# OPTIMUM SAMPLE DESIGNS FOR SKEWED POPULATIONS

Patrick M. McCarthy and Robert P. Clickner
APPLIED MANAGEMENT SCIENCES

## 1. INTRODUCTION

Survey sample practitioners are often required to select probability samples from populations in which the variable to be used in the construction of the design is some measure of unit size and has a right-skewed distribution. For example, in sample surveys for electric utility companies, one often uses billed electrical energy use (in kWh) over some period of time as a design variable. Exhibits 1 and 2 display the distribution of kWh for two utility populations, commercial customers of Northeast Utilities in Connecticut and Massachusetts in Exhibit 1 and residential customers of Consolidated Edison Company of New York in Exhibit 2. Both of these distributions are clearly highly skewed. In the Northeast Utilities population, the 75th percentile is nearly four times as far from the 50th percentile as the 25th percentile. Distributions of size-related variables typically are right-skewed. Exhibit 3 displays the distribution of total assets reported on Federal corporate income tax returns (Forms 1120 and 1120S) for 1980. Total assets is a commonly used measure of corporate size, and as Exhibit 3 shows, its distribution is right-skewed.

Survey statisticians typically take either of two approaches when required to sample from a population like those of Exhibits 1, 2 or 3. The first approach consists of sampling with certainty all units above a certain arbitrarily selected cutoff point, and then constructing a stratified design on the remaining units. The stratified design is typically constructed using Dalenius-Hodges cutpoints and Neyman allocation. (See Cochran (1977) for discussions of these techniques.) This technique is employed by the Internal Revenue Service in its annual Corporate Statistics of Income Sample. Currently all corporate returns with total assets above $25,000,000 are sampled with certainty. (See, e.g., Internal Revenue Service (1983) for the details of the Corporate Statistics of Income sample design.) This technique amounts to the creation of a stratum to be sampled with certainty. It has the virtue of yielding zero sampling error in the stratum with the largest unit variance. However, the cutoff between the certainty and noncertainty strata is judgemental and is not based on any optimality criterion.

The second approach frequently employed consists of applying the Dalenius-Hodges and Neyman techniques directly to the complete population. This is a technique that is based entirely on optimality criteria and does not necessarily result in certainty selection for the largest units. It is therefore useful to develop a technique that selects the largest units with certainty and optimally determines the cutpoint between the certainty and noncertainty strata.

## 2. OPTIMIZED CERTAINTY STRATA

The proposed procedure for developing optimized certainty strata under a fixed sample size constraint is iterative. It proceeds as follows. Let N denote the total population size and n denote the fixed total sample size. If the $N_c$ largest units are selected with certainty, then $n_c = N_c$ sample units are "used up" to census the large units. The remaining $n_s = n - N_c$ sample units are then available to take a probability sample of size $n_s$ from the remaining population units $N_s = N - N_c$.

Using this notation, the benefit of the scheme can be examined intuitively by considering the variance of the estimates from the sample. The variance of an estimate from any sample depends on the variance of the population and the size of the sample taken.

For simple random samples from infinite populations, the formula is

$$\sigma^2(N_s)/\ n_s \qquad (1)$$

where $\sigma^2(N_s)$ is the population unit variance among the $N_s$ smallest units. More complex designs have more complex varance formulas. However, they typically contain terms of the form of expression (1).

If one unit from n is used to take the largest unit from N, with certainty, then we have $n_s = n - 1$ and $N_s = N - 1$. This reduction in sample size will increase the sampling variance. That is, $1/n_s$ will increase. However, at the same time, the population to be sampled, now of size $N_s$, has a lower variance than it originally had, because its largest member was removed. Reducing the population variance will reduce sampling variance. This variance reduction often exceeds the variance increase caused by decreasing the sample size, and the net effect is decreased sampling variance.

