Survey Methodology, December 1996
Vol. 22, No. 2, pp. 139-146
Statistics Canada

A Transformation Method for Finite Population Sampling Calibrated With Empirical Likelihood

GEMAI CHEN and JIAHUA CHEN[1]

ABSTRACT

In this paper, we study a confidence interval estimation method for a finite population average when some auxiliary information is available. As demonstrated by Royall and Cumberland in a series of empirical studies, naive use of existing methods to construct confidence intervals for population averages may result in very poor conditional coverage probabilities, conditional on the sample mean of the covariate. When this happens, we propose to transform the data to improve the precision of the normal approximation. The transformed data are then used to make inference on the original population average, and the auxiliary information is incorporated into the inference directly, or by calibration with empirical likelihood. Our approach is design-based. We apply our approach to six real populations and find that when transformation is needed, our approach performs well compared to the usual regression method.

KEY WORDS: Finite population; Sampling; Confidence interval; Transformation; Empirical likelihood.

1. INTRODUCTION

Let (x_i, y_i), $i = 1, 2, ..., N$ be values associated with N units in a finite population. For unit i, y_i is the variable of interest and x_i is an auxiliary variable. One of the most extensively studied finite population problems is the estimation of the population average $\bar{y} = (y_1 + ... + y_N)/N$ (or total $N\bar{y}$) under various sampling schemes. We shall focus on the simple random sampling scheme in this paper, because the nature of the problems we want to study can be better seen from this scheme and the results obtained here can be easily generalized into other sampling schemes of which the simple random sampling scheme is the building block.

It is often true that some information about the auxiliary variable x is known and can be used to make inference about $\bar{y}$. For example, let $S = \{1, ..., i, ..., N\}$ and let $s \subset S$ be a simple random sample of size n. When $\bar{x} = (x_1 + ... + x_N)/N$ is known, and x and y are correlated, the population average $\bar{y}$ can be estimated by the ratio estimator $\hat{\bar{y}} = (\bar{y}_s/\bar{x}_s)\bar{x}$, or by the regression estimator $\hat{\bar{y}} = \bar{y}_s + b(\bar{x} - \bar{x}_s)$, where $\bar{x}_s$ and $\bar{y}_s$ are the sample averages of x and y, respectively, and $b = \sum(x_i - \bar{x}_s)(y_i - \bar{y}_s)/\sum(x_i - \bar{x}_s)^2$.

Under very general conditions, both the ratio estimator and the regression estimator are asymptotically normal; see Scott and Wu (1981), Bickel and Freedman (1984), and Theorem 2.1 of Section 2. Hence, if v is a carefully chosen estimator of the variance of $\hat{\bar{y}}$, the standardized variable $(\hat{\bar{y}} - \bar{y})/\sqrt{v}$ is customarily treated to have the standard normal distribution. Therefore, if z_α denotes the upper α-percentile of the standard normal distribution, then

$$(\hat{\bar{y}} - z_\alpha\sqrt{v}, \hat{\bar{y}} + z_\alpha\sqrt{v}) \qquad (1.1)$$

will produce an approximate 100 $(1 - 2\alpha)$% confidence interval for $\bar{y}$.

Confidence interval (1.1) is widely used in practice. However, problems arise when it is applied to certain populations. Royall and Cumberland (1981a, 1981b, 1985) studied the ratio and regression estimators and applied them to six real populations where strong correlations between x and y seemed to exist. (See Section 3 for a summary of the six populations.) Various estimators of the variance of $\hat{\bar{y}}$ were used. It was found that the actual conditional coverage rate of the confidence interval (1.1), conditional on $\bar{x}_s$, depended heavily on the size of $\bar{x}_s$ and were usually much lower than the claimed coverage rate, even with the most adaptive variance estimator. For example, the 95% confidence interval for a population named Counties 70 had a conditional coverage rate 76% with the jackknife variance estimator when $\bar{x}_s$ was small, and the conditional coverage rate could go as low as 50% with other variance estimators.

The above mentioned studies point to the need to construct confidence intervals that "will live up to their name" (Royall and Cumberland 1985, p. 359). However, up to now there has been little progress made in this direction. In this paper, we present some results from studying an alternative procedure for constructing confidence intervals and from applying it to the six populations studied by Royall and Cumberland and many others. As will be shown in Section 3, the conditional coverage rate of our confidence intervals is more accurate.

Two important ideas, namely, transformation and empirical likelihood, are used simultaneously to attack the problems encountered by Royall and Cumberland in particular, and to develop a new procedure in general. As explained in Cochran (1977, p. 150), the preference in sample survey theory is to make, at most, limited assumptions about the frequency

[1] Gemai Chen, Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada, S4S 0A2; Jiahua Chen, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1.

distribution followed by the data in the sample. However, ratio or regression estimator can help obtain increased precision by taking advantage of the correlation between y_i and x_i. This, of course, can be described by some assumption(s), such as an approximate linear relationship between y and x. Although almost no further assumptions are necessary to use the ratio or regression approach, the procedure (1.1) is clearly based on an normal approximation. But as it is well known, the normal approximation can be very poor when the population distribution is severely skewed and the sample size is small. In terms of procedure (1.1), the closer the estimator distribution is to the normal, the better one can construct confidence intervals. If the population distribution is severely skewed, a transformation may produce a population distribution that is at least more symmetric, so that the normal approximation for the estimator is more accurate.

