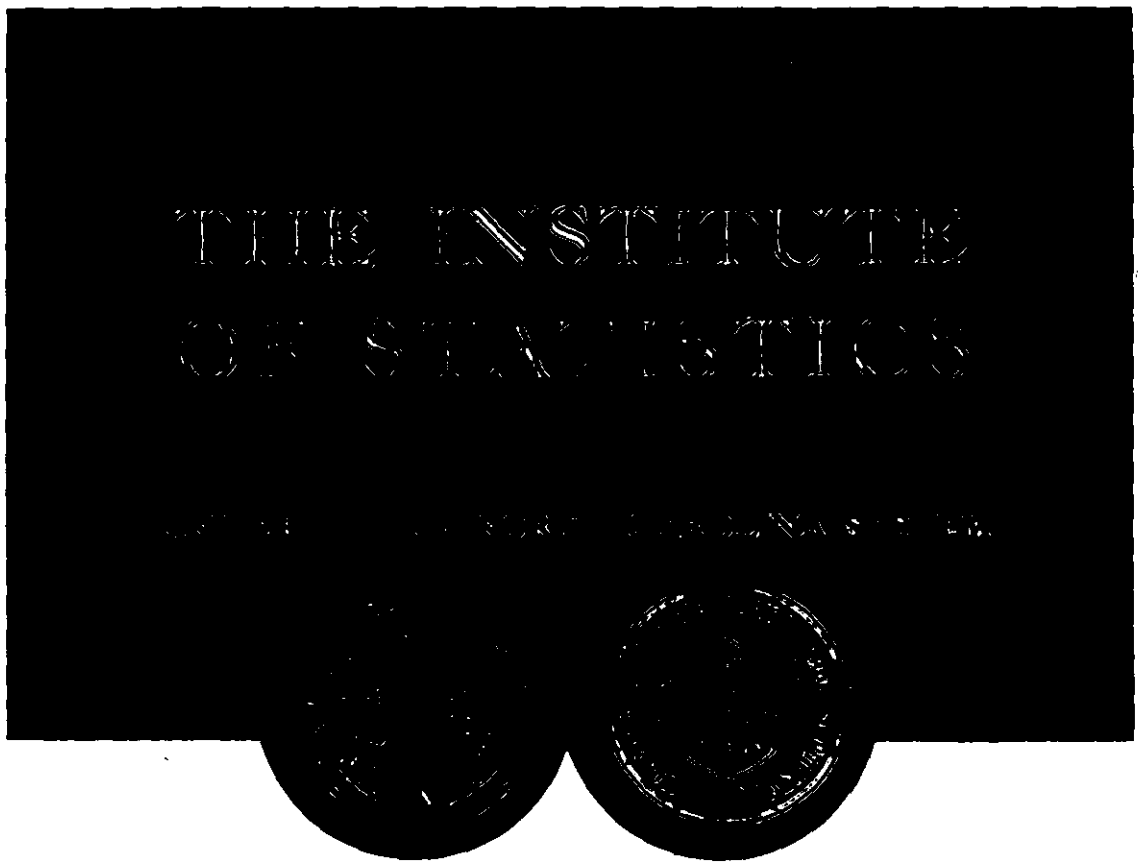


STATISTICAL CONSIDERATIONS IN BULK SAMPLING

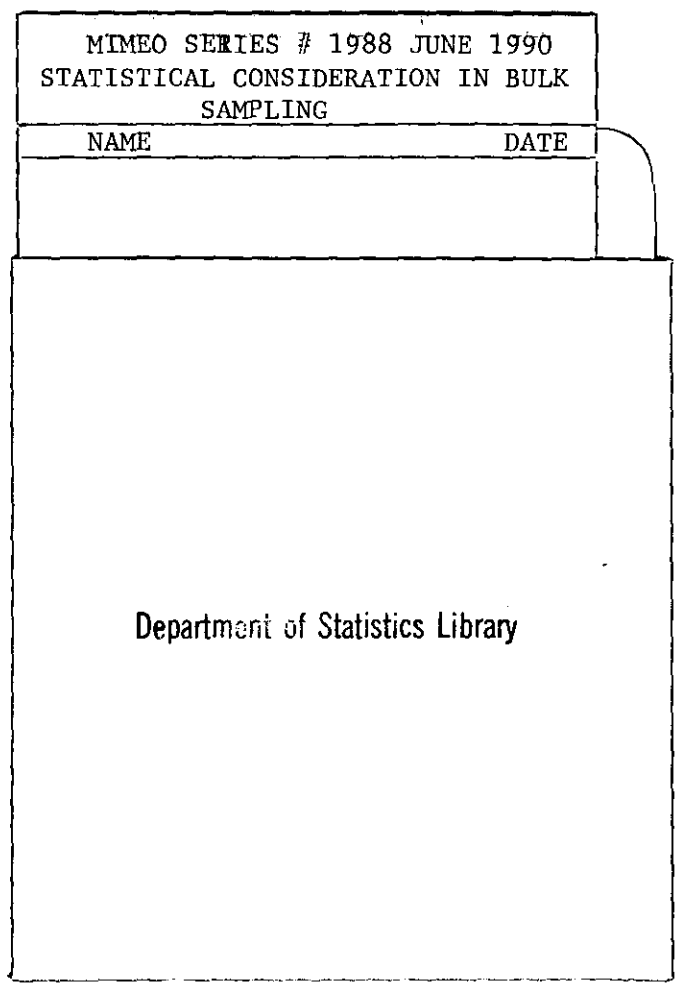
by

Charles H. Proctor

Institute of Statistics Mimeograph Series No. 1988

June 1990

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina


MIMEO SERIES # 1988 JUNE 1990
STATISTICAL CONSIDERATION IN BULK SAMPLING
NAME
DATE
Department of Statistics Library

Statistical Considerations in Bulk Sampling

C. H. Proctor

June 1990, NCSU

1. Introduction

The sampling of a bulk involves four features that, while sometimes present in other types of sampling, make it worthy of special attention by statisticians. These are:

1) Distinctions among (a) the carrier or bulk material itself, (b) the coordinate frame for the bulk, and (c) the sampling frame.

2) Systematic spacing and compositing.

3) Nonpartition extraction of subsamples.

4) Comminution, mixing, segregation and riffling.

Generally there is a bulk that is finite in extent and constitutes the population to be surveyed. For example, the coal on a conveyor belt passing a given place over an eight hour period makes a one dimensional bulk, an automobile exterior body surface is a two dimensional bulk, and a pile of fish meal is a three dimensional bulk. Realizing that the coal on a conveyor belt has a cross-sectional area at any point allows one also to define a three dimensional bulk for this case if it becomes useful to do so. Thus dimension is not an inherent property of a bulk.

A mass of material is described as a bulk when it cannot be or is not conveniently partitioned. For example, an entire book would not be thought of as a bulk of paper since the separate pages form a convenient partition. Even a single page need not be treated as a bulk, if it can easily be cut up into

pieces, each one of the size used for some laboratory determination. The pieces could be numbered and conveniently sampled by conventional methods. As another example, thread wound on a spool need not be considered a bulk since it could be nicely cut into pieces.

Either the page of paper or the spool of thread could become a bulk if it were comminuted (ground, crushed, or otherwise made into very small bits). Thus a common mark of a bulk is its particulate nature, in which the particles are highly variable in size and shape. Liquids, slurries, gases, and gelatinous masses are also bulks by virtue of the difficulty of partitioning them.

When surveying a bulk there is always some characteristic of interest, a constituent or a property, whose average or total is the objective of the survey. It may be the amount of sulfur per pound of coal or the average breaking strength of the fibers.

There is also usually some minimum amount of the material needed for a laboratory determination. This smallest measurable amount of the bulk or carrier material will define an element or elementary cluster. It may be a gram of coal in a railcar or a drop of water in a lake.

In theory it is conceivable that the entire bulk could be partitioned into elements, and the position of (the center of) each element recorded along with the determined level of the variable of interest. These values would then constitute the finite population of interest. In theory one can imagine repeated determinations performed on each element which would show variability around a, so called, true value. The finite population results based on a

single determination are called an equivalent complete coverage, while the complete collection of true values may be said to give the idealized complete coverage.

