Jackknife and Bootstrap in multinomial situations

Jan de Leeuw

Resampling methods

A very short introduction

Jan de Leeuw
Datatheory FSW/RUL
21-12-83

1: Introduction

In this note we discuss a number of nonparametric methods to compute bias, standard error, or (more generally) sampling distributions of estimates. We only discuss them in the very simple case of a binomial proportion (a Bernouilli variable). Extensions to more complicated situations are quite straightforward, however. First we can extend to a multinomial vector, which in a sense already gives maximum generality in the idenependent, identically distrbuted case. Then we can extend to functions of the empirical distribution function, which gives some additional elegance to the formulations, but not much more generality.

2: Delta method

Suppose $\underline{p}_n$ is a binomial proportion, based on n replications. We imbed it in a sequence $\underline{p}_n$, such that $n^{\frac{1}{2}}(\underline{p}_n - \pi) \xrightarrow{L} N(0,\pi(1-\pi))$, where $\xrightarrow{L}$ denotes convergence in law (= weak convergence of measures). This can be done by the classical de Moivre-Laplace central limit theorem.

Now suppose ϕ is a real-valued function which is differentiable in π. Then $n^{\frac{1}{2}}(\phi(\underline{p}_n) - \phi(\pi)) \xrightarrow{L} N\{0,\pi(1-\pi)(\phi'(\pi))^2\}$. This gives a (first order) approximation to the sampling distribution of $\phi(\underline{p}_n)$. It also implies that for the variance of the estimate

$$n.\mathrm{VAR}(\phi(\underline{p}_n)) = \pi(1-\pi)(\phi'(\pi))^2 + o(1).$$

This requires more assumptions on ϕ, of course (basically uniform integrability, i.e. boundedness conditions on ϕ). The delta method estimates the standard error by using

$$\underline{p}_n(1-\underline{p}_n)(\phi'(\underline{p}_n))^2 = \pi(1-\pi)(\phi'(\pi))^2 + o_p(1),$$

which of course requires continuous differentiability of ϕ near π.

The delta method can also be used for bias correction. We use the expansion

$$\phi(\underline{p}_n) = \phi(\pi) + n^{-\frac{1}{2}}\phi'(\pi)\underline{z}_n + \tfrac{1}{2}n^{-1}\phi''(\pi)\underline{z}_n^2 + o_p(n^{-1}),$$

where $\underline{z}_n = n^{\frac{1}{2}}(\underline{p}_n - \pi)$. This expansion has been justified in a famous 1941 paper by Mann and Wald (Annals of Mathemtical Statistics), also by Chernoff (1956, same journal). Taking expectations on both sides gives

$$E(\phi(\underline{p}_n)) = \phi(\pi) + \tfrac{1}{2}n^{-1}\phi''(\pi)\pi(1 - \pi) + o(n^{-1}).$$

Thus

$$n\ \mathrm{BIAS}(\phi(\underline{p}_n)) = \tfrac{1}{2}\phi''(\pi)\pi(1 - \pi) + o(1),$$

and we can estimate the bias by using

$$\tfrac{1}{2}\phi''(\underline{p}_n)\underline{p}_n(1 - \underline{p}_n) = \tfrac{1}{2}\phi''(\pi)\pi(1 - \pi) + o_p(1),$$

using two times continuous differentianility.

All this extremely simple and straightforward. It can be applied very easily to correspondence analysis (cf Gifi), and it is implemented in our program ANACOR. A possible disadvantage of the delta method is that it requires computation of the first and second derivatives of ϕ, which can be quite demanding computationally in more complicated situations. Another disadvantage is that the results of the delta method are tied to the binomial (or multinomial) model. Although this model is nonparametric, it does suppose the framework of repeated independent trials. If this framework does not apply, the delta method does not make much sense.

3: The jack-knife

The idea behind the jackknife is to delete one observation at a time, and recompute the estimate for each of the resulting n reduced samples of size n - 1. One possibility is approximate the sampling distribution of

the statistic by the empirical distribution of the pseudo-values (cf below). More modestly we can use the pseudo-values to estimate standard error and bias.

In the binomial case we must compute the n values obtained by deleting one Bernouilli trial. But we can only delete either a success or a failure, thus the resulting value can be either $\phi((n\underline{p}_n - 1)/(n - 1))$ or $\phi(n\underline{p}_n/(n - 1))$. Or, to put it differently, we observe $\underline{p}_n$ times $\phi(\underline{p}_n + \frac{1}{n-1}(\underline{p}_n - 1))$ and $(1 - \underline{p}_n)$ times $\phi(\underline{p}_n + \frac{1}{n-1}(\underline{p}_n - 0))$. The pseudo-values are $\underline{p}_n$ times $n\phi(\underline{p}_n) - (n - 1)\phi(\underline{p}_n + \frac{1}{n-1}(\underline{p}_n - 1))$ and $(1 - \underline{p}_n)$ times $n\phi(\underline{p}_n) - (n - 1)\phi(\underline{p}_n + \frac{1}{n-1}\underline{p}_n)$.