When using the ratio and regression estimators, knowing $\bar{x}$ is crucial to gain improvement over the use of sample mean. In our proposed procedure, the complete information about the auxiliary variable x can be incorporated. But if $\bar{x}$ is the only auxiliary information available, it is difficult to use this information directly when a transformation is involved, because any non-linear transformation obscures the link between $\bar{x}$ and $\bar{y}$. In this second case, we find the method of empirical likelihood very helpful in solving our problem; see particularly Owen (1988, 1990) and Chen and Qin (1992) for references. The empirical likelihood method in this situation can also be regarded as a calibration method as discussed in Deville and Särndal (1992). This approach rescues us from losing information about x after transforming the data.

There have been many discussions on how to use transformations to make better inference on the transformed scale (Box and Cox 1964; Carroll and Ruppert 1988; Calvin and Sedransk 1991, and the references therein). There have also been some studies on how to make inference on the original scale, after a transformation is applied (Carroll and Ruppert 1984; Elliott 1977). What is new with our procedure is the attempt to link the above two steps by combining transformation with auxiliary information and/or by applying empirical likelihood method when necessary.

The details of our procedure are given in Section 2. Then our procedure is applied to the six populations studied by Royall and Cumberland in Section 3. The validity of our procedure in an arbitrary setting is demonstrated in Section 4 and some comments are made at the end of the paper.

2. THE NEW PROCEDURE

As mentioned in the last section, a problem with the confidence interval (1.1) is that it will fail if the distribution of $(\hat{\bar{y}} - \bar{y})/\sqrt{v}$ is severely asymmetric and far from the normal distribution. The problem can be inherited from the skewness of the population distribution. When the skewness is severe, a central confidence interval procedure like (1.1) is doomed to fail. The basic model employed by Royall and Cumberland (1981a, 1981b, 1985) is

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{2.1}$$

with $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$ and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, for $i \neq j$. It is easy to find that for the six real populations studied by Royall and Cumberland, the corresponding error distributions are very skewed. These observations lead us to consider transforming the variables y and/or x, and consider the model

$$h(y_i) = \alpha + \beta g(x_i) + \sigma \epsilon_i, \tag{2.2}$$

where $h(\cdot)$ and $g(\cdot)$ are two monotone functions. There are many families of transformations suggested in the literature. One commonly used family is the Box-Cox power transformation family defined by

$$f(x,\lambda) = \begin{cases} (x^\lambda - 1)/\lambda & \text{when } \lambda \neq 0, \\ \log(x) & \text{when } \lambda = 0. \end{cases}$$

Model (2.1) is a special case of (2.2) when both h and g equal $f(x, 1)$.

The choice of transformations in model (2.2) might be suggested by an examination of the sample x's and y's based on a possible model relationship, or by our subject knowledge about the population under investigation. For example, for the six populations discussed in Royall and Cumberland, the population distributions are severely skewed towards the right which can be learned from the nature of the finite populations. Therefore, a log transformation may make them all less skewed. Other more objective methods of choosing transformations are discussed in Section 4.

We emphasize that models (2.1) and (2.2) are used here to motivate transformations, point estimators, or confidence interval procedures. Our study of conditional coverage rates will, however, be based on the probability measure generated by the design, as in Royall and Cumberland (1985). For this purpose, we embed our finite population in a sequence of populations indexed by k. This means that a sub-index k is needed to write $N = N_k$ and $n = n_k$, *etc.*, but for simplicity, we will suppress the index k if there is no possibility for confusion.

Let $v_i = h(y_i)$, $u_i = g(x_i)$, $\bar{v}_N = N^{-1}\sum_{i=1}^N v_i$ and $\bar{u}_N = N^{-1}\sum_{i=1}^N u_i$. Define

$$\beta_N = \frac{\sum_{i=1}^N (u_i - \bar{u}_N) v_i}{\sum_{i=1}^N (u_i - \bar{u}_N)^2},$$

$$\alpha_N = \bar{v}_N - \beta_N \bar{u}_N,$$

$$e_i = v_i - (\alpha_N + \beta_N u_i),$$

$$\sigma_N^2 = \frac{1}{N-1}\sum_{i=1}^N e_i^2.$$

Suppose $s \subset S$ is a simple random sample of size n. We similarly define

$$\hat{\beta} = \frac{\sum_{i\in s}(u_i - \bar{u}_s)v_i}{\sum_{i\in s}(u_i - \bar{u}_s)^2},$$

$$\hat{\alpha} = \bar{v}_s - \hat{\beta}\,\bar{u}_s,$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i\in s}(v_i - \hat{\alpha} - \hat{\beta}u_i)^2,$$

where $\bar{u}_s$ and $\bar{v}_s$ are the sample averages.

Denote the inverse function of $h(\cdot)$ by $h^{-1}(\cdot)$. Then the fitted value of y_i is

$$\hat{y}_i = h^{-1}(\hat{\alpha} + \hat{\beta}\,u_i). \tag{2.3}$$

We discuss confidence interval estimation of $\bar{y}$ in two cases. In the first case where all x_i ($i = 1, ..., N$) are known, a natural estimator of $\bar{y}$ is $(\sum_{i\in s} y_i + \sum_{i\notin s}\hat{y}_i)/N$. However, for the purpose of constructing confidence intervals for $\bar{y}$, we study the distribution of

$$\hat{\bar{y}}(\hat{\alpha},\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i = \int_{-\infty}^{\infty} h^{-1}(\hat{\alpha} + \hat{\beta}u)\,dF_N(u) \tag{2.4}$$

instead, where $F_N(u)$ is the empirical distribution function of the u_i ($i = 1, ..., N$). Clearly, the distribution of $\hat{\bar{y}}(\hat{\alpha},\hat{\beta})$ is determined by the distribution of $(\hat{\alpha},\hat{\beta})$ which is descibed in the following design-based theorem.

