NUREG/CR-4268
PNL-5156

Ratio Methods for Cost-Effective Field Sampling of Commercial Radioactive Low-Level Wastes

Prepared by L. L. Eberhardt, M. A. Simmons, J. M. Thomas

Pacific Northwest Laboratory
Operated by
Battelle Memorial Institute

Prepared for
U.S. Nuclear Regulatory Commission

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability of responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

NOTICE

Availability of Reference Materials Cited in NRC Publications

Most documents cited in NRC publications will be available from one of the following sources:

1. The NRC Public Document Room, 1717 H Street, N.W.
   Washington, DC 20555
2. The Superintendent of Documents, U.S. Government Printing Office, Post Office Box 37082,
   Washington, DC 20013-7082
3. The National Technical Information Service, Springfield, VA 22161

Although the listing that follows represents the majority of documents cited in NRC publications, it is not intended to be exhaustive.

Referenced documents available for inspection and copying for a fee from the NRC Public Document Room include NRC correspondence and internal NRC memoranda; NRC Office of Inspection and Enforcement bulletins, circulars, information notices, inspection and investigation notices; Licensee Event Reports; vendor reports and correspondence; Commission papers; and applicant and licensee documents and correspondence.

The following documents in the NUREG series are available for purchase from the NRC/GPO Sales Program: formal NRC staff and contractor reports, NRC sponsored conference proceedings, and NRC booklets and brochures. Also available are Regulatory Guides, NRC regulations in the *Code of Federal Regulations*, and *Nuclear Regulatory Commission Issuances*.

Documents available from the National Technical Information Service include NUREG series reports and technical reports prepared by other federal agencies and reports prepared by the Atomic Energy Commission, forerunner agency to the Nuclear Regulatory Commission.

Documents available from public and special technical libraries include all open literature items, such as books, journal and periodical articles, and transactions. *Federal Register* notices, federal and state legislation, and congressional reports can usually be obtained from these libraries.

Documents such as theses, dissertations, foreign reports and translations, and non-NRC conference proceedings are available for purchase from the organization sponsoring the publication cited.

Single copies of NRC draft reports are available free, to the extent of supply, upon written request to the Division of Technical Information and Document Control, U.S. Nuclear Regulatory Commission, Washington, DC 20555.

Copies of industry codes and standards used in a substantive manner in the NRC regulatory process are maintained at the NRC Library, 7920 Norfolk Avenue, Bethesda, Maryland, and are available there for reference use by the public. Codes and standards are usually copyrighted and may be purchased from the originating organization or, if they are American National Standards, from the American National Standards Institute, 1430 Broadway, New York, NY 10018.

NUREG/CR-4268
PNL-5156
RW

Ratio Methods for Cost-Effective Field Sampling of Commercial Radioactive Low-Level Wastes

Manuscript Completed: May 1985
Date Published: July 1985

Prepared by
L. L. Eberhardt, M. A. Simmons, J. M. Thomas

Pacific Northwest Laboratory
Richland, WA 99352

Prepared for
Division of Radiation Programs and Earth Sciences
Office of Nuclear Regulatory Commission
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555
NRC FIN B2461

ABSTRACT

In many field studies to determine the quantities of radioactivity at commercial low-level radioactive waste sites, preliminary appraisals are made with field radiation detectors, or other relatively inaccurate devices. More accurate determinations are subsequently made with procedures requiring chemical separations or other expensive analyses. Costs of these laboratory determinations are often large, so that adequate sampling may not be achieved due to budget limitations. In this report, we propose double sampling as a way to combine the expensive and inexpensive approaches to substantially reduce overall costs. The underlying theory was developed for human and agricultural surveys, and is partially based on assumptions that are not appropriate for commercial low-level waste sites. Consequently, extensive computer simulations were conducted to determine whether the results can be applied in circumstances of importance to the Nuclear Regulatory Commission. This report gives the simulation details, and concludes that the principal equations are appropriate for most studies at commercial low-level waste sites. A few points require further research, using actual commercial low-level radioactive waste site data. The final section of the report provides some guidance (via an example) for the field use of double sampling. Details of the simulation programs are available from the authors. Major findings are listed in the Executive Summary.

PREFACE

The results discussed in this interim report are based on a range of correlations, sample sizes, measurement error variances, statistical models and coefficients of variation that we believe might be encountered at commercial low-level radioactive waste sites. At this point in our research, we could not investigate all possibilities, nor explain all our observations. However, it is clear that the results to date are useful, and if applied, the methodology would result in more cost-effective estimates of mean or total quantities of radionuclides.

EXECUTIVE SUMMARY

The purpose of this study was to investigate a statistical technique known as double sampling as a cost-effective means to estimate the mean or total amount of radioactivity which might be found in the environs of a commercial low-level radioactive waste site. Double sampling is derived from a broader classification of methods called ratio estimation. In ratio estimation, the entire population of interest (e.g. a commercial low-level waste site) is measured using an inexpensive method, such as a portable radiation detector, so that a site mean or total (for radioactivity) can be estimated. This estimate by itself is often either not very accurate, unacceptable for regulatory purposes or of little use for health effects assessment. In order to obtain a more precise and cost-effective estimate, a subset of site samples may be analyzed by very accurate and usually expensive methods. The ratio of results obtained using the expensive and inexpensive methods on the same sample can then be used to estimate a site total or mean based on the measurements over the entire site using the inexpensive method.

Unless measurements for the entire site are "free" (i.e. already available), ratio estimation methodology may not be cost-effective. However, an alternative approach, called double sampling, has been developed for use in human and agricultural surveys, wherein only a large sample of the population must be measured using the inexpensive method. In this report, we have used computer simulation to investigate many of the assumptions, models, and sample size requirements involved in order to assess applicability to commercial low-level radioactive waste site needs.

The report is arranged so that a non-statistically inclined reader can use Section 1.0 to gain an understanding of ratio methods (by example), decide whether double sampling would be cost-effective (Section 2.0), and perform the needed computations (Section 7.0). Section 5.0 briefly discusses the research results and provides the basis which indicates that double sampling can be cost-effective for a variety of commercial low-level waste site problems. Section 3.0 provides the technical basis for the simulation results found in Section 4.0 and is written for the more statistically inclined reader. Some of our principal results are summarized below and are referenced to specific report sections to provide easy access for readers with a particular interest.

1. Double-sampling regression and ratio estimates (based on a selected range of models and parameters) resulted in acceptable estimates when confidence limits were restricted to plus or minus 40 percent of the mean. Thus, we conclude that double sampling is "robust" in that it

gives accurate results even when there are substantial departures from the theoretical models and parameter values upon which it was developed. It also appears that the technique will be very cost-effective at commercial low-level radioactive waste sites. However, since the results are based on computer simulations, field testing is essential. (Section 4.0 and page 41).

2. Confidence limits greater than plus or minus 40 percent for an estimated mean or total are probably useless for regulatory action. When it can be shown a priori that such a result is likely, field studies probably should not be conducted. (pages 21 and 41).

3. Double sampling may be useful when sampling to detect a spill or migration. The research on estimating a mean or total reported herein should be extended to double sampling for stratification to make double sampling applicable to a wider range of likely problems. (Section 6.0).

4. Double sampling becomes very cost-effective when the inexpensive and expensive methods are highly correlated and the ratio of costs (expensive/inexpensive) is very high. (page 6).

5. When the correlation between the expensive and inexpensive measurement is poor, additional research to improve the relationship may result in substantial long-term cost reductions. (page 5).

6. When inexpensive measurements have already been made over the entire population of interest (they are "free" of cost), ratio approaches should always be considered for applicability. (page 5).

7. When the cost of the expensive method is high relative to the inexpensive technique, double sampling is worthwhile even when the two measurements are poorly correlated. (page 6).

8. The ratio of small to large sample sizes studied in our research ranged from 0.025 to 0.20. It appears that ranges of from 0.015 to 0.55 can be encountered in field studies. Double sampling simulation studies using these ratios may be needed. (page 16).

9. When the inexpensive method includes background, extraneous noise, shielding or other statistical contamination, linear regression methods may be preferable to ratio estimates. (page 17).

10. Simulations using multiplicative errors for the inexpensive method resulted in about 2 percent more of the calculated means outside

theoretical confidence limits compared to limits based on additive normal errors often assumed to apply (usually an erroneous assumption for radionuclides). We conclude that this difference is not of practical significance. (pages 25-29).

11. Field data should be developed to define the relative magnitude of measurement errors for the inexpensive method (often a field instrument) for circumstances and radionuclides of interest at commercial low-level radioactive waste sites. Our simulations were based on errors we assumed to be "likely". Measured values are needed. (page 43).

12. The prospects of using deliberately selected samples for expensive analysis (from a random sample of the large sample selected for inexpensive evaluation) should be investigated. Considerable additional cost reduction may thus be possible. (page 43).

13. Research on mathematical derivations of appropriate double sampling equations is needed to avoid testing all possible values of parameters as well as additional plausible models. Such results would have much wider applicability. (page 43).

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

In an appraisal of statistical and sampling needs for environmental monitoring of commercial low level radioactive waste (CLLRW) disposal facilities, Eberhardt and Thomas (1983:pp. 4-11 and 4-12) pointed out the potential advantages of combining two methods of measuring concentrations of radionuclides in various kinds of samples to estimate totals or mean concentrations. Assessments using portable radiation detectors are quick and inexpensive, but ordinarily are not very accurate. More exacting procedures usually involve chemical separations and other kinds of expensive measurements. In the usual approach, the inexpensive methods are used to do a preliminary survey, which is then followed up by collection of samples for laboratory assays using chemical procedures. The two surveys are almost always done separately, and it is often concluded that the budget would not permit enough chemical separations for a desirable level of accuracy in the final survey. Modern sampling methods are available to combine the two approaches in order to achieve suitably accurate results in a cost-effective manner.

In the present report, we consider the use of "ratio methods" for improving sampling for commercial low-level radioactive wastes. These methods have been developed and widely used in sampling human and agricultural populations for a variety of purposes, but have been largely neglected in other fields. Unfortunately, this leaves some gaps in the underlying theory with respect to the applications of primary interest here. Also, the lack of experience with the method in waste management contexts leaves some uncertainties about the appropriate models for waste site problems. Thus, there is a need for field testing and field research on methodology. In this report, we provide a description of the methods and explore the theoretical problems by using computer simulations. Further investigations are needed to develop the full potential of the method for CLLRW applications.

There are several different approaches which can be identified by simple examples. The original method is known as ratio estimation. As an illustration, assume that a very large quantity of low-level waste has been accumulated in containers of quite different sizes in temporary storage, and that the individual containers were all accurately weighed, with the weights recorded in a card file. Also suppose that, as sometimes happens, it is later decided that an accurate estimate of the total quantity of some particular component radionuclide in the waste is needed. Assaying all of the containers is ordinarily not feasible, and a few must be selected by sampling. If the substance of concern is fairly evenly distributed through

the containers, a sensible approach is to assay a sample of containers and to expand the sample results to cover the total population of containers by using the known weights. A very simple equation can be used:

$$\hat{Y}_R = (\bar{y}/\bar{x})X_T \qquad (1.1)$$

Here $\hat{Y}_R$ denotes the estimated total inventory of the particular waste component being measured, $\bar{y}$ is the average quantity in the samples assayed, $\bar{x}$ is the average weight of those containers in the sample, and X_T is the total weight of all containers (obtained by summing the weights on the card file). Note that the weights used are the original values. If a random sample is used to assay current radioactivity in a sample of containers, changes in weight during storage will usually not have any significant effect on the outcome.

The term, ratio estimation, comes from the ratio used in eq. (1.1) above, and the approach implies the assumption that the underlying relationship will pass through the origin on a plot of the data points, i.e., if a container weight approaches zero, then the radioactivity burden of that container necessarily also approaches zero. In other circumstances, discussed below, this assumption may be unsatisfactory so that the alternative of fitting a straight line not restricted to passing through the origin may be utilized in regression estimation. Instead of the simple ratio of eq. (1.1), we now fit a regression line and use it to estimate a total inventory.

An undesirable feature of the current textbook approach is that the x-values (known also as auxiliary variables) are assumed to be measured with little or no error, as is possible in weighing or in making various kinds of counts (number of people in households, etc.). There are, however, many circumstances in which this assumption is unrealistic. A relevant example might be the case where it is desired to estimate the quantity of some specific radionuclide passing through an air sampler over a long time period. Usually, only gross radioactivity measurements are made on air sampler filters. If, however, more accurate determinations are desired, some small fraction of the many filters used may be subjected to more accurate analysis. One then could use the gross reading of the filters as an auxiliary variable [(x) in eq. (1.1)] to estimate a total based on the accurate method. In this case, it is unrealistic to consider X_T as known without error, since replicate readings of the same filter give results that differ due to "measurement errors" or "instrument errors". A few theoretical results have been given for this case (Cochran 1977:158-160).

In both of the above examples, a total is available for the entire population of interest. In the first case, the total is assumed to be known

without error, and in the second case, it may be subject to instrument errors, but is nonetheless based on a measurement made on every unit in the population. In a great many instances, it would be desirable to use the approach when a total is not available. One then takes two samples, one large sample measured using a relatively inexpensive method (the auxiliary variable, x) and a second much smaller sample on which x is measured but an expensive and accurate assay (y) is also done. The same procedure as before is followed, but instead of a known total one only has a large sample of the population. In survey sampling literature, the approach is known as double sampling while a quite different approach is widely recognized in quality control circles under the same cognomen.

A good example of a textbook model, without errors in either x or y, is the use of ratio estimation to estimate populations in years between the decennial censuses. If we know the population of all U.S. cities in 1980, then a relatively small sample of cities might be censused in 1985 and used to estimate the total urban population in that year. Since there is a very good correlation between the size of a city in one year and its size in a subsequent year, the method gives excellent results. It can be especially accurate if provisions are made to include disproportionately more of the largest cities, inasmuch as they contain a sizable fraction of the population. This suggests yet another possible effective use of ratio methods at CLLRW sites, i.e., deliberate selection of units to include in the sample.

It can realistically be supposed that both y and x are measured with little error in this example, so that the variation in the outcome is not associated with the measurement technique at all, but is instead due to differences in growth rates of cities. These chance fluctuations are then assumed to be independent of the actual measurement, (i.e., large and small cities can grow faster or slower independently of one another) and considered to be "additive", so that the underlying model is:

$$y_i = Rx_i + e_i \qquad (1.2)$$

where e_i denotes the "sampling error", and R is the true ratio of mean city populations in 1985 to those in 1980.

In dealing with radioactivity, there will almost always be "instrument errors" to contend with and these errors will affect both of the variables being considered. Usually, they will be smaller for the "accurate" measurement (y), but nonetheless replicate determinations on the same sample (when feasible) will be expected to give at least slightly different results. Sometimes large differences can result (Eberhardt and Thomas, 1983:p. 3-7)

especially when "mixing" is a problem or when particles are present. For the accurate measurement (y) it may be reasonable to assume that the measurement (instrument) error is additive. Whether or not this is true for the auxiliary variable (x) is open to debate, and may depend on the instrumentation and counting procedures used (e.g., fixed time counts versus counting to a predetermined total). Hence, we have assumed both an additive error and a multiplicative measurement error for the x-variables utilized in our simulations.

The basic situation considered in this report is one in which a large number of samples containing radioactivity are identified and subjected to some rapid and inexpensive measurement. A much smaller subset of these samples is also measured by using an accurate method, and the two sets of data are combined to estimate a grand total (or, usually, an overall mean) which is expressed in units of the accurate measurement. For concreteness, one might think of sampling surface soils for a particular radionuclide. A field measuring instrument might be used to take readings on several hundred (or more) individual sites. At a much smaller sample of these sites (perhaps 10 to 30), samples of soil are removed and taken to a laboratory for accurate analyses, which may be expressed in picocuries or microcuries per gram. The double sampling procedure then serves to estimate a total (or mean) for the entire area surveyed. If the area is not too large, it may be feasible to cover it completely with the field instrument, thus approximating the second example of the ratio estimation approach described above.

Much of the balance of this report is concerned with checking the utility of standard textbook equations used for estimating means and variances in the face of circumstances likely at CLLRW sites, where the underlying model may be inappropriate (as indicated above). However, we first need to briefly discuss a major question, i.e., when is the double sampling approach worthwhile? It has a corollary issue which is "What is the appropriate ratio of small to large sample sizes?"

2.0 WHEN SHOULD RATIO METHODS BE USED?

The cost-effective use of ratio methods depends on the existence of some rapid and inexpensive analytical procedure that is correlated with a more accurate method (i.e., the "standard" method), which yields widely accepted results. An essential question then is one of how well correlated the two methods need to be to make the ratio approach worthwhile. The correlation coefficient considered here is Pearson's product-moment correlation, which is defined as:

$$\rho = \sqrt{COV^2(X,Y)/VAR(X)VAR(Y)}$$

If two measurements give almost identical results, the correlation coefficient (ρ) will be nearly unity, and one would routinely use the less expensive method, while the expensive method would be relegated to very limited use as a confirmatory tool. Conceivably, the inexpensive method might give a value that is a constant multiple of the expensive method, so that there might be a need for a single "calibration" study, but in most practical instances, this calibration will have been done when the method was developed.

If the correlation is very poor (nearly zero), one probably would not consider the less expensive method, or would conduct studies aimed at improving the correlation. We believe that this is an area needing a good deal of further attention, especially in circumstances where sample analysis is very expensive. Many opportunities for using ratio methods probably exist, but have gone unrecognized simply because investigators have not known about the approach.