In the more general multinomial case $\underline{p}_n$ is a vector with elements $\underline{p}_{jn}$. The pseudo-values are $n\phi(\underline{p}_n) - (n - 1)\phi(\underline{p}_n + \frac{1}{n-1}(\underline{p}_n - e_j))$, which are observed $\underline{p}_{jn}$ times. Here the e_j are the unit vectors of length m, and the pseudo-values assume only m different values. This may be a bit small for a satisfactory approximation to the sampling distribution of $\phi(\underline{p}_n)$.

To introduce the jack-knife estimates of standard error and bias we define $\phi_1(\underline{p}_n)$ and $\phi_0(\underline{p}_n)$ as the two possible reduced values (deleting a success or a failure). The jack-knife estimate of the bias is

$$\underline{\text{BIAS}}_J(\phi(\underline{p}_n)) = (n - 1)\{\underline{p}_n\phi_1(\underline{p}_n) + (1 - \underline{p}_n)\phi_0(\underline{p}_n) - \phi(\underline{p}_n)\},$$

and the jack-knife estimate of the sampling variance is

$$\underline{\text{VAR}}_J(\phi(\underline{p}_n)) = (n - 1)\ \text{VAR(pseudovalues)}.$$

In the Bernouilli case these are asymptotically correct estimates of the population quantities, which (a) do not require computation of derivatives, (b) have some data analytic value. By this last statement we

mean that the question 'what happend to our results if one observation is deleted' is interesting, even if we do not want to commit ourselves to the model of repeated independent trials. Of course in the Bernouilli case there are only two pseudo-values, which is not interesting. In the multivariate case the number of possible profiles is usually even larger than the number of observations, so the situation is quite different.

Gifi also discusses a randomized version of the jack-knife. There are also more complicated versions which eliminate higher-order bias.

4: The bootstrap

The idea behind the bootstrap is that we approximate the sampling distribution of the estimate by sampling with replacement from the empirical distribution function of the observations. That this approximation works has been proved, for example, in a interesting recent paper by Bretagnolle (Ann Inst H Poincare, 1983). Again we are mainly interested in using the bootstrap as a bias and sampling variance estimator, and as a data analytic tool.

Some recent references, not mentioned in Gifi, are Efron (Biometrika-1981, 68, 589-99), Efron (The jackknife, the bootstrap, and other resampling plans, Philadelphia, SIAM, 1982), Parr (Biometrika, 1983, 70, 719-722).

It is clear that the bootstrap interpolates by using many more points that the jack-knife. This makes it more reliable, and also more costly, in most cases.

Again we concentrate on the Bernouilli case. The bootstrap distribution gives resample value $\phi(m/n)$ probability $\binom{n}{m}\underline{p}_n^m(1 - \underline{p}_n)^{n-m}$. Thus, in general, there are n + 1 different resample values, and not just two as for the jackknife.

The bootstrap bias estimate is

$$\underline{\text{BIAS}}_B(\phi(\underline{p}_n)) = \frac{n}{n-1}\{\sum_{m=0}^{n} \phi(\frac{m}{n})\binom{n}{m}\underline{p}_n^m(1 - \underline{p}_n)^{n-m} - \phi(\underline{p}_n)\},$$

and the bootstrap sampling variance estimate is

$$\underline{\text{VAR}}_B(\phi(\underline{p}_n)) = \frac{n}{n-1}\,\text{VAR(resample values)}.$$

For the jackknife we need a randomized version if either n or the number of cells in the multinomial is too large. For the bootstrap we need a randomized version is almost all practical situations. The idea is to estimate mean and variance of resample values by actually drawing independent samples from the multinomial with observed proportions.

Of course in the randomized case there are three parameters which must be large. We have n, the sample size. We have N, the parameter of the bootstrap distribution. And we have M, the number of random samples of the bootstrap distribution. In the classical bootstrap n = N, and M is infinity. In the paper by Bretagnolle N = O(n) and M is again infinity.

It seems to us that the randomized bootstrap is preferable to the jackknife in typical multiple correspondence analysis situations, because it is more smooth. But many theoretical and practical questions must still be answered in this context.

