THE BOOTSTRAP AND QUANTITATIVE STRUCTURE ACTIVITY RELATION MULTIPLE LINEAR REGRESSION MODELS FOR NONPARAMETRIC ESTIMATES OF STANDARD ERROR

by

Efraim Halfon

Aquatic Physics and Systems Division
National Water Research Institute
Canada Centre for Inland Waters
Burlington, Ontario, Canada L7R 4A6

November 1984
NWRI Contribution . 85-21

ABSTRACT

A recently invented statistical method, the bootstrap, is used to verify whether a multiple regression model, developed from a limited data set, produces reliable predictions of toxicity if all possible data would have been available to develop the model. This statistical method allows generalization to chemicals of the same class not included in the original analysis. An example is taken from the literature and the errors associated with the coefficients of a multiple regression are computed using the bootstrap. Also, the multiple regression model is used to estimate the toxicity of a contaminant not used in model development and estimates with their relative probability are graphically displayed.

RÉSUMÉ

Une méthode statisitque de facture récente, le bootstrapping, est utilisée afin de vérifier si un modèle de régression multiple élaboré à partir d'un ensemble de données limité permet de prédire les niveaux de toxicité comme si l'on avait disposé d'un ensemble de données complet pour créer le modèle. Cette méthode statistique permet d'appliquer les résultats obtenus à des produits chimiques appartenant à la même catégorie sans qu'ils aient fait partie de l'analyse initiale. On présente un exemple tiré de la documentation. Les erreurs liées aux coefficients d'une régression multiple sont calculées à l'aide de cette méthode. On utilise également le modèle de régression multiple pour estimer la toxicité d'un contaminant qui n'a pas été inclus dans le modèle initial. Les probabilités relatives des niveaux estimatifs ainsi obtenus sont présentées sous forme de graphiques.

EXECUTIVE SUMMARY

Laboratory tests for determining chemical and environmental properties of toxic contaminants are time consuming and expensive. Over the years much research has been performed to predict the relation between certain easily measurable properties and others more difficult to obtain. Statistical relations, such as multiple linear regression models, have been developed for this purpose. The problem is, however, that no test has been developed to assure that the predictions of these statistical models are valid also for new chemicals. The bootstrap is a new statistical procedure that allows generalization of the results to chemicals not used in the development of the original mode; the bootstrap is used to verify whether a hypothesis developed from a limited data set would be valid, if all possible data would have been available. Thus, the method is very useful to reduce the amount of data to be collected from laboratory experiments to evaluate the toxicity and environmental hazard of toxic contaminants. When the statistical models are used for prediction of new chemical properties, the bootstrap allows an estimate of the probability and range of the chemical property, such as toxicity, bioconcentration, etc.

RÉSUMÉ ADMINISTRATIF

Les essais en laboratoire visant à déterminer les propriétés chimiques et environnementales des contaminants toxiques nécessitent beaucoup de temps et d'argent. Depuis plusieurs années, des travaux sont en cours afin de prédire les rapports entre certaines propriétés qui peuvent être mesurées sans difficulté et d'autres qui, au contraire, se mesurent difficilement. Les outils statistiques tels que la régression multiple ont été mis au point pour mettre de tels liens en évidence. Il n'existe toutefois aucune technique de vérification qui permette de transposer de façon valide les prédictions de modèles statistiques à de nouveaux produits chimiques. Le bootstrapping est une nouvelle méthode statistique qui permet d'appliquer les résultats obtenus à des produits chimiques n'étant pas compris dans le modèle initial. Le bootstrapping sert à vérifier si les hypothèses s'appuyant sur un ensemble de données restreint sont aussi valides que si elles avaient été formulées à la lumière de données absolument complètes. Cette méthode permet donc d'évaluer le degré de toxicité et les risques que posent pour le milieu les contaminants toxiques, tout en limitant la quantité de données à recueillir en laboratoire à cette fin. Lorsque les modèles statistiques s'appliquent à la prédiction de propriétés chimiques nouvelles, le bootstrapping permet d'évaluer la probabilité et l'étendue des propriétés telles que la toxicité, la bio-accumulation, etc.

