THE JACKKNIFE, THE BOOTSTRAP, AND OTHER RESAMPLING PLANS

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 63
DECEMBER 1980

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 2 RO1 GM21215-06

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

THE JACKKNIFE, THE BOOTSTRAP, AND OTHER RESAMPLING PLANS

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 63
December 1980

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 2 R01 GM21215-06

Also prepared under National Science Foundation Grant MCS 77-16974, and issued as Technical Report No. 163, Stanford University, Department of Statistics.

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

TABLE OF CONTENTS

I.	Introduction	1
II.	The Jackknife Estimate of Bias	3
	1. Quenouille's Bias Estimate	4
	2. The Grouped Jackknife	6
	3. A Picture	6
	4. Aitken Acceleration	8
	5. The Law School Data	9
	6. What Does $\widehat{\text{BIAS}}$ Really Estimate?	11
III.	The Jackknife Estimate of Variance	14
	1. The Expectation	15
	2. The Unbiased Estimate of Variance	16
	3. Trimmed Means	17
	4. The Sample Median	19
	5. Ratio Estimation	19
	6. Functions of the Expectation	21
	7. The Law School Data	21
	8. Linear Regression	22
IV.	Bias of the Jackknife Variance Estimate	25
	1. ANOVA Decomposition of $\hat{\theta}$	26
	2. Proof of the Main Result	28
	3. Influence Functions	30
	4. Quadratic Functionals	30
	5. Sample Size Modification	32
V.	The Bootstrap	35
	1. Monte Carlo Evaluation of $\hat{\text{SD}}$	36
	2. Parametric Bootstrap	40
	3. Smoothed Bootstrap	41
	4. Bootstrap Methods for More General Problems	43
	5. The Bootstrap Estimate of Bias	44
	6. Finite Sample Spaces	46
	7. Regression Models	48
VI.	Infinitesimal Jackknife, Delta Method, and the Influence Function	51
	1. Resampling Procedures	51
	2. Relation Between the Jackknife and Bootstrap Estimates of Standard Deviation	54
	3. Jaeckel's Infinitesimal Jackknife	56
	4. Influence Function Estimates of Standard Deviation	59
	5. The Delta Method	60
	6. Estimates of Bias	63
	7. More General Random Variables	65

VII. Cross-Validation, The Jackknife, and The Bootstrap 68
1. Excess Error 69
2. Bootstrap Estimate of Expected Excess Error 74
3. Jackknife Approximation to the Bootstrap Estimate 75
4. Cross-Validation Estimate of Excess Error 76
5. Relationship Between the Cross-Validation and Jackknife Estimates 80
6. A Complicated Example 82

VIII. Balanced Repeated Replications (Half-Sampling) 86
1. Bootstrap Estimate of Standard Deviation 87
2. Half-Sample Estimate of Standard Deviation 88
3. Balanced Repeated Replications 90
4. Complementary Balanced Half-Samples 92
5. Some Possible Alternative Methods 94

IX. Random Subsampling 97
1. M-Estimates 97
2. The Typical Value Theorem 98
3. Random Subsampling 100
4. Resampling Asymptotics 102
5. Random Subsampling for Other Problems 103

X. Nonparametric Confidence Intervals 105
1. The Median 105
2. Typical Value Theorem for the Median 106
3. Bootstrap Theory for the Median 109
4. The Percentile Method 110
5. Percentile Method for the Median 113
6. Bayesian Justification of the Percentile Method 116
7. The Bias-Corrected Percentile Method 118
8. Typical Value Theory and the Percentile Method 121
9. The Percentile Method for M-Estimates 125
10. Bootstrap-t and Tilting 126

References 132

THE JACKKNIFE, THE BOOTSTRAP, AND OTHER RESAMPLING PLANS

Bradley Efron

I. INTRODUCTION

Our goal is to understand a collection of ideas concerning the nonparametric estimation of bias, variance, and more general measures of error. Historically the subject begins with the Quenouille-Tukey jackknife, which is where we will begin also. In fact it would be more logical to begin with the bootstrap, which most clearly exposes the simple idea underlying all of these methods. (And in fact underlies many common parametric methods as well, such as Fisher's information theory for assigning a standard error to a maximum likelihood estimate.) *Good* simple ideas, of which the jackknife is a prime example, are our most precious intellectual commodity, so there is no need to apologize for the easy mathematical level. The statistical ideas run deep, sometimes over our head at the current level of understanding. Chapter 10, on nonparametric confidence intervals, is particularly speculative in nature.

Some material has been deliberately omitted for these notes. This includes most of the detailed work on the jackknife, especially the asymptotic theory. Miller (1974) gives an excellent review of the subject.

From a traditional point of view, all of the methods discussed here are prodigious computational spendthrifts. We blithely ask the reader to consider techniques which require the usual statistical calculations

to be multiplied a thousand times over. None of this would have been feasible twenty five years ago, before the era of cheap and fast computation. An important theme of what follows is the substitution of computational power for theoretical analysis. The payoff, of course, is freedom from the constraints of traditional parametric theory, with its overreliance on a small set of standard models for which theoretical solutions are available. In the long run, understanding the limitations of the nonparametric approach should make clearer the virtues of parametric theory, and perhaps suggest useful compromises Some hints of this appear in Chapter 10, but so far as these are only hints and not a well developed point of view.

II. THE JACKKNIFE ESTIMATE OF BIAS

Quenouille (1949) invented a nonparametric estimate of bias, subsequently named the jackknife, which is the subject of this chapter. Suppose that we sample independent and identically distributed random quantities $X_1, X_2, X_3, \ldots, X_n \overset{iid}{\sim} F$, where F is an unknown probability distribution on some space $\mathcal{X}$. Often $\mathcal{X}$ will be the real line, but all of the methods discussed here allow $\mathcal{X}$ to be completely arbitrary. Having observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, we compute some statistic of interest, say

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n) .$$

We are interested in the bias of $\hat{\theta}$ for estimating a true quantity θ.

For now we concentrate on *functional statistics*: $\theta(F)$ is some real-valued parameter of interest, such as an expectation, a quantile, a correlation, etc., which we estimate by the statistic

$$\hat{\theta} = \theta(\hat{F}) , \tag{2.1}$$

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

$$\hat{F}\text{: mass } \frac{1}{n} \text{ at } x_1, x_2, \ldots, x_n . \tag{2.2}$$

Form (2.1) guarantees that $\hat{\theta}(x_1, x_2, \ldots, x_n)$ is invariant under permutations of the arguments, which we use below, and more importantly that the concept of bias is well-defined,

$$\text{Bias} \equiv E_F \theta(\hat{F}) - \theta(F) . \tag{2.3}$$

Here "E_F" indicates expectation under $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$. Three familiar examples of functional statistics are

Example 1. The expectation; $\mathcal{X} = \mathcal{R}^1$, the real line; $\theta(F) = \int_{\mathcal{X}} x dF = E_F X$; $\hat{\theta} = \int_{\mathcal{X}} x d\hat{F} = \frac{1}{n}\Sigma x_i = \bar{x}$, the average, or sample expectation.

Example 2. The correlation; $\mathcal{X} = \mathcal{R}^2$, the plane; $\theta(F)$ = Pearson's product-moment correlation coefficient, Cramer (1946), p. 265; $\hat{\theta} = \theta(\hat{F})$ = sample correlation coefficient.

Example 3. Ratio estimation; $\mathcal{X} = \mathcal{R}^{2+}$, the positive quadrant of the plane; denoting $X = (Y,Z)$, then $\theta(F) = E_F(Z)/E_F Y$, the ratio of expectations for the two coordinates; $\hat{\theta} = \theta(\hat{F}) = \bar{z}/\bar{y}$, the ratio of corresponding averages.

1. Quenouille's Bias Estimate. Quenouille's method is based on sequentially deleting points x_i, and recomputing $\hat{\theta}$. Removing point x_i from the data set gives a different empirical probability distribution,

$$\hat{F}_{(i)}: \text{ mass } \frac{1}{n-1} \text{ at } x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n , \tag{2.4}$$

and a corresponding recomputed value of the statistic,

$$\hat{\theta}_{(i)} = \theta(\hat{F}_{(i)}) = \hat{\theta}(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) . \tag{2.5}$$

Let

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)} . \tag{2.6}$$

Quenouille's estimate of bias is

$$\widehat{\text{BIAS}} \equiv (n-1)\,(\hat{\theta}_{(\cdot)} - \hat{\theta}) , \tag{2.7}$$

leading to the bias-corrected "jackknifed estimate" of θ,

$$\tilde{\theta} \equiv \hat{\theta} - \widehat{\text{BIAS}} = n\hat{\theta} - (n-1)\,\hat{\theta}_{(\cdot)} . \tag{2.8}$$

The usual rationale for $\widehat{\text{BIAS}}$ and $\tilde{\theta}$ goes as follows. If E_n denotes the expectation for sample size n, $E_n \equiv E_F\hat{\theta}(X_1, X_2, \ldots, X_n)$, then for many common statistics, including most maximum likelihood estimates,

$$E_n = \theta + \frac{a_1(\theta)}{n} + \frac{a_2(\theta)}{n^2} + \cdots , \tag{2.9}$$

where the functions $a_1(\theta)$, $a_2(\theta)$, ... do not depend upon n, see Schucany, Gray, and Owen (1971). Notice that

$$E_F\hat{\theta}_{(\cdot)} = E_{n-1} = \theta + \frac{a_1(\theta)}{n-1} + \frac{a_2(\theta)}{(n-1)^2} + \cdots$$

and so

$$\begin{aligned} E_F\tilde{\theta} &= nE_n - (n-1)E_{n-1} \qquad (2.10)\\ &= \theta - \frac{a_2(\theta)}{n(n-1)} + a_3(\theta)\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \cdots . \end{aligned}$$

We see that $\tilde{\theta}$ is biased $O\left(\frac{1}{n^2}\right)$ compared to $O\left(\frac{1}{n}\right)$ for the original estimator .

Example 1 continued. For $\hat{\theta} = \bar{x}$ we calculate $\hat{\theta}_{(\cdot)} = \bar{x} = \hat{\theta}$, and $\widehat{\text{BIAS}} = 0$. Of course $\hat{\theta}$ is unbiased for $\theta = E_F X$ in this case, so $\widehat{\text{BIAS}} = 0$ is the correct bias estimate.

Example 4. The variance; $\mathcal{X} = \mathcal{R}^1$, $\theta(F) = \int_{\mathcal{X}} (x - E_F X)^2 \, dF$; $\hat{\theta} = \sum_{i=1}^{n} (x_i - \bar{x})^2/n$. A simple calculation shows that

$$\widehat{\text{BIAS}} = -\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)} ,$$

yielding

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

the usual unbiased estimate of θ. In this case the expansion (2.9) is

$$E_n = \theta - \frac{\theta}{n},$$

so $a_1(\theta) = -\theta$, $a_j(\theta) = 0$ for $j > 1$. The motivating formula (2.10) shows why $\tilde{\theta}$ is exactly unbiased in this situation.

2. The Grouped Jackknife. Suppose $n = gh$ for integers g and h. We can remove observations in blocks of size h, e.g. first remove $x_1, x_2, \ldots, x_h$, second remove $x_{h+1}, x_{h+2}, \ldots, x_{2h}$, etc. Now define $\hat{\theta}_{(i)}$ as the statistic recomputed with the i-th block removed, and $\hat{\theta}_{(\cdot)} = \frac{1}{g}\Sigma\hat{\theta}_{(i)}$. Then

$$\tilde{\theta} = g\hat{\theta} - (g-1)\,\hat{\theta}_{(\cdot)}$$

also removes the first order of the bias term, as at (2.10). Quenouille's 1949 paper considered the "half-sample" case $g=2$.

If computationally feasible, it is preferable to define $\hat{\theta}_{(\cdot)} = \sum_{\tilde{i}} \hat{\theta}_{(\tilde{i})}/\binom{n}{h}$, where $\tilde{i}$ indicates a subset of size h removed from $\{1, 2, \ldots, n\}$, and $\sum_{\tilde{i}}$ is the sum over all such subsets. Then $\tilde{\theta} = g\hat{\theta} - (g-1)\,\hat{\theta}_{(\cdot)}$ has the same expectation and smaller variance than in the blocked case above, by a sufficiency argument. We consider only the ungrouped jackknife, $h=1$, in what follows, except for a few occasional remarks.

3. A Picture. Figure 2.1 shows E_n graphed versus $1/n$. The notation $\theta = E_\infty$ is based on (2.9), with $n=\infty$. Assuming perfect linearity of the bias in $1/n$ implies

$$\frac{E_n - E_\infty}{E_{n-1} - E_n} = \frac{1/n}{1/(n-1) - 1/n}$$

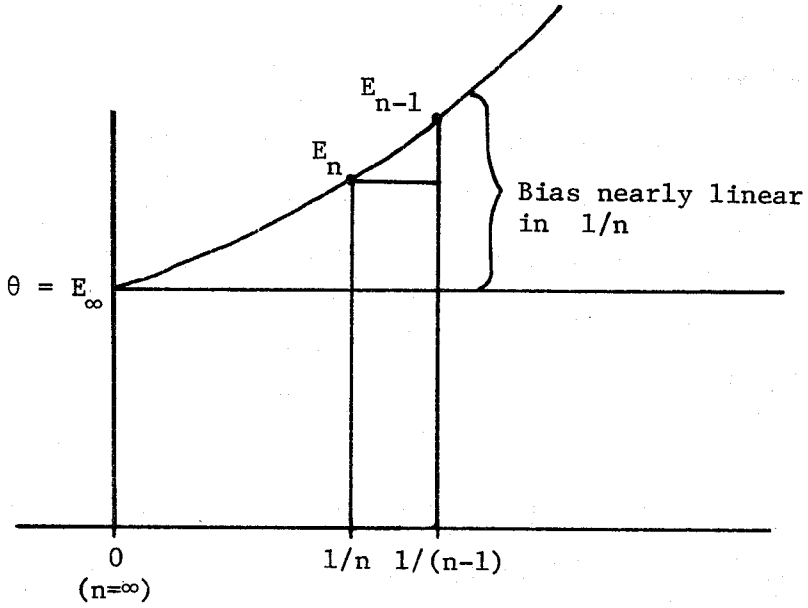


Figure 2.1. The expectation E_n as a function of $1/n$.

which gives

$$\text{Bias} = E_n - E_\infty = (n-1)\,(E_{n-1} - E_n)$$

and

$$\theta = E_\infty = nE_n - (n-1)E_{n-1}\,. \tag{2.11}$$

The jackknife formulas (2.7), (2.8) simply replace E_n and E_{n-1} on the right side of (2.11) by their unbiased estimates, $\hat{\theta}$ and $\hat{\theta}_{(\cdot)}$, respectively.

Removing two data points at a time and averaging the resulting recomputed values of $\hat{\theta}$ gives an unbiased estimate of E_{n-2}, say $\hat{\theta}_{(\cdot\cdot)}$. Looking at Figure 2.1, it seems reasonable to use quadratic extrapolation to predict $\theta = E_\infty$ from $\hat{\theta}$, $\hat{\theta}_{(\cdot)}$, $\hat{\theta}_{(\cdot\cdot)}$. A neat way of deriving all such higher-order bias correction formulas is given in Schucany, Gray, and Owen (1971).

In practice there has been little use made of higher-order bias correction. Even the first order bias correction (2.8) may add more to the mean square error in variance than it removes in bias^2. Hinkley (1978) discusses this effect for the case of the correlation coefficient. However it can still be interesting to compute $\widehat{\text{BIAS}}$, even if the bias correction isn't made, especially in conjunction with $\widehat{\text{VAR}}$, an estimate of variance such as that discussed in Section 3. Very often it turns out that $\widehat{\text{BIAS}}/\sqrt{\widehat{\text{VAR}}}$ is small, say $< 1/4$, in which case bias is probably not a serious issue.

4. Aitken Acceleration. The extrapolation method underlying the jackknife has a long history in numerical analysis, two examples being Aitken acceleration and Richardson extrapolation. (Here we discuss only the former.) The connection is as follows. Letting $E_n - E_\infty = \text{Bias}_n$, the bias for sample size n, simple algebra yields

$$E_\infty = \frac{E_n - (\text{ratio})\, E_{n-1}}{1 - (\text{ratio})}$$

where

$$\text{ratio} = \frac{\text{Bias}_n}{\text{Bias}_{n-1}} .$$

The approximation $\text{Bias}_n/\text{Bias}_{n-1} \doteq (n-1)/n$ gives (2.11), and hence the jackknife results (2.7), (2.8).

Now suppose we wish to approximate an infinite sum $S_\infty = \sum_{k=0}^{\infty} b_k$ on the basis of finite sums $S_n = \sum_{k=0}^{n} b_k$. Letting $B_n = \sum_{n+1}^{\infty} b_k$, the "bias" in S_n for approximating S, the same simple algebra yields

$$S_\infty = \frac{S_n - (\text{ratio})\, S_{n-1}}{1 - (\text{ratio})}$$

where

$$\text{ratio} = \frac{B_n}{B_{n-1}} .$$

Aitken acceleration replaces B_n/B_{n-1} with $b_n/b_{n-1} = (S_n - S_{n-1})/(S_{n-1} - S_{n-2})$, which is exactly right for a geometric series $b_k = cr^k$. The series transformation

$$\tilde{S}_n = \frac{S_n - \dfrac{S_n - S_{n-1}}{S_{n-1} - S_{n-2}} S_{n-1}}{1 - \dfrac{S_n - S_{n-1}}{S_{n-1} - S_{n-2}}} \tag{2.12}$$

can be applied repeatedly to speed up convergence. As an example, borrowed from Gray, Watkins, and Adams (1972), consider the series $n = 4 - 4/3 + 4/5 - 4/7 \ldots$. Taking only seven terms of the original series, and applying (2.12) three times, gives 3.14160, with deviation less than .00001 from $S_\infty = 3.141593 \ldots$:

n	$\tilde{S}_n$	$\tilde{\tilde{S}}_n$	$\tilde{\tilde{S}}_n^{(2)}$	$\tilde{\tilde{S}}_n^{(3)}$
0	4.00000			
1	2.66667			
2	3.46667	3.16667		
3	2.89524	3.13334		
4	3.33968	3.14524	3.14211	
5	2.97605	3.13968	3.14145	
6	3.28374	3.14271	3.14164	3.14160

Iterating (2.12) three times amounts to using a cubic extrapolation formula in Figure 2.1. This is more reasonable in a numerical analysis setting than in the noisy world of statistical estimation.

5. The Law School Data. Table 2.1 gives the average LSAT and GPA for the 1973 entering classes of 15 American law schools. (LSAT is a national test for prospective lawyers, GPA the undergraduate grade point average; see

Efron (1979b) for details.) The data is plotted in Figure 2.2. Consider Pearson's correlation coefficient, as in example 2. Denoting the statistic by $\hat{\rho}$, rather than $\hat{\theta}$, the value of the sample correlation coefficient for the 15 schools is $\hat{\rho} = .776$. The quantities $\hat{\rho}_{(i)} - \hat{\rho}$, also given in Table 2.1, yield $\widehat{\text{BIAS}} = 14(\hat{\rho}_{(\cdot)} - \hat{\rho}) = -.007$. As a point of comparison, the normal theory estimate of Bias is $-.011$, obtained by substituting $\hat{\rho}$ in formula 10.2, page 225, Johnson and Kotz. We will return to this example several times in succeeding chapters.

School #i	1	2	3	4	5	6	7	8
LSAT	576	635	558	578	666	580	555	661
GPA	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
$\hat{\rho}_{(i)}-\hat{\rho}$	.116	-.013	-.021	-.000	-.045	.004	.008	-.040

	9	10	11	12	13	14	15
LSAT	651	605	653	575	545	572	594
GPA	3.36	3.13	3.12	2.74	2.76	2.88	2.96
$\hat{\rho}_{(i)}-\hat{\rho}$	-.025	-.000	.042	.009	-.036	-.009	.003

Table 2.1. Average LSAT and GPA for the 1975 entering classes of 15 American law schools. Values of $\hat{\rho}_{(i)} - \hat{\rho}$ are used to compute $\widehat{\text{BIAS}}$ for the correlation coefficient.

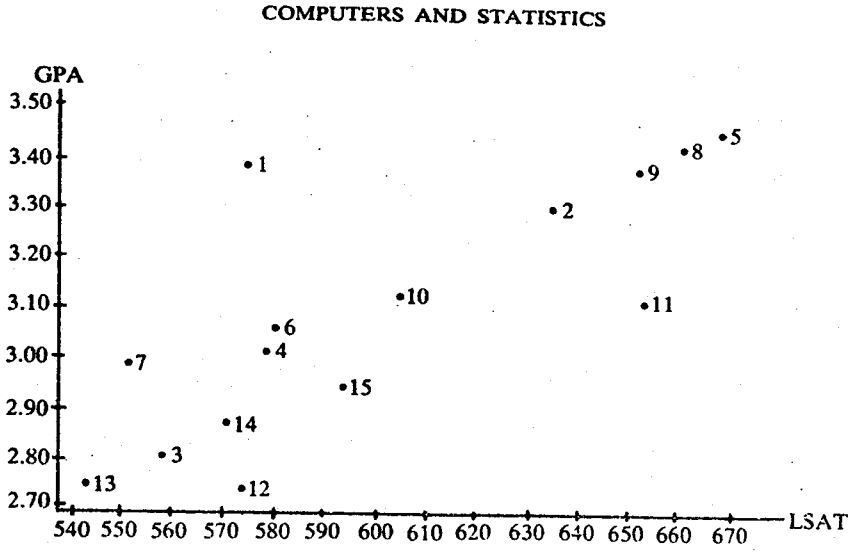


Figure 2.2. A plot of the law school data given in Table 2.2.

6. What Does $\widehat{\text{BIAS}}$ Really Estimate? The motivation for Quenouille's estimate $\widehat{\text{BIAS}} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$ based on (2.9) collapses under closer scrutiny. Notice that (2.9) can always be rewritten as

$$E_n = \theta + \frac{A_1(\theta)}{n+1} + \frac{A_2(\theta)}{(n+1)^2} \cdots$$

where $A_1(\theta) = a_1(\theta)$, $A_2(\theta) = a_1(\theta) + a_2(\theta)$, $A_3(\theta) = a_1(\theta) + 2a_2(\theta) + a_3(\theta)$, ... The same reasoning that gave (2.10) now gives

$$E_F\{(n+1)\ \hat{\theta} - n\hat{\theta}_{(\cdot)}\} = \theta - \frac{A_2(\theta)}{n(n+1)} + A_3(\theta)\left(\frac{1}{(n+1)^3} - \frac{1}{n^3}\right) \cdots .$$

This would suggest that $\widehat{\text{BIAS}} = n(\hat{\theta}_{(\cdot)} - \hat{\theta})$ is the correct formula for removing the $1/n$ bias term, not Quenouille's formula (2.7).

The real justification for Quenouille's formula is contained in Example 4. The statistic $\hat{\theta} = \Sigma(x_i - \bar{x})^2/n$ is an example of a *quadratic functional*: $\hat{\theta}$ is of the functional form (2.1), $\hat{\theta} = \theta(\hat{F})$, and $\hat{\theta}$ can be expressed as

$$\hat{\theta} = \mu^{(n)} + \frac{1}{n}\sum_{i=1}^{n} \alpha^{(n)}(x_i) + \frac{1}{n^2}\sum_{1\le i_1 < i_2 \le n} \beta(x_{i_1}, x_{i_2}) , \qquad (2.13)$$

i.e. in a form which involves the x_i one and two at a time, but in no higher-order interactions. Quadratic functionals, which are closely related to Hoeffding's (1948) U-statistics, are discussed in Chapter IV. The proof of the following theorem is given there.

Theorem 2.1. For a quadratic functional, $\widehat{\text{BIAS}} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$ is unbiased for estimating the true bias $E_F\theta(\hat{F}) - \theta(F)$.

It is easy to think of functional statistics $\hat{\theta} = \theta(\hat{F})$ for which $\widehat{\text{BIAS}}$ is useless or worse than useless. For example, let $\mathcal{X} = \mathbb{R}^1$, and $\theta(F) = 0$ if F has no discrete probability atoms, $\theta(F) = 1$ if it does. If F has no atoms, so $\theta = 0$, then $\hat{\theta} = \theta(\hat{F}) = 1$ has true bias 1, while $\widehat{\text{BIAS}} = 0$, for any sample size n. This points out that the concept (2.1) of a functional statistic is itself useless without some notion of continuity in the argument $\hat{F}$, see Chapter II of Huber (1974). Its only purpose in this chapter was to give an unambiguous meaning to the concept of bias. We will see what $\widehat{\text{BIAS}}$ actually estimates, whether or not the statistic is funtional, at the end of Chapter VI. Roughly

speaking, $\widehat{\text{BIAS}}$ is the true bias of $\hat{\theta}$ if F were actually equal to the observed $\hat{F}$. That is, $\widehat{\text{BIAS}}$ itself is (approximately) a functional statistic, the function being the bias of $\hat{\theta}$. All of this will be made clear in Chapter V when we discuss the bootstrap.

III. THE JACKKNIFE ESTIMATE OF VARIANCE

Tukey (1958) suggested how the recomputed statistics $\hat{\theta}_{(i)}$ could also provide a nonparametric estimate of variance. Let

$$\text{Var} = E_F[\hat{\theta}(X_1, X_2, \ldots, X_n) - E_F\hat{\theta}]^2 , \tag{3.1}$$

where as before E_F indicates expectation with $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$, F an unknown probability distribution on some space $\mathcal{X}$. (In general, "E_F" means that all random variables involved in the expectation are independently distributed according to F.) Tukey's formula for estimating Var is[†]

$$\widehat{\text{VAR}} = \frac{n-1}{n} \sum_{i=1}^{n} [\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}]^2 , \tag{3.2}$$

$\hat{\theta}_{(\cdot)} = \frac{1}{n} \Sigma\hat{\theta}_{(i)}$. We will often be more interested in standard deviations than variances, since the standard deviation relates directly to accuracy statements about $\hat{\theta}$, in which case we will use the notation

$$\text{Sd} = \sqrt{\text{Var}} , \qquad \widehat{\text{SD}} = \sqrt{\widehat{\text{VAR}}} . \tag{3.3}$$

Considerable effort has gone into verifying, and in some cases disverifying, the usefulness of $\widehat{\text{VAR}}$ as an estimate of Var, see Miller (1974). This chapter presents several examples typifying $\widehat{\text{VAR}}$'s successes and failures. The theoretical basis of (3.2) is discussed in Chapters IV, V, and VI.

[†]Some writers consider $\widehat{\text{VAR}}$ as an estimate of $\text{Var}(\tilde{\theta})$, rather than of $\text{Var}(\hat{\theta})$, but in fact it seems to be a better estimator of the latter, see Hinkley (1978). This will be our point of view.

1. The Expectation. As in example 1 of Chapter I, $\hat{\theta} = \bar{x} = \Sigma x_i/n$. Then

$$\hat{\theta}_{(i)} = \frac{n\hat{\theta}-x_i}{n-1}, \qquad \hat{\theta}_{(\cdot)} = \hat{\theta}, \qquad \hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} = \frac{\bar{x}-x_i}{n-1}$$

and so

$$\widehat{VAR} = \frac{\Sigma(x_i-\bar{x})^2}{n(n-1)},$$

the usual nonparametric estimate for the variance of an average $\bar{X}$. This is the motivating example behind Tukey's formula (3.2).

"*Pseudo-values*": Tukey called

$$\tilde{\theta}_i = \hat{\theta} + (n-1)\,(\hat{\theta} - \hat{\theta}_{(i)})$$

the "i-th pseudo-value". For a general statistic $\hat{\theta}$, the jackknife estimate $\tilde{\theta}$ equals $\Sigma\tilde{\theta}_i/n$, and $\widehat{VAR} = \Sigma(\tilde{\theta}_i-\tilde{\theta})^2/[\,n \cdot (n-1)]$. This makes the $\tilde{\theta}_i$ look like they're playing the same role as do the x_i in the case $\hat{\theta} = \bar{x}$. (Indeed $\tilde{\theta}_i = x_i$ when $\hat{\theta} = \bar{x}$.)

Unfortunately the analogy doesn't seem to go deep enough. Attempts to extract additional information from the $\tilde{\theta}_i$ values, beyond the estimate $\widehat{VAR}$, have been disappointing. For example Tukey's original suggestion was to use

$$\tilde{\theta} \pm t_\alpha^{n-1}\,\hat{SD} \tag{3.4}$$

as a $1-2\alpha$ confidence interval for θ, where t_α^{n-1} is the α upper percentile point of a t distribution with $n-1$ degrees of freedom. Verification of (3.4) as a legitimate confidence interval, as in Miller (1964), has been successful only in the asymptotic case $n\to\infty$, for which the "t" effect disappears, and where instead of (3.4) we are dealing with the

comparatively crude limiting normal theory. (Small sample nonparametric confidence intervals are discussed in Chapter X.) The pseudo-value terminology is slightly confusing, and will not be used in our discussion.

2. The Unbiased Estimate of Variance. Let $\mathcal{X} = \mathcal{R}^1$ and $\hat{\theta} = \sum_{i=1}^{n} (x_i-\bar{x})^2/(n-1)$. Define the k-th central moments of F and $\hat{F}$ to be, respectively,

$$\mu_k = E_F[X-E_F X]^k , \qquad \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} [x_i-\bar{x}]^k . \tag{3.5}$$

For this $\hat{\theta}$, (3.2) has the simple expression

$$\widehat{\mathrm{VAR}} = \frac{n^2}{(n-1)(n-2)^2}(\hat{\mu}_4-\hat{\mu}_2^2) \tag{3.6}$$

agreeing nicely with the true variance of $\hat{\theta}$,

$$\mathrm{Var} = \frac{n^2}{n(n-1)^2}(\mu_4-\mu_2^2) . \tag{3.7}$$

Hint: it helps to work with $y_i = (x_i-\bar{x})^2$ in deriving (3.6).

Notice that this $\hat{\theta}$ is not a functional statistic since, for example, the doubled data set $x_1, x_1, x_2, x_2, \ldots, x_n, x_n$ has the same value of $\hat{F}$ but a different value of $\hat{\theta}$. Variance is simpler than bias in that it refers only to the actual sample size n. Bias refers to sample size n and also to sample size ∞, i.e. the true θ. The concept of a functional statistic plays no role in (3.2). We only assume that $\hat{\theta}(x_1, x_2, \ldots, x_n)$ is symmetrically defined in its n arguments. Chapter VI shows that $\widehat{\mathrm{VAR}}$ is based on a simpler idea than is $\widehat{\mathrm{BIAS}}$, the difference being the use of a linear rather than quadratic extrapolation formula. In the author's experience, $\widehat{\mathrm{VAR}}$ tends to yield more dependable estimates than does $\widehat{\mathrm{BIAS}}$.

3. Trimmed Means. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the order statistics of a sample on the real line $\mathcal{X} = \mathbb{R}^1$, and let α be the proportion of points we "trim" from each end of the sample. The α trimmed mean is the average of the remaining $n(1-2\alpha)$ central order statistics. For instance if $n = 10$, $\alpha = .1$, then $\hat{\theta} = \sum_2^9 x_{(i)}/8$. (The two end points of the trimming region are counted partially, in the obvious way, if $n\alpha$ is not an integer.) If $(n-1)\alpha = g$, an integer, then

$$\widehat{VAR} = \frac{1}{n(n-1)(1-2\alpha)^2} \sum_{i=1}^{n} [x_{(W_i)} - \bar{x}_W]^2 \tag{3.8}$$

where

$$W_i = \begin{cases} g+1 & i \leq g+1 \\ i & g+1 < i < n-g \\ n-g & i \geq n-g \end{cases}$$

and $\bar{x}_W = \Sigma x_{(W_i)}/n$, see Huber (1974), p. 27. The letter "W" stands for "Winsorized" since (3.8) is proportional to the variance of what is called the Winsorized sample, i.e., the sample where the end order statistics are not trimmed off, but rather changed in value to $x_{(g+1)}$ or $x_{(n-g)}$. The expression for $\widehat{VAR}$ is only slightly more complicated if $(n-1)\alpha$ is not an integer.

Formula (3.8) has proved reasonably dependable for $\alpha \leq .25$, see Carroll (1979). Table 3.1 shows the results of a small Monte Carlo experiment: 200 trials, each consisting of a sample of size $n = 15$, $X_1, X_2, \ldots, X_{15} \overset{iid}{\sim} F$. Two cases were investigated, the normal case $F \sim h(0,1)$, and the negative exponential case $F \sim G_1$ (i.e. $f(x) = e^{-x}$ for $x > 0$, $f(x) = 0$ for $x \leq 0$). The trimming proportion was $\alpha = .25$.

	$F \sim \eta(0,1)$			$F \sim G_1$		
	Ave	Std Dev	Coeff Var	Ave	Std Dev	Coeff Var
Jackknife	.280	.084	.30	.224	.085	.38
Bootstrap, 200 bootstrap reps per trial	.287	.071	.25	.242	.078	.32
True Sd [minimum possible cv]	.286		[.19]	.232		[.27]

Table 3.1. Estimates of standard deviation for the 25% trimmed mean using the jackknife and the bootstrap: 200 trials of $X_1, X_2, \ldots, X_{15} \overset{iid}{\sim} F$. The averages and standard deviations of $\hat{SD}$ for the 200 trials show a moderate advantage for the bootstrap.

For $F \sim \eta(0,1)$, the 200 jackknife estimates of standard deviation, each essentially the square root of (3.8) with $n = 15$ and $\alpha = .25$, averaged .280 with sample standard deviation .084. Typically $\hat{SD}$ was between .200 and .400. The true standard deviation of $\hat{\theta}$ is .286 in this case, so the jackknife estimate is nearly unbiased, though quite variable, having coefficient of variation .084/.280 = .30.

As a point of comparison, Table 3.1 also gives summary statistics for the bootstrap estimate of standard deviation introduced in Chapter V. The bootstrap estimate is also nearly unbiased but with noticeably smaller variability from trial to trial. The figures in brackets show the minimum possible coefficient of variation, in each case, for any estimate of standard deviation which is scale invariant. In the normal case, for example, .19 is the coefficient of variation of $[\Sigma(x_i-\bar{x})^2/14]^{1/2}$.