The coordinate frame used to record the positions of the elements is a theoretical concept. In actual practice a more concretely attainable sampling frame of some sort will be used. A grid of possible locations created by sighting posts placed along the shore line may be used on the surface of a lake. A moving conveyor belt may be stopped at certain times or a beam of light may be turned on at certain times to detect sulfur level in the stack gasses so illuminated.

The sampling frame may either consist of all locations where some material can be drawn out, or a random selection may be made of certain frame locations for drawing the sample. It is very important to distinguish those cases where some randomizing device (coin flip, or random number table) is used from cases where none is used. Actually there are a variety of sample designs (e.g., centered systematic, multi-start systematic or two-per-zone) used in bulk sampling. It seems that nonrandom systematic designs are more often used than probability sample designs.

At each of the selected locations some of the bulk material is drawn out and the sum of this material becomes the primary sample. In the usual case, this primary sampled material is riffled and a secondary subsample is thereby drawn out. A riffle consists of ordering or segregating the original material and then distributing it (partitioning it) over a number of subsamples. There are several machines for doing this for various materials.

Comminutions, whereby the particle sizes are reduced, are oftentimes alternated with riffles. It is frequently noticed that the constituent of interest occurs in some particles more than in others so that reducing the particle size and then mixing (the opposite of segregating) causes the variability in level of constituent from one compact subsample to the next to be made smaller. This allows the same precision of estimation to be attained with smaller sample size.

The riffling after segregating and comminuting with mixing are relatively straightforward operations in so far as the subsampling is done on a partitioning of the primary sample. There can, however, be problems of loss of material as dust or of a change in moisture level due to the handling of the material, and these need to be addressed. The end result is one or more laboratory samples or specimens on which a determination is carried out.

Two aspects of getting the primary sample that need special attention in bulk sampling are: (1) increment extraction and (2) increment spacing. The material drawn at each sampled point is called the increment. Increment extraction refers primarily to the size (the amount extracted) but also to the method of extraction or the shape. This may be done, for example, by dropping a frame on a stopped conveyor belt and scraping away the material thus bounded, or by opening a slot in the floor of the belt or by passing a pail under a falling stream (a cutter), or by digging a 6-inch diameter core into the soil, or by raising an opened bottle from the bottom to the surface of a lake.

The only way to construct increments so that any one of them drawn at random will be unbiased is to partition the bulk, but this, by definition of a

"bulk," is almost never possible and thus there is almost always the possibility for bias due to increment extraction. There would seem to be two ways to discover this bias. The rather costly method is to compare the increment extraction method to a partitioned subsample on similar material. The alternative way is to note the changes in level of the constituent of interest as one reduces the size of the increment. Whatever are the causes of bias they are usually aggravated by smaller increment sizes. Although this second method does not directly estimate bias, it can show that there is (or is not) little change in level so long as increment size is large enough.

The final distinctive feature of bulk sampling is the spacing of the increments. This is usually in a lattice pattern or as a, so called, systematic sample. The purpose of the uniform spacing is to make the estimate more precise. This will occur if the constituent of interest rises and falls over rather large sectors of the bulk. The variable is then said to have positive autocorrelation or positive adjacency correlation, and the uniform spacing will furnish a nice cross section of high, low and middle values of the variable of interest.

2. Survey Frames, Superpopulations and Random Sampling

We now turn to the kinds of statistical theory used to aid in solving the problems of bulk sampling. As has been discussed elsewhere (*Sampling Symposium*) two main aspects of the sampling govern the choice of statistical theory. One is the type of frame and the other is the patterning versus randomness of the sample selection. If there is a frame covering the target population and if random numbers are used to select the sample then much of conventional sampling

theory is applicable, and it would seem that this is the direction in which new advances in bulk sampling should be guided. At present, however, the frames in use are themselves more akin to samples, and if random selection is used it may consist of only a single start number, while the sample remains highly patterned.

There is a much more detailed theory available for one-dimensional frames than for higher dimensional frames, so we will first derive some rather extensive results for the case of sampling a linear frame. We will provide the necessary notation and stochastic process assumptions for the case of the linear frame, but these should be capable of generalization to more dimensions. To keep our discussion in contact with reality a linear frame will be constructed for the fuel stream of a coal-fired power plant indexed by time. Consider two situations: the coal on the conveyor belt heading for the furnace is one case, while the stack gasses on their way out to the atmosphere is the other. In the first case, sampling is done by a cutter that passes across the stream and collects an increment of coal. In the second, a beam of light of a continuous emissions monitor (CEM) illuminates a sector of the stream that is thus captured for analysis.

In both cases the extraction apparatus sets a lower limit on the amount of material extracted and, when this is expressed as a time slice, it defines an element of the bulk frame. The element will be 30 seconds for the cutter, while it will be 1/10 sec. for the light beam of the CEM. The extraction apparatus can be set to remove larger increments either by increasing the width and slowing its passage, in the case of the cutter, or by leaving the light on longer for the CEM. In order to define the survey frame from the bulk frame,

one adopts some particular size of increment. In general each sampling increment will consist of M successive elements from the bulk frame.

In addition to the size of increment there may also be some mechanically set minimum period between increments. That is, the cutter could be used perhaps every five minutes but not more often, while the CEM could be left on essentially all the time or "continuously." These considerations allow the survey frame to be defined over the bulk frame. Let's consider for concreteness an 8 hour period as the population of interest and we will now define the bulk frame and survey frame for the two cases.

For the CEM the 1/10 sec. interval will be taken as the sampling unit as well as the element of the bulk frame. We will denote by T the number of elements on the bulk frame and by N the number of sampling units so that, with M = 1, T = N = 288,000. For the cutter we will consider a 1-minute pass with a minimum frequency of one pass every five minutes. Thus N = 96, T = 960 and M = 2 for the cutter.

During a 1/10 second interval there is a given amount of stack gas passing a point such that 288,000 times this amount gives the total amount exiting the stack during the 8 hour period. Likewise during 30 seconds the amount of coal passing a point on the belt can be multiplied by T = 960 to give the total amount burned in the 8 hour period. In practice, the molecules hit by the CEM's light beam represent only a portion of the stack gas passing the whole cross section of the stack, and the coal falling into the cutter is only a part of that passing a point on the belt.

Thus the level of the variable of interest takes on two values for every element of the bulk frame: one is the value exhaustively determined for the total bulk material passing and the other is the value exhaustively determined for just the material in the element trapped or exposed. Let us use ξ_i for the first and η_i for the second quantity, where i = 1, 2, ..., T. It may happen that all material is exposed and $\xi_i = \eta_i$. The population quantities of interest will be denoted $\bar{\xi} = \Sigma\, \xi_i/T$ and $\bar{\eta} = \Sigma\, \eta_i/T$. Separate studies will be required to determine how $\bar{\xi}$, the total population mean, differs from $\bar{\eta}$ the measured population mean, but for present purposes we will be modestly concerned only with estimating $\bar{\eta}$.

For the CEM the recorded values that could be obtained by an actual survey run over all N = 288,000 times and will be denoted Y_i for i = 1, 2, ..., 288,000. These recorded values may suffer measurement inaccuracies. That is, Y_i may equal η_i but it generally does not. We will let Y_i' represent the level of the variable of interest when determined without measurement error. In the case of the CEM example $Y_i' = \eta_i$.

For the cutter example, the first 10 values of η_i cover the first 5 minute interval. If the sampling frame is centered then $Y_1' = (\eta_5 + \eta_6)/2$. This is a true value. In practice the material extracted by the cutter in the 1-minute pass will be subsampled and analyzed to give the determination as Y_i.

If we let ϵ_i represent measurement error then $Y_i = Y_i' + \epsilon_i$. This notation for the measurement error process will be used generally. For our example of the cutter the one minute pass may collect 100 lbs. of coal in which the level of the variable of interest may be denoted Y_i'. After crushing and

segregating and then riffling, followed by further crushing and mixing and finally by laboratory analysis of a 1 gm. laboratory sample, we obtain the value Y_i. The difference $Y_i - Y_i'$ defines ϵ_i. Notice that in this illustration ϵ_i includes subsampling or reduction error as well as analytical or laboratory errors. It is not uncommon to suppose the ϵ_i are independent from one increment to another and even to know σ_ϵ^2 the measurement error variance. If so then one can discuss the difference between $\bar{Y}$ and $\bar{Y}'$.

