Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

WORKING PAPER NO. BSMD-92-003E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-92-003E

DIRECTION DE LA MÉTHODOLOGIE

COMPARISON OF CONFIDENCE INTERVALS DERIVED USING CLASSICAL TAYLOR SERIES EXPANSION AND THE METHOD OF TEST INVERSION - A SIMULATION STUDY

by

Zdenek Patak
March 1992

RÉSUMÉ

Ce rapport compare deux méthodes de construction d'intervalles de confiance pour des paramètres d'enquêtes complexes. L'une des méthodes utilise l'approche habituelle par développement en série de Taylor, et l'autre, l'approche par fonctions estimantes. Une simulation a été mise au point pour étudier les est estimations par quotient et par régression. L'estimation par régression illustre les difficultés associées à l'élimination des paramètres de nuisance dans les cas multi-dimensionnels. Les résultats de cette étude indiquent que la méthode des fonctions estimantes produit de meilleurs intervalles de confiance que la méthode de développement en série de Taylor.

Comparison of confidence intervals derived using classical Taylor series expansion and the method of test inversion - a simulation study

by Zdenek Patak

1 Introduction

Binder (1991) proposed the method of test inversion using estimating functions for constructing confidence intervals for a wide variety of population parameters. This is in contrast to the more common method of first computing the variance $\hat{\sigma}^2(\theta)$ of the estimated population parameter of interest θ and then, assuming normality, computing the confidence interval given by $(\theta - z_{1-\alpha/2}\hat{\sigma}(\theta), \theta + z_{1-\alpha/2}\hat{\sigma}(\theta))$, where z_α is the α^{th} percentile of the standard normal distribution. Binder's method advocates the use of estimating functions to obtain the confidence intervals. The confidence region is defined by the equation

$$\frac{\hat{U}(\theta)^2}{mse\{\hat{U}(\theta)\}} \leq z^2_{1-\frac{\alpha}{2}},$$

where $\hat{U}(\theta)$ is the estimating function for the population parameter θ. The estimator θ is defined as the point where $\hat{U}(\theta)=0$. Unlike the classical approach, the method of test inversion leads to non-symmetric intervals which may be a contributing factor to improved coverage probabilities as will be shown later.

The first use of estimating functions on survey data appears in Woodruff (1952), where the methodology is used for constructing confidence intervals for percentiles. Since then, this approach has been extended to ratio and regression estimates to name but a few.

In this paper we discuss a simulation study which compares the coverage probabilities of the confidence intervals obtained under various approaches. In the next section we describe the data used for the study and the method of sampling. Section 3 follows up with simulation details and some results obtained for the ratio estimate. In Section 4 we consider the regression estimate and how it can be reduced to the ratio estimate by a transformation of the data, whereby nuisance parameters are eliminated. An explanation of the graphical comparison of the proposed method with the Taylor series expansion method can be found in Section 5. A short discussion of some of the difficulties encountered when dealing with small samples follows in Section 6. Evaluation of the proposed method with emphasis on coverage probability can be found in Section 7.

2 Data and Sample Selection

For the simulation study we have used real data from a business survey to compare the coverage probabilities of the proposed method with the usual approach.

The data set selected for this study was the September 1990 Survey of Employees, Payroll and Hours (SEPH). The data consisted of 66,752 observations, where 44,149 were take-alls ($P_{selection} = 1$) and 22,603 were take-somes ($P_{selection} < 1$). The SEPH universe from which the sample was drawn contained 691,450 sampling units and was stratified by province, industry group based on 3-digit Standard Industrial Classification and employment size ranging from 1 to 4, where size group 1 is 1-19 employees, size group 2 is 20-49 employees, size group 3 is 50-199 employees and size group 4 is 200+ employees, based on the Statistics Canada Business Register. It should be noted that these size group codes could be out of date.

In order to create a sampling frame which mimics a real population, the SEPH sample was reduced. The *pseudo-frame* or subuniverse was constructed so that it had equal probabilities of selection from the original SEPH universe. This was accomplished by dividing the size of the SEPH universe by the largest stratum weight, resulting in a reduced universe size of approximately 6,000 sampling units. The subuniverse was selected by taking a simple random sample without replacement within each stratum from the original SEPH sample. The reduced sample size from this subuniverse was determined by dividing the size of the SEPH sample by the largest weight. This subsample contained approximately 600 units for an overall sampling fraction of 10%. We generated samples which maintained the original probabilities of selection.

