COMMUN. STATIST.—THEORY METH., 24(3), 829-844 (1995)

IMPROVED BOOTSTRAP CONFIDENCE INTERVALS IN CERTAIN TOXICOLOGICAL EXPERIMENTS

Chris C. Frangos

Department of Statistics
University of the Witwatersrand
P O Wits 2050
Johannesburg, South Africa

William R. Schucany

Department of Statistical Science
Southern Methodist University
Dallas, Texas 75275
USA

Key Words and Phrases: influence function; jackknife; litter effect; overdispersion; studentized pivotal quantity.

ABSTRACT

The bootstrap, the jackknife, and classical methods are compared through their confidence intervals for the proportion of affected fetuses in a common type of animal experiment. Specifically, suppose that for the ith of M pregnant animals, there are x_i affected fetuses out of n_i total in the litter. The conditional distribution of x_i given n_i is sometimes modeled as binomial (n_i, p_i), where p_i is a realization from some unknown continuous density. The p_i are not observable and it is of interest in some toxicological experiments to find confidence intervals for $E(p)$. Theory suggests that the proposed parametric bootstrap should produce higher order agreement between the nominal and actual coverage than that exhibited by the usual

nonparametric bootstrap. Some simulation results provide additional evidence of this superiority of the modified parametric bootstrap over the jackknife and classical approaches. The proposed resampling is flexible enough to handle a more general model allowing correlation between p_i and n_i.

1. INTRODUCTION

Some investigators commonly carry out a toxicological experiment, where drugs or chemical agents are given to animals, these animals are mated, and the endpoint of interest is the rate of some effect on the fetuses or offspring. Usually the aim of such an experiment is to discover the proportion of fetuses that die or of live offspring that are malformed. Statistical analysis is relatively straightforward, but some confusion has occurred about the appropriate methods of analysis of the proportion affected; see Haseman and Hogan (1975). In the notation of Frangos and Stone (1984), a treatment is given to M pregnant animals and the ith animal has x_i affected fetuses out of n_i total in the litter. The conditional distribution of x_i given n_i and p_i is modeled as binomial (n_i, p_i), where the unobservable p_i has a marginal distribution, $g(p)$, with mean μ. For example, g might be a Beta density with unknown praameters a and b so that $\mu = a/(a+b)$. By assuming that p_i and n_i are uncorrelated, the general model is of the form:

$$\begin{aligned} E(x|n_i) &= \mu n_i \\ \text{var}(x|n_i) &= \delta(n_i)\ , \end{aligned} \tag{1.1}$$

for some specific function $\delta(n)$ of n.

The more general probability model for M independent litters is that (X_i, n_i, p_i) $i = 1, \ldots, M$ are independent triples from the same population. The complicating factor is that the p_i are not directly observable and arise from densities, g_i, that may differ depending on litter size, n_i. To make this explicit we write $g(p_i|n_i)$. This represents a flexible model for litter effect. For a review of a variety of analyses that have been proposed see Haseman and Kupper (1979). Many of these earlier approaches require an assumption

that the response probabilities, p_i, are not dependent upon n_i. There are certainly biological outcomes for which such a rate will be correlated with litter size. Our proposed resampling incorporates this feature by respecting the joint of (n_i, p_i) rather assuming that they are not correlated.

In this article we will be concerned with interval estimation of the proportion of affected fetuses in the population $\mu = E(x)/E(n)$. This is more general than model (1.1), but the definitions of μ agree when the more restrictive assumption is true. Specifically, we compare the classical large-sample interval estimators with interval estimators based on resampling methods, particularly the parametric bootstrap method of Efron (1982, 1987) and the jackknife procedure. For some discussion of classical approaches in similar contexts, there are obvious relationships to the topic of estimation in survey sampling when the sampling units are obtained from cluster sampling and the clusters are of random size.

The more challenging problem in this setting is to assess the significance of dose-response trends. Carr and Portier (1993) review a number of approaches to this, including a jackknife and a bootstrap. We believe that progress in this area will come easier after we have a clear understanding of inference for the appropriate response rate at a single fixed dose. Our proposed bootstrap resampling has several advantages over the ordinary bootstrap. We discuss these in the final section.

