor sample from the overall mean, $B_{j(i)}$ = deviation of the j^{th} subsample from the mean of the i^{th} reactor sample, and $\varepsilon_{k(ij)}$ = deviation of the k^{th} field count from the mean of the j^{th} subsample from reactor sample i, where i = 1,...,a; j = 1,...,b; and k = 1,...,n. For the experiment we conducted, a = number of reactor samples = 3, b = number of subsamples per reactor sample = 3, and n = number of microscopic field counts per subsample = 10.

In statistical science, Eq. (2) is called a 'components of variance' model for a completely nested design ([1], pp. 643–654). The standard model requires that the random variables, A_i, $B_{j(i)}$ and $\varepsilon_{k(ij)}$ are mutually uncorrelated with means equal to zero and variances denoted by σ_A^2, σ_B^2, and σ_ε^2, respectively.

The model also requires homogeneous variances, i.e., σ_ε^2 the variance of $\varepsilon_{k(ij)}$, is the same for all $a \times b \times n$ combinations of i, j, and k. This latter requirement would be fulfilled if the variance of microscopic field counts within a subsample was essentially constant from subsample to subsample, even if the subsamples were from different reactor samples. The microscopic field counts (log-transformed counts) conform reasonably well to the homogeneous variance requirement. For the nontransformed

counts, however, the variances are heterogeneous in that the sample variance is larger whenever the sample mean is larger. This is often the case with such count data ([1], pp. 615–616).

The model and standard requirements imply that an individual microscopic field count has mean equal to the constant, μ, and variance equal to the sum of the three 'variance components.' In notation, the variance is

$$\begin{aligned} \mathrm{Var}[Y_{ijk}] &= \sigma_A^2 + \sigma_B^2 + \sigma_\varepsilon^2 \\ &= \text{variance among reactor samples} + \text{variance among subsamples} \\ &\quad + \text{variance among microscopic field counts} \end{aligned} \tag{3}$$

The corresponding variance of the observed overall sample mean can be derived as a weighted sum of the variance components. Denoting the overall sample mean as $\overline{Y}$, and the variance of the sample mean as $var(\overline{Y})$,

$$var(\overline{Y}) = \frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{ab} + \frac{\sigma_\epsilon^2}{abn} \tag{4}$$

To put the mathematics in perspective, Eqs. (1)–(3) define a statistical model for the observable quantity Y_{ijk}. The components of the model, μ, A_i, $B_{j(i)}$, $\varepsilon_{k(ij)}$, σ_A^2, σ_B^2 and σ_ε^2 are conceptual, not directly measurable. The goal of the statistical analysis is to estimate μ. Statistical theory shows that the sample mean is a good estimator. Statistical theory also shows that an analysis of variance (ANOVA) applied to the Y_{ijk}'s can be used to estimate σ_A^2, σ_B^2, and σ_ε^2, the unknown constituents of the variance of the sample mean (4). The model has made it possible to attach a variance to the sample mean, and by comparing the relative sizes of the estimated σ_A^2, σ_B^2, and σ_ε^2, to determine which sampling stage contributes most to that variance.

Data analysis

An analysis of variance (ANOVA) was performed on the data for each of the twelve sampling days. The analysis was conducted on a desktop computer using Procedure NESTED in the SAS statistical software package [3]. The same analysis could have been accomplished with other statistics software packages. It is possible, but tedious, to do the calculations without special software, by following the formulas in Snedecor and Cochran ([6], pp. 285–288), for example. (We feel that this analysis is sufficiently complicated that the assistance of a statistician would be helpful.)

The SAS printout included estimates of the overall mean, μ, and the variance components, σ_A^2, σ_B^2, and σ_ε^2. The estimator of μ is the overall sample mean, $\overline{Y}$. The formulas for estimating $var(\overline{Y})$ and for calculating confidence intervals for μ are given in Neter et al. ([1], p. 992). The estimators are known to have good statistical properties [4].

Sample design optimization

Let T denote the total sampling cost for a specific sampling design. Then

$$T = aC_A + abC_B + abnC_N \tag{5}$$

where C_A = cost of a reactor sample, C_B = cost of a subsample, including filter, and C_N = cost of counting bacteria in a microscopic field. A standard calculus technique known as the method of Lagrange multipliers [5] can be used to determine the sample design that either attains the minimal $var(\overline{Y})$ within a fixed cost T or achieves the minimal cost T among designs for which the associated $var(\overline{Y})$ is no larger than specified [6]. (Many scientists prefer to describe variability with the standard error of the mean rather than the variance. Because the standard error is the square root of the variance of $\overline{Y}$, the smallest possible standard error occurs for the same sampling design that produces the smallest possible variance.)