4. The Sample Median. The trimmed mean with $\alpha \to .5$ is the sample median: $\hat{\theta} = x_{(m)}$ if $n = 2m-1$, $\hat{\theta} = (x_{(m)} + x_{(m+1)})/2$ if $n = 2m$. Formula (3.2) fails in this case. In the even case $n = 2m$, (3.2) gives

$$\widehat{\mathrm{VAR}} = \frac{n-1}{4}\,[x_{(m+1)} - x_{(m)}]^2 \,. \tag{3.9}$$

Standard asymptotic theory, Pyke (1965), shows that if F has a density function f, then

$$n\,\widehat{\mathrm{VAR}} \xrightarrow{\mathcal{L}} \frac{1}{4f^2(\theta)}\left[\frac{\chi_2^2}{2}\right]^2$$

as $n\to\infty$, where $f(\theta)$ is the density at the true median θ, $f(\theta)$ assumed > 0, and $[\chi_2^2/2]^2$ is a random variable with expectation 2, variance 20. The true variance of $\hat{\theta}$ goes to the limit

$$n\,\mathrm{Var} \to \frac{1}{4f^2(\theta)} \,,$$

see Kendall and Stuart (1958). In this case $\widehat{\mathrm{VAR}}$ is not even a consistent estimator of Var. An explanation of what goes wrong is given in Chapter VI. The bootstrap estimate of variance performs reasonably well for the sample median, see Chapter X.

5. Ratio Estimation. As in example 3 of Chapter 2, $\mathcal{X} = \mathcal{R}^{2+}$, $X = (Y,Z)$, but here we consider the statistic $\hat{\theta} = \log \bar{z}/\bar{y}$. Table 3.2 reports the results of a Monte Carlo experiment: 100 trials of (Y_1,Z_1), (Y_2,Z_2), ..., $(Y_{10},Z_{10}) \overset{\text{iid}}{\sim} F$. The two cases considered for F both had $Y \sim U(0,1)$, the uniform distribution on $[0,1]$; Z was taken independent of Y, in case 1 $Z \sim G_1$, in case 2 $Z \sim G_1^2/2$. The summary statistics for the 100 trials show that SD, the jackknife estimate of standard deviation, is

nearly unbiased for the true Sd. Once again the estimates $\hat{SD}$ are quite variable for trial to trial, perhaps not surprising given a sample size of only $n = 10$.

	$Y \sim U(0,1),\ Z \sim G_1$			$Y \sim U(0,1),\ Z \sim G_1^2/2$		
	Ave	Std Dev	Coeff Var	Ave	Std Dev	Coeff Var
Jackknife	.37	.11	.30	.70	.33	.47
Bootstrap, 1000 reps per trial	.37	.10	.27	.64	.23	.36
delta method	.35	.09	.26	.53	.14	.26
True Sd	.37			.67		

Table 3.2. Estimates of standard deviation for $\hat{\theta} = \log \bar{z}/\bar{y}$; 100 trials, sample size $n = 10$ for each trial. Summary statistics of $\hat{SD}$ for the 100 trials show that the jackknife is more variable than the bootstrap or the delta method. However the delta method is badly biased in the second case.

Table 3.2 also presents results for the bootstrap, Chapter V, and the delta method, Chapter VI. The delta method is best known of all the techniques we will discuss. In the present case it consists of approximating the Sd of $\hat{\theta} = \log \bar{z}/\bar{y}$ by the Sd of its first order Taylor series expansion, $\hat{\theta} \doteq \log(\mu_z/\mu_y) + (\bar{z}-\mu_z)/\mu_z - (\bar{y}-\mu_y)/\mu_y$, where $\mu_z = E_F Z$, $\mu_y = E_F Y$. The resulting approximation, $Sd(\hat{\theta}) \doteq [\mu_{yy}/\mu_y^2 + \mu_{zz}/\mu_z^2 - 2\mu_{yz}/\mu_y\mu_z]/n$, where $\mu_{yy} = E_F[Y - EF_Y]^2$, etc., is then estimated by substituting sample moments $\hat{\mu}_y$, $\hat{\mu}_z$, $\hat{\mu}_{yy}$, $\hat{\mu}_{zz}$, $\hat{\mu}_{yz}$, for the unknown true population

moments. In the present situation the delta method has the lowest coefficient of variation, but is badly biased downwards in the second case. The delta method is discussed further in Chapter VI.

6. Functions of the Expectation. Suppose $\mathcal{X} = \mathcal{R}^1$, $\hat{\theta} = g(\bar{x})$, where g is some nicely behaved function such as $\sin(\bar{x})$ or $1/(1+\bar{x})$. (For what follows we need that the derivative g' exists continuously.) Then a first order Taylor series expansion gives

$$\hat{\theta}_{(i)} = g\left(\frac{n\bar{x}-x_i}{n-1}\right) \doteq g(\bar{x}) + g'(\bar{x})\,\frac{\bar{x}-x_i}{n-1}\,,$$

so

$$\widehat{\mathrm{VAR}} \doteq \frac{n-1}{n}\,[g'(\bar{x})]^2\,\frac{\Sigma(x_i-\bar{x})^2}{(n-1)^2} \qquad (3.9)$$

$$= [g'(\bar{x})]^2\,\frac{\hat{\sigma}^2}{n}\,,$$

where $\hat{\sigma}^2 = \Sigma(x_i-\bar{x})^2/(n-1)$ is the usual unbiased estimate of variance.

The variance of $\hat{\theta} = g(\bar{x})$ is usually obtained by the delta method: $g(\bar{x}) \doteq g(\mu) + g'(\mu)(\bar{x}-\mu)$, where $\mu = E_F X$, so $\mathrm{Var}\{\hat{\theta}\} \doteq [g'(\mu)]^2\,\mathrm{Var}\{\bar{X}\}$. Estimating μ by $\bar{x}$ and $\mathrm{Var}\{\bar{X}\}$ by $\hat{\sigma}^2/n$ gives (3.9). We have shown that the delta method gives the same estimate of variance as the jackknife, if a linear approximation is used to simplify the latter. Chapter VI discusses the "infinitesimal jackknife", a variant of the jackknife which gives *exactly* the same estimates as the delta method, for both bias and variance, whenever the delta method applies.

7. The Law School Data. For the correlation coefficient $\hat{\rho}$ based on the law school data, Table 2.1 and Figure 2.2, we calculate $\widehat{\mathrm{SD}} = .142$.

This might be compared with the normal theory estimate $(1-\hat{\rho}^2)/\sqrt{n-3}$ $= (1 - .776^2)/\sqrt{12} = .115$, see Johnson and Kotz (1970), p. 229. A disturbing feature of $\widehat{VAR}$ can be seen in Table 2.1: 55% of $\Sigma[\hat{\rho}_{(i)} - \hat{\rho}_{(\cdot)}]^2$ comes from data point 1. $\widehat{VAR}$ is not robust in the case of $\hat{\rho}$, a point discussed in Hinkley (1978).

	$\hat{\rho}$			$\tanh^{-1}\hat{\rho}$		
	Ave	Std Dev	Coeff Var	Ave	Std Dev	Coeff Var
Jackknife	.223	.085	.38	.314	.090	.29
Bootstrap, 512 reps per trial	.206	.063	.31	.301	.062	.21
Delta Method	.175	.058	.33	.244	.052	.21
True Sd	.221			.299		

Table 3.3. Estimates of standard deviation for $\hat{\rho}$, the correlation coefficient, and $\tanh^{-1}\hat{\rho}$: 200 trials, $X_1, X_2, \ldots, X_{14} \overset{iid}{\sim} F$, F bivariate normal with true $\rho = .5$. The jackknife is more variable than the bootstrap or the delta method, but the latter is badly biased downward.

Table 3.3 reports the results of a Monte Carlo experiment: 200 trials of $X_1, X_2, \ldots, X_{14} \sim F$ bivariate normal, true $\rho = .50$. Two statistics were considered, $\hat{\rho}$ and Fisher's transformation $\tanh^{-1}\hat{\rho} = \frac{1}{2}\log(1+\hat{\rho})/(1-\hat{\rho})$. In this case the jackknife $\widehat{SD}$ is considerably more variable than the bootstrap $\widehat{SD}$.

8. Linear Regression. Consider the linear regression model

$$y_i = c_i\beta + \varepsilon_i , \qquad i=1, 2, \ldots, n ,$$

where $\varepsilon_i \overset{iid}{\sim} F$, F an unknown distribution on $\mathcal{R}^1$ having $E_F\varepsilon = 0$. Here c_i is a known $1\times p$ vector of covariates while β is a $p\times 1$ vector of unknown parameters. The statistic of interest is the least squares estimate of β,

$$\hat{\beta} = \underset{\sim}{G}^{-1}\underset{\sim}{C}'\underset{\sim}{y} \ , \tag{3.10}$$

$\underset{\sim}{y} = (y_1, y_2, \ldots, y_n)'$, $\underset{\sim}{C}' = (c_1', c_2', \ldots, c_n')$, $\underset{\sim}{G} = \underset{\sim}{C}'\underset{\sim}{C}$. We assume that the $p\times n$ matrix $\underset{\sim}{C}$ is nonsingular, so that the $p\times p$ matrix $\underset{\sim}{G}$ has an inverse.

The usual estimate of $\mathrm{Cov}(\hat{\beta})$, the covariance matrix of $\hat{\beta}$, is

$$\hat{\sigma}^2 \underset{\sim}{G}^{-1} \ , \qquad \hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/(n-p) \ , \tag{3.11}$$

$\hat{\varepsilon}_i$ being the estimated residual $y_i - c_i\hat{\beta}$. We don't need the jackknife in this situation, but it is interesting to compare $\widehat{\mathrm{VAR}}$ with (3.11).

The statistic $\hat{\beta}$ is not a symmetric function of $y_1, y_2, \ldots, y_n$, but it is symmetrically defined in terms of the vectors (c_1,y_1), (c_2,y_2), $\ldots$, (c_n,y_n). What we have called x_i before is now the $p+1$ vector (c_i,y_i). Let $\hat{\beta}_{(i)}$ be the statistic (3.10) computed with (c_i,y_i) removed, which turns out to be

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{\underset{\sim}{G}^{-1}c_i\hat{\varepsilon}_i}{1 - c_i\underset{\sim}{G}^{-1}c_i'} \ , \tag{3.12}$$

see Miller (1974b), Hinkley (1977). The multivariate version of Tukey's formula (3.2) is

$$\widehat{\mathrm{COV}} = \frac{n-1}{n} \sum_{i=1}^{n} [\hat{\beta}_{(i)} - \hat{\beta}_{(\cdot)}] [\hat{\beta}_{(i)} - \hat{\beta}_{(\cdot)}]' \qquad (3.13)$$

$$\doteq \frac{n-1}{n} \underset{\sim}{G}^{-1} [\Sigma c_i' c_i \hat{\varepsilon}_i^2] \underset{\sim}{G}^{-1} .$$

This last formula ignores the factor $1 - c_i \underset{\sim}{G}^{-1} c_i' = 1 - O(\frac{1}{n})$ in the denominator of (3.12), which turns out to be equivalent to using the infinitesimal jackknife - delta method.

If all the $\hat{\varepsilon}_i^2$ are identical in value then (3.13) is about the same as the standard answer (3.11), but otherwise the two formulas are quite different. We will see why when we apply the bootstrap to regression problems in Chapter VII.

IV. BIAS OF THE JACKKNIFE VARIANCE ESTIMATE

This section shows that the jackknife variance estimate tends to be conservative in the sense that its expectation is greater than the true variance. The actual statement of the main theorem given below is necessarily somewhat different, but all of our Monte Carlo results, for example Tables 3.1 - 3.3, confirm that $\widehat{\mathrm{VAR}}$ is, if anything, biased moderately upward[†]. This contrasts with the delta method, which we have seen to be capable of severe downward biases.

The material of this chapter is somewhat technical, and can be skipped by readers anxious to get on with the main story. On the other hand it is nice to have a precise result in the midst of so much approximation and heuristic reasoning. A fuller account of these results is given in Efron and Stein (1981). Section 4.3, concerning influence functions, is referred to in Chapter VI.

Once again let $\hat{\theta}(X_1, X_2, \ldots, X_n)$ be a statistic symmetrically defined in its n arguments, these being an i.i.d. sample from an unknown distribution F on an arbitrary space $\mathcal{X}$, $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$. In order to use the jackknife formula (3.2), it is necessary that $\hat{\theta}$ also be defined for sample size $n-1$. Let Var_n be the variance of $\hat{\theta}(X_1, X_2, \ldots, X_n)$, Var_{n-1} the variance of $\hat{\theta}(X_1, X_2, \ldots, X_{n-1})$, and define

$$\widetilde{\mathrm{VAR}} = \sum_{i=1}^{n} [\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}]^2 \,. \tag{4.1}$$

[†]Notice that in Table 3.1, F normal, the 200 jackknife estimates $\widehat{\mathrm{VAR}}$ averaged .0854 $(= .280^2 + (.199/200)\ .084^2)$ compared to the true variance .0816 $(= .286^2)$.

It is useful to think of VAR, formula (3.2), as estimating the true variance Var_n in two distinct steps:

(i) A direct estimate of Var_{n-1}, namely $\widetilde{\mathrm{VAR}}$, and

(ii) a sample size modification to go from $n-1$ to n,

$$\widehat{\mathrm{VAR}} = \frac{n-1}{n}\,\widetilde{\mathrm{VAR}} \ . \tag{4.2}$$

The main result of this chapter concerns (i). We show that

$$E_F\,\widetilde{\mathrm{VAR}} \geq \mathrm{Var}_{n-1} \ . \tag{4.3}$$

$\widetilde{\mathrm{VAR}}$ always overestimates Var_{n-1} in expectation. We also discuss, briefly, the sample size modification (ii), which is based on the fact that for many familiar statistics

$$\mathrm{Var}_n = \frac{n-1}{n}\,\mathrm{Var}_{n-1} + O\left(\frac{1}{n^3}\right) \ , \tag{4.4}$$

the $O(1/n^3)$ term being negligible compared to the modification $\frac{1}{n}\,\mathrm{Var}_{n-1} = O(1/n^2)$. For $\hat{\theta} = \bar{x}$, $\mathrm{Var}_n = \frac{n-1}{n}\,\mathrm{Var}_{n-1}$ exactly.

1. ANOVA Decomposition of $\hat{\theta}$. This will be the main tool in proving (4.3). It is a decomposition of $\hat{\theta}(X_1, X_2, \ldots, X_n)$ based on the ANOVA decomposition of a complete n-way table. Assume that $E_F\hat{\theta}^2 < \infty$, and define

$$\mu = E_F\hat{\theta} \tag{4.5}$$

$$\alpha_i = \alpha(X_i) = n[E_F\{\hat{\theta}|X_i\} - \mu]$$

$$\beta_{ii'} = \beta(X_i,X_{i'}) = n^2[E_F\{\hat{\theta}|X_i,X_{i'}\} - E_F\{\hat{\theta}|X_i\} - E_F\{\hat{\theta}|X_{i'}\} + \mu] \ ,$$

etc., the last definition being

$$\eta_{123\ldots n} = \eta(X_1, X_2, X_3, \ldots, X_n)$$

$$= n^n[\hat{\theta} - E_F\{\hat{\theta}|X_1, X_2, \ldots, X_{n-1}\} - E_F\{\hat{\theta}|X_1, X_2, \ldots, X_{n-2}, X_n\}$$

$$\ldots + (-1)^n\mu] \ .$$

In the usual ANOVA terminology μ is the grand mean, $n\alpha_i$ is the i-th main effect, $n^2\beta_{ii'}$ is the ii'-th interaction, etc. The reason for multiplying by powers of n is discussed below.

There are 2^n random variables $\mu, \alpha_i, \beta_{ii'}, \ldots, \eta_{12\ldots n}$ defined above, corresponding to the 2^n possible subsets of $\{1, 2, \ldots, n\}$. They have three salient properties:

Property 1. Each random variable is a function only of the X_i indicated by its subscripts (e.g. β_{37} is a function of X_3 and X_7).

Property 2. Each random variable has conditional expectation 0, when conditioned upon all but one of its defining X_i (e.g. $E_F\alpha_1 = 0$, $E_F\{\beta_{12}|X_2\} = 0$).

Property 3. $\hat{\theta}$ decomposes into a sum of $\mu, \alpha_i, \beta_{ii'}, \ldots, \eta_{123\ldots n}$ as follows,

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \mu + \frac{1}{n}\sum_i \alpha_i + \frac{1}{n^2}\sum_{i<i'} \beta_{ii'} + \ldots + \frac{1}{n^n}\eta_{123\ldots n} \ . \quad (4.6)$$

("$i < i'$" is short for $1 \leq i < i' \leq n$, etc.)

The proofs of properties 2 and 3 are essentially the same as those for the standard ANOVA decomposition of an n-way table, see Scheffe (1959), Section 4.5, and are given in Efron and Stein (1981). Notice that property 2 implies

Property 2'. The 2^n random variables μ, α_i, $\beta_{ii'}$, ..., $\eta_{123\ldots n}$ are mutually uncorrelated (e.g. $E_F\alpha_1\alpha_2 = 0$, $E_F\alpha_1\beta_{12} = 0$).

2. Proof of the Main Result. First of all notice that the main result (4.3) concerns only samples of size n-1. It is really a statement about $\hat{\theta}(X_1, X_2, \ldots, X_{n-1})$, and there is no need to define the original statistic $\hat{\theta}(X_1, X_2, \ldots, X_n)$. We will need both $\hat{\theta}(X_1, X_2, \ldots, X_{n-1})$ and $\hat{\theta}(X_1, X_2, \ldots, X_n)$ when we discuss (4.4).

Consider decomposition (4.6) for $\hat{\theta}(X_1, X_2, \ldots, X_{n-1})$, and define

$$\sigma_\alpha^2 = \mathrm{Var}_F\, \alpha_i\ , \qquad \sigma_\beta^2 = \mathrm{Var}_F\, \beta_{ii'}\ , \quad \text{etc.} \tag{4.7}$$

Using Property 2', we can immediately calculate $\mathrm{Var}\{\hat{\theta}(X_1, X_2, \ldots, X_n\} = \mathrm{Var}_{n-1}$ simply by counting the terms in (4.6),

$$\mathrm{Var}_{n-1} = \frac{\sigma_\alpha^2}{n-1} + \binom{n-2}{1}\frac{\sigma_\beta^2}{2(n-1)^3} + \binom{n-2}{2}\frac{\sigma_\alpha^2}{3(n-1)^5}\ . \tag{4.8}$$

(Remember we are applying (4.6) to $\hat{\theta}(X_1, \ldots, X_{n-1})$, so, for example, the α term is $\sum_{i=1}^{n-1}\alpha_i/(n-1)$.) There are (n-2) terms on the right side of (4.8).

The statistic $\hat{\theta}(X_1, X_2, \ldots, X_{n-1})$ is what we have previously called $\hat{\theta}_{(n)}$. We can also apply (4.6) to $\hat{\theta}_{(n-1)} = \hat{\theta}(X_1, X_2, \ldots, X_{n-2}, X_n)$, take the difference, and calculate

$$E_F[\hat{\theta}_{(n)} - \hat{\theta}_{(n-1)}]^2 = 2\left[\frac{\sigma_\alpha^2}{(n-1)^2} + \binom{n-2}{1}\frac{\sigma_\beta^2}{(n-1)^2} + \ldots\right]\ . \tag{4.9}$$

Since $\widetilde{\mathrm{VAR}} = \sum_{i=1}^{n}[\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}]^2 = \frac{1}{n}\sum_{1\le i<i'\le n}[\hat{\theta}_{(i)} - \hat{\theta}_{(i')}]^2$, and all the terms in this last experssion have expected value (4.9), we get

$$E_F \widetilde{VAR} = \frac{\sigma_\alpha^2}{n-1} + \binom{n-2}{1} \frac{\sigma_\beta^2}{(n-1)^3} + \binom{n-2}{2} \frac{\sigma_\gamma^2}{(n-1)^5} + \cdots . \tag{4.10}$$

Comparing (4.8) with (4.10) gives

Theorem 4.1. $E_F \widetilde{VAR}$ exceeds Var_{n-1} by an amount

$$E_F \widetilde{VAR} - Var_{n-1} = \binom{n-2}{1} \frac{\sigma_\beta^2}{2(n-1)^3} + 2\binom{n-2}{2} \frac{\sigma_\gamma^2}{3(n-1)^5} + \cdots , \tag{4.11}$$

there being $n-2$ terms on the right side of (4.11).

Several comments are pretinent: (1) All the terms on the right side of (4.11) are positive, so this proves the main result (4.3). (2) A *linear functional* is by definition a statistic of the form

$$\hat{\theta} = \mu + \frac{1}{n} \sum_i \alpha(X_i) . \tag{4.12}$$

If $\hat{\theta}(X_1, X_2, \ldots, X_{n-1})$ is linear, the right side of (4.11) is zero, and $E_F \widetilde{VAR} = Var_{n-1}$, otherwise $E_F \widetilde{VAR} > Var_{n-1}$. (3) Comparing (4.10) with (4.8) shows that $E_F \widetilde{VAR}$ doubles the quadratic (σ_β^2) term in Var_{n-1}, triples the cubic term, etc. If $\widetilde{VAR}$ is not to be badly biased upwards, then most of the variance of $\hat{\theta}(X_1, \ldots, X_{n-1})$ must be contained in the linear term $\sum_i \alpha_i/(n-1)$. This seems usually to be the case. The theory of influence functions discussed below shows that it is asymptotically true under suitable regularity conditions on $\hat{\theta}$. (4) Tables 3.1 - 3.3 allow us to estimate the proportion of the variance of $\hat{\theta}(X_1, X_2, \ldots, X_{n-1})$ contained in the linear term. The estimated proportions are

Table 3.1: 96%, 93%

Table 3.2: 91%, 67%

Table 3.3: 83%, 81%

(For example, .96 = 1 - (.0854 - .0816)/.0816, see the footnote at the beginning of this chapter.) (5) A version of (4.3) applicable for $S(X_1, X_2, \ldots, X_{n-1})$ any function, symmetric or not, of any n-1 independent arguments, identically distributed or not, is given in Efron and Stein (1981). (6) The terms σ_α^2, σ_β^2, etc. depend on the sample size, see the discussion of quadratic functionals below. (7) Steele (1980) has used (4.3) in the proof of certain conjectures concerning subadditive functionals in the plane.

3. Influence Functions. The influence function IF(x) of a functional statistic $\hat{\theta} = \theta(\hat{F})$ evaluated at the true probability distribution F, is defined as

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

where δ_x is a unit probability mass on the point x. Under reasonable conditions, see Huber (1974) or Hampel (1974),

$$\theta(\hat{F}) = \theta(F) + \frac{1}{n}\sum_{i=1}^{n} IF(X_i) + o_p(\tfrac{1}{n}) . \qquad (4.13)$$

This formula looks like the beginning of (4.6).

As a matter of fact, the function $\alpha(x)$ converges to IF(x) as $n \to \infty$, again under suitable regularity conditions. That is the reason for multiplying by n in (4.5). Likewise $\beta(x,x')$ converges to the second order influence function, etc. In a sense $\alpha(x)$ deserves to be called the finite sample influence funct on, a point discussed by Mallows (1974).

4. Quadratic Functionals. Consider the statistic $\hat{\theta}(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} [X_i - \bar{X}]^2/n$, where the X_i are i.i.d. with expectation ξ and variance σ^2, $X_i \overset{iid}{\sim} (\xi, \sigma^2)$. Expansion (4.6) can be evaluated explicitly in this case:

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \mu^{(n)} + \frac{1}{n}\sum_i \alpha^{(n)}(X_i) + \frac{1}{n^2}\sum_{i<i'} \beta(X_i, X_n') \qquad (4.14)$$

where

$$\mu^{(n)} = \frac{n-1}{n}\sigma^2 , \quad \alpha^{(n)}(x) = \frac{n-1}{n}[(x-\xi)^2 + \sigma^2] , \qquad (4.15)$$

$$\beta(x,x') = -2(x-\xi)(x'-\xi) .$$

This is an example of a *quadratic functional:* the ANOVA decomposition (4.14) terminates at the quadratic term, and $\hat{\theta}$ is a functional statistic, $\hat{\theta} = \theta(\hat{F})$.

Notice in (4.15) that $\beta(\cdot,\cdot)$ does not depend upon n, while $\mu^{(n)}$ and $\alpha^{(n)}(\cdot)$ both involve $1/n$ terms. ($\lim_{n\to\infty} \alpha^{(n)}(x) = \alpha^{(\infty)}(x) = (x-\xi)^2 + \sigma^2$, the influence function of $\hat{\theta}$, as mentioned above; $\lim_{n\to\infty} \mu^{(n)} = \mu^{(\infty)} = \theta(F)$.) This turns out to be true in general: $\hat{\theta}$ is a quadratic functional if and only if it can be written in form (4.14) with

$$\mu^{(n)} = \mu^{(\infty)} + \frac{E_F\beta(X,X)}{2n}$$

and (4.16)

$$\alpha^{(n)}(x) = \alpha^{(\infty)}(x) + \frac{\beta(x,x) - E_F\beta(X,X)}{2n} .$$

See Efron and Stein (1981).

Quadratic functionals are the simplest nonlinear functional statistics, and as such they can be used to understand the problems caused by nonlinearity, as demonstrated at the end of this chapter. Theorem 2.1 shows that they justify the jackknife estimate of bias. We can now prove theorem 2.1: from (4.14) we get

$$\hat{\theta}_{(\cdot)} = \mu^{(n-1)} + \frac{1}{n}\sum_{1\le i\le n} \alpha_i^{(n-1)} + \frac{1}{n^2}\frac{n(n-2)}{(n-1)^2}\sum_{1\le i<i'\le n} \beta_{ii'} . \qquad (4.17)$$

Define

$$\Delta_i = \Delta(X_i) = \frac{\beta(X_i,X_i)}{n}, \qquad E_F\Delta = E_F\Delta(X) . \tag{4.18}$$

Then (4.16) can be rewritten as

$$\mu^{(n-1)} - \mu^{(n)} = \frac{E_F\Delta}{n-1}, \qquad \alpha^{(n-1)}(x) - \alpha^{(n)}(x) = \frac{\Delta(x) - E_F\Delta}{n(n-1)} . \tag{4.19}$$

Subtracting (4.14) from (4.17) gives

$$\widehat{\text{BIAS}} = (n-1)\ (\hat{\theta}_{(\cdot)} - \hat{\theta}) = (n-1)[\mu^{(n-1)} - \mu^{(n)}] + \frac{n-1}{n}\sum_i [\alpha_i^{(n-1)} - \alpha_i^{(n)}]$$

$$+ \frac{1}{n^2}\left[\frac{n(n-2)}{n-1} - (n-1)\right] \sum_{i<i'} \beta_{ii'}$$

$$= \frac{E\Delta}{n} + \frac{1}{n}\sum_i \frac{\Delta_i - E_F}{n} - \frac{1}{n^2(n-1)^2} \sum_{i<i'} \beta_{ii'} .$$

Taking the expectation of this last expression gives

$$E_F\ \widehat{\text{BIAS}} = \frac{E_F\Delta}{n}$$

since $\Delta_i - E_F\Delta$ and $\beta_{ii'}$ all have expectation zero. However, the first equation in (4.16) shows that

$$\frac{E_F\Delta}{n} = \mu^{(n)} - \mu^{(\infty)} = E_F\theta(\hat{F}) - \theta(F) ,$$

which is the statement of Theorem (2.1).

5. Sample Size Modification. It is *not* always true that $E_F\ \widehat{\text{VAR}} \geq \text{Var}_n$. Arbitrarily bad counterexamples can be constructed, which is not surprising since $\widehat{\text{VAR}}$ is defined entirely in terms of samples of size $n-1$, while Var_n is the variance for sample size n. However for many reasonable classes of

statistics we do have $E_F \widehat{VAR} \geq Var_n$, either asymptotically or for all n. Three examples are discussed in Efron and Stein (1981):

(1) *U Statistics*. A U statistic is a statistic of the form

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \sum_{i_1<i_2<\ldots<i_k} g(X_{i_1}, X_{i_2}, \ldots, X_{i_k})/\binom{n}{k} ,$$

k some fixed integer and g some fixed symmetric function of k arguments. Hoeffding (1948) showed that for $n-1 \geq k$, the smallest possible sample size, U statistics satisfy $\frac{n-1}{n} Var_{n-1} \geq Var_n$. Combining this with Theorem (4.1) and definition (4.2) gives

$$E_F \widehat{VAR} = \frac{n-1}{n} E_F \widetilde{VAR} \geq \frac{n-1}{n} Var_{n-1} \geq Var_n ,$$

the desired result. Note: Hoeffding's important paper uses what we have called the ANOVA decomposition, though in rearranged form.

(2) *Von Mises Series*. This is a term coined by C. Mallows for statistics of the form

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \mu + \frac{1}{n}\sum_i \alpha(X_i) + \frac{1}{n^2}\sum_{i<i'} \beta(X_i, X_{i'}) + \ldots$$

$$+ \frac{1}{n^k}\sum_{i_1<i_2<\ldots<i_k} \phi(X_{i_1}, X_{i_2}, \ldots, X_{i_k}) .$$

Here k is a fixed integer, μ, α, β, ..., ϕ fixed functions not depending on n, and $n \geq k$. A Von Mises series is not quite a U statistic. (It becomes one if we divide by n, $n(n-1)$, ... instead of n, n^2, ... in the definition.) Efron and Stein (1981) show that in this case it is *not* always true that $\frac{n-1}{n} Var_{n-1} \geq Var_n$, but it is still true that $E_F \widehat{VAR} \geq Var_n$, the desired result.

(3) *Quadratic Functionals*. For a quadratic functional,

$$E_F \widehat{VAR} = Var_n + \frac{1}{n(n-1)} \left\{ \frac{n^3-n^2-3n+1}{n^3-n^2} \frac{\sigma_\beta^2}{2} + \frac{2\sigma_{\alpha\Delta}}{n} + \frac{\sigma_\Delta^2}{n^2(n-1)} \right\} , \qquad (4.20)$$

where $\sigma_{\alpha\Delta} = E_F\alpha(X)\Delta(X)$, $\sigma_\Delta^2 = E_F[\Delta(X) - E_F\Delta]^2$. By constructing examples with $\sigma_{\alpha\Delta}$ sufficiently negative, we can force $E_F \widehat{VAR} < Var_n$ for small n. For n large, a sufficient condition being $n > -4\sigma_{\alpha\Delta}/\sigma_\beta^2$, we again have $E_F \widehat{VAR} \geq Var_n$.

V. THE BOOTSTRAP

The bootstrap, Efron (1979), is conceptually the simplest of all the techniques considered here. We begin our discussion with the bootstrap estimate of standard deviation, which performed well in Tables 3.1 - 3.3, and then go on to more general problems. The connection with the jackknife is made in Chapter VI.

Given a statistic $\hat{\theta}(X_1, X_2, \ldots, X_n)$ defined symmetrically in $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, write the standard deviation of $\hat{\theta}$ as

$$Sd = \sigma(F, n, \hat{\theta}) = \sigma(F) \ . \tag{5.1}$$

This last notation emphasizes that given the sample size n and the form of the statistic $\hat{\theta}(\cdot, \cdot, \cdot, \cdot)$, the standard deviation is a function of the unknown probability distribution F. The bootstrap estimate of standard deviation is simply $\sigma(\cdot)$ evaluated at $F=\hat{F}$,

$$\hat{SD} = \sigma(\hat{F}) \ . \tag{5.2}$$

Since $\hat{F}$ is the nonparametric maximum likelihood estimate of $\hat{F}$, another way to say (5.2) is that $\hat{SD}$ is the nonparametric MLE of Sd.

Example 1. The Average. The sample space $\mathcal{X} = \mathcal{R}^1$ and $\hat{\theta}(X_1, X_2, \ldots, X_n) = \bar{X}$. In this case we know that the standard deviation of $\bar{X}$ is

$$\sigma(F) = [\mu_2/n]^{1/2} \ ,$$

where $\mu_2 = E_F[X - E_F X]^2$, the second central moment of F, as in (3.5). Therefore

$$\hat{SD} = \sigma(\hat{F}) = [\hat{\mu}_2/n]^{1/2} \ , \tag{5.3}$$

$\hat{\mu}_2 = \sum_{i=1}^{n} [x_i - \bar{x}]^2/n$ being the second central moment of $\hat{F}$, i.e. the sample value of μ_2. This is not quite the usual estimate of Sd since it uses the maximum likelihood rather than the unbiased estimate of μ_2. The bootstrap variance estimate $\widehat{VAR} = \hat{\mu}_2/n$ is biased downward,

$$E_F \widehat{VAR} = E_F \hat{\mu}_2/n = \frac{n-1}{n}\frac{\mu_2}{n} = \frac{n-1}{n} \mathrm{Var}\{\bar{X}\} \ .$$

We could rescale (5.2) to make VAR unbiased in the case $\hat{\theta}=\bar{X}$, i.e. define $\hat{SD} = [n/(n-1)]^{1/2} \sigma(\hat{F})$, but this doesn't seem to give better Sd estimates. The jackknife estimate of standard deviation *is* rescaled in this way, as will be made clear in Section 6.

1. Monte Carlo Evaluation of $\hat{SD}$. Usually the function $\sigma(F)$ cannot be written down explicitly. In order to carry out the calculation of $\hat{SD}$, (5.2), it is then necessary to use a Monte Carlo algorithm.

1. Fit the nonparametric MLE of F,

$$\hat{F}\text{: mass } \frac{1}{n} \text{ at } x_i \ , \quad i=1, 2, \ldots, n \ . \tag{5.4}$$

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

$$X_1^*, X_2^*, \ldots, X_n^* \overset{iid}{\sim} \hat{F} \ , \tag{5.5}$$

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

3. Independently repeat step (2) a large number "B" of times, obtaining "bootstrap replications" $\theta^{*1}, \theta^{*2}, \ldots, \theta^{*B}$, and calculate

$$\hat{SD} = \left\{ \frac{\sum_{b=1}^{B} [\hat{\theta}^{*b} - \hat{\theta}^{*\cdot}]^2}{B-1} \right\}^{1/2} . \qquad (5.6)$$

The dot notation indicates averaging as before, $\hat{\theta}^{*\cdot} = \sum_{b=1}^{B} \hat{\theta}^{*b}/B$.

