Division of Research
Graduate School of Business Administration
The University of Michigan

January, 1980

SAMPLE DESIGN FOR STRATIFIED RATIO ESTIMATION

Working Paper No. 199

Roger L. Wright

The University of Michigan

ABSTRACT

This paper develops simple, easily implemented rules for designing stratified sampling plans for combined ratio estimation. The analysis is based on a superpopulation model and on approximations that hold when strata are constructed to tightly control the variation of the auxiliary variable. The proposed techniques are illustrated using a utility load research example, and are related to the somewhat different designs obtained by the cum $\sqrt{f}$ rule developed by Dalenius and Hodges.

CONTENTS

1. Introduction

Ratio estimators are widely used in sampling studies of finite populations. In many of these applications, a stratified random sampling design can be developed using the known population distribution of the positive auxiliary variable x together with information (or assumptions) about the relationship between x and the target variable y. Using a superpopulation model of this relationship together with suitable approximations, easily applied rules can be formulated for planning all aspects of the sampling design: total sample size, strata cutpoints, and strata allocation.

The emphasis in this paper is on conventional sample designs utilizing simple random sampling within prescribed strata. The sample design is chosen to optimize the expected statistical precision of the conventional combined ratio estimator. The principal innovation offered here is a simple procedure for choosing the strata cutpoints, and for planning the total sample size and the sample allocation among the strata. The proposed method is analogous to the Dalenius-Hodges procedure [5, pp. 127-135; 6; 7; 16] for choosing strata cutpoints when the mean or finite population total of x is to be estimated. However, the interest here is in stratification of the distribution of x for ratio estimation of the finite population total of y.

Any comparison of sampling designs for ratio estimation

should focus on the joint distribution of x and y. The practice of using the Dalenius-Hodges "cum $\sqrt{f}$" rule for stratification relies solely on the marginal distribution of x and yields designs that can be inefficient for ratio estimation. When a suitable previous sample of x and y is available, the conventional sample-based formula for the precision of the combined ratio estimator [5, pp. 165-167] can be used to develop a design. However, the problem of choosing the strata cutpoints to minimize the sample-based expected mean square error bears disquieting similarities to the computationally vexing traveling-salesman and bin-packing problems. As strata cutpoints are tentatively adjusted, the datapoints in the available sample migrate among strata and often cause abrupt changes to the within-strata sample variances and estimated precision. The resulting designs often seem to be overly tuned to the realized values of x and y in the available sample and not sufficiently based on the underlying joint population distribution.

To avoid this problem, this paper uses a superpopulation model to represent the relationship between x and y, together with the known finite population distribution of x. Section 2 develops this model. Section 3 considers the advantages of stratified random sampling for ratio estimation; the discussion borrows heavily from Royall and Herson's concepts of balanced sampling and robustness [12, 13, 14, 15]. The present paper suggests that stratified random

sampling with the combined ratio estimator gives much of the robustness of a balanced unstratified sample but with often-substantial gains in efficiency. The concept of strong stratification is defined in Subsection 3.1.

Section 4 uses the superpopulation model and certain approximations derived from strong stratification to develop a design rule for stratification, sample allocation, and total sample size. Section 4.1 develops the model-based expected mean square error of the ratio estimator of the finite population total of y. Section 4.2 examines the question of efficient allocation. The results developed in these two sections are used in Section 4.3 to formulate a specific design rule. A very simple expression for the gain in efficiency from stratification is developed in Section 4.4. Finally, Section 4.5 completes the model-based analysis by proposing simple estimators for the model parameters.

Section 5 presents an application and numerical example that arises in electric utility management [1, 2, 9]. Section 6 completes the paper with a discussion of the relationship between the proposed design rule and the cum $\sqrt{f}$ rule developed by Dalenius and Hodges [5, 6, 7, 16].

2. The Superpopulation Model

How should the joint distribution of x and y be specified at the planning stage? In the applications of interest, the finite population consists of N units labelled

1, 2, $\cdots$, N and x_k is known for each unit k. Much less is known about the conditional distribution of y given x so that it is appropriate and convenient to utilize a superpopulation heteroscedastic regression model. In this formulation y_k is regarded as the realized value of a random variable denoted Y_k which is determined from x_k and a random disturbance ε_k following the regression equation

$$Y_k = h(x_k) + \varepsilon_k [v(x_k)]^{1/2} \qquad k = 1, 2, \cdots, N. \qquad (2.1)$$

Here the expected value and variance of Y_k depend on x_k and are denoted as $h(x_k)$ and $\sigma^2 v(x_k)$ respectively. The disturbances $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N$ are independent random variables with mean zero and variance σ^2. This model, notation, and many of the analytical techniques follow Royall and Herson [13]. A more comprehensive presentation of this approach is given in [4].

At the design stage it is helpful to adopt a specific form of (2.1) that combines reasonable accuracy, parsimony and analytical convenience. In this paper we will consider the superpopulation model ξ given by (2.1) with the very simple specifications $h(x_k) = \beta x_k$ and $v(x_k) = x_k^{\gamma}$. In many of the applications of the combined ratio estimator, this model provides a sufficiently realistic basis for sample design. The three parameters β, γ and σ can be assessed fairly easily even when relevant data is severely limited. Using this model sample designs can be developed following simple

and sensible rules which extend the conventional sample-based formulas. However the robustness of these rules to model misspecification remains to be investigated.

