DOCUMENT RESUME

ED 372 124 TM 021 969

AUTHOR	Fan, Xitao
TITLE	Does Bootstrap Procedure Provide Biased Estimates? An Empirical Examination for a Case of Multiple Regression.
PUB DATE	Apr 94
NOTE	31p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Estimation (Mathematics); Monte Carlo Methods; *Regression (Statistics); *Sample Size; *Sampling; Simulation; Statistical Studies; *Validity
IDENTIFIERS	*Bootstrap Methods; Population Parameters

ABSTRACT

This paper empirically and systematically assessed the performance of bootstrap resampling procedure as it was applied to a regression model. Parameter estimates from Monte Carlo experiments (repeated sampling from population) and bootstrap experiments (repeated resampling from one original bootstrap sample) were generated and compared. Sample sizes of 20, 30, 50, and 100 were considered in the simulation. Ten independent Monte Carlo experiments and 10 independent bootstrap experiments were conducted respectively for each sample size condition, with 1,000 samples (resamples for bootstrap) for each experiment. Estimates for standardized regression coefficients were obtained from each sample, and the mean estimates across samples were evaluated in relation to the population parameters. The results indicate that, as the number of resamples increases, the mean bootstrapped estimates did not show a clear tendency to converge on the population parameters. But, with the increase of the original bootstrap sample size, the quality of the bootstrapped estimates improved. For the case of regression analysis, the results raise some concern about the validity of the assumption underlying the bootstrap procedure. (Contains 27 references and 9 figures.) (Author)

..oes Bootstrap Procedure Provide Biased Estimates? An Empirical Examination for a Case of Multiple Regression

Xitao Fan

Utah State University

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 7, 1994, Session # 40.51.

Abstract

This paper empirically and systematically assessed the performance of bootstrap resampling procedure as it was applied to a regression model. Parameter estimates from Monte Carlo experiments (repeated sampling from population) and bootstrap experiments (repeated resampling from <u>one</u> original bootstrap sample) were generated and compared. Sample sizes of 20, 30, 50, and 100 were considered in the simulation. Ten independent Monte Carlo experiments and ten independent bootstrap experiments were conducted respectively for each sample size condition, with 1,000 samples (resamples for bootstrap) for each experiment. Estimates for standardized regression coefficients were obtained from each sample, and the mean estimates across samples were evaluated in relation to the population parameters. The results indicate that, as the number of resamples increases, the mean bootstrapped estimates did not show a clear tendency to converge on the population parameters. But, with the increase of the original bootstrap sample size, the quality of the bootstrapped estimates improved. For the case of regression analysis, the results raise some concern about the validity of the assumption underlying the bootstrap procedure.

Background

In educational and psychological research, the overreliance on statistical significance testing has been challenged on several grounds, including issues related to sample size and to the validity of theoretical assumptions underlying parametric statistical techniques (Carver, 1978; Shaver, 1993; Thompson, 1989). The sample size issue becomes prominent due to the fact that any null hypothesis can be rejected (statistically significant) when sample size is large enough, and the importance of statistical significance tends to be greatly exaggerated in research practice. As to the assumptions required for parametric statistical techniques, often, these assumptions are difficult, if not impossible, for research practitioners to meet or assess.

To avoid the blind reliance on statistical significance testing, some researchers have turned to methods which are more empirically grounded. Bootstrap procedure, which is computing-intensive in nature, has become prominent in recent years as a complement to the traditional statistical significance testing, or an alternative approach to making statistical decisions (Thompson, 1993). Instead of relying on the theoretical sampling distribution and sample sizes, bootstrap procedure, through repeated resampling with replacement from the original sample, empirically generates estimated sampling distribution, upon which our statistical decisions can be based (Diaconis & Efron, 1983; Efron, 1979; Lunneborg, 1990; Thompson, 1992). In this sense, bootstrap procedure attempts to avoid the pitfalls associated with the traditional statistical significance testing, such as the sample size issue, the concern for the validity of the theoretical assumptions for our data in hand, etc. Since its debut in the late 70's (Efron, 1979), bootstrap method has gradually attracted the attention of the researchers in the educational and psychological research arena (Lunneborg, 1983; Lunneborg, 1987). With increasingly easier access to powerful computing facilities, this method becomes more attractive than before.

Researchers in the educational and psychological research arena have applied the bootstrap technique to a variety of research problems, ranging from measurement issues, such as item discrimination and item bias indices, to multivariate statistical techniques, such as principal component analysis, factor analysis, and structural equation modeling (Bentler, 1992; Daniel, 1992; Harris & Kolen, 1988, 1989; Lambert, Wildt & Durand, 1990, 1991; Mendoza, Hart & Powell, 1991; Thompson, 1988; Thompson & Melancon, 1990). Some researchers also provided theoretical rationale or simulation results attesting to the applicability of this procedure to some widely used statistical methods (Bickel & Freedman, 1981; Freedman, 1981; Wu, 1986).

Though bootstrap procedure is promising as a complement or an alternative to the traditional statistical significance testing, it may have its own weaknesses which have not been adequately investigated empirically. One potential weakness was pointed out by Bollen and Stine (1993) in the area of structural equation modeling research, which indicated that conventional bootstrap resampling might fail to generate the intended empirical distributions for some sample statistics. Another example was from a study by Tryon (1984) which showed that bootstrapped estimates failed to converge on such basic population parameters as means and standard deviations.

One potential problem with bootstrap procedure may be related to the underlying assumption of bootstrap resampling itself. Bootstrap resampling solely relies on one original sample. In the argument of Diaconis and Efron (1983), the underlying assumption for bootstrap procedure was that, large sample or population results could be approached or reconstructed by repeated random resampling from a small sample. (It is in the sense of using this sample at hand and pulling oneself up by one's own bootstraps that the procedure acquired its name.) As empirical support for the validity of this assumption, Diaconis and Efron (1983) empirically demonstrated that the population correlation between Grade Point Average (GPA) and average Law School Admission Test (LSAT) for all 82 American law schools in 1973 could be approached by repeated random resampling from a bootstrap sample of 15 law schools. Lunneborg (1983) stated this assumption more clearly, "The bootstrap conjecture is that the sampling distribution of the statistic being studied and the sampling distribution found from this iterative process are essentially identical for a wide variety of statistics." (p. 1)

The underlying assumption that, through repeated resampling, enough information can be extracted from a small sample to reconstruct large sample or population results, may not have been subjected to rigorous empirical scrutiny. In sampling, even assuming that sound sampling techniques are used, a particular sample may have its own idiosyncracies due to sampling fluctuations. If random samples are repeatedly drawn from a population, these sampling idiosyncracies tend to be cancelled out. In bootstrap resampling, however, repeated random samples are drawn from one original sample, and potentially, we may be capitalizing on those sampling idiosyncracies associated with our particular sample. Consequently, the resultant picture of the empirical sampling distribution provided by bootstrap procedure may be a distorted one (Fan, 1993).

