

**BOOTSTRAPPING UNREPLICATED  
TWO-LEVEL DESIGNS WITH  
MISSING RESPONSES**

**MAHER QUMSIYEH  
and GERALD SHAUGHNESSY**

Department of Mathematics  
University of Dayton  
300 College Park  
Dayton OH 45469-2316  
USA  
e-mail: qumsiyeh@notes.udayton.edu

**Abstract**

This paper will demonstrate how the bootstrap can be used to analyse unreplicated two-level designs with some missing responses. Also, it will be shown how the bootstrap can be used to construct confidence intervals for the effect size, and how it can be used to estimate the missing values.

**1. Introduction**

Missing values in factorial and fractional factorial designs destroys the orthogonal structure of the design and leads to relatively complicated least square analysis. Several papers exist on this subject. Draper and Stoneman [12] give a method to estimate the missing values, but their method depends on sacrificing some of the effects to estimate the missing

---

2010 Mathematics Subject Classification: 62K15.

Keywords and phrases: bootstrap, unreplicated factorial experiments, effect estimation, confidence intervals.

Received November 10, 2010

values. Wilkinson [23, 24] gives a method that can require considerable computations. Shearer [19] gives a new procedure to use with factorial designs by using an iterated method and convergence of such iteration.

In this paper, we propose the bootstrap approximation of Efron [13] for selecting the active factors (factors that have influence on the response) and for approximating the size of the effect. After doing this, we will show how the bootstrap can be used to estimate the missing values, and we will compare our estimates of missing values with others, also we will construct confidence intervals for the effect size.

Recently, the bootstrap was shown by Qumsiyeh and Shaughnessy [18] to be a very effective method in selecting active factors in unreplicated two-level factorial design, but without missing values. Benski [3] compares nine different methods for determining the active factors, among those methods are Box and Meyer [8], La Pena and La Pena [15], Lenth [16], and Voss [22].

The bootstrap is proved to provide “better than normal” estimates of distribution functions of studentized statistics, see, for example, Singh [20], Bickle and Freedman [5], and Babu and Singh [1, 2]. Qumsiyeh [17] proved 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 [14] show the superiority of the bootstrap for approximating the distribution of  $M$ -estimators. Bhattacharya and Qumsiyeh [4] do an  $L^p$ -comparison between the bootstrap and Edgeworth expansions.

## **2. Bootstrap Method to Select Active Factors**

Unreplicated factorial experiments are commonly employed in industrial settings effect sparsity is a common assumption. To identify the active factors in such experiments, especially in the case of missing values, we propose the following procedure:

**First.** Two level full factorial unreplicated experiments with missing data points.

Assume the data set has  $N$  total responses and that one response is missing, let us assume that we want to test, if some effect  $L$  is active or not ( $L$  could be the interaction of other effect, for example,  $ABC$ ):

- Sample  $N/2 - 1$  responses with replacement from data at the +1 level of a given effect.
- Sample  $N/2 - 1$  responses with replacement from data at the -1 level of a given effect.
- Estimate the effect of that factor using the difference between the average at the +1 level and -1 level.
- Repeat the sampling procedure a large number of times (1,000 in our example).
- Determine the upper  $(1 - \alpha / 2)$  and lower  $\alpha / 2$  percentile points of the resampled effect values.
- Use these values to construct the  $(1 - \alpha) \times 100$  percent confidence interval for the effect size.
- If the confidence interval does not contain zero, then the factor is identified as an active factor (have influence on the response), otherwise, it is an inactive factor.
- We propose estimating the missing value by using the least effective factor (the factor for which its confidence interval definitely contains zero and the distance from zero to the closest end point is larger than the same distance for all other factors), and having the difference between the average at the (+) setting and the (-) setting be zero, and solve for the missing value.

**Notes.** (1) Sampling is done with replacement. It can be done by using SAS, EXCEL, or any resampling software available, in our case, it was done by SAS.

(2) The above method would work best, if your total runs of the experiment are at least  $2^3 = 8$  with one missing value.

(3) If you have two or more missing values, then we can do the same thing, but the number of runs should be  $2^4 = 16$  runs; For the case of two missing values, if both from the (+) setting of  $L$ , then you would resample  $(N/2 - 2)$  at the (+) setting and  $(N/2 - 2)$  at the (-) setting. If the missing responses are one at the (+) setting of  $L$  and other at the (-) setting, then you resample  $N/2 - 1$  responses at each setting. You will do something similar for three missing data points as we will see later.