The superpopulation model ξ has been widely used to explore the properties of ratio estimation. If the heteroscedasticity parameter γ is one, the Gause-Markov theorem implies that the simple ratio estimator is the best linear unbiased estimator of the superpopulation parameter β. Brewer [3] and Royall [11] have extended this result to prediction of the finite population ratio $\Sigma_1^N Y_k / \Sigma_1^N x_k$ and the population total $\Sigma_1^N Y_k$. If γ is different from one, the simple ratio estimator is still unbiased but not the most efficient estimator. However, in the absence of complete confidence in the accuracy of the superpopulation model ξ and perfect knowledge of γ, the ratio estimator is often chosen in practice because of its robustness [13].

3. Stratified Sampling

Stratified sampling usually provides two major advantages over simple random selection: control over the sample distribution of x, and more efficient allocation of the sample among the population units to reflect differences in the conditional variance of Y. Royall and Herson [13] have clarified the advantages of a sample with a balanced x-distribution. They have shown that balanced samples provide protection against bias arising from misspecification of h(x) in the

superpopulation model (2.1). Unfortunately, their unstratified balanced samples often suffer a substantial loss of efficiency compared to samples that are optimal for a particular specification of h(x) and v(x) [13, p. 885].

Stratified sampling regains some of the lost efficiency with little sacrifice of robustness. The balance is achieved by using an estimator that weighs observations according to the known population distribution of x. Substantial gains in efficiency can be achieved by allocating the sample observations appropriately. Although the stratified sample estimator may still not be as efficient as the optimal-sample ratio estimator for a particular superpopulation, it generally seems to be much more robust.

Ratio estimators can be adopted to stratified sampling either by using a separate ratio estimator within each stratum or by using a single combined ratio determined from the stratified-sample estimators of the means of y and x [5, pp. 164-169]. Under the model ξ of Section 2 the combined ratio estimator would usually be recommended for small samples while for larger samples the choice between the combined and separate ratio estimators would depend upon the credibility of the specification of the regression function h(x). If the model ξ is plausible at the planning stage, it is convenient to consider the combined estimator for planning even though the separate estimator might be used in the analysis for added robustness. In [14] Royall and Herson

have provided certain results useful for designing stratified samples for the separate ratio estimator under a broad class of superpopulation models. The present paper will propose simpler and more prescriptive sampling rules for the combined estimator under the model ξ.

3.1 Strong Stratification According to Size

The design rule developed in this paper is developed from a concept called strong stratification. In order to develop this idea in adequate detail, some notation is required. Suppose that the population of N units is divided into H strata with N_h units in strata h. The population is stratified according to x if stratum one contains the N_1 units that are smallest as measured by x, and in general, if stratum h contains the N_h smallest units excluded from strata 1, 2, $\cdots$, h-1. Each such stratification S is essentially determined by the number of strata H and the strata sizes N_1, N_2, $\cdots$, N_H.

Each stratification S determines within-strata population moments of x. The population mean of x^c (c = 1, 2, $\gamma/2$, γ, etc.) within stratum h will be denoted as $\bar{x}_h^{(c)}$, i.e. $\bar{x}_h^{(c)} = \Sigma x_k^c/N_h$ where the summation is over the N_h units in stratum h.

The absence of the subscript h will denote the overall population mean, so that

$$\bar{x}^{(c)} = \Sigma_1^N x_k^c/N = \Sigma_1^H N_h \bar{x}_h^{(c)}/N.$$

The absence of the superscript (c) will indicate the first moment (c = 1), either overall or within strata.

A very simple rule for design can be formulated by concentrating on stratifications that tightly control the variance of x within each stratum. A stratification S will be called strong for a given number c if the variance of x^c is small within each stratum, i.e. if $\bar{x}_h^{(2c)} \doteq (\bar{x}_h^{(c)})^2$ for $1 \leq h \leq H$.

3.2 Overall Balance of a Stratified Sample

Royall and Herson's concept of a balanced sample can be adopted to combined ratio estimation. Consider a stratification S and a particular sample s comprised of n units from the population. Let n_h be the number of sample units from stratum h and let $\bar{x}_{sh}^{(c)}$ denote the sample mean of x^c within stratum h so that $\bar{x}_{sh}^{(c)} = \Sigma x_k^c / n_h$. The overall stratified sample mean of x^c is $\Sigma_1^H N_h \bar{x}_{sh}^{(c)} / N$ which will be denoted $\bar{x}_s^{(c)}$.

The stratified sample s has overall balance of degree J if the overall stratified sample moment $\bar{x}_s^{(c)}$ is equal to the population moment $\bar{x}^{(c)}$ for all c = 1, 2, $\cdots$, J. For overall balance it is sufficient but not necessary that the sample is balanced within each stratum, i.e. that s is a balanced stratified sample of degree J as defined in [14, p. 890].

Royall and Herson [13, p. 883] showed that for a

simple balanced sample of degree J the simple ratio estimator is unbiased under any superpopulation model (2.1) for which h(x) is a polynomial of degree J. A similar approach shows that the combined ratio estimator is unbiased under the same family of superpopulation models if the stratified sample has overall balance of degree J.

3.3 Random Sampling Within Strata

If a stratification S is strong for given c (Section 3.1), and if a sample is obtained by randomly selecting n_h units from each stratum then conventional sampling theory shows that $\bar{x}_s^{(c)}$ has high probability of being close to $\bar{x}^{(c)}$. This implies that a random, strongly stratified sampling plan is likely to lead to an approximately balanced sample. Under these conditions, the combined ratio estimator is likely to be quite robust. In particular if the stratification is strong at least for c = 1, then approximate first degree overall balance can be expected to provide protection against bias from a nonzero intercept β_0 in $h(x) = \beta_0 + \beta_1 x$.

