Applied Stochastic Models in Business and Industry

Special Issue Paper

Received 13 December 2013, Accepted 15 December 2013 Published online 4 February 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2017

George Box and the design of experiments: statistics and discovery

David M. Steinberg*[†]

George Box was fascinated with how we make discoveries. His path-breaking contributions to experimental design made statistics an active partner in the process of discovery. Box introduced us to response surface methods, evolutionary operation, resolution and rotatability, projective properties and design robustness. He developed popular experimental plans like the central composite and Box-Behnken designs. He explored the consequences of imperfect models and derived D-optimal designs for experiments to estimate mechanistic models. Box's ideas grew from close collaborations with scientists and engineers and have been applied successfully in a wide range of disciplines. He has left an indelible stamp on the field of experimental design and on the practice of scientific investigation. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: response surface methods; evolutionary operation; resolution; iterative learning; factorial design; central composite design; Box–Behnken design

1. Introduction

George Box was a pioneer in the statistical design of experiments. He introduced new strategies and concepts, such as response surface methods (RSM), sequential experimentation, rotatability, projectivity and robustness. He was fascinated by the process of scientific discovery and was a keen observer of how scientists and engineers engaged in discovery. His technical contributions to design paralleled significant ideas on the philosophy of experimentation and how experimental science could be made more efficient by exploiting statistical principles. His paradigm for experimental optimization and the tools he developed to implement it have found widespread application spanning the realm of science and technology.

Many scientists and engineers learned the fundamentals of experimental design from his classic textbook, *Statistics for Experimenters*, written jointly with William G. Hunter and J. Stuart Hunter [1, 2].

Throughout Box's career, there was a constant dialog and interplay between theory and practice, with challenging real problems stimulating the development of new statistical theory, which in turn was presented with a focus on relevance to statistical practice. The importance of merging theory and application will be evident throughout this article.

I will focus on Box's early research in experimental design, in particular RSM, evolutionary operation (EVOP), factorial designs and design for mechanistic models. Later research on robust parameter design and on Bayesian methods for the design and analysis of experiments are discussed in the articles by Jones, Meyer and Fung in this volume.

2. Historical background

George Box was initially trained as a chemist. His background in the physical sciences and first-hand involvement in experiments helped to shape his thinking on the statistical aspects of experimentation. He was fond of relating how his career in statistics began during World War II (Box, [3], pages 28–31). At that time, he served with a unit working to develop and test antidotes that could be used if England were attacked with poison gas. He was convinced that the unit needed a statistician who could help them better understand their test data and, upon presenting that need to his commanding officer, found himself appointed to the position. During the remainder of his army service, he designed and analyzed

Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel
**Correspondence to: David M. Steinberg, Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel.*
[†]*E-mail: dms@post.tau.ac.il*

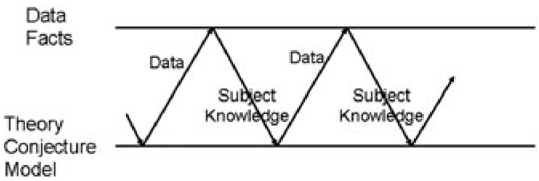

Figure 1. George Box's graphic description of the iterative nature of learning and of statistical investigation.

hundreds of experiments using methods that had been developed by R.A. Fisher [4]. Thus, he first learned about block and factorial designs in the laboratory, planning and running the experiments and analyzing the resulting data.

After the war, Box studied statistics and began to work with the Statistical Methods Group at Imperial Chemical Industries (ICI). One of his main tasks was to assist in improving the yield of chemical processes, and it was there that he developed the main elements of response surface methodology. Here again, Box's ideas coalesced in collaboration with the chemists and engineers on the yield improvement teams. The applied problems that they needed to solve were the stimulus for the statistical theory that he developed and those ideas found immediate application in real experiments.

3. Iterative learning

George Box was extremely sensitive to the scientific context of experimentation. The scientists with whom he worked in industry typically carried out experiments in an iterative manner – the results of initial experimental runs affected the decisions made about how to proceed with subsequent runs. As Box [5] wrote, 'In any realistic view of the process of investigation, the dimensions, identity, location, and metrics of measurement of regions of interest in the experimental space are all *iteratively evolving*.'

The factorial plans that had been developed by Fisher and his colleagues included many runs and were not attractive to industrial scientists. An important contrast between the industrial environment in which Box worked and the agricultural experiments that Fisher designed was what Box has called 'immediacy', the fact that the results in industry were often available quickly. Box was convinced that the efficiency of industrial experiments could be improved by applying the principles of experimentation set forth by Fisher, but that adaptation was needed to take advantage of the rapid feedback and sequential framework found in industry.

The fruitful marriage of design principles to industrial experiments is effectively illustrated in Figure 1 (adapted from Box [6]) and is succinctly described in Box [7]. One of Box's most important contributions to experimental design is the framework of iterative experimentation in Figure 1. The cycle begins with a model, a conceptual framework of what is driving the system under study. Many aspects of the system will not be fully understood, and this sets the stage for collecting data to improve understanding. Once the data have been analyzed, new questions will arise, suggesting the need for further experimentation. The procedure continues, much in the spirit of the Deming cycle of plan, do, check and act.

