Quality Engineering
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597292

Must a Process Be in Statistical Control before Conducting Designed Experiments?

Søren Bisgaard [a]
[a] Eugene M. Isenberg School of Management, University of Massachusetts Amherst, Amherst, Massachusetts, and Institute for Business and Industrial Statistics, University of Amsterdam, The Netherlands

To cite this Article Bisgaard, Søren(2008) 'Must a Process Be in Statistical Control before Conducting Designed Experiments?', Quality Engineering, 20: 2, 143 — 150
To link to this Article: DOI: 10.1080/08982110701826721
URL: http://dx.doi.org/10.1080/08982110701826721

Quality Engineering, 20:143–150, 2008

ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982110701826721

Taylor & Francis
Taylor & Francis Group

Must a Process Be in Statistical Control before Conducting Designed Experiments?

Søren Bisgaard

Eugene M. Isenberg School of Management, University of Massachusetts Amherst, Amherst, Massachusetts, and Institute for Business and Industrial Statistics, University of Amsterdam, The Netherlands

ABSTRACT Fisher demonstrated three quarters of a century ago that the three key concepts of randomization, blocking, and replication make it possible to conduct experiments on processes that are not necessarily in a state of statistical control. However, even today there persists confusion about whether statistical control is a necessary prerequisite for conducting valid experiments in industry. In this article we revisit and extend Fisher's original argument. Reusing his 1925 examples, we demonstrate that the need for statistical control as a prerequisite for conducting industrial experiments is misconceived. Clarifying this issue may help quality practitioners identify new and wider opportunities for the use of designed experiments in industrial practice.

KEYWORDS analysis of variance, blocking, design of experiments, randomization

INTRODUCTION

It is widely believed among certain groups of quality practitioners that statistical control is a necessary prerequisite for conducting valid experiments. We believe this issue merits careful discussion and subsequent refutation. Our current interest is provoked primarily by repeated exposure to this opinion through Six Sigma consulting. However, this view is not just industrial folklore. It has also been promulgated in scholarly publications. For example, in a recent discussion accompanying an article by Woodall (2000), Ryan (2000) wrote, "it is . . . important to remember that processes must be in a state of statistical control when designed experiments are performed." As we will demonstrate, this is a problematic position. If true, it would render design of experiments impotent as a tool for bringing industrial processes into a state of statistical control. Indeed, this kind of myth may prevent well-intended practitioners from potentially benefiting from the powerful tool of design of experiments in situations where such methods would be advantageous or the only economical way to tackle a particular problem.

The irony is that three quarters of a century ago, Ronald A. Fisher, the originator of modern methods of statically designed experimenters, introduced the three key concepts of randomization, blocking, and replication

Address correspondence to Søren Bisgaard, Eugene M. Isenberg School of Management, University of Massachusetts, Amherst, 121 President Drive, Amherst, MA 01003, USA. E-mail: bisgaard@som.umass.edu

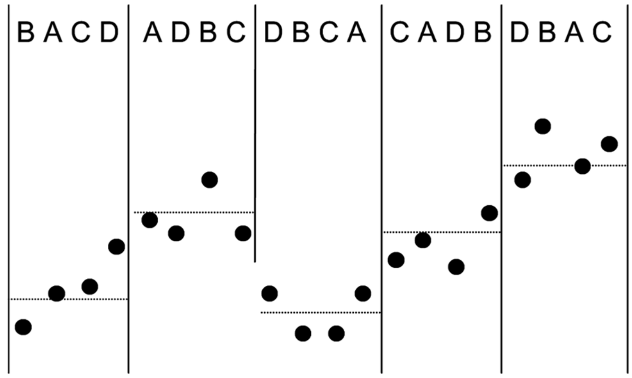

FIGURE 1 A reproduction of the symbolic diagram from Box et al. (1990, p. 190) referred to by Ryan (2000).

precisely for the reason of making it possible to conduct experiments on processes that are not necessarily in a state of statistical control. Nevertheless, Ryan (2000) continued, "Box, Bisgaard and Fung (1990, p. 190) show a diagram [see Figure 1] in which experimentation is performed during (short) time intervals when a process appears to be stable. This ideal may be difficult to achieve, however, as these 'blocks' can be constructed *only* if process shifts can be identified [emphasis added]." Again this is fortunately not true. The idealized diagram in Figure 1 was not intended to imply that the shifts necessarily need to be identified. Of course, it is helpful if they can. But it is not a necessary prerequisite.

