73797

(6)

# Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods

Bradley Efron

*Biometrika*, Volume 68, Issue 3 (Dec., 1981), 589-599.

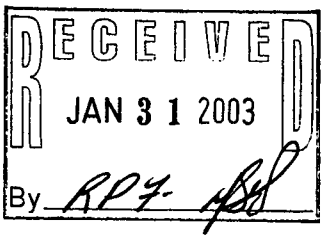
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198112%2968%3A3%3C589%3ANEOSET%3E2.0.CO%3B2-2

---



---








*Biometrika* (1981), **68**, 3, *pp.* 589–99
*Printed in Great Britain*



# Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods

BY BRADLEY EFRON
*Department of Statistics, Stanford University*

SUMMARY

We discuss several nonparametric methods for attaching a standard error to a point estimate: the jackknife, the bootstrap, half-sampling, subsampling, balanced repeated replications, the infinitesimal jackknife, influence function techniques and the delta method. The last three methods are shown to be identical. All the methods derive from the same basic idea, which is also the idea underlying the common parametric methods. Extended numerical comparisons are made for the special case of the correlation coefficient.

*Some key words*: Balanced repeated replications; Bootstrap; Delta method; Half-sampling; Jackknife; Infinitesimal jackknife; Influence function.

## 1. OUTLINE

We wish to attach a standard error to some point estimate $\hat{\rho}$. The standard error itself must be estimated from the data, and this is usually done by parametric modelling methods. Here we discuss several nonparametric methods for estimating standard errors; the jackknife, the bootstrap, half-sampling, subsampling, balanced repeated replications, the infinitesimal jackknife, influence function techniques and the delta method. The discussion is built around a single numerical example, the correlation coefficient from a bivariate normal model.

The purpose of the discussion is fourfold:

(i) to describe concisely the various methods;

(ii) to show how all these methods derive from the same basic idea, which is also the idea underlying common parametric methods;

(iii) to draw more specific connexions between certain of the techniques. For example, § 6 shows that the ordinary delta method is exactly the same as the infinitesimal jackknife;

(iv) to show how differently the various methods perform in the numerical example, even though they are asymptotically equivalent. The bootstrap performs notably best.

The various methods are defined and described in §§ 3–8, but in a brief manner which omits much of their practical and theoretical motivation. The reader is referred to Miller (1974) for a neat review of the jackknife and infinitesimal jackknife, Hartigan (1969) and Maritz (1979) on subsampling theory, McCarthy (1969) and Kish and Frankel (1974) for half-sampling and balanced repeated replications, Hampel (1974) and Huber (1972) for influence function methods, and Efron (1979a, 1981) for the bootstrap. Hinkley (1978) specifically discusses the jackknife for the correlation coefficient.

Section 9 briefly discusses why standard errors are of interest. More ambitious nonparametric accuracy statements, such as confidence intervals, are mentioned, though no satisfactory general theory yet exists.

## 2. A MONTE CARLO EXPERIMENT

Table 1 shows the results of a Monte Carlo experiment. There were 200 trials, each of which involved 14 independently drawn bivariate normal points $X_i = (U_i, V_i)$ with

$$E(U_i) = E(V_i) = 0, \quad \text{var}(U_i) = \text{var}(V_i) = 1, \quad \text{cov}(U_i, V_i) = \tfrac{1}{2}. \tag{2·1}$$

The statistic of interest, for which an estimated standard error is desired, is the Pearson correlation coefficient $\hat{\rho}$,

$$\hat{\rho}(X_1, \ldots, X_n) = \frac{\Sigma U_i V_i - \Sigma U_i \Sigma V_i/n}{[\{\Sigma U_i^2 - (\Sigma U_i)^2/n\}\{\Sigma V_i^2 - (\Sigma V_i)^2/n\}]^{\frac{1}{2}}}. \tag{2·2}$$

We also consider $\hat{\phi} = \tanh^{-1}\hat{\rho} = \frac{1}{2}\log\{(1+\hat{\rho})/(1-\hat{\rho})\}$, Fisher's variance-stabilizing transformation.

From data $X_1 = x_1, \ldots, X_n = x_n$ $(n = 14)$, fifteen different methods were used to construct estimated standard errors for $\hat{\rho}$ and $\hat{\phi}$. Table 1 shows summary statistics over the 200 trials. For example a normal theory estimate of $\sigma(\hat{\rho})$, the true standard error of $\hat{\rho}$, is $\sigma_N(\hat{\rho}) = (1-\hat{\rho}^2)/11^{\frac{1}{2}}$; see §4. Line 15 of the table shows that the normal theory estimates for the 200 trials averaged 0·217, with sample standard deviation 0·056, and coefficient of variation 0·26 = 0·056/0·217. The true value $\sigma(\hat{\rho})$ is 0·218 in situation (2·1) (Johnson and Kotz, 1970, p. 225). The root mean squared error, from 0·218, was $\text{MSE}^{\frac{1}{2}} = 0·056$.