Additional Jackknife and bootstrap results

In this note we extend the results of Parr (Biometrika, 70, 1983, 719-722) to the multinomial case. Thus $\underline{p}_n$ is an m-vector of observed proportions, with expectation π. We usually omit subscript n, the number of observations. The elements of $\underline{p}_n$ are thus simply $\underline{p}_\alpha$ (α=1,...,m). We study $\phi(\underline{p}_n)$, where ϕ is a real valued function having sufficiently many continuous and bounded derivatives to make our expansions valid.

Jackknife results

For the jackknife we compute ϕ in the m points $(n\underline{p} - e_\alpha)/(n-1) = \underline{p} + \frac{1}{n-1}(\underline{p} - e_\alpha)$. Here e_α is the unit vector with +1 as element α, and zero everywhere else. Write $\underline{g}_\alpha$ for the first partials of ϕ, evaluated at $\underline{p}$, $\underline{g}_{\alpha\beta}$ for the second partials, and so on. Then

$$\phi(\underline{p} + \tfrac{1}{n-1}(\underline{p} - e_\alpha)) = \underline{\phi} + (n-1)^{-1}\sum \underline{g}_\beta(\underline{p}_\beta - \delta^{\alpha\beta}) + \tfrac{1}{2}(n-1)^{-2}\sum\sum \underline{g}_{\beta\gamma}(\underline{p}_\beta - \delta^{\alpha\beta})(\underline{p}_\gamma - \delta^{\alpha\gamma}) + \tfrac{1}{6}(n-1)^{-3}\sum\sum\sum \underline{g}_{\beta\gamma\delta}(\underline{p}_\beta - \delta^{\alpha\beta})(\underline{p}_\gamma - \delta^{\alpha\gamma})(\underline{p}_\delta - \delta^{\alpha\delta}) + o((n-1)^{-3}).$$

Here $\underline{\phi}$ is short for $\phi(\underline{p}_n)$. The pseudo-values are

$$n\phi(\underline{p}) - (n-1)\phi(\underline{p} + \tfrac{1}{n-1}(\underline{p} - e_\alpha)) = \underline{\phi} - \sum \underline{g}_\beta(\underline{p}_\beta - \delta^{\alpha\beta}) - \tfrac{1}{2}(n-1)^{-1}\sum\sum \underline{g}_{\beta\gamma}(\underline{p}_\beta - \delta^{\alpha\beta})(\underline{p}_\gamma - \delta^{\alpha\gamma}) - \tfrac{1}{6}(n-1)^{-2}\sum\sum\sum \underline{g}_{\beta\gamma\delta}(\underline{p}_\beta - \delta^{\alpha\beta})(\underline{p}_\gamma - \delta^{\alpha\gamma})(\underline{p}_\delta - \delta^{\alpha\delta}) + o((n-1)^{-2})$$

If we write this pseudo-value as $\phi_\alpha(\underline{p})$, then the jackknife estimate is

$$\sum \underline{p}_\alpha\phi_\alpha(\underline{p}) = \underline{\phi} - \tfrac{1}{2}(n-1)^{-1}\sum\sum \underline{g}_{\beta\gamma}\underline{\lambda}_{\beta\gamma} + \tfrac{1}{6}(n-1)^{-2}\sum\sum\sum \underline{g}_{\beta\gamma\delta}\underline{\lambda}_{\beta\gamma\delta} + o((n-1)^{-2}).$$

Here $\underline{\lambda}_{\beta\gamma}$ are the observed second order moments, and $\underline{\lambda}_{\beta\gamma\delta}$ are the observed third order moments. Thus $\underline{\lambda}_{\beta\gamma} = \delta^{\beta\gamma}\underline{p}_\beta - \underline{p}_\beta\underline{p}_\gamma$ and $\underline{\lambda}_{\beta\gamma\delta} = \delta^{\beta\gamma\delta}\underline{p}_\beta - \underline{p}_\beta\delta^{\gamma\delta} -$

$\underline{p}_\gamma \delta^{\beta\delta} - \underline{p}_\delta \delta^{\beta\gamma} + 2\underline{p}_\beta \underline{p}_\gamma \underline{p}_\delta$. The bias-estimate of the delta-method is

$$\underline{BIAS}_D = -\tfrac{1}{2}(n-1)^{-1} \sum\sum \underline{g}_{\beta\gamma} \underline{\lambda}_{\beta\gamma}.$$

Thus

$$(n-1)^2(\underline{BIAS}_J - \underline{BIAS}_D) = \frac{1}{6} \sum\sum\sum \underline{g}_{\beta\gamma\delta} \underline{\lambda}_{\beta\gamma\delta} + o(1).$$

