Journal of Modern Applied Statistical Methods

Volume 12 | Issue 1 Article 9

5-1-2013

Using the Bootstrap for Estimating the Sample Size in Statistical Experiments

Maher Qumsiyeh
University of Dayton

Recommended Citation

Qumsiyeh, Maher (2013) "Using the Bootstrap for Estimating the Sample Size in Statistical Experiments," *Journal of Modern Applied Statistical Methods*: Vol. 12 : Iss. 1 , Article 9.
DOI: 10.22237/jmasm/1367381280

Using the Bootstrap for Estimating the Sample Size in Statistical Experiments

Cover Page Footnote

This research is supported by University of Dayton Seed Grant.

Journal of Modern Applied Statistical Methods
May 2013, Vol. 12, No. 1, 45-53

Using the Bootstrap for Estimating the Sample Size in Statistical Experiments

Maher Qumsiyeh
University of Dayton,
Dayton OH

Efron’s (1979) Bootstrap has been shown to be an effective method for statistical estimation and testing. It provides better estimates than normal approximations for studentized means, least square estimates and many other statistics of interest. It can be used to select the active factors - factors that have an effect on the response - in experimental designs. This article shows that the bootstrap can be used to determine sample size or the number of runs required to achieve a certain confidence level in statistical experiments.

Key words: Efron’s bootstrap, experimental factors, statistical estimation, confidence level.

Introduction

Traditional methods of finding sample sizes depend on knowing the underlying distribution. For example, to determine a sample size that will result in (1-α)×100% confidence that the sample mean is within E units from the population mean the following is used:

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2} \tag{1}$$

assuming normality, or using the central limit theorem, and determining an approximate value for σ. For a multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \tag{2}$$

a (1−α)×100% confidence interval for β_j is given by

$$\widehat{\beta_j} \pm t_{\propto/2} \sqrt{\frac{1-R^2}{(1-R_j^2)(N-p)}}. \tag{3}$$

where, R^2 is the multiple coefficient of determination and R_j^2 is the same when x_j is predicted from the remaining $k-1$ regressors.

Using equation (3), the sample size to predict the standardized coefficient within E units of the true value (replacing t with normal) is given by

$$n = \frac{z_{\propto/2}^2}{E^2}\left(\frac{(1-R^2)}{\left(1-R_j^2\right)}\right) + p. \tag{4}$$

However with this *n* there is approximately a 50% chance that the interval will be longer than 2E (Kelly & Maxwell, 2003).

Hahn and Meker (1991) provide a value for *N* for which the confidence is (1-δ)×100% that the interval obtained is of a length less than or equal 2E. The value of such *N* is

$$N = \frac{z_{\propto/2}^2}{E^2}\left(\frac{(1-R^2)}{\left(1-R_j^2\right)}\right)\left(\frac{\chi_\delta^2(n-1)}{n-p}\right) + p. \tag{5}$$

where *n* is the value found in equation (4).

An alternative is to use the bootstrap method to determine if the sample size calculated using equations (1) and (5) is necessary or if it is larger than what is needed to achieve a certain confidence. The bootstrap has been shown to provide better than normal estimates of distribution functions of studentized

Maher Qumsiyeh is an Assistant Professor in the Department of Mathematics. Prior to joining the University of Dayton he was a Professor at Bethlehem University for over 25 years, with 10 years as Department Chair. Email him at: qumsiyeh@udayton.edu.

statistics (Singh, 1981; Bickle & Freedman, 1980; Babu and Singh, 1983, 1984). Qumsiyeh (1994) showed that bootstrap approximation for the distribution of the studentized least square estimate is asymptotically better, not only than the normal approximation, but also than the two-term Edgeworth expansion. Lahiri (1992) showed the superiority of the bootstrap for approximating the distribution of M-estimators. Bhattacharya and Qumsiyeh (1989) preseneted an L^p -comparison between the bootstrap and Edgeworth expansions. Finally, Qumsiyeh and Shaughnessy (2008, 2010) showed that the bootstrap can be used to determine the active factors in two level designs and how to estimate missing responses in those designs. In this study the bootstrap was applied to three data sets; SAS and the SQL procedure in SAS were used to perform calculations and resampling.