Theorem 2.1 Suppose that when $k \to \infty$, both $n = n_k$ and $N - n = N_k - n_k$ go to ∞ and

1. $\bar{u} = \lim_{k\to\infty} N^{-1}\sum_{i=1}^{N} u_i$ exists.
2. $N^{-1}\sum_{i=1}^{N} u_i^4 = O(1)$.
3. $\sigma_u^2 = \lim_{k\to\infty}\sigma_{u,N}^2 = \lim_{k\to\infty}(N-1)^{-1}\sum_{i=1}^{N}(u_i - \bar{u}_N)^2$ exists and is greater than zero.
4. $\sigma^2 = \lim_{k\to\infty}\sigma_N^2 = \lim_{k\to\infty}(N-1)^{-1}\sum_{i=1}^{N} e_i^2$ exists and is greater than zero.
5. $N^{-1}\sum_{i=1}^{N}|e_i|^3 = O(1)$, $N^{-1}\sum_{i=1}^{N}|(u_i - \bar{u}_N)e_i|^3 = O(1)$.
6. $r = \lim_{k\to\infty}(\sigma_{u,N}^2\sigma_N^2)^{-1}N^{-1}\sum_{i=1}^{N}(u_i - \bar{u}_N)^2 e_i^2$ exists and is greater than zero.
7. $f = \lim_{k\to\infty} n/N$ exists and is less than 1.

Then

(1) $\sqrt{n}(\hat{\alpha} - \alpha_N, \hat{\beta} - \beta_N)'$ converges in distribution to the bivariate normal distribution $N_2(0,\Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 + \frac{\bar{u}^2}{\sigma_u^2}r & -\frac{\bar{u}}{\sigma_u^2}r \\ -\frac{\bar{u}}{\sigma_u^2}r & \frac{1}{\sigma_u^2}r \end{pmatrix}(1-f)\,\sigma^2.$$

(2) Let B_n be any joint $100(1-\gamma)\%$ confidence region for (α_N, β_N) and define G_n by

$$G_n = \{\hat{\bar{y}}(\alpha,\beta) : (\alpha,\beta) \in B_n\}, \tag{2.5}$$

then,

$$\text{Prob}\{\bar{y}(\alpha_N, \beta_N) \in G_n\} \geq 1 - \gamma,$$

where $\bar{y}(\alpha_N, \beta_N) = \sum_{i=1}^{N} h^{-1}(\alpha_N + \beta_N u_i)/N$.

The proof is deferred to the Appendix.

We note that without underlying normality on the errors, it is not easy to get an exact confidence region B_n for (α_N, β_N) for a specified confidence level $1 - \gamma$. The B_n used in the following discussion and the expressions built upon it are, therefore, approximate.

Theorem 2.1 allows us to construct confidence intervals for $\bar{y}(\alpha_N,\beta_N)$, but $\bar{y}(\alpha_N,\beta_N)$ is not equal to $\bar{y}$ in general. This is an intrinsic problem as long as a non-linear transformation is used. If only a point estimator is needed, we would use the regression estimator currently, and we suggest that the methodology developed in this paper be used for interval estimation. Bias corrections for $\hat{\bar{y}}(\hat{\alpha},\hat{\beta})$ are, however, possible, and a specific one is used in our simulation study. Work on general corrections is under study.

According to Theorem 2.1, G_n is a conservative confidence interval for $\bar{y}(\alpha_N,\beta_N)$, which can also be regarded as an approximate confidence interval for $\bar{y}$. To improve the coverage rate of G_n, observe that the contours of $\hat{\bar{y}}(\alpha,\beta)$ in a small neighborhood of $O = (\hat{\alpha},\hat{\beta})$ are approximately parallel straight lines on the $\alpha\beta$ plane; see Figure 1. Let (a, b) be the

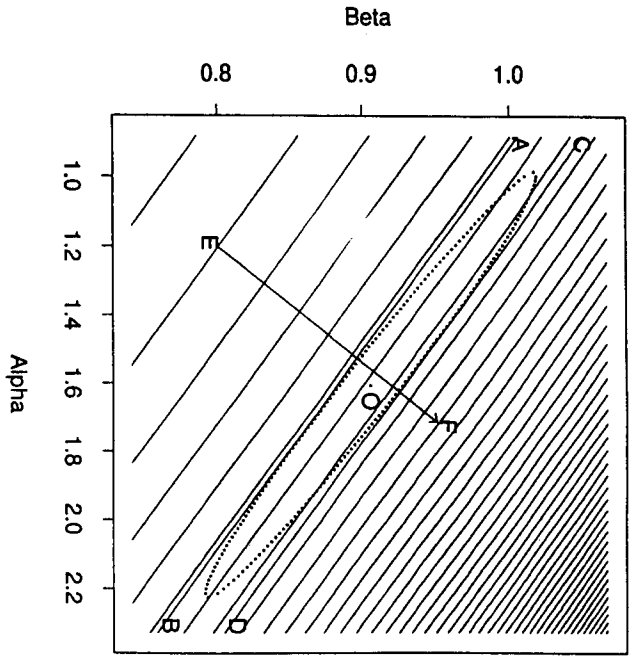

Figure 1. Contour plot of the bi-variate function $\hat{\bar{y}}(\alpha,\beta)$ in the neighbourhood of $O = (\hat{\alpha},\hat{\beta})$, based on a random sample of size 32 taken from population Cancer