The Lagrange solutions for the number of samples at the last two stages of a three-stage sampling design are

$$b = \sqrt{\frac{\sigma_B^2}{\sigma_A^2}\frac{C_A}{C_B}} \tag{6}$$

and

$$n = \sqrt{\frac{\sigma_\epsilon^2}{\sigma_B^2}\frac{\sigma C_B}{C_N}} \tag{7}$$

To minimize $var(\overline{Y})$ for specified T, set the right-hand side of (5) equal to that T, substitute the b and n from Eqs. (6) and (7), then solve for a. To minimize T for specified $var(\overline{Y})$, use the same approach with equation (4). These solutions for a are, respectively,

$$a = \frac{T}{C_A + \sqrt{\frac{\sigma_B^2}{\sigma_A^2}C_A C_B} + \sqrt{\frac{\sigma_\epsilon^2}{\sigma_A^2}C_A C_N}} \tag{8}$$

and

$$a = \frac{\sigma_A^2 + \sqrt{\sigma_A^2\sigma_B^2\frac{C_B}{C_A}} + \sqrt{\sigma_A^2\sigma_\epsilon^2\frac{C_N}{C_A}}}{var(\bar{Y})} \tag{9}$$

For a two-stage sampling design, the total cost is given by

$$T = aC_A + anC_N \tag{10}$$

The variance of the sample mean is

$$var(\overline{Y}) = \frac{\sigma_A^2}{a} + \frac{\sigma_\epsilon^2}{an} \tag{11}$$

The optimal number of replicates for the second stage is

$$n = \sqrt{\frac{\sigma_\epsilon^2}{\sigma_A^2}\frac{C_A}{C_N}} \tag{12}$$

To find the optimal number of primary samples, substitute n from Eq. (12) into either

Eq. (10) or (11), depending on whether total cost or the variance has been specified, and solve for a. These solutions for a are, respectively,

$$a = \frac{T}{C_A + \sqrt{\frac{\sigma_\epsilon^2}{\sigma_A^2}}\, C_A C_N} \tag{13}$$

and

$$a = \frac{\sigma_A^2 + \sqrt{\sigma_A^2 \sigma_\epsilon^2 \frac{C_N}{C_A}}}{var(\bar{Y})} \tag{14}$$

These equations for optimizing sampling designs are presented in statistical texts (e.g. [6]). The solutions provided by Eqs. (6)–(9) and (12)–(14) will usually not be whole numbers; hence, they must be treated as approximate optimal sample sizes.

Cost analysis

We used the estimated data collection costs presented in Table 1 for the sampling design optimization. The cost assessment was conducted by one of us (R.W.L.), based on his perspective as a graduate research assistant at the time the pilot experiment was conducted. The costs are provided to illustrate the methodology; they are not necessarily relevant to any other laboratory setting.

Results

Variance components estimates

Variance components estimates for the suspended cells data, expressed as a percentage of the total variance, are presented in Fig. 2. In general, variability among the reactor samples and among the field counts for a given subsample were the primary sources of variability in the data. There was little variability among subsamples of a given reactor sample.

Fig. 3 provides visual support to the calculated percentages shown in Fig. 2. In Fig. 3, each row of data points shows the ten microscopic field counts for a subsample; the three subsamples for each reactor sample are grouped together. The data points were randomly jittered vertically to separate overlapping points slightly. Within a reactor

TABLE 1

Cost analysis for collection of bulk fluid suspended cell data

Sample level	Materials	Labor (min)	Total cost ($)
Reactor	Syringe, glassware, preservative	3	0.80
Subsample	Microscope slide/slip, filter, stain, filtration equipment rental	4	1.60
Microscopic field	Microscope rental, UV light	2	0.80

Assumptions: microscope rental fee: \$10 h^{-1}; filtration equipment rental fee: \$ 2 h^{-1}; labor fee: \$14 h^{-1}.

