TRUMP: Tuned Ratio Unbiased Mean Predictor

Sarjinder Singh and Stephen A. Sedory

Department of Mathematics
Texas A&M University-Kingsville
Kingsville, TX 78363, USA
E-mail: kuss2008@tamuk.edu

Abstract
In this paper, we introduce what we call a tuned ratio unbiased mean predictor (TRUMP) and which we show that has smaller variance than the ratio predictor for simple random and with replacement sampling. The proposed TRUMP can be made even more efficient than the Best Linear Unbiased Predictor (BLUP) by appropriate choice of what we call a TRUMP Care coefficient. The generalized regression (GREG) predictor and linear regression predictors are also considered in the comparison while doing simulation study.

Key Words: TRUMP Cuts, TRUMP Care Coefficient, First Basic Information (FBI), Jackknifing, Model unbiased, Relative efficiency.

1. Introduction

Let y_i and x_i, $i=1,2,...,N$, be the values of the study variable and auxiliary variable, respectively, of the i[th] unit in the population Ω. Here we consider the problem of estimating the population mean

$$\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i \tag{1.1}$$

by assuming that the population mean

$$\bar{X} = N^{-1}\sum_{i=1}^{N} x_i \tag{1.2}$$

of the auxiliary variable is known.

Let (y_i, x_i), $i=1,2,...,n$, be the values of the study variable and auxiliary variable of the i[th] unit in the sample s drawn using the simple random and with replacement sampling (SRSWR) scheme.

Let

$$\bar{y}_n = n^{-1}\sum_{i\in s} y_i \tag{1.3}$$

and

$$\bar{x}_n = n^{-1}\sum_{i\in s} x_i \tag{1.4}$$

be the sample means for the study variable and the auxiliary variable respectively.

Cochran (1940) defined a ratio estimator of the population mean given by

$$\bar{y}_{Rat} = \bar{y}_n\left(\frac{\bar{X}}{\bar{x}_n}\right) \tag{1.5}$$

When the regression line passes through the origin, the well known mean predictor model is given by:

$$y_i = Rx_i + e_i \tag{1.6}$$

where

$$R = \frac{\bar{Y}}{\bar{X}} \tag{1.7}$$

is the ratio of the population mean of the study variable to that of the auxiliary variable;

$$E_m(e_i \mid x_i) = 0 \tag{1.8}$$

$$E_m(e_i^2 \mid x_i) = V_m(e_i \mid x_i) = \sigma^2 \tag{1.9}$$

and

$$E_m(e_i e_j \mid x_i x_j) = C_m(e_i, e_j \mid x_i x_j) = 0 \tag{1.10}$$

where E_m, V_m, and C_m denote the model expectation, variance and covariance, respectively.

Under model (1.6), the Best Linear Unbiased Predictor (BLUP), see Singh (2003, page 222), $\bar{y}_{BLUP}$, is given by

$$\bar{y}_{BLUP} = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right)\bar{X} \tag{1.11}$$

It is easy to verify that both the ratio and BLUP estimators are unbiased and the $\bar{y}_{\mathrm{BLUP}}$ is always more efficient than $\bar{y}_{\mathrm{Rat}}$. The readers should note that only final results are given in this paper, and that detailed results will be available in the near future (See Singh and Sedory 2017b).

In this paper, in section 2 we define a few new terms such as TRUMP Cuts, TRUMP Care Coefficient, First Basic Information (FBI), and finally introduce the TRUMP Predictor. In section 3, we look for what is behind TRUMP? It is shown that there is a ratio which helps the proposed TRUMP to perform better than BLUP. In section 4, we attempt to look for which family is supportive of the proposed TRUMP?

2. TRUMP: Tuned Ratio Unbiased Mean Predictor

Consider a sample s of n observations taken by the simple random and with replacement (SRSWR) design and the observed values are (y_i, x_i), $i = 1,2,...,n$. Following Singh, Sedory, Rueda, Arcos and Arnab (2015), we now consider a new estimator of the population mean $\bar{Y}$ defined as:

$$\bar{y}_{\text{TRUMP}} = \sum_{j \in s} \left\{ (n-1)^2 \bar{w}_n(j) - (n-2) \right\} \bar{y}_n(j)_{TC} \tag{2.1}$$

where

$$\bar{y}_n(j)_{TC} = \frac{n^g y_j - \bar{y}_n}{n^g - 1} \tag{2.2}$$

are called TRUMP Cuts (TC), and $\bar{w}_n(j)$ are called tuned calibrated weighs to be determined based on certain criterion.

