Analysis of Performance Measures in Experimental Designs Using Jackknife

ASOKAN MULAYATH VARIYATH and BOVAS ABRAHAM

University of Waterloo, Waterloo, ON N2L 3G1, Canada

JIAHUA CHEN

University of Waterloo, Waterloo, ON N2L 3G1, Canada and Nankai University, Tianjin, China

Experimental designs with performance measures as responses are common in industrial applications. The existing analysis methods often regard performance measures as sole response variables without replicates. Consequently, no degrees of freedom are left for error variance estimation in these methods. In reality, performance measures are obtained from replicated primary-response variables. Precious information is hence lost. In this paper, we suggest a jackknife-based approach on the replicated primary responses to provide an estimate of error variance of performance measures. The resulting tests for factor effects become easy to construct and more reliable. We compare the proposed method with some existing methods using two real examples and investigate the consistency of the jackknife variance estimate based on simulation studies.

KEY WORDS: Lenth's Method; Performance Measure; Pseudo Standard Error; Unreplicated Factorial Experiments.

Introduction

EXPERIMENTAL designs are widely used in industries to control and improve the quality of products. The basic purpose of these experiments is to arrive at a combination of factor levels that optimize the response or to identify the important factors that control the characteristic of interest. Designed experiments are also used to reduce the variation of the response by identifying critical factors (Taguchi (1986)).

There are many situations where analysis is

Mr. Variyath is a PhD Student in the Department of Statistics and Actuarial Science. He is a member of ASQ. His email address is mvasokan@uwaterloo.ca.

Dr. Abraham is a Professor in the Department of Statistics and Actuarial Science. He is a Fellow of ASQ. His email address is babraham@uwaterloo.ca.

Dr. Chen is a Professor in the Department of Statistics and Actuarial Science, University of Waterloo and Department of Statistics, Nankai University. His email address is jhchen@uwaterloo.ca.

performed on summary statistics of the primary-response variables. Hereafter, we refer to these summary statistics as performance measures. For example, in an experiment discussed in Wu and Hamada (2000, p. 124), the primary-response variable is the thickness of epitaxial layer on a silicon wafer. The aim of the experiment is to find the level combinations of the 4 factors such that its variation is minimized. In this case, the performance measure is chosen as the log(sample variance) of the replicated observations at each level combination. If the performance measure is regarded as our response, for the purpose of data analysis, then no degrees of freedom are left for error variance estimation. In this situation, the general practice is to use the analysis methods for unreplicated factorial experiments. See Wu and Hamada (2000).

A detailed review of analysis of unreplicated factorial experiments is available in Hamada and Balakrishnan (1998). Some widely used methods are (i) normal/half normal probability plots (Daniel (1959)) to identify the active effects and then to pool the inactive effects to arrive at an estimate of error variance and (ii) the Pseudo standard error (PSE) estima-

tion method by Lenth (1989). These simple methods will be discussed in more detail later with examples and/or in simulations.

The problem we are interested in is different from the analysis of unreplicated factorial experiments. We have replications at each design setting, but our focus is on a performance measure of these replications. When we use analysis methods for unreplicated factorial experiments for these performance measures, precious information in the replicated observations (for the primary response) is lost and, hence, the opportunity to obtain a proper estimate of the error variance is also lost. We suggest a method based on jackknife to recoup the information to estimate the error variance.

This paper is organized as follows. First, we discuss the most frequently used performance measures and introduce the jackknife method for analyzing the performance measures, which will be explained by two real examples. In the subsequent section, we explore the consistency of the jackknife method and compare its performance with that of Lenth's method. Some concluding remarks are given in the last section.

Jackknife Method for Performance Measures

Consider an experiment with n runs, each replicated m times. Let y_{ij} be the jth replicate of the ith experimental run, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. Most commonly used performance measures include $\eta_1 = \bar{y}_i$ and $\eta_2 = \log_e(s_i^2)$, where $\bar{y}_i =$

$(1/m)\sum_{j=1}^m y_{ij}$ and $s_i^2 = [1/(m-1)]\sum_{j=1}^m (y_{ij} - \bar{y}_i)^2$. Taguchi (1986) proposed a number of performance measures in the context of quality engineering based on the response of interest. These performance measures, referred to as signal-to-noise (SN) ratios are also discussed in Box and Meyer (1986), Leon et al. (1987), Box (1988), and Phadke (1989). For some explanation of these performance measures and the selection of appropriate ones in a given context, one may refer to these papers. Some frequently used performance measures are given in Table 1.