Table 1. *Nonparametric estimates of standard error for $\hat{\rho}$ and $\hat{\phi} = \tanh^{-1}\hat{\rho}$; 200 trials of 14 independent, bivariate normal pairs with true correlation* 0·5

| | $\hat{\rho}$ | | | | $\hat{\phi} = \tanh^{-1}\hat{\rho}$ | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Mean | SD | CV | $\text{MSE}^{\frac{1}{2}}$ | Mean | SD | CV | $\text{MSE}^{\frac{1}{2}}$ |
| 1. Bootstrap (BST), $N = 128$ | 0·206 | 0·066 | 0·32 | 0·067 | 0·301 | 0·065 | 0·22 | 0·065 |
| 2. Bootstrap, $N = 512$ | 0·206 | 0·063 | 0·31 | 0·064 | 0·301 | 0·062 | 0·21 | 0·062 |
| 3. Normal smoothed BST, $N = 128$ | 0·200 | 0·060 | 0·30 | 0·063 | 0·296 | 0·041 | 0·14 | 0·041 |
| 4. Uniform smoothed BST, $N = 128$ | 0·205 | 0·061 | 0·30 | 0·062 | 0·298 | 0·058 | 0·19 | 0·058 |
| 5. Uniform smoothed BST, $N = 152$ | 0·205 | 0·059 | 0·29 | 0·060 | 0·296 | 0·052 | 0·18 | 0·052 |
| 6. Jackknife | 0·223 | 0·085 | 0·38 | 0·085 | 0·314 | 0·090 | 0·29 | 0·091 |
| 7. Infinitesimal jackknife | 0·175** | 0·058 | 0·33 | 0·072 | 0·244* | 0·052 | 0·21 | 0·076 |
| 8. Half-samples, all 128 | 0·244* | 0·083 | 0·34 | 0·087 | 0·364** | 0·099 | 0·27 | 0·118 |
| 9. Random HS, $N = 128$ | 0·248* | 0·079 | 0·32 | 0·085 | 0·368** | 0·084 | 0·23 | 0·109 |
| 10. Balanced HS, 8 | 0·244* | 0·095 | 0·39 | 0·098 | 0·366** | 0·111 | 0·30 | 0·129 |
| 11. Complementary HS, all 128 | 0·223 | 0·079 | 0·35 | 0·079 | 0·336* | 0·099 | 0·30 | 0·105 |
| 12. Complementary bal. HS, 16 | 0·222 | 0·081 | 0·36 | 0·081 | 0·335* | 0·100 | 0·30 | 0·106 |
| 13. Random subsampling, $N = 128$ | 0·267** | 0·080 | 0·30 | 0·094 | 0·423*** | 0·089 | 0·21 | 0·153 |
| 14. Random subsampling, Range est. SD | 0·242 | 0·092 | 0·38 | 0·095 | 0·354* | 0·077 | 0·27 | 0·111 |
| 15. Normal theory | 0·217 | 0·056 | 0·26 | 0·056 | 0·302 | 0 | 0 | 0·003 |
| Theoretical value | 0·218 | | | | 0·299 | | | |

The true standard errors are $\sigma(\hat{\rho}) = 0·221$, $\sigma(\hat{\phi}) = 0·299$. Large biases are indicated by asterisks: *Relative bias $\geqslant$ 0·10, **Relative Bias $\geqslant$ 0·20, ***Relative Bias $\geqslant$ 0·40. HS, half-samples.

The true standard error for $\hat{\phi}$ is $\sigma(\hat{\phi}) = 0{\cdot}299$ in situation (2·1). A normal theory estimate is $1/(n-3)^{\frac{1}{2}} = 0{\cdot}302$ (Johnson and Kotz, 1970, p. 229). In this case the normal theory estimate of standard deviation has zero sample-to-sample variation, which is the underlying reason for the $\tanh^{-1}$ transformation.

Root mean squared error is a convenient criterion for comparing how closely the various estimates of standard error cluster about the true values. Large biases are unpleasant though, even if root mean squared error is low, and these are indicated by asterisks in the table.

In what follows, the various methods will be described in terms of estimating $\sigma(\hat{\rho})$, the corresponding details for $\sigma(\hat{\phi})$ or for the standard error of any other statistic then being obvious. Mnemonic subscripts $\sigma_N(\hat{\rho})$, $\sigma_B(\hat{\rho})$, etc., identify the different estimates.

## 3. The bootstrap

All the methods in this paper assume independent identical sampling from an unknown distribution $F$ on an arbitrary sample space $\mathcal{X}$. In the correlation example $\mathcal{X} = R^2$, the plane. The bootstrap estimate of standard error for $\hat{\rho}$, denoted by $\sigma_B(\hat{\rho})$ (Efron 1979a), is easy to describe:

(i) let $\hat{F}$ be the empirical probability distribution

$$\hat{F} \text{ having mass } 1/n \text{ at each observed } x_i \quad (i = 1, \ldots, n); \tag{3·1}$$

(ii) let $X_1^*, \ldots, X_n^*$ be a random sample from $\hat{F}$, i.e. $n$ independent draws each with distribution $\hat{F}$, and let $\hat{\rho}^* = \hat{\rho}(X_1^*, \ldots, X_n^*)$;

(iii) the bootstrap estimate is $\sigma_B(\hat{\rho}) = \{\mathrm{var}_*(\hat{\rho}^*)\}^{\frac{1}{2}}$, where $\mathrm{var}_*(\hat{\rho}^*)$ indicates the variance of $\hat{\rho}^*$ under the probability mechanism (ii), with $\hat{F}$ fixed at its observed value (3·1).

In other words, the bootstrap estimate $\sigma_B(\hat{\rho})$ is simply the standard deviation of the quantity of interest, $\hat{\rho}(X_1, \ldots, X_n)$, if the unknown distribution $F$ is taken equal to the observed distribution $\hat{F}$. Theoretical calculation of $\sigma_B(\hat{\rho})$ is impossible, but Monte Carlo simulation yields a quick approximation: step (ii) is repeated independently $N$ times, yielding $N$ independent realizations of $\hat{\rho}^*$, say $\hat{\rho}^*(1), \ldots, \hat{\rho}(N)$. Then $\sigma_B(\hat{\rho})$ is approximated by the sample standard deviation $[\Sigma\{(\hat{\rho}^*(j) - \hat{\rho}^*(.)\}^2/(N-1)]^{\frac{1}{2}}$, where $\hat{\rho}^*(.) = \Sigma\hat{\rho}^*(j)/N$. Note that in what follows the dot always indicates averaging over the collection of recalculated values $\hat{\rho}^*(j)$.