The correlation between inexpensive and expensive methods is most important in double sampling. This is because the ratio methods, in which known totals are used, are likely to be only employed when the auxiliary measurement is "free", i.e., is collected for some other purpose. In the first introductory examples (pages 1-2), the container weights would have been collected for other reasons, and the air filter gross radioactivity readings (page 2) would have been obtained in the course of routine monitoring. When such data are available, it is always worthwhile to consider ratio estimation. Cochran (1977:157) gives a simple rule, which states that, for simple random samples, the ratio estimate has a smaller variance than results from an independent estimate based only on the expensive method if the correlation coefficient (ρ) exceeds one-half the coefficient of variation (standard deviation divided by mean, C.V.) of the

auxiliary variable (x) divided by the coefficient of variation of the accurate but expensive variable (y) (i.e., $\rho > 1/2C.V._x/C.V._y$).

For double sampling Cochran (1977:341) utilizes a simple cost function:

$$C = cn + c'n' \tag{2.1}$$

where c = cost per unit of the expensive analysis (y) for which n samples are taken, and c' = cost per unit of the inexpensive determination (x) for which n' samples are taken. In most practical situations, the total cost (C) is fixed, so that it is possible to calculate the ratio of individual costs for which double sampling is worthwhile, given a particular correlation or, conversely, the minimum required correlation when an established ratio of sampling costs is available. This result can be expressed as (Cochran 1977:341):

$$\rho^2 > 4(c/c')/(1 + c/c')^2 \tag{2.2}$$

Thus if the cost ratio (c/c') is 10, the critical correlation is $\rho = [4(10)/(11)^2]^{1/2} = 0.58$. If the cost ratio is 100, ρ is 0.2, while a cost ratio of 2 requires ρ to be at least 0.94. Clearly, a high cost ratio, as would exist for, say, "wet chemistry" versus use of a field radiation detector makes double sampling seem worthwhile even if there is a relatively poor correlation between the two methods.

Judging cost ratios and correlations will often depend on experience and limited data, so that an allowance needs to be made for errors in guessing at some of the values. Hence we might consider a reduction of expected variance to, say, 80 percent of the variance obtained by simple random sampling (using the expensive method) as a possible criterion for employing double sampling. Figure 2.1 gives a "break-even" line (equal variances), and curves showing expected variances of 80 percent, 66 percent, and 50 percent of those achieved without double sampling (i.e., if all of the available resources were spent on samples using the accurate but expensive method). For example, at a cost ratio of about 20 and a correlation of just over 0.4, all the available resources can be devoted to an analysis of expensive samples since double sampling will not reduce the variance of the estimated total. Conversely if the correlation between the inexpensive and expensive methods is about 0.7, a cost ratio of about 60 will result in a 66 percent reduction on the variance of an estimated total. A worked-out example is presented in section 7.0.

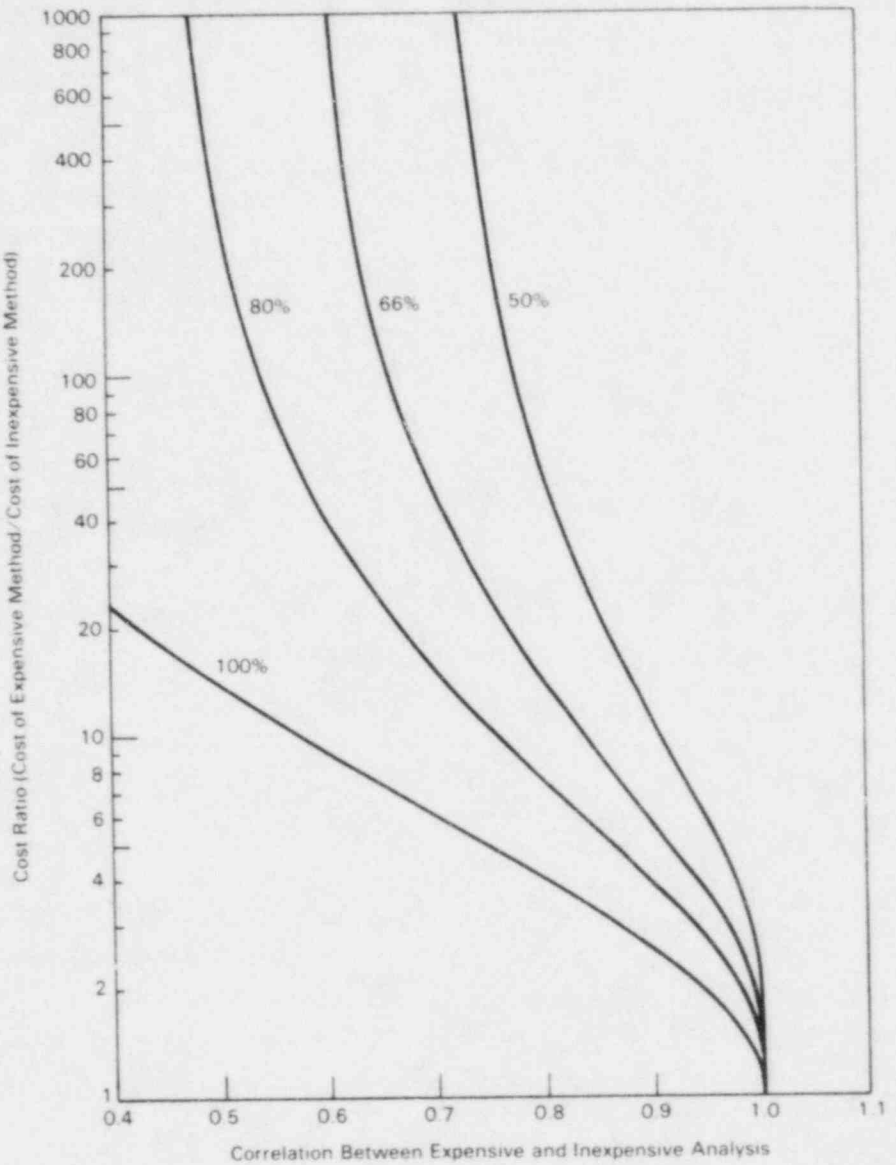

Figure 2.1 Relation between cost ratios and the correlation for 4 percentage reductions of variance in sample estimates when double sampling is used. The 100 percent curve gives the "break-even" situation, i.e., no gain from double sampling, the 80 percent curve denotes situations where the double sampling variance is 80 percent of that for simple random sampling with the expensive measurement only, while the 66 percent curve represents a variance reduction of 66 percent, and the 50 percent curve represents cases where double sampling reduces the variance to one-half that for simple random sampling.

3.0 SIMULATING DOUBLE SAMPLING

3.1 A Model for Double Sampling

Our objective in simulating double sampling is to check on the performance of the method under circumstances likely to be appropriate for sampling radioactivity at commercial low-level radioactive waste sites (CLLRWS). To do this, we had to specify a model for the situation to be simulated. As indicated in Section 1.0, virtually all applications involving radioactivity can be expected to involve instrument errors (measurement errors). We thus need to make these errors an explicit component of the model. A second need is to define the population to be simulated. In actual applications, it is only necessary to be able to identify all of the elements in a given population, and one does not need to specify much more about structure of the population being sampled. However, it is essential to have an advance estimate of population variability in order to use a realistic sample size in actual applications. Hence, we need to specify something about variability in the simulated population. Also, field experience with radioactivity and other trace substances shows that the frequency distributions are almost invariably skewed, that is, there is a small proportion of high values, while the bulk of the population takes much smaller values.

Realistic assumptions about the variance and the distribution are sufficient to construct a suitable population for simulations (i.e., to sample from). However, it is useful to generate the population from a theoretical model having known parameters, both for ease in programming and to repeat the processes used, but also because some of the relevant sampling theory rests on a "superpopulation" approach. That is, it is assumed that the actual finite population being sampled was originally generated by some process that could produce a very large universe of similar populations. This view is increasingly prevalent in some of the earth sciences, principally as a model for geological features. The superpopulation approach has been criticized by some statisticians as being inappropriate for specific kinds of problems. Here, we are only considering it as a way to avoid specifying the size of a very large finite population.

Since it is well established that skewed populations are present in sampling for radioactivity, we adopted a gamma distribution as the "parent" for the simulations. The gamma distribution used here has two parameters, denoted by k and α. The mean and variance are:

$$E(z) = k/\alpha \qquad V(z) = k/\alpha^2 \tag{3.1}$$

so that the squared coefficient of variation is:

$$(C.V.)^2 = V(z)/[E(z)]^2 = 1/k$$

The coefficient of variation is the parameter of main interest here, since the fairly substantial evidence available on variability of radionuclides in field settings (see, for example, Eberhardt et al.(1976)) indicates that it is relatively constant for a particular radionuclide and situation. Values of the coefficient of variation encountered in practice range from roughly unity (some transuranic data) down to 20 to 30 percent or thereabouts. We have consequently selected values of k of 2, 4, and 11, giving coefficients of variation of about 70 percent, 50 percent, and 30 percent. Sample sizes required for a given level of confidence about a mean or total will vary inversely with the coefficients of variation of the population, that is, the largest sample will be required for the highest coefficient of variation. Consequently, we will mainly be interested in simulations where the coefficients of variation are 70 and 50 percent (i.e., k is 2 or 4).

In many respects, the parameter α is a "nuisance parameter" in the simulations. One approach is to set it equal to unity, but this results in different mean values for the several coefficients of variation, so we have chosen to use an arbitrary mean value of 5.0 [E(z)] for the simulations, assuming that the average of the "expensive" determinations will always be 5 units (nannocuries, picocuries, microcuries, etc.). The three values of α (0.4, 0.8 and 2.2) are thus determined from eq. (3.1). Simulated values for three "true" (gamma-distributed) population were generated from the expression:

$$z = \frac{-1}{\alpha} \sum_{i=1}^{k} \log_e(1 - r_i) \qquad (3.2)$$

where r_i is obtained from a random number generator. The rationale is that a gamma distribution can be obtained as the sum of random variables drawn from an exponential distribution, which can in turn be generated from $-1/\alpha \log_e(1-r_i)$, as shown by Bratley et al. (1983:157). The uniform random number generator used here has been described by Simpson, Harkins, and Watson (1979). Figure 3.1 shows the gamma distributions used in the simulations.

The remaining components required for the simulation model are the measurement errors. The simplest model for measurement or instrument errors is:

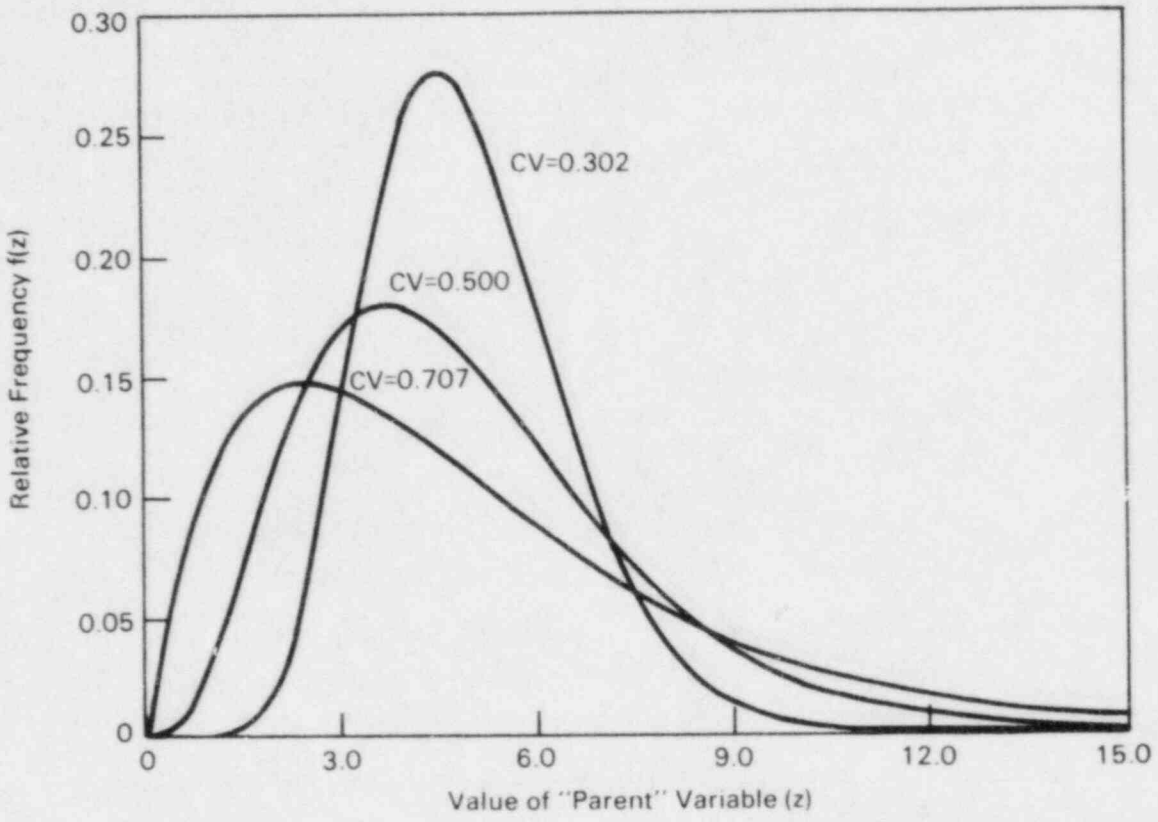

Figure 3.1 Gamma distributions for three coefficients of variation (0.707, 0.500, and 0.302)

$$y_i = z_i + e_{2i} \tag{3.3}$$

$$x_i = \beta z_i + e_{1i} \tag{3.4}$$

where the random errors are additive and denoted by e_1 and e_2, while β denotes a "scaling factor" between the two classes of measurement. Presumably β will depend on the nature of the measurement systems and it may incorporate aspects of "dimensionality", i.e., x may be reported in counts per minute, which will be proportional to radionuclide concentration in units of curies. For convenience, we have arbitrarily set $\beta = 2$, but it might very well be some other number, depending on the field instruments, the field situation and the radionuclide.

As previously indicated, we initially consider additive and normal errors, with means zero and variances $V(e_1)$ and $V(e_2)$. These errors were generated by summing 12 uniform random variables (to simulate a unit normal distribution) and then transforming to a normal distribution with the appropriate variance (cf. Naylor et al. 1968:93). For a model with multiplicative errors in the x-variable, we use:

$$x_i = \beta z_i e_{1i} \tag{3.5}$$

where the error e_{1i} now has a mean of unity and is lognormally distributed, being generated by converting a unit normal random variable to have the appropriate variance, as discussed Section 3.5, below. The procedure followed here has four steps:

(1) Draw N values of z_i from a gamma distribution corresponding to a particular coefficient of variation.

(2) Draw a large random sample (n') from the population (N) and incorporate the "instrument" errors and "scaling factor" (β) of equations (3.3) and (3.4) or (3.5), and record the x_i to calculate $\bar{x}'$.

(3) Draw a random subsample (n) of observations from the large sample (n'), and record the y_i and x_i.

(4) Estimate overall mean concentrations, variances, etc., and compare with "true" values for the finite population (N) or the parent population (gamma distribution).

3.2 Estimating Equations and Simulation Criteria

The underlying model for these simulations assumes that a large population of N potential sampling units contains quantities (z_i) of radionuclides. A sample of n' of these units is selected and measured with the inexpensive technique, yielding the x_i values (in the actual programming, we also construct a y_i value for each of the n' elements selected both as a means of checking accuracy of the simulations, and to permit testing various alternative sampling schemes, although only n of these values are actually used in calculations). A subsample of n of the n' units is "measured" by the expensive method and the y_i values recorded.

Two estimating equations are then used:

$$\bar{y}_R = (\bar{y}/\bar{x})\bar{x}' \tag{3.6}$$

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x}) \qquad (3.7)$$

These equations are those for the ratio and regression methods, respectively. It should be noted that here we estimate a mean rather than the total as in eq. (1.1). This is because we do not now have the total of the auxiliary variable (X_T), but must instead work with the mean ($\bar{x}'$) of a large sample. A total is easily estimated by multiplying by N, e.g., $\hat{Y}_R = N\bar{y}_R$. In the second equation, b denotes the usual linear regression coefficient, estimated from the n pairs of observations on y_i and x_i.

Variance estimates for these two means follow equations given by Cochran (1977:343-344). For the ratio method:

$$V(\bar{y}_R) \doteq \left(S_y^2 - 2RS_{yx} + R^2S^2_x\right)/n + \left(2RS_{yx} - R^2S^2_x\right)/n' - S^2_y/N \qquad (3.8)$$

This expression is given in terms of the population values of the several quantities, for which sample values were utilized in our calculations (R is estimated by $\bar{y}/\bar{x}$). The equation (based on sample values) used for the regression method is:

$$v(\bar{y}_{lr}) = s^2_{y.x}\left\{1/n + \left[(\bar{x}' - \bar{x})^2/\sum(x_i - \bar{x})^2\right]\right\} + \left(s^2_y - s^2_{y.x}\right)/n' - s^2_y/N \qquad (3.9)$$

where $s^2_{y.x}$ denotes the variance about regression obtained from linear regression calculations, and thus should be distinguished from the sample value of S_{yx}, the covariance of y and x. The corresponding variance estimate using population parameters is:

$$V(\bar{y}_{lr}) \doteq \left[S^2_y(1 - \rho^2)/n\right] + \rho^2S^2_y/n' - S^2_y/N \qquad (3.10)$$

The criteria used for judging outcomes of the simulations included comparison of calculated means with "true" values and "coverage" of confidence intervals calculated using the variance estimates (eqs. 3.8 and 3.9). Two sets of "true" values were used, one being simply the means of the N random variables in the population (denoted here as "finite" values), and the second was expected values obtained from the underlying gamma distribution, e.g., E(z) of eq. (3.1). "Coverage" is determined by recording whether or not the calculated confidence limits do include the true value, using the 5 percent and 10 percent values from the t-distribution in confidence limit calculations, e.g. $\bar{y}_R \pm t_{.05}[v(\bar{y}_R)]^{1/2}$.