(4) Again, if we have two missing values, we can estimate them by using the two least effective factors as described above and having the difference between the average at the (+) setting and the (-) setting for these be set to zero, and solve for the two equations with two unknowns for missing value.

(5) Some might think that this is similar to a  $t$ -test. However, after we estimate the missing responses and produce the normal or half normal plots, it will consider all the different interactions.

**Second.** Two level fractional factorial unreplicated experiments:

The previous procedure will also work with fractional factorials under the previous conditions, however, we have to realize that some factors will be confounded with others and that we should be left with at least three data points at the (+) setting and three data points at the (-) setting.

### 3. The Data Sets

Two data sets will be considered to do the above analysis. The results will be compared with the results obtained by others.

#### Data set I

Our first data set is an example of a  $27^{-4}$  fractional factorial design given in Box and Hunter [6] and used by Draper and Stoneman [12].

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>Y</i>
-1	-1	-1	-1	1	1	1	68.4
1	-1	-1	-1	-1	-1	1	77.7
-1	1	-1	-1	-1	1	-1	(66.4) <i>M</i>
1	1	-1	-1	1	-1	-1	81.0
-1	-1	1	-1	1	-1	-1	78.6
1	-1	1	-1	-1	1	-1	41.2
-1	1	1	-1	-1	-1	1	68.7
1	1	1	-1	1	1	1	38.7

*M* here indicate the missing value, which was 66.4 in the original data set, when the observations are complete, seven combinations of effects can be estimated, in addition to the mean. Here  $D = ABC$ ,  $E = AB$ ,  $F = AC$ , and  $G = BC$ .

**Data set II**

Our second examples are four data sets of a 16 run two-level design from a paper by Box and Meyer [8]:

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>Y1</i>	<i>Y2</i>	<i>Y3</i>	<i>Y4</i>
-1	-1	-1	-1	0.23	43.7	14	0.08
1	-1	-1	-1	0.3*	40.2	16.8	0.04
-1	1	-1	-1	0.52	42.4	15	0.53
1	1	-1	-1	0.54	44.7	15.4	0.43
-1	-1	1	-1	0.7*	42.4	27.6	0.31
1	-1	1	-1	0.76	45.9	24	0.09
-1	1	1	-1	1	42.2	27.4*	0.12
1	1	1	-1	0.96	40.6	22.6	0.36
-1	-1	-1	1	0.32	42.4	22.3	0.79
1	-1	-1	1	0.39	45.5	17.1	0.68
-1	1	-1	1	0.61	43.6	21.5	0.73
1	1	-1	1	0.66	40.6	17.5*	0.08
-1	-1	1	1	0.89	44	15.9	0.77
1	-1	1	1	0.97	40.2	21.9	0.38
-1	1	1	1	1.07	42.5*	16.7*	0.49
1	1	1	1	1.21	46.5*	20.3	0.23

For this data set, we are going to assume that we have two missing data points from each data set with response  $Y_1$  or  $Y_2$ , and three missing data points from the data set with response  $Y_3$ . For the data set with response  $Y_4$ , we did the experiment assuming one, two, and three missing data points. Assume the ones with \* are the ones assumed missing, they were picked out at random, the experiment was repeated by using other missing data points and yielded the same result.

#### 4. Data Analysis

##### Data set I

The average response for each of the 1,000 runs, for each effect  $A-G$ , at the high and low setting was obtained resampling three of the 4 responses at each setting. A factor effect is judged to be *active*, if the bootstrap confidence interval for the effect does not contain zero. As before, the effect with confidence interval that definitely contain the zero and for which, the zero is farthest from the closer end points among all other effects, is the factor we will use to estimate the missing value.

The effect size (the difference between the average of the +1 setting and the -1 setting) and the confidence interval for the effect are given for each of the 7 factors in Table 1:

**Table 1**

Effect	Effect Estimate	95%Confidence Interval		Estimate of missing value Assuming the effect inactive
		Lower	Upper	
$A$	- 12.8	- 32.78	10.2	22.8
$B$	- 3.5	- 29.6	23.62	77.8
$C$	- 19.05	- 39.27	0.5	0
$D$	0.66	- 24.82	18.00	64.4
$E$	4.60	- 24.9	29.83	79.2
$F$	- 27.34	- 39.7	- 7.4	157.6
$G$	- 3.55	- 29.3	20.32	52.8

From the above table, we can see that only effect  $F$  does not contains 0, and is definitely active.