Line 1 of Table 1 used $N = 128$, a convenient number for comparison with other techniques; line 2 used $N = 512$, which performed only slightly better. A components of variance analysis of all the data going into lines 1 and 2 of the table showed that further increases of $N$ would be pointless. The MSE$^{\frac{1}{2}}$ for $N = \infty$ is no more than 0·001 smaller than that for $N = 512$, for either $\sigma_B(\hat{\rho})$ or $\sigma_B(\hat{\phi})$. In a real situation, where there is only one set of observations, choosing $N$ is more problematical, but routine error analyses give a rough idea of when to stop the bootstrap sampling. Usually the choice of $N$ seems not to be crucial, past $N = 50$ or 100.

## 4. Normal theory and the smoothed bootstrap

The standard normal theory estimate $\sigma_N(\hat{\rho})$ can itself be thought of as a bootstrap estimate, carried out in a parametric framework. The maximum likelihood estimate for

the unknown sampling distribution $F$, assuming bivariate normality, is

$$\hat{F}_N \sim \mathcal{N}_2(\bar{x}, \hat{\Omega}), \tag{4·1}$$

$\bar{x} = \Sigma x_i/n$, $\hat{\Omega} = \Sigma (x_i - \bar{x})(x_i - \bar{x})'/n$. Replacement of $\hat{F}$ with $\hat{F}_N$ in the bootstrap algorithm, otherwise proceeding exactly as described in steps (ii)–(iii), gives, apart from degrees of freedom, the estimate $\sigma_N(\hat{\rho})$. In other words, $\sigma_N(\hat{\rho})$ is the standard deviation of $\hat{\rho}(X_1, \ldots, X_n)$ if the unknown distribution $F$ is taken equal to $\hat{F}_N$.

Theoretical calculation of $\sigma_N(\hat{\rho})$ as described above is impossible, but Taylor series methods give the approximation $(1-\hat{\rho}^2)/\sqrt{n}$. Higher order calculations show that $(1-\hat{\rho}^2)/(n-3)^{\frac{1}{2}}$ is a better approximation to $\sigma_N(\hat{\rho})$. More details are given in § 6; see also Johnson and Kotz (1970, p. 229).

A smoothed bootstrap is produced by compromising between $\hat{F}_N$, the normal theory maximum likelihood estimate of $F$, and $\hat{F}$, the nonparametric maximum likelihood estimate. Define $\hat{F}_{0·5} = \hat{F} * (0·5\hat{F}_N)$, where $0·5\hat{F}_N$ represents the distribution $\mathcal{N}_2(0·5\bar{x}, 0·25\Omega)$, and "$*$" indicates convolution. The distribution $\hat{F}_{0·5}$ has the same correlation as both $\hat{F}$ and $\hat{F}_N$, namely the observed value $\hat{\rho}$, but is intermediate in smoothness between the two. To use more familiar terminology, it is a smoothed window estimate of the unknown $F$.

The normal smoothed bootstrap, line 3 of Table 1, generated the $X_i^*$ from $\hat{F}_{0·5}$, at step (ii) of the bootstrap algorithm. It performed somewhat better than the unsmoothed bootstrap for estimating $\sigma(\hat{\rho})$, and much better for estimating $\sigma(\hat{\phi})$. However, normal smoothing is suspiciously selfserving here, since the true distribution of the $X_i$ is itself normal. Lines 4 and 5 of the table used uniform smoothing. The $X_i^*$ are drawn from $\hat{F} * (0·5\hat{F}_U)$, where $\hat{F}_U$ is the uniform distribution over a rhombus selected such that $\hat{F}_U$ has the same covariance matrix as $\hat{F}$.

## 5. THE JACKKNIFE

Tukey's jackknife estimate of standard error (Miller, 1974) is defined in terms of the quantities $\hat{\rho}_{(i)} = \hat{\rho}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$:

$$\sigma_J(\hat{\rho}) = \left[\frac{n-1}{n} \sum_{i=1}^{n} \{\hat{\rho}_{(i)} - \hat{\rho}_{(\cdot)}\}^2\right]^{\frac{1}{2}}, \quad \hat{\rho}_{(\cdot)} = \sum_{i=1}^{n} \hat{\rho}_{(i)}/n. \tag{5·1}$$

In our Monte Carlo experiment the jackknife results, line 6, have considerably larger MSE$^{\frac{1}{2}}$ than does the bootstrap, for both $\sigma_J(\hat{\rho})$ and $\sigma_J(\hat{\phi})$.

There is a simple connexion between the jackknife and the bootstrap. This is easiest to see for statistics like the correlation coefficient which are functionals. For example, $\hat{\rho} = \rho(\hat{F})$, where $\rho(F)$ is the function which assigns any bivariate distribution $F$, with finite second moments, the value of its Pearson correlation coefficient.

The jackknife and the bootstrap both work by evaluating $\rho(F)$ for other values of $F$ besides $\hat{F}$. Let $P = (P_1, \ldots, P_n)$ be a probability vector, having nonnegative coordinates summing to one, and define the reweighted empirical probability distribution

$$\hat{F}(P)\text{: mass } P_i \text{ on } x_i \quad (i = 1, \ldots, n). \tag{5·2}$$

Corresponding to each 'resampling vector' $P$ is a resampled value of the statistic of interest,

$$\hat{\rho}(P) = \rho\{F(P)\}. \tag{5·3}$$

For instance, $P^0 = (1, \ldots, 1)/n$ corresponds to the observed value $\hat{\rho} = \rho(F)$, while

