New Zealand Statistician, Vol. 11, No. 2. 1976

APPROXIMATE ESTIMATION OF THE STANDARD ERRORS OF COMPLEX STATISTICS BASED ON SAMPLE SURVEYS

David R. Brillinger
University of California, Berkeley and University of Auckland

Nowadays quite complex statistics are routinely computed for sample surveys with nonelementary probability structure. In order to be able to interpret these statistics, it is necessary to have some idea of their sampling variability. In this paper we survey some general methods of estimating standard errors including: balanced repeated replication, the jackknife, Taylor series expansion and elementary perturbation. In an appendix we provide a justification of the jackknife procedure for M-estimates.

INTRODUCTION

To set the scene, let me describe a continuing survey that I am presently associated with. The National Assessment of Educational Progress (NAEP) is an information gathering project which surveys the educational attainments of 9-year-olds, 13-year-olds, 17-year-olds and young adults (aged 26-35) in ten different learning areas. The areas are periodically reassessed in order to measure educational progress or regress. Exercises are prepared by test developers, groups of exercises are collected into packages and the packages are administered to probability samples of various populations. Each exercise is attempted by about 2500 individuals and approximately 100,000 persons participate annually. Not all exercise results are released for publication, because some will be readministered in the future in order to measure change.

The results are reported in various categories including: age, region, sex, colour, level of parental education, size-and-type of community. The basic measure of educational achievement is the percent of people who can perform a task, or group of tasks. It is especially interesting to measure "group effects"—the difference between the percentage achieved by a group and the corresponding percentage in the nation.

Because a population is not homogeneous across a nation, observed group percentages can be misleading; for example, parental education effects may be masquerading as size-and-type of community effects. In consequence, balanced group effects are computed by fitting a linear model and generally used for interpretation. Balancing is carried out with respect to the categories: region, sex, size-and-type of community, colour, level of parental education. This procedure is meant to "make" the mixture of characteristics in each group, the same as for the whole population. See Maxwell and Jones (1976).

As an example of a specific question and results we mention the following question put to 9 year olds in 1970 and 1973. "Putting sand and salt together makes : 1. a chemical; 2. a compound; 3. an element; 4. a mixture; 5. a solution; 6. I don't know." The answer sought was 4. "a mixture". The results found were

---

This research was supported by the J.S. Guggenheim Memorial Foundation and the National Science Foundation Grant MCS76-06117.

	1970	1973	Change
All 9 year olds	61.5%	61.7%	+ .2%
South East 9 year olds	55%	59%	+ 4%
Difference	−7%	−3%	+ 4%

Altogether 92 science exercises were repeated in the 1970 and 1973 surveys. The average change measured was −1.7%. The estimated standard error of this last statistic was .6%, as determined by a procedure to be described later.

The sampling of the survey began with a selection of 2 primary units (PSU) from each of 104 strata. Stratification variables were: region, size of community, socio-economic-status. The primary units consisted of clusters of schools within selected listing units. The listing units were counties or parts of counties. Within each selected listing unit, a cluster of schools was selected. Students were then selected at the schools. There were about 12 packages of questions, each administered to groups of about 12 students at a time. About 150 students were required in each PSU. Each package of questions appeared in each PSU.

Correlations are introduced into the data through students being at the same school and questions appearing in the same package, among other things. A further complication occurs because questions change packages from survey to survey.

It is not at all clear what the best summary statistics are for a single survey or change. The questions are not meant to be a sample of questions. Atypical results can creep in through local conditions. The distributions observed are not normal. Means, medians and bi-weights are computed generally. These in turn are functions of ratio and regression estimates involving unequal weights. A basic statistic reported is the median of the difference observed between a subgroups performance and the national performance for exercises of a similar sort.

We now turn to a discussion of some estimates of the standard errors of the complicated statistics described above. We begin with some formulas for a particular age group and exercise. Set

Y_{hijk} = 1 if student k of school j of PSU i in stratum h answers correctly

= 0 otherwise

The estimate of the proportion correct is

$$P = \Sigma\ W_{hijk} Y_{hijk} / W$$

where the sum is over all subscripts,

$$W_{hijk} = 1\ /[\text{Prob(PSU hi)} \cdot \text{Prob(Sch j} \mid \text{hi)} \cdot \text{Prob(child k} \mid \text{hij)}]$$

and W denotes the sum of all the W_{hijk}. We are also interested in subgroup P-values (subgroup performances) and delta-P-values (differences between a subgroup's performances and the national performance). More complicated statistics are required for estimating the balanced effects.