Thus, if $N_c$ is iteratively increased, and each time $N_c$ is increased the largest member of the remaining $N_s$ is removed, then each subsequent removal will cause a population variance reduction among the remaining $N_s$ units, but this variance reduction will get smaller and smaller. At the same time, each iteration results in a variance increase associated with reducing $n_s$. Eventually, the variance reduction due to increasing $N_c$ will not exceed the variance increase due to reducing $n_s$. At this point, it no longer pays to census additional members.

These intuitive considerations are illustrated with a test population consisting of monthly consumption (kWh/month) for 50 "utility accounts". Exhibit 4 is a histogram of the test population. Some of the basic statistics are:

| | | |
|---|---|---|
| Minimum | = | 400 kWh/month |
| Maximum | = | 60,000 kWh/month |
| Mean | = | 6,692 kWh/month |
| Standard Deviation | = | 12,199 kWh/month |
| Coefficient of Variation | = | 182.3% |

(See ADM Associates (1984) for further details on this test population.) The iterative procedure described above was applied to the test population, starting with $N_c = 1$ and incrementing by 1 until $N_c = 10$. We computed variances for $N_s = 50, 49, \ldots, 40$. The variances for each successive value of $N_c$ are plotted in Exhibit 5. As noted above, once $N_c$ exceeds 4, the reduction in variance diminishes rapidly. Also plotted in Exhibit 5 is $1/n_s$. The plot shows its increase, at an increasing rate. Exhibit 5 shows the inverse relationship between the two factors in the sampling variance that necessitate the trade-offs discussed above.

Next we assumed that available resources limited the sample size for this group to $n = 10$. Using this resource constraint, we evaluated ten possible schemes corresponding to $N_c = 0, 1, \ldots, 9$. We did not analyze $N_c = 10$ because that would be a probability sample of size zero from the $N_s$ of 40, resulting in a biased scheme. Assuming simple random sampling for simplicity, we computed the relative error at 95 percent confidence

$$\varepsilon = (1.96/6{,}692)\ \sigma(N_s)/n_s$$

for each value of $N_c$. Exhibit 6 is a plot of $\varepsilon$ against $N_c$ that shows quite clearly the value of using a certain amount of the study resources to census a number of the largest members of the population. For this illustration, the lowest relative error is encountered when the six largest units are selected with certainty. The relative error is reduced by more than half, from 101 percent to 47.1 percent, for no additional cost, when a portion of the study resources are used to select the largest members with certainty.

Two points are made relative to our analysis thus far. First, while our specific results depend on the specific test population we have created, the general conclusions and approach will be valid for any population which exhibits the extreme skewness we have encountered. This point will be discussed further in Section 4. Second, the simple random sampling assumption

does not alter or distort the results. More complex sampling schemes would produce $\varepsilon$'s lower in absolute terms, but exhibiting the same relative relationships with $N_c$.

## 3. APPLICATIONS

We discuss here an application of the optimized certainty stratum techniques to a real population--the Northeast Utilities population of Exhibit 1. The application is described in detail in McCarthy, et al. (1984). Other applications of the technique to utility customer populations are described in ADM Associates (1984). Because of the size of the population, the iterative procedure was modified slightly and proceeded as follows. The total sample size was fixed at 300 units due to resource constraints. The split between the certainty and probability groups was successively set at 29,000, 25,000, ..., 9,000, 5,000, 4,500, ..., 3,000, and 2,500 kWh/day. All accounts with average daily kWh above these bounds were set aside to be selected with certainty. Their number was deducted from the 300 total sample size. The remainder were allocated across 24 strata defined by state, industrial classification, and average daily kWh, using Neyman allocation and Dalenius-Hodges stratum boundaries.