Monte Carlo Simulation and Bootstrap Resampling

In this study, a clear distinction is made between Monte Carlo simulation and bootstrap resampling. In Monte Carlo experiment, new random samples are repeatedly drawn from a population with known parameters. From each new sample, the statistic is obtained as the estimate of the population parameter. The performance of these statistics from all the random samples are then examined relative to the known population parameter. In bootstrap experiment, however, one original random sample is drawn from a population with known parameters. Random samples are then repeatedly drawn from this original bootstrap sample using the technique of sampling with replacement. From each resampling, the statistic for the population parameter is obtained. The performance of these statistics over all the resamples are then examined relative to the known population parameter.

Monte Carlo experiment represents the classical approach for estimating probability of certain event. The underlying assumption for this approach is obvious: sampling fluctuations tend to be cancelled out over repeated random sampling from the population, and as the number of samples increases, the mean statistic over the samples will converge on the population parameter. An intuitive example may help shed some light on this logic. Suppose we know that for a good and even coin, the likelihood of obtaining either the head or the tail from each flip is 0.5. In an empirical experiment to check one particular coin, we flip it 10 times. We may not obtain five heads and five tails exactly due to sampling fluctuations, and probably we will NOT be surprised to see eight heads and two tails. But, if we repeat our experiment 1000 times with 10 flippings in each experiment, we have reason to believe that the average number of heads and tails over 1000 repeated experiments will approach the population value of 5, instead 8 or 2 as in the first experiment. If the average numbers of heads and tails over 1000 experiments turn out to be 8 and 2 respectively, we may become suspicious about the quality of our coin, because over repeated sampling, a good and even coin is highly unlikely to give us estimate so far off from the theoretical population parameters.

If we have a coin which has been proven previously to be good and even, and we want to test a new coin-flipping device (sampling procedure). If, from our repeated sampling, the estimate for the probability of head or tail to occur converges on values of, say, 0.2 or 0.8, rather than on the known parameter of 0.5, then we have reason to call into question the integrity of our new flipping device itself.

The purpose of the present study is to examine empirically the performance of the bootstrap resampling procedure as applied to regression analysis. For this purpose, computer simulation is used to examine the characteristics of standardized regression coefficient estimates from both the Monte Carlo and the bootstrap experiments.

Methods

To assess the performance of bootstrap resampling, bootstrap experiments were conducted, and the estimates from the bootstrap experiments were compared with the known population parameters, and with those from Monte Carlo experiments. Since sample size may affect how well a statistic converges on its parameter, several sample size conditions were considered in the study.

A regression model of one dependent variable (Y) regressing on two independent variables (X1, X2) was used in the simulation. The population correlations among the three variables (Y, X1, X2) are specified as follows:

	Y	X1	X2
Y	1.00	0.40	0.60
X1	0.40	1.00	0.20
X2	0.60	0.20	1.00

Through the procedures proposed by Kaiser and Dickman (1962), samples were generated from the population with the intercorrelations as specified above.

Since the population intercorrelations are known for the variables in the regression model below, the standardized

$$Y = X\beta + \varepsilon$$

regression coefficients for the population (population parameters) are fully specified to be (Johnson & Wichern, 1988):

$$\begin{aligned}\beta_{(standardized)} &= \begin{bmatrix} Beta1 \\ Beta2 \end{bmatrix} \\ &= \rho_{x_1 x_2}^{-1}\ \rho_{xy} \\ &= \begin{bmatrix} 1.00 & 0.20 \\ 0.20 & 1.00 \end{bmatrix}^{-1} \begin{bmatrix} 0.40 \\ 0.60 \end{bmatrix} \\ &= \begin{bmatrix} 0.291 \\ 0.541 \end{bmatrix}\end{aligned}$$

Four sample size conditions of 20, 30, 50, and 100 were used in the simulation. In order to reduce the likelihood for haphazard chance discovery, ten independent Monte Carlo experiments and ten independent bootstrap experiments were conducted for each sample size condition. (Total number of Monte Carlo experiments: 4×10=40; total number of bootstrap experiments: 4×10=40.) For each experiment (Monte Carlo or bootstrap), 1,000 samples (for bootstrap experiment, 1,000 resamples) were drawn and sample standardized regression coefficients were obtained from each sample. Altogether, for each sample size condition, a total of 10,000 samples were generated for Monte Carlo and bootstrap procedures respectively. A schematic representation of the study for one sample size condition is presented in Figure 1.

Since the individual samples (for Monte Carlo experiments) and the original bootstrap samples (one for each bootstrap experiment) were randomly drawn from a population with known parameters (standardized regression coefficients), the estimates for the parameters obtained from these samples were expected to converge on the parameters as the number of replications increased. The failure in this respect was indication of problems with sampling procedures. In other words, if the sampling procedure works the way as it should, logic dictates that, the mean statistic based on ten samples will tend to be closer to the parameter than the statistic based on a single sample. In the same vein, the mean statistic over 100 samples will tend to be closer to the parameter than that over 10 samples, etc. As the number of samples increases, the mean statistic will tend to approach the population parameter, or converge on the parameter.

Bootstrap Experiments

For each of the four sample size conditions (n=20, 30, 50, 100), ten independent bootstrap experiments were conducted. For each experiment, first, one original bootstrap sample was drawn from the population with the known parameters (standardized regression coefficients). From this one bootstrap sample, 1,000 resamples were drawn by sampling with replacement. From each bootstrap resample, the sample standardized regression coefficients for the two independent variables (Beta1 for X1, and Beta2 for X2) were obtained. So for each independent bootstrap experiment, a total of 1,000 estimates were obtained for Beta1 (X1) and Beta2 (X2) respectively.

As reasoned above, unbiased estimate for population parameter should possess the property of convergence on population parameter over repeated sampling. Thus, as a general tendency, the mean statistic based on several samples is expected to be closer to the parameter than the statistic based on a single sample; and the mean statistic based on a large number of samples is expected to be closer to the parameter than that based on a small number of samples. To check this important property expected for unbiased estimate, for each bootstrap experiment, the following indices were obtained:

1) the parameter estimate from the first bootstrap resample;
2) the mean parameter estimate based on the first ten bootstrap resamples;
3) the mean parameter estimate based on the first 100

bootstrap resamples;

4) the mean parameter estimate based on the first 500 bootstrap resamples; and

5) the mean parameter estimate over all the 1,000 bootstrap resamples.

These indices were compared with the known population parameters to see how well these statistics would converge on the population parameters.

Monte Carlo Experiments

Monte Carlo experiments were conducted in the same fashion as the bootstrap experiments. The only difference, as explained previously, is in how the samples were drawn. For each independent Monte Carlo experiment, the same statistics were obtained, and the same five indices were calculated to check how well the statistics converged on the population parameters:

1) the parameter estimate from the first Monte Carlo sample;

2) the mean parameter estimate over the first ten Monte Carlo samples;

3) the mean parameter estimate over the first 100 Monte Carlo samples;

4) the mean parameter estimate over the first 500 Monte Carlo samples; and

5) the mean parameter estimate over all the 1,000 Monte Carlo samples.

Data Generation and Calculation of Statistics

All data generation, sampling (including sampling with replacement for bootstrap), and calculations were accomplished by using the Interactive Matrix Language (PROC IML) under the Statistical Analysis System (SAS). Random normal samples were generated by using the random number generator for normal distribution (RANNOR under SAS). Random sampling with replacement for bootstrap procedure was accomplished by generating a vector of random numbers using random number generator for uniform distribution (RANUNI under SAS). Each element of the vector was independently generated (to accomplish the feature of "with replacement" required by bootstrap), and constrained to be integers between *1* and *m*, with *m* being the original bootstrap sample size. This vector of integers was then used as the index numbers to draw row vectors (samples) from the original bootstrap sample data matrix of m×3 dimensions (m: original bootstrap sample size).

