critical, but so it is in nonparametric methods in which statisticians are quite happy to consider parameter dependent reference sets. Here is our recipe for straight line regression, referring to Professor Wu's paper.

First set (on the computer) the value of β. Then *evaluate* the

$$z_i = y_i - x_i^T\beta.$$

Bootstrap the z_i values (keeping the x_i values fixed). Regress each bootstrap z_i^* set back on the x_i values to obtain a $\hat{\beta}^*$ value for each bootstrap. Smooth the set of $\hat{\beta}^*$ to obtain $\hat{f}(\hat{\beta}|\beta)$. Note that $\hat{f}$ depends on the set value β. Put $\hat{\beta} = \hat{\beta}_0$ the value obtained from the original (unbootstrapped) z_i values and we have our generated likelihood $L(\beta)$. Here $\hat{\beta}$ plays the role of the statistic T.

REFERENCES

DAVISON, A., HINKLEY, D. and SCHECHTMAN, E. (1986). Efficient bootstrap simulation. *Biometrika* **73**. To appear.

NIEDERREITER, H. (1978). Quasi-Monte Carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc.* **84** 957–1041.

TUKEY, J. W., BRILLINGER, D. R. and JONES, L. V. (1978). *The Management of Weather Resources* **2**. Weather Modification Advisory Board, Statistical Task Force, U.S. GPO, Washington.

DEPARTMENT OF MATHEMATICS
UNIVERSITY STATISTICAL LABORAOTRY
CITY UNIVERSITY
NORTHAMPTON SQUARE
LONDON EC1V 0HB
ENGLAND

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF BENIN
PMB 1154, EKENWAN ROAD
BENIN CITY
NIGERIA

REJOINDER

C. F. J. WU

University of Wisconsin-Madison

The overwhelming response to the paper reflects great interest and perhaps confusion in the bootstrap and the jackknife. The contributions of the discussants make the discussion informative, valuable and diverse. Their comments, even though I do not always agree with them, help clarify certain points, suggest new ideas and results, and in some cases prompt me to study the issues more carefully. Most of these comments can be grouped into five broad categories. My reply will concentrate on the major points of interest in each category.

Among the new ideas and results to which my response will not be directed, let me mention: robustification of resampled values (*Beran*), two interesting applications from genetics (*Felsenstein* and *Mitchell-Olds*), examples of inconsistency of bootstrap estimators (*Olshen* and *Srivastava*), use of weighted jackknife in variance components model (*Rao and Prasad*), results on the

stochastic order of some bias estimators (*Shao*), and an efficient simulation code (*Wynn and Ogbonmwan*). The counterexamples given or alluded to by *Olshen* and *Srivastava* may be taken as a caveat concerning the indiscriminate use of resampling methods.

1. A sharpened jackknife. The classical jackknife can be made more versatile in several ways.

(i) The delete-d jackknife, with $d \to \infty$ as $n \to \infty$, can handle function estimation (as well as moment estimation) and nonsmooth parameters.

(ii) The subsets can be chosen in a more efficient way by using Hadamard matrices and other combinatorial techniques.

(iii) Proper weighting allows the jackknife to handle nonexchangeable problems.

With this in mind, let me now respond to some of the comments.

Felsenstein argues that, in the construction of evolutionary trees the extrapolation factor $(r - k + 1)/(n - r)$ should be avoided. This is in agreement with our finding (Rao and Wu (1986)) in a different context. If the parameter β is confined to a particular region (e.g., positive), the resampled estimate $\tilde{\beta}_s$ can be outside the region, which may cause serious problems. The subset size $(n + k - 1)/2$ is a sensible choice here. He mentions the possibility of dropping half the observations *at random* as an alternative to the bootstrap but cannot see any advantages in doing so. I think there may be advantages in numerical efficiency if the half-samples are dropped in a *balanced* manner such as in (7.5)–(7.8).

Freedman's example shows that the jackknife, weighted or not, does not provide a good guide to the *distribution* of the estimator if the errors are heteroscedastic, skewed and long-tailed. The jackknife is not alone in this regard. Even for i.i.d. errors (but with heavy tails), the bootstrap can have similar difficulties according to the results of Athreya (1987) and of Ghosh et al. (1984) (see also Olshen's discussion.) A similar phenomenon was observed for plain i.i.d. errors with finite second moments (Wu (1986)), in which I prove that the histogram of the delete-d jackknife consistently estimates the asymptotic distribution of the estimator (say, the one-sample mean) iff d and $n - d \to \infty$. Therefore, only the delete-d jackknife with *unbounded* d and $n - d$ provides a good guide to the distribution of the estimator.