In our example of the cutter there are 96 values in both $\bar{Y}$ and $\bar{Y}'$, as in general there will be N values in both $\bar{Y}$ and $\bar{Y}'$. If measurement errors are independent then the average distance between $\bar{Y}$ and $\bar{Y}'$ is $\sigma_\epsilon/\sqrt{N}$. Thus when N is large enough, an estimate of $\bar{Y}$ is essentially the same as an estimate of $\bar{Y}'$, and one may forget about measurement error. Then random sampling methods can be used to estimate $\bar{Y}$ and also can be used to estimate standard errors based on the observations themselves.

A simple random sample, for example, can be selected in n draws with each of the N survey frame units equally likely to be picked at each draw. If any unit is drawn a second time this is ignored and the draw is done again until the resulting sample has n distinct units. The observed values are denoted y_1, y_2, ..., y_n in order of their being drawn. These are random variables whose distributions are created by the random numbers or by coin flips used to make the selections.

Notice that, for example, y_1 has the same distribution as y_2 but they are slightly negatively correlated. For a given 8 hour period there are fixed

values of the Y_i and one can in theory calculate the finite population variance:

$$S^2 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2/(N-1) .$$

This governs the sampling variance of the sample mean $\bar{y}_{sr} = \sum_{i=1}^{n} y_i/n$, in that:

$$V(\bar{y}_{sr}) = S^2(n^{-1} - N^{-1}) .$$

This sampling variance can be estimated from the observations by replacing S^2 by

$$s^2 = \Sigma(y_i - \bar{y})^2/(n-1) ,$$

to get:

$$v(\bar{y}_{sr}) = s^2(n^{-1} - N^{-1}) .$$

In the CEM example it would not be reasonable to suppose the string of 288,000 ϵ_i's to be independent. There would likely be some autocorrelation of measurement errors running along until the monitor was recalibrated or replaced. The average distance between $\bar{Y}$ and $\bar{Y}'$ when there is correlated measurement errors is $\sigma_\epsilon(1 + \bar{\rho})/\sqrt{N}$ where $\bar{\rho}$ is an average correlation among the measurement errors. This average correlation will be found to decrease with N and since $\sqrt{N}$ is also a divisor it turns out that $\bar{Y}$ and $\bar{Y}'$ will be virtually identical if N is large enough. Again, if N is large enough and $\bar{Y}$ equals $\bar{Y}'$, then conventional sampling theory will be useful in characterizing $\bar{y}$ as an estimate of $\bar{Y}$.

In practice probability samples are not taken, as they could be by using random numbers, and so we will now indicate how statistical analysis can be applied in finding sampling variances for the methods actually used. We will particularize to a sample of size n = 4 -- again to maintain closer contact with reality.

Rather than select four times (to make a pass with the cutter) at random from the 96 possible, it is more likely that the cutter would be turned on at four evenly spaced times in the range i=1 to i=96. In particular, the observations could be made at i = 13, i = 37, i = 61 and i = 85. Again in practice the coal would be composited from these four passes and one measurement operation of preparation and analysis would be carried out. We can thus express this estimator, as a statistic, as $\hat{\hat{Y}} = (Y'_{13} + Y'_{37} + Y'_{61} + Y'_{85})/4 + \epsilon_i$. The four locations are a systematic sample of the 96 and also of the 960, while the 96 are themselves a systematic sample of the 960. What is needed from statistical theory is a variance expression for a systematic sample estimator such as $\bar{y}_{sy} = (Y'_{13} + Y'_{37} + Y'_{61} + Y'_{85})/4$. Then we can just add σ^2_ϵ, assuming we know it, to get the variance of the estimator, $\hat{\hat{Y}}_{sy}$.

The variance expression for $\bar{y}_{sy}$ is derived from knowledge of how the η_i values were generated. In particular the η_i values may all be taken to have a common mean, a common variance and some pattern of correlations that depends only on distances. That is,

$$E(\eta_i) = \mu \text{ for all } i,$$
$$V(\eta_i) = \sigma^2 \text{ for all } i \text{ and}$$
$$\text{Cov}(\eta_i, \eta_j) = \sigma^2 \rho_d \text{ where } d = |i-j| ,$$

so ρ_d is the correlation at distance d apart. A convenient and fairly realistic functional form for ρ_d is:

$$\rho_d = \sum_{i=1}^{k} \pi_i \rho_i^d , \text{ where } \sum_{i=1}^{k} \pi_i = 1.$$

If $k = 1 = \pi_1$ then the process is a first order autoregressive or Markov process, otherwise it is a weighted sum of such processes.

The quantity $\bar{y}_{sy}$ is a linear function of the η_i, while the objective of the survey is to estimate the population quantity $\bar{\eta}$ which is also a linear combination of the η_i. In the cutter example the value Y'_{13} is the average of η_{125} and η_{126}, the value $Y'_{37} = (\eta_{365} + \eta_{366})/2$, and so forth. Thus:

$$\bar{y}_{sy} = (\eta_{125} + \eta_{126} + \eta_{365} + \eta_{366} + \eta_{605} + \eta_{606} + \eta_{845} + \eta_{846})/8$$

The difference $(\bar{y}_{sy} - \bar{\eta})$ is a linear combination of the 960 values of the η_i and can be written

$$(\bar{y}_{sy} - \bar{\eta}) = \underline{a}' \underline{\eta} ,$$

where, for example, $a_1 = -1/960$ and $a_{125} = (1/8 - 1/960)$ are the first and 125th components of $\underline{a}$.

Under the assumptions, the variance-covariance matrix of the 960 η_i's has diagonal entries of σ^2 and dth off-diagonal of $\sigma^2\rho_d$. The matrix as a whole can be written Σ_η. Thus

$$V(\bar{y}_{sy}) = E(\bar{y}_{sy} - \bar{\eta})^2 = \underset{\sim}{a}' \, \Sigma_\eta \, \underset{\sim}{a} \, .$$

This expression gives the variance of a centered systematic sample expected under the superpopulation process.

3. Numerical Illustrations Comparing Random to Patterned Sampling

It may be helpful in fixing ideas to use the above variance expressions in discussing the relative variances of a simple random sample design compared to a centered systematic design for the example of the cutter frame with N = 96 and T = 960. We will suppose the η_i are governed by $\mu = 25$, $\sigma^2 = 4$ and

$$\rho_d = .9 \times .9^d + .1 \times .999^d \, ,$$

or $\pi_1 = .9$, $\pi_2 = .1$, $\rho_1 = .9$ and $\rho_2 = .999$. Here the distances are measured on the bulk frame over the range i=1 to i=960.

On the survey frame the true values are given by $Y'_1 = (\eta_5 + \eta_6)/2$, $Y'_2 = (\eta_{15} + \eta_{16})/2, \cdots, Y'_{96} = (\eta_{955} + \eta_{956})/2$; while the finite population values are $Y_i = Y'_i + \epsilon_i$. We will further suppose that the ϵ_i have zero mean (no bias) and variance .1. Notice that the process variances of the Y'_i are:

$$V(Y'_i) = 2\sigma^2[1 + (.9 \times .9 + .1 \times .999)]/4$$

$$= .95495 \, \sigma^2 = 3.81980 \, .$$

This value, while less than $\sigma^2 = 4$, is nowhere near as small as $\sigma^2/2 = 2$, the variance of the mean of two uncorrelated random values -- the two adjacent values of η_i are highly correlated.

The process covariance between Y_1' and Y_2' can be found as:

$$E(Y_1' - 25)(Y_2' - 25) = E(\eta_5 - 25 + \eta_6 - 25)(\eta_{15} - 25 + \eta_{16} - 25)/4$$

$$= \sigma^2(2\rho_{10} + \rho_9 + \rho_{11})/4$$

$$= .413686804\ \sigma^2 .$$

Similar computations can be used to find all process covariances among the Y_i'. An approximate covariance is $\rho_{10}\sigma^2 = .412815084\ \sigma^2$, and similar approximations can be used to get the other covariances among Y_i'. Let us denote the correlations among the Y_i' as ρ_d'. For example, $\rho_1' \doteq \rho_{10}$.