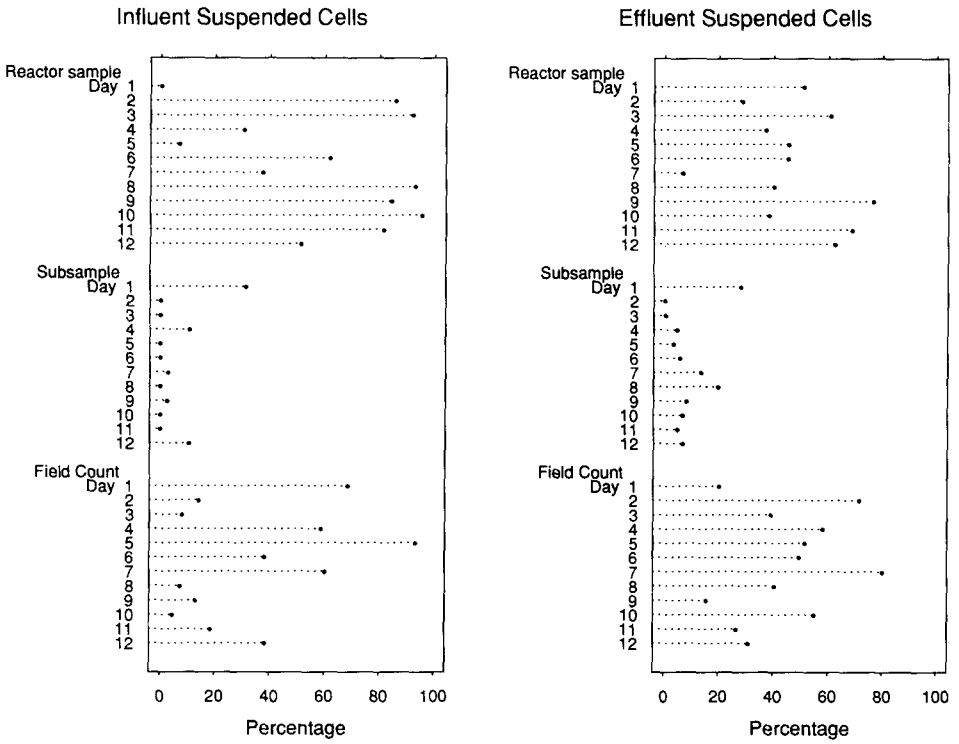

Fig. 2. Percentage of the total variation of Y_{ijk} attributable to each of the three sampling stages, estimated separately for each day. For any particular day, the three percentages sum to 100.

sample, approximate by eye the mean of each row of subsample points. Note that the three subsample means are not far apart on the horizontal scale. Now approximate by eye the mean of the thirty points (ten points for each of three subsamples) for each reactor sample. Note that these reactor sample means are further apart than is typical for the three subsample means. This visual analysis shows, as did the ANOVA calculations, that the means for the subsamples within a reactor sample vary relatively little; i.e., σ_B^2 is small.

Optimal sample design

All sample design optimization calculations in this paper are based on the estimates of σ_A^2, σ_B^2, and σ_ε^2 shown in Table 2. As indicated by Fig. 2, we calculated a set of variance components estimates for each observation day. These variance estimates differed from day to day; the differences were due in part to having few observations on which to base day-specific estimates. To increase the reliability of each variance component estimate, we averaged the eight corresponding variance components calculated for the final eight days of observation (when the reactor had reached 'steady-state' with respect to bulk fluid suspended cells). In most applications, such averaging will not be possible because there will be only one set of variance component estimates.