3.3 Expected Values

The simple linear structure of the model used in eq. (3.3) and (3.4) makes it easy to calculate expected variances for the population of N random variables:

$$E\left(S^2_x\right) = \beta^2 V(z) + V(e_1) \tag{3.11}$$

$$E\left(S^2_y\right) = V(z) + V(e_2) \tag{3.12}$$

$$E\left(S_{xy}\right) = \beta V(z) \tag{3.13}$$

where $V(z)$ is the variance of the underlying gamma distribution, and $V(e_1)$ and $V(e_2)$ are the respective measurement or instrument error variances. We can substitute these expected values in the definitions of the regression and correlation coefficients to obtain:

$$b = S_{xy}/S^2_x = \left[\beta + V\left(e_1\right)/\beta V(z)\right]^{-1} \tag{3.14}$$

$$\rho^2 = S^2_{xy}/S^2_x S^2_y = \left\{1 + \left[V\left(e_1\right)/\beta^2 V(z)\right]\right\}^{-1}\left[1 + V\left(e_2\right)/V(z)\right]^{-1} \tag{3.15}$$

Using the above expected values, we also calculated expectations for variances calculated according to equations (3.8) and (3.10).

3.4 Simulation Parameters

As noted above, the simulations are based on parent gamma populations having coefficients of variation of about 70, 50, and 30 percent. Finite populations of N = 1000 were used, although we would expect such populations to be much larger in practice. When the actual populations are larger, the results will essentially correspond to the "theoretical" outcomes for the parent gamma distributions. Hence, the small finite populations (N = 1000) serve to represent those situations where a relatively small finite population might be encountered. Large samples of n'=100 and n'=200 were used, and subsamples (n) of 5, 10, and 20 were taken from these. Larger subsamples may sometimes be taken in practice, but these can be expected to behave according to the variance equations if the smaller subsamples do so.

The number of runs in each simulation was set at 2,000, in order to have enough replications to check agreement with expected values. This gives

approximate binomial confidence limits (CL) on the "coverage" calculation of about ± 0.01 since

$$\text{Standard Error (SE)} = (pq/n)^{1/2} = [(.05)(.95)/2000]^{1/2} = 0.0049$$

and the approximate 95% CL = 2(SE) ≈ 0.01, that is, if our empirical calculation of the coverage of a confidence limit calculation for the 5 percent level of significance comes out within about ± 0.01 of the expected value, we have little reason to suppose that it differs significantly from that level. However, chi-square calculations were also used to check the correspondence between expected and simulated coverage.

The choice of measurement or instrument errors can be considered by referring to eq. (3.15), which shows that these errors determine the correlation coefficients for our basic model (given fixed parameters β and $V(z)$). We assume measurement errors for the accurate method should be a relatively small component of the overall variation of the population (if they are not relatively small, then there is little point in doing much sampling). Hence, we set the measurement error at about 10 and 25 percent of the variance in the parent population, i.e.,

$$V(e_2)/V(z) = 0.10 \text{ and } 0.25.$$

We then introduce these values in eq. (3.15) and vary the relative errors $[V(e_1)/V(z)]$ for the inexpensive method so as to cover the range of correlation coefficients that might be expected in practice. The expected correlation coefficients will, of course, depend on the relative costs of the two methods, and we suspect these will usually be a factor of 10 or more, so that the underlying correlation coefficients will mostly be on the order of 0.5 and larger (see eq. (2.2) and Figure 2.1). By using ratios of $V(e_1)$ to $V(z)$ of 1, 2, 4, and 8, we obtained correlations (eq. (3.15)) of about 0.55 to 0.85, which should cover a sizable fraction of the cases where the method is cost effective and thus likely to be of practical interest (cf. Figure 2.1).

It should be noted that the actual field application of double sampling will usually be done in circumstances where the correlation between y and x and the cost ratio (c'/c) are more or less fixed in advance. In designing a survey, one thus has to use the approximate values of correlation and costs to determine the relative sizes of the large and small samples. Cochran (1977:341) gives the optimum ratio as:

$$n/n' = [(c'/c)(1 - \rho^2)/\rho^2]^{1/2}$$

In order to make our simulations broadly representative, we have used various combinations of n and n'. The corresponding ratios (n/n') range from 5/200 = 0.025 to 20/100 = 0.20. If we consider a range of correlations (0.5 to 0.9) similar to that used in the actual simulations and cost ratios (c'/c) ranging from 1/10 to 1/1000, the corresponding optimum ratios (n/n') will range from 0.015 to 0.55, a somewhat wider range than covered by the simulations.

3.5 Lognormal Measurement Errors

In order to simulate multiplicative, lognormal errors, as in eq. (3.5), we calculated parameter values such that the overall expected values of eqs. (3.11) to (3.13) would remain unchanged, giving the same numerical expectations for regression slopes (eq. (3.14)) and correlations (eq. (3.15)) as were used with the additive normal errors of eq. (3.4). This means that $V(e_1)$, the variance of the lognormal, multiplicative error, has to be calculated so that $E(S^2_x)$, as computed previously from eq. (3.11), has the same numerical value. We thus need an expression for the variance of x_i computed from eq. (3.5). Since β is a constant, we can use the rule for product of a variance of two independent random variables (Goodman 1960) to obtain:

$$S^2_x = V(\beta z e_1) = \beta^2\{[E(z)]^2 V(e_1) + [E(e_1)]^2 V(z) + V(z)V(e_1)\} \quad . \qquad (3.16)$$

In order to use a multiplicative error, we set $E(e_1) = 1.0$. We can now rearrange eq. (3.16) and equate it to the "old" variance term of eq. (3.11), obtaining:

$$V\, e_1 = V'\, e_1 \,/\, \left(\beta^2\{[E(z)]^2 + V(z)\}\right) \qquad (3.17)$$

where $V'(e_1)$ denotes the (numerical) value of $V(e_1)$ previously used for normal additive errors in eq. (3.4).

To generate the actual lognormal errors we recall that the expected value of a lognormal distribution is $E(e_1) = \exp(\mu + \sigma^2/2)$. Since $E(e_1)$ is to be unity in the simulations, we set $\mu = -\sigma^2/2$ (i.e, $e^0 = 1.$). Because this lognormal distribution is to have a mean of unity, the variance reduces to:

$$V(e_1) = \exp(\sigma^2) - 1$$

Equating this to eq. (3.17), we can calculate the value of σ^2 for the parameters previously determined for each case. The individual error terms are then calculated from:

$$e_{1i} = \exp\left[-\sigma^2/2 + \sigma r_i\right] \tag{3.18}$$

where r_i is a random variable drawn from a unit normal distribution. The resulting simulations thus have the same set of expected values given by eqs. (3.11) through (3.15) for the case of a normal, additive error.

3.6 Alternative Models

The simulation models considered thus far are essentially those for which ratio estimators were devised, i.e., straight lines through the origin. In practice, one cannot always expect such a simple relationship between two variables. The simplest alternative model is one in which a regression relationship holds, i.e., instead of eq. (3.4) we have:

$$x_i = \alpha + \beta z_i + e_{1i} \tag{3.19}$$

Such a relationship might be generated if the auxiliary measurement includes "background" counts or contamination of some kind. In this case, the extraneous source may be removed in the expensive analyses, which usually include chemical separations to remove contaminants. These background counts could thus give positive values for x_i in cases where y_i is essentially zero. Hence, a regression model is appropriate.

A variety of other alternative models might also be considered to result from various kinds of inhomogeneities in the material sampled, or contamination, inadequate shielding, and the like. Determining the most likely candidates among the large number of possible such models calls for information not presently available. To explore some of the prospective effects, we used a non-linear model, widely encountered in studies of radioactivity:

$$x_i = A\left[1 - \exp\left(-Bz_i\right)\right] + e_{1i} \tag{3.20}$$

An example of this model appears in Figure 3.2. Note that this is a model to generate the x_i from a parent distribution of z_i, and is not the relationship

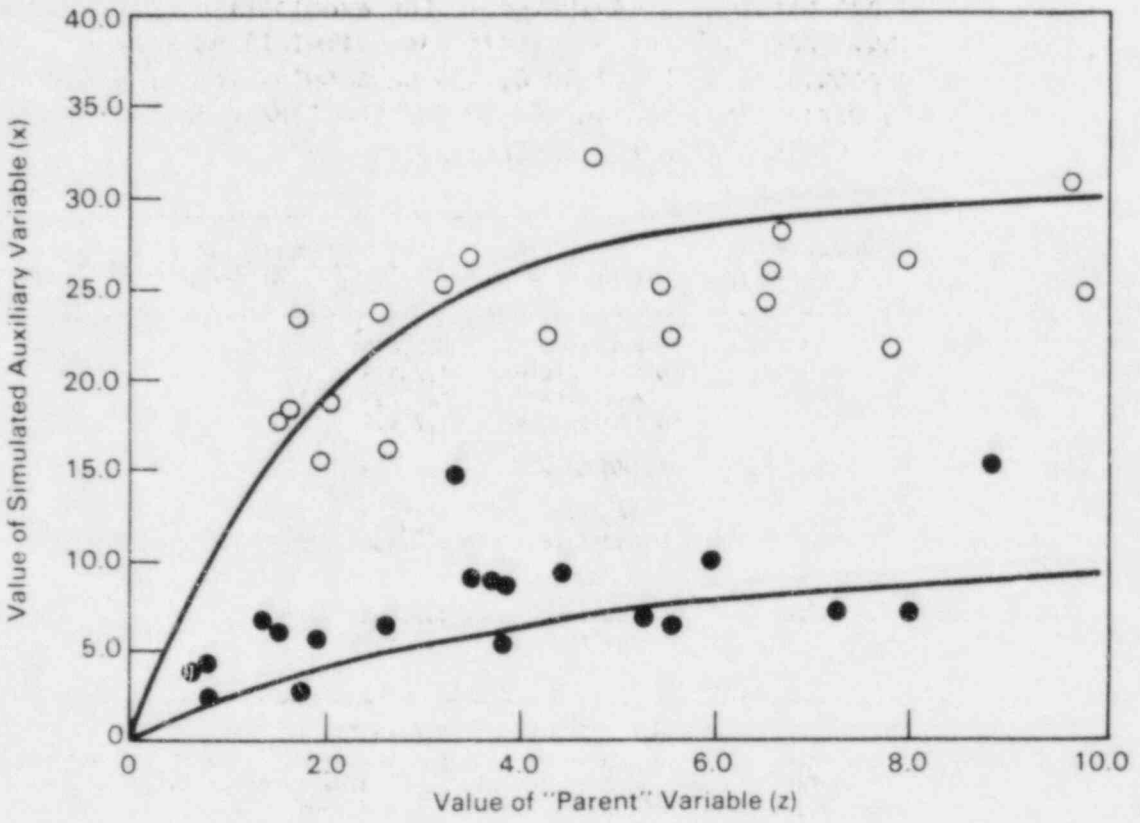

Figure 3.2 Examples of auxiliary variables generated using the non-linear model (eq. (3.20)). Data points and curves have parameters A = 10, B = 0.25 (●) and A = 30, B = 0.50 (○). Measurement errors are normal and additive, with error of the auxiliary variable equal to variance of parent population. Note the non-linear model serves to transform variables drawn from a gamma distribution (z) into the auxiliary variable (x) used in the simulations described in Section 4.0.

that would be observed in practice, i.e., between y_i and x_i. This relationship is illustrated in Figure 4.6.

The various models and parameters evaluated in this progress report are given in Table 3.1.

Table 3.1 Models and parameters evaluated in the examination of double sampling. For each set of conditions, simulations were run for finite populations (N) of 1000, and subsamples (n) of 5, 10 and 20 for the ratio model and 10 and 20 for the linear and non-linear models. n' is the large sample size.

C.V.	n'	Measurement Error for y	Error on x[a]	Model for x		
				Ratio	Linear	Non-linear
.707 (k=2, α=.4, V(z)=12.5)	100	.10	Additive	1,2,4,8[b]	1,4	1,2,4
			Multiplicative	1,2,4,8		1,4
		.25	Additive	1,2,4,8		
			Multiplicative	1,2,4,8		
	200	.10	Additive	1,2,4,8		
.500 (k=4, α=.8, V(z)=6.25)	100	.10	Additive	1,2,4,8		1,2,4
		.25	Additive	1,2,4,8		
.302 (k=11, α=2.2, V(z)=2.273)	100	.10	Additive	1,2,4,8		
		.25	Additive	1,2,4,8		

[a] Errors on x were additive, normal or multiplicative, lognormal.

[b] 1, 2, 4, 8 refer to the ratio of the variance in x to the variance in the finite population (i.e, $V(e_1)/V(z)$). These ratios correspond to correlations between the primary (y) and auxiliary (x) variables of: .853, .778, .674, .550 (.10) and .800, .730, .632, .516 (.25).

4.0 SIMULATION RESULTS

4.1 An Upper Limit for Variability

Since we cannot, at present, accurately predict which of the parameter combinations and sample sizes are most likely to be encountered under field conditions, it is necessary to cover a fairly wide range of situations. In order to do so, we have used some examples in which the ratio of small to large sample size is far from its optimal value, and we thus would obtain unacceptably large standard errors in an actual survey. There are no hard and fast rules as to just what constitutes an "unacceptable" result, but, on the basis of experience in various kinds of sampling, we believe that most investigators would prefer to have confidence limits no wider than ±40 percent of the estimate for a total or mean. Where it is at all possible, we recommend that confidence limits no wider than ±20 percent be sought. However, cost considerations almost always dictate the sample sizes that can be achieved in a given study.

If we suppose that confidence limits of ±40 percent constitute an approximate upper bound on allowable variability, then the maximum coefficient of variation for $\bar{y}_R$ or $\bar{y}_{lr}$ can be no more than 20 percent (presumably it will be somewhat smaller, depending on actual sample sizes and the significance level selected). Since N, the size of the finite population being considered, will nearly always be large, the last term in eq. (3.10) becomes very small (relative to the other terms) and can be dropped, giving on factoring out S^2_y:

$$V(\bar{y}_{lr}) \doteq S^2_y\left[(1 - \rho^2)/n + \rho^2/n'\right]$$

Eq. (3.12) gives the expected value of S^2_y:

$$E\left(S^2_y\right) = V(z) + V(e_2)$$

The value of V(z) is determined in the simulations by the coefficient of variation of the "parent population" (the gamma distribution generating z_i), and two values of $V(e_2)$ were used for each coefficient of variation, i.e., $V(e_2) = 0.1\ V(z)$ and $V(e_2) = 0.25\ V(z)$ (Section 3.4). Consequently, we can write:

$$CV^2(\bar{y}) = [V(z) + V(e_2)]/[E(z)]^2 = aV(z)/[E(z)]^2 = aCV^2(z)$$

Table 4.1 Means of $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement errors were normal and additive, and measurement error of primary variable ($\bar{y}$) was 10 percent of that for the parent population. The true value of $\bar{y}$ (E(z)) is 5.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
.707	100	5	$\bar{y}_R$	5.20	5.62*	6.11*	1.50*
		10		5.10	5.13	5.39	5.61*
		20		5.04	5.04	5.10	5.26
	100	5	$\bar{y}_{lr}$	4.87	4.77*	4.78*	4.81*
		10		4.94	4.88	4.91	4.89*
		20		4.98	4.94	4.95	4.96
.500	100	5	$\bar{y}_R$	5.08	5.11	5.32	5.94*
		10		5.04	5.08	5.12	5.51
		20		5.01	5.03	5.04	5.10
	100	5	$\bar{y}_{lr}$	4.92	4.88	4.89	4.92*
		10		4.97	4.97	4.93	4.98
		20		4.98	4.98	4.97	4.97
.302	100	5	$\bar{y}_R$	5.03	5.05	5.09	5.22
		10		5.02	5.02	5.04	5.10
		20		5.01	5.01	5.01	5.06
	100	5	$\bar{y}_{lr}$	4.98	4.97	4.96	4.99
		10		4.99	4.99	4.97	5.00
		20		5.00	5.00	4.99	5.00

* Coefficient of variation of estimated mean greater than 20 percent.

where a = 1.1 or 1.25, depending on the relative value of $V(e_2)$ selected. Since $CV^2(z) = 1/k$ (Eq.(3.1)), and assuming $E(\bar{y}_{lr}) = E(z)$, we have:

$$CV^2\,\bar{y}_{lr} \doteq (a/k)\left[(1 - \rho^2)/n + \rho^2/n'\right] \qquad (4.1)$$

This result lets us approximate the criterion of confidence limits less than ±40 percent for any parameter set. Values of outcomes exceeding this value are identified (by an asterisk) in the tables of results. We believe that these particular outcomes can largely be neglected in judging the utility of double sampling for practical applications, since reasonable care in planning the survey will avoid these outcomes. We have used the same criterion for the ratio estimator, since a similarly convenient equation is not available. As a second check, Appendix A, (Table A.1) gives coefficients of variation based on expected variances and expected means for $\bar{y}_R$ and $\bar{y}_{lr}$ for cases corresponding to parts of Table 4.1 and Table 4.5. The results suggest that

eq. (4.1) underestimates somewhat for $\bar{y}_R$, so that a few more cases might possibly be excluded from consideration.