The variance of the centered systematic sample mean $\bar{y}_{sy}$ is now found as:

$$\underset{\sim}{a}' \Sigma_\eta \underset{\sim}{a} = .196751\ \sigma^2$$

$$= .787004 .$$

This "variance" represents the average squared distance between $\bar{y}_{sy}$ and $\bar{Y}'$ over many realizations of the process -- that is, over many, many possible finite populations. When we suppose $\sigma_\epsilon^2 = .1$ the variance of the estimator $\hat{\bar{Y}}$ is .887004.

Any one finite population would have a particular value of S^2 and thus a particular value of $V(\bar{y}_{sr})$. It may be of interest to calculate the process average value of $V(\bar{y}_{sr})$ to compare to $V(\bar{y}_{sy}) + \sigma_\epsilon^2 = .787004 + .1 = .88700$.

To begin we will notice that S^2 derives from both true values and from measurement errors as:

$$S^2 = \Sigma(Y_i - \bar{Y})^2/(N-1)$$

$$= \Sigma(Y'_i - \bar{Y}' + \epsilon_i - \bar{\epsilon})^2/(N-1)$$

$$= \Sigma(Y'_i - \bar{Y}')^2/(N-1) + (\text{cross products}) + \Sigma(\epsilon_i - \bar{\epsilon})^2/(N-1) \ .$$

We can now average over the random generation of the Y'_i's (from the η_i process) and over the measurement operations. The first sum is rewritten as:

$$S'^2 = \sum_{i=1}^{N} (Y'_i - \bar{Y}')^2/(N-1) = \sum_{i<j=1}^{N\,N} (Y'_i - Y'_j)^2/N(N-1)$$

so we can use $E(Y'_i - Y'_j)^2 = 2\sigma^2(1 - \rho_{|i-j|})$. This shows that

$$E(S'^2) = \sigma'^2(1 - \bar{\rho}') \ ,$$

where $\bar{\rho}'$ is the average of the off-diagonal entries in the matrix of process correlations among the 96 Y'_i, and σ'^2 is the variance of a Y'_i. The average over the cross product term is zero because the two processes are assumed to be independent. The average of the last term is σ^2_ϵ itself.

In order to calculate $\bar{\rho}'$ we use the approximation giving $\rho'_d = \pi_1\rho_1^{10d} + \pi_2\rho_2^{10d}$, and find $\bar{\rho}' = .084117$, while the variance of the Y'_i is, as found above, $\sigma'^2 = 3.8198$. It turns out that $E(S'^2) = 3.49849$ and thus $E(S^2) = 3.59849$. This shows that $V(\bar{Y}_{sr}) = 3.59849(4^{-1} - 96^{-1}) = .86214$. Thus, for our illustration, the process expected variance of a simple random sample of four increments

with laboratory determination on all four increments is less than the variance of a centered systematic sample with compositing and one laboratory determination.

Our illustration is probably most unrealistic in the large relative size of measurement error. For the true values, the centered systematic sample variance is .787004 as compared to the value of $E(S'^2)(4^{-1} - 96^{-1}) = .83818$ for the simple random sample, and if only one composite had been analyzed from the random sample its variance would have been .93818 as compared to the .887004 of the systematic samples with only one determination. This comparison is more honest since the cost of these two methods will generally be more nearly equal.

Of course, the advantage of doing all four laboratory analyses with the random sample is that $V(\bar{y}_{sr})$ can be calculated to give information on sampling variability as well as to furnish the estimate. This points up the very large conceptual distinction between having both $\bar{y}_{sr}$ and $v(\bar{y}_{sr})$ on one hand and having just $\hat{\bar{Y}}_{sy}$ on the other. Along with $\hat{\bar{Y}}_{sy}$ one may have a process expected value for $V(\hat{\bar{Y}}_{sy})$ but there is always some doubt as to how correct are the assumed values of σ^2 and ρ_d for the current material.

4. Sample Preparation

A number of approaches have been taken to deal with the particulate nature of a bulk and we will outline one such. The approach we have chosen uses conventional analysis of variance on a hierarchical subdivision of the bulk to estimate a quantity called Smith's b which is then tied to the various operations of comminuting, mixing, segregating and riffling.

Let's consider as examples of the population bulk the coal in a primary sample, the fertilizer in a bag, the shelled corn in a truckload or some

similar undifferentiated three-dimensional particulate. All particles, above the size of fines, are indexed from i = 1 to i = P. Let W_i be the weight and X_i be the amount (by weight) of the constituent of interest in the i<u>th</u> particle. Also, let $\underset{\sim}{x}_i$ be the three-dimensional coordinates of the i<u>th</u> particle, although we have no immediate need of such a refinement. Additionally, one may record properties such as specific gravity, hardness, "sphericity," etc. in a vector $\underset{\sim}{Z}_i$, but as yet we have not been using this information.

By combining adjacent particles one forms elementary clusters. We suppose all elementary clusters weigh w' and that there are T of them, thus $\Sigma W_i/w' = T$. Next, one forms clusters by combining a number, M, of elementary clusters. The bulk sample is drawn by randomly selecting a number of these clusters. The way one puts together the elementary clusters into a cluster is determined by what mixing, segregating and riffling is used.

Let's begin by considering the case of no mixing, segregating nor riffling. The bulk has been divided into clusters, with elementary clusters nested in clusters and particles in elementary clusters. This is three-stage nesting and we will adopt the conventional sampling notation of N_1 clusters in the bulk, N_2 elementary clusters in each cluster and N_3 particles in each elementary cluster. Notice that $N_3 = P/T$ and $M = N_2$. Of course, in reality not all elementary clusters may have exactly N_3 particles but this will not upset the results very much.

This balanced and nested arrangement of units will be indexed by i_1, i_2 and i_3 where $i_1 = 1, 2, \ldots, N_1$, $i_2 = 1, 2, \ldots, N_2$ and $i_3 = 1, 2, \ldots, N_3$. Thus $X_{i_1 i_2 i_3}$ is the amount of the constituent of interest in the i_3<u>th</u> particle, of

the i_2th elementary cluster in the i_1th cluster. Let's now define $Y_{i_1 i_2 i_3} = X_{i_1 i_2 i_3}/W_{i_1 i_2 i_3}$. We will argue that $\bar{Y} = \Sigma\Sigma\Sigma\, Y_{i_1 i_2 i_3}/N_1 N_2 N_3$ may be taken as a legitimate goal of our sample survey, from the standpoint of the study of variances.

In fact there is a numerical difference between

$$\bar{Y}_{i_1 i_2} = \sum_{i_3=1}^{N_3} Y_{i_1 i_2 i_3}/N_3$$

and $\sum_{i_3=1}^{N_3} X_{i_1 i_2 i_3} / \sum_{i_3=1}^{N_3} W_{i_1 i_2 i_3}$, so that there will be a difference between $\bar{X}/\bar{W}$ and $\bar{Y}$. However, this is a constant or bias and does not invalidate the close correspondence between variances calculated from Y-values and those encountered in estimates made by dividing sums of X-values by sums of W-values. The use of a single series of Y-values is obviously a much more convenient task than attempting to work out the variances of ratios of X to W values.

As a sampling method we first consider a simple random sample of n_1 clusters from the N_1. The estimator would be written $\bar{y} = (y_1 + y_2 + \ldots + \bar{y}_{n_1})$ with sampling variance:

$$V(\bar{y}) = S_{c_1}^2 (n_1^{-1} - N_1^{-1}) \ ;$$

where $S_{c1}^2 = \Sigma(\bar{Y}_{i_1} - \bar{Y})^2/(N_1 - 1)$ is first stage cluster variance. Remember that S_{c1}^2 is based on a cluster of size M. That is, if M were increased (and N_1 thus will decrease) then S_{c1}^2 would change -- it would generally become smaller.

If the M elementary clusters going into every cluster were selected at random then one would expect that S^2_{c1} would nearly equal $S^2_{c2}M^{-1}$ where S^2_{c2} is the variance among elementary clusters. In fact one often finds $S^2_{c1} = S^2_{c2}M^{-b}$, where b is a number smaller than 1. This is because the elementary clusters going into a cluster are more alike than would be a randomly selected set of M. This will be true when particles are naturally segregated or when they have been comminuted without mixing. The relationship $S^2_{c1} = S^2_{c2}M^{-b}$ is called Smith's empirical law and is just that -- an empirical generalization that needs to be checked in each setting.