Resampling methods are commonly used for constructing simple and efficient variance estimators. Our idea here is to obtain an appropriate estimator of the variance of a performance measure by resampling methods. Jackknife and bootstrap are two widely used resampling methods. In general, both jackknife and bootstrap lead to similar variance estimates. For small sample sizes, jackknife is simple and quicker to use. Moreover, jackknife produces nonrandom estimates and it is easy to work with its closed-form solution. For small sample sizes, such as $m = 3$ or 4, there is high probability that the bootstrap samples contain only one distinct unit, which leads to η_2 and η_4 not being defined. Thus, we will only pursue jackknife in this paper.

Assume that we have n experimental runs and each run is replicated m times. Let $y_i = (y_{i1}, y_{i2}, \ldots, y_{im})$ be the vector of replications from the ith experimental run, with $c(y_i)$ being the corresponding performance measure. By deleting y_{ij} from y_i for $j = 1, \ldots, m$, we obtain m delete-one jackknife repli-

TABLE 1. Frequently Used Performance Measures

Response type	Notation	Formulas
Average	η_1	$\bar{y}_i$
Log variance	η_2	$\log_e(s_i^2)$
Smaller the better	η_3	$-10\log_{10}\left(\dfrac{1}{m}\sum_{j=1}^m y_{ij}^2\right)$
Nominal the better	η_4	$10\log_{10}\left(\dfrac{\bar{y}_i^2}{s_i^2}\right)$
Larger the better	η_5	$-10\log_{10}\left(\dfrac{1}{m}\sum_{j=1}^m y_{ij}^{-2}\right)$

cates of size $(m - 1)$, $y_{i(j)}$. Hence, we obtain m jackknife replications of the performance measure $c(y_{i(j)})$, $j = 1, 2, \ldots, m$. The jackknife variance estimate of the performance measure $c(y_i)$ is given by

$$\widehat{V}_{ja}\left(c\left(y_i\right)\right) = \frac{m-1}{m} \sum_{j=1}^{m} \left(c\left(y_{i(j)}\right) - c\left(y_{i.}\right)\right)^2, \quad (1)$$

where $c(y_{i.}) = (1/m)\sum_{j=1}^{m} c(y_{i(j)})$. A pooled estimate of the error variance is

$$\widehat{V}_{pja}\left(c\left(y\right)\right) = \frac{1}{n} \sum_{i=1}^{n} \widehat{V}_{ja}\left(c\left(y_i\right)\right). \quad (2)$$

Let us consider the F statistic,

$$F = \frac{\text{Mean Square for the Factor Effect}}{\widehat{V}_{pja}(c(y))}.$$

We suggest using this F to test the significance of factor effects in the ANOVA, where the variance estimate has $(m-1)n$ degrees of freedom. Some theoretical aspects of jackknife together with justifications for the F test are discussed in the Appendix.

In principle, one can obtain delete-d $(d = 1, 2, \ldots)$ jackknife replicates and a corresponding estimate of variance of $c(y_i)$. However, in the experimental design context, m is reasonably small and deleting more than one observation at a time can be problematic. Thus, we consider only the delete-one jackknife method.

Next we illustrate this method with two real examples from the literature together with existing analysis methods.

Example 1

This example is from Wu and Hamada (2000, p. 124). The nominal value of the thickness of the epitaxial layer on a silicon wafer is 14 μm, with a specification of ± 0.5 μm. The current process setting leads to excessive variation, and a 2^4 factorial experiment is conducted with four process factors (A, B, C, and D), and the experiment is replicated $m = 6$ times. In this experiment, we would like to identify the important factors that could be used to minimize epitaxial layer nonuniformity while maintaining average thickness close to the nominal value. Let us consider the performance measure $\eta_2 = \log_e(s^2)$. The design matrix together with values of the performance measure are given in Table 2.

Traditional half-normal plots are used to judge the significance of the factor effects. Factor A is judged as significant based on these plots. A formal test

TABLE 2. Design Matrix, Performance Measure, and the Jackknife Variance Estimates for Example 1

| | Factors | | | | | |
Run	A	B	C	D	η_2	$\widehat{V}_{ja}(\eta_2)$
1	−	−	−	+	−5.77	0.6904
2	−	−	−	−	−5.31	0.1665
3	−	−	+	+	−5.70	0.6371
4	−	−	+	−	−6.98	0.8964
5	−	+	−	+	−5.92	0.4658
6	−	+	−	−	−5.49	0.9030
7	−	+	+	+	−4.11	0.1596
8	−	+	+	−	−6.24	0.5398
9	+	−	−	+	−1.54	0.2893
10	+	−	−	−	−2.12	0.1446
11	+	−	+	+	−1.58	0.1155
12	+	−	+	−	−1.49	0.2961
13	+	+	−	+	−1.92	0.2711
14	+	+	−	−	−2.43	0.2231
15	+	+	+	+	−1.12	0.1129
16	+	+	+	−	−2.65	0.1816
	Pooled					0.3803