Also, the above table agrees with results obtained by Box and Hunter [6] and used by Draper and Stoneman [12], actually Draper and Stoneman indicate clearly that we can not use  $C$  or  $F$  to estimate the missing value, which is what we came up with since  $F$  the active effect and 0 is very close to the confidence interval for effect  $C$ .

The most inactive effect from the above would be effect  $E$  and  $B$ , because the distance from the end points to 0 is at least 23. We can use these to estimate the missing value; we can, for example, average the two estimates  $((77.8 + 79.22)/2)$ , which is 78.5.

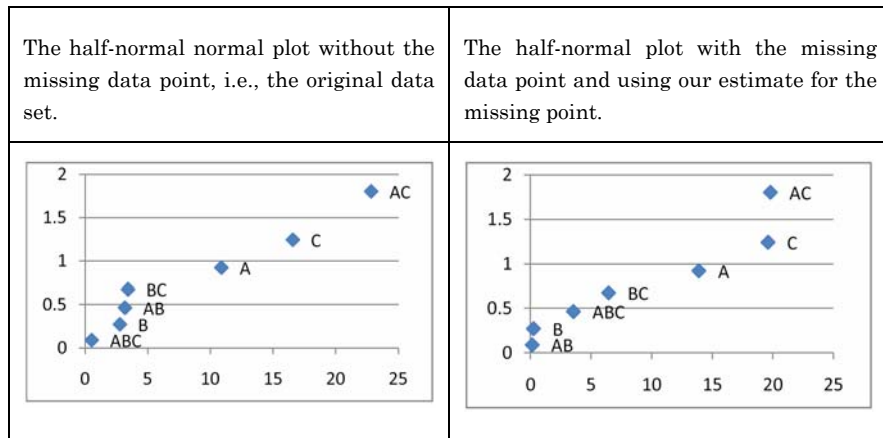
Now using this estimate of the missing value, we can do the bootstrapping again, with 4 sample points at each setting, and as in Qumsiyeh and Shaughnessy [18]. The results are given in Table 2:

**Table 2**

Effect	Effect Estimate	95%Confidence Interval	
		Lower	Upper
$A$	- 13.7	- 33.43	5.91
$B$	- 0.28	- 19.55	19.90
$C$	- 19.64	- 36.28	- 2.55
$D$	0.66	- 24.82	18.00
$E$	1.31	- 21.06	24.55
$F$	- 19.38	- 36.14	- 3.29
$G$	- 6.37	- 25.38	15.55

Again, this table shows that  $F$  is the only active factor.

The half normal plot for the original data set and the one using our estimate for the missing data point are given in Table 3.

**Table 3**

As we can see from the half normal plots, that effect  $F = AC$  is clearly active when we substituted 78.5 for the missing value. It is not clear from the original data if  $AC$  is active.

### Data set II

The average response for each of the 1,000 runs, for each effect  $A$  through  $D$  and their interactions are performed, at the high and low setting was obtained. Since we are assuming two missing data points, we are resampling 6 or 7 of the 8 responses (or 8-the number missing) at each setting depending on whether at a certain setting, we have one or two missing data points. A factor effect is judged to be *active*, if the bootstrap confidence interval for the effect does not contain zero. For the case of two missing data points, the two most inactive effects as described before can be used to estimate the missing values.

The effect size (the difference between the average of the +1 setting and the -1 setting) and the confidence interval for the effect is given for each of the 15 factors. Results are given below for each of the responses  $Y_1$  through  $Y_4$ .



**Results for Y1**

Table 4 below, gives the bootstrap results for the response Y1:

**Table 4**

Effect	Effect Estimate	95%Confidence Interval	
		Lower	Upper
<i>A</i>	0.1300	- 0.171	0.43
<i>B</i>	0.2132	- 0.06	0.51667
<i>C</i>	0.5136	0.378	0.655
<i>D</i>	0.0948	- 0.215	0.41
<i>AB</i>	- 0.0702	- 0.3614	0.2614
<i>AC</i>	- 0.0615	- 0.3692	0.2383
<i>AD</i>	- 0.0286	- 0.3271	0.27
<i>BC</i>	0.0436	- 0.28	0.3443
<i>BD</i>	0.0593	- 0.3058	0.38
<i>CD</i>	0.0938	- 0.195	0.3564
<i>ABC</i>	0.0705	- 0.245	0.41
<i>ABD</i>	0.08442	- 0.2129	0.3821
<i>ACD</i>	0.0916	- 0.2291	0.4317
<i>BCD</i>	- 0.0744	- 0.3507	0.235
<i>ABCD</i>	- 0.0479	- 0.3492	0.2983