The sample intercorrelations among the three variables of Y, X1, and X2, as samples from a population with the specified intercorrelations as in Table 1, were accomplished through implementation of the procedures proposed by Kaiser and Dickman (1962). All calculations were accomplished using matrix language programming under PROC IML of SAS. For quality control purpose, before iterations began, results of every step of matrix language programming for calculations were compared, and found to be in agreement, with the results from regular SAS procedures such as

PROC REG, PROC MEANS, etc. The whole simulation process for both Monte Carlo and bootstrap experiments was accomplished by using SAS Version 6.08 for Microsoft Window on an IBM Computer with 486DX 66 MHz CPU with built-in math co-processor.

Results and Discussions

Due to the huge number of sample estimates which makes it difficult to tabulate,fficult to tabulate, a graphic approach is used to data. Figure 2 to Figure 5 depict the convergence tendency of the estimates for the four sample size conditions respectively. In these figures, to check the property of convergence, the mean statistics over consecutively larger number of samples were plotted in relation to the parameter values:

1) the regression coefficient from the first sample (resample for bootstrap);
2) the mean regression coefficient of the first ten samples;
3) the mean regression coefficient of the first 100 samples;
4) the mean regression coefficient of the first 500 samples;
5) the mean regression coefficient of all the 1,000 samples.

The figures show that the estimates from Monte Carlo experiments (represented by "o" in the graphs) consistently converge on the population parameters as the number of samples increased, and this tendency is clear for both Beta1 and Beta2, and for all sample size conditions. The estimates from bootstrap experiments (represented by "*" in the graphs), however, do not fare as well, since they do not seem to continue to approach parameter values as the number of resamples increases. Especially, when the original bootstrap sample size is small (e.g., n=20), the divergence of bootstrap statistics from the population parameters seems to be very substantial.

Although bootstrap estimates do not seem to exhibit the tendency to converge on population parameters, a closer examination of the figures reveals that, in terms of degree of divergence of the statistics from the population parameters, bootstrap estimates show clear improvement as the original bootstrap sample size increases. This is obvious when we compare Figure 2 (sample size: 20) with Figure 5 (sample size: 100). The improvement pattern is the same for both Beta1 and Beta2. This indicates that, for the regression model, larger sample size for the original bootstrap sample may be necessary in order to avoid potential excessive divergence of bootstrap estimates from the population parameter values. This may also partially explain the non-convergence patterns as exhibited by bootstrap estimates for means and standard deviations as reported in Tryon's study (Tryon, 1983), since only small sample sizes (15 and 25) were used in the simulation in that study. It is possible that if larger sample sizes had been considered in that study, the quality of bootstrap estimates would have improved.

The fact that the bootstrap estimates exhibit little tendency to converge on population parameter value has some interesting implications. As explained by Diaconis and Efron

(1983), bootstrap procedure assumes that population results could be approached or reconstructed through repeated sampling from the small sample at hand. This assumption may have been accepted prematurely without rigorous empirical scrutiny.

Conclusions

Based on the results from these systematic simulation experiments, three tentative conclusions may be drawn with regard to the bootstrap resampling procedure. First, the assumption for bootstrap procedure, i.e., large sample or population results may be approached or reconstructed through intensive resampling of a small sample at hand, may not be capable of withstanding rigorous empirical test, and the validity of this assumption may be called into question. Although Diaconis and Efron (1983) conceded that the procedure might not work for a few samples, and one could not know in advance which they were, the results from this study indicate that their view about the applicability of the procedure still might have been overly sanguine. Especially considering the generality of regression analysis as general linear model which subsumes a variety of parametric tests, the somewhat disappointing performance of the procedure in this study may become a source for concern.

Second, the bootstrap estimates for regression coefficients are biased to a certain extent in the sense that they do not continue to approach population parameters with the increase of the number of resamples. In other words, in many cases, these estimates may approach some value other than the population parameter, and it is not known what the value is. The divergence pattern of the bootstrap estimates is especially striking when the original bootstrap sample size was small.

Last, there is clear indication that the quality of bootstrap estimates improves substantially with the increase of the original bootstrap sample size, since when sample size gets larger, the estimates tend to diverge much less from the population parameters. If this can be replicated for other types of statistical analysis, it may imply that it would be advantageous or even necessary to use larger bootstrap sample size so as to reduce the potential divergence of bootstrap estimates from the population parameters.

References

Bentler, P. M. (1992). EQS Structural Equations Program Manual. Los Angeles, CA: BMDP Statistical Software.

Bickel, P. & Freedman, D. (1981). Some asymptotic theory for the bootstrap. Annals of Statistics, 9, 1196-1217.

Bollen, K.A., & Stine, R. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K.A. Bollen & J.S. Long (Eds.), Testing structural equation models. Newbury Park, California: Sage Publications, Inc.

Carver, R.P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Daniel, L. G. (1992). Bootstrap methods in the principal component case. Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992). ERIC #: ED 346135.

Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, May, 116-130.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. the Annals of Statistics, 7, 1-26.

Fan, X. (1993). Statistical significance testing and alternatives: A critique. Unpublished manuscript.

Freedman, D. A. (1981). Bootstrapping regression models. Annals of Statistics, 9, 1218-1228.

Harris, D. J. & Kolen, M. J. (1988). Bootstrap and traditional standard errors of the point-biserial. Educational and Psychological Measurement, 48, 43-51.

Harris, D. J. & Kolen, M. J. (1989). Examining the stability of Angoff's delta item bias statistic using bootstrap. Educational and Psychological Measurement, 49, 81-87.

Johnson, R. A. & Wichern, D. W. (1988). Applied multivariate statistical analysis (2nd Ed.). Englewood Cliffs, New Jersey: Prentice Hall.

Kaiser, H.F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. Psychometrika, 27, 179-182.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1990). Assessing sampling variation relative to number-of-factor criteria.

Educational and Psychological Measurement, 50, 33-48.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence interval for factor loadings. Multivariate Behavioral Research, 26, 421-434.

Lunneborg, C. E. (1983). Efron's bootstrap with some applications in psychology. Paper presented at the Annual Meeting of the American Psychological Association, Anaheim, California, August 26-30. ERIC#: ED 241554.

Lunneborg, C.E. (1987). Bootstrap applications for the behavioral sciences. Seattle: University of Washington.

Lunneborg, C.E. (1990). [Review of computer intensive methods for testing hypotheses]. Educational and Psychological Measurement, 50, 441-445.

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. Multivariate Behavioral Research, 26, 255-269.

Shaver, J.P. (1993). What significance testing is, and what it isn't. The Journal of Experimental Education, ??

Thompson, B. (1988). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. Educational and Psychological Measurement, 48, 681-686.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B. (1992). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC No. ED 344895)

Thompson, B. (1993). The use of statistical significance tests i research: Bootstrap and other alternatives. The Journal of Experimental Education, ??