To compute the jackknife variance estimate we first square the pseudo-values $\phi_\alpha(\underline{p})$. This gives

$$\phi_\alpha^2(\underline{p}) = \underline{\phi}^2 + \sum\sum \underline{g}_\beta \underline{g}_\gamma (\underline{p}_\beta - \delta^{\alpha\beta})(\underline{p}_\gamma - \delta^{\alpha\gamma}) - (n-1)^{-1} \underline{\phi} \sum\sum \underline{g}_{\beta\gamma} (\underline{p}_\beta - \delta^{\alpha\beta})(\underline{p}_\gamma - \delta^{\alpha\gamma})$$
$$+ (n-1)^{-1} \sum\sum\sum \underline{g}_\beta \underline{g}_{\gamma\delta} (\underline{p}_\beta - \delta^{\alpha\beta})(\underline{p}_\gamma - \delta^{\alpha\gamma})(\underline{p}_\delta - \delta^{\alpha\delta}) + o((n-1)^{-1}).$$

Thus

$$\sum \underline{p}_\alpha \phi_\alpha^2(\underline{p}) = \underline{\phi}^2 + \sum\sum \underline{g}_\beta \underline{g}_\gamma \underline{\lambda}_{\beta\gamma} - (n-1)^{-1} \underline{\phi} \sum\sum \underline{g}_{\beta\gamma} \underline{\lambda}_{\beta\gamma} + (n-1)^{-1} \sum\sum\sum \underline{g}_\beta \underline{g}_{\gamma\delta} \underline{\lambda}_{\beta\gamma\delta} + o((n-1)^{-}$$

The variance of the pseudo-values consequently is

$$\underline{VAR}(\phi_\alpha(\underline{p})) = \sum\sum \underline{g}_\beta \underline{g}_\gamma \underline{\lambda}_{\beta\gamma} + (n-1)^{-1} \sum\sum\sum \underline{g}_\beta \underline{g}_{\gamma\delta} \underline{\lambda}_{\beta\gamma\delta} + o((n-1)^{-1}).$$

Because the delta-method variance estimate is

$$\underline{VAR}_D = (n-1)^{-1} \sum\sum \underline{g}_\beta \underline{g}_\gamma \underline{\lambda}_{\beta\gamma},$$

and the jackknife variance estimate is

$$\underline{VAR}_J = (n-1)^{-1} \underline{VAR}(\phi_\alpha(\underline{p})),$$

we find

$$(n-1)^2(\underline{VAR}_J - \underline{VAR}_D) = \sum\sum\sum \underline{g}_\beta \underline{g}_{\gamma\delta} \underline{\lambda}_{\beta\gamma\phi} + o(1).$$

Bootstrap results

For the bootstrap (at least the *theoretical bootstrap*) we compute ϕ in all vectors q, where nq is a vector of integers adding up to n. Define $\underline{z} = n^{\frac{1}{2}}(q - \underline{p})$. Then $\phi(q) = \phi(\underline{p} + n^{-\frac{1}{2}}\underline{z})$, and thus

$$\phi(q) = \underline{\phi} + n^{-\frac{1}{2}}\sum \underline{g}_{\alpha}\underline{z}_{\alpha} + \tfrac{1}{2}n^{-1}\sum\sum \underline{g}_{\alpha\beta}\underline{z}_{\alpha}\underline{z}_{\beta} + \frac{1}{6}n^{-3/2}\sum\sum\sum \underline{g}_{\alpha\beta\gamma}\underline{z}_{\alpha}\underline{z}_{\beta}\underline{z}_{\gamma} +$$

$$\frac{1}{24}\, n^{-2}\sum\sum\sum\sum \underline{g}_{\alpha\beta\gamma\delta}\underline{z}_{\alpha}\underline{z}_{\beta}\underline{z}_{\gamma}\underline{z}_{\delta} + o(n^{-2}).$$

If $w(q)$ is the multinomial probability of observing nq if sampling n times from a multinomial with probability vector $\underline{p}$, then the bootstrap estimate is

$$\sum w(q)\phi(q) = \underline{\phi} + \tfrac{1}{2}n^{-1}\sum\sum \underline{g}_{\alpha\beta}\underline{\lambda}_{\alpha\beta} + \frac{1}{6}\, n^{-2}\sum\sum\sum \underline{g}_{\alpha\beta\gamma}\underline{\lambda}_{\alpha\beta\gamma} +$$

$$\frac{1}{24}\, n^{-2}\sum\sum\sum\sum \underline{g}_{\alpha\beta\gamma\delta}(\underline{\lambda}_{\alpha\beta}\underline{\lambda}_{\gamma\delta} + \underline{\lambda}_{\alpha\gamma}\underline{\lambda}_{\beta\delta} + \underline{\lambda}_{\alpha\delta}\underline{\lambda}_{\beta\gamma}) + o(n^{-2}).$$