Ghosh questions the use of a delete-d jackknife for nonsmooth parameters. It is shown in Shao and Wu (1986) that, for a class of nonsmooth functionals, including the sample median, the jackknife variance estimators with $d = O(n)$ are consistent. It does rectify the deficiency of the delete-1 jackknife. He is right in questioning the robustness of the jackknife (or other resampling) estimators against correlated errors. Estimators that are consistent for correlated errors are available in the econometrics literature.

Singh points out that the nonsingularity requirement of $X_s^T X_s$ for every s is too restrictive. It is only required for proving exact results. In fact any method (including the bootstrap) that resamples from (y_i, x_i) will have the same restric-

tion. This is not a serious problem if r is substantially bigger than k and the problem is not ill-posed. Since the proportion of such subsets will then be small, they can be discarded in the resampling without affecting the estimator's performance. His question about the jackknife histograms can be readily answered by a result from Wu (1986). In addition to the asymptotic normality of the jackknife histogram (see the above response to Freedman), I show that the jackknife histogram possesses a desirable second-order property (i.e., capturing the second-order term of the Edgeworth expansion) if r is chosen to be approximately $0.724n$.

2. Bootstrap and modeling. A main theme of the paper, which most discussants seem to agree on, is that "no resampling methods, no matter what the computing power is, can replace good work in modeling and analysis." This is in contrast to the euphoria in the earlier literature on the bootstrap (Efron and Gong (1983), Section 1). My example ((6.14)–(6.17)) on bootstrapping the μ parameter illustrates the kind of difficulties that can result from *routine* and *blind* use of a resampling method. Efforts in understanding the parameters of the model will alleviate such problems, which is also pointed out by *Efron* and *Tibshirani*.

Another example provided by *Tibshirani* is the misspecification of the mean component of the regression model. This is a more serious violation, because the bias it induces dominates the variance. His defense of the unweighted bootstrap as a way of estimating the variability of the (inconsistent) estimator fails to address the more important issue of bias.

Carroll and Ruppert and *Srivastava* emphasize the importance of carefully modeling the heteroscedasticity of errors. Resampling methods such as the weighted jackknife can be used to take care of the residual heteroscedasticity.

I find *Efron's* schematic diagram very useful in relating resampling to the original sampling. What the diagram depicts is the simulation approach to statistical inference, which predates the bootstrap. Such a simulation approach includes the bootstrap, the jackknife and many others. There are, however, situations in which this approach is not applicable. One such example is when y is obtained from unequal probability selection without replacement. Another example, provoked by the comments of *Tibshirani* and *Weber*, is the familiar heteroscedastic linear model with a small number of replications for each x_i. Since there are not enough observations for estimating each error distribution, one may have to be content with estimating the moments (e.g., variance) rather than the distribution. Methods based on matching the first two moments such as those in Section 7 are intended only for this.

Several of the jackknife variance estimates can be expressed as

$$(X^TX)^{-1}\sum_1^n \hat{\sigma}_i^2 x_i x_i^T (X^TX)^{-1}.$$

Beran and *Efron* point out that this can also be obtained by a heteroscedastic bootstrap of the residuals. However, this equivalence between the jackknife and

the bootstrap breaks down in nonlinear situations. Take, for example, the binary regression problem. The weighted jackknife is applicable (see Section 8), whereas the heteroscedastic bootstrap, which is based on resampling the residuals, is not.

3. Conditional and unconditional inference. This issue is raised by *Efron*, *Hinkley*, *Olshen* and *Tibshirani*. The justification of the unweighted bootstrap as an unconditional procedure requires the strong assumption that the x_i's are a *random* sample from the population. Is this assumption tenable or verifiable? In data analysis, how often do analysts bother to find out what the sampling design is? On the other hand, a conditionally valid procedure such as the weighted jackknife does not require such a stringent condition on the sampling design.

Olshen questions the appropriateness of the exchangeability assumption (A) for nonexchangeable models such as those in the paper. In the proposed weighted approach, the nonexchangeability in the model is accounted for by weighting adjustment, not by resampling. *Shao's* proposal does it the other way. *Hinkley's* conditional bootstrap, which I find very interesting, is another way of handling nonexchangeability.