There are some strong arguments against random sampling when a superpopulation model can be assumed. For example, under ξ with γ equal to one, the model-based mean square error of the simple ratio estimator can be minimized by observing Y_k for the n largest units in the population [13, p. 883]. However, this design could give a badly biased estimator if ξ is inaccurate. If complete confidence in ξ is

lacking, an approximately balanced sample may be preferred. One convenient way of providing approximate balance is to use a strongly stratified design.

With strong stratification, the arguments against randomization are almost mute. As long as v(x) is continuous, there will be little heteroscedasticity in the superpopulation model within strata and little preference among units. At a small cost, randomization will provide protection against model misspecification. For example randomization reduces the concern about systematic selection of outliers. It will be seen that randomization also contributes to the simplicity of the design rules.

It is important to note that the overall balance of a stratified sample does not require proportional sample allocation among strata since the population sizes are used in the overall sample moments. This means that robustness can be retained even while the sample allocation is chosen to maximize the expected precision of the combined ratio estimator.

4. Sample Design for Ratio Estimation With Strong Stratification

A stratified random sampling design p has three components: a stratification S determined by $N_1, N_2, \cdots, N_H$, the overall sample size n, and the sample allocation $n_1, n_2, \cdots, n_H$. The sampling plan p determines a sample

space of $\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_H}{n_H}$ equally likely stratified samples. For each such sample s, $x_1, x_2, \cdots, x_h$ are fixed but $Y_1, Y_2, \cdots, Y_h$ are regarded as random variables specified by the model ξ.

The criterion for selecting p is the model-based expected mean square error of the combined ratio estimator $\hat{T} = (\bar{Y}_s/\bar{x}_s)\Sigma_1^N x_k$. Here $\hat{T}$ is regarded as a predictor of the finite population total $T = \Sigma_1^N Y_k$ which is also a random variable under ξ. Initially the conditional mean square error $E_s(\hat{T} - T)^2$ is evaluated using ξ with a fixed sample s. Then the unconditional mean square error $E(\hat{T} - T)^2$ is obtained by averaging $E_s(\hat{T} - T)^2$ over the sample space determined by the design p.

Very simple rules for selecting a good design can be followed if two approximations are reasonable: Condition A: The desired design is strongly stratified for $c = \gamma/2$ so that $\bar{x}_h^{(\gamma)} \doteq (\bar{x}_h^{(\gamma/2)})^2$ for all strata h. Here γ is the heteroscedasticity parameter of the superpopulation model ξ. Condition B: The desired design is strongly enough stratified for c = 1 to be confident that a randomly selected sample s will have approximate overall balance of degree one, i.e. $\bar{x}_s \doteq \bar{x}$. Condition B is used as in conventional analysis to neglect the sampling variation of the denominator of $\hat{T}$.

4.1 The Expected Mean Square Error

Using the model ξ, the combined ratio estimation $\hat{T}$ may be written as $(\beta + \Sigma_1^H N_h \bar{u}_{sh}/N\bar{x}_s)\Sigma_1^N x_k$ where $u_k = Y_k - \beta x_k = \varepsilon_k x_k^{\gamma/2}$. Similarly, the finite population total T is $(\beta + \Sigma_1^H N_h \bar{u}_h/N\bar{x})\Sigma_1^N x_k$. So if ξ is accurate, $\hat{T}$ is an unbiased predictor of T for any sample s since

$$E_s(\hat{T}) = E_s(T) = \beta\Sigma_1^N x_k.$$

Moreover, the expected mean square error given s, $E_s(\hat{T}-T)^2$, is

$$\bar{x}^2\Sigma_1^H N_h^2 E_s(\bar{u}_{sh}/\bar{x}_s - \bar{u}_h/\bar{x})^2$$

which is equal to

$$\sigma^2\bar{x}^2\Sigma_1^H N_h^2[\bar{x}_{sh}^{(\gamma)}/n_h\bar{x}_s^2 + \bar{x}_h^{(\gamma)}/N_h\bar{x}^2 - 2\bar{x}_{sh}^{(\gamma)}/N_h\bar{x}_s\bar{x}] \qquad (4.1)$$

Condition B implies that $\bar{x}_s \doteq \bar{x}$ so that $E_s(\hat{T} - T)^2$ is approximately

$$\sigma^2\Sigma_1^H N_h^2[\bar{x}_{sh}^{(\gamma)}/n_h + \bar{x}_h^{(\gamma)}/N_h - 2\bar{x}_{sh}^{(\gamma)}/N_h]$$

When this expression is averaged over the sample space of a specific design p, the unconditional expected mean square error can be approximated as

$$E(\hat{T} - T)^2 \doteq \sigma^2\Sigma_1^H N_h^2\bar{x}_h^{(\gamma)}(n_h^{-1} - N_h^{-1}). \qquad (4.2)$$

4.2 Approximately Optimal Allocation

For a given stratification S and a given total sample size n, the expression (4.2) for the approximate expected mean square error of $\hat{T}$ can be minimized by choosing the sample allocation

$$n_h/n = N_h(\bar{x}_h^{(\gamma)})^{1/2}/\Sigma_1^H N_h(\bar{x}_h^{(\gamma)})^{1/2}. \tag{4.3}$$

This follows from the argument commonly used to demonstrate the optimality of Nehman allocation [5, pp. 96-98].

With the allocation (4.3), the approximate mean square error (4.2) becomes

$$\sigma^2[\Sigma_1^H N_h(\bar{x}_h^{(\gamma)})^{1/2}]^2/n - \sigma^2\Sigma_1^H N_h\bar{x}_h^{(\gamma)}.$$

The preceding expression can be considerably simplified under Condition A. Note that the term $\Sigma_1^H N_h\bar{x}_h^{(\gamma)}$ is equal to $N\bar{x}^{(\gamma)}$ where $\bar{x}^{(\gamma)}$ is the overall population mean of x^γ. Moreover, Condition A implies that $\Sigma_1^H N_h(\bar{x}_h^{(\gamma)})^{1/2}$ can be approximated as $\Sigma_1^H N_h\bar{x}_h^{(\gamma/2)}$ which is equal to $N\bar{x}^{(\gamma/2)}$.