2. 'CLASSICAL' METHODS OF ESTIMATION

We consider two 'classical" estimators of μ. The pure-binomial estimator is denoted by

$$\hat{\mu}_0 = \sum_{i=1}^{M} x_i/N\ , \qquad \text{where} \quad N = \sum_{i=1}^{M} n_i\ , \tag{2.1}$$

and the average of rates within each litter by

$$\hat{\mu}_1 = \frac{1}{M} \cdot \sum_{i=1}^{M} \frac{x_i}{n_i} = \frac{1}{M} \cdot \sum_{i=1}^{M} \hat{p}_i = \bar{p} \ . \tag{2.2}$$

The distinction between $\hat{\mu}_0$ and $\hat{\mu}_1$ is discussed in Frangos and Stone (1984). If one assumes that there is a single common value of p, then confidence intervals for μ, called CLS_1, may be based on the easily derived result that $(\hat{\mu}_0 - \mu)/\hat{\sigma}_0$ is asymptotically $N(0,1)$ as $M \to \infty$ with the n_i values bounded, where $\hat{\sigma}_0^2 = \hat{\mu}_0(1-\hat{\mu}_0)/N$. A more general motivation for $\hat{\mu}_0$ is that $\bar{x}/\bar{n}$ results from the plug-in principle. Less restrictively, one may also use the asymptotic normality of $(\hat{\mu}_1 - \mu)/\hat{\sigma}_1$, where $\hat{\sigma}_1^2 = \sum_{i=1}^{M}(\hat{p}_i - \bar{p})^2/\{M(M-1)\}$, to derive alternative approximate confidence intervals that weight litters equally. These are centered on $\hat{\mu}_1$ and are denoted by CLS_2.

These intervals and resampling intervals will be compared in a simulation in Section 7. The jackknife methods are explained in the next section. Several bootstrap intervals are described in Sections 4 through 6.

3. JACKKNIFE METHODS OF ESTIMATION

Using the estimator (2.1) as an initial estimator of μ, Gladen (1979) investigates the jackknife estimator

$$\hat{\mu}_{0j} = \sum_{i=1}^{M} K_i \frac{x_i}{n_i} \ , \tag{3.1}$$

where

$$K_i = \frac{n_i}{N} + \frac{(M-1)}{M} \left\{ \frac{n_i}{N-n_i} - \frac{n_i}{N} \sum_{j=1}^{M} \left(\frac{n_j}{N-n_j} \right) \right\} .$$

Confidence intervals for μ may be based on the familiar result that the statistics

$$(\hat{\mu}_{0j} - \mu)/\hat{\sigma}_{0j} \qquad \text{or} \qquad (\hat{\mu}_0 - \mu)/\hat{\sigma}_{0j} \tag{3.2}$$

are asymptotically $N(0,1)$ as $M \to \infty$, where $\hat{\sigma}_{0j}$ is the jackknife estimator of the standard deviation of either $\hat{\mu}_0$ or $\hat{\mu}_{0j}$. The $100(1-2\alpha)\%$ confidence limits of the form $\hat{\mu}_{0j} \pm z_\alpha \hat{\sigma}_{0j}$ and $\hat{\mu}_0 \pm z_\alpha \hat{\sigma}_{0j}$ are denoted by JKN_1, which was called "G" in Frangos and Stone (1984), and JKN_2, respectively. Results here and in previous studies indicate that the confidence intervals JKN_1 and JKN_2 are almost identical with respect to coverage and expected length.

Hinkley and Wei (1984) apply an Edgeworth expansion to studentized parameter estimates when the standard error has been computed by the jackknife method. In this way they improve the normal approximation for an estimate $\hat{\theta}$ of a parameter θ. In the context of the present problem these improved jackknife confidence limits can be described as follows. Consider the estimator $\hat{\mu}_0$ of μ, and denote by $\hat{\mu}_{0,-i} (i = 1,2,\ldots,M)$, the same estimator calculated from the sample with (x_i, n_i) omitted. The jackknife estimate of the standard error for $\hat{\mu}_0$ is

$$\tilde{S} = \left[\sum_{i=1}^{M} \tilde{I}_i^2/\{M(M-1)\} \right]^{\frac{1}{2}} \ , \tag{3.3}$$

where $\tilde{I}_i$ is a finite sample estimate of the influence function of $\hat{\mu}_0$ at the point (x_i, n_i), given by

$$\tilde{I}_i = (M-1)(\hat{\mu}_0 - \hat{\mu}_{0,-i}) \ , \qquad (i = 1,2,\ldots,M) \ . \tag{3.4}$$

