Resampling Methods

Herwig Friedl Erwin Stampfer

3rd May 2001

Statistical inference is based generally on some statistics (estimates) that are functions of the data. Their sampling distributions depend on the underlying population and therefore are unknown. Resampling methods offer strategies to estimate or approximate the sampling distribution of a statistic or its characteristics. These include accuracy measures of an estimate like its bias or the mean squared error, or relevant **Quantiles** of a test statistic, among others.

Traditional approaches, e.g. the likelihood method, rely strongly on postulated probability models. Even if they hold, the conclusions are frequently based on asymptotical or approximate properties. The reliability of the statistical analysis depends therefore on the validity of these assumptions and on the sample size.

In resampling we consider the observed data as a representative picture of the entire population. The essential idea then is to make statistical inference based on an artificial resample, which is drawn from the data. Additional to calculating the statistic as a function of the sample, we also consider this statistic evaluated at the resample. The benefit of such a procedure is the knowledge of the resampling probability mechanism used, implying that the characteristics of a resampled statistic are known theoretically. Its distribution is utilized to describe the unknown distribution of the statistic with respect to the population. This has particular advantages for statistics based on small samples, where asymptotics do not work properly.

Different resampling plans result in different resampling techniques. The term plan refers to the set of all possible resamples to be taken and their weighting. There are deterministic and stochastic resampling plans, the former leading deterministically to a set of resamples, the latter to a random set, given by all possible resamples along with probabilities of their occurrence.

We may produce resamples by repeatedly leaving out one observation from the data, known as the **Jackknife resampling**. If the data is of size n, this leads deterministically to a set of n resamples, each consisting of $n-1$ observations. The jackknife was introduced by Quenouille [16], and is a resampling method that aims primarily at the calculation of the bias [17] and the standard error [20] of estimates, without making distributional assumptions. It is nonparametric and easy to apply in a quite general setting.

Miller [15], who has discussed the jackknife in the context of maximum likelihood estimation, U-statistics, regression, stochastic processes and multivariate analysis, gives a review. Efron [4] has compared the **Jackknife resampling** to some alternative resampling schemes. Wu [21] has proposed generalized weighted jackknife variance estimators in the context of regression. Shao and Tu [18] is an excellent theoretical reference discussing many applications.

Cross validation is another technique that is designed especially to select and assess models (prediction rules). Applying the estimated model on new data from the same source can validate it. However, in many cases one is forced to use the same data for model building and assessment. The idea of cross validation is very similar to the jackknife idea with respect to the resampling plan. The data is partitioned into K pieces of a certain length. One piece can now be used to test the model that was trained on the remaining $K-1$ pieces. To remove the effect of a particular division, this is repeated for all K pieces of the data. The results of the K testing procedures are then combined suitably. However, cross validation should not be mixed up with jackknife. Both of these resampling procedures are technically quite different. From a practical point of view, the major difference lies in their applications. Cross validation is used for model selection and assessment, whereas the jackknife provides bias and variance estimates.

Further details can be found in Stone [19]. Cross validation has been used widely in the context of smoothing for purposes of optimal bandwidth selection. It is a quite universal and easy to handle tool, and has therefore been applied especially to complex models, e.g. in **Neural network**, tree structured classifiers and discriminant analysis. Efron and Gong [6] have compared cross-validation with other resampling techniques by citing the respective error rate estimators as an example.

Another resampling procedure is the **Bootstrap resampling**, where the resample consists of n elements that are drawn randomly from the n original data points with replacement. In this case the random resample is defined on the set of all variations of the data with repetitions allowed. This resampling strategy was introduced by Efron [4], and allows also for the assessment of estimators. It has found a large number of applications, ranging from estimating bias or standard errors to calculating confidence intervals and testing hypotheses. The technique has been applied successfully in various fields. Among these, some strategies within generalized linear models have been summarized in Friedl and Stadlober [8]. However, unlike the jackknife the bootstrap is applicable more widely, and gives better results in many situations. The major difference between both methods is that the jackknife works with subsets of data whereas the bootstrap takes into account some special features of the postulated model. Hence, the jackknife may not be as sufficient as the bootstrap, but it is often more robust against misspecifications of the model. Jackknifing is recommended mainly if one needs only a variance estimator.

Papers about the bootstrap at an introductory level are Efron and Gong [6] or Diaconis and Efron [3]. The bootstrap and its relations to other resampling techniques are explored in the monograph by Efron [5] and in Hjorth [12]. Besides the theoretical books of Hall [11] and Shao and Tu [18], there have been published a number of books providing know how for the statistical practitioner. These are Efron and Tibshirani [7], Davison and Hinkley [2], Good [10], and Chernick [1]. All of these include extensive bibliographies, as a guide through the vast literature on this topic.

