EVALUATION OF BALANCED HALF-SAMPLE AND JACKKNIFE ESTIMATES OF COMBINED RATIO ESTIMATES FOR NON-NORMALLY DISTRIBUTED POPULATIONS

David Hislop, U.S. Bureau of the Census
Stanley Lemeshow, University of Massachusetts/Amherst

Abstract

The performance of the Balanced Half-Sample and Jackknife methods for estimating the variance of the combined ratio estimate is evaluated using artificially generated non-normally distributed populations. In a Monte-Carlo design two variations of the balanced half-sample technique and three variations of the jackknife are examined within a framework which permits the manipulation of the underlying distributions of the random variables. The variance estimates are empirically evaluated using one symmetric and two skewed non-normal distributions which are related to the well documented results based on the normal distribution.

The results of this investigation demonstrate that the variance estimates of the combined ratio estimate are highly biased and quite unstable when the underlying distribution is non-normal and the balanced half-sample method is used. The jackknife estimates are shown to be considerably better, particularly when estimates are desired for domains of interest containing few observations.

This paper examines the performance of the performance of the Balanced Half-Sample and Jackknife techniques for estimating the variance of the combined ratio estimate when the underlying distributions of the random variables under consideration are non-normal. Previous work by McCarthy (1966), Frankel (1971), Bean (1975), Lemeshow and Epp (1977), and Lemeshow and Levy (1977) concerned the ability of these techniques to accurately estimate the variance of both linear and non-linear estimates from complex multi-stage survey designs such as the Health Examination Survey (HES) and Health Interview Survey (HIS) of the National Center for Health Statistics (NCHS). To date, all research in this area has dealt with populations whose distributional characteristics were either unknown or specifically normal. This study will evaluate the two variance estimation techniques by means of Monte-Carlo methods in which samples are selected from populations whose parameters are precisely specified.

1. Background

The Balanced Half-Sample technique is currently used by the NCHS for variance estimation of population estimates from the HES and HIS. The Jackknife, originally due to Quenouille (1956), has been gaining popularity in recent years. Both methods have been thoroughly examined by Lemeshow and Epp (1977) and Lemeshow and Levy (1977) in Monte-Carlo sampling experiments under the assumption of normality.

The nature of much of the data collected by sample survey methods, especially in the health sciences, is well documented. Data gathered in the HES, for example, are in many instances found to be non-normally distributed. Clearly, evaluation of techniques designed specifically for data from such complex sample surveys as HES should include examination of specifically non-normal populations.

Research into the Balanced Half-Sample and Jackknife variance estimation methods has been in response to the fact that precise formulae for the variance of non-linear parameters in highly complex surveys do not exist. The effect of non-normality on the ability of the techniques under consideration to provide precise variance estimates is, to date, unexplored.

2. The Sampling Experiment

To obtain the sample, observations are drawn at random from L strata of infinite size. The distribution of these observations is known and specified. This sample is used to estimate the population ratio. Subsequent samples are drawn and from them estimates are made of this population parameter. This process is repeated $M=1000$ times and the distribution of the sample estimates is studied. This Monte-Carlo computer simulation is patterned after the work of Lemeshow and Levy (1977).

For the two variations of the Balanced Half-Sample technique considered, half-sample estimates are constructed such that,

$$\hat{R}_p = \frac{\sum_{h=1}^{L} (\delta_{ph} X_{h1} + (1-\delta_{ph}) X_{h2})}{\sum_{h=1}^{L} (\delta_{ph} Y_{h1} + (1-\delta_{ph}) Y_{h2})} ,$$

is the p^{th} half-sample estimate of R, the population ratio, where δ_{ph} is an element from the p^{th} row and h^{th} column of the appropriate matrix given by Plackett and Burman (1946), and (X_{hi}, Y_{hi}) is the i^{th} sample observation from the h^{th} stratum.