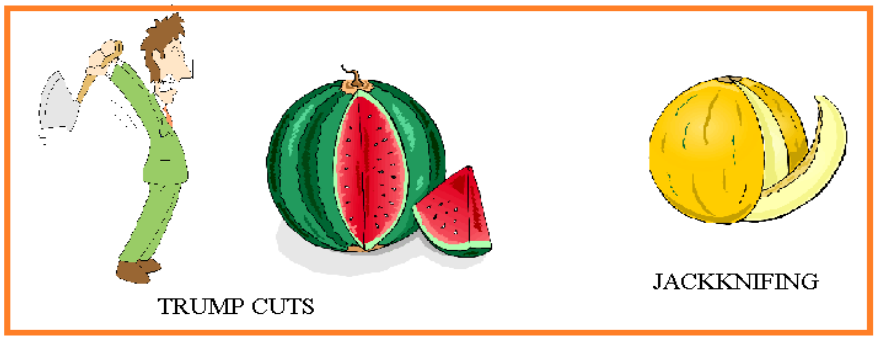

Fig.2.1. TRUMP Cuts versus Jackknifing

The TC is obtained by calibrating the jth sampled observation y_j by n^g, and then subtracting the sampled mean value $\bar{y}_n$. The value of $g \neq 0$ is called TRUMP Care Coefficient and its value depends on the First Basic Information (FBI) in hands based on past experience or otherwise. For example, if $g = -1$, then

$$\bar{y}_n(j)_{TC} = \frac{y_j - n\bar{y}_n}{1-n} = \frac{n\bar{y}_n - y_j}{n-1} = \bar{y}_n(j) \tag{2.3}$$

which is the usual jackknifing due to Quenouille (1956) and was first used by Tukey (1958) to estimate the variance.

Singh and Sedory (2017a) made use of the jackknifing idea to create what they called a Tuned Ratio Unbiased Mean Predictor (TRUMP) and which was presented at the Seventh Annual Statistics Day, hosted on the campus of Texas A&M University-Kingsville. Improvements were made and a later version was presented at the Joint Statistical Meting 2017, Baltimore, Maryland, USA.

Singh and Sedory (2017b) constructed a new calibration constraint

$$\sum_{i=1}^{n} \overline{w}_n(j)\overline{x}_n(j)_{TC} = \frac{\overline{X} - n(2-n)\overline{x}_n}{(n-1)^2} \tag{2.4}$$

which is same constraint as that used in Deville and by Särndal (1992) and Singh *et al.* (2015), except that the jackknifed means are replaced by "TRUMP Cuts", given by

$$\overline{x}_n(j)_{TC} = \frac{n^g\, x_j - \overline{x}_n}{n^g - 1} \tag{2.5}$$

Next they considered minimizing the model variance

$$V_m(\overline{y}_{TRUMP}) = \frac{\sigma^2}{n}\left(\frac{n^{2g+1} + 1 - 2n^g}{(n^g - 1)^2}\right)\sum_{j=1}^{n}\left\{(n-1)^2\,\overline{w}_n(j) - (n-2)\right\}^2 \tag{2.6}$$

They developed the TRUMP weights which lead to the estimator $\overline{y}_{\text{TRUMP}}$ given by:

$$\overline{y}_{\text{TRUMP}} = \left(\frac{\sum_{j=1}^{n}\overline{y}_n(j)_{TC}\,\overline{x}_n(j)_{TC}}{\sum_{j=1}^{n}\{\overline{x}_n(j)_{TC}\}^2}\right)\overline{X} \tag{2.7}$$

which became the Tuned Ratio Unbiased Mean Predictor (TRUMP) under the "TRUMP Cuts" model;

$$\overline{y}_n(j)_{TC} = R\overline{x}_n(j)_{TC} + \overline{e}_n(j)_{TC} \tag{2.8}$$

They also showed that the variance of the proposed $\overline{y}_{\text{TRUMP}}$ is given by:

$$V(\overline{y}_{\text{TRUMP}}) = \frac{\sigma^2}{n}\left(\frac{n^{2g+1} + 1 - 2n^g}{(n^g - 1)^2}\right)E_p\left[\frac{\overline{X}^2}{\sum_{j=1}^{n}\{\overline{x}_n(j)_{TC}\}^2}\right] \tag{2.9}$$

where E_p denotes the design expectation.