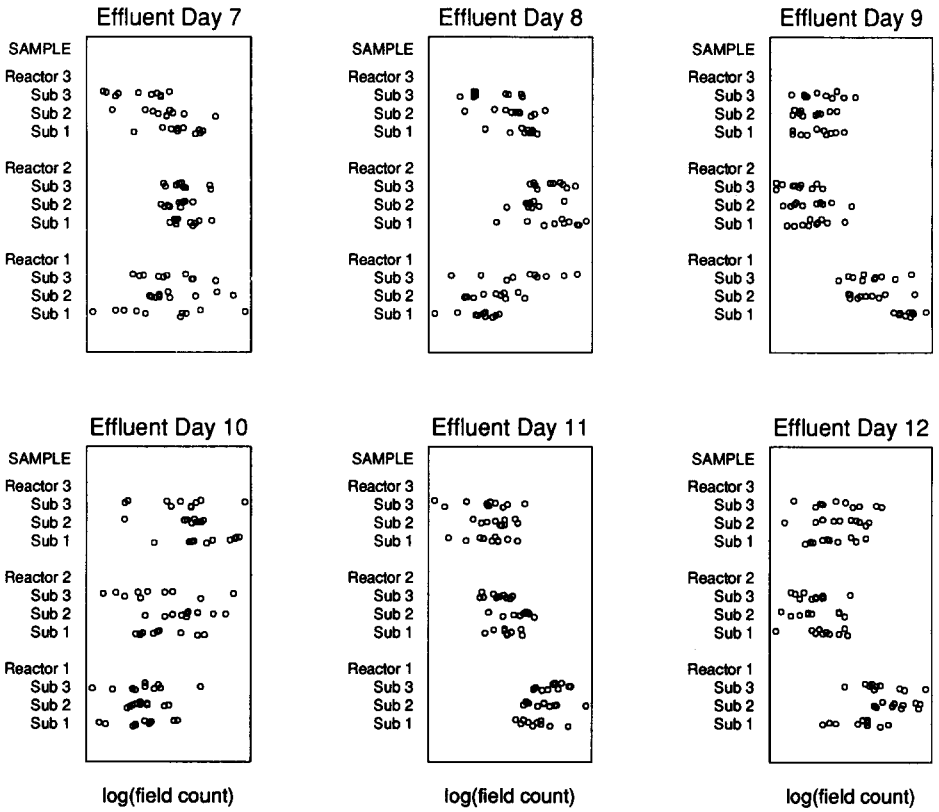

Fig. 3. Individual microscopic field counts within subsamples. The points have been jittered slightly in the vertical direction to separate overlapping points.

To illustrate the optimization calculations, we chose target specifications based on the pilot experiment for which $T = 88.8$ and $var(\overline{Y}) = 0.005592$. First, we shall find the design minimizing $var(\overline{Y})$ among designs with total cost less than 88.8. Second, we shall find the design minimizing the total cost among designs for which $var(\overline{Y})$ is no larger than 0.005592.

The optimization began by substituting the variance components estimates (Table

TABLE 2

Variance components estimates based on data from days 5–12

Sample stage	Variance component parameter	Variance component estimate	% of total variance
Reactor	σ_A^2	0.01552	51.7
Subsample	σ_B^σ	0.00258	8.6
Field count	$\sigma_\varepsilon^\sigma$	0.01192	39.7

2) and the costs for each sample stage (Table 1), into Eqs. (6)–(7). The results are $b = 0.29$ and $n = 3.0$. Ordinarily, the solutions to Eqs. (6)–(7) could be rounded to whole numbers to give the optimal design; e.g., because the experiment requires at least one subsample (at least one prepared filter) per reactor sample, we could round up and set $b = 1$. In this instance, however, the extremely small result $b = 0.29$ indicated that our model could be simplified. In essence, there was so little variation attributable to subsampling that replication of subsamples is not necessary. Therefore, for purposes of the optimization analysis, we consolidated the filter preparation step into the reactor sampling step. This consolidation, in effect, reduced the number of sampling stages to two, (i) take a reactor sample and prepare one filter and (ii) observe microscopic field counts. The variance attributable to stage (i) is $\sigma_A^2 + \sigma_B^2$. Using the entries in Table 2, the variance estimate was $0.01552 + 0.00258 = 0.01810$. The variance for stage (ii), σ_ε^2, was estimated to be 0.01192. Similarly, the cost for stage (i) was $0.80 + 1.60 = 2.40$ and for stage (ii) was 0.80 (Table 1). For the reduced two-stage model, Eq. (12) was solved using these variances and costs. The result was $n = 1.4$. (This solution must be converted to a whole number, either 1 or 2. We have chosen $n = 2$ to demonstrate the procedure, but one should carry out parallel calculations for $n = 1$.)

Setting T in Eq. (13) equal to 88.8, substituting $n = 2$, and solving, we found $a = 22.2$. Setting $var(\bar{Y})$ in Eq. (14) equal to 0.005592, substituting $n = 2$, and solving, we found $a = 4.3$. The designs $a = 22, n = 2$ and $a = 5, n = 2$ are candidates for optimal designs of the two-stage type, in which the first stage entails both sampling from the reactor and preparing a filter. For the original three-stage design, in which sampling from the reactor and filter preparation are modeled as distinct steps, the equivalent designs are $a = 22, b = 1, n = 2$ and $a = 5, b = 1, n = 2$.

