UNIVERSITY OF CALIFORNIA

Los Angeles

An Evaluation of the Performance of the Balanced Half-Sample and Jackknife Variance Estimation Techniques

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Biostatistics

by

Stanley Lemeshow

1976

Biomed
W4C
L54re
1976
copy 2

The dissertation of Stanley Lemeshow is approved, and it is acceptable in quality for publication on microfilm.

John M. Chapman
John M. Chapman

Virginia A. Clark

Robert Epp
Robert Epp

Max R. Mickey, Jr
Max R. Mickey, Jr.

F. J. Massey Jr
F. J. Massey, Jr., Committee Chairman

University of California, Los Angeles

1976

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Professor Frank J. Massey, Jr., from whom I have learned so much since beginning my studies at U.C.L.A. His great patience, understanding, and insight were greatly appreciated throughout the long hours he spent working with me on this dissertation.

I am very much indebted to Dr. Robert Epp for his many valuable suggestions and observations, as well as to Dr. Virginia Clark for her encouragement throughout my years at U.C.L.A. and her helpful comments during the preparation of this document. Sincere thanks also goes to Dr. Max R. Mickey and to Dr. John M. Chapman for their interest and assistance in this research.

I owe a special measure of gratitude to Dr. Peter V.V. Hamill of the National Center for Health Statistics for introducing me to the HES, for motivating my continued interest in it, and for offering much good advice throughout our association. I also thank the staff members of the NCHS for their interest in this research and their willingness to assist me throughout it.

I would also like to thank my fellow students in the biostatistics division at U.C.L.A. for their reactions and comments pertaining to certain aspects of this investigation. I am especially grateful to Linda Bundick for typing this manuscript despite her extremely busy schedule.

Most importantly, I would like to thank my wife, Elaine, for her love, understanding and encouragement which were always so important to me.

Computing assistance was obtained from the Health Sciences Computing Facility, U.C.L.A., supported by NIH Special Resources Grant RR-3.

VITA

January 29, 1948 --- Born, Brooklyn, New York

1964-1969 --- B.B.A., City College of New York

1969-1970 --- M.S.P.H., University of North Carolina

1970-1972 --- U.S. Public Health Service, Analytical Statistician, National Center for Health Statistics, Rockville, Md.

1972-1976 --- Teaching Assistant, Division of Biostatistics, School of Public Health, University of California, Los Angeles, California

Fields of Study

Studies in Biostatistics:

Professors A. Afifi, P. Chang, V. Clark, W. Dixon, O. Dunn, D. Hosmer, F. Massey, Jr.

Studies in Mathematical Statistics:

Professors R. Epp, S. Port, T. Ferguson

Studies in Epidemiology:

Professors J. Chapman, R. Detels

Publications

An application of multivariate analysis to complex sample survey data, *Journal of the American Statistical Association,* Vol. 67, No. 340, pp. 780-782, December, 1972, with Gary G. Koch.

Skinfold thickness in a national probability sample of U.S. males and females aged 6 through 17 years, *American Journal of Physical Anthropology*, pp. 321-324, May, 1974, with F. E. Johnston and P.V.V. Hamill.

ABSTRACT OF THE DISSERTATION

An Evaluation of the Performance of the
Balanced Half-Sample and Jackknife Variance Estimation Techniques

by

Stanley Lemeshow

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 1976

Professor Frank J. Massey, Jr., Chairman

In recent years, methods for approximating the variances of estimates computed from complex sample surveys have received increased attention since the precise expressions for such variances are usually unknown. The balanced half-sample and jackknife are two such methods.

Considered in this dissertation are two balanced half-sample estimates $(\hat{V}_{B1}(\hat{W}),\hat{V}_{B2}(\hat{W}))$ and three jackknife estimates $(\hat{V}_{J1}(\hat{W}),\hat{V}_{J2}(\hat{W}),$ $\hat{V}_{J3}(\hat{W}))$ of the variance of $\hat{W}$, where $\hat{W}$ is the estimate of some parameter of interest. The particular situation considered is one in which the population is subdivided into L strata of known sizes or weights from which n_i observations are selected. Cases considered are such that, for each stratum, the n_i observations may be collapsed into k equal sized groups. For the balanced half-sample technique, $k = 2$ groups of $r_i = n_i/2$ observations are always established, whereas for the jackknife method $2 \leq k \leq n_i$, each group containing $r_i = n_i/k$ observations.

The variance estimation techniques are introduced and theorems

are presented for the case where $\hat{W}$ is a linear combination of the observations. In this linear situation the three jackknife estimators are identical for all values of k and L and are unbiased estimates of the true variance of $\hat{W}$. When L is a multiple of 4 and $k = 2$, the value of $\hat{V}_{B2}(\hat{W})$ is identical to the three jackknife values but $\hat{V}_{B1}(\hat{W})$ differs from the rest and is negatively biased as an estimate of the true variance. However, its variance and mean square error is less than the corresponding values of any of the other estimates. When L is not a multiple of 4, all five estimates are identical when $k = 2$. The variances and mean square errors of the jackknife estimates decrease as k increases. With $k > 2$, the variances and mean square errors of the jackknife estimates are less than the corresponding values for either of the balanced half-sample estimates.

As examples of non-linear $\hat{W}$, we consider the combined ratio estimate, the estimate of the slope in a linear regression situation, and the estimate of the correlation coefficient. The five variance estimators are evaluated by means of sampling experiments using computer generated data from populations with specified values for the strata parameters. The sampling experiment also produces estimates of the target variances so that bias and mean square error can be assessed.

The sampling experiments for the combined ratio estimate, $\hat{R}$, indicate that either the balanced half-sample or jackknife methods may be effectively used for the estimation of variances. There was close correspondence to theorems which were derived for the linear case. $\hat{V}_{B1}(\hat{R})$ was shown, on the average, to underestimate the target variance when L was a multiple of 4. However, this estimate had lower variance and mean square error than the other four estimates when $k = 2$. With $k > 2$, the variance and mean square error of the jackknife variance estimates were smaller than the corresponding values of either balanced half-sample estimate.

Sampling experiments for the slope, β, and correlation coefficient, ρ, demonstrate that it is necessary to select a moderately large number of observations from each stratum in order to produce acceptable variance estimates. The jackknife estimates for $k > 2$ are again shown to have smaller variance and mean square error than the corresponding balanced half-sample estimates. As was the case for the linear estimate or combined ratio estimate, $\hat{V}_{B1}(\hat{\beta})$ and $\hat{V}_{B1}(\hat{\rho})$ are, on the average, underestimates of the target variance but have smaller variance and mean square error than the other estimates when $k = 2$ and L is a multiple of 4. However, contrary to the linear and combined ratio situations, there does not exist the high degree of similarity among the other variance estimation techniques. Using any of these methods, satisfactory estimates were made of the variance of the estimated slope or correlation coefficient provided that a large enough sample was selected from each stratum. Because of the flexibility afforded by the jackknife method to establish $k > 2$ groups per stratum, a jackknife method would be preferred over a balanced half-sample method.

Also considered was the use of the variance estimation procedures when each sample individual is assigned a unique statistical weight. These weights presumably bring the demographic breakdown in the sample into closer alignment with known demographic characteristics of the

population. Two weighting schemes are considered for the purpose of estimating the variance of the estimated population mean. One method uses the same weights for estimates based on subgroups of the sample individuals (a half-sample is an example of such a subgroup) as are used for the estimate based on the entire sample. The other method assigns weights to the individuals in the subgroup so that the resulting estimate better reflects the population parameter. Examples are given of situations in which variance estimates produced using the latter method had markedly smaller bias and variance.