SAMPLE RE-USE

by

William R. Schucany

Technical Report No. SMU-DS-TR-186
Department of Statistics ONR Contract

September 1984

Research sponsored by the Office of Naval Research
Contract N00014-82-K-0207
Project NR 042-479

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

SAMPLE RE-USE

by
William R. Schucany
Department of Statistics
Southern Methodist University
Dallas, Texas 75275

Summary

This paper is a survey of statistical methodologies that can be categorized as involving a reuse of the sample. The historical roots are found in standard error estimation and randomization tests. The modern developments consist of jackknife, bootstrap, cross-validation and adaptive procedures. The article has been prepared as an entry for the Encyclopedia of Statistical Sciences.

INTRODUCTION

The meaning of the term "sample re-use" rests on the idea that a sample is used to evaluate a statistic and then may be used again to assess or improve the performance of the basic statistical procedure. Some applications of the concept fit quite obviously into this sample re-use framework, while some others that belong in this category are not quite so apparent. The varieties range from the overt resampling that is characteristic of the bootstrap* to more subtle manifestations of the principle such as a cross-validatory choice of bandwidth for a kernel density estimator.

STANDARD ERROR AND RERANDOMIZATION

Several of the cornerstones of statistical inference that were laid by Karl Pearson*, 'Student' (see GOSSET, WILLIAM SEALY). R.A. Fisher* and E.J.G. Pitman may be viewed as early applications of the notion that the data set could profitably be used a second time. The fundamental concept of the sampling distribution of a sample mean led almost immediately to the process of re-using the sample to estimate the standard error. Strictly speaking the first use of a random sample X_1, $X_2, \ldots, X_n$ would be to compute $\bar{X}$. The re-use would be to estimate σ^2/n with the unbiased estimator s^2/n. In a very real sense this represents a return to the sample to assess an average squared deviation about the statistic which was calculated initially.

A second landmark development that fits into this same framework is the method of randomization* due to Fisher [7] and Pitman [18]. The essence of this technique, which yields exact tests of hypothesis in some settings, is a permutation argument first conceived by Fisher. It is founded upon the idea that given a particular random sample it is often the case that the algebraic signs, labels or positions in a table are features of the data set that would tend to differ under competing hypotheses even if the magnitudes of all of the numerical outcomes were not different. The logical basis for the test is that given the validity of the null hypothesis, each of these different configurations of signs or positions would be equally likely. Consequently, this classical testing procedure utilizes a null distribution generated by the set of all reevaluations of a statistic, T, for a specific reference set of permutations. This set of values obtained by re-using the sample provide a frame of reference that allows one to judge whether the originally observed value of T is extreme.

To illustrate this approach to the calculation of such a conditional p-value consider testing the hypothesis, H, of independence of absolutely continuous random variables X and Y based upon a random sample of size n from their joint distribution. Given the observed nx2 table the correlation coefficient, r, is calculated. Under H each of the n! permutations of the column of y values are equally likely and corresponding to for each there would be a distinct value of r. If all n! outcomes were tabulated, this would represent the null distribution. Conceptually one would have enumerated all of the potential realizations that are more extreme in some predetermined sense than the actual outcome. In practice the calculations of this p-value may be formidable if the size of the data set is large. Monte Carlo sampling techniques to approximate the procedure were first discussed by Dwass [4]. For recent computational developments, extensions to interval estimation and a good review see Gabriel, et al. [8]. The relatively recent and appropriate nomenclature that identifies these as"rerandomization" procedures is due to Brillinger, et al. [2].

JACKKNIFE AND BOOTSTRAP

In situations that are more complex than using $\bar{X}$ to estimate the population mean the estimators of standard error are typically more complicated than the sum of squared deviations of individual observations. In many such cases it is still feasible to use a direct estimate of the standard error through expressions derived for the pertinent parametric family. Occasionally, these must be only large sample approximations, The primary function of the jackknife* and bootstrap* is to provide nonparametric estimates of standard errors. It follows that approximate tests and confidence intervals can be and are constructed from these in nonparametric settings and for some parametric problems when an analytical solution is difficult. Even though these procedures often also yield an improved estimator, for example one with a reduced bias*, the main emphasis is on the re-use of the sample to obtain estimates of the variability that is present in the sampling distribution of a statistic.

The related topics of an empirical influence function* and the delta method* are examined in systematic treatments by Efron [5,6]. Still another resampling plan that was introduced by McCarthy [15] is entitled balanced half-sample pseudo-replication*. This procedure is employed mainly for highly stratified survey samples. See JACKKNIFE METHODS for detailed descriptions of jackknife, bootstrap and pseudo-replication.

CROSS-VALIDATION AND PREDICTIVE SAMPLE RESUSE

Cross-validation as a means of assessing the quality of statistical predictions is a simple idea. Stone [19] cites examples from the 1930's. The essence of the process is a partitioning of the data into two subsamples. One portion is used to construct the predictor, the other to measure the performance of the rule when it is applied to data that were not used in selecting the specific prediction rule. At this stage of development sample re-use is not in evidence. For the simple situation involving a construction sample and a separate validation sample the approach does not actually warrant the name cross-validation either. This scheme has an advantage over the naive alternative involving direct resubstitution, which is usually optimistically biased. Clearly, any prediction rule should appear to do better with predictions of cases that were used to construct that rule than it could be expected to do generally.