Thompson, B. & Melancon, J.G. (1990). Bootstrap versus statistical effect corrections: A comparison with data from the finding embedded figure test. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, LA. ERIC #: ED 325509.

Tryon, W. W. (1984). An evaluation of the bootstrap hypothesis using computer simulation. Paper presented at the Annual Meeting of the American Psychological Associ. ion (Toronto, Canada, August 24-28, 1984). ERIC #: ED 260093.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. The Annals of Statistics, 14, 1261-1294.

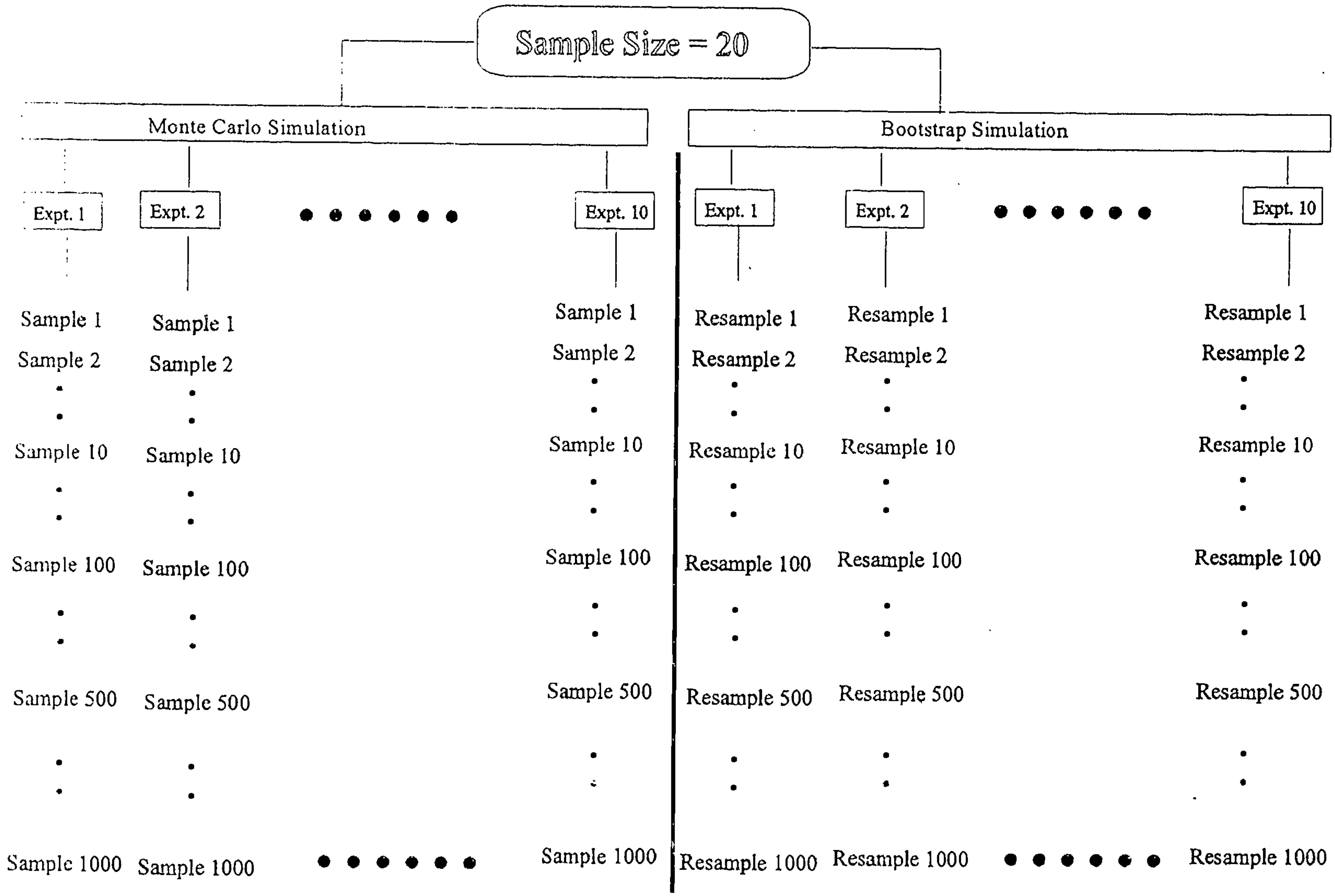

Figure 1: Schematic Representation of the Monte Carlo and Bootstrap Experiments for One Sample Size Condition