Experimental design principles are relevant at each of the data collection steps. For example, the team would do well to heed Fisher's exhortation to run factorial experiments rather than proceeding one-factor-at-a-time. Small fractional factorial plans will often be ideal choices to take advantage of immediate feedback. Each step should be considered as part of the complete picture of experimentation. The team can adapt to what they have learned in previous steps. Often this will mean changing the factors that are being studied, changing the range of the factors or perhaps adding intermediate levels.

4. Response surface methods

The paradigm of iterative learning is clearly evident in RSM. The ideas were first described in a path-breaking paper by Box and Wilson [8]. Two sequels (Box [9], Box and Youle [10]) elaborated on them and provided additional tools for design and analysis.

Response surface methods grew out of Box's work with yield optimization teams at ICI. A typical application would have a list of factors, say $x_1, \ldots, x_k$, that might affect the yield (y). The goal is to learn about the relationship between the yield and the factors and, thereby, to identify factor settings that maximize the yield. At the start of experimentation, often little is known about the relationship between y and $x_1, \ldots, x_k$. In that case, a first-order polynomial approximation might reflect the main structure of the relationship in the experimental region,

$$E\{y\} = \beta_0 + \sum_{j=1}^{k} \beta_j x_j . \tag{1}$$

A two-level fractional factorial design is a natural candidate to estimate this model. This will enable the team to identify which of the factors are most influential and, perhaps, to point toward higher or lower levels of those factors than were originally considered. The results might lead the team to drop some of the factors or to add other factors.

The first-order model can be used to find a direction of steepest ascent, which may lead to quick gains in yield via further experiments conducted along that direction. When yield no longer increases, another two-level fractional factorial can be used, and the process repeated. Center points are usually included in the two-level designs to help check whether the first-order approximation is a plausible model. When that is not the case, alternative designs will be needed, as described later.

Response surface methods are widely used by scientists and engineers. There are several excellent books on the topic, including Box and Draper [11] and Myers, Montgomery and Anderson-Cook [12].

5. Central composite design

At some point, the team is likely to approach the region of optimal yield. This could be a sharp optimum but often involves a ridge in which a number of conditions in the factor space can be 'traded off' against one another with little change in yield. Box showed that these relationships can often be usefully represented by treating expected yield as a second-order polynomial of the factors,

$$E\{y\} = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 + \sum_{j<j'} \beta_{jj'} x_j x_{j'}. \tag{2}$$

It is important to emphasize that models (1) and (2) are meant as *local approximations* that will be useful in the current region of experimentation; they should not be viewed as exact descriptions of the factor–yield relationship over a broad range of factor settings.

Box realized that new designs were needed to fit quadratic models like (2). Box and Wilson [8] proposed the central composite design (CCD) for this purpose. The CCD is composed of three basic building blocks:

(1) A two-level (fractional) factorial, with all factors set at (coded) levels of ±1.
(2) Center points in which all factors are at a middle level of 0.
(3) Axial (or star) points, in which all factors but one are at the middle level, and the remaining factor takes levels $\pm\alpha$, to be chosen sensibly.

The design permits estimation of all the two-factor interaction terms, usually achieved by selecting a sufficiently large two-level fractional factorial. The CCD can be conveniently blocked by running the factorial points and some center points in one block and the axial points and some center points in the second block. In many response surface studies, a fractional factorial with center points suggests that a second-order approximation is needed, and the design is then augmented by the second block of points. The full design then permits estimation of the full quadratic model and a block effect. The constant α could also be 1, so that each factor is run at three distinct levels or could be chosen to meet some other criteria.

Macedo *et al.* [13] described a response surface study in biotechnology. Their goal was to improve the medium used to grow cells for a new microbial source of transglutaminase. Transglutaminase (TGase) is an enzyme that catalyses an acyl transfer reaction using peptide-bond glutamine residues as acyl donors and several primary amines as acceptors. Reactions catalyzed by TGase can be broadly used in food processing industries. The project began with a factor screening experiment in which six factors were studied in a 2^{6-2} design, with three center points. Four of the factors had statistically significant effects on TGase activity, and there were several important chains of two-factor interactions. The lack of fit test was not statistically significant, but suggested that a higher order model might be needed.

In the next phase of experimentation, the team retained the four important factors, shifting their ranges in the direction of steepest ascent. The other two factors were dropped. They employed a CCD, which is shown in Table I along with the results of the experiment. The first 16 rows are a 2^4 experiment, the next eight rows are axial points and the final three rows are center points. The value of α in this experiment was 2. A second-order model fits the data well. The fitted surface has an optimum near the center of the design region. The coefficients of all the quadratic terms are negative and large, suggesting a well-defined optimum. The predicted value at the estimated optimal setting is 1.39, close to the average of 1.37 at the three center points. See Steinberg and Bursztyn [14] for a detailed analysis of the data from these experiments.