To bolster his argument, Ryan (2000) wrote, "Deming strongly believed that processes should be in a state of statistical control." That Deming should have expressed such opinions has some merit. For example, Deming (1986) states that "Students, are not warned in classes nor in the books that for analytic purposes (such as to improve a process), distributions and calculations of mean, mode, standard deviation, chi-square, *t*-test, etc., serve no useful purpose for improvement of a process unless the data were produced in a state of statistical control" (p. 312). When making such statements we tend to think Deming aimed at other issues than the blanket condemnation of the use of design of experiments. We certainly have not been able to find any explicit statement in his writing to that effect and doubt he would have been that dogmatic. Nevertheless, our consulting experience indicates that many practitioners have interpreted Deming's statements to imply that he also condemned the use of design of experiments unless the process is in statistical control. It is sobering to note that W. S. Gosset invented the paired *t*-test for the explicit purpose of applying it to agricultural field trials that he well knew were not to be in a state of statistical control; see Student (1911). What makes this application legitimate and useful is, of course, the clever application of blocking, randomization, and replication.

In this article we review Fisher's (1925) original discussion of the fundamental principles of design of experiments and in particular randomization, replication, and blocking. Central to our discussion is a review of Fisher's examples with added statistical graphics and commentary. We then extend Fisher's examples to further illustrate the application of these principles to processes that are not in a state of statistical control. In the final section of the article we discuss the implication of these fundamental principles for industrial quality improvement experiments.

EARLY UNIFORMITY TRIALS

Although quality engineers primarily are concerned with industrial experiments, it is historically interesting, we believe, to revisit the agricultural example used by R. A. Fisher when he introduced the fundamental principles of design of experiments. Indeed we are going to show how similar the data structure of his experiment is to those we often encounter today in industry. Fisher, as indicated above, the primary inventor of modern statistical principles of design of experiments, began employment in 1919 as a statistician at Rothamsted Experimental Station in England. In this capacity, he was intimately involved with agricultural scientists, first in analyzing the results of their often not so well-planned field trials and later also in the design of new experiments.

In earlier work at Rothamsted, uniformity trials were common. The purpose was to understand

TABLE 1 Twenty Observations from Adjacent Plots of the Yield of Mangold Roots from Uniformity Trials Reported by Mercer and Hall (1911) and Used by Fisher (1925)

3504	3430	3376	3334	3253	3314	3287	3361	3404	3366
3416	3291	3244	3210	3168	3195	3330	3118	3029	3085

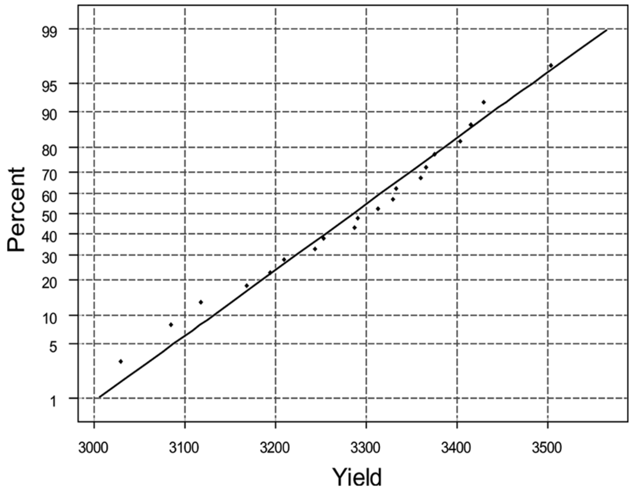

FIGURE 2 **A normal probability plot of Mercer and Hall's data in Table 1.**

the natural variability between plots treated alike. It was found that plots, the basic units in agricultural experiments, were heterogeneous across fields and that the yield from adjacent plots was highly correlated. For example, a particularly important set of uniformity trials on growing mangold roots reported by Mercer and Hall (1911) was the focus of much discussion at that time. Among other data, Mercer and Hall published the yield (in pounds) from 20 adjacent strips of land across an experimental field. These 20 observations are shown (row wise) below in Table 1.