of effect significance using Lenth's method (1989) was also provided in Wu and Hamada (2000). The method uses a robust estimator of the standard deviation of the factor effect, θ_i. It is called pseudo standard error (PSE) and is defined as

$$\text{PSE} = 1.5 \; \text{Median}_{(|\theta_i| < 2.5s_0)}|\theta_i|,$$

where the median is computed among the $|\theta_i|$ with $|\theta_i| < 2.5s_0$ and $s_0 = 1.5 \; \text{Median}|\theta_i|$.

An effect θ_i is declared significant if $|\theta_i/\text{PSE}|$ exceeds the critical values, which can be found in Wu and Hamada (2000), reproduced from Ye and Hamada (2000). In this example, factor A is found to be significant.

Using jackknife replicates for each experimental run, we obtain a variance estimate $\widehat{V}_{ja}(\eta_2)$ of the performance measure as per Equation (1). For the first run, the primary responses are (14.812, 14.774, 14.772, 14.794, 14.860, 14.914) (see Wu and Hamada (2000)). By deleting one observation at a time, we create 6 jackknife samples of size 5. The performance measures, $\log_e(s_i^2)$, for these jackknife samples are (−5.5537, −5.7338, −5.7518, −5.6052, −5.6719, −6.6432). Then the jackknife variance estimate of

TABLE 3. Analysis of Variance Table for η_2

Source	d.f.	MS	F
A	1	58.8135	154.45
B	1	0.0245	0.06
AB	1	0.7321	1.92
C	1	0.0236	0.06
AC	1	0.1829	0.48
BC	1	0.4394	1.15
ABC	1	0.4479	1.18
D	1	1.5961	4.19
AD	1	0.0000	0.00
BD	1	0.3721	0.98
ABD	1	0.0295	0.08
CD	1	1.3535	3.55
ACD	1	0.9758	2.56
BCD	1	0.3947	1.04
ABCD	1	0.0472	0.12
$\widehat{V}_{pja}$	80	0.3808	

$\log_e(s_i^2)$ for run 1 is computed as 0.6904. Similarly, we compute the jackknife variance estimate for each run, and they are given in the last column of Table 2. The pooled jackknife variance estimate $\widehat{V}_{pja} = 0.3808$ by Equation (2), which is then used to test the factor effects. The ANOVA table for η_2 is given in Table 3.

We find that factors A and D are significant at the 5% level. The analysis based on Lenth's method finds only A being significant. Note that the effect for A is extremely large compared with other effects. Almost any method will declare it to be significant. The smaller effect D was not judged significant by Lenth's method.

Example 2

This example is from Taguchi (1986, p. 127). An experiment was conducted to identify the factors that have strong effects on the wear on a slider pump. Five factors (A, B, C, D, and E) and two interaction effects (AB and AC) were thought to influence the wear, and an orthogonal array experiment $L_8(2^7)$ with $m = 8$ replicates was performed. The data consist of wear (in microns) at eight points on the slider of a pump, and the goal is to reduce both the mean and variation of the wear. According to Taguchi, the SN ratio of smaller-the-better type, η_3 (see Table 1), is an appropriate choice for a performance measure. The design matrix and the values of η_3 for this experiment are given in Table 4. The corresponding ANOVA table is given in Table 5.

TABLE 4. Design Matrix, Performance Measure, and the Jackknife Variance Estimates for Example 2

| Run | Factors | | | | | η_3 | $\widehat{V}_{ja}(\eta_3)$ |
	A	B	C	D	E		
1	−	−	−	−	−	−21.8717	1.8395
2	−	−	+	+	+	−20.6023	5.6720
3	−	+	−	+	+	−14.7712	4.9053
4	−	+	+	−	−	−16.1278	1.3237
5	+	−	−	−	+	−24.1539	7.0389
6	+	−	+	+	−	−21.7136	9.9465
7	+	+	−	+	−	−22.9584	2.6745
8	+	+	+	−	+	−23.2710	7.1220
	Pooled						5.0663

In Taguchi's (1986) analysis, the sum of squares due to the small effects C, AC, and E are pooled to obtain the sum of squares of error, which is given as Error in Table 5. The contribution percentages ($\rho\%$) of the remaining effects are then calculated. The effects of A, B, and AB are found significant by examining the magnitude of their contribution percentages.