Box [15] related how the development of the CCD resulted from practical needs. An initial factorial design at ICI to study yield had produced surprising results: all the two-factor interactions were large but all the main effects were small, contrary to common wisdom that low-order effects will typically dominate. The explanation for this was that 'one-factor-at-a-time' optimization had been carried out in earlier experiments, conducted before Box became involved. Consequently, the current process conditions had approximately 0 derivatives along each factor axis, hence the small main effects. The

Table I. Coded and actual values of the central composite design for optimization of microbial source of transglutaminase (MTGase) activity.

	Powered soy	KH_2PO_4	$MgSO_4.7H_2O$	Peptone	MTGase activity (U/ml)
Run	X1	X2	X3	X4	120 hrs
1	−1 (1.5)	−1 (0.2)	−1 (0.1)	−1 (0.5)	0.87
2	1 (3.5)	−1 (0.2)	−1 (0.1)	−1 (0.5)	0.74
3	−1 (1.5)	1 (0.6)	−1 (0.1)	−1 (0.5)	0.51
4	1 (3.5)	1 (0.6)	−1 (0.1)	−1 (0.5)	0.99
5	−1 (1.5)	−1 (0.2)	1 (0.3)	−1 (0.5)	0.67
6	1 (3.5)	−1 (0.2)	1 (0.3)	−1 (0.5)	0.72
7	−1 (1.5)	1 (0.6)	1 (0.3)	−1 (0.5)	0.81
8	1 (3.5)	1 (0.6)	1 (0.3)	−1 (0.5)	1.01
9	−1 (1.5)	−1 (0.2)	−1 (0.1)	1 (1.5)	1.33
10	1 (3.5)	−1 (0.2)	−1 (0.1)	1 (1.5)	0.70
11	−1 (1.5)	1(0.6)	−1 (0.1)	1 (1.5)	0.82
12	1 (3.5)	1(0.6)	−1 (0.1)	1 (1.5)	0.78
13	−1 (1.5)	−1 (0.2)	1 (0.3)	1 (1.5)	0.36
14	1 (3.5)	−1 (0.2)	1 (0.3)	1 (1.5)	0.23
15	−1 (1.5)	1(0.6)	1 (0.3)	1 (1.5)	0.21
16	1 (3.5)	1(0.6)	1 (0.3)	1 (1.5)	0.44
17	−2 (0.5)	0 (0.4)	0 (0.2)	0 (1.0)	0.56
18	2 (4.5)	0 (0.4)	0 (0.2)	0 (1.0)	0.49
19	0 (2.5)	−2 (0.0)	0 (0.2)	0 (1.0)	0.57
20	0 (2.5)	2 (0.8)	0 (0.2)	0 (1.0)	0.81
21	0 (2.5)	0 (0.4)	−2 (0.0)	0 (1.0)	0.90
22	0 (2.5)	0 (0.4)	2 (0.4)	0 (1.0)	0.65
23	0 (2.5)	0 (0.4)	0 (0.2)	−2 (0.0)	0.91
24	0 (2.5)	0 (0.4)	0 (0.2)	2 (2.0)	0.49
25	0 (2.5)	0 (0.4)	0 (0.2)	0 (1.0)	1.43
26	0 (2.5)	0 (0.4)	0 (0.2)	0 (1.0)	1.17
27	0 (2.5)	0 (0.4)	0 (0.2)	0 (1.0)	1.50

initial experiments had not studied interactions and so had missed the gains in yield that could result from changing several factors simultaneously. Rather than extend the design from a two-level factorial to a three-level factorial (which would have greatly increased the sample size), Box realized that the important issue was to move from a first-order to a second-order graduating function. The modest addition of the axial points, with some additional center runs to guard against a block effect, was all that was needed to estimate this model. The CCD has since been used in an enormous number of scientific and engineering problems.

There were other major contributions of the Box and Wilson paper. It was the first paper to espouse Box's vision of sequential, iterative experimentation. It developed analysis methods for the data from designs chosen to fit first-order and second-order models.

6. Effect aliasing and fold-over designs

Box and Wilson also introduced the idea of aliasing of effects and showed how some aliasing could be eliminated by 'folding over' the points in a design. Aliasing refers to settings in which a design, due to fractionation, cannot provide separate estimates of all the effects that might be present. For example, consider a 2^{7-4} design, which has seven factors at two levels each but only eight of the 128 possible factor combinations. This design can estimate all the main effects if we are prepared to assume that all the interactions are negligible. Suppose, though, one of the two-factor interactions is *not* negligible. Then that interaction will match up exactly with one of the seven main effects. Not only will we miss the interaction, we may erroneously conclude that the matching main effect is important. Box and Wilson described this by saying the interaction is *aliased* with the main effect.