It may be useful to describe briefly how the value of b is estimated. One draws a sample of clusters say 20 of them of the maximum size of interest. Each of these large clusters is subdivided into 2 clusters half the size of the largest ones, and these are further divided into 2 until the elementary cluster size is reached. For example, with 20 large clusters and 40 of the next size and 80 of the next size and 160 elementary clusters one is left to do 160 laboratory analyses. There are ways of dropping certain of the 160 and still being able to estimate the components of variance -- one uses a, so called, staggered design. From the nested ANOVA on these determinations one can estimate Smith's b.

For naturally occurring material or material after comminution b would range from .5 to .7, reflecting some adjacency correlation. Knowing b permits us to find an optimum value for M in some cases. Consider using variable sizes of probes to sample bags of fertilizer or truckloads of shelled corn. From

studies of enumerators in action we may be able to express the time costs of using the differing sizes of probes as:

$$C_T = C_1 n_1 + C_2 n_1 M$$

where C_1 is the time taken to locate and extract a probe while C_2 is the per-elementary-cluster cost of handling the material once it is removed.

Recall that the variance is $S_{c1}^2 n_1^{-1}$ where we ignore the term in N_1^{-1} which will be, generally, very small even for large M. Also, by Smith's empirical law the variance is $S_{c2}^2 M^{-b} n_1^{-1}$. Substituting for n_1 from the cost function and finding the minimum variance leads to:

$$M_{opt} = bC_1/(1-b)C_2 \ .$$

Notice how this result is free from the choice of size of elementary cluster.

Also, notice that if b can be made to approach 1 then M_{opt} becomes very large. That is to say, one needs only one cluster. It is precisely the objective of segregating followed by riffling to increase b close to 1, thus insuring that a single cluster even of small size will be enough. However, one comes up against the fact that the minimum size of elementary cluster must be larger than the largest particle size. This leads one to comminute and thus reduce elementary cluster size. Let's consider these effects separately.

A riffle operates essentially by lining up the elementary clusters in serpentine order and distributing them in turn over the subsamples or parts which now become the clusters. If the material can be caused to segregate by vibrating or tumbling or allowing it to settle then the serpentine order may

show considerable adjacency correlation. Since the riffled portions resemble systematic samples, they will have a reduced cluster-to-cluster variance. The Smith's b may even be found to exceed 1.

It can be seen that sampling variance will decrease as b is increased and as M_{opt} is adjusted to whatever b is in force. In the neighborhood of b = .8 an increase to b = .9 can cause a 16% reduction in variance but it would likely be less since it would not be expected that the optimum M would actually be attained.

The effect of comminution is, quite naturally, difficult to predict theoretically and requires that experiments be conducted. If the largest particle size after comminution is a proportion, say π of the largest particle size we will say that the bulk after comminution is a π^{-1}-fold comminution of the original bulk. After comminution the elementary cluster size becomes $\pi w'$ and there are T/π of them, etc. There is still no way to predict what will turn out to be the new variance among the mean Y-values from elementary cluster to elementary cluster. If a Smith's b value can reasonably be adduced, then the variance of the new and smaller elementary clusters will be, say, S^{*2}_{c2} where $S^{*2}_{c2} = S^{2}_{c2}\pi^{-b}$. Notice that since $\pi < 1$ this new cluster variance will increase. In order to estimate the new elementary cluster variance and the Smith's b, it will be necessary to conduct an experiment on nested clusters just as was described for the original bulk.

5. A Review of Some Variance Expressions in Bulk Sampling

It has been of interest to compare variance expressions appearing in the bulk sampling literature with those in the general sampling literature. In some cases essential novelty appears in the bulk sampling field and in other cases the notation is different but the formula is familiar. A case of the latter is Gy's (1982) formula for $\sigma^2(FE)$, the variance of what he calls fundamental error.

A comparable formula in the general sampling literature is the variance expression for an estimated ratio. The formula that Cochran (1952) gives for the variance of a ratio is:

$$V(\hat{R}) = \frac{1-f}{n\,\bar{X}^2} \sum_{i=1}^{N} \frac{(Y_i - RX_i)^2}{(N-1)}$$

where: X_i = weight of i<u>th</u> particle, Y_i = weight of ash in i<u>th</u> particle, i=1,2,..., N indexes the N particles in the population. $\bar{X} = \Sigma X_i/N$, $R = \Sigma Y_i/\Sigma X_i$ n is the number of particles in the sample, f = n/N and $\hat{R} = \bar{y}/\bar{x}$ is the sample ratio.

The Gy formula is:

$$\sigma^2(FE) = \frac{1-P}{Pa_L^2 M_L^2} \sum_{\alpha} \sum_{\beta} (a_{\alpha\beta} - a_L)^2 M_{\alpha\beta} M_{F\alpha\beta}$$

where α indexes particle diameter classes and β indexes specific gravity classes, $a_{\alpha\beta}$ is the ratio of ash to total weight among particles in the $(\alpha\beta)$<u>th</u> cell, a_L is the population proportion ash, $M_{\alpha\beta}$ is total weight of particles in the $(\alpha\beta)$<u>th</u> cell and $M_{F\alpha\beta}$ is the weight per particle in the $(\alpha\beta)$<u>th</u> cell, M_L is the weight of the population and P is the ratio of sample weight to population

weight. The Gy formula is a rel-variance, which means that $\sigma^2(FE)$ corresponds to $V(\hat{R})/R^2$ in Cochran's notation. It differs from Cochran's by collapsing all particles into cells of a size by specific gravity cross-classification. All particles within a cell are treated as identical in ash content. Thus the Gy formula has an additional approximation step beyond the Cochran one. Since collapsing will reduce the variance, the Gy expression is understating the variance relative to the formula based on individual particles.

To see this, let's expand the total sum of squares in the Cochran formula into a between sum of squares plus a within sum of squares, that is TSS = BSS + WSS. Let $N_{\alpha\beta} = M_{\alpha\beta}/M_{F\alpha\beta}$ be the number of particles in the $(\alpha\beta)\underline{th}$ cell.

$$
\begin{aligned}
TSS &= \sum_{i=1}^{N} (Y_i - RX_i)^2 \\
&= \sum_{\alpha}\sum_{\beta}\sum_{i\in(\alpha,\beta)}\left[(Y_i - RX_i) - (\bar{Y}_{\alpha\beta} - R\bar{X}_{\alpha\beta}) + (\bar{Y}_{\alpha\beta} - R\bar{X}_{\alpha\beta})\right]^2 \\
&= \sum\sum N_{\alpha\beta}(\bar{Y}_{\alpha\beta} - R\bar{X}_{\alpha\beta})^2 + \sum\sum\sum\left[Y_i - \bar{Y}_{\alpha\beta} - R(X_i - \bar{X}_{\alpha\beta})\right]^2 \\
&= \sum_{\alpha}\sum_{\beta}(M_{\alpha\beta}/M_{F\alpha\beta})(M_{F\alpha\beta}a_{\alpha\beta} - a_L M_{F\alpha\beta})^2 + WSS \\
&= \sum_{\alpha}\sum_{\beta}(a_{\alpha\beta} - a_L)^2 M_{\alpha\beta}M_{F\alpha\beta} + WSS \\
&\equiv \eta^2 TSS + (1 - \eta^2)\ TSS\ .
\end{aligned}
$$

Here we have used η to represent the correlation ratio (Yule and Kendall, 1968, p. 256) of ash weight on total weight over the cells. This correlation ratio would be expected to be fairly large (well above .5) when the classes involve specific gravity. That is, rock with lots of ash has higher specific gravity than particles with more carbon. Other constituents such as sulfur

may not show such large values of η and the variance may then be more seriously underestimated by $\sigma^2(FE)$.

It can be seen from the above formulas that setting $M_L = N\bar{X}$, $P = n/N$, and $a_L = R$ makes $\sigma^2(FE) = V(\hat{R})/R^2$. Gy's derivation was based on assumptions of binomial sampling in which each and every particle of the N was assumed to have been separately subject to a random trial to decide its inclusion or not in the sample. Such a supposition is unrealistic, but does allow for representing the action of increment extraction bias as inequalities among particle inclusion probabilities. However, after making the usual approximations found in theoretical treatments of ratio estimation, the Gy result became the usual variance expression.