If we could let $B \to \infty$ then (5.6) would exactly equal (5.2). In practice we have to stop the bootstrap process sooner or later, sooner being preferable in terms of computational cost. Tables 3.1, 3.2, 3.3 used $B = 200, 1000, 512$ respectively. These values were deliberately taken large, for investigating quantities other than $\hat{SD}$, and $B = 100$ performed almost as well in all three situations. In most cases there is no point in taking B so large that (5.6) agrees very closely with (5.2), since (5.2) itself will be highly variable for estimating the true Sd. This point will be discussed further as we go through the examples.

Example 2. Switzer's Adaptive Trimmed Mean. The charm of the jackknife and the bootstrap is that they can be applied to complicated situations where parametric modelling and/or theoretical analysis is hopeless. As a relatively simple "complicated situation", consider Switzer's adaptive trimmed mean $\hat{\theta}(x_1, x_2, \ldots, x_n)$, defined in Carroll (1979):

(i) given the data $x_1, x_2, \ldots, x_n$, compute the jackknife estimate of variance for the 5%, 10%, and 25% trimmed means, and

(ii) let $\hat{\theta}$ be the value of that trimmed mean corresponding to the minimum of the three variance estimates.

The results of a large Monte Carlo study are shown in Table 5.1. Two sample sizes, $n = 10, 20$, and three distributions, $F \sim h(0,1)$, G_1, $e^{h(0,1)}$, were investigated. $B = 200$ bootstrap replications were taken for each trial.† The bootstrap clearly outperforms the jackknife, except for the

†"Trial" always refers to a new selection of the original data $X_1, X_2, \ldots, X_n \sim F$, while "replication" refers to a selection of the bootstrap data $X_1^*, X_2^*, \ldots, X_n^* \sim \hat{F}$.

case $n = 10$, $F \sim e^{h(0,1)}$, for which both are ineffective. The bootstrap results are surprisingly close to the theoretical optimum for a scale invariant Sd estimator, assuming full knowledge of the parametric family, when $F \sim h(0,1)$ and $F \sim G_1$.

		Sample Size n = 10			Sample Size n = 20		
		$F \sim h(0,1)$	G_1	$e^{h(0,1)}$	$h(0,1)$	G_1	$e^{h(0,1)}$
Jackknife $\hat{SD}$	Ave	.327	.296	.421	.236	.234	.324
	Std Dev	.127	.173	.335	.070	.143	.228
	[Coeff Var]	[.39]	[.58]	[.80]	[.30]	[.61]	[.70]
Bootstrap $\hat{SD}$, B = 200	Ave	.328	.310	.541	.236	.222	.339
	Std Dev	.081	.123	.310	.047	.072	.142
	[Coeff Var]	[.25]	[.40]	[.57]	[.20]	[.32]	[.42]
True Sd		.336	.306	.483	.224	.222	.317
[Min possible CV]		[.24]	[.33]		[.16]	[.23]	
No. of Trials		1000	3000	1000	1000	3000	1000

Table 5.1. Estimates of standard deviation for Switzer's adaptive trimmed mean using the jackknife and the bootstrap. The minimum possible Coeff of Variation for a scale invariant estimate of standard deviation, assuming full knowledge of the parametric family, is shown for $F \sim h(0,1)$ and G_1.

Example 3. The Law School Data. Again referring to Table 2.1 and Figure 2.2, B = 1000 bootstrap replications of the correlation coefficient were generated. Each of the 1000 bootstrap camples consisted of drawing 15 points *with* replacement from the 15 original data points shown in Figure

2.2, the corresponding bootstrap replication being the correlation coefficient of the resampled points. A typical bootstrap sample might consist of law school 1 twice, law school 2 zero times, law school 3 once, etc. (Notice that the expected proportion of points in the original sample absent from the bootstrap sample is $(1 - 1/15)^{15} = .36 \doteq e^{-1}$.)

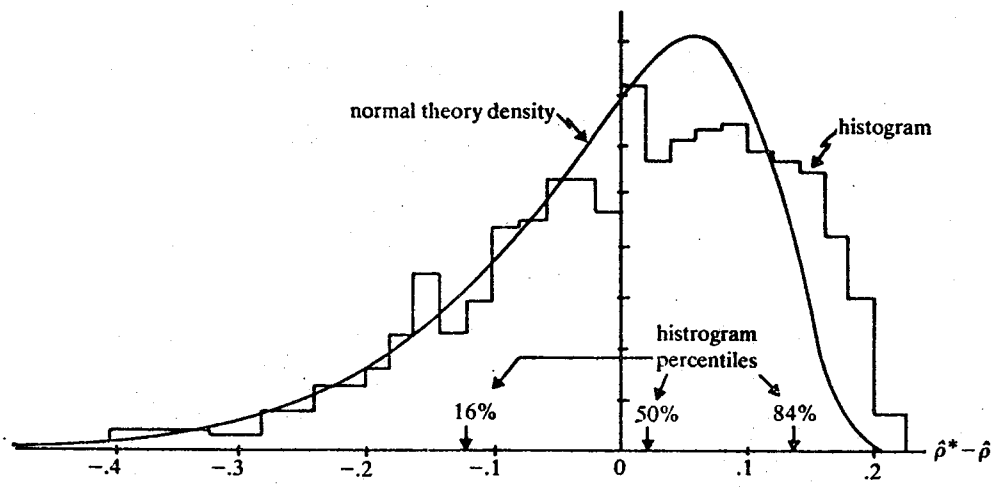


Figure 5.1. Histogram of 1000 bootstrap replications of $\hat{\rho}^* - \hat{\rho}$ for the law school data. The smooth curve is the normal theory density of $\hat{\rho}^*$, centered at $\hat{\rho}$, when the true correlation is $\hat{\rho} = .776$.

The bootstrap estimate of $\widehat{SD}$, (5.6), equaled .127, intermediate between the normal theory estimate .115 and the jackknife estimate .142. Figure 5.1 displays the histogram of the 1000 differences $\hat{\rho}^{*b} - \hat{\rho}$. Also shown is the normal theory density of $\hat{\rho}^*$, centered at $\hat{\rho}$, if the true

correlation coefficient is $\rho = \hat{\rho} = .776$, Johnson and Kotz (1970), p. 222. The similarity between the histogram and the density curve suggest that the bootstrap replications may contain information beyond that used in calculating (5.6). We consider this point in Chapter X where we try to construct nonparametric confidence intervals.

2. Parametric Bootstrap. Fisher's familiar theory for assigning a standard error to a maximum likelihood estimate is itself a "bootstrap theory", carried out in a parametric framework. Consider the law school example again, and suppose we are willing to accept a bivariate normal model for the data. The parametric maximum likelihood estimate of the unknown F is

$$\hat{F}_{NORM} \sim n_2\left(\begin{pmatrix}\hat{\mu}_y\\ \hat{\mu}_z\end{pmatrix}, \begin{pmatrix}\hat{\mu}_{yy} & \hat{\mu}_{yz}\\ \hat{\mu}_{yz} & \hat{\mu}_{zz}\end{pmatrix}\right) , \tag{5.7}$$

where, if $X = (Y,Z)$ denotes a typical bivariate data point, $\hat{\mu}_y = \bar{y}$, $\hat{\mu}_{yz} = \Sigma(y_i-\bar{y})(z_i-\bar{z})/n$, etc.

We could now execute the bootstrap algorithm exactly as before, except starting with $\hat{F}_{NORM}$ in place of $\hat{F}$ at (5.4). In fact, we don't carry out the Monte Carlo sampling (5.5) - (5.6). Theoretical calculations show that if we did, and if we let $B\to\infty$, $\hat{SD}$ as calculated at (5.6) would approximately[†] equal $[(1-\hat{\rho}^2)/(n-3)]^{1/2}$. Theoretical calculation is impossible outside a narrow family of parametric models, but the bootstrap algorithm (5.4) - (5.6) can always be carried out, given enough

[†]The approximation involved is very much like the approximation of the bootstrap by the jackknife or infinitesimal jackknife, as discussed in Chapter VI.

raw computing power. The bootstrap is a theory well-suited to an age of plentiful computation, see Efron (1979b).

3. Smoothed Bootstrap. We might wish to attribute some smoothness to F, without going all the way to the normal model (5.7). One way to do this is to use a smoothed estimate of F in place of $\hat{F}$ at step (5.4). In the law school problem, for example, we might use

$$\hat{F}_{.5} = \hat{F} * (.5\ \hat{F}_{NORM}) , \tag{5.8}$$

the convolution of $\hat{F}$ with a version of $\hat{F}_{NORM}$ scaled down by factor .5. This amounts to smearing out the atoms of $\hat{F}$ into half sized versions of $\hat{F}_{NORM}$, each centered at an x_i. Notice that $\hat{F}_{.5}$ has the same correlation coefficient as does $\hat{F}$ and $\hat{F}_{NORM}$, namely the observed value $\hat{\rho} = .776$. If we were bootstrapping a location parameter, say $\hat{\mu}_y$, instead of $\hat{\rho}$, we would have to rescale $\hat{F}_{.5}$ to have the same covariance matrix as $\hat{F}$. Otherwise our Sd estimate would be biased upward.

Example 4. The Correlation Coefficient. Table 5.2, taken from Efron (1980b), is a comparative Monte Carlo study of 15 estimators of standard deviation for the correlation coefficient $\hat{\rho}$, and also for $\hat{\phi} = \tanh^{-1} \hat{\rho} = (1/2) \log (1+\hat{\rho})/(1-\hat{\rho})$. The study comprised 200 trials of $X_1, X_2, \ldots, X_{14} \sim$ Bivariate normal with true correlation coefficient $\rho = .5$. Four summary statistics of how the Sd estimates performed in the 200 trials, the average, standard deviation, coefficient of variation, and root mean square error (of estimated minus true standard deviation), are presented for each of the 15 estimators, for both $\hat{\rho}$ and $\hat{\phi}$. We will refer back to this table in later chapters as we introduce the estimators of lines 7-14.

Method	$\hat{\rho}$ AVE	$\hat{\rho}$ SD	$\hat{\rho}$ CV	$\hat{\rho}$ $\sqrt{MSE}$	$\hat{\phi} = \tanh^{-1}\hat{\rho}$ AVE	$\hat{\phi}$ SD	$\hat{\phi}$ CV	$\hat{\phi}$ $\sqrt{MSE}$
1. Bootstrap, B = 128	.206	.066	.32	.068	.301	.065	.22	.065
2. Bootstrap, B = 512	.206	.063	.31	.065	.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	.063	.298	.058	.19	.058
5. Uniform Smoothed Bootstrap, B = 512	.205	.059	.29	.061	.296	.052	.18	.052
6. Jackknife	.223	.085	.38	.085	.314	.090	.29	.091
7. Infinitesimal Jackknife (Delta Method)	.175**	.058	.33	.074	.244*	.052	.21	.076
8. Halfsamples, all 128	.244*	.083	.34	.086	.364**	.099	.27	.118
9. Random HS, B = 128	.248*	.079	.32	.083	.368**	.084	.23	.109
10. Balanced HS, 8	.244*	.095	.39	.098	.366**	.111	.30	.129
11. Complementary HS, all 128	.223	.079	.35	.079	.336*	.099	.30	.105
12. Complementary Bal. HS, 16	.222	.081	.36	.081	.335*	.100	.30	.106
13. Random Subsampling, B = 128	.267**	.080	.30	.092	.423***	.089	.21	.153
14. Random Subsampling, Range $\hat{SD}$	.242	.092	.38	.094	.354*	.077	.27	.111
15. Normal Theory	.217	.056	.26	.056	.302	0	0	.003
True Value	.221				.299			

Table 5.2. A comparison of 15 methods of assigning standard deviation estimates to $\hat{\rho}$ and $\hat{\phi} = \tanh^{-1}\hat{\rho}$. The Monte Carlo experiment consisted of 200 trials of $X_1, X_2, \ldots, X_{14} \sim$ Bivariate Normal, true $\rho = .5$. The true standard deviations are $Sd(\hat{\rho}) = .221$, $SD(\hat{\phi}) = .299$. Large biases are indicated by asterisks: *Relative Bias $\geq .10$, **Relative Bias $\geq .20$, ***Relative Bias $\geq .40$.

Lines 1 and 2 of Table 5.2 refer to the bootstrap estimate of standard deviation (5.6), with $B = 128$ and 512 respectively. A components of variance analysis revealed that taking $B=\infty$ would not further decrease the root mean square error below .064 for $\hat{SD}(\hat{\rho})$ or .061 for $\hat{SD}(\hat{\phi})$. Line 3 of the table used the smoothed bootstrap (5.8). Lines 4 and 5 used uniform smoothing: the X_i^* were selected from $\hat{F} * (.5\ \hat{F}_{UNIF})$, where $\hat{F}_{UNIF}$ is the uniform distribution over a rhombus selected such that it has the same covariance matrix as $\hat{F}_{NORM}$. The jackknife results, line 6, are substantially worse, as we have already seen in Table 3.3.

Line 15 gives summary statistics for the parametric bootstrap, i.e. the standard normal theory estimates $\hat{SD}(\hat{\rho}) = [(1-\hat{\rho}^2)/(n-3)]^{1/2}$, $\hat{SD}(\hat{\phi}) = [1/(n-3)]^{1/2}$. The ordinary bootstrap performs surprisingly close to the normal theory estimate for $\hat{SD}(\hat{\rho})$, so it is not surprising that smoothing doesn't help much here. The motivation behind the $\tanh^{-1}$ transform is to stabilize the variance, that is to make $\hat{SD}$ a constant. In this case smoothing is quite effective. Notice that if the constant .5 in (5.8) were increased toward ∞, the normal smoothed bootstrap would approach the normal theory estimate of Sd, so that the root mean square error would approach zero.

4. Bootstrap Methods for More General Problems. The standard deviation plays no special role in (5.2), or in anything else having to do with the bootstrap. We can consider a perfectly general one-sample problem. Let

$$R(\underset{\sim}{X},F)$$

be a random variable of interest, where $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ indicates the entire i.i.d. sample $X_1, X_2, \ldots, X_n$. On the basis of having observed $\underset{\sim}{X}=\underset{\sim}{x}$, we wish to estimate some aspect of R's distribution, for example $E_F R$ or $\text{Prob}_F\{R < 2\}$.

The bootstrap algorithm (5.4) = (5.6) is carried out exactly as before except that at step 2 we calculate

$$R^* = R(\underset{\sim}{X}^*, \hat{F})$$

instead of $\hat{\theta}^*$, and at step 3 we calculate whichever aspect of R's distribution we are interested in, rather than $\hat{SD}$. For instance, if we wish to estimate $E_F R$ we calculate

$$E_* R^* = \frac{1}{B} \sum_{b=1}^{B} R^{*b} , \tag{5.9}$$

while if we are interested in $\text{Prob}_F\{R < 2\}$ we calculate

$$\text{Prob}_*\{R^* < 2\} = \frac{\#\{R^{*b} < 2\}}{B} .$$

In all cases, we are calculating a Monte Carlo approximation to the nonparametric MLE for the quantity of interest, the approximation being that B is finite rather than infinite.

Notation. "$E_* R^*$" indicates the expectation of $R^* = R(\underset{\sim}{X}^*, \hat{F})$ under the bootstrap sampling procedure $X_1^*, X_2^*, \ldots, X_n^* \overset{iid}{\sim} \hat{F}$, $\hat{F}$ fixed as at (5.4); likewise the notations "Prob_*", "Var_*", "SD_*", etc. Expression (5.9), like (5.6), ignores the fact that B is finite.

5. The Bootstrap Estimate of Bias. Suppose we wish to estimate the bias of a functional statistic, $\text{Bias} = E_F\theta(\hat{F}) - \theta(F)$ as at (2.3). We can take $R(\underset{\sim}{X},F) = \theta(\hat{F}) - \theta(F)$, and use the bootstrap algorithm to estimate $E_F R = \text{Bias}$. In this case

$$R^* = R(\underset{\sim}{X}^*, \hat{F}) = \theta(\hat{F}^*) - \theta(\hat{F}) = \hat{\theta}^* - \hat{\theta} ,$$

where $\hat{\theta}^* = \theta(\hat{F}^*)$, $\hat{F}^*$ being the empirical probability distribution of the bootstrap sample: $\hat{F}^*$ puts mass M_i^*/n on x_i, where M_i^* is the number of times x_i appears in the bootstrap sample.

The bootstrap estimate of bias is $\widehat{BIAS} = E_* R^*$, approximated by

$$\widehat{BIAS} = \frac{1}{B}\sum_{b=1}^{N} \hat{\theta}^{*b} - \hat{\theta} = \hat{\theta}^{*\cdot} - \hat{\theta} \ . \tag{5.10}$$

The 1000 bootstrap replications for the law school data yielded $\hat{\rho}^{*\cdot} = .779$, so BIAS = .779 - .776 = .003, compared to $-.007$ for the jackknife and $-.011$ for normal theory. In the Monte Carlo experiment of Table 5.2, the 200 bootstrap estimates of bias, B = 512, averaged $-.013$ with standard deviation .022. The jackknife bias estimates averaged $-.018$ with standard deviation .037. The true bias is $-.014$.

We don't need $\hat{\theta}$ to be a functional statistic to apply (5.10), and as a matter of fact we don't need $\hat{\theta}$ to be "the same statistic as θ" in any sense. We could just as well take $\theta(F) = E_F X$ and $\hat{\theta}$ = sample median. To state things as in (5.1), (5.2), write the bias $E_F\hat{\theta} - \theta(F)$ as

$$\text{Bias} = \beta(F, n, \hat{\theta}, \theta) = \beta(F) \ ,$$

a function of the unknown F, once the sample size n and the forms $\hat{\theta}(\cdot, \cdot, \ldots, \cdot)$ and $\theta(\cdot)$ are fixed. Then the bootstrap estimate is simply

$$\widehat{BIAS} = \beta(\hat{F}) \ .$$

Chapter VI indicates the connection between $\widehat{BIAS}_{BOOT}$ and $\widehat{BIAS}_{JACK}$, the bootstrap and jackknife estimates of Bias. The following theorem is proved in Chapter VI, Section 6:

Theorem 5.1. If $\hat{\theta} = \theta(\hat{F})$ is a quadratic functional, then

$$\widehat{\text{BIAS}}_{\text{BOOT}} = \frac{n-1}{n} \widehat{\text{BIAS}}_{\text{JACK}} .$$

6. Finite Sample Spaces. The rationale for the bootstrap method is particularly evident when the sample space $\mathcal{X}$ is finite, say $\mathcal{X} = \{1, 2, \ldots, L\}$. Then we can express F as $\underset{\sim}{f} = (f_1, f_2, \ldots, f_L)$, where $f_\ell = \text{Prob}_F\{X=\ell\}$, and $\hat{F}$ as $\hat{\underset{\sim}{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_L)$, where $\hat{f}_\ell = \#\{x_i = \ell\}/n$. The random variable $R(\underset{\sim}{X},F)$ can be written as

$$R(\underset{\sim}{X},F) = Q(\hat{\underset{\sim}{f}},\underset{\sim}{f}) , \tag{5.11}$$

some function of $\hat{\underset{\sim}{f}}$ and $\underset{\sim}{f}$, assuming that $R(\underset{\sim}{X},F)$ is invariant under permutations of the X_i.

The distribution of $\hat{\underset{\sim}{f}}$ given $\underset{\sim}{f}$ is a rescaled multinomial, L categories, n draws, true probability vector $\underset{\sim}{f}$,

$$\hat{\underset{\sim}{f}} | \underset{\sim}{f} \sim \text{Mult}_L(n,\underset{\sim}{f})/n . \tag{5.12}$$

The bootstrap distribution of $X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F}$ can be described in terms of $\hat{\underset{\sim}{f}}^* = (\hat{f}_1^*, \hat{f}_2^*, \ldots, \hat{f}_L^*)$, where $\hat{f}_\ell^* = \#\{X_i^* = \ell\}/n$. It is the same as (5.12), except with $\hat{\underset{\sim}{f}}$ playing the role of $\underset{\sim}{f}$,

$$\hat{\underset{\sim}{f}}^* | \hat{\underset{\sim}{f}} \sim \text{Mult}_L(n,\hat{\underset{\sim}{f}})/n . \tag{5.13}$$

The bootstrap method estimates the unobservable distribution of $Q(\hat{\underset{\sim}{f}},\underset{\sim}{f})$ under (5.12) by the observable distribution of $Q^* = Q(\hat{\underset{\sim}{f}}^*, \hat{\underset{\sim}{f}})$ under (5.13).

Example 5. Binomial Probability. $\mathcal{X} = \{1,2\}$, $\theta(\underset{\sim}{f}) = f_2 = \text{Prob}_{\underset{\sim}{f}}\{X=2\}$, $\hat{\theta} = \theta(\hat{\underset{\sim}{f}}) = \hat{f}_2$, and $R(\underset{\sim}{X},F) = Q(\hat{\underset{\sim}{f}},\underset{\sim}{f}) = \hat{f}_2 - f_2$, the difference between the

observed and theoretical frequency for the second category. From (5.13) we see that Q^* is a standardized binomial,

$$Q^* = \hat{f}_2^* - f_2 \overset{*}{\sim} \frac{Bi(n, \hat{f}_2)}{n} - \hat{f}_2$$

with first two moments

$$Q^* \overset{*}{\sim} \left(0, \frac{\hat{f}_1\hat{f}_2}{n}\right) .$$

(The notation "$\overset{*}{\sim}$" indicates the bootstrap distribution.) The implication from the bootstrap theory is that $E_{\hat{f}}Q = 0$, i.e. that $\hat{f}_2$ is unbiased for f_2, and $Var\{Q\} = Var\{\hat{f}_2^* - \hat{f}_2\} = (\hat{f}_1\hat{f}_2)/n$, which of course is the standard binomial estimate.

Asymptotics. As the sample size $n\to\infty$, both $\hat{\underset{\sim}{f}} - \underset{\sim}{f}$ under (5.12) and $\hat{\underset{\sim}{f}}^* - \hat{\underset{\sim}{f}}$ under (5.13) approach the same L dimensional normal distribution, $h_L(0, \underset{\sim}{\ddagger}_f/n)$, where $\underset{\sim}{\ddagger}_f$ has diagonal elements $f_\ell(1-f_\ell)$, off-diagonal elements $-f_\ell f_m$. If $Q(\cdot,\cdot)$ is a well-behaved function, as described in Remark G of Efron (1979a), then the bootstrap distribution of Q^* is asymptotically the same as the true distribution of Q. This justifies bootstrap inferences, such as estimating $E_F R$ by $E_* R^*$, at least in an asymptotic sense.

What is easy to prove for χ finite is quite difficult for χ more general. Recently Freedman and Bickel (1980) and Singh (1980) have separately demonstrated the asymptotic validity of the bootstrap for χ infinite. They consider statistics like the average U-statistics, t-statistics, and quantiles, and show that the bootstrap distribution of R^* converges to the true distribution of R. The convergence is generally quite good, faster than the standard convergence results for the central limit theorem.

7. Regression Models. So far we have only discussed one-sample situations, where all the random quantities X_i come from the same distribution F. Bootstrap methods apply just as well to many-sample situations, and to a variety of other more complicated data structures. For example, Efron (1980a) presents bootstrap estimates and confidence intervals for censored data. We conclude this chapter with a brief discussion of bootstrap methods for regression models.

A reasonably general regression situation is the following: independent real-valued observations $Y_i = y_i$ are observed, where

$$Y_i = g_i(\beta) + \varepsilon_i , \qquad i=1, 2, \ldots, n . \tag{5.14}$$

The functions $g_i(\cdot)$ are of known form, usually depending on some observed vector of covariates c_i, while β is a $p\times 1$ vector of unknown parameters. The ε_i are i.i.d. for some distribution F on $\mathcal{R}^1$,

$$\varepsilon_i \overset{iid}{\sim} F , \qquad i=1, 2, \ldots, n , \tag{5.15}$$

where F is assumed to be centered at zero in some sense, perhaps $E_F\varepsilon = 0$ or $\text{Prob}_F\{\varepsilon < 0\} = .5$.

Having observed the $n\times 1$ data vector $\underset{\sim}{Y} = \underset{\sim}{y} = (y_1, y_2, \ldots, y_n)'$, we estimate β by minimizing some measure of distance $D(\underset{\sim}{y},\underset{\sim}{\eta})$ between $\underset{\sim}{y}$ and the vector of predictors $\underset{\sim}{\eta}(\beta) = (g_1(\beta), g_2(\beta), \ldots, g_n(\beta))'$,

$$\hat{\beta}: \min_{\beta} D(\underset{\sim}{y}, \underset{\sim}{\eta}(\beta)) . \tag{5.16}$$

The most common choice of D is $D(\underset{\sim}{y},\underset{\sim}{\eta}) = \sum_{i=1}^{n} (y_i-\eta_i)^2$.

Suppose model (5.14) - (5.16) is too complicated for standard analysis, but we need an assessment of $\hat{\beta}$'s sampling properties. For example, we

might have $g_i(\beta) = e^{c_i\beta}$, F of unknown distributional form, and $D(\underset{\sim}{y},\underset{\sim}{\eta}) = \Sigma|y_i-\eta_i|$. The bootstrap algorithm (5.4) - (5.6) can be modified as follows:

1) Construct $\hat{F}$ putting mass $\frac{1}{n}$ at each observed residual,

$$\hat{F}\text{: mass } \frac{1}{n} \text{ at } \hat{\varepsilon}_i = y_i - g_i(\hat{\beta}) \ . \tag{5.17}$$

2) Draw a bootstrap data set

$$Y_i^* = g_i(\hat{\beta}) + \varepsilon_i^* \ , \qquad i=1, 2, \ldots, n \ , \tag{5.18}$$

where the ε_i^* are i.i.d. from $\hat{F}$, and calculate

$$\hat{\beta}^*\text{: } \min_{\beta} D(\underset{\sim}{Y}^*, \underset{\sim}{\eta}(\beta)) \ . \tag{5.19}$$

3) Independently repeat step (2) B times, obtaining bootstrap replications $\hat{\beta}^{*1}, \hat{\beta}^{*2}, \ldots, \hat{\beta}^{*B}$.

As an estimate of $\hat{\beta}$'s covariance matrix, for example, we can take

$$\widehat{\mathrm{COV}} = \frac{\sum_{b=1}^{B} (\hat{\beta}^{*b}-\hat{\beta}^*)(\hat{\beta}^{*b}-\hat{\beta}^{*\cdot})'}{B-1} \ . \tag{5.20}$$

Example 6. Linear Regression. The ordinary linear regression situation is $g_i(\beta) = c_i\beta$, c_i a $1\times p$ vector of known covariates, and $D(\underset{\sim}{y},\underset{\sim}{\eta}) = \Sigma(y_i-\eta_i)^2$. Let $\underset{\sim}{C}$ be the $n\times p$ matrix with c_i as the i-th row, and $\underset{\sim}{G} = \underset{\sim}{C}'\underset{\sim}{C}$. For convenience assume that the first element of each c_i is 1, and that $\underset{\sim}{G}$ is of full rank p.

In this case we can evaluate (5.20) without recourse to Monte Carlo sampling. Notice that $\hat{F}$ has expected value 0 and variance $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/n$, and that $Y_i^* = c_i\hat{\beta} + \varepsilon_i^*$ is a standard linear model written in unusual notation. Standard linear model theory shows that $\hat{\beta}^* = \underset{\sim}{G}^{-1}\underset{\sim}{C}'\underset{\sim}{Y}^*$, and that

$$\widehat{\mathrm{COV}} = \hat{\sigma}^2 \, \underset{\sim}{G}^{-1} \, . \tag{5.21}$$

In other words, the bootstrap gives the standard estimate of covariance in the linear regression case, except for the use of $\Sigma \hat{\varepsilon}_i^2/n$ rather than $\Sigma \hat{\varepsilon}_i^2/(n-p)$ to estimate σ^2. This contrasts with the jackknife result (3.13).

The algorithm (5.17) - (5.19) depends on $\hat{F}$ being a reasonable estimate of F, and can give falsely optimistic results if we are fitting highly overparameterized models in hopes of finding a good one. As an example, consider ordinary polynomial regression on the real line. The observed data is of the form (t_1,y_1), (t_2,y_2), ..., (t_n,y_n), where t_i is the value of the predictor variable for y_i. If $n = 20$ and we fit a 19th degree polynomial ($g_i(\beta) = c_i\beta$ where $c_i = (1, t_i, t_i^2, \ldots, t_i^{19})$), then $\hat{\sigma}^2 = \Sigma \hat{\varepsilon}_i^2/20$ is likely to be very small and (5.21) will probably give a foolishly optimistic assessment of $\mathrm{Cov}(\hat{\beta})$. In this case the trouble can be mitigated by using the unbiased estimate of σ^2, $\Sigma \hat{\varepsilon}_i^2$, instead of $\hat{\sigma}^2$, but the general situation is unclear.

As a more cautious alternative to (5.17) - (5.19) we can use the one-sample bootstrap (5.4) - (5.6), with the individual data points being $x_i = (t_i,y_i)$. This method appears to give reasonable answers in model selection situations, as discussed in Chapter VII. On the other hand, it gets us back to results more like (3.13) in the standard linear model.

VI. INFINITESIMAL JACKKNIFE, DELTA METHOD, AND THE INFLUENCE FUNCTION

In this section we show the connection between the jackknife and the bootstrap, using a simple picture. The picture suggests another version of the jackknife, Jaeckel's (1972) "infinitesimal jackknife". The infinitesimal jackknife turns out to be exactly the same as the ordinary delta method, when the latter applies, and also the same as methods based on the influence function, Hampel (1974). We begin with a brief discussion of *resampling procedures*, a generic name for all methods which evaluate $\hat{\theta}$ at reweighted versions of the empirical probability distribution $\hat{F}$.

1. Resampling Procedures. For simplicity, consider a functional statistic $\hat{\theta} = \theta(\hat{F})$, (2.1). The data $x_1, x_2, \ldots, x_n$ is thought of as observed and fixed in what follows. A *resampling vector* †

$$\underset{\sim}{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$$

is any vector on the n dimensional simplex

$$\mathcal{S}_n = \{\underset{\sim}{P}^* : \; P_i^* \geq 0, \; \sum_{i=1}^{n} P_i^* = 1\} \; , \tag{6.1}$$

in other words, any probability vector. Corresponding to each $\underset{\sim}{P}^*$ is a reweighted empirical probability distribution $\hat{F}^*$,

$$\hat{F}^* : \text{ mass } P_i^* \text{ on } x_i, \quad i=1, 2, \ldots, n \; , \tag{6.2}$$

and a "resampled" value of $\hat{\theta}$, say $\hat{\theta}^*$,

$$\hat{\theta}^* = \theta(\hat{F}(\underset{\sim}{P}^*)) = \hat{\theta}(\underset{\sim}{P}^*) \; . \tag{6.3}$$

†It is notationally convenient to consider P^*, as well as some of the other vectors introduced later, as rows rather than columns.

Some of the resampling vectors play special roles in the bootstrap and jackknife theories. In particular

$$\underset{\sim}{P}^* = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) \tag{6.4}$$

corresponds to $\hat{F}$ itself, and to the observed value of the statistic $\hat{\theta} = \hat{\theta}(P^*)$. The jackknife considers vectors

$$\underset{\sim}{P}_{(i)} = (\frac{1}{n-1}, \frac{1}{n-1}, \ldots, 0, \frac{1}{n-1}, \ldots, \frac{1}{n-1}) \ , \quad (0 \text{ in i-th place}) \ , \tag{6.5}$$

with corresponding values $\hat{\theta}_{(i)}$ of the statistic, i=1, 2, ..., n. The bootstrap considers all $\underset{\sim}{P}^*$ vectors of the form $\underset{\sim}{M}^*/n$, $\underset{\sim}{M}^*$ having non-negative integer coordinates adding to n.

Another way to describe the bootstrap algorithm is to say that the resampling vectors are selected according to a rescaled multinomial distribution,

$$\underset{\sim}{P}^* \overset{*}{\sim} \text{Mult}_n(n, \underset{\sim}{P}^\circ)/n \ , \tag{6.6}$$

n independent draws on n categories each having probability 1/n, rescaled by factor 1/n. The symbol "$\overset{*}{\sim}$" is a reminder that the statistician, not nature, induces the randomness in $\underset{\sim}{P}^*$. Here

$$P_i^* = \frac{\#\{X_j^* = x_i\}}{n} \ ,$$

the proportion of the bootstrap sample equal to x_i. The bootstrap standard deviation and bias estimates are simply

$$\hat{SD} = SD_* \ \hat{\theta}(\underset{\sim}{P}^*)$$

and

$$\widehat{\text{BIAS}} = E_* \hat{\theta}(\underset{\sim}{P}^*) - \hat{\theta} \ ,$$

SD_* and E_* indicating standard deviation and expectation under (6.6). For future reference note that distribution (6.6) has mean vector and covariance matrix

$$\underset{\sim}{P}^* \ \tilde{*} \ (\underset{\sim}{P}^\circ, \ \underset{\sim}{I}/n^2 - \underset{\sim}{P}^{\circ\prime}\underset{\sim}{P}^\circ/n) \ , \tag{6.7}$$

$\underset{\sim}{I}$ the nxn identitiy matrix.

Figure 6.1 shows a schematic representation of the function $\hat{\theta}(\underset{\sim}{P}^*)$ as a curved surface over the simplex $\mathcal{S}_n$. The vertical direction can be taken along the n-th coordinate axis since P_n^* is redundant, $P_n^* = 1 - \sum_{i=1}^{n-1} P_i^*$.