Mathematically, aliasing has the following description. Suppose k factors are being studied, and the model currently under consideration is a regression model with p terms. For a first-order polynomial model, $p = 1 + k$; for a second-order model, $p = 1 + k + k(k+1)/2$. In matrix notation, the model for the data vector $\boldsymbol{Y}$ can be written as

$$Y = X\beta + \varepsilon, \tag{3}$$

where β is the parameter vector. Now, suppose that additional regression terms, not in the model, permit a much closer approximation to the expected responses, so that in fact

$$E\{Y\} = X\beta + X_1\beta_1. \tag{4}$$

Here the matrix X_1 includes all the additional regression functions, evaluated at the design points. For the fitted model (3), we now find that the expected values of the least squares parameter estimates are

$$E\{\hat{\beta}\} = \beta + A\beta_1, \tag{5}$$

where the *alias matrix* A is given by $A = (X^T X)^{-1} X^T X_1$. The (i, j) entry in this matrix relates how the ith coefficient in the fitted model is biased by the jth regression function among those that were left out of the fitted model.

A desirable design property is to minimize aliasing of main effects. For example, if the fitted model is a first-order regression (equation 1), a good design should have little bias due to second-order terms (the additional terms that appear in equation 2). Box and Wilson showed that the bias of fitted first-order terms resulting from second-order terms could be completely eliminated by *folding over* an orthogonal design for a first-order model, that is, by reflecting each of the design points through the origin. In the previous example, folding over the 2^{7-4} design results in a 16-run 2^{7-3} design.

7. Rotatable designs

Box and Hunter [16] studied the general structure of response surface designs. For first-order models, the standard designs were the two-level fractional factorials. When used to fit first-order models, these designs generate linear models of the form

$$Y = X\beta + \varepsilon \tag{6}$$

for which the matrix $X^T X$ is diagonal. Thus, the designs are *orthogonal*. This is a desirable design property; it implies that the effect estimates are uncorrelated with one another, so that the design provides independent information on each effect. However, there is no immediate extension of orthogonality to designs for second-order models.

Box and Hunter sought a more fundamental principle for characterizing designs. They observed that a further feature of the orthogonal designs for first-order models is related to the variance of the predicted values. Let $x = (x_1, \ldots, x_k)^T$ be any point in the factor space. Then,

$$\mathrm{Var}(\hat{y}(x)) = (\sigma^2/n)\left(1 + x_1^2 + \cdots + x_k^2\right) \tag{7}$$

and depends on x only through its distance from the origin, $d(x) = \left(x_1^2 + \cdots + x_k^2\right)^{0.5}$. They introduced the term *rotatable* to refer to designs with this property and argued that it was intuitively reasonable to look for designs that had such a balanced information pattern over the factor space.

Box and Hunter derived necessary conditions for rotatability of designs for second-order models. They presented several classes of such designs, including the CCDs, as well as orthogonal blocking schemes for the designs. Rotatable CCD's are obtained by setting the axial points at $\alpha = n_f^{0.25}$ (where n_f is the number of factorial points in the design). The variance profile for the design can be adjusted via the number of center points. The design in Table I is an example of a rotatable CCD. Box and Hunter also devoted attention to computational issues (important in the pre-computer days when the paper was published); for example, calculation of a confidence region for the stationary point of a second-order model was greatly simplified when a rotatable design was used.

Subsequent research has explored a number of metrics that can be useful for describing designs when rotatability cannot be achieved. These include variance dispersion graphs [17] and some indices of rotatability [18–21].

8. Box–Behnken designs

Box and Behnken [22,23] discovered two ingenious construction methods that could be used to design an experiment when it is desired to fit a second-order model. Each of these new classes had the advantage of being able to reduce the sample size by comparison with corresponding CCD's for various numbers of experimental factors. A further advantage was that they typically required only three levels for each factor, meeting a practical concern common to many experiments.

Box and Behnken [22] presented *simplex-sum designs*. The basis for these designs is a first-order orthogonal design. Denote by $\boldsymbol{D}$ the n by k design matrix for such a design, with n runs and k factors. New design points are derived from the initial points by computing and scaling the vector sum of pairs, triplets, and so on of design points in $\boldsymbol{D}$. For example, if two of the rows in $\boldsymbol{D}$ were $(-1,-1,-1)$ and $(-1,-1,1)$, the vector sum, divided by 2, gives the new design point $(-1,-1,0)$. Box and Behnken derived useful choices for the scaling factors, including choices for which the resulting design is rotatable. They also showed how, for some values of k, the design could be limited to subsets of these points, producing very economical designs.

Box and Behnken [23] found a clever construction method in which pieces of a factorial design are embedded in an incomplete block design (IBD). The latter is a design for studying k treatments in b blocks, each of which contains only $c<k$ treatments. Box and Behnken showed how an IBD could be used to generate a design for fitting a second-order model when k factors are under investigation. The idea is to associate each factor with one of the treatments in the block design. Each block is converted into a set of runs in the Box–Behnken design, as follows. The factors that correspond to the c treatments in the block are converted into a 2^c factorial design (or in some cases to a smaller fraction of the full factorial). All factors matching treatments not in the block are set at zero. Thus, each factor is limited to just three levels: ± 1(in the blocks that include the matching treatment) or zero (in the other blocks). The design is completed by adding a modest number of center points. Box and Behnken showed that the resulting designs are rotatable for second-order models. Moreover, they are easy to divide into blocks, using the blocks from the IBD. For most of the designs, orthogonal blocking is possible by combining blocks in the original set.

