Estimation of Co-efficient of Variation in PPS sampling.

Archana. V (1 st Author)
Department of Statistics, Mangalore University
Mangalagangotri, Konaje- 574199
Karnataka, India.
Email: archana_stkl@yahoo.com

Aruna Rao K (2nd Author)
Department of Statistics, Mangalore University
Mangalagangotri, Konaje- 574199
Karnataka, India.
Email: arunaraomu@yahoo.com

1. Intoduction.

Inference concerning the population coefficient of variation (C.V) has attracted the attention of several eminent researchers in the past. It dates back to the works of Mckay(1932) and Pearson(1932). At that time the primary concern was for the estimation and confidence interval for the C.V of the normal distribution. Subsequent works concentrated on the improved estimation of the confidence interval for the normal C.V. They include the works of Mahmoudvand and Hassani(2007) and Panchkitkosotkul(2009). Subsequently the focus was to develop tests for equality of the population C.V's of two or more groups when the underlying distribution is normal. Likelihood ratio, wald and score test principles are used to develop the tests. The research papers in this direction includes Ahmed (2002), Nairy and Rao(2003) and Jafari and Behboodian(2010) and the reference cited therein.

For a long time inference regarding C.V did not draw the attention of survey statisticians. The first reported work is due to Das and Tripathi (1980) who proposed estimation of C.V when the design is simple random sampling without replacement (SRSWOR). Subsequently several papers have addressed the estimation of finite population C.V. One of the pioneering work in this direction is due to Tripathi et al., (2002) where in they proposed a class of estimators for the population C.V. This class includes the ratio and regression estimators of population C.V using sample C.V and certain hybrid estimators where regression estimators are used to propose a new ratio estimator. Patel and Shah (2007) examined the small sample mean square error (MSE) of several estimators including some estimators belonging to the hybrid class. All these estimators refers to the case of simple random sampling. Rajyaguru and Gupta (2006) proposed three classes of estimators of population C.V under stratified random sampling. Probability proportional to size sampling (PPS) is widely used in sample surveys. Thus it becomes important to propose estimators of C.V under PPS sampling. In this paper we have proposed an estimator of C.V under probability proportional to size with replacement sampling(PPSWR). The mean square error of the estimator is derived to the order of $O(n^{-1})$. The performance of the proposed estimator is compared with the estimator under simple random sampling using simulations. It is common practice in the applied research is to draw the samples using a particular sampling design but to use normal theory estimators for the analysis of the data. Thus the performance of the proposed estimator is also compared for the sample C.V when the samples are drawn using PPSWR scheme. The sample C.V under PPSWR scheme turns out to be the best estimator in terms of the MSE and justifies the practice followed by the applied researchers.

The rest of the paper is organized as follows: In section 2, estimators of C.V under PPSWR scheme is proposed. Small sample comparison of the performance of estimators of C.V is presented in section 3. The results and discussions are given in section 4 and the paper concludes in section 5.

2. Estimators of C.V under PPSWR sampling.

In theory of sampling it is customary to denote the study variable by 'y' and the auxiliary variable by 'x'. Let (X_i, Y_i), i=1, .., N be the values of the variables 'x' and 'y' for the ith unit in the population. Let $\overline{Y}$ and $\sigma_y^{\,2}$ denote the population mean and population variance for the study variable where,

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad \text{and} \quad \sigma_y^{\,2} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2 .$$

The primary focus of interest is to estimate $\theta_y = \dfrac{\sigma_y}{\overline{Y}}$.

Under PPSWR scheme an unbiased estimator of σ_y^2 and $\overline{Y}$ is given by,

$$\hat{\overline{Y}}_{HH} = \frac{1}{Nn}\sum_{i=1}^{n}\frac{y_i}{p_i} \qquad \text{\_\_\_\_\_\_\_\_\_\_\_\_\_\_(2.1)}$$

$$v(\hat{\overline{Y}}_{HH}) = \frac{1}{N^2 n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{HH}\right)^2 \qquad \text{\_\_\_\_\_\_\_\_\_\_\_\_(2.2)}$$

$$where \qquad \hat{Y}_{HH} = \frac{1}{n}\sum_i \frac{y_i}{p_i}$$

Here $p_i = \dfrac{x_i}{X}$, $\mathrm{X} = \sum \mathrm{X}_\mathrm{i}$ i.e. probability of selecting ith unit in the sample.

