Quality Engineering, 12(3), 463–469 (2000)

A p CHART FOR MONITORING CAPABILITY USING SENSITIVITY DATA

Bruce E. Ankenman and William R. McDaniel
Department of Industrial Engineering and Management Science
Northwestern University
Evanston, Illinois 60208

Key Words

Power of the test; Statistical process control; Destructive testing; Binomial distribution.

Introduction

A special type of data, called sensitivity data, presents a unique set of problems for those wishing to monitor the performance of a process. Sensitivity data refer to data collected as a pass or fail of a sample at a certain level or intensity of exposure to a control variable. Dixon and Mood (1) originally presented this type of data in reference to testing explosives at different drop heights. The difficulty with the data arises because a continuous variable, such as the minimum drop height at which the explosive will detonate, is the measure of interest but cannot be measured directly because of the destructive nature of the testing.

Modern quality practices present further problems with this type of testing because customers expect statistical assurance that the product they receive will not be defective. One common measure, process capability or C_{pk}, measures the relative distance from the process mean to the nearest specification limit (2). C_{pk} is measured in units equal to three times the process standard deviation, so $C_{pk} = 1$ indicates that the process mean is three standard deviations from the closest specification limit. Similarly, $C_{pk} = 2$ indicates the mean is six standard deviations from the closest limit. When a Normal distribution is assumed, $C_{pk} = 1$ implies that the 0.00135th quantile of the process distribution lies on the nearest specification limit and, thus, no more than 3 in 1000 (2×0.00135) products are out of specification. This common quality measure has been used extensively in practice because it is applicable to many situations and gives a good indication of the acceptability of incoming product to customers.

Customers also expect the capability to be monitored over time to be sure the process quality does not drift. Monitoring low quantiles of a distribution is expensive because it requires a high number of samples. Thus, capability for continuous variables is usually monitored by assuming normally distributed data and then using $\overline{X}$ and R control charts for monitoring changes in the process mean or standard deviation. If either chart gives an out-of-control signal, the process is investigated, a cause is assigned, and appropriate adjustments are made. When only sensitivity data are available, these types of chart are not applicable because the variable of interest cannot be directly measured. For example, when collecting sensitivity data on a threshold-failure height, the drop heights must be chosen by the experimenter and each test results in a pass or fail rather than a numerical reading. Thus, for a small sample, the estimate of the mean failure height may be biased toward the heights chosen for testing. In addition, the range is completely dependent on the range between the test heights chosen. Because a range chart for

Copyright © 2000 by Marcel Dekker, Inc. www.dekker.com

sensitivity data measures the distance between experimental points chosen by the user instead of the sample range of a continuous variable, an R chart is unusable in this situation. These problems mean that the $\overline{X}$ and R charts give the user little relevant information, so an alternative needs to be found.

In the next section, the example that led us to this research is presented with some necessary notation. Then, we provide theory and methods for constructing a p chart that can be used as a surrogate for monitoring process capability with sensitivity data.

Motivation

This research was inspired by a problem encountered in a plastic blow-molding process for the manufacture of plastic fluid containers. The measure of interest was the threshold drop height at which a filled bottle would fail, where failure is indicated by any leaks in the bottle. This is an obvious example of sensitivity data because each bottle can only be dropped from a certain height and the only information available is whether or not the bottle failed. Once a bottle is dropped, it cannot be tested or used again, whether or not it fails. The customers for the bottles had a lower specification limit for the threshold failure height and wanted both a C_{pk} for the process and a control chart for monitoring process capability over time. Thus, there was a need to develop a control chart for monitoring this process using the sensitivity data.

As discussed earlier, knowing certain quantiles of the distribution is equivalent to knowing the C_{pk}, so it seems natural to devise a way to monitor the quantiles of the distribution. In addition, the pass–fail nature of the testing indicates that a p chart might be appropriate. A standard p chart is designed to monitor the proportion of a sample that fails a test, which is exactly how sensitivity data must be taken. Thus, devising a p chart to monitor a quantile of the distribution is an effective way to monitor changes in the process capability.

Suppose that H_q is the qth quantile of the threshold failure height distribution of the process. If we drop a number of bottles from height H_q, we can expect $(q)(100\%)$ of them to break. For a given H_q, a p chart can be constructed to track the proportion of failures at that height for a given sample size. If the proportion of failures deviates significantly from q, then the process has gone out of control (i.e., the qth quantile has changed and, in turn, the process capability has changed). In the next section, the context of the blow-molding example is used to present details for the construction of a p chart to monitor process capability. Although the work that we present concentrates on the situation where there is a lower specification on the sensitivity data, the method could be adapted to an upper limit or to a nearest specification limit with some straightforward modifications.