A large number of strata in the subsample contained less than two observations and could not be used in the estimation because stratum variance would not be defined. To exclude as few strata as possible, the subuniverse was collapsed over size to increase the number of units in each stratum. This step reduced the number of take-some strata from 89 to 47. In the process, few strata were removed due to insufficient size. The final subuniverse size was 5,654 sampling units; the final subsample consisted of 439 take-alls and 149 take-somes. The sampling fractions $f_h = n_h/N_h$ varied between 0.0117 and 0.50.

The variables used in the simulation study were the following: total number of employees and total wages and salaries. Our goal was to produce ratio estimates of average wages per employee and the corresponding confidence intervals. As well we considered the regression coefficient resulting from regressing wages and salaries on number of employees.

3 Confidence Intervals for Ratio Estimates

In the simulation study an estimate of the population ratio was computed. The main objective of this study was to compare coverage probabilities of the classical Taylor series expansion method and the estimating function approach. We first computed the ratio estimate defined as

$$\Psi_c = \frac{\sum_{h=1}^{L} W_h \bar{y}_h}{\sum_{h=1}^{L} W_h \bar{x}_h}, \tag{1}$$

where L is the number of strata in the sample, $\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{ih}$ and similarly $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{ih}$. The weighting factor for the stratum h is $W_h = \frac{N_h}{N}$. An estimate of the variance of Ψ_c is given by

$$\hat{\sigma}^2(\Psi_c) = \frac{1}{\bar{x}_{ST}^2}\sum_{h=1}^{L} \frac{W_h^2(1-f_h)}{n_h}\,(s_{yh}^2 + \Psi_c^2 s_{xh}^2 - 2\Psi_c s_{xyh}), \tag{2}$$

where $\bar{x}_{ST} = \sum_{h=1}^{L} W_h \bar{x}_h$ and $f_h = \frac{n_h}{N_h}$ is the sampling fraction in the stratum (see, for example, Cochran; p.166). The sample variance of the x-variable is $s_{xh}^2 = \frac{\sum_{i=1}^{n_h}(x_{ih}-\bar{x}_h)^2}{n_h - 1}$. The sample variance of the y-variable and the sample covariance, s_{yh}^2 and s_{xyh} respectively, can be obtained in similar fashion. Assuming that Ψ_c is normally distributed with mean $\Psi_c = \bar{Y}/\bar{X}$ and variance $\sigma^2(\Psi_c)$, the classical confidence region for Ψ_c, which is symmetric about Ψ_c, is

$$\Psi_c \pm z_{1-\frac{\alpha}{2}}\hat{\sigma}(\Psi_c).$$

To assess the coverage characteristics of the classical approach, we generated 5,000 random subsamples from the SEPH subuniverse and computed Ψ_c and $\hat{\sigma}(\Psi_c)$ for each subsample. Then we compared the empirical distribution $\hat{F}$ of the standardized ratio estimates to the theoretical distribution $F_{N(0,1)}$. Ideally, the quantiles of the two distributions would be equal. However, the 95th percentile of $F_{N(0,1)}$ corresponds to the 90th percentile of $\hat{F}$. This means that the estimated 95% confidence interval is undercovering the true confidence region by 5%.

One feature of the classical method that may be contributing to the undercoverage of the confidence region is its symmetric nature. Sometimes a nonsymmetric confidence interval may increase the coverage probability; the estimating function approach with its roots in the log-likelihood theory yields such intervals. Computationally this alternative to constructing confidence intervals is more expensive; however, there is evidence showing that improved coverage probability may be realized by using it.

Following the notation in Binder (1991),the parameter of interest Ψ can be defined as the solution to the equation

$$U(\Psi) = \int_{-\infty}^{\infty} u(y, \Psi)\, dF(y) = 0.$$

For the ratio estimate $\overline{Y}/\overline{X}$ the estimating function is $u(y, x, \Psi) = y - \Psi x$. For our pseudo-frame, we obtained $\Psi = 557.95$. It is assumed that the estimator of $U(\Psi)$, $\hat{U}(\Psi)$, is approximately normal with mean $U(\Psi)$ and variance $\sigma\{\hat{U}(\Psi)\}$. Confidence intervals may be obtained by solving

$$\frac{\hat{U}(\Psi)^2}{\hat{\sigma}^2\{\hat{U}(\Psi)\}} \leq z^2_{1-\alpha/2}, \tag{3}$$

where $\hat{\sigma}^2\{\hat{U}(\Psi)\}$ is a consistent estimator of $\sigma^2\{\hat{U}(\Psi)\}$. In the ratio estimate setup (3) becomes

$$\frac{\left\{\sum_{h=1}^{L} W_h \bar{y}_h - \Psi \sum_{h=1}^{L} W_h \bar{x}_h\right\}^2}{\sum_{h=1}^{L} \frac{W_h^2(1-f_h)}{n_h}\left(s^2_{xh} - 2\Psi s_{xyz} + \Psi^2 s^2_{yh}\right)} \leq z^2_{1-\alpha/2}. \tag{4}$$