3. What is behind TRUMP?

Note that the $\bar{y}_{BLUP}$ can be written as:

$$\bar{y}\text{BLUP} = \frac{\bar{y}_n}{\bar{x}_n}\left[\frac{1+\frac{(n-1)s_{xy}}{n\bar{x}_n\bar{y}_n}}{1+\frac{(n-1)s_x^2}{n\bar{x}_n^2}}\right]\bar{X} \tag{3.1}$$

which is similar to the unbiased ratio estimator found in Beale (1962). Note that the Beale (1962) estimator is design unbiased, but $\bar{y}_{BLUP}$ is a model unbiased predictor, see Singh (2003, page 222).

The proposed $\bar{y}_{TRUMP}$ can be written as:

$$\bar{y}_{TRUMP} = \frac{\bar{y}_n}{\bar{x}_n}\left[\frac{1+\frac{(n-1)n^{2g}s_{xy}}{n(n^g-1)^2\bar{x}_n\bar{y}_n}}{1+\frac{(n-1)n^{2g}s_x^2}{n(n^g-1)^2\bar{x}_n^2}}\right]\bar{X} \tag{3.2}$$

On comparing (3.1) with (3.2), one can see that the ratio:

$$\text{Ratio} = \frac{n^{2g}}{(n^g-1)^2} \tag{3.3}$$

is playing some role to make TRUMP more efficient or less efficient.

If $g = \ln(0.5)/\ln(n)$, then the BLUP is a special case of the proposed TRUMP. Note that the "Ratio" is in the middle of the estimator, thus it will not be as so easy to beat BLUP for an obvious choice of TRUMP Care Coefficient g.

In the next section, we perform a simulation study to see if there is any First Basic Information (FBI) about the value of TRUMP Care Coefficient (g) that could help the performance of the proposed TRUMP.

4. Which family is supporter of TRUMP?

The generator of data (GOD) produces the set of all data sets, including the "big" ones, which are surely subsets of it. To discover which family of distributions might support

TRUMP, we borrowed a bivariate dataset from GOD, and did a simulation study. The bivariate data set is found using the model

$$y_i = R\, x_i + e_i \tag{4.1}$$

where we generated $x_i \sim G(a,b)$ and $e_i \sim N(0,1)$. In the simulation study, we have compared six estimators including those considered above, which we define again for the convenience of the readers:

$$\hat{\theta}_0 = \bar{y}_n \quad \text{(Sample mean)} \tag{4.2}$$

$$\hat{\theta}_1 = \bar{y}_n \left(\frac{\overline{X}}{\bar{x}_n} \right) \quad \text{(Ratio predictor)} \tag{4.3}$$

$$\hat{\theta}_2 = \bar{y}_n + \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \left(\overline{X} - \bar{x}_n \right) \quad \text{(GREG predictor)} \tag{4.4}$$

$$\hat{\theta}_3 = \bar{y}_n + \frac{s_{xy}}{s_x^2} \left(\overline{X} - \bar{x}_n \right) \quad \text{(Regression predictor)} \tag{4.5}$$

$$\hat{\theta}_4 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \overline{X} \quad \text{(BLUP)} \tag{4.6}$$

and

$$\hat{\theta}_5 = \frac{\sum_{j=1}^{n} \{\bar{x}_n(j)_{TC}\, \bar{y}_n(j)_{TC}\}}{\sum_{j=1}^{n} \{\bar{x}_n(j)_{TC}\}^2} \overline{X} \quad \text{(TRUMP)} \tag{4.7}$$

We note that the GREG predictor is due to Deville and Sárndal (1992), and the linear regression predictor is due to Hansen, Hurwitz and Madow (1953).

For different sample sizes, n, we computed the percent relative efficiency of the jth predictor $\hat{\theta}_j$ over the sample mean predictor $\hat{\theta}_0$ as:

$$\mathrm{RE}(\hat{\theta}_j) = \frac{\sum_{k=1}^{NITR}\left(\hat{\theta}_{0|k} - \bar{Y}\right)^2}{\sum_{k=1}^{NITR}\left(\hat{\theta}_{j|k} - \bar{Y}\right)^2} \times 100\% \qquad (4.8)$$

where $NITR$ stands for the number of iterations, that is the number of times we make a prediction of the population mean $\bar{Y}$ using the particular predictor.

For the simulation study, we set $N = 1050$ $R = 0.8$, $a = 3.0$, and $b = 0.8$, where N represents population size, R represent the ratio of two population means, and a and b are the shape and scale parameters in a Gamma distribution. Such choice of parameters results in a bivariate population with correlation coefficient value of $\rho_{xy} = 0..8595791$. A pictorial presentation of such a population is given in Figure 4.1.