directional cosines of the direction $\vec{EF}$ along which the contours increase. Then $\hat{\bar{y}}(\alpha,\beta)$ is (approximately) a monotone function of $T_n = a(\alpha - \hat{\alpha}) + b(\beta - \hat{\beta})$, where T_n is the corresponding change along the direction $\vec{EF}$ to the changes in α and β. A natural choice of B_n is

$$B_n = \{(\alpha,\beta) : |a(\alpha - \hat{\alpha}) + b(\beta - \hat{\beta})| \le c\hat{\sigma}t(\gamma/2; n-2)\},$$

where $c^2 = \mathrm{Var}(T_n)/\sigma^2$, $\mathrm{Var}(T_n)$ is the variance of T_n, and $t(\gamma/2; n-2)$ is the upper $\gamma/2$-percentile of the t distribution with $n-2$ degrees of freedom. This B_n is the region between two parallel straight lines AB and CD in Figure 1.

A drawback of the above B_n is that it is an *unbounded* region. If the contours of $\hat{\bar{y}}(\alpha,\beta)$ are not close to be parallel and/or straight, this B_n will lead to very conservative confidence intervals. To guard against this possibility, we construct a bounded elliptic region C_n defined by those (α, β) that satisfy

$$\left\{ n(\alpha - \hat{\alpha})^2 + 2n\bar{u}_s(\alpha - \hat{\alpha})(\beta - \hat{\beta}) + n\left(\bar{u}_s^2 + r_s^{-1}\frac{\sum_{i\in s}(u_i - \bar{u}_s)^2}{n-1} \right)(\beta - \hat{\beta})^2 \right\} \le \left(1 - \frac{n}{N}\right)\hat{\sigma}^2 t^2(\gamma/2; n-2),$$

where $(1 - n/N)$ is part of the variances of $\hat{\alpha}$ and $\hat{\beta}$, because we are doing sampling without replacement from a finite population, and

$$r_s = \frac{n^{-1}\sum_{i\in s}(u_i - \bar{u}_N)^2(v_i - \hat{\alpha} - \hat{\beta}u_i)^2}{\left\{n^{-1}\sum_{i\in s}(u_i - \bar{u}_N)^2\right\}\left\{(n-2)^{-1}\sum_{i\in s}(v_i - \hat{\alpha} - \hat{\beta}u_i)^2\right\}} \tag{2.6}$$

is a sample estimate of the quantity r in Theorem 2.1. The C_n thus defined is represented by the region inside the ellipse in Figure 1 and has the property that it touches both boundary lines of B_n regardless of the direction (a, b). Therefore, when $\hat{\bar{y}}(\alpha,\beta)$ is indeed a monotone function of T_n, C_n produces the same confidence interval for $\bar{y}$ as B_n does. However, C_n is less vulnerable than B_n if the contours of $\hat{\bar{y}}(\alpha,\beta)$ are not close to be parallel and/or straight, because C_n shrinks to one point as n increases. A confidence interval for $\bar{y}$ corresponding to C_n is defined as

$$I_n = \{\hat{\bar{y}}(\alpha,\beta) : (\alpha,\beta) \in C_n\}. \tag{2.7}$$

As the error distributions are more symmetric after the transformation, the new confidence interval based on C_n is therefore expected to be better than the confidence interval without transformation. Note that since all x_i are known, other approaches, such as optimal stratification and post-stratification, may be better. However, optimal stratification may not be possible in some cases as discussed in Cochran (1977, p. 134). Also research is needed on the use of post-stratification when the error distributions are severely skewed.

We now turn to the discussion of the second case where $\bar{x} = (x_1 + \ldots + x_N)/N$ is known, but x_i, $i = 1, \ldots, N$, are unknown. If we want to proceed as in the first case, one approach is to estimate $F_N(u)$ and somehow make use of the information in $\bar{x}$. The following empirical likelihood methodology is found to be an effective way of doing this. We outline the main ideas here; the interested reader should consult Owen (1988, 1990) and Chen and Qin (1992) for more details. The key idea is to maximize the (empirical) likelihood functions under various restrictions formed by the knowledge about some aspects of the parameters. For example, in our problem, the knowledge is $\bar{x}$. It is shown by Chen and Qin (1992) that the resulting estimators with the presence of restrictions are asymptotically more efficient than those without restrictions.

Specifically, we estimate $F_N(u)$ in (2.4) by

$$\hat{F}_N(u) = \sum_{i\in s} p_i I[u_i \le u], \tag{2.8}$$

where the p_i are chosen by maximizing

$$\prod_{i\in s} p_i \tag{2.9}$$

subject to

$$p_i \ge 0, \quad \sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i x_i = \bar{x}. \tag{2.10}$$

If y_i, $i \in s$ are regarded as realizations of the random variables Y_i, $i \in s$, with distribution function F, the p_i in (2.9) can be defined by $p_i = F(Y_i) - F(Y_i-)$, and (2.9) is called the empirical likelihood function in Owen (1990).

Deville and Särndal (1992) look at the above approach from a calibration point of view. They suggest using unequal weights for different units sampled to reflect their different contributions, while keeping $\sum p_i x_i = \bar{x}$. It is believed that if these weights give a perfect estimate of $\bar{x}$, they should also be good for estimating $\bar{y}$.