Notice that to solve (4) for Ψ results in solving a quadratic inequality in Ψ. The two roots will define a non-symmetric confidence region about the parameter of interest. To assess the performance of the estimating function method, 5,000 random subsamples were used to produce a relatively smooth version of the empirical distribution $\hat{F}$ of the left hand side of (4). $\hat{F}$ was plotted against $F_{N(0,1)}$ to compare empirical confidence region coverage with its theoretical counterpart. The 95th percentile of the theoretical distribution was equal to the \$91.5th percentile of the empirical distribution. This means that a 95% empirical confidence interval undercovers the theoretical confidence region by 3.5%, a 1.5% improvement over the classical method.

4 Confidence Intervals for Regression Estimates in the Presence of Nuisance Parameters

In the simulation study we considered a simple regression analysis for *y* (wages) on *x* (number of employees) with an intercept term, which is taken to be the nuisance parameter. The complete theory behind removing the nuisance paramater by integrating it out of the estimating equations is described in Binder (1991). We transformed the data to reduce the original problem to a one-dimensional situation similar to making a ratio estimate.

The SEPH sample that was used for the simulations contained approximately 20% of zeros. These cases were removed prior to the estimation of Ψ because they caused the distribution of the residuals to be skewed to the right. An x-y plot of the remaining observations was made to determine whether classical regression assumptions were being satisfied. In the x-y plot we noticed a fanning out of the data points suggesting that the variance was increasing with the x-variable. Therefore, the data were reweighted to stabilize the variance of the regression by dividing the value of the response variable (*y*) by the value of the explanatory variable (*x*).

The transformed x- and y-variables were the following

$$y^* = \left(1 - \frac{\hat{\bar{X}}_v}{x}\right)\left(y - \hat{\bar{Y}}_u\right)$$

and

$$x^* = x\left(1 - \frac{\hat{\bar{X}}_v}{x}\right)^2,$$

where

$$\hat{\bar{X}}_v = \frac{\sum_{h=1}^{L} \frac{N_h}{n_h} m_h}{\sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i:x_{ih} \neq 0} x_{ih}^{-1}}$$

and

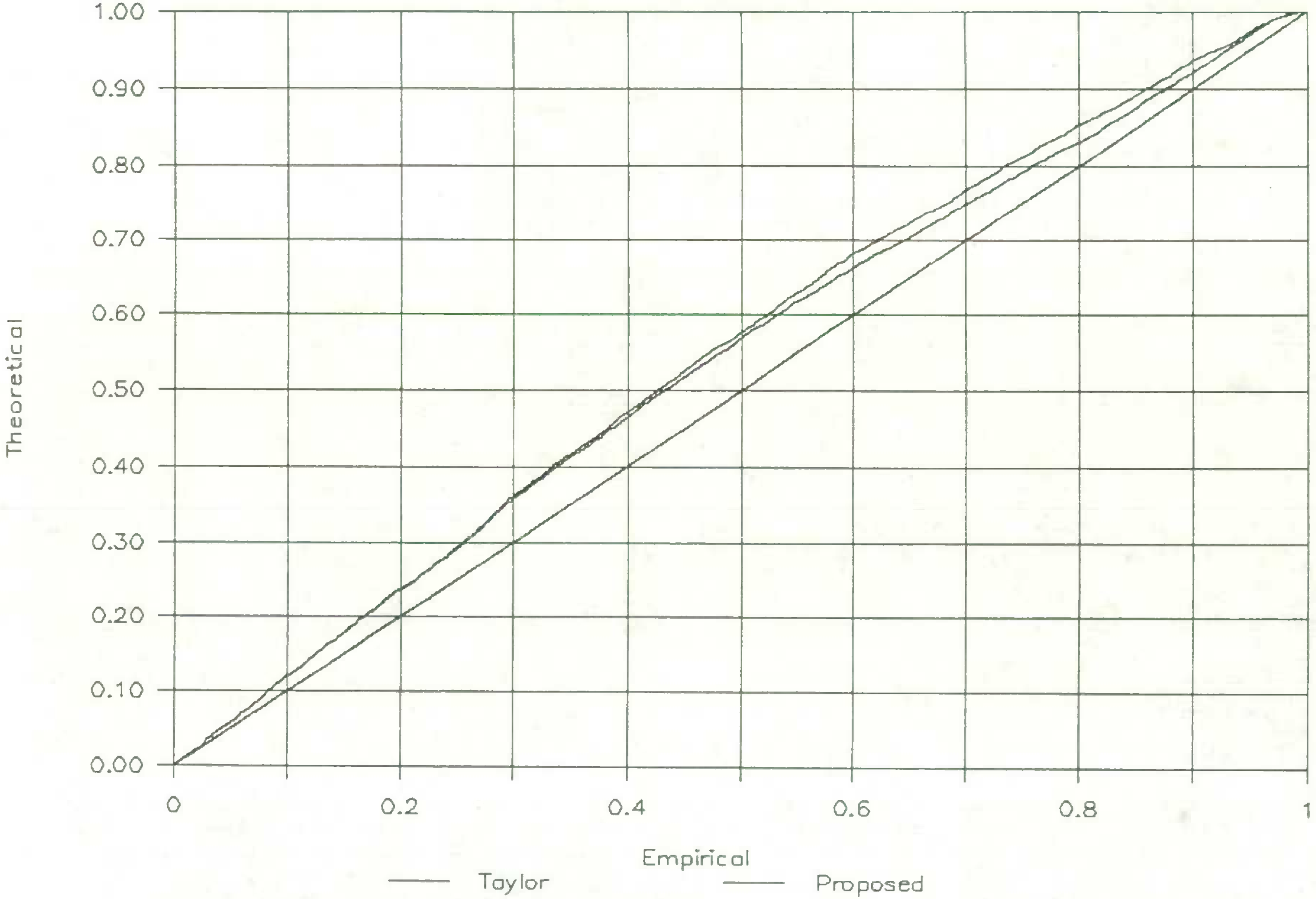
$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i:x_{ih} \neq 0} \frac{y_{ih}}{x_{ih}}}{\sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i:x_{ih} \neq 0} x_{ih}^{-1}}$$