From the population of $N = 1050$ units, we select $NITR = 10{,}000$ samples each of sizes $n = 30$ (say) and then change the value of the TRUMP care coefficient from $g = 0.5$ to 2.5 with a step of 0.05. Next we changed the sample sizes from 30 to 55 with a step of 5. Using R-programming, we retained those results when RE(5) is greater than or equal to RE(4), that is, when the proposed TRUMP is at least as efficient as the BLUP estimator.

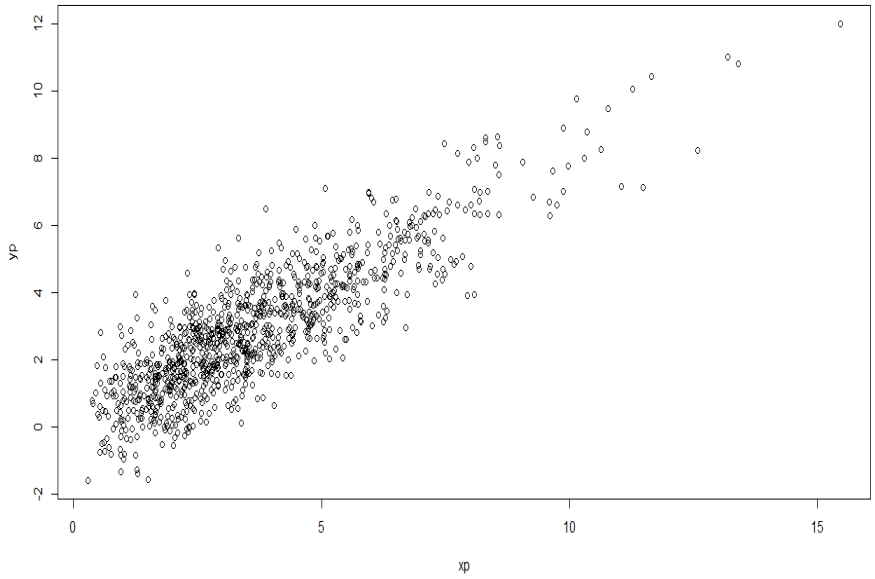

Fig. 4.1. Scatter plot of 1050 data values

Similarly we retained the results for each of the other estimators had relative efficiencies (RE(1), RE(2), RE(3), and RE(4)) greater than 100%. This was done limit ourselves to cases where the counterparts - ratio, regression, GREG and BLUP estimators are more efficient than the sample mean estimator. Under the homoscedastic model considered here, it is not easy to beat or develop another predictor which could beat the BLUP. However, here the proposed TRUMP shows efficient results for a good choice of TRUMP Care Coefficient g. Thus the First Basic Information (FBI) in hands about the choice of g is helpful. There are many choices of values of g where the value of $RE(5) \geq RE(4)$. We first provide a range of First Basic Information (FBI) which could be useful for selecting a value of the TRUMP Care Coefficient (g) for different sample sizes in Table 4.1.

Table 4.1. Range of the TRUMP Care Coefficient vs sample size.

	Value of g			
n	Minimum	Median	Maximum	Frequency
30	0.50	1.50	2.50	41
35	0.50	1.50	2.50	41
40	0.50	1.50	2.50	41
45	0.50	1.50	2.50	41
50	0.50	1.50	2.50	41
55	0.50	1.50	2.50	41

There is a huge number of results in favour of TRUMP; we provide for each of the sample sizes considered, the average relative efficiency values over all values of g. (See Table 4.2) The last column in Table 4.2 provides only standard deviation of the TRUMP $RE(5)$. Note that there is no value of SD for $RE(j)$, $j = 1,2,3,4$, because those estimator are free from the value of g. The values of standard deviations of $RE(5)$ are not very small indicating that there is a variation in the value of $RE(5)$, thus choosing a value of TRUMP care coefficient g close of its median values listed in Table 4.1 will be a safe value for different sample sizes. It is interesting to note that for each sample size between 30 to 55 with a step of five, there was always a value of g between 0.5 to 2.5 such that the condition $RE(5) \geq RE(4)$ holds.

Table 4.2. Average $RE(j)$, $j = 1,2,3,4,5$ for different sample sizes n over all possible values of g considered.