Table 3 shows characteristics (total cost, $var(\bar{Y})$, standard error of the sample mean, and total time) of optimal three-stage sampling designs and of the design used for the pilot study. The design $a = 22, b = 1, n = 2$ is Example 1; the design $a = 5, b = 1, n = 2$ is Example 2; and $a = 6, b = 1, n = 1$ is Example 3. The variance of the mean shown in Table 3 was found by substituting the variance components estimates

TABLE 3

Examples of optimal and near optimal sampling designs

Example	(1) Sample design	(2) Total cost (T)	(3) Variance $var(\bar{Y})$	(4) Standard error	(5) Total time
Pilot study	3,3,10	88.8	0.005592	0.075	225
1	22,1,2	88.0	0.001094	0.033	242
2	5,1,2	20.0	0.004812	0.069	55
3	6,1,1	19.2	0.005003	0.071	54

Col. (1): Three values are a, b, n, which are the numbers of reactor samples, subsamples per reactor sample, and microscopic field counts per subsample, respectively.

Col. (2): Found by substituting a, b, n, and the costs from Table 1 into Eq. (5).

Col. (3): Found by substituting a, b, n, and the variances from Table 2 into Eq. (4).

Col. (4): Square root of column 3.

Col. (5): Total time in minutes; based on the time estimates in Table 1.

and sample sizes into Eq. (4). The standard error is the square root of the variance. Total cost was calculated using Eq. (5). For Example 1, the variance of the sample mean is (0.01552/22) + (0.00258/22) + (0.01192/44) = 0.001094. The standard error of the sample mean is $\sqrt{0.001094} = 0.033$, which is substantially smaller than the standard error for the pilot study design, 0.075. The total cost is 22(0.80) + 22(1.60) + 44(0.80) = 88.0, which is less than the specified 88.8.

We entered Eqs. (4) and (5) in a computer spread sheet and evaluated T and $var(\overline{Y})$ for all combinations of a, b, and n that achieve either the specified $T = 88.8$ or the specified $var(\overline{Y})$ of 0.005592. These calculations covered all the possibilities arising when the solutions to the optimizing equations are rounded to whole numbers. The results of the spread sheet analyses showed that Example 1 is truly the optimal sampling design for specified $T = 88.8$. Example 3 is actually the optimal sampling design for specified $var(\overline{Y}) = 0.005592$. But, because taking duplicate microscopic field counts per filter is good experimental practice, we prefer the design of Example 2, which is almost as effective as Example 3. Notice that the design of Example 2 produces an estimate as reliable as that for the pilot study design (S.E. of 0.069 versus 0.075), but the total cost is much less, 20 compared to 88.8. These examples indicate the gain in experimental efficiency that can be achieved by concentrating sampling effort on the major source of variability, reactor samples.

Dissolved constituents

We have illustrated the design strategy using the effluent suspended cell data, where the sampling design involved three stages. The strategy can be applied if other constituents are of interest or if the sampling design involves a different number of stages.

Two stages of sampling were executed for the dissolved constituents (TOC, SOC and glucose): reactor samples and replicate subsamples. For each of these three response variables, the primary source of variability was in sampling from the reactor; there was very little variation among subsamples from a given reactor sample. Typically, the reactor sample component of variance was 90% of the total variance for glucose concentrations and nearly 100% of the total variance for SOC and TOC concentrations. Replicate reactor samples, but not replicate subsamples, are required to decrease the uncertainty of the estimated mean concentrations.

Discussion

The methods illustrated in this article can make a practical difference. The case study was an experiment concerning cellular detachment in a bioreactor. Efficient sample designs were quite different from the sampling strategy originally suggested when the project was conceived. Admittedly, the method requires some initial extra effort in observing replications at each sample step, estimating sampling costs and conducting the statistical analysis. But such initial extra effort is required to determine the design that will attain the most reliable results possible within time and budget constraints.