These data shown as a normal plot in Figure 2 appear to be roughly normally distributed. However,

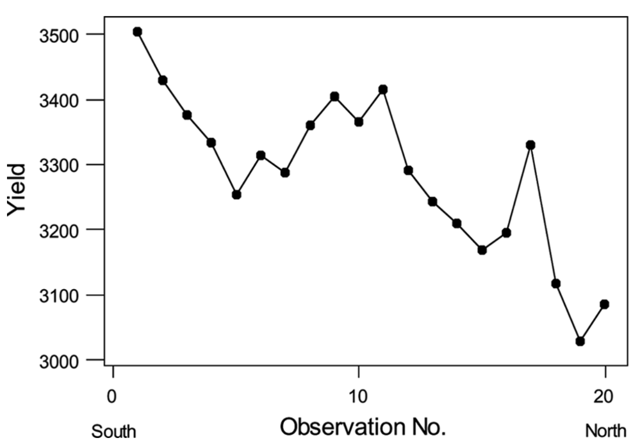

FIGURE 3 **Mercer and Hall's yield data from narrow adjacent plots of land plotted in south-north direction across an agricultural field. Note that if the directional scale was substituted with a time scale, the appearance could easily resemble the typical nonstationarity exhibited by an industrial process that is not in statistical control.**

plotted in order of observation across the field as in Figure 3, we see that the data are highly nonstationary. Note that if the horizontal axis was time, Figure 3 would be indistinguishable from a typical nonstationary industrial process. Although the number of observations is too small for serious time series analysis, we find that the data fit well a first-order autoregressive AR(1) model, the very same model typically used as the immediate alternative to assuming independence when checking the performance of control charts. Another time series model that fits perhaps even better is an integrated moving average IMA(1,1) model, the model typically used to model a wide class of industrial processes that are not in statistical control.

FISHER'S DEMONSTRATION OF THE BASIC PRINCIPLES OF DESIGN OF EXPERIMENTS

From the careful study of such data as those of Mercer and Hall (1911), Fisher knew that field data were never in a state of statistical control. Given that fact, he considered how valid experimentation could be made possible. So inspired, Fisher developed during the early part of the 1920s modern principles of design of experiments, in particular the concepts of randomization, blocking, and factorial experiments as well as the analysis of variance (ANOVA), the latter logically tying together the former three concepts. Specifically, in Fisher (1925) he explained:

> The peculiarity of agricultural field experiments lies in the fact, verified in all careful uniformity trials, that the area of ground chosen for the experimental plots may be assumed to be markedly heterogeneous, in that its fertility varies in a systematic, and often a complicated manner from point to point. For our test of significance to be valid the difference in fertility between plots chosen as parallels must be truly representative of the differences between plots with different treatments; and we cannot assume that this is the case if our plots have been chosen in any way according to a prearranged system; for the systematic arrangement of our plots may have, and tests with the results of uniformity trials show that it often have, features in common with the systematic variation of fertility, and thus the test of significance is wholly vitiated.

Hence, we see that Fisher explicitly recognized from the outset that the challenge was to develop methods for conducting experiments under conditions

TABLE 2 **Random Allocation of Five Treatments to the 20 Uniformity Trials on Mangold from Table 1 (this is Fisher's Table 57, p. 268)**

B	*C*	*A*	*C*	*E*	*E*	*E*	*A*	*D*	*A*
3504	3430	3376	3334	3253	3314	3287	3361	3404	3366
B	*C*	*B*	*D*	*D*	*B*	*A*	*D*	*C*	*E*
3416	3291	3244	3210	3168	3195	3330	3118	3029	3085

where the process exhibits systematic variation and variance heterogeneity or using industrial terminology not in a state of statistical control.

