The Canadian Journal of Statistics
Vol. 9, No. 2, 1981, Pages 139–172
La Revue Canadienne de Statistique

Nonparametric standard errors and confidence intervals*

Bradley EFRON

Stanford University

Key words and phrases: Bootstrap, jackknife, delta method, nonparametric confidence intervals, nonparametric standard errors.
AMS 1980 *subject classification:* Primary 62G05.

ABSTRACT

We investigate several nonparametric methods; the bootstrap, the jackknife, the delta method, and other related techniques. The first and simplest goal is the assignment of nonparametric standard errors to a real-valued statistic. More ambitiously, we consider setting nonparametric confidence intervals for a real-valued parameter. Building on the well understood case of confidence intervals for the median, some hopeful evidence is presented that such a theory may be possible.

1. INTRODUCTION

This article concerns nonparametric methods for estimating standard errors and confidence intervals. The discussion focuses on the bootstrap (Efron 1979a), which is easy to motivate and connects nicely to the jackknife, the delta method, and other more or less familiar nonparametric techniques.

We begin on relatively firm ground—assigning a nonparametric estimate of standard error to a real-valued statistic. Whether or not there exists a useful theory of nonparametric small-sample confidence intervals is still a matter of speculation. Building on the well-understood case of confidence intervals for the median, we offer some hopeful, if not conclusive, evidence that such a theory may be possible. Several different methods, all related to the bootstrap, are compared in two simple situations.

The author appreciates the invitation to present these ideas for discussion. There has been surprisingly little public discussion of jackknife-related methods, considering their potential importance to the practicing statistician. A much longer review of the subject, including related topics such as cross-validation, resampling, subsampling, half-sampling, and influence functions, appears in Efron (1980c). Most of the presentation here is drawn from that report, and also Efron (1980b, 1980a, 1979a), some additional material appearing in the latter sections.

2. THE BOOTSTRAP ESTIMATE OF THE STANDARD ERROR

We have a real-valued statistic $\hat{\theta}(X_1, X_2, \ldots, X_n)$, which is a function of n independent identically distributed observations

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F, \tag{2.1}$$

* Presented as an invited discussion paper at the 1981 Annual Meeting of the Statistical Society of Canada, Dalhousie University, Halifax, Nova Scotia, May 24 & 25, 1981. The texts of the discussants' comments are presented along with written submissions from George A. Barnard, V.T. Farewell, Ross L. Prentice, Paul Switzer, and W.G. Warren, who were unable to attend the meeting.

F being an unknown probability distribution on a space $\mathscr{X}$. Having observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, we wish to attach an estimate of standard error to $\hat{\theta}$.

EXAMPLE 1. $\mathscr{X} = \mathbb{R}^1$, the real line, and $\hat{\theta} = 25\%$ trimmed mean, i.e., the average of the central 50% of the sample.

EXAMPLE 2. $\mathscr{X} = \mathbb{R}^2$, the plane, and $\hat{\theta} = \hat{\rho}(X_1, X_2, \ldots, X_n)$, the Pearson product moment correlation coefficient for the observed sample.

The true standard error of $\hat{\theta}$ is a function of F, n, and the form of the statistic $\hat{\theta}$, say

$$\sigma(F, n, \hat{\theta}(\cdot, \cdot, \ldots, \cdot)) = \sigma(F). \tag{2.2}$$

This last notation emphasizes that, knowing n and the form of $\hat{\theta}$, the true standard error is only a function of the unknown distribution F.

The bootstrap estimate of the standard error, $\hat{\sigma}_B$, is simply

$$\hat{\sigma}_B = \sigma(\hat{F}), \tag{2.3}$$

where $\hat{F}$ is the empirical probability distribution

$$\hat{F}\!: \text{mass} \frac{1}{n} \text{ on } x_i, \qquad i = 1, 2, \ldots, n. \tag{2.4}$$

As a simple example suppose $\mathscr{X} = \mathbb{R}^1$ and $\hat{\theta} = \bar{X} = \sum X_i/n$, the average, in which case $\sigma(F) = [\mu_2(F)/n]^{1/2}$, where $\mu_2(F) = \int_{-\infty}^{\infty} (x - \mathscr{E}_F X)^2 dF(x)$. Then $\hat{\sigma}_B = [\hat{\mu}_2/n]^{1/2}$, where $\hat{\mu}_2 = \sum(x_i - \bar{x})^2/n$.

In fact the function $\sigma(F)$ is usually impossible to express in simple form, and $\hat{\sigma}_B$ must be evaluated using a Monte Carlo algorithm:

Step 1. Construct $\hat{F}$ as at (2.4).

Step 2. Draw a *bootstrap sample* from $\hat{F}$,

$$X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F}, \tag{2.5}$$

and calculate $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)$.

Step 3. Independently do Step 2 some number B times, obtaining *bootstrap replications* $\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)$, and calculate

$$\hat{\sigma}_B = \left[\sum_{b=1}^{B} \frac{[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}\right]^{1/2}, \tag{2.6}$$

where $\hat{\theta}^*(\cdot) = \sum \hat{\theta}^*(b)/B$.

As $B \to \infty$, the right-hand side of (2.6) converges to $\hat{\sigma}(F)$. In practice, the author has found B in the range 50–200 adequate for estimating standard errors. The point is discussed further below. Larger values of B are required for the confidence interval calculations of Sections 4–10.

Tables 1 and 2 report on sampling experiments relating to Examples 1 and 2 respectively. Two distributions F are investigated in Table 1, a standard normal, $F \sim \mathbf{N}(0, 1)$, and a standard negative exponential $F \sim G_1$. The sample size is $n = 15$; $\hat{\sigma}_B$ is based on $B = 200$ bootstrap replications in each trial. ["Trial" refers to a new choice of the data $X_1, X_2, \ldots, X_n \sim F$. "Replication" refers to a selection of the bootstrap sample $X_1^*, X_2^*, \ldots, X_n^* \sim F$. By $\sim$ we here mean "independently and identically distributed as".] Summary statistics are given for 200 trials, both for $\hat{\sigma}_B$ and the jackknife estimate of standard error

TABLE 1: Estimates of standard error for the 25% trimmed mean using the jackknife and the bootstrap: 200 trials of $X_1, X_2, \ldots, X_{15} \overset{\text{iid}}{\sim} F$. The standard deviations of $\hat{\sigma}_B$, $\hat{\sigma}_J$ for the 200 trials show a moderate advantage for the bootstrap.

	Summary statistics for 200 trials					
	$F \sim \mathbf{N}(0, 1)$			$F \sim G_1$		
	Ave	Std Dev	CV	Ave	Std Dev	CV
Jackknife	.280	.084	.30	.224	.085	.38
Bootstrap (B = 200)	.287	.071	.25	.242	.078	.32
True standard error [minimum possible CV]	.289		[.19]	.232		[.27]

$$\hat{\sigma}_J = \left[\frac{n-1}{n}\sum(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2\right]^{1/2}, \tag{2.7}$$

$\hat{\theta}_{(i)} = \hat{\theta}(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, $\hat{\theta}_{(\cdot)} = \sum\hat{\theta}_{(i)}/n$. Both the bootstrap and the jackknife are nearly unbiased. The bootstrap performs better in that its coefficient of variation is lower. The bracketed figures show the minimum possible coefficient of variation for a scale-invariant estimate of standard error, assuming full knowledge of F. In the normal case, for example, 0.19 is the coefficient of variation of $[\sum(x_i - \bar{x})^2/14]^{1/2}$. (See Table 1.)

Table 2 compares several nonparametric estimates for the standard error of $\hat{\rho}$, the correlation coefficient, and also for $\hat{\phi} = \tanh^{-1}\hat{\rho}$. The true distribution F is bivariate normal with $\rho = 0.5$; the sample size is $n = 14$. The true standard errors are $\sigma\{\hat{\rho}\} = 0.218$, $\sigma\{\hat{\phi}\} = 0.299$ in this situation. Summary statistics for the estimates of the standard error are presented, based on 200 trials.

The bootstrap was run with $B = 128$ and also with $B = 512$. The increased effort of the latter value yielded only slightly less variable estimates $\hat{\sigma}_B$. A standard argument based on components of variance shows that further increases of B would be pointless. Taking $B = \infty$ would reduce the root-mean-square error of $\hat{\sigma}_B\{\hat{\rho}\}$, as an estimator of the true standard error $\sigma\{\hat{\rho}\}$, to 0.063. [The notation $\hat{\sigma}_B\{\hat{\theta}\}$ indicates the bootstrap estimate of $\sigma\{\hat{\theta}\}$, the true standard deviation for the statistic $\hat{\theta}$.] Likewise the value of $\sqrt{\text{MSE}}$ for $\hat{\sigma}_B\{\hat{\phi}\}$ would be reduced to 0.061. As a point of comparison, the normal-theory estimate $\hat{\sigma}_N\{\hat{\rho}\} = (1 - \hat{\rho}^2)/(n-3)^{1/2}$ has $\sqrt{\text{MSE}} = 0.056$.

Why not generate the bootstrap observations from a smoother estimate of F than $\hat{F}$? This is done in lines 3, 4, and 5 of Table 2. Let $\hat{\Sigma} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'/n$, the sample covariance matrix of the observed data. The *normal smoothed bootstrap* uses $X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F} \oplus \mathbf{N}_2(0, 0.25\,\hat{\Sigma})$, $\oplus$ indicating convolution. This amounts to estimating F by a normal window estimate, that is, by an equal mixture of the n distributions $\mathbf{N}_2(x_i, 0.25\,\hat{\Sigma})$. Smoothing gives a moderate improvement for estimating $\sigma\{\hat{\rho}\}$, and an enormous improvement for estimating $\sigma\{\hat{\phi}\}$. The latter result is suspect, since the true sampling situation is itself bivariate normal, and the function $\hat{\phi} = \tanh^{-1}\hat{\rho}$ is chosen to have $\sigma\{\hat{\phi}\}$ nearly constant in the bivariate normal family. The *uniform smoothed bootstrap* uses $X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} F \oplus \mathbf{U}(0, 0.25\,\hat{\Sigma})$, where $\mathbf{U}(0, 0.25\,\hat{\Sigma})$ is the uniform distribution on a rhombus selected to have mean vector 0 and covariance matrix $0.25\,\hat{\Sigma}$.