The debiased bootstrap is

$$\frac{2n-1}{n-1}\,\underline{\phi} - \frac{n}{n-1}\sum w(q)\phi(q) = \underline{\phi} - \tfrac{1}{2}\,(n-1)^{-1}\sum\sum \underline{g}_{\alpha\beta}\underline{\lambda}_{\alpha\beta} - \frac{1}{6}\,\frac{1}{n(n-1)}\sum\sum\sum \underline{g}_{\alpha\beta\gamma}\underline{\lambda}_{\alpha\beta\gamma}$$

$$- \frac{1}{24}\,\frac{1}{n(n-1)}\sum\sum\sum\sum \underline{g}_{\alpha\beta\gamma\delta}(\underline{\lambda}_{\alpha\beta}\underline{\lambda}_{\gamma\delta} + \underline{\lambda}_{\alpha\gamma}\underline{\lambda}_{\beta\delta} + \underline{\lambda}_{\alpha\delta}\underline{\lambda}_{\beta\gamma}) + o(n^{-1}(n-1)^{-1}).$$

Thus

$$n(n-1)(\underline{\mathrm{BIAS}}_B - \underline{\mathrm{BIAS}}_D) = -\frac{1}{6}\sum\sum\sum \underline{g}_{\alpha\beta\gamma}\underline{\lambda}_{\alpha\beta\gamma} - \frac{1}{24}\sum\sum\sum\sum \underline{g}_{\alpha\beta\gamma\delta}(\underline{\lambda}_{\alpha\beta}\underline{\lambda}_{\gamma\delta} +$$

$$\underline{\lambda}_{\alpha\gamma}\underline{\lambda}_{\beta\delta} + \underline{\lambda}_{\alpha\delta}\underline{\lambda}_{\beta\gamma}) + o(1).$$

For the Bootstrap variance estimate we need

$$\phi^2(q) = \underline{\phi}^2 + n^{-1}\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta}\underline{z}_{\alpha}\underline{z}_{\beta} + \tfrac{1}{4}n^{-2}\sum\sum\sum\sum \underline{g}_{\alpha\beta}\underline{g}_{\gamma\delta}\underline{z}_{\alpha}\underline{z}_{\beta}\underline{z}_{\gamma}\underline{z}_{\delta} +$$

$$+ 2n^{-\frac{1}{2}}\underline{\phi}\sum \underline{g}_{\alpha}\underline{z}_{\alpha} + n^{-1}\underline{\phi}\sum\sum \underline{g}_{\alpha\beta}\underline{z}_{\alpha}\underline{z}_{\beta} + \frac{1}{3}n^{-3/2}\underline{\phi}\sum\sum\sum \underline{g}_{\alpha\beta\gamma}\underline{z}_{\alpha}\underline{z}_{\beta}\underline{z}_{\gamma} +$$

$$+ \frac{1}{12}\, n^{-2}\underline{\phi}\sum\sum\sum\sum \underline{g}_{\alpha\beta\gamma\delta}\underline{z}_{\alpha}\underline{z}_{\beta}\underline{z}_{\gamma}\underline{z}_{\delta} + n^{-3/2}\sum\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta\gamma}\underline{z}_{\alpha}\underline{z}_{\beta}\underline{z}_{\gamma} +$$

$$+ \frac{1}{3}n^{-2}\sum\sum\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta\gamma\delta}\underline{z}_{\alpha}\underline{z}_{\beta}\underline{z}_{\gamma}\underline{z}_{\delta} + o(n^{-2}).$$

Thus

$$\sum w(q)\phi^2(q) = \underline{\phi}^2 + n^{-1}\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta}\underline{\lambda}_{\alpha\beta} + n^{-1}\underline{\phi}\sum\sum \underline{g}_{\alpha\beta}\underline{\lambda}_{\alpha\beta} +$$

$$\tfrac{1}{4}n^{-2}\sum\sum\sum\sum \underline{g}_{\alpha\beta}\underline{g}_{\gamma\delta}\underline{\lambda}_{\alpha\beta\gamma\delta} + \frac{1}{3}n^{-2}\underline{\phi}\sum\sum\sum \underline{g}_{\alpha\beta\gamma}\underline{\lambda}_{\alpha\beta\gamma} +$$

$$\frac{1}{12}n^{-2}\underline{\phi}\sum\sum\sum\sum \underline{g}_{\alpha\beta\gamma\delta}\underline{\lambda}_{\alpha\beta\gamma\delta} + n^{-2}\sum\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta\gamma}\underline{\lambda}_{\alpha\beta\gamma} +$$

$$\tfrac{1}{3}n^{-2}\sum\sum\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta\gamma\delta}\underline{\lambda}_{\alpha\beta\gamma\delta} + o(n^{-2}).$$