So under the model ξ and under the Conditions A and B of strong stratification,* then the expected mean square error of the combined ratio estimator $\hat{T}$ is approximately

$$E(\hat{T} - T)^2 \doteq N^2\sigma^2[(\bar{x}^{(\gamma/2)})^2/n - \bar{x}^{(\gamma)}/N]. \tag{4.4}$$

Condition A also yields two very helpful approximate reformulations of the allocation rule (4.3), namely

---

*and under the approximately optimal allocation (4.3),

$$n_h/n \doteq N_h \bar{x}_h^{(\gamma/2)} / N\bar{x}^{(\gamma/2)} \tag{4.5a}$$

$$= \Sigma_1^{N_h} x_k^{\gamma/2} / \Sigma_1^{N} x_k^{\gamma/2} \tag{4.5b}$$

The first of these relationships implies that the within-strata sampling fraction n_h/N_h is proportional to the within-stratum mean $\bar{x}_h^{(\gamma/2)}$. In the case that the superpopulation model ξ is homoscedastic, γ is zero and our allocation gives a constant sampling fraction. If $\gamma > 0$, then the sampling fraction increases from stratum to stratum in proportion to the mean $\bar{x}_h^{(\gamma/2)}$. The larger the heteroscedasticity parameter γ, the more heavily oversampled are the units with large x. The case that $\gamma = 2$ is closely related to sampling with probability proportional to size.

The second formulation given above, (4.5b), implies that the sample allocation n_h/n is proportional to the within strata population totals of $x^{\gamma/2}$. In the next section, this approximate characterization of the allocation (4.3) will be used for a convenient stratification rule.

4.3 Design Rules

In the expressions to be developed in this section, the ratio $(\bar{x}^{(\gamma/2)})^2/\bar{x}^{(\gamma)}$ occurs repeatedly and will be referred to as the design effect and denoted $de(\gamma)$. The term design effect will be justified in the next section.

The design effect $de(\gamma)$ can be conveniently characterized in terms of the coefficient of variation of $x^{\gamma/2}$, defined as

$$cv(x^{\gamma/2}) = [\bar{x}(\gamma) - (\bar{x}(\gamma/2))^2]^{1/2}/\bar{x}(\gamma/2).$$

In fact

$$de(\gamma) = [cv^2(x^{\gamma/2}) + 1]^{-1}. \tag{4.6}$$

Using the design effects, (4.4) may easily be restated. Under the superpopulation model ξ, and using a sample design satisfying the Conditions A and B of strong stratification together with the approximately optimal allocation (4.3), the expected mean square error of $\hat{T}$ is approximately

$$E(\hat{T} - T)^2 \doteq N^2\sigma^2_{\bar{x}}(\gamma)[de(\gamma)/n - 1/N]. \tag{4.7}$$

This result gives the following rule for choosing the sample size n with a strongly stratified design and approximately optimal allocation. For expected mean square error $E(\hat{T} - T)^2$ approximately equal to N^2s^2, choose the sample size n equal to n_{cr} where

$$n_{cr} = de(\gamma)n_0/[1 + n_0/N], \tag{4.8a}$$

with

$$n_0 = \sigma^2_{\bar{x}}(\gamma)/s^2. \tag{4.8b}$$

Here σ^2 and γ are the parameters of the superpopulation

model ξ defined in Section 2. Estimation of these parameters is considered in Section 4.5.

One surprising implication of the proceeding result is that the expected mean square error is insensitive to the choice of stratification as long as approximately optimal allocation, (4.3) or (4.5), is used. However, in most applications it is desirable to choose a design giving equal subsamples from each stratum. If (4.5b) is used for allocation, the strata sample sizes will be equal if the population totals of $x^{\gamma/2}$ are equal within all strata. This suggests the following design rule for combined ratio estimation, subsequently called the cr-rule. Choose the number of strata H large enough to provide Conditions A and B of strong stratification, and choose the strata sizes $N_1, N_2, \cdots, N_H$ to equalize the population totals of $x^{\gamma/2}$ within all strata. Determine the required sample size from (4.8) and allocate the sample equally among strata.

In some applications the cr-rule may prescribe a sample size n_h exceeding the population size N_h for some stratum h. If $\gamma > 0$, it is sufficient to consider stratum H. In this case, the cr-rule can be modified to provide a 100 percent sample of stratum H while maintaining the efficiency of the allocation and the validity of our approximation (4.7) for the mean square error. Simply decrease the lower boundary of stratum H until $\bar{x}_h^{(\gamma/2)} \doteq N\bar{x}^{(\gamma/2)}/n$. Then use the cr-rule to stratify the rest of the population and

to allocate the remainder of the sample. If $\gamma < 0$, a similar adjustment might be required in stratum 1.