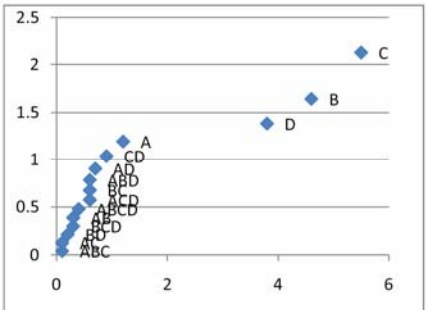
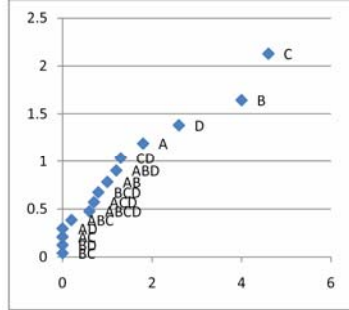
It can be seen from the above table that factor *C* is definitely active since its confidence interval does not contain zero, also it is good to note that factor *B* is almost active, it would be considered active, if we use a 90% confidence interval. These results agree with the results of Box and Meyer [8], Daniel [9, 10], and the half- normal plot method.

The two most inactive factors from the above table are *BD* and *ABCD*, they both definitely contain 0 and the distance from zero to the closest end point is at least 0.2983 for *ABCD* and 0.3058 for *BD*. However, we can not use these two factors to estimate the missing values because we will

end up with no solution, so we use  $ABCD$  and the next most inactive factor, which is  $BC$  and use these two to estimate the missing values. We get two equations with two unknowns; the estimates for the missing points would be in this case, 0.41 and 0.64.

The half normal plot for the original data without missing data points and the one using our estimate for the missing data points are given in Table 5.

**Table 5**

<p>The half-normal normal plot without the missing data point, i.e., the original data set.</p>	<p>The half-normal plot with the missing data points and using our estimate for the missing points.</p>
	

When two data values are missing, both graphs are similar in showing that  $B$  and  $C$  stand out, so they are identified as the active factors.

**Results for Y2**

Table 6 gives the bootstrap results for the response Y2:

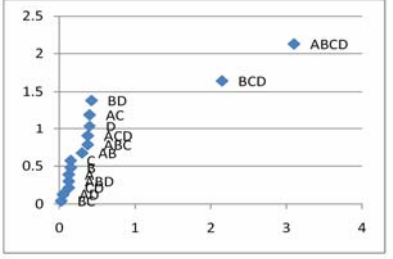
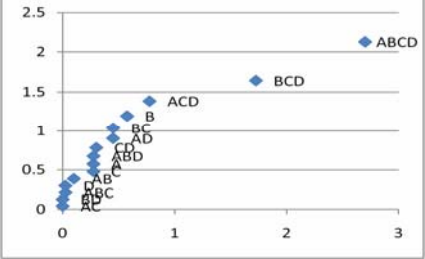
**Table 6**

Effect	Effect Estimate	95%Confidence Interval	
		Lower	Upper
<i>A</i>	- 0.3788	- 2.2357	1.5429
<i>B</i>	- 0.6351	- 2.575	1.5333
<i>C</i>	- 0.4226	- 2.3833	1.7833
<i>D</i>	- 0.0487	- 2.0583	1.85
<i>E = AB</i>	- 0.2231	- 2.2357	1.6929
<i>F = AC</i>	- 0.1195	- 2.1357	1.8857
<i>G = AD</i>	- 0.6156	- 2.4714	1.2857
<i>H = BC</i>	- 0.5544	- 2.633	1.3833
<i>J = BD</i>	- 0.170	- 2.1333	2.2833
<i>K = CD</i>	- 0.4073	- 2.6833	1.6167
<i>L = ABC</i>	- 0.1345	- 2.0286	1.8071
<i>M = ABD</i>	- 0.3923	- 2.2357	1.5
<i>N = ACD</i>	- 1.0192	- 2.85	0.75
<i>O = BCD</i>	- 2.0217	0.2667	3.7
<i>P = ABCD</i>	2.95	1.8143	3.9929

It can be seen from the above table that factor *BCD* and *ABCD* are definitely active since both confidence intervals do not contain zero. These results agree with the results of Box and Meyer [8], Taguchi and Wu [21], and the half-normal plot method. The two most inactive factors from the above table are *F* and *J*; they both definitely contain 0 and the distance from zero to the closest end point is at least 1.8857. We can use these two factors to estimate the missing values and we end up with two equations and two unknowns; the estimates for the missing points would be in this case, 42.4 and 43.2.