Data Set 1

Data set 1 is comprised of 1,000 randomly selected samples of size 61 each from a normal distribution with mean 20 and standard deviation 2 (61 is the number n obtained using Equation (1) with E = 0.5 and α = 0.05). An example of one such sample of size 61 is:

21.39	19.92	19.08	19.86
19.98	20.47	22.84	16.87
20.86	21.29	22.25	20.14
23.20	19.86	21.95	19.11
19.74	23.06	17.06	19.06
19.92	21.22	25.37	21.60
19.06	20.87	22.99	21.77
22.14	21.83	19.61	17.87
19.59	18.42	17.43	18.98
19.16	20.49	19.19	19.07
18.10	19.12	21.01	19.69
19.13	19.20	19.55	18.51
17.66	20.90	21.88	21.09
17.53	20.97	20.41	21.68
18.93	19.56	19.56	19.17
15.91			

The mean for this sample was 20.068 and the standard deviation was 1.75.

Data Set 2

Data set 2 is real data that correlates the GPA y (out of 4) of 194 students from Bethlehem University, with their high school Math (x_1) and English (x_2) scores (out of 100 points). The first few observations are shown in Table 1. The model for data set two is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Table 1: Data Set 2 Example

y	x_1	x_2
2.55	75	77
3.69	87	99
2.48	80	70
1.90	70	65
2.07	70	89
2.73	72	64
1.81	80	66
2.30	71	67
1.76	83	66
2.17	78	89
1.77	65	60

Data Set 3

Data set 3 is an example provided by Bisgaard and Fuller (1995). It is a 2^4 full factorial experiment to determine if blade size (A), centering (B), leveling (C) and speed (D) had an effect on the occurrence of undesirable marks on a steel sample. The design matrix is shown in Table 2, where Y represents the number of defective (undesirable marks) among 20 samples at each setting and $\widehat{\mathrm{P}}$ is the proportion of defects at each setting.

The Bootstrap

The bootstrap was used to analyze the three data sets. SAS programming and the SQL procedure in SAS were used to perform the analyses. Resampling with replacement was conducted 1,000 times based on Efron and Tibshirani (1993) finding that 1,000-2,000 works best. The SAS program used for data set 1

Table 2: Data Set 3 Design Matrix

Run	A	B	C	D	Y	$\widehat{P}$
1	-1	-1	-1	-1	0	0
2	1	-1	-1	-1	16	0.8
3	-1	1	-1	-1	0	0
4	1	1	-1	-1	20	1
5	-1	-1	1	-1	0	0
6	1	-1	1	-1	10	0.5
7	-1	1	1	-1	0	0
8	1	1	1	-1	14	0.7
9	-1	-1	-1	1	0	0
10	1	-1	-1	1	10	0.5
11	-1	1	-1	1	0	0
12	1	1	-1	1	20	1
13	-1	-1	1	1	1	0.05
14	1	-1	1	1	12	0.6
15	-1	1	1	1	0	0
16	1	1	1	1	20	1

is provided in Appendix A; due to the length of the programs for data sets two and three they are not provided. A different procedure was used for each data set.

Data Set 1

For the first example the sample size was 61, this is the sample size necessary for 95% confidence that the sample mean is within 0.5 units from the population mean using the equation: $n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$. Using resampling and taking a random sample of size 61 from a N(20, 2^2) distribution, it is resampled 1,000 times with replacement, the mean of each of the 1,000 samples is calculated and half the difference between the 2.5 and 97.5 percentiles of the 1,000 means is found. This should be the value of E. One sample of size 61 from a N(20, 2^2) distribution to another the value of such E will vary to a great degree, thus, this procedure is repeated several times (500 in this case) and an interval for the values of such E's is listed. The sample size continued to decrease and the values of E continued to be recorded. Results are shown in Table 3.

Table 3: Data Set 1 Results

n	*E*
61	0.403-0.458
53	0.451-0.478
48	0.463-0.509
40	0.538-0.567

Table 3 shows that a sample of size 61 was not necessary; 48 would have been sufficient. The bootstrap was repeated 500 times, each resampling 1,000 times using a different 61 randomly selected data points with replacement from a $N(20,2^2)$ distribution and the values of all 500 replications were in the interval given which is 0.403-0.458 for n = 61. (See Appendix A for the SAS program used for data set 1 with n = 48.)

Data Set 2

The second data set, which correlates the university GPA to high school English and math scores, the model is: $y= \beta_0+ \beta_1 x_1+ \beta_2 x_2+\varepsilon$. The initial results using SAS for the whole data set are shown in Table 4.