	Ratio	GREG	LR	BLUP	TRUMP	SD of
n	$RE(1)$	$RE(2)$	$RE(3)$	$RE(4)$	$RE(5)$	$RE(5)$
30	964.29	1142.10	2463.20	3402.30	3805.00	788.00
35	508.30	606.81	1388.40	1591.80	1728.10	263.20
40	651.28	771.18	1818.90	1754.70	1863.20	208.10
45	883.15	1077.70	3466.40	3623.50	3994.00	765.00
50	564.13	621.64	923.48	1116.90	1150.00	62.70
55	467.36	523.86	788.51	1711.10	1824.80	230.00

A pictorial presentation of all $RE(j)$, $j = 1,2,3,4,5$ values for different sample sizes are shown in Fig. 4.2.

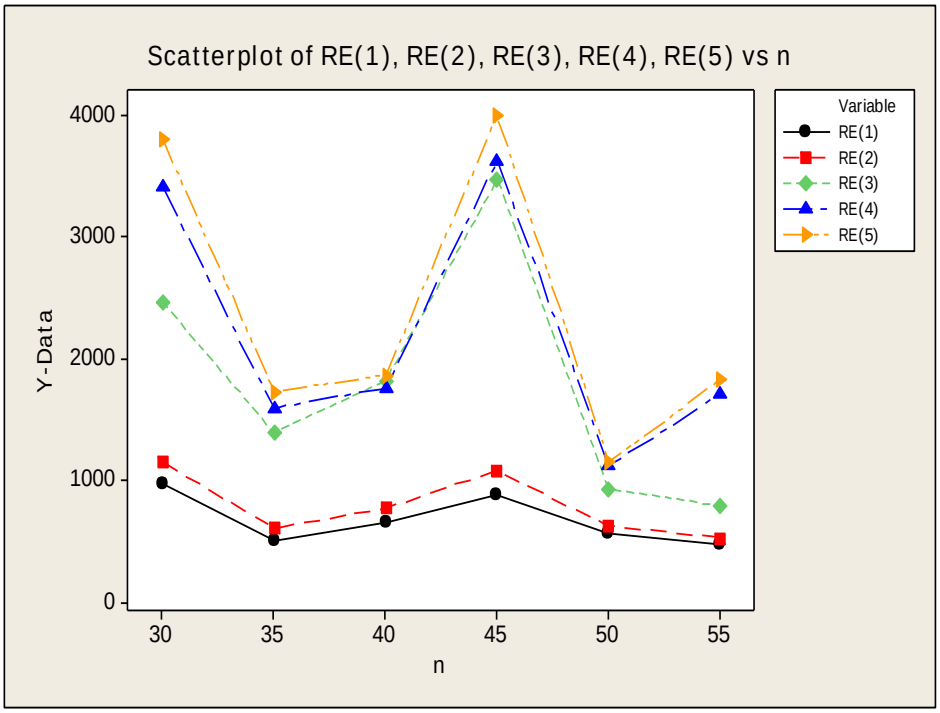

Fig. 4.2. Comparison of the five predictors with mean.

From Fig 4.2, one can see that the ratio, and GREG fall into one category and, the Linear Regression, the BLUP and TRUMP form another efficient category. We conclude that the proposed TRUMP can be made at least as efficient as the natural BLUP when the regression line passes through the origin.

Figure 4.3 shows the actual values of RE(1), RE(2), RE(3), RE(4) and RE(5) versus the value of g for a sample of size $n = 30$.

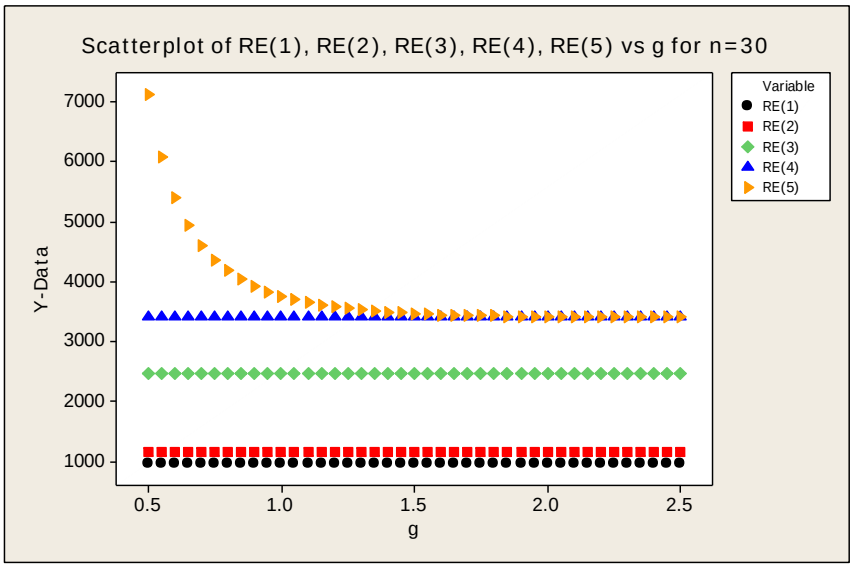

Fig. 4.3. RE(j), $j = 1,2,3,4,5$ values for sample size of 30 units.