4.4 Gain from Stratification

This section examines the gain in efficiency of stratification and allocation following the cr-rule. Our comparisons will be with a design using a simple random sample of size n and the ordinary ratio estimator $(\Sigma_1^n y_k/\Sigma_1^n x_k)\Sigma_1^N x_k$. Conditional on any sample s the expected mean square error of the ordinary ratio estimator can be found from (4.1) with H = 1 to be

$$(N\sigma_{\bar{x}})^2[\bar{x}_s^{(\gamma)}/n\bar{x}_s^2 + \bar{x}^{(\gamma)}/N\bar{x}^2 - 2\bar{x}_s^{(\gamma)}/N\bar{x}_s\bar{x}].$$

Here $\bar{x}_s$ and $\bar{x}_s^{(\gamma)}$ are the unstratified sample moments, $\bar{x}_s = \Sigma_1^n x/n$ and $\bar{x}_s^{(\gamma)} = \Sigma_1^n x^\gamma/n$. Assume that s is a large simple random sample so that $\bar{x}_s \doteq \bar{x}$. Then the unconditional expected mean square error of the ordinary ratio estimator is approximately

$$N^2\sigma^2_{\bar{x}^{(\gamma)}}(1/n - 1/N).$$

This will be equal to N^2s^2 if the size of the simple random sample n is equal to n_r where

$$n_r = n_0/(1 + n_0/N). \tag{4.9}$$

Here n_0 is $\sigma^2_{\bar{x}^{(\gamma)}}/s^2$ as in (4.8b), and n_0 is the simple random sample size required if the population size N is large.

A comparison of (4.9) with (4.8a) gives the following result. Given the superpopulation model ξ, approximately the same expected mean square error can be achieved by either the ordinary ratio estimator with a simple random sample of size n_r or by the combined ratio estimator with an appropriately allocated, strongly stratified sample of size $n_{cr} = de(\gamma)n_r$. So $de(\gamma)$, (4.6), gives the reduction in the sample size achieved by strong stratification with allocation following the cr-rule.

It is interesting to note that if the superpopulation relationship is homoscedastic, then γ is equal to zero and the design effect is one so that there is no gain from stratifying. For a fixed population distribution of x, the greater γ is, the greater is the gain from stratification.

4.5 Estimation of Model Parameters

In order to implement the cr-rule for a strongly stratified sample, the parameters γ and σ of the superpopulation model ξ must be estimated at the design stage. A full analysis of the estimation problem might address many rather complex issues perhaps including a Bayesian analysis of inference featuring the value of sample information, methods of pooling information drawn from various more or less relevant populations, and a comparison of model-efficient estimators with robust estimators. In this paper, estimation is not the main focus and it may suffice to suggest simple es-

timators that minimize reliance on superpopulation assumptions and that maximize consistancy with common sampling-based theory and practice.

Consider first the estimation of the heteroscedasticity parameter γ from m observations of Y_k generated according to ξ for given x_k. A simple procedure for estimating the heteroscedasticity parameter of a regression model has been proposed by Park [10] and developed further by Harvey [8]. Let u_k be the deviation $Y_k - \beta x_k$ which is equal to $\sigma\varepsilon_k x_2^{\gamma/2}$ under ξ, and assume that the first two moments of $\ell n(\varepsilon_k^2)$ exist. Then

$$\ell n(u_k^2) = \alpha + \gamma \ell n(x_k) + v_k \tag{4.10}$$

where $\alpha = E[\ell n(\sigma\varepsilon_k^2)$ and $v_k = \ell n(\sigma\varepsilon_k^2) - \alpha$. Under ξ, the disturbances v_k are independent and identically distributed so that the coefficient γ could be estimated using ordinary least squares if the u_k were observable. An asymptotically unbiased predictor of u_k is $\hat{u}_k = Y_k - \hat{\beta}x_k$ where $\hat{\beta} = \Sigma_1^m y_k / \Sigma_1^m x_k$. This leads us to the least squares regression estimator

$$\hat{\gamma} = (\overline{zw}_s - \bar{z}_s\bar{w}_s)/[\bar{w}_s^{(2)} - (\bar{w}_s)^2]. \tag{4.11}$$

Here $z_k = \ell n(\hat{u}_k^2)$, $w_k = \ell n(x_k)$ and $\overline{zw}_s$ is the sample moment $\Sigma_1^m z_k w_k/m$, $\bar{w}_s^{(2)}$ is $\Sigma_1^m w_k^2/m$ and so on.

If the available sample was randomly selected, the sample deviations $\hat{u}_k$ can also be used to estimate the quantity

$\sigma^{2}_{\bar{x}}(\gamma)$ which arises in (4.7)-(4.9). Define

$$s^2 = \Sigma_1^m \hat{u}_k^2/(m-1). \tag{4.12}$$

Conditional on the Y_k of the finite population, conventional sampling theory shows that s^2 is an asymptotically unbiased estimator of $\bar{u}^{(2)} = \Sigma_1^N u_k^2/N$. But under ξ, the model-based expectation of $\bar{u}^{(2)}$ is $\sigma^{2}_{\bar{x}}(\gamma)$. This implies that s^2 is a model-based asymptotically unbiased estimation of $\sigma^{2}_{\bar{x}}(\gamma)$. This estimator is closely related to robust estimators considered in [12].

If s^2 is used to estimate $\sigma^{2}_{\bar{x}}(\gamma)$ and if $de(\hat{\gamma})$ is used to estimate $de(\gamma)$, then equations (4.8) and (4.9) give the following estimates of n_0, n_r, and n_{cr}:

$$n_0 = s^2/s^2 \tag{4.13a}$$

$$n_r = n_0/(1 + n_0/N) \tag{4.13b}$$

$$n_{cr} = de(\hat{\gamma})n_r. \tag{4.13c}$$

Here N^2s^2 is the expected mean square error required for the estimator of the population total T, n_0 is the sample size required using the ordinary ratio estimator with a large population, n_r is the sample size required using the ordinary ratio estimator with the finite population correction, and n_{cr} is the sample size required using the combined ratio estimator with allocation following the cr-rule. One noteworthy implication of this use of s^2 is that (4.13a) and

(4.13b) are completely consistent with conventional sampling-based analysis.