The next stage of development occurred during the 1960's. Mosteller and Wallace [17] suggested the approach and the first distinct appearances of cross-validation were published by Lachenbruch and Mickey [13] and Mosteller and Tukey [16]. The heart of the procedure is that one individual case is set aside as the validation portion and the estimation or construction is carried out with the subset of (n-1) cases. After the performance of this rule is tested upon the held-out case, the process is repeated for each possible case. In other words, in turn the sample has been partitioned and re-used in all n possible ways of applying a leave-out-one rule. The applications of these early developments were to discriminant analysis*, but the basic ingredient is that of a prediction for which there is an observable measure of error or error rate.

In the mid-1970's another extension surfaced. Both Stone [19] and Geisser [9] raised the issue of the choice of the predictor in addition to assessing predictor performance. They then proposed a merger of the cross-validation assessment component and the choice component. Subsequent reference to cross-validation tends to pertain to this major refinement which has shifted the methodology from assessment to construction. Wahba and Wold [20] formulated and used cross-validation to select the degree of smoothing in spline* fitting. Some asymptotic optimality results have been established for spline smoothing with generalized cross validation of Craven and Wahba [3].The application of cross-validation to kernel density estimation has yielded a mixture of encouraging results with occasional failures. Hall [10,11] summarizes these findings.Properties of cross-validated nearest neighbor nonparametric regression are given by Li [14]. Early independent application of this same principle, which is called PRESS in the regression setting,can be found in Allen [1]. This alternative to least squares basically selects predictor variables that minimize the prediction error sum of squares. For more discussion of predictive aspects in general see PREDICTIVE ANALYSIS. These sample re-use techniques orginated from attempts to validate assumptions or particular statistical models. As such, the calculation and display of residuals qualify as among the earliest and most widely used applications. Yet, residuals as regression diagnostics do not exhibit the two primary ingredients of cross-validation, namely a summary assessment and a choice of the specific prediction model. However, deleted residuals do require calculations that parallel those in the cross-validation methodology.

ADAPTIVE PROCEDURES

As another distinct application of sample re-use the topic of adaptive estimation actually appears to warrant the term "pre-use" instead. Again, the general idea can be discussed in the context of the specific problem of estimating the mean of a distribution. If one were reasonably certain that the data were drawn from a normal distribution then $\bar{X}$ would be an efficient* estimator. On the other hand if the probability model were known to be the Laplace distribution, the median would be optimal in this sense. Clearly, if one could reliably decide about the population actually sampled, assuming for the sake of simplicity that it is one or the other of these two parametric families, then one could select the better measure of location. So in its most elementary form adaptive* estimation uses the sample first to make a choice and then uses it a second time to evaluate the appropriate estimator. Hogg [12] summarizes numerous developments along these lines.

Example applications range from a dichotomous choice, e.g., a preliminary-test estimator, to continuously adaptive estimators. An example of the latter is a trimmed mean in which the fraction to be trimmed is estimated from the same data. Both of these procedures have a characteristic two-stage nature. This is in contrast with the simultaneous character of the cross-validatory choice paradigm.

REFERENCES

[1] Allen, D. M. (1974). *Technometrics*, **16**, 125-127.

[2] Brillinger, D. R., Jones, L. V. and Tukey J. W. (1978). *Report to the Secretary of Commerce, Statistical Task Force to the Weather Modification Advisory Board*, U.S. Govt. Printing Office, Washington, D.C.

[3] Craven, P. and Wahba, G. (1979). *Numer. Math.*, **31**, 377-404. Introduces the method of generalized cross-validation to estimate the degree of smoothing with splines.

[4] Dwass, M. (1957). *Ann. Math. Statist.*, **28**, 181-187.

[5] Efron, B. (1981). *The Jackknife, the Bootstrap and Other Resampling Plans.* CBMS Monogr. No. 38, SIAM, Philadelphia.

[6] Efron, B. and Gong, G. (1983). *The Amer. Statist.*, **37**, 36-48. An expository article explaining relationships among several resampling schemes.

[7] Fisher, R. A. (1935). *The Design of Experiments*, Oliver and Boyd, London.

[8] Gabriel, K. R. and Hall, W. J. (1983). *J. Amer. Statist. Ass.*, **78**, 827-836. This and two other articles coauthored by Gabriel in the same issue of the journal all contain the term "rerandomization" in the title.

[9] Geisser, S. (1975). *J. Amer. Statist. Ass.*, **70**, 320-328.

[10] Hall, P. (1982). *Biometrika*, **69**, 383-390.

[11] Hall, P. (1983). *Ann. Statist.*, **11**, 1156-1174.

[12] Hogg, R. V. (1982). *Commun. Statist.*, **11**, 2531-2542.

[13] Lachenbruch, P. A. and Mickey, M. R. (1968). *Technometrics*, **10**, 1-11.

[14] Li, K.C. (1984). *Ann. Statist.*, **12**, 230-240.

[15] McCarthy, P.J. (1969). *Rev. Inst. Statist. Int.*, **37**, 239-264.

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER SMU/DS/TR-186	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* SAMPLE RE-USE		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) WILLIAM R. SCHUCANY		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0207
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University Department of Statistics Dallas, Texas 75275		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-479
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA. 22217		12. REPORT DATE September 1984
		13. NUMBER OF PAGES 9
14. MONITORING AGENCY NAME & ADDRESS*(If different from Controlling Office)*		15. SECURITY CLASS. *(of this report)*
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any putpose of The United States Government.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

Same as above

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This paper is a survey of statistical methodologies that can be categorized as involving a reuse of the sample. The historical roots are found in standard error estimation and randomization tests. The modern developments consist of jackknife, bootstrap, cross-validation and adaptive procedures. The article has been prepared as an entry for the Encyclopedia of Statistical Sciences.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*