Constructing a p Chart for Monitoring a Quantile

When using a p chart to monitor the process capability, the problem to address is what quantile to monitor. Because most customers require at least a $C_{pk} = 1$, we would like to use the 0.00135th quantile. However, to monitor this quantile would require a prohibitively high sample size. Because about 1 item in 1000 would fail at $H_{0.00135}$, almost no failures would occur and too many products would need to be tested. Instead, a set of typical control chart subgroup sizes (5–15) is considered. The statistical power of the p chart is used to choose the best quantile to monitor.

Before discussing the selection of q, we assume that once q is selected, an estimate of H_q can be determined by some initial testing procedure that may use many more samples than the subgroup size of the p chart. Dixon and Mood (1) and Little and Thomas (3) provide an "up-and-down" method for determining the mean and standard deviation of a distribution from sensitivity data. The "up-and-down" method begins by recording a pass or fail from an initial height. The height at which the next sample is dropped is lower if the initial sample fails and higher if the initial sample passes. The most recent success or failure determines the heights for all subsequent tests. The result is a series of tests that tend toward the mean of the threshold-failure height distribution. This method has been adapted for many ASTM standards (e.g., Ref. 4). By assuming a distribution for the process, the quantiles can be determined. Often the Normal or log-normal distribution is assumed, but in some cases, other distributions may be applicable. Wu (5,6) also presents a method for estimating a specified quantile of a distribution using sensitivity data. Thus, methods are currently available for obtaining an initial estimate of H_q.

Using the blow-molding example for ease of discussion, suppose that all bottles in a subgroup are dropped from the height H_q; then the true proportion of bottles that fail should be q. We will use a p chart, centered at q, to monitor the actual proportion p of the bottles that fail at height H_q and determine if the actual proportion differs significantly from q. Although an estimate of H_q can be found for any q, only certain quantiles are easily monitored due to the relatively small size of the subgroups used in most control chart applications. We will choose the value of q to maximize the power of the p chart with a specified false alarm rate of α. The number of bottles that will fail in a given subgroup will follow a binomial distribution if we assume that all the threshold-failure heights are independent and identically distributed. If the qth quantile is to be monitored, then a p chart with an upper control limit, UCL, will have a false alarm rate of α if

$$\alpha = \sum_{i=\mathrm{UCL}}^{n} \binom{n}{i} q^i (1 - q)^{n-i}, \tag{1}$$

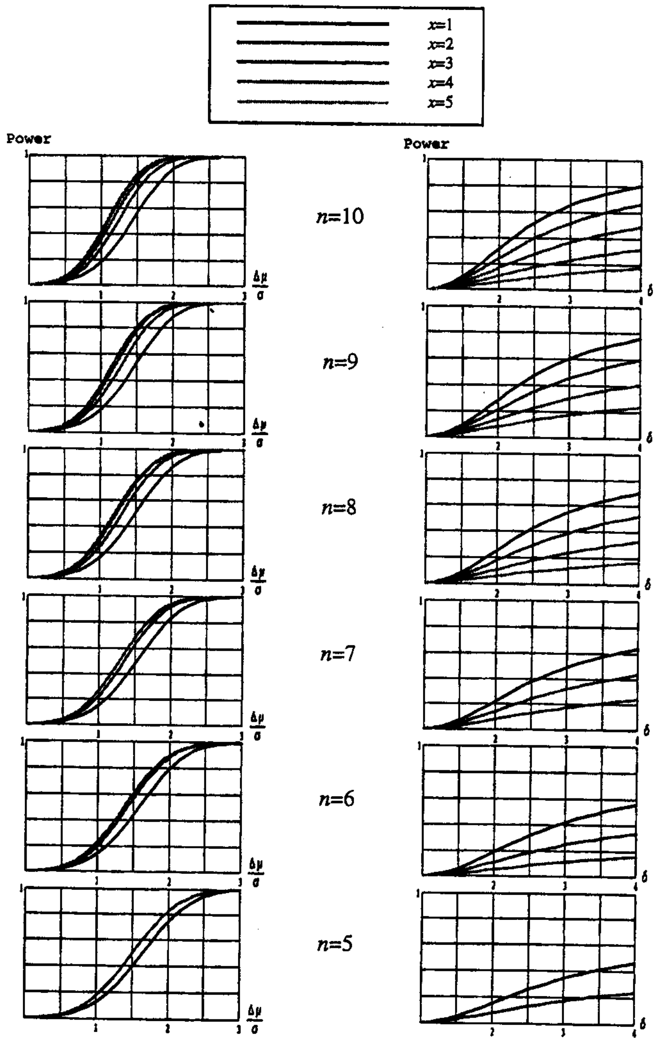
where n is the size of each subgroup. The upper control limit will be set to some x/n, where x is an integer less than n. Thus, if x or fewer of the n products in the subgroup fail when dropped from height H_q, there is no out-of-control signal. If $x + 1$ or more products fail, there is an out-of-control signal. Due to the discrete nature of the data, any control limit between x/n and $(x + 1)/n$ would have the same effect as this control limit.