Bilonick (1989) gives data on 6 volume classes by 5 specific gravity classes for $a_{\alpha\beta}$ (his Table 1, when divided by 100), $M_{\alpha\beta}$ (his Table 2) and $M_{F\alpha\beta}$ (his Table 4). We took the volume per particle in the smallest size class to be .005 cm^3 and multiplied it by the densities to give values for $M_{F\alpha\beta}$ in his Table 4. Dividing Table 2 entries by Table 4 entries will furnish numbers of particles in each cell. The total number of particles in the data is then found to be N = 30, 550, 396. The variance of deviations of ash weight is found to be .00114456. Dividing this by $\bar{X}^2 = (.026638)^2$ gives 1.613 so that

$$V(\hat{R}) = \left(\frac{1}{n} - \frac{1}{N} \right) 1.613$$

For these data R appears to be .0458 (i.e., $R = \bar{Y}/\bar{X} = .001220/.02664$) and so a population coefficient of variation would be found as $\sqrt{1.613/.0458} = 27.7$. In order to achieve a sampling CV of 5% one needs a sample size of $n = (27.7/.05)^2 = 306{,}916$ particles or a sample size weighing 306,916 x .026638 = 8.2 kilograms.

In his article, Bilonick suggests taking a sample of size 1000 lbs. or of n = 17,028,000 particles which would have a sample CV of .0067. This agrees with his result $\sigma^2(FE) = 4.5018 \times 10^{-5}$ in that $.0067^2 = 4.5 \times 10^{-5}$, that is, a rel-variance is a squared coefficient of variation.

Let's now pursue some novelties in the bulk sampling literature and, in particular , let's consider the issue of "convenience" of the formula. In sampling theory when a sample size is required it is often convenient to use the sample coefficient of variation in setting this. We gave an illustration of the usual reasoning when we found that n = 306,916 particles would give a 5% sampling CV. It is, however, more convenient to learn that a sample of "8.2 kilograms" is needed, than to hear one needs one of "306,916 particles."

The formula for $V(\hat{R})$ can be rewritten for convenience by omitting the fpc and by replacing $n\bar{X}$ by w say where w is sample weight. Thus,

$$V(\hat{R}) = w^{-1}\left[\sum_{i=1}^{N}(Y_i - RX_i)^2/X\right],$$

where $X = N\bar{X}$ is the total weight of the coal in the experiment. As a formula for a rel-variance we find:

$$V(\hat{R})/R^2 = w^{-1}\left[\Sigma(Y_i - RX_i)^2/R^2X\right]$$

or

$$\text{Sample CV} = \left[\Sigma(Y_i - RX_i)^2/R^2X\right]^{1/2}/\sqrt{w}$$

$$= 4.526/\sqrt{w},$$

for the present problem with w in grams and CV as a ratio (not a percentage). With sample CV in percent and w in kilograms, we can use $14.31/\sqrt{w}$. Notice that to bring this to 5% we need w = 8.19 kilograms as found before.

By way of comparison we will now recalculate the "true unit increment variance" from data furnished by Bertholf and Webb (1954, in their Table X, p. 109). They furnish a cross-tabulation of 5 size classes by 6 specific gravity classes from which we can calculate:

$$V(\hat{R}) = \left(\frac{1}{w} \right)(.0392) ,$$

where w is sample size in pounds. The ratio found in the experimental data was R = .1007 or 10.07% ash in their coal. When the formula is brought to percent sampling CV in terms of kilograms of sample we find

$$\% \text{ Sample CV} = 13.24/\sqrt{w} .$$

Comparing 13.24 to 14.31 one must marvel at the similarity in variabilities of ash over distinct coals. The quantity .0392 in $V(\hat{R})$ is for the variance of proportion ash and it would be 3.92 for percent ash. This is exactly the value given as "true unit increment variance" which Bertholf and Webb calculated using the Kassel-Guy (1935) formula. It turns out that, aside from the finite population correction factor of (1-f), the Kassel-Guy formula is Gy's formula and is thus the grouped version of the sampling formula for the variance of a sample ratio.

Actually the article by L. S. Kassel and T. W. Guy appeared in 1935 so their formulation has priority among the three mentioned thus far. The formula for the variance of a ratio is undoubtedly quite old, but I'm not sure when it

originated -- in the least square literature of the 20's is my conjecture, although Gauss may very well have used it.

A different material was studied by Gy to illustrate the calculation of $\sigma^2(FE)$. It was a magnetite ore crushed to 20 mm. and the iron content was given in percent for the average particle in 7 size classes by 5 specific gravity classes (3 cells were empty). The total weight was 100 kg. We used SAS procedures to do the calculations as follows:

1) Enter the data by cells with X = weight of average particle, Y = weight of constituent of interest and N = No. of particles in cell. A regression weight is calculated as W = N/X for each cell.

2) Run the regression of Y on X without an intercept with regression weight W. The ratio of Y to X should appear as the estimated regression coefficient. Save the regression residuals on a data set.

3) Square the regression residuals, multiply by N, and add to get $\Sigma(Y_i - \hat{R}X_i)^2$. Divide by the total weight to get $\Sigma(Y_i - \hat{R}X_i)^2/\bar{X}N$. This can be divided by $n\bar{X}$ = sample weight to give the forecasted sampling variance.

For Gy's data in his Table 22.3 one finds $\Sigma(Y_i - \hat{R}X_i)^2/\bar{X}N = .01635$. Expressed as a rel-variance this becomes $.01635/(.5517)^2 = .054$ which Gy writes as "Z." Expressed as a sampling CV for a 1 kg. sample we find $\sqrt{.054/1000} = .0073$ and comparing it to the above values for coals of 13.24 and 14.31 the distinctive feature of this finely crushed and relatively homogeneous material becomes evident. To attain a 5% sampling CV a sample weight of 22 grams would appear to be adequate.

Gy also uses data on individual particles to get $\sigma^2(FE)$ and it may help to review this approach by way of understanding some of the more novel aspects of the bulk sampling literature. The data were obtained on "... a working batch of 46 fragments extracted one by one at random among the coarsest fragments of the material." (p. 276) Notice that the words "at random" do not here suggest any reference to a table of random numbers, but to behavior supposed to be haphazard.

The calculations now proceed by using $\Sigma_{i=1}^{46}(Y_i - RX_i)^2$, since the data are not grouped into cells. Gy introduces a division by the volumes of the individual particles and a multiplication by the average volume which in effect cancel. This he does to avoid having to guess at a "shape factor," although there is in this case no need to consider volumes at all. The rel-variance is found to be:

$$V(\hat{R})/R^2 = \frac{1}{w}\,(19619)/(.3791^2 \times 5371) = 25.4/w\ .$$

However, this sample was drawn "among the coarsest fragments," and this translates to the largest 5% of the particles. The variance for this subpopulation has been found, by Gy, to give a rel-variance about 4 times larger than that from a sample from all particle sizes. Thus the 25.4 needs to be multiplied by .25, -- the value g = .25 is Gy's "size range factor."

When I calculated the rel-variance on the largest size class for the Bertholf-Webb coal data and compared it to the rel-variance for the whole range of classes, the ratio was .20, so we see that .25 may not be unreasonable.

An additional way that Gy illustrates to calculate a rel-variance is to guess a constant C and multiply it by the average diameter of the top size

particle. The appropriate multiplier is based on judgements for four factors: g = .25 as above, f a shape factor reflecting sphericity of the particles, and the product of a c by an ℓ reflecting judged variability in content from particle to particle.

This procedure of judging various constants and plugging them into a formula to find a rel-variance may seem to some a bit mysterious. It is somewhat novel to classical sampling, but not wholly so. For example, Cochran (1977) gives a number of ways to judge the value of S^2, the population variance, in calculating required sample size, and among these one finds: "... it was decided to assume $S^2 = 1.2\bar{Y}$, the factor 1.2 being an arbitrary safety factor." Cochran's choice of words may disparage the method more than Gy's, but the approach is the same. Unfortunately, the method is subjective and it depends on the particular experience of particular workers to arrive at numerical values.