4.2 Normal and Additive Errors

As described in Section 3.1, our initial model is that of normal and additive measurement errors, yielding the values of y_i and x_i given by eqs. (3.3) and (3.4). Table 4.1 contains the average values of $\bar{y}_R$ and $\bar{y}_{lr}$ for the various parameter combinations used in the simulations and covers those simulations for which the variance of measurement errors in the "accurate" measurements (y) was 10 percent of that in the parent population [$V(e_1) = 0.1V(z)$]. The outcomes here are quite straightforward, with $\bar{y}_{lr}$ yielding a small underestimate of the true value (5.0), and $\bar{y}_R$ giving a small overestimate, excepting those cases where confidence limits on the estimates would exceed ±40 percent of the estimate.

The variance estimates used here (eqs. (3.8) using sample values and (3.9)) are based on approximations of one sort or another, as discussed by Cochran(1977). Consequently, we examined the ratio of the average computed variance to expected values, finding that the computed values are good approximations (Table 4.2), with a small positive bias (overestimating slightly). Cases where confidence limits exceed ±40 percent are more seriously biased.

The results discussed above pertain to cases where the large sample size (n') is 100. A subset of cases for the largest C.V. (0.707) was studied for n' = 200, giving very similar results (Table 4.3).

A third check on the simulation results is provided by examining "coverage" of confidence limits obtained from the simulated outcomes. That is, we calculated separate confidence limits for the means of each one of the 2,000 simulated sets of data, and then determined whether or not these confidence limits included the expected values. If one chooses the 95 percent "level of confidence" (sometimes expressed as a 5 percent "Type I error"), then, in about 5 percent of the cases, the simulated value should fall outside of the calculated limits. The outcomes (Table 4.4) are generally quite close to 5 percent. A chi-square value was calculated in each case, and those exceeding the one percent level of significance are identified in the tables. Since we are using approximate equations, and there are small biases in the estimating equations (discussed below), it is not surprising that there are more significant values of chi-square than would be expected if all of the theoretical results held exactly. The important point here is that between 4 to 6 percent of the confidence limits do not contain the expected value, which is quite acceptable in planning for a 5 percent level.

Table 4.2 Ratios of the mean calculated variances to expected variance for $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement errors were normal and additive, and measurement error of primary variable (y) was 10 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
.707	100	5	$\bar{y}_R$	1.22	9.50*	15.71*	9759.*
		10		1.05	1.14	1.58	3.95*
		20		1.03	1.04	1.08	1.30
	100	5	$\bar{y}_{lr}$	1.39	1.42*	1.40*	1.36*
		10		1.08	1.09	1.06	1.09*
		20		1.03	1.02	1.02	1.02
.500	100	5	$\bar{y}_R$	1.08	1.16	1.35	4.66*
		10		1.05	1.04	1.10	4.22
		20		1.01	1.02	1.04	1.07
	100	5	$\bar{y}_{lr}$	1.42	1.40	1.40	1.45*
		10		1.10	1.09	1.09	1.10
		20		1.04	1.02	1.03	1.05
.302	100	5	$\bar{y}_R$	1.03	1.03	1.09	1.12
		10		1.01	1.02	1.03	1.06
		20		1.01	1.02	1.03	1.03
	100	5	$\bar{y}_{lr}$	1.49	1.48	1.42	1.45
		10		1.11	1.10	1.10	1.10
		20		1.03	1.04	1.05	1.05

* Coefficient of variation of estimated mean greater than 20 percent.

"Coverage" for the 90 percent level of confidence was also examined in the simulations, as was coverage for the finite population of N = 1000 for both 95 percent and 90 percent levels. These data sets behaved nearly the same as the coverage shown in Table 4.4, so they are given in Appendix A (Tables A.2 to A.4).

Results for simulations in which the error term for the "expensive" measurement (y) was 25 percent of that of the parent population [$V(e_2) = 0.25$ $V(z)$] were very similar to those already given for the case where this error was 10 percent of the variability for the parent population (Tables 4.1 and

Table 4.3 Means and ratios of the mean calculated variance to expected variance for $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement errors were normal and additive, and measurement error of primary variable (y) was 10 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
		MEANS					
.707	200	5	$\bar{y}_R$	5.20	6.14*	3.45*	7.85*
		10		5.06	5.20	5.39	5.77*
		20		5.03	5.07	5.18	5.32
		5	$\bar{y}_{lr}$	4.85	4.79*	4.76*	4.80*
		10		4.91	4.90	4.88	4.89
		20		4.95	4.95	4.94	4.94
		VARIANCE RATIOS					
		5	$\bar{y}_R$	1.24	23.2*	123.*	214.*
		10		1.07	1.34	1.38	9.82*
		20		1.03	1.06	1.14	1.63
		5	$\bar{y}_{lr}$	1.51	1.38*	1.31*	1.43*
		10		1.09	1.09	1.06	1.10*
		20		1.04	1.03	1.02	1.03

* Coefficient of variation of estimated mean greater than 20 percent.

4.2). Thus, mean values (Table 4.5) and variance ratios (Table 4.6) are presented only for coefficients of variation of 70 and 50 percent. Again, ratio estimates provided slight overestimates, while regression estimates were slightly low. It should be noted that the correlations are reduced by increasing the measurement error.

4.3 Lognormal Measurement Errors

Since our prior experience suggests that measurement errors for at least the auxiliary variable (x) may be multiplicative and non-normal, we used a lognormal distribution to generate errors for eq. (3.5) as described in Section 3.5. These simulations were limited to coefficients of variation (of the parent population) of 70 percent and 50 percent. Results for a C.V. of 30 percent were generally less variable, as evident from the tables in Section 4.2.

Table 4.4 "Coverage" of calculated confidence limits for mean $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Values shown are proportions of the total simulations for which the 95 percent confidence limits constructed from simulated data do not include the "true" mean [E(z), expected mean for the gamma distribution]. Measurement errors were normal and additive, and measurement error of the primary variable (y) was 10 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
.707	100	5	$\bar{y}_R$	.050	.063	.055	.064*
		10		.053	.058	.057	.052
		20		.057	.052	.055	.043
	100	5	$\bar{y}_{lr}$	.051	.063*	.069*	.082*
		10		.057	.073*	.072*	.068*
		20		.061	.064*	.058	.064*
.500	100	5	$\bar{y}_R$	.045	.049	.035*	.043
		10		.043	.049	.059	.053
		20		.053	.053	.041	.054
	100	5	$\bar{y}_{lr}$	.057	.058	.058	.071*
		10		.050	.056	.059	.059
		20		.053	.048	.052	.048
.302	100	5	$\bar{y}_R$	.040	.049	.040	.057
		10		.043	.048	.044	.051
		20		.049	.046	.045	.062
	100	5	$\bar{y}_{lr}$	.058	.051	.059	.063*
		10		.053	.055	.051	.054
		20		.046	.045	.051	.047

* Calculated chi-square exceeds one percent level of significanc (6.63).

One of the interesting outcomes of these simulations is that the averages of $\bar{y}_R$ and $\bar{y}_{lr}$ (Table 4.7) now both exceed the expected values [only $\bar{y}_R$ exceeded the expected value for additive errors (Table 4.1)], but by amounts that are largely tolerable in practical situations, being at most within a few percentage points of the means (E(z) = 5). In addition, the variance ratios for $\bar{y}_R$ are now less than unity (Table 4.8), i.e., eq. (3.8) for $V(\bar{y}_R)$ now underestimates somewhat, where it previously provided an overestimate (Table 4.2) for the normal and additive measurement error on the auxiliary variable (x). The net effect of these changes on "coverage" (Table 4.9) is that we now are operating at somewhat higher "error rates" than assumed by

Table 4.5 Means of $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement errors were normal and additive, and measurement error of primary variable (y) was 25 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .800	.730	.632	.516
.707	100	5	$\bar{y}_R$	5.17*	5.62*	6.40*	7.16*
		10		5.08	5.16	5.39*	5.51*
		20		5.04	5.06	5.11	5.28
	100	5	$\bar{y}_{lr}$	4.79*	4.79*	4.80*	4.79*
		10		4.95	4.90	4.90*	4.91*
		20		4.97	4.96	4.93	4.96
.500	100	5	$\bar{y}_R$	5.07	5.15	5.40*	6.19*
		10		5.03	5.07	5.16	5.26
		20		5.01	5.04	5.05	5.12
	100	5	$\bar{y}_{lr}$	4.91	4.89	4.88*	4.93*
		10		4.98	4.96	4.97	4.95
		20		4.98	4.98	4.98	4.99

* Coefficient of variation of estimated mean greater than 20 percent.

Table 4.6 Ratios of the mean calculated variances to expected variance for $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement errors were normal and additive, and measurement error of primary variable (y) was 25 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .800	.730	.632	.516
.707	100	5	$\bar{y}_R$	1.19*	10.6*	313.*	204.*
		10		1.08	1.11	1.57*	9.15*
		20		1.03	1.04	1.09	1.34
	100	5	$\bar{y}_{lr}$	1.40*	1.39*	1.75*	1.46*
		10		1.13	1.09	1.08*	1.09*
		20		1.04	1.03	1.03	1.02
.500	100	5	$\bar{y}_R$	1.10	1.14	1.73*	10.1*
		10		1.04	1.06	1.12	1.55
		20		1.01	1.02	1.04	1.09
	100	5	$\bar{y}_{lr}$	1.46	1.49	1.44*	1.41*
		10		1.13	1.10	1.12	1.11
		20		1.03	1.02	1.03	1.03

* Coefficient of variation of estimated mean greater than 20 percent.

Table 4.7 Means of $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement error of auxiliary variable (x) was multiplicative and lognormal, and measurement error of primary variable (y) was 10 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
.707	100	5	$\bar{y}_R$	5.12	5.18*	5.41*	5.70*
		10		5.05	5.09	5.18	5.44*
		20		5.03	5.07	5.08	5.20
	100	5	$\bar{y}_{lr}$	5.15	5.22*	5.37*	5.47*
		10		5.07	5.11	5.14	5.30*
		20		5.03	5.08	5.07	5.12
.500	100	5	$\bar{y}_R$	5.06	5.17	5.18	5.38*
		10		5.04	5.07	5.08	5.19
		20		5.01	5.03	5.06	5.13
	100	5	$\bar{y}_{lr}$	5.09	5.16	5.16	5.23*
		10		5.06	5.07	5.07	5.09
		20		5.01	5.03	5.04	5.08

* Coefficient of variation of estimated mean greater than 20 percent.

Table 4.8 Ratios of the mean calculated variances to expected variance for $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement error of auxiliary variable (x) was multiplicative and lognormal, and measurement error of primary variable (y) was 10 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
.707	100	5	$\bar{y}_R$	0.87	0.76*	0.68*	0.55*
		10		0.94	0.88	0.79	0.69*
		20		0.96	0.92	0.88	0.78
	100	5	$\bar{y}_{lr}$	1.42	1.47*	1.66*	2.06*
		10		1.05	1.03	1.06	1.12*
		20		0.99	0.99	0.99	0.99
.500	100	5	$\bar{y}_R$	0.89	0.85	0.80	0.68*
		10		0.95	0.92	0.87	0.80
		20		0.98	0.96	0.92	0.89
	100	5	$\bar{y}_{lr}$	1.32	1.42	1.38	1.48*
		10		1.07	1.08	1.10	1.10
		20		1.01	1.01	1.01	1.02

* Coefficient of variation of estimated mean greater than 20 percent.

Table 4.9 "Coverage" of calculated confidence limits for mean $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Values shown are proportions of the total simulations for which the 95 percent confidence limits constructed from simulated data do not include the "true" mean [E(z), expected mean for the gamma distribution]. Measurement error of auxiliary variable (x) was multiplicative and lognormal, and measurement error of the primary variable (y) was 10 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
.707	100	5	$\bar{y}_R$	.088*	.086*	.119*	.146*
		10		.071*	.089*	.099*	.132*
		20		.063	.080*	.087*	.108*
	100	5	$\bar{y}_{lr}$	.068*	.077*	.077*	.089*
		10		.059	.072*	.076*	.085*
		20		.062	.078*	.073*	.070*
.500	100	5	$\bar{y}_R$	.068*	.083*	.088*	.106*
		10		.062	.065*	.087*	.095*
		20		.060	.062	.062	.088*
	100	5	$\bar{y}_{lr}$	.066*	.065*	.078*	.076*
		10		.062	.052	.071*	.063
		20		.063	.058	.055	.071*

* Calculated chi-square exceeds one percent level of significance (6.63).

the usual confidence limit calculations. Nonetheless, a shift from an assumed 5 percent rate to an actual level of 7 percent or so is not a matter that should be of much practical concern.

Again, changing the value of the measurement error in the principal variable (y) from 10 percent to 25 percent of the variance of the parent population yields results very close to those in Tables 4.7 to 4.9, so the outcomes appear in Appendix A (Tables A.5 to A.7).

4.4 Bias in Estimators

Estimates of the regression slope (b) of the primary variable (y) on the auxiliary variable (x) are known to be biased. The regression slope is "attenuated" by measurement errors, having an expected value approximately described by eq. (3.14). An approximation is involved inasmuch as we use the ratio of expected values of S_{xy} to S^2_x rather than E(b), which would be very difficult to determine analytically. The simulation outcomes for the case

with normal and additive errors (Figure 4.1) indicate that eq. (3.14) provides an adequate model for the relationship, with the means of slopes calculated from 2,000 simulations being a little less than predicted by eq. (3.14). The deviations from the expected outcome are most pronounced for the more variable data sets. The outcomes for larger samples (n = 20) and the smallest coefficients of variation (0.30) were nearly indistinguishable from expected values. We have thus shown only the smallest samples (n = 5) from

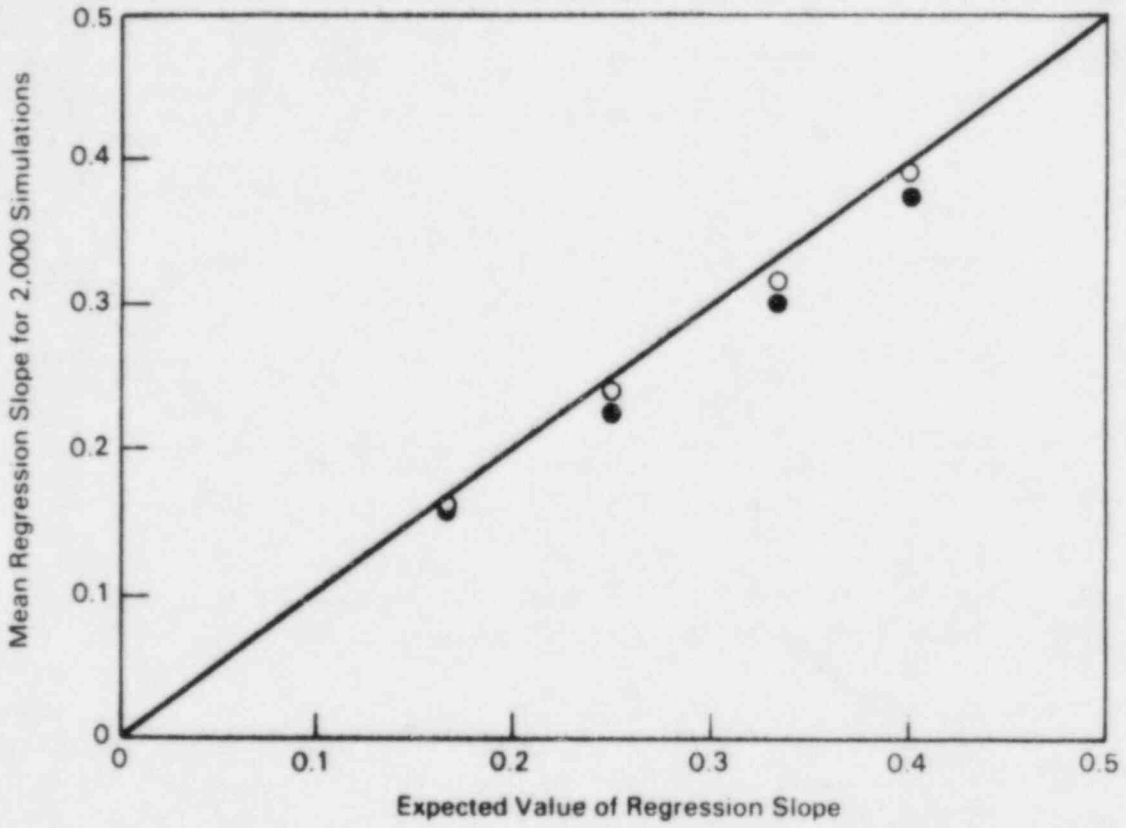

Figure 4.1 Mean values of regression slopes from 2,000 simulations plotted against calculated expected values for normal and additive errors. Data selected from Appendix Table A.8; only values for the more variable data sets [n = 5, coefficients of variation 0.707 (•) and 0.500 (○)] are shown. Less variable outcomes were very close to the expected values. The straight line shows the 1:1 relationship. Measurement error of the primary (y) variable was 10 percent of the variance of the parent population.

the more variable parent populations (coefficients of variation 0.500 and 0.707) in Figure 4.1.