The two variations of the Balanced Half-Sample variance estimate considered here are,

$$(1) \qquad \hat{V}_{B1}(\hat{R}) = \frac{1}{\ell} \sum_{p=1}^{\ell} (\hat{R}_p - \bar{R})^2$$

where $\bar{R} = \frac{1}{\ell} \sum_{p=1}^{\ell} \hat{R}_p$ and ℓ is the number of half-samples formed, and $\hat{R}$ is the sample estimate of R, and

$$(2) \qquad \hat{V}_{B2}(\hat{R}) = \frac{1}{\ell} \sum_{p=1}^{\ell} (\hat{R}_p - \hat{R})^2$$

where $\hat{R} = \sum_{h=1}^{L}\sum_{j=1}^{2} X_{hj} / \sum_{h=1}^{L}\sum_{j=1}^{2} Y_{hj}$.

In the sampling experiment the observations in each stratum are grouped into two primary units of equal size.

For the combined ratio estimate the jackknifed estimate of R are,

$$\hat{R}_{hj} = \frac{\sum_{u=1}^{L}\sum_{v=1}^{2} X_{uv} - (X_{hj} - X'_{hj})}{\sum_{u=1}^{L}\sum_{v=1}^{2} Y_{uv} - (Y_{hj} - Y'_{hj})} ,$$

where (X'_{hj}, Y'_{hj}) is the observation left in the h^{th} stratum following the delection of (X_{hj}, Y_{hj}).

The three variations of the jackknife variance estimate considered are,

(1) $$\hat{V}_{J1}(\hat{R}) = \frac{1}{2}\sum_{h=1}^{L}\sum_{j=1}^{2} (\hat{R}_{hj} - \bar{R})^2$$

where $\bar{R} = \sum_{h=1}^{L}\sum_{j=1}^{2} \hat{R}_{hj}/2L$,

(2) $$\hat{V}_{J2}(\hat{R}) = \frac{1}{2}\sum_{h=1}^{L}\sum_{j=1}^{2} (\hat{R}_{hj} - \hat{R})^2$$

where $\hat{R} = \sum_{h=1}^{L}\sum_{j=1}^{2} X_{hj} / \sum_{h=1}^{L}\sum_{j=1}^{2} Y_{hj}$,

and

(3) $$\hat{V}_{J3}(\hat{R}) = \frac{1}{2}\sum_{h=1}^{L}\sum_{j=1}^{2} (\hat{R}_{hj} - \hat{R}_{i\cdot})^2$$

where $\hat{R}_{i\cdot} = \frac{1}{2}\sum_{j=1}^{2} \hat{R}_{hj}$.

Two situations are considered:

I. L=3 strata with n=2 observations per strata.

II. L=3 strata with n=10 observations per strata.

In naturally occuring health related data sets one may find cases in which the ratio of the variables under consideration differs greatly in each stratum. Conversely, it is possible to find data in which virtually no spread across strata ratios occurs. Into this experiment are designed two cases: "No Spread" and "High Spread." No Spread is the case where the probability distribution in each stratum is precisely the same yielding equal location parameters. High Spread is characterized by large differences between strata with respect to the stratum ratios.

Four families of distributions are considered in this experiment: the Uniform, the Chi-Square, the F and the Normal Distributions. Note that two are skewed and two are symmetric. Figure 1 presents the parameters chosen for each distribution by spread.

FIGURE 1

PARAMETERS OF THE DISTRIBUTIONS TO BE CONSIDERED

		SPREAD		
		NO	LOW	HIGH
DISTRIBUTION	(Parameters)	(a,b)	(a,b)	(a,b)
U(a,b)	Stratum 1	(100,150)	(90,140)	(60,110)
	Stratum 2	(100,150)	(100,150)	(100,150)
	Stratum 3	(100,150)	(110,160)	(140,190)
	(Parameter)	(n)	(n)	(n)
χ^2 n=df	Stratum 1	10	9	2
	Stratum 2	10	10	10
	Stratum 3	10	11	18
	(Parameters)	(ν_1,ν_2)	(ν_1,ν_2)	(ν_1,ν_2)
$F_{(\nu_1,\nu_2)}$	Stratum 1	(6,14)	(6,12)	(6,10)
	Stratum 2	(6,14)	(6,14)	(6,14)
	Stratum 3	(6,14)	(6,16)	(6,18)
	(Parameters)	(μ,σ^2)	(μ, σ^2)	(μ, σ^2)
$N(\mu,\sigma^2)$	Stratum 1	(50,5)	(45,5)	(30,5)
	Stratum 2	(50,5)	(50,5)	(50,5)
	Stratum 3	(50,5)	(55,5)	(70,5)