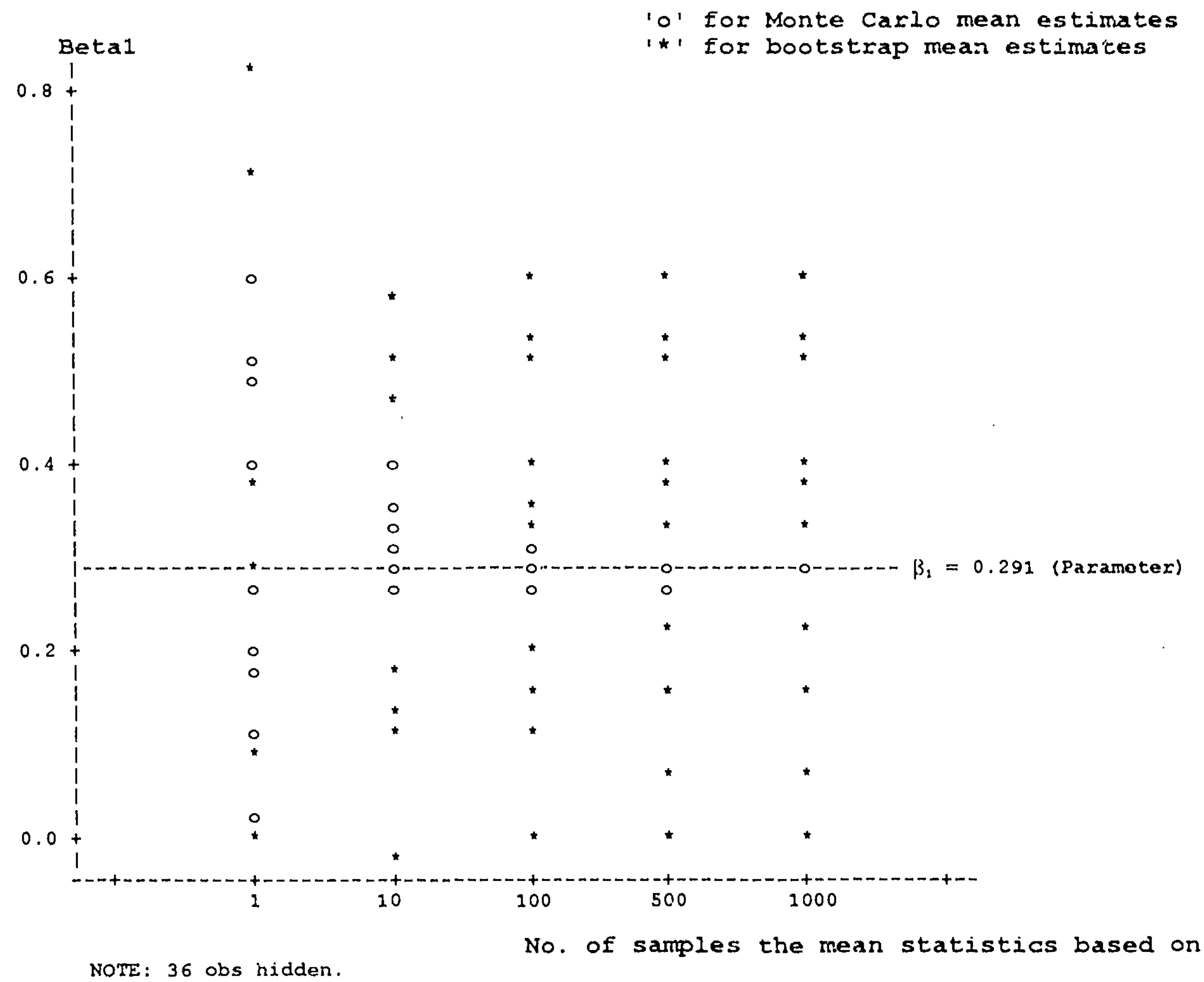

Figure 2: Mean Statistics and Parameter: Convergence Tendency for Beta1 (Sample Size 20)

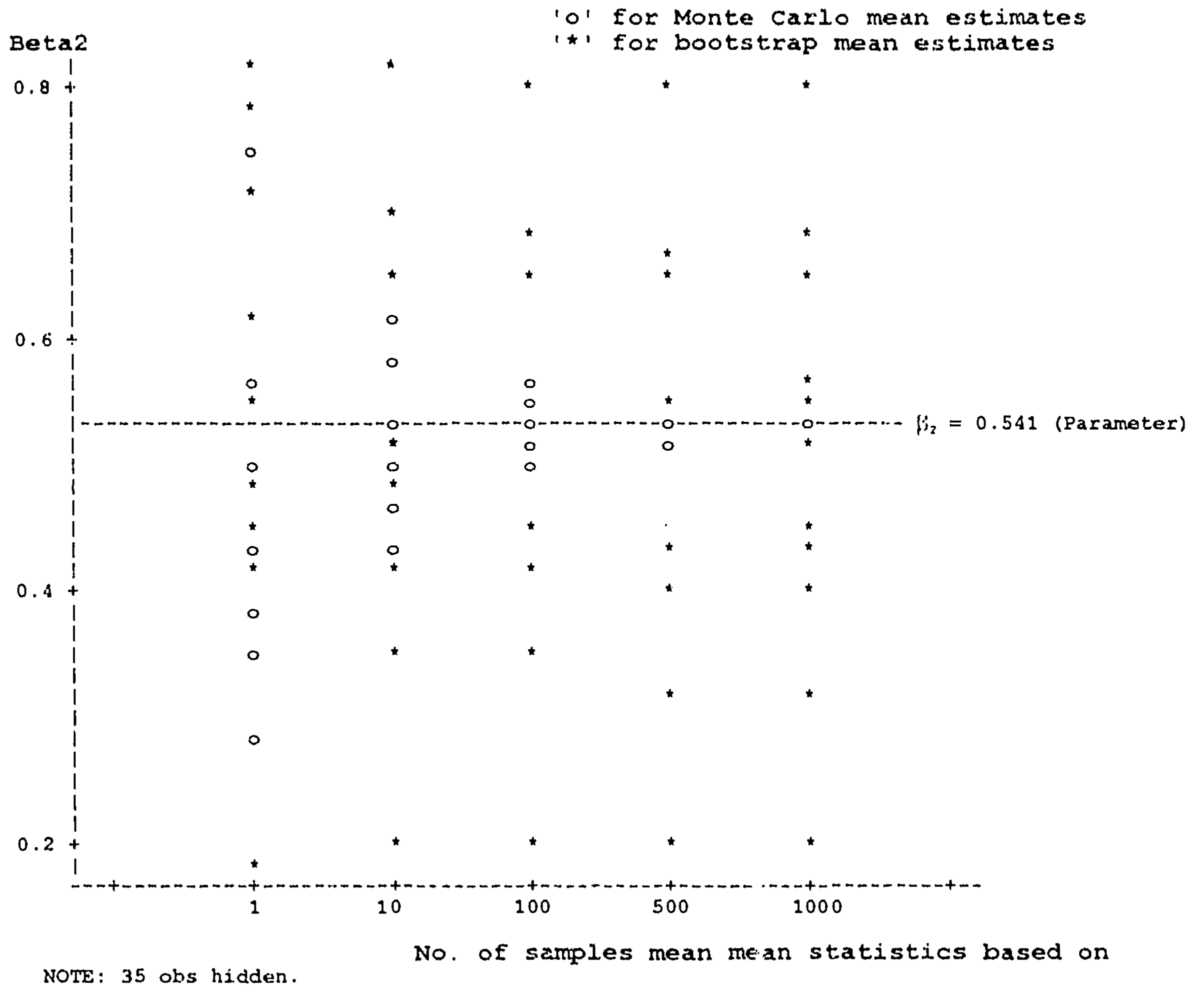

Figure 3: Mean Statistics and Parameter: Convergence Tendency for Beta2 (Sample Size 20)