9. Robustness to model errors

The iterative approach to learning shown in Figure 1 describes a steady refinement of knowledge. The bottom axis is described as theory, conjecture or model. For Box, this often meant a summary of the phenomenon under study in terms of which factors were most important, over what ranges they should be considered and what sort of functional relationship linked the key process outputs to those inputs. These latter elements are the basics to how statistical models are expressed. Box was quite clear that these models should be regarded as useful approximations to reality – this is evident in his oft quoted aphorism, ‘all models are wrong but some models are useful’.

This view of the model as a useful approximation influenced Box’s thinking about how to choose an informative experimental design. An immediate concern was that additional terms that were ignored in the approximation would cause bias. Box and Wilson had already derived the alias matrix (equation 5) to describe the bias in parameter estimates. Box and Draper [24, 25] carried the idea a step further and looked at the bias in estimated values of the response function.

As in equation (4), suppose that the true response function $\eta(x)$ can be approximated with high accuracy in the current region of interest by

$$E\{Y(x)\} = \eta(x) \approx f^T(x)\beta + f_1^T(x)\beta_1. \tag{8}$$

Here x designates a particular factor combination, $f(x)$ is the vector of p regression functions in the model that will be fitted, evaluated at x, and $f_1(x)$ is the vector of extra regression functions that are left out of the fitted model. Once the data are collected and analyzed, the expected value will be estimated by

$$\hat{Y}(x) = f^T(x)\hat{\beta} \tag{9}$$

where $\hat{\beta}$ is the least squares estimator of β. The mean squared error (MSE) of $\hat{Y}(x)$ is given by

$$MSE(x) = E\left\{\left[\hat{Y}(x) - \eta(x)\right]^2\right\} = \mathrm{Var}\left(\hat{Y}(x)\right) + \left[E\{\hat{Y}(x)\} - \eta(x)\right]^2 = V(x) + B(x). \tag{10}$$

The final term $B(x)$ is the square of the bias and, treating the right-hand side of (8) as the true expected value, is given by

$$B(x) = \beta_1^T(A^T f - f_1)\left(f^T A - f_1^T\right)\beta_1. \tag{11}$$

Box and Draper argued that a useful summary of design quality for the current fitted model is the integrated MSE, with integration over a current region of interest. Using this as a design criterion naturally leads to consideration of the *scale* of the design; that is, to how widely the factor levels should be spread apart. An exact solution to this problem depends on the value of β_1, which of course is not known. So Box and Draper made the reasonable argument that the fitted model would likely be chosen to roughly balance the contributions of variance and squared bias over the current region of interest. Under that constraint, they derived the optimal design scale. Box and Draper [24] studied the implications of this scheme when the fitted model is first-order, and the bias is due to second-order terms. Box and Draper [25] extended the analysis to settings where the fitted model is second-order, and the bias is from third-order terms.

There is an interesting counterpoint in these papers of Box and Draper to the ideas of optimal experimental design, which were the subject of much debate at the time (see, e.g., Kiefer [26] and the subsequent discussion). Researchers in optimal design typically took as fixed inputs the form of the statistical model and the domain of experimentation. The model was assumed known, rather than an approximation, so bias was not considered; and the region was treated as perfectly known, rather than being somewhat amorphous and subject to discussion. For optimal design, then, the goal was to minimize variance, and the essential question was where to place the design points in the specified region. The particular criterion might be the determinant of the variance matrix of $\hat{\beta}$ (the D-optimality criterion) or the integrated variance of the predictions (the I-optimality criterion), or others. Invariably minimizing variance leads to placing most of the design points at the border of the experimental region.

Box and Draper turned the optimal design problem on its head. In their view, guided by the notion that the fitted model is just a local approximation, they fixed the placement of the design points and saw the main question as being how far to spread them out in the region of interest. (It is interesting that, for the variance only design criteria, the spread of the points has a much greater impact on design quality than their placement within a fixed region. Nonetheless, discussion of how to determine a good experimental region has attracted little attention in optimal design research.) Moreover, the tendency of optimal designs to push points to extreme settings was regarded with suspicion, as that would likely degrade the quality of the local approximation. Box and Draper [24, 25] derived a principled way to balance minimization of variance (which leads to extreme settings) with minimization of bias (which penalizes extreme settings).

Subsequent research on variance based criteria has used a variety of methods to merge the joint concerns over variance and bias. See, for example, [27–31].

10. Robustness to outliers

Box and Draper [32] studied the problem of how to choose a design that would limit the damage caused by outliers. They looked at the effect of outliers on the predicted values at the design points and at arbitrary points in the experimental region. They took as an initial case the setting where a single observation deviates from its expected value by a random error plus an additional error equal to c. The vector $\hat{Y}$ of predicted values at the design points will then have additive errors of $\delta = Hc$, where $H = X(X^T X)^{-1} X^T$ is the 'hat' matrix for the regression model. Using the fact that H is symmetric and idempotent, and assuming the u'th observation is the outlier, Box and Draper found that the sum of the squared errors is $\delta^T \delta = c^2 H_{u,u}$. Further assuming that the probability that any particular observation will be the outlying one is $1/n$, the expected sum of squares of the additive errors is $c^2 p/n$ and depends on the number of predictors (p) but not on the design itself.