This procedure was repeated for each of the bounds given above. Exhibits 7 and 8 summarize the results in tabular and graphic form, respectively. From Exhibits 7 and 8, it can be seen that the minimum relative error occurred when 157 of the 300 sampled accounts are selected with certainty. Further, the minimum relative error, 9.32 percent, is less than half the error of selecting no accounts with certainty. As can be seen in Exhibits 7 and 8, the optimum is flat with a range of less than 0.2 percent as the number of certainty accounts range from 133 to 188. For reasons detailed in McCarthy, et al. (1984), the lower end of the range, corresponding to 4,500 kWh/day was selected as the certainty/non-certainty cutpoint. Thus, the 133 largest accounts were selected with certainty.

## 4. THEORETICAL PROPERTIES

Thus far, the optimality claims for the iterative certainty/noncertainty cutpoint determination procedure have been based upon intuitive arguments and empirical evidence. This section presents a theoretical basis for the procedure. To develop the theory, we need some notation. Let $X_1 \leq X_2 \ldots \leq X_N$ be the ordered values for a finite population of size N. The objective is to estimate the population total $T = X_1 + \ldots + X_N$ with the data from a probability sample of size n. Denote the ordered sample values by $x_1 \leq x_2 \leq \ldots \leq x_n$. Assuming a simple random sample, the mean per unit estimator of T is:

$$T' = (N/n)\ (x_1 + \ldots + x_n)$$

which is unbiased and has variance

$$V_o = N\ \sigma^2\ (N-n)/n,$$

where $\sigma^2$ is the population unit variance. If the largest member of the population is selected with certainty then the estimator of T is

$$T'_1 = X_n + [(N-1)/(n-1)]\ (x_1 + \ldots + x_{n-1})$$

which has variance

$$V_1 = (N-1)\ \sigma_1^2\ (N-n)/(n-1)$$

where $\sigma_1^2$ is the population unit variance computed without $X_N$.

Selection of $X_N$ with certainty decreases variance if and only if $V_o - V_1 > 0$. This is equivalent to

$$X_n - (T/N) > \sigma\ [(N(N-2) + n)/nN]^{1/2} \qquad (2)$$

By redefining N to be N - k + 1, n to be n - k + 1, T and $\sigma$ to be computed without $X_{N-k+2}, \ldots, X_N$, we can apply (2) to the

k-th iteration for k = 1, 2, ..., n - 1. At the k-th iteration the decision is made to add or not add the k-th largest unit $X_{N-k+1}$ to the certainty group.

In order to establish conditions under which (2) holds or doesn't hold we will postulate a superpopulation model. That is, we assume the finite population of interest is a random sample from a continuous distribution with density f(x). The ordered finite population values $X_1, \ldots, X_N$ then become the ordered statistics for a simple random sample from f(x). This assumption permits the application of the theory of order statistics. (See David (1970) for an exposition of the theory of order statistics.) We further assume that the finite population $X_1, \ldots X_N$ represents a "typical" sample from the superpopulation, so that each $X_i$ is the expected value of the i-th order statistic for a sample of size N from f(x). For most right-skewed distributions such as the exponential, Pareto, or lognormal, the left-hand side of (1) will considerably exceed the right hand side at the initial iteration. Then as k increases, the LHS decreases and the RHS increases. For typical values of N and n, equality will be achieved before k = n - 1 and a unique optimal breakpoint obtained. For example, for the exponential distribution, the LHS is initially:

$$\sigma(1/2 + 1/3 + \ldots + 1/N) = \sigma V \text{ (say)}$$

which exceeds the RHS unless n is too small. Thus, provided n is large enough, a finite population from an exponential superpopulation will always have an optimal cutpoint between the certainty and noncertainty strata.

## 5. REFERENCES


ADM Associates (1984), Sampling Methodologies for the Commercial Sector, Electric Power Research Institute, EPRI EA3688, Palo Alto, California.

Cochran, W.G. (1977) Sampling Techniques, Third Edition, John Wiley and Sons, New York.

David, H.A. (1970), Order Statistics, John Wiley and Sons, New York.

Internal Revenue Service (1983), Statistics of Income - 1980, Corporation Income Tax Returns, U.S. Government Printing office, Washington, DC.