3. Evaluation of the Variance Estimators

To evaluate the performance of the Balanced Half-Sample and Jackknife estimators of $V(\hat{R})$ one would like a precise value for $V(\hat{R})$. For the purpose a "target value," $\hat{V}(\hat{R})$, is used. This value is the variance of the M=1000 values of $\hat{R}$ as computed in the sampling experiment. Also estimated from the sampling experiment are the expected values, variances, and absolute relative biases of the variance estimation techniques under consideration.

4. Results

Since the populations used in the experiment were artifically generated several checks were implemented to verify the performance of the computer processes. First a goodness of fit test provided information confirming that the basic U(0,1) numbers generated were random. Subsequent goodness of fit tests supported the claim that the transformation utilized provided populations having the specified F, Chi-Square and Normal Distributions.

As a check on the validity of the experiment the final results are presented only after several independent trails, each using a different set of random numbers were done. On each occasion the results were comparable.

As one check on the operation of the sampling experiment the expected value of the combined ratio estimate using all 2L observations over the M=1000 trails, $\hat{E}(\hat{R})$, was compared to the theoretical value. The two were in close agreement confirming the reliability of the simulation. First the case where n=2 will be examined.

In this research a criterion is established for considering the magnitude of the estimated absolute relative bias to be "acceptable" at 10%. Table 1 shows that for n=2 when Y, the variable in the denominator of the combined

ratio estimate, has the uniform distribution, both the Jackknife and the Balanced Half-Sample method yield estimates which have low bias. The absolute relative bias, ARB, was less than or equal to 9% regardless of the distribution of X, the numerator variable. However, in virtually every other instance a pattern was found to develop. The jackknife estimates were consistently less biased than the Balanced Half-Sample estimates and yielded values which were acceptable with ARB$\leq$9%. In each of the four situations with skewed, non-normally distributed variables in the denominator, the Balanced Half-Sample estimates were shown to be highly biased. For example, when the denominator distribution was $F_{(\nu_1,\nu_2)}$ the Jackknife produced estimates generally within acceptable bounds while the balanced half-sample proved to be highly biased with ARB ranging from 37% to 69% regardless of the distribution of the variable in the numerator.

Table 2 presents $\hat{V}[\hat{V}_I(\hat{R})]$, I=B1, B2, J1, J2, J3, for selected representative distributional combinations for n=2. Clearly, the three jackknife estimates of the variance of the combined ratio estimate are less variable than either of the balanced half-sample estimates.

When there are n=2 observations per stratum and the distribution is non-normal the three jackknife estimates are shown to provide better estimates of $V(\hat{R})$, with respect to amount of bias and variability, than the two balanced half-sample estimates. This is not surprising since the jackknife techniques use more of the available information from a stratified sample in constructing estimates of the variance of the combined ratio estimate than does the balanced half-sample method. Each of the 2L jackknife estimates of the population ratio omits only one observation from a specified stratum adding twice the value of the observation left in that stratum to all the information contained in the remaining strata. This should be compared to a balanced half-sample estimate which uses one of the two observations in each stratum to estimate the combined ratio estimate. Also note that for the L=3 strata case considered here only ℓ=4 half-sample estimes of the combined ratio estimate are used for $\hat{V}_I(R)$, I=B1, B2, as opposed to the 2L=6 jackknife estimates that are used for $\hat{V}_I(\hat{R})$, I=J1, J2, J3. The next result is that, for n=2, the Jackknife technique produced a more stable estimate of the variance of the combined ratio estimate than is possible using the Balanced Half-Sample method, particularly when the data are from dispersed or highly skewed distributions.

Hislop (1977) demonstrated that the Spread factor has little effect upon the results of this investigation and, therefore, for brevity, only one case was selected for presentation.