For processes with a lower specification such as a threshold-failure height, the monitored quantile, H_q, will always be below the median because the procedure should be sensitive to an increase in the number of failures. Keeping H_q below the median means that, on average, there will be fewer failures than successes, so an increase in the number of failures will be signaled more clearly. Also, the quantile of interest for the C_{pk} is below the median, so we will be measuring a quantile relatively close to it. Therefore, we will consider all values of x up to $\frac{1}{2}n$ for each subgroup size.

Once a value for n is chosen based on the difficulty and cost of the testing, the appropriate control limit x/n can be determined by comparing the power curves for different choices of x. For each x, we can substitute UCL $= x/n$ into Eq. (1) and then obtain the value q numerically for a selected α.

Then, the p chart is constructed from a series of subgroups of size n selected from the process on a regular basis. All products are tested at H_q (to be estimated from testing after x is chosen). If more than x failures occur in any subgroup, the process is said to be out of control.

As mentioned previously, we use the power of the statistical test to determine which quantile, H_q, to monitor for each subgroup size. The power of the test measures the probability that the test will detect a given change in conditions. It is usually displayed as a curve of detection probability versus change in the process mean assuming that the shape of the distribution remains unchanged. The units for the change in mean are such that 1 unit = 1 standard deviation, so the results are displayed in terms of a standard Normal distribution. Power curves for changes in the mean for $n = 5$ to 10 and $x = \frac{1}{2}n$ are shown in Figure 1a. Power curves for changes in the mean for $n = 10$ to 15 and $x = 1$ to $\frac{1}{2}n$ are shown in Figure 2a.