First using equation (4),

$$n = \frac{z^2_{\propto/2}}{E^2}\left(\frac{(1-R^2)}{\left(1-R_j^2\right)}\right)+p\ ,$$

to find an initial n using α as 0.05 and E as 0.2 for the standardized betas (the value of E depends on the type of data at hand). Using all 194 data points and the values of R^2 and R_j^2 from the data set ($R^2 \approx 0.4192$ and $R_j^2 \approx 0.0163$) and equation (4), the value of *n* is 60 (approximating to the next integer), however, with this *n* there is approximately a 50% chance that the interval will be longer than 2E. (Kelly & Maxwell, 2003). By contrast, using equation (5),

$$N = \frac{z^2_{\propto/2}}{E^2}\left(\frac{(1-R^2)}{\left(1-R_j^2\right)}\right)\left(\frac{\chi^2_\delta\,(n-1)}{n-p}\right)+p,$$

δ = 0.05 and j = 2 the value of N is determined to be 78; this results in 95% confidence that the interval obtained is of a length less than or equal 2E (Hahn & Meker, 1991). Note that the value of E used is for the standardized betas; thevalue of E for the non-standardized betas (EN) will be approximately

$$\text{EN} \approx \frac{S_y}{S_{x_2}}\text{E} = \frac{0.51651973}{10.5807703} = 0.0097\,.$$

Next, the half-length of the confidence interval for β_2 is determined using the bootstrap method and a random sample of size 78 from the 194 original observations. The procedure is as follows:

1. Select 78 points using random sampling without replacement from the 194; name this as subset and perform a regular regression procedure obtaining ($X_{1,1}$, $X_{2,1}$, Y_1 ,E_1), ($X_{1,2}$, $X_{2,2}$, Y_2 ,E_2), ..., ($X_{1,78}$, $X_{2,78}$, Y_{78} ,E_{78}). Here, $E_i = Y_i - \widehat{Y}_i$, is the residual for the ith observation.

2. Select 1,000 samples with replacement from the subset, this is the bootstrap sample. Each sample has 78 points and samples are designated as {sample$_1$}, {sample$_2$}, ..., {sample$_{1000}$}.

3. Examine {sample$_1$}; n=78 points taken with replacement from the subset. We have sets of points ($X_{1,1}$*, $X_{2,1}$*, Y_1*, E_1*), ($X_{1,2}$*, $X_{2,2}$*, Y_2*, E_2*), ..., ($X_{1,78}$*, $X_{2,78}$*, Y_{78}*,E_{78}*). Each ($X_{1,j}$*, $X_{2,j}$*, Y_j*, E_j*) can be any of the ($X_{1,1}$, $X_{2,1}$, Y_1,E_1), ($X_{1,2}$, $X_{2,2}$, Y_2, E_2), ..., ($X_{1,78}$, $X_{2,78}$, Y_{78}, E_{78}) with probability 1/78.

4. Find the average of the E_i*'s, and name this ME_1 . Due to the fact that the mean of the errors is assumed to be 0, standardize the errors by subtracting ME_1 from each of them.

5. Still using {sample$_1$}, the new Y's are obtained by the adding the respective standardized error term to the predicted values and these are termed as new Y_j's.

6. Continue to examine {sample$_1$}; using the least-square method, find the slope and the intercept, (slope$_1$, slope$_2$, intercept$_1$), based on ($X_{1,1}$*, $X_{2,1}$*, , newY_1^*), ($X_{1,2}$*, $X_{2,2}$*, newY_2^*), ..., ($X_{1,78}$*,$X_{2,78}$*, newY_{78}*).

7. Repeat steps 3-6 for the other 999 samples to obtain 1,000 estimates for the intercept β_0 and the slopes β_1 and β_2. Interest is in β_2.

8. Estimate the value of β_2 by averaging the 1,000 estimates of β_2 and calculate a 95% CI for β_2 by finding the 2.5 and 97.5 percentile

of those 1,000 values. Half the length of this interval, E^*, will be compared with the EN = 0.0097 previously obtained.

Based on this procedure, how is it known that there is 95% confidence that half the length of the interval will not exceed EN? The answer is that by repeating steps 1-8, 1,000 times to obtain 1,000 EN's and then finding the top 95 percentile, it should not exceed EN.

The estimate for β_2 from one random subset of 78 points was $\widehat{\beta_2}$ =0.01655 and a 95% CI for β_2 was (0.0089, 0.0242). This assumes that all conditions, such as normal residuals and constant variances, hold; in addition, half the length of the interval is 0.0077, which is smaller than expected (0.0097). The n = 78 guarantees that 95% of the cases will result in smaller half lengths.