The standard-normal-theory estimates of line 8, Table 2, are themselves bootstrap estimates, carried out in a parametric framework. At step 1 of the bootstrap algorithm,

TABLE 2: Estimates of standard error for the correlation coefficient $\hat{\rho}$ and for $\hat{\phi} = \tanh^{-1} \hat{\rho}$. The sampling experiment consisted of 200 trials of $X_1, X_2, \ldots, X_{14} \sim$ bivariate normal, true $\rho = 0.5$. The delta-method estimates are badly biased downward. The jackknife estimates are substantially more variable than the bootstrap. From a larger table in Efron (1980b).

	Summary statistics for 200 trials							
	Standard error estimates for $\hat{\rho}$				Standard error estimates for $\hat{\phi}$			
	Ave	Std Dev	CV	$\sqrt{\text{MSE}}$	Ave	Std Dev	CV	$\sqrt{\text{MSE}}$
1. Bootstrap $B = 128$	.206	.066	.32	.067	.301	.065	.22	.065
2. Bootstrap $B = 512$	.206	.063	.31	.064	.301	.062	.21	.062
3. Normal smoothed bootstrap $B = 128$	.200	.060	.30	.063	.296	.041	.14	.041
4. Uniform smoothed bootstrap $B = 128$	.205	.061	.30	.062	.298	.058	.19	.058
5. Uniform smoothed bootstrap $B = 512$	.205	.059	.29	.060	.296	.052	.18	.052
6. Jackknife	.223	.085	.38	.085	.314	.090	.29	.091
7. Delta method (infinitesimal jackknife)	.175	.058	.33	.072	.244	.052	.21	.076
8. Normal theory	.217	.056	.26	.056	.302	0	0	.003
True standard error	.218				.299			

the fitted distribution is taken to be the parametric MLE $\hat{F}_N = \mathbf{N}_2(\bar{x}, \hat{\boldsymbol{\Sigma}})$, rather than the nonparametric MLE $\hat{F}$. The bootstrap observations $X_1^*, X_2^*, \ldots, X_n^*$ are drawn independently from $\hat{F}_N$ rather than from $\hat{F}$, and the algorithm proceeds as described in steps 2 and 3. This process isn't actually carried out. If it were, and if $B \to \infty$, then a high-order Taylor-series approximation shows that $\hat{\sigma}_N\{\hat{\rho}\} \simeq (1 - \hat{\rho}^2)/[n - 3]^{1/2}$, $\hat{\sigma}_N\{\hat{\phi}\} \simeq 1/(n - 3)^{1/2}$; see Johnson and Kotz (1970, p. 229). Notice that the normal smoothed bootstrap can be thought of as a compromise between using $\hat{F}$ and using $\hat{F}_N$ to begin the bootstrap process.

3. THE JACKKNIFE AND THE DELTA METHOD

We can restate the bootstrap idea in a way which makes clear its connection with the jackknife and the delta method. Suppose $\hat{\theta}$ is a *functional statistic*, i.e., of the form $\hat{\theta} = \theta(\hat{F})$, where $\theta(F)$ is a functional assigning a real number to any distribution F on $\mathscr{X}$. Both examples in Section 2 are of this form. Let $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ be a probability vector, having nonnegative weights summing to one, and define the reweighted empirical distribution

$$\hat{F}(\mathbf{P})\text{: mass } P_i \text{ on } x_i, \qquad i = 1, 2, \ldots, n. \tag{3.1}$$

Corresponding to $\mathbf{P}$ is a *resampled value* of the statistic, say $\hat{\theta}(\mathbf{P}) = \theta(\hat{F}(\mathbf{P}))$. The shorthand notation $\hat{\theta}(\mathbf{P})$ assumes that $x_1, x_2, \ldots, x_n$ are fixed at their observed values.

Another way to describe the bootstrap estimate $\hat{\sigma}_B$ is as follows. Let $\mathbf{P}^*$ indicate a vector drawn from the rescaled multinomial distribution

$$\mathbf{P}^* \sim \frac{\text{Mult}_n(n, \mathbf{P}^0)}{n} \qquad [\mathbf{P}^0 = (1, 1, \ldots, 1)/n], \tag{3.2}$$

meaning the observed proportions from n draws on n categories, with equal probability $1/n$ for each category. Then

$$\hat{\sigma}_B = [\mathscr{Var}_* \hat{\theta}(\mathbf{P}^*)]^{1/2}. \tag{3.3}$$

$\mathscr{Var}_*$ indicates variance under (3.2). (P_i^* equals $\#\{X_j^* = x_i\}/n$ in step 2 of the bootstrap algorithm.)

The jackknife values $\hat{\theta}_{(i)}$ equal $\hat{\theta}(\mathbf{P}_{(i)})$, where $\mathbf{P}_{(i)} = (1, 1, \ldots, 0, 1, 1)/(n - 1)$, 0 in the ith place. We can approximate $\hat{\theta}(\mathbf{P})$ by the linear function of $\mathbf{P}$, say $\hat{\theta}_L(\mathbf{P})$, having $\hat{\theta}_L(\mathbf{P}_{(i)}) = \hat{\theta}(\mathbf{P}_{(i)})$ for $i = 1, 2, \ldots, n$. It is easy to see that

$$\hat{\theta}_L(\mathbf{P}) = \hat{\theta}_{(\cdot)} + (\mathbf{P} - \mathbf{P}^0)\mathbf{U}, \tag{3.4}$$

where $\mathbf{U}$ is the column vector having ith coordinate

$$U_i = (n - 1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}). \tag{3.5}$$

THEOREM (Efron 1980b). *The jackknife estimate of standard error for $\hat{\theta}$ is*

$$\hat{\sigma}_J = \left[\frac{n}{n - 1} \mathscr{Var}_* \hat{\theta}_L(\mathbf{P}^*)\right]^{1/2}, \tag{3.6}$$

which is $[n/(n - 1)]^{1/2}$ times the bootstrap estimate of standard error for $\hat{\theta}_L$.

In other words the jackknife is, almost, a bootstrap itself. The factor $[n/(n - 1)]^{1/2}$ makes $\hat{\sigma}_J^2$ unbiased for σ^2 if $\hat{\theta}$ is a linear statistic, e.g., $\hat{\theta} = \bar{X}$. We could multiply $\hat{\sigma}_B$ by this same factor, but there doesn't seem to be any particular advantage to doing so.

NOTATION. *We shall use* $\mathcal{Var}_*$ *and* Prob$_*$ *to indicate variances and probabilities calculated from* (2.5), *which is equivalent to* (3.2). *In both* (2.5) *and* (3.2), *the observations* $x_1, x_2, \ldots, x_n$ *are fixed.*

The advantage of working with the linear approximation $\hat{\theta}_L$, rather than $\hat{\theta}$, is that there is no need for Monte Carlo calculations: $\mathcal{Var}_* \, \hat{\theta}_L(\mathbf{P}^*) = \mathcal{Var}_*(\mathbf{P}^* - \mathbf{P}^0)\mathbf{U} = \sum U_i^2/n^2$, using the known covariance of (3.2) and the fact that $\sum U_i = 0$. The disadvantage is usually increased error of estimation, as evidenced in Tables 1 and 2.

Instead of approximating $\hat{\theta}(\mathbf{P})$ by $\hat{\theta}_L(\mathbf{P})$, we could use the first-order Taylor-series expansion for $\hat{\theta}(\mathbf{P})$ about the point $\mathbf{P} = \mathbf{P}^0$,

$$\hat{\theta}_T(\mathbf{P}) = \hat{\theta}(\mathbf{P}^0) + (\mathbf{P} - \mathbf{P}^0)\mathbf{U}^0, \qquad U_i^0 = \lim_{\varepsilon \to 0} \frac{\hat{\theta}(\mathbf{P}^0 + \varepsilon(\boldsymbol{\delta}_i - \mathbf{P}^0)) - \hat{\theta}(\mathbf{P}^0)}{\varepsilon}, \tag{3.7}$$

$\boldsymbol{\delta}_i$ being the ith coordinate vector. This suggests the standard-error estimate

$$\hat{\sigma}_{\mathrm{IJ}} = [\mathcal{Var}_* \, \hat{\theta}_T(\mathbf{P}^*)]^{1/2} = [\textstyle\sum U_i^{0\,2}/n^2]^{1/2}, \tag{3.8}$$

with $\mathcal{Var}_*$ still indicating variance under (3.2). The initials IJ stand for *infinitesimal jackknife*, as defined by Jaeckel in his highly original 1972 paper. The ordinary jackknife takes $\varepsilon = -1/(n-1)$ in (3.7), while the infinitesimal jackknife lets $\varepsilon \to 0$, thereby earning its name.

The U_i^0 are values of the *empirical influence function* (Mallows 1974), their definition being an obvious nonparametric estimate of the true influence function

$$\mathrm{IF}(x) = \lim_{\varepsilon \to 0} \frac{\theta((1-\varepsilon)F + \varepsilon\delta_x) - \theta(F)}{\varepsilon}, \tag{3.9}$$

where δ_x is the degenerate distribution putting mass 1 on point x. The right side of (3.8) is then the obvious estimate of the influence-function approximation to the standard error of $\hat{\theta}$, due to Hampel (1974): $[\int IF^2(x)\,dF(x)/n]^{1/2}$. The empirical influence-function approach and the infinitesimal jackknife give identical estimates of standard error.