Using multiplicative, lognormal errors (eq. (3.5)) changes the relationship appreciably (Figure 4.2), yielding mean values of the regression coefficient appreciably larger than expected. This change presumably is a consequence of the multiplicative model. While we maintained the same numerical values for the variance of the auxiliary variable (cf. Section

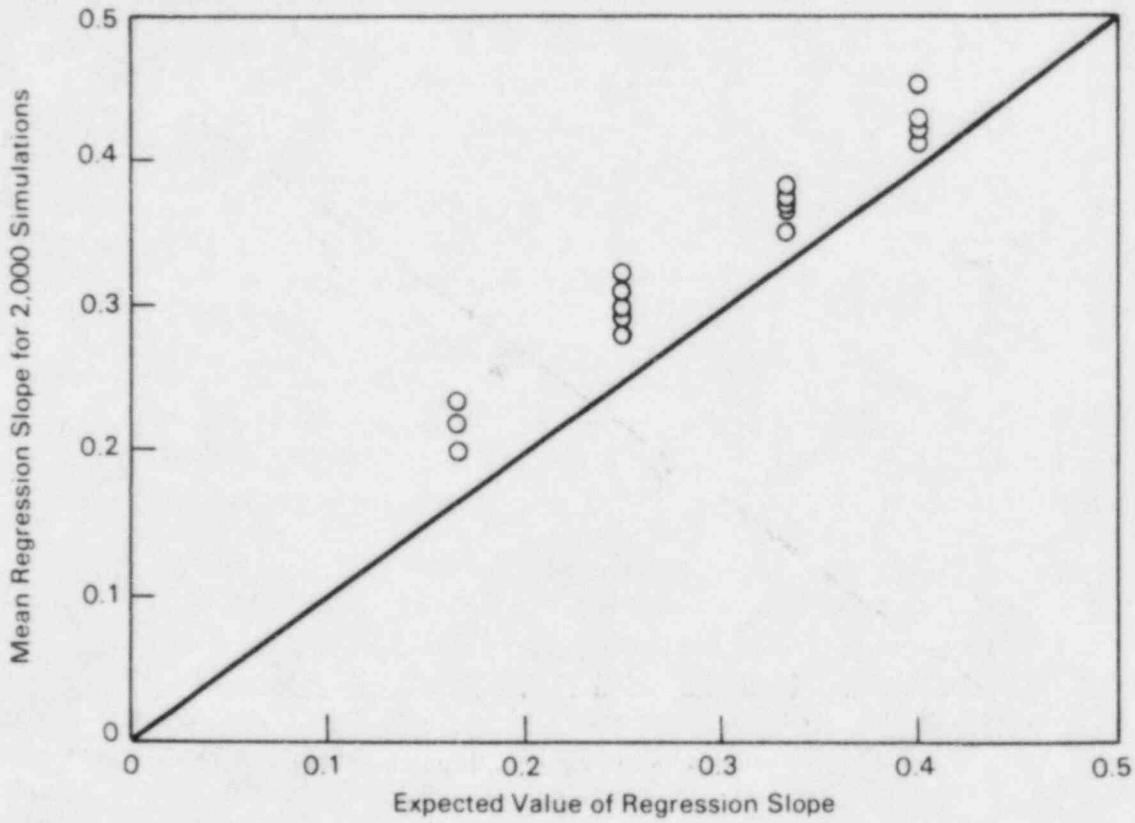

Figure 4.2 Mean values of regression slopes from 2,000 simulations plotted against calculated expected values for multiplicative and lognormal errors. The straight line shows the 1:1 relationship. Measurement error of the primary (y) variable was 10 percent of the variance of the parent population. Data selected from Appendix Table A.10, for expected coefficients of variation less than approximately 20 percent.

3.5), the expected value presumably will not be the same as for the normal, additive case. Average estimates of b for the various cases are listed in Tables A.8 to A.11.

The bias of the mean ratio ($\hat{R} = \bar{y}/\bar{x}$) can be approximated as follows (Cochran 1977:161):

$$\hat{E}(\hat{R} - R) \doteq 1/n(C_{xx} - C_{xy})R \qquad (4.2)$$

(we have dropped a finite population correction term since it is not important here). In this expression, C_{xx} and C_{xy} denote coefficients of variation and can be calculated from eq. (3.11) and (3.13) as:

$$C_{xx} = S^2_x/[E(x)]^2 = CV^2(z) + V(e_1)/\beta^2[E(z)]^2 = 1/k(1 + a/\beta^2)$$

$$C_{xy} = S_{xy}/E(x)E(y)$$

where $a = V(e_1)/V(z) = 1$, 2, 4, or 8, the ratios used (cf. Section 3.4) to obtain the range of correlations between y and x in this report, and $CV^2(z) = 1/k$. Substituting $\beta = 2$, and $R = 0.5$ (Section 3.1), we have:

$$E(\hat{R} - R) \doteq a/8nk$$

where n denotes small sample size and k determines the coefficient of variation of the parent population (0.707 for k = 2, 0.500 for k = 4, and 0.302 for k = 11).

A plot of the data for the normal and additive case (Figure 4.3) indicates relatively large deviations for the larger expected values. These again result from the more variable data sets. The simulation data (Appendix A, Tables A.12 and A.13) suggest that the more variable data sets (large C.V. of parent population, small n, lower correlations) tend to have the larger expectations and deviations from expectation. Those results where the estimated coefficient of variation was 20 percent or more are not included in Figure 4.3.

Although the prediction of bias for the ratio estimator for the normal and additive case is not as accurate as might be desirable, the relative magnitudes are nonetheless not particularly disturbing, being at most about 10 percent of the true ratio (R = 0.5) for those cases of main interest here (Figure 4.3). The multiplicative case with lognormal errors somewhat surprisingly yields a closer relationship between deviations and expected

values (Figure 4.4). The corresponding data are in Appendix A, Tables A.14 and A.15.

4.5 Linear Regression Model

We have thus far only considered a basic model in which the expected value of the auxiliary variable (x) is directly proportional to the underlying "true" value (z). As noted in Section 3.6, it is desirable to also consider linear and nonlinear alternative models. In the linear regression model of eq. (3.19), we add a constant quantity (α) to each observation. Doing so shifts the values of x_i to the right on a plot, so that the

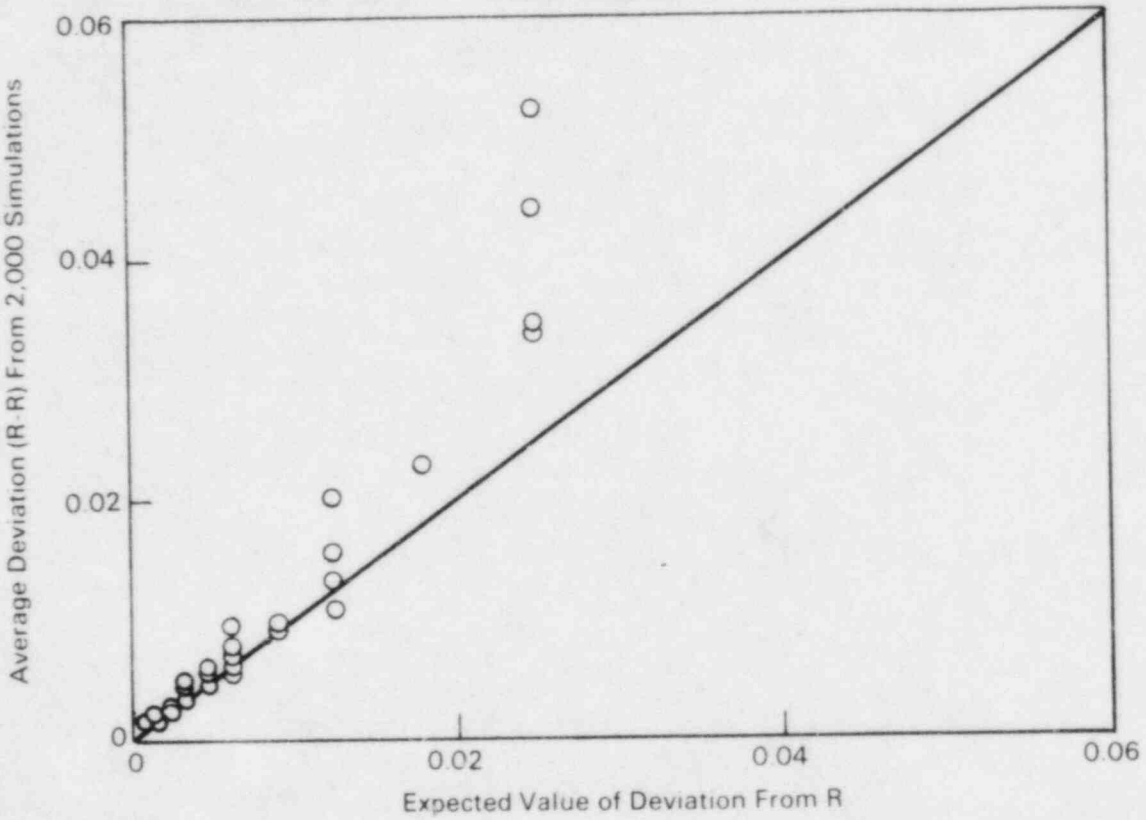

Figure 4.3 Mean values of deviation of average value of R from the true value plotted against calculated expected deviations for normal and additive errors. Measurement error of the primary (y) variable was 10 percent of the variance of the parent population. Data selected from Appendix Table A.12, for expected coefficients of variation less than approximately 20 percent.

relationship between y and x no longer goes through the origin, but is instead moved to the right, giving a negative y-intercept. A plot of a set of simulated data appears in Figure 4.5. Inasmuch as the added constant does not change the expected values of eq. (3.11) to (3.13), the various other expectations remain unchanged, and the only substantial change in results that one might expect is that the ratio estimate should not perform as well as previously, since the underlying model is now a regression model.

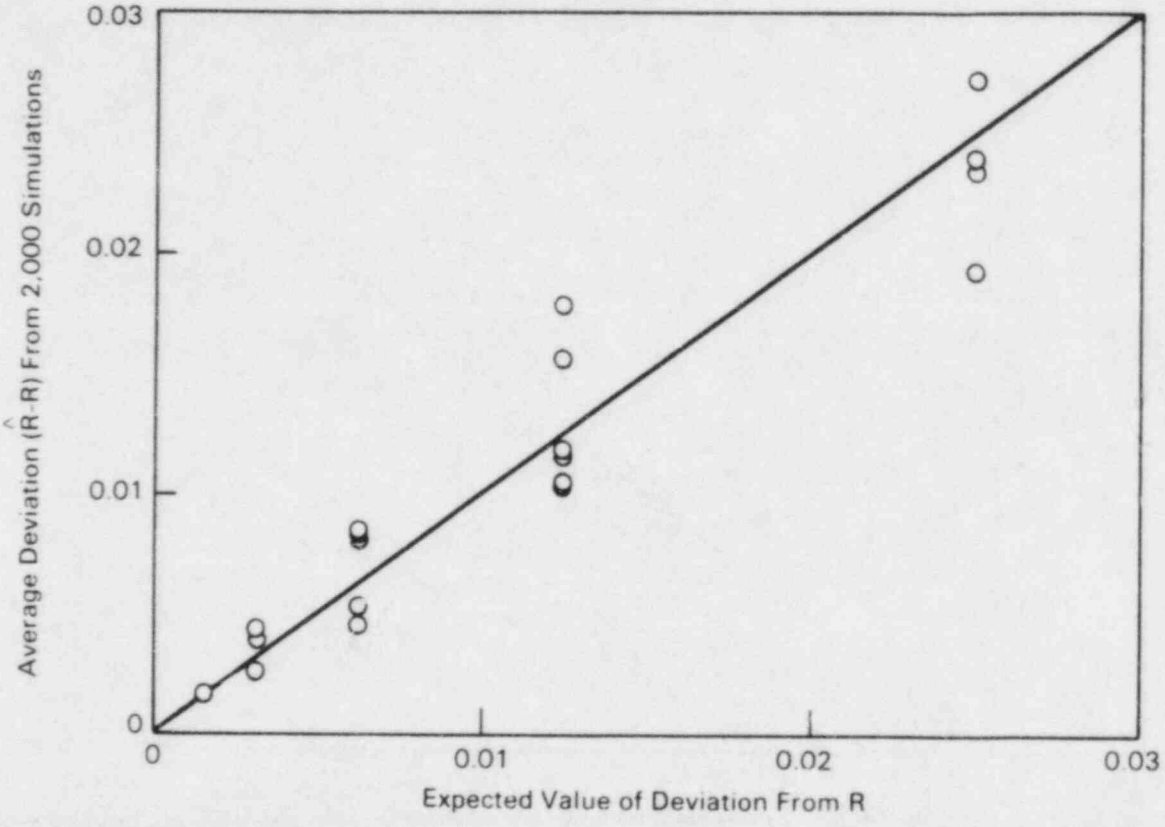

Figure 4.4 Mean values of deviation of average value of R from the true value plotted against calculated expected deviations for lognormal and multiplicative errors. Measurement error of the primary (y) variable was 10 percent of the variance of the parent population. Data selected from Appendix Table A.14, for expected coefficients of variation less than approximately 20 percent.

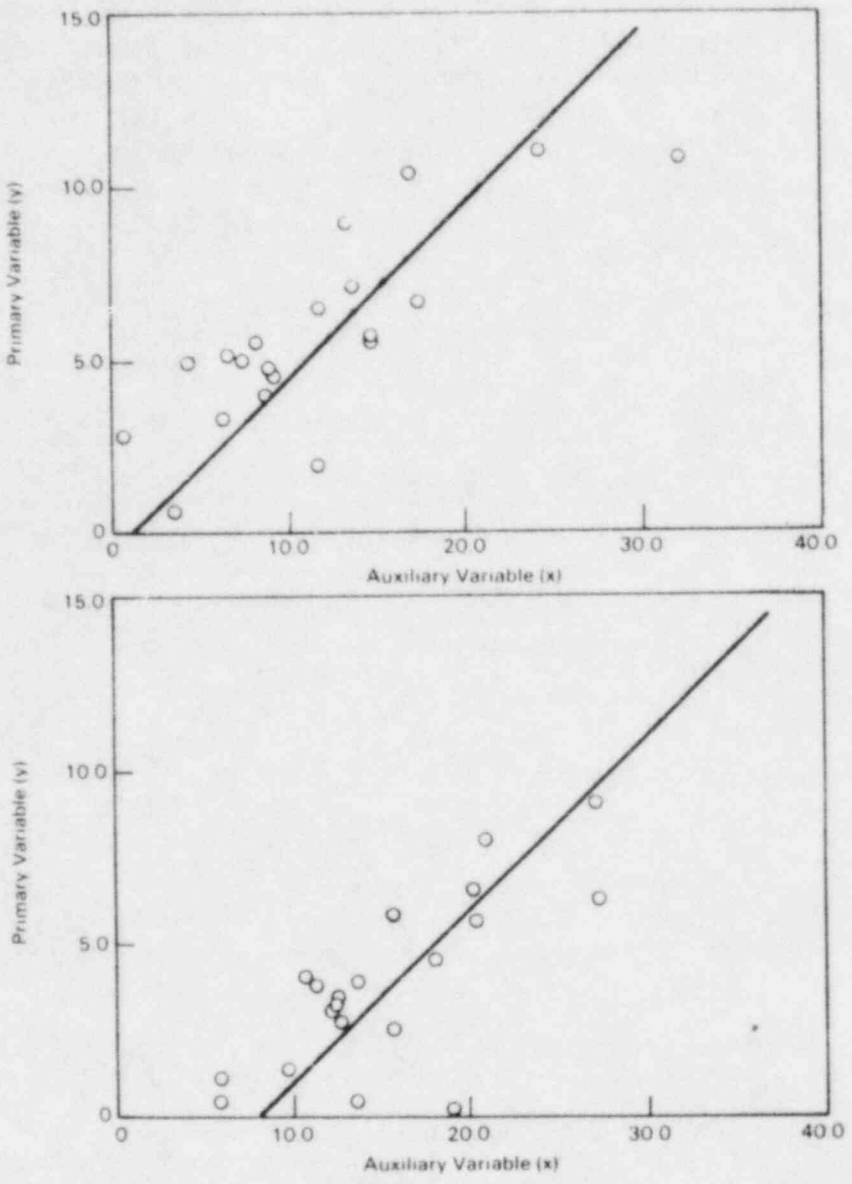

Figure 4.5 Relationship between primary (y) and auxiliary (x) variables when x is generated by linear regression [eq. (3.19)]. Measurement errors were normal and additive, with variance selected to produce a correlation of $\rho = 0.853$, and the error of the primary (y) variable 10 percent of that of the parent population. Upper set of data generated by eq. (3.19) with $\alpha = 8$ and eq. (3.3). In the lower set, $\alpha = 1.0$.

Table 4.10 Outcomes for 2,000 double sampling simulations in which the auxiliary variable (x) was generated by a linear regression model [eq. (3.19)]. Measurement errors were normal and additive, and measurement error of the primary variable (y) was 10 percent of that for the parent population. Large sample size (n') was 100. Coefficient of variation for the parent population was 0.707.

