CHAPTER 14

Tying It All Together

In statistics books in general, a method is presented and an illustrative example shortly follows. Simple exercises are given at the end of sections and the student/reader knows that they are to be worked using one of the methods just presented in the section.

The real world is not so accommodating. Specifically, an experimenter has a very large assortment of designs from which to choose and, in the absence of expert systems software, may have difficulty making a good choice. Frequently, the physical scenario is "bent" to make it conformable to published designs, just as an unscrupulous mechanic may try to bend a tailpipe assembly to make it fit a car it is not designed to fit. The bending of scenarios has been done often to allow the use of Taguchi designs, in particular.

Since the analysis of data from a designed experiment is relatively straightforward when the experiment is well designed, the focus in training for experimentation with statistical designs should be on gaining expertise in the selection of good designs. The end-of-chapter exercises provide an opportunity to move in that direction.

14.1 TRAINING FOR EXPERIMENTAL DESIGN USE

Before we look at some scenarios and try to decide how to proceed, it is worth noting that experience in experimentation can be gained "off-line" in various ways. As the late statistician Bill Hunter often said, one should "do statistics," and this certainly applies to engineering students, medical students and researchers, and students in other fields in which experimentation is routinely performed.

An excellent way to learn experimental design, especially some of the nuances that don't come through strongly in textbooks, is to conduct actual, simple experiments that do not require much in the way of materials or time. The primary motivation for students conducting such simple experiments was Hunter (1977), and there have been various other articles on the subject written since then, especially

Modern Experimental Design By Thomas P. Ryan

articles appearing in the *Journal of Statistics Education*, an online publication (see http://www.amstat.org/publications/jse/).

One example of a hands-on exercise that I have used in training industrial personnel in experimental design is commonly referred to as the catapult experiment, with a small catapult used to catapult a ping-pong ball, which is measured for distance traveled. Factors that can be varied include the rubber band type, position of the arm, type of ball used, and so on. Data are generated from the experiment, which would be analyzed using the methods given in various chapters such as Chapter 4 if each factor has two levels, as is desirable for the exercise. Of course an important part of the exercise is trying to figure out why certain effects (possibly including interactions) are significant. A good description of an actual experiment of this type is given in the paper "Training for design of experiments using a catapult" by J. Anthony (*Quality and Reliability Engineering International*, **18**, 29–35, 2002). Quoting from that article, "The engineers in the company felt that the experiments were useful in terms of formulating the problem, identifying the key control factors, determining the ranges of factor settings, selecting the experimental layout, assigning the control factors to the design matrix, conducting the experiment, and analysing and interpreting the results of the experiment."

There are variations of the catapult experiment that have been presented in the literature and one variation is to use it to try to determine the factor settings to hit a target value, as in multiresponse optimization that was presented in Chapter 12. An example of this is given in Section 5.4.7.2 of the *NIST/SEMATECH e-Handbook of Statistical Methods* (Croarkin and Tobias, 2002), which can be viewed at http://www.itl.nist.gov/div898/handbook/pri/section4/pri472.htm. In that example, a 2_{IV}^{5-1} design was used with four centerpoints and the objective was to determine the factor settings that should be used to cause the ping-pong ball to reach three different distances—30, 60, and 90 inches. Note that there is only a single response variable but a desirability function approach can still be used, as was used at the end of that section.

REFERENCES

Bjerke, F., A. H. Aastveit, W. W. Stroup, B. Kirkhus, and T. Naes (2004). Design and analysis of storing experiments: A case study. *Quality Engineering*, **16**(4), 591–611.

Box, G. E. P. and S. Jones (1992). Split-plot designs for robust product experimentation. *Journal of Applied Statistics*, **19**, 3–26. (This is available as Report No. 61, Center for Quality and Productivity Improvement, University of Wisconsin-Madison (see http://www.engr.wisc.edu/centers/cqpi.)

Croarkin, C. and P. Tobias, eds. (2002). *NIST/SEMATECH e-Handbook of Statistical Methods* (http://www.itl.nist.gov/div898/handbook), a joint effort of the National Institute of Standards and Technology and International SEMATECH.

Hunter, W. G. (1977). Some ideas about teaching design of experiments with 2^5 examples of experiments conducted by students. *The American Statistician*, **31**(1), 12–17.