The *nonparametric delta method* applies to statistics of the form $t(\bar{Q}_1, \bar{Q}_2, \ldots, \bar{Q}_A)$, where $t(\cdot, \cdot, \ldots, \cdot)$ is a known function and each $\bar{Q}_a$ is an observed average, $\bar{Q}_a = \sum_{i=1}^n Q_a(X_i)/n$. For example, the correlation $\hat{\rho}$ is a function of $A = 5$ such averages,

$$\hat{\rho}(X_1, \ldots, X_n) = \frac{\bar{Q}_4 - \bar{Q}_1\bar{Q}_2}{[\bar{Q}_3 - \bar{Q}_1^2]^{1/2}[\bar{Q}_5 - \bar{Q}_2^2]^{1/2}}, \tag{3.10}$$

$Q_1(X) = Q_1(Y, Z) = Y$, $Q_2 = Z$, $Q_3 = Y^2$, $Q_4 = YZ$, $Q_5 = Z^2$. The method works by (i) expanding t in a linear Taylor series about the expectations of the $\bar{Q}_a$; (ii) evaluating the standard error of the Taylor series, using the usual expressions for variances of averages; and (iii) substituting $\gamma(\hat{F})$ for any unknown quantity $\gamma(F)$ occurring in (ii). For example, the delta-method estimate of the standard error of $\hat{\rho}$

is

$$\left\{ \frac{\hat{\rho}^2}{4n}\left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} - \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}} \right]\right\}^{1/2}, \tag{3.11}$$

where if $x_i = (y_i, z_i)$, then $\hat{\mu}_{gh} = \sum(y_i - \bar{y})^g(z_i - \bar{z})^h/n$. See Cramér (1945, p. 359).

Efron (1980b) proves that for any statistic of the form $\hat{\theta} = t(\bar{Q}_1, \ldots, \bar{Q}_A)$, the nonparametric delta method and the infinitesimal jackknife give the same estimate of standard error. The infinitesimal jackknife, the delta method, and the empirical influence-function approach are three names for the same method, and all are Taylor-series approximations to the bootstrap, as in (3.8). Notice that the results reported in line 7 of Table 2 show a serious downward bias. Efron and Stein (1981) prove that the ordinary jackknife estimate of standard error will tend to be biased upward, in a sense made precise in that paper. In the author's opinion, the ordinary jackknife is the method of choice for estimating standard errors if one doesn't want to do extensive bootstrap computations.

4. THE PERCENTILE METHOD

Estimated standard errors, along with assumptions of approximate normality, provide the rough confidence intervals so widely used in applied statistics: $\hat{\theta} \pm z_\alpha\hat{\sigma}$, with z_α being the α-point of a standard normal distribution, e.g., $z_{0.05} = -1.645$. In small-sample parametric situations, where we can do exact calculations, confidence intervals are often highly asymmetric about the best point estimate $\hat{\theta}$. The asymmetry, which is $O(1/\sqrt{n})$ in magnitude, is substantially more important than the Student t-correction (replacing $\hat{\theta} \pm z_\alpha\hat{\sigma}$ by $\hat{\theta} \pm t_\alpha\hat{\sigma}$, with t_α the α-point of an appropriate t-distribution), which is only $O(1/n)$. We will discuss some nonparametric methods of assigning confidence intervals which attempt to capture the correct degree of asymmetry. This is a brave venture, since the problem isn't well understood even in parametric settings; but the results are mildly encouraging.

We begin with an example described more fully in Efron (1979b). The data consist of $n = 15$ pairs of points: (576, 3.39), (635, 3.30), (558, 2.81), (578, 3.03), (666, 3.44), (580, 3.07), (555, 3.00), (661, 3.43), (661, 3.36), (605, 3.13), (653, 3.12), (575, 2.74), (545, 2.76), (572, 2.88), (594, 2.96). Each pair is two statistics describing the 1973 entering class at an American law school. The observed correlation between the two statistics is $\hat{\rho} = 0.776$. Figure 1 shows the histogram of $B = 1000$ bootstrap replications $\hat{\rho}^*$, with the abscissa plotted in terms of $\hat{\rho}^* - \hat{\rho}$. Also shown is the normal-theory density curve $f_{\hat{\rho}}(\hat{\rho}^*)$ for $\hat{\rho} = 0.776$, plotted versus $\hat{\rho}^* - \hat{\rho}$ (Johnson and Kotz 1970, p. 222).

The similarity between the histogram and $f_{\hat{\rho}}(\hat{\rho}^*)$ is striking. It suggests, in a manner motivated in Sections 5–7, a simple method for constructing nonparametric confidence intervals. Let $\widehat{\mathrm{CDF}}(t)$ represent the cumulative of the bootstrap distribution for some real-valued functional statistic $\hat{\theta} = \theta(\hat{F})$:

$$\widehat{\mathrm{CDF}}(t) = \mathrm{Prob}_*\{\hat{\theta}^* < t\} = \frac{\#\{\hat{\theta}^*(b) < t\}}{B}. \tag{4.1}$$

"Prob$_*$" indicates the bootstrap probability, as induced by the mechanism (2.5); the last expression equals this probability as $B \to \infty$, a distinction we shall ignore. For a given value of α between 0 and 1 define

$$\hat{\theta}(\alpha) = \widehat{\mathrm{CDF}}^{-1}(\alpha). \tag{4.2}$$

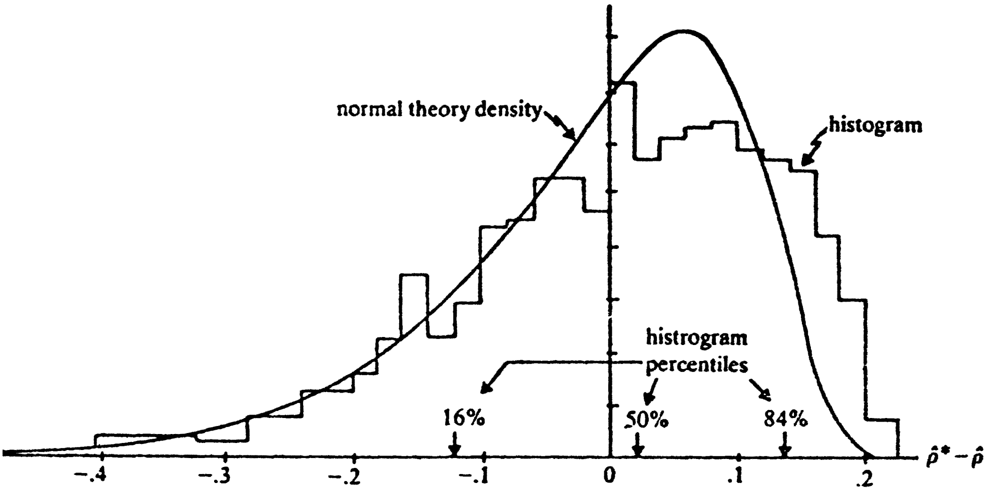

FIGURE 1: Histogram of 1000 bootstrap replications $\hat\rho^* - \hat\rho$ for the law-school data. The smooth curve is the normal-theory density of $\hat\rho^*$, centered at $\hat\rho$, when the true correlation is $\hat\rho = 0.776$.

The *percentile method* assigns

$$\theta \in [\hat\theta(\alpha), \hat\theta(1-\alpha)] \tag{4.3}$$

as a putative $1 - 2\alpha$ central confidence interval for the parameter $\theta = \theta(F)$. With $\alpha = 0.16$, the 1000 bootstrap replications for the law-school data gave $\rho \in [0.654, 0.908] = [\hat\rho - 0.12, \hat\rho + 0.13]$ as a central 68% interval for ρ, compared to the standard-normal-theory interval $[\hat\rho - 0.16, \hat\rho + 0.09]$, obtained by inverting

$$\hat\phi \sim \mathbf{N}\left(\phi + \frac{\rho}{2(n-1)}, \frac{1}{n-3}\right).$$

We shall also consider the *bias-corrected percentile method*

$$\theta \in [\widehat{\mathrm{CDF}}^{-1}(\Phi(2z_0 + z_\alpha)), \widehat{\mathrm{CDF}}^{-1}(\Phi(2z_0 + z_{1-\alpha}))], \tag{4.4}$$

where $z_0 = \Phi^{-1}\,\widehat{\mathrm{CDF}}(\hat\theta)$ and $\Phi(z) = (1/\sqrt{2\pi}) \int_{-\infty}^{z} e^{-s^2/2}\, ds$. For the law-school data, $\widehat{\mathrm{CDF}}(\hat\theta) = 0.433$ (433 out of 1000 bootstrap replications $\hat\theta^*$ less than 0.776), so $z_0 = \Phi(0.433) = -0.17$. The bias-corrected central 68% interval was

$$\rho \in [\widehat{\mathrm{CDF}}^{-1}\,\Phi(-1.34), \widehat{\mathrm{CDF}}^{-1}\,\Phi(0.66)] = [\hat\rho - 0.17, \hat\rho + 0.10],$$

almost the same as the normal-theory interval.

Table 3 shows the results of 10 Monte Carlo trials with $X_1, X_2, \ldots, X_{15}$ bivariate normal, true correlation $\rho = 0.5$. For each trial, central 68% intervals were constructed using normal theory, the percentile method, the bias-corrected percentile method, and a smoothed version of this last approach. The smoothed version is based on (4.4), but with $\widehat{\mathrm{CDF}}(t)$ being the cumulative of smoothed bootstrap replications $\hat\theta^*$, the smoothing mechanism being the same one used in line 3 of Table 2. The resemblance between the normal intervals and the smoothed bias-corrected intervals is impressive, though again one must be suspicious of a smoothing mechanism which obviously relates to the true distribution. The unsmoothed bias-corrected intervals are more variable, but still give an excellent suggestion of the correct left-right asymmetry in the normal-theory intervals. The percentile method doesn't do this

TABLE 3: Central 68% confidence intervals, 10 trials of $X_1, X_2, \ldots, X_{15}$ bivariate normal, true $\rho = 0.5$. Each interval has $\hat\rho$ subtracted from both endpoints.