If $\tilde{V} = M\tilde{S}^2$, then the improved jackknife $(1-2\alpha)$ confidence intervals are

$$(\hat{\mu}_0 - z^*_{1-\alpha}\tilde{S} \ , \quad \hat{\mu}_0 - z^*_\alpha \tilde{S}) \ , \tag{3.5}$$

where

$$z^*_{1-\alpha} = z_{1-\alpha} + M^{-1/2} \left\{ \tilde{g}_{11} + \frac{1}{6} \cdot \tilde{g}_{32}(z^2_{1-\alpha} - 1) \right\} \ ,$$

$$\tilde{g}_{11} = -\sum_i \tilde{I}_i/\sqrt{\tilde{V}} - \frac{1}{2}(M^{-1}\sum_i \tilde{I}_i^3 + 2M^{-2}\sum_{i \neq k}\sum \tilde{I}_i \tilde{I}_k \tilde{Q}_{ik})/\tilde{V}^{3/2}$$

$$\tilde{g}_{32} = -(2M^{-1}\sum_i \tilde{I}_i^3 + 3M^{-2}\sum_{i\neq k}\sum \tilde{I}_i\tilde{I}_k\tilde{Q}_{ik})/\tilde{V}^{3/2}$$

and

$$\tilde{Q}_{ik} = M\{M\hat{\mu}_0 - (M-1)(\hat{\mu}_{0,-i} - \hat{\mu}_{0,-k}) + (M-2)\cdot\hat{\mu}_{0,-(i,k)}\}\ , \qquad (i \neq k)\ .$$

Beran (1984) finds the first-order Edgeworth expansion for the distribution of $n^{1/2}(\hat{\theta}-\theta)$. He then estimates the coefficients of this expansion by jackknifing and obtains the interesting result that the Edgeworth expansion is asymptotically equivalent to the corresponding bootstrap distribution estimate of $\hat{\theta}$, to be defined in Section 4. Hence, he suggests some corrections for skewness and bias to confidence limits for θ. Considering $\hat{\mu}_0$ as the estimate of μ, Beran's results lead to $100(1-2\alpha)\%$ confidence intervals of the form

$$\hat{\mu}_0 - M^{-1}\{BI\hat{A}S_j + \hat{S}_j SK\hat{E}W_j(z_\alpha^2 - 1)/6\} \pm M^{-1/2}\hat{S}_j \cdot z_\alpha\ , \qquad (3.6)$$

where

$$BI\hat{A}S_j = M^{-1}\sum_i F_i \quad , \qquad \hat{S}_j^2 = M^{-2}(M-1)^{-1}\sum_i \bar{F}_i^2\ ,$$

$$SK\hat{E}W_j = \frac{M^{-4}\sum_i \bar{F}_i^3 + 3M^{-3}(M-1)^{-1}\sum_{i\neq j}\sum \bar{F}_{ij}\bar{F}_i\bar{F}_j}{[M^{-3}\sum_i \bar{F}_i^2]^{3/2}}\ ,$$

$$F_i = (M+1)^2(\hat{\mu}_{0,+i} - \hat{\mu}_0)\ ,$$

$$F_{ij} = (M+2)^2\{\hat{\mu}_{0,+(i,j)} - \hat{\mu}_0\} - F_i - F_j\ ,$$

$$\bar{F}_i = F_i - F_{ii}/2\ , \quad \text{and} \quad \bar{F}_{ij} = F_{ij} - (F_i + F_j)/M\ .$$

The quantities $\hat{\mu}_{0,+i}$ and $\hat{\mu}_{0,+(i,j)}$ are estimates of μ based on all the observations including additional observations (x_i, n_i) and $\{(x_i, n_i),\ (x_j, n_j)\}$, respectively. The confidence limits given by (3.5) and (3.6) are called HWJ and JKN_3, respectively.