McCarthy, P.M., Clickner, R.P., Zabek, E., and Goksel, H. (1984), End Use Survey for Northeast Utilities Offices, Final Report, Applied Management Sciences, Silver Spring, Maryland

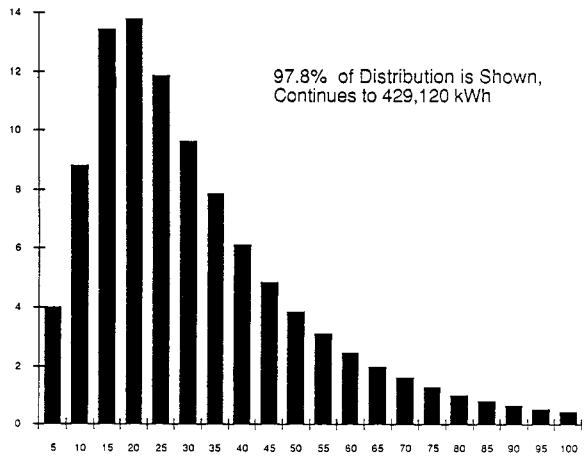
EXHIBIT 1: DISTRIBUTION OF CONSUMPTION FOR NORTHEAST UTILITIES COMMERCIAL SECTOR (1983)

| Percentile | Average Daily kWh |
|---|---|
| Minimum | .003 |
| 5% | 1.17 |
| 25% | 5.83 |
| 50% | 20.3 |
| 75% | 70.9 |
| 95% | 500.2 |
| Maximum | 32,846 |
| Population Size | 21,014 |

EXHIBIT 3: DISTRIBUTION OF TOTAL ASSETS REPORTED ON FEDERAL CORPORATE INCOME TAX RETURNS FOR 1980

| Asset Range (000) | Number of Returns (000) | Percent of Total |
|---|---|---|
| $0 under $100 | 1,504 | 55.9 |
| $100 under $250 | 495 | 18.4 |
| $250 under $500 | 285 | 10.6 |
| $500 under $1,000 | 184 | 6.8 |
| $1,000 under $5,000 | 167 | 6.2 |
| $5,000 under $10,000 | 22 | 0.8 |
| $10,000 under $25,000 | 16 | 0.6 |
| $25,000 under $100,000 | 12 | 0.4 |
| $100,000 or more | 6 | 0.2 |
| Total | 2,689 | 100.0 |

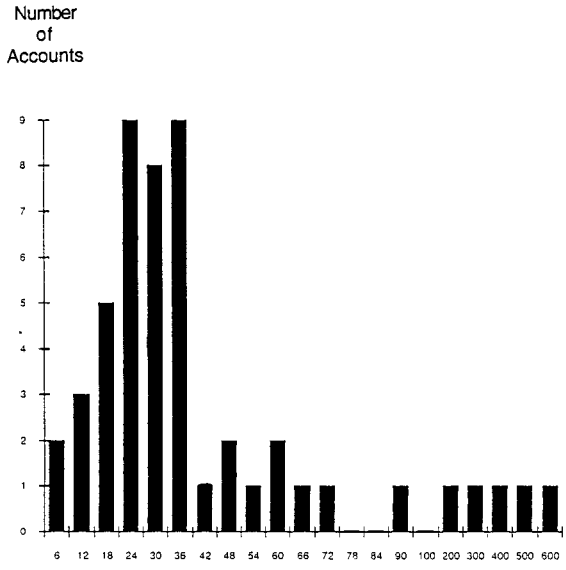
EXHIBIT 2: DISTRIBUTION OF CONSUMPTION FOR CON EDISON RESIDENTIAL SECTOR (1984), TRUNCATED AT 10,000 kWh




Total Annual kWh (x100)

Population Size = 2,460,901 Customers

EXHIBIT 4: HISTOGRAM OF TEST POPULATION




AVG. kWh/Mo. (x100)