Variances for P— and delta—P— values have been estimated as follows:

PSU totals are taken as basic statistics and there are 2 PSU's per stratum. For h = 1, ...,H and i = 1,2

$$Y_{hi} = \sum_j \sum_k W_{hijk}\, Y_{hijk} \qquad Y = \sum_h \sum_i Y_{hi}$$

$$W_{hi} = \sum_j \sum_k W_{hijk} \qquad W = \sum_h \sum_i W_{hi}$$

then

$$P = \sum_h (Y_{h1} + Y_{h2}) \,/\, \sum_h (W_{h1} + W_{h2}) = Y \,/\, W$$

Next replicate P–values are computed as

$$P_{-h1} = (Y - Y_{h1} + Y_{h2}) \,/\, (W - W_{h1} + W_{h2})$$

(and an analagous formula for P_{-h2}) corresponding to discarding a PSU and replacing its values by its companions. Denote the mean of the P_{-hi} by $P_{.}$.

The jackknife P–value is then defined to be

$$P_J = (1 + H)P - H\ P_{.}$$

The jackknife estimate of the variance of P and P_J is now

$$\sum_h (P_{-h1} - P_{-h2})^2/4$$

THE JACKKNIFE

The above is an example of the following general procedure for dealing with data based on a design of H strata having n_h PSU's in stratum h. Let A denote a statistic based on all the data. Let A_{-hi} be the same statistic based on all the data, but with the hi values replaced by missing value estimates of the same and let $A_{-h.}$ be the mean of the A_{-hi} . The jackknife estimate is

$$A_J = A + \sum_h (n_h - 1)\,(A - A_{-h.})$$

and the variance of it and A may be estimated by

$$\sum_h (1/n_h - 1/N_h)\, s^2_{-h}$$

where

$$s^2_{-h} = (n_h - 1) \sum_i (A_{-hi} - A_{-h.})^2 \ .$$

In the case of a multidimensional statistic, A, the last expression is replaced by the sample covariance matrix

$$s^2_{-h} = (n_h - 1) \sum_i (A_{-hi} - A_{-h.})\,(A_{-hi} - A_{-h.})^T \ .$$

The above jackknife procedure is one general method of constructing approximate estimates of standard errors. We now describe other procedures that have found some use.

SIMPLE REPLICATED (INTERPENETRATING) SAMPLES

In this procedure one constructs a number of disjoint replicates of the survey, that mimic the structure of the overall design. Its great advantage is that it provides one with a number of independent estimates of the parameter of interest, and classical variance

estimation formulas may be employed. Its great disadvantage is that individual replicates may not be able to be based on sufficient data.

This type of procedure has a long history as the following quote from Guy (1839) illustrates, "where our means or instruments of observation are imperfect, or the things observed differ widely in numerical value, a large number of observations is necessary, in other cases a smaller number can suffice. Perhaps the best rule which can be given for ascertaining whether the observations which we have collected are sufficiently numerous to yield a true average, is to divide the whole number of observations into groups of equal size and compare them the one with the other ..."

TAYLOR EXPANSION (LINEARIZATION)

Suppose that the statistic of interest has the form

$$A = A(T_{hi}\ ;\ i = 1, \ldots, n_h\ ;\ h = 1, \ldots, H) = A(T_{hi})$$

$$= A(t_{hi}) + \Sigma\ \Sigma\ (T_{hi} - t_{hi})^T A_h(t_{hi}) + \ldots$$

where T_{hi} is a p–dimensional vector of sample values, t_{hi} is the associated population mean value, where A_h denotes the (vector-valued) derivative of A with respect to values of stratum h and the error in the final expression is assumed to be not too large.

Then the classic result (of Gauss) is that the variance of A may be estimated by

$$\sum_h A_h(T_{hi})^T\ s_{hh}\ A_h(T_{hi})$$

assuming derivatives are equal within a stratum and with

$$s_{hh} = \Sigma\ (T_{hi} - T_{h.})\ (T_{hi} - T_{h.})^T/(n_h - 1)\ .$$

Forming this estimate involves a knowledge of derivatives.