Example: Quadratic Functionals. A quadratic functional as defined at (4.14), (4.16) is also quadratic as a function of $\underset{\sim}{P}^*$,

$$\hat{\theta}(\underset{\sim}{P}^*) = \hat{\theta}(\underset{\sim}{P}^\circ) + (\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{U} + \frac{1}{2}(\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{V}\ (\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)' \ . \tag{6.8}$$

The column vector $\underset{\sim}{U}$ and symmetric matrix $\underset{\sim}{V}$ are expressed, after some algebraic manipulation, in terms of $\alpha_i = \alpha^{(\infty)}(x_i)$ and $\beta_{ii'} = \beta_{i'i} = \beta(x_i, x_{i'})$,

$$U_i = \alpha_i - \alpha_\cdot + \beta_{i\cdot} - \beta_{\cdot\cdot} \ , \quad V_{ii'} = \beta_{ii'} - \beta_{i\cdot} - \beta_{\cdot i'} + \beta_{\cdot\cdot} \ , \tag{6.9}$$

the dot indicating averages, $\alpha_\cdot = \sum_i \alpha_i/n$, $\beta_{i\cdot} = \sum_{i'} \beta_{ii'}/n$, $\beta_{\cdot\cdot} = \sum_i \beta_{i\cdot}/n$. Expressions (6.9) satisfy the side conditions

$$\sum_i U_i = 0, \quad \sum_{i'} V_{ii'} = \sum_i V_{ii'} = 0 \ , \tag{6.10}$$

which defines the quadratic form (6.8) uniquely. (The possibility of non-uniqueness arises because $\hat{\theta}(\underset{\sim}{P}^*)$ is defined only on $\mathcal{S}_n$, lying in an n-1 dimensional subspace of R^n.) In the special case $\beta(\cdot,\cdot) = 0$, $\hat{\theta}$ is a *linear functional statistic*,

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = \mu + \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i) \ , \tag{6.11}$$

and $\hat{\theta}(\underset{\sim}{P}^*)$ is a linear function of $\underset{\sim}{P}^*$, $\hat{\theta}(\underset{\sim}{P}^*) = \hat{\theta}(\underset{\sim}{P}^\circ) + (\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{U}$, $U_i = \alpha_i - \alpha_\cdot$.

2. Relation Between the Jackknife and Bootstrap Estimates of Standard Deviation. There is a unique linear function $\hat{\theta}_{LIN}(\underset{\sim}{P}^*)$ agreeing with $\hat{\theta}(\underset{\sim}{P}^*)$ at $\underset{\sim}{P}_{(i)}$, i=1, ..., n ,

$$\begin{aligned} &\hat{\theta}_{LIN}(\underset{\sim}{P}^*) = \hat{\theta}_{(\cdot)} + (\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{U} \\ &U_i = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}), \qquad i=1, 2, \ldots, n \ . \end{aligned} \tag{6.12}$$

Theorem 6.1. The jackknife standard deviation estimate for $\hat{\theta}$, $\hat{SD}_{JACK}(\hat{\theta})$, is

$$\hat{SD}_{JACK}(\hat{\theta}) = \sqrt{\frac{n}{n-1}}\, SD_*(\hat{\theta}_{LIN}(\underset{\sim}{P}^*)) \ . \tag{6.13}$$

In other words, $\hat{SD}_{JACK}(\hat{\theta})$ is itself almost a bootstrap Sd estimate, equaling $[\frac{n}{n-1}]^{1/2}\, \hat{SD}_{BOOT}(\hat{\theta}_{LIN})$. The factor $[\frac{n}{n-1}]^{1/2}$ makes $[\hat{SD}_{JACK}(\hat{\theta})]^2$ unbiased for $[Sd(\hat{\theta})]^2$ if $\hat{\theta}$ is a linear functional. As remarked in Chapter III, we could but don't use the same factor to make $[\hat{SD}_{BOOT}(\hat{\theta})]^2$ unbiased in the linear case.

Proof. From (6.12), (6.7), and the fact that $\underset{\sim}{P}^\circ\underset{\sim}{U} = \Sigma U_i/n = 0$,

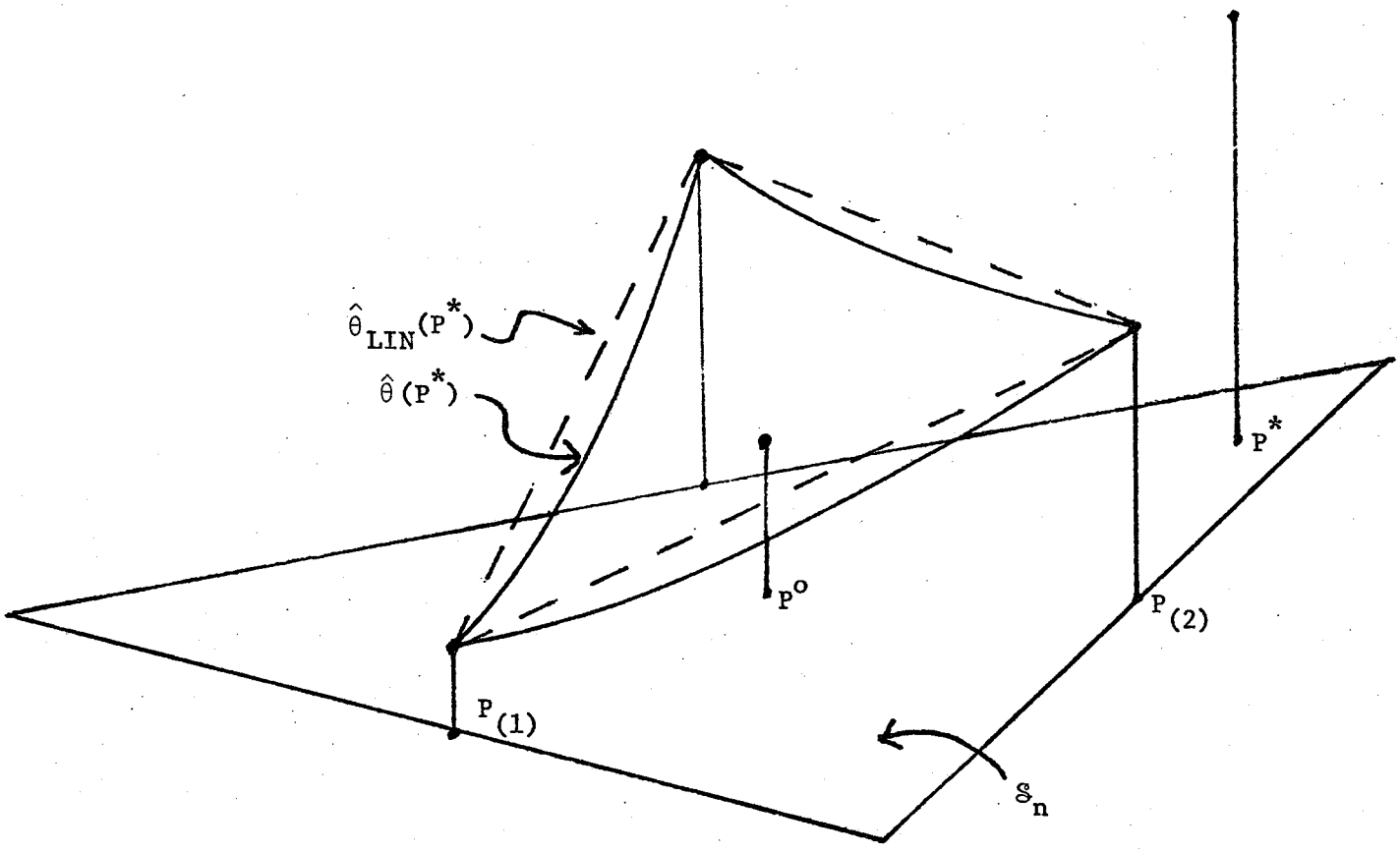


Figure 6.1. A schematic representation of $\hat{\theta}(\underset{\sim}{P}^*)$ as a function on $\mathcal{S}_n$. The curved surface $\hat{\theta}(\cdot)$ is approximated by the linear function $\hat{\theta}_{LIN}(\cdot)$. The bootstrap standard deviation estimate is $SD_*\hat{\theta}(\underset{\sim}{P}^*)$, the jackknife Sd estimate is $[\frac{n}{n-1}]^{1/2}\ SD_*\ \hat{\theta}_{LIN}(\underset{\sim}{P}^*)$, where SD_* indicates standard deviation under the multinomial distribution (6.6).

$$\mathrm{SD}_* \hat{\theta}_{\mathrm{LIN}}(\underset{\sim}{P}^*) = \left[\frac{\sum_i U_i^2}{n^2}\right]^{1/2} = \left[\left(\frac{n-1}{n}\right)^2 \Sigma[\hat{\theta}_{(i)}-\hat{\theta}_{(\cdot)}]^2\right]^{1/2} \qquad (6.14)$$

$$= \left[\frac{n-1}{n}\right]^{1/2} \hat{\mathrm{SD}}_{\mathrm{JACK}}(\hat{\theta}) \ \square$$

Notice that for linear functionals $\hat{\theta}$, we can evaluate $\mathrm{SD}_*\hat{\theta}(\underset{\sim}{P}^*)$ directly from (6.7), without resorting to Monte Carlo calculations. The jackknife requires less computation than the bootstrap because it approximates any $\hat{\theta}$ with a linear functional.

Figure 6.1 is misleading in one important sense. Under (6.6), $\|\underset{\sim}{P}^* - \underset{\sim}{P}^\circ\| = O_p(\frac{1}{\sqrt{n}})$, while $\|\underset{\sim}{P}_{(i)} - \underset{\sim}{P}^\circ\| = O(\frac{1}{n})$. The bootstrap resampling vectors tend to be much further away from the central value $\underset{\sim}{P}^\circ$ than are the jackknife resampling vectors $\underset{\sim}{P}_{(i)}$. This is what cuases trouble for markedly nonlinear statistics like the median. The grouped jackknife, Chapter II, Section 2, resamples at distance $\frac{1}{n}[h\,\frac{g}{g-1}]^{1/2}$ from $\underset{\sim}{P}^\circ$. Taking the group size $h = n^{1/2}$ gives distance $O(1/\sqrt{n})$, as with the bootstrap, and gives an asymptotically correct Sd estimate for the sample median.

3. Jaeckel's Infinitesimal Jackknife. Figure 6.1 suggests another estimate of standard deviation: instead of approximating $\hat{\theta}(\underset{\sim}{P}^*)$ by $\hat{\theta}_{\mathrm{LIN}}(\underset{\sim}{P}^*)$, why not use $\hat{\theta}_{\mathrm{TAN}}(\underset{\sim}{P}^*)$, the hyperplane tangent to the surface $\hat{\theta}(\underset{\sim}{P}^*)$ at the point $\underset{\sim}{P}^* = \underset{\sim}{P}^\circ$? The corresponding estimate of standard deviation is then

$$\hat{\mathrm{SD}} = \mathrm{SD}_* \hat{\theta}_{\mathrm{TAN}}(\underset{\sim}{P}^*) \ , \qquad (6.15)$$

SD_* indicating standard deviation under distribut on (6.6) as before. Formula (6.15) is Jaeckel's (1972) *infinitesimal jackknife* estimate of standard deviation. In other words, $\hat{\mathrm{SD}}_{\mathrm{IJ}}(\hat{\theta}) = \hat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{\theta}_{\mathrm{TAN}})$.

The function $\hat{\theta}_{TAN}(\cdot)$ equals

$$\hat{\theta}_{TAN}(\underset{\sim}{P}^*) = \hat{\theta}(\underset{\sim}{P}^\circ) + (\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{U} ,$$
$$U_i = \lim_{\varepsilon \to 0} \frac{\hat{\theta}(\underset{\sim}{P}^\circ + \varepsilon(\underset{\sim i}{\delta} - \underset{\sim}{P}^\circ)) - \hat{\theta}(\underset{\sim}{P}^\circ)}{\varepsilon} , \quad i=1, 2, \ldots, n , \tag{6.16}$$

$\underset{\sim i}{\delta}$ being the i-th coordinate vector. The U_i are directional derivatives. Suppose we extend the definition of $\hat{\theta}(\underset{\sim}{P}^*)$ to values of $\underset{\sim}{P}^*$ outside $\mathcal{S}_n$ in any reasonable way, for example by the homogeneous extension $\hat{\theta}(\underset{\sim}{P}^*) = \hat{\theta}(\underset{\sim}{Q}^*)$, $\underset{\sim}{Q}^* = \underset{\sim}{P}^* / \sum_{i=1}^n P_i^*$. (If $\underset{\sim}{P}^*$ has nonnegative coordinates summing to any positive value then $\underset{\sim}{Q}^* \in \mathcal{S}_n$; the homogeneous extension assigns $\hat{\theta}(\underset{\sim}{P}^*)$ the same values along any ray from the origin.) Let $\underset{\sim}{D}$, a column vector, be the gradient vector of $\hat{\theta}(\underset{\sim}{P}^*)$ evaluated at $\underset{\sim}{P}^* = \underset{\sim}{P}^\circ$, $D_i = \frac{\partial}{\partial P_i^*} \hat{\theta}(P^*)\Big|_{\underset{\sim}{P}^* = \underset{\sim}{P}^\circ}$. This definition makes sense because $\hat{\theta}(\underset{\sim}{P}^*)$ is now defined in an open neighborhood of $\underset{\sim}{P}^\circ$. Then $U_i = (\underset{\sim i}{\delta} - \underset{\sim}{P}^\circ)\underset{\sim}{D}$, and we see that

$$\sum_{i=1}^n U_i = \underset{\sim}{P}^\circ \underset{\sim}{U} = 0 . \tag{6.17}$$

Therefore, applying (6.7) to (6.15), (6.16) shows that the infinitesimal jackknife estimate of standard deviation is

$$\hat{SD}_{IJ}(\hat{\theta}) = [\sum_{i=1}^n U_i^2/n^2]^{1/2} . \tag{6.18}$$

The infinitesimal jackknife resamples $\hat{\theta}$ at $\underset{\sim}{P}^*$ values infinitesimally close to $\underset{\sim}{P}^\circ$, rather than $O(1/n)$ away as with the ordinary jackknife - hence the name "infinitesimal". Looking at the last equation in (6.16), take $\varepsilon = -1/(n-1)$ instead of letting $\varepsilon \to 0$. Then $U_i = [\hat{\theta}(\underset{\sim}{P}^\circ) + \varepsilon(\underset{\sim i}{\delta} - \underset{\sim}{P}^\circ)) - \hat{\theta}(\underset{\sim}{P}^\circ)]/\varepsilon = (n-1)(\hat{\theta} - \hat{\theta}_{(i)})$, similar to definition (6.12). (Compare (6.14) with (6.18).)

We can use other values of ϵ to define other jackknifes. For example, taking $\epsilon = 1/(n+1)$ in (6.16) makes $U_i = (n+1)(\hat{\theta}_{[i]}-\hat{\theta})$, where $\hat{\theta}_{[i]} = \hat{\theta}(x_1, x_2, \ldots, x_i, x_i, x_{n+1}, \ldots, x_n)$, i.e. the value of $\hat{\theta}$ when x_i is repeated rather than removed from the data set. The corresponding standard deviation estimate $[\Sigma(U_i-U_\cdot)^2/n^2]^{1/2}$, which might be called the "positive jackknife", is the bootstrap Sd of the linear function having value $\hat{\theta}_{[i]}$ at $P_{[i]} = (1/n, 1/n, \ldots, 2/n, \ldots, 1/n)$, $2/n$ in the i-th place, i=1, 2, ..., n. The positive jackknife was applied to $\hat{\theta}$ the correlation coefficient in Hinkley (1978), and produced estimates of standard deviation with extreme downward biases.

Getting back to the infinitesimal jackknife, consider a linear functional statistic $\hat{\theta} = \mu + \frac{1}{n}\Sigma\alpha(X_i)$; having representation (6.3), $\hat{\theta}(\underset{\sim}{P}^*) = \mu + \Sigma P_i^*\alpha_i$. Then U_i as defined in (6.16) equals $\alpha_i-\alpha_\cdot$, $\alpha_\cdot = \Sigma\alpha_i/n$, so

$$\hat{SD}_{IJ} = \left[\sum_{n=1}^{n} (\alpha_i-\alpha_\cdot)^2/n^2\right]^{1/2} .$$

Definition (6.15) does not include the bias correction factor $\sqrt{n/(n-1)}$, so for linear functionals $E_F[\hat{SD}_{IJ}]^2 = \frac{n-1}{n}[Sd]^2$. Multiplying $\hat{SD}_{IJ}$ by $\sqrt{14/13} = 1.038$ helps correct the severe downward bias of the infinitesimal jackknife evident in line 7 of Table 5.2, but not by much.

The directional derivatives U_i in (6.16) can be calculated explicitly for many common statistics. For example, if the statistic is the sample correlation coefficient $\hat{\rho}$, then

$$U_i = -\frac{1}{2}\hat{\rho}\left[\left(\frac{y_i-\hat{\mu}_y}{\sqrt{\hat{\mu}_{yy}}}\right)^2 + \left(\frac{z_i-\hat{\mu}_z}{\sqrt{\hat{\mu}_{zz}}}\right)^2\right] + \left(\frac{y_i-\hat{\mu}_y}{\sqrt{\hat{\mu}_{yy}}}\right)\left(\frac{z_i-\hat{\mu}_z}{\sqrt{\hat{\mu}_{zz}}}\right) , \qquad (6.19)$$

where $x_i = (y_i, z_i)$, $\hat{\mu}_y = \Sigma y_i/n$, $\hat{\mu}_{yy} = \Sigma(y_i - \hat{\mu}_y)^2/n$, etc. In fact, it is usually more convenient to evaluate the U_i numerically: simply substitute a small value of ε into definition (6.16). The value $\varepsilon = .001$ was used in Table 5.2.

4. Influence Function Estimates of Standard Deviation. The influence function expansion (4.13), $\theta(\hat{F}) = \theta(F) + \Sigma\ IF(X_i)/n + o_p(1/n)$, approximates an arbitrary functional statistic $\theta(\hat{F})$ by an average of i.i.d. random variables $\Sigma\ IF(X_i)/n$. This immediately suggests the standard deviation approximation

$$Sd(\hat{\theta}) \doteq [Var_F\ IF(X)/n]^{1/2}\ , \tag{6.20}$$

where, since $E_F\ IF(X) = 0$ (basically the same result as (6.17), see Section 3.5 of Hampel (1974)),

$$Var_F\ IF(X) = \int_{\mathcal{X}} IF^2(x)\ dF(x)\ .$$

In order to use (6.20) as an Sd estimator, we have to estimate $Var_F\ IF(X)$. The definition $IF(x) = \lim_{\varepsilon \to 0} [\theta((1-\varepsilon)F + \varepsilon\delta_x) - \theta(F)]/\varepsilon$ is obviously related to the definition of U_i in (6.16). As a matter of fact, U_i is the influence function of $\theta(F)$ for $F=\hat{F}$, evaluated at $x=x_i$. Mallows (1974) aptly calls U_i the *empirical influence function*, denoted $\hat{IF}(x_i)$. The obvious estimate of $Var_F\ IF(X)$ is $\int_{\mathcal{X}} \hat{IF}^2(x)\ d\hat{F}(x) = \Sigma\ U_i^2/n$. Plugging this into (6.20) gives the Sd estimate $\left[\frac{\Sigma\ U_i^2}{n^2}\right]^{1/2}$, which is exactly the infinitesimal jackknife estimate (6.18).

The ordinary jackknife and positive jackknife can also be thought of as estimates of (6.20). They use the influence function estimates

$$\hat{IF}(x_i) = \begin{cases} (n-1)\ (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}) & \text{(ordinary jackknife)} \\ (n+1)\ (\hat{\theta}_{[i]} = \hat{\theta}_{[\cdot]}) & \text{(positive jackknife)} \ . \end{cases}$$

All these methods give asymptotically correct results if $\theta(\cdot)$ is sufficiently smooth, but perform quite differently in small samples, as illustrated by Table 5.2. In the author's experience, the ordinary jackknife is the only jackknife which can be trusted not to give badly underbiased estimates of standard deviation. (This is the import of Theorem 4.1.) If one isn't going to use the bootstrap, because of computational costs, the ordinary jackknife seems to be the method of choice.

5. The Delta Method. Many statistics are of the form

$$\hat{\theta}(X_1, X_2, \ldots, X_n) = t(\bar{Q}_1, \bar{Q}_2, \ldots, \bar{Q}_A) \ , \tag{6.21}$$

where $t(\cdot, \cdot, \cdot\cdot, \cdot)$ is a known funition and each $\bar{Q}_a$ is an observed average,

$$\bar{Q}_a = \frac{1}{n} \sum_{i=1}^{n} Q_a(X_i) \ .$$

For example the correlation coefficient $\hat{\rho}$ equals

$$t(\bar{Q}_1, \bar{Q}_2, \bar{Q}_3, \bar{Q}_4, \bar{Q}_5) = \frac{\bar{Q}_4 - \bar{Q}_1\bar{Q}_2}{[\bar{Q}_3 - \bar{Q}_1^2]^{1/2}\ [\bar{Q}_5 - \bar{Q}_2^2]^{1/2}} \ ,$$

with $Q_1(X) = Q_1((Y,Z)) = Y$, $Q_2 = Z$, $Q_3 = Y^2$, $Q_4 = YZ$, $Q_5 = Z^2$.

Suppose that the vector $\underset{\sim}{Q}(X) = (Q_1(X), Q_2(X), \ldots, Q_A(X))$ corresponding to one observation of $X \sim F$ has mean vector $\underset{\sim F}{\alpha}$ and covariance matrix $\underset{\sim F}{\beta}$, and let $\underset{\sim F}{\nabla}$ be the gradient vector with a-th component

$\partial t/\partial Q_a \Big|_{\underset{\sim}{Q}=\underset{\sim}{\alpha}_F}$. Expanding $\hat{\theta} = t(\underset{\sim}{\hat{Q}})$ in a first order Taylor series about $\underset{\sim}{\alpha}_F$ gives the approximation

$$\mathrm{Sd}(\hat{\theta}) \doteq [\underset{\sim}{\nabla}_F \underset{\sim}{\beta}_F \underset{\sim}{\nabla}_F'/n]^{1/2} . \tag{6.22}$$

In the case of the correlation coefficient, somewhat tedious calculations show that (6.22) becomes

$$\mathrm{Sd}(\hat{\rho}) \doteq \left\{ \frac{\rho^2}{4n} \left[\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}} \right] \right\}^{1/2} , \tag{6.23}$$

where, denoting $X = (Y,Z)$, $\mu_{gh} = E_F[Y - E_F Y]^g\,[Z - E_F Z]^h$.

Substituting $\hat{F}$ for F in (6.22) gives the *nonparametric delta method* estimate of standard deviation,

$$\hat{SD} = [\underset{\sim}{\nabla}_{\hat{F}} \underset{\sim}{\beta}_{\hat{F}} \underset{\sim}{\nabla}_{\hat{F}}'/n]^{1/2} . \tag{6.24}$$

For example (6.23) would be estimated by the same expression with $\hat{\rho}$ replacing ρ and the sample moments $\hat{\mu}_{gh}$ replacing the $\hat{\mu}_{gh}$.

Theorem 6.2. For any statistic of form (6.21), the nonparametric delta method and the infinitesimal jackknife give identical estimates of standard deviation.

Proof. For statistics $\hat{\theta}$ of form (6.21), the directional derivatives U_i in (6.16) are

$$U_i = \underset{\sim}{\nabla}_{\hat{F}} \cdot [\underset{\sim}{Q}(x_i) - \underset{\sim}{\bar{Q}}]$$

since

$$\begin{aligned} \hat{\theta}(\underset{\sim}{P}^\circ + \varepsilon(\underset{\sim}{\delta}_i - \underset{\sim}{P}^\circ)) &= t(\underset{\sim}{\bar{Q}} + \varepsilon(\underset{\sim}{Q}(x) - \underset{\sim}{\bar{Q}})) \\ &\doteq t(\underset{\sim}{\bar{Q}}) + \varepsilon \underset{\sim}{\nabla}_{\hat{F}} \cdot [\underset{\sim}{Q}(x) - \underset{\sim}{\bar{Q}}] . \end{aligned}$$

Therefore (6.18) gives

$$\widehat{SD}_{IJ} = \left[\nabla_{\hat{F}} \frac{\Sigma[Q(x_i) - \bar{Q}]' [Q(x_i) - \bar{Q}]}{n} \nabla_{\hat{F}}' / n \right]^{1/2}$$

$$= [\nabla_{\hat{F}} \beta_{\hat{F}} \nabla_{\hat{F}}' / n]^{1/2} ,$$

agreeing with the delta method estimate (6.24). Here we have used the fact that

$$\frac{\Sigma[Q(x_i) - \bar{Q}]' [Q(x_i) - \bar{Q}]}{n} = \mathrm{Cov}_{F=\hat{F}} Q ,$$

the covariance of Q under $\hat{F}$, and so must equal $\beta_{\hat{F}}$. Likewise

$$(\ldots, \partial t/\partial Q_a, \ldots)\Big|_{Q=\bar{Q}} = (\ldots, \partial t/\partial Q_a, \ldots)\Big|_{Q=\alpha_{\hat{F}}} = \nabla_{\hat{F}} . \square$$

By comparing (6.18), (6.19) with (6.23), (6.24) the reader can see how Theorem 6.2 works for the correlation coefficient $\hat{\rho}$. Both (6.19) and (6.23) are difficult calculations, fraught with the possibility of error, and it is nice to know that they can be circumvented by numerical methods, as commented at the end of Chapter VI, Section 3.

The infinitesimal jackknife works by perturbing the weights $1/n$ which define $\hat{F}$, keeping the x_i fixed, and seeing how $\hat{\theta}$ varies. The delta method perturbs the x_i (which only affect $\hat{\theta}$ through the $\bar{Q}_a$ in form (6.21)), keeping the weights constant. It is reassuring to see that the results are identical. In this sense there is only one nonparametric delta method.

So far we have discussed the delta method in a nonparametric framework. Suppose though that F is known to belong to a parametric family, say $\mathcal{F}$,

$$\mathcal{F} = \{F_\theta:\ \theta \in \Theta\}\ , \tag{6.25}$$

Θ a subset of R^p. Write $\underset{\sim}{\alpha}_\theta$, $\underset{\sim}{\beta}_\theta$, and $\underset{\sim}{\nabla}_\theta$ in place of $\underset{\sim}{\alpha}_{F_\theta}$, $\underset{\sim}{\beta}_{F_\theta}$, $\underset{\sim}{\nabla}_{F_\theta}$. The *parametric delta method* estimate of standard deviation for a statistic $t(\underset{\sim}{\bar{Q}})$ is

$$\hat{SD} = [\underset{\sim}{\nabla}_{\hat\theta}\underset{\sim}{\beta}_{\hat\theta}\underset{\sim}{\nabla}'_{\hat\theta}/n]^{1/2}\ , \tag{6.26}$$

and is closely related to the parametric bootstrap of Chapter V, Section 2. Here $\hat\theta$ is the MLE of θ. Without pursuing the details, we mention that (6.26) can be applied to the case where $t(\underset{\sim}{\bar{Q}})$ is the MLE of some parameter, and results in the familiar Fisher information bound for the standard deviation of a maximum likelihood estimator. Jaeckel (1972) discusses this point.

6. Estimates of Bias. Figure 6.1 also helps relate the jackknife and bootstrap estimates of bias.

Theorem 6.3. Let $\hat\theta_{QUAD}(\underset{\sim}{P}^*)$ be any quadratic function $a + (\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{b} + \frac{1}{2}(\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{c}(\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)'$, a a constant, $\underset{\sim}{b}$ a column vector, $\underset{\sim}{c}$ a symmetric matrix, satisfying

$$\hat\theta_{QUAD}(\underset{\sim}{P}^\circ) = \hat\theta(\underset{\sim}{P}^\circ) \quad \text{and} \quad \hat\theta_{QUAD}(\underset{\sim}{P}_{(i)}) = \hat\theta(\underset{\sim}{P}_{(i)}\ , \quad i=1,\ 2,\ \ldots,\ n\ . \tag{6.27}$$

Then

$$\widehat{BIAS}_{JACK}(\hat\theta) = \frac{n}{n-1}\,[E_*\hat\theta_{QUAD}(\underset{\sim}{P}^*) - \hat\theta_{QUAD}(\underset{\sim}{P}^\circ)]\ . \tag{6.28}$$

In other words, the jackknife estimate of bias equals $\frac{n}{n-1}$ times the bootstrap estimate of bias for any quadratic function agreeing with $\hat\theta$ at $\underset{\sim}{P}^\circ$ and $\underset{\sim}{P}_{(1)}, \underset{\sim}{P}_{(2)}, \cdots, \underset{\sim}{P}_{(n)}$.

Proof. We can always rewrite a quadratic function $\hat\theta_{QUAD}(\underset{\sim}{P}^*) = a + (\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{b} + \frac{1}{2}(\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{c}(\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)'$ so that

$$\underset{\sim}{P}^{\circ}\underset{\sim}{b} = 0 , \quad \underset{\sim}{P}^{\circ}\underset{\sim}{c} = 0 . \tag{6.29}$$

(If (6.29) is not satisfied, replace $\underset{\sim}{b}$ with $\underset{\sim}{b}-b_{\cdot}\underset{\sim}{1}$, where $b_{\cdot} = \underset{\sim}{P}^{\circ}\underset{\sim}{b}$ and $\underset{\sim}{1} = (1, 1, \ldots, 1)'$; replace $\underset{\sim}{c}$ with $\underset{\sim}{c} - \underset{\sim}{c}_{\cdot}\underset{\sim}{1}' - \underset{\sim}{1}\underset{\sim}{c}_{\cdot}' + \underset{\sim}{1}c_{\cdot\cdot}\underset{\sim}{1}'$, where $\underset{\sim}{c}_{\cdot} = \underset{\sim}{c}\underset{\sim}{P}^{\circ\prime}$, $c_{\cdot\cdot} = \underset{\sim}{P}^{\circ}\underset{\sim}{c}\underset{\sim}{P}^{\circ\prime}$.) From (6.7) we compute

$$E_{*}\hat{\theta}_{QUAD}(\underset{\sim}{P}^{*}) - \hat{\theta}_{QUAD}(\underset{\sim}{P}^{\circ}) = \frac{1}{2}\operatorname{tr}\underset{\sim}{c}(\underset{\sim}{I}/n^2 - \underset{\sim}{P}^{\circ\prime}\underset{\sim}{P}^{\circ}/n) = \frac{1}{2}\operatorname{tr}\underset{\sim}{c}/n^2 \tag{6.30}$$

$$= \frac{1}{2}\sum_{i=1}^{n} c_{ii}/n^2 .$$

By (6.27)

$$\hat{\theta}(\underset{\sim}{P}_{(i)}) - \hat{\theta}(\underset{\sim}{P}^{\circ}) = \hat{\theta}_{QUAD}(\underset{\sim}{P}_{(i)}) - \hat{\theta}_{QUAD}(\underset{\sim}{P}^{\circ}) = -\frac{b_i}{n-1} + \frac{1}{2}\frac{c_{ii}}{(n-1)^2} , \tag{6.31}$$

the last result following from $\underset{\sim}{P}_{(i)} - \underset{\sim}{P}^{\circ} = (\underset{\sim}{P}^{\circ} - \underset{\sim}{\delta}_i)/(n-1)$ and (6.29). Averaging (6.31) over i, and using $b_{\cdot} = \underset{\sim}{P}^{\circ}\underset{\sim}{b} = 0$ again, gives

$$\widehat{\mathrm{BIAS}}_{JACK}(\hat{\theta}) = \frac{1}{2}\frac{\Sigma\, c_{ii}}{n(n-1)} . \tag{6.32}$$

Comparing (6.32) with (6.30) verifies (6.28). □

We have seen, at (6.8), (6.9), that a quadratic functional statistic $\theta(\hat{F})$ is also a quadratic function $\hat{\theta}(\underset{\sim}{P}^{*})$. In this case we can take $\hat{\theta}_{QUAD} = \hat{\theta}$, so (6.28) becomes

$$\widehat{\mathrm{BIAS}}_{JACK}(\hat{\theta}) = \frac{n}{n-1}\,\mathrm{BIAS}_{BOOT}(\hat{\theta})$$

for $\hat{\theta}$ quadratic, which is Theorem 5.1. The factor $\frac{n}{n-1}$ makes $\widehat{\mathrm{BIAS}}_{JACK}$ unbiased for the true bias of a quadratic functional, Theorem 2.1. Notice the similarity of this result to Theorem 6.1.

The infinitesimal jackknife and nonparametric delta method give identical estimates of bias, just as in Theorem 6.2. These estimates are $\Sigma V_{ii}/2n^2$ and $\frac{1}{2}\,\mathrm{tr}\,\underset{\sim}{\hat{\beta}}_F\underset{\sim}{\hat{\nabla}}^2_F$ respectively, where $\underset{\sim}{\nabla}^2_F$ is the A×A matrix with ab-th element

$$\left.\frac{\partial^2 t}{\partial Q_a \partial Q_b}\right|_{\underset{\sim}{Q}=\underset{\sim}{\alpha}_F} ,$$

and

$$V_{ii} = \left.\frac{d^2\hat{\theta}(\underset{\sim}{P}^\circ + \varepsilon(\delta_i - \underset{\sim}{P}^\circ))}{d\varepsilon^2}\right|_{\varepsilon=0} .$$

A proof is given in Gray, Schucany, and Watkins (1975). If we can expand $\hat{\theta}(\underset{\sim}{P}^*)$ in a Taylor series about $\underset{\sim}{P}^\circ$,

$$\hat{\theta}(\underset{\sim}{P}^*) = \hat{\theta}(\underset{\sim}{P}^\circ) + (\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{U} + \frac{1}{2}(\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)\underset{\sim}{V}(\underset{\sim}{P}^* - \underset{\sim}{P}^\circ)' + \ldots ,$$

then stopping the series after the quadratic term gives a quadratic approximation $\hat{\theta}_Q(\underset{\sim}{P}^*)$ to $\hat{\theta}(\underset{\sim}{P}^*)$. The infinitesimal jackknife estimate of bias equals the bootstrap estimate of bias for $\hat{\theta}_Q$,

$$\widehat{\mathrm{BIAS}}_{IJ}(\hat{\theta}) = \mathrm{BIAS}_{BOOT}(\hat{\theta}_Q) ,$$

just as at (6.15).