Suppose now that instead of a simple random sample, we have a stratified random sample with m_h observations in each stratum h = 1, 2, ···, H. Then equation (4.12) defining S^2 must be modified to use the population-weighted average of within-strata sample variances. Let $S_h^2 = \Sigma\hat{u}_k^2/(m_h-1)$, the sample variance of $\hat{u}_k$ within stratum h. In place of (4.12), define

$$S^2 = \Sigma_1^H (N_h/N) S_h^2. \qquad (4.14)$$

With this redefinition of S^2, equations (4.13a-c) remain appropriate.

These procedures will be illustrated in the following section.

5. Sampling in Electric Utility Load Research

The sampling rules discussed in this paper have been developed for a specific application in electric utility load research. Interruptions of electric power service endanger the health of many people, disrupt business, and inconvenience everyone. Utility planners, regulators and managers must provide enough generating and transmitting capacity to instantaneously meet their customers' greatest demands for electricity. To minimize the cost of service, managers try to maintain an efficient balance of base generating units

which have high capital costs but low operating costs, and peak generating units which have low capital costs but high operating costs. Regulators try to establish rates that fairly allocate both capital and operating costs among users.

These various concerns -- capacity planning, power production, and rate setting -- all require accurate assessment of the timing of the customers' demand for electricity. Most utilities measure their power production almost continuously and these data can be adjusted for transmission losses to estimate past demand on an almost instantaneous system-wide basis. In addition, customer billing procedures generally measure each customer's use of electricity monthly or bi-monthly.

Despite this abundance of accurate data describing the entire generating system and population of customers, utility managers and regulators need additional data describing the timing of demand for electricity within certain customer subpopulations, especially present or proposed rate-groups. These data are usually obtained by utilizing special time-of-day meters for a sample of customers. These meters generally measure an individual customer's usage of electricity during each consecutive fifteen minute period. The data are continuously recorded on magnetic tapes which are periodically returned to the utility or to a service bureau for editing and transcription. The expenses of equipment acquisition and

maintenance, meter installation, data collection, and data processing usually add up to several hundred dollars per sample customer, so sample planning is important. A discussion of this application is especially timely because the Public Utility Regulatory Policy Act of 1978 will greatly expand this activity. In fact this act requires most utilities to develop sampling plans to estimate demand characteristics of specified customer subpopulations with 90 percent probability of less than 10 percent error.

Current practices and issues in electric utility load research are described in detail in various publications and presentations of the Load Research Committee of the Association of Edison Illuminating Companies, especially [2, 9]. Aigner's work [1] is also relevant.

While the details vary, the methods of sample design discussed in the present paper are generally applicable to load research. Consider a specific population of customers, let x_k be the use of electricity of customer k during a specific month as recorded by the customer billing procedure, and let Y_k be his use of electricity during a specific interval of time within the month, perhaps the hour of peak system usage. For convenience, the variable x_k will be called "use" and Y_k will be called "demand." We regard use x_k as nonstochastic, but demand Y_k as a random variable. The utility wants to predict the total demand $T = \Sigma_1^N Y_k$ of all N

customers. It is equivalent to predict the ratio $B = \Sigma_1^N Y_k / \Sigma_1^N x_k$ or, as is more common in the industry, to predict the reciprocal B^{-1} which is called the load factor. In the latter case x_k is usually redefined as the average hourly use during the month.

The utility needs to plan a procedure for selecting n customers for whom demand Y_k will be measured using time-of-day meters. Since these data will be used to determine electricity rates for the population of customers, any design involving non-random sample selection is politically unattractive and random sampling is almost always used. Royall and Herson's concept of balanced sampling also helps to justify the preference of utility managers and regulators for random sampling.

Many load research designs stratify the population according to use, x. Stratified random sampling lets the researcher oversample the large customers while preserving the impartiality of random selection. This practice is supported by the results of this paper which show that an allocation following the cr-rule usually provides greater efficiency than a simple random sample and also much of the robustness of a balanced sample.

A stratified sampling design must specify strata boundaries, the total sample size, and the sample allocation among strata. If the superpopulation model ξ (Section 2) is believed to be reasonably accurate, and if the parameters γ

and σ^2 can be assessed, then an efficient sampling design can be developed following the rules given in Section 4.3. If a relevant sample is available, the estimation techniques of Section 4.5 are applicable.

A numerical example may be helpful. In [9], Higgins provides data describing use x_k (mwh) and demand y_k (kw) for each of 210 customers. Figure 1 shows a scatterplot of the data, and Table 1 provides sample statistics.