EXERCISES

14.1 Assume that an experimenter has a strong need to conduct an experiment but knows that a factor that will affect the response variable cannot be maintained in a state of statistical control, with the latter interpreted to mean fixed parameter values. He has a few factors that he wants to study and will probably use two levels for each factor. What would you advise him to do? (*Note*: A possible design approach was covered in one of the chapters but was not emphasized.)

14.2 An experimenter wants to use a replicated 2^3 design to investigate the three factors of interest. He wants to be able to detect main effects of a certain minimum magnitude and can afford several replicates but suspects that the *AB* and *AC* interactions may be significant. He intends to consult tables to determine how many replicates he should use. What advice would you give him for determining the number of replicates to use when these interactions are expected to be significant. Is there any other advice that you would give him?

14.3 An experimenter is considering using a $2^2 \times 3^2$ design because there are three levels of interest for two of the factors. A 2^4 design may be used instead, however, because of a desire to use a simpler design, for which the analysis would also be simpler. A 2^4 design with two replicates and four centerpoints would have the same number of runs as the mixed factorial. Could the replicated 2^4 design be used as a suitable substitute for the mixed factorial? What, if anything, would be lost with this substitution? Explain.

14.4 The case study given by Bjerke, Aastveit, Stroup, Kirkhus, and Naes (2004, references) is a good example of how experimental settings can have various complications, which was discussed in Example 9.1. The authors described an experiment that was performed to investigate the effects of starch concentration on the quality of low-fat mayonnaise during different processing and storing conditions. A face-centered central composite design for three factors was used and there was a split-plot structure with repeated measures. The three factors had been selected in close cooperation with the manufacturer, as well as the levels of those factors.

There were 18 experimental runs and the response variable, which was (Brookfield) viscosity, was measured after 7, 14, and 38 weeks. It is worth noting, as the authors did, that these are not equally spaced time intervals.

At most six runs could be made per day, so the experiment was run over three days, with the axial points run on one day and a half fraction of the factorial points run on each of the other two days, with the four centerpoints scattered among the three days.

All the production samples were split into two halves for storing, with one half stored at room temperature (21°C), and the other half stored in a refrigerator at 4°C. This was labeled factor "D."

The manufacturer decided to use the measurements obtained from the second storage temperature because those most closely resembled the real-life storage of mayonnaise.

The following statement by the authors is worth noting:

> Throughout the production of the samples, some relevant process parameters were monitored to ensure stable process conditions. Observing these parameters during the production stage of the experiment did not reveal anything unusual. Unfortunately, these data were not recorded for further use (p. 599).

Of course such monitoring is important, as was stressed in Section 1.7, although it is undoubtedly not considered in most experiments.

One complication was that the Brookfield viscometer had an upper limit of 100. The viscometer cannot measure the viscosity if the samples are too thick. If the samples are too thick for the viscometer to perform properly, then the viscosity must be greater than 100. During the experiment, some of the mayonnaise samples became so thick that the spindle of the Brookfield meter could not rotate as required. Three of the data values listed in the authors' Table 2 were 100, including two of the three response values when all three factors were set at their highest level. Thus, part of the data was censored, which raises the question of whether the censored (i.e., incorrect) values should be used, should the true values be estimated, or should the censored values simply be treated as missing values.

Another problem is that since the three production factors were difficult to vary, the run order that was used each day "... was partly due to ease rather than randomization."

The authors considered three modeling approaches, which included a mixed model approach and a robustness approach. Address the following questions.

(a) How would you deal with the censored data problem, recognizing that using the recorded values could introduce spurious nonlinear effects?

(b) Are you concerned about the fact that there were hard-to-change factors (apparently all of them) and so randomization was not used? The authors analyzed the data, specifically their Table 5 analysis, as if complete randomization had been performed during the experiment. Would you have analyzed the data in some other way rather than constructing a standard ANOVA table and looking at p-values, which are not valid, strictly speaking, when there is restricted randomization, as was discussed in Section 4.19.

(c) The authors stated, "A more detailed analysis would include modeling of the time data," which they then proceeded to perform. Why would this be better than a repeated measures approach?

(d) Time was actually involved in two different ways as the experimental runs were made over three consecutive days. These would logically constitute three blocks and the blocking factor would be analyzed, but this was not performed in the authors' analyses. Can you think of a way of justifying this, or do you believe that a possible block effect should have been investigated? Explain.

(e) Read the article, if possible, and comment on approaches used with which you either agree and disagree.