Trial	$\hat\rho$	Normal theory	Percentile method	Bias-corrected percentile method	Smoothed and bias-corrected percentile method
1	.16	(−.29, .26)	(−.29, .24)	(−.28, .25)	(−.28, .24)
2	.75	(−.17, .09)	(−.05, .08)	(−.13, .04)	(−.12, .08)
3	.55	(−.25, .16)	(−.24, .16)	(−.34, .12)	(−.27, .15)
4	.53	(−.26, .17)	(−.16, .16)	(−.19, .13)	(−.21, .16)
5	.73	(−.18, .10)	(−.12, .14)	(−.16, .10)	(−.20, .10)
6	.50	(−.26, .18)	(−.18, .18)	(−.22, .15)	(−.26, .14)
7	.70	(−.20, .11)	(−.17, .12)	(−.21, .10)	(−.18, .11)
8	.30	(−.29, .23)	(−.29, .25)	(−.33, .24)	(−.29, .25)
9	.33	(−.29, .22)	(−.36, .24)	(−.30, .27)	(−.30, .26)
10	.22	(−.29, .24)	(−.50, .34)	(−.48, .36)	(−.38, .34)
Ave	.48	(−.25, .18)	(−.21, .19)	(−.26, .18)	(−.25, .18)

consistently. On the other hand, 100 trials of the same experiment showed the percentile method giving correct average coverage rates: in 13 out of the 100 trials $\rho = 0.5$ was less than $\widehat{\text{CDF}}^{-1}(0.10)$, compared to the theoretical expected value 10 out of 100; likewise 16 trials had $\widehat{\text{CDF}}^{-1}(0.10) \le \rho < \widehat{\text{CDF}}^{-1}(0.25)$, 22 had $\widehat{\text{CDF}}^{-1}(0.25) \le \rho < \widehat{\text{CDF}}^{-1}(0.50)$, 27 had $\widehat{\text{CDF}}^{-1}(0.50) \le \rho < \widehat{\text{CDF}}^{-1}(0.75)$, 12 had $\widehat{\text{CDF}}^{-1}(0.75) \le \rho < \widehat{\text{CDF}}^{-1}(0.90)$, and 10 had $\widehat{\text{CDF}}^{-1}(0.90) \le \rho$.

5. BAYESIAN JUSTIFICATION FOR THE PERCENTILE METHOD

Suppose the sample space $\mathscr{X}$ is discrete, say $\mathscr{X} = \{1, 2, \ldots, L\}$. This is no real restriction, since we can discretely approximate most situations if L is taken sufficiently large. Let $f_l = \text{Prob}_F\{X = l\}$ and $\hat f_l = \#\{x_i = l\}/n$ be the true and observed frequencies for category l, and denote $\mathbf{f} = (f_1, f_2, \ldots, f_L)$, $\hat{\mathbf{f}} = (\hat f_1, \hat f_2, \ldots, \hat f_L)$.

Consider the symmetric Dirichlet prior distribution with parameter a,

$$\mathbf{f} \sim \text{Di}_L(a\mathbf{1}), \tag{5.1}$$

i.e., take the prior density function of $\mathbf{f}$ proportional to $\prod_l f_l^{a-1}$. Having observed $\hat{\mathbf{f}}$, the *a posteriori* density $\mathbf{f} \mid \hat{\mathbf{f}}$ is proportional to $\prod f_l^{n\hat f_l + a - 1}$. Letting $a \to 0$ to represent prior ignorance gives the well-known result

$$\mathbf{f} \mid \hat{\mathbf{f}} \sim \text{Di}_L(n\hat{\mathbf{f}}). \tag{5.2}$$

The result (5.2) is quite similar to the bootstrap distribution when $\mathscr{X}$ is discrete,

$$\hat{\mathbf{f}}^* \mid \hat{\mathbf{f}} \sim \frac{\text{Mult}_L(n, \hat{\mathbf{f}})}{n}, \tag{5.3}$$

where $\hat f_l^* = \#\{X_i^* = l\}/n$: (i) both distributions are supported entirely on those categories in which data were observed, i.e., those l for which $\hat f_l > 0$; (ii) both distributions have expectation vector $\hat{\mathbf{f}}$; (iii) the covariance matrices are also nearly equal, $\mathscr{C}ov(\mathbf{f} \mid \hat{\mathbf{f}}) = \Sigma_{\hat f}/(n+1)$, $\mathscr{C}ov_*(\hat{\mathbf{f}}^* \mid \hat{\mathbf{f}}) = \Sigma_{\hat f}/n$, where $\Sigma_{\hat f}$ has diagonal elements $\hat f_l(1 - \hat f_l)$, and off-diagonals $-\hat f_l \hat f_m$. A continuity correction for the discreteness of the bootstrap distribution makes the covariance matrices agree even more closely.

The point here is that the *a posteriori* distribution of $\theta(\mathbf{f}) \mid \hat{\mathbf{f}}$ is likely to be well approximated by the bootstrap distribution of $\theta(\hat{\mathbf{f}}^*) \mid \hat{\mathbf{f}}$, whenever $\theta(\mathbf{f})$ is a reasonably smooth function of f. If this is true, the percentile-method $1 - 2\alpha$ central confidence interval will be a good approximation to the central Bayes interval of probability $1 - 2\alpha$.

Rubin (1981) has criticized the bootstrap for agreeing with a silly prior, a form of guilt by association, and suggests doing a genuine Bayesian analysis for these problems. On the other hand, we shall see in Section 7 that the uninformative Dirichlet approach gives a quite reasonable answer in the case where $\theta(F)$ is the median.

6. TRANSFORMATIONS AND PIVOTAL QUANTITIES

The argument supporting the bias-corrected percentile method (4.4) is based on hypothesizing a transformation to a *normal pivotal quantity*. Suppose there exists a monotonic increasing function $g(\cdot)$ such that the transformed quantities

$$\phi = g(\theta), \qquad \hat\phi = g(\hat\theta), \qquad \hat\phi^* = g(\hat\theta^*) \tag{6.1}$$

satisfy

$$\hat\phi - \phi \sim \mathbf{N}(-z_0\sigma, \sigma^2) \quad \text{and} \quad \hat\phi^* - \hat\phi \underset{*}{\sim} \mathbf{N}(-z_0\sigma, \sigma^2) \tag{6.2}$$

for some constants z_0 and σ. The symbol $\underset{*}{\sim}$ indicates distribution under the bootstrap sampling (2.5). In other words, $\hat\phi - \phi$ is a normal pivotal having the same distribution under both F and $\hat F$.

We shall see that (6.2) leads easily to (4.4), and also (4.3). The interesting aspect of this argument is that the bias-corrected percentile method requires no knowledge of the transformation $g(\cdot)$ or the constants z_0, σ. All we need to know to construct the appropriate interval for $\theta = \theta(F)$ is $\hat\theta = \theta(\hat F)$ and $\widehat{\text{CDF}}(t)$, the cumulative distribution of the statistic of interest assuming $F = \hat F$. This is particularly useful in nonparametric contexts, where it is difficult to imagine interesting alternative distributions to the MLE $\hat F$. (An attempt at imagining such alternatives is made in Section 11.)

In parametric contexts, (6.2) is a device frequently used to obtain confidence intervals. Fisher's transformation $\phi = \tanh^{-1} \rho$ is a classic example. Within the class of bivariate normal distributions it produces a good approximation to (6.2), with $\sigma^2 = 1/(n - 3)$ and $z_0 = -\rho\sqrt{n-3}/2(n - 1)$. It is interesting to apply (4.4) directly to this parametric situation. To do so it is only necessary to redefine $\widehat{\text{CDF}}(t) = \int_{-1}^{t} f_{\hat\rho}(\hat\rho^*)\, d\hat\rho^*$, where $f_{\hat\rho}(\hat\rho^*)$ is the normal-theory density of the correlation coefficient when the true correlation equals $\hat\rho$. Then $\widehat{\text{CDF}}(t)$ is still the bootstrap CDF of $\hat\rho^*$, but working off the parametric maximum-likelihood estimate $F = \hat F_N$. For the case $n = 15$, $\hat\rho = 0.5$, the 95% central interval obtained from (4.4) is $\rho \in [-0.039, 0.798]$. Again, it should be emphasized that this calculation requires no knowledge of the $\tanh^{-1}$ transformation or of z_0 and σ^2, only the density $f_{\hat\rho}(\cdot)$. The exact 95% interval obtained from the *Biometrika Tables* (1954) is $\rho \in [-0.024, 0.790]$.

Now to verify (4.4) from (6.2). Notice that the middle statement in (6.1) is actually a definition of the estimator $\hat\phi$. The last relationship in (6.1) then follows from $\hat\phi^* = \hat\phi(X_1^*, X_2^*, \ldots, X_n^*) = g(\hat\theta(X_1^*, X_2^*, \ldots, X_n^*)) = g(\hat\theta^*)$. It implies that the bootstrap $\widehat{\text{CDF}}$ of $\hat\phi^*$, say

$$\widehat{\text{CDG}}(s) = \text{Prob}_* \{\hat\phi^* \leq s\}, \tag{6.3}$$

is the obvious transformation of $\widehat{\text{CDF}}(t)$,

$$\widehat{\text{CDG}}(g(t)) = \widehat{\text{CDF}}(t). \tag{6.4}$$

The first relationship in (6.2) gives

$$\phi \in [\hat{\phi} + z_0\sigma \pm z_\alpha\sigma] \tag{6.5}$$

as the $1 - 2\alpha$ central confidence interval for ϕ. Transforming (6.5) back to the θ-scale, by the transformation $g^{-1}(\cdot)$, turns out to give (4.4). First notice that (6.2) and (6.4) imply

$$\text{Prob}_*\{\hat{\phi}^* \leq \hat{\phi}\} = \Phi(z_0) = \widehat{\text{CDG}}(g(\hat{\theta})) = \widehat{\text{CDF}}(\hat{\theta}), \tag{6.6}$$

which gives $z_0 = \Phi^{-1}\,\widehat{\text{CDG}}(\hat{\theta})$ as in (4.4).

Using (6.2) again,

$$\text{Prob}_*\{\hat{\phi}^* < \hat{\phi} + z_0\sigma \pm z_\alpha\sigma\} = \text{Prob}_F\{\hat{\phi} < \phi + z_0\sigma \pm z_\alpha\sigma\} = \Phi\{2z_0 \pm z_\alpha\}.$$

This can be written as $\widehat{\text{CDG}}(\hat{\phi} + z_0\sigma \pm z_\alpha\sigma) = \Phi(2z_0 \pm z_\alpha)$, or

$$\hat{\phi} + z_0\sigma \pm z_\alpha\sigma = \widehat{\text{CDG}}^{-1}[\Phi(2z_0 \pm z_\alpha)].$$

Transforming (6.5) back to the θ-scale by the mapping $g^{-1}(\cdot)$ gives the interval with endpoints $g^{-1}(\hat{\phi} + z_0\sigma \pm z_\alpha\sigma) = g^{-1}\,\widehat{\text{CDG}}^{-1}[\Phi(2z_0 \pm z_\alpha)] = \widehat{\text{CDF}}^{-1}\Phi(2z_0 \pm z_\alpha)$, the last equality following from (6.4): $\widehat{\text{CDF}}^{-1} = [\widehat{\text{CDG}}\,g]^{-1} = g^{-1}\,\widehat{\text{CDG}}^{-1}$. We now have derived (4.4), the bias-corrected percentile interval, from (6.2).