As a secondary criterion, Box and Draper argued that it would be reasonable to make the consequences relatively uniform with respect to which observation is the outlier. This leads to the idea of minimizing the variance of the $H_{u,u}$. They showed that this continues to be a good criterion when additional additive outliers may be present. For first-order models, two-level factorial and fractional factorial designs optimize the Box–Draper robustness criterion. For second-order models, Box and Draper derived optimal configurations of CCDs. For most values of p and n, these designs had axial points similar to those needed to achieve rotatability.

The paper by Box and Draper [32] is also often noted for its list of 14 desirable properties of a design, which appears at the beginning of the article. Box and Draper exhorted experimenters to be conscious of these various, and sometimes contradictory, goals when planning an experiment. Robustness to outliers is one of the properties on the list.

11. Two-level factorial designs

Two-level factorial and fractional factorial designs had become a staple of the statistical tool kit by the early 1950s. Yates [33] pioneered the study of full factorials and Finney [34] developed the theory of regular two-level fractional factorials. Plackett and Burman [35] had published their war time research, which led to a large collection of orthogonal two-level designs, many of them not regular fractions.

Two-level designs were ideally suited to fitting first-order models and so became a standard choice for the initial stages of response surface studies. This also led Box and Hunter [36, 37] to take a much more in-depth look at the properties of these designs. This remarkably lucid pair of articles presents the theory of 2^{k-p} designs in a highly accessible manner that was ideal to promote their expanded use in industrial experimentation. The articles emphasized the use of design resolution to distinguish between alternate fractions. In the words of Box and Hunter [36], 'a design of resolution R is one in which no p factor effect is confounded with any other effect containing less than $R - p$ factors'. Box and Hunter discussed the choice of fractions, how to block 2^{k-p} designs and sequential strategies based on combining different fractions to disentangle aliased effects.

Box and Hunter [36] also briefly discussed the *projective properties* of the 2^{k-p} designs; that is, the structure of the design when some factors are ignored. For example, when a 2^{4-1} design with generator $4 = 123$ is projected onto any three factors, the resulting design is a full factorial. Projective aspects are especially relevant in screening contexts, when it is likely that only a subset of the factors under study will prove to have dominant effects, and it would be useful to have all combinations of those primary factors present in the experiment.

Some years later Box came to view projective properties as a key justification for using fractional designs and studied them in detail. These ideas were set out by Box and Tyssedal [38, 39] . Other related research was performed by Lin and Draper [40], Wang and Wu [41] and Cheng [42]. An interesting aspect is that non-regular designs (such as the 12-run Plackett and Burman design) proved to have advantages over regular fractions in terms of projectivity. Consider, for example, screening designs with a number of factors each at two levels. The 12-run design can accommodate 11 factors and the projection on any three factors includes all eight-factor combinations. The 16-run regular design loses this projective property as soon as more than eight factors are used.

12. Evolutionary operation

We generally think of using experiments in research and development, but not in full scale production, where careful control will be exerted to keep processes at target levels. Box [43] argued that there could also be great gains from exploiting simple experimental schemes, such as two-level factorials, to study and improve on-going production processes. In a nutshell, his idea was that 'a process should be run so as to generate product *plus information on how to improve the product*' (Box [43], p. 82; italics in the original). Modifying production conditions according to a small factorial design would provide constant feedback, enabling the process to evolve to more favorable production conditions. Box [43] called this approach *EVOP*. He pointed to a number of different settings in which EVOP would be a natural improvement strategy. These included, for example, changes in optimal process settings related to scaling up from a lab to plant level production, adjusting to drift of conditions in the operating environment and adapting a process to new equipment.

For details on the use of EVOP, including applications in industry, see Box and Draper [44].

13. Design for mechanistic models

From his background in the physical sciences, Box was familiar with many models whose basis was in the underlying physics, chemistry or biology of the problem. These models were derived mathematically from a theoretical description of the system under study. Box's approach to experimental design for such *mechanistic models* was quite different from his approach when the model was a convenient empirical graduating function. The designs he proposed in this setting focus on parameter estimation or model discrimination and not on approximation accuracy and bias.

Box and Lucas [45] were the first to address this problem. They considered settings in which the experimenter needs to fix the values of k input factors, $x_1, \ldots, x_k$, and the expected value of the response follows a nonlinear statistical model involving p parameters,

$$E\{Y(X)\} = f(x_1, \ldots x_k; \theta_1, \ldots, \theta_p). \tag{12}$$

They proposed finding designs that minimized the volume of an approximate confidence ellipsoid for the parameter vector θ, a goal that is equivalent to maximizing the determinant of the matrix $F^T F$, where F is the $n \times p$ matrix with $F_{i,j} = \partial f / \partial \theta_j$, evaluated at the ith design point and at the parameter vector. This is the well-known D-optimality criterion for experimental designs.