Note: Change in Scale at 9,000 kWh

EXHIBIT 5: VARIANCE $\sigma^2(N_S)$ OF kWh/MONTH IN NONCERTAINTY PORTION OF TEST POPULATION AND $1/n_S$ VERSUS NUMBER OF UNITS SELECTED WITH CERTAINTY ($N_C$)

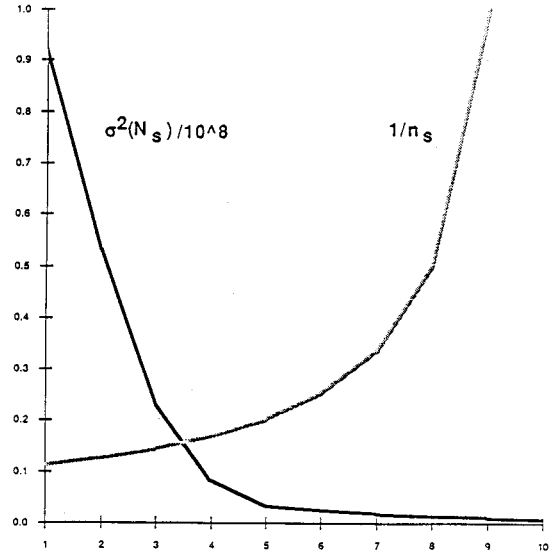



EXHIBIT 6: PLOT OF RELATIVE ERROR IN PERCENT AGAINST $N_C$ FOR TEST POPULATION

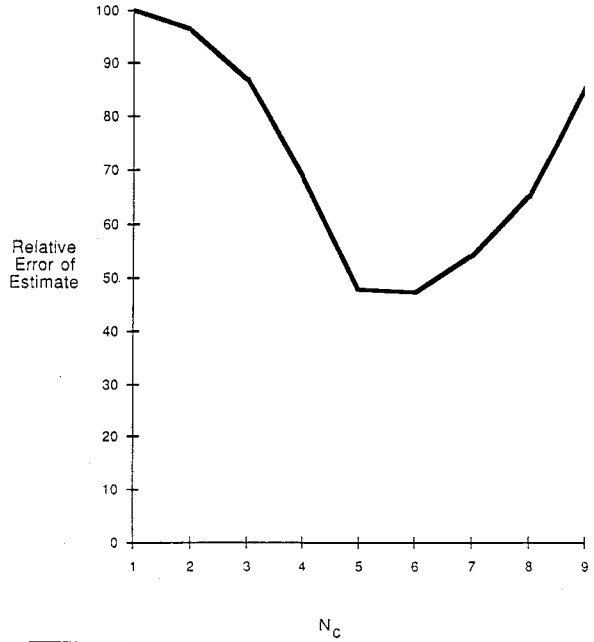



EXHIBIT 7: RELATIVE ERROR ε AS A FUNCTION OF $N_C$ (TOTAL SAMPLE SIZE = 300), NORTHEAST UTILITIES POPULATION

| $N_C$ | Certainty/ Probability Breakpoint | ε At 95 Percent Confidence |
|---|---|---|
| 1 | 27,113 | 20.0% |
| 6 | 24,425 | 18.3 |
| 13 | 19,902 | 16.3 |
| 20 | 16,881 | 14.8 |
| 29 | 12,864 | 13.4 |
| 52 | 8,864 | 11.6 |
| 118 | 4,941 | 9.56 |
| 133 | 4,468 | 9.40 |
| 157 | 3,995 | 9.32 |
| 188 | 3,493 | 9.45 |
| 229 | 2,980 | 10.36 |
| 272 | 2,499 | 14.5 |

EXHIBIT 8: RELATIVE ERROR IN THE ESTIMATOR OF MEAN AVERAGE DAILY kWh VERSUS NUMBER OF ACCOUNTS SAMPLED WITH CERTAINTY, NORTHEAST UTILITIES POPULATION