Here $\underline{\lambda}_{\alpha\beta\gamma\delta}$ is short for $\underline{\lambda}_{\alpha\beta}\underline{\lambda}_{\gamma\delta} + \underline{\lambda}_{\alpha\gamma}\underline{\lambda}_{\beta\delta} + \underline{\lambda}_{\alpha\delta}\underline{\lambda}_{\beta\gamma}$. Moreover

$$(\sum w(q)\phi(q))^2 = \underline{\phi}^2 + \tfrac{1}{4}n^{-2}\sum\sum\sum\sum \underline{g}_{\alpha\beta}\underline{g}_{\gamma\delta}\underline{\lambda}_{\alpha\beta}\underline{\lambda}_{\gamma\delta} + n^{-1}\underline{\phi}\sum\sum \underline{g}_{\alpha\beta}\underline{\lambda}_{\alpha\beta} +$$
$$\tfrac{1}{3}n^{-2}\underline{\phi}\sum\sum\sum \underline{g}_{\alpha\beta\gamma}\underline{\lambda}_{\alpha\beta\gamma} + \tfrac{1}{12}n^{-2}\underline{\phi}\sum\sum\sum\sum \underline{g}_{\alpha\beta\gamma\delta}\underline{\lambda}_{\alpha\beta\gamma\delta} + o(n^{-2}).$$

Thus

$$\underline{\mathrm{VAR}}_B = \frac{n}{n-1}\{\sum w(q)\phi^2(q) - (\sum w(q)\phi(q))^2\} =$$
$$= (n-1)^{-1}\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta}\underline{\lambda}_{\alpha\beta} + \tfrac{1}{4}\frac{1}{n(n-1)}\sum\sum\sum\sum \underline{g}_{\alpha\beta}\underline{g}_{\gamma\delta}(\underline{\lambda}_{\alpha\beta\gamma\delta} - \underline{\lambda}_{\alpha\beta}\underline{\lambda}_{\gamma\delta}) +$$
$$+ \frac{1}{n(n-1)}\sum\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta\gamma}\underline{\lambda}_{\alpha\beta\gamma} + \frac{1}{3}\frac{1}{n(n-1)}\sum\sum\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta\gamma\delta}\underline{\lambda}_{\alpha\beta\gamma\delta} + o(n^{-1}(n-1)^{-1}).$$

Thus

$$n(n-1)(\underline{\mathrm{VAR}}_B - \underline{\mathrm{VAR}}_D) = \tfrac{1}{4}\sum\sum\sum\sum \underline{g}_{\alpha\beta}\underline{g}_{\gamma\delta}(\underline{\lambda}_{\alpha\gamma}\underline{\lambda}_{\beta\delta} + \underline{\lambda}_{\alpha\delta}\underline{\lambda}_{\beta\gamma}) + \sum\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta\gamma}\underline{\lambda}_{\alpha\beta\gamma} +$$
$$+ \tfrac{1}{3}\sum\sum\sum\sum \underline{g}_{\alpha}\underline{g}_{\beta\gamma\delta}\underline{\lambda}_{\alpha\beta\gamma\delta} + o(1).$$

Remarks

The derivations above are heuristic (or formal), and not properly justified. Justifications are possible by using the work of Bhattacharya and Rao on Edgeworth expansions, or the work of Hurt on expansions of moments. There may still be computational errors in the formulas, of course. The formulas do not apply to the empirical bootstrap, which uses samples to estimate the conditional expectations.

Validity of the bootstrap

Validity can be defined in various ways. We have shown above that the bootstrap and jackknife are valid, in the sense that they give the same bias correction and variance estimates as the delta method to a high degree of approximation. It also follows from the results above that bootstrap and jackknife estimates have the same asymptotic distribution as the original estimate, which could also be interpreted as a form of validity.