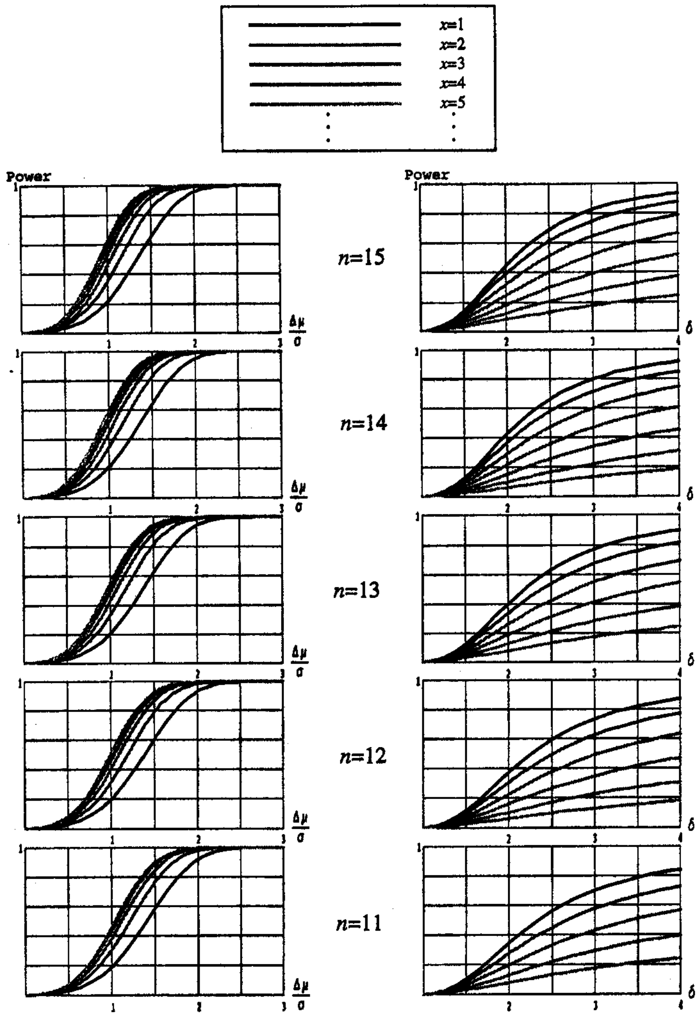


(a) Changes in μ (b) Changes in σ

Figure 1. Power curves for $n = 5$ to 10 and $x = 1$ to $\frac{1}{2}n$.

As an example of the calculations made to create these power curves, consider the case where the sample size is $n = 10$ and the maximum number of failures before an out-of-control condition is signaled is $x = 1$, as shown by the bold line on the top graph of Figure 1a. From Eq. (1) and using $\alpha = 0.0027$, we can calculate that $q = 0.0079$. We assume that the threshold-failure height distribution initially has a Normal distribution with mean μ and standard deviation σ. If the mean drops by $\Delta\mu$, the new probability that any one sample will fail is now

$$q^* = \Phi\left(\frac{H_q - (\mu - \Delta\mu)}{\sigma}\right) = \Phi\left(\frac{H_q - \mu}{\sigma} + \frac{\Delta\mu}{\sigma}\right), \tag{2}$$



(a) Changes in μ (b) Changes in σ

Figure 2. Power curves for $n = 11$ to 15 and $x = 1$ to $\frac{1}{2}n$.

where Φ indicates the cumulative distribution function (CDF) of the standard Normal distribution. Using $\Phi^{-1}(q) = (H_q - \mu)/\sigma = -2.41$, if the mean drops by one standard deviation, then $\Delta\mu = \sigma$, $[(H_q - \mu)/\sigma + \Delta\mu/\sigma] = -1.41$; thus, $q^* = 0.08$. From this, the binomial probability, ρ, that more than 1 of 10 samples will fail can be calculated using the CDF of the binomial distribution,

$$\rho = 1 - \sum_{i=0}^{x} \binom{n}{i} (q^*)^i (1 - q^*)^{n-i}, \tag{3}$$

where $x = 1$ and $n = 10$. The probability that the test will detect this out-of-control condition is calculated to be 0.18. This is the power of the test for a change in the mean of one standard deviation. The power curves show the power of the test for changes in the mean up to 3σ. For all sample sizes, the power of the test increases as x is increased to $\frac{1}{2}n$. However, the gains in power for increasing the value of x are relatively small, especially for $x \geq 2$.

The threshold-failure heights are assumed to be normally distributed for the power curves in Figures 1 and 2. Dixon and Mood (1) suggest using a log-normal distribution for

Table 1. Power of the Test

		CHANGE IN μ		CHANGE IN σ				CHANGE IN μ		CHANGE IN σ	
n	x	$\Delta\mu = 1.0\sigma$	$\Delta\mu = 1.5\sigma$	$\delta = 2$	$\delta = 3$	n	x	$\Delta\mu = 1.0\sigma$	$\Delta\mu = 1.5\sigma$	$\delta = 2$	$\delta = 3$
5	*1	0.13	0.40	0.15	0.34	12	1	0.20	0.61	0.38	0.74
	2	0.18	0.50	0.08	0.18		*2	0.32	0.79	0.31	0.63
6	*1	0.14	0.44	0.19	0.42		3	0.41	0.87	0.23	0.49
	2	0.20	0.57	0.11	0.25		4	0.46	0.91	0.16	0.35
	3	0.23	0.61	0.06	0.12		5	0.50	0.92	0.10	0.23
7	*1	0.15	0.48	0.22	0.49		6	0.52	0.93	0.06	0.13
	2	0.23	0.62	0.14	0.32	13	1	0.21	0.63	0.40	0.77
	3	0.27	0.68	0.08	0.18		*2	0.33	0.81	0.34	0.67
8	*1	0.17	0.51	0.26	0.55		3	0.43	0.89	0.26	0.55
	2	0.25	0.67	0.18	0.39		4	0.49	0.92	0.19	0.41
	3	0.30	0.74	0.11	0.24		5	0.53	0.94	0.13	0.28
	4	0.33	0.76	0.06	0.12		6	0.55	0.95	0.08	0.18
9	1	0.18	0.54	0.29	0.61	14	1	0.21	0.64	0.43	0.80
	*2	0.27	0.71	0.21	0.46		*2	0.35	0.83	0.37	0.71
	3	0.33	0.78	0.14	0.30		3	0.45	0.90	0.29	0.60
	4	0.37	0.81	0.08	0.18		4	0.52	0.94	0.22	0.47
10	1	0.18	0.56	0.32	0.66		5	0.56	0.95	0.15	0.34
	*2	0.29	0.74	0.24	0.52		6	0.59	0.96	0.10	0.22
	3	0.36	0.82	0.17	0.37		7	0.60	0.96	0.06	0.13
	4	0.40	0.85	0.11	0.23	15	1	0.22	0.66	0.45	0.82
	5	0.42	0.86	0.06	0.13		2	0.36	0.84	0.40	0.75
11	1	0.19	0.59	0.35	0.70		*3	0.47	0.92	0.32	0.65
	*2	0.30	0.77	0.28	0.58		4	0.54	0.95	0.25	0.52
	3	0.38	0.85	0.20	0.43		5	0.59	0.96	0.18	0.39
	4	0.44	0.88	0.13	0.29		6	0.62	0.97	0.13	0.28
	5	0.46	0.90	0.08	0.18		7	0.64	0.97	0.08	0.18