The half-normal plot for the original data without missing data points and the one using our estimate for the missing data points are given in Table 7.

**Table 7**

The half-normal normal plot without the missing data point, i.e., the original data set.	The half-normal plot with the missing data points and using our estimate for the missing points.
	

Again, with two missing data points, the graphs are similar in showing that *BCD* and *ABCD* can be identified as active factors.

### Results for *Y3*

With three missing data points, we have to do a resampling with replacement at each setting for each factor; if that factor has, for example, 3 missing data points at the (+) setting, then you would resample 5 out of 5 at the (+) setting and 5 out of 8 at the (-) setting. If the missing data points for some factor were two at the (+) setting and one at the (-) setting, then you would resample 6 out of 6 at the (+) setting and 6 out of 7 at the (-) setting and so on. We have four different situations for a specific factor with three missing data points:

1. All missing data points are at the (+) setting.
2. Two missing data points at the (+) setting and one at the (-) setting.
3. One missing data points at the (+) setting and two at the (-) setting.
4. All missing data points are at the (-) setting.

The results are given in Table 8 below:

**Table 8**

Effect	Effect Estimate	95%Confidence Interval	
		Lower	Upper
<i>A</i>	0.3502	- 4.15	4.7166
<i>B</i>	- 0.9839	- 5.94	3.79
<i>C</i>	4.6139	0.74167	8.475
<i>D</i>	0.4622	- 4.3333	4.85
<i>E = AB</i>	0.3583	- 3.9	4.9667
<i>F = AC</i>	1.6541	- 3.43	6.14
<i>G = AD</i>	- 0.5019	- 4.8333	3.8583
<i>H = BC</i>	- 1.3725	- 6.1	2.9583
<i>J = BD</i>	2.0225	- 2.5167	6.3917
<i>K = CD</i>	- 5.4561	- 8.5667	- 1.9
<i>L = ABC</i>	0.35	- 4.31	5.2
<i>M = ABD</i>	- 0.9699	- 5.325	3.2667
<i>N = ACD</i>	2.7595	- 1.32	7.27
<i>O = BCD</i>	1.4169	- 3.0667	6.1333
<i>P = ABCD</i>	- 2.6895	- 7.075	1.55

Again here, it can be seen from the above table that factor *C* and *CD* are definitely active since both confidence intervals do not contain zero. These results agree with the results of Box and Meyer [8], Box et al. [7], and the half-normal plot method. Box and Meyer indicate that factor *ACD* is also active and looking at our above results, 0 is the closest to one of the end points of the interval. If we decrease our confidence to 90%, it will be considered as active.

The three most inactive factors from the above table are *A*, *D*, and *ABC*; they both definitely contain 0 and the distance from zero to the closest end point is at least 4.15. We can use these three factors to estimate the missing values and we end up with three equations and three unknowns; the estimates for the missing points would be in this case, 42.4 and 43.2.

**Note.** If the three equations are unsolvable, then we have to use the next most inactive factor, in our case, they are solvable.

The half-normal plot for the original data without missing data points and the one using our estimate for the missing data points are given in Table 9.

**Table 9**

<p>The half-normal normal plot without the missing data point, i.e., the original data set.</p>	<p>The half-normal plot with the missing data points and using our estimate for the missing points.</p>

You can clearly see again that even with three missing data points, the graphs are similar in showing that *C*, *CD*, and *ACD* stand out, so they are the active factors.

**Results for Y4**

The results for Y4 agreed with the results of Box and Meyer [8] and with the results of Davis [11]. This was checked assuming one and two missing data points. However to avoid repetition, the results will not be displayed in this paper.