Using the bootstrap method discussed results in a mean half-length of 1,000 runs of the bootstrap method of 0.0095 and a 95% confidence interval of (0.00915, 0.0102): this is without any assumptions on the model. An estimate for β_2 was calculated as an average of the 1,000 bootstrap sample estimate for β_2 it was $\widehat{\beta_2}$ = 0.01643 for one run with a 95% CI of (0.0.0162, 0.0167). The bootstrap was repeated 1,000 times and the average value for the estimated values of β_2 was 0.01633 with a 95% CI of (0.01607. 0.01662). Without any assumption on the model, the bootstrap produced an estimate for β_2 that was close to that produced assuming the regular model assumptions hold, in addition, the length of the 95% CI was a little shorter than expected using a sample of size 78 (0.0095 vs. 0.0097). It is important to note that the calculations were carried out without assuming the error terms to be normal, however, it is valuable to understand what will happen if the error terms in the example are exactly normal.

Table 4: SAS Results for Data Set 2

Model: Model1							
Dependent Variable: y							
Number of Observations Read: 194							
Number of Observations Used: 194							
Analysis of Variance							
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F		
Model	2	21.58370	10.79185	68.92	<0.0001		
Error	191	29.90728	0.15658				
Corrected Total	193	51.49098					
Root MSE		0.39571		R-Square	0.4192		
Dependent Mean		2.65381		Adj R-Sq	0.4131		
Coeff Var		14.91081					
Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t value	Pr > |t|	95% Confidence Limits	
Intercept	1	-0.73516	0.29076	-2.53	0.0123	-1.30868	-0.16163
x1	1	0.02696	0.00294	9.16	<.0001	0.02116	0.03277
x2	1	0.01659	0.00271	6.11	<.0001	0.01124	0.02194

2^{nd} Data Set (Normal Errors)

For a random sample of size 78, the predicted values of the y_i's, $\widehat{y_i}$ were calculated and a new variable w_i was defined as:

$$w_i = \widehat{y_i} + c\varepsilon_i$$

where ε_i is randomly chosen from a N(0,1) distribution. The new w_i's with the original x's will have normal errors with constant variance. Note that the variance of the w_i's must be the same as the original y_i's to be able to compare the length of the confidence interval for the new β_2 with the previous one. The solver function in Microsoft Excel was used to provide a value for c to achieve this; this value of c was calculated to be 0.4123. This improved the previous results and half the length of the 95% CI for β_2 using the bootstrap method was much smaller (0.0089 vs. 0.0097). This shows that using the bootstrap requires a smaller sample size than the previous estimate of n = 78.

Data Set 3

Bisgaard and Fuller (1995) provided a table that gives estimated sample sizes (n) for the number of runs at each setting for two level full factorial experiments using proportions as a response. Their estimate for n which represents the number of runs needed to detect an error of size Δ in the untransformed scale is given by

$$n = \frac{(z_{\propto/2} - z_{\beta})^2}{N\delta^2} \tag{6}$$

where N is the total number of basic runs in a 2^k factorial experiment (4, 8, 16, …), α and β are the probabilities of type I and type II errors, 0.05 and 0.1 respectively, and δ is the expected value of the effect (Bisgaard & Fuller, 1995). Bisgaard and Fuller's table presents values of Δ that vary from 10% to 90% of the proportion of defective (p_0) and shows that sample size depends on the average defective level. If the average defective level is low, for example 5%, a larger sample size is needed to indicate that a change has truly occurred.

For the 3^{rd} data set the current level of defective (p_0) was not known, it was approximated with the average proportion of defective in the sample at each setting which is $\hat{\bar{p}} \approx 0.384$. Because n is given in this experiment as 20, the method described in Bisgaard and Fuller (1995) or the table they provide can be used to determine the minimum size of detectable error. For α to be at most 0.05, the minimum size is $|\Delta|>0.185$; calculating the effect size of each factor, it was found that factors A, B and the AB interaction have effect sizes larger than this (0.984, 0.208, 0.247 respectively). This agrees with the half normal plot (Daniel, 1959) which states that factors B and the AB interaction appear to be slightly active (not very clear) and that factor A is a definitely active factor (see Figure 1).

In the calculations described it is not certain that a 95% confidence interval for the effect size will have its lower bound less than 0.185 for those factors (A, B and AB).