ρ for y and x	Small Sample Size (n)	Intercept for x on z (α)	Mean Value from Ratio Estimator ($\bar{y}_R$)	Mean Value from Linear Regression of y on x ($\bar{y}_{lr}$)	Ratios of Calculated Variance to Expected Value ($\bar{y}_R$)	Ratios of Calculated Variance to Expected Value ($\bar{y}_{lr}$)	Calculated Ratio	Calculated Reg. Coefficient
.853	10	1	5.05	4.94	0.95	1.07	.460	.388
		2	5.03	4.93	0.90	1.07	.419	.388
		4	4.97	4.91	0.91	1.10	.355	.389
		8	4.96	4.92	1.04	1.10	.275	.389
	20	1	5.03	4.98	0.96	1.03	.457	.392
		2	5.00	4.97	0.93	1.03	.417	.392
		4	4.99	4.96	0.92	1.02	.356	.391
		8	4.98	4.96	1.04	1.03	.276	.393
.674	10	1	5.26	4.91	1.07	1.05	.481	.240
		2	5.20	4.91	0.88	1.06	.435	.239
		4	5.10	4.90	0.69	1.05	.366	.239
		8	5.05	4.92	0.58	1.06	.281	.239
	20	1	5.10	4.95	0.95	1.02	.466	.241
		2	5.09	4.96	0.84	1.02	.426	.241
		4	5.05	4.96	0.71	1.02	.362	.242
		8	5.01	4.96	0.64	1.03	.278	.244

This expectation is borne out by a limited series of trials (Table 4.10), with the intuitively somewhat surprising result that the estimates of $\bar{y}_R$ are still quite accurate. The noticeable differences are in estimates of R and of variances (compare with Tables 4.1 and 4.2). These quantities shift with α, simply because the fitted relationship is constrained to go through the origin. Hence the estimated ratio ($\hat{R}$) decreases with increasing α and influences the variance calculation (eq. (3.8)). The fact that the estimated means ($\bar{y}_R$) remain reasonably accurate is a consequence of the nature of the estimator (eq. (3.6)). Although the ratio is distorted from its true value, it nonetheless produces an appropriate correction. If we rearrange eq. (3.6):

$$\hat{y}_R = \bar{y}(\bar{x}'/\bar{x})$$

it is apparent that, if $\bar{x}$ results from a random sample of n', then the ratio of $\bar{x}'$ to $\bar{x}$ can be expected to be approximately unity. The main question then

has to do with variances and efficiency of the estimator. However, in the present situation, we need only note that the linear regression estimator is the appropriate technique, since the underlying relationship does not go through the origin.

4.6 Nonlinear Models

Introducing a nonlinear model brings in a variety of complications, including the fact that expected values may be difficult to obtain. More importantly, it is not now known just what kinds of nonlinearities may exist in practice. Consequently, we have only conducted a limited investigation, using the model of eq. (3.20), with the two sets of parameters illustrated in Figure 3.2. Figure 4.6 gives an example of a data set generated for each of the two sets of parameters. Although the underlying curves deviate quite sharply from linearity, it is evident (Table 4.11) that the estimates of $\bar{y}_R$ and $\bar{y}_{lr}$ are nonetheless quite accurate. However, "coverage" suffers somewhat from nonlinearity but is adequate for all practical purposes.

Very similar results (Table 4.12) were obtained when the multiplicative lognormal error structure replaced the additive, normal structure used to generate the data in Table 4.11. In this case, coverage is a little less satisfactory, but still quite acceptable, and the mean calculated regression coefficients show a wider range than before. The sets of data discussed here (Table 4.11 and 4.12) were generated from a parent distribution having a coefficient of variation of 0.707. As in the other examples previously discussed, less variable results can be expected for smaller coefficients of variation. One example (normal, additive error, C.V. = 0.500) is in Appendix A (Table A.16).

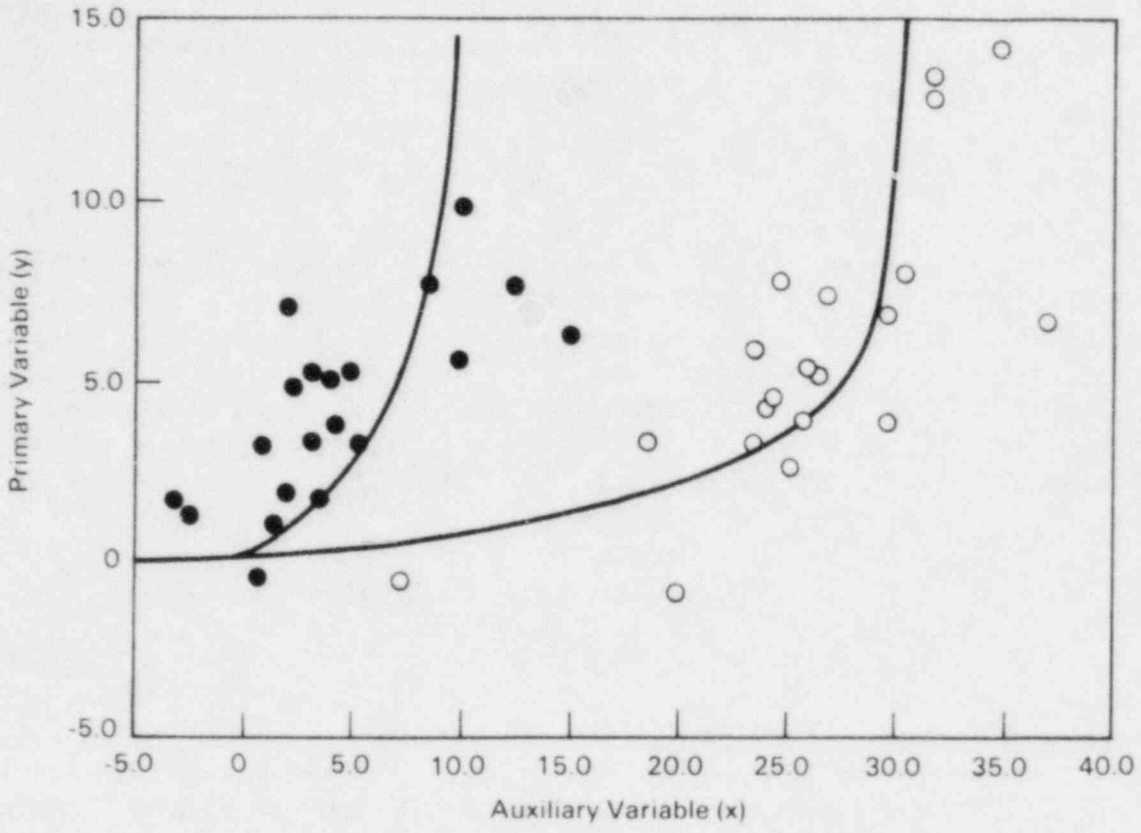

Figure 4.6 Relationship between primary (y) and auxiliary (x) variables when x is generated by a non-linear regression [eq. (3.2)]. Errors were normal and additive. Measurement error of the primary (y) variable was 10 percent that of the parent population and errors in the auxiliary (x) variable were selected to produce a correlation of $\rho = 0.853$. Data points on the left were generated using A = 10, B = 0.25 (●); those on the right with A = 30, B = 0.50 (○). Plotted lines represent the relationship when no error is included.

Table 4.11 Outcomes for 2,000 double sampling simulations in which the auxiliary variable (x) was generated by a non-linear regression model [eq. (3.20)]. Measurement errors were normal and additive, and measurement error of the primary variable (y) was 10 percent of that for the parent population. Large sample size (n') was 100. Coefficient of variation for the parent population was 0.707.

ρ for y and x	Small Sample Size (n)	Nonlinear Parameters B	A	Mean Value from Ratio Estimator ($\bar{y}_R$)	Mean Value from Linear Regression of y on x ($\bar{y}_{lr}$)	Coverage for 95% Level ($\bar{y}_R$) ($\bar{y}_{lr}$)	Calculated Ratio	Calculated Reg. Coefficient
.853	10	.25	10	5.08	4.94	.074*.075*	.820	.415
			20	5.00	4.94	.075*.073*	.402	.440
			30	4.96	4.89	.079*.081*	.266	.365
		.50	10	5.03	4.99	.079*.075*	.627	.341
			20	4.97	4.97	.080*.077*	.310	.389
			30	4.95	4.90	.079*.087*	.205	.346
	20	.25	10	5.06	4.99	.060 .058	.816	.414
			20	5.04	5.02	.055 .055	.405	.432
			30	5.02	4.99	.065*.067*	.269	.361
		.50	10	5.05	5.03	.058 .058	.630	.329
			20	5.01	5.02	.060 .059	.312	.378
			30	5.00	4.98	.067*.068*	.207	.336
.778	10	.25	10	5.35	4.99	.061 .068*	.865	.239
			20	5.02	4.96	.074*.075*	.404	.316
			30	5.00	4.94	.081*.082*	.268	.302
		.50	10	5.18	5.02	.070*.064*	.647	.186
			20	5.01	5.01	.075*.069*	.312	.263
			30	5.01	5.01	.081*.081*	.208	.269
	20	.25	10	5.19	5.04	.055 .052	.842	.242
			20	5.03	5.00	.058 .060	.405	.317
			30	4.99	4.97	.063 .065*	.268	.301
		.50	10	5.13	5.05	.059 .051	.642	.187
			20	5.03	5.03	.062 .058	.313	.262
			30	5.01	5.00	.060 .062	.208	.266
.674	10	.25	10	5.91	5.00	.067*.064*	.957	.131
			20	5.13	4.96	.069*.075*	.413	.202
			30	5.02	4.95	.062 .064*	.270	.220
		.50	10	5.33	5.01	.063*.066*	.670	.101
			20	5.06	5.02	.069*.065*	.316	.164
			30	5.02	5.03	.078*.072*	.208	.188
	20	.25	10	5.42	5.04	.052 .059	.890	.131
			20	5.10	5.04	.055 .056	.412	.204
			30	5.03	5.00	.058 .061	.270	.225
		.50	10	5.22	5.01	.052 .057	.656	.100
			20	5.07	5.04	.065*.057	.316	.163
			30	5.03	5.04	.062 .056	.209	.188

*Calculated chi-square exceeds one percent level of significance (6.63).

Table 4.12 Outcomes for 2,000 double sampling simulations in which the auxiliary variable (x) was generated by a non-linear regression model [eq. (3.20)]. Measurement errors for auxiliary variable (x) were multiplicative and lognormal, and measurement error of the primary variable (y) was 10 percent of that for the parent population. Coefficient of variation for the parent population was 0.707.

ρ for y and x	Small Sample Size (n)	Nonlinear Parameters B	Nonlinear Parameters A	Mean Value from Ratio Estimator ($\bar{y}_R$)	Mean Value from Linear Regression of y on x ($\bar{y}_{lr}$)	Coverage for 95% Level ($\bar{y}_R$)	Coverage for 95% Level ($\bar{y}_{lr}$)	Calculated Ratio	Calculated Reg. Coefficient
.853	10	.25	10	4.99	5.05	.082*	.075*	.804	.835
			20	4.96	5.01	.087*	.087*	.399	.418
			30	5.01	5.07	.080*	.074*	.269	.284
		.50	10	5.04	5.12	.083*	.074*	.628	.581
			20	5.02	5.09	.080*	.077*	.313	.292
			30	5.01	5.08	.079*	.081*	.208	.196
	20	.25	10	5.01	5.05	.074*	.071*	.807	.823
			20	5.00	5.03	.075*	.071*	.403	.414
			30	4.99	5.02	.079*	.078*	.268	.275
		.50	10	4.98	5.02	.070*	.066*	.621	.560
			20	5.02	5.05	.062	.062	.313	.288
			30	4.97	5.01	.064*	.060	.207	.192
.674	10	.25	10	5.13	5.16	.068*	.068*	.829	.448
			20	5.09	5.11	.087*	.083*	.411	.232
			30	5.12	5.15	.073*	.076*	.275	.153
		.50	10	5.13	5.11	.076*	.071*	.644	.264
			20	5.14	5.12	.077*	.069*	.321	.127
			30	5.10	5.11	.077*	.070*	.212	.086
	20	.25	10	5.02	5.03	.076*	.065*	.812	.429
			20	5.06	5.09	.079*	.068*	.409	.209
			30	5.08	5.07	.059	.057	.274	.142
		.50	10	5.03	5.02	.068*	.067*	.630	.237
			20	5.04	5.05	.066*	.063	.316	.119
			30	5.05	5.06	.078*	.064*	.211	.078

*Calculated chi-square exceeds one percent level of significance (6.63).

5.0 DISCUSSION

For the most part, the simulation results described in Section 4.0 indicate that the double sampling approach will function adequately under the range of models and parameters investigated here. One necessary caveat is that the sampling be planned so that the resulting confidence limits can be expected to be less than ±40 percent of the estimate. This realistic requirement should easily be satisfied in practice. If costs appear to preclude obtaining confidence limits narrower than ±40 percent, then serious consideration should be given to doing no sampling at all.

Our major concern at this point is the apparent lack of data on measurement errors for radionuclides likely to be of interest in a commercial radioactive low-level waste management context. We cannot determine whether our results are wholly appropriate without better knowledge of circumstances prevailing in field sampling. Further discussion of the needed data appears in Section 6.0 (Research Needs). An important point, currently widely neglected, is that additive error terms are quite unrealistic in a context of low level radioactive waste sites. To understand this, one only needs to write down a few equivalents, e.g., 5 millicuries equals 5,000 microcuries, and 5,000,000,000 picocuries. Quite obviously a realistic additive error term at one scale (e.g., microcuries) will be meaningless at another scale. This is the reason for considering multiplicative errors, and provides an explanation for the observation that the coefficient of variation is relatively constant over a wide range of concentrations. Consequently, we believe that the lognormal, multiplicative error model [eq. (3.5), Section 3.1] provides the most realistic model for the auxiliary variable (x). As previously remarked, the nature of the error term for the primary ("accurate") measurement may depend on the way in which measurement errors are controlled.

We emphasize that our simulations serve mainly to show that the double sampling approach is "robust" over the range of models and parameters tested. The variance equations (3.8 and 3.9, Section 3.2) generally give satisfactory results for small samples, ranging down to as few as $n = 5$ for the "expensive" method, in the sense that "coverage" of calculated confidence limits using these variances is close to the assumed levels (5 and 10 percent). In actual applications one can thus safely use these variance equations (as given by Cochran, 1977:343-344) to check whether double sampling is worthwhile or not, along with the available information on correlation between the primary (expensive) and auxiliary (inexpensive) variables. Simulations like those used here can only demonstrate feasibility under assumed conditions, and both experience and preliminary field data must serve to determine whether or not a given technique should actually be used in practice.

6.0 RESEARCH NEEDS

The outcomes of the simulations presented in this report serve to establish the feasibility of using double sampling as a cost-efficient tool in low level waste management. Effective use of this methodology will, however, need to be guided by a detailed study of actual data on measurement errors. Our preliminary investigations suggest that little of this data is available, and that some field and laboratory studies may be needed. Such studies would also be very valuable in establishing other details of applications of double sampling and of general efficiency in sampling. We have been using what is, in effect, a two stage procedure in other sampling contexts, that may be very useful for double sampling applications. The procedure is essentially to collect a large random sample of, say, soils, and analyze portions of the sample sequentially. In a double sampling context, if certain conditions for ratio estimation are met, it might be possible to deliberately select the samples for the expensive measurement (y) on the basis of the observed values of the auxiliary variable (x). This approach has the potential of being very efficient, and thus reducing sampling costs substantially.

A different area of needed research concerns the use of double sampling in defining the area and extent of a "spill" or other source of contamination and in efficiently guiding any procedures used to contain or "clean up" such a contaminated area. Skalski and Thomas (1984) discuss an approach using compositing procedures. When a useful auxiliary variable is available (such as data from a field detection instrument), it may be quite feasible to utilize double sampling for stratification (see Eberhardt and Thomas, 1983, pp. 4-9 and 4-11) as an alternative approach. In this case, double sampling would be used to define strata (we used double sampling in this report for estimation purposes), and thus might well serve to define areas of different levels of contamination.

Most of the available textbook results on double sampling are based on approximations of one kind or another. Many of these start with Taylor series approximations to a ratio, since as Cochran (1977:153) remarked, "The distribution of the ratio estimate has proved annoyingly intractable because both y and x vary from sample to sample." Since the results of the present simulations suggest that Cochran's (1977) basic equations perform reasonably well under assumptions quite different from those normally given for double sampling, it may well be that the various approximations have a considerable degree of generality. Hence, another area of needed research is to re-examine the available derivations to see whether they can be considered appropriate under the revised assumptions of the present study. The need for

this research stems from the fact that simulations are essentially "narrow-minded" procedures, when compared to mathematical analysis. For example, we found it necessary to use specific values for certain parameters, and these parameters may not be appropriate in all CRLLWS circumstances.

7.0 INTERIM GUIDANCE FOR USING RATIO METHODS

A large part of this report has dealt with the technique known as double sampling, which is a special case of a broader classification described here as ratio methods. The important distinction is that ratio methods were developed to fit circumstances where a total (or overall mean) value was available for an auxiliary (inexpensive) variable measured over the entire population being studied. In many commercial low-level radioactive waste site (CRLLWS) situations, an auxiliary variable can readily be measured (e.g., gross beta or gamma), but a population total is not available. Double sampling was developed to permit substituting data from a large sample of the population for knowledge of the overall mean or total of the auxiliary variable. Since it is the main practical use of ratio methods at CRLLWS, the following example concerns double sampling. As noted in Section 2.0, the original ratio methods will mainly be used when information on the auxiliary variable is already on hand, and is thus essentially "free" as far as the field work is concerned. An evaluation as to whether such "free" data will reduce the variance of an estimated mean or total can be obtained by calculating a correlation coefficient, and coefficients of variation for the expensive and inexpensive methods (see Section 2.0).