For the bootstrap another form of validity has been studied quite thoroughly. If $n\underline{q}$ is multinomial, with parameters $\underline{p}_n$ and n, then the distribution of $\phi(\underline{q})$ is called the bootstrap distrbution. We take a random sample of size n from the empirical distrbution. The bootstrap distribution is computed conditionally on the data, i.e. $\underline{p}_n$ is considered as fixed. It was already proved by Efron (Ann Stat, 1979, 1-26) that the conditional distribution of $\phi(\underline{q}_n)$, given $\underline{p}_n$, converges to the same limit as the distribution of $\phi(\underline{p}_n)$, given π. This result has been extended to t-statistics, the empirical process, the quantile process, von Mises functionals by Bickel and Freedman (Ann Stat, 1981, 1196-1217) and by Bretagnolle (Ann Inst Henri Poincaré, 1983, 281-296). In the case of bootstrapping a mean or studenticized mean rate of convergence results have been proved by Singh (Ann Statist, 1981, 1187-1195) and by Babu and Singh (Ann Statist, 1983, 999-1003). They show that the deviation between the distributions is $O(n^{-\frac{1}{2}})$. In fact this remains true for the B-sample empirical or Monte Carlo bootstrap, provided that $B/n \ln n \to \infty$. We conjecture that it is sufficient for validity in this case that $B/n^{\frac{1}{2}} \ln n \to \infty$ or even that $B/n^{\frac{1}{2}} \to \infty$ while $B = O(n^{\frac{1}{2}} \ln n)$. Compare Freedman (Z. Wahrscheinlichkeitstheorie, 1977, 1-11).

A different, and perhaps equally important result, has been proved by Beran (Ann Statist, 1982, 212-225). He shows that bootstrap estimates (by which we mean the bootstrap distribution and transforms of the bootstrap distrbution) are not only consistent, but actually efficient. Here efficiency is in the LAM (local asymptotically minimax) sense, familiar from the work of Le Cam, Hajek, Ibragimov, Hasminskii, Hannan, Fabian, Beran, Millar, Levit and so on.

Leave-one-out cross-validation techniques

Multinomial maximum likelihood estimation

We study models of the form $\pi \in \Omega$, with Ω an s-dimensional differentiable manifold in R^m. We use a best asymptotically normal estimation technique Φ to estimate π. Thus Φ associates an estimate $\Phi(p)$ with each probability vector p in S^{m-1}, the unit simplex in R^m.

Now suppose $\underline{p}$ is a vector of relative frequencies, constructed on the basis of a simple random sample of size n. If we leave out one observation we change $\underline{p}$ to $(n\underline{p} - e_j)/(n-1) = \underline{p} + \frac{1}{n-1}(\underline{p} - e_j)$. Here e_j is unit vector j, and the observation we left out was in category j. The corresponding estimate of π is $\Phi(\underline{p} + \frac{1}{n-1}(\underline{p} - e_j))$, and the predicted likelihood of the omitted observation is $\phi_j(\underline{p} + \frac{1}{n-1}(\underline{p} - e_j))$. Because there are $\underline{p}_j$ observations with 'value' e_j the total predicted log-likelihood is

$$L_\Omega(n,\underline{p}) = \sum_{j=1}^{m} \underline{p}_j \ln \phi_j(\underline{p} + \frac{1}{n-1}(\underline{p} - e_j)).$$

Now

$$\phi_j(\underline{p} + \frac{1}{n-1}(\underline{p} - e_j)) = \underline{\pi}_j + \frac{1}{n-1}\ \underline{g}_j'(\underline{p} - e_j) + o((n-1)^{-1}).$$

Here $\underline{\pi}_j$ is $\phi_j(\underline{p})$, and $\underline{g}_j$ are the partials of ϕ_j, evaluated at $\underline{p}$. It follows that

$$L_\Omega(n,\underline{p}) = \sum_{j=1}^{m} \underline{p}_j \ln \underline{\pi}_j + \frac{1}{n-1} \sum_{j=1}^{m} (\underline{p}_j/\underline{\pi}_j)\ \underline{g}_j'(\underline{p} - e_j) + o((n-1)^{-1}).$$

If π is the 'true' value $E(\underline{p})$, then $\underline{\pi}_j = \pi_j + o_p(1)$ and $\underline{p}_j = \pi_j + o_p(1)$. If the partials of Φ are continuous at π, then also $\underline{G} = \Gamma + o_p(1)$, with Γ the partials of Φ at π. If we use, in addition, the fact that the rows of Γ sum to zero, we see that

$$L_\Omega(n,\underline{p}) = L_\Omega(\underline{p}) - \frac{1}{n-1} \operatorname{tr} \Gamma + o_p((n-1)^{-1}).$$

But, by consistency, $\operatorname{tr} \Gamma = s$, the number of free parameters, or the dimensionality of the manifold. In particular for a saturated model $\pi \in S$ we have

$$L_S(n,\underline{p}) = L_S(\underline{p}) - \frac{m-1}{n-1} + o_p((n-1)^{-1}).$$

If we have to choose from a number of different models $\Omega_1, \Omega_2, \ldots$ we choose the one with the highest value of $L_\Omega(n,\underline{p})$. It is convenient to use the saturated model for normalization purposes. Thus

$$2(n-1)(L_S(n,\underline{p}) - L_\Omega(n,\underline{p})) = 2(n-1)(L_S(\underline{p}) - L_\Omega(\underline{p})) - 2(m - s - 1) + o_p(1).$$