It is easy to visualize that as the value of the TRUMP Care Coefficient varies between 0.5 and 1.5, the proposed TRUP is better than all the other four competitors. As soon as the value of g reaches 1.5 then the proposed TRUMP and BLUP are showing almost equal efficiency. However, both BLUP and TRUMP are better than the linear regression estimator, which in turn is more efficient than the both ratio and GREG estimator. The graphs for RE(1), RE(2), RE(3), and RE(4) are straight horizontal lines because these value are not dependent on the value of g. As the value of g increases from 0.5 towards 2.5, the RE(5) value seems to decrease exponentially.

In the same way, the Figures 4.4 through 4.8 are devoted to displaying the values of RE(j), $j = 1,2,3,4,5$ for different sample sizes from 35 to 55 with a step of 5.

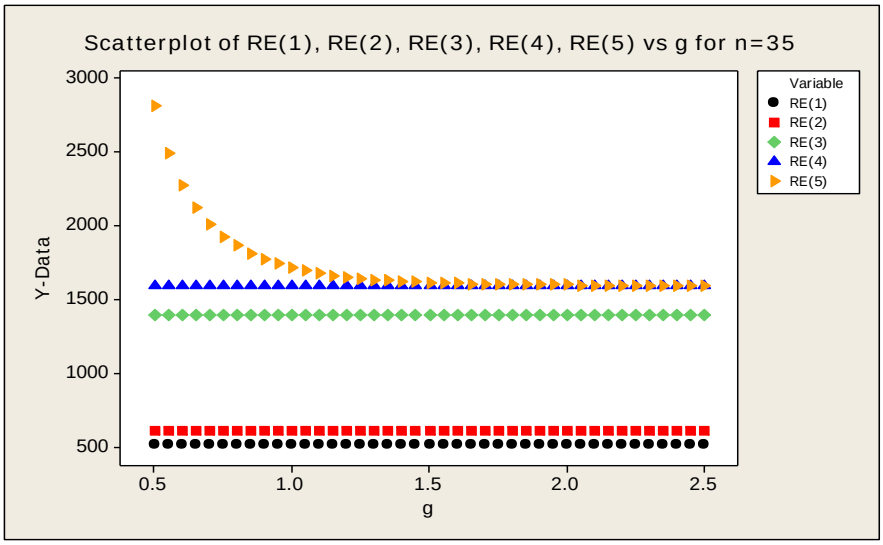

Fig. 4.4. RE(j), $j = 1,2,3,4,5$ values for sample size of 35 units.

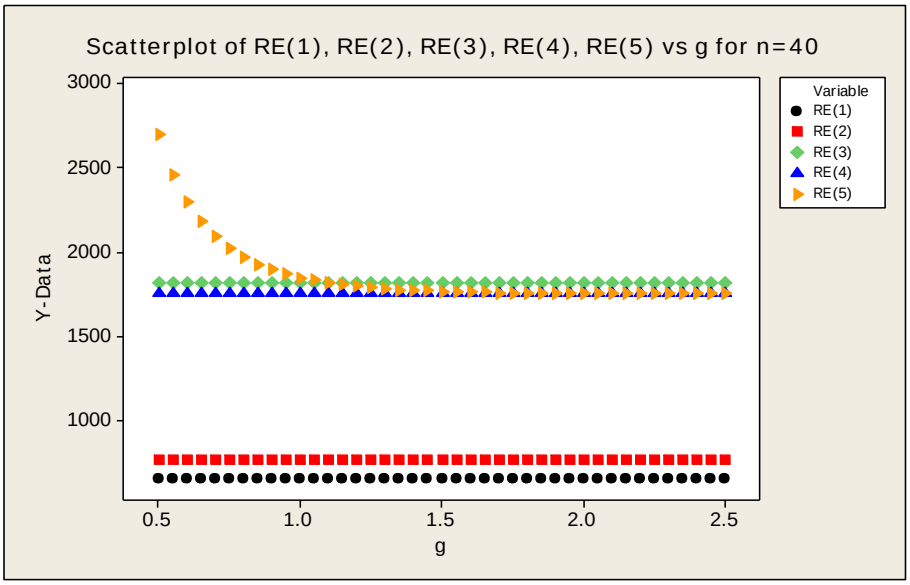

Fig. 4.5. RE(j), $j = 1,2,3,4,5$ values for sample size of 40 units.