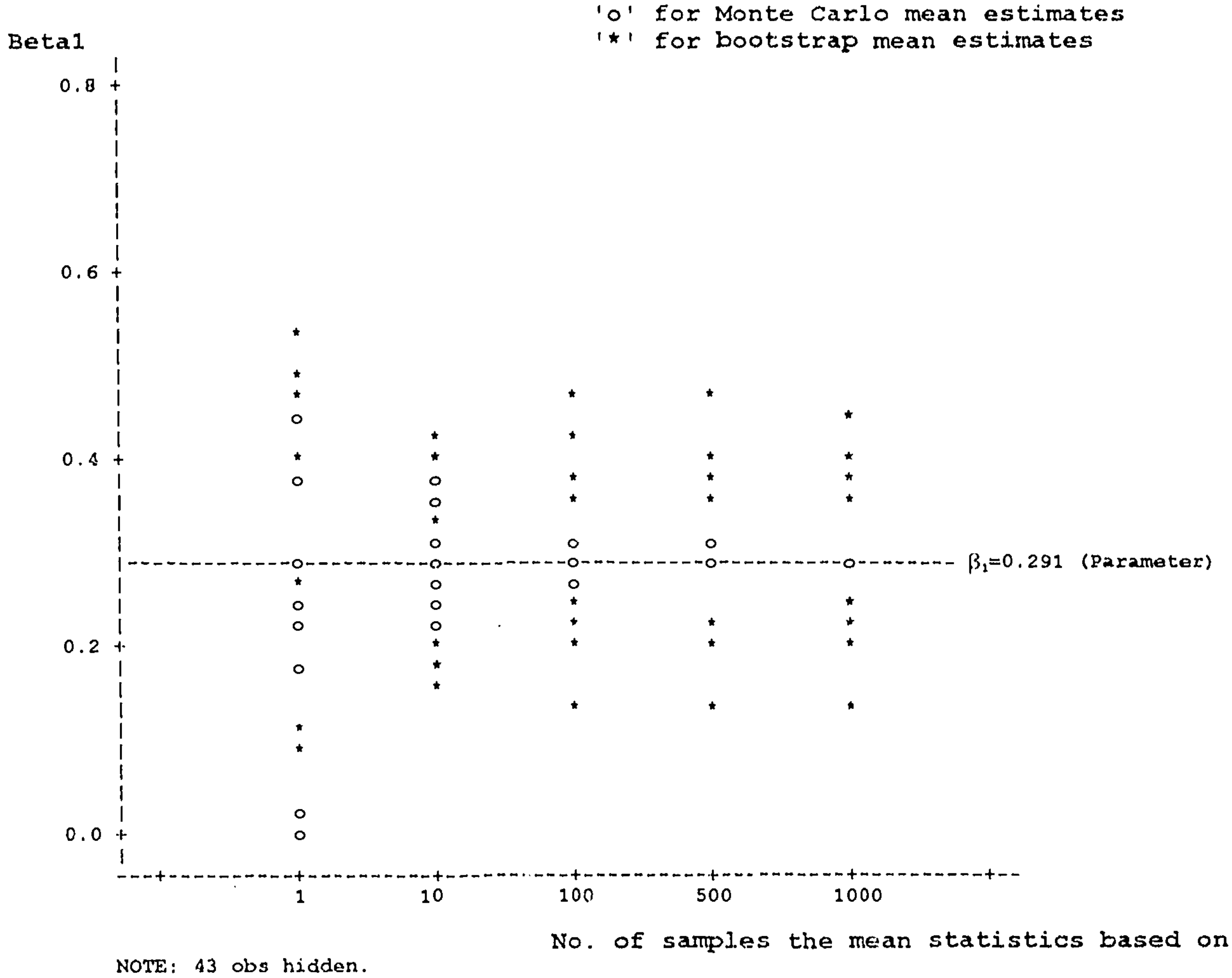

Figure 4: Mean Statistics and Parameter: Convergence Tendency for Beta1 (Sample Size 30)

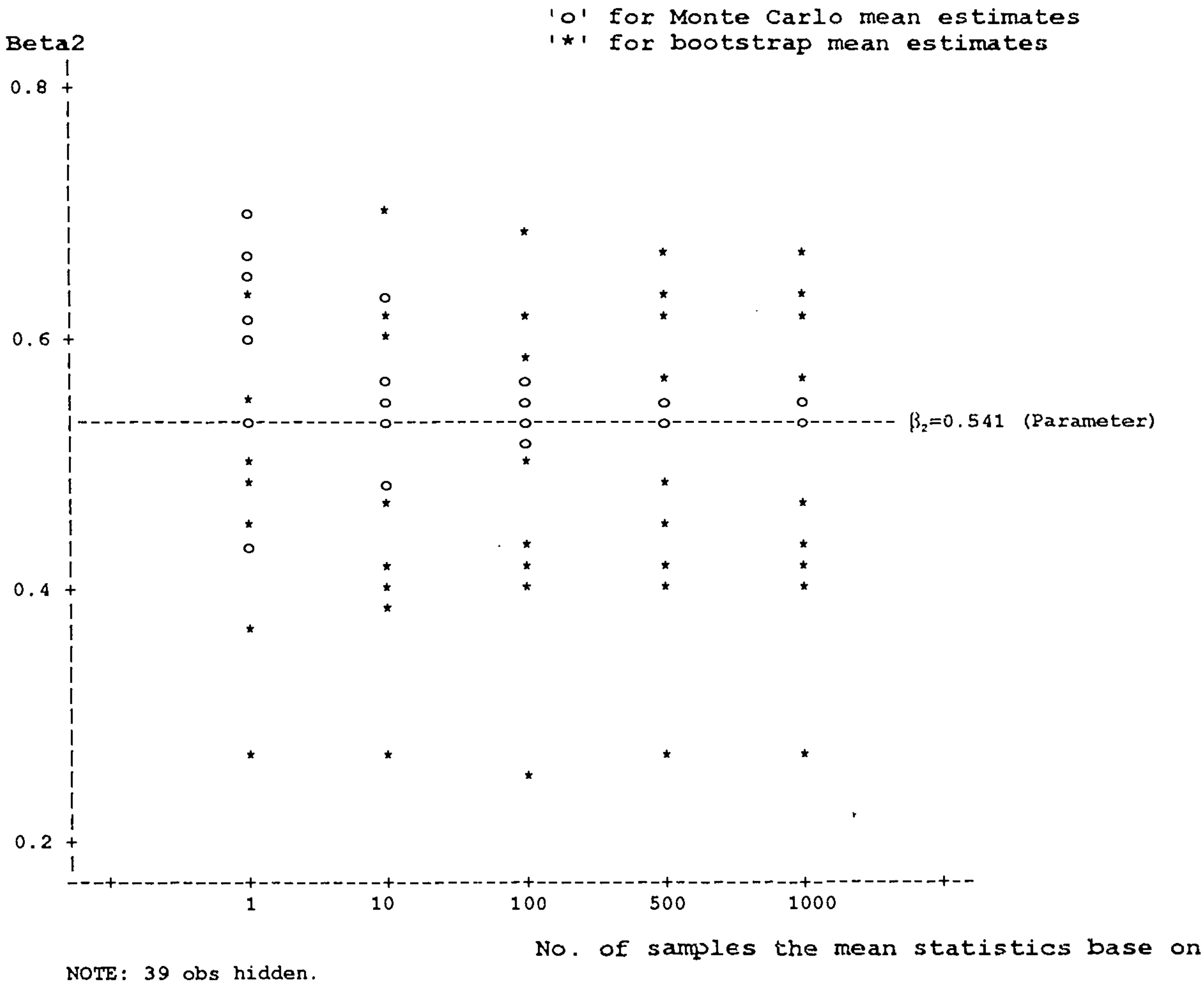

Figure 5: Mean Statistics and Parameter: Convergence Tendency for Beta2 (Sample Size 30)