**Conclusion**

The results of this paper, with one, two or three missing data points agree totally with the results of Draper and Stoneman [12], Box and Meyer [8], and the original authors, who used these data sets. Also, the bootstrap estimates of the active effects agree well with the original

normal plot method and Box-Meyer method. Using the bootstrap estimates provide confidence interval for the effect size. The level of confidence can be adjusted to make the selection of active factors more or less stringent as was seen in the previous example. Also, these confidence intervals can be used to decide, which the most inactive factors are; this will help in estimating the missing responses.

Our choice was to do the resampling 1,000 times. Resampling much less than a thousand times will not provide as good results. It is not clear how many times you should resample at each level.

### References

- [1] G. J. Babu and K. Singh, Inference on means using the bootstrap, *The Annals of Statistics* 11 (1983), 999-1003.
- [2] G. J. Babu and K. Singh, On one term Edgeworth correction by Efrons bootstrap, *Sankhya* 46, Ser. A (1984), 219-232.
- [3] C. Benski, Applicability of nine numerical techniques for detecting active factors in unreplicated experimental designs, *ASA Proceedings of the Statistical Computing Section*, 214-217, American Statistical Association (Alexandria, VA) CISid: 154639 (1994).
- [4] R. N. Bhattacharya and M. Qumsiyeh, Second order and  $L^P$ -comparison between the bootstrap and empirical Edgeworth expansion methodologies, *Ann. Stat.* 17 (1989), 160-169.
- [5] P. J. Bickel and D. A. Freedman, On Edgeworth expansions for the bootstrap, Unpublished (1980).
- [6] G. Box and J. Hunter The  $2^{k-p}$  Fractional Factorial Designs, *Technometrics* 3(311-351) (1961), 449-458.
- [7] G. Box, W. Hunter and J. Hunter, *Statistics for Experiments*, John Wiley, New York, 1978.
- [8] G. Box and D. Meyer, An analysis for unreplicated fractional factorials, *Technometrics* 28(1) (1986), 11-18.
- [9] C. Daniel, Using of half normal plots in interpreting factorial two-level experiments, *Technometrics* 1 (1959), 311-341.
- [10] C. Daniel, *Applications of Statistics to Industrial Experimentation*, John Wiley, New York, 1976.
- [11] O. Davis, *The Design and Analysis of Industrial Experiments*, Oliver and Boyd, London, 1954.

- [12] N. Draper and D. Stoneman, Estimating missing values in unreplicated two-level factorial and fractional factorial designs, *Biometrics* (1964), 443-458.
- [13] B. Efron, Bootstrap methods: Another look at Jackknife, *The Annals of Statistics* 7 (1979), 1-26.
- [14] S. Lahiri, Bootstrapping  $M$ -estimators of a multiple linear regression parameter, *The Annals of Statistics* 20(3) (1992), 1548-1570.
- [15] J. La Pena and D. La Pena, A simple method to identify significant effects in unreplicated two-level factorials, *Comm. Stat. Theory and Methods* 21 (1992), 1383-1403.
- [16] R. Lenth, Quick and easy analysis of unreplicated factorials, *Technometrics* 31 (1989), 469-473.
- [17] M. Qumsiyeh, Bootstrapping and empirical Edgeworth expansions in multiple linear regression models, *Comm. Stat. Theory and Methods* 23(11) (1994), 3227-3239.
- [18] M. Qumsiyeh and G. Shaughnessy, Using the Bootstrap to Select Active Factors in Unreplicated Factorial Experiment, *JSM Proceedings, Statistical Computing Section, American Statistical Association, Alexandria, VA, 2008*.
- [19] P. Shearer, Missing data in quantitative designs, *Applied Statistics* 22 (1973), 135-140.
- [20] K. Singh, On the asymptotic accuracy of Efron's bootstrap, *The Annals of Statistics* 9 (1981), 1187-1195.
- [21] G. Tagushi and Y. Wu, *Introduction to Off-line Quality Control*, Central Japan Quality Control Association, Nagoya, Japan, 1980.
- [22] D. Voss, Generalized moulus-ratio test for analysis of fractional designs with zero degrees of freedom for error, *Comm. Stat. Theory and Methods* 17 (1988), 3345-3359.
- [23] G. Wilkinson, Estimation of missing values for the analysis of incomplete data, *Biometrika* 14 (1958), 257-286.
- [24] G. Wilkinson, A general recursive procedure for analysis of variance, *Biometrika* 57 (1970), 19-46.