Qumsiyeh and Shaughnessy (2008, 2010) showed that the bootstrap can be used (under no assumptions) to determine active factors in factorial experiments, to estimate the size of the effect and to determine a confidence interval for the effect size. The method can be described with the following steps using factor A for illustration purposes:

1. Sample N/2 responses with replacement from data at the +1 level of the given factor A.

2. Sample N/2 responses with replacement from data at the −1 level of the given factor A.

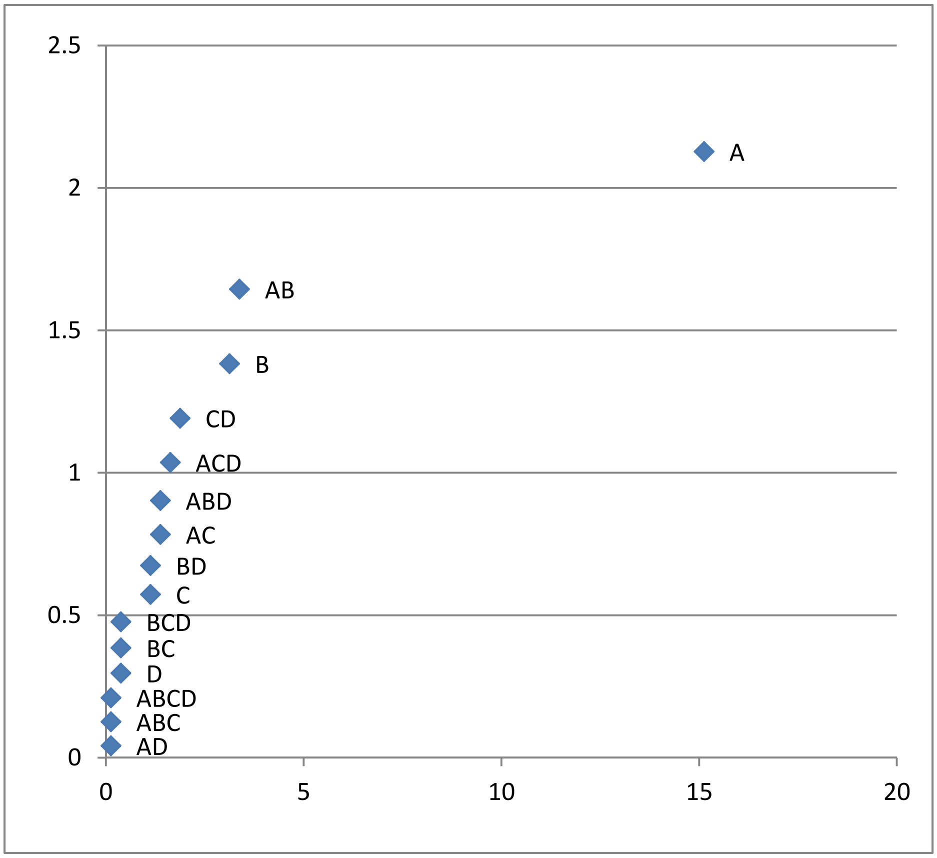
3. Estimate the effect of that factor using the difference between +1 level and −1 level.

4. Repeat the sampling procedure a large number of times (1,000 in this example).

5. Find the average of the 1,000 values; this is an estimate of the effect size of factor A.

Determine the upper (1−α/2) and lower α/2 percentile points of the resampled effect values found in step 4. Use these values to construct the

Figure 1: Half-Normal Plot for the Effects in Data Set 3



effect size. If the confidence interval doesn't contain 0 then this factor is an active factor – a factor that has an effect on the response.

Using the procedure described previously for data set 3, the following were determined: All confidence intervals for the effect size for all factors except factor A contained 0, therefore they must be assumed as inactive factors. For factors A, B and the AB which appear to be effective using the normal plot and are reported as active factors by Bisgaard and Fuller (1995), using the proportion of defectives, the results were as follows:

- Factor A: The mean effect size for the 1,000 runs was 0.7566 and a 95% confidence interval for the effect size was (0.6188, 0.8875).

- Factor B: The mean effect size for the 1,000 runs was 0.1500 and a 95% confidence interval for the effect size was (–0.2313, 0.5406).

- Factor AB (The AB interaction): The mean effect size for the 1,000 runs was 0.1776 and a 95% confidence interval for the effect size was (–0.2406, 0.5894).

Results show that only factor A can be considered active. If this is the case, the confidence interval for effect A has a lower bound of 0.618 which leads to a much higher value than the least expected of 0.185. This indicates that a sample size smaller than 20 would have been sufficient.

