BOOTSTRAP CONFIDENCE INTERVALS FROM ADAPTIVE SAMPLING OF AN INSECT POPULATION

Jeffrey S. Pontius

Mary C. Christman

BOOTSTRAP CONFIDENCE INTERVALS FROM ADAPTIVE SAMPLING OF AN INSECT POPULATION

Jeffrey S. Pontius
Department of Statistics
Kansas State University
Manhattan KS 66506-0802

Mary C. Christman
Department of Mathematics and Statistics
The American University
Washington D.C. 20016-8050

Abstract We construct 90% normal, percentile, and bias-corrected and accelerated confidence intervals using a finite population bootstrapping algorithm based on adaptive sampling in an agroecosystem. We evaluate the interval estimates based on sampling simulations of a spatially arranged population of plots that contain counts of beet webworms and based on an adaptive condition that generates small networks. The sampling distributions of the original sample estimates and of the bootstrap estimates were generally similar and symmetric. The simulation coverages were from 84% to 90% and similar under any of the sample sizes and any of the three confidence interval types. This study also serves as an example of how adaptive sampling may be used to estimate population characteristics of insects in agroecosystems.

Keywords: beet webworms, coverage, Horvitz-Thompson estimation, Sitter bootstrapping

1. Introduction

Adaptive sampling (Thompson and Seber 1996) is a design where units not in an initial sample can be added to the (final) sample if units in the initial sample meet a condition of interest (defined on the variable of interest, *Y*). Adaptive sampling can improve the efficiency of estimation when the units possessing the condition of interest are aggregated and occur infrequently in the population. The Horvitz-Thompson estimator of the parametric total

$$\hat{\tau} = \sum_{j=1}^{\nu} \frac{y_j}{\pi_j}$$

is usually used, where y_j is the total of the unit values (y_i) in network j,

$$\pi_j = 1 - \frac{\binom{N - m_j}{n}}{\binom{N}{n}}$$

is the probability that, with an initial without replacement simple random sample (srswor), network j is encountered by at least one unit in the initial sample (m_j is the number of units in network j), and v is the number of distinct networks in the adaptive (final) sample.

For example, consider a census of the counts of beet webworms per 3-foot of row (a plot) from a grid of plots (population) in a beet field (Beall 1940, Table VI; see Figure 1). If the counts are high in certain plots we might want to observe the adjoining plots if we suspect that they will contain high counts. To do this we set the *condition* of interest at, e.g., 3 or more webworms per plot ($y_i \geq 3$). The shaded areas in Figure 1 label those plots that meet the condition. If the condition is satisfied for a plot in the initial sample, then we could observe the 4 nondiagonal adjacent plots (the *neighborhood*) and count the number of webworms in each of those plots. If any of those plots meet the condition, then we observe their neighborhoods, and so on until no units satisfy the condition in neighborhoods being inspected. All of the units in the neighborhood that satisfy the condition are a *network*.

For example, if the plot in row 3, column 4 ($y_i = 4$) of Figure 1, is in the initial sample (it meets the condition $y_i \geq 3$), then we inspect the four adjacent plots (above, right, below, and left) and observe that two of the four plots (left and below) satisfy $y_i \geq 3$. Then we inspect each of their neighborhoods, adding two more plots (each with $y_i = 3$) to the sample. Because no other plots in the third series of inspections satisfy $y_i \geq 3$, we have found all of the plots in that network given the defined neighborhood. That network contains the five plots, as indicated by their y_i , {4, 3, 3, 3, 3} (the plot in row 1, column 4 is not in the neighborhood of the plot in row 2, column 3). Then the next plot in the initial sample is observed and its neighborhood observed *if* that plot satisfies the condition.

Confidence intervals for τ are typically based on the normal distribution. However, the sampling distribution of $\hat{\tau}$ can be very asymmetric (Christman, 1997), which casts doubt on the usefulness of using normal distribution based confidence intervals. An alternative is to use bootstrapping to generate an empirical sampling distribution of $\hat{\tau}$, and compute interval estimates from the $\hat{\tau}$ sampling distribution. Brown (1994) used bootstrapping to assess the bias of $\hat{\tau}$ in adaptive sampling. However, the bootstrapping procedure was the usual one (see Efron and Tibshirani 1993), which is not appropriate for finite populations.