7. More General Random Variables. So far we have considered functional statistics $\theta(\hat{F})$. More generally we might be interested in a random quantity

$$R(\hat{F},F) , \tag{6.33}$$

for example the Kolmogorov-Smirnov test statistic, on $\mathcal{X} = \mathcal{R}^1$,

$$\sup_x \left| \frac{\#\{X_i \le x\}}{n} - \mathrm{Prob}_F\{X \le x\} \right| .$$

The resampled quantity R^* corresponding to R is

$$R^* = R(\hat{F}^*, \hat{F}) = R(P^*) \ . \tag{6.34}$$

Here $\hat{F}^*$ is the reweighted empirical distribution (6.2). The shorthand notation $R(\underset{\sim}{P}^*)$ tacitly assumes that $x_1, x_2, \ldots, x_n$ are fixed at their observed values.

The curved surface in Figure 6.1 is now $R(\underset{\sim}{P}^*)$ rather than $\hat{\theta}(\underset{\sim}{P}^*)$. The bootstrap estimate of any quantity of interest, such as $E_F R(\hat{F},F)$ or $\text{Prob}_F\{R(\hat{F},F) > \frac{2}{\sqrt{n}}\}$, is the corresponding quant ty computed under (6.6), e.g. $E_* R(\underset{\sim}{P}^*)$ or $\text{Prob}_*\{R(\underset{\sim}{P}^*) > 2/\sqrt{n}\}$. Jackknife approximations can be used as before, to reduce the bootstrap computations,

$$SD_*(R(\underset{\sim}{P}^*)) \doteq \left[\frac{n-1}{n} \sum_{i=1}^{n} [R_{(i)} - R_{(\cdot)}]^2\right]^{1/2} \ , \tag{6.35}$$

and

$$E_* R(\underset{\sim}{P}^*) - R(\underset{\sim}{P}^\circ) \doteq (n-1)(R_{(\cdot)} - R(\underset{\sim}{P}^\circ)) \ , \tag{6.36}$$

$R_{(i)} = R(\underset{\sim}{P}_{(i)})$, $R_{(\cdot)} = \Sigma R_{(i)}/n$.

The justification of (6.35), (6.36) is the same as in Theorems 6.1 and 6.3: If $R_{LIN}(\underset{\sim}{P}^*)$ is the linear function of $\underset{\sim}{P}^*$ agreeing with $R(\underset{\sim}{P}^*)$ for $\underset{\sim}{P}^* = \underset{\sim}{P}_{(i)}$, i=1, 2, ..., n, then the right side of (6.35) equals $[\frac{n}{n-1}]^{1/2}$ $SD_*(R_{LIN}(\underset{\sim}{P}^*))$. Likewise, the right side of (6.36) equals $\frac{n}{n-1}$ $[E_* R_{QUAD}(\underset{\sim}{P}^*) - R(\underset{\sim}{P}^\circ)]$, where $R_{QUAD}(\underset{\sim}{P}^*)$ is any quadratic function agreeing with $R(\underset{\sim}{P}^*)$ at $\underset{\sim}{P}^\circ$, $\underset{\sim}{P}_{(1)}$, $\underset{\sim}{P}_{(2)}$, ..., $\underset{\sim}{P}_{(n)}$. We could use the more direct approximation

$$SD_*(R(\underset{\sim}{P}^*)) \doteq SD_*(R_{LIN}(\underset{\sim}{P}^*)) = \left[(\frac{n-1}{n})^2 \Sigma[R_{(i)} - R_{(\cdot)}]^2\right]^{1/2} \ ,$$

instead of (6.35), and make the corresponding change in (6.36), but won't do so in what follows.

The infinitesimal jackknife also yields approximations to the bootstrap standard deviation and expectation,

$$SD_*(R(\underset{\sim}{P}^*)) \doteq [\Sigma U_i^2/n^2]^{1/2} , \quad E_*R(\underset{\sim}{P}^*) - R(\underset{\sim}{P}^\circ) = \Sigma V_{ii}/2n . \tag{6.37}$$

Here U_i and V_{ii} are defined as before, with $R(\underset{\sim}{P}^\circ + \varepsilon(\delta_i - \underset{\sim}{P}^\circ))$ replacing $\hat{\theta}(\underset{\sim}{P}^\circ + (\delta_i - \underset{\sim}{P}^\circ))$.

What happens if we consider variables not of the functional form $R(\hat{F},F)$? As an example, consider $R = a_n + b_n\hat{\theta} - \theta$, where $\hat{\theta} = \theta(\hat{F})$ is a functional statistic, $\theta = \theta(F)$, and a_n and b_n are constants, $\lim a_n = 0$, $\lim b_n = 1$. The bootstrap estimate of bias (for $a_n + b_n\hat{\theta}$ as an estimate of θ), as discussed in Chapter V, Section 5, is

$$E_*R^* = a_n + b_n E_*\hat{\theta}^* - \hat{\theta} = a_n + (b_n - 1)\hat{\theta} + b_n(E_*\hat{\theta}^* - \hat{\theta})$$

$$= a_n + (b_n - 1)\hat{\theta} + b_n \widehat{BIAS}_{BOOT}(\hat{\theta}) ,$$

compared with the true bias

$$a_n + (b_n - 1)\theta + b_n(E_F\hat{\theta} - \theta) = a_n + (b_n - 1)\theta + b_n \, Bias(\hat{\theta}) . \tag{6.38}$$

Define

$$A_n = na_n , \quad B_n = nb_n .$$

Then the jackknife estimate of bias reduces to

$$a_n + (b_n - 1)\hat{\theta} + b_{n-1} \widehat{BIAS}_{JACK}(\hat{\theta}) + [(A_{n-1} - A_n) + (B_{n-1} - B_n + 1)\hat{\theta}] . \tag{6.39}$$

The situation is quite delicate: if $a_n = c/n$ and $b_n = 1 - d/n$, then $A_{n-1} - A_n = 0$, $B_{n-1} - B_n = 0$, and (6.39) agrees nicely with (6.38). On the other hand, if $a_n = (-1)^n a/n$, $b_n = 1 - d/n$, expansion (6.39) can completely disagree with (6.38). The jackknife estimate of bias is not recommended for statistics other than functional form.

VII. CROSS-VALIDATION, THE JACKKNIFE, AND THE BOOTSTRAP

Cross-validation is an old idea whose time seems to have come again with the advent of modern computers. The original method goes as follows. We are interested in fitting a regression model to a set of data, but are not certain of the model's form, e.g. which predictor variables to include, whether or not to make a logarithmic transform on the response variable, which interactions to include if any, etc. The data set is randomly divided into two halves, and the first half used for model fitting. Anything goes during this phase of the procedure, including hunches, preliminary testing, looking for patterns, trying large numbers of different models, and eliminating "outliers".

The second phase is the cross-validation: the regression model fitted to the first half of the data is used to predict the second half. Typically the model does less well predicting the second half than it did predicting the first half, upon which it was based. The first half predictions are over optimistic, often strikingly so, because the model has been selected to fit the first half data.

There is no need to divide the data set into equal halves. These days it is more common to leave out one data point at a time, fit the model to the remaining points, and see how well the fited model predicts the excluded point. The average of the prediction errors, each point being left out once, is the cross-validated measure of prediction error. See Stone (1974) or Geisser (1975) for a full description, and Wahba and Wold (1975) for a nice application.

This form of cross-validation looks like the jackknife in the sense that data points are omitted one at a time. However, there is no obvious statistic $\hat{\theta}$ being jackknifed, and any deeper connection between the two data ideas has been firmly denied in the literature.

This chapter discusses cross-validation, the jackknife, and the bootstrap, in the regression context given above. It turns out that all three ideas are closely connected in theory, though not necessarily in their practical consequences. The concept of "excess error", vaguely suggested above, is formally defined in Chapter VII, Section 1. (In a discriminant analysis, for example, excess error is the difference between the true and apparent rate of classification errors.) The bootstrap estimate of excess error is easily obtained. Then the jackknife approximation to the bootstrap estimate is derived, and seen to be closely related to the cross-validated estimate.

1. Excess Error. In a regression problem the data consists of pairs (T_1,Y_1), (T_2,Y_2), ..., (T_n,Y_n), where T_i is a $1\times p$ vector of predictors and Y_i is a real-valued response variable. In this discussion we take the simpler point of view mentioned at the end of Chapter V, Section 7: that the $X_i = (T_i,Y_i)$ can be thought of as independent random quantities from an unknown distribution F on $\mathcal{X} = \mathcal{R}^{p+1}$,

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F . \tag{7.1}$$

We observe $X_1=x_1$, $X_2=x_2$, ..., $X_n=x_n$, and denote $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$, $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$.

Having observed $\underset{\sim}{X}=\underset{\sim}{x}$, we have in mind some method of fitting a regression surface, which will then be used to predict future values of the response variable, say

$$\text{predicted value of } y_o = \eta_{\underset{\sim}{x}}(t_o) \ . \tag{7.2}$$

(The subscript "o" indicates a new point $x_o = (t_o, y_o)$ distinct from $x_1, x_2, \ldots, x_n$.) For example, ordinary linear regression fits the surface $\eta_{\underset{\sim}{x}}(t_o) = t_o\hat{\beta}$ where $\hat{\beta} = (\underset{\sim}{t}'\underset{\sim}{t})^{-1}\,\underset{\sim}{t}'\underset{\sim}{y}$, $\underset{\sim}{t}' = (t_1', t_2', \ldots, t_n')$, $\underset{\sim}{y} = (y_1, y_2, \ldots, y_n)'$. Logistic regression, in which the y_i all equal 0 or 1, fits the surface $\eta_{\underset{\sim}{x}}(t_o) = [1 + e^{-t_o\hat{\beta}}]^{-1}$, where $\hat{\beta}$ maximizes the likelihood function $\prod_{i=1}^{n} [e^{-t_i\beta(1-y_i)}/(1 + e^{-t_i\beta})]$.

Section 6 gives an example of a much more elaborate fitting procedure, involving sequential decisions and a complicated model building process. Cross-validation and the bootstrap are unfazed by such complications. The only restriction we impose is that $\eta_{\underset{\sim}{x}}(\cdot)$ be a functional statistic, which means that it depends on $\underset{\sim}{x}$ through $\hat{F}$, the empirical distribution (2.2); in other words, there exists $\eta(t_o, F)$, not depending on n, such that

$$\eta_{\underset{\sim}{x}}(t_o) = \eta(t_o, \hat{F}) \ . \tag{7.3}$$

All of the common fitting methods, including linear regression and logistic regression, satisfy (7.3). In fact we only need (7.3) to establish the connection between the bootstrap and the other methods. The bootstrap itself requires the weaker condition that $\eta_{\underset{\sim}{x}}(\cdot)$ be symmetrically defined in $x_1, x_2, \ldots, x_n$, and similarly for cross-validation.

Let $Q[y, \eta]$ be a measure of *error* between an observed value y and a prediction η. In ordinary regression theory the common measure is $Q[y,\eta] = (y-\eta)^2$. For logistic regression, in which we are trying to assign probabilities η to dichotomous events y, a familiar choice is

$$Q[y,\eta] = \begin{cases} 0 & \text{if } y=1,\ \eta > \frac{1}{2} \text{ or} \\ & \quad y=0,\ \eta \leq \frac{1}{2} \\ 1 & \text{otherwise} \end{cases} .$$

If events $y_1, y_2, \ldots, y_n$ have been assigned probabilities $\eta_1, \eta_2, \ldots, \eta_n$, then $\Sigma Q[y_i,\eta_i]$ is the number of prediction errors. Efron (1978) discusses other Q functions for this situation.

We will be interested in est mating a quantity called the "expected excess error". Define excess error as the random variable

$$R(\underset{\sim}{X},F) = E_{oF}\, Q[Y_o,\ \eta_{\underset{\sim}{X}}(T_o)] - E_{o\hat{F}}\, Q[Y_o,\ \eta_{\underset{\sim}{X}}(T_o)] . \tag{7.4}$$

The symbol "E_{oF}" indicates expectation over a single new point

$$X_o = (T_o,Y_o) \sim F , \tag{7.5}$$

independent of $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, the data which determines $\eta_{\underset{\sim}{X}}(\cdot)$. ($\underset{\sim}{X}$ is called the "training set" in the discrimination literature.) Likewise "$E_{o\hat{F}}$" indicates expectation over

$$X_o \sim \hat{F} \tag{7.6}$$

independent of $\underset{\sim}{X}$. Neither E_{oF} nor $E_{o\hat{F}}$ averages over $\underset{\sim}{X}$, which is why $R(\underset{\sim}{X},F)$ is written as a function of $\underset{\sim}{X}$. It is a function of F through the term $E_{oF}Q$.

The second term on the right of (7.4) equals

$$E_{o\hat{F}}\, Q[Y_o,\ \eta_{\underset{\sim}{X}}(T_o)] = \frac{1}{n} \sum_{j=1}^{n} Q[Y_j,\ \eta_{\underset{\sim}{X}}(T_j)] ,$$

since $\hat{F}$ puts mass $\frac{1}{n}$ at each point (T_j,Y_j). This is a statistic for which we observe the realization

$$\frac{1}{n}\sum_{j=1}^{n} Q[y_j,\ \eta_{\underset{\sim}{x}}(t_j)] = \frac{1}{n}\sum_{j=1}^{n} Q[y_j,\hat{\eta}_j] \ , \tag{7.7}$$

using the notation

$$\hat{\eta}_j = \eta_{\underset{\sim}{x}}(t_j) \ . \tag{7.8}$$

Statistic (7.7) is the "apparent error". Typically, since $\eta_{\underset{\sim}{x}}(\cdot)$ is fitted to the observed data $\underset{\sim}{x}$, this will be smaller than the "true error" $E_{oF}\, Q[Y_o,\ \eta_{\underset{\sim}{x}}(T_o)]$, which is the expected error if $\eta_{\underset{\sim}{x}}(\cdot)$ is used to predict a new Y_o from its T_o value. We are interested in estimating the expected excess error $E_F R(\underset{\sim}{X},F)$, the expected amount by which the true error exceeds the apparent error. A subtle point arises here: $E_F R$ is not the expected excess error for the regression surface $\eta_{\underset{\sim}{x}}(\cdot)$ actually fitted, but rather the expectation over all potential regression surfaces $\eta_{\underset{\sim}{X}}(\cdot)$. $E_F R$ is like the bias $E_F\hat{\theta}-\theta$, which is an average property of $\hat{\theta}(\underset{\sim}{X})$, not $\hat{\theta}(\underset{\sim}{x})-\theta$ for the particular $\underset{\sim}{x}$ observed. This point is discussed again in Section 4.

Example. Linear Discriminant Analysis. Suppose that in the training set y_j equals 1 or 2 as t_j comes from population 1 or population 2. For example, the t_j may be diagnostic variables on patients who do $(y_j = 1)$ or don't $(y_j = 2)$ require surgery. Given a new t_o we wish to predict the corresponding y_o. Fisher's (estimated) linear discriminant function is

$$\eta_{\underset{\sim}{x}}(t_o) = \hat{\alpha} + t_o\hat{\beta} \ , \tag{7.9}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are calculated as follows. Let $n_1 = \#\{y_j = 1\}$, $n_2 = \#\{y_j = 2\}$,

$$\bar{t}_1 = \sum_{y_j=1} t_j/n_1 , \qquad \bar{t}_2 = \sum_{y_j=2} t_j/n_2 ,$$

and

$$S = [\sum_{j=1}^{n} t_j t_j' - n_1 \bar{t}_1' \bar{t}_1 - n_2 \bar{t}_2' \bar{t}_2]/n .$$

Then

$$\hat{\alpha} = [\bar{t}_1 S^{-1} \bar{t}_1' - \bar{t}_2 S^{-1} \bar{t}_2']/2$$

and

$$\hat{\beta} = (\bar{t}_2 - \bar{t}_1) S^{-1} . \tag{7.10}$$

The linear discriminant function (7.9) divides $\mathcal{R}^p$ into two sets

$$\begin{aligned} G_1(\underset{\sim}{x}) &= \{t_o : \eta_{\underset{\sim}{x}}(t_o) \leq 0\} \\ G_2(\underset{\sim}{x}) &= \{t_o : \eta_{\underset{\sim}{x}}(t_o) > 0\} . \end{aligned} \tag{7.11}$$

If $t_o \in G_2(x)$ then the prediction is made that $y_o = 2$, while if $t_o \in G_1(x)$ the prediction is $y_o = 1$. Why use this procedure? If the two populations are a priori equally likely,

$$\text{Prob}_F\{Y=1\} = \text{Prob}_F\{Y=2\} = \frac{1}{2} , \tag{7.12}$$

and if given $Y=y$, T is multivariate normal with mean vector μ_y and covariance matrix Σ,

$$T|y \sim n_p(\mu_y, \Sigma) , \tag{7.13}$$

then the linear discriminant estimates the Bayes classification procedure. See Efron (1975).

The apparent error in a discriminant analysis problem is the proportion of misclassified points in the training set,

$$\frac{\#\{j:\ t_j \in G_1(\underset{\sim}{x}),\ y_j=2 \quad \text{or} \quad t_j \in G_2(\underset{\sim}{x}),\ y_j=1\}}{n} . \tag{7.14}$$

In other words, we are using the error measure

$$Q[y,\eta] = \begin{cases} 0 & \text{if} \quad y=2,\ \eta > 0 \quad \text{or} \quad y=1,\ \eta \leq 0 \\ 1 & \text{otherwise} . \end{cases} \tag{7.15}$$

Suppose $p=2$, $n = 14$, and the multivariate normal model (7.12), (7.13) is correct: $\mu_1 = -(\frac{1}{2},0)$, $\mu_2 = (\frac{1}{2},0)$, $\Sigma=I$. A Monte Carlo experiment of 1000 trials showed that the expected apparent error was .262 (about 1/4 of the 14 points misclassified) but that the expected true error was .356. In other words, the expected excess error $E_F R = .356 - .262 = .094$. This experiment is discussed further in Section 4; see Table 7.1.

2. Bootstrap Estimate of Expected Excess Error. For convenience we denote an estimate of expected excess error by $\widehat{EEE}$. The bootstrap estimate is easy to write down. Consider a single bootstrap sample $\underset{\sim}{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$, selected as at (5.5). The resampling vector $\underset{\sim}{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$, $P_i^* = \#\{X_j^* = x_i\}/n$ has multinomial distribution (6.6). The bootstrap realization of random variables (7.4) is

$$R^* = R(\underset{\sim}{X}^*,\hat{F}) = E_{o\hat{F}}\, Q[Y_o, \eta_{\underset{\sim}{X}^*}(T_o)] - E_{o\hat{F}^*}\, Q[Y_o, \eta_{\underset{\sim}{X}^*}(T_o)] ,$$

where $\hat{F}^*$ is the distribution putting mass P_i^* on x_i, and $\eta_{\underset{\sim}{X}^*}(\cdot)$ is the regression surface determined by $\underset{\sim}{X}^*$. Writing

$$\hat{\eta}_j^* = \eta_{\underset{\sim}{X}^*}(t_j) \ , \tag{7.16}$$

it is easy to see that

$$R^* = \sum_{j=1}^{n} (P_j^\circ - P_j^*)\, Q[y_j, \hat{\eta}_j^*] \ , \tag{7.17}$$

$P_j^\circ = \frac{1}{n}$ as before. The bootstrap estimate of expected excess error is

$$\widehat{EEE}_{BOOT} = E_* R^* = E_* \sum_{j=1}^{n} (P_j^\circ - P_j^*)\, Q[y_j, \hat{\eta}_j^*] \ , \tag{7.18}$$

$E_* R^*$ indicating expectation under bootstrap resampling (6.6). (Since the data $\underset{\sim}{x}$ is fixed, the $\hat{\eta}_j^*$ are functions of $\underset{\sim}{P}^*$, under the assumption that $\eta_{\underset{\sim}{X}^*}(\cdot)$ is symmetrically defined in $X_1^*, X_2^*, \ldots, X_n^*$.)

Example: Ordinary Linear Regression. As before, following (7.2), $\eta_{\underset{\sim}{x}}(t_o) = t_o(\underset{\sim}{t}'\underset{\sim}{t})^{-1}\,\underset{\sim}{t}'\underset{\sim}{y}$. Define $\underset{\sim}{D}_{\underset{\sim}{P}^*}$ to be the diagonal matrix with ii-th element P_i^*. Then

$$\eta_j^* = t_j(\underset{\sim}{t}'\underset{\sim}{D}_{\underset{\sim}{P}^*}\underset{\sim}{t})^{-1}\,(\underset{\sim}{t}'\underset{\sim}{D}_{\underset{\sim}{P}^*}\underset{\sim}{y}) \ , \tag{7.19}$$

and $R^* = E_* \sum_j (P_j^\circ - P_j^*)(y_j - \eta_j^*)^2$. Notice that $\widehat{EEE}_{BOOT} = E_* R^*$ is minus the correlation, under (6.6), between P_j^* and $(y_j - \eta_j^*)^2$. Since large values of P_j^* tend to decrease $(y_j - \eta_j^*)^2$, it is clear that $\widehat{EEE}_{BOOT}$ should be positive.

3. Jackknife Approximation to the Bootstrap Estimate. The excess risk random variable is of form $R(\hat{F},F)$, (6.33), under assumption (7.3). We can use (6.36) to get a jackknife approximation of $E_* R^*$. Since $R(\underset{\sim}{P}^\circ) = 0$ in this case, (6.36) gives

$$\widehat{EEE}_{JACK} \doteq E_* R^* = (n-1)\, R_{(\cdot)} = \frac{n-1}{n} \sum_{i=1}^{n} R(\underset{\sim}{P}_{(i)}) \ . \tag{7.20}$$

Define

$$\hat{\eta}_{(i)j} = \eta_{\underset{\sim}{x}_{(i)}}(t_j) \ , \tag{7.21}$$

where $\underset{\sim}{x}_{(i)}$ is the data set $x_1, x_2, \ldots, x_{n-1}, x_{n-1}, \ldots, x_n$. Following through (7.4) gives

$$R(P_{(i)}) = \sum_{j=1}^{n} \left(\frac{1}{n} - P_{(i)j}\right) Q[y_j, \hat{\eta}_{(i)j}] = \frac{Q[y_i, \hat{\eta}_{(i)i}]}{n} - \frac{\sum_{j\neq i} Q[y_j, \hat{\eta}_{(i)j}]}{n(n-1)}$$

so

$$\begin{aligned}\widehat{EEE}_{JACK} &= \frac{n-1}{n^2} \sum_i Q[y_i, \hat{\eta}_{(i)i}] - \frac{1}{n^2} \sum_i \sum_{j\neq i} Q[y_j, \hat{\eta}_{(i)j}] \\ &= \frac{1}{n} \sum_i Q[y_i, \hat{\eta}_{(i)i}] - \frac{1}{n^2} \sum_i \sum_j Q[y_j, \hat{\eta}_{(i)j}] \qquad (7.22) \\ &= \frac{1}{n} \sum_i Q[y_i, \hat{\eta}_{(i)i}] - \frac{1}{n} \sum_i \frac{\sum_j Q[y_i, \hat{\eta}_{(j)i}]}{n} \ .\end{aligned}$$

4. Cross-Validation Estimate of Excess Error. The cross-validation estimate of expected excess error is

$$\widehat{EEE}_{CROSS} = \frac{1}{n} \sum_i Q[y_i, \hat{\eta}_{(i)i}] - \frac{1}{n} \sum_i Q[y_i, \hat{\eta}_i] \ , \tag{7.23}$$

the difference in observed error when we don't or do let x_i assist in its own prediction. Notice the similarity between (7.22) and (7.23). Before presenting any theoretical calculations, we give further Monte Carlo results from the experiment described at the end of Section 1.

Table 7.1 shows the first 10 trials of the linear discriminant problem, true distributions normal as at (7.12), (7.13), and summary statistics for 100 and 1000 trials. The sample size is $n = 14$, dimension $p=2$. We

see that $\widehat{EEE}_{JACK}$ and $\widehat{EEE}_{CROSS}$ are almost the same, except in trial 2, with correlation .93 over 1000 trials. Neither method yields useful estimates. The values of $\widehat{EEE}$ are capriciously large or small, with coefficients of variations .073/.091 = .80 and .068/.093 = .73 in 100 trials. The bootstrap estimates, B = 200 bootstrap replications per trial, are much less variable, with coefficients of variation only .028/.080 = .35.

The actual excess error, $R(\underset{\sim}{x},F)$, is given in column A. It is quite variable from trial to trial. In 5 of the first 10 trials (and 22 of the first 100 trials) it is negative, the apparent error being *greater* than the true error. This is not a good situation for bias correction, i.e. for adjusting the apparent error rate by adding an estimate of expected excess error. The last column gives the bootstrap estimates of Sd(R) = .114,

$$SD_{*}R^{*} = \left[\sum_{b=1}^{B} (R^{*b} - R^{**})^2/(B-1)\right]^{1/2} .$$

These estimates are seen to be quite dependable.

In Table 7.2 the situation is more favorable to bias correction. The actual excess error $R(\underset{\sim}{x},F)$ was positive in 98 out of 100 trials, averaging .184. $\widehat{EEE}_{JACK}$ and $\widehat{EEE}_{CROSS}$ are even more highly correlated, .98, and again both are too variable from trial to trial to be useful. The bootstrap estimates $\widehat{EEE}_{BOOT}$ are much less variable from trial to trial, but are biased downward. Adding $\widehat{EEE}_{BOOT}$ to the apparent error rate (7.14) substantially improves estimation of the true error rate in this case. The root mean square error of estimation for the true error rate decreases from .189 (using just (7.14) to .133 (using (7.14) plus $\widehat{EEE}_{BOOT}$) for the first 10 trials. The comparable values for cross-validation and the jackknife are .190 and .183 respectively.

Trial #	n_1	Apparent Error Rate (7.14)	A Actual Excess $R(\underset{\sim}{x},F)$	B Bootstrap Estimate (B=200) $\widehat{EEE}_{BOOT}$	C Cross-Val Estimate $\widehat{EEE}_{CROSS}$	D Jackknife Estimate $\widehat{EEE}_{JACK}$	Bootstrap Std Dev Estimate SD_*R^*
1	9	.286	.172	.083	.214	.214	.117
2	6	.357	-.045	.098	.000	.066	.118
3	7	.357	-.044	.110	.071	.066	.108
4	8	.429	-.078	.107	.071	.066	.111
5	8	.357	-.027	.102	.143	.148	.120
6	8	.143	.175	.073	.214	.194	.094
7	8	.071	.239	.047	.071	.066	.077
8	6	.286	.094	.097	.071	.056	.109
9	7	.429	-.069	.127	.071	.087	.101
10	8	.143	.192	.048	.000	.010	.090
100 Trials Ave		.264	.096	.080	.091	.093	.104
100 Trials (Sd)		(.123)	(.114)	(.028)	(.073)	(.068)	(.014)
100 Trials Ave		.262	.094		.097	.095	
100 Trials (Sd)		(.118)			(.085)	(.074)	

Table 7.1. Expected excess error estimated by the bootstrap, cross-validation, and the jackknife for the linear discriminant problem, n = 14, true situation as in (7.12), (7,13), p=2, $\mu_1 = (-\frac{1}{2}, 0)$, $\mu_2 = (\frac{1}{2}, 0)$, $\sharp$=I. Results for first ten trials, summary statistics for first 100 trials and 1000 trials. Correlation (C,D) = .93, Corr(A,C) = -.07, Corr(A,D) = -.23 (1000 trials); Corr(A,B) = -.64 (100 trials).

Trial #	n_1	Apparent Error Rate (7.14)	A Actual Excess $R(\underset{\sim}{x},F)$	B Bootstrap Estimate (B=200) $\widehat{EEE}_{BOOT}$	C Cross-Val Estimate $\widehat{EEE}_{CROSS}$	D Jackknife Estimate $\widehat{EEE}_{JACK}$	Bootstrap Std Dev Estimate SD_*R^*
1	7	.071	.135	.124	.357	.321	.112
2	5	.214	.010	.159	.357	.342	.102
3	8	.000	.247	.040	.000	.000	.064
4	6	.143	.098	.126	.143	.168	.098
5	6	.143	.101	.132	.286	.276	.090
6	4	.071	.229	.107	.143	.143	.106
7	8	.000	.236	.073	.143	.133	.070
8	8	.071	.142	.120	.357	.342	.082
9	5	.000	.269	.086	.214	.189	.068
10	8	.000	.239	.054	.071	.066	.080
100 Trials { Ave		.069	.184	.103	.170	.167	.087
100 Trials { (Sd)		(.076)	(.100)	(.031)	(.094)	(.089)	(.012)

Table 7.2. Same as Table 7.1, except dimension p=5, $\mu_1 = (-1, 0, 0, 0, 0)$, $\mu_2 = (1, 0, 0, 0, 0)$. Correlation (C,D) = .98, Corr(A,C) = -.15, Corr(A,D) = -.26, Corr(A,B) = -.58 (100 trials).

It would be nice if the estimates $\widehat{EEE}$ correlated well with $R(\underset{\sim}{x},F)$, i.e. if the suggested bias corrections were big or small as the situation called for. In fact somewhat the opposite happens: the correlations are all negative, the bootstrap being most markedly so. Situations that produce grossly overoptimistic apparent errors, such as trial 10 of Table 7.2, tend to have the smallest estimated $\widehat{EEE}$. The author can't explain this phenomenon. Current research focuses on this problem, and on the downward bias of $\widehat{EEE}_{BOOT}$ evident in Table 7.2.

5. Relationship Between the Cross-Validation and Jackknife Estimates.

Under reasonable conditions the expected excess error and the estimates $\widehat{EEE}_{BOOT}$, $\widehat{EEE}_{JACK}$, and $\widehat{EEE}_{CROSS}$ are order of magnitude $O_p(1/n)$, while $\widehat{EEE}_{JACK} - \widehat{EEE}_{CROSS} = O_p(1/n^2)$. We briefly discuss the case of ordinary linear regression with quadratic error, $\eta_{\underset{\sim}{x}}(t_o) = t_o(\underset{\sim}{t}'\underset{\sim}{t})^{-1}\,\underset{\sim}{t}'\underset{\sim}{y}_*$, $Q[y,\eta] = (y-\eta)^2$.

Define

$$r_i = y_i - \hat{\eta}_i \quad \text{and} \quad a_i = t_i(\underset{\sim}{t}'\underset{\sim}{t})^{-1}\,t_i' \,. \tag{7.24}$$

Under the usual regression assumptions r_i is $O_p(1)$, while a_i is $O_p(\frac{1}{n})$. (Notice that $\sum_i a_i = \text{tr } I = p$, so by symmetry $E_F\, a_i = p/n$.) Using the matrix identity $[I - v'v]^{-1} = I + v'v/[1 - vv']$, v a row vector, we can express $\hat{\eta}_{(j)i} = t_i(\underset{\sim}{t}'\underset{\sim}{t} - t_j't_j)^{-1}\,(\underset{\sim}{t}'\underset{\sim}{y} - t_j'y_j)$ in a simple form,

$$\hat{\eta}_i - \hat{\eta}_{(j)i} = t_i(\underset{\sim}{t}'\underset{\sim}{t})^{-1}\,t_j'\,\frac{r_j}{1-a_j} = O_p(\tfrac{1}{n}) \,. \tag{7.25}$$

In particular

$$\hat{\eta}_i - \hat{\eta}_{(i)i} = \frac{a_i}{1-a_i}\,r_j \,. \tag{7.26}$$

Letting $\hat{\eta}_{(\cdot)i}$ denote $\sum_j \hat{\eta}_{(j)i}/n$, (7.25) gives

$$\begin{aligned}\hat{\eta}_i - \hat{\eta}_{(\cdot)i} &= t_i (\underset{\sim}{t}'\underset{\sim}{t})^{-1} \frac{1}{n} \sum_j t_j' \left(1 + \frac{a_j}{1-a_j}\right) r_j \\ &= t_i (\underset{\sim}{t}'\underset{\sim}{t})^{-1} \frac{1}{n} \sum_j t_j' \frac{a_j r_j}{1-a_j} = O_p\left(\frac{1}{n^2}\right) .\end{aligned} \tag{7.27}$$

(We have used the orthogonality condition $\Sigma\, t_j' r_j = 0$.)

From (7.23),

$$\widehat{EEE}_{CROSS} = \frac{1}{n} \sum_i [(r_i + \hat{\eta}_i - \hat{\eta}_{(i)i})^2 - r_i^2] = \frac{2}{n} \sum_i r_i (\hat{\eta}_i - \hat{\eta}_{(i)i}) + \frac{1}{n} \sum_i (\hat{\eta}_i - \hat{\eta}_{(i)i})^2$$

$$= \frac{2}{n} \sum_i \frac{a_i}{1-a_i} r_i^2 + \frac{1}{n} \sum_i \left(\frac{a_i}{1-a_i} r_i\right)^2 = O_p(\tfrac{1}{n}) .$$

Comparing (7.23) with (7.22) gives

$$\widehat{EEE}_{JACK} - \widehat{EEE}_{CROSS} = \frac{1}{n} \sum_i (y_i - \hat{\eta}_i)^2 - \frac{1}{n} \sum_i (y_i - \hat{\eta}_{(\cdot)i})^2$$

$$= -\frac{2}{n} \sum_i r_i (\hat{\eta}_i - \hat{\eta}_{(\cdot)i}) + \frac{1}{n^2} \sum_i \sum_j (\hat{\eta}_i - \hat{\eta}_{(j)i})^2$$

$$= O_p\left(\frac{1}{n^2}\right) .$$

There is nothing special about ordinary linear regression in these calculations, except the tractability of the results. It seems likely that $(\widehat{EEE}_{JACK} - \widehat{EEE}_{CROSS})/\widehat{EEE}_{CROSS}$ is $O_p(\frac{1}{n})$ under quite general conditions, but this remains to be proved. The results in Table 7.1, 7.2 are encouraging.