Table 1
Sample Statistics for Example

$n = 210$

$\bar{x}_s = 1589$ mwh

$\bar{y}_s = 3757$ kw

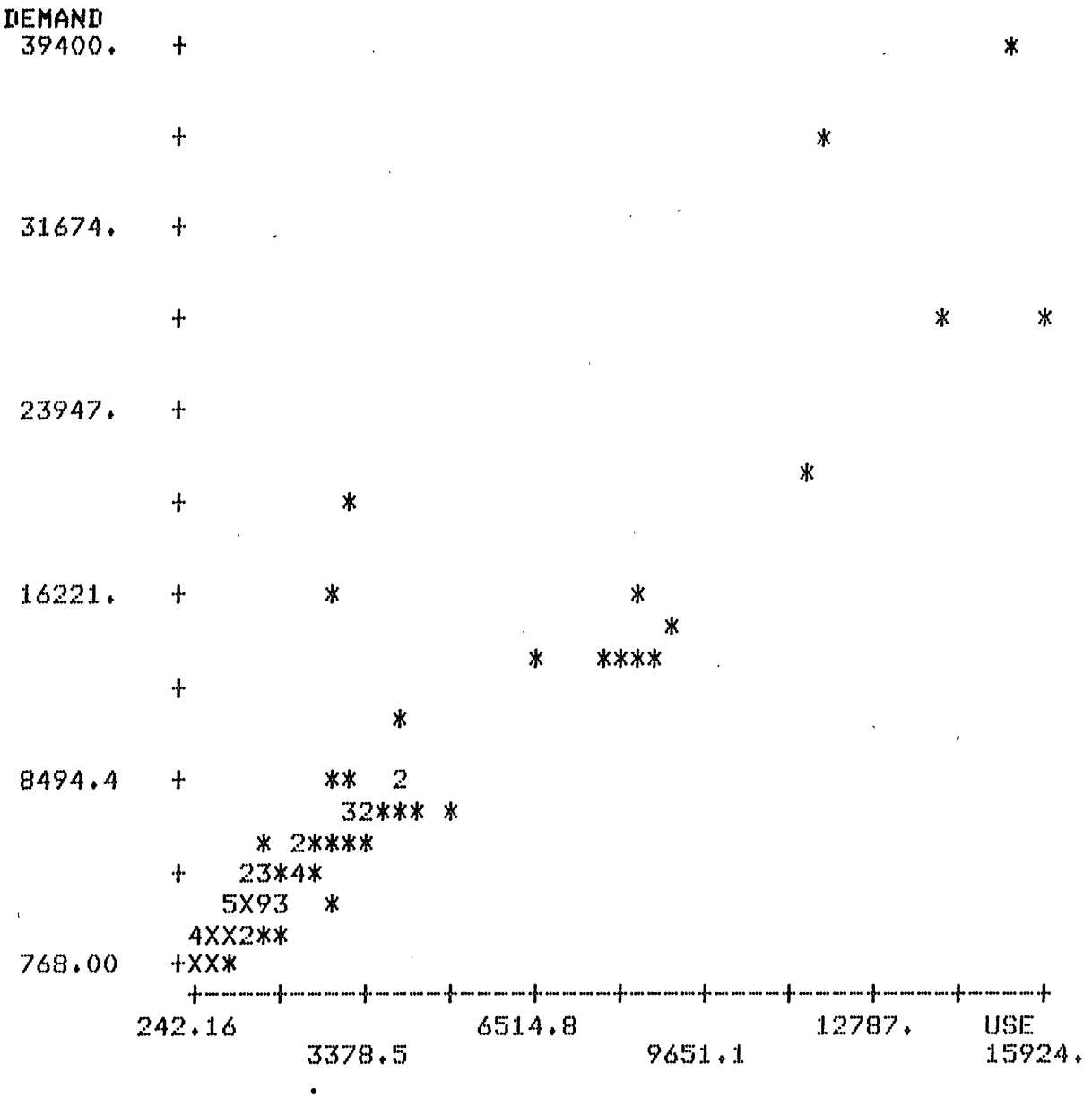
$\hat{\gamma} = 1.704$

$S = 1920$ kw

While population statistics were not published, the population size N and the population distribution of use would ordinarily be readily available. For the sake of illustration we assume $N = 740$, $\bar{x} = 1400$ mwh, and $cv(x^{\hat{\gamma}/2}) = 1.238$ so that $de(\hat{\gamma}) = .3948$. These figures are consistent with the available sample.

Suppose that the design criterion is 10 percent relative error with 90 percent probability as required for the Public Utility Regulatory Policy Act. Then the expected

Figure 1. Scatterplot of Demand and Usage for Example



mean square error N^2s^2 should satisfy

$$1.645s = (.10)\hat{\beta}\bar{x} = (.10)(3757/1589)(1400)\text{kw}$$

so that s = 201.2 kw. Then (4.13a-c) imply

$$n_0 = (1920/201.2)^2 = 91 \text{ units},$$
$$n_r = 91/(1 + 91/740) = 81 \text{ units, and}$$
$$n_{cr} = (.3948)(81) = 32 \text{ units.}$$

A practical design might use H = 8 strata with four units per strata. Following the cr-rule, the eight strata are chosen to equalize the population totals of $x^{\hat{\gamma}/2}$ within all strata. Using the available sample data, the resulting stratification is shown in Table 2.

Table 2

Choice of Stratification for Example

Strata h	Size N_h	Upper Boundary x_h	Sampling Fraction (%) n_h/N_h	$\bar{x}_h^{(\hat{\gamma}/2)}$	$(\bar{x}_h^{(\hat{\gamma})})^{1/2}$
1	268	575	1.5	170.0	173.0
2	173	840	2.3	261.3	262.3
3	113	1483	3.5	403.0	407.2
4	74	2522	5.4	603.0	608.3
5	49	3814	8.2	931.4	934.6
6	28	7955	14.3	1398.0	1432.0
7	21	11600	16.7	2347.0	2360.0
8	14	16000	28.6	3457.0	3473.0

Despite the comparatively small number of strata, Conditions A and B of strong stratification seem well satisfied. The consistency between the last two columns of Table 2 confirms Condition A. Condition B can be verified by showing that the coefficient of variation of $\bar{x}$ is small. In fact, it turns out to be 2.8 percent.

An interesting alternative design would be to use 32 strata with one unit selected at random from each stratum.

6. Comparison with Dalenius-Hodges Stratification

Rules for constructing strata boundaries have been considered by Dalenius and others [5, pp. 129-131; 6; 7; 16]. Cochran summarizes this work by saying that "the cum $\sqrt{f}$ rule applied to x should give an efficient stratification for another variable y that has a linear regression on x with high correlation," [5, p. 131]. However, the cr-rule for stratification proposed in this paper is quite different than Dalenius and Hodges' cum $\sqrt{f}$ rule.