14.5 Assume that you have been drawn into a debate about hierarchical versus nonhierarchical models, with one person arguing in favor of the former and another person arguing in favor of the latter. Assume initially that interest is focused only on effect estimates. What would you say to the hierarchical model proponent who forces main effect terms into models when the terms are not statistically significant, and what would you say to the person who favors nonhierarchical models? Now assume that the emphasis is on fitted values rather than effect estimates. Would you adopt a different position in regard to each of these two people? Explain. (In answering the fitted values question, you may wish to consider a configuration similar to Figure 4.2, except that there are two replicates and the replicated values at each design point differ only slightly.)

14.6 Assume that there are at most eight factors that are believed to be related to a response variable. You can make at most 100 runs in one or more experiments with an end objective of identifying the important factors and the levels of those factors that are necessary to hit a target value for the response, or at least come as close as possible to doing so. How will you proceed?

14.7 Would you advocate the use of supersaturated designs? What is one very important thing to guard against in selecting a supersaturated design?

14.8 Assume that you have limited resources and intend to conduct an experiment in a somewhat unusual way: by making runs until you have enough information to identify important effects. Would a one-factor-at-a-time (OFAT) design be a viable alternative to a more conventional design? Explain.

14.9 Assume that you have a scenario for which experimental runs are very inexpensive, as in many computer experiments. If a 2^4 design is to be used, why would it necessarily *not* be a good idea to use, say, 10 replicates?

14.10 What are the advantages of ANOM over ANOVA? Can you think of a design scenario for which an ANOM display could be constructed (such as for a 2^3 design) but it would be better to use ANOVA? More specifically, could ANOM and ANOVA give different results for a 2^3 design for any of the seven estimable effects? If so, give an example in which this occurs. If not, explain why it could not happen.

14.11 Consider a single factor with five levels and an analyst uses both ANOM and ANOVA in analyzing the data. The F-test in ANOVA has a p-value of .031

but all of the plotted points on the ANOM display are within the .05 decision lines (two of the points are barely inside the decision lines in one dataset that can be constructed). If you were handed these data, what general relationship would you expect between the five means? Guided by your answer to this question, construct an example with four observations for each of the five levels that results in the ANOVA p-value being less than .05 but all of the five means falling within the ANOM .05 decision limits.

14.12 Critique the experimenter's following statement: "I routinely use 2_{IV}^{k-p} designs because although I know that two-factor interactions are confounded, this generally doesn't bother me because I usually know which two-factor interactions are likely to be real, so it is simply a matter of selecting a particular design such that each two-factor interaction that is likely to be real is confounded with a two-factor interaction that probably isn't a real effect."

14.13 There was no discussion in any of the chapters about possibly computing conditional effects for a Plackett–Burman design. Could those effects be computed for that type of design? What would be the motivating factor for computing the conditional effects, or would there even be a motivating factor? Explain.

14.14 Assume that you have a need to run an experiment using a simple 2^2 design, but just before the experiment is to begin, a scientist who is part of the experimental team states that, although unlikely, an explosion could occur if the (1, 1) treatment combination were used, because of how extreme the high level was for each of the factors, pointing out that it would be much safer if the (1, 1) design point were replaced by something like (0.8, 0.7). What are you going to do? Do you have all the information that you need to make a decision? If yes, what is your decision? If not, what additional information do you need?

14.15 An experimenter is advised to consider a uniform design rather than selecting a central composite design. The experimenter responds by stating "No, I will never consider using a design that is not an orthogonal design." Respond to that statement.

14.16 Assume that you have encountered an experimenter who has studied several books on experimental design and is convinced that OFAT designs should never be used. The person needs to estimate interactions and each experimental run is quite expensive. You are serving as a consultant to the company and you are working with this person. What will be your recommendation?

14.17 Assume that you have used a factorial design with a mixture of fixed and random factors. Why is it important to label each one correctly when the data are analyzed with statistical software?

14.18 As discussed in Chapter 9, split-plot design configurations that result from hard-to-change factors have often been ignored in the analysis of data, with the data analyzed as if there were no restrictions on randomization. Consider the cake mix data given in Exercise 9.5 in Chapter 9.

(a) Analyze the data as a completely randomized experiment and compare with the analysis of the data as a split-plot experiment.

(b) There was no discussion of conditional effects in Chapter 9. However, if you look at the split-plot analysis given by Box and Jones (1992, references), you will see a two-factor interaction estimate that is 70% of one of the main effect estimates and almost eight times the other main effect estimate. Do you believe it would practical, or even possible, to perform a conditional effects analysis for that dataset? If so, perform such an analysis and comment. If it is not possible, explain why.