Sitter (1992a, b) developed a bootstrapping algorithm for sampling from a finite population of size N with a fixed sample size n. In his algorithm, the original sample is resampled according to the sampling design that was used to select the original sample from the population. Specifically, if the original sampling fraction is $1/k = n/N$, then take k independent subsamples of size $n' = n/k = n^2/N$ from the original sample. The union of the k subsamples generates a bootstrap sample of size n. Sitter (1992a) showed that for linear estimators the bootstrap estimates are second order correct. For example, take a srswor of $n = 65$ plots from the N = 325 plots. Using Sitter's algorithm, we take $k = N/n = 5$ independent subsamples of size $n' = 13$ units and combine them to obtain a bootstrap sample of size $kn' = 65$.

Because the adaptive sample contains units from the initial sample and units adaptively observed, we use a data transformation that allows us to account for the spatial aspects imposed on the sampling and to treat the adaptive sample as a srswor of size n. Each observed y_i is replaced with the sum of the y_i in the network to which unit i belongs. Hence, sampling any single unit in the population is equivalent to sampling the entire network to which it belongs, which is exactly how adaptive sampling behaves.

In adaptive sampling, each network is used once in $\hat{\tau}$ even though a network can appear in the adaptive sample more than once. Under the data transformation, the actual sampling design is that networks of size 1 are selected according to srswor and networks containing more than one unit are sampled with replacement. Hence, the sample may contain multiple copies of a network, but a network is used only once in $\hat{\tau}$. We modified Sitter's algorithm so that each subsample is selected srswor from the initial sample, and in each subsample duplicates of each network are removed so that each subsample contains only unique networks. Note that the bootstrap sample may contain multiple copies of a network. Hence, the bootstrap estimate is computed using these multiple copies.

2. Simulation Procedure

The population is the set of N = 325 plots (Figure 1) with associated values of interest, y_i, being the counts of beet webworms per plot. Our statistical goal is to estimate the total number of beet webworms in the field (τ = 277 webworms) using a confidence interval estimate. We use adaptive sampling with the condition $y_i \geq 3$ webworms in a plot, and use the symmetric four unit neighborhood. The initial design is srswor with sample sizes n = 18, 26, 36, and 65. The number of subsamples and subsample sizes are (k, n'): (18, 1), (13, 2), (9, 4), and (5, 13), respectively on n. Note that some of the k or n' are not exact integers, but are very close approximations, hence the selected n. The noninteger k or n' will invoke a small amount of bias in the estimation.

For each n, generate 500 adaptive samples using the population of plots and, for each generated sample, take 1000 bootstrap samples using Sitter's algorithm for srswor. For each subsample of each bootstrap sample remove units so that unique networks remain within each subsample. Compute $\hat{\tau}$, say $\hat{\tau}_b$, on each bootstrap sample (of size $\leq n$). Then compute percentile, normal, and bias-corrected and accelerated (BCa) interval estimates (Efron and Tibshirani 1993) based on the 1000 $\hat{\tau}_b$. The interval endpoints in the percentile method are computed by evaluating the ordered sample of 1000 $\hat{\tau}_b$ at the 0.05 and 0.95 percentiles. The normal interval endpoints are computed using $\hat{\tau}_b \pm 1.645\text{s.d.}(\hat{\tau}_b)$. The BCa method is used to adjust percentile endpoints for bias in $\hat{\tau}_b$ relative to τ and for changes in the standard deviation of $\hat{\tau}_b$. An advantage of BCa intervals over percentile intervals is that they are second order accurate (percentile intervals are first order accurate). That is, in general, BCa intervals give more accurate coverage than percentile intervals. We used S+ (version 3.4, release 1 for Sun SPARC 1966) (see Venables and Ripley 1994 for S+) to implement the simulations. BCa was implemented using the S+ function `bcanon` in Efron and Tibshirani (1993).

As an example, suppose we take an initial sample of n = 26 plots by srswor using the population in Figure 1. Then we use k = 13 and n' = 2. With r labeling a row and c labeling a column, suppose our (partially displayed) initial sample is $\{(r, c; y_i) : (3, 4; 4), (27, 4; 3), (5, 1; 1), (48, 3; 0), (48, 2; 2), (3, 3; 3), ... \}$ of n = 26 plots. In one of the 1000 bootstrap samples of this original sample, suppose three of our thirteen subsamples are $s_1 = \{(48, 3; 0), (5, 1; 1)\}$, $s_2 = \{(3, 4; 4), (3, 3; 3)\}$, and $s_3 = \{(3, 4; 4), (27, 4; 3)\}$ (there are ten other subsamples). In s_1 each plot is a network of size one (they are in the initial sample), so both are retained for the final bootstrap sample. In s_2 both plots are members of the same network, so one copy of that network is retained. In s_3 both plots are in separate networks, so both networks are retained. This approach is applied to all of the other ten subsamples. Then $\hat{\tau}_b$ is computed on that bootstrap sample by combining the thirteen subsamples.