4. THE PERCENTILE PARAMETRIC BOOTSTRAP METHOD

Let (x_i, n_i), $(i = 1, 2, \ldots, M)$ be a random sample from the general probability model described in Section 1. From this random sample draw a bootstrap sample parametrically as follows:

Let (x_i^*, n_i^*), $(i = 1, 2, \ldots, M)$, be a random sample with replacement from (x_i, n_i), $(i = 1, 2, \ldots, M)$. Draw a sample (x_i^{**}, n_i^{**}), where $n_i^{**} = n_i^*$ and x_i^{**} is distributed according to the binomial $B(n_i^*, p_i^*)$ with $p_i^* = x_i^*/n_i^*$, $(i = 1, 2, \ldots, M)$. From the bootstrap sample $(x_1^{**}, n_1^{**}), \ldots, (x_M^{**}, n_M^{**})$, one finds the estimate

$$\hat{\mu}_0^1 = \sum_i x_i^{**} / \sum_i n_i^{**}\ . \qquad (4.1)$$

The above procedure is repeated independently B times and the bootstrap histogram of the estimates $\hat{\mu}_0^1, \hat{\mu}_0^2, \ldots, \hat{\mu}_0^B$ is constructed. Note that an important feature of this bootstrap resampling is that the pairing of (n_i, p_i) is retained. By not imposing a model for n, p or their joint behavior, this resampling scheme still has a strong nonparametric flavor even though it involves the binomial for conditioning x_i on n_i.

Confidence intervals for μ are derived by the percentile method using the 100α and $100(1-\alpha)$ percentiles of the bootstrap histogram of $\hat{\mu}_0^b$, $(b = 1, 2, \ldots, B)$. This yields the $1-2\alpha$ central confidence interval (BPC)

$$\mu \in [\hat{G}^{-1}(\alpha)\ ,\ \hat{G}^{-1}(1-\alpha)]\ , \qquad (4.2)$$

where

$$\hat{G}(t) = \frac{\#\{\hat{\mu}_0^b \leq t\}}{B} \qquad (4.3)$$

is the estimated bootstrap distribution function.

5. THE ACCELERATED BOOTSTRAP METHOD

Efron (1987) introduced an improved version of the bias-corrected

bootstrap, called BC_a, that incorporates both a bias and skewness correction. In this section we define the method and some variations of it and construct three BC_a confidence limits for μ. For more explanation and illustrations of the method see Chapter 14 of the recent monograph by Efron and Tibshirani (1993).

We resample parametrically, as in Section 4, B bootstrap samples $(x_{ij}^{**}, n_{ij}^{**})\,(i = 1, 2, \ldots, M), (j = 1, 2, \ldots, B)$, from the original data (x_i, n_i), $(i = 1, 2, \ldots, M)$. The central $100(1 - 2\alpha)\%$ confidence interval for μ by the BC_a method is given by

$$\mu \in \{\hat{G}^{-1}(\Phi(z[\alpha])) \;,\; \hat{G}^{-1}(\Phi(z[1-\alpha]))\} \;, \tag{5.1}$$

where

$$z[\alpha] = z_0 + \frac{(z_0 - z_\alpha)}{1 - a(z_0 - z_\alpha)} \;, \qquad z_0 = \Phi^{-1}[\hat{G}(\hat{\mu}_0)] \;, \tag{5.2}$$

and

$$a = \frac{1}{6} SKEW_{\mu=\hat{\mu}_0}(\dot{\ell}_\mu) \;. \tag{5.3}$$

Here $SKEW_{\theta=\hat{\theta}}(X)$ is the skewness at $\theta = \hat{\theta}$ of a random variable X, and statistic $\hat{\mu}_0$.

Efron (1987), approximates a by

$$\hat{a} = \frac{1}{6} \cdot \frac{\sum\{I_i\}^3}{\{\sum I_i^2\}^{3/2}} \;, \tag{5.4}$$

where I_i is the influence function of $\hat{\mu}_0$ at the point (x_i, n_i). He further suggests approximating I_i by the infinitesimal jackknife

$$\tilde{I}_i = \lim_{\epsilon \to 0} \frac{t[(1-\epsilon)\hat{F} + \epsilon \delta x_i] - t(\hat{F})}{\epsilon} \;, \tag{5.5}$$

where $\hat{\theta} = t(\hat{F})$ is the estimate of $\theta = t(F)$, F is the cumulative distribution function, $\hat{F}$ is the empirical distribution function and δx_i is the degenerate distribution at the point x_i.