$P_{(i)} = (1, 1, \ldots, 0, \ldots, 1)/(n-1)$, 0 in the $i$th place, corresponds to the jackknife quantity $\hat{\rho}_{(i)}$. Definitions (5·2)–(5·3) emphasize the crucial feature of all the methods considered here: the data $x_1, \ldots, x_n$ are held fixed, the weights $P_1, \ldots, P_n$ are varied, and $\sigma(\hat{\rho})$ estimated from the variation in $\hat{\rho}(P)$. The methods differ in their choice of resampling vectors. For example, another way to describe the bootstrap estimate $\sigma_B(\hat{\rho})$ is to say that the resampling vectors are drawn multinomially,

$$P \sim \text{Mult}_n(n, P^0)/n, \tag{5·4}$$

$n$ draws on $n$ categories, probability $1/n$ for each category on each draw. Here $P_i$ equals card $\{X_j^* = x_i\}/n$ at step (ii) of the bootstrap procedure. Then $\sigma_B(\hat{\rho}) = [\text{var}_*\{\hat{\rho}(P)\}]^{\frac{1}{2}}$, where $\text{var}_*$ denotes variance under distribution (5·4).

There is a unique linear function of $P$, say $\hat{\rho}_L(P)$, which has values $\hat{\rho}_L(P_i) = \hat{\rho}_{(i)}$ $(i = 1, \ldots, n)$; $\hat{\rho}_L(P)$ is a convenient linear approximation to $\hat{\rho}(P)$. A simple calculation shows that $\sigma_J(\hat{\rho}) = \{n/(n-1)\}^{\frac{1}{2}}[\text{var}_*\{\hat{\rho}_L(P)\}]^{\frac{1}{2}}$ with $\text{var}_*$ still indicating variance under (5·4). In other words, except for the factor $\{n/(n-1)\}^{\frac{1}{2}}$, the jackknife estimate is itself a bootstrap estimate, but applied to $\hat{\rho}_L(P)$ rather than $\hat{\rho}(P)$. The jackknife formula involves less computation because the variance of a linear function can be calculated, without Monte Carlo, from the known covariance matrix of the multinomial (5·4).

For statistics which are averages, say $\hat{\rho} = \Sigma\, Y(X_i)/n$, where $Y(X)$ is real-valued, the factor $n/(n-1)$ makes the jackknife variance estimate $\{n/(n-1)\}\,\text{var}_*\{\hat{\rho}_L(P)\}$ unbiased for the true variance of $\hat{\rho}$. In this case, $\hat{\rho}_L(P) = \hat{\rho}(P)$. Multiplying the bootstrap variance estimate $\text{var}_*\{\hat{\rho}(P)\}$ by $n/(n-1)$ also makes it unbiased for the variance of an average, but does not seem to improve estimation in general.

The bootstrap vectors tend to be much further away from the central value $P^0$ than are the jackknife resampling vectors: $\| P - P^0 \| = O_p(1/\sqrt{n})$ from (5·4), compared to $\| P_{(i)} - P^0 \| = O(1/n)$. The jackknife estimate $\sigma_J(\hat{\rho})$ involves extrapolating from the local behaviour of $\hat{\rho}(P)$ near $P^0$, and this can cause trouble for 'unsmooth' statistics $\hat{\rho}$ such as the sample median (Miller, 1974).

## 6. Infinitesimal jackknife, influence function and the delta method

Rather than approximating $\hat{\rho}(P)$ by the linear function $\hat{\rho}_L(P)$, it seems more natural to approximate it by $\hat{\rho}_T(P)$, the first-order Taylor series for the function $\hat{\rho}(P)$ expanded about the central point $P = P^0$. The obvious estimate of standard error is then $\sigma_{IJ}(\hat{\rho}) = [\text{var}_*\{\hat{\rho}(P)\}]^{\frac{1}{2}}$, with $\text{var}_*$ indicating variance under distribution (5·4). This is exactly Jaeckel's infinitesimal jackknife (Miller, 1974; Efron, 1979a, § 5).

The infinitesimal jackknife replaces the finite differences $\hat{\rho}_{(i)} - \hat{\rho}_{(\cdot)}$ used in the ordinary jackknife by derivatives

$$\hat{d}_i = \lim_{\varepsilon \to 0}[\hat{\rho}\{P^0 + \varepsilon(\delta_i - P^0)\} - \hat{\rho}(P^0)]/\varepsilon,$$

where $\delta_i$ is the $i$th coordinate vector. Jaeckel's estimate can be written as $\sigma_{IJ}(\hat{\rho}) = (\Sigma\, \hat{d}_i^2/n^2)^{\frac{1}{2}}$. Three facts should be noted:

(i) The $\hat{d}_i$ are values of what C. L. Mallows, in an unpublished paper, calls the 'empirical influence function'; Jaeckel's formula is the obvious finite sample estimate based on the asymptotic expression for $n$ times the variance

$$\int \text{IF}^2(x)\, dF(x),$$

where $\text{IF}(x)$ is the influence function (Hampel, 1974; Huber, 1972).

(ii) For a linear statistic, i.e. an average, $\Sigma\hat{d}_i^2/n^2$ must be multiplied by $n/(n-1)$ to give an unbiased estimate of variance. Multiplying the estimates $\sigma_{IJ}$ by $\{n/(n-1)\}^{\frac{1}{2}} = (14/13)^{\frac{1}{2}} = 1{\cdot}038$ helps correct the severe downwards bias evident in line 7 of the table, but not by much.

(iii) Closed-form expressions for $\hat{d}_i$ can be computed for many statistics, including the correlation coefficient (Devlin, Gnanadesikan and Kettenring, 1975), but it is usually easier just to substitute a small value of $\varepsilon$ into

$$\hat{d}_i \doteq [\hat{\rho}\{P^0 + \varepsilon(\delta_i - P^0)\} - \hat{\rho}(P^0)]/\varepsilon.$$

The value $\varepsilon = 0{\cdot}001$ was used for line 7 of Table 1.

