DEGREES OF FREEDOM AND CONFIDENCE INTERVAL COVERAGE IN COMPLEX MODEL BASED SAMPLING AND ESTIMATION

Wendy Rotz, Archana Joshee and Jinhee Yang,
Ernst & Young LLP, 1225 Connecticut Ave., NW, Washington, DC 20036

Key Words: Confidence Interval, Degrees of Freedom, Model Based Sampling

1. Introduction

The allocation of fixed assets to different depreciation categories for tax purposes can be costly since it involves site visits, blueprints, engineers, architects, lawyers and tax experts. Therefore, small samples sizes are essential. Fortunately, business data generally includes one or more strong covariates, allowing the use of model-based sampling and estimation to improve precision in small samples.

Occasionally there are different subgroups in the population that are expected to have varying compositions of fixed assets and do not belong in the same model. In these circumstances, we stratify based on the anticipated mix of assets. Separate models are then developed in each stratum to estimate the asset amounts that can be moved to a shorter depreciable life category. The stratum estimates are summed and the corresponding variances of the estimates are summed.

This paper assesses the appropriate degrees of freedom to use when constructing a confidence interval of the overall total estimate and explores how well confidence intervals cover the true values in our setting.

2. Background

It is important to distinguish the type of stratification we are considering in this study.

Figure 1 illustrates an example when two subgroups of a population have two distinct depreciation compositions, for which we will stratify. By contrast, Figure 2 shows stratification based on the design variable, or the independent variable.

In settings illustrated by Figure 1, separate models are built for each stratum. However, when stratum data fall on the same line, as in Figure 2, we use a single model incorporating sampling weights to account for the stratification. The degrees of freedom in this latter setting are well established, and not a subject of study in this paper.

Figure 1. Stratification by Anticipated Trend

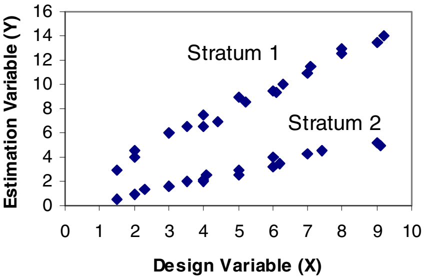


Figure 2. Stratification by Size

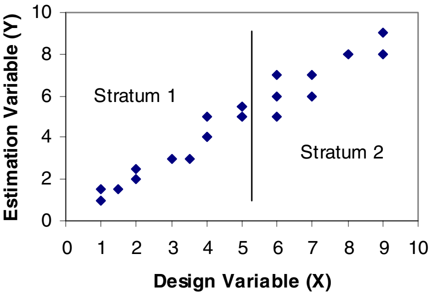
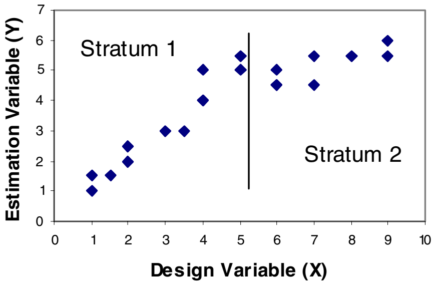


Figure 3. Stratification by Size and Trend



Occasionally, data may level off somewhat in the second stratum as shown in Figure 3. In these situations, we may use a transformation on x and a single model. However, we sometimes obtain a smaller variance by building separate models for each stratum. In this case, we again have a question regarding the degrees of freedom.

When calculating the correct total degrees of freedom (***df***) for a confidence interval of the sum of two model based estimates, it is tempting to use the sum of each model's degrees of freedom:

$$df = df_1 + df_2 .$$

However, we know summing degrees of freedom over strata is inappropriate in design based estimation where theoretically:

$$min(df_1, df_2) \leq df \leq df_1 + df_2 .$$

In design-based estimation, we commonly use Satterthwaite's approximation to ***df***. If ***Y is*** a linear combination of normally distributed Y_j, then, Satterthwaite's ***df*** is a linear combination of the variances of Y_j divided by the sum of the weighted variances squared:

$$df = \frac{\left(\sum_j w_j S_{yj}^2\right)^2}{\sum \frac{w_j^2 S_{yj}^4}{(n_j - 1)}} .$$

Is Satterthwaite's approximation applicable to model-based estimation as well? Consider that our model-based sum is a linear combination of $\hat{Y}_j$, with the variance of the total given by the sum of the stratum variances:

$$\sum_j Var(\hat{Y}_j) = \sum_j w_j MSE_j$$

where w_j is a constant related to ***X*** values, regression weights, the sampling fraction, and sample sizes. Therefore, it appears as though Satterthwaite's ***df*** would be applicable with the Mean Square Error (MSE) in place of S_{yj}^2

In this paper, through simulation, we studied the confidence interval coverage using Satterthwaite's ***df*** for our sum of model-based estimates. For benchmarking, we compared it to $min(df_1, df_2)$ and $df_1 + df_2$ controlling for different factors.