In the simulation study of Section 7 we investigate confidence limits, called BC_{a1}, which are derived from (5.1) using, as an estimate of I_i, the negative jackknife

$$\tilde{I}_i = (M-1)(\hat{\mu}_0 - \hat{\mu}_{0,-i}) \;. \tag{5.6}$$

Similarly, confidence intervals, denoted by BC_{a2}, use (5.1) and the positive jackknife

$$\tilde{I}_{i+} = (M+1)(\hat{\mu}_{0,+i} - \hat{\mu}_0) \;. \tag{5.7}$$

By analyzing the estimate $\hat{\theta}$ in an expansion for differentiable statistical functions, von Mises (1947), one can find a higher-order approximation of I_i. Thus,

$$\hat{\theta} \sim \theta + M^{-1} \sum_i I_i + 2^{-1} M^{-2} \sum_j \sum_k I_{jk} \;, \tag{5.8}$$

where I_i and I_{jk} are the first and second-order influence functions of $\hat{\theta}$ at x_i and (x_j, x_k) respectively. From Hampel (1974)

$$\frac{d^2}{d\epsilon^2} t[(1-\epsilon)F + \epsilon G]_{\epsilon=0} = \int\int I_{ij}(x,y)\, dG_i(x)\, dG_j(y) \;, \tag{5.9}$$

where $G_i(x)$ is the degenerate distribution at the point x_i. A detailed analysis of $\hat{\theta}$ and $\hat{\theta}_{-i}$ using all the terms given in (5.8) gives

$$I_i^* = \tilde{I}_i + \frac{2M \sum_k \tilde{I}_{ik}(x_i, x_k) - M\tilde{I}_{ii}(x_i, x_i) - \sum_j \sum_k \tilde{I}_{jk}}{2M(M-1)} \;. \tag{5.10}$$

Using the definition of I_{ij} in (5.9) and the approach of Hinkley and Wei (1984), we find the estimate $\tilde{I}_{ij}$ of the second-order influence function for $\theta = \hat{\mu}_0$, and substitute it together with $\tilde{I}_i$ from (5.6) in (5.10). Hence, we find a second-order approximation of I_i and thus a second-order approximation for a in (5.4). Substituting this in (5.1) produces the confidence intervals BC_{a3} which are second-order corrected by means of influence functions.

6. BOOTSTRAP CONFIDENCE INTERVALS USING A STUDENTIZED PIVOTAL QUANTITY

It has been shown by Abramovitch and Singh (1985), that bootstrapping statistics of the form $T = (\hat{\theta} - \theta)/\sqrt{\hat{var}(\hat{\theta})}$, improves the normal approximation. Therefore, it is of interest to examine confidence intervals for μ which are of the form

$$[\hat{\mu}_0 - \hat{G}_s^{-1}(1-\alpha)\sqrt{\hat{var}(\hat{\mu}_0)}\ ,\quad \hat{\mu}_0 - \hat{G}_s^{-1}(\alpha)\sqrt{\hat{var}(\hat{\mu}_0)}] \ , \qquad (6.1)$$

where $\hat{G}_s(t)$ is the estimated distribution of a studentized pivotal quantity

$$T = \frac{(\hat{\mu}_0 - \mu)}{\sqrt{\hat{var}(\hat{\mu}_0)}} \ .$$

The distribution of T is estimated by the bootstrap method. Generate B bootstrap samples, as in Section 4, and calculate the quantities

$$T^{*b} = \frac{\hat{\mu}_0^b - \hat{\mu}_0}{\sqrt{\hat{var}(\hat{\mu}_0^b)}} \ , \qquad b = 1, 2, \ldots, B \ .$$

Hence the distribution of T is estimated by $\hat{G}_s(t) = \#\{T^{*b} \le t\}/B$. The variance of $\hat{\mu}_0$ is estimated by $\hat{var}(\hat{\mu}_0) = \sum_i \hat{I}_i^2/M^2$, where $\hat{I}_i = \tilde{I}_i$ from (5.6) or $\hat{I}_i = \tilde{I}_{i+}$ from (5.7). Therefore the $100(1-2\alpha)\%$ studentized bootstrap confidence intervals, (BST) are given by (6.1). If $Var(\hat{\mu}_0)$ is estimated with $\tilde{I}_i$, the confidence intervals are called BST_1, and if it is estimated with $\tilde{I}_{i+}$, the confidence intervals are called BST_2.