When n=10 observations per stratum are used with two primary units in each stratum the results appear similar to those obtained in the linear case insofaras the ARB falls within acceptable bounds, (ARB$\leq$10%). This is seen in Table 3. A possible explanation for this may be attributed to the central limit theorem, since summary measures are calculated in each stratum yielding two primary units per stratum when the number of observations exceeds two. Each primary unit is the mean of half the observations in the stratum. Thus, regardless of the distribution of the original observations, as n increases, results much like those obtained when the underlying distribution is normal are expected. Table 3 shows, for the normal case, ARB$\approx$.8%.

Table 4 presents the target value as well as the expected value of the estimates over the M=1000 trials, $\hat{E}[\hat{V}_I(\hat{R})]$, I=B1, J3, for n=10 observations per stratum. Upon visual inspection it is clear that in many cases the balanced half-sample and jackknife methods are producing estimates of the target value which are strikingly similar to the findings for the normal case regardless of the distribution of the variables comprising the random pair. For several particular cases, notably when the numerator distribution is U(100,150) and the denominator distribution is $F_{(6,14)}$, the variability of the estimates was found to be high. The most variable families of distributions considered in this work are the uniform, U(a,b), and the $F_{(\nu_1,\nu_2)}$. This is shown in Table 5.

5. Conclusions

It is proposed that, as n increases, no matter what the distribution of the original observations, one may appeal to the central limit theorem and the estimates under consideration will yield values similar to those found when the distribution is normal.

For most situations considered, however, with n=10 the two techniques under consideration are shown to yield estimates whose variability is of the magnitude found in previous research for the case where the underlying distribution is normal. This is a key point for it is supportive of the use of the balanced half-sample techniques for estimating the variance of the combined ratio estimate regardless of the underlying distribution when the number of observations per stratum is equal to 10. This implies that surveys such as the HES are correct in using balanced replication since in most cases, the sample size is much larger than n=10. Notably, the Jackknife once again out performs the Balanced Half-Sample but the difference is not as pronounced.

In the complex multi-stage surveys presently in use, comparisons within domains of interest many times effectively reduce the sample size under consideration. In these cases, when the distribution of the variables of interest are non-normal or unknown, with n<10 observations per stratum, the jackknife estimate of the variance of the combined ratio estimate is to be preferred. As brought out in this research, the effect of small stratum sample size and non-normally distributed populations on the Balanced Half-Sample technique is quite serious producing estimates which are highly biased and

unstable. When n is large, however, both techniques considered here are shown to perform well regardless of the distribution of the variables under consideration.

Acknowledgements

The authors would like to acknowledge a grant from the University of Massachusetts/ Amherst Computer Center which made the programming and computations of this research possible.

References

Bean, Judy A. (1975). Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples: An Empirical Distribution. Data Evaluation and Methods Research. NCHS. Series 2, Number 65. DHEW Publication No. (HRA) 75-1339.

Frankel, Martin R. (1971). Inference from Survey Samples. Ann Arbor: Institute for Social Research, The University of Michigan.

Hislop, David A. (1977). Evaluation of the Balanced Half-Sample and Jackknife Variance Estimates for Linear and Combined Ratio Estimates for Non-Normal Populations. Biostatistics-Epidemiology Program Series No. 77-6. University of Massachusetts, Amherst.

Lemeshow, Stanley and Epp, Robert (1977). Properties of the Balanced Half-Sample and Jackknife Variance Estimation Techniques in the Linear Case. Communications in Statistics A, Vol. 6, Issue 13.

Lemeshow, Stanley A. and Levy, Paul S. (1977). Estimating the Variance of Ratio Estimates in Complex Sample Surveys with Two Primary Sampling Units per Stratum - A Comparison of Balanced Replication and Jackknife Techniques. Submitted for publication.

Marsaglia, G., Ananthanarayanan, K. and Paul, N. (1973). How to Use the McGill Random Number Package 'SUPER-DUPER,' four page typed description, School of Computer Sciences, McGill University, Montreal, Quebec.

McCarthy, Philip J. (1966). Replication. An Approach to the Analysis of Data from Complex Surveys. Vital and Health Statistics. NCHS. Series 2, No. 14.

_____ (1969). Pseudoreplication: Half-Samples. Review of the International Statistical Institute, Vol. 37, No. 3: 239-264.