Another approach is to use the traditional analysis where the error sum of squares is constructed by pooling the small effects after a visual inspection of the normal probability plot. Here it has 3 degrees of freedom, and the F tests confirm that A, B, and AB are significant at the 5% level. Analysis based on Lenth's method indicates that only factor A is significant at that level.

TABLE 5. Analysis of Variance Table for η_3

Source	d.f.	SS	MS	F_{pooled}	$\rho \%$	F_{jack}
A	1	43.82	43.82	115.30	52.90	8.65
B	1	15.72	15.72	41.40	18.70	3.10
AB	1	17.81	17.81	46.90	21.20	3.51
C	1	0.52	0.52	1.37	0.00	0.10
AC	1	0.61	0.61	1.60	0.00	0.12
D	1	3.61	3.61	9.50	3.90	0.71
E	1	0.00	0.00	0.00	0.00	0.00
(Error)	(3)	1.13	0.38	−	3.30	
Total	7	82.09			100.0	

Using the jackknife replicates for each experiment, we obtain jackknife variance $\widehat{V}_{ja}(\eta_3)$ (see Table 4). The pooled jackknife variance estimate is $\widehat{V}_{pja} = 5.06$. This is used to test the significance of the factor effects in the ANOVA table (see Table 5). We find that only the main effect A is significant at the 5% level since our jackknife error variance estimate for η_3 is larger than the estimate obtained by pooling the 3 smaller effects.

Performance of Jackknife

Adapting methods from unreplicated factorial experiments to analyze performance measures requires the pooling of sum of squares for smaller effects, as indicated in Examples 1 and 2. However, tests based on jackknife variance estimate make use of the information from the primary response variables and, thus, is a better approach.

Consistency and Relative Bias

It is well known that the jackknife variance estimator is consistent under very general conditions (Shao and Tu (1995)). In our context, the number of replicates is typically small at each design setting, and the jackknife variance estimate may be biased. However, this bias can be corrected easily if the degree of bias is known. In this section, we investigate the relative bias by simulation.

The jackknife variance estimator for the performance measure η is strongly consistent if

$$\frac{\widehat{V}_{ja}}{\sigma_n^2(\eta)} \underset{\text{a.s.}}{\to} 1$$

as $n \to \infty$, where $\sigma_n^2(\eta)$ is the variance of η based on n observations and a.s. denotes almost sure convergence (Shao and Tu (1995, p. 25)).

We postulate a simple linear model for a 2^3 full factorial design with all three factors having significant effects:

$$y_{ij} = \mu + \alpha A + \beta B + \gamma C + e_{ij}, \qquad (3)$$

with $\mu = 10, \alpha = \beta = \gamma = 1$, $e_{ij} \sim N(0, 0.3^2)$, and A, B, C take values of ± 1 depending on the levels of the three factors. The chosen values of α, β, and γ do not play any role in the consistency of the variance estimators of the performance measures.

For each experimental run, we generate m replicates based on the above model and compute all five performance measures given in Table 1. We construct m jackknife samples from each run and compute the jackknife variance for each of the perfor-

mance measures. This procedure is repeated 10,000 times. The average of the jackknife variance estimates over 10,000 simulations is computed and denoted as $\widehat{V}_{ja}(\eta)$. The variances of performance measures based on 10,000 repetitions are denoted as $\widehat{V}(\eta)$.

Let us define the relative bias as

$$\widehat{R}(\eta) = \frac{\widehat{V}_{ja}(\eta)}{\widehat{V}(\eta)}. \qquad (4)$$

If the jackknife variance estimator provides a sensible error variance estimate, we should have $\widehat{R}(\eta)$ close to 1. The jackknife variance estimator overestimates when $\widehat{R}(\eta) > 1$, and it underestimates when $\widehat{R}(\eta) < 1$.

For each performance measure, simulation studies were carried out for $m = 3, 4, 5, 6, 10, 20$, and 50. The results for η_1, η_3, and η_5 (not shown here) indicate that relative bias is within the 3% range, which implies that jackknife variance estimate is nearly unbiased. Simulation results for η_2 and η_4 are given in Tables 6 and 7.

Tables 6 and 7 indicate that some adjustments are necessary for small m for η_2 and η_4. For example, when $m = 6$, the relative bias of η_2 is 1.55. Hence, the jackknife variance estimator should be divided by 1.55 when $m = 6$ before it is used to compute the F statistic. For small m, the relative bias can be large and the adjustment is crucial. When m is large ($m \geq 20$), an adjustment is not crucial.