For the delta method we express the correlation coefficient (2·2) as

$$\hat{\rho}(X_1, \ldots, X_n) = t(\bar{Q}_1, \ldots, \bar{Q}_R), \tag{6·1}$$

where $t$ is a known function and each $\bar{Q}_r$ is an observed average,

$$\bar{Q}_r = \frac{1}{n}\sum_{i=1}^{n} Q_r(X_i). \tag{6·2}$$

For the correlation coefficient $R = 5$ and, in terms of $X = (U, V)$, $Q_1 = U$, $Q_2 = V$, $Q_3 = U^2$, $Q_4 = UV$, $Q_5 = V^2$, with $t = (\bar{Q}_4 - \bar{Q}_1\bar{Q}_2)/\{(\bar{Q}_3 - \bar{Q}_1^2)^{\frac{1}{2}}(\bar{Q}_5 - \bar{Q}_2^2)^{\frac{1}{2}}\}$.

Suppose that the vector $Q = \{Q_1(X), Q_2(X), \ldots, Q_R(X)\}$, corresponding to one observation of $X \sim F$, has mean vector $\alpha_F$ and covariance matrix $\beta_F$, and let $\nabla_F$ be the gradient vector with $r$th component $[\partial t/\partial Q_r]_{Q=\alpha_F}$. A first-order Taylor series expansion of (6·1) gives the approximation

$$\sigma(\hat{\rho}) \doteq (\nabla_F \beta_F \nabla_F'/n)^{\frac{1}{2}}. \tag{6·3}$$

In the specific case of the correlation coefficient (6·3) gives

$$\sigma(\hat{\rho}) \doteq \left[\frac{\rho^2}{4n}\left\{\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}}\right\}\right]^{\frac{1}{2}}, \tag{6·4}$$

where $\mu_{gh} = E_F[(U - E_F(U))^g\{V - E_F(V)\}^h]$; see Cramér (1946, p. 359). For $F$ bivariate normal, (6·4) reduces to $(1-\rho^2)/\sqrt{n}$, as in §3.

To use the delta method in a practical problem it is necessary to estimate $F$ in (6·3). Substitution of $\hat{F}$ for $F$ gives the nonparametric delta method estimate of standard error,

$$\sigma_D(\hat{\rho}) = (\nabla_{\hat{F}} \beta_{\hat{F}} \nabla_{\hat{F}}'/n)^{\frac{1}{2}}. \tag{6·5}$$

In the case of the correlation coefficient, for example, (6·5) is (6·4) with $\hat{\rho}$ replacing $\rho$ and the sample moments $\hat{\mu}_{gh}$ replacing the $\mu_{gh}$.

THEOREM. *For any statistic $\hat{\rho}$ of form* (6·1), *the nonparametric delta method and the infinitesimal jackknife give identical estimates of standard error.*

*Proof.* The derivatives

$$\hat{d}_i = \lim_{\varepsilon \to 0}[\hat{\rho}\{P^0 + \varepsilon(\delta_i - P^0)\} - \hat{\rho}(P^0)]/\varepsilon,$$

for $\hat{\rho} = t(\bar{Q})$ as in (6·1), are $\hat{d}_i = \nabla_{\hat{F}}\{Q(x_i) - \bar{Q}\}'$, since

$$\hat{\rho}\{P^0 + \varepsilon(\delta_i - P^0)\} = t[\bar{Q} + \varepsilon\{Q(x_i) - \bar{Q}\}] \doteq t(\bar{Q}) + \varepsilon\nabla_{\hat{F}}\{Q(x_i) - \bar{Q}\}'.$$

Therefore the infinitesimal jackknife estimate of standard error is

$$(\Sigma \hat{d}_i^2/n^2)^{\frac{1}{2}} = \left[\nabla_{\hat{F}} \frac{\Sigma \{Q(x_i)-\bar{Q}\}' \{Q(x_i)-\bar{Q}\}}{n} \nabla'_{\hat{F}}/n\right]^{\frac{1}{2}} = (\nabla_{\hat{F}} \beta_{\hat{F}} \nabla'_{\hat{F}}/n)^{\frac{1}{2}}, \tag{6·6}$$

agreeing with the delta method estimate (6·5). Here we have used the fact that $\Sigma \{Q(x_i)-\bar{Q}\}' \{Q(x_i)-\bar{Q}\}/n = \mathrm{cov}_{F=\hat{F}}(Q)$ the covariance matrix of the random vector $Q$ under the distribution $\hat{F}$, and so must equal $\beta_{\hat{F}}$, and likewise

$$[(\ldots, \partial t/\partial Q_r, \ldots)]_{Q=\bar{Q}} = [(\ldots, \partial t/\partial Q_r, \ldots)]_{Q=\alpha_{\hat{F}}} = \nabla_{\hat{F}}.$$

Deriving expressions like (6·4) involves considerable theoretical labour, especially for more complicated statistics. According to the theorem, this can be avoided by using Jaeckel's formula, with numerical evaluation of the $\hat{d}_i$, as in remark (iii) above.

Line 7 of Table 1 shows that the delta method can be badly biased. Better estimates than simply substituting $\hat{F}$ for $F$ might help, but no general theory exists. The ordinary jackknife has superior bias properties (Efron and Stein, 1981).

An intriguing question is 'why not perturb the $x_i$ and keep the weights $1/n$ constant, instead of vice versa as with the infinitesimal jackknife?' The theorem shows that the results will be the same for statistics of form (6·1). In this sense there is only one delta theory.