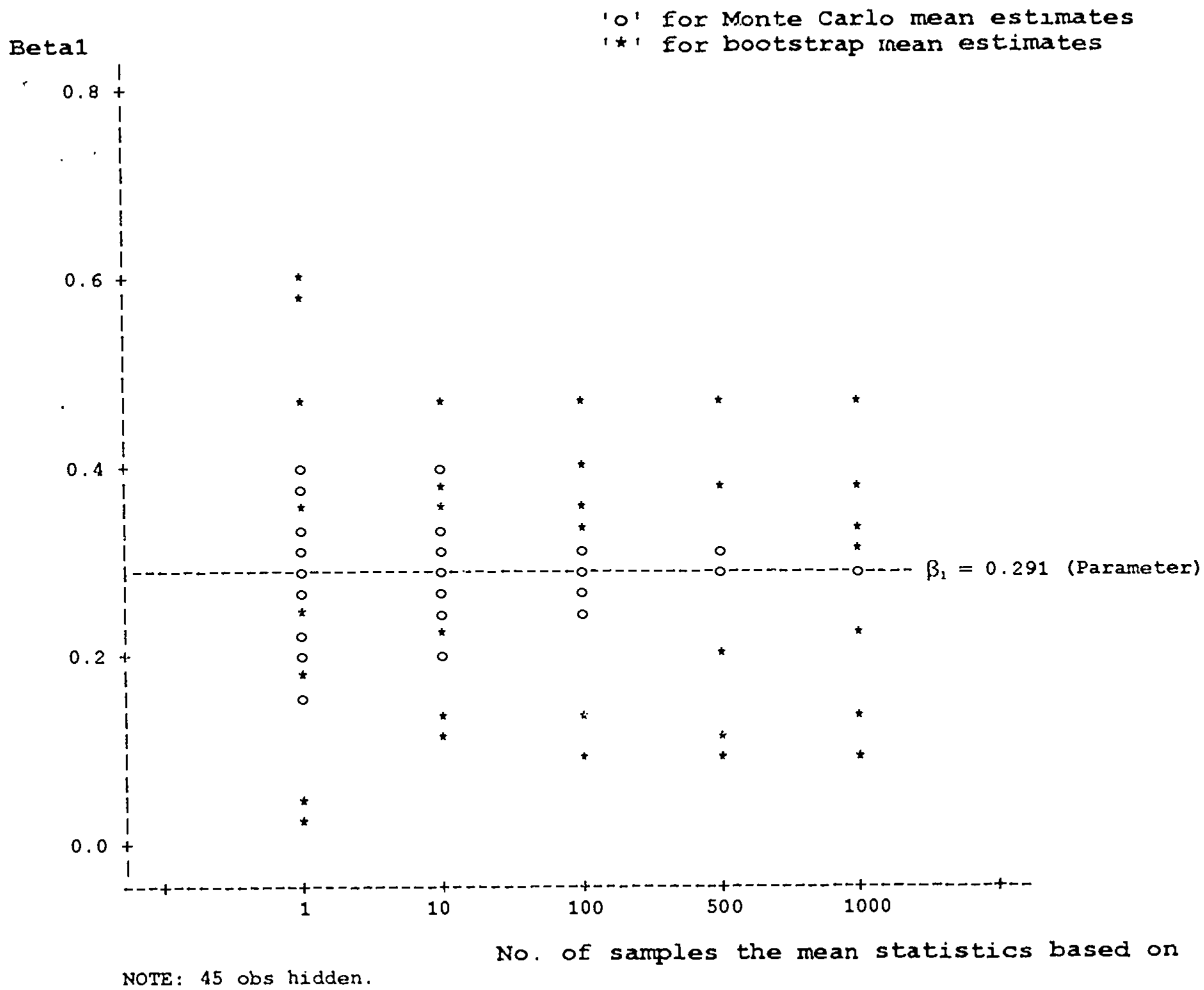

Figure 6: Mean Statistics and Parameter: Convergence Tendency for Beta1 (Sample Size 50)

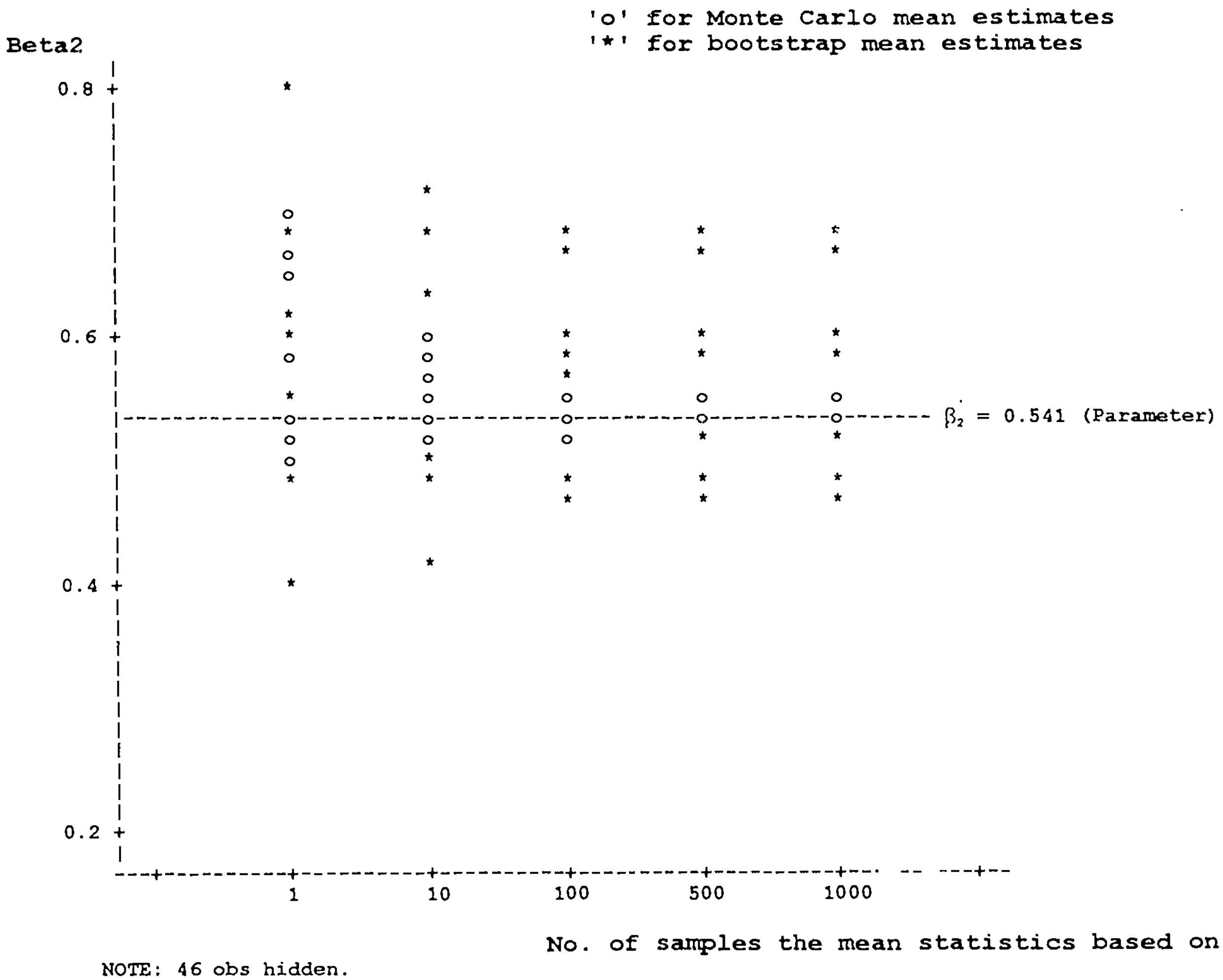

Figure 7: Mean Statistics and Parameter: Convergence Tendency for Beta2 (Sample Size 50)