In design-based estimation, we know that Satterthwaite's approximation may overestimate ***df*** in the presence of positive kurtosis in ***Y***, causing under-coverage of confidence intervals.

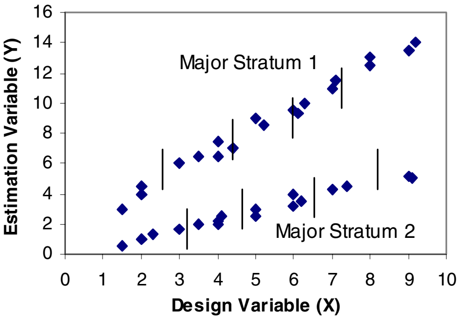
In model-based estimation, ***Y*** is assumed to be normally distributed about ***X***. The variance of the model, MSE, would appear to be more critical than the variance of ***Y*** when considering kurtosis.

Yet, our application is to data with highly skewed distributions. Our independent (design) variable, X, has long tails in the larger values, resulting in similar long tails in the distribution of the dependent variable Y.

Thus in our study, in accordance with the model assumptions, we simulated ***Y*** normally distributed about ***X***, but chose several increasing levels of kurtosis in ***X*** in order to mirror the data we find in application. The corresponding distributions of ***Y*** therefore had positive kurtosis as well.

In addition, deep stratification within each major stratum was used to draw the sample selections. Deep stratification is a sample selection method that obtains a representative sample, reduces sampling error, and may improve confidence interval coverage. Each major stratum is divided into numerous (deep) substrata of equal counts. See Figure 4 below. Within each major stratum, the deep strata are sampled at a constant rate, so that within each major stratum, every record has the same probability of selection.

Figure 4. Deep Stratification Illustrated



Thus our study of confidence interval coverage included deep stratification, methods of computing degrees of freedom, and levels of kurtosis in highly skewed data.

3. Methodology

We performed simulations to assess confidence interval coverage under Satterthwaite's ***df*** with two benchmark comparisons:

1. $min(df_1, df_2)$
2. Satterthwaite's approximation
3. $df_1 + df_2$

We generated several population data sets similar to Figure 1 creating both an ***X*** and ***Y*** value for the entire population, with the two strata differing in the relation between ***X*** and ***Y***. Actual population values of the total ***Y*** over the two strata were known in these simulations.

For each population, we conducted numerous simulations using deep stratification to draw samples in each stratum. We estimated $\boldsymbol{Y}$, and created confidence intervals using the three ***df*** approaches above. Finally, we noted the percent of confidence intervals containing the actual $\boldsymbol{Y}$ values to assess the confidence interval coverage.

The $\boldsymbol{X}$ values were generated using a gamma distribution in SAS. Eight populations with varying degrees of kurtosis in $\boldsymbol{X}$ were considered. For each distribution, 400 records were created in both major strata 1 and 2.

The $\boldsymbol{Y}$ values were simulated according to a heteroscedastic model without an intercept:

$$Y_i = \beta X_i + \varepsilon_i \text{, where}$$

$$Var(\varepsilon_i) = X_i \sigma^2 .$$

The values of ε_i were generated by multiplying $\sigma\sqrt{X_i}$ by a normal ***N(0,1)*** variable generated by the Rannor function in SAS. Note that no overt steps were taken to create kurtosis of epsilon in this process. However, small levels of kurtosis in epsilon were created by chance in the generation of normal ***N(0,1)*** variables.

The value of $\boldsymbol{\sigma}$ was set to the same constant across all populations and strata in this study. The value of β was held constant within each major stratum across all populations studies.

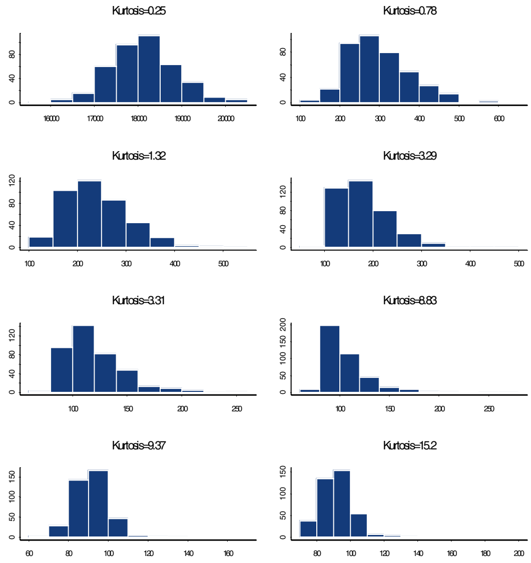
The only factor that changed between populations was the degree of kurtosis in $\boldsymbol{X}$, which accordingly formed kurtosis in $\boldsymbol{Y}$. See Figure 5 for the differing distributions of $\boldsymbol{Y}$ depicted here just for stratum 1.