_____ (1969). Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique. Vital and Health Statistics. Series 2, No. 31.

Plackett, R.L. and Burman, J.P. (1946). The Design of Optimum Multifactoral Experiments. Biometrika, Vol. 33 (Pt. IV): 305.325.

Quenouille, M.H. (1956). Notes on Bias in Estimation. Biometrika, Vol. 43: 353-360.

TABLE 1. Absolute Relative Bias, ARB, as estimated from a sampling experiment with m=1000 trials. Values are given by distribution and spread for $\hat{V}_I(\hat{R})$, I=B1, B2, J1, J2, J3, for n=2.

Absolute Relative Bias

Numerator* Distribution and Spread		B1	B2	J1	J2	J3
		Denominator Distribution χ^2_n No spread				
χ^2_n	High	.175247E 00	.217068E 00	.355795E-01	.454862E-01	.217410E-01
N	No	.248767E 00	.319661E 00	.611118E-01	.803036E-01	.409610E-01
U	No	.399185E-01	.928543E-01	.101506E-01	.871287E-01	.115853E-01
F	No	.197346E 00	.226171E 00	.904778E-01	.981040E-01	.805292E-01
		Denominator Distribution $F_{(\nu_1,\nu_2)}$ No spread				
χ^2_n	High	.372152E 00	.525033E 00	.518855E-01	.875958E-01	.203471E-01
N	No	.494171E 00	.695482E 00	.496440E-01	.937074E-01	.708737E-02
U	No	.466603E 00	.658083E 00	.442817E-01	.873108E-01	.114631E-01
F	High	.464392E 00	.588241E 00	.118826E 00	.145909E 00	.805166E-01
		Denominator Distribution U(a,b) No spread				
χ^2_n	High	.670399E-02	.762607E-02	.129730E-02	.994899E-03	.180999E-02
N	No	.731554E-01	.765164E-01	.490608E-01	.501436E-01	.483511E-01
U	No	.903523E-01	.925498E-01	.801824E-01	.809013E-01	.795514E-01
F	No	.423911E-03	.106765E-02	.223983E-02	.202627E-02	.263540E-02

* χ^2_n: Chi square U: uniform(a,b)
N: normal(μ,σ) F: $F_{(\nu_1,\nu_2)}$

TABLE 2. Variance of the variance estimates of the combined ratio estimate from a sampling experiment, $\tilde{V}[\hat{V}_I(\hat{R})]$, I=B1, B2, J1, J2, J3. Values are given by distribution of the random variables and spread for n=2.

$\tilde{V}[\hat{V}_I(\hat{R})]$

Numerator* Distribution and Spread		B1	B2	J1	J2	J3
		Denominator Distribution χ^2_n No spread				
N	No	.226662E 01	.274572E 01	.120096E 01	.128264E 01	.110110E 01
U	No	.864569E 02	.103471E 03	.403767E 02	.426828E 02	.380012E 02
F	No	.242117E-04	.255654E-04	.197749E-04	.200323E-04	.194153E-04
χ^2	High	.126978E-01	.142676E-01	.750100E-02	.775114E-02	.722465E-02
		Denominator Distribution $F_{(\nu_1,\nu_2)}$ No spread				
N	No	.290601E 06	.416092E 06	.731203E 05	.838236E 05	.580908E 05
U	No	.968188E 07	.138746E 08	.291988E 07	.354533E 07	.231239E 07
F	High	.148071E 01	.173652E 01	.768299E 00	.706854E 00	.726055E 00
χ^2	No	.115364E 04	.157357E 04	.297568E 03	.331110E 03	.250521E 03
		Denominator Distribution U(a,b) No spread				
N	No	.118490E-06	.119759E-06	.103981E-06	.104347E-06	.103793E-06
U	No	.132358E-04	.133280E-04	.126253E-04	.126540E-04	.126062E-04
F	No	.265109E-08	.265154E-08	.264356E-08	.264371E-08	.264328E-08
χ^2	High	.580932E-07	.582082E-07	.565221E-07	.565594E-07	.565493E-07

* N: Normal(μ,σ), U: Uniform(a,b), F: $F_{(\nu_1,\nu_2)}$, χ^2: χ^2_n

TABLE 3. Absolute Relative Bias, ARB, as estimated from a sampling experiment with m=1000 trials. Values are given by distribution for $\hat{V}_I(\hat{R})$, I=B1, B2, J1, J2, J3, n=10 and No spread.