3. Results and Discussion

3.1 Performance

First we look at the performance of the adaptive bootstrapping approach. In general, the means of the bootstrapped estimates were close to the sample estimates (Figure 2). The deviates from the line are probably because of some bootstrap samples that contained a high proportion of networks repeated in the bootstrap sample. Sampling distributions for sample estimates (Figure 3) and bootstrapped estimates (Figure 4) are reasonably symmetric, with the variability being slightly less for the bootstrapped sample estimates (means over 1000 bootstrap samples) than for the sample estimates (Table 1). Note that these results are not unexpected because the network sizes are small relative to the size of the population (Figure 1).

3.2 Confidence Intervals

The adaptive bootstrap approach is a method of calculating confidence interval estimates that does not rely on large sample theory for the sampling distribution of $\hat{\tau}$. Because of the potential bias introduced by the modified use of Sitter's algorithm, we considered two types of confidence interval estimation methods: the percentile and the BCa. The results of the simulations are shown in Figures 5-1 and 5-2. For all three confidence interval types, the biases are minimal and the interval widths are similar for a given sample size, n. In fact, the average widths of the three interval types are similar as are the standard deviations of those widths (Table 2). Not surprisingly, for each of the types, the average widths and standard deviations of the widths decrease as the sample sizes increase.

All interval types for estimating the total abundance somewhat underestimated the actual 90% confidence interval (Table 3). Coverage ranged from a low of 84% for the intervals based on the normal approximation to a high of 90% for the BCa intervals. If one were to use the adaptive bootstrap approach to estimate the actual number of beet webworms in a field similar in environment to the one used in this study, the interval estimate would probably be slightly too narrow. We think that the undercoverage may be because of the use of Sitter's algorithm for taking bootstrap samples. The algorithm was originally constructed for use with sampling strategies that do not rely on either random sample sizes nor deletion of some of the sample data under the estimation procedure. Both of these are inherent in adaptive sampling designs which use the Horvitz-Thompson estimator of τ. Although similar in their coverages, the confidence interval based on the normal approximation had the worst coverage for all of the sample sizes, whereas the BCa interval always had the best coverage.

In our example, the sampling distributions of the estimator were reasonably symmetric and hence confidence intervals based on the normal approximation are appropriate. However, in most situations in which adaptive sampling is more efficient than srswor, it is likely that the sampling distributions will be highly skewed. For populations in which individual networks comprise a large portion of the total size, we would expect the sampling distributions to be asymmetric. In such cases, the normal approximation would be invalid.

4. Summary

Sitter's bootstrap algorithm worked reasonably well given the population and the network sizes used. We would anticipate that with larger networks (proportional to the size of the population) that the variability in interval estimates would increase. The bootstrapped estimates were generally close to the sample estimates, and the simulated sampling distributions were reasonably symmetric.

Simulated confidence interval coverages were from 84% to 90% of the 90% target, close to the parametric total (essentially no bias), and similar over sample sizes and confidence interval types. Interval estimates became narrower as sample sizes increased.

We are extending this initial research to look at the effect of proportional increases in networks (relative to the size of the population), bias, and second-order accuracy to further explore the application of Sitter's bootstrapping algorithm to interval estimation in adaptive sampling.

5. References

Beall, G. 1940. The fit and significance of contagious distributions when applied to observations on larval insects. *Ecology*. 21:460-474.

Brown, J.A. 1994. The application of adaptive cluster sampling to ecological studies. In *Statistics in Ecology and Environmental Monitoring*, D.J. Fletcher & B.F.J. Manly, editors. Otago Conference Series #2, University of Otago Press, Dunedin: New Zealand. pp.86-97.

Christman, M. C. 1997. Efficiency of some sampling designs for spatially clustered populations. *Environmetrics*. 8:145-166.

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

Sitter, R.R. 1992a. A resampling procedure for complex survey data. *Journal of the American Statistical Association*. 87:755-765.

Sitter, R.R. 1992b. Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*. 20:135-154.

Thompson, S.K. and G.A.F. Seber. 1996. *Adaptive Sampling*. John Wiley & Sons, Inc., New York.

Venables, W.N. and B.D. Ripley. 1994. *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York.

Acknowledgement We thank Tom Loughin for his suggestions in constructing the coverage density plots. This work was partially supported by the National Science Foundation under a grant (Number DMS-9631318) made to Mary C. Christman.