Our example for application of double sampling is based on data analyzed by Gilbert and Eberhardt (1976). Those authors evaluated the prospective use of double sampling on previously collected data on concentrations of plutonium in soil and readings made with either a field instrument ("FIDLER") or a laboratory instrument using a lithium-drifted germanium crystal [Ge(Li) detector] to measure soil concentrations of americium-241. Cost ratios were based on the analytical cost (at that time) of a single soil sample for plutonium by a commercial laboratory. These single sample costs were about double those for using a laboratory Ge(Li) detector, and about 50 times the cost of using the field instrument (FIDLER) to make one measurement. We caution that cost ratios will vary with circumstances and substance being studied, and must be determined independently for each study. In this example, we assume that the ratio of cost (c) for the expensive measurement (y) to the cost (c') for the auxiliary measurement (x) is either 2 or 50, depending on the analytical method used.

In our example application we start with an overall budget (C) which will be allocated to a sample of laboratory analyses (n) and to making n' instrument readings in the field or laboratory. We then must estimate the correlation between the two kinds of measurement and use either eq. (2.2) or Figure 2.1 to determine whether double sampling is cost effective. From eq. (2.2), we find that we need:

$$\rho > \left[4(2)/(1+2)^2\right]^{1/2} = 0.94 \text{ for the Ge(Li) detector, and}$$

$$\rho > \left[4(50)/(1+50)^2\right]^{1/2} = 0.28 \text{ for the FIDLER (field instrument).}$$

If the actual correlation is greater than the right hand side of eq. (2.2) then we would allocate sampling effort to the two methods according to an equation given by Cochran (1977:341):

$$n/n' = \left\{(c'/c)\left[\left(1 - \rho^2\right)/\rho^2\right]\right\}^{1/2}$$

If we assume, for example, that the correlation between results from the field instrument and soil plutonium is $\rho = 0.50$ [Gilbert and Eberhardt (1976:Table 1) give data on observed correlations], then:

$$n/n' = \left\{1/50[(1-0.25)/0.25]\right\}^{1/2} = 0.245.$$

Hence we should take about $1/0.245 = 4$ instrument readings for each laboratory determination. Consequently, if the overall budget is, say, 10,000 dollars, and a laboratory determination costs 100 dollars (c), and the field reading thus costs 2 dollars (c'), then we can use eq. (2.1) (Section 2.0):

$$10{,}000 = 50(n) + 2(n')$$

to express our overall budget, and substitute the ratio $n/n' = 0.245$ to calculate the two respective sample sizes as follows:

$$10{,}000 = 50(0.245\ n') + 2(n')$$

$$702 = n'$$

since $n = 0.245n'$

then $n = 172.$

Given the two sample sizes, one would then actually collect the samples and take the measurements, after which eq. (3.6) and (3.7) can be used to estimate mean concentrations, while eq. (3.8) and (3.9) serve to provide variance estimates, from which confidence limits can be calculated. We illustrate these calculations from a set of the data studied by Gilbert and Eberhardt (1976). Table 7.1 shows the FIDLER readings and corresponding laboratory determinations of plutonium concentration. The overall total number of FIDLER readings (n') was taken to be 165 with a mean ($\bar{x}'$) of 10. In practice, these data would be obtained by going to n' randomly selected field locations and taking FIDLER readings. From these n' locations, n would

Table 7.1 An example (from Gilbert and Eberhardt, 1976) illustrating double sampling computations. The auxiliary variable (x) is FIDLER counts of ^{241}Am, while the primary variable (y) results from a radiochemical analysis for plutonium. For the calculations the overall mean ($\bar{x}'$) was taken to be 10.0, and the sample size (n') was 165. s_y^2 and s_x^2 are the variances of y and x respectively, s_{yx} is the covariance.

FIDLER (x)	Radio-Chemical (y)	FIDLER (x)	Radio-Chemical (y)	FIDLER (x)	Radio-Chemical (y)
5.2	17.056	1.4	4.277	10.8	64.336
18.6	58.383	19.6	136.610	9.8	51.081
7.8	67.579	7.2	61.097	10.8	24.626
10.8	36.824	6.2	39.119	16.2	64.640
3.2	2.939	3.8	13.115	3.2	14.834
8.8	0.886	6.2	19.567	23.0	172.310
15.2	15.749	11.8	30.885	10.2	7.622
15.0	30.467				

$\bar{y} = 42.45$ $s_y^2 = 1814.76$ $s_{yx} = 179.41$

$\bar{x} = 10.22$ $s_x^2 = 33.37$

be randomly selected and at each of these a soil sample would also be collected and analyzed for plutonium in the laboratory. Table 7.1 thus pertains to the data collected at these n locations.

Both ratio and regression estimates are calculated since our simulations suggest that either may be appropriate. When enough data are available, one might choose between the two on the basis of statistical theory. If the actual relationship goes through the origin and the variance increases approximately in proportion to x, one would use a ratio calculation. If the actual relationship clearly does not go through the origin, one would likely prefer the regression calculation. In practice, there may not be enough data or experience for a clear cut choice, so both calculations are carried out here.

Ratio calculations using data from Table 7.1 are:

$$\bar{y}_R = (\bar{y}/\bar{x})\bar{x}'$$
$$= (42.45/10.22)10$$
$$\bar{y}_R = 41.54$$

For the variance estimate the total parent population is assumed to contain 1,000 units (N), and R is 4.15 (i.e., 42.45/10.22). The estimated variance is calculated as follows:

$$v(\bar{y}_R) = \left(s_y^2 - 2Rs_{yx} + R^2s^2_x\right)/n + \left(2Rs_{yx} - R^2s^2_x\right)/n' - s^2_y/N$$

$$= [1814.76 - 2(4.15)(179.41) + 17.22(33.37)]/22 + [2(4.15)(179.41 - 17.22(33.37)]/165 - 1814.76/1000$$

$$v(\bar{y}_R) = 44.63$$

Calculations for the regression method are:

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x})$$

$$= 42.45 + 5.38(10 - 10.22)$$

$$\bar{y}_{lr} = 41.27$$

where b = 5.38 (i.e., $b = s_{yx}/s_x^2 = 179.41/33.37$). For the variance estimate, the calculations are:

$$v(\bar{y}_{lr}) = s^2_{y.x}\left\{1/n + \left[(\bar{x}' - \bar{x})^2/\Sigma(\bar{x}_i - \bar{x})^2\right]\right\} + \left(s^2_y - s^2_{y.x}\right)/n' - s^2_y/N$$

$$= 892.56\{1/22 + [(10-10.22)^2/700.71]\} + [(1814.76 - 892.56)/165] - 1814.76/1000$$

$$v(y_{lr}) = 44.41$$

where $s_{y.x}^2 = 892.56$ (i.e., $s_{y.x}^2 = \left[s_y^2(n - 1) - b^2s_x^2(n - 1)\right]/(n - 2)$

$$= [1814.76(21) - 28.94(33.37)(21)]/20)$$

8.0 LITERATURE CITED

Bratley, P., B. L. Fox, and L. E. Schrage. 1983. A Guide to Simulation. Springer-Verlag, New York. xix+383pp.

Cochran, W. G. 1977. Sampling Techniques, 3rd. ed. John Wiley and Sons, New York. xvi+429pp.

Eberhardt, L. L., and J. M. Thomas. 1983. Survey of Statistical and Sampling Needs for Environmental Monitoring of Commercial Low-Level Radioactive Waste Disposal Facilities: A Progress Report in Response to Task 1. PNL-4804 Pacific Northwest Laboratory, Richland, Washington.

Eberhardt, L. L., R. O. Gilbert, H. L. Hollister, and J. M. Thomas. 1976. "Sampling for Contaminants in Ecological Systems." Environ. Sci. and Technol. 10:917-925.

Gilbert, R. O. and L. L. Eberhardt. 1976. "An Evaluation of Double Sampling for Estimating Plutonium Inventory in Surface Soil. In Radioecology and Energy Resources, eds. C. E. Cushing, Jr., et al., pp. 157-163. Dowden, Hutchinson and Ross Inc., Stroudsburg, Pennsylvania.

Goodman, L. A. 1960. "On the Exact Variance of Products". J. Amer. Stat. Assoc. 55:708-713.

Naylor, T. H., J. L. Balintfy, D. S. Burdick, and K. Chu. 1968. Computer Simulation Techniques. J. Wiley and Sons, New York. xiii+352 pp.

Simpson, J. C., R. S. Harkins, and C. R. Watson. 1979. "Evaluation of the Multiplier in the Multiplicative Congruential Pseudo-Random Number Generator". Proc. Dig. Equip. Users Soc. 705-710.

Skalski, J. R., and J. M. Thomas. 1984. Improved Field Sampling Design and Compositing Schemes for Cost Effective Detection of Migration and Spills at Commercial Low-Level Radioactive or Chemical Waste Sites. PNL-4935, Pacific Northwest Laboratory, Richland, Washington.

APPENDIX A

ADDITIONAL SIMULATION RESULTS

APPENDIX A

ADDITIONAL SIMULATION RESULTS

Since a number of the simulation outputs are quite similar to data already presented in the tables of Section 4.0, they are tabulated and presented in this Appendix.

Table A.1 Coefficients of variation based on means and expected variances for $\bar{y}_R$ and $\bar{y}_{lr}$ for selected portions of Tables 4.1 and 4.5.

Error in Primary Variable (y)	Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853*	.778	.674	.550
.10	.707	100	5	$\bar{y}_R$	19.54	24.89	33.08	45.22
			10		14.39	17.87	23.33	31.54
			20		10.93	13.02	16.42	21.67
		100	5	$\bar{y}_{lr}$	18.29	21.47	24.89	27.89
			10		13.58	15.64	17.87	19.86
			20		10.46	11.67	13.02	14.24
	.500	100	5	$\bar{y}_R$	13.82	17.60	23.39	31.97
			10		10.17	12.64	16.50	22.30
			20		7.73	9.21	11.61	15.32
		100	5	$\bar{y}_{lr}$	12.93	15.18	17.60	19.72
			10		9.60	11.06	12.64	14.04
			20		7.40	8.25	9.21	10.07
.25	.707	100	5	$\bar{y}_R$	23.05	27.73	35.27	46.84
			10		16.77	19.84	24.87	32.69
			20		12.50	14.36	17.50	22.50
		100	5	$\bar{y}_{lr}$	21.99	24.71	27.73	30.45
			10		16.09	17.85	19.84	21.65
			20		12.09	13.15	14.36	15.48
	.500	100	5	$\bar{y}_R$	16.30	19.61	24.94	33.12
			10		11.86	14.03	17.59	23.12
			20		8.84	10.16	12.37	15.91
		100	5	$\bar{y}_{lr}$	15.55	17.47	19.61	21.53
			10		11.37	12.62	14.03	15.31
			20		8.55	9.30	10.16	10.94

*Correlation values are .800, .730, .632, .516 for lower half of table (error in primary variable = .25).

Table A.2 "Coverage" of calculated confidence limits for mean $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Values shown are proportions of the total simulations for which the 90 percent confidence limits constructed from simulated data do not include the "true" mean [E(z), expected mean for gamma distribution]. Measurement errors were normal and additive, and measurement error of the primary variable (y) was 10 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
.707	100	5	$\bar{y}_R$	.104	.106	.110	.113
		10		.102	.114	.112	.104
		20		.107	.110	.105	.095
	100	5	$\bar{y}_{lr}$	.097	.121*	.128*	.146*
		10		.108	.122*	.125*	.127*
		20		.118	.120*	.120*	.117
.500	100	5	$\bar{y}_R$	.086	.092	.086	.087
		10		.095	.099	.104	.100
		20		.102	.110	.091	.097
	100	5	$\bar{y}_{lr}$	.105	.120*	.113	.129*
		10		.103	.108	.113	.114
		20		.109	.102	.099	.094
.302	100	5	$\bar{y}_R$	.087	.106	.082*	.112
		10		.094	.095	.095	.088
		20		.093	.089	.091	.111
	100	5	$\bar{y}_{lr}$	.104	.108	.112	.112
		10		.103	.105	.105	.103
		20		.087	.098	.107	.103

* Calculated chi-square exceeds one percent level of significance (6.63).

Table A.3 "Coverage" of calculated confidence limits for mean $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Values shown are proportions of the total simulations for which the 95 percent confidence limits constructed from simulated data do not include the "true" mean [i.e., the mean of the finite population (N) of 1,000]. Measurement errors were normal and additive, and measurement error of the primary variable (y) was 10 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.674	.550
.707	100	5	$\bar{y}_R$	.047	.060	.054	.063
		10		.051	.059	.054	.051
		20		.051	.052	.050	.042
	100	5	$\bar{y}_{lr}$	.048	.061	.070*	.082*
		10		.053	.070*	.068*	.071*
		20		.054	.059	.060	.058
.500	100	5	$\bar{y}_R$	.044	.049	.035*	.041
		10		.046	.050	.056	.054
		20		.051	.046	.039	.052
	100	5	$\bar{y}_{lr}$	.058	.059	.060	.076*
		10		.050	.057	.057	.057
		20		.052	.043	.045	.046
.302	100	5	$\bar{y}_R$	.041	.049	.040	.056
		10		.041	.047	.043	.050
		20		.047	.043	.046	.061
	100	5	$\bar{y}_{lr}$	.056	.051	.059	.065*
		10		.049	.053	.048	.053
		20		.041	.040	.047	.049

* Calculated chi-square exceeds one percent level of significance (6.63).

Table A.4 "Coverage" of calculated confidence limits for mean $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Values shown are proportions of the total simulations for which the 90 percent confidence limits constructed from simulated data do not include the "true" mean (i.e., the mean of the finite population (N) of 1,000]. Measurement errors were normal and additive, and measurement error of the primary variable (y) was 25 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .853	.778	.670	.550
.707	100	5	$\bar{y}_R$	.105	.107	.107	.111
		10		.096	.108	.115	.105
		20		.101	.101	.107	.093
	100	5	$\bar{y}_{lr}$	.099	.121*	.121*	.146*
		10		.106	.122*	.122*	.126*
		20		.111	.111	.110	.116
.500	100	5	$\bar{y}_R$	.086	.096	.089	.088
		10		.090	.093	.104	.100
		20		.094	.105	.086	.097
	100	5	$\bar{y}_{lr}$	.106	.120*	.111	.127*
		10		.098	.110	.112	.109
		20		.102	.095	.099	.096
.302	100	5	$\bar{y}_R$	.085	.104	.082*	.110
		10		.091	.094	.093	.087
		20		.090	.086	.088	.110
	100	5	$\bar{y}_{lr}$	.101	.107	.109	.112
		10		.101	.102	.106	.101
		20		.091	.087	.096	.098

* Calculated chi-square exceeds one percent level of significance (6.63).

Table A.5 Means of $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement error of auxiliary variable (x) was multiplicative and lognormal, and measurement error of primary variable (y) was 25 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .800	.730	.632	.516
.707	100	5	$\bar{y}_R$	5.10	5.20	5.34	5.69
		10		5.04	5.09	5.21	5.35
		20		5.04	5.03	5.12	5.17
	100	5	$\bar{y}_{lr}$	5.14*	5.21*	5.38*	5.50*
		10		5.04	5.10	5.17*	5.21*
		20		5.06	5.03	5.11	5.12
.500	100	5	$\bar{y}_R$	5.07	5.10	5.22	5.40
		10		5.04	5.06	5.09	5.23
		20		5.10	5.04	5.05	5.10
	100	5	$\bar{y}_{lr}$	5.09	5.10	5.17	5.22*
		10		5.04	5.05	5.09	5.14
		20		5.01	5.04	5.04	5.05

* Coefficient of variation of estimated mean greater than 20 percent.

Table A.6 Ratios of the mean calculated variances to the expected variance for $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Measurement error of auxiliary variable (x) was multiplicative and lognormal, and measurement error of primary variable (y) was 25 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .800	.730	.632	.516
.707	100	5	$\bar{y}_R$	0.90	0.84	0.72	0.59
		10		0.95	0.89	0.81	0.71
		20		0.97	0.94	0.91	0.80
	100	5	$\bar{y}_{lr}$	1.54*	1.49*	1.89*	1.93*
		10		1.07	1.06	1.08*	1.14*
		20		1.01	1.00	1.00	1.01
.500	100	5	$\bar{y}_R$	0.96	0.88	0.81	0.75
		10		0.98	0.93	0.89	0.82
		20		0.99	0.97	0.94	0.88
	100	5	$\bar{y}_{lr}$	1.49	1.34	1.43	1.67*
		10		1.09	1.08	1.11	1.15
		20		1.03	1.02	1.03	1.03

* Coefficient of variation of estimated mean greater than 20 percent.

Table A.7 "Coverage" of calculated confidence limits for mean $\bar{y}_R$ and $\bar{y}_{lr}$ estimated by double sampling (2,000 simulations). Values shown are proportions of the total simulations for which the 95 percent confidence limits constructed from simulated data do no include the "true" mean [E(z), expected mean for gamma distribution]. Measurement error of auxiliary variable (x) was multiplicative and lognormal, and measurement error of primary variable (y) was 25 percent of that for the parent population.

Coefficient of Variation of Parent Population	Large Sample Size (n')	Small Sample Size (n)	Estimate Evaluated	Correlation (ρ) Between Primary (y) and Auxiliary (x) Variable .800	.730	.632	.516
.707	100	5	$\bar{y}_R$	.074*	.087*	.102*	.140*
		10		.072*	.082*	.096*	.127*
		20		.067*	.064*	.072*	.094*
	100	5	$\bar{y}_{lr}$	.064*	.061	.077*	.087*
		10		.061	.066*	.067*	.074*
		20		.060	.058	.062	.064*
.500	100	5	$\bar{y}_R$	.060	.066*	.086*	.098*
		10		.055	.060	.071*	.093*
		20		.048	.078*	.070*	.076*
	100	5	$\bar{y}_{lr}$	.058	.064*	.070*	.068*
		10		.049	.054	.057	.059
		20		.047	.060	.061	.055

* Calculated chi-square exceeds one percent level of significance (6.63).