7. SIMULATION COMPARISONS

Using the IBM 3081-D and the IMSL library (as well as the Convex computer with NAG routines), random samples were generated for a variety of combinations. To conserve journal space only one is presented. The general patterns of the results are essentially the same for the others. Specifically, we examine

(i) sample sizes $M = 15, 20, 30$,

(ii) each value of n is equiprobably 5, 10, 15,

(iii) given n, each x is binomial (n, p) with p independently from a Beta (a, b) distribution with $a = 0,5$, $b = 5$, and

(iv) a run of 1 000 independent samples for each of the three sample sizes.

For each of the bootstrap confidence intervals, the same $B = 1\,000$ bootstrap replications were used to estimate each of the relevant percentiles, following the general guidelines in Efron and Tibshirani (1993) (pages 162, 188, 275).

In designing the study we were interested in practical sample sizes. The litter sizes (n) were set at an extreme case of a discrete uniform over nonadjacent values. Comparable results were obtained for n chosen from (8, 10, 12). Initially, we ran the experiment with larger litter sizes, but reduced them at a referee's suggestion that they could be more realistic. The patterns in the current table are more pronounced than for the larger n's. The specific beta density is monotone decreasing with an expected value near .09. We were interested in rates away .50, but not approaching Poisson limiting cases either. The beta (.5, 5) has adequate variability to yield noticeably different results from the more simplistic model with constant p. Comparable results were obtained for expected values of .077 and .067. Naturally, we recommend that one investigates combinations that differ markedly from these, rather than assuming that our findings extrapolate to those experiments.

For each sample, the confidence intervals CLS_1, CLS_2, JKN_1, JKN_2, JKN_3, HWJ, BPC, BC_{a1}, BC_{a2}, BC_{a3}, BST_1, BST_2, were calculated. The observed coverage and the average length for some of the above confidence intervals are shown in Table 1 for $1 - 2\alpha = 0,90$. The patterns are almost identical for 95% and 99% and therefore they are not reported here. The validity of the nominal confidence coefficient is our primary criterion. An interval that has actual coverage closer to the nominal will be judged better. It may be of interest to know at what cost the interval is better in terms of average length. A reasonable secondary consideration might be variability of these lengths.

TABLE I

Approximate Confidence Intervals for Expected Proportions
Estimates of the actual coverage in percent are on the first line and the average length of intervals are indented on the second line.

	Sample Size		
Method	$M = 15$	$M = 20$	$M = 30$
CLS_2	81	83	86
	.12	.10	.08
JKN_1	82	84	86
	.12	.10	.08
JKN_3	80	82	85
	.11	.10	.08
HWJ	83	85	87
	.12	.10	.08
BPC	88	88	92
	.13	.12	.10
BC_{a1}	91	91	92
	.14	.12	.10
BC_{a3}	91	91	92
	.14	.12	.10
BST_1	91	92	92
	.17	.14	.10

The standard errors of the tabled percentages are approximately .95. The results involving the positive and negative jackknife are virtually identical and thus JKN_2, BC_{a2} and BST_2 are not included in the table. The CLS_1 and CLS_2 methods yield poor results, see Frangos and Stone (1984). Only the results for CLS_2 are reported in Table 1, because CLS_1 had significantly lower undercoverage.

From Table 1 one sees the parametric bootstrap with "corrections" and BST_1, which uses studentized pivotal quantities, produce the most trustworthy confidence intervals with respect to coverage and length and specifically outperform the classical methods. Hence the inferiority of the original bootstrap relative to classical methods, reported by Frangos and Stone (1984) in an investigation of this same problem, is not present for more refined applications of bootstrap methodology. We find, also, that there is not much difference between the Percentile Bootstrap, BPC, and the two versions of the "accelerated" Bootstrap, BC_{a1} and BC_{a2}. Furthermore, the second order influence functions do not appear to improve BC_{a3} enough to matter in this setting. The exception to this occurred for 95% and 99% confidence intervals where the second-order intervals were significantly better than BPC.