Note: An asterisk indicates the recommended plan for each subgroup size.

drop height distributions because the drop height is bounded by zero. Some preliminary analysis showed that the results are fairly robust to skewness in the data. In addition, the log-normal distribution has a higher tail probability above the mean than below it. Thus, for the quantiles that we are measuring, the effect of skewness is likely to be small and the Normal assumption only serves to make the power curves more conservative.

In addition to comparing the effect of changes in mean on the power of the procedure, the effect of changes in the standard deviation was also considered. We compare the performance of the different control limits when the standard deviation is inflated by a factor δ, but the mean stays fixed. Assuming a Normal distribution, the probability of detecting an increase of δ in the standard deviation is $q^{**} = \Phi[(H_q - \mu)/\delta\sigma]$. As earlier, we use $\Phi^{-1}(q) = [(H_q - \mu)/\sigma] = -2.41$, so if the standard deviation doubles, then for $\delta = 2$, $[(H_q - \mu)/\delta\sigma] = -1.20$, and thus $q^{**} = 0.114$. The power of the test can be calculated from Eq. (3) to be 0.318.

When changing the mean of a Normal variable, the CDF of a fixed point can take any value between 0 and 1. However, when inflating the standard deviation with a fixed mean, the CDF of a fixed point below the mean cannot exceed 0.5. This is because at least one-half of the probability will always be above the mean. Thus, the power of this test can only reach the limit $\rho = 1 - (0.5)^n \sum_{i=0}^{x} \binom{n}{i}$. The power curves in Figures 1b and 2b compare the various choices for x with respect to changes in the standard deviation. In contrast to the changes in the mean, the power of this test decreases as the value of x is increased. Table 1 chooses a few points from each graph for easy comparison. An asterisk designates the recommended choices for x for each subgroup size. Because the gains in power for changes in the mean are very small above $x = 2$ or 3, we recommend $x = 1$, $x = 2$, or $x = 3$ for all values of n between 5 and 15 so that the power for detecting changes in the standard deviation is as high as possible.

Numerical Example

Take as an example the threshold height of dropped plastic bottles mentioned earlier. In this problem, the customer has a lower specification limit (LSL) for the threshold-failure height of a filled bottle, such that any bottle dropped at a height lower than LSL should not break. Assume that the LSL = 3 ft and the customer wants to maintain a C_{pk} of at least 1 for the process. Now, suppose that the manufacturer has done some preliminary testing on the process and has found that the process has a mean threshold-failure height of 5 ft with an estimated standard deviation of 0.67 ft, giving the process a C_{pk} of $(5 - 3)/3(0.67) = 1.00$. The manufacturer wants to monitor the process three times an hour to look for changes, but the tests are expensive, so only six bottles will be dropped every 20 min. The above-described procedure and the power curves can be used to set up a p chart for monitoring the process.