Conclusion

The bootstrap method can be used to determine sample sizes in statistical experiments and to check whether a certain sample size used is more than is needed by examining the length of the confidence interval resulting from using the bootstrap method. The bootstrap is also good for selecting active factors and in constructing confidence intervals for effect size. The availability of computers and statistical software make using re-sampling (bootstrap) easy and fast and provides good predictions.

Acknowledgment

This research is supported by University of Dayton Seed Grant.

References

Babu, G., & Singh, K. (1983). Inference on means using the bootstrap. *The Annals of Statistics*, *11*, 999-1003

Babu, G. J., & Singh, K. (1984). On one term Edgeworth correction by Efrons bootstrap. *Sankhya*, *46*, *Series A*, 219-232.

Bhattacharya, R. N., & Qumsiyeh, M. (1989). Second order and L^p- comparison between the bootstrap and empirical Edgeworth expansion methodologies. *Annals of Statistics*, *17*, 160-169.

Bickel, P. J., & Freedman, D. A. (1980). On Edgeworth expansions for the bootstrap. Unpublished.

Bisgaard, S., & Fuller, H. (1995). Sample size estimate for 2^{k-p} designs with binary responses. *Quality Engineering*, *27*(*4*), 344-354.

Daniel, C. (1959). Using of half normal plots in interpreting factorial two-level experiments. *Technometrics*, *1*, 311-341.

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

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. New York, NY: Chapman and Hall.

Hahn, G. J., & Meeker, W. Q. (1991). Statistical intervals: A guide for practitioners. New York, NY: Wiley.

Kelley, K., & Maxwell, E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(*3*), 305-321.

Lahiri, S. (1992). Bootstrapping M-estimators of a multiple linear regression parameter. *The Annals of Statistics*, *20*(*3*), 1548-1570.

Qumsiyeh, M. (1994). Bootstrapping and empirical Edgeworth expansions in multiple linear regression models. *Communications in Statistical Theory and Methods*, *23*(*11*), 3227-3239.

Qumsiyeh, M., & Shaughnessy, G. (2008). Using the bootstrap to select active factors in unreplicated factorial experiment. In *JSM Proceedings*, *Statistical Computing Section*. Alexandria, VA: American Statistical Association.

Qumsiyeh, M., & Shaughnessy, G. (2010). Bootstrapping Un-replicated two-level designs with missing responses. *Journal of Statistics: Advances in Theory and Applications*, *4*, 91-106.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, *9*, 1187-1195.

Appendix A: SAS Program Used for Data Set 1

```
%macro rep;
%do rep=1 %to 500;
data ma (drop=i);
*%let num=48;
*do i=1 to &num; ;
do i=1 to 48;
x=2*rannorm(56367)+20;
output;
end;
run;
%macro numbering(N);
data numbering;
do i=1 to &N; output; end; run;
quit;
data ma1;
set numbering;
set ma;
run;
%mend;
%numbering(48);
%macro repeat;
```

Appendix A (continued): SAS Program Used for Data Set 1

```
%do repeat=1 %to 1000;
%macro distinct(nThrow, N);
data table2;
      do j = 1 to &nThrow;
                  pt = int(int(ranuni(0) * &N) + 1.5);
    set ma1 point=pt;
                  output;
  end;
stop;     * required for point= ;
run;
%mend;
%distinct(48,48);
run;
proc means data=Table2 n mean noprint;
var x;
output out=Table3 n=n mean=mx;
run;
quit;
proc sql;
create table tableF as
select a1.mx as mx1
from table3 as a1;
run;
proc append data=TableF base=summary force;
run;
%end;
%mend;
%repeat;
proc univariate data=summary noprint; var mx1;
output out=z1 mean=tm pctlpts = 2.5, 97.5 pctlname=p25 p975 pctlpre = mx1;
run;
data E;
set z1;
E=(mx1p975-mx1p25)/2;
run;
proc append data=E base=E1 force;
run;
proc sql;
drop table  E ;
drop table  Ma ;
drop table  Ma1 ;
drop table  Numbering ;
drop table  Summary ;
drop table  Table2 ;
drop table  Table3 ;
drop table  Tablef ;
drop table  z1 ;
run;
quit;
%end;
%mend;
%rep;
proc univariate data=E1 noprint; var E;
output out=Length mean=tm pctlpts = 2.5, 97.5 pctlname=p25 p975 pctlpre = E;
run; quit;
```