where the quantity m_h is the number of non-zero units in sample in stratum h. After this transformation of the x- and y-variables we proceed as in the ratio estimate case. We here see that $\Psi=581.08$ (the intercept term was $\beta_0=-527.53$).To obtain confidence intervals, we solve equation (4) with $\Psi, \bar{y}_h, \bar{x}_h, s_{xh}^2, s_{xyh}$ and s_{yh}^2 replaced with $\Psi^\bullet, \bar{y}_h^\bullet, \bar{x}_h^\bullet, (s_{xh}^\bullet)^2, s_{xyh}^\bullet, (s_{yh}^\bullet)^2$. Coverage probabilities for both methods were then compared. In the regression case with a nuisance parameter the classical approach to constructing confidence intervals led to an undercoverage of the 90% confidence interval by 5%. The alternative based on estimating functions resulted in a 2% undercoverage, an improvement of 3%.

5 Plots for the Confidence Regions

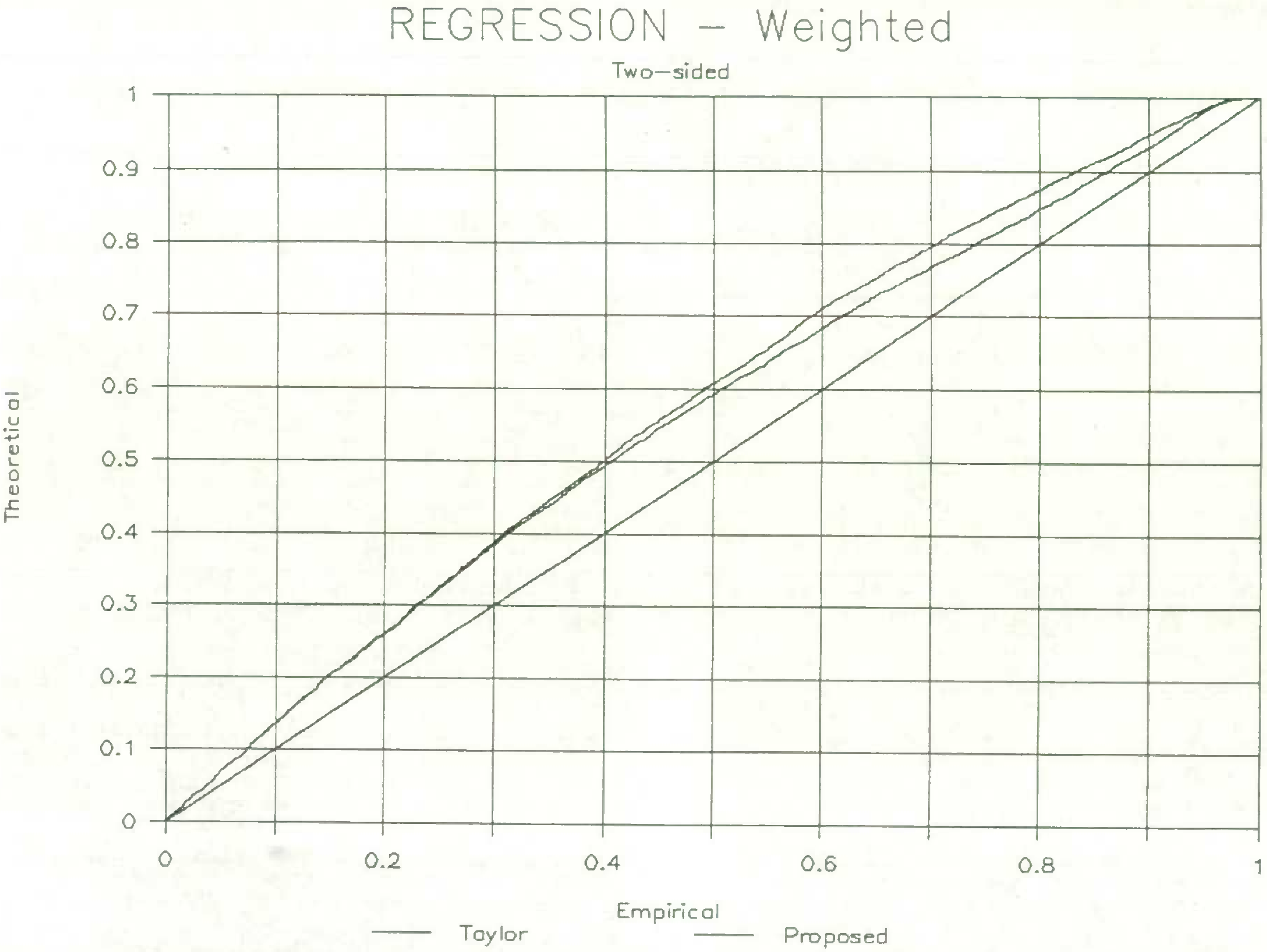
The coverage probabilities of two-sided confidence regions for ratio and regression estimates were graphically compared for the methods of Taylor series expansion, and test inversion based on estimating functions using P-P plots. The probabilities for the estimates being contained in the sample confidence interval were plotted against their theoretical counterparts.

The probability plots reveal that both methods fall short of the claimed confidence level; however, the proposed method does so to a lesser extent. In the regression setup, the proposed method exhibits visibly better coverage characteristics for high probability confidence regions than the classical method. The improvement can be detected even at 95%.



RATIOS

Two-sided


REGRESSION - Weighted
Two-sided
Theoretical
Empirical
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Taylor
Proposed

6 Discussion