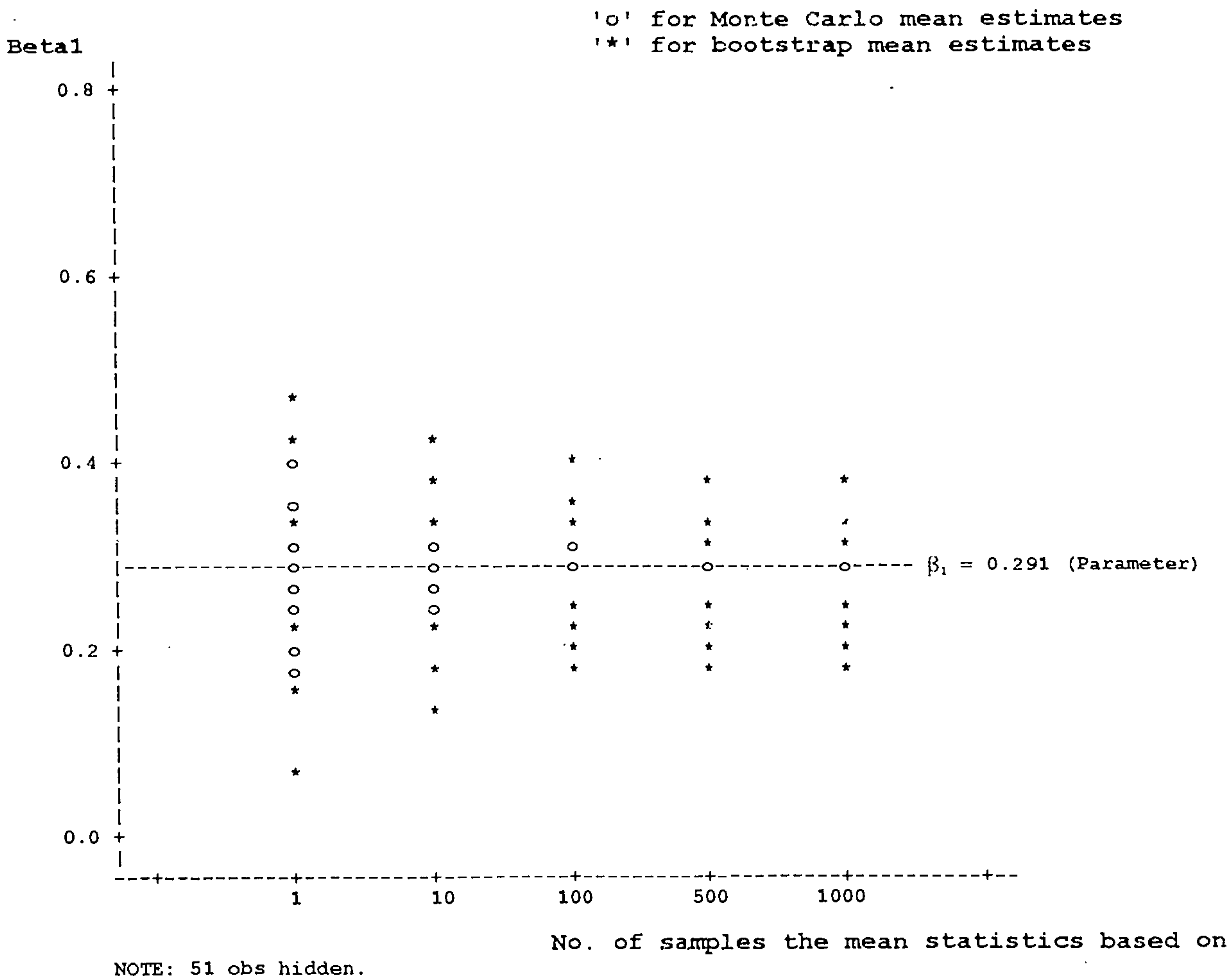

Figure 8: Mean Statistics and Parameter: Convergence Tendency for Beta1 (Sample Size 100)

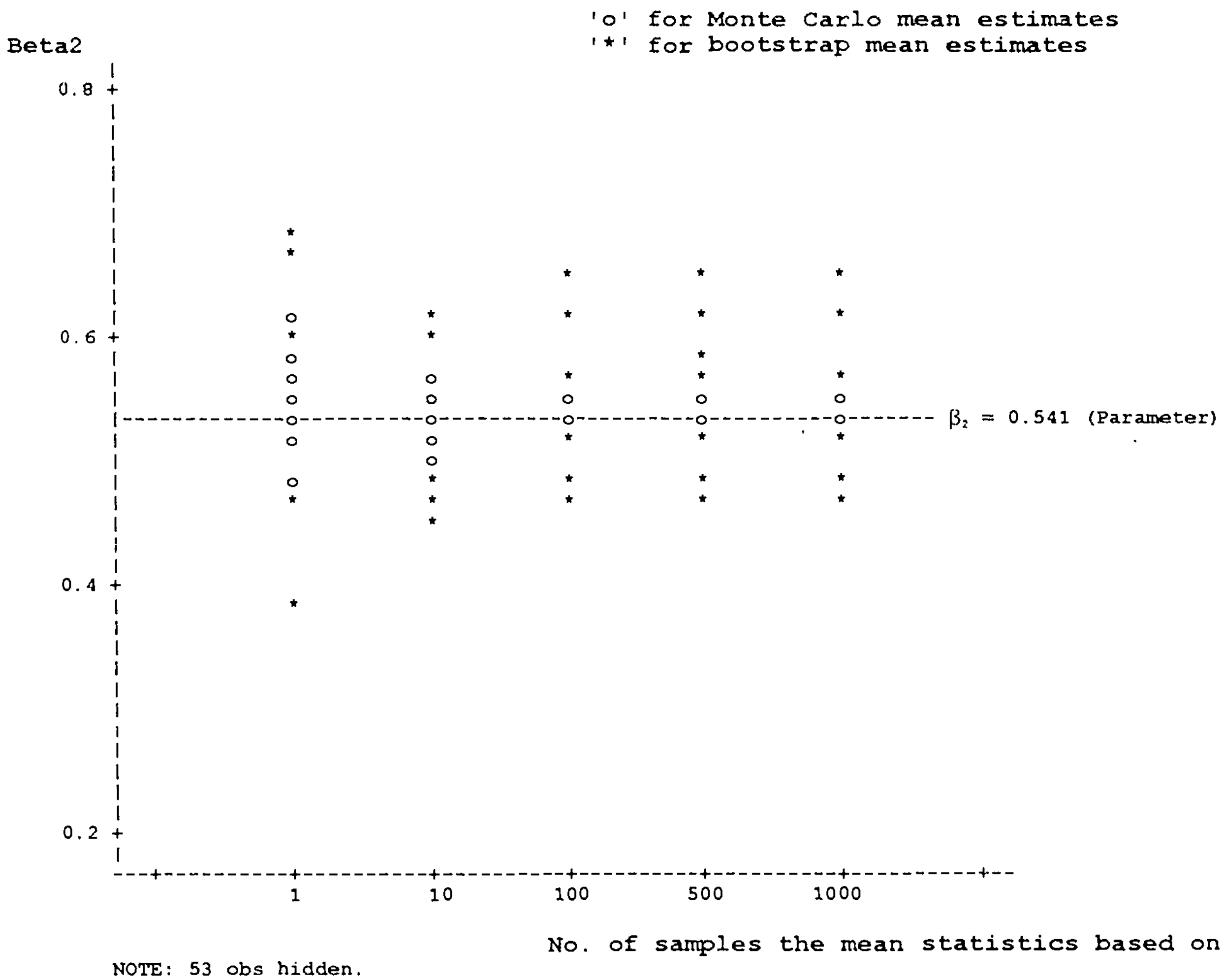

Figure 9: Mean Statistics and Parameter: Convergence Tendency for Beta2 (Sample Size 100)