The normal distribution plays no special role in this argument. We could assume that the pivotal quantity has some other symmetric distribution than the normal, in which case "Φ" would have a different meaning in (4.4). In the unbiased case, $z_0 = 0$, the normal distribution plays no role at all, since we get the uncorrected percentile interval (4.3). This is worth stating separately: *if we assume there exists a monotonic mapping $g(\cdot)$ such that $\hat{\phi} - \phi$ and $\hat{\phi}^* - \hat{\phi}$ have the same distribution, symmetric about the origin, then the percentile interval* (4.3) *has the correct coverage probability.*

7. THE PERCENTILE METHOD FOR THE MEDIAN

We now consider a parameter, almost the only one, for which exact nonparametric confidence intervals exist, namely, the median $\mu(F)$ of a continuous distribution F on the real line, $\mu = \inf_t \{\text{Prob}_F\{X \leq t\} = 0.5\}$. Define

$$b_{k,n}(p) = \binom{n}{k} p^k (1 - p)^{n-k}. \tag{7.1}$$

The random variable $Z = \#\{X_i < \mu\}$ is a pivotal for μ, always having the binomial distribution $Z \sim B_i(n, \frac{1}{2})$. Given the observed order statistics $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ of an independently and identically distributed sample from F, we can make the confidence statement

$$\text{Prob}_F\{\mu \in [x_{(k_1)}, x_{(k_2)}]\} = \sum_{k=k_1}^{k_2-1} b_{k,n}(0.5), \tag{7.2}$$

since $\{x_{(k_1)} \leq \mu \leq x_{(k_2)}\}$ has the same probability as $\{x_{(k_1)} < \mu \leq x_{(k_2)}\}$, which is equivalent to $\{k_1 \leq Z < k_2\}$.

As an example, take $n = 13$, $k_1 = 4$, $k_2 = 10$. Then a binomial table shows that

$$\mu \in [x_{(4)}, x_{(10)}] \tag{7.3}$$

is a central 0.908 confidence interval for μ.

It turns out, happily enough, that the percentile method agrees closely with the binomial intervals (7.2) in the case of the median. First of all we notice that the bootstrap distribution of the sample median can be calculated exactly without resorting to Monte Carlo. Assuming odd sample size, say $n = 2m - 1$, then $\hat{\mu} = x_{(m)}$, the middle order statistic, and

$$p_{(k)} \equiv \text{Prob}_* \{\hat{\mu}^* = x_{(k)}\} = \sum_{j=0}^{m-1} \left\{ b_{j,n}\left(\frac{k-1}{n}\right) - b_{j,n}\left(\frac{k}{n}\right) \right\} \tag{7.4}$$

(Efron 1979, p. 6). For $n = 13$ the bootstrap distribution is as follows:

$p_{(k)}$	.0000	.0015	.0142	.0550	.1242	.1936	.2230
k	1	2	3	4	5	6	7

(7.5)

with $p_{(7+i)} = p_{(7-i)}$ for $i = 1, 2, \ldots, 6$, by symmetry.

In the example just given, $[x_{(4)}, x_{(10)}]$ is a central $1 - 2\alpha$ percentile interval for μ, with $\alpha = (0.0000 + 0.0015 + 0.0142 + 0.0550/2) = 0.0432$. Here we have split the bootstrap probability at the endpoint $\hat{\mu}^* = x_{(4)}$, for reasons having to do with the "grainy" nature of the simple median: The bootstrap median $\hat{\mu}^*$ takes on only n possible values, compared to $\binom{2n-1}{n}$ possible bootstrap values for $\hat{\theta}^*$ if $\hat{\theta}$ is smoothly defined, as is the correlation coefficient. Splitting the endpoint probabilities can be justified by approximating $\hat{\mu}$ with a series of less grainy statistics. With the endpoint probabilities split, the percentile method gives

$$\mu \in [x_{(4)}, x_{(10)}] \tag{7.6}$$

as a central 0.914 confidence interval, agreeing remarkably well with (7.3). Numerical investigation confirms that the agreement is always excellent as long as $\alpha \geq 0.01$. Some theoretical reasons are given in Section 10.5 of Efron (1980c). The bootstrap distribution (7.4) is median unbiased, $\text{Prob}_* \{\hat{\mu}^* \leq \hat{\mu}\} = 0.50$ splitting the probability $\text{Prob}_* \{\hat{\mu}^* = \hat{\mu}\}$, so the bias correction (4.4) has no effect.

The Bayesian statement (5.2) gives

$$\text{Prob}\{\mu = x_{(k)} \mid \hat{\mathbf{f}}\} = b_{k-1,n-1}(0.5), \tag{7.7}$$

assuming that $\mathscr{X} = \mathscr{R}^1$ has been partitioned so finely that $\hat{f}_l$ equals 1 or 0 for every category l. [Equation (7.7) makes use of a well-known relationship between the beta and binomial distributions.] Realistically we would never believe that the *a posteriori* distribution for μ concentrates exclusively on the observed values x_i. Interpreting "$\mu = x_{(k)}$" in (7.7) as "$\mu = x_{(k)} + \varepsilon$", where ε has any distribution symmetric about 0 [for example, $\varepsilon \sim \mathbf{N}(0, 0.01)$], gives

$$\begin{aligned} \text{Prob}\{\mu \in [x_{(k_1)}, x_{(k_2)}] \mid \hat{\mathbf{f}}\} &= \tfrac{1}{2} b_{k_1-1,n-1}(0.5) + \sum_{k=k_1}^{k_2-2} b_{k,n-1}(0.5) + \tfrac{1}{2} b_{k_2-1,n-1}(0.5) \\ &= \sum_{k=k_1}^{k_2-1} b_{k,n}(0.5). \end{aligned} \tag{7.8}$$

Splitting the endpoint probabilities makes the Bayesian coverage probability for $[x_{(k_1)}, x_{(k_2)}]$ agree exactly with that for the classical confidence level (7.2). From Section 5 we then expect the percentile method to agree well with (7.2), which is indeed the case.

8. CONFIDENCE INTERVALS FOR THE MEAN

The rest of this paper concerns setting nonparametric confidence intervals for the mean $\mu = \mathscr{E}_F X$ of a distribution F on the real line, on the basis of observing X_1, X_2, $\dots$, $X_n \overset{\text{iid}}{\sim} F$. In one sense this problem is impossible, since modifying F with a tiny probability of X being enormous, say $\text{Prob}_F\{X = 10^{100}\} = 10^{-10}$, can totally change μ without ever showing up in most samples of size $n < 10^5$. On the other hand, the problem is "solved" every day using the standard Student t-intervals, $\bar{x} \pm t_{\alpha,n-1}\hat{\sigma}$ with $\hat{\sigma}^2 = \sum(x_i - \bar{x})^2/n(n-1)$. Genuine parametric intervals for μ are often strikingly nonsymmetric about $\hat{\mu} = \bar{x}$. We shall consider several nonparametric methods which attempt to capture this asymmetry.

The discussion of general methods will be illustrated on the specific problem of Table 4. Ten independently and identically distributed samples $X_1, X_2, \dots, X_{15}$ were obtained from the negative-exponential distribution centered at 0, $\text{Prob}\{X > x\} = e^{-(x+1)}$, $x > -1$. Each sample $x_1, x_2, \dots, x_{15}$ was standardized: translated to have $\hat{\mu} = \bar{x} = 0$ and scaled to have $\sum(x_i - \bar{x})^2/14 = 1$. This stabilized the entries of Table 4 without affecting comparisons between the various methods of setting confidence intervals, all of which scale and translate in the obvious way.

Four different methods are illustrated in Table 4: (1) the percentile method based on the bootstrap distribution of $\hat{\mu}^* = \bar{X}^*$, $B = 1000$; (2) random subsampling, $B = 1000$, explained in the next paragraph; (3) the bias-corrected percentile method; and (4) the Pitman intervals. The latter assume that we are sampling from a translated and rescaled negative exponential, say $\mu + \sigma X$, X as above, and are the confidence intervals based on the *conditional* distribution of the usual t-statistic, *given* the standardized version of the sample. They are also the Bayes posterior intervals versus the uninformative prior $d\mu\ d\sigma/\sigma$. The Pitman intervals would usually be considered the correct parametric solution, and we shall use them here as a standard of comparison. [This is an arguable point. The Pitman intervals based on the translation model $\mu + \sigma X$, *σ known*, are completely different than (4). The "actual T" intervals of Section 9 are another reasonable standard of comparison.] Notice that they extend much further to the right of $\hat{\mu}$ (=0 in the standardized samples) than to the left.

Random subsampling is based on Hartigan's typical-value theory (1969, 1971,

TABLE 4: Nonparametric and parametric confidence intervals for the expectation, negative-exponential distribution; 10 standardized samples, $n = 15$. Confidence limits are listed in the order 5%, 10%, 90%, 95%, so the outer [inner] two numbers are an approximate 90% [80%] interval.