4. Bias and variance of a variance estimator. The paper's overemphasis on bias and lack of other theoretical properties have invited criticism (*Efron*, *Ghosh*, *Rao and Prasad*, *Singh* and *Tibshirani*). Some of these properties have already been studied in Shao and Wu (1985). Let me mention results that are relevant to the discussion. Consider the heteroscedastic case. It is obvious that the usual estimator $\hat{v}$ $(= v_b)$ is inconsistent. For the consistency of $v_{H(1)}$ and v_J, the condition

$$h_n = \max_{1 \le i \le n} x_i^T (X^T X)^{-1} x_i \to 0 \tag{1}$$

is necessary. On the other hand, the consistency of $v_{J(1)}$ does not necessarily require (1) (Shao and Wu (1985), Theorems 6 and 3). Next consider the orders of (bias)^2 and variance. The variances of the four estimators are of the order $n^{-2}h_n$. The squared biases of $v_{J(1)}$, v_J and $v_{H(1)}$ are of the order $n^{-2}h_n^2$, but the squared bias of $\hat{v}$ is generally of the higher order n^{-2}. Therefore, the bias of the ordinary estimator $\hat{v}$ (and the bootstrap estimator v_b) plays a more dominant role and should be given greater attention. *Tibshirani* comments that the biases of v_J and $v_{J(1)}$ are of the same order, which is in agreement with the preceding results. A more refined analysis (Shao and Wu (1985), Theorem 7) shows that, in a qualitative sense, v_J has a bigger bias than $v_{J(1)}$. In summary, $v_{J(1)}$ has stronger theoretical justifications than the others.

Simonoff and Tsai suggest the estimator RLQM, which performs well in their simulation study. For linear models, RLQM reduces to

$$v_{\mathrm{RLQM}} = \hat{\sigma}^2 (X^T X)^{-1} \sum_1^n (1 - w_i)^2 x_i x_i^T (X^T X)^{-1}, \tag{2}$$

which is not consistent. Additional comments will be given later. Another estimator $v_{*,\mathrm{new}}$, proposed by *Srivastava*, is also inconsistent.

The small sample behavior of variance estimators is a more difficult subject. An attempt is made here to understand qualitatively the stability of $v_{J(1)}$, $v_{H(1)}$ and $\hat{v}$. From the form of $v_{J(1)}$,

$$v_{J(1)} = (X^TX)^{-1}\sum_1^n \frac{r_i^2}{1-w_i}x_ix_i^T(X^TX)^{-1},$$

there are two contributing factors to its possibly big variance: (i) w_i that are close to 1 and (ii) variability in r_i^2. To bound the influence of the large w_i's, $1 - w_i$ is replaced by its average $1 - k/n$ in $v_{H(1)}$ (2.6). To bound the influence of the variability of each r_i^2, albeit at the expense of a bigger bias,

$$\sum_1^n \frac{r_i^2}{1-k/n}x_ix_i^T \tag{3}$$

in $v_{H(1)}$ is replaced by

$$\frac{1}{n-k}\sum_1^n r_i^2 \sum_1^n x_ix_i^T,$$

which results in the estimator $\hat{v}$. The bigger root mean squared error (rmse) of $v_{J(1)}$ (than that of $v_{H(1)}$) appears to be caused by a few big values of w_i. Another way to stabilize $v_{J(1)}$ is to bound the w_i values. That is, define

$$w_i' = \begin{cases} w_i & \text{if } w_i \le c \\ c & \text{if } w_i > c \end{cases}$$

and

$$v_{J(1)}(c) = \frac{n-k}{\sum(1-w_i')}(X^TX)^{-1}\sum \frac{r_i^2}{1-w_i'}x_ix_i^T(X^TX)^{-1}. \tag{4}$$

Consider again the form of v_{RLQM} (2). The influence of the w_i is diminished by using the weight $(1 - w_i)^2$ and, like $\hat{v}$, $\hat{\sigma}^2$ is used instead of (3). A simple approximation to v_{RLQM} is

$$\bar{u}\hat{v},\ \bar{u} = \text{average of } (1-w_i)^2, \tag{5}$$

which is obtained by taking the average of $(1 - w_i)^2$ from the middle matrix of v_{RLQM}. In the simulation study (Table 1A) $\bar{u}\hat{v}$ outperforms v_{RLQM}. Their small variances in this situation appear to come from the smallness of the coefficients $(1 - w_i)^2$ rather than the justification given in Simonoff and Tsai (1986).