The classical method CLS_2 is as reliable as the jackknife methods. This is somewhat surprising since the experiment has relatively less between batch variability, a situation which usually favors $\hat{\mu}_0$ over $\hat{\mu}_1$. One possible explanation is that in these skewed cases of practical interest, where μ is near 0 or 1, the jackknife in other settings has been found to need symmetrizing transformations. Hinkley and Wei's (HWJ) "skewness-adjusted" method represents the most improved jackknife confidence intervals. However, they are not comparable to the parametric bootstrap without corrections, BPC.

It is noteworthy that the average lengths of the BST confidence intervals are slightly greater than those produced by the other methods. This is a reasonable price to pay for bringing the true coverage up to the nominal. However, the sampling variance of the lengths of the confidence intervals over simulations is much greater for BST than for the other methods. In Table 1, its variance was around 3 times as great as for all the others, which were approximately the same.

8. DISCUSSION

It has been shown that reliable confidence intervals for an expected proportion can be constructed using percentile bootstrap methods. This represents a substantial improvement over the nonparametric bootstrap of

a cross-validation combination of $\hat{\mu}_0$ and $\hat{\mu}_1$, reported by Frangos and Stone (1984). For discrete data some adjustment to a in (5.4) may be appropriate. The theoretical foundation for this value involves Edgeworth expansions for absolutely continuous distributions. The relevant expansions for Bernoulli trials yield a Berry-Esseen bound of the same order, but we did not investigate a different a or whether this might improve the coverage of the BC_a methods. Further improvements may be achieved by the prepivoting or iteration that Beran (1987) or Hall (1986) advocate, assuming that the greater computing time is warranted.

The key features of the successful bootstrap here are

1) resampling the full known range of the x_i, namely $(0, 1, \ldots, n)$,
2) the percentile method,
3) a studentized pivotal quantity or other second-order corrections.

It is appropriate to question the robustness of the present approach if the binomial model does not hold. It should be conservative if the conditional distribution is hypergeometric, as may well be the case in some survey sampling or quality control applications. We recommend retaining feature (1) even if one is reluctant to impose the binomial model.

ACKNOWLEDGEMENTS

We are grateful to Professor Mervyn Stone for comments on an earlier version. The careful critique of two anonymous referees allowed us to make substantial improvements in the paper.

BIBLIOGRAPHY

Abramovitch, L. and Singh, K. (1985). "Edgeworth corrected pivotal statistics and the bootstrap," *Ann. Statist.*, **13**, 116–132.

Beran, R. (1984). "Jackknife approximations to bootstrap estimates," *Ann. Statist.*, **12**, 101–118.

Beran, R. (1987). "Prepivoting to reduce level error of confidence sets," *Biometrika*, **74**, 457–468.

Carr, G.J. and Portier, C.J. (1993). "An evaluation of some methods for fitting dose-response models to quantal-response developmental toxicology data," *Biometrics*, **49**, 779–791.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society of Industrial and Applied Mathematics, Philadelphia.

Efron, B. (1987). "Better bootstrap confidence intervals (with discussion)," *J. Amer. Statist. Assoc.*, **82**, 171–200.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Frangos, C.C. and Stone, M. (1984). "On jackknife, cross-validatory and classical methods of estimating a proportion with batches of different sizes," *Biometrika*, **71**, 361–366.

Gladen, B. (1979). "The use of the jackknife to estimate proportions from toxicological data in the presence of litter effects," *J. Amer. Statist. Assoc.*, **74**, 278–283.

Hall, P. (1986). "On the bootstrap and confidence intervals," *Ann. Statist.*, **14**, 1431–1452.

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

Haseman, J.K. and Hogan, M.D. (1975). "Selection of the experimental unit in teratology studies," *Teratology*, **12**, 165–172.

Haseman, J.K. and Kupper, L.L. (1979). "Analysis of dichotomous response data from certain toxicological experiments," *Biometrics*, **35**, 281–293.

Hinkley, D.V. and Wei, B. (1984). "Improvements of jackknife confidence limit methods," *Biometrika*, **71**, 331-339.

Von Mises, R. (1947). "On the asymptotic distributions of differentiable statistical functions," *Ann. Math. Statist.*, **18**, 309-348.

Received September, 1994; Revised, November, 1994.