Trial	Percentile method ($B = 1000$)	Random subsampling ($B = 1000$)	Bias corrected percentile method ($B = 1000$)	Pitman intervals
1	−.38, −.31, .34, .41	−.44, −.34, .33, .43	−.34, −.27, .38, .48	−.31, −.25, .45, .60
2	−.39, −.34, .34, .45	−.47, −.36, .37, .46	−.36, −.27, .38, .54	−.34, −.27, .48, .64
3	−.44, −.35, .30, .40	−.42, −.36, .36, .46	−.42, −.32, .32, .41	−.42, −.34, .56, .66
4	−.38, −.32, .33, .45	−.44, −.35, .36, .47	−.38, −.32, .33, .45	−.30, −.24, .44, .58
5	−.37, −.32, .34, .44	−.42, −.36, .33, .46	−.35, −.28, .39, .49	−.25, −.20, .37, .50
6	−.37, −.31, .34, .44	−.47, −.36, .36, .48	−.34, −.27, .39, .50	−.41, −.33, .55, .65
7	−.42, −.34, .31, .39	−.45, −.36, .35, .46	−.38, −.29, .34, .46	−.40, −.32, .54, .65
8	−.35, −.30, .35, .46	−.42, −.35, .36, .48	−.32, −.27, .40, .50	−.32, −.26, .46, .62
9	−.40, −.32, .33, .43	−.48, −.37, .34, .42	−.38, −.30, .34, .45	−.33, −.27, .47, .62
10	−.38, −.31, .32, .41	−.42, −.36, .32, .43	−.37, −.30, .33, .42	−.32, −.26, .46, .61
Average	−.39, −.32, .33, .43	−.44, −.36, .35, .46	−.36, −.29, .36, .47	−.34, −.27, .48, .61

1975). A set S is chosen randomly from the $2^n - 1$ nonempty subsets of $\{1, 2, \ldots, n\}$, $n = 15$ here, and the ***subsample mean*** $\bar{x}_S = \sum_{i \in S} x_i / \sum_{i \in S} 1$ calculated. For column (2) of Table 4 this process was independently repeated $B = 1000$ times for each of the standardized samples. The percentile method (4.2), (4.3) provides the confidence limits, but with $\widehat{\text{CDF}}(t)$ now being the cumulative distribution of the 1000 subsample means. The typical-value theorem (Hartigan 1969) says that these confidence intervals will have exactly the claimed coverage probabilities if F is continuous and symmetrically distributed about μ.

Random subsampling, like the bootstrap, is a resampling plan, as described at the beginning of Section 3. Instead of the multinomial distribution (3.2), the resampling vector has components $P_i = I_i / \sum_{j=1}^{n} I_j$, where the I_j independently equal 0 or 1 with probability $\frac{1}{2}$, and I_i indicates whether or not x_i is included in the random subsample. An easy calculation shows that, given $x_1, x_2, \ldots, x_n$, a random subsample mean has expectation and variance

$$\bar{X}_S \sim \left(\bar{x}, \frac{n+2}{n-1} \frac{\sum (x_i - \bar{x})^2}{n^2} \left[1 + o\left(\frac{1}{n}\right)\right]\right), \tag{8.1}$$

compared to the bootstrap expectation and variance

$$\bar{X}^* \sim \left(\bar{x}, \frac{\sum (x_i - \bar{x})^2}{n^2}\right). \tag{8.2}$$

Comparing the averages of columns (1) and (2) of Table 4, we see that the random subsample intervals are just about $\sqrt{(n+2)/(n-1)} = 1.10$ times as wide as the percentile intervals, as suggested by (8.1), (8.2). For $1 - 2\alpha = 0.90$, the ratio of widths is $(0.46 + 0.44)/(0.43 + 0.39) = 1.10$. Neither of the methods shows the right-left asymmetry of the Pitman intervals. The bias correction of column (3) is helpful in this regard, shifting all 10 percentile intervals rightward, though not far enough so. A similar bias correction was tried on the subsample intervals, but had little effect, often moving the intervals slightly leftwards.

9. BOOTSTRAP T

Let $Z_1, Z_2, \ldots, Z_n$ be independently and identically distributed from a known continuous distribution F on the real line, and suppose we observe $X_1, X_2, \ldots, X_n$ where $X_i = \mu + \sigma Z_i$, μ and σ unknown. A confidence interval for μ can be based on the pivotal quantity

$$T = \frac{\hat{\mu} - \mu}{\hat{\sigma}} \qquad (\hat{\mu} = \bar{X}, \quad \hat{\sigma} = [\sum (X_i - \bar{X})^2 / n(n-1)]^{1/2}). \tag{9.1}$$

For example, if $n = 15$ and F is negative exponential centered at 0, then T has (5%, 10%, 90%, 95%) percentile points $(-2.67, -1.94, 1.08, 1.39)$, compared to the t_{14} percentiles $(-1.76, -1.34, 1.34, 1.76)$ appropriate for $n = 15$, $F \sim \mathbf{N}(0, 1)$. [Quadruples will always refer to these four percentiles.]

If t_α, $t_{1-\alpha}$ are defined by $\text{Prob}\{T < t_\alpha\} = \alpha = \text{Prob}\{T > t_{1-\alpha}\}$, then

$$\mu \in [\hat{\mu} - t_{1-\alpha}\hat{\sigma}, \hat{\mu} - t_\alpha \hat{\sigma}] \tag{9.2}$$

is a $1 - 2\alpha$ central confidence interval for μ. In the negative-exponential case, $[\hat{\mu} - 1.39\,\hat{\sigma}, \hat{\mu} + 2.67\,\hat{\sigma}]$ is a central 90% interval. It extends nearly twice as far to the right as to the left of $\hat{\mu}$.

We can apply (9.2) just as well if $\hat{\mu}$ is any *translation statistic*, $\hat{\mu}(a\mathbf{1} + b\mathbf{X}) = a + b\hat{\mu}(\mathbf{X})$, and $\hat{\sigma}$ is any *scale statistic* $\hat{\sigma}(a\mathbf{1} + b(\mathbf{X}) = b\hat{\sigma}(\mathbf{X}))$. Then $T = (\hat{\mu} - \mu)/\hat{\sigma}$ is still pivotal, though of course its distribution, and the values of t_α, $t_{1-\alpha}$, depend upon the specific forms of $\hat{\mu}$, $\hat{\sigma}$. It is reasonably straightforward to prove the following theorem: (9.2) *is the central a posteriori interval, probability* $1 - 2\alpha$, *of* μ *given* $(\hat{\mu}, \hat{\sigma})$, *versus the uninformative prior* $d\mu d\sigma/\sigma$. In other words, (9.2) is the appropriate Bayes interval for μ based on a reduced amount of information, the values of $\hat{\mu}$, $\hat{\sigma}$ rather than the entire sample $x_1, x_2, \ldots, x_n$.

In a nonparametric setting we don't know the distribution of T, but we can use the bootstrap to estimate it. The entries in column 1 of Table 5 were obtained in this way. For each sample, $B = 1000$ bootstrap values of T,

$$T^* = \frac{\hat{\mu}^* - \hat{\mu}}{\hat{\sigma}^*} \qquad (\hat{\mu}^* = \bar{X}^*, \quad \hat{\sigma}^* = [\textstyle\sum (X_i^* - \bar{X}^*)^2/n(n-1)]^{1/2}) \tag{9.3}$$

were generated. The $1 - 2\alpha$ central confidence interval (9.2) was estimated by $[\hat{\mu} - \hat{t}_{1-\alpha}\hat{\sigma}, \hat{\mu} - \hat{t}_\alpha\hat{\sigma}]$ with $\hat{t}_\alpha$, $\hat{t}_{1-\alpha}$ defined by $\text{Prob}_*\{T^* < \hat{t}_\alpha\} = \alpha = \text{Prob}_*\{T^* > \hat{t}_{1-\alpha}\}$. Since $\hat{\mu} = 0$ and $\hat{\sigma} = 1/\sqrt{15}$ for each trial in Table 5 (because the samples were standardized), the interval is simply $[-\hat{t}_{1-\alpha}/\sqrt{15}, -\hat{t}_\alpha/\sqrt{15}]$. Notice how closely the average endpoints for the 10 trials approximate the actual T-values, $[-t_{1-\alpha}/\sqrt{15}, -t_\alpha/\sqrt{15}\}$. The bootstrap t-intervals are highly asymmetric about $\hat{\mu} = 0$.

Fraser (1976) has suggested nonparametrically estimating the actual Bayes intervals, those given $x_1, x_2, \ldots, x_n$ rather than those given only $\hat{\mu}$, $\hat{\sigma}$. The actual Bayes intervals, versus $d\mu\, d\sigma/\sigma$, are the Pitman intervals of column (4), Table 4. They are of the form $[\hat{\mu} - t_{1-\alpha}^{\mathbf{x}}\hat{\sigma}, \hat{\mu} - t_\alpha^{\mathbf{x}}\hat{\sigma}]$, where $t_\alpha^{\mathbf{x}}$ is defined by

$$\alpha = \int_{\hat{\mu}-t_\alpha^{\mathbf{x}}\hat{\sigma}}^{\infty} \int_0^\infty \frac{1}{\sigma} L_{\mathbf{x}}(\mu, \sigma)\, d\alpha\, d\mu \bigg/ \int_{-\infty}^{\infty} \int_0^\infty \frac{1}{\sigma} L_{\mathbf{x}}(\mu, \sigma)\, d\sigma\, d\mu,$$
$$L_{\mathbf{x}}(\mu, \sigma) = \prod_{i=1}^{n} f_{\mu,\sigma}(x_i), \tag{9.4}$$

and similarly for $t_{1-\alpha}^{\mathbf{x}}$, with $1 - \alpha$ replacing α in (9.4). Here $f_{\mu,\sigma}(x_i)$ is the density function.

TABLE 5: Three more nonparametric confidence-interval methods applied to the 10 standardized negative-exponential samples of Table 4. The averages of the bootstrap t endpoints almost equal the actual T distribution limits for the negative exponential, $n = 15$. Methods (2) and (3) are explained in Sections 10 and 11.