We reran the simulation study reported in Table 1 for these new estimators. Their rmse's are given in Table 1A. (The rmse's of the variance estimators considered in the paper are independent of the β values.) The simple estimator $\bar{u}\hat{v}$ outperforms v_{RLQM} ($\bar{u} = 0.596$) and the other estimators. In fairness to $v_{H(1)}$ and $v_{J(1)}$, it is possible to reduce their rmse's in this particular situation by using the modification $cv_{H(1)}$ and $cv_{J(1)}$, $c < 1$. By bounding the w_i values in $v_{J(1)}$, the resulting estimator $v_{J(1)}(0.4)$ is comparable to $v_{H(1)}$. The message is clear: Improvement in rmse in this situation is mainly achieved by *downweighting*.

TABLE 1A
Root mean squared errors of several variance estimators (3000 simulations). Unequal variances ($\sigma_i^2 = x_i/2$).

	(0,0)	(0,1)	(0,2)	(1,1)	(1,2)	(2,2)
$\hat{v}$	1.24	0.46	0.038	0.23	0.022	0.0024
$v_{J(1)}$	0.99	0.52	0.052	0.31	0.033	0.0038
$v_{J(1)}(0.4)$	0.80	0.41	0.041	0.25	0.026	0.0030
$v_{H(1)}$	0.77	0.40	0.040	0.24	0.026	0.0029
v_{RLQM}	0.74	0.46	0.049	0.30	0.033	0.0038
$\bar{u}\hat{v}$	0.71	0.39	0.040	0.26	0.028	0.0031

TABLE 1B
Root mean squared errors of several variance estimators (500 simulations). Unequal variances ($\sigma_i^2 = 2.5/|x_i - 5.5|$).

	(0,0)	(0,1)	(0,2)	(1,1)	(1,2)	(2,2)
$\hat{v}$	1.22	0.46	0.038	0.22	0.019	0.0019
$v_{J(1)}$	0.82	0.44	0.038	0.26	0.024	0.0022
$v_{J(1)}(0.4)$	0.74	0.39	0.033	0.24	0.020	0.0018
$v_{H(1)}$	0.76	0.42	0.036	0.26	0.022	0.0020

Efron's simulation produces dramatic results in favor of $\hat{v}$ $(= v_b)$ for the particular variance pattern $|x_i - 5.5|$. Here high leverage (big w_i) goes with big σ_i. However, the conclusion depends on the setting chosen. We repeated his study by changing the variance pattern to

$$e_i \sim N\left(0, \frac{2.5}{|x_i - 5.5|}\right),$$

where small leverage goes with big σ_i. The results are given in Table 1B. The picture here is different from the one painted by Efron. The estimator $\hat{v}$ for $\text{Var}(\hat{\beta}_0)$ now trails behind the other three. The modified estimator $v_{J(1)}(0.4)$ is slightly better than $v_{H(1)}$, which in turn is slightly better than $v_{J(1)}$.

It is clear that simulation alone does not provide a reliable guide to the relative performance of various variance estimators. The small-sample behavior of a variance estimator depends on x_i and e_i in a complicated manner. A thorough theoretical investigation is called for. Some of the questions are already outlined in *Efron's* discussion. Such a study should include MINQUE and related methods. *Rao and Prasad* note one such result for a special case (Rao (1973)).

5. Asymptotics and coverage probability. At the suggestions of *Beran*, *Efron*, *Hall* and *Hinkley*, I have included in the simulation study the studentized bootstrap, where $(v_{\text{lin}})^{1/2}$ is used as the standard error estimate. The results are given under TBOOT in the last row of Tables 3 and 4. Except for

TABLE 3A
Error rates for six methods (3000 simulations). Nominal rate = 0.05.