To verify that the adjustment factor is not sensitive to σ^2 chosen in the model of Equation (3), we repeat the simulation study with $\sigma = 0.3, 0.6, 1, 1.5$, and 2 for $m = 3, 4, 5, 6, 10, 20$, and 50. We found that the ratio $\widehat{R}(\eta)$ does not depend on the value of σ and takes very similar values among the experimental runs. For brevity, we only present the ratio for pooled variance in Table 8 for η_2 and η_4.

Continuing Example 1 and incorporating the adjustment factor 1.55 for $m = 6$ for η_2, the jackknife variance estimate is reduced from 0.3808 to 0.2457. This results in interaction effect CD being significant at the 5% level.

For the performance measures η_1, η_3, and η_5, the jackknife variance estimates are approximately unbiased and no adjustments are required. For η_2 and η_4, an adjustment factor needs to be used to estimate the jackknife variance as indicated in Tables 6 and 7. For any other performance measure not listed in

TABLE 6. Comparison of $\widehat{R}(\eta)$ for η_2

Run	$m=3$	$m=4$	$m=5$	$m=6$	$m=10$	$m=20$	$m=50$
1	3.45	2.12	1.78	1.56	1.26	1.15	1.01
2	3.60	2.14	1.69	1.62	1.27	1.15	1.08
3	3.67	2.12	1.81	1.50	1.27	1.11	1.02
4	3.55	2.15	1.69	1.58	1.26	1.12	1.04
5	3.54	2.19	1.75	1.61	1.27	1.14	1.07
6	3.57	2.14	1.64	1.50	1.26	1.14	1.05
7	3.58	2.03	1.71	1.55	1.25	1.11	1.05
8	3.46	2.18	1.78	1.53	1.29	1.08	1.05
Pooled	3.55	2.13	1.73	1.55	1.27	1.12	1.05

TABLE 7. Comparison of $\widehat{R}(\eta)$ for η_4

Run	$m=3$	$m=4$	$m=5$	$m=6$	$m=10$	$m=20$	$m=50$
1	3.62	2.21	1.70	1.57	1.23	1.15	1.03
2	3.62	2.18	1.71	1.54	1.27	1.07	1.05
3	3.72	2.13	1.65	1.57	1.30	1.10	1.04
4	3.36	2.18	1.75	1.51	1.31	1.09	1.02
5	3.58	2.20	1.68	1.47	1.28	1.12	1.06
6	3.57	2.28	1.66	1.55	1.29	1.11	1.03
7	3.43	2.15	1.75	1.54	1.27	1.11	1.04
8	3.55	2.11	1.74	1.50	1.23	1.08	1.07
Pooled	3.55	2.18	1.71	1.53	1.27	1.10	1.04

Table 1, simulation studies are recommended to decide whether an adjustment factor is necessary.

Comparison of Jackknife Method with Lenth's Method

Among the methods designed for unreplicated factorial experiments, Lenth's method is most preferable due to its efficiency and simplicity (Hamada and Balakrishnan (1998)). Hence, Lenth's method is commonly used for analyzing performance measures (Wu and Hamada (2000)), and we now compare it with the jackknife method. Other commonly used methods for analyzing the performance measures are normal probability plots and the pooling method (see Examples 1 and 2). These methods are subjective in nature and require human intervention for each simulation run. Thus, they are not included here in the comparisons.

Comparison for Performance Measure η_1

We compare the jackknife method to Lenth's method for η_1, where the true variance is known. Let us consider a 2^3 factorial experiment with factors A, B, and C. Let y_{ij} be the response for the jth replicate of the ith run, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$, which is modeled as

$$y_{ij} = 10 + 0.2A + 0.05B + 0.1C$$
$$+ 0.1AB + 0.075AC + e_{ij}, \qquad (5)$$

where A, B, and C take values of ± 1 depending on the levels, the e_{ij} are normally distributed with mean 0, and $\sigma = 0.5$, and AB and AC represent the values of the variables associated with the interactions. The coefficients in Equation (5) are chosen to represent a range of factor effects. This choice enables us to study the sensitivity of the two methods in different