The solution to (2.9) and (2.10) will not exist if either the minimum x value in a sample is greater than or equal to $\bar{x}$, or the maximum x value in a sample is less than or equal to $\bar{x}$. When this happens, one remedy is to replace (2.9) with

$$\sum_{i\in s}(np_i - 1)^2, \tag{2.11}$$

subject to a milder constraint

$$\sum_{i\in s} p_i = 1, \quad \sum_{i\in s} p_i x_i = \bar{x}. \tag{2.12}$$

Under (2.11) and (2.12), we have

$$p_i = \frac{1}{n} + (\bar{x} - \bar{x}_s)(x_i - \bar{x}_s)\Big/\sum_{i\in s}(x_i - \bar{x}_s)^2, \tag{2.13}$$

which always exists unless all the x_i in the sample are the same. The latter situation corresponds to the lack of a covariate, which implies $p_i = n^{-1}$ if $\bar{x} = x_i$, or the solution does not exist if $\bar{x} \neq x_i$. The function given in (2.11) is called the Euclidean likelihood, which is asymptotically equivalent to the empirical likelihood (2.9) (Owen 1990).

For our simulation study in Section 3, we suggest a bias correction to be used in our computation. If $h(w) = g(w) = \log(w)$, we suggest a corrected estimator of $\bar{y}$ as

$$\hat{\bar{y}}^*(\hat{\alpha},\hat{\beta}) = \int_{-\infty}^{\infty} \exp\left\{\hat{\alpha} + \hat{\beta} u_i + \frac{1}{2}\hat{\sigma}^2\right\} F_N(u), \qquad (2.14)$$

if all u_i, $i = 1, ..., N$ are known, and replace $F_N(u)$ by $\hat{F}_N(u)$ and $\bar{u}_N$ in (2.6) by $\bar{u}_s$ when only $\bar{x}$ is known. This correction is motivated by model-based considerations under a normality assumption. Correspondingly, I_n of (2.7) is corrected as

$$I_n^* = \{\hat{\bar{y}}^*(\alpha,\beta) : (\alpha,\beta) \in C_n\}. \qquad (2.15)$$

When other power transformations are used, similar corrections can be made using the results in Pankratz and Dudley (1987).

3. APPLICATION TO SIX REAL POPULATIONS

The six real populations studied by Royall and Cumberland (1981a, 1981b, 1985) are summarized in Table 1. Attention was given to the variety in the type of data (demographic, economic, *etc.*), and in the logical relationship between the x and y variables, when these populations were chosen. Note that we have added 1 to the y values in population Cancer in order to take the log transformation.