Having considered increment variance, or the small scale fluctuations, let's now look at the variance deriving from increment spacing or from long range fluctuations. The major styles in bulk sampling have been the variance components approach versus the time series correlations and variogram approach. Let's look first then at the components of variance approach.

In sampling theory, the multi-stage design is essentially based on a variance components model. For multi-stage sampling of any frame, one makes elements and then groups adjacent elements into somewhat larger units, then groups these in turn into even larger aggregates, and so keeps increasing the size of units until the largest ones are reached. These are first stage units; their immediate subdivision are second stage units, and so on back down the

scale. The classical multi-state sampling design is a simple random sample of primary units, followed by simple random samples of secondary units in the selected primary units, and so on.

In coal sampling the first stage units may be sub-lots and all are selected into the sample. Then the second stage units may be called zones [see Deming (1960) for this usage] and again all are usually selected. The third stage units may be called local neighborhoods and they will be sampled (one from each zone) or, more commonly, the same position within each zone is selected (as when using a systematic sample). Finally a particular increment is drawn from the possible increments within the local neighborhood.

Since the sub-lots and zones are completely sampled they are effectively strata. The sub-zone or within-zone sampling variability may be represented by a component of variance, written A and called (somewhat misleadingly) the trend variance in the Bertholf and Webb (1954) formulation. The within local neighborhood variance component may be written B and is the unit increment variance discussed above.

An experiment to estimate the quantities A and B has been designed for coal sampling that is quite innovative. The conveyor belt is stopped at intervals, and adjacent increments of various sizes within a local neighborhood are removed and analyzed separately. In Table IV of Bertholf and Webb one finds data for three sizes of increments (w_1 = 15 lbs., w_2 = 40 lbs. and w_3 = 100 lbs.) for each of 79 times the belt was stopped. The observed value for he i<u>th</u> stop of the belt and the j<u>th</u> size of increment may be viewed as:

$$y_{ij} = \mu_i + e_i + d_{ij} ,$$

where μ_i is the zone mean, e_i is the within zone effect and d_{ij} is the increment effect.

To analyze these data one begins by computing the three variances for each of the increment sizes as:

$$s_i^2 = \sum_{i=1}^{n} (y_{ij} - \bar{y}_{\cdot j})^2/(n-1) ,$$

for i = 1, 2 and 3, and where n = 79. Now, under the model equation:

$$E(s_i^2) = \Sigma(\mu_i - \bar{\mu})^2/(n-1) + \sigma_e^2 + \sigma_d^2/w_i + \sigma_{ra}^2$$

where $\Sigma(\mu_i - \bar{\mu})^2/(n-1)$ shows the variability among the zone means, σ_e^2 is the within zone variance component, σ_d^2 is the constant B, and σ_{ra}^2 is the variance of reduction and analysis.

In order to estimate B we subtracted an estimate of σ_{ra}^2(.060) from each of the s_i^2 and regressed these on $1/w_i$. We also omitted stops 56 through 59 which gave s_1^2 = .742, s_2^2 = .455 and s_3^2 = .444. Thus we regressed .68, .40 and .38 on 1/15, 1/40 and 1/100. This gave $\hat{B}$ = 5.6, and agrees fairly well with Bertholf and Webb's value of B = 6.80, although our methods of estimation differ.

Our interpretations also differ. They notice that the Kassel-Guy formula gives 3.92 as increment variance rather than 5.60 (or 6.80) and they decide that the Kassel-Guy formula is correct. It would seem more reasonable to recognize that the grouping of particles into cells has made the 3.92 too small and the 5.60 is the more realistic value.

Now we come to the trend variance, which refers to the uncertainty caused by positioning of the increment within a zone. Since only one increment was

drawn in each zone there is no immediately apparent basis for estimating this within zone variance component. In the general sampling literature one would use what Cochran calls a "Stratification Effects Only" variance estimator based on successive differences. This is called the "time series variance" in the bulk sampling field (Merks, 1985, p. 117).

The 74 successive differences of the weighed means of the three increments show a variance of .3747 which when halved becomes .1874. This variance also reflects the reduction and analysis as well as the increment variance. The increment variance may be estimated as 6/(155) = .039 and we already have σ^2_{ra} = .060. The trend variance is then estimated as .1874 - .039 - .060 = .088. Sampling variance thus is calculated, in Bertholf and Webb notation, as:

$$s_o^2 = .09/N + 5.6/wN + .060 ,$$

where N is the number of increments and w is the weight of each increment.

While we estimate $A \hat{=} .09$ and $B \hat{=} 5.6$, Bertholf and Webb use A = 0.2 and B = 4. It is obvious that some variability among zones is entering into their estimate of A and, since all zones were sampled, this should not have happened. This variance expression would appear to be reasonably valid for values of w in the range 10 lbs. to 200 lbs. and for spacings between increments similar to those in the experiment. In order to forecast the effects on sampling variance from changes in spacings one must adopt the correlogram or variogram approach that we will describe shortly.

When one examines an even more recent report (Visman, 1969) on the estimation of A it is apparent that this component is the sum of zone-to-zone variability plus within-zone variance. Thus it is a serious overestimate (in

most cases). In a comment on Visman's paper, Duncan (1971) questions the assumption of the reciprocal 1/w. However, when we used the three sizes to see if any power other than -1 on w would improve the fit, we found that the power -1 does seem to be most reasonable for the Bertholf-Webb data.

The, so called, time series variance or successive differences calculation can be made more elaborate by more complicated contrasts than just the (+1, -1) coefficients applied to y_i and y_{i+1}. Yates (1981) suggests (.5, -1, +1, -1, +1, -1, +1, -1, .5) applied to y_i, y_{i+1}, ..., y_{i+8} to give d_i and then the estimate of the within zones variance becomes $\Sigma_{i=1}^{n-8} d_i^2/7.5(n-8)$.

Extensions of this method of estimating the within zones variance can be made to 2-dimensional and 3-dimensional systematic or patterned sampling plans. The schemes are called lattice sampling by Yates (1981). The choice of contrasts should, in the best of circumstances, be based on knowledge of the process covariance function but I believe the methods are fairly robust so that any simple contrast should work well.

Finally we come to the variogram and correlogram literature. The formulation of a population stochastic process with covariance matrix Σ_η (as on page 13 above) has been part of sampling theory, at least since Cochran's (1946) paper and the one by Matern (1947). As we have seen, once the sampling pattern is chosen one can easily calculate the variance of an estimate as a quadratic form in Σ_η. There are two practical problems. One is choice of a relatively simple function for the elements of Σ_η, and the other is to design an experiment to collect needed data for estimating parameters of this function.

Concerning choice of a covariance function the most primitive method (an possibly the most applicable) is one described by Jowett (1952) and uses the variogram (rather than the correlogram). It consists of drawing an eye-fitted curve through points giving average squared differences at various distances separating the observations. A short step up the ladder of objectivity would be to fit a polynomial to the first few points and this is what I believe Gy (1982) does, although his book doesn't seem to have an example of this curve fitting so I'm not sure.

In my (Proctor (1981)) work on CEM data I used a mixture of exponentials to fit the empirical correlogram. The correlogram is a plot of correlations between observations separated by a given distance against those distances. This exponentials function was chosen because it allowed us to fit a range of distances from half-seconds as read from a stylus recording to weeks and months. A number of serial correlations (32 actually) were used in the fit at distances d where d = 1(1) 15(5) 100 [1 to 15 by unit (15 min.) jumps and then to 100 by jumps of 5]. The calculation was by non-linear generalized least squares based on some formulas for covariances among serial correlations that had been furnished by Bartlett (1946). In most sets of CEM data two exponentials seemed to be needed -- a short term and a long term component -- although there were differences among sets (i.e., the process was not stationary).

An exciting proposal has been made by Rose (1983) to fit a correlation function derived by him from analogy to an empirical law of H. Fairfield Smith (1938). The Rose correlogram function turns out to be:

$$\rho_d = \left(\frac{1}{2}\right)\left[(d+1)^{2-b} - 2d^{2-b} + (d-1)^{2-b}\right]$$
$$= \frac{1}{2}\left[(u+1)^{2-g} - 2u^{2-g} + (u-1)^{2-g}\right],$$

where d = u is distance between points and b = g is Smith's b. Although the function has been fit (Rose, 1987) to three serial correlations at selected distances for a batch of coal, there is as yet no general method of estimating g from observed values of r_d.