TABLE 8. Sensitivity of $\widehat{R}(\eta)$ for Different White Noise (σ) for η_2 and η_4

| | η_2 | | | | | η_4 | | | | |
σ	0.3	0.6	1	1.5	2	0.3	0.6	1	1.5	2
$m = 3$	3.55	3.59	3.55	3.57	3.53	3.55	3.53	3.57	3.48	3.45
$m = 4$	2.13	2.12	2.11	2.14	2.11	2.18	2.17	2.13	2.10	2.08
$m = 5$	1.73	1.74	1.75	1.74	1.75	1.71	1.75	1.76	1.72	1.67
$m = 6$	1.55	1.58	1.54	1.59	1.60	1.51	1.60	1.56	1.54	1.53
$m = 10$	1.25	1.28	1.22	1.25	1.26	1.23	1.23	1.30	1.24	1.26
$m = 20$	1.16	1.12	1.13	1.11	1.10	1.13	1.10	1.07	1.10	1.12
$m = 50$	1.06	1.04	1.03	1.03	1.08	1.04	1.08	1.07	1.05	1.03

situations. $\text{Var}(\bar{y}_i)$ for the ith run is σ^2/m, and the variance of an estimated effect is $\sigma^2/2m$. The jackknife error variance estimate for each run, $\widehat{V}_{ja}(\bar{y}_i)$, is unbiased for $\text{Var}(\bar{y}_i) = \sigma^2/m$. Thus, $\widehat{V}_{pja}/2$ is an unbiased estimate of $\sigma^2/2m$. Because the expected value of PSE is complicated, we carry out 1,000 simulations according to model of Equation (5) to approximate its expected value. For each simulation, we estimate PSE and $\widehat{V}_{pja}$ and perform 5% significance tests based on both jackknife and Lenth's methods. Based on the 1,000 simulations, we compute the percentage of times each effect is declared significant at the 5% level of significance. The results are summarized in Table 9.

Table 9 shows that Lenth's method overestimates the standard error of the factor effects for this performance measure, but the jackknife variance estimate is almost the same as the true variance. For instance, when $m = 6$, the true standard deviation of a factor effect is 0.1443. The jackknife variance estimate is 0.1436 while the estimate by Lenth's method is much larger at 0.2467. This is the reason why Lenth's method is more conservative in picking up the significant effects. When $m = 6$, factor A is found significant 76.6% of the time by the jackknife method but only 26.2% by Lenth's method. Factor effects not in the model (say BC) are found significant at 5% and 1.1% of the times by jackknife and Lenth's methods, respectively, compared with the 5% specified by the significance level. For larger sample sizes, the jackknife performs superior to Lenth's method.

TABLE 9. Percent of Times a Factor is Significant for the Jackknife Method and Lenth's Method for η_1

| Factor | True effect | $m = 3$ | | $m = 4$ | | $m = 5$ | | $m = 6$ | |
		Jack	Lenth	Jack	Lenth	Jack	Lenth	Jack	Lenth
A	0.200	45.5	20.6	57.7	21.5	70.8	23.1	76.6	26.2
B	0.050	6.5	2.3	9.0	2.3	9.5	2.6	11.1	1.4
AB	0.100	15.1	4.8	20.2	6.4	20.9	5.5	29.2	4.7
C	0.100	15.4	6.0	20.2	5.5	22.5	4.9	25.6	5.7
AC	0.075	11.4	4.7	13.6	3.7	15.4	3.4	19.3	2.3
BC	0.000	4.8	1.4	4.3	1.8	4.8	1.5	5.0	1.1
ABC	0.000	4.8	1.8	4.6	0.9	4.7	1.9	4.4	0.8
Standard error of factor effect		0.2033	0.2891	0.1750	0.2643	0.1577	0.2534	0.1436	0.2467
True standard error of factor effect		0.2041	0.2041	0.1768	0.1768	0.1581	0.1581	0.1443	0.1443

TABLE 10. Percent of Times a Factor is Significant for the Jackknife Method and Lenth's Method for η_2

Factor	True effect	$m = 3$		$m = 4$		$m = 5$		$m = 6$	
		Jack	Lenth	Jack	Lenth	Jack	Lenth	Jack	Lenth
A	1	56.5	30.0	79.1	42.3	90.8	58. 5	96.5	66.2
B	0	7.2	2.1	5.9	1.5	6.7	2.7	6.5	2.7
AB	0	6.1	1.2	6.7	1.9	6.0	2.6	6.5	2.6
C	1	54.2	29.5	79.9	40.2	91.7	58. 7	96.9	65.7
AC	0	5.6	1.7	7.1	2.7	6.2	3.6	5.3	2.4
BC	0	6.3	1.3	6.4	1.5	6.3	1.9	6.7	2.6
ABC	0	6.1	2.3	5.7	1.5	5.4	2.1	5.5	2.3