Trial	(1) Bootstrap t ($B = 1000$)	(2) Johnson's t	(3) Exponential tilting	Sample skewness
1	−.36, −.29, .53, .71	−.37, −.29, .46, .70	−.35, −.27, .42, .56	1.40
2	−.37, −.28, .52, .65	−.38, −.29, .46, .66	−.35, −.26, .41, .53	1.30
3	−.42, −.31, .39, .51	−.44, −.34, .36, .47	−.38, −.29, .33, .43	0.15
4	−.37, −.28, .51, .70	−.37, −.29, .46, .70	−.31, −.26, .42, .53	1.40
5	−.34, −.27, .62, .84	−.36, −.28, .57, *	−.31, −.26, .43, .57	1.86
6	−.39, −.30, .45, .59	−.39, −.30, .42, .59	−.34, −.30, .39, .50	1.04
7	−.39, −.29, .41, .61	−.41, −.32, .38, .52	−.38, −.28, .36, .47	0.62
8	−.33, −.27, .63, .77	−.35, −.28, .64, *	−.33, −.24, .45, .62	1.98
9	−.37, −.30, .45, .60	−.39, −.30, .42, .58	−.34, −.26, .38, .48	1.02
10	−.34, −.27, .56, .79	−.38, −.29, .45, .66	−.32, −.27, .39, .52	1.32
Average	−.38, −.29, .51, .68	−.38, −.30, .46, *	−.34, −.27, .40, .50	1.21
Actual T	(−.36, −.28, .50, .69)			

One can replace $L_{\mathbf{x}}(\mu, \sigma)$ by $\hat{L}_{\mathbf{x}}(\mu, \sigma) = \prod \hat{f}_{\mu,\sigma}(x_i)$, where $\hat{f}_{\mu,\sigma}(\cdot)$ is a window estimate of the unknown density function, and thereby estimate $t_\alpha^{\mathbf{x}}$, $t_{1-\alpha}^{\mathbf{x}}$ from the upper equation of (9.4). This was attempted for the 10 samples of Table 5, but with unsatisfactory results. The estimates were extremely sensitive to the way the window estimate $\hat{f}_{\mu,\sigma}(\cdot)$ was constructed. Fraser (1976) obtained better results, but in his examples the true f was symmetric and $n = 100$.

10. JOHNSON'S t-STATISTIC

Johnson (1978) uses a Cornish-Fisher expansion to suggest that a certain function of (9.1),

$$g_\gamma(T) = T + \frac{\gamma T^2}{3\sqrt{n}} + \frac{\gamma}{6\sqrt{n}}, \tag{10.1}$$

is more nearly distributed as a standard t_{n-1}-variable than is T itself. Here γ is the skewness of the distribution F, $\gamma = \mu_3/\mu_2^{3/2}$, $\mu_k = \mathscr{E}_F(X - \mathscr{E}_F X)^k$.

Column (2) of Table 5 was obtained by estimating γ by the sample skewness $\hat{\gamma} = \hat{\mu}_3/s^3$, where $s^2 = \sum (x_i - \bar{x})^2/(n - 1)$; assuming that $g_{\hat{\gamma}}(T)$ was distributed as t_{n-1}; and solving for the values of μ which gave $g_{\hat{\gamma}}((\hat{\mu} - \mu)/\hat{\sigma})$ in the central $1 - 2\alpha$ region for t_{n-1}. For example, in trial 1, μ in the interval $[-0.37, 0.70]$ gave $g_{1.40}((\hat{\mu} - \mu)/\hat{\sigma}) = g_{1.40}(-\sqrt{15}\mu)$ in the interval $[-1.76, 1.76]$.

The results are in close agreement with column (1), the bootstrap t. In trials 5 and 8, which had large skewness, the upper 95% point could not be calculated. This is because (10.1) describes a parabolic curve, and in these two trials no value of μ gave $g_{\hat{\gamma}}(T) < -1.76$. [Column (2) of Table 5 ignores those values of μ far from 0 which also give $g_{\hat{\gamma}}(-\sqrt{15}\ \mu)$ in the proper regions, the solutions on the wrong arm of the parabolic function (10.1).]

It is not surprising that Johnson's t and the bootstrap t tend to agree. Both methods can be described as follows: (i) the percentile points of $T = (\hat{\mu} - \mu)/\hat{\sigma}$ are functions of F, say $t_\alpha(F)$; (ii) if we knew F, we would assign μ central $1 - 2\alpha$ interval $[\hat{\mu} - \hat{\sigma} t_{1-\alpha}(F), \hat{\mu} - \hat{\sigma} t_\alpha(F)]$; (iii) we don't know F, so we assign the interval $[\hat{\mu} - \hat{\sigma} t_{1-\alpha}(\hat{F}), \hat{\mu} - \hat{\sigma} t_\alpha(\hat{F})]$. The bootstrap evaluates $t_\alpha(\hat{F})$ directly, while Johnson's method uses an asymptotic formula to approximate $t_\alpha(\hat{F})$, $t_\alpha(\hat{F}) \simeq g_{\hat{\gamma}}^{-1}(t_{\alpha,n-1})$.

Johnson's t is much easier to use than the bootstrap t. Its unpleasant features, as shown in Tables 5 and 8, can be mitigated by replacing $g_\gamma(T)$ with a monotonic function having the same first and second derivatives at $T = 0$, and perhaps, by estimating γ more robustly. It certainly deserves further attention as a device for assigning confidence intervals to a location parameter. [The expression (10.1) is appropriate only for T as defined in (9.1). For each choice of $\hat{\mu}$, $\hat{\sigma}$ other than the sample mean and standard error, an appropriate transformation of $T = (\hat{\mu} - \mu)/\hat{\sigma}$ must be derived.]

The bootstrap t, and by implication Johnson's method also, performed poorly when used to set confidence intervals for the correlation coefficient. In this case the T-statistic was taken to be

$$T = \frac{\hat{\rho} - \rho}{\hat{\sigma}_J}, \tag{10.2}$$

$\hat{\sigma}_J$ being the jackknife estimate of standard error for $\hat{\rho}$. A bootstrap replication was $T^* = (\hat{\rho}^* - \hat{\rho})/\hat{\sigma}_J^*$, where $\hat{\sigma}_J^*$ was the jackknife estimate of standard error in the bootstrap sample. The bootstrap distribution of T^* was obtained by Monte Carlo, its percentiles $\hat{t}_\alpha$ calculated, and $[\hat{\rho} - \hat{t}_{1-\alpha}\hat{\sigma}_J,\ \hat{\rho} - \hat{t}_\alpha\hat{\sigma}_J]$ assigned as a central $1 - 2\alpha$ interval for ρ. The results varied eccentrically with a tendency toward occasional extremely long intervals.

11. NONPARAMETRIC TILTING

A parametric confidence set consists of those members of the parametric family which cannot be soundly discredited as having generated the observed data. "Soundly discredited" means rejected by a hypothesis test. The hypothesis test is chosen to minimize, at least approximately, the length of the confidence interval for the parameter of interest. If the parameter is the expectation μ (for example), the test might be based on the sampling distribution of $\hat{\mu}$, the MLE for μ.

This program can be difficult to carry out, even in well-defined parametric situations. We have a trial value of μ, say μ^t, which we want to test for inclusion or exclusion from the confidence interval. However, the distribution of the test statistic $\hat{\mu}$ depends upon nuisance parameters as well as μ^t, so it may be difficult to guarantee the significance level of the hypothesis test.

In a nonparametric situation there are an infinity of nuisance parameters, which makes the program described above impossible for most parameters μ. This section discusses a method, *nonparametric tilting*, which is a less ambitious version of this same basic idea. First we give an operational description of how the entries in column (3) of Table 5 were obtained, followed by a more general discussion of the method, and its connection with the previous results.

For a given value of the real number t, define weights

$$w_i^t = \frac{e^{tx_i}}{\sum_{j=1}^{n} e^{tx_j}}, \qquad i = 1, 2, \ldots, n, \tag{11.1}$$

the data $x_1, x_2, \ldots, x_n$ ($n = 15$ in our example) being fixed as observed. The trial value of the expectation μ, which we shall test for inclusion or exclusion from the confidence interval, is

$$\mu^t = \sum_{i=1}^{n} w_i^t x_i. \tag{11.2}$$

Notice that $t = 0$ gives $\mu^t = \hat{\mu} = \bar{x}$. Instead of (3.2), consider choosing resampling vectors $\mathbf{P}^* = (P_1^*, \ldots, \mathrm{P}_n^*)$ according to

$$\mathbf{P}^* \sim \frac{\mathrm{Mult}_n(n, \mathbf{w}^t)}{n}, \tag{11.3}$$

$\mathbf{w}^t = (w_1^t, w_2^t, \ldots, w_n^t)$. If $t = 0$ then (11.3) is the same as (3.2), but otherwise the bootstrap sample is selected nonuniformly from $\{x_1, x_2, \ldots, x_n\}$: $\mathrm{Prob}_* \{X_j^* = x_i\} = w_i^t$. This defines a "tilted" bootstrap distribution for $\hat{\mu}^* = \sum_{i=1}^n P_i^* x_i$, with expectation $\sum w_i^t x_i = \mu^t$ and variance $\mathbf{x}\ \mathbf{\Sigma}^t \mathbf{x}'/n$, where $\mathbf{\Sigma}^t$ has diagonal elements $w_i^t(1 - w_i^t)$, off-diagonals $-w_i^t w_j^t$. We shall write Prob_*^t to indicate probabilities under this distribution.

Define

$$\alpha^t = \mathrm{Prob}_*^t \{\hat{\mu}^* < \hat{\mu}\}, \tag{11.4}$$

the achieved significance level of the observed value $\hat{\mu} = \bar{x}$ under the bootstrap distribution (11.3) of $\hat{\mu}^*$. The upper 95% point of the tilted confidence interval (=0.56 for trial 1 of Table 5) is the value of μ^t corresponding to that t having $\alpha^t = 0.05$, and similarly for the other percentile points. The results reported in Table 5 are intermediate to the Pitman intervals and the bias-corrected percentile intervals (Table 4), lying somewhat closer to the latter.

The distributions (11.3), thought of as a family indexed by t, form a one-parameter exponential family with sufficient statistic $\hat{\mu}^* = \sum_{i=1}^n P_i^* x_i$. (The x_i are fixed at their observed values, as before.) An easy calculation shows that the probability density function of $\hat{\mu}^*$, say $f_t(\hat{\mu}^*)$, satisfies

$$\frac{f_t(\hat{\mu}^*)}{f_0(\hat{\mu}^*)} = e^{n[t(\hat{\mu}^* - \hat{\mu}) - \phi(t)]}, \qquad \text{where} \quad \phi(t) = \log \frac{1}{n} \sum_{i=1}^{n} e^{t(x_i - \bar{x})}. \tag{11.5}$$

In other words, the bootstrap distribution of $\hat{\mu}^*$ under (11.4) is an exponential tilt of the bootstrap distribution under (3.2). The calculations in column 3 of Table 5 were obtained using this shortcut trick. Tilting is a useful tool in large-deviations theory; see Chernoff (1972, p. 45).