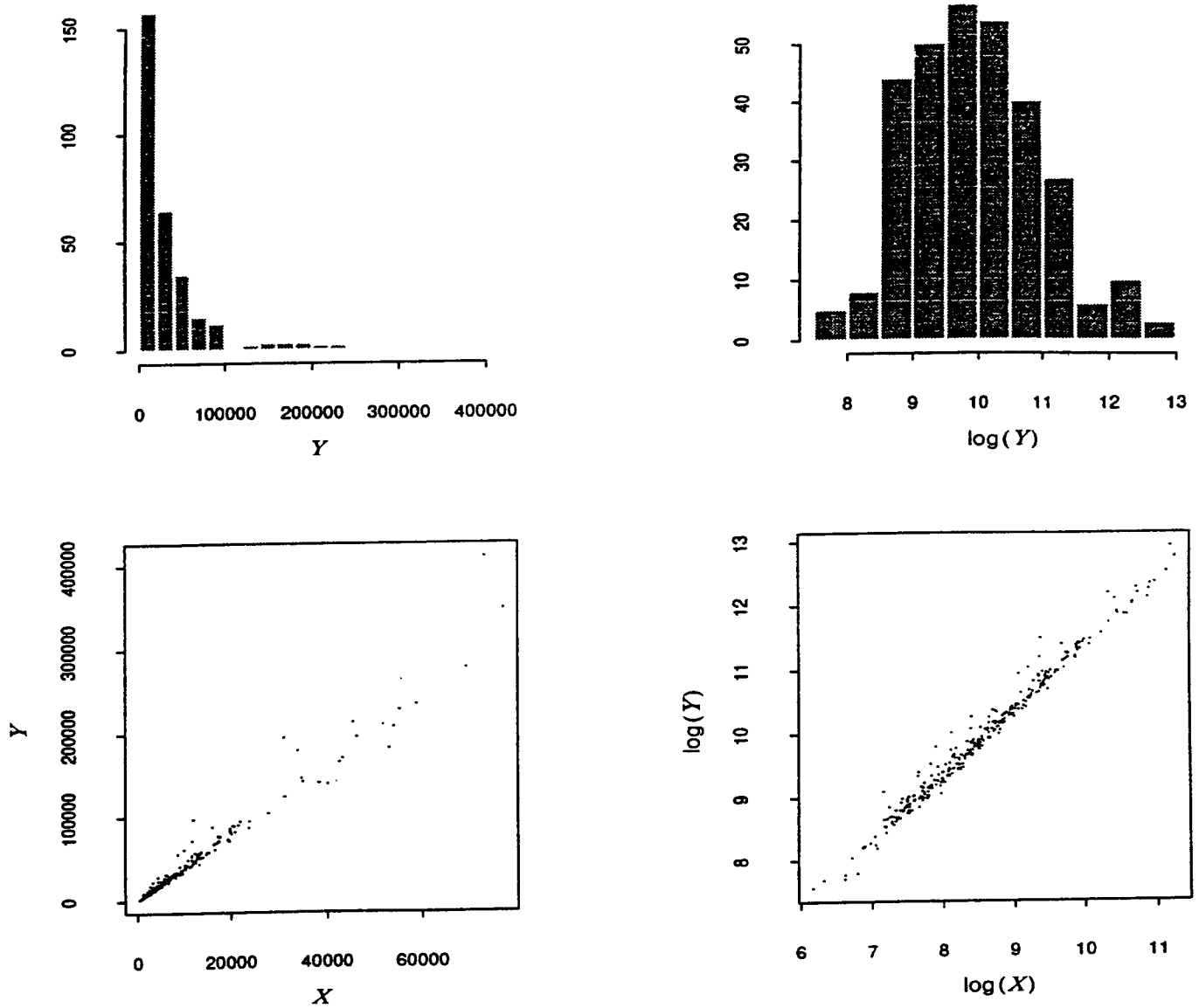

Figure 2. Histograms and scatter plots for the population Counties 70 before and after taking the log transformation

Table 1
Summaries of the Six Populations

Population	N	$\bar{x}$	$\bar{y}$	$\rho(x, y)$	$\rho(\log(x), \log(y))$
Cancer	301	1.1288×10^4	4.0847×10^1	0.967	0.948
Cities	125	2.6602×10^5	2.8553×10^5	0.947	0.953
Counties 60	304	8.9312×10^3	3.2916×10^4	0.998	0.998
Counties 70	304	8.9312×10^3	3.6984×10^4	0.982	0.991
Hospitals	393	2.7470×10^2	8.1465×10^2	0.911	0.943
Sales	331	2.3164×10^9	2.4078×10^9	0.997	0.985

The Counties 70 data are plotted in Figure 2. The histogram of y clearly indicates that the population distribution is severely skewed, while the same plot for $\log(y)$ shows a substantial improvement. Also, the scatter plot of $\log(y)$ *vs.* $\log(x)$ shows a better linear relationship than the scatter plot of y *vs.* x. The need and the benefit of taking transformation is therefore obvious. Similar comments can also be made for populations Cities, Counties 60 and Hospitals. For populations Cancer and Sales, the log transformation (or any other power transformations) seem to weaken the linear relationship that exists between x and y.

Now, we illustrate our new procedure by assuming $h = g = \log$ in (2.2). Equations (2.9) to (2.15) are used to perform the calculations. As in Royall and Cumberland (1981b, 1985), for each of the six populations, we take a simple random sample of size 32 and calculate $\bar{x}_s$, $\hat{\bar{y}}^*(\hat{\alpha}, \hat{\beta})$ and construct a 95% confidence interval I_{32}^*. We repeat this process 10,000 times for each population. The results are reported in Table 2 under the title "Transformation Method" when all x values are known, and under the title "Empirical Likelihood Method" when only $\bar{x}$ is known. The term ratio denotes the average length of the confidence intervals divided by the root mean square error for each population. The non-coverage rate (Ncr) is the proportion of intervals that fail to contain the population average $\bar{y}$. The quantities under the titles "Regression Method (regression variance)" and "Regression Method (jackknife variance)" are obtained using the same method of Royall and Cumberland (1981b) when the usual regression variance and the jackknife variance of $\hat{\bar{y}}$ are used, respectively, but for 10,000 random samples instead of the original 1,000 samples. The results under "Empirical Likelihood Method (created population)" are to be explained in the next Section.

Next, we follow Royall and Cumberland to make *design based* inference and to study the *conditional* coverage properties of several interval estimation procedures. Specifically, we divide the confidence intervals into 20 groups according to the size of $\bar{x}_s$, and plot the proportions of intervals in each group that fail to contain the population average $\bar{y}$. For each specific group, the proportion of those intervals that lay above (below) $\bar{y}$ is plotted above (below) the horizontal line. Figure 3 contains such plots for the Counties 70 data. The top two plots show the non-coverage rates of the regression method using the usual regression variance and the jackknife variance for $\hat{\bar{y}}$; the middle two plots show the non-coverage rates of our new procedure. The bottom left plot will be explained in Section 4. As can be seen clearly, our new procedure with a log transformation produces substantial improvement. For populations Cities, Counties 60 and Hospitals, our new procedure also produces some improvement (plots are not shown here). For populations Cancer and Sales, the new procedure produces very conservative results. This is likely due to the fact that the log transformation (or any power transformation) actually weakens the linear relationship between x and y.

Table 2
Simulation results based on 10,000 simple random samples of size 32

	Cancer	Cities	Counties 60	Counties 70	Hospitals	Sales
Regression Method (regression variance)						
Ratio	3.26	3.65	3.05	2.90	3.62	2.94
Ncr	0.141	0.116	0.146	0.271	0.098	0.176
Regression Method (jackknife variance)						
Ratio	4.03	3.88	4.03	3.57	3.93	3.95
Ncr	0.081	0.102	0.083	0.192	0.068	0.079
Transformation Method (all x values are known)						
Ratio	5.08	4.00	3.75	3.76	4.04	5.41
Ncr	0.018	0.074	0.053	0.069	0.042	0.001
Empirical Likelihood Method (only $\bar{x}$ is known)						
Ratio	5.12	3.74	3.37	3.69	4.15	4.90
Ncr	0.017	0.082	0.081	0.082	0.037	0.006
Empirical Likelihood Method (created population)						
Ratio	3.92	3.92	3.97	3.96	3.90	3.99
Ncr	0.057	0.059	0.055	0.058	0.059	0.059

We have also performed simulations for sample sizes 16 and 64, and/or for target coverage rate 90%. The results are very similar to what we have presented.