Comparison for Performance Measure η_2

Our next simulation compares the jackknife method to Lenth's method for performance measure η_2. We construct a linear model for a 2^3 factorial experiment with performance measure η_2 as response, with factors A and C having significant effects on η_2, and the other factor effects, including interactions, being negligible. That is, we use a linear model,

$$y_{ij} = \mu_i + e_{ij}, \qquad (6)$$

for the primary response of the ith run and the jth replicate. Suppose that we are interested in the effect of the factors on the performance measure $\log_e(s_i^2)$. Let μ represents the overall mean, α the effect of factor A, and β the effect of factor C. A suitable model for the performance measure is to let $\sigma_i^2 = \text{var}(e_{ij})$ such that

$$\log\left(\sigma_i^2\right) = \mu + \alpha A + \beta C + \epsilon_i, \qquad (7)$$

where A and C take values of ± 1 depending on their levels in the ith run, and ϵ_i is white noise such that $\epsilon_i \sim N(0, \sigma^2)$.

In our simulation, we set $\sigma = 0.05$, $\alpha = 1$, $\beta = 1$, and $\mu = 1$. With these values, σ_i^2 will change substantially at different design setting. In addition, $\sigma = 0.05$ adds some external variation from one experiment to another. We first generate the ϵ_i and then determine a value of σ_i^2 according to Equation (7). We then obtain m primary responses according to Equation (6) by generating the e_{ij} from $N(0, \sigma_i^2)$ for the ith run. The performance measure $\log_e(s_i^2)$ and its jackknife variance estimate are computed thereafter. We have used the adjustment factor to estimate the jackknife variance as per Table 6. We then test the significance of factor effects at a 5% level based on the jackknife variance estimate and Lenth's method. The whole process is repeated 1,000 times. The percentage of times each factor is significant at a 5% level is shown in Table 10.

According to Table 10, Lenth's method has lower power in detecting the effect of factors A and C than the jackknife method. For instance, when $m = 6$, factor A is found significant 96.5% of the time by using the jackknife method while it is judged significant only 66.2% of the time by Lenth's method. In addition, the factor effects not in the model are (falsely) significant around 6% of the time, which is in line with the 5% significance level, whereas the significance level is around 2.5% by Lenth's method. We note that the adjusted jackknife variance is used to identify the significant factor effects. If one does not adjust the jackknife variance appropriately, the conclusions can be misleading.

Discussion and Conclusions

We have demonstrated that the jackknife method is useful in analyzing performance measures in factorial experiments. However, it can be used to analyze performance measures in any other designs where replicates are available on the primary response variable for each design setting. When replications are available, it is better to use this information to estimate the variance of the performance measures of interest. The proposed jackknife method is simple to use and efficient in estimating the error variance. It gives an opportunity to obtain an estimate of the within-run variance, which leads to a more reliable test for the factorial effects. Other usual methods (including Lenth's) use variance estimators based on between-effects variation. In fact, Lenth's method is devised for the analysis of unreplicated factorials when there is effect sparsity. If the ratio of number of

active factors to the number of runs is large, then the commonly used methods for analyzing unreplicated factorials are not very effective. Increasing the number of replications in each run improves the power of detecting significant effects in the proposed method while it has less impact in the methods commonly used. We feel that the proposed method can be a great tool whenever one is analyzing performance measures such as the average or $\log_e(s^2)$.

Appendix

Let $y = (y_1, y_2, \ldots, y_m)$ be a sample of size m. Suppose that the parameter ξ is estimated by $\widehat{\xi} = c(y)$, with c being a continuous function of y. We wish to estimate the bias and variance of $\widehat{\xi}$. The jackknife focuses on the construction of pseudo samples by leaving out one observation at a time. The ith jackknife sample consists of data with the ith observation removed,

$$y_{(i)} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m),$$
$$i = 1, 2, \ldots, m. \qquad (8)$$

Let $\widehat{\xi}_{(i)} = c(y_{(i)})$ be the ith jackknife replication of the estimate $\widehat{\xi}$. The estimates of the bias and variance of this estimator are

$$\widehat{\text{bias}}_{ja} = (m-1)\left(\widehat{\xi}_{(\cdot)} - \widehat{\xi}\right),$$
$$\widehat{V}_{ja} = \frac{m-1}{m}\sum_{i=1}^{m}\left(\widehat{\xi}_{(i)} - \widehat{\xi}_{(\cdot)}\right)^2, \qquad (9)$$

where $\widehat{\xi}_{(\cdot)} = \sum_{i=1}^{m}\widehat{\xi}_{(i)}/m$ (Efron and Tibshirani (1993)).