6. A Complicated Example. Figure 7.1 shows[†] a decision tree for classifying heart attack patients into low risk of dying (population 1) or high risk of dying (population 2) categories. A series of binary decisions brings a patient down the tree to a terminal node, which predicts either class 1 (e.g. node T1) or class 2 (e.g. node T9). For example, a patient with small PKCK value, small MNSBP value, and finally large PKCK value ends up at T3, and is predicted to be in population 1. The numerical definition of "small" and "large" changes from node to node, so there is no contradiction between the first and third criterion of the previous sentence.

The decision tree, which is a highly nonlinear discriminant function, was based on a training set of 389 patients, 359 of whom survived their heart attacks at least 30 days (population 1), and 30 of whom didn't (population 2). Without going into details the tree's construction followed these rules:

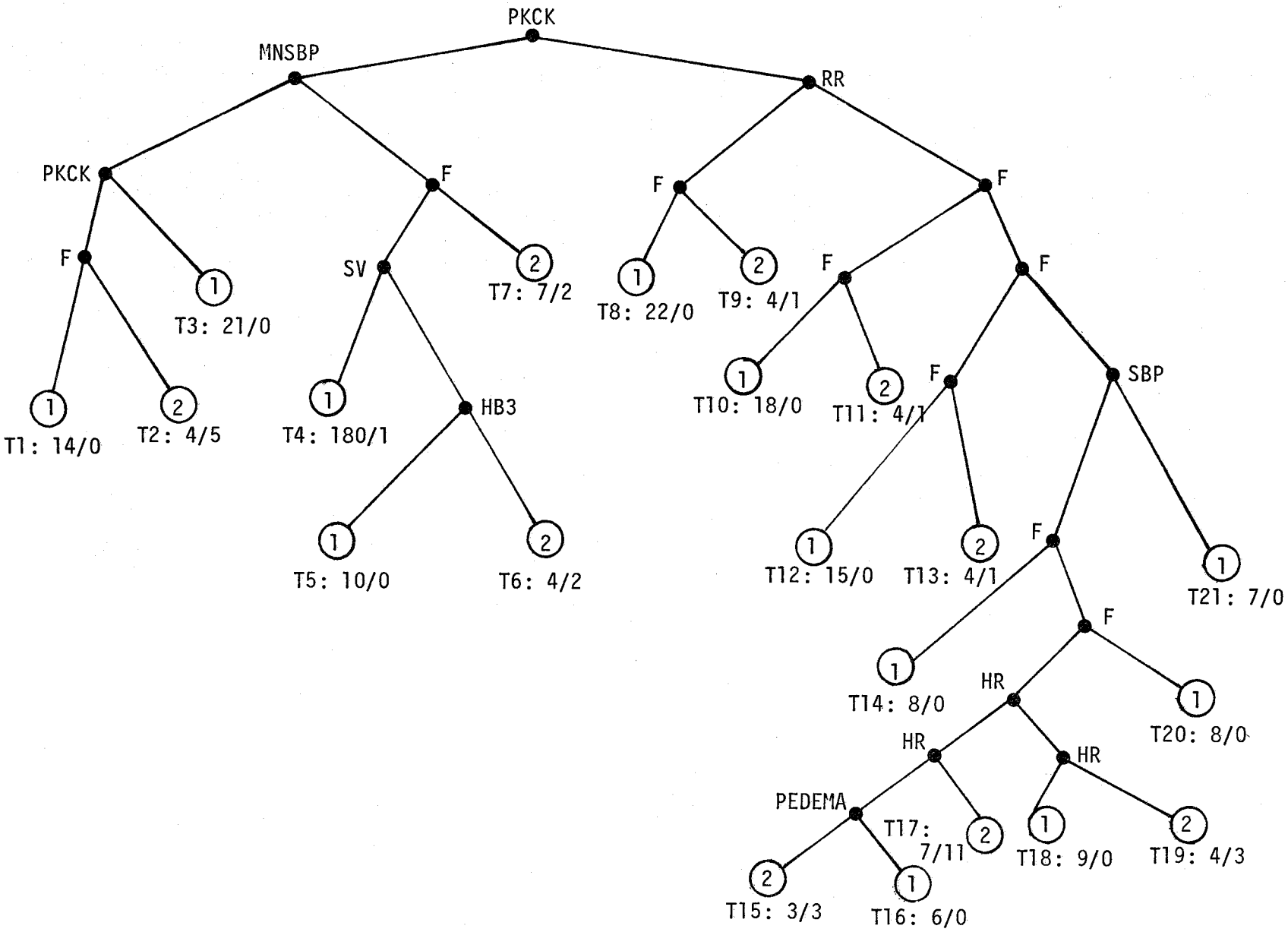
1) The original 389 patients were divided into low and high groups using the best decision variable and the best dividing point for that variable. "Best" here means maximizing the separation between the two populations in a certain quantitative sense. (The ultimate best division would have all the "low" group in population 1 and all the "high" group in population 2, or vice-versa, in which case we could predict the training

[†]This data comes from the Special Center of Research on Ischemic Heart Disease, UC San Diego, investigators John Ross, Jr., Elizabeth Gilpin, Richard Olshen, and H. E. Henning, University of British Columbia. The tree was constructed by R. Olshen, who also performed the bootstrap analysis. It is a small part of a more extensive investigation.

(Please see following page for figure.)

Figure 7.1. A decision tree for classifying heart attack patients into low risks of dying (population 1) or high risk of dying (population 2). Smaller values of the decision variables go to the left. Circled numbers at terminal nodes indicate population prediction. For example, 6 of the 389 patients in the training set end up at T6, 4 from population 1, 2 from population 2; these patients would all be predicted to be in population 2.

Abbreviations: PKCK - peak creatinine kinease level; MNSBP - minimum systolic blood pressure; SBP - systolic blood pressure; RR - respiration rate; HR - average heart rate; SV - superventricular arrhythemia; HB3 - heart block 3rd degree; PEDEMA - peripheral edema; F - Fisher linear discriminant function, differing from node to node.


PKCK
MNSBP
RR
PKCK
F
F
F
F
F
SV
F
F
SBP
HB3
F
HR
HR
HR
PEDEMA
T1: 14/0
T2: 4/5
T3: 21/0
T4: 180/1
T5: 10/0
T6: 4/2
T7: 7/2
T8: 22/0
T9: 4/1
T10: 18/0
T11: 4/1
T12: 15/0
T13: 4/1
T14: 8/0
T15: 3/3
T16: 6/0
T17: 7/11
T18: 9/0
T19: 4/3
T20: 8/0
T21: 7/0

set perfectly on the basis of this one division, but of course that wasn't possible.) A computer search of all possible division points on each of 19 division variables selected PKCK, peak creatiame kinease level, to make the first division. Certain linear combinations of the variables, labelled "F" in Figure 7.1, were also examined in the search for the best division.

2) Step 1 was repeated, separately, to best subdivide the high and low PKCK groups. (The low PKCK group was divided into low and high MNSBP groups; the high PKCK group into low and high RR groups.) The process was iterated, yielding a sequence of "complete" trees, the k-th of which had 2^k terminal nodes. All subtrees of the complete trees were also considered. For example, Figure 7.1 is a subtree of the 10th complete tree; it has 21 terminal nodes rather than the complete set of 2^{10}.

3) A terminal node was said to predict population 1 if $n_1/n_2 \geq 8$, where n_i was the number of members of population i, in the training sample, at that node. If $n_1/n_2 < 8$ the prediction was population 2.

4) The tree in Figure 7.1 predicts 41 population 1 patients into population 2 (nodes T2, T6, T7, T9, T11, T13, T15, T17, T19), and 1 population 2 patient into population 1 (node T4). The apparent error rates are 41/359 = 11.5% for population 1, 1/30 = 3.33% for population 2, and 42/389 = 10.8% overall. This tree was selected as best because it minimized the quantity {overall apparent error rate + k · number of terminal nodes}, k a certain constant.

These rules are not ad hoc; they are based on considerable theoretical work, see Gordon and Olshen (1978). On the other hand, they are far too complicated for standard analysis. Instead, a bootstrap analysis was run, as in Section 2. Only B=3 bootstrap replications of (7.17) were

generated, but these agreed closely with each other: the estimate $\widehat{EEE}_{BOOT}$ equaled 6.1%, making the bias corrected estimate of error rate 10.8% + 6.1% = 16.9%. More seriously, the bias corrected estimated error rate for the population 2 patients was 30%, compared to the apparent error rate of 3.33%! The tree in Figure 7.1 does not predict population 2 patients, those with a high risk of dying, nearly as well as it appears to.

VIII. BALANCED REPEATED REPLICATIONS (HALF-SAMPLING)

Half-sampling methods come from the literature of sampling theory. The basic idea is almost identical to the bootstrap estimate of standard deviation, but with a clever shortcut method, balanced repeated replications, that we haven't seen before. Kish and Frankel (1974) give a thorough review of the relevant sample survey theory.

In sampling theory it is natural to consider stratified situations where the sample space $\mathcal{X}$ is a union of disjoint strata $\mathcal{X}_h$,

$$\mathcal{X} = \bigcup_{h=1}^{H} \mathcal{X}_h \ .$$

(For example, $\mathcal{X}$ = United States, and $\mathcal{X}_h$ = State h, h=1, 2, ..., 50.) The data consists of separate i.i.d. samples from each stratum,

$$X_{hi} \overset{iid}{\sim} F_h \ , \quad i=1, 2, \ldots, n_h \ , \quad h=1, 2, \ldots, H \ , \tag{8.1}$$

where F_h is an unknown probability distribution on $\mathcal{X}_h$. Having observed $X_{hi} = x_{hi}$ i=1, ..., n_h, h=1, ..., H, define

$$\hat{F}_h\text{: mass } \frac{1}{n_h} \text{ on } x_{hi} \ , \tag{8.2}$$

the empirical probability distribution for stratum h, h=1, 2, ..., H.

The goal of half-sampling theory is to assign an estimate of standard deviation to a functional statistic

$$\hat{\theta} = \theta(\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_H) \ . \tag{8.3}$$

For example $\hat{\theta}$ might be a linear functional statistic

$$\hat{\theta} = \sum_{h=1}^{H} \{\mu_h + \frac{1}{n_h} \sum_{i=1}^{n_h} \alpha_h(X_{hi})\} , \tag{8.4}$$

μ_h and $\alpha_h(\cdot)$ known, though possibly different for different strata. As another example, suppose $\mathcal{X}_h = \mathcal{R}^2$, h=1, ..., H, and that π_h are known probabilities, $\sum_{h=1}^{H} \pi_h = 1$. The *mixture* $\sum_h \pi_h \cdot \hat{F}_h$ is a distribution on $\mathcal{R}^2$ putting mass π_h/n_h on each x_{hi}. In particular, if $\pi_h = n_h/n$ then $\sum_h \pi_h \cdot \hat{F}_h = \hat{F}$, (2.2). The sample correlation can be written $\rho(\hat{F}) = \rho \sum_h \frac{n_h}{n} \cdot \hat{F}_h$. In this way any functional statistic $\hat{\theta} = \theta(\hat{F})$ can be written in form (8.3). In the case of the correlation, we might prefer the statistic $\rho(\sum_h n_h/n \cdot \tilde{F}_h)$, where $\tilde{F}_h$ puts mass $1/n_h$ on $x_{hi}-\bar{x}_h$, $\bar{x}_h = \sum_i x_{hi}/n_h$. This is of form (8.3), but not of form (2.2).

1. Bootstrap Estimate of Standard Deviation. The obvious bootstrap estimate of $Sd(\hat{\theta})$ is obtained, as at (5.4) - (5.6), by the following algorithm:

1. Construct the $\hat{F}_h$, (8.2),

2. Draw independent bootstrap samples $X^*_{hi} \overset{iid}{\sim} \hat{F}_h$, i=1, ..., n_h, h=1, ..., H; let $\hat{F}^*_h$ be the distribution putting mass P^*_{hi} on x_{hi}, where $P^*_{hi} = \#\{X^*_{hj} = x_{hi}\}/n_h$; and let $\hat{\theta}^* = \theta(\hat{F}^*_1, \hat{F}^*_2, \ldots, \hat{F}^*_H)$.

3. Independently repeat step (2) B times, obtaining bootstrap replications $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$, and estimate $Sd(\hat{\theta})$ by (5.6),

$$\hat{SD}_{BOOT} = \left\{ \frac{\sum_{b=1}^{B} [\hat{\theta}^{*b} - \hat{\theta}^{*\cdot}]^2}{B-1} \right\}^{1/2} . \tag{8.5}$$

As before, $\hat{SD}_{BOOT}$ is really defined as the limit of (8.5) as $B \to \infty$, but in most cases we have to settle for some finite value like B = 50 or 100.

In the case of a linear statistic (8.4), where we can take $B=\infty$ without actually using Monte Carlo sampling, standard theory shows that

$$\hat{SD}_{BOOT} = \left[\sum_{h=1}^{H} \frac{\hat{\sigma}_h^2}{n_h}\right]^{1/2} , \tag{8.6}$$

$\hat{\sigma}_h^2$ being the h-th sample variance

$$\hat{\sigma}_h^2 = \sum_{i=1}^{n_h} [\alpha_{hi}-\alpha_{h\cdot}]^2/n_h , \tag{8.7}$$

$\alpha_{hi} = \alpha_h(x_{hi})$, $\alpha_{h\cdot} = \sum_i \alpha_{hi}/n_h$. This compares with the true standard deviation

$$Sd(\hat{\theta}) = \left[\sum_{h=1}^{H} \frac{\sigma_h^2}{n_h}\right]^{1/2} , \quad \sigma_h^2 = Var\ \alpha_h(X_{hi}) . \tag{8.8}$$

2. Half-Sample Estimate of Standard Deviation. The expected value of $\widehat{VAR}_{BOOT} = \hat{SD}_{BOOT}^2$ equals

$$E_{F_1,F_2,\ldots,F_H} \sum_h \frac{\hat{\sigma}_h^2}{n_h} = \sum_h \frac{n_h-1}{n_h}\frac{\sigma_h^2}{n_h} , \tag{8.9}$$

compared to the true value $Var\ \hat{\theta} = \Sigma\ \sigma_h^2/n_h$. Previously we have ignored the downward bias in (8.9), but in sampling theory the n_h are often small, and the bias can be severe. In particular, if all the $n_h = 2$, the case most often considered in half-sampling theory, then $E\ \widehat{VAR}_{BOOT} = \frac{1}{2} Var(\hat{\theta})$.

The *half-sample*, or *repeated replications*, estimate of standard deviation, $\hat{SD}_{HS}$, is the same as the bootstrap estimate, except that at step (2) of the algorithm we choose samples of size n_h-1 instead of n_h, $X_{hi}^* \overset{iid}{\sim} \hat{F}_h$, i=1, 2, ..., n_h-1, h=1, ..., H. Reducing the size of the bootstrap samples by 1 removes the bias in $\widehat{VAR}$ when $\hat{\theta}$ is linear since then

$$\hat{SD}_{HS} = \left[\sum_{h=1}^{H} \frac{\hat{\sigma}_h^2}{n_h - 1} \right]^{1/2} ,$$

and so $E\,\widehat{VAR}_{HS} = \sum_h \sigma_h^2/n_h = \text{Var}\,\hat{\theta}$.

Suppose all the $n_h = 2$. Then the half-sample method really chooses half-samples. One of the two data points from each stratum is selected to be the bootstrap observation, independently and with equal probability,

$$X_{h1}^* = \begin{cases} x_{h1} & \frac{1}{2} \\ & \text{prob}_* \\ x_{h2} & \frac{1}{2} \end{cases} , \quad \text{independently} \quad h=1, 2, \ldots, H . \qquad (8.10)$$

In other words, each $\hat{F}_h^*$ is a one-point distribution putting all of its mass at either x_{h1} or x_{h2}, with equal probability.

Henceforth we will only discuss the situation where $n_h = 2$ for $h=1, 2, \ldots, H$. There are $n = 2H$ observations in this case and 2^H possible half-samples. Let $\underset{\sim}{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_H)$ be a vector of ± 1's, indicating a half-sample according to the rule

$$X_{h1}^* = \begin{cases} x_{h1} & \text{if} \quad \varepsilon_h = +1 \\ x_{h2} & \text{if} \quad \varepsilon_h = -1 \end{cases} . \qquad (8.11)$$

The set $\mathcal{J}_o$ of all possible vectors $\underset{\sim}{\varepsilon}$ has $J_o = 2^H$ members, each of which is selected with equal probability under half-sampling. Since $\underset{\sim}{\varepsilon}$ determines all the $\hat{F}_h^*$, by (8.11), we can write $\hat{\theta}(\underset{\sim}{\varepsilon})$ in place of $\hat{\theta}(\hat{F}_1^*, \hat{F}_2^*, \ldots, \hat{F}_H^*)$. In this notation, the half-sample estimate of standard deviation is

$$\hat{SD}_{HS} = \{ \sum_{\underset{\sim}{\varepsilon} \in \mathcal{J}_o} [\hat{\theta}(\underset{\sim}{\varepsilon}) - \hat{\theta}(\cdot)]^2/J_o \}^{1/2} , \qquad (8.12)$$

where $\hat{\theta}(\cdot) = \sum_{\underset{\sim}{\varepsilon} \in \mathcal{J}_o} \hat{\theta}(\underset{\sim}{\varepsilon})/J_o$.

Notice that J_o is not what we called "B" before. In fact $B=\infty$, since we have considered all 2^H possible outcomes of $\hat{F}_1^*, \hat{F}_2^*, \ldots, \hat{F}_H^*$. That is why we divide by J_o rather than J_o-1 in (8.12).

Line 8 of Table 5.2 shows half-sampling applied to $\hat{\rho}$ the correlation coefficient, and to $\hat{\phi} = \tanh^{-1} \hat{\rho}$. The strata were defined artificially, (x_1, x_2) representing stratum 1, (x_3, x_4) stratum 2, ..., (x_{13}, x_{14}) stratum 7. For each of the 200 Monte Carlo trials, all 128 half-sample values were evaluated, and $\hat{SD}_{HS}$ calculated according to (8.12). The numerical results are discouraging. Both bias and root mean square error are high, for both $\hat{SD}_{HS}(\hat{\rho})$ and $\hat{SD}_{HS}(\hat{\phi})$. Of course this is not a naturally stratified situation, so there is no particular reason to do half-sampling. On the other hand, it would be nice if the method worked well since, as we shall see, it can be implemented with considerably less computation than the bootstrap.

In line 10 of Table 5.2, the Sd estimates for each of the 200 trials were constructed using 128 randomly selected (out of all $14!/(7!)^2$ possible) half-samples. This method removes the component of variance in $\hat{SD}_{HS}$ due to the artificial creation of strata, but the numerical results are still poor compared to the bootstrap results of line 1.

3. Balanced Repeated Replications. Suppose that $\mathcal{J}$ is a subset of $\mathcal{J}_o$ containing J vectors $\underset{\sim}{\varepsilon}$, say $\mathcal{J} = \{\underset{\sim}{\varepsilon}^1, \underset{\sim}{\varepsilon}^2, \ldots, \underset{\sim}{\varepsilon}^J\}$, and that these vectors satisfy

$$\sum_{j=1}^{J} \epsilon_h^j \epsilon_k^j = 0 , \qquad 1 < h < k \leq H . \tag{8.13}$$

McCarthy (1969) calls $\mathcal{J}$ a *balanced set* of half-samples. We will also require that a balanced set satisfy

$$\sum_{j=1}^{J} \varepsilon_h^j = 0 , \qquad 1 \le h \le H . \tag{8.14}$$

The complete set $\mathcal{J}_o$ is itself balanced.

We define the *balanced half-sample* estimate[†] of standard deviation

$$\hat{SD}_{BHS} = \{ \sum_{j=1}^{J} [\hat{\theta}(\underset{\sim}{\varepsilon}^j) - \hat{\theta}(\cdot)]^2/J \}^{1/2} , \tag{8.15}$$

$\hat{\theta}(\cdot) = \sum_{j=1}^{J} \hat{\theta}(\underset{\sim}{\varepsilon}^j)/J$. This is McCarthy's method of *Balanced Repeated Replications*, and has the advantage of requiring only J instead of $J_o = 2^H$ recomputations of $\hat{\theta}$, while still giving the same Sd estimate for linear statistics.

Theorem 8.1 (McCarthy). For a linear statistic (8.4), $\hat{SD}_{BHS} = \hat{SD}_{HS}$.

Proof. With $\alpha_{hi} = \alpha_h(x_{hi})$, $\alpha_{h\cdot} = (\alpha_{h1}+\alpha_{h2})/2$ as before,

$$\hat{\theta}(\underset{\sim}{\varepsilon}^j) = \sum_{h=1}^{H} \left\{ \mu_h + \alpha_{h\cdot} + \varepsilon_h^j \frac{(\alpha_{h1}-\alpha_{h2})}{2} \right\} = \hat{\theta} + \sum_h \varepsilon_h^j \frac{(\alpha_{h1}-\alpha_{h2})}{2} , \tag{8.16}$$

so $\hat{\theta}(\cdot) = \sum_j \hat{\theta}(\underset{\sim}{\varepsilon}^j)/J = \hat{\theta}$ because of (8.14). Then

$$\hat{SD}_{BHS} = \left\{ \sum_{j=1}^{J} \left[\sum_{h=1}^{H} \varepsilon_h^j \frac{(\alpha_{h1}-\alpha_{h2})}{2} \right]^2 /J \right\}^{1/2} \tag{8.17}$$

$$= \left\{ \sum_{h=1}^{H} \sum_{k=1}^{H} \sum_{j=1}^{J} \varepsilon_h^j \varepsilon_k^j \frac{(\alpha_{h1}-\alpha_{h2})(\alpha_{k1}-\alpha_{k2})}{4} /J \right\}^{1/2} = \left\{ \sum_{h=1}^{H} \left(\frac{\alpha_{h1}-\alpha_{h2}}{2} \right)^2 \right\}^{1/2} ,$$

[†]McCarthy's result is stated with $\hat{\theta}$ replacing $\hat{\theta}(\cdot)$ in (8.12) and (8.15), in which case condition (8.14) is not required. This replacement makes almost no difference in Table 5.2, and probably not in most cases, but if there were a substantial difference, definitions (8.12), (8.15) would be preferred for estimating standard deviation. McCarthy's definition is more appropriate for estimating root mean square error.

using (8.13). This last expression doesn't depend on $\mathcal{J}$, so $\hat{SD}_{BHS}$ for linear statistics is the same for all balanced sets, including $\mathcal{J}_o$, which proves the theorem. □

Line 10 of Table 5.2 gives summary statistics for $\hat{SD}_{BHS}$ applied to the correlation experiment. A balanced set $\mathcal{J}$ with J=8 members was used, so each $\hat{SD}_{BHS}$ required only 8 recomputations of $\hat{\rho}$ (or $\hat{\phi}$), instead of 128 as at line 8. The vectors $\underset{\sim}{\varepsilon}^j$ were the rows of this matrix:

$$\begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 & -1 \\ -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix} . \qquad (8.18)$$

The results are quite similar to those of line 8, as Theorem 8.1 would suggest, though $\sqrt{MSE}$ is somewhat increased.

4. Complementary Balanced Half-Sample. The *complementary half-sample* to that represented by $\underset{\sim}{\varepsilon}$ is $-\underset{\sim}{\varepsilon}$, i.e. the other half of the data. Suppose now that a balanced set $\mathcal{J}$ is also *closed under complementation*, so that if $\underset{\sim}{\varepsilon} \in \mathcal{J}$ then $-\underset{\sim}{\varepsilon} \in \mathcal{J}$. Then J is even, and we can index $\mathcal{J}$ so that each $\underset{\sim}{\varepsilon}^j$, j=1, 2, ..., J/2, is complementary to an $\underset{\sim}{\varepsilon}^{j+J/2}$. (In other words, the second half of $\mathcal{J}$ is complementary to the first half.) The complete balanced set $\mathcal{J}_o$ is closed under complementation.

The *complementary balanced half-sample* estimate of standard deviation is

$$\hat{SD}_{CBHS} = \left\{ \frac{2}{J} \sum_{j=1}^{J/2} \left[\frac{\hat{\theta}(\underset{\sim}{\varepsilon}^j) - \hat{\theta}(-\underset{\sim}{\varepsilon}^j)}{2} \right]^2 \right\}^{1/2} . \tag{8.19}$$

The advantage of SD_{CBHS} is that Theorem 8.1 can now be extended to quadratic statistics. First we have to define what we mean by a quadratic statistic in the stratified context (8.1). The easiest definition is by analogy with (6.8). Let $\underset{\sim}{P}^\circ = (\frac{1}{2}, \frac{1}{2})$ and

$$\underset{\sim h}{P}^j = \begin{cases} (1, 0) & \text{if } \varepsilon_h^j = 1 \\ (0, 1) & \text{if } \varepsilon_h^j = -1 \end{cases} . \tag{8.20}$$

A statistic $\hat{\theta}$ is quadratic if its half-sample values can be expressed as

$$\hat{\theta}(\underset{\sim}{\varepsilon}^j) = \hat{\theta} + \sum_{h=1}^{H} (\underset{\sim h}{P}^j - \underset{\sim}{P}^\circ)\underset{\sim h}{U} + \frac{1}{2} \sum_{h=1}^{H} \sum_{k=1}^{H} (\underset{\sim h}{P}^j - \underset{\sim}{P}^\circ)\underset{\sim hk}{V}(\underset{\sim k}{P}^j - \underset{\sim}{P}^\circ)' , \tag{8.21}$$

where the $\underset{\sim h}{U}$ are 1×2 vectors and the $\underset{\sim hk}{V}$ are 2×2 matrices. Quadratic functional statistics (4.14), (4.15) can be rewritten in this form. A linear functional statistic (8.4) is of form (8.21) with $\underset{\sim hk}{V} = \underset{\sim}{0}$, $\underset{\sim h}{U} = (\alpha_{h1}, \alpha_{h2})$.

Theorem 8.2. For a quadratic statistic (8.21), $\hat{SD}_{CBHS}$ has the same value for all balanced sets $\mathcal{J}$ closed under complementation, including the complete set $\mathcal{J}_o$.

Proof. $\underset{\sim h}{P}^j - \underset{\sim}{P}^\circ = \varepsilon_h^j \underset{\sim}{d}$, where $\underset{\sim}{d} = (\frac{1}{2}, -\frac{1}{2})$, so

$$\frac{\hat{\theta}(\underset{\sim}{\varepsilon}^j) - \hat{\theta}(-\underset{\sim}{\varepsilon}^j)}{2} = \sum_{h=1}^{H} \varepsilon_h^j \underset{\sim}{d}\underset{\sim h}{U} = \sum_{h=1}^{H} \varepsilon_h^j \frac{U_{h1} - U_{h2}}{2} , \tag{8.22}$$

the quadratic terms in (8.21) cancelling out. The same calculation as in the proof of Theorem 8.1 shows that

$$\hat{SD}_{CBHS} = \left\{ \sum_{h=1}^{H} \left[\frac{U_{h1} - U_{h2}}{2} \right]^2 \right\}^{1/2} . \quad \square \tag{8.23}$$

Line 11 of Table 5.2 refers to $\hat{SD}_{CBHS}$ for $\mathcal{J} = \mathcal{J}_o$, the complete set of 128 half-samples. Line 12 refers to $\hat{SD}_{CBHS}$ for $\mathcal{J}$ consisting of 16 half-samples, the 8 displayed in (8.18) plus their complements. The results shown on lines 11 and 12 are remarkably similar, much more so than lines 8 and 10. Root mean square error is reduced, compared to $\hat{SD}_{HS}$, though the results are still disappointing compared to the bootstrap, especially for $\hat{\phi}$.

The averages for $\hat{SD}_{CBHS}$ shown in Table 5.2 are smaller than those for $\hat{SD}_{BHS}$. This must always be the case:

Theorem 8.3. For any statistic $\hat{\theta}$ and any balanced set $\mathcal{J}$ closed under complementation, $\hat{SD}_{BHS} \geq \hat{SD}_{CBHS}$.

Proof. From definition (8.15),

$$\begin{aligned} \hat{SD}^2_{BHS} &= \frac{1}{J} \sum_{j=1}^{J} [\hat{\theta}(\underset{\sim}{\varepsilon}^j) - \hat{\theta}(\cdot)]^2 \\ &= \frac{2}{J} \sum_{j=1}^{J/2} \frac{[\hat{\theta}(\underset{\sim}{\varepsilon}^j) - \hat{\theta}(\cdot)]^2 + [\hat{\theta}(-\underset{\sim}{\varepsilon}^j) - \hat{\theta}(\cdot)]^2}{2} \\ &\geq \frac{2}{J} \sum_{j=1}^{J/2} \left[\frac{\hat{\theta}(\underset{\sim}{\varepsilon}^j) - \hat{\theta}(-\underset{\sim}{\varepsilon}^j)}{2} \right]^2 = SD^2_{CBHS} . \end{aligned}$$

Here we have used the elementary inequality $(a^2+b^2)/2 \geq [(a-b)/2]^2$. $\square$

5. Some Possible Alternative Methods. The half-sample form of the bootstrap, in which each stratum's bootstrap sample size is reduced by 1, is not the only way to correct the bias in the linear case. Still

considering only the situation where all $n_h = 2$ we could, for example, simply multiply formula (8.5) by $\sqrt{2}$, i.e. estimate Sd by $\sqrt{2}\ SD_{BOOT}$.

Even if we wish to use half-sampling, we might prefer to half-sample from distributions other than $\hat{F}_h$, (8.2). Suppose for instance that $\mathcal{X}_h = \mathcal{R}^1$, h=1, 2, ..., H, and let $\hat{\mu}_h = (x_{h1}+x_{h2})/2$, $\hat{\sigma}_h^2 = (x_{h1}-\hat{\mu}_h)^2$, the sample mean and variance of $\hat{F}_h$. Define

$$\tilde{F}\text{: mass } \frac{1}{n} \text{ at } \frac{x_{hi}-\hat{\mu}_h}{\hat{\sigma}_h}, \quad i=1, 2, \quad h=1, \ldots, H,$$

and let $\tilde{F}_h$ be the distribution of $\hat{\mu}_h + \hat{\sigma}_h\tilde{X}$, where $\tilde{X} \sim \tilde{F}$. Then $\tilde{F}_h$ has the same mean and variance as $\hat{F}_h$, but makes use of information from all strata to estimate the distribution in stratum h. Half-sampling from the $\tilde{F}_h$, (i.e. independently selecting $X_{h1}^* \sim \tilde{F}_h$, h=1, 2, ..., H and computing the standard deviation of $\hat{\theta}(\hat{F}_1^*, \hat{F}_2^*, \ldots, \hat{F}_H^*)$ where $\hat{F}_h^*$ puts all its mass on X_{h1}^*) might be better than half-sampling from the $\hat{F}_h$.

Balanced half-sampling and complementary balanced half-sampling are not the only ways to cut down the amount of computation needed to estimate a standard deviation. Let $\hat{\theta}^h = \hat{\theta}(x_{11}, x_{12}, x_{21}, x_{22}, \ldots, x_{h1}, x_{h1}, \ldots, x_{H1}, x_{H2})$, the value of the statistic when x_{h2} is replaced by a duplicate of x_{h1}, but no other changes are made in the data set; and let $\hat{\theta}^{-h}$ be the value when instead x_{h1} is replaced by a duplicate of x_{h2}. Then it can be shown that

$$\hat{SD} = \left\{ \sum_{h=1}^{H} \left[\frac{\hat{\theta}^h - \hat{\theta}^{-h}}{2} \right]^2 \right\}^{1/2}$$

equals $\hat{SD}_{CBHS}$, (8.23), for quadratic statistics (8.21). Evaluating this $\hat{SD}$ requires only n recomputations of $\hat{\theta}$, which is the minimum possible

number requried for $\hat{SD}_{CBHS}$. The relationship between this $\hat{SD}$ and $\hat{SD}_{CBHS}$ is analogous to the relationship between $\hat{SD}_{JACK}$ and $\hat{SD}_{BOOT}$.

Looking in the other direction, the clever idea underlying balanced repeated replications might be extended to reduce the number of calculations necessary for $\hat{SD}_{BOOT}$. Artificial stratification into pairs is not a good general answer, as we saw in Table 5.2, but more ambitious stratification schemes seem promising.

IX. RANDOM SUBSAMPLING

Hartigan (1969) introduced another resampling plan which we will call *random subsampling*. It is designed to give exact confidence intervals, rather than just standard deviations, but in a special class of problems: that of estimating the center of a symmetric distribution on the real line. We begin with a description of the problem and Hartigan's "typical value theorem", which very neatly gives the desired confidence intervals. Then we go on to show the connection between random subsampling and the bootstrap, in terms of large sample theory. Chapter X concerns the important problem of small sample nonparametric confidence intervals for nonsymmetric problems.

1. M Estimates. We consider the case of i.i.d. observations from a symmetric distribution on the real line

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta , \quad \text{Prob}_\theta\{X \in A\} = \int_A f(x-\theta)dx , \tag{9.1}$$

where $f(\cdot)$ is a symmetric density function, $\int_{-\infty}^{\infty} f(x)dx = 1$, $f(x) \geq 0$, $f(-x) = f(x)$. The unknown translation parameter θ is the center of symmetry of F_θ.

An "M-estimate" $\hat{\theta}(x_1, x_2, \ldots, x_n)$ for θ is any solution to the equation

$$\sum_{i=1}^{n} \psi(x_i - t) = 0 . \tag{9.2}$$

Here the observed data $X_i = x_i$, $i=1, 2, \ldots, n$ is fixed while t is the variable. The kernel $\psi(\cdot)$ is assumed to be anti-symmetric and strictly increasing

$$\text{i)}\ \ \psi(-z) = -\psi(z)\ , \quad \text{ii)}\ \ \psi(z) \uparrow z\ . \tag{9.3}$$

This last condition is not usually imposed, but is necessary for the development which follows. It guarantees that $\hat{\theta}$ is defined uniquely for any data set $x_1, x_2, \ldots, x_n$. Notice that $\hat{\theta}$ is a functional statistic, $\hat{\theta} = \theta(\hat{F})$, since (9.2) can be written as $\int_{-\infty}^{\infty} \psi(x-t)\, d\hat{F}(x) = 0$.