To simplify the discussion suppose that the population distribution of x is continuous with the probability density function $f(x)$. The cum $\sqrt{f}$ rule is to choose strata boundaries, say x_{h-1} and x_h, to equalize the integrals $\int_{x_{h-1}}^{x_h} \sqrt{f(x)}\, dx$ between all strata, and then to allocate the sample equally among strata. Under the cr-rule, on the other hand, the integrals $\int_{x_{h-1}}^{x_h} x^{\gamma/2} f(x)dx$ are equalized. Both rules balance integrals of the form $\int_{x_{h-1}}^{x_h} w(x)f(x)dx$ where the weight func-

tion $w(x)$ is $f(x)^{-1/2}$ for the cum $\sqrt{f}$ rule and $x^{\gamma/2}$ for the cr-rule.

This comparison can be reformulated in terms of the sampling fractions n_h/N_h. If we ignore 100 percent sampling constraints, both rules divide the total sample n equally among the strata. This implies that the within-strata sample size n_h is proportional to the integral $\int_{x_{h-1}}^{x_h} w(x)f(x)dx$ so that the sampling fraction n_h/N_h is proportional to the within-strata population mean of $w(x)$, $\int_{x_{h-1}}^{x_h} w(x)f(x)dx/\int_{x_{h-1}}^{x_h} f(x)dx$. In the case of the cr-rule, the sampling fractions are proportional to the within strata means of $x^{\gamma/2}$ as was pointed out in Section 4.2, so that the sampling fractions increase with size as long as $\gamma > 0$. However, in the case of the cum $\sqrt{f}$ rule, the sampling fractions are proportional to the means of $f(x)^{-1/2}$ within each stratum, so that the sampling fractions decrease as the density of x increases.

The difference between the two rules is most dramatic for strata below the mode of a unimodel distribution. The cr-rule will give sampling fractions which increase with x if $\gamma > 0$, while the cum $\sqrt{f}$ rule will give sampling fractions which decrease with increasing x.

The two rules will be equivalent only if their weight functions $f(x)^{-1/2}$ and $x^{\gamma/2}$ are proportional, i.e. if $f(x)$ is proportional to $x^{-\gamma}$ wherever $f(x)$ is nonzero.

The results given in this paper appear to contradict Cochran's evaluation of the cum $\sqrt{f}$ rule, quoted above. The

difference in designs derives from different choices of estimator. Dalenius' work dealt with the stratified sampling estimator of the mean or total of y while the cr-rule relates to the combined ratio estimator. The cum $\sqrt{f}$ rule is appropriate if x is to be used for stratification but not for estimation. The cr-rule seems to be appropriate in those cases in which x is to be used both for stratification and in ratio estimation.

REFERENCES

[1] Aigner, D.J., "Bayesian Analysis of Optimal Sample size and a Best Decision Rule for Experiments in Direct Load Control," Journal of Econometrics, 9 (1979), 209-22.

[2] Brandenburg, L. and Higgins, C.E., Jr., "Stratified Random Sampling Methods for Class Load Surveys for Electric Utilities," Applied Statistics for Loan Research, Vol. III, Association of Edison Illuminating Companies, New York (July 1974).

[3] Brewer, K.W.R., "Ratio Estimation in Finite Populations: Some Results Deducible from the Assumptions of an Underlying Stochastic Process," Australian Journal of Statistics 5 (1963), 93-105.

[4] Cassel, C.M., Sarndal, C.E., and Wretman, J.H., Foundations of Inference in Survey Sampling, John Wiley & Sons, New York, 1977.

[5] Cochran, W.G., Sampling Techniques, Third Edition, John Wiley & Sons, New York, 1977.

[6] Dalenius, T., Sampling in Sweden, Contributions to the Methods and Theories of Sample Survey Practice, Alquist and Wicksell, Stockholm, 1957.

[7] Dalenius, T. and Hodges, J.L., Jr., "Minimum Variance Stratification," Journal of the American Statistical Association, 54 (1959), 88-101.

[8] Harvey, A.C., "Estimating Regression Models with Multiplicative Heteroscedasticity," Econometrica, 44 (1976), 461-64.

[9] Higgins, C.E., Jr., "Stratified Random Sampling Methods for Class Load Surveys for Electric Utilities," Unpublished Mimeograph, Virginia Electric & Power Company.

[10] Park, R.E., "Estimation with Heteroscedastic Error Terms." Econometrica, 34 (1966), 888.

[11] Royall, R.M., "On Finite Population Sampling Theory Under Certain Linear Regression Models," Biometrika, 57 (1970), 377-87.

[12] Royall, R.M., and Cumberland, W.G., "Variance Estimation in Finite Population Sampling," Journal of the American Statistical Association, 73 (1978), 351-58.

[13] Royall, R.M. and Herson, J., "Robust Estimation in Finite Populations I," Journal of the American Statistical Association, 68 (1973), 880-89.

[14] Royall, R.M. and Herson, J., "Robust Estimation in Finite Populations II: Stratification on a Size Variable," Journal of the American Statistical Association, 68 (1973), 890-93.

[15] Scott, A.J., Brewer, K.R.W., and Ho, E.W.H., "Finite Population Sampling and Robust Estimation," Journal of the American Statistical Association, 73 (1978), 359-61.

[16] Singh, R., "Approximately Optimal Stratification on the Auxiliary Variable," Journal of the American Statistical Association, 66 (1971), 829-30.