As it is clear from Equation (8) that two jackknife samples differ only by two data points, and, hence, are very similar. Thus, the simple standard deviation of the jackknife replications does not represent the standard error of the original estimator. Because of this, the total variation in jackknife replicates provides an approximate estimate of the variance of $\widehat{\xi}$ as given in Equation (9). For a class of estimators, it can be shown that $\widehat{V}_{ja}$ is asymptotically unbiased as $m \to \infty$ (Theorem 2.1, Shao and Tu (1995, p. 25)). One may refer to Efron and Tibshirani (1993) and Shao and Tu (1995) for a more detailed discussion.

In many applications, the unknown parameter ξ is a smooth function of other easily estimable parameters such as the mean and the second moment. Then, the corresponding estimator is a function of the sample moments. This includes all the perfor-

mance measures listed in Table 1, for example. Let $y_{i1}, y_{i2}, \ldots, y_{im}$ be a set of independent and identically distributed observations. For simplicity, let $\xi_1 = g(\mu_1)$, and $\widehat{\xi}_1 = g(\bar{y}_1)$, with $\mu_1 = E(y_{11})$ and $\bar{y}_1 = (1/m)\sum_{j=1}^{m}y_{1j}$. Let $s_1^2 = [1/(m-1)]\sum_{j=1}^{m}(y_{1j} - \bar{y}_1)^2$. Approximately, we have,

$$\widehat{\xi}_1 - \xi_1 = g'(\mu_1)\left[\bar{y}_1 - \mu_1\right] + o_p\left(\bar{y}_1 - \mu_1\right), \qquad (10)$$

where $o_p(\cdot)$ is a negligible quantity when m is large, and the jackknife variance estimate of $\widehat{\xi}_1$ is approximately

$$\widehat{V}_{ja} = [g'(\mu_1)]^2\sum_{i=1}^{m}\left(y_{1(i)} - \bar{y}_1\right)^2 = \frac{1}{m}[g'(\mu_1)]^2 s_1^2. \qquad (11)$$

Under normality, $\bar{y}_1 - \mu_1$ and s_1^2 are independent. Without the normality assumption, they are approximately independent when m becomes large. Hence,

$$\frac{(\widehat{\xi}_1 - \xi_1)^2}{\widehat{V}_{ja}} \approx \frac{m(\bar{y}_1 - \mu_1)^2}{s_1^2},$$

which has an F distribution with degrees of freedom 1 and $m-1$. Applied to our performance measure, the F statistic is approximately the ratio of two independent quadratic forms. Thus, the F distribution with 1 and $n(m-1)$ degrees of freedom is a reasonable choice.

When the distribution of individual observations is nearly symmetric, the normal approximation is very good even when the sample size is not large. However, the Taylor expansion in Equation (10) may not be very precise for the typical size of m in experimental designs. This may result in some bias in $\widehat{V}_{ja}$ and some adjustment may be required, as seen in Tables 6 and 7.

Acknowledgment

References

Box, G. E. P. (1988). "Signal-to-Noise Ratios, Performance Criteria, and Transformation" (with discussion). *Technometrics* 30, pp. 1–38.

Box, G. E. P. and Meyer, R. D. (1986). "Dispersion Effects from Fractional Designs". *Technometrics* 28, pp. 19–27.

DANIEL, C. (1959). "Use of Half-Normal Plots in Interpreting Factorial Two Level Experiments". *Technometrics* 1, pp. 311–340.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

HAMADA, M. and BALAKRISHNAN, N. (1998). "Analyzing Unreplicated Factorial Experiments: A Review with Some New Proposals". *Statistica Sinica* 8, pp. 1–41.

LENTH, R. V. (1989). "Quick and Easy Analysis of Unreplicated Factorials". *Technometrics* 31, pp. 469–473.

LEON, R. V.; SHOEMAKER, A. C.; and KACKER, R. N. (1987). "Performance Measures Independent of Adjustment" (with discussion). *Technometrics* 29, pp. 253–285.

PHADKE, S. M. (1989). *Quality Engineering Using Robust Design.* Englewood Cliffs, New Jersey: Prentice Hall.

SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap.* Springer, New York.

TAGUCHI, G. (1986). *Introduction to Quality Engineering: Designing Quality into Products and Processes.* Asian Productivity Organization, Tokyo.

WU, J. C. F. and HAMADA, M. (2000). *Experiments: Planning, Analysis and Parameter Design Optimization.* Wiley, New York.

YE, K. and HAMADA, M. (2000). "Critical Values of the Lenth Method for Un-Replicated Factorial Designs". *Journal of Quality Technology* 32, pp. 57–66.

∼