Permutation test are nonparametric procedures that calculate an attained significance level of a test statistic by comparing it with its resampled values. In contrary to the bootstrap, which is based on resampling with replacement, permutation tests utilize resamples that are drawn without replacement (permutation) from the observations. The distribution of the resampled values is called permutation distribution, and plays the role of the null distribution in parametric testing problems. The applicability of permutation tests relies on the property that some observations are exchangeable under the null hypothesis, whereas unter the alternative hypothesis they are not. Deeper insight into this topic is provided by Lehmann [13].

Permutation tests can be employed for continuous as well as for ordinal and nominal data. They have found many applications, ranging from anthropology to geology. Manly [14] focuses on various environmental applications, including models for spatial data and time series analyses. An extensive bibliography may be found in Good [9].

Whenever we talk about resampling, Monte Carlo is not far away. Although this alliance is not mandatory, it is the main reason for resampling being so successful and prominent. Very often it is not possible to calculate a statistic for all possible resamples. This happens, for example, when n is large for deterministic resampling plans, but even when n is small for stochastic resampling plans. In this case one can resort to an empirical simulation, called Monte Carlo. What Monte Carlo does is randomly choose say B out of a huge number of resamples, recalculate the estimator for each one, and use the empirical distribution of these replicates as an approximation to the exact resampling distribution. The validity of this procedure can be justified by the law of large numbers. The question of how large B should be can be answered only in the context of the application at hand. Typically, B varies from 50 up to some thousands. An excellent application-oriented approach to such computer-intensive methods is provided by Manly [14].

Consequently, resampling requires massive computations. Hence, it has emerged mainly with the advance of modern and powerful computers. Software with built-in resampling techniques is already available. The macros in the jackboot.sas package, available from the **SAS** homepage, allow for jackknife and bootstrap analyses. S-plus can be extended by packages provided by Efron and Tibshirani [7] or Davison and Hinkley [2]. Additional to general bootstrap and jackknife routines, these include also procedures for time series, survival, and censored data, as well as strategies for improved Monte Carlo simulation, like balanced or importance resampling.

After the pioneering work of Quenouille, Tukey, Stone, Efron and others, these methods have reached a large community of researchers, going beyond statisticians, and have caused a vast amount of research thereafter. Nowadays, all these techniques are in widespread use and have proven to be essential tools of statistical inference.

References

[1] Chernick, M. R. (1999). *Bootstrap Methods. A Practitioner's Guide,* Wiley, New York.

[2] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Applications,* Cambridge University Press, Cambridge.

[3] Diaconis, P. and Efron, B. (1983). Statistics on the computer. The bootstrap, *Scientific American* **5**, 116-130.

[4] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *The Annals of Statistics* **7**, 1-26.

[5] Efron, B. (1982). *The Jackknife, the Bootstrap, and other Resampling Plans,* SIAM, Philadelphia.

[6] Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, *The American Statistician* **37**, 36-48.

[7] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap,* Chapman & Hall, London.

[8] Friedl, H. and Stadlober, E. (1997). Resampling methods in generalized linear models useful in environmetrics, *Environmetrics* **8**, 441-457.

[9] Good, P. (1994). *Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypothesis,* Springer-Verlag, New York.

[10] Good, P. (1999). *Resampling Methods. A Practical Guide to Data Analysis,* Birkhäuser, Boston.

[11] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion,* Springer-Verlag, New York.

[12] Hjorth, J. S. U. (1994). *Computer Intensive Statistical Methods–Validation Model Selection and Bootstrap,* Chapman & Hall, London.

[13] Lehmann, E. L. (1997). *Testing Statistical Hypothesis,* 2nd Edition, Springer-Verlag, New York.

[14] Manly, B. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology,* 2nd Edition, Chapman & Hall, London.

[15] Miller, R. G. (1974). The jackknife–a review, *Biometrika* **61**, 1-15.

[16] Quenouille, M. H. (1949). Approximate tests of correlation in time-series, *Journal of the Royal Statistical Society, Series B* **11**, 68-84.

[17] Quenouille, M. H. (1956). Notes on bias in estimation, *Biometrika* **43**, 353-360.

[18] Shao, J. and Tu, D. (1995). *The Jackknife and the Bootstrap,* Springer-Verlag, New York.

[19] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *Journal of the Royal Statistical Society, Series B* **36**, 111-147.

[20] Tukey, J. (1958). Bias and confidence in not quite large samples, *Annals of Mathematical Statistics* **29**, 614.

[21] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion), *The Annals of Statistics* **14**, 1261-1350.

Herwig Friedl, Erwin Stampfer
Institute of Statistics
Technical University Graz
Steyrergasse 17
A-8010 Graz, Austria.