In each simulation, a random sample of 20 items was selected from stratum 1 and a sample of 5 items was selected from stratum 2.

The estimates were calculated according to the ratio method as described in Lohr.[1]

For each population, 10,000 simulations were conducted. Using the three methods for determining degrees of freedom, 90, 95, and 99 percent confidence intervals were calculated. We determined the percent of intervals containing the true value of Y and compared the percentage to the ascribed confidence level.

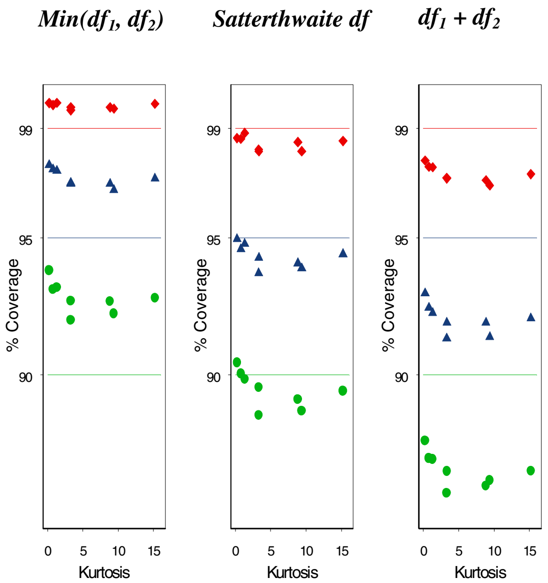
Figure 5. Kurtosis of Y in Stratum 1



4. Results

The results of the simulation are summarized in Figure 6 below in three adjacent plots, one for each *df* method. The vertical axis is the actual percent coverage found in the simulations, where the red diamonds, blue triangles and green circles show the coverage of 99, 95, and 90% confidence intervals respectively. The horizontal axis is the varying levels of kurtosis on $\boldsymbol{Y}$. The results show that kurtosis has a minimal level of impact in our setting.

Figure 6. Confidence Coverage Results



It can be seen in the first plot that the confidence interval when using the minimum degrees of

[1] Lohr, S. (1999) Sampling: Design and Analysis, Duxbury Press: Pacific Grove, CA, pages 81-83

freedom over covers and is too conservative in this situation. The minimum degrees of freedom are too small creating confidence intervals that are conservative, but wider than necessary, and contain the actual value more than their ascribed percent.

The last plot demonstrates that the sum of the two degrees of freedom consistently under covers. Thus, while tempting to use, df_1+df_2 is not conservative leading to confidence intervals that are narrower than they should be.

Satterthwaite's approximation to the degrees of freedom, in the middle plot, has the closest coverage to the ascribed values of the confidence levels.

However, note that Satterthwaite's approximation resulted in some mild under coverage with kurtosis levels between 4 and 10. This needs further exploration. It is not clear whether kurtosis or some other factor caused these findings, since the coverage improved at a kurtosis of 15.

The symmetry of the confidence intervals was evaluated by examining the simulations where the true **Y** value fell outside the confidence interval. The percentages where the actual value was above and below the confidence interval were determined and plotted in Figures 7 through 9. Figures 7 and 8 show that the percent of times the actual value fell below the lower limit for the 90 and 95 percent coverage is higher than the percent that fell above the upper limit in cases where the kurtosis of Y is low. When the kurtosis of Y is on the higher side, the percent that fell above the upper limit is higher than the percentage that fell below the lower limit. However, there is less obvious asymmetry for the 99 percent confidence coverage as can be seen in Figure 9.

Figure 7. Percent Above and Below the Upper and Lower Limit for 90 Percent Coverage

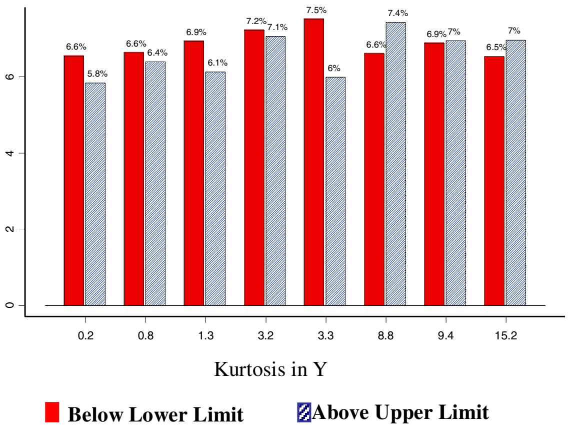


Figure 8. Percent above and Below the Upper and Lower Limit for 95 Percent Coverage