Following this explanation, Fisher proceeded to demonstrate, via numerical examples, the principles of randomization, blocking, and ANOVA. Specifically, he simulated two experimental designs with five treatments *A, B, C, D*, and *E* using the 20 observations from Mercer and Hall (1911) shown in Table 1. Fisher's first design was an unrestricted random layout and the second a randomized block design. For the random layout Fisher shuffled a deck of 20 cards with the five letters *A, B, C, D*, and *E* each repeated four times and assigned these five dummy treatments to the 20 observations in Table 1. He obtained the assignment shown in Table 2 and illustrated graphically in Figure 4. Fisher proceeded to perform an analysis of variance here reproduced in Table 3 (except he computed the log of what we now call the *F*-ratio).

From the analysis of variance in Table 3, we see that the treatments are not significant. As Fisher (1925) noted, this "illustrates the manner in which a purely random arrangement of plots ensures that the experimental error calculated shall be an unbiased estimate of the errors actually present." In particular, notice that the mean squares for treatment and for error are almost identical. Indeed, the process of randomization ensures that the long run means of the mean squares of the two are identical and that the test is unbiased. Hence, should a real treatment effects be present, the error mean squares will provide an appropriate estimate of the treatment differences; see Welch (1937), Pitman (1937), and Box and Andersen (1955).

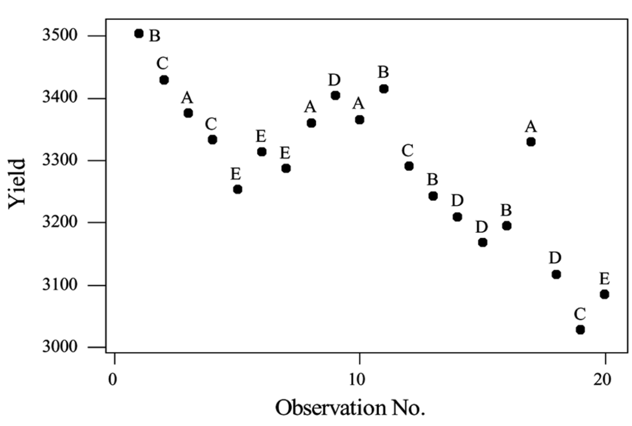

FIGURE 4 **Mercer and Hall's yield data plotted in the south-north direction with superimposed labels for five dummy treatments *A, B*, . . . , *E* from an unrestricted random layout.**

After the simulation of the unrestricted random layout, Fisher proceeded to demonstrate the benefits of blocking when processes are heterogeneous or not in a state of statistical control. Specifically, he explained that "it is still possible to eliminate much of the effect of the soil heterogeneity, and so increase the accuracy of our observations, by laying restrictions on the order in which the strips are arranged."

Dividing the 20 observations up in four blocks each of five observations, Fisher continued, "If the five treatments are arranged at random within each block, our estimate of the experimental error will be an unbiased estimate of the actual errors in the differences due to treatment." The specific randomized block arrangement Fisher used is given in Table 4 and illustrated graphically in Figure 5. The corresponding analysis of variance is shown in Table 5.

From the analysis of variance in Table 5, we see that the block differences are highly significant, but the treatments are not. Further, the residual error standard deviation is square root $(7869) = 88.7$. The true error variance can be calculated by adding the sum of squares for treatments and error divided

TABLE 3 **Analysis of Variance for the Unrestricted Random Layout**

Source	Degrees of freedom	Sum of squares	Mean squares	*F*	*p*
Treatments	4	58,726	14,681	0.95	0.461
Error	15	231,040	15,403		
Total	19	289,766			

TABLE 4 Random Block Allocation of Five Treatments to Four Blocks

AECDB	CBEDA	ADEBC	CEBAD

by the combined degrees of freedom: $df = 4 + 12 = 16$. Hence, the analysis of variance again provides an estimate of the residual standard deviation that is close to the "true" error standard deviation of $\sqrt{(40,859 + 94,424)/16} = 91.95$. By comparison with the random layout, notice that blocking has cut the residual error mean square approximately in half, improving the designs ability to pick up real effects should such be present.