This estimate may be related to the jackknife as follows; suppose

$$(\ T_{hi}\ ;\ i = 1, \ldots, n_h\ ) = \sum_i S_{hi}/n_h = X_{h.}$$

then $A = A(X_{1.}, \ldots, X_{H.})$. It is clear that

$$X_{-hi} = \sum_{j \neq i} X_{hj}\ /(n_h - 1) = X_{h.} + (X_{h.} - X_{hi})\ /\ (n_h - 1)$$

and so

$$A_{-hi} = A(T_{hi}) + (X_{h.} - X_{hi})^T\ A_h(X_{h.})\ /(n_h - 1) + \ldots$$

$$A_{-h.} = A(T_{hi}) + \ldots$$

giving

$$s^2_{-h} = A_h\ (X_{h.})^T \sum_h\ (X_{hi} - X_{h.})\ (X_{hi} - X_{h.})^T\ A_h(X_{h.})\ /(n_h - 1) + \ldots$$

that is, approximately the variance of the Taylor expansion method.

REPEATED REPLICATION

Suppose that 2 PSU's are selected per stratum ($n_h = 2$). Form a replicate of the whole sample by randomly selecting one of the 2 PSU's in each of the H strata. Let F denote the set of values in the whole sample and F_j refer to those of the j–th replicate. In the repeated replication procedure, the variance estimate employed is

$$\sum_{j=1}^{J} (A(F) - A(F_j))^2 / J$$

assuming J replicates have been selected.

This procedure may be related to the Taylor procedure as follows:

suppose $(T_{hi} ; i = 1,2) = (X_{h1} + X_{h2})/2 = X_{h.}$. Then $F = (X_{1.}, ...,X_{H.})$

and $F_1 = (X_{11}, ...,X_{H1})$ say . It follows that

$$A(F_1) = A(F) + \Sigma (X_{h1} - X_{h.})^T A_h(X_{h.}) + ...$$

and

$$(A(F_1) - A(F))^2 = \Sigma \Sigma A_h(X_{h.})^T(X_{h1} - X_{h.}) (X_{j1} - X_{j.})^T A_h(X_{j.}) + ...$$

The Taylor approximation to the variance appears once again when one averages over many selected replications for the averaging of $(X_{h1} - X_{h.}) (X_{j1} - X_{j.})^T$ leads to s_{hh} if $j = h$ and to 0 if $j \neq h$. We remark that McCarthy has proposed that the replicates be selected in a balanced manner, for example in accordance with pieces of Plackett-Burman designs.

SIMPLE PERTURBATION

Suppose that the estimate is given by $A = A(T_1, ...,T_H)$ where an estimate s_{hh} of the covariance matrix is available and where the T_h are approximately uncorrelated. For e small, form the perturbed statistics

$$A_j = A(T_1 + eB_1 C_{1j}, ...,T_H + eB_H C_{Hj}) \qquad j = 1, ...,J$$

where $C = [C_{hj}]$ is a design matrix satisfying $CC^T = 1$, the identity, and $B_h B_h^T = s_{hh}$. Now estimate the variance of A by

$$\sum_h \sum_j \sum_k C_{hj}^T C_{hk} (A_j - A) (A_k - A) / e^2$$

Making a Taylor expansion, this last expression may be seen to be

$$\sum_h A_h(T_1, ...,T_H)^T s_{hh} A_h (T_1, ...,T_H) + ...$$

that is the result of the Taylor procedure.

This method is seen to relate to the jackknife procedure when $n_h = 2$ via the correspondence $B_h = (X_{h1} - X_{h2}) / 2^{1/2}$, $C_{hj} = \pm 2^{-1/2}$ or 0, and $e = 1$ for now $X_{h.} + eB_h C_{hj} = X_{h1}$, X_{h2} or $X_{h.}$. Likewise it relates to repeated replication by taking $C_{hj} = \pm 2^{-1/2}$.

A COUNTEREXAMPLE TO THE JACKKNIFE

The jackknife is not always a justifiable procedure. Consider a simple random sample of size n taken from a population with median m . Suppose $n = 2k + 2$ is even and that the sample values are ordered with x_i the i—th largest. The median may be estimated by $(x_{m+2} + x_{m+1}) /2$. The jackknife estimate of variance using groups of size n—1 may be determined as

$$s^2_{-h} = n(x_{m+2} - x_{m+1})^2/8.$$

The variance of the sample median is known to be asymptotically $1/[4n\ f(m)^2]$ where $f(x)$ is the density of the x's. The variate s^2_{-h} may be seen to be tending to $1/[16n\ f(m)^2]$ which is too small. The difficulty here seems to be caused by the poor degree of approximation of the sample median by a linear statistic. The jackknife procedure may be reasonable if the sample is split into r groups of size (r−1)s, where n = rs and s is larger than r . (This particular counterexample is due to Lincoln Moses.)