Now we can define the estimator of C.V under PPSWR as,

$$\hat{\theta}_{y_1} = \frac{\left(\hat{\sigma}_y^{\,2}\right)^{\frac{1}{2}}_{HH}}{\hat{\overline{Y}}_{HH}} \qquad \text{\_\_\_\_\_\_\_\_\_\_\_\_\_(2.3)}$$

$$\text{where} \qquad \left(\hat{\sigma}_y^{\,2}\right)_{HH} = \left\{\frac{1}{Nn}\sum_i \frac{y_i^{\,2}}{p_i} - \frac{1}{N^2}\hat{Y}^2{}_{HH} + \frac{1}{N^2}\hat{v}(\hat{Y}_{HH})\right\}$$

Another possible estimate of C.V under PPSWR is the sample C.V and is given by,

$$\hat{\theta}_{y_2} = \frac{\left(s_y^{\,2}\right)^{\frac{1}{2}}}{\overline{y}} \qquad \text{\_\_\_\_\_\_\_\_\_\_\_\_\_(2.4)}$$

where $\overline{y}$ is the sample mean using the PPS sample and $s_y^{\,2}$ is the sample variance using the PPS sample.

The sample C.V under simple random sampling scheme is defined as,

$$\hat{\theta}_{y_3} = \frac{s_y}{\bar{y}} \qquad \text{\_\_\_\_\_\_\_\_\_\_\_\_(2.5)}$$

where $\bar{y}$ is the sample mean and $s_y^{\ 2}$ is the sample variance and are given by,

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad s_y^{\ 2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}.$$

The expressions for the bias and MSE of the estimators $\hat{\theta}_{y_1}, \hat{\theta}_{y_2}$ and $\hat{\theta}_{y_3}$ are derived by the authors to the order of O(n^{-1}). The derivations of MSE's of the above estimators under PPSWR scheme is algebraically tedious, but since it is not used anywhere in the paper to save space these expressions are not reported here.

3. Small sample comparison of the estimators of C.V.

3.1. Simulation Experiment.

A simulation experiment is carried out to compare the MSE's of the various estimators. The expressions for the MSE of the estimators derived in the previous section were to the order of O(n-1) and thus are the asymptotic MSE's. In order to estimate the exact MSE, it is necessary to go for a simulation experiment. The estimation of mean under PPSWR sampling scheme has got a smaller MSE compared to the estimator under SRSWR/WOR if the size variable is highly correlated with the study variable. Thus it is necessary to incorporate the correlation coefficient in the simulation experiment. For this purpose 2000 samples were generated from a bivariate normal distribution. Care is taken to ensure that all the values of the size (auxiliary) variable is positive. From this population a sample of size 'n' is selected using PPSWR and SRSWR schemes. From the PPS sample the estimators of population C.V as given by equation (2.3) is computed along with the sample C.V. Using SRS samples, sample C.V is also computed. The MSE of these three estimators are estimated using 10,000 simulations. The configurations used in the simulations are the following.

The values of the C.V of the study variable used in the simulations were 0.1, 0.3, 0.5, 0.8, 1.0 and 2.0. For each fixed value of the study variable a set of four values of C.V of the auxiliary variable are considered. They are 0.5, 1.0, 1.5 and 2 times the C.V of the study variable. The correlation coefficients 'r' used in the simulation study are -0.9, -0.7,-0.5,-0.3,-0.1, 0, 0.1, 0.3, 0.5, 0.7, 0.9. The sample sizes considered are n=100, 200.

The total number of configurations works out to be=6*4*11*2=528.

3.2 Small sample MSE.

Table (3.1) represents the ratio of the MSE of the sample C.V under PPS sample and SRS sample compared to the estimator of C.V under PPS sample. For bravity the latter estimator is referred as the PPS estimator of C. V. Generally when comparing the MSE of the two estimators, the MSE of the estimator having smaller value is taken in the denominator. However since our motivation is compare the MSE of the sample C.V's to the MSE of the PPS estimator, the MSE of the latter is taken in the denominator. Thus a little caution is needed to interpret the results. The result is presented only for n=100 and the pattern remains the same for the other sample sizes. Before preparing this table, the MSE of the estimate of C.V is carefully examined for the 4 values of the C.V of the auxiliary variable for each of the estimators. The MSE almost remains the same irrespective of the C.V of the auxiliary variable and in the table (3.1) the ratio's refers to the ratio of the average MSE of the estimator in the numerator to the average MSE of the estimator in the denominator; the average is taken over the four values of the MSE corresponding to the values of C.V of the auxiliary variable.

Table 3.1: Relative MSE (%) of the sample C.V under PPS and SRS sample compared to the estimator of C.V under PPSWR.