1	1	2	3	1
3	2	3	1	0
0	1	3	4	0
1	0	0	3	1
1	2	0	3	2
0	2	0	0	0
0	1	1	1	0
1	0	0	1	0
0	0	2	0	0
0	0	1	0	0
0	0	0	0	2
1	0	0	1	0
4	2	1	1	0
1	2	0	3	1
0	0	0	1	0
4	0	0	0	0
1	0	0	0	2
0	0	0	2	0
1	0	0	0	0
0	0	0	2	1
3	0	3	1	0
1	0	0	2	1
0	0	0	0	0
0	1	2	0	0

0	1	0	0	0
1	1	0	1	0
1	0	2	3	0
2	1	0	2	0
2	0	0	4	4
0	0	0	0	3
1	2	2	3	1
1	1	5	0	2
3	3	1	5	0
1	2	0	1	2
0	0	0	3	2
0	3	1	2	1
1	1	0	2	3
2	2	0	1	0
1	2	0	1	2
1	1	1	0	1
0	2	1	0	0
0	1	1	0	0
1	0	0	2	0
0	0	1	3	0
0	1	1	0	0
2	2	1	0	0
1	0	2	1	0
0	2	0	0	0

1	2	1	0	1
3	1	1	3	3
0	0	0	1	2
1	0	1	1	1
3	3	0	2	0
1	0	0	0	0
1	1	0	0	2
0	0	0	0	0
1	2	2	0	0
0	0	1	1	0
1	0	0	2	0
2	1	0	1	0
0	0	0	0	0
1	1	0	0	1
0	1	0	1	1
1	0	0	0	0
0	2	0	0	0

Figure 1. Spatial layout of the treatment 3 beet field in Beall (1940, Table VI). Squares are 3-foot of row of beets, and numbers are counts of beet webworms. Shaded plots meet the condition of interest $y_i \geq 3$.

Table 1. Standard deviations of the sample estimates ($\hat{\tau}$) and of the averages of bootstrapped estimates.

n	sample estimates ($\hat{\tau}$)	means of bootstrapped estimates
18	81.2	78.8
26	67.9	66.3
36	56.0	55.4
65	42.8	40.7

Table 2. Average widths and standard deviations of the widths of estimated confidence intervals.

		type	
n	percentile	normal	BCa
18	257 (57)	259 (56)	263 (59)
26	217 (40)	218 (40)	218 (41)
36	182 (29)	182 (27)	183 (27)
65	134 (15)	134 (15)	133 (15)

Table 3. Estimated confidence interval coverages from simulations.

		type	
n	percentile	normal	BCa
18	.85	.84	.87
26	.86	.85	.86
36	.87	.87	.90
65	.85	.86	.88