The possible values for the lower control limit of the process are 0.167, 0.333, and 0.5, corresponding to $x = 1$, 2, and 3 failures out of $n = 6$ bottles tested. Now, we examine the power curves for $n = 6$ samples and find that $x = 3$ is the best for detecting changes in the mean, but it performs very poorly for detecting changes in the standard deviation. On the other hand, $x = 1$ gives almost as much power for detecting changes in the mean and much more power for detecting changes in the standard deviation. Therefore, we chose to use a p chart with a control limit of 0.167, so that if more than one of the six bottles break, the chart will signal an out-of-control condition. For this control limit (again selecting $\alpha = 0.0027$), the value of q is calculated from Eq. (1) to be $q = 0.0137$. Then, using this q value and assuming a Normal distribution, the drop height, H_q, can be calculated from the standard Normal curve. Using $\mu = 5$ and $\sigma = 0.67$, $\Phi^{-1}(0.0137) = -2.21 = (H_q - 5)/0.67$, which gives $H_q =$ 3.52 ft. Therefore, we choose to use a drop height of 3.52 ft, and if more than one of the six bottles tested at this height fails, the process is considered out of control and may no longer meet the C_{pk} criterion. The power of the test for this chart is 0.14 if the mean of the distribution drops to 4.33 ft. The power is 0.18 if the standard deviation doubles to 1.33 ft.

Another common measure for the effectiveness of a control chart is the average run length (ARL) before an out-of-control signal. It is easily calculated because

$$\text{ARL} = \frac{1}{P(\text{a subgroup gives an out of control signal})}.$$

When the process is in control, the ARL measures the average frequency of false alarms and is $1/\alpha$, where α is the probability of type I error that is selected. Because $\alpha =$ 0.0027, we can expect 1 false out-of-control signal every 370 subgroups or every 123 h. When the process goes out of control, the ARL measures the average number of subgroups tested until the out-of-control condition is signaled. If the mean drops to 4.33 ft (one standard deviation), then the ARL = 1/Power of the test = 1/0.14 = 7 subgroups. So we would expect an out-of-control signal in about 2.3 h. Similarly, if the standard deviation doubles to 1.33 ft, then we would expect a signal in about 1/0.18 = 5 subgroups or 1.66 h. These calculations can be used to balance the cost of testing with the cost of producing defective product.

Conclusions

A p chart has been adapted for monitoring the process capability when sensitivity data are all that are available for measuring the quantity of interest. The method uses the same sort of distributional assumptions that most control charts rely on and is fairly robust to changes in the underlying distribution. Instead of trying to adapt the "up-and-down" method (1) to a control chart, this method uses a constant drop height, making testing simpler and doing away with the forced step-size choices. In addition, the method has a low probability of false signals, and by changing subgroup sizes, it can be adapted to many different requirements for the power of the testing. Power curves and recommendations for setting up the p chart for subgroup sizes between 5 and 15 have been provided.

Acknowledgments

The authors would like to thank Sushant Jain and Les Kaffko of Continental Plastic Containers for proposing this problem and for their help and cooperation on the research project.

References

1. Dixon, W. J. and Mood, A. M., A Method for Obtaining and Analyzing Sensitivity Data, *J. Am. Statist. Assoc.*, 60, 967–978 (1948).
2. Devor, R. E., Chang, T., and Sutherland, J. W., *Statistical Quality Design and Control*, MacMillan, New York, 1992.
3. Little, R. E. and Thomas, J. J., Up-and-Down Methodology Applied to Statistically Planned Experiments, *J. Test. Eval.*, 21(1), 14–20 (1993).

4. ASTM D2463, Standard Test Method for Drop Impact Resistance of Blow-Molded Thermoplastic Containers, American Society for Standards and Testing, Philadelphia, 1974.
5. Wu, C. F. J., Efficient Sequential Designs with Binary Data, *J. Am. Statist. Soc.*, 80, 974–984 (1985).
6. Wu, C. F. J., Optimal Design for Percentile Estimation of a Quantal Response Curve, *Optimal Design and Analysis of Experiments*, edited by Y. Dodge, V. V. Fedorov, and H. P. Wynn, Elsevier Science/North-Holland, Amsterdam, 1988.

About the Authors: William R. McDaniel is a doctoral candidate in industrial engineering at Northwestern University. His research interests include experimental design for the latter stages of prototype testing, regression analysis, and simulation.

Bruce E. Ankenman is an assistant professor of industrial engineering at Northwestern University. His research interests are response surface methodology, experimental design, and design and analysis for estimating variance components.