Example 1. $\psi(z) = z$. Then $\hat{\theta}=\bar{x}$, the sample mean.

Example 2. $\psi(z) = \text{sgn}(z)\ [1 - e^{-c|z|}]$ for some constant $c > 0$. As $c\to\infty$, $\hat{\theta}_c(x_1, x_2, \ldots, x_n) \to$ sample median, the middle order statistic if n is odd, the average of the middle two order statistics if n is even.

Example 3. $\psi(z) = -\frac{f'(z)}{f(z)} = -\frac{d}{dz}\log f(z)$. Then (9.2) says that $\hat{\theta}$ is the solution to $\sum_i \frac{\partial}{\partial t}\log f(x_i-t) = 0$, i.e. the maximum likelihood estimate of θ. (This is the origin of the name "m-estimator".) If $f(z)$ is the normal density than $\psi(z) = z$ and $\hat{\theta}=\bar{x}$. If $f(z) = \frac{1}{2}e^{-|z|}$, the double exponential, then $\psi(z) = \text{sgn}(z)$ and $\hat{\theta}$ = sample median. If $f(z) = \frac{1}{\pi}(1+z^2)^{-1}$, the Cauchy distribution, then $\psi(z) = 2z/(1+z^2)$. In the last two examples, condition (ii) of (9.3) isn't satisfied.

The influence function (4.13) of an m-estimate based on $\psi(\cdot)$ is $IF(x) = c\psi(x-\theta)$, c some positive constant. Robustness theory focuses on choices of $\psi(\cdot)$ which have bounded influence function, $\sup|\psi(z)| < \infty$, but still give reasonably high estimation efficiency for standard families like the normal, double exponential, and Cauchy. Huber (1974) gives a thorough review of this theory.

2. The Typical Value Theorem. There are 2^n-1 nonempty subsets of $\{1, 2, \ldots, n\}$. If S is such a subset, define $\hat{\theta}_S$ as the m-estimate based on $\{x_i;\ i\in S\}$,

$$\hat{\theta}_S: \sum_{i\in S} \psi(x_i-t) = 0 \ . \tag{9.4}$$

The 2^n-1 values of $\hat{\theta}_S$ will be distinct with probability one, under assumptions (9.3). Their ordered values partition the line into 2^n intervals, say $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_{2^n}$. For instance $\mathcal{I}_1 = (-\infty, x_{(1)})$, where $x_{(1)}$ is the smallest value of $x_1, x_2, \ldots, x_n$.

Typical Value Theorem[†]. The true value θ has probability $1/2^n$ of being in any interval $\mathcal{I}_j$, $j=1, 2, \ldots, 2^n$.

Proof. There are 2^n vectors $\underset{\sim}{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$ having components $\delta_i = \pm 1$. For each $\underset{\sim}{\delta}$ define

$$Q(\underset{\sim}{\delta},t) = \sum_{i=1}^{n} \psi(\delta_i(x_i-t)) = \sum_{i=1}^{n} \delta_i \psi(x_i-t) \ . \tag{9.5}$$

In particular $Q(\underset{\sim}{1},t) = \sum_{i=1}^{n} \psi(x_i-t)$, the defining function for $\hat{\theta}$ in (9.2). The quantity

$$Q(\underset{\sim}{\delta},t)-Q(\underset{\sim}{1},t) = -2 \sum_{i\in S(\underset{\sim}{\delta})} \psi(x_i-t) \ , \tag{9.6}$$

where

$$S(\underset{\sim}{\delta}) = \{i: \delta_i = -1\} \ , \tag{9.7}$$

is strictly increasing in t, if $\underset{\sim}{\delta} \neq \underset{\sim}{1}$.

The following two statements are seen to be equivalent:

$$\begin{aligned} &\text{(i)} \quad Q(\underset{\sim}{\delta},\theta) > Q(\underset{\sim}{1},\theta) \\ &\text{(ii)} \quad \hat{\theta}_{S(\underset{\sim}{\delta})} < \theta \end{aligned} \tag{9.8}$$

[†]From Hartigan (1969), who also credits J.Tukey and C. Mallows. The derivation here follows Maritz (1979).

(since by (9.6) $Q(\underset{\sim}{\delta},\theta) - Q(\underset{\sim}{1},\theta) > 0$ implies that $\sum_{i\in S(\underset{\sim}{\delta})} \psi(x_i-t)$ has its root to the left of θ.) But by the symmetry of $\psi(\cdot)$ and $f(\cdot)$ the 2^n random variables $Q(\underset{\sim}{\delta},\theta) = \sum_{i=1}^n \delta_i \psi(X_i-\theta)$ are exchangeable. Thus

$$\text{Prob}_\theta\{Q(\underset{\sim}{1},\theta) \text{ is j-th largest among the } Q(\underset{\sim}{\delta},\theta)\} = \frac{1}{2^n}, \qquad (9.9)$$

$j=1, 2, \ldots, 2^n$, so by (9.8),

$$\text{Prob}_\theta\{\text{exactly } j-1 \text{ of the } \hat\theta_S < \theta\} = \frac{1}{2^n}, \quad j=1, 2, \ldots, 2^n, \qquad (9.10)$$

which is the statement of the Theorem. □

The typical value theorem is used to set confidence intervals as follows: suppose we observe $n = 10$ observations from a symmetric density on the real line. Let $\hat\theta_{(1)} < \hat\theta_{(2)} < \cdots < \hat\theta_{(1023)}$ be the ordered subsample m-estimates. Then $(\hat\theta_{(51)}, \hat\theta_{(973)}) = \bigcup_{j=52}^{973} \vartheta_j$ is a $922/1024 = .900$ central confidence interval for θ. *Warning*: an exact confidence interval is not necessarily a good one. For instance if $\psi(z) = z$, so $\hat\theta=\bar x$, and $f(z)$ is Cauchy, then $(\bar x_{(51)}, \bar x_{(973)})$ is a 90% central confidence interval for the center of the Cauchy distribution, $\bar x_{(j)}$ being the j-th ordered subsample average. This interval can be absurdly long, compared to the optimum interval for the Cauchy, if the sample includes an outlying observation.

3. Random Subsampling. One needn't evaluate all 2^n-1 subsample values $\hat\theta_S$ in order to use the typical value theorem. *Random subsampling* provides a convenient shortcut:

Corollary. Let $S_1, S_2, \ldots, S_{B-1}$ be chosen randomly and without replacement from the 2^n-1 nonempty subsets of $\{1, 2, \ldots, n\}$, and let $\vartheta_1, \vartheta_2, \ldots, \vartheta_B$ be the intervals determined by the ordered values of $\hat\theta_{S_j}$. Then θ has probability $1/B$ of being in any interval ϑ_j, $j=1, 2, \ldots, B$.

The proof, which appears in Hartigan (1969), is left as a pleasant exercise for the reader. *Note*: "probability" has a different meaning in the corollary than in the theorem. Write $x_i = \theta + \varepsilon_i |x_i - \theta|$, so that

$$\varepsilon_i \Big| |x_i - \theta| = \begin{cases} 1 & \frac{1}{2} \\ -1 & \frac{1}{2} \end{cases} \text{ prob} \quad \text{independently} \quad i=1, 2, \ldots, n . \tag{9.11}$$

"Probability" in the theorem refers to the conditional distribution (9.11) of the ε_i given the $|x_i - \theta|$. The probability statement in the corollary also averages over the random choice of the subsets S_j. This can be less satisfactory. Suppose, as a rather farfetched example, that $B=n$ and that we happen to choose $S_1 = \{x_1\}$, $S_2 = \{x_2\}$, ..., $S_n = \{x_n\}$. Then $\hat{\theta}_{S_j} = x_j$ for any m-estimate, and so $(x_{(1)}, x_{(n)})$ is an $(n-1)/(n+1)$ central confidence interval for θ, by the corollary. However, since θ is the median of the distribution generating the data, $(x_{(1)}, x_{(n)})$ is a $1 - 1/2^{n-1}$ central confidence interval for θ! (See Chapter X, Section 2.)

Hartigan also provides a more satisfactory version of the corollary, in which "probability" refers only to mechanism (9.11). This involves choosing $S_1, S_2, \ldots, S_{B-1}$ in a balanced way, "balance" referring to a symmetry condition between the selected subsets. (Essentially, the set of vectors $\underset{\sim}{1}, \underset{\sim}{\delta}_1, \underset{\sim}{\delta}_2, \ldots, \underset{\sim}{\delta}_{B-1}$, relating to the subsets as at (9.7), has to be such that any one vector has the same set of angles with the remaining B-1.) In practice the advantage of balanced subsets over randomly selected ones seems modest, and we will consider only the latter.

Random subsampling is a resampling plan, as described in Chapter VI, Section 1. The choice of a single random subsample S amounts to the choice of a resampling vector $\underset{\sim}{P}^*$ as follows: let

$$I_i^* = \begin{cases} 1 & \text{prob } \frac{1}{2} \\ 0 & \text{prob } \frac{1}{2} \end{cases} \quad \text{independently} \quad i=1, 2, \ldots, n .$$

Then, conditional on the event $\sum_{i=1}^n I_i^* > 0$,

$$P_i^* = \frac{I_i^*}{\sum_{j=1}^n I_j^*} . \tag{9.12}$$

4. Resampling Asymptotics. Random subsampling belongs to a large class of resampling methods, including the bootstrap and half-sampling, which have identical asymptotic properties, at least to a first order of approximation. Consider an arbitrary resampling plan in which we only assume that the components of P^* are exchangeable. Let $\mathrm{Var}_* P_1^*$ be the variance of any one component under the resampling plan. Notice that

$$0 = \mathrm{Var}_* \sum_{i=1}^n P_i^* = n \, \mathrm{Var}_* P_1^* + n(n-1) \, \mathrm{Cov}_*(P_1^*, P_2^*) .$$

We see that $\underset{\sim}{P}^*$ has mean vector and covariance matrix

$$\underset{\sim}{P}^* \sim (\underset{\sim}{P}^\circ, \frac{n}{n-1} (\underset{\sim}{I} - n\underset{\sim}{P}^{\circ\prime} \underset{\sim}{P}^\circ) \mathrm{Var}_* P_1^*) , \tag{9.13}$$

$\underset{\sim}{P}^\circ = (1, 1, \ldots, 1)/n$ as before.

Suppose that $\mathcal{X} = \mathcal{R}^1$ and that we are resampling the average, $\bar{X}$. Given $\underset{\sim}{X}=\underset{\sim}{x}$, the resampled average $\bar{X}^* = \underset{\sim}{P}^* \underset{\sim}{x}'$ has mean and variance

$$\bar{X}^* \sim (\bar{x}, \frac{n}{n-1} \Sigma(x_i - \bar{x})^2 \mathrm{Var}_* P_1^*) . \tag{9.14}$$

under the resampling distribution.

The Bootstrap. $\mathrm{Var}_* P_1^* = (n-1)/n^3$, so $\mathrm{Var}_* \bar{X}^* = \Sigma(x_i-\bar{x})^2/n^2$.

Random Subsampling. It is easy to show, using (9.11), that $\mathrm{Var}_* P_1^* = (n+2)/n^3\ [1 + O(\frac{1}{n})]$, so

$$\mathrm{Var}_* \bar{X}^* = \frac{n+2}{n-1}\ \frac{\Sigma(x_i-\bar{x})^2}{n^2}\ [1 + O(\tfrac{1}{n})]\ . \tag{9.15}$$

Random Half-Sampling. Randomly choosing subsamples of size $n/2$, as in line 9 of Table 5.2, gives $\mathrm{Var}_* P_1^* = 1/n^2$ and so $\mathrm{Var}_* \bar{X}^* = \Sigma(x_i-\bar{x})^2/(n \cdot (n-1))$, the usual estimate of variance for a sample average.

The point here is that any resampling plan having $\underset{\sim}{P}^*$ exchangeable and $\lim n^2 \mathrm{Var}_* P_1^* = 1$ gives, asymptotically, the same value of $\mathrm{Var}_* \bar{X}^*$. This equivalence extends beyond averages to a wide class of smoothly defined random quantities. Working in the finite sample space context of Chapter V, Section 6, the resampling distribution of f^* is approximately normal,

$$\hat{f}^* - \hat{f} \to_L \mathcal{N}\left(0,\ \frac{\dot{\hat{f}}\,\underset{\sim}{\dot{\hat{f}}}}{n-1}\ (n^2\ \mathrm{Var}_*\ P_1^*)\right),$$

(the expressions for the mean vector and covariance matrix being exact). Smoothly defined random quantities $Q(\hat{f}^*, \hat{f})$ have the same limiting distribution under any resampling plan satisfying $\lim_{n\to\infty} n^2\ \mathrm{Var}_*\ P_1^* = 1$.

5. Random Subsampling for Other Problems. Line 13 of Table 5.2 shows random subsampling applied to estimate $\mathrm{Sd}(\hat{\rho})$ and $\mathrm{Sd}(\hat{\phi})$ in the correlation experiment. For each of the 200 trials, $B = 128$ random subsamples were generated as at (9.11), the corresponding values $\hat{\rho}^{*1}, \ldots, \hat{\rho}^{*B}$ computed, and the standard deviation of $\hat{\rho}$ estimated by

$$\widehat{\mathrm{SD}}_{\mathrm{SUB}}(\hat{\rho}) = \{\sum_{b=1}^{B} [\hat{\rho}^{*b} - \hat{\rho}^{*\cdot}]^2/[B-1]\}^{1/2} \tag{9.16}$$

(with a similar calculation for $\widehat{\mathrm{SD}}_{\mathrm{SUB}}(\hat{\phi})$).

The results border on the disasterous, especially for $\hat{\phi}$. They would have been even worse if we had not placed a restriction on the subsampling: only subsamples of size ≥ 4 were allowed. Asymptotically, we know that $\hat{SD}_{SUB}$ is equivalent to $\hat{SD}_{BOOT}$. Obviously the asymptotics cannot be trusted to predict small sample behavior, at least not in this problem.

Line 14 of Table 5.2 used the same data as line 13, the random subsample values of $\hat{\rho}$ and $\hat{\phi}$, but calculated standard deviations in a more robust way,

$$\hat{SD}(\hat{\rho}) = \frac{\hat{\rho}^*_{(B_2)} - \hat{\rho}^*_{(B_1)}}{2}, \tag{9.17}$$

$B_1 = [.16(B+1)]$, $B_2 = [.84(B+1)]$, where $\hat{\rho}^*_{(j)}$ is the j-th ordered value of $\hat{\rho}^{*1}, \hat{\rho}^{*2}, \ldots, \hat{\rho}^{*B}$. In other words, (9.17) is one-half the length of what would be the central 68% confidence interval for ρ, if the typical value theorem applied to this case. (We could just as well apply (9.17) to the bootstrap if we thought that occasional outlying values of $\hat{\rho}^{*j}$ were having an inordinate effect on formula (5.6).) The results are better, but still not encouraging. Another correction, suggested by comparison of (9.15) with the corresponding bootstrap calculation, is to multiply (9.17) by $[(n-1)/(n+2)]^{1/2} = .901$. This gives $\sqrt{MSE}$ of .083 for estimating $Sd(\hat{\rho})$, and .072 for $Sd(\hat{\phi})$, quite reasonable results, but suspect because of the special "corrections" required.

The problem of choosing among asymptotically equivalent resampling plans is of considerable practical importance. The author feels that the bootstrap has demonstrated some measure of superiority, probably because it is the nonparametric MLE, but the question is still far from settled.

X. NONPARAMETRIC CONFIDENCE INTERVALS

So far we have mainly concentrated on estimating the bias and standard deviation of a point estimator $\hat{\theta}$ for a real parameter θ. This is often all that is needed in applications. However a *confidence interval* for θ is usually preferable. This section, which is highly speculative in content, concerns setting approximate confidence intervals in small sample nonparametric situations.

We begin on familiar ground: setting a confidence interval for the median of a distribution F on the real line. The typical value theorem reduces to the standard order statistic intervals for the median in this case. The bootstrap distribution of the sample median is derived, this being one case where theoretical calculation of the bootstrap distribution is possible. It is shown that the percentiles of the bootstrap distribution also provide (almost) the classical confidence intervals for the median. The method of using the bootstrap distribution, called the *percentile method*, is justified from various theoretical points of view, and improvements suggested. The section ends with a brief discussion of more adventurous bootstrap methods for obtaining confidence intervals.

1. The Median. Let F be a distribution on $\mathcal{R}^1$ with median θ, defined as $\theta = \inf_t[\text{Prob}_F\{X \leq t\} = .5]$. For convenience we assume that F is continuous. Having observed an i.i.d. sample $X_i = x_i$, $i=1, 2, \ldots, n$ from F, we can construct exact confidence intervals for θ using the order statistics $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. Define

$$b_{k,n}(p) = \binom{n}{k} p^k(1-p)^{n-k} , \qquad (10.1)$$

the binomial probability of observing k heads in n independent flips of a coin having probability p of heads. The random variable

$$Z = \#\{X_i < \theta\} \tag{10.2}$$

has a binomial distribution with $p = \frac{1}{2}$, $Z \sim Bi(n, \frac{1}{2})$. Therefore

$$\text{Prob}_F\{x_{(k_1)} < \theta \leq x_{(k_2)}\} = \sum_{k=k_1}^{k_2-1} b_{k,n}(.5) \tag{10.3}$$

since the event $\{x_{(k_1)} < \theta \leq x_{(k_2)}\}$ is the same as the event $\{k_1 \leq Z < k_2\}$.

As an example of the use of (10.3), take $n = 13$, $k_1 = 4$, $k_2 = 10$. Then a binomial table gives

$$\text{Prob}_F\{x_{(4)} < \theta \leq x_{(10)}\} = .908 \ . \tag{10.4}$$

The two tail probabilities are equal, $\text{Prob}_F\{\theta \leq x_{(4)}\} = \text{Prob}\{Z \leq 3\} = .046$, and $\text{Prob}_F\{\theta > x_{(10)}\} = \text{Prob}\{Z \geq 10\} = .046$. In this case $(x_{(4)}, x_{(10)}]$ is a central 90.8% confidence interval for θ.

2. Typical Value Theorem for the Median. The median is an M-estimator as described in Chapter IX, Section 1, with $\psi(z) = \text{sgn}(z)$. Since $\psi(z)$ is not strictly monotonic, and we have made no assumptions about the symmetry of F, the typical value theorem of Chapter IX, Section 2 does not apply. However a version of the theorem does hold in this case, as we shall see, and in fact reduces exactly to the binomial interval (10.3).

For any nonempty subset S of $\{1, 2, \ldots, n\}$ let $\hat{\theta}_S$ be the sample median of $\{x_i, i \in S\}$,

$\hat{\theta}_S$ = middle order statistic if $\#S$, the number of elements in S, is odd ,

$\hat{\theta}_S$ = any number between the two middle order statistics if $\#S$ is even .

("Between" $x_{(a)}$ and $x_{(b)}$ means lying in the interval $(x_{(a)}, x_{(b)}]$.) Define the random variable

$$Y = \#\{S: \hat{\theta}_S < \theta\} , \tag{10.5}$$

the number of nonempty subsets S for which the sample median is less than the true median. Definition (10.5) assigns Y a *range* of integer values depending on how the even sized cases are assigned.

Example: Suppose $n=4$ and $Z = \#\{X_i < \theta\} = 2$, i.e. $\theta \in (x_{(2)}, x_{(3)}]$. Then there are $2^4-1 = 15$ nonempty subsets S, of which 5 have $\hat{\theta}_S < \theta$, namely $\{x_{(1)}\}$, $\{x_{(1)}, x_{(2)}\}$, $\{x_{(1)}, x_{(2)}, x_{(3)}\}$, $\{x_{(1)}, x_{(2)}, x_{(4)}\}$, and $\{x_{(2)}\}$. There are 5 sets S for which $\hat{\theta}_S \geq \theta$, namely $\{x_{(3)}\}$, $\{x_{(3)}, x_{(4)}\}$, $\{x_{(2)}, x_{(3)}, x_{(4)}\}$, $\{x_{(1)}, x_{(3)}, x_{(4)}\}$, and $\{x_{(4)}\}$. Finally, there are 5 ambiguous sets S, namely $\{x_{(1)}, x_{(3)}\}$, $\{x_{(2)}, x_{(3)}\}$, $\{x_{(1)}, x_{(4)}\}$, $\{x_{(2)}, x_{(4)}\}$, and $\{x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}\}$, for which we can have either $\hat{\theta}_S < \theta$ or $\hat{\theta}_S \geq \theta$. In this case Y takes on the range of values $\{5, 6, 7, 8, 9, 10\}$, depending on how the ambiguous cases are assigned. In general we have the following relation between the random variables $Z = \#\{X_i < \theta\}$ and $Y = \#\{S: \hat{\theta}_S < \theta\}$:

Theorem 10.1. The event $\{Z=z\}$ is equivalent to the event

$$\sum_{j=0}^{z-1} \binom{n}{j} \leq Y < \sum_{j=0}^{z} \binom{n}{j} . \tag{10.6}$$

(Proof left to the reader.) in other words, Y is really the same random variable as Z, except it takes its values in a more complicated space. With $n=4$ for example, the values $Z = 0, 1, 2, 3, 4$ correspond to

$Y = \{0\}, \{1, 2, 3, 4\}, \{5, 6, 7, 8, 9, 10\}, \{11, 12, 13, 14\}, \{15\}$, respectively.

Theorem 10.1 is a version of the typical value theorem. To see this, notice that any postulated value of θ assigns a value to Z, say $z = \#\{x_i < \theta\}$. The attained one-sided significance level of $Z=z$, according to the binomial distribution $Z \sim Bi(n, \frac{1}{2})$, satisfies

$$\sum_{j=0}^{z-1} \binom{n}{j} \frac{1}{2^n} < \text{sig level} \leq \sum_{j=0}^{z} \binom{n}{j} \frac{1}{2^n}, \tag{10.7}$$

depending on how the atom at z is assigned.

Going back to the situation described in Chapter IX, Sections 1 and 2, let $Y = \#\{\hat{\theta}_S < \theta\}$. Any postulated value of θ assigns a value to Y, say $y = \#\{S: \hat{\theta}_S < \theta\}$. The attained one-sided significance level according to the typical value theorem is

$$\frac{y}{2^n} < \text{sig level} \leq \frac{y+1}{2^n}, \tag{10.8}$$

since y of the intervals θ_j lie to the left of θ.

In the case of the median we can't observe $Y=y$ exactly, but rather a range of values y depending on $Z=z$, namely $\sum_{j=0}^{z-1} \binom{n}{j} \leq y < \sum_{j=0}^{z} \binom{n}{j}$. If (10.8) applied here, this range of y values would correspond to the range of significance levels (10.7), which is exactly what we get from the binomial theory. In other words, the typical value theorem does apply to the median, in somewhat coarser form. This coarseness is due to the fact that for the median the 2^n-1 choices of S don't correspond to 2^n-1 different values of $\hat{\theta}_S$.

It is not surprising that the typical value theorem for the median does not require F to be symmetric about its true median θ. If F is not

symmetric about θ to begin with, we can symmetrize it with a monotonic transformation of the data. This transformation doesn't effect the value of Z or Y.

3. Bootstrap Theory for the Median. The bootstrap distribution for the sample median can be caluclated theoretically, without recourse to Monte Carlo methods. It is convenient to consider odd sample sizes, say $n = 2m-1$. Then the sample median $\hat{\theta}$ equals $x_{(m)}$, the middle order statistic.

The bootstrap sample $X_1^*, X_2^*, \ldots, X_n^* \overset{iid}{\sim} \hat{F}$ has bootstrap sample median $\hat{\theta}^* = X_{(m)}^*$, the m-th ordered value of the X_i^*. (Notice that this is true even though there are ties among the bootstrap observations.) Define

$$M_j^* = \#\{X_i^* = x_{(j)}\} , \tag{10.9}$$

$j=1, 2, \ldots, n$. The event $\{X_{(m)}^* > x_{(k)}\}$ is equivalent to $\{\sum_{j=1}^{k} M_j^* \leq m-1\}$, so that

$$\begin{aligned} \text{Prob}_*\{\hat{\theta}^* \geq x_{(k)}\} &= \text{Prob}_*\{\sum_{j=1}^{k} M_j^* \leq m-1\} \\ &= \text{Prob}\{\text{Bi}(n, \tfrac{k}{n}) \leq m-1\} = \sum_{j=0}^{m-1} b_{j,n}(\tfrac{k}{n}) . \end{aligned} \tag{10.10}$$

Here we are using $\sum_{j=1}^{k} M_j^* \sim \text{Bi}(n, \frac{k}{n})$, see (6.6), and definition (10.1). Therefore the bootstrap distribution of $\hat{\theta}^*$ is concentrated on the values $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, say with bootstrap probability $p_{(k)}$ of equaling $x_{(k)}$,

$$p_{(k)} \equiv \text{Prob}_*\{\hat{\theta}^* = x_{(k)}\} = \sum_{j=0}^{m-1} \{b_{j,n}(\tfrac{k-1}{n}) - b_{j,n}(\tfrac{k}{n})\} . \tag{10.11}$$

Example. 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

$p_{(k)}$	.1936	.1242	.0550	.0142	.0015	.0000
k	8	9	10	11	12	13

The bootstrap estimate of standard deviation is $\hat{\sigma}_{BOOT} = [\Sigma p_{(k)} x_{(k)}^2 - (\Sigma p_{(k)} x_{(k)})^2]^{1/2}$. This can be shown to be asymptotically consistent for the true standard deviation of $\hat{\theta}$, in contrast to the jackknife Sd estimate, Chapter III, Section 4.

4. The Percentile Method. We now discuss a simple method for assigning approximate confidence intervals to any real-valued parameter $\theta = \theta(F)$, based on the bootstrap distribution of $\hat{\theta} = \theta(\hat{F})$. Once we have introduced the method, we will apply it to the case of the median and obtain, almost, the binomial result (10.3).

Let

$$\widehat{CDF}(t) = \text{Prob}_*\{\hat{\theta}^* \leq t\} \tag{10.12}$$

be the cumulative distribution function of the bootstrap distribution of $\hat{\theta}^*$. (If the bootstrap distribution is obtained by Monte Carlo then $\widehat{CDF}(t)$ is approximated by $\#\{\hat{\theta}^{*b} \leq t\}/B$.) For a given α between 0 and .5 define

$$\hat{\theta}_{LOW}(\alpha) = \widehat{CDF}^{-1}(\alpha) \ , \qquad \hat{\theta}_{UP}(\alpha) = \widehat{CDF}^{-1}(1-\alpha) \ , \tag{10.13}$$

usually denoted simply $\hat{\theta}_{LOW}$, $\hat{\theta}_{UP}$. The *percentile method* consists of taking

$$[\hat{\theta}_{LOW}(\alpha), \ \hat{\theta}_{UP}(\alpha)] \tag{10.14}$$

as an approximate $1-2\alpha$ central confidence interval for θ. Since $\alpha = \widehat{CDF}(\hat{\theta}_{LOW})$, $1-\alpha = \widehat{CDF}(\hat{\theta}_{UP})$, the percentile method interval consists of the central $1-2\alpha$ proportion of the bootstrap distribution.

As an example, consider the law school data, as given in Table 2.1. The bootstrap distribution is displayed in Figure 5.1. A large number of bootstrap replications, B = 1000 in this case, is necessary to get reasonable accuracy in the tails of the distribution. For $\alpha = .16$, the central $1-2\alpha = .68$ percentile confidence interval for ρ is $[.66, .91] = [\hat{\rho} - .12, \hat{\rho} + .13]$. This differs noticeably from the standard normal theory confidence interval for ρ, $[.62, .87] = [\hat{\rho} - .16, \hat{\rho} + .09]$, which is skewed to the left relative to the observed value $\hat{\rho} = .78$. (The normal theory interval has endpoints $\tanh[\hat{\phi} - \frac{\hat{\rho}}{2(n-1)} \pm z_\alpha/\sqrt{n-3}]$ where $\hat{\phi} = \tanh^{-1}\hat{\rho}$ and z_α is the $1-\alpha$ point for a standard normal, $z_{.16} = 1$; it is an approximate inversion of the confidence interval for ϕ based on $\hat{\phi} \sim \eta(\phi + \frac{\rho}{2(n-1)}, \frac{1}{n-3})$.) Section 7 suggests a bias correction for the percentile method which rectifies this disagreement.

The results of a small Monte Carlo experiment are reported in Table 10.1. 100 trials of $n = 15$ independent bivariate normal observations were generated, true $\rho = .5$. For each trial the bootstrap distribution of $\hat{\rho}^*$ was approximated by $B = 1000$ bootstrap replications. We see, for example, that in 22 of the 100 trials the true value .5 lay in the region $[\widehat{CDF}^{-1}(.25), \widehat{CDF}^{-1}(.5)]$, compared to the expected number 25 if the percentile method were generating exact confidence intervals.

Table 10.1 is reassuring, perhaps misleadingly so when viewed in conjunction with Table 10.2. Central 68% confidence intervals ($\alpha = .16$), with $\hat{\rho}$ subtracted from each endpoint, are presented for the first ten

Region	0-10%	10-25%	25-50%	50-75%	75-90%	90-100%
Expected #	10	15	25	25	15	10
Observed #	13	16	22	27	12	10

Table 10.1. 100 trials of $X_1, X_2, \ldots, X_{15} \sim$ bivariate normal with $\rho = .5$. For each trial the bootstrap distribution of $\hat{\rho}^*$ was calculated, based on $B = 1000$ bootstrap replications. In 13 of the 100 trials, the true value .5 lay in the lower 10% of the bootstrap distribution, etc.

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, .18)	(-.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)

Table 10.2. Central 68% confidence intervals for the first ten trials of the Monte Carlo experiment, each interval having $\hat{\rho}$ subtracted from both endpoints.

trials of the Monte Carlo experiment. Compared to the normal theory intervals, which are correct here, the percentile method gives somewhat erratic results, both in terms of the length of the intervals and of their skewness about $\hat{\rho}$. (Four of the ten percentile intervals are symmetric or even skewed to the right.) The bias corrected percentile method of Section 7 performs better. The final column, which combines smoothing, as at (5.8), with the bias correction is more satisfactory still, but is suspect since the smoothing biases the answers toward what we know is the correct model in this situation.

The percentile method is not as trustworthy as $\hat{\sigma}_{BOOT}$, which, in the author's experience, can be relied upon to automatically give quite reasonable estimates. On the other hand, setting confidence intervals is a harder problem than estimating standard deviations. The percentile method, perhaps modified as in Table 10.2, is usually more informative than the naive interval $\hat{\theta} \pm c_\alpha \hat{\sigma}$, where c_α is a number taken from the normal or t-tables, though it requires more bootstrap sampling than does the estimation of $\hat{\sigma}$. Some theoretical justification for the percentile intervals is given in Sections 5 - 9. More ambitious methods are discussed in Section 10.

5. Percentile Method for the Median. In the case where $\theta(F)$ is the median of a distribution F on the real line, and $\hat{\theta}$ is the sample median, the percentile method comes very close to giving the classical binomial intervals, (10.3). For instance consider the case $n = 13$. The bootstrap distribution of $\hat{\theta}^*$ is supported on the order statistics $x_{(k)}$, as shown at the end of Chapter X, Section 3, so any percentile interval will be of the form $[x_{(k_1)}, x_{(k_2)}]$. Take $k_1 = 4$ and $k_2 = 10$. The interval $[x_{(4)}, x_{(10)}]$ is a central $1-2\alpha$ interval of the bootstrap

distribution with $\alpha = (.0000 + .0015 + .0142 + .0550/2) = .0432$. Here we have split the bootstrap probability at the endpoint of the interval, for reasons discussed at the end of this section. The percentile method assigns

$$[\hat{\theta}_{LOW}, \hat{\theta}_{UP}] = [x_{(k_1)}, x_{(k_2)}] \tag{10.15}$$

approximate confidence level $1-2\alpha = .914$ for θ. This compares remarkably well with (10.4). Numerical investigation confirms that the agreement is always very good as long as $\alpha \geq .01$. A theoretical reason for this agreement is given next.

We consider just the lower limit of the interval, the argument for the upper limit being the same. For the classical binomial interval of Section 1, the α-level connected with the event $\{\theta \leq x_{(k_1)}\}$ is

$$\alpha = \text{Prob}\{Z \leq k_1 - 1\} = \text{Prob}\{\text{Bi}(n, \tfrac{1}{2}) \leq k_1 - 1\} \ . \tag{10.16}$$

Looking at (10.10), notice that $\text{Prob}_*\{\hat{\theta}^* \leq x_{(k_1)}\} = \text{Prob}\{\text{Bi}(n, \frac{k}{n}) \geq m\}$. The percentile method for the median, taking into account splitting the boctstrap probability at the endpoint, assigns approximate significance level $\hat{\alpha}$ to $\{\theta \leq x_{(k_1)}\}$,

$$\hat{\alpha} = \tfrac{1}{2}[\text{Prob}\{\text{Bi}(n, \tfrac{k}{n}) \geq m\} + \text{Prob}\{\text{Bi}(n, \tfrac{k-1}{n}) \geq m\}] \ . \tag{10.17}$$

If the reader replaces (10.16) and (10.17) with their usual normal approximations, making the continuity corrections, but ignoring the difference in denominators, he will see why $\hat{\alpha}$ approximates α. (Actually the approximation is mysteriously better than this computation suggests, especially when k_1/n is much less than $1/2$.)

Let

$$D(t) = \text{Prob}_F\{X < t\} , \quad \hat{D}(t) = \frac{\#\{x_i < t\}}{n} \tag{10.18}$$

be the cumulative and empirical cumulative distribution functions. Then

$$\hat{D}(t) \sim Bi(n, D(t))/n , \tag{10.19}$$

so if $D(t) = \frac{1}{2}$ then $\hat{D}(t) \sim Bi(n, \frac{1}{2})/n$. According to (10.16), another way to describe the binomial α-level interval is the following: the lower limit of the interval is the smallest value of t for which we can accept the null hypothesis $D(t) = \frac{1}{2}$ with one-sided significance level α. The lower limit of the percentile interval has this interpretation: it is the smallest value of t for which the α-level upper confidence interval for $D(t)$, based on (10.19), includes the value $1/2$. In other words, the binomial interval checks whether $\hat{D}(t)$ is too small compared to expectation $1/2$, while the percentile interval checks whether $1/2$ is too large compared to expectation $\hat{D}(t)$.