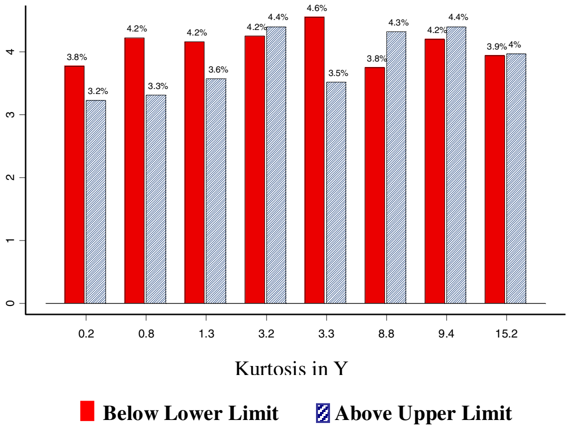
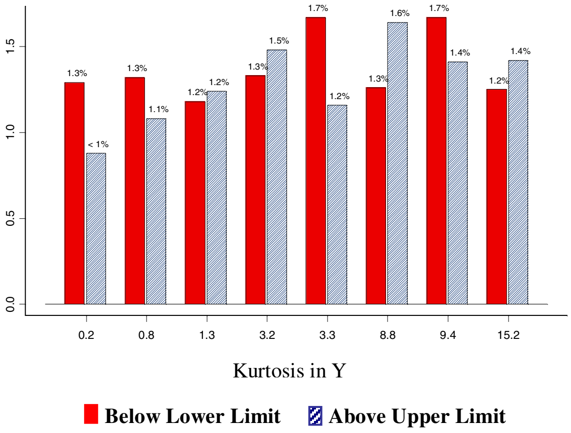


Figure 9. Percent Above and Below the Upper and Lower Limit for 99 Percent Coverage



5. Conclusion and Next Steps.

Our findings show that the conservative approach of using the minimum of the degrees of freedom is too conservative. The sum of stratum ***df***s overstates the appropriate degrees of freedom.

Satterthwaite's approximation to the degrees of freedom appears to be the most accurate method in the scenarios we tested.

More testing is needed to understand the factors related to the slight under-coverage that we found in our study. Also, different population and sample sizes should be tested for a variety of variance levels in the strata with large variances in small strata and vice versa.

Finally, generalized linear models should also be considered as described by Dorfman, Valliant and Royall.[2]

In conclusion, in the simulated populations we've tested, it appears as though Satterthwaite's *df* applied to model-based estimation behaves similarly to design-based estimation. Therefore, Satterthwaite's approximation to the *df* can be considered a practical alternative to either using the sum or the minimum degrees of freedom.

6. References

[1] Batcher, M. & Liu, Y. (2003), Ratio Estimation of small samples using deep stratification, *Proceedings of the 2003 Joint Statistical Meetings, Methodology Section*

[2] Brewer, K.R.W. (1999). Design-based or Prediction-based Inference? Stratified Random vs. Stratified Balanced Sampling. *International Statistical Review*, 67, 1, 35-47.

[3] Cochran, W.G. (1977), *Sampling Techniques,* 3rd ed., New York: Wiley.

[4] Dorfman, A, Valliant, R. & Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, New York: John Wiley.

[5] Lewis, P.A.W., & Orav, E.J. (1989) Simulation Methodology for Statisticians, Operations Analysists, and Engineers, Wadsworth, Inc.

[6] Lohr, S. (1999) *Sampling: Design and Analysis*, Duxbury Press

[7] Rotz, W., Joshee A., Yang, J. (2006) Confidence Interval Coverage in Complex Model Based Estimation, *Proceedings of the Joint Statistical Meetings at 2006, Methodology Section*

[8] Royall, R.M. & Herson, J. (1973). Robust Estimation in Finite Populations. *Journal of the American Statistical Association* 68, 344, 880-889.

[9] Royall, R.M. & Cumberland, W.G. (1981). An Empirical Study of Ratio Estimator and Estimators of Its Variance. *Journal of the American Statistical Association* 76, 373, 66-77.

[10] Royall, R.M. & Cumberland, W.G. (1981). The Finite-Population Linear Regression Estimator and Estimators of Its Variance – An Empirical Study. *Journal of the American Statistical Association* 76, 376, 924-930.

[11] Sarndal, C., Swenson B., and Wretman J. (1992) *Model Assisted Survey Sampling*, Springder-Verlag

[12] Tam, S.M. & Chan, N.N. (1984). Screening of Probability Samples. *International Statistical Review*, 52, 3, 301-308.

[2] Dorfman, A, Valliant, R. & Royall, R.M. (2000) *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley: New York, page 113.