The SEPH sample contains approximately 44,000 take-alls and 22,000 take-somes. Our subsample had 439 take-alls and only 149 take-somes because several strata had to be ignored due to insufficient size. This resulted in a higher-than-normal proportion of take-alls in our subsample.

The increased number of take-alls in the subsample causes the variance of the estimated parameter to underestimate the true variance and therefore overestimate the true coverage probability. Also, the sample distributions are somewhat skewed to the right. For larger samples with a smaller ratio of take-alls to take-somes variances will be estimated with greater accuracy resulting in more symmetric coverage probability distributions.

7 Conclusion

In both the ratio estimate and the regression cases the proposed method based on test inversion proved to be superior to the classical, Taylor series expansion method in terms of coverage probabilities of the corresponding confidence regions. Furthermore, the proposed method provides a unifying framework for a wide class of problems dealing with the estimation of parameters in finite populations whereas the classical approach has to be tailored to a specific problem.

The estimating functions may be computationally more involved but their inherent adaptability to specific situations as evidenced by the non-symmetric nature of the confidence regions they yield marks a significant improvement over existing methodology.

8 Acknowledgements

I would like to express my thanks to Dr. David Binder for giving me the opportunity to work with him on an exciting research project, for his guidance during the simulation study and for his helpful comments during the preparation of this paper.

References

Binder, D.A., (1991). Use of Estimating Functions for Interval Estimation from Complex Surveys.

Cochran, W.G., (1977). Sampling Techniques. John Wiley & Sons. 3rd edition.

Woodruff, R.S., (1952). Confidence Intervals for Medians and other Position Measures. JASA, 47, 635-646.