INTRODUCTION

Lack of data for the recognition of environmental pollutants was pointed out as a problem by Kaiser et al. (10) who stated that "as a general rule, toxicological data on specialty chemicals are non-existent and even for some compounds produced in large quantities, little information is available on their sublethal effects on aquatic and terrestrial species, including man". In their analysis on the procedures to screen out hazardous compounds from those which are safe, Kaiser et al. also stated that "given the large number of chemicals and formulations involved, and the enormous resources and time necessary to actually test each, it is evident that routes must be found to predict at least the type and magnitude of hazard associated with their use and release to the environment". With the large number of chemicals that may pose an environmental hazard, the derivation of the prediction equations and the extrapolation of such predictions to chemical compounds of the same class as those used to derive the model is of primary importance.

In 1977, Efron (4-6) invented a new statistical test, the bootstrap, which generalizes the jackknife (15) and uses information from a given data set to estimate a statistic, for example correlation with its confidence limits, if all possible data from a population, or distribution F, would have been available. The answer to the question "what would we see if we had much more data?" was until recently very hypothetical because no accepted statistical procedure was available.

THE DATA

Kaiser et al. (10) provide measures of physico-chemical properties of 20 chlorophenol isomers including log P, or the logarithm of the octanol-water partition coefficient, pK, and melting point (mp). Toxicity of the chlorophenol isomers was measured by the Microtox 30EC50 test (1,2,13,14), which quantifies the effects of toxicants by the decrease in light emission by photoluminescent bacteria.

THE LINEAR MODEL

In quantitative structure activity relation (QSAR) tests, multiple regression is sometimes used to predict the toxicity of a chemical compound based on information on the chemical structure and on several physical-chemical properties. For the above mentioned data the multiple regression model is

$$y = -3.347 + .907 \log P + .235 pK + .003 mp, \; n = 20 \qquad (1)$$

where y is the measured Microtox 30EC50; Halfon (8) pointed out that both the measured toxicity (y) and the independent structural variables (x's) are usually measured or estimated with some error; for the linear regression case he suggested that the geometric mean (GM) functional regression method should be used to estimate the

coefficients and also suggested (7) the use of the bootstrap for computing error estimates. The same reasoning on data uncertainty is also valid for multiple regression models. In fact, even if some parameters may be well known, for example the melting point and pK, others such as solubility and the octanol water partition coefficient (P or K_{ow}) are often measured with errors since the experimental set up is very important. No straightforward theoretical method exists to compute the errors of the multiple linear regression coefficients comparable to the geometric mean regression method available for the simple linear regression case, thus, a numerical method, such as the bootstrap, must be used.

THE BOOTSTRAP

The bootstrap can be used to seek generalizations by estimating the unknown distribution F from data; the frequency distribution F does not have to be assumed normal, which is very useful since chemicals with different structures and chemical properties are used in this model. The bootstrap, as mentioned before, assumes that the unknown distribution F can be estimated from the observed distribution F, i.e., we can infer from the observed data the validity of the hypothesis for all other chemicals with similar properties without having to perform more experiments; the generality of the hypothesis can be inferred from the standard errors associated with the correlation and with the coefficients of the linear model.

To perform the bootstrap test, each of the 20 data points is replicated a very large number of times, e.g., one billion times, and then this large amount of data is sampled 100 to 1000 times, the bootstrap samples. From a practical point of view the data are not really replicated a billion times, but a random number generator is used. The statistics of interest, in this case the standard errors and the confidence limits of the slopes, of the intercept and of the correlation coefficient are computed for each such bootstrap sample. Since the assumption of normality has been abandoned, the confidence limits may not be symmetrical around the mean, if the probability density function is skewed.

RESULTS

In Table 1 the coefficients of the multiple regression model (Eq. 1 above) are computed with the assumption of errors in the independent variables (the structural properties of the compound) and in the independent variable (the results of the toxicity tests). The bootstrap estimates are similar to the standard estimates with the additional information that some coefficients are very uncertain.

The Bootstrap for Toxicity Prediction

The bootstrap might be considered useful only to compute realistic estimates of the coefficients with their relative errors;

this is not so, the bootstrap allows the computation of uncertainty, associated with toxicity prediction and therefore a more precise estimate of hazard, when multiple regression is routinely used to estimate the toxicity associated with new chemicals, given the enormous resources and time necessry to actually test each. The method of Halfon (9) can then be used to rank each chemical according to its toxicity and to compare it with other chemicals also present in the environment.