THE NON-NULL CASE

To evaluate how randomization, blocking, and replication help detect real effects of treatments, we now take Fisher's simulations one step further. Suppose the effects were $A = -100$, $B = -50$, $C = 0$, $D = +50$ and $E = +100$. The yield is then the sum of the uniformity data provided in Table 1 and the treatment effects given above. The simulated data for the random layout and the randomized block design are all shown Table 6.

For the random layout, the analysis of variance is shown in Table 7. Notice that, despite the presence of real effects, the treatments are not significant. This is, of course, because the random layout does not help remove the pronounced nonstationarity of the data. Notice also that to make it difficult we deliberately simulated small treatment effects relative to the size of the noise.

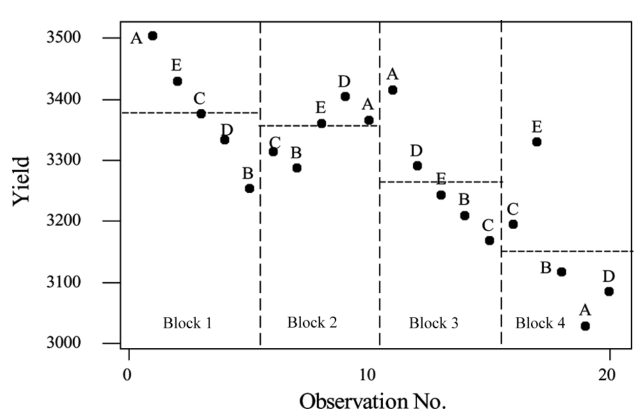

FIGURE 5 Fisher's randomized block design in Table 4 applied to Mercer and Hall's uniformity trial data plotted in south-north direction. Superimposed labels for five dummy treatments *A*, *B*, …, *E*. The horizontal lines are block averages.

Let us now turn to the randomized block design. The analysis of variance is shown in Table 8. Both blocks and treatments now are very significant. Hence, by using blocking and randomization, we now detect the presence of the simulated effects despite the severe nonstationarity of the process.

To sum up, the examples demonstrated that it is possible to perform experiments on processes not in a state of statistical control. With appropriate use of randomization, the experiment will provide a correct unbiased error estimate. Further, blocking, when used appropriately, helps reduce the error and hence improves the discriminative power of the experiment. Blocking is essentially a high-pass filter that filters out low-frequency changes but passes through high-frequency changes, the changes we impose with the treatments. For a theoretical discussion and underpinning, see Welch (1937), Pitman (1937), and Box and Andersen (1955). See also Box et al. (1978, chapter 4, for a more accessible discussion.

Note that the blocks in these examples were not necessary associated with "process shifts [that] can be identified" (Ryan, 2000). Rather, because of the relatively high positive correlation between neighboring plots, blocks were formed as compact groups of experimental units close together in space (but could as well have been in time) so that the process is relatively more homogeneous within the blocks.

In general, blocks can be formed under any circumstances and for any number of reasons. However, blocking will only be effective if smaller subgroups of experimental units exhibit less variability than the experimental units taken as a whole; see Bisgaard (1997). If process shifts can be identified prior to the execution, that is obviously a good reason to block. However, as illustrated with Fisher's examples, the criterion for subdividing the units is not limited to identifiable process shifts. The forming of blocks may be based on a wide variety of knowledge. In the above example, it was the knowledge that processes not in a state of statistical control frequently exhibit positive autocorrelation. Ironically, if the data in the above example had been in a state of statistical control, blocking would not have improved the experiments ability to detect significant effects.