	Unequal variances β_2				Equal variances β_2	
	−0.25		−0.35		−0.25	
	left	right	left	right	left	right
VHJ(1)	0.013	0.121	0.067	0.088	0.020	0.082
VHJ8	0.003	0.066	0.031	0.061	0.006	0.053
VBOOT	0.002	0.111	0.026	0.072	0.004	0.040
VLIN	0.005	0.130	0.034	0.075	0.005	0.046
PBOOT	0.071	0.100	0.089	0.097	0.047	0.041
TBOOT	0.147	0.098	0.096	0.087	0.049	0.034

$\beta_2 = -0.25$ and unequal variances, it gives improvement over the bootstrap percentile method (PBOOT). However, it still performs worse than VLIN and other t-intervals. In an unpublished study I have found the studentized bootstrap intervals to be too liberal for heavy-tailed distributions. The study of Loh (1987) shows that in some situations it can give unduly large intervals.

So far I have considered only two-sided intervals. *Singh* and *Hall* point out the difference in asymptotic coverage probabilities between one-sided and two-sided intervals. According to Singh's formulas, the studentized bootstrap should give the best one-sided intervals. Is this supported by its small-sample performance? To answer this, the error rates of the various intervals in Table 3 are separated into two parts, those to the left and those to the right of the intervals. The results for selected parameter values and methods are given in Table 3A.

Singh's asymptotics correctly predict that the two (left and right) error rates are more evenly distributed for the histogram-based methods PBOOT and TBOOT. But on more important grounds his asymptotics seem to fail completely. Indeed, TBOOT has very high error rates on both sides, contrary to his formula $P(\mu < L_{B,st}) = \frac{1}{2}\alpha + o(n^{-1/2})$. Here the lengths of the TBOOT intervals are comparable to the others. Asymmetry is a requirement for good confidence intervals. It *alone* does not guarantee that the intervals are good. I had no misunderstanding when I wrote "This is very disappointing in view of the second-order asymptotic results on the bootstrap." I understand and appreciate the large sample validity of these results. The question is whether they can deliver their promise in the small- or moderate-sample situations. I am not surprised that they do not. The coefficients of an Edgeworth expansion, on which the asymptotic justification is based, are functions of moments. Can moments be used to adequately describe the delicate behavior in the tails? Before these issues are properly addressed, I will stick to my original claim that "theoretical results that can explain small-sample behavior are needed."

The problem of finding nonparametric confidence intervals is a difficult one. Many methods have been proposed. It is fair to say that so far no clear winner has emerged (see, for example Loh (1987); Loh and Wu (1986)). This goal is

perhaps too ambitious. Bahadur and Savage (1956) showed that it is impossible to find nonparametric confidence intervals without any restriction on the distribution F. In practice one should have some knowledge about F (I sound more like a Bayesian now!). By narrowing the class of distributions F belongs to, such intervals may be obtainable. As Efron says, the problem is far from being solved.

Acknowledgment. I sincerely thank the Editor and the Associate Editor for their efforts in organizing the discussion.

REFERENCES

ATHREYA, K. B. (1987). Bootstrap of the mean in the infinite variance case. To appear in *Ann. Statist.*

BAHADUR, R. R. and SAVAGE, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27** 1115–1122.

EFRON, B. and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. *Amer. Statist.* **37** 36–48.

GHOSH, M., PARR, W. C., SINGH, K. and BABU, G. J. (1984). A note on bootstrapping the sample median. *Ann. Statist.* **12** 1130–1135.

LOH, W. Y. (1987). Calibrating confidence coefficients. To appear in *J. Amer. Statist. Assoc.*

LOH, W. Y. and WU, C. F. J. (1986). Discussion of "Better confidence intervals" by B. Efron. To appear in *J. Amer. Statist. Assoc.*

RAO, J. N. K. (1973). On the estimation of heteroscedastic variances. *Biometrics* **29** 11–24.

RAO, J. N. K. and WU, C. F. J. (1986). Resampling inference with complex survey data. Unpublished.

SHAO, J. and WU, C. F. J. (1985). Heteroscedasticity-robustness of jackknife variance estimators in linear models. Technical report 778, Univ. of Wisconsin-Madison.

SHAO, J. and WU, C. F. J. (1986). Jackknifing a general statistical functional. Technical report, Univ. of Wisconsin-Madison.

SIMONOFF, J. S. and TSAI, C. L. (1986). Jackknife-based estimators and confidence regions in nonlinear regression. *Technometrics* **28** 103–112.

WU, C. F. J. (1986). On the asymptotic properties of the jackknife histograms. Technical report, Univ. of Wisconsin-Madison.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
1210 W. DAYTON ST.
MADISON, WISCONSIN 53706