C. V of the study variable		**Correlation co-efficient 'r'** (n=100)										
		-0.9	**-0.7**	**-0.5**	**-0.3**	**-0.1**	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
0.1	$\hat{\theta}_1$	100	100	100	100	100	100	100	100	100	100	100
	$\hat{\theta}_2$	4.76	5.00	5.56	5.00	4.54	4.76	5.26	5.00	4.76	5.26	4.76
	$\hat{\theta}_3$	23.81	25.00	27.78	25.00	27.27	28.57	26.32	25.00	23.81	26.32	23.81
0.3	$\hat{\theta}_1$	100	100	100	100	100	100	100	100	100	100	100
	$\hat{\theta}_2$	1.98	2.18	2.35	2.46	2.50	2.57	2.63	2.71	2.74	2.20	2.01
	$\hat{\theta}_3$	3.28	3.34	3.44	3.53	3.49	3.68	3.94	3.80	4.10	3.37	3.32
0.5	$\hat{\theta}_1$	100	100	100	100	100	100	100	100	100	100	100
	$\hat{\theta}_2$	5.28	5.17	6.08	5.68	6.19	6.00	6.18	6.02	5.90	5.30	5.15
	$\hat{\theta}_3$	6.38	6.89	7.22	7.55	7.44	7.60	8.11	8.43	7.38	6.71	6.19
0.8	$\hat{\theta}_1$	100	100	100	100	100	100	100	100	100	100	100
	$\hat{\theta}_2$	17.82	17.86	18.43	18.78	19.15	20.32	18.89	19.53	18.08	17.87	18.01
	$\hat{\theta}_3$	20.14	19.87	20.97	21.33	23.18	25.48	23.92	23.41	22.89	21.28	20.06
1.0	$\hat{\theta}_1$	100	100	100	100	100	100	100	100	100	100	100
	$\hat{\theta}_2$	33.82	36.57	35.68	36.83	36.60	36.72	36.98	37.15	36.82	36.67	34.11
	$\hat{\theta}_3$	38.41	39.81	37.88	39.06	38.99	38.87	39.14	39.49	39.33	38.67	36.46
2.0	$\hat{\theta}_1$	100	100	100	100	100	100	100	100	100	100	100
	$\hat{\theta}_2$	45.70	46.96	47.01	47.89	47.67	48.04	48.43	47.98	47.78	46.54	46.01
	$\hat{\theta}_3$	54.57	55.97	56.08	56.24	56.89	56.64	57.29	56.94	57.03	56.10	55.87

From the table it is clear that the estimator which has got the smallest MSE corresponds to the sample C.V under PPS sampling followed by the MSE of the sample C.V under SRS scheme. This is true for various values of the correlation coefficient. For moderate values of C.V, the relative efficiency (i.e., inverse of the values reported here) of these 2 estimators is very high compared to the PPS estimator of C.V. As the value of C.V of the study variable increases, although the efficiency is higher for these 2 estimators the gain in sample size is reduced. For eg: when θ_y =0.1, r=-0.9 the relative MSE of these 2 estimators are respectively 4.76 and 23.81. This indicates that when the PPS estimator requires a sample of size 100 to attain the same MSE, the sample C.V requires the sample size of 5 and 24 under PPS and SRS scheme. When θ_y = 0.8, r=-0.9 the relative MSE for these 2 estimators are 17.82 and 20.14 respectively indicating that a sample sizes of 18 and 20 are required to attain the same MSE of PPS estimator which requires a sample size 100.

We have examined the MSE of the PPS estimator across the correlation coefficients for each value of the C.V of the study variable. The results are not fully reported here to save space. Table (3.2) presents the MSE of the PPS estimator for 0.1, 0.5 and 1.0 for selected values of θ_y. The results indicate that the MSE of the PPS estimator is relatively higher when the correlation co-efficient is low.

Table 3.2: The MSE's of the proposed three estimators for selected value of C.V of the study variable.

C. V of the study variable		**Correlation co-efficient 'r' (n=100)**										
		-0.9	**-0.7**	**-0.5**	**-0.3**	**-0.1**	**0**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
0.1	$\hat{\theta}_1$	0.0021	0.0020	0.0018	0.0020	0.0022	0.0021	0.0019	0.0020	0.0021	0.0019	0.0021
	$\hat{\theta}_2$	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	$\hat{\theta}_3$	0.0005	0.0005	0.0005	0.0005	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005
0.5	$\hat{\theta}_1$	0.0284	0.0290	0.0263	0.0265	0.0243	0.0269	0.0258	0.0249	0.0271	0.0283	0.0291
	$\hat{\theta}_2$	0.0015	0.0015	0.0016	0.0015	0.0015	0.0016	0.0016	0.0015	0.0016	0.0015	0.0015
	$\hat{\theta}_3$	0.0017	0.0020	0.0019	0.0020	0.0018	0.0021	0.0021	0.0021	0.0020	0.0019	0.0018
1.0	$\hat{\theta}_1$	0.0479	0.0432	0.0454	0.0448	0.0459	0.0463	0.0465	0.0471	0.0478	0.0450	0.0469
	$\hat{\theta}_2$	0.0162	0.0158	0.0162	0.0165	0.0168	0.0170	0.0172	0.0175	0.0176	0.0165	0.0160
	$\hat{\theta}_3$	0.0184	0.0172	0.0173	0.0175	0.0179	0.0180	0.0182	0.0186	0.0188	0.0174	0.0171