4. DISCUSSION

We use the log transformation in some of our discussions because it is perhaps the most frequently used transformation in practice. Nevertheless, there exist more objective methods to select transformations. One such a method is the well known Box-Cox power transformation which we have mentioned; see Box and Cox (1964), Box and Tidwell (1962), Carroll and Ruppert (1988). Another recent method is based on a procedure called alternating conditional expectation (ACE) (Breiman and Friedman 1985, De Veaux and Steele 1989).

There are other possibilities to improve conditional coverage rate. One such a possibility is to employ asymmetrical error distributions such as the inverse Gaussian family (Whitmore 1983). Another possibility is to adopt quasi-likelihood (Nelder and Pregibon 1987) to finite population problems.

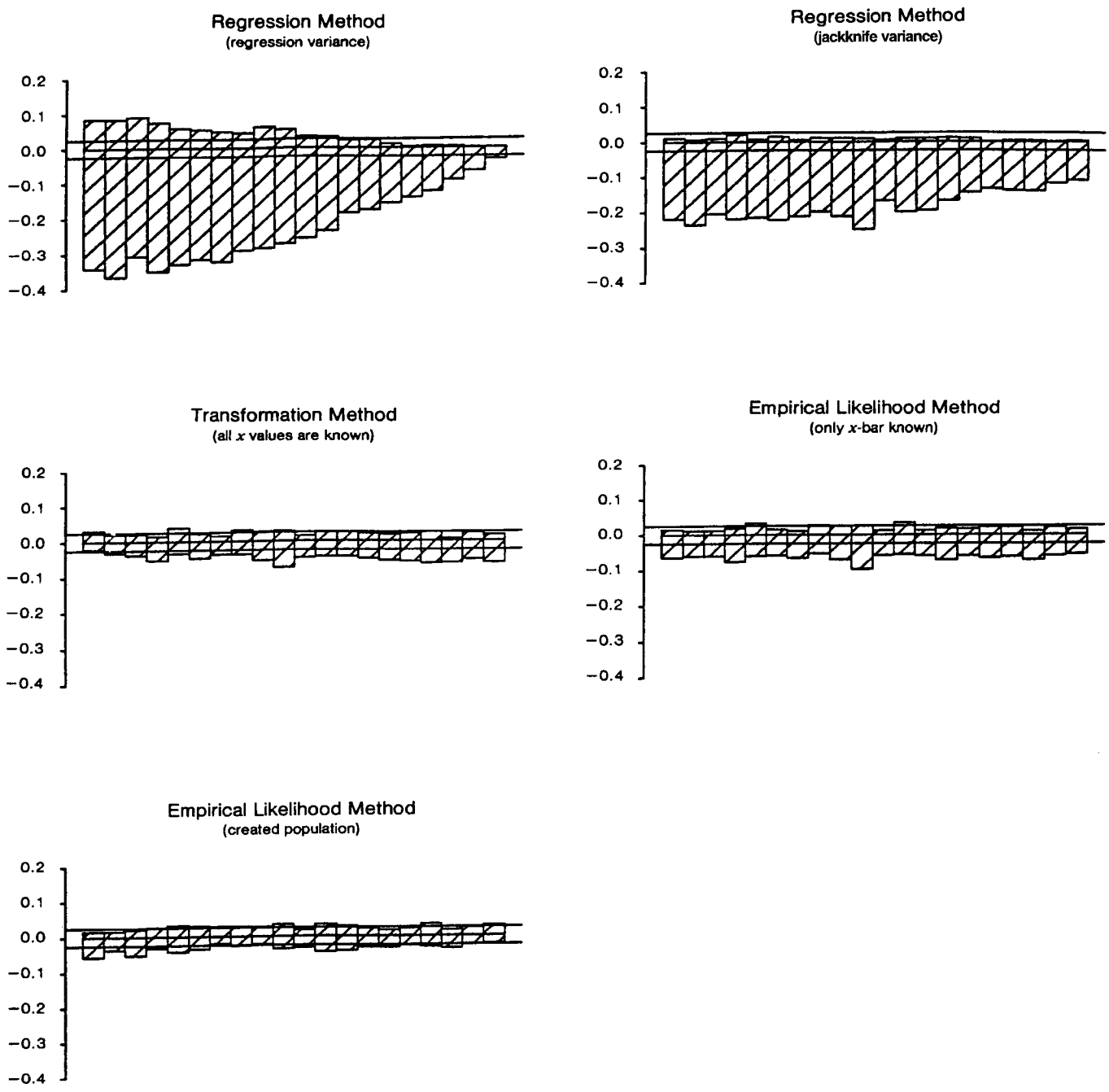

Figure 3. Plots of conditional non-coverage rates for the population Counties 70 based on 10,000 simple random samples of size 32. Reference lines are drawn at 2.5% and the expected non-coverage rate is 5%