It is not surprising that the percentile method for the median requires splitting the bootstrap probability at the endpoints. The problem is the same as in the typical value theory for the median, namely that the sample median takes on only n possible different values under bootstrap sampling. This contrasts with smoothly defined statistics, such as the correlation, for which the bootstrap distribution is effectively continuous when $n \geq 10$. Suppose that instead of the sample median we were considering the m-estimator $\hat{\theta}_c$ of Example 2, Chapter IX, Section 1, with c very large. Then if a bootstrap median $\hat{\theta}^*$ equals $x_{(k)}$, the corresponding value of $\hat{\theta}_c^*$ will be *almost but not quite* equal to $x_{(k)}$. Take $n = 13$ and $\alpha = .0432$ as

at (10.15). Then $\lim_{c\to\infty} \hat{\theta}_{LOW,c} = x_{(4)}$, and $\lim_{c\to\infty} \hat{\theta}_{UP,c} = x_{(10)}$. In this sense, $[x_{(4)}, x_{(10)}]$ is a .914 confidence interval for the median, as claimed. The Bayesian arguments of the next section provide another justification for splitting the endpoint probabilities.

6. Bayesian Justification of the Percentile Method. We assume that the sample space $\mathcal{X}$ is discrete as in Chapter V, Section 6. This is no real restriction since we can take the number L of discrete categories arbitrarily large. If $\mathcal{X}$ is the real line, for instance, we might partition $[-10^{10}, 10^{10}]$ into $2 \cdot 10^{20}$ intervals of length 10^{-10}. Then $L = 2 \cdot 10^{20} + 2$, counting the semi-infinite end intervals, and for practical purposes the discretization will have no effect on our inferences. As in Chapter V, Section 6, we let f_ℓ equal the probability that X occurs in category ℓ, with $\hat{f}_\ell$ equal the corresponding observed frequency $\#\{x_i \in \text{category } \ell\}/n$, and denote $\underset{\sim}{f} = (f_1, f_2, \ldots, f_L)$, $\hat{\underset{\sim}{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_L)$.

We take the prior distribution on $\underset{\sim}{f}$ to be a symmetric Dirichlet distribution with parameter a,

$$\underset{\sim}{f} \sim Di_L(a\underset{\sim}{1}) \ , \tag{10.20}$$

i.e. the prior density of $\underset{\sim}{f}$ is taken proportional to $\prod_\ell f_\ell^{a-1}$. Having observed $\hat{\underset{\sim}{f}}$, the a posteriori density of $\underset{\sim}{f}$ is

$$\underset{\sim}{f}\,|\,\hat{\underset{\sim}{f}} \sim Di_L(a\underset{\sim}{1} + n\hat{\underset{\sim}{f}}) \ ,$$

with density function proportional to $\prod_\ell f_\ell^{n\hat{f}_\ell + a - 1}$. Letting $a \to 0$ to represent prior ignorance gives the well known result

$$\underset{\sim}{f}\,|\,\hat{\underset{\sim}{f}} \sim Di_L(n\hat{\underset{\sim}{f}}) \ . \tag{10.21}$$

Distribution (10.21) is quite similar to the bootstrap distribution (5.13),

$$\underset{\sim}{f}^{*}|\hat{\underset{\sim}{f}} \sim \text{Mult}_L(n,\hat{\underset{\sim}{f}})/n \ . \tag{10.22}$$

(i) Both distributions are supported entirely on those categories having $\hat{f}_\ell > 0$, i.e. on those categories in which data was observed. (ii) Both distributions have expectation vector $\hat{\underset{\sim}{f}}$. (iii) The covariance matrices are also nearly equal, $\text{Cov}(\underset{\sim}{f}|\hat{\underset{\sim}{f}}) = \hat{\underset{\sim}{\ddagger}}_{\hat{f}}/(n+1)$, $\text{Cov}_*(\hat{\underset{\sim}{f}}^{*}|\underset{\sim}{f}) = \hat{\underset{\sim}{\ddagger}}_{\hat{f}}/n$, where $\hat{\underset{\sim}{\ddagger}}_{\hat{f}}$ has diagonal elements $\hat{f}_\ell(1-\hat{f}_\ell)$ and off-diagonal elements $-\hat{f}_\ell\hat{f}_m$.

The point here is that the a posteriori distribution of $\theta(\underset{\sim}{f})|\hat{\underset{\sim}{f}}$ is likely to be well approximated by the bootstrap distribution of $\theta(\hat{\underset{\sim}{f}}^{*})|\hat{\underset{\sim}{f}}$, if $\theta(\underset{\sim}{f})$ is any reasonably smooth function of $\underset{\sim}{f}$. 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 $1-2\alpha$.

The prior distribution $\text{Di}_L(\alpha\underset{\sim}{1})$, $\alpha\to 0$, may seem unreasonable[†], but it gives a reasonable answer when $\theta(\underset{\sim}{f})$ is the median of a distribution on the real line. In this case, letting the discretization of $\mathcal{X}$ become infinitely fine, it can be shown that

$$\text{Prob}\{x_{(k_1)} < \theta(\underset{\sim}{f}) \leq x_{(k_2)}|\hat{\underset{\sim}{f}}\} = \sum_{k=k_1}^{k_2-1} b_{k,n-1}(.5) \ , \tag{10.23}$$

comparing nicely with the classic binomial interval (10.3).

[†]In a recent paper, Rubin (1979), this criticism is made, with the suggestion that it would be better to do the Bayesian analysis using a more informative prior distribution.

Realistically we would never believe that our a posteriori distribution for $\underset{\sim}{f}$ concentrates exclusively on just the data points already seen. Smoothing the distribution (10.21) even slightly splits the endpoint probabilities in (10.23), as in the application of the percentile method to the median.

7. The Bias-Corrected Percentile Method. The bootstrap distribution for the sample median, Chapter X, Section 3, is *median unbiased* in the sense that $\text{Prob}_*\{\hat{\theta}^* \leq \theta\} = .50$ (splitting the probability at $\hat{\theta} = x_{(m)}$). The argument which follows suggests that if $\text{Prob}_*\{\hat{\theta}^* \leq \hat{\theta}\} \neq .50$ then a bias correction to the percentile method is called for.

To be specific, define

$$z_o = \Phi^{-1}(\widehat{\text{CDF}}(\hat{\theta})) \ , \tag{10.24}$$

where $\text{CDF}(t) = \text{Prob}_*\{\hat{\theta}^* \leq t\}$ as at (10.12), and Φ is the cumulative distribution function for a standard normal variate. The *bias corrected percentile method* consists of taking

$$[\widehat{\text{CDF}}^{-1}(\Phi(2z_o - z_\alpha)), \widehat{\text{CDF}}^{-1}(\Phi(2z_o + z_\alpha))] \tag{10.25}$$

as an approximate $1-2\alpha$ central confidence interval for θ. Here z_α is the upper α point for a standard normal, $\Phi(z_\alpha) = 1-\alpha$.

Notice that if $\text{Prob}_*\{\hat{\theta}^* \leq \theta\} = .50$ then $z_o = 0$ and (10.25) reduces to (10.14), the uncorrected percentile interval. However even small differences of $\text{Prob}_*\{\hat{\theta}^* \leq \hat{\theta}\}$ from .50 can make (10.25) much different than (10.14). In the law school data, for example, $\widehat{\text{CDF}}(\hat{\rho}) = .433$ (i.e. 433 out of 1000 bootstrap replications $\hat{\rho}^*$ were less than $\hat{\rho} = .776$). Therefore $z_o = \Phi^{-1}(.433) = -.17$, and taking $\alpha = .16$, $z_\alpha = 1$, in (10.25) gives the approximate 68% interval

$$[\widehat{CDF}^{-1}(\Phi(-1.34)),\ \widehat{CDF}^{-1}(\Phi(.66))] = [\widehat{CDF}^{-1}(090),\ \widehat{CDF}^{-1}(.745)]$$

$$= [\hat{\rho} - .17,\ \hat{\rho} + .10]$$

for ρ. This compares with the uncorrected percentile interval $[\widehat{CDF}^{-1}(.16),\ \widehat{CDF}^{-1}(.84)] = [\hat{\rho} - .12,\ \hat{\rho} + .13]$ and the normal theory interval $[\hat{\rho} - .16,\ \hat{\rho} + .09]$.

The argument supporting (10.25) is based on hypothesizing a transformation to a normal pivotal quantity. Suppose there exists some monotonic increasing function $g(\cdot)$ such that the transformed quantities

$$\phi = g(\theta),\ \hat{\phi} = g(\hat{\theta}),\ \hat{\phi}^* = g(\hat{\theta}^*) \tag{10.26}$$

satisfy

$$\begin{aligned} \hat{\phi} - \phi &\sim \mathcal{N}(-z_o\,\sigma,\ \sigma^2) \\ \hat{\phi}^* - \hat{\phi} &\underset{*}{\sim} \mathcal{N}(-z_o\,\sigma,\ \sigma^2) \end{aligned} \tag{10.27}$$

for some constants z_o and σ. In other words, $\hat{\phi} - \phi$ is a *normal pivotal quantity*, having the same normal distribution under F and $\hat{F}$. (Remember, the distribution of $R = \hat{\phi} - \phi$ under $\hat{F}$ is what we call the bootstrap distribution of $R^* = \hat{\phi}^* - \hat{\phi}$, "$\underset{*}{\sim}$" indicating the distribution under i.i.d. sampling from $\hat{F}$.)

In parametric contexts, (10.27) is a device frequently used to obtain confidence intervals. Fisher's transformation $\phi = \tanh^{-1}\rho$ or the correlation coefficient is the classical example. Within the class of bivariate normal distributions it produces a good approximation to (10.27), with $\sigma^2 = 1/(n-3)$ and $z_o = -\frac{\rho\sqrt{n-3}}{2(n-1)}$. The distributions are not perfectly normal, and z_o is not perfectly constant, but the theory still produces useful intervals, as described in Section 10.4.

The middle statement in (10.26) is actually a definition of the estimator $\hat{\phi}$. If $\hat{\theta} = \theta(\hat{F})$ is a functional statistic, then $\hat{\phi} = g(\theta(\hat{F}))$ is the nonparametric maximum likelihood estimator for ϕ. The last relationship in (10.26) 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 distribution of $\hat{\phi}^*$ is the obvious mapping of the bootstrap distribution of $\hat{\theta}^*$. Letting

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

we have

$$\widehat{CDG}(g(t)) = CDF(t) \tag{10.28}$$

for all t.

The standard $1-2\alpha$ confidence interval for ϕ is

$$\phi \in [\hat{\phi} + z_o\sigma - z_\alpha\sigma,\ \hat{\phi} + z_o\sigma + z_\alpha\sigma] . \tag{10.29}$$

Using (10.27), we will see that mapping (10.29) back to the θ scale gives (10.25). First notice that (10.27) and (10.28) imply

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

which gives $z_o = \Phi^{-1}(\widehat{CDF}(\hat{\theta}))$ as at (10.24).

Using (10.27) again,

$$\text{Prob}_*\{\hat{\phi}^* < \hat{\phi} + z_o\sigma \pm z_\alpha\sigma\} = \text{Prob}_F\{\hat{\phi} < \phi + z_o\sigma \pm z_\alpha\sigma\} = \Phi(2z_o + z_\alpha) ,$$

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

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

Transforming (10.29) back to the θ scale by the inverse mapping $g^{-1}(\cdot)$ gives the interval with endpoints $g^{-1}(\hat{\phi} + z_o\sigma \pm z_\alpha\sigma) = g^{-1}\,\widehat{CDG}^{-1}[\Phi(2z_o \pm z_\alpha)]$ $= \widehat{CDF}^{-1}[\Phi(2z_o \pm z_\alpha)]$, the last equality following from (10.28), $[\widehat{CDF}]^{-1} = [\widehat{CDG}\ g]^{-1} = g^{-1}\,\widehat{CDG}^{-1}$. We have now derived (10.25) from (10.27) □

The normal distribution plays no special role in this argument. Instead of (10.27) we could assume that the pivotal quantity has some other symmetric distribution than normal, in which case "Φ" would have a different meaning in (10.25). In the unbiased case, where $z_o = 0$, the normal distribution plays no role at all since we get the uncorrected percentile interval (10.14). 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 (10.14) has the correct coverage probability.*

None of these arguments require knowing the form of the transformation g, only of its existence. Consider the correlation coefficient again, assuming that the true distribution F is bivariate normal. Applying the parametric bootstrap of Section 5.2, and following definition (10.25), will automatically give almost exactly the normal theory interval $\tanh[\hat{\phi} - \frac{\hat{\rho}}{2(n-1)} \pm \frac{z_\alpha}{\sqrt{n-3}}]$, *without any knowledge of the transformation* $\hat{\phi} = \tanh^{-1}\hat{\rho}$.

8. Typical Value Theory and the Percentile Method. The uncorrected percentile method (10.13), (10.14) is a direct analogue of typical value theory, as described in Section 9.2. In fact, if we let $\widehat{CDF}(t)$ be the cumulative distribution function of the subsample values rather than of the bootstrap values, $\widehat{CDF}(t) = \#\{\hat{\theta}_\delta \le t\}/(2^n-1)$, then (10.13), (10.14) gives the $1-2\alpha$ central subsample interval. The same connection holds

for the Monte Carlo versions of the two methods, where $\widehat{CDF}$, either subsample or bootstrap defined, is approximated by Monte Carlo simulations.

The asymptotic considerations of Section 9.4 suggest that the typical value intervals will be wider than the percentile intervals by a factor of about $\sqrt{(n+2)/(n-1)}$. We see this effect in Table 10.3. Ten i.i.d. samples $X_1, X_2, \ldots, X_{15}$ were obtained from the negative exponential distribution, $\text{Prob}\{X>x\} = e^{-x}$, $x > 0$. Four different methods were used to obtain confidence intervals for the expectation θ: (1) the percentile method, $B = 1000$, based on the bootstrap distribution of $\bar{X}^*$; (2) random subsampling, $B = 1000$ subsample values of $\bar{X}_S$; (3) the bias-corrected percentile method based on $\bar{X}^*$, $B = 1000$; (4) and the Pitman intervals. The latter assume that we are sampling from a translated and rescaled negative exponential, say $\theta + \sigma(X-1)$, X as above, and are the Bayes posterior intervals versus the uninformative prior $d\theta\, d\sigma/\sigma$, $\sigma > 0$, see Pitman (1938). The Pitman intervals are a parametric technique making full use of the negative exponential form, and as such give the "correct" answer. We will use them as the standard here, even though it is not clear that they are the optimum result.†

Before computing the confidence limits, each sample $x_1, x_2, \ldots, x_{15}$ was translated to have $\bar{x} = 0$ and scaled to have $\Sigma(x_i - x)^2/14 = 1$. This stabilized the entries of Table 10.3, without affecting comparisons between the different methods. With $n = 15$, $\sqrt{(n+2)/(n-1)} = 1.10$, and we see that this is just about the ratio of the widths of random subsample

†The Pitman intervals based on the translation model $\theta + \sigma(X-1)$, X negative exponential, $\sigma = 1$ known, uniform prior distribution $d\theta$, are completely different. They are longer toward the left than toward the right of $\bar{x}$.

Trial	Percentile Method (B=1000)	Random Subsample (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
AVE	-.39, -.32, .33, .43	-.44, -.36, .35, .46	-.36, -.29, .36, .47	-.34, -.27, .48, .61

Table 10.3. 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.

intervals compared to the percentile method. For $1-2\alpha = .90$, the outer two numbers in each quadruple, the ratio obtained from the AVE row is $(.46 + .44)/(.43 + .39) = 1.10$.

Both of these methods are disappointing in one major aspect: their intervals are not nearly as extended toward the right as the Pitman intervals. The bias correction is helpful in this regard, shifting all 10 intervals rightwards, though not far enough so. The methods discussed in Section 10.10 are more drastic.

The central limit theorem implies that the bootstrap distribution of $\bar{X}^*$ will be approximately $\eta(\bar{x}, \hat{\sigma}^2/n)$ when $\hat{\sigma}^2 = \Sigma(x_i - \bar{x})^2/n$. If this is an accurate approximation, then the percentile interval will equal, approximately, $\bar{x} \pm z_\alpha \hat{\sigma}/\sqrt{n}$. This is narrower than the standard t interval $\bar{x} \pm t_{\alpha,n-1} \hat{\sigma}/\sqrt{n-1}$; $t_{\alpha,f}$ the upper α point of a student's t distribution with f degrees of freedom, and it is reasonable to suggest widening the percentile interval by factor $\frac{t_{\alpha,n-1}}{z_\alpha} [\frac{n}{n-1}]^{\frac{1}{2}}$. Interestingly, this factor is quite close to $[\frac{n+2}{n-1}]^{\frac{1}{2}}$ for $\alpha = .05$. However, using this correction factor is not universally helpful, a counterexample being the case of the median. More pertinently, there are other defects of the percentile method, and of the other methods so far introduced, demanding greater attention. The worst of those is discussed next. A more direct method of correcting for the "t effect" is introduced in Section 10.

The bias corrected interval (10.25), based on the typical value distribution rather than the bootstrap, was tried here, but had little effect, often moving the typical value intervals slightly leftwards. In this example the typical value method appeared insensitive to asymmetry

in the observed sample. This suggests that it should be used with caution, or not at all, when distributional asymmetry is a definite possibility.

9. The Percentile Method for M Estimates. We consider m-estimates, as defined in (9.2), (9.3). For such estimates there are two seemingly distinct ways that the percentile method can be used to construct approximate confidence intervals for the true θ. These two ways turn out to give the same answer. So we see that whether or not the percentile method is any good in general, it is at least consistent with itself for m-estimates.

Define

$$M(t) = \int_{-\infty}^{\infty} \psi(x-t)\, dF(x),\ \hat{M}(t) = \frac{1}{n}\sum_{n=1}^{n} \psi(x_i - t)\ . \qquad (10.30)$$

The "true θ" is defined as those values of t satisfying $M(t) = 0$. For each value of t we can use (10.14) to construct an approximate central $1-2\alpha$ interval for $M(t)$,

$$[\hat{M}_{LOW}(t),\ \hat{M}_{UP}(t)]\ . \qquad (10.31)$$

This leads to an approximate $1-2\alpha$ region for θ, namely

$$\{t : 0 \in [\hat{M}_{LOW}(t),\ \hat{M}_{UP}(t)]\}\ . \qquad (10.32)$$

If (10.31) gave exact $1-2\alpha$ intervals for $M(t)$, then (10.32) would be exactly $1-2\alpha$ for θ.

Theorem 10.2. The region (10.32) is the same as the $1-2\alpha$ percentile interval for θ, $[\hat{\theta}_{LOW}(\alpha),\ \hat{\theta}_{UP}(\alpha)]$.

Proof. Theorem 10.2 follows from the equivalence of these two events,

$$\{\hat{M}^*(t) > 0\} \Leftrightarrow \{\hat{\theta}^* > t\}\ ,$$

for every bootstrap sample $\square$

Boos (1980) has suggested constructing confidence intervals for m-estimates using an interesting variation of (10.31), (10.32): for each t the interval (10.31) is replaced by an interval for M(t) constructed in the standard way, using a student's t approximation for the sampling distribution of $\hat{M}(t)$. Boos' method requires less computation than the percentile method, at the expense of greater reliance on approximation theory. It is nice to see that the two methods agree in principle.

Efron (1980A) applies (10.14) to obtain confidence intervals for trimmed and winsorized means, including the median, in a censored data situation. Brookmeyer and Crowley (1978) and Emerson (1979) use a technique similar to Boos' to obtain confidence intervals for the median, with censored data. The connection with the bootstrap is the same as above.

10. Bootstrap t and Tilting. Table 10.4 shows the application of two new methods to the 10 negative exponential samples of Table 10.3, again with the goal of providing approximate confidence intervals for the expectation. Only a brief description will be given of each method.

The true distribution of

$$T = \frac{\bar{X}-\theta}{S} ,$$

where $X_1, X_2, \ldots, X_{15}$ are independent negative exponentials, $\theta = 1$ is the expectation, and $S^2 = \Sigma(X_i - \bar{X})^2/14$, does not look anything like the normal theory distribution, which is a student's t with 14 degrees of freedom, divided by $\sqrt{15}$. Instead, it is sharply skewed left, $\text{Prob}\{T < -.69\} = \text{Prob}\{T > .36\} = .05$, $\text{Prob}\{T < -.50\} = \text{Prob}\{T > .28\} = .10$.

Trial	Bootstrap t (B=1000)	Exponential Tilting	Pitman Intervals	Sample Skewness
1	-.36, -.29, .53, .71	-.35, -.27, .42, .56	-.31, -.25, .45, .60	1.40
2	-.37, -.28, .52, .65	-.35, -.26, .41, .53	-.34, -.27, .48, .64	1.30
3	-.42, -.31, .39, .51	-.38, -.29, .33, .43	-.42, -.34, .56, .66	0.15
4	-.37, -.28, .51, .70	-.31, -.26, .42, .53	-.30, -.24, .44, .58	1.40
5	-.34, -.27, .62, .84	-.31, -.26, .43, .57	-.25, -.20, .37, .50	1.86
6	-.39, -.30, .45, .59	-.34, -.30, .39, .50	-.41, -.33, .55, .65	1.04
7	-.39, -.29, .41, .61	-.38, -.28, .36, .47	-.40, -.32, .54, .65	0.62
8	-.33, -.27, .63, .77	-.33, -.24, .45, .62	-.32, -.26, .46, .62	1.98
9	-.37, -.30, .45, .60	-.34, -.26, .38, .48	-.33, -.27, .47, .62	1.02
10	-.34, -.27, .56, .79	-.32, -.27, .39, .52	-.32, -.26, .46, .61	1.32
AVE	-.38, -.29, .51, .68	-.34, -.27, .40, .50	-.34, -.27, .48, .61	1.21
("True T")	(-.36, -.28, .50, .69)			

Table 10.4. Two more nonparametric confidence interval methods applied to the 10 negative exponential samples of Table 10.3. The averages of the bootstrap t endpoints almost equal the actual "T" distribution limits for the negative exponential, n = 15.

Knowing this distribution, we could construct confidence intervals for θ based on the observed values of $\bar{x}$ and s. For example, $[\bar{x} - .36s, \bar{x} + .69s]$ is a central 90% confidence interval for θ. (It can be shown that this is the .90 Bayes a posteriori interval for θ versus the uninformative prior $d\theta\, d\sigma/\sigma$ assuming the same model as for the Pitman interval: a translation-scale family based on the negative exponential, but where only $\bar{x}$ and s are observed, rather than the entire sample $x_1, x_2, \ldots, x_{15}$.) These intervals would all equal the entries given by "True T" in Table 10.4, since we have standardized each sample to have $\bar{x} = 0$, $s = 1$.

In a nonparametric problem we *don't* know the true distribution of T, but we can use the bootstrap to estimate it. The *bootstrap t* entries in Table 10.4 were obtained in this way. For each sample, B=1000 bootstrap values $T^* = (\bar{X}^* - \bar{x})/S^*$ were generated. The 90% central interval was then $[\bar{x} - \hat{T}_{UP}s, \bar{x} - \hat{T}_{LOW}s]$, where $\text{Prob}_*\{T^* < \hat{T}_{LOW}\} = .05$, $\text{Prob}_*\{T^* > \hat{T}_{UP}\} = .05$. Since we set $\bar{x} = 0$, $s = 1$, the bootstrap t entries for each trial are negatives of the T^* percentiles, -95%, -90%, -10%, -5%. Notice how closely the averages for the 10 trials approximate the True T values.

This method gives more realistic answers than any of the nonparametric techniques reported in Table 10.3, though the upper 95% point is a bit wild in trials 5, 8, and 10. There is an interesting connection of the bootstrap t with Johnson's (1978) work on Cornish-Fisher approximations for T, which we will not present here. A drawback is that the method seems specific to translation problems. An attempt to use it for the correlation coefficient, now defining

$$T = \frac{\hat{\rho} - \rho}{\hat{SD}(\hat{\rho})} ,$$

$\hat{SD}(\hat{\rho})$ being the jackknife standard deviation estimate, gave poor results.

Exponential Tilting in Table 10.4 refers to the following method applied to the observed sample $x_1, x_2, \ldots, x_n$, $n = 15$:

1) For a given value of t define the weights

$$w_i^t = e^{tx_i} \Big/ \sum_{j=1}^{n} e^{tx_j} \qquad i = 1,2,\ldots,n , \tag{10.33}$$

and the "trial value" of the expectation

$$\hat{\theta}^t = \sum_{i=1}^{n} w_i^t x_i . \tag{10.34}$$

2) Generate resampling vectors $\underset{\sim}{P}^*$ according to the multinomial distribution

$$\underset{\sim}{P}^* \overset{*}{\sim} \text{Mult}_n(n, \underset{\sim}{w}^t)/n . \tag{10.35}$$

(For $t = 0$ this is the distribution (6.6), but otherwise it puts different probabilities in the categories 1, 2, 3, ..., n.) Define the significance level $\hat{\alpha}^t$ corresponding to t as the bootstrap probability

$$\hat{\alpha}^t = \text{Prob}_{\underset{\sim}{w}^t}\{\hat{\theta}^* \leq \hat{\theta}\} , \tag{10.36}$$

where $\hat{\theta}^* = \sum_{i=1}^{n} P_i^* x_i$ and $\text{Prob}_{\underset{\sim}{w}^t}$ indicate probability under distribution (10.35).

3) The upper 95% point of the approximate confidence interval (= .56 for trial 1 of Table 10.4) is the value of $\hat{\theta}^t$ for t such that $\hat{\alpha}^t = .05$, etc.

Without going into details, the distribution $\hat{F}_{\underset{\sim}{w}^t}$ putting mass w_i^t on x_i is the closest distribution to $\hat{F}$ (in a certain plausible metric) which is supported entirely on the observed data points x_i, and which has expectation equal to the trial value $\hat{\theta}^t$. If $\hat{\alpha}^t$ is small, then the value $\hat{\theta}^t$ is excluded from the confidence interval for θ because even the distribution $\hat{F}_{\underset{\sim}{w}^t}$ would be unlikely to yield the observed value $\hat{\theta}$. Exponential tilting is similar to parametric techniques of obtaining confidence intervals for a real function of a vector of parameters.

The term "tilting" comes from exponential family theory. The tilted density $f_t(x)$ corresponding to a given density $f_o(x)$ on the real line is $f_t(x) = e^{tx-\psi(t)} f_o(x)$, $\psi(t)$ being chosen so that $\int_{-\infty}^{\infty} f_t(x)dx = 1$. The tilted bootstrap distribution of $\hat{\theta}^*$ under (10.35) is in fact obtained by tilting the usual bootstrap distribution of $\hat{\theta}^*$, (6.6) in just this way. The entries in Table 10.4 were obtained by first approximating the usual bootstrap distribution with B=1000 replications of $\hat{\theta}^*$, and then tilting this distribution to obtain the values $\hat{\alpha}^t$ = .95, .90, .10, .05.

Comparing Tables 10.4 and 10.3, exponential tilting gave results intermediate between the bias-corrected percentile intervals with the bootstrap t. Unlike the latter method, it performed well when applied (with suitable alterations) to the correlation coefficient problem.

Whether or not tilting is a useful approach, it emphasizes a principle limitation of the jackknife, the bootstrap, and the other methods we have discussed. They are nonparametric in not making specific model assumptions, but they all tacitly assume that the true distribution

is supported on the observed data points $x_1, x_2, \ldots, x_n$. This seems to be harmless enough in estimating a standard deviation, but may be seriously misleading in more delicate problems such as setting confidence intervals. Smoothing, as introduced in Section (5.3), helps overcome this objection, at the expense of introducing a parametric element into the estimation procedure.

References

Bickel P., and Freedman, D. (1980). "Some asymptotic theory for the bootstrap", Technical Report, University of California, Berkeley.

Boos, D. (1980). "A new method for constructing approximate confidence intervals from M-estimates", Journal of the American Statistical Association 75, 142-145.

Brookmeyer, R., and Crowley, J. (1978). "A confidence interval for the median survival time", Technical Report No. 2, Wisconsin Clinical Cancer Center, University of Wisconsin.

Carroll, R. (1979). "On estimating variances of robust estimates when the errors are asymmetric", Journal of the American Statistical Association 74, 674-679.

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

Efron, B. (1975). "The efficiency of logistic regression compared to normal discriminant analysis", Journal of the American Statistical Association 70, 892-898.

Efron, B. (1978). "Regression and ANOVA with zero-one data: measures of residual variation", Journal of the American Statistical Association 73, 113-121.

Efron, B. (1979A). "Bootstrap methods: another look at the jackknife", The Annals of Statistics 7, 1-26.

Efron, B. (1979B). "Computers and the theory of statistics: thinking the unthinkable", SIAM Review 21, 460-480.

Efron, B. (1980A). "Censored data and the bootstrap", Technical Report No. 53, Department of Statistics, Stanford University. (To appear in the Journal of the American Statistical Association.)

Efron, B. (1980B). "Nonparametric estimates of standard error: the jackknife, the bootstrap, and other methods", Technical Report No. 53, Department of Statistics, Stanford University.

Efron, B., and Stein, C. (1978). "The jackknife estimate of variance". (To appear in the Annals of Statistics.)

Emerson, J. (1979). "Nonparametric confidence intervals for quantiles in the presence of partial right censoring", Technical Report No. 50Z, Sidney Farber Cancer Institute, Boston, Massachusetts.

Geisser, S. (1975). "The predictive sample reuse method with applications", Journal of the American Statistical Association 70, 320-328.

Gordon, L., and Olshen, R. (1978). "Asymptotically efficient solutions to the classification problem", Annals of Statistics 6, 515-533.

Gray, H., Watkins, T., and Adams, J. (1972). "On the generalized jackknife, its extensions, and its relation to e_n-transformations", Annals of Mathematical Statistics 43, 1-30.

Gray, H., Schucany, W., and Watkins, T. (1975). "On the generalized jackknife and its relation to statistical differentials", Biometrika 62, 637-642.

Hampel, F. R., (1974). "The influence curve and its role in robust estimation", Journal of the American Statistical Association 69, 383-393.

Hartigan, J. A. (1969). "Using subsample values as typical values", Journal of the American Statistical Association 64, 1303-1317.

Hartigan, J. A. (1971). "Error analysis by replaced samples", Journal of the Royal Statistical Society Series B 33, 98-110.

Hartigan, J. A. (1975). "Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values", Annals of Statistics 3, 573-580.

Hartigan, J. A., and Forsythe, A. (1970). "Efficiency and confidence intervals generated by repeated subsample calculations", Biometrika 57, 629-640.

Hinkley, D. V. (1977). "Jackknifing in unbalanced situations", Technometrics 19, 285-292.

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

Hoeffding, W. (1948). "A class of statistics with asymptotically normal distributions", Annals of Mathematical Statistics 19, 293-325.

Huber, P. J. (1974). "Robust statistical procedures", Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Jaeckel, L. (1972). "The infinitesimal jackknife", Bell Labs. Memorandum #MM 72-1215-11.

Johnson, N. J. (1978). "Modified t tests and confidence intervals for asymmetrical populations", Journal of the American Statistical Association 73, 536-544.

Johnson, N., and Kotz, S. (1970). Continuous univariate distributions - 2, Houghton Mifflin, Boston, Massachusetts.

Kendall, M., and Stuart, A. (1958). The advanced theory of statistics, Griffin, London.

Kish, L., and Frankel, M. (1974). "Inference from complex samples (with discussion)", Journal of the Royal Statistical Society, Series B 36, 1-37.

Mallows, C. (1973). "Some comments on Cp", Technometrics 15, 661-675.

Mallows, C. (1974). "On some topics in robustness", Bell Labs. Memorandum.

Maritz, J. S. (1979). "A note on exact robust confidence intervals for location", Biometrika 66, 163-166.

McCarthy, P. J. (1969). "Pseudo-replication: half-samples", Review of the ISI 37, 239-263.

Miller, R. G. (1964). "A trustworthy jackknife", Annals of Mathematical Statistics 39, 1594-1605.

Miller, R. G. (1974A). "The jackknife - a review", Biometrika 61, 1-17.

Miller, R. G. (1974B). "An unbalanced jackknife", Annals of Statistics 2, 880-891.

Olshen, R. (1980). Private communication reporting on research done in collaboration with John Ross, Jr., Elizabeth Gilpon, and H. E. Henning.

Pitman, E. (1938). "The estimation of location and scale parameters of a continuous population of any given form", Biometrika 30, 391-421.

Pyke, R. (1965). "Spacings", Journal of the Royal Statistical Society, Series B 27, 395-449.

Quenouille, M. (1949). "Approximate tests of correlation in time series", Journal of the Royal Statistical Society, Series B 11, 18-84.

Quenouille, M. (1956). "Notes on bias in estimation", Biometrika 43, 353-360.

Rubin, D. B. (1979). "A Bayesian bootstrap", unpublished report, Princeton, Educational Testing Service.

Scheffé, H. (1959). The Analysis of Variance, Wiley, New York.

Schucany, W., Gray, H., and Owen, O. (1971). "On bias reduction in estimation", Journal of the American Statistical Association 66, 524-533.

Singh, K. (1980). "On asymptotic accuracy of Efron's bootstrap", Technical Report No. 158, Department of Statistics, Stanford University.

Steele, J. M. (1980). "Optimum triangulation of random samples in the plane", Technical Report #157, Department of Statistics, Stanford University.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions", Journal of the Royal Statistical Society, Series B 36, 111-147.

Tukey, J. (1958). "Bias and confidence in not quite large samples", (Abstract), Annals of Mathematical Statistics 29, 614.

Wahba, G., and Wold, S. (1975). "A completely automatic French curve: fitting spline functions by cross-validation", Communications in Statistics 4, 1-17.