The delta method and infinitesimal estimates of bias are also identical. These estimates are $\frac{1}{2}\,\mathrm{tr}\,(\beta_{\hat{F}} \nabla^2_{\hat{F}})$ and $\Sigma \hat{e}_{ii}/(2n^2)$ respectively, where $\nabla^2_F$ is the matrix with $(rs)$th element $[\partial^2 t/\partial Q_r\, \partial Q_s]_{Q=\alpha_F}$, and $\hat{e}_{ii} = [\partial^2 \hat{\rho}\{P^0+\varepsilon(\delta_i-P^0)\}/\partial\varepsilon^2]_{\varepsilon=0}$. See Gray, Schucany and Watkins (1975), who also provide results closely related to the theorem above.

## 7. Half-sampling: Balanced repeated replications

Half-sampling methods come from the literature of sampling theory (Kish & Frankel, 1974), where it is natural to consider stratified situations. We suppose that the sample space is

$$\mathscr{X} = \bigcup_{h=1}^{H} \mathscr{X}_h,$$

where the $\mathscr{X}_h$ are disjoint strata; that there is an unknown probability distribution $F_h$ defined on each $\mathscr{X}_h$; that the data consist of independent random samples

$$X_{hi} \sim F_h \quad (i = 1, \ldots, n_h;\ h = 1, \ldots, H); \tag{7·1}$$

that the statistic of interest is of the form $\hat{\rho} = \rho(\hat{F}_1, \ldots, \hat{F}_H)$, where $\hat{F}_h$ is the distribution putting mass $1/n_h$ at each observed $x_{hi}$; and finally that we wish to attach a standard error to $\hat{\rho}$.

The obvious bootstrap algorithm is

(i) construct the $\hat{F}_h$;

(ii) draw $X_{hi}$ independently from $\hat{F}_h$ $(i = 1, \ldots, n_h;\ h = 1, \ldots, H)$ and let $\hat{\rho}^* = \rho(\hat{F}^*_1, \ldots, \hat{F}^*_H)$, where $\hat{F}^*_h$ is the empirical distribution of $X^*_{h1}, \ldots, X^*_{hn_h}$;

(iii) estimate $\sigma(\hat{\rho})$ by $\sigma_{\mathrm{HS}}(\hat{\rho}) = \{\mathrm{var}_*(\hat{\rho}^*)\}^{\frac{1}{2}}$, $\mathrm{var}_*$ indicating variance under probability mechanism (ii).

If $\hat{\rho}$ is a linear statistic, say

$$\hat{\rho} = \sum_{h=1}^{H} w_h \bar{Y}_h,$$

where the $w_h$ are fixed weights and $\bar{Y}_h = \Sigma\, Y(X_{hi})/n_h$, for some attribute of interest $Y(\,.\,)$, then the bootstrap estimate of variance is biased downward; if $\mathrm{var}_{F_h}\{Y(X_h)\} = \beta_h^2$, then

$$E\{\sigma^2_{\mathrm{HS}}(\hat{\rho})\} = \Sigma\,\{(n_h-1)/n_h\}\, w_h^2\, \beta_h^2/n_h$$

compared to the true variance $\Sigma\, w_h^2\, \beta_h^2/n_h$.

Half-sampling corrects this bias by reducing each bootstrap sample from size $n_h$ to size $n_h-1$, at step (ii) of the algorithm. This makes $E\{\sigma^2_{\mathrm{HS}}(\hat{\rho})\} = \Sigma\, w_h^2\, \beta_h^2/n_h$, the correct answer. In the most commonly considered case, where all the $n_h = 2$, each reduced bootstrap sample is indeed a half-sample, consisting of one of the two observed values $x_{h1}$ or $x_{h2}$ for each stratum $h$.

Line 8 of Table 1 shows half-sampling applied to $\hat{\rho}$, the correlation coefficient. The strata were defined artificially, $(x_1, x_2)$ represented stratum 1, $(x_3, x_4)$ stratum 2, ..., $(x_{13}, x_{14})$ stratum 7. For each of the 200 trial samples $x_1, \ldots, x_{14}$, all $2^7 = 128$ half-samples were constructed, yielding half-sample correlations $\hat{\rho}^*(1), \ldots, \hat{\rho}^*(128)$, and standard error estimate

$$\sigma_{\mathrm{HS}}(\hat{\rho}) = \left[\sum_{j=1}^{128} \{\hat{\rho}^*(j) - \hat{\rho}^*(\,.\,)\}^2/128\right]^{\frac{1}{2}}.$$

Notice that this is actually $\sigma_{\mathrm{HS}}(\hat{\rho})$, and not a Monte Carlo approximation, since we have considered all 128 possible reduced bootstrap samples. Usually $\hat{\rho}^*(\,.\,)$ is replaced by $\hat{\rho}$ in the formula for $\sigma_{\mathrm{HS}}(\hat{\rho})$, but this has almost no effect on the numbers reported here.

The numerical results shown in line 8 are discouraging. Both bias and root mean squared error are high, for both $\sigma_{\mathrm{HS}}(\hat{\rho})$ and $\sigma_{\mathrm{HS}}(\hat{\phi})$. Of course this is not a naturally stratified situation, so there is no particular reason to do half-sampling, but J. W. Tukey, in unpublished notes and lectures, has advocated half-sampling for this type of problem, particularly for statistics like the median where the jackknife fares poorly. In line 9 of the table, the standard error estimates for each of the 200 trials were constructed using 128 randomly selected half-samples, out of all $14!/(7!)^2$ possible ones. This method removes the component of variance in $\sigma_{\mathrm{HS}}(\hat{\rho})$ due to the artificial creation of strata, but the numerical results are still poor.

McCarthy (1969) suggested an interesting shortcut method for reducing the number of half-sample calculations, balanced repeated replication. Rather than look at all $2^{\frac{1}{2}n}$ possible half-samples, he pointed out that a 'balanced' subcollection of the half-samples gives exactly the same estimate of standard error when $\hat{\rho}$ is a linear statistic. Balanced here has a technical definition related to orthogonality. Line 10 of Table 1 is based on the eight balanced half-samples defined by the first seven rows of the matrix of McCarthy (1969, p. 243). For each of the 200 samples,

$$\sigma_{\mathrm{BHS}}(\hat{\rho}) = \left[\sum_{j=1}^{8} \{\hat{\rho}^*(j) - \hat{\rho}^*(\,.\,)\}^2/8\right]^{\frac{1}{2}},$$

Line 10 shows results similar to lines 8 and 9, with somewhat worse root mean squared error.