Table A.8 Expected values of regression slope [E(b)] and mean values (b) calculated from 2,000 double sampling simulations. Measurement error of the auxiliary variable (x) was normal and additive, and measurement error of the primary variable (y) was 10 percent of that for the parent population. Large sample size (n') was 100.

Correlation Between Primary (y) and Auxiliary (x) Variable	Coefficient of Variation of Parent Population	Small Sample Size (n)	Expected Value of Regression Coefficient E(b)	Average Regression Coefficient from from Simulations (b)
.550	.707	5	.1667	.1580
		10	.1667	.1554
		20	.1667	.1627
	.500	5	.1667	.1614
		10	.1667	.1621
		20	.1667	.1672
	.302	5	.1667	.1631
		10	.1667	.1623
		20	.1667	.1675
.670	.707	5	.2500	.2289
		10	.2500	.2356
		20	.2500	.2443
	.500	5	.2500	.2423
		10	.2500	.2389
		20	.2500	.2457
	.302	5	.2500	.2369
		10	.2500	.2429
		20	.2500	.2460
.778	.707	5	.3333	.3044
		10	.3333	.3140
		20	.3333	.3241
	.500	5	.3333	.3171
		10	.3333	.3262
		20	.3333	.3272
	.302	5	.3333	.3279
		10	.3333	.3254
		20	.3333	.3299
.853	.707	5	.4000	.3770
		10	.4000	.3852
		20	.4000	.3944
	.500	5	.4000	.3915
		10	.4000	.3934
		20	.4000	.3974
	.302	5	.4000	.3945
		10	.4000	.3960
		20	.4000	.3999

Table A.9 Expected values of regression slope [E(b)] and mean values (b) calculated from 2,000 double sampling simulations. Measurement error of the auxiliary variable (x) was normal and additive, and measurement error of primary variable (y) was 25 percent of that for the parent population. Large sample size (n') was 100.

Correlation Between Primary (y) and Auxiliary (x) Variable	Coefficient of Variation of Parent Population	Small Sample Size (n)	Expected Value of Regression Coefficient E(b)	Average Regression Coefficient from from Simulations (b)
.550	.707	5	.1667	.1511
		10	.1667	.1627
		20	.1667	.1640
	.500	5	.1667	.1650
		10	.1667	.1646
		20	.1667	.1653
.670	.707	5	.2500	.2200
		10	.2500	.2355
		20	.2500	.2406
	.500	5	.2500	.2304
		10	.2500	.2405
		20	.2500	.2480
.778	.707	5	.3333	.3022
		10	.3333	.3193
		20	.3333	.3254
	.500	5	.3333	.3147
		10	.3333	.3236
		20	.3333	.3262
.853	.707	5	.4000	.3791
		10	.4000	.3931
		20	.4000	.3928
	.500	5	.4000	.3916
		10	.4000	.3932
		20	.4000	.3972

Table A.10 Expected values of regression slope [E(b)] and mean values (b) calculated from 2,000 double sampling simulations. Measurement error of the auxiliary variable (x) was multiplicative and lognormal, and measurement error of the primary variable (y) was 10 percent of that for the parent population. Large sample size (n') was 100.

Correlation Between Primary (y) and Auxiliary (x) Variable	Coefficient of Variation of Parent Population	Small Sample Size (n)	Expected Value of Regression Coefficient E(b)	Average Regression Coefficient from from Simulations (b)
.550	.707	5	.1667	.3087
		10	.1667	.2598
		20	.1667	.2320
	.500	5	.1667	.2436
		10	.1667	.2197
		20	.1667	.2004
.670	.707	5	.2500	.3554
		10	.2500	.3218
		20	.2500	.2980
	.500	5	.2500	.3091
		10	.2500	.2904
		20	.2500	.2794
.778	.707	5	.3333	.4008
		10	.3333	.3812
		20	.3333	.3709
	.500	5	.3333	.3766
		10	.3333	.3648
		20	.3333	.3532
.853	.707	5	.4000	.4522
		10	.4000	.4278
		20	.4000	.4208
	.500	5	.4000	.4291
		10	.4000	.4188
		20	.4000	.4120

Table A.11 Expected values of regression slope [E(b)] and mean values (b) calculated from 2,000 double sampling simulations. Measurement error of the auxiliary variable (x) was multiplicative and lognormal, and measurement error of the primary variable (y) was 25 percent of that for the parent population. Large sample size (n') was 100.

Correlation Between Primary (y) and Auxiliary (x) Variable	Coefficient of Variation of Parent Population	Small Sample Size (n)	Expected Value of Regression Coefficient E(b)	Average Regression Coefficient from from Simulations (b)
.550	.707	5	.1667	.3231
		10	.1667	.2632
		20	.1667	.2282
	.500	5	.1667	.2385
		10	.1667	.2212
		20	.1667	.2016
.670	.707	5	.2500	.3628
		10	.2500	.3248
		20	.2500	.2963
	.500	5	.2500	.3003
		10	.2500	.2907
		20	.2500	.2801
.778	.707	5	.3333	.4085
		10	.3333	.3824
		20	.3333	.3656
	.500	5	.3333	.3595
		10	.3333	.3599
		20	.3333	.3528
.853	.707	5	.4000	.4480
		10	.4000	.4259
		20	.4000	.4241
	.500	5	.4000	.4235
		10	.4000	.4177
		20	.4000	.4114

Table A.12 Deviations of calculated ratio from true value ($\hat{R} - R$) compared to expected value calculated from eq. (4.2). Measurement errors were normal and additive and measurement error of primary variable (y) was 10 percent of that for the parent population. Large sample size (n') was 100.

Correlation Between Primary (y) and Auxiliary (x) Variable	Coefficient of Variation of Parent Population	Small Sample Size (n)	Expected Deviation Between Estimated and Actual R $E(\hat{R} - R)$	Average Deviation Between Calculated and True R $(\hat{R} - R)$
.853	.707	5	.0125	.0203
		10	.0063	.0094
		20	.0031	.0048
	.500	5	.0063	.0070
		10	.0031	.0045
		20	.0016	.0013
	.302	5	.0023	.0027
		10	.0011	.0022
		20	.0006	.0015
.778	.707	5	.0250	.0646
		10	.0125	.0156
		20	.0063	.0056
	.500	5	.0125	.0115
		10	.0063	.0080
		20	.0031	.0036
	.302	5	.0045	.0061
		10	.0023	.0021
		20	.0011	.0014
.670	.707	5	.0500	.1174
		10	.0250	.0441
		20	.0125	.0131
	.500	5	.0250	.0344
		10	.0125	.0133
		20	.0063	.0063
	.302	5	.0091	.0092
		10	.0045	.0046
		20	.0023	.0024
.550	.707	5	.1000	-.5093
		10	.0500	.0664
		20	.0250	.0339
	.500	5	.0500	.0989
		10	.0250	.0523
		20	.0125	.0131
	.302	5	.0182	.0231
		10	.0091	.0097
		20	.0045	.0058

Table A.13 Deviations of calculated ratio from true value ($\hat{R} - R$) compared to expected value calculated from eq. (4.2). Measurement errors were normal and additive and measurement error of primary variable (y) was 25 percent of that for the parent population. Large sample size (n') was 200.

Correlation Between Primary (y) and Auxiliary (x) Variable	Coefficient of Variation of Parent Population	Small Sample Size (n)	Expected Deviation Between Estimated and Actual R $E(\hat{R} - R)$	Average Deviation Between Calculated and True R $(\hat{R} - R)$
.853	.707	5	.0125	.0194
		10	.0063	.0074
		20	.0031	.0036
	.500	5	.0063	.0077
		10	.0031	.0032
		20	.0016	.0015
.778	.707	5	.0250	.0642
		10	.0125	.0185
		20	.0063	.0064
	.500	5	.0125	.0162
		10	.0063	.0087
		20	.0031	.0046
.670	.707	5	.0500	.1945
		10	.0250	.0411
		20	.0125	.0144
	.500	5	.0250	.0421
		10	.0125	.0169
		20	.0063	.0064
.550	.707	5	.1000	.2581
		10	.0500	.0542
		20	.0250	.0361
	.500	5	.0500	.1241
		10	.0250	.0260
		20	.0125	.0156

Table A.14 Deviations of calculated ratio from true value ($\hat{R} - R$) compared to expected value calculated from eq. (4.2). Measurement error of the auxiliary variable (x) was multiplicative and lognormal and measurement error of primary variable (y) was 10 percent of that for the parent population. Large sample size (n') was 100.

Correlation Between Primary (y) and Auxiliary (x) Variable	Coefficient of Variation of Parent Population	Small Sample Size (n)	Expected Deviation Between Estimated and Actual R $E(\hat{R} - R)$	Average Deviation Between Calculated and True R $(\hat{R} - R)$
.853	.707	5	.0125	.0120
		10	.0063	.0047
		20	.0031	.0027
	.500	5	.0063	.0054
		10	.0031	.0039
		20	.0016	.0016
.778	.707	5	.0250	.0183
		10	.0125	.0103
		20	.0063	.0084
	.500	5	.0125	.0178
		10	.0063	.0080
		20	.0031	.0043
.670	.707	5	.0500	.0445
		10	.0250	.0239
		20	.0125	.0119
	.500	5	.0250	.0193
		10	.0125	.0106
		20	.0063	.0078
.550	.707	5	.1000	.0757
		10	.0500	.0488
		20	.0250	.0271
	.500	5	.0500	.0402
		10	.0250	.0235
		20	.0125	.0156

Table A.15 Deviations of calculated ratio from true value ($\hat{R} - R$) compared to expected value calculated from eq. (4.2). Measurement errors of the auxiliary variable (x) was multiplicative and lognormal and measurement error of primary variable (y) was 25 percent of that for the parent population. Large sample size (n') was 100.

Correlation Between Primary (y) and Auxiliary (x) Variable	Coefficient of Variation of Parent Population	Small Sample Size (n)	Expected Deviation Between Estimated and Actual R $E(\hat{R} - R)$	Average Deviation Between Calculated and True R $(\hat{R} - R)$
.853	.707	5	.0125	.0096
		10	.0063	.0046
		20	.0031	.0051
	.500	5	.0063	.0062
		10	.0031	.0032
		20	.0016	.0013
.778	.707	5	.0250	.0221
		10	.0125	.0099
		20	.0063	.0040
	.500	5	.0125	.0091
		10	.0063	.0061
		20	.0031	.0044
.670	.707	5	.0500	.0380
		10	.0250	.0236
		20	.0125	.0116
	.500	5	.0250	.0215
		10	.0125	.0114
		20	.0063	.0079
.550	.707	5	.1000	.0760
		10	.0500	.0399
		20	.0250	.0228
	.500	5	.0500	.0437
		10	.0250	.0274
		20	.0125	.0122

Table A.16 Outcomes for 2,000 double sampling simulations in which the auxiliary variable (x) was generated by a non-linear regression model [eq. (3.20)]. Measurement errors were normal and additive, and measurement error of the primary variable (y) was 10 percent of that for the parent population. Large sample size (n') was 100. Coefficient of variation for the parent population was 0.500.

ρ for y and x	Small Sample Size (n)	Nonlinear Parameters B	Nonlinear Parameters A	Mean Value from Ratio Estimator ($\bar{y}_R$)	Mean Value from Linear Regression of y on x ($\bar{y}_{lr}$)	Coverage for 95 Level ($\bar{y}_R$)	Coverage for 95 Level ($\bar{y}_{lr}$)	Calculated Ratio	Calculated Reg. Coefficient
.853	10	.25	10	5.03	4.99	.049	.061	.759	.436
			20	4.99	4.96	.066*	.061	.376	.450
			30	5.00	4.95	.062	.062	.251	.376
		.50	10	5.01	5.01	.063	.067*	.584	.337
			20	4.99	5.00	.071*	.065*	.292	.415
			30	4.97	4.95	.076*	.071*	.194	.383
	20	.25	10	5.01	5.00	.055	.052	.758	.441
			20	4.99	4.98	.060	.058	.377	.444
			30	4.99	4.97	.054	.050	.251	.369
		.50	10	4.99	5.00	.056	.059	.583	.338
			20	5.00	5.01	.054	.054	.292	.414
			30	4.98	4.97	.063*	.062	.194	.375
.778	10	.25	10	5.12	5.03	.057	.058	.771	.274
			20	5.00	4.98	.061	.059	.377	.330
			30	4.98	4.96	.065*	.062	.251	.310
		.50	10	5.06	5.01	.053	.059	.593	.193
			20	5.02	5.04	.062	.057	.293	.286
			30	4.99	5.00	.073*	.072*	.194	.292
	20	.25	10	5.03	5.00	.054	.053	.761	.264
			20	5.00	4.99	.054	.054	.378	.330
			30	4.99	4.98	.057	.056	.251	.310
		.50	10	5.02	5.01	.058	.051	.588	.195
			20	4.99	5.00	.062	.057	.291	.278
			30	5.00	5.00	.057	.055	.195	.291
.674	10	.25	10	5.26	4.98	.046	.057	.800	.145
			20	5.03	4.99	.050	.061	.380	.222
			30	5.01	4.99	.066*	.059	.252	.239
		.50	10	5.15	5.01	.057	.065*	.601	.099
			20	5.01	5.01	.059	.060	.293	.174
			30	5.04	5.06	.058	.053	.196	.202

Table A.16, continued

ρ for y and x	Small Sample Size (n)	Nonlinear Parameters B	A	Mean Value from Ratio Estimator ($\bar{y}_R$)	Mean Value from Linear Regression of y on x ($\bar{y}_{lr}$)	Coverage for 95 Level ($\bar{y}_R$)	($\bar{y}_{lr}$)	Calculated Ratio	Calculated Reg. Coefficient
	20	.25	10	5.10	4.99	.054	.058	.773	.143
			20	5.02	5.00	.052	.054	.379	.221
			30	4.99	4.98	.055	.055	.251	.233
		.50	10	5.04	5.00	.056	.051	.590	.105
			20	5.01	5.01	.057	.055	.293	.169
			30	4.99	5.01	.055	.055	.194	.203

* Calculated chi-square exceeds one percent level of significance (6.63).

DISTRIBUTION

No. of
Copies

OFFSITE

U. S. Nuclear Regulatory
Commission
Division of Technical Information
and Document Control
7920 Norfolk Avenue
Bethesda, MD 20014

5 Edward O'Donnell
U. S. Nuclear Regulatory
Commission
Mail Stop 1130 SS
Washington, DC 20555

ONSITE

50 Pacific Northwest Laboratory

K. E. Byers
L. L. Cadwell
D. W. Carlile
D. W. Dragnich
L. L. Eberhardt (8)
R. O. Gilbert
J. M. Hales
D. H. McKenzie
T. L. Page
G. L. Poole
L. A. Prohammer
J. V. Ramsdell
W. H. Rickard
R. G. Riley
R. G. Schreckhise
M. A. Simmons (8)
J. R. Skalski
J. A. Stottlemyre
J. A. Strand
J. M. Thomas (8)
B. E. Vaughan
R. E. Wildung
Publishing Coordination (2)
Technical Information (5)

*U.S. GOVERNMENT PRINTING OFFICE:1985-461-721:20238

NRC FORM 335 (2 84) NRCM 1102, 3201, 3202

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

SEE INSTRUCTIONS ON THE REVERSE

1. REPORT NUMBER (Assigned by TIDC, add Vol. No., if any)
NUREG/CR-4268
PNL-5156

2. TITLE AND SUBTITLE
Ratio Methods for Cost-Effective Field Sampling of Commercial Radioactive Low-Level Wastes

3. LEAVE BLANK

4. DATE REPORT COMPLETED
MONTH: May YEAR: 1985

5. AUTHOR(S)
L. L. Eberhardt
M. A. Simmons
J. M. Thomas

6. DATE REPORT ISSUED
MONTH: July YEAR: 1985

7. PERFORMING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)
Pacific Northwest Laboratory
P.O. Box 999
Richland, Washington 99352

8. PROJECT/TASK/WORK UNIT NUMBER

9. FIN OR GRANT NUMBER
B2461

10. SPONSORING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)
Division of Radiation Programs and Earth Sciences
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555

11a. TYPE OF REPORT
Topical Technical Report

11b. PERIOD COVERED (Inclusive dates)

12. SUPPLEMENTARY NOTES

13. ABSTRACT (200 words or less)

An investigation of cost-effective methods for sampling at commercial radioactive low-level waste sites has been one goal of this project. To that end, double sampling was investigated, and we found that the method appears useful when estimating total radionuclide inventory in waste site environs.

The methods are explained, decision criteria for cost effectiveness presented, and a worked example based on field data is provided. The statistical basis for the conclusion that double sampling appears to be robust and cost-effective is in separate sections. Field tests and additional estimates of "field instrument" errors are needed to substantiate the findings.

14. DOCUMENT ANALYSIS - a. KEYWORDS/DESCRIPTORS
Double sampling
Survey statistics
Cost effectiveness
Low-level waste

b. IDENTIFIERS/OPEN-ENDED TERMS

15. AVAILABILITY STATEMENT
Unlimited

16. SECURITY CLASSIFICATION
(This page) Unclassified
(This report) Unclassified

17. NUMBER OF PAGES

18. PRICE

NUREG/CR-4268

RATIO METHODS FOR COST-EFFECTIVE FIELD SAMPLING OF COMMERCIAL RADIOACTIVE LOW-LEVEL WASTES

JULY 1985