TABLE 5 Analysis of Variance for the Randomized Block Design

Source	Degrees of freedom	Sum of squares	Mean squares	F	p
Blocks	3	154,493	51,494	6.54	0.007
Treatments	4	40,859	10,215	1.30	0.325
Error	12	94,424	7,869		
Total	19	289,766			

TABLE 6 The Simulated Data for the Random Layout and for the Randomized Block Design With Simulated Real Effects. The First Column, Y, is the Mercer and Hall Data. The Two Last Columns Show the Simulated Yield Observations from a Random and Randomized Block Experiment Respectively

Obs. no.	Y	Treatments, random	Blocks	Treatments, blocked	Effects, random	Effects, blocked	Yield, random	Yield, blocked
1	3504	*B*	1	*A*	−50	−100	3454	3404
2	3430	*C*	1	*E*	0	100	3430	3530
3	3376	*A*	1	*C*	−100	0	3276	3376
4	3334	*C*	1	*D*	0	50	3334	3384
5	3253	*E*	1	*B*	100	−50	3353	3203
6	3314	*E*	2	*C*	100	0	3414	3314
7	3287	*E*	2	*B*	100	−50	3387	3237
8	3361	*A*	2	*E*	−100	100	3261	3461
9	3404	*D*	2	*D*	50	50	3454	3454
10	3366	*A*	2	*A*	−100	−100	3266	3266
11	3416	*B*	3	*A*	−50	−100	3366	3316
12	3291	*C*	3	*D*	0	50	3291	3341
13	3244	*B*	3	*E*	−50	100	3194	3344
14	3210	*D*	3	*B*	50	−50	3260	3160
15	3168	*D*	3	*C*	50	0	3218	3168
16	3195	*B*	4	*C*	−50	0	3145	3195
17	3330	*A*	4	*E*	−100	100	3230	3430
18	3118	*D*	4	*B*	50	−50	3168	3068
19	3029	*C*	4	*A*	0	−100	3029	2929
20	3085	*E*	4	*D*	100	50	3185	3135

TABLE 7 Analysis of Variance for the Unrestricted Random Layout with Simulated Real Effects

Source	Degrees of freedom	Sum of squares	Mean squares	F	p
Treatments	4	14,026	3,506	0.23	0.919
Error	15	231,040	15,403		
Total	19	245,066			

TABLE 8 Analysis of Variance for the Randomized Block Design with Simulated Real Effects

Source	Degrees of freedom	Sum of squares	Mean squares	F	p
Blocks	3	154,483	51,494	6.54	0.007
Treatments	4	175,459	43,865	5.57	0.009
Error	12	94,424	7,869		
Total	19	424,366			

CONCLUSIONS

The state of statistical control is a fiction but often a convenient mathematic assumption for certain statistical derivations. As pointed out by Box and Luceno (1997), this assumption contradicts the Second Law of Thermodynamics. Taken to the extreme the logic that statistical control is a necessary prerequisite for conducting valid experiments would imply that we never would be able to conduct real-life experiments. However, as demonstrated by countless practitioners in industry as early as the 1920s, designed experiments can indeed be successfully executed even when the underlying process is not in a state of statistical control provided that replication, blocking, and randomization are appropriately used.

Since the early work by Fisher and Gosset, statisticians working in the area of experimental design have understood that to be of any use in agriculture, it is a necessary prerequisite that designed experiments can be applied to experimental units that are not in a state of statistical control. A survey of the early literature suggests that this was the main impetus for Fisher's and Gosset's work. We may even go as far as suggesting that the reason modern theory of design of experiments developed in the agricultural context was that agriculture for that very reason is so much more challenging. Physicists and chemists were, for centuries prior, successfully able to perform carefully controlled informative experiments in their laboratories without much need for replication, blocking, and randomization. However, agriculture was different. As Fisher and Gosset realized and frequently discussed, if agricultural experimenters artificially tried to make the experimental units resemble a state statistical control, the resulting conclusions might be of little practical value; see Box (1980). For example, if agricultural experiments were conducted in the controlled conditions of a greenhouse, the experiments may likely exhibit little if any valid inferential relevance to real farm production where crops must compete with weeds and be able to grow in a variety of soil conditions, etc. It was therefore necessary to develop techniques that could cope with the type of nonstationarity exemplified by Mercer and Hall's (1911) data. As pointed out above, that was Fisher's goal. Unfortunately, this is often overlooked in the industrial context with the notable exception of Davies (1954).