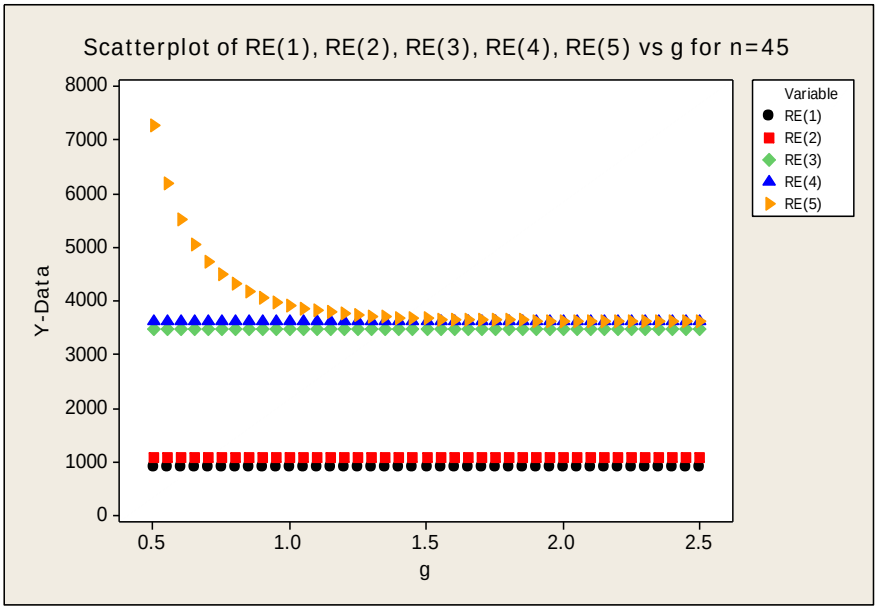

Fig. 4.6. RE(j), $j = 1,2,3,4,5$ values for sample size of 45 units.

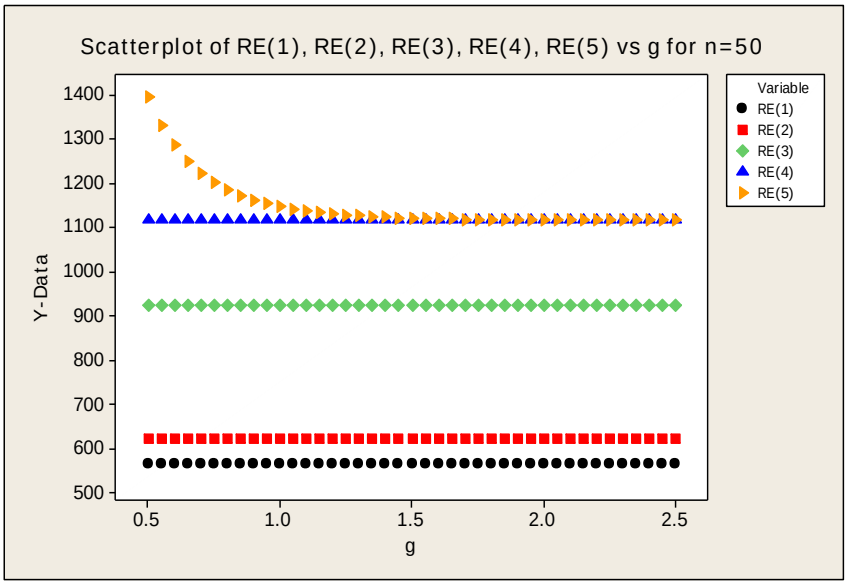

Fig. 4.7. RE(j), $j = 1,2,3,4,5$ values for sample size of 50 units.