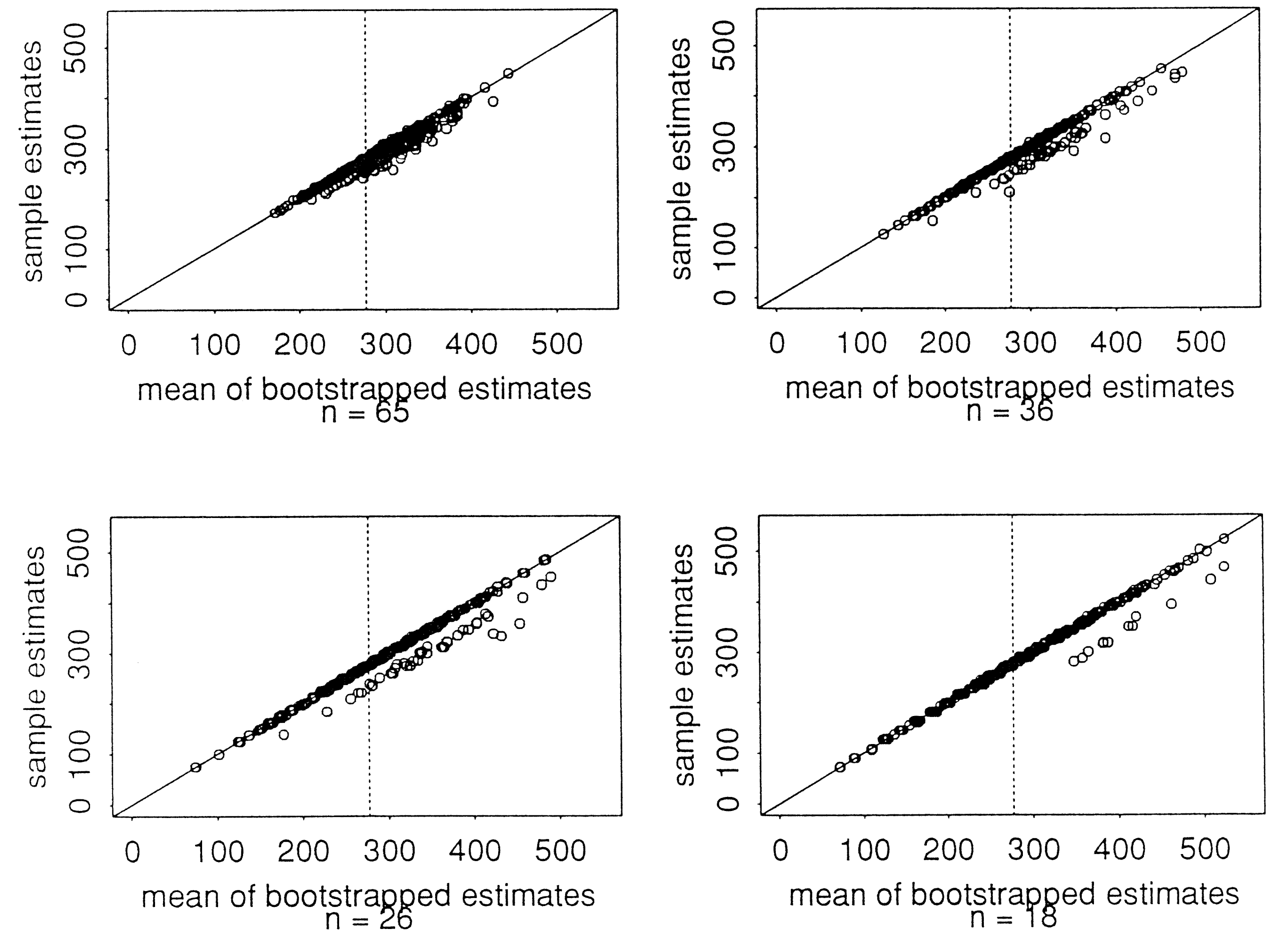


Figure 2. Sample estimates, $\hat{\tau}$, and means of 1000 bootstrap estimates. A circle on the line indicates that the mean of the bootstrap estimates agrees with the sample estimate.