Two procedures, one step and two step, can be used to estimate prediction errors: both take only a few seconds of computer time and therefore both should be performed and results compared. Figure 1 shows the formal steps in the two procedures.

In the one step procedure, 400 bootstrap samples of the data are taken by the computer and the coefficients of the multiple regression are estimated assuming no error in the independent variables; thus the error in the predictions is computed assuming a correct multiple regression model but using uncertain data. Table 2 shows the results of this analysis including the standard deviation of the estimated toxicity, its coefficient of variation and the range of the 400 values; the ranges are very large in relation to the standard deviation. The standard deviation statistic assumes a normal unskewed frequency distribution in the data and, since the bootstrap does not assume normality, the results of the analysis are better displayed as histograms: Figure 2 shows the prediction of data points as with its confidence limits.

The second procedure is a two step process, first a linear regression is fitted between a structural property and the toxicity, the toxicity now being the independent variable. The GM linear regression method (8) is used to compute the coefficients since we know that both data sets are measured with uncertainty, especially the toxicities. This process is repeated on 400 bootstrap samples to estimate the uncertainty in the structural variables. These 400 samples are used as data in multiple linear regression models to produce estimates of toxicity: Since uncertain data were used, the models produce data toxicity with uncertainty. Table 3 shows the results for this procedure. Figure 3 presents the same results for the point #5.

Comparisons of Tables 2 and 3 show that the coefficient of variation of the estimated toxicity are usually similar for the two procedures but sometimes the prediction error estimates are different, for example points #1 and #2.

DISCUSSION

The bootstrap is a computer intensive statistical method that uses Monte Carlo simulations to provide information when theoretical analytical solutions are not possible, for example if the original frequency distribution is not normal. The present analysis was performed on a CDC Cyber 171 computer and it took 27 CPU seconds for 400 replications or bootstrap samples; Efron (4,5) suggests 128 to

512 replications since the method converges asymptotically. The method is numerically simple enough to be programmed on a microcomputer such as an Atari 400 or a Commodore 64. The application to ecotoxicological problems is intriguing since very often correlation and regression models are published in the literature based on few data and the reliability of the results is usually difficult to establish given the diversity of chemicals. The bootstrap is an interesting method that should be often used to establish the uncertainty of the proposed hypothesis. A FORTRAN program with a test run is available upon request.

REFERENCES

(1) Beckman Instruments Inc. "Microtox system: operational manual". (1982).

(2) Bulich, A.A. In "Aquatic Toxicology". 2nd Conference ASTM STP 667, 98-106 (1969).

(3) Diaconis, P. and B. Efron. Scientific American, 248: 116-130 (1983).

(4) Efron, B. "The jackknife, the bootstrap and other resampling methods". SIAM, 92 pp. (1982).

(5) Efron, B. Can. J. of Statistics, 9: 139-172 (1981).

(6) Efron, B. Biometrika, 68: 589-599 (1981).

(7) Halfon, E. "The bootstrap and the jackknife in ecotoxicology or nonparametric estimates of standard error". Chemosphere (submitted).

(8) Halfon, E. "The regression method in ecotoxicology: a better formulation using the geometric mean functional regression". Environmental Science and Technology (submitted).

(9) Halfon, E. "On ranking chemicals for environmental hazard". NWRI unpublished report.

(10) Kaiser, K.L., D.G. Dixon and P.V. Hodson. In K.E. Kaiser (Ed.) "QSAR in Environmental Toxicology". D.E. Reidel Publishing Co., Dordrecht and Boston, 189-206 (1984).

(11) Kolata, G. Science, 225: 156-158 (1984).

(12) Quenouille, M. J. Royal Statist. Soc. Ser. B., 11: 18-84 (1949).

(13) Ribo, J.M. and K.L.E. Kaiser. Chemosphere 12: 1421-1442 (1983).

(14) Serat, W.F., F.E. Budinger and P.K. Mueller. J. Bact., 90: 832-833 (1965).

(15) Teisser, G. Biometrics, 4: 14-48 (1948).

(16) Tukey, J. Abstract, Ann. Math. Statist., 29: 614 (1958).