By way of illustration of what could be done fairly objectively with the variogram we will report on the results of fitting a quadratic variogram function to the first five points using the Bertholf-Webb data from their Table IV. We will use the three values at each stop weighted so as to represent a single increment of 155 lbs. For such observations the variance of reduction and analysis arises from a weighted average of three terms and is .492 x .060 = .030, where .060 = σ^2_{ra} from above and .492 = $(15^2 + 40^2 + 100^2)/155^2$.

The halved average squared differences at the first five distances are (after rounding and subtracting .030): .1548, .2332, .2636, .3012 and .2974. These are the empirical variogram ordinates. By fitting a quadratic function having no intercept through these points, using ordinary least squares, we found the variogram function as $v(d) = .1496d - .0183d^2$. The units of distance are here in 3 hour periods. An approximate variance for a systematic sample of size n at a spacing of D units between successive increments is thus given by

$$\sigma^2_x = \frac{1}{n}\left[.1496D/4 + (-.0183)D^2/12\right],$$

where this function is derived from formula (6) of Jowett (1952).

We are here supposing that nD is larger than 30 or so, so that quantities such as 1/nD and 2/nD which appear in the original formula can be omitted. As an example, consider a sample of size n = 20 from N = 80 so that D = 4. The variance of the value determined by compositing the 20 increments would be forecast to be:

$$\frac{1}{20}\left[.1496 - .0183 \times 16/12\right] + .060 = .066 .$$

Now compare this to the forecast, based on use of a 155 lb. increment taken from 20 locations, based on the (.09 + 5.6/155)/N + .060 formula. This is found to be .066 as well. This latter formula has the feature of forecasting the effect of using a 15 lb. increment as (.09 + 5.6/15)/20 + .060 = .083 which would seem a useful property.

A portion of the statistically oriented literature in bulk sampling appears as brief chapters or paragraphs in more applied books. In one such case, the statistical content almost seemed to dominate. This was the book by Sommer (1986). Here two-dimensional distribution functions, for example, were discussed at great length but used hardly at all for solving sampling questions. A table of random numbers was included but then the text reads, "The proposed method of random selection frequently requires too much effort in practice." (p. 82) This is an especially disconcerting statement in view of some loosely worded earlier ones such as "... in statistics a finite subset which consists of elements of the population is called a random sample." (p. 47)

The major mistake that I found in the book was one following a variance formula for the mean from a belt length L as an estimator of a hypothetically infinite population mean: "Although this [$E(\sigma^2_L(\bar{X})) = 0$] means that

unambiguous statements can be made about the values P_L on the basis of means $\bar{X}$, they cannot be made about the expectation P of the population of the belt, which is what we are interested in." p. 247. The value P is a hypothetical construct and only of theoretical interest; estimating it has nowhere near the importance of knowing P_L, the mean of the particular lot.

I recommend for their practicality the books by Smith and James (1981) and the one already cited by Merks (1985). I find my assessment of the Merks book differs considerably from that in a review by Bilonick (1988). Bilonick criticized Merks' use of "time series variance," which I feel was unwarranted, and then repeated the same mistake as Sommer did in saying that it is an estimate of the "global average" which is usually desired in particulate material sampling (p. 132).

In this more speculative vein let's mention some treatments of particle models which may become useful in studying the comminution, mixing, seiving, riffling, etc., operations, but so far there is a shortage of experimental data. These include a textbook by Herdan (1960) and articles by Grant and Pelton (1973) and by Elder, Thompson and Myers (1978). All of these cite earlier literature and results such as the succinct Benedetti-Pichler formula [Grant and Pelton (1973), p. 20] for mixtures of particles of two types and the potentially important Rittinger's law [Herdan (1960, p. 234] that holds grinding time to be proportional to the new surface area created.

Elegant theories have also arisen in spatial processes that are concerned with existence, with continuity, with aggregation and with other somewhat esoteric, although at times important, topics. Let me mention as pertinent to systematic sampling the papers by Jones (1948), Blight (1972) and Heilborn

(1978). The first two are on the optimum properties and the third is on variance estimation.

REFERENCES

Bartlett, M. S. (1946) "On the theoretical specification and sampling properties of autocorrelated time series," *Jour. of the Royal Statistical Society B*, 8:27- (Corrigenda, 10).

Bertholf, W. M. and W. L. Webb (1954) "Tests of the Geary-Jennings sampler at Cabin Creek," *Symposium on Bulk Sampling*, ASTM STP 114, Am. Society for Testing and Materials.

Bicking, Charles A. (1967) "The sampling of bulk materials," *Materials Research and Standards* 7:95-116.

Bilonick, R. A. (1988) "Review of J. W. Merks' 'Sampling and Weighing of Bulk Solids'," *Technometrics* 30:132-133.

Bilonick, R. A. (1989) "Quantifying sampling precision for coal ash using Gy's discrete model of the fundamental errors," *Journal of Coal Quality* 8:33-39.

Blight, B. J. N. (1972) "Sampling from an autocorrelated finite population," *Biometrika* 60:375-385.

Cochran, W. G. (1946) "Relative accuracy of systematic and stratified random samples for a certain class of populations," *Annals of Math. Stat.* 17:164-177.

Cochran, W. G. (1977) *Sampling Techniques*. Third Ed., John Wiley, NY.

Deming, W. E. (1960) *Sample Design in Business Research*. John Wiley, NY.

Duncan, A. J. (1971) "Comments on 'A General Theory of Sampling'," *Materials Research and Standards* 11:25.

Elder, R. S., W. O. Thompson and R. H. Myers (1978) "Properties of composite sampling procedures," *Technometrics* 22:179-186.

Grant, C. L. and P. A. Pelton (1973) "Role of homogeneity in powder sampling," *Sampling, Standards and Homogeneity* ASTM STP 540, Am. Soc. for Testing and Materials, pp. 16-29.

Gy, P. M. (1982) *Sampling of Particulate Materials: Theory and Practice.* Elsevier Scientific Publishing Co., NY.

Herdan, G. (1960) *Small Particle Statistics.* Academic Press, NY.

Heilborn, D. C. (1978) "Comparison of estimators of the variance of systematic sampling," *Biometrika* 65:429-433.

Jones, A. E. (1948) "Systematic sampling of a continuous parameter population," *Biometrika* 35:283-290.

Jowett, G. H. (1952) "The accuracy of systematic sampling from conveyor belts," *Applied Statistics* 1:50-59.

Kassell, L. S. and T. W. Guy (1935) "Determining the correct weight of sample in coal sampling," *Industrial and Engineering Chemistry, Analytical Edition* 7:112-115.

Matern, B. (1947) "Methods of estimating the accuracy of line and sample plot surveys," *Medd. for Statens Skogsforskings Institut* 49:1-144.

Merks, J. W. (1985) *Sampling and Weighing of Bulk Solids.* Gulf Publishing Co., Houston.

Proctor, C. H. (1985) "Fitting H. F. Smith's empirical law to cluster variances for use in designing multi-stage sample surveys," *Jour. of the Am. Stat. Assoc.* 80:294-300.

Rose, C. D. (1983) "Variances in sampling streams of coal," *Journal of Testing and Evaluation* 11:320-326.

Rose, C. D. (1987) "Coal sampling: fundamentals and new applications - belt conveyor systems," *Fact* 1:7-15.

Smith, H. F. (1938) "An empirical law describing heterogeneity in the yields of agricultural crops," *Journal of Agricultural Science* 28:1-23.

Smith, R. and G. V. James (1981) *The Sampling of Bulk Materials*. The Royal Society of Chemistry, Burlington House, London WIV OBN.

Visman, V. (1969) "A general sampling theory," *Materials Research and Standards* 9:8-13.

Yates, Frank (1981) *Sampling Methods for Censuses and Surveys*. Fourth Ed., MacMillan Pub. Co., NY.

Yule, G. U. and M. G. Kendall (1968) *An Introduction to the Theory of Statistics*. Hafner Pub. Co., NY.