For nonlinear models, the partial derivatives in F, and hence the efficiency of any proposed design, depend on the unknown parameter vector. Box and Lucas note this dependence and advise choosing a design that maximizes the determinant when it is evaluated at θ^*, the experimenters' best guess of the parameter vector. A design that is tuned in this manner to a particular parameter vector is known as *locally D-optimal* [46]. Box and Lucas also note the likely benefits of a sequential approach, in which data from initial experiments are used to refine the information about θ and, in turn, to improve the sensitivity of the design.

Box also proposed methods that could be used when there was uncertainty as to the precise form of the mechanistic model. Box and Lucas [45] specifically suggested the idea of embedding the presumed model in a larger model, with additional parameters. A design for the larger model would then serve the dual purpose of good estimation for the original model (if it were supported by the data) and of checking whether the model was valid for use.

Box and Hill [47] studied the more general problem of finding a good design when there are a number of candidate models, say $f_1, \ldots, f_m$, each with its own set of parameters, linking the outcome(s) to the factors. Box and Hill exploited a Bayesian framework in which prior probabilities $\pi_1, \ldots, \pi_m$ are assigned to each model and are updated as new data are observed. They proposed two design criteria. The first of these was to choose the next design point(s) to maximize the expected change in the entropy of the model probabilities. The second was to maximize the expected change in the Kullback–Leibler divergence,

$$MD = \sum_{1 \leqslant i \neq j \leqslant n} \pi_i \pi_j I(p_i, p_j), \tag{13}$$

where all quantities are updated to reflect the data in hand, p_i is the predictive density for the new observation, conditional on f_i being the correct model and on the new design site, and $I(p_i, p_j) = \int p_i \ln(p_i/p_j)$ is the Kullback–Leibler divergence between the predictive densities from the ith and jth models. Box and Hill chose the latter as their working method due to its computational advantages. The MD criterion is appealing intuitively. In selecting design sites, it focuses on models that are currently considered most likely. In comparing those models, it looks for design sites at which the responses are expected to differ, as measured by the Kullback–Leibler divergence between their predictive densities.

14. Scientific impact

Simply put, George Box changed experimental research. His strategy of sequential learning and the tools that he developed to implement it have achieved widespread use in diverse fields. For example, a Google scholar search on Box–Behnken designs produces more than 12,000 hits, most of them articles by scientists and engineers who have put these designs to use for addressing their experimental needs. In some biotechnology journals, a majority of the articles present studies that feature response surface methods.

The remarkable impact of George Box's research on experimental design is intimately linked to his view of statistics as an integral part of science and of the central role of application in driving statistical research (see [6, 48]). His own research typically took root in real scientific problems, and it was always developed and presented in a way that facilitated use by others.

Box's ideas have also left an indelible stamp on statistical research in experimental design. Much further research has built on the concepts and methods that he developed. I have cited only a small fraction of that work in this article.

George Box was a giant figure in the world of experimental design. As an insightful observer, he saw the important questions that needed framing. His research stands out, not just for its breadth and its depth, but also for the clarity with which he presented his ideas. His many contributions to experimental design have become an integral part of the scientific repertoire and that is precisely the legacy to which George Box aspired.

15. Personal note

I had the good fortune to carry out my PhD research on experimental design under the supervision of George Box. Working with him was immensely rewarding, both professionally and personally. To his students, he was both an inspiration and a good friend.

Acknowledgements

I wish to thank Norman Draper and Stu Hunter for their careful reading and comments on an earlier draft of this paper.

References

1. Box GEP, Hunter JS, Hunter WG. *Statistics for Experimenters: Design, Innovation and Discovery*, Second Edition. Wiley-Interscience: New York, 2005.
2. Box GEP, Hunter JS, Hunter WG. *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building*, First Edition. Wiley-Interscience: New York, 1978.
3. Box GEP. *An Accidental Statistician: The Life and Memories of George E. P. Box*. John Wiley & Sons: New York, 2013.
4. Fisher RA. *The Design of Experiments*. Oliver and Boyd: Edinburgh and London, 1935.
5. Box GEP. Choice of response surface design and alphabetic optimality. *Utilitas Mathematica* 1982; **21B**:11–55.
6. Box GEP. Science and statistics. *Journal of the American Statistical Association* 1976; **71**:791–199.
7. Box GEP. Statistics as a catalyst to learning by scientific method, part II – a discussion. *Journal of Quality Technology* 1999; **31**:16–29.
8. Box GEP, Wilson KB. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B* 1951; **13**:1–45 (with discussion).
9. Box GEP. The exploration and exploitation of response surfaces: some general considerations and examples. *Biometrics* 1954; **10**:16–60.