TABLE 1. Multiple regression coefficients with the assumption of no errors and errors in the independent variables (bootstrap). Data from (10).

n = 20 Coefficient		Standard Estimate	Bootstrap Estimate	(S.D.)
a_1	log P	0.907	0.897	.200
a_2	pK	0.235	0.248	.129
a_3	mp	0.003	0.004	.005
Intercept		-3.347	-3.474	1.490
Correlation		0.915	0.921	.039

TABLE 2. Measured and estimated 30EC50 Microtox toxicity using the one step bootstrap method

Measured	Bootstrap Estimate	Coefficient of Variation	Range	
.42	0.43	67.79	-1.14	1.18
.58	0.53	51.38	-1.27	1.31
.96	1.08	8.74	0.64	1.53
1.09	0.97	30.19	-0.52	2.05
1.19	1.02	10.60	0.67	1.44
1.19	1.94	6.31	1.44	2.48
1.24	1.48	4.33	1.25	1.83
1.41	1.65	9.53	1.08	2.05
1.47	1.39	7.04	0.86	1.78
1.52	1.50	4.31	1.27	1.66
1.77	1.99	7.42	1.35	2.76
2.00	1.89	9.20	1.45	3.14
2.02	2.81	5.67	2.30	3.23
2.19	1.83	1.37	1.62	2.03
2.20	2.07	5.78	1.76	2.35
2.25	2.11	10.11	1.42	2.74
2.26	1.59	12.29	0.41	2.31
2.71	3.16	20.98	0.58	5.42
2.74	2.39	6.54	1.91	2.82
3.12	2.63	7.05	2.18	3.22

TABLE 3. Measured and estimated 30EC50 Microtox toxicity using the two step bootstrap method

Measured	Bootstrap Estimate	Coefficient of Variation	Range	
.42	0.59	30.42	0.10	1.06
.58	0.73	22.24	0.28	1.16
.96	1.06	11.93	0.72	1.45
1.09	1.18	9.97	0.87	1.55
1.19	1.26	8.81	0.96	1.62
1.19	1.26	8.82	0.96	1.62
1.24	1.31	8.33	1.00	1.66
1.41	1.45	7.13	1.55	1.79
1.47	1.51	6.85	1.21	1.84
1.52	1.55	6.66	1.25	1.87
1.77	1.76	6.22	1.47	2.08
2.00	1.97	6.33	1.62	2.34
2.02	1.98	6.36	1.63	2.34
2.19	2.13	6.61	1.71	2.57
2.20	2.14	6.63	1.71	2.58
2.25	1.18	6.72	1.74	2.65
2.26	2.19	6.74	1.74	2.66
2.71	2.58	7.63	1.94	3.21
2.74	2.61	7.69	1.96	3.24
3.12	2.94	8.41	2.12	3.70