Absolute Relative Bias

Distribution*	B1	B2	J1	J2	J3
U/U	.702727E-01	.707533E-01	.686316E-01	.687913E-01	.684745E-01
χ^2/U	.507675E-01	.509781E-01	.482317E-01	.483012E-01	.481224E-01
F/F	.254325E-01	.431291E-01	.274835E-01	.220894E-01	.334131E-01
χ^2/F	.481465E-01	.229696E-01	.116021E 00	.108459E 00	.124244E 00
U/F	.593404E-01	.929300E-01	.203678E-01	.101377E-01	.304936E-01
χ^2/χ^2	.725569E-02	.115231E-02	.196865E-01	.176952E-01	.217388E-01
U/χ^2	.573058E-01	.683600E-01	.300167E-01	.335846E-01	.269165E-01
F/χ^2	.546991E-02	.506174E-01	.761548E-01	.748622E-01	.777854E-01
N/N	.796754E-02	.790266E-02	.841999E-02	.839838E-02	.844212E-02

* U/U: Uniformly distributed variable in both numerator and denominator, U(100,150).

χ^2: Chi Square n=10 df, F: $F_{(6,14)}$, N: Normal(50,5).

TABLE 4. Expected value of the estimate of the variance of the combined ratio, $\tilde{E}[\hat{V}_I(\hat{R})]$, I=B1, J3 and the Target value for those estimates, $\tilde{V}(\hat{R})$. Results of a sampling experiment are given by distribution for No spread and n=10.

Distributions*	$\tilde{V}(\hat{R})$	$\tilde{E}[\hat{V}_I(\hat{R})]$ B1	J3
U/U	.887925E-03	.950520E-03	.948723E-03
χ^2/U	.464299E-04	.487870E-04	.486642E-04
F/F	.427233E-01	.438099E-01	.412958E-01
χ^2/F	.213145E 01	.202883E 01	.186663E 01
U/F	.258163E 03	.243701E 03	.236270E 03
χ^2/χ^2	.140054E-01	.139037E-01	.137009E-01
U/χ^2	.118276E 01	.125054E 01	.121460E 01
F/χ^2	.389343E-03	.368046E-03	.354058E-03
N/N	.133671E-03	.132606E-03	.132543E-03

* U/U: Uniformly distributed variable in numerator and denominator, U(100,150).

χ^2: Chi Square n=10 df, F: $F_{(6,14)}$, N: Normal(50,5).

TABLE 5 Variance of the variance estimates of the combined ratio estimates from a sampling experiment, $\tilde{V}[\hat{V}_I(\hat{R})]$, I=B1, B2, J1, J2, J3. Values are given by distribution for No spread and for n=10.

$\tilde{V}[\hat{V}_I(\hat{R})]$

Distributions*	B1	B2	J1	J2	J3
U/U	.613459E-06	.614459E-06	.614383E-06	.614729E-06	.614092E-06
χ^2/U	.147141E-08	.147224E-08	.145501E-08	.145528E-08	.145268E-08
F/F	.195193E-02	.205731E-02	.161293E-02	.163940E-02	.158894E-02
χ^2/F	.401505E 01	.441704E 01	.303707E 01	.313801E 01	.291710E 01
U/F	.667917E 05	.747136E 05	.478116E 05	.496628E 05	.457670E 05
χ^2/χ^2	.149780E-03	.153090E-03	.139113E-03	.140117E-03	.138225E-03
U/χ^2	.106813E 01	.110353E 01	.923549E 00	.940041E 00	.918151E 00
F/χ^2	.171114E-06	.175258E-06	.153689E-06	.154274E-06	.153101E-06
N/N	.123622E-07	.123651E-07	.123188E-07	.123197E-07	.123179E-07

* U/U: Uniformly distributed variable in numerator and denominator, U(100,150).

χ^2: Chi Square, n=10 df, F: $F_{(6,14)}$, N: Normal(50,5).