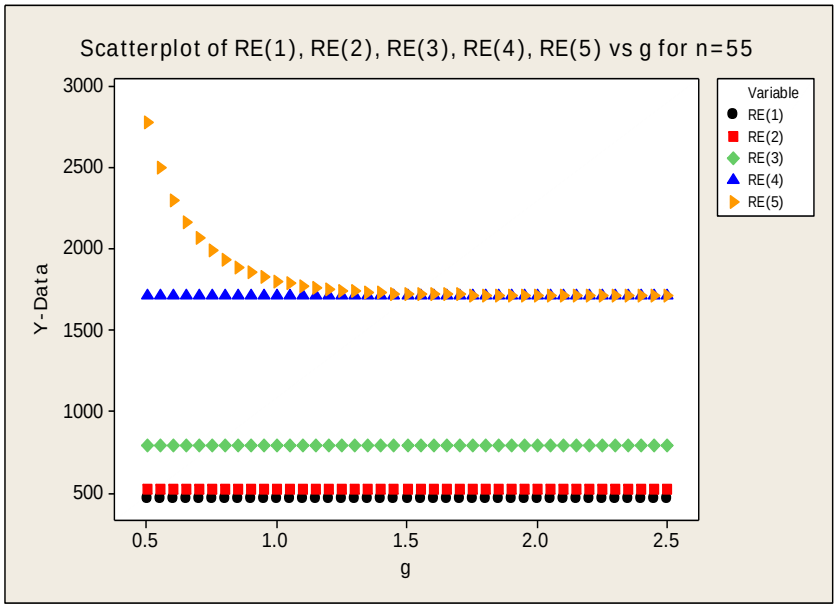

Fig. 4.8. RE(j), $j = 1,2,3,4,5$ values for sample size of 55 units.

One straightforward conclusion can be made from these figures that the trend of the percent relative efficiency values as a function of TRUMP Care Coefficient remains same for different sample sizes in the range 30 to 55 with a step of 5.

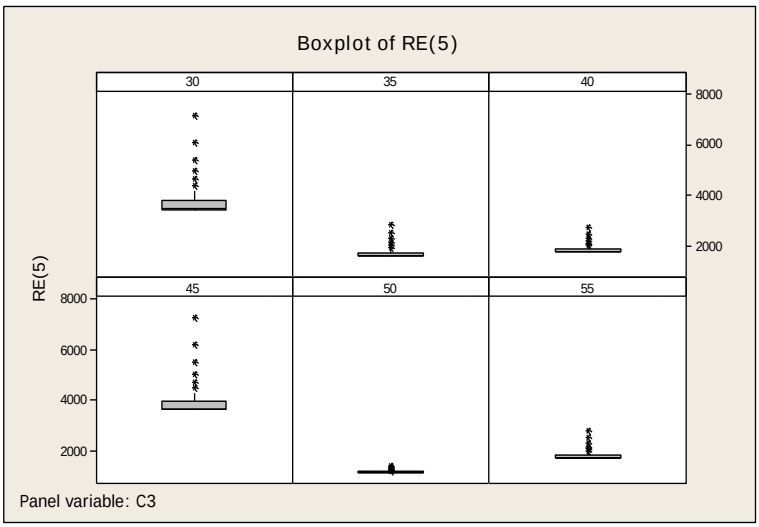

Fig. 4.9. Box plots for TRUMP

Figure 4.9 shows the box plots obtained from the relative efficiency RE(5) for different sample sizes for different values of the TRUMP Care Coefficient. More outliers are observed for sample sizes 30 and 45 in comparison other sample sizes considered.

Acknowledgements

The use of ARTEXPLOSION 600,000 in sketching pictures is duly acknowledged. We also acknowledge the use of R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Disclaimer

The opinion and results discussed in this contribution are solely the authors' own and do not reflect these of any associated institute or organization.

References

Cochran, W.G. (1940). Some properties of estimators based on sampling scheme with varying probabilities. *Austral. J. Statist.,* 17, 22--28.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc*., 87, 376-382.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory.* New York, John Wiley and Sons, 456--464.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

Singh, S. (2003). *Advanced Sampling Theory with Applications: How Michael Selected Amy*. Kluwer.

Singh, S. and Sedory, S.A. (2017a). TRUMP: Tuned Ratio Unbiased Mean Predictor. *Presented at the Seventh Annual Statistics Day-Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX (April 22, 2017)*.

Singh, S. and Sedory, S.A. (2017b). TRUMP: Tuned Ratio Unbiased Mean Predictor. *Monograph in Progress, Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX*

Singh, S., Sedory, S.A., Rueda, M.M, Arcos, A., and Arnab R. (2015). *A new concept for tuning design weights in survey sampling*. Elsevier.

Tukey, J.W. (1958). Bias and confidence in not-quite large samples (abstract). *Ann. Math. Statist*., 29, 61-75.