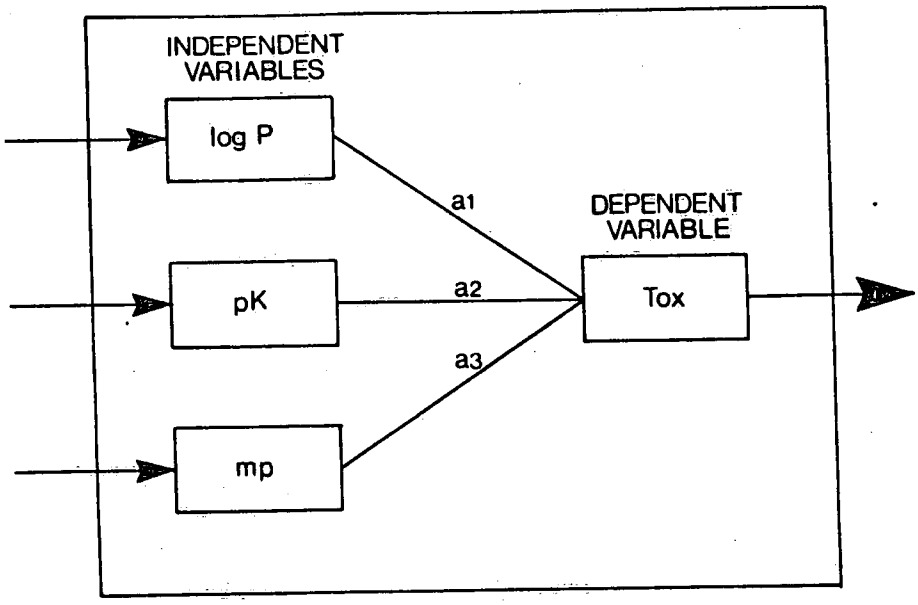
FIGURE LEGENDS

Figure 1 This diagram shows the two procedures that can be followed to estimate the prediction error for a new chemical. In the first procedure 400 bootstrap samples are taken of the data, in this case 19 points, the one that has to be predicted blindly is not included, and 400 equations are derived. These are used to predict the average and prediction error of the unknown chemical. In the second procedure, as a first step, the measured toxicity is considered an independent variable to predict the structural properties; the GM linear method is used, since both independent and dependent variables contain errors; the estimated structural properties are then considered independent variables to estimate the range of toxicity of the new chemical.

Figure 2 Measured and estimated Microtox toxicity values of point no. 5, by using the one step procedure. One standard deviation and 95% confidence limits are shown. Note that the frequency distribution of the 400 bootstrap estimates is skewed.

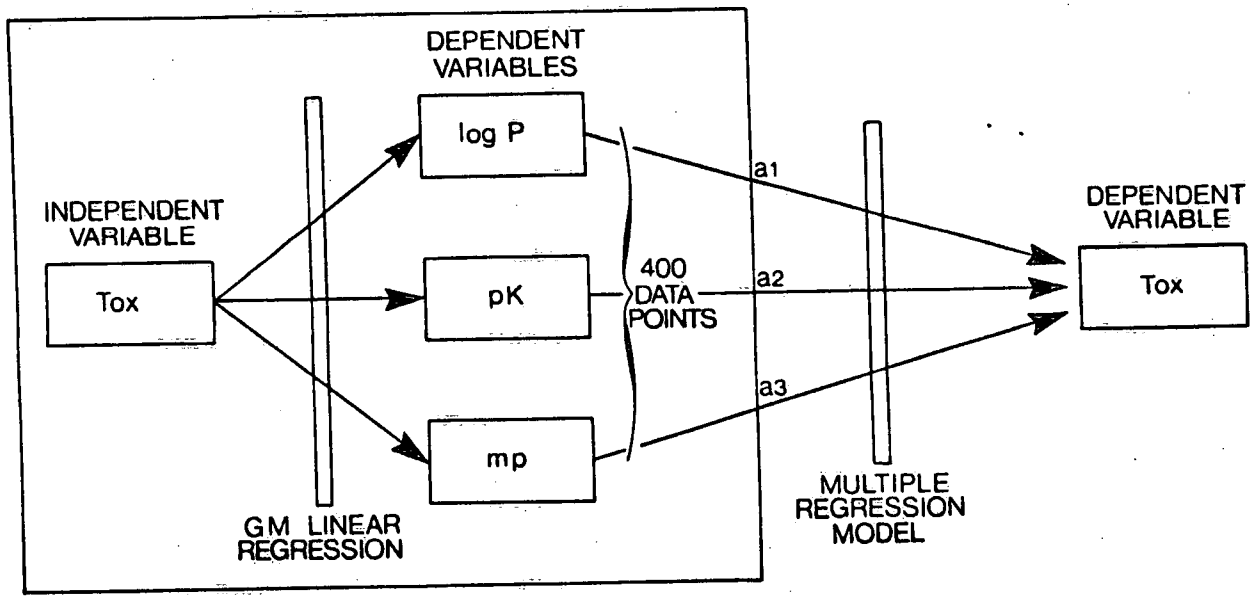
Figure 3 Measured and estimated Microtox toxicity values of point no. 5, by using the two-step procedure. One standard deviation and 95% confidence limits are shown. Note that the frequency distribution of the 400 bootstrap estimates is skewed.