For the cellular detachment project, the components of variance analysis showed that the first sampling step (reactor sample) contributed significant variability to the

final result. In such a situation, the variance of the sample mean can be reduced only by taking replicates at that first step. This conclusion is evident from Eq. (4) which shows that, unless σ_A^2 is negligible, large values for a are necessary to reduce the variance of the mean. Increasing subsample replicate numbers (b and n in the case study) will not affect the σ_A^2 component of the variance of the overall mean. Moreover, if σ_A^2 is not negligible, inference about the overall mean is accomplished with statistical procedures (e.g., t-tests or F-tests) that use $a - 1$ as the relevant degrees of freedom ([1], p. 992). In this sense, the effective sample size is a, not the total number of observations $a \times b \times n$.

Regardless of the sampling design, it is important to assess correctly the uncertainty in the results. In the bioreactor case study, for example, if one wished to compare effluent cell count means for two different experimental conditions, the correct assessment of the variance of the mean might well include a component for run-to-run variability. Replicate runs of the bioreactor under the same experimental conditions would be required to estimate that variance component. (A run-to-run component of variability was not included in our case study because our goal was to show how to find the optimal sampling design for a single run.)

If the investigator fails to include an important variance component when calculating the variance of the result, the result will appear more reliable than it really is. This type of error commonly occurs when the levels of subsampling are nested and the variance component due to the primary level sample is not recognized. In this circumstance, the investigator probably will take replicate subsamples, mistaking them for the effective sample size. Such replicates have been called *pseudoreplicates* [7].

Although replicate subsamples proved ineffective for reducing the uncertainty of the overall mean in our case study, replicates may well be useful for routine monitoring of laboratory procedures and instrumentation. For example, the TOC analyzer can be checked for stability by comparing readings of duplicate TOC subsamples. On the other hand, such monitoring can reasonably be based on replicate reactor samples (single subsamples), although the sensitivity is reduced.

The optimization method requires two types of quantitative information for the sampling stages, (i) estimates of the variance components and (ii) costs. Variance components estimates can be calculated from historical data, pilot experimentation, or the first run of the main experiment. It may be difficult to assign accurate sampling costs because necessary data (e.g., deprecation costs for instruments) are not always readily available. It may be worthwhile to do multiple analyses using various cost assignments to determine the sensitivity of the optimal design.

The Lagrange optimization method does not take into account the fact that the number of samples at any level must be a positive integer; the formulas can give fractional sample numbers. Two approaches can be used to obtain integer-valued sample numbers. First, a formal, mathematical integer-optimization technique may be substituted for the Lagrange multiplier method. This technique gives an exact solution, but is iterative in nature and usually is accomplished with a computer program. We used a second approach, which is to conduct calculations of the total cost and the variance of the mean for integer-valued sample sizes near the Lagrange multiplier solution. Solutions thus obtained may not be exact; nonetheless, substan-

tial improvements in sampling design efficiency may still be realized. A computer spreadsheet program can facilitate trial-and-error calculations.

We recommend that experimenters routinely plot all data they submit to statistical analysis. A display such as Fig. 3 provides not only a visual confirmation of the analysis but also a quality control check on the data. In fact, our initial plots (not shown) displayed outliers that, upon checking, were data entry errors. We produced our plots using the statistical programming language S-PLUS [8].

Acknowledgment

The authors acknowledge support from Cooperative Agreement ECD-8907039 between the National Science Foundation and Montana State University and from the Industrial Associates of the Center for Interfacial Microbial Process Engineering. The perceptive questions and comments of the referees substantially improved the paper.

References

1 Neter, J., Wasserman, W. and Kutner, M.H. (1985) Applied Linear Statistical Models. Irwin, Homewood, IL.
2 Griebe, T. (1991) Experimentelle Untersuchungen zur Aggregatbildung. Diplomarbeit, Institut für Hydrobiologie und Fischereiwissenschaft, Universität Hamburg.
3 SAS/STAT™ User's Guide, Release 6.03 edn. (1988) SAS Institute, Cary, NC.
4 Searle, S.R. (1971) Linear Models, pp. 404–406, Wiley, New York, NY.
5 Boyce, W.E. and DiPrima, R.C. (1988) Calculus, pp. 861–868, Wiley, New York, NY.
6 Snedecor, G.W. and Cochran, W.G. (1967) Statistical Methods, Iowa State UP, Ames, IA. pp. 528–534.
7 Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments, Ecol. Monogr. 54, 187–211.
8 S-PLUS User's Manual, Version 3.0, 1991, Statistical Sciences, Seattle, WA.