4. Discussions:

In this paper an estimator of population C.V is derived when the sampling design is PPSWR. Generally it is expected that this estimator performs well compared to the estimator of C.V namely sample C.V under SRS scheme. The numerical results indicate that for the estimation of C.V, SRS is more efficient compared to the PPSWR scheme. At the time of numerical computation it was observed that the estimator of σ_y^2 using PPS scheme turns out to be negative in several cases. The percentage is approximately 20%. Thus the PPS estimator is not defined in such cases. A larger question that needs to be addressed is "Whether SRS sampling is more efficient for the estimation of compound parameters like C.V, skewness and kurtosis compared to the PPS scheme".

The results indicate that the applied researchers are justified in using sample C.V as an estimator of population C.V when the samples are drawn using PPRWR design. If this estimator is taken as the estimator of C.V under PPS design, then the PPS design turns out to be more efficient compared to the SRS design.

5. Conclusion:

Two estimators are proposed for population C.V when the samples are selected using PPSWR scheme. The first estimator is the PPS estimator obtained by unbiasedly estimating σ_y^2 (population variance) and $\bar{Y}$ (population mean) using PPS scheme. The second estimator is the sample C.V. The sample C.V under PPS sampling turns out to be the best estimator compared to the PPS estimator and sample C.V under SRSWR. We advocate that for greater efficiency it is advisable to select the sample using PPS scheme and use the sample C.V as an estimator of the population C.V. Further research is needed to shed light on the more general question namely "Whether using the estimator derived under SRS scheme turns out to be more efficient under PPS scheme for complex parameters like skewness, kurtosis and gini index". Deriving estimators for these under PPS scheme is algebraically tedious while their estimators in SRS scheme turn out to be simple.

SELECTED REFERENCES

1. Ahmed, S.E (2002): Simultaneous estimation of Co-efficient of Variation. Journal of Statistical Planning and Inference. 104, 31-51.

2. Das, A.K. and Tripathi, T.P. (1981 a): Sampling strategies for coefficient of variation using knowledge of the mean using an auxiliary character. Tech. Rep. Stat. Math. 5/81. ISI, Calcutta.

3. Das, A.K. and Tripathi, T.P. (1981 b): A class of estimators for co-efficient of variation using knowledge on coefficient of variation of an auxiliary character. Paper presented at annual conference of Ind. Soc. Agricultural Statistics. Held at New Delhi, India.

4. Mahmoudvand, R. and Hassani, H.(2007): Two new confidence intervals for the C.V in a normal distribution. Journal of Applied Statistics, 36: 4,429-442.

5. Murthy M.N. (1967): Sampling Theory and Methods, Statistical Publishing Society, Calcutta.

6. Nairy, K.S. and Rao, K.A. (2003): Tests of coefficient of variation in normal population. Communications in Statistics: Simulations and Computation 32 (3), 641 – 661.

7. Patel, P. A. and Shah Rina. (2009): A Monte Carlo comparison of some suggested estimators of Co-efficient of variation in finite population. Journal of Statistics sciences, 1(2), 137-147.

8. Panichkitkosolkul, W. (2009): Improved confidence intervals for a Co-efficient of variation of a normal distribution. Thailand statistician, 7(2), 193-199.

9. Rajyaguru, A. and Gupta, P. (2006): On the estimation of the co-efficient of variation from finite population-II, Model Assisted Statistics and application, 1(1), 57-66.

10. Rajyaguru, A. and Gupta, P. (2002): On the estimation of the co-efficient of variation from finite population-I, Model Assisted Statistics and application, 36(2), 145-156.

11. Tripathi, T.P. Singh, H.P. and Upadhyaya, L.N. (2002): A general method of estimation and its application to the estimation of co-efficient of variation. Statistics in Transition, 5(6), 887-909.

Abstract: Co-efficient of variation (C.V) is a unitless measure of dispersion and is widely used rather than standard deviation. In the recent years estimation of C.V in finite populations has drawn the attention of several researchers, however many of these works on C.V confined to the case of simple random sampling. In this paper an estimator of C.V is proposed in PPS sampling with replacement along with its asymptotic mean square error. The expressions for the design efficiency are also derived. The proposed estimator is compared with the estimator of C.V in simple random sampling with replacement using several real life data settings. By treating the finite population as a sample from an infinite normal population, the efficiency of the estimators in PPS with replacement sampling is compared with simple random sampling with replacement (SRSWR) using extensive simulation. The results indicate that for the estimation of C.V, PPSWR sampling is more efficient compared to SRSWR.

Keywords: Co-efficient of variation, PPS sampling, Simple random sampling, Design efficiency, Model-based comparison.