10. Box GEP, Youle PV. The exploration and exploitation of response surfaces: an example of the link between the fitted surface and the basic mechanism of the system. *Biometrics* 1955; **11**:287–323.
11. Box GEP, Draper NR. *Empirical Model Building and Response Surfaces*. John Wiley & Sons: New York, 1987.
12. Myers RH, Montgomery DC, Anderson-Cook CM. *Response Surface Methodology: Product and Process Optimization Using Designed Experiments*, 3rd Edition. John Wiley & Sons: Hoboken NJ, 2009.
13. Macedo JA, Sette LD, Sato HH. Optimization of medium composition for transglutaminase production by a Brazilian soil *Streptomyces* sp. *Electronic Journal of Biotechnology* 2007; **10**(4). DOI: 10.2225/vol10-issue4-fulltext-10. Available from: http://www.ejbiotechnology.info/content/vol10/issue4/full/10/index.html [Accessed on 15 October 2007].
14. Steinberg DM, Bursztyn D. Response surface methodology in biotechnology. *Quality Engineering* 2010; **22**:78–87.
15. Box GEP. The invention of the composite design. *Quality Engineering* 1999; **12**:119–122.
16. Box GEP, Hunter JS. Multifactor experimental designs for exploring response surfaces. *Annals of Mathematical Statistics* 1957; **28**:195–241.
17. Giovannitti-Jensen A, Myers R. Graphical assessment of the prediction capability of response surface designs. *Technometrics* 1989; **31**:159–171.
18. Draper NR, Guttman I. An index of rotatability. *Technometrics* 1988; **30**:105–111.
19. Draper NR, Pukelsheim F. Another look at rotatability. *Technometrics* 1990; **32**:195–202.
20. Park SH, Kim HJ. A measure of slope-rotatability for second order response surface experimental designs. *Journal of Applied Statistics* 1992; **19**:391–404.
21. Park SH, Lim JH, Baba Y. A measure of rotatability for second order response surface designs. *Annals of the Institute of Statistical Mathematics* 1993; **45**:655–664.
22. Box GEP, Behnken DW. Simplex-sum designs: a class of second order rotatable designs derivable from those of first order. *Annals of Mathematical Statistics* 1960; **31**:838–864.
23. Box GEP, Behnken DW. Some new three level designs for the study of quantitative variables. *Technometrics* 1960; **2**:455–475.
24. Box GEP, Draper NR. A basis for the selection of a response surface design. *Journal of the American Statistical Association* 1959; **54**:622–654.
25. Box GEP, Draper NR. The choice of a second order rotatable design. *Biometrika* 1963; **50**:335–352.
26. Kiefer J. Optimum experimental designs. *Journal of the Royal Statistical Society, Series B* 1959; **21**:272–319.
27. Steinberg DM. Model robust response surface designs: scaling two-level factorials. *Biometrika* 1985; **72**:513–526.
28. Draper NR, Guttman I. Treating bias as variance for experimental design purposes. *Annals of the Institute of Statistical Mathematics* 1992; **44**:659–671.
29. DuMouchel W, Jones B. A simple Bayesian modification of d-optimal designs to reduce dependence on an assumed model. *Technometrics* 1994; **36**:37–47.
30. Voelkel JG. The efficiencies of fractional factorial designs. *Technometrics* 2005; **47**:488–494.
31. Bursztyn D, Steinberg DM. Comparison of designs for computer experiments. *Journal of Statistical Planning and Inference* 2006; **136**:1103–1119.
32. Box GEP, Draper NR. Robust designs. *Biometrika* 1975; **62**:347–352.
33. Yates F. *The Design and Analysis of Factorial Experiments*. Imperial Bureau of Soil Science: Harpenden, UK, 1937.
34. Finney DJ. The fractional replication of factorial arrangements. *Annals of Eugenics* 1945; **12**:291–301.
35. Plackett RL, Burman JP. The design of optimum multifactorial experiments. *Biometrika* 1946; **33**:305–325.
36. Box GEP, Hunter JS. The 2^{k-p} fractional factorial designs, Pt. I. *Technometrics* 1961; **3**:311–351.
37. Box GEP, Hunter JS. The 2^{k-p} fractional factorial designs, Pt. II. *Technometrics* 1961; **3**:449–458.
38. Box G, Tyssedal J. Projective properties of certain orthogonal arrays. *Biometrika* 1996; **83**:950–955.
39. Box GEP, Tyssedal J. Sixteen run designs of high projectivity for factor screening. *Communications in Statistics – Simulation and Computation* 2001; **30**:217–228.
40. Lin DKJ, Draper NR. Projection properties of Plackett and Burman designs. *Technometrics* 1992; **34**:423–428.
41. Wang JC, Wu CFJ. A hidden projection property of Plackett-Burman and related designs. *Statistica Sinica* 1995; **5**:235–250.
42. Cheng CS. Some hidden projection properties of orthogonal arrays with strength three. *Biometrika* 1998; **85**:491–495.
43. Box GEP. Evolutionary operation: a method for increasing industrial productivity. *Applied Statistics* 1957; **6**:81–101.
44. Box GEP, Draper NR. *Evolutionary Operation*. John Wiley & Sons: New York, 1969.
45. Box GEP, Lucas H. Design of experiments in non-linear situations. *Biometrika* 1959; **46**:77–90.
46. Chernoff H. Locally optimal designs for estimating parameters. *Annals of Mathematical Statistics* 1953; **24**:586–602.
47. Box GEP, Hill WJ. Discrimination among mechanistic models. *Technometrics* 1967; **9**:57–71.
48. Box GEP. The importance of practice in the development of statistics. *Technometrics* 1984; **26**:1–8.