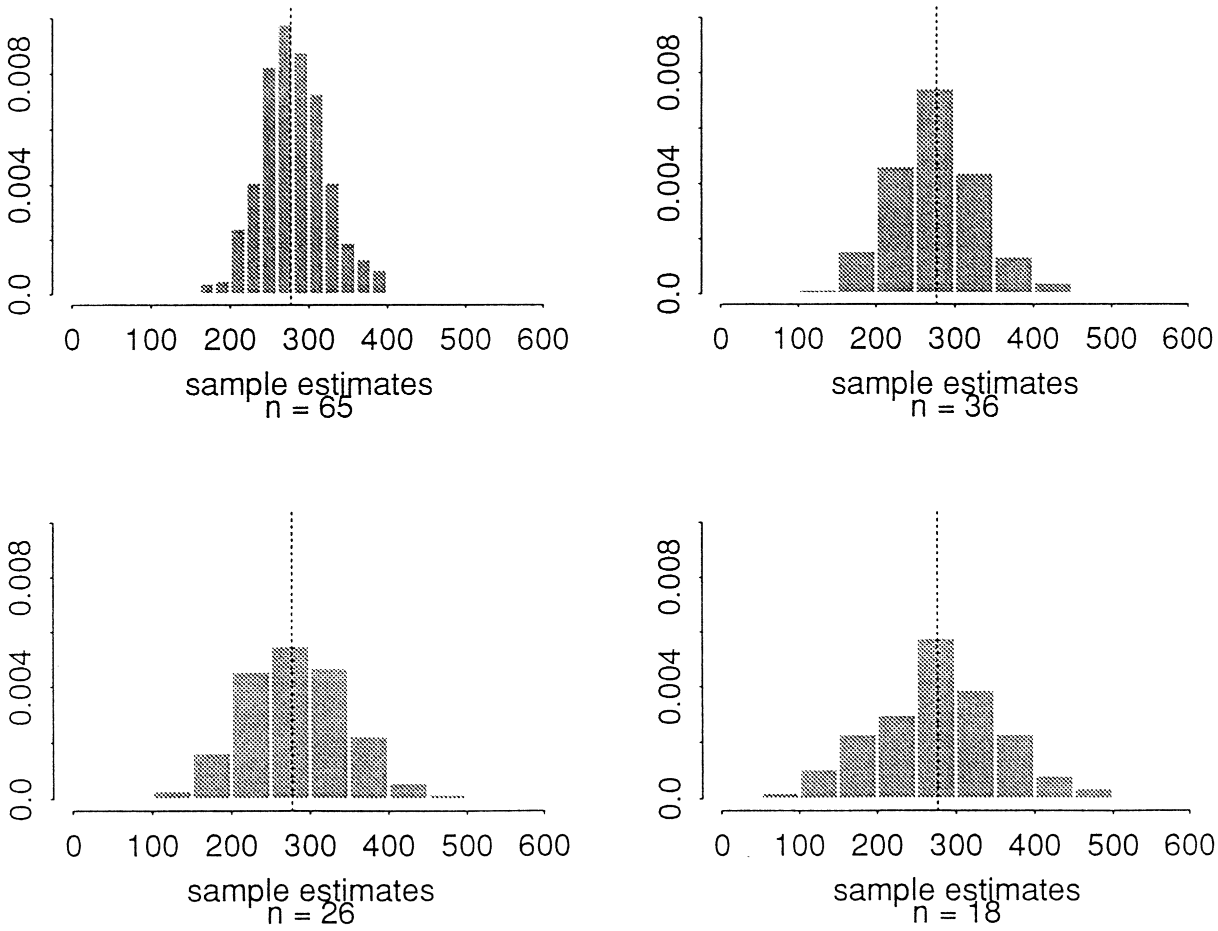


Figure 3. Simulation sampling distributions of 500 sample estimates.