The validity of our new procedure is also demonstrated in the following simulation study. For each of the six real populations, we create a new population by replacing the original y_i values with

$$y_i^* = \exp\{\hat{\alpha} + \hat{\beta}\log(x_i)\,\hat{\sigma}\,\epsilon_i\},$$

where $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}$ are the parameter estimates from fitting model (2.2) with $h = g =$ log to the old population, and ϵ_i are generated as i.i.d. standard normal variates. Using the six created populations which are fixed, we repeat the simulations as in Section 3 for the case where only $\bar{x}$ is known. Table 2 contains the summary of this simulation study, and the non-coverage plot for the Counties 70 data is shown at the bottom left corner of Figure 3. (Non-coverage plots for other populations look very similar to this plot.) It is clear from this study that when the finite population is generated from a super-population model like (2.2) with a normal error distribution, our new procedure gives the correct conditional coverage rates. Furthermore, we decrease the correlation between

x and y to as low as 0.5 for each of the six populations by increasing $\hat{\sigma}$ and repeat the above simulations. The results are as good as those shown in Table 2 and Figure 3.

Although only the simple random sampling scheme is considered in this paper, the proposed procedure is applicable as long as (i) there is a linear correlation between $h(y)$ and $g(x)$ for some monotone functions h and g, and (ii) either $F_N(u)$ or $\hat{F}_N(u)$ can be found. Since the six populations studied here are carefully chosen to be representative, our new procedure is expected to be useful to study other finite populations.

ACKNOWLEDGEMENTS

We would like to thank the referee and the Editor for their comments which greatly improved the presentation. Both authors are supported by grants from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

Proof of Theorem 2.1 (1). For any given real numbers t_1 and t_2, we have

$$t_1(\hat{\alpha} - \alpha_N) + t_2(\hat{\beta} - \beta_N) =$$

$$t_1 n^{-1} \sum_{i \in s} e_i + \frac{t_2 - t_1 \bar{u}_s}{\sum_{i \in s}(u_i - \bar{u}_s)^2} \sum_{i \in s} (u_i - \bar{u}_s) e_i.$$

From Conditions 1, 2 and 3, we have

$$\bar{u}_s \to \bar{u}, \quad n^{-1} \sum_{i \in s} (u_i - \bar{u}_s)^2 \to \sigma_u^2.$$

Therefore, we can write

$$t_1(\hat{\alpha} - \alpha_N) + t_2(\hat{\beta} - \beta_N) =$$

$$t_1 n^{-1} \sum_{i \in s} e_i + \frac{t_2 - t_1 \bar{u}}{\sigma_u^2} n^{-1} \sum_{i \in s} (u_i - \bar{u}) e_i + o_p(n^{-1/2}).$$

The Lindeberg-Hájek condition is satisfied for $t_1 e_i + t_2 - t_1 \bar{u}/\sigma_u^2 (u_i - \bar{u}) e_i$ under the moment condition 5, see Hájek (1960), Scott and Wu (1981) and Bickel and Freedman (1984). Together with Conditions 4, 6 and 7, the desired result follows by using the Cramér-Wold device.

Proof of Theorem 2.1 (2). Because there may be other values $(\alpha', \beta') \notin B_n$ for which $\hat{\bar{y}}(\alpha', \beta') = \hat{\bar{y}}(\alpha, \beta)$ for some $(\alpha, \beta) \in B_n, G_n$ is always conservative.

REFERENCES

BICKEL, P.J., and FREEDMAN, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics,* 12, 470-482.

BOX, G.E.P., and TIDWELL, P.W. (1962). Transformations of the independent variables. *Technometrics*, 4, 531-550.

BOX, G.E.P., and COX, D.R. (1964). An analysis of transformations. *Journal of The Royal Statistical Society*, Series B, 26, 211-243, discussions 244-252.

BREIMAN, L., and FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-597.

CARROLL, R.J., and RUPPERT, D. (1981). On prediction and power transformation family. *Biometrika*, 68, 609-615.

CARROLL, R.J., and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.

CALVIN, J.A., and SEDRANSK, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *Journal of the American Statistical Association*, 86, 36-54.

CHEN, J., and QIN, J. (1992). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

COCHRAN, W.G. (1977). *Sampling Techniques*. (3rd ed.) New York: John Wiley.

DE VEAUX, R.D., and STEELE, J.M. (1989). ACE guided-transformation method for estimation of the coefficient of soil-water diffusivity. *Technometrics*, 31, 91-98.

DEVILLE, J., and SÄRNDAL, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ELLIOTT, J.M. (1977). Statistical analysis of samples of benthic invertebrates. *Freshwater Biological Scientific Publication*, No. 25, (2nd ed.).

HÁJEK, J. (1960). Limiting Distributions in Simple Random Sampling From a Finite Population. Publications in Mathematics of the Hungarian Academy of Science, 5, 361-374.

NELDER, J.A., and PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika*, 74, 221-232.

OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

OWEN, A. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*, 18, 90-120.

PANKRATZ, A., and DUDLEY, U. (1987). Forecasts of power-transformed series. *Journal of Forecasting*, 6, 239-248.

ROYALL, R.M., and CUMBERLAND, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

ROYALL, R.M., and CUMBERLAND, W.G. (1981b). The finite-population linear regression estimator and estimators of its variance – An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

ROYALL, R.M., and CUMBERLAND, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

SÄRNDAL, C., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SCOTT, A., and WU, C.F. (1981). On the asymptotic distribution of the ratio and regression estimators. *Journal of the American Statistical Association*, 76, 98-102.

WHITMORE, G.A. (1983). A regression method for censored inverse-gaussian data. *The Canadian Journal of Statistics*, 11, 305-315.