Corresponding to each half-sample is the complementary half-sample consisting of those elements which do not appear in the former. If $\hat{\rho}^*(j)$ is the half-sample value of the statistic, let $\tilde{\rho}^*(j)$ be its value for the complementary half-sample. Line 11 of the table is based on the complementary half-sample estimate of standard error,

$$\sigma_{\mathrm{CHS}}(\hat{\rho}) = \left(\sum_{j=1}^{64} [\{\hat{\rho}^*(j) - \tilde{\rho}^*(j)\}/2]^2/64\right)^{\frac{1}{2}}.$$

Here the indexing is such that none of the first 64 half-samples is complementary to another. It is easy to show that the formula is always numerically smaller than the noncomplementary version used in line 8. This reduces the bias in our particular case, though not necessarily in general.

Complementation and balancing are combined in line 12. Here

$$\sigma_{\mathrm{CBHS}}(\hat{\rho}) = \left[\sum_{j=1}^{8} \{[\hat{\rho}^*(j) - \tilde{\rho}^*(j)]/2\}^2/8\right]^{\frac{1}{2}},$$

where the $\hat{\rho}^*(j)$ refer to the 8 balanced half-samples discussed previously. This method involved 16 $\hat{\rho}$ recomputations for each trial, rather than 128, and gave almost the same results as those of line 11. The author has shown that the results would be exactly the same for 'quadratic functionals', defined by Efron and Stein (1981), which are the next step past linear statistics.

## 8. Random subsampling (typical values)

Hartigan (1969, 1971) and Hartigan and Forsythe (1970) discuss another interesting resampling plan which may be described as random subsampling: from the collection of $2^n - 1$ nonempty subsets of $\{x_1, \ldots, x_n\}$, draw subsets $S(1), \ldots, S(N)$ randomly and without replacement, $N \leqslant 2^n - 1$. Each subset $S$ determines an empirical distribution $\hat{F}_S$, putting mass $1/n_S$ on each element of $S$, where $n_S$ is the number of such elements. These in turn determine the resampled values of the statistic of interest, $\hat{\rho}^*(j) = \rho(\hat{F}_{S(j)})$ for $j = 1, \ldots, N$.

Consider the case where $\mathscr{X}$ is the real line, $F$ is a distribution known to be symmetric about an unknown central point $\rho$, and $\hat{\rho}$ is an $M$-estimate, i.e. the solution to $\Sigma\psi(x_i - \rho) = 0$, where $\psi(t)$ is strictly monotonic and $\psi(-t) = -\psi(t)$. Hartigan (1969) demonstrated the following result, called the typical value theorem: the ordered values of the $\hat{\rho}^*(j)$, say $\hat{\rho}^*[1] < \hat{\rho}^*[2] < \ldots < \hat{\rho}^*[N]$, divide $\mathscr{X}$ into $N+1$ intervals, each of which has probability $1/(N+1)$ of containing the true value $\rho$. See also Maritz (1979). The interval $(\hat{\rho}^*[N_1], \hat{\rho}^*[N_2])$, where $N_1 = [\alpha(N+1)]$, $N_2 = [(1-\alpha)(N+1)]$, is then a $1-2\alpha$ central confidence interval for $\rho$.

Given such a neat result, there is a temptation to use random subsampling to obtain accuracy estimates in more general problems. Hartigan (1969) considers setting confidence intervals for sample variances and eigenvalues. Line 13 of the table shows its application to standard error estimation for the correlation coefficient; $N = 128$ and $\sigma_{\mathrm{RS}}(\hat{\rho}) = \Sigma\{\hat{\rho}^*(j) - \hat{\rho}^*(\,.\,)\}^2$. The results are badly biased upwards, especially for $\hat{\phi}$. In order to avoid even worse biases, only subsets $S$ with $n_S \geqslant 4$ were allowed. Somewhat better results are obtained by using a more robust estimate of standard error, $\frac{1}{2}(\hat{\rho}^*[N_2] - \hat{\rho}^*[N_1])$, where $N_1 = [0{\cdot}16(N+1)]$, $N_2 = [0{\cdot}84(N+1)]$, as shown on line 14 of the table. This version of $\sigma_{\mathrm{RS}}(\hat{\rho})$ is one half the length of the putative 68% central confidence interval for $\rho$; see Efron (1979b).

The random subsample method belongs to a large class of resampling techniques, including the bootstrap and half-sampling, which have equivalent asymptotic properties, at least to a first order of approximation. Consider an arbitrary resampling plan in which we assume only that the vector $P$ is selected randomly, according to a distribution invariant under permutations of its coordinates. Then $P$ has mean vector and covariance matrix

$$P \sim \left(\frac{e}{n}, \frac{n}{n-1}\left(I - \frac{e'e}{n}\right)\mathrm{var}_*(P_1)\right), \tag{8·1}$$

where $e = (1, \ldots, 1)$, and $\mathrm{var}_*(P_1)$ is the variance of $P_1$ under the resampling distribution. This follows from symmetry and the equality

$$0 = \mathrm{var}_*(\Sigma P_i) = n\,\mathrm{var}_*(P_1) + n(n-1)\,\mathrm{cov}_*(P_1, P_2).$$

Relation (8·1) implies that a resampled average $\bar{Y}^* = \Sigma P_i Y(x_i) = \Sigma P_i y_i$ has mean and variance

$$\bar{Y}^* \sim \left(\bar{y}, \frac{1}{n-1}\Sigma(y_i-\bar{y})^2\,\mathrm{var}_*(P_1)\right) \tag{8·2}$$

under the resampling distribution. We consider three cases.