The motivation for the weights (11.1) is as follows. Among all distributions putting mass only on the observed data points $x_1, x_2, \ldots, x_n$,

$$\mathbf{w}\colon \text{mass } w_i \text{ on } x_i \qquad \left(w_i \geq 0, \quad \sum_{i=1}^{n} w_i = 1 \right), \tag{11.6}$$

The choice $\mathbf{w} = \mathbf{w}^t = (w_1^t, \ldots, w_n^t)$ minimizes the Kullback-Leibler distance from $\mathbf{w}^0 = \mathbf{1}/n$,

$$D(\mathbf{w}, \mathbf{w}^0) = \sum_{i=1}^{n} w_i \log(n w_i), \tag{11.7}$$

subject to the constraint $\mu(\mathbf{w}) = \sum w_i x_i = \mu^t$. In this sense $\mathbf{w}^t$ is the closest distribution to the observed data, subject to $\mu = \mu^t$. Testing the observed value $\hat{\mu}$ versus the $\mathbf{w}^t$ distribution of $\hat{\mu}^*$, to see whether or not to include μ^t in the confidence interval, is similar to parametric techniques. In effect we are estimating the nuisance parameters (everything about F except μ) as well as possible, subject to $\mu = \mu^t$, and then using this estimated distribution to assign a significance value to the observed value $\hat{\mu}$.

In a parametric problem, with parameters (μ, η), it is common to estimate the nuisance parameters η by $\hat{\eta}^t$, the MLE subject to $\mu = \mu^t$. We can do the same thing here. Instead of (11.1), the weights turn out to be

$$w_i^t = \frac{[1 + t x_i]^{-1}}{\sum_{j=1}^{n} [1 + t x_j]^{-1}}. \tag{11.8}$$

Then $\mathbf{w}^t$ is the nonparametric maximum-likelihood estimate of F subject to $\mu = \mu^t = \sum w_i^t x_i$, with the understanding that we are only considering distributions supported on $x_1, \ldots, x_n$. The distance minimized is $D(\mathbf{1}/n, \mathbf{w}) = \sum (1/n) \log(1/n w_i)$ rather than (11.7). Our method of assigning confidence intervals, comparing the observed $\hat{\mu}$ with the distribution of $\hat{\mu}^*$ under $(\mu^t, \hat{\eta}^t)$, is now quite similar to standard exponential-family testing theory, except that the latter would usually be based on conditional rather than unconditional distributions. The most compelling reason for using (11.1) rather than (11.8) in Table 5 was the computational advantage of the tilting argument (11.5).

Nonparametric tilting is fundamentally more ambitious than anything else we have considered. The bootstrap, and the other methods, replace the true distribution F by an estimate $\hat{F}$. Tilting replaces the entire family of possible distributions we might consider, say $F \in \mathscr{F}$, by an estimated family $\hat{\mathscr{F}}$. The estimated family is

$$\hat{\mathscr{F}}: \mathbf{m} \sim \text{Mult}_n(n, \mathbf{w}^t) \qquad (w_i^t = e^{tx_i}/\textstyle\sum e^{tx_j}, \quad -\infty < t < \infty). \tag{11.9}$$

The entries in column (3) of Table 5 are the confidence limits for $\mu(\mathbf{w}) = \sum w_i x_i$ in the family $\hat{\mathscr{F}}$, having observed $\mathbf{m} = (1, 1, \ldots, 1)$. As before, the data $x_1, \ldots, x_n$ are considered fixed.

It is easy to verify that the Cramér-Rao variance bound for the unbiased estimation of $\mu(\mathbf{w})$ in $\hat{\mathscr{F}}$, evaluated at $\mathbf{w} = \mathbf{w}^0 = (1/n, \ldots, 1/n)$, equals

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n^2}, \tag{11.10}$$

which is the bootstrap estimate of variance for $\hat{\mu} = \bar{x}$, i.e., the nonparametric maximum-likelihood estimate of variance. The estimated variance is not made smaller by restricting attention to the one-parameter family $\hat{\mathscr{F}}$, rather than considering the problem of estimating μ in a full nonparametric setting. In this sense at least, the restriction to $\hat{\mathscr{F}}$ isn't spuriously helpful. This "least favourable" property of $\hat{\mathscr{F}}$ can be shown to hold everywhere, not just at $\mathbf{w} = \mathbf{w}^0$, but won't be discussed further here.

Closed-form expressions for the endpoints of the tilting intervals can be obtained if one is willing to accept certain approximations: $\phi(t)$, (11.5), is approximated by a Taylor series about 0, beginning $\phi(t) = \hat{\mu}_2 t^2/2 + \hat{\mu}_3 t^3/6 + \ldots$, and the distribution of $\hat{\mu}^* = \sum m_i x_i/n$ under (11.9) is approximated by an Edgeworth series. The crudest such approximation gives the interval $\hat{\mu} \pm z_\alpha \sqrt{\hat{\mu}_2/n}$, the standard large-sample result. Going to the next level of approximation gives, to a reasonable degree of accuracy, Johnson's t-interval, as described in Section 10.

As a final point, suppose we are interested in a parameter $\theta(F)$ other than the expectation. Let $\theta(\mathbf{w})$ be the value of θ for that F putting mass w_i on x_i, $i = 1, 2, \ldots, n$, and define

$$U_i(\mathbf{w}) = \lim_{\varepsilon \to 0} \frac{\theta((1 - \varepsilon)\mathbf{w} + \varepsilon \boldsymbol{\delta}_i) - \theta(\mathbf{w})}{\varepsilon}. \tag{11.11}$$

In particular $U_i(\mathbf{w}^0)$ is the empirical influence function of θ, called U_i^0 in (3.7). For a trial value $\theta = \theta^t$, we can now look for the vector $\mathbf{w}^t$ minimizing $D(\mathbf{w}, \mathbf{w}^0) = \sum w_i \log(nw_i)$, as at (11.7), among all $\mathbf{w}$ satisfying $\theta(\mathbf{w}) = \theta^t$. Standard calculations show that the solution satisfies

$$w_i^t = \frac{e^{tU_i(\mathbf{w}^t)}}{\sum_{j=1}^{n} e^{tU_j(\mathbf{w}^t)}}, \qquad i = 1, 2, \ldots, n. \tag{11.12}$$

Equation (11.1) is a special case of (11.12), since it can be expressed as $w_i^t = \exp\{t(x_i - \bar{x}^t)\}/\sum \exp\{t(x_j - \bar{x}^t)\}$, $\bar{x}^t = \sum w_j^t x_j$, and $U_i(\mathbf{w}) = x_i - \sum w_j x_j$ for $\theta(\mathbf{w}) = \sum w_i x_i$.

Confidence intervals for θ can now be constructed as before. They are the limits for $\theta(\mathbf{w})$ in $\hat{\mathscr{F}} = \{\mathbf{m} \sim \text{Mult}_n(n, w^t)\}$, $\mathbf{w}^t$ defined by (11.12), having observed $\mathbf{m} =$

$(1, 1, \ldots, 1)$. The Cramér-Rao variance bound for $\theta(\mathbf{w})$ in $\hat{\mathscr{F}}$, evaluated at $\mathbf{w} = \mathbf{w}^0$, equals $\sum U_i^2(\mathbf{w}^0)/n^2$. This is the infinitesimal-jackknife, influence-function, delta-method nonparametric estimate for the variance of $\hat{\theta} = \theta(\hat{F})$, (3.8), in analogy to (11.10). Unfortunately, the tilting property (11.5) no longer applies, so that the actual computation of the tilted intervals appears difficult.

RÉSUMÉ

Plusieurs méthodes non paramétriques ont été étudiées, en particulier le «bootstrap», le «jackknife», et la méthode de «delta». Dans un premier temps, on attribue des erreurs types non paramétriques à une statistique à valeurs réelles. On considère ensuite le problème plus ambitieux de construire des intervalles de confiance non paramétriques pour un paramètre à valeurs réelles. Partant du cas bien connu des intervalles de confiance pour la médiane, des indications sont fournies à l'effet qu'une telle théorie semble possible.

REFERENCES

Chernoff, H. (1972). *Sequential Analysis and Optimal Design.* Society for Industrial and Applied Mathematics, Philadelphia.

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

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 Rev.*, 21, 460–480.

Efron, B. (1980a). Censored data and the bootstrap. *J. Amer. Statist. Assoc.*, 76, 312–319.

Efron, B. (1980b). Nonparametric estimates of standard error: The jackknife, the bootstrap, and other methods. Technical Report No. 56, Dept. of Statistics, Stanford University. *Biometrika,* to appear.

Efron, B. (1980c). The jackknife, the bootstrap, and other resampling plans. Technical Report No. 63, Dept. of Statistics, Stanford University.

Efron, B., and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.*, 9, 586–596.

Fraser, D.A.S. (1976). Necessary analysis and adaptive inference. *J. Amer. Statist. Assoc.*, 71, 99–113.,

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69, 383–393.

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

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

Hartigan, J.A. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Ann. Statist.*, 3, 573–580.

Jaeckel, L. (1972). The infinitesimal jackknife. Bell Laboratories Memorandum No. 72-1215-11.

Johnson, N.J. (1978). Modified *t* tests and confidence intervals for asymmetrical populations. *J. Amer. Statist. Assoc.*, 73, 536–544.

Johnson, N.J., and Kotz, S. (1970). *Continuous Univariate Distributions—2.* Houghton Mifflin, Boston.

Mallows, C.L. (1974). On some topics in robustness. Bell Laboratories Memorandum.

Pearson, E.S., and Hartley, H.O., *eds.* (1954). *Biometrika Tables for Statisticians. Volume I.* Cambridge Univ. Press.

Rubin, D.B. (1981). The Bayesian bootstrap. *Ann. Statist.*, 9, 130–134.

Received 27 March 1981
Revised 24 June 1981

Department of Statistics
Stanford University
Sequoia Hall
Stanford, California 94305, U.S.A.

Discussion

George A. BARNARD, *University of Waterloo*

Efron's paper gives us a very useful survey of some old and some new answers to the question: What should we do when we do not know the form of the observational distribution?