One of the important conceptual developments that occurred in the past 20 years in quality is the change of focus from quality control to quality improvement. We interpret Deming's (1986) admonition of "Cease dependence on inspection to achieve quality" as a "call to arms" for the use of designed experiments as a tool for discovering and eliminating causes for poor quality in general and lack of statistical control in particular. It is to a large extent this shift of emphasis toward proactive improvement of quality that economically legitimizes large-scale applications of statistics in industry as exemplified by the Six Sigma program. Thus, it is counterproductive if it is assumed that statistical control is a prerequisite for the use of designed experiments.

Design of experiments is perhaps the most potent tool available for the quality engineer to get a process in control. Any alternative noninvasive observational approach to process improvement is fraught with the possibility of unknown biases and confounding. As explained by Box et al. (1978), only a carefully designed and properly randomized experiment can prevent the danger of confusing causation with non-sense correlation.

When Fisher first advanced his ideas of designed experiments in the 1920s, the ideas were by no means readily accepted. In the decades that followed, statistical and agricultural journals hosted numerous discussions on the topic, some of them contentious. However, today, with a considerable literature to back up the theory, it would be considered irrational to require stable uniformity of experimental units in agriculture. We hope this article will contribute to a similar development for industrial experimentation.

ACKNOWLEDGMENTS

The author thanks George Box for helpful comments on a earlier draft of this manuscript and Carla Vivacqua and Andre Pinho for their careful proofreading of the final draft. The work on this article was supported by the Isenberg Program for Technology Management, the Isenberg School of Management, University of Massachusetts Amherst.

ABOUT THE AUTHOR

Dr. Søren Bisgaard is the Eugene M. Isenberg Professor of Technology Management at University of Massachusetts Amherst and Professor or Business and Industrial Statistics at University of Amsterdam. Originally trained as a machinist/toolmaker, he holds degrees in industrial and manufacturing engineering from Denmark and MS and Ph.D. degrees in statistics from university of Wisconsin–Madison. Previously he worked as a full time management consultant, held positions as Professor and a Director in the Institute for Technology Management at University of St. Gallen, Switzerland, as Professor of Industrial Engineering and Director of the Center for Quality and Productivity Improvement at University-Madison. Professor Bisgaard is a Fellow of the American Society for Quality (ASQ) and the American Statistical Association, an Academician of the International Academy for Quality, and the author of more than 100 publications. He has received the Shewhart Medal, the Hunter Award, the Ott Award, the Wilcoxon Prize, the Shewell Award, and the Brumbaugh Award and serves on the editorial board of *Journal of Quality Technology, Quality Engineering*, is on the management board of *Technometrics* and on the publications Management Board of ASQ.

REFERENCES

Bisgaard, S. (1997). Blocking in factorial experiments. *Quality Engineering*, 9(4):753–759.

Box, G., Bisgaard, S., Fung, C. (1990). *Designing Industrial Experiments*. Madison, WI: BBBF Books.

Box, G., Luceno, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment*. New York: John Wiley & Sons.

Box, G. E. P., Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumptions. *Journal of the Royal Statistical Society, B*, 17(1):1–34.

Box, G. E. P., Hunter, W. G., Hunter, J. S. (1978). *Statistics for Experimenters*. New York: John Wiley & Sons.

Box, J. F. (1980). R. A. Fisher and the design of experiments, 1922–1926. *The American Statistician*, 34:1–7.

Davies, O. L. Ed. (1954). *Design and Analysis of Industrial Experiments*. London: Hafner Publishing.

Deming, W. E. (1986). *Out of the Crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Fisher, R. A. (1935). *Design of Experiments*. London: Oliver and Boyd.

Mercer, W. A., Hall, A. D. (1911). The experimental error of field trials. *The Journal of Agricultural Science*, 4(2):8–127.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, 29(4):322–335.

Ryan, T. P. (2000). Discussion: Controversies and contradictions in statistical process control. *Journal of Quality Technology*, 32(4):367–369.

Student. [W. S. Gosset, pseud.]. (1911). Appendix to Mercer and Hall's paper on ''The experimental error of field trials.'' *Journal of Agricultural Science*, 4:128–131.

Welch, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika*, 29(1):21–52.

Woodall, W. H. (2000). Controversies and contradictions in statistical process control. *Journal of Quality Technology*, 32(4):341–350.