ONE STEP PROCEDURE

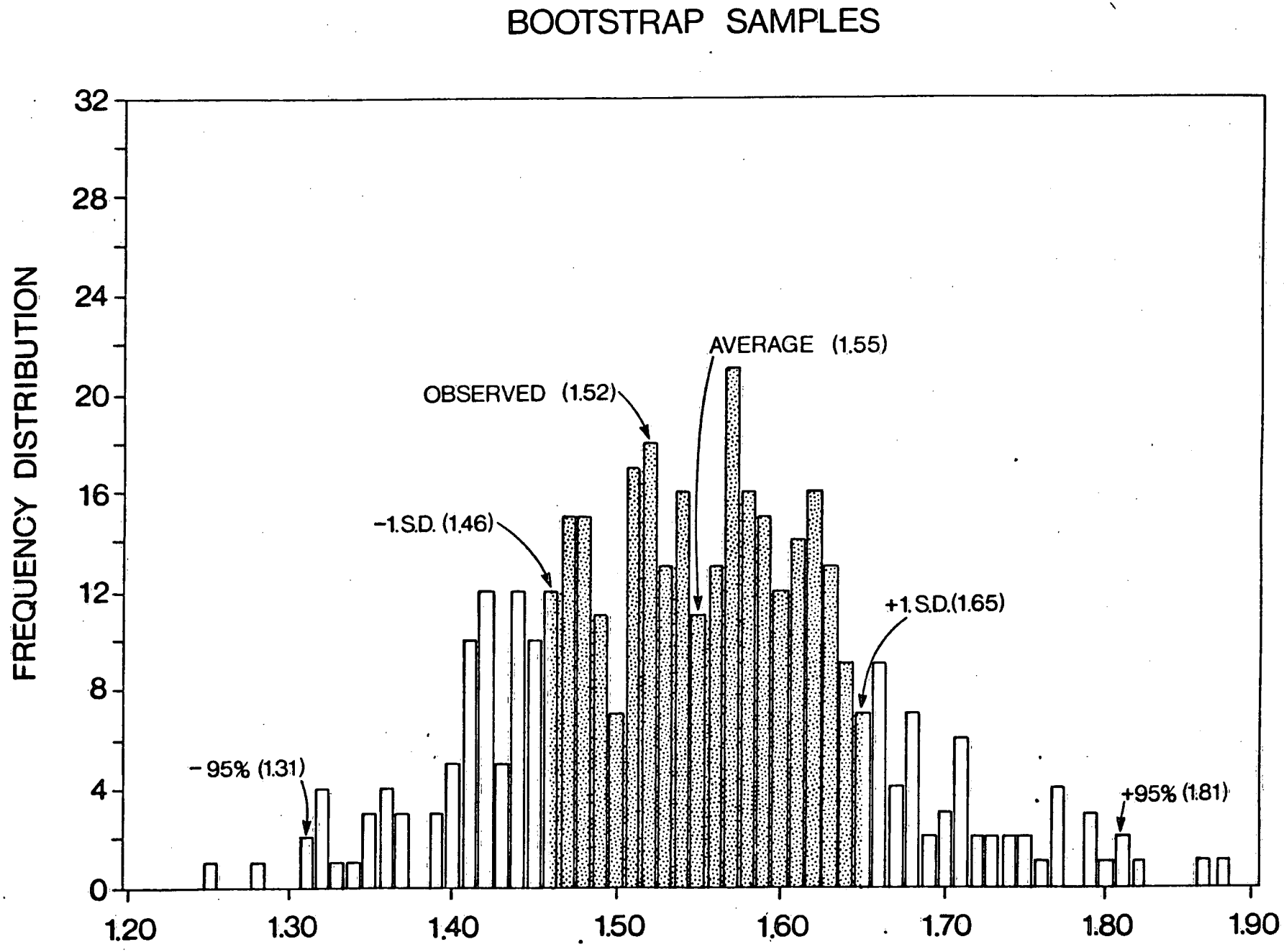


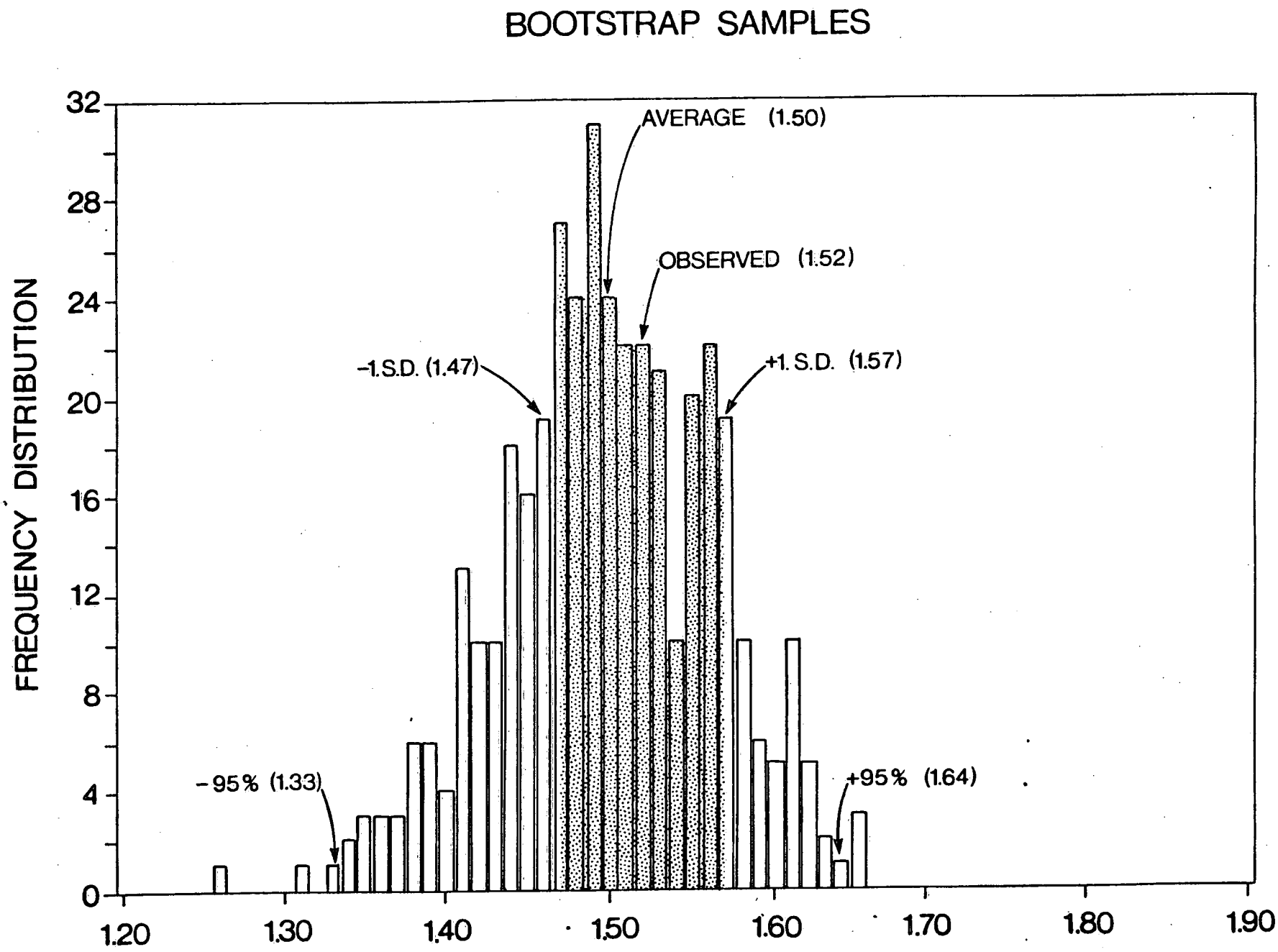
400 BOOTSTRAP SAMPLES

TWO STEP PROCEDURE



400 BOOTSTRAP SAMPLES


BOOTSTRAP SAMPLES
FREQUENCY DISTRIBUTION
32
28
24
20
16
12
8
4
0
1.20
1.30
1.40
1.50
1.60
1.70
1.80
1.90
AVERAGE (1.55)
OBSERVED (1.52)
-1.S.D. (1.46)
+1.S.D.(1.65)
- 95% (1.31)
+95% (1.81)


BOOTSTRAP SAMPLES
FREQUENCY DISTRIBUTION
32
28
24
20
16
12
8
4
0
1.20
1.30
1.40
1.50
1.60
1.70
1.80
1.90
AVERAGE (1.50)
OBSERVED (1.52)
-1.S.D. (1.47)
+1. S.D. (1.57)
-95% (1.33)
+95% (1.64)