Another difficulty that can sometimes arise with the jackknife is that the statistic of interest may become singular when based on the reduced number of observations.

CONCLUDING REMARKS

In conclusion we mention that there are connections between the procedures discussed in this paper and cross-validation analysis, the sensitivity analysis of control engineers, procedures for computing system performance or "tolerancing", Mickey's procedure for constructing unbiased estimates, influence curves and Hartigan's technique.

The procedures would appear to be especially useful in dealing with M—estimates and robust regression estimates. For this reason a brief justification of the jackknife for M—estimates is provided in the Appendix.

REFERENCES

ARVESEN, J. N. (1969). Jackknifing U—statistics. *Ann. Math. Statist.*, 40, 2076 – 2100.

BISSELL, A.F. and FERGUSON, R.A. (1975). The jackknife – toy, tool or two-edged weapon. *The Statistician,* 24, 79 – 100.

CHIBISOV, D.M. (1973). An asymptotic expansion for a class of estimators containing maximum likelihood estimators. *Theory Prob. Appl.,* 18, 295 – 303.

GUY, W.A. (1839). On the value of the numerical method as applied to science, but especially to physiology and medicine. *J. Statist. Soc. London,* 2, 34.

HARTIGAN, J.A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.,* 64, 1303 – 1317.

KISH, L. and FRANKEL, M.R. (1974). Inference from complex samples. *J. Roy. Statist. Soc. Ser. B,* 36, 1 – 37.

MAXWELL, S.E. and JONES, L.V. (1976). Female and male admission to graduate school. *J. Educat. Statist.,* 1, 1–37.

MICKEY, M.R. (1974). Some finite population unbiased ratio and regression estimators. *J. Amer. Statist. Assoc.,* 54, 594 – 612.

MILLER, R.G. (1974). The jackknife – a review. *Biometrika,* 61, 1 – 15.

APPENDIX

A class of complex estimates, coming into common use, is the class of M—estimates. Given sample values $X_1, \ldots, X_n$ these are defined as a statistic, A, providing the minimum value of a penalty function

$$M(X_1,A) + \ldots + M(X_n,A) \quad .$$

Least squares estimates are examples of M—estimates. So are maximum likelihood estimates. If $m(X,A)$ denotes the derivative of $M(X,A)$ with respect to A, then the M—estimate A is a solution of the equation

$$m(X_1,A) + \ldots + m(X_n,A) = 0 .$$

Suppose A_0 is the true value of A, satisfying $E\, m(X,A_0) = 0$. Suppose m' and m'' denote the first and second derivatives of m with respect to A. Let a' and a'' denote $E\, m'(X,A_0)$ and $E\, m''(X,A_0)$ respectively. Then Chibisov (1973) develops the asymptotic expansion

$$A = A_0 - \Sigma\, m(X_i,A_0) / na' + (\Sigma\, m(X_i,A_0) / n)(\Sigma\, (m'(X_i,A_0) - a')) / na'^2 - (\Sigma\, m(X_i,A_0) / n)^2\, a'' / 2a''^3 + \ldots$$

This expansion may be manipulated to find an expression for A_{-i} the M—estimate based on the data excluding the observation X_i. At this point it is important to note the expansion is being carried out in terms of polynomials of the means $\Sigma\, m(X_i,A_0) / n$, $\Sigma\, (m'(X_i,A_0) - a') / n$, $\Sigma\, (m''(X_i,A_0) - a'') / n$, ... and to note that Arvesen (1969) shows how to justify the jackknife for regular functions of means. We can now set down,

Theorem. Suppose the conditions of the theorem of Chibisov (1973) are satisfied with $k = 3$, $r = 8$, $m = 4$ (in his notation.) Let $A_{-.} = \Sigma\, A_{-i} / n$, $A_J = n\,A - (n-1)A_{-.}$, $s_{-}^2 = (n-1)\, \Sigma\, (A_{-hi} - A_{-h.})^2$. Then $n^{1/2}(A - A_0)$ and $n^{1/2}(A_J - A_0)$ are both asymptotically normal with mean 0 and variance $E\, m(X,A_0)^2 / a'^2$. Also s_{-}^2 tends to $E\, m(X,A_0)^2 / a'^2$ in probability and $n^{1/2}(A_J - A_0) / s_{-}$ is asymptotically standard normal.

This theorem provides a justification for the use of the variance estimate s_{-}^2 for the statistics A and A_J in the case of M—estimates and when one observation is dropped out at a time in forming the jackknifed estimate.