This can be written as

$$\Delta_\Omega(n,\underline{p}) = \Delta_\Omega(\underline{p}) - 2K_\Omega + o_p(1).$$

Here $\Delta_\Omega(\underline{p})$ converges in law to chi square with $K = m - s - 1$ degrees of freedom. Thus we can compare the predictive qualities of models by subtracting two times their degrees of freedom from their chi squares. It is irrelevant which best asymptotically normal estimate we use

Use of random sub-sampling for bias-correction,
estimation of standard error, and cross-validation.

We restrict ourselves again to multinomial situations. Random variables $\underline{x}_i$ (i=1,2,...) are independent and identically distributed. They take the unit vectors of length m as their values, and prob($\underline{x}_i = e_j$) = π_j. The relative frequencies in a sample of size n are $\underline{p}_j$.

Now suppose $\underline{\varepsilon}_i$ is another sequence of independent and identically distributed random variables, moreover independent of all $\underline{x}_i$. Variable $\underline{\varepsilon}_i$ takes the value one with probability ξ and the value zero with probability $1 - \xi$. To estimate the distribution of $\Phi(\underline{p})$ we use the subsampling distribution, which is the distribution of $\Phi(\sum\underline{\varepsilon}_i x_i/\sum\underline{\varepsilon}_i)$. Observe that we use x_i, without underlining here. This means that the subsampling distribution is defined conditionally on the data, which are fixed at their observed values $\underline{x}_i$.

Let $\underline{q} = (\sum\underline{\varepsilon}_i x_i)/(\sum\underline{\varepsilon}_i)$. Then $n^{\frac{1}{2}}(\underline{q} - p)$ is asymptotically normal, with mean zero, and with dispersion

$$V = \frac{\xi(1-\xi)}{\xi^2}(P - pp').$$

Thus if we want to continue without unnecessary corrections we must take $\xi = \frac{1}{2}$. This is also very convenient in practice: we can do stability analysis with a fair coin. It is clear that subsampling provides us with a valid approximation to the sampling distribution. It is also clear that the subsampling bias correction and standard error coincide with that provided by the bootstrap.

Contrary to the bootstrap, the subsample method can also be used nicely for cross validation (because it leaves out a portion of the data). We

illustrate this for multinomial maximum likelihood estimation. The maximum likelihood estimate is computed as $\Phi(\underline{q})$, and evaluated using $\underline{r} = (\sum(1 - \underline{\varepsilon}_i)x_i)/(\sum(1 - \underline{\varepsilon}_i))$. Thus the predicted log-likelihood is

$$L_{\Omega}(n,p) = \sum_{j=1}^{m} \underline{r}_j \ln \phi_j(\underline{q}).$$

Simple asymptotic approximations, as for the leave-one-out method, are not directly available, however.

On the multinomial jackknife

If we delete one observation we replace p by $p + \frac{1}{n-1}(p - e_k)$. The jackknife value is

$$\phi(p + \frac{1}{n-1}(p - e_k)) \approx \phi(p) + \frac{1}{n-1} g'(p - e_k), \tag{1}$$

with g the vector of partials of ϕ at p. If we define

$$h_k = (n-1)\{\phi(p + \frac{1}{n-1}(p - e_k)) - \phi(p)\}, \tag{2}$$

then we can rewrite the equations (1) as

$$h \approx Sg, \tag{3}$$

with

$$S = up' - I. \tag{4}$$

If $V = P - pp'$, as usual, then $VS = -V$. Thus

$$(h - Sg)'V(h - Sg) = h'Vh + 2g'Vh + g'Vg = (h + g)'V(h + g), \tag{5}$$

and $PS = -V$, which implies

$$(h - Sg)'P(h - Sg) = h'Ph + 2g'Vh + g'Vg = (h + g)'V(h + g) + (h'p)^2. \tag{6}$$

Both least squares criteria (5) and (6) are minimized by taking $g = -h$. It follows that we can estimate the delta method variance estimate by using $h'Vh$. This can be interpreted by first estimating g by finite differences methods, and by using the estimate of g in the delta method formula.

The jackknife variance estimate is the variance of the jackknife pseudo-values. The vector of pseudo values f is given by

$$f_k = \phi(p) - h_k, \tag{7}$$

and thus $f'Vf = h'Vh$. Thus the jackknife variance estimate can be interpreted as a two-step estimate of the delta method variance estimate, using finite differences to approximate derivatives.

Jan de Leeuw
may 13, 1985.