*Case 1, bootstrap.* Here $\mathrm{var}_*(P_1) = (n-1)/n^3$, by (5·4), so that (8·2) gives $\mathrm{var}_*(\bar{Y}^*) = \Sigma(y_i-\bar{y})^2/n^2$.

*Case 2, random half-sampling.* Randomly chosen subsamples of size $\frac{1}{2}n$, as in line 9 of Table 1, give $\mathrm{var}_*(P_1) = 1/n^2$ and $\mathrm{var}_*(\bar{Y}) = \Sigma(y_i-\bar{y})^2/\{n(n-1)\}$, the usual estimate of variance for an average.

*Case 3, random subsampling.* Here $\mathrm{var}_*(P_1) = (n+2)n^{-3}\{1+o(1/n)\}$, so that

$$\mathrm{var}_*(\bar{Y}^*) = \frac{n+2}{n}\frac{\Sigma(y_i-\bar{y})^2}{n(n-1)}\left\{1+o\left(\frac{1}{n}\right)\right\}.$$

The point here is that any resampling plan having $P$ exchangeable and $\mathrm{var}_*(P_1) = n^{-2}\{1+O(1/n)\}$ gives asymptotically the same value of $\mathrm{var}_*(\bar{Y}^*)$. Following Efron (1979a, § 8, remark G), result (8·1) shows that this asymptotic equivalence extends beyond linear statistics $\bar{Y}$ to a wide class of smoothly defined random quantities. However, as we have seen in the case of the correlation coefficient, the asymptotics cannot be completely trusted; the different methods can lead to quite different results in small samples.

## 9. FINAL COMMENTS

We conclude with four miscellaneous comments.

What is the purpose of estimating a standard error? At the most basic level, the concept of root mean squared error conveys its own meaning about the accuracy of a point estimate $\hat{\rho}$, useful for comparing error distributions which are roughly normal. Asymptotic normal theory leads to approximate confidence intervals of the form $\hat{\rho} \pm z_\alpha \hat{\sigma}$, where $\hat{\sigma}$ is the estimated standard error and $z_\alpha$ is taken from the normal table. Student's $t$ theory gives confidence intervals of the form $\hat{\rho} \pm t_{\alpha,n}\hat{\sigma}$. Most of the jackknife literature is phrased in terms of these last, but in fact no general theory has been verified beyond the normal level; see Miller (1974).

It would be nice to have a more satisfactory theory of small-sample nonparametric confidence intervals. The main problem is to capture correctly the asymmetry about $\hat{\rho}$ exhibited by parametric confidence intervals, which is usually of greater magnitude than the $t$ effect mentioned above, $O(1/n^{\frac{1}{2}})$ compared to $O(1/n)$. The author discusses one approach, a bootstrap method similar to Hartigan's typical value theorem, in §§ 5 and 6 of Efron (1981), but the problem is still largely untouched.

It is not surprising that the bootstrap performs best among the genuinely nonparametric methods in Table 1, since the bootstrap estimate is the nonparametric maximum likelihood estimate of the standard error. If we want to do better, we have to use some form of estimation which is not truly nonparametric. The smoothed bootstraps,

lines 3–5, bias the estimation process towards smooth underlying models. This produces substantial gains in the present case, especially for $\hat{\phi}$, but again no general theoretical guidelines exist.

The biases reported in Table 1 look quite different when reported in terms of estimating variances rather than estimating standard errors. For example, in line 1, the bootstrap estimates $\sigma_B^2(\hat{\rho})$ averaged 0·0468 compared to the true value $\mathrm{var}(\hat{\rho}) = 0{\cdot}0488$. The jackknife variance estimates $\sigma_J(\hat{\rho})^2$ averaged 0·0569. In terms of the first comment, it seems more meaningful to discuss the results in terms of standard errors rather than variances.

## References


Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Devlin, S., Gnanadesikan, R. & Kettenring, J. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531–46.

Efron, B. (1979a). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Efron, B. (1979b). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review* **21**, 460–80.

Efron, B. (1981). Censored data and the bootstrap. *J. Am. Statist. Assoc.* **76**, 312–9.

Efron, B. & Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9**, 586–96.

Gray, H., Schucany, W. & Watkins, T. (1975). On the generalized jackknife and its relation to statistical differentials. *Biometrika* **62**, 637–42.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69**, 383–93.

Hartigan, J. A. (1969). Using subsample values as typical values. *J. Am. Statist. Assoc.* **64**, 1303–17.

Hartigan, J. A. (1971). Error analysis by repeated samples. *J. R. Statist. Soc.* B **33**, 98–110.

Hartigan, J. A. & Forsythe, A. (1970). Efficiency and confidence intervals generated by repeated subsample calculations. *Biometrika* **57**, 629–40.

Hinkley, D. V. (1978). Improving the jackknife with special reference to correlation estimation. *Biometrika* **65**, 13–22.

Huber, P. J. (1972). Robust statistics: A review. *Ann. Math. Statist.* **43**, 1041–67.

Johnson, N. L. & Kotz, S. (1970). *Continuous Univariate Distributions* **2**. Boston: Houghton Mifflin.

Kish, L. & Frankel, M. (1974). Inference from complex samples (with discussion). *J. R. Statist. Soc.* B **36**, 1–37.

McCarthy, P. J. (1969). Pseudo-replication: Half-samples. *Rev. I.S.I.* **37**, 239–63.

Maritz, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66**, 163–6.

Miller, R. G. (1974). The jackknife—a review. *Biometrika* **61**, 1–16.


[*Received June* 1980. *Revised January* 1981]