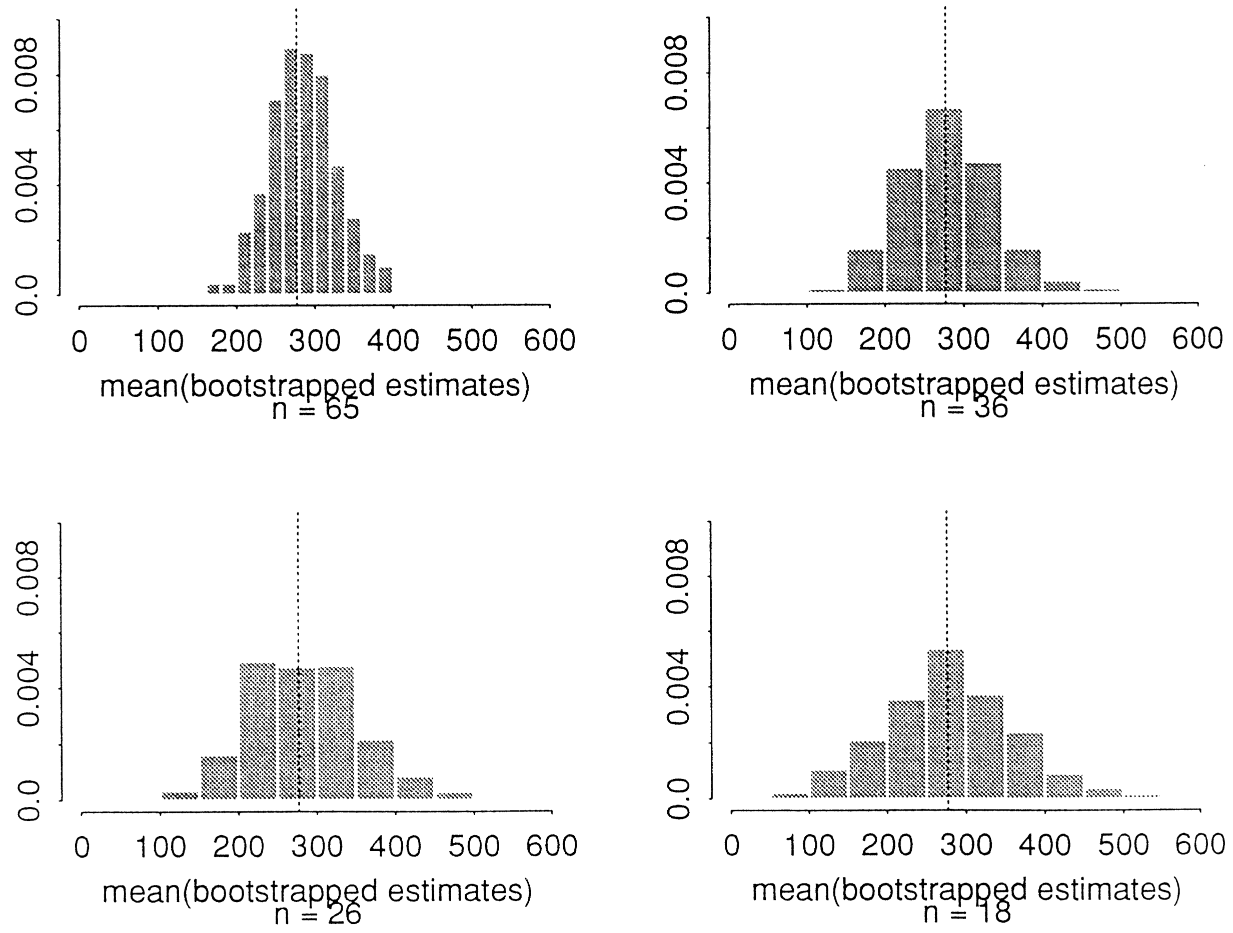
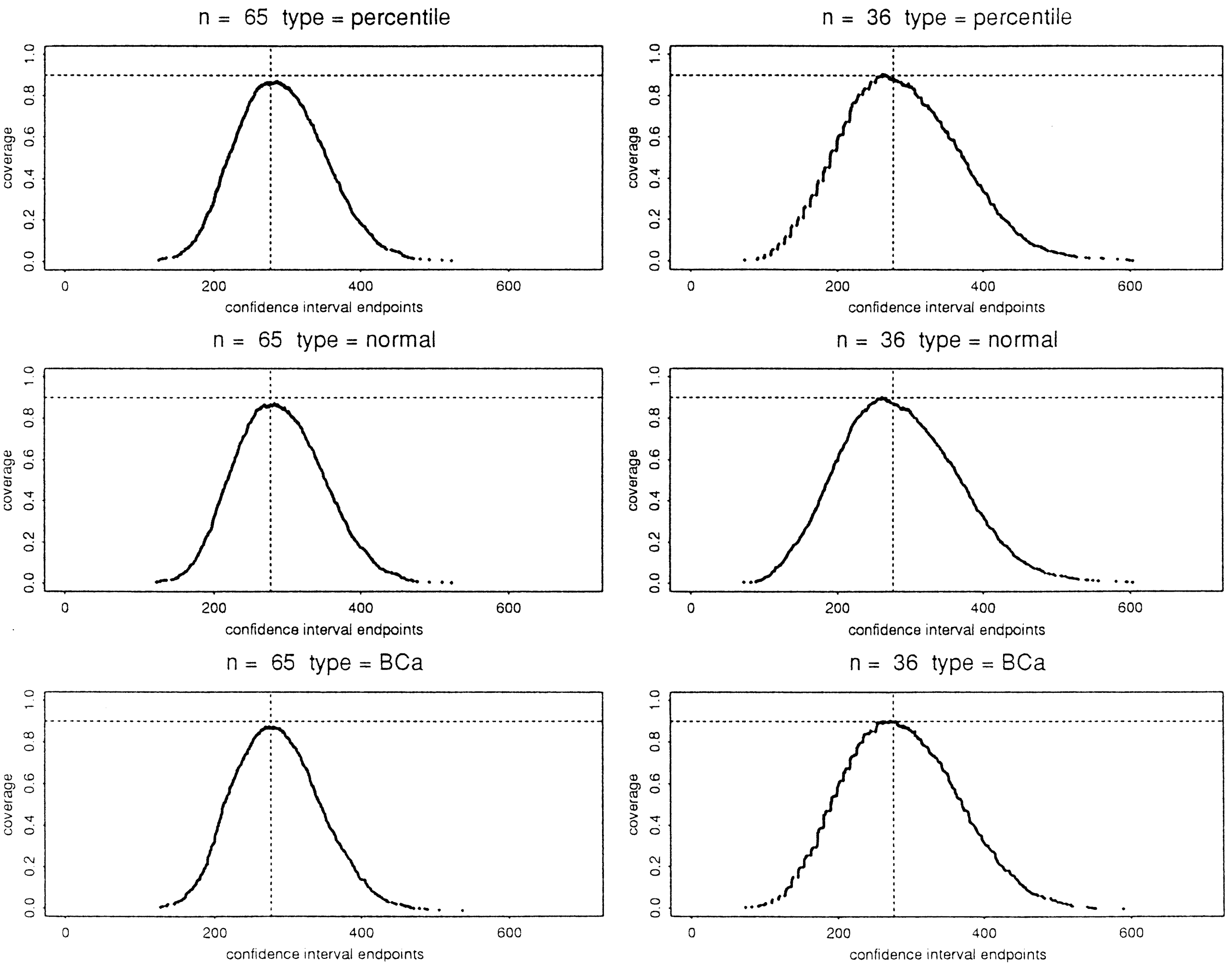


Figure 4. Simulation sampling distributions of 500 means of 1000 bootstrap estimates.


n = 65 type = percentile

n = 36 type = percentile

n = 65 type = normal

n = 36 type = normal

n = 65 type = BCa

n = 36 type = BCa

Figure 5-1. Coverage plots (n = 65 and 36) of percentile, normal and BCa confidence interval estimates. The horizontal dashed lines are at .9 confidence coefficient, and the vertical dashed lines are at τ = 277. The ideal plot should have the peak of a symmetric curve at the intersection of the two dashed lines, and small width at the baseline.

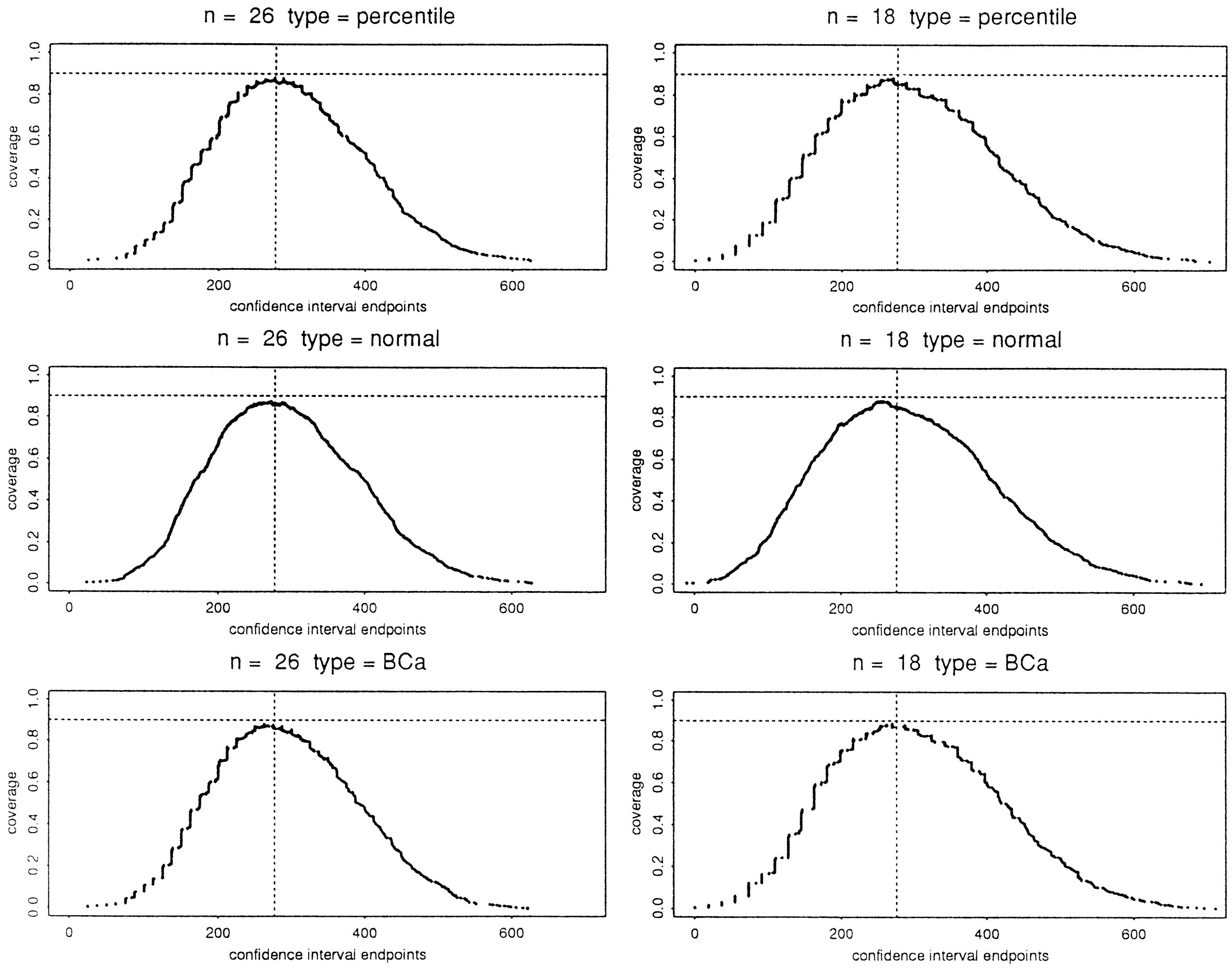


Figure 5-2. Coverage plots (n